DOKTORANDSKÉ DNY 2010

sborník workshopu doktorandů FJFI
oboru Matematické inženýrství

19. a 26. listopadu 2010

P. Ambrož, Z. Masáková (editoři)

Doktorandské dny 2010
sborník workshopu doktorandů FJFI oboru Matematické inženýrství

P. Ambrož, Z. Masáková (editoři)
Kontakt petr.ambroz@fjfi.cvut.cz / 224 358 569

Vydalo České vysoké učení technické v Praze
Zpracovala Fakulta jaderná a fyzikálně inženýrská
Vytisklo Nakladatelství ČVUT-výroba, Zikova 4, Praha 6
Počet stran 272, Vydání 1.

ISBN 978-80-01-04644-9

Seznam příspěvků

Předmluva

Hlavní náplní workshopu Doktorandské dny je prezentace práce doktorandů oboru Matematické inženýrství doktorského studijního programu Aplikace přírodních věd, který zajišťují katedry matematiky, fyziky a softwarového inženýrství v ekonomii na FJFI ve spolupráci s podobně zaměřenými ústavy Akademie věd ČR. I proto je škála témat příspěvků v tomto sborníku tak široká. Sahá od výpočetních metod pro řešení problémů mechaniky kontinua přes statistickou analýzu dat, nestandardní reprezentace čísel či rovnice matematické fyziky až po rozpoznávání obrazu v biomedicínských aplikacích.

Letošní, již pátý ročník workshopu probíhá ve dnech 19. a 26. listopadu 2010 a koná se, jako již tradičně, v prostorách FJFI. K jeho úspěšnému konání přispívá podpora Katedry matematiky a Dopplerova ústavu pro matematickou fyziku a aplikovanou matematiku při FJFI. Za finanční zabezpečení děkujeme také Studentské grantové soutěži při ČVUT v rámci grantu SVK 15/10/F4.

Editoři

Kernel PCA in Alzheimer's Disease Diagnosis*

Jakub Adamec

2nd year of PGS, email: adamec.jakub@email.cz
Department of Software Engineering in Economy
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Jaromír Kukal, Department of Software Engineering in Economy,
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. The paper deals with the problem of automatic identification of pacients with Alzheimer's disease. The identification process is based on data mining method using the Kernel PCA algorithm. I determine the recommended value of parameter σ to the test data set of CT scans of the human brain, which is crucial for the proper functioning of the algorithm for that problem. Some results visualized using the programming environment MATLAB are presented at the end of the contribution.

Keywords: data mining, kernel PCA, Alzheimer's Disease, MATLAB

Abstrakt. Příspěvek se zabývá problémem strojové identifikace pacientů s Alzheimerovou chorobou. Proces identifikace je založen na data miningové metodě využívající algoritmu Kernel PCA. Na množině testovacích dat CT snímků mozku jsem určil doporučenou hodnotu parametru σ, který je klíčovým pro správné fungování algoritmu pro danou úlohu. Nakonec jsou prezentovány výsledky vizualizované pomocí programovacího prostředí MATLAB.

Klíčová slova: dolování dat, kernel PCA, Alzheimerova nemoc, MATLAB

1 Introduction

Data mining involves a wide range of methodologies for obtaining hidden and potentially useful information from data sets. It is difficult to give clear guidance on the process of data mining.

During the 90th years, data mining evolved into two general methodologies that at least can be roughly described by the steps of SEMMA methodology (Sample, Explore, Modify, Model, Assess) and the CRISP-DM (CRoss-Industry Standard Process for Data Mining). The common methodology is the essence of all sequence of several steps:

1. Practical (Business) - the role of formulation and understanding of the problem. Even the automatic search of knowledge can not be done completely blind.

2. Data - search for and preparation of data for analysis. Statistical algorithms usually require data ready in some form, and therefore can not be applied directly to raw data from databases.

3. Analytical - searching for information in the data and producing statistical models. These use a variety of methods from simple tabulation and visualization to

*This work has been supported by the grant SGS 10/092/OHK4/1T/14.

sophisticated approaches such as genetic programming. The most commonly used methods, however, the logistic regression with automatic variable selection, decision trees and neural networks. The output of this phase would be general knowledge and mathematical models.

4. Application - findings and models can be put into practice.

5. Control - the need for feedback and to check whether the model is not too aged and retains its effectiveness.

2 Theoretical Background

2.1 Principal Component Analysis

Principal **C**omponent **A**nalysis involves a mathematical procedure that transforms a number of possibly correlated variables into a smaller number of uncorrelated variables called principal components. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible. Depending on the field of application, it is also named the discrete Karhunen–Loeve transform (KLT), the Hotelling transform or proper orthogonal decomposition (POD).

PCA was invented in 1901 by Karl Pearson. Now it is mostly used as a tool in exploratory data analysis and for making predictive models. PCA involves the calculation of the eigenvalue decomposition of a data covariance matrix or singular value decomposition of a data matrix, usually after mean centering the data for each attribute. The results of a PCA are usually discussed in terms of component scores and loadings.

PCA is the simplest of the true eigenvector-based multivariate analyses. Often, its operation can be thought of as revealing the internal structure of the data in a way which best explains the variance in the data. If a multivariate dataset is visualised as a set of coordinates in a high-dimensional data space (1 axis per variable), PCA supplies the user with a lower-dimensional picture, a "shadow" of this object when viewed from its (in some sense) most informative viewpoint.

PCA is closely related to factor analysis; indeed, some statistical packages deliberately conflate the two techniques. True factor analysis makes different assumptions about the underlying structure and solves eigenvectors of a slightly different matrix.

2.2 Kernel Principal Component Analysis

Kernel Principal **C**omponent **A**nalysis is an extension of principal component analysis (PCA) using techniques of kernel methods. Using a kernel, the originally linear operations of PCA are done in a reproducing kernel Hilbert space with a non-linear mapping.

Principal component analysis (PCA) projects high - dimensional data onto a lower - dimensional subspace by seeking a linear combination of a set of projection vectors that can best describe the variance of data in a sum of squared - error sense. Kernel PCA extends the capability of linear PCA by capturing nonlinear structure in the data, since

a linear PCA performance in the feature space corresponds to a nonlinear projection in the original data space.

For a set of data points $\mathbf{x}_j \in \Re^d, j = 1, ..., N$, we map them into an arbitrary high - dimensional feature space with the nonlinear function $\mathbf{\Phi} : \Re^d \rightarrow F$. The transformed data are centered, i.e., the mean is 0. This can be achieved by using the substitute kernel matrix,

$$\mathbf{k} = \mathbf{k} - \mathbf{1}_N\mathbf{k} - \mathbf{k}\mathbf{1}_N + \mathbf{1}_N\mathbf{k}\mathbf{1}_N \tag{2.1}$$

where $\mathbf{k} = \{k(\mathbf{x}_i, \mathbf{x}_j)\}$ is the kernel matrix, $k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2})$ and $(\mathbf{1}_N)_{i,j} = 1/N, \mathbf{1}_N \in \Re^{N\times N}$.

Similar to linear PCA, the principal components are obtained by calculating the eigenvectors $\mathbf{e}$ and eigenvalues $\lambda > 0$ of the covariance matrix

$$\sum{}^{\mathbf{\Phi}} = \frac{1}{N}\sum_{j=1}^{N} \mathbf{\Phi}(\mathbf{x}_j)\mathbf{\Phi}(\mathbf{x}_j)^T,$$

$$\lambda\mathbf{e} = \sum{}^{\mathbf{\Phi}}\mathbf{e}. \tag{2.2}$$

By multiplying with $\mathbf{\Phi}(\mathbf{x}_i)$ from the left and noticing that $\mathbf{e} = \sum_{l=1}^{N} \alpha_l \mathbf{\Phi}(\mathbf{x}_l)$, straightforward manipulation of Eq. 2.2 yields

$$\lambda\sum_{l=1}^{N} \alpha_l(\mathbf{\Phi}(\mathbf{x}_l) \cdot \mathbf{\Phi}(\mathbf{x}_l)) = \frac{1}{N}\sum_{l=1}^{N} \alpha_l \left(\mathbf{\Phi}(\mathbf{x}_l) \cdot \sum_{j=1}^{N} \mathbf{\Phi}(\mathbf{x}_j)\right)(\mathbf{\Phi}(\mathbf{x}_j) \cdot \mathbf{\Phi}(\mathbf{x}_l)). \tag{2.3}$$

for all $l = 1, \ldots, N$

Using the kernel function, Eq. 2.3 can be written as

$$\lambda\boldsymbol{\alpha} = \mathbf{k}\boldsymbol{\alpha}, \tag{2.4}$$

where $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_N)^T$. The achieved solutions must be normalized following the condition,

$$\lambda_i(\boldsymbol{\alpha}_i \cdot \boldsymbol{\alpha}_i) = 1. \tag{2.5}$$

Given a new data point, its projection can be calculated as

$$(\boldsymbol{\alpha}_i \cdot \mathbf{\Phi}(\mathbf{x})) = \sum_{l=1}^{N} \alpha_{il} k(\mathbf{x_l}, \mathbf{x}). \tag{2.6}$$

3 Implementation in MATLAB

Method using kernel PCA algorithm is programmed in MATLAB programming environment that allows easy entering and processing of matrix computations and their subsequent visualization. Most interesting are the following two functions.

The first function prepares the matrix K - there is an important dependence on the parameter σ, which affects the quality of the result set in next function.

```
1  function [crit,Y,W,eff]=KERNELPCAGAUSS(X,iseuclid,sigma,beta,d,p)
2  if iseuclid
3      m=size(X,1);
4      K=zeros(m);
5      for i=1:m-1
6          for j=i+1:m
7              K(i,j)=norm(X(i,:)-X(j,:));
8              K(j,i)=K(i,j);
9          end
10     end
11 else
12     K=X;
13 end
14 if sigma>0
15     if beta>0
16         K=1./(1+0.5/beta*(K/sigma).^2).^beta;
17     else
18         K=exp(-0.5*(K/sigma).^2);
19     end
20 else
21     K=-K.^2;
22 end
```

The second function performs the actual calculation of the functional value of PCA. The parameter σ is used in matrix K - see equation 2.1 vs. line 5.

```
1  function [crit,Y,W,eff]=KERNELPCANALYZER(K,d,p)
2  eps=1e-100;
3  m=length(K);
4  ONEM=ones(m)/m;
5  K=K-ONEM*K-K*ONEM+ONEM*K*ONEM;
6  [E,LAMBDA]=eig(K);
7  E=real(E);LAMBDA=real(LAMBDA);
8  lambda=diag(LAMBDA);aaa=trace(K);
9  U=[lambda E LAMBDA];
10 U=sortrows(U,1);
11 lambda=U(:,1);
12 E=U(:,2:m+1);
13 LAMBDA=U(:,m+2:end);

14 eff=cumsum(lambda(end:-1:end-d+1)/(aaa+eps));
15 W=[];
16 for k=1:d
17     W=[W E(:,end-k+1)/sqrt(LAMBDA(end-k+1,end-k+1)+eps)];
18 end
19 Y=K*W;
```

```
20 Y=Y*diag(1./(std(Y)+eps));
21 dcrit=chi2inv(1-p,d);
22 crit=sum(Y.^2,2)>=dcrit;
```

4 Results

In experiments I ivestigated that with increasing parameter σ incereases the number of correctly classified patients. This growth is reflected in the figures below - initially in 2D then in 3D. It is obvious that increasing sigma value from some value do not increase the function value of PCA. Growth performance values stops for parameter sigma more than 5000. For comparison, the last figure shows the result with the $\sigma = 0$.

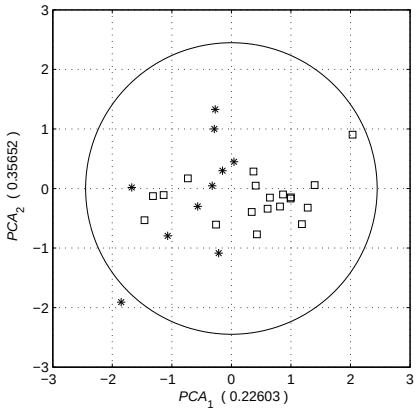


Figure 1: 2D view, $\sigma = 100$

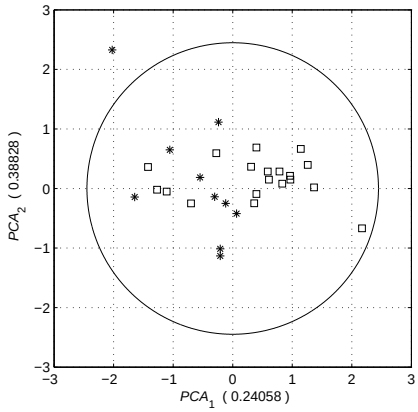


Figure 2: 2D view, $\sigma = 200$

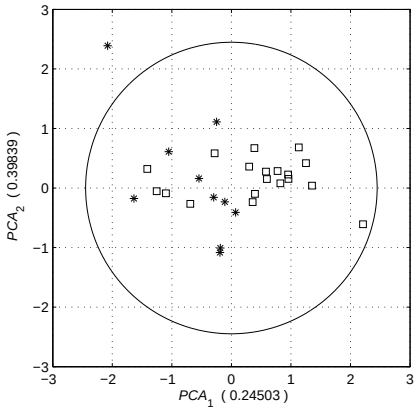


Figure 3: 2D view, $\sigma = 500$

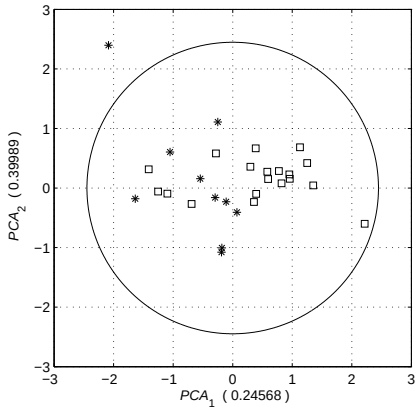


Figure 4: 2D view, $\sigma = 1000$

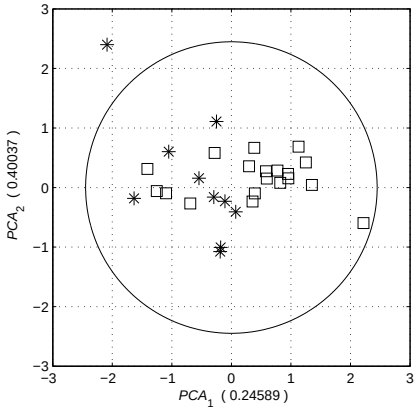


Figure 5: 2D view, $\sigma = 5000$

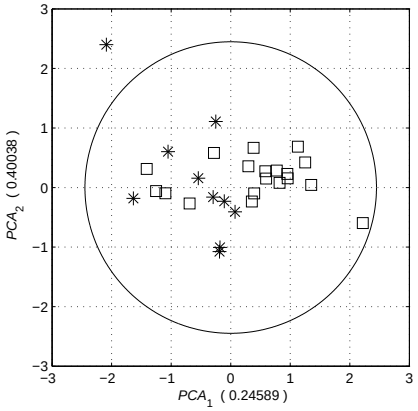


Figure 6: 2D view, $\sigma = 10^4$

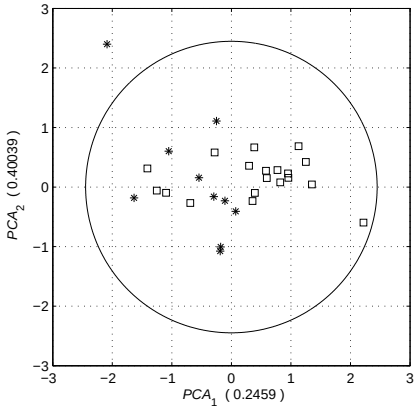


Figure 7: 2D view, $\sigma = 10^6$

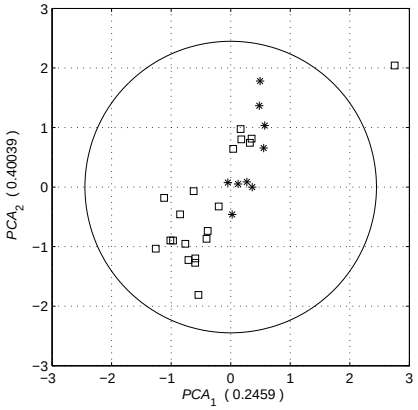


Figure 8: 2D view, $\sigma = 0$

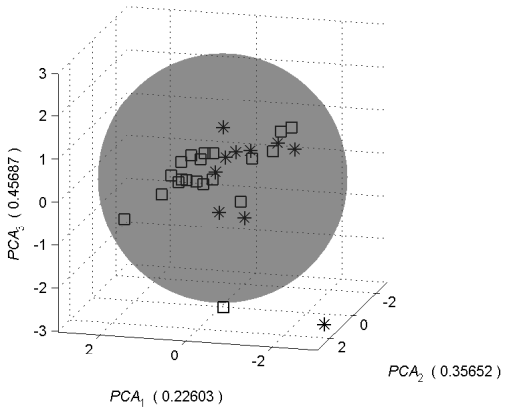


Figure 9: 3D view, $\sigma = 100$

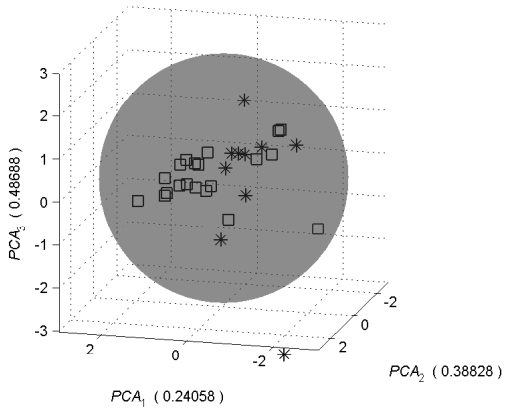


Figure 10: 3D view, $\sigma = 200$

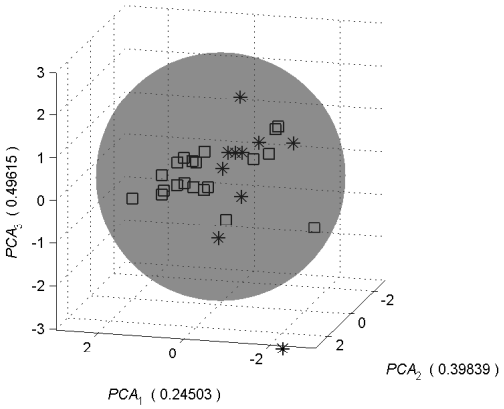


Figure 11: 3D view, $\sigma = 500$

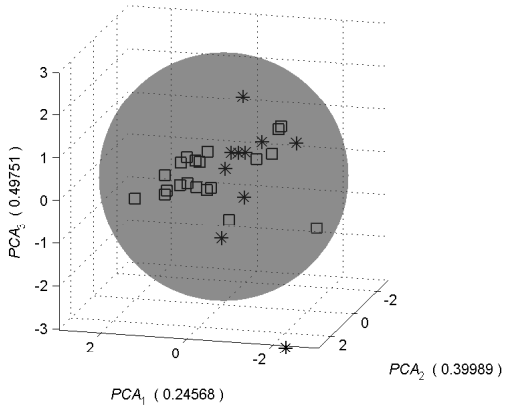


Figure 12: 3D view, $\sigma = 1000$

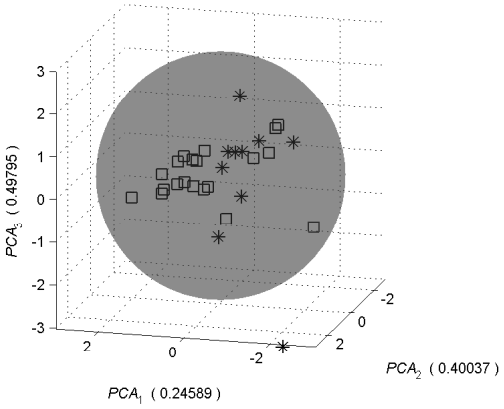


Figure 13: 3D view, $\sigma = 5000$

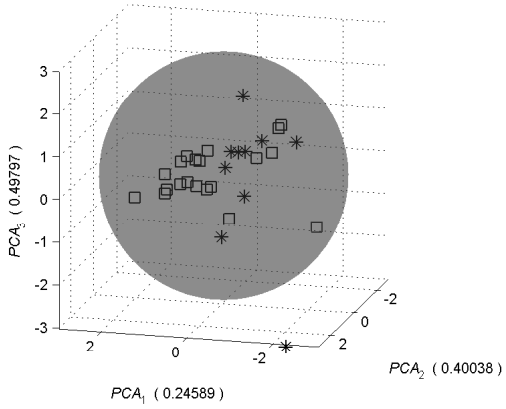


Figure 14: 3D view, $\sigma = 10^4$

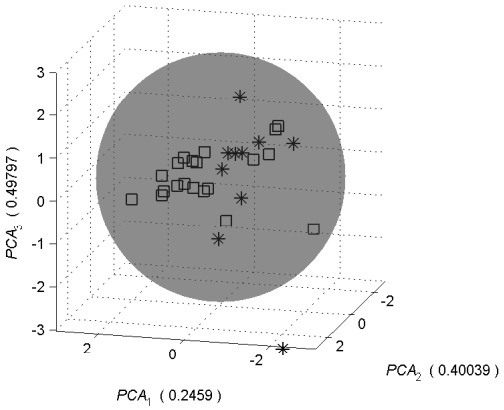


Figure 15: 3D view, $\sigma = 10^6$

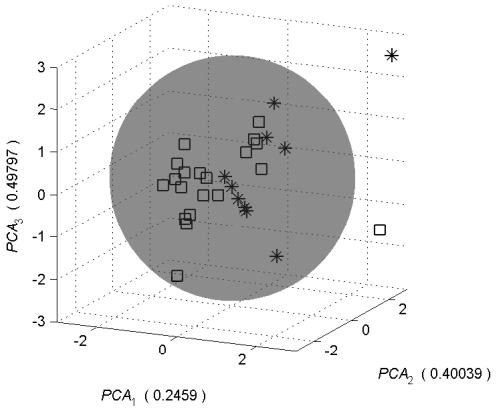


Figure 16: 3D view, $\sigma = 0$

5 Conclusion

The methodology allows to visually study the separation of patients with ALD from normal patients. The system can be set so that the first three components carry more than 75% of information about all patients. The first three components do not allow linear separation, but can be set so that the error rate is $10 - 25\%$, which corresponds to common clinical practice in the diagnosis. I assume, that in the future I will be using linear classifiers working with more dimensions. I also assume that the application of cluster analysis will allow me to distinguish several types of normality, respectively several types of dimension.

References

[1] N. Cristianini, J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press, Cambridge, (2000).

[2] K. Nováková. *Využití transformací pri rozpoznávání objektu*. Disertační práce, FJFI, ČVUT, Praha, (2008).

[3] B. Scholkopf, A. J. Smola. *Learning with Kernels - Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press, Cambridge, Massachusetts, London, England, (2002).

[4] R. Xu, D. Wunsch. *Clustering (IEEE Press Series on Computational Intelligence)*. Wiley-IEEE Press, (2008).

Modifications of Hausdorff Distance as Dissimilarity Measurement Tools

Kamil Barbierik

2nd year of PGS, email: kamil.barbierik@fjfi.cvut.cz
Department of Software Engineering in Economy
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Jaromír Kukal, Department of Software Engineering in Economy,
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. Object registration is a common problem in many fields. Several methods for registration have been developed to efficiently register two sets on each other. One approach is to use the Hausdorff distance to measure the degree of dissimilarity of two sets and minimize it in registration process. To increase the robustness of Hausdorff distance to noise and outliers, several modification of this measure were introduced. This paper deals with three modifications: Partial Hausdorff distance, Modified Hausdorff distance, and local Windowed Hausdorff distance, and discusses their properties and usability.

Keywords: Set distance, Hausdorff distance, Hausdorff distance modifications, partial HD, modified HD, windowed HD, local dissimilarity map

Abstrakt. S automatizovaným rozpoznávaním objektov sa už v dnešnej dobe môžeme stretnúť v najrôznejších odvetviach priemyslu, či vedy. Za účelom čo najpresnejšie identifikovať objekt, či odlišnosti určitého objektu od daného vzoru, boli vyvinuté mnohé metódy, ktoré tento problém viac či menej efektívne riešia. Jedným z prístupov k riešeniu tohto problému je použitie metód založených na Hausdorffovej vzdialenosti (HD). Vzhľadom k extrémnej citlivosti HD na šum, je vhodné uvažovať o modifikáciách tejto vzdialenosti, ktoré sú voči šumu robustnejšie a stabilnejšie. Tento článok predstavuje 3 takéto modifikácie, konkrétne sa jedná o Čiastočnú HD, Modifikovanú HD a Okienkovú HD. Sú diskutované ich vlastnosti a výhody. Nakoniec sa článok zameriava najmä na Okienkovú HD a lokálne vzdialenostné mapy, ktoré sú výsledkom merania vzdialenosti objektov práve okienkovou HD a to v 2D a 3D.

Kľúčové slová: vzdialenosť množín, Hausdorffova vzdialenosť, Modifikácie Hausdorffovej vzdialenosti, čiastočná HD, modifikovaná HD, okienková HD, lokálna vzdialenostná mapa

1 Introduction

Now days, the image recognition and registration is a common problem in many fields from flow production where the quality of products may be controlled by a system based on shape recognition, through satellite photographs enhancement, up to medical image processing. Especially in medicine, the outputs of various diagnostic tools are 3D images where registration of such data sets is very common problem. The term registration refers here to a transformation that maps the points of one coordinate system onto corresponding points in another coordinate system. The purpose is to fit pictures of the same thing

captured by various technologies on each other. Thus, the resulting "multi" picture provides more information and helps doctors with the diagnosis of potential disease. Other application may be comparing probably sick parts of body against healthy etalons. Again, registration of two images is necessary to help doctors to detect changes and interpret them. The registration is always necessary, because it is practically impossible to place an examined patient in the same position every time and in every diagnostic tool. What is more, the human body is not a rigid structure, but is subject to slight deformations caused by heart beat, breathing, or other slight movements.

Several methods for object matching and object registration have been developed. They enable to determine the similarity between compared objects, and in the process of registration minimize this dissimilarity. In this paper we introduce *Hausdorff distance (HD)* as a good tool for measuring degree of dissimilarity between two sets. Further, we discuss some modifications of Hausdorff distance that improve its properties and performance when considering real world noisy images. Although these alternations may violate some rules of well-mannered distance measure (a metric), they generally yield better results over degraded images where standard HD is unsuitable.

1.1 Distance measure

Let M be a set of points. In the following discussion we assume an Euclidean metric space $\{M, \varrho\}$ where $\vec{x}, \vec{y} \in M$ and metric ϱ is defined as follows:

$$\varrho(\vec{x}, \vec{y}) = \sqrt{\sum_{k=1}^{r} (x_k - y_k)^2} \tag{1}$$

Equation (1) is a mathematical tool for expressing the distance between two elements of metric space. In the following sections we present methods for determining the distance between two sets of elements. We mention trivial set distance measures, but due to their incapability to measure degree of mismatch of two sets, we will not discuss them in detail. We rather focus on Hausdorff distance and its modifications that have, on the contrary, very good ability to measure dissimilarity of two sets, i.e. distance between two sets.

Well-mannered measure of distance satisfies axioms of metric: explicitly *identity*, *symmetry*, and *triangle inequality*. It can be shown that only the Hausdorff distance in its basic not modified definition satisfies all axioms under certain general conditions (refer to [1], [2]). To be exact, Hausdorff distance satisfies metric axioms over the set of all closed and bounded sets. Over such sets the Hausdorff distance is a metric, and therefore, has very favorable mathematical properties as a distance measure.

From among closed and bounded sets we restrict ourselves to finite point sets. In the following discussion we will consider only this subset, because finite point sets are very common output of available technologies for image capturing.

1.2 Trivial set distance measures

Trivial measures between two sets are *nearest points distance*, *farthest points distance*, or *center of gravity distance*. These measures of set distance are not metrics, because they

violate some of metric axioms, and also, they have a very limited discriminatory power, so they are unusable in determining the degree of dissimilarity of two sets. However, they may have some useful properties in some applications. They are very easy to implement and algorithms using these measures take relatively small computation times.

2 Hausdorff distance

Hausdorff distance is a max-min distance defined by the following definition.

Definition: Let $\{M, \varrho\}$ be a metric space where M is a finite set of points, and metric ϱ is defined by equatoin (1). Let $A = \{\vec{a_1}, \ldots, \vec{a_p}\}$ and $B = \left\{\vec{b_1}, \ldots, \vec{b_p}\right\}$ be two subsets of M. We define Hausdorff distance $H(A, B)$ by:

$$H(A,B) := \max\left\{\max_{\vec{a}\in A}\min_{\vec{b}\in B}\varrho\left(\vec{a},\vec{b}\right), \max_{\vec{b}\in B}\min_{\vec{a}\in A}\varrho\left(\vec{a},\vec{b}\right)\right\} \tag{2}$$

Note: The definition of Hausdorff distance can be derived by a series of steps naturally extending the distance function ϱ in the underlying metric space $\{M, \varrho\}$ as follows: Let $\{M, \varrho\}$ be a metric space. Given $\vec{a} \in M$ and non-empty set $B \subset M$ we define a distance $dist\,(\vec{a}, B)$ between point $\vec{a}$ and the set B by:

$$dist\,(\vec{a}, B) := \min_{\vec{b}\in B}\varrho\left(\vec{a},\vec{b}\right) \tag{3}$$

Using this distance we define $h\,(A, B)$, the distance between A and B where $A, B \subset M$:

$$h\,(A,B) := \max_{\vec{a}\in A} dist\,(\vec{a}, B) \tag{4}$$

$h\,(A, B)$ is called the *directed Hausdorff distance*. If A and B are compact sets, then $h\,(A, B)$ will be finite. Triangular inequality property of $h\,(A, B)$ is inherited from metric ϱ. Directed Hausdorff distance is not a metric yet because although fact that $A = B$ implies $h\,(A, B) = 0$, $h\,(A, B) = 0$ does not imply that $A = B$. $h\,(A, B) = 0$ only implies that $A \subseteq B$. What is more, directed Hausdorff distance does not obey even the symmetry property of metric. To be precise, $h\,(A, B)$ is not always equal to $h\,(B, A)$. However, we can create a metric using the directed Hausdorff distance by defining undirected distance called Hausdorff distance as follows:

$$H\,(A,B) := \max\left\{h\,(A,B), h\,(B,A)\right\} \tag{5}$$

Hausdorff distance is a very powerful tool for measuring dissimilarity between two sets. The set can represent some graphics in 2D or various 3D pictures in medicine, for instance captured by technologies like MRI or CT. Using Hausdorff distance, we can measure a degree of mismatch between two object shapes very precisely. Unlike feature based methods, Hausdorff distance is zero if and only if the shapes of objects are exactly the same and increases with growing dissimilarity. What is more, if we need to minimize the Hausdorff distance over the space of some transformation parameters, any

transformation of object (rotation, affine transform...) can be taken into consideration. An advantage is also the possibility of independently using the undirected distances that Hausdorff distance is composed of.

On the other hand, the major disadvantage is computation burden of discussed measure. $H(A,B)$, where A,B are two sets of size p,q can be trivially computed in time $O(pq)$. In the last decades several methods have been proposed that allow speeding up the computation. By approximation of objects by polygons, for instance, the computation time can be improved to $O((p+q)\log(p+q))$ [3].

Besides computation troubles, there is another fact that is necessary to take into consideration. As was mentioned above, Hausdorff distance measures difference between two sets very precisely. Consequently, it is extremely sensitive to outliers. Therefore, if we want to obtain satisfactory results, measured sets have to be without any random disturbance or noise, what is difficult to achieve in the real world. Hausdorff distance as defined by eq. (2) or (5) is not capable to distinguish between what are data of interest and what is a noise. It simply processes noise as it was a part of processed set. Without any additional information about noise or the set itself and without any modification to the measure, the resulting distance between sets can differ from an intuitive notion.

Imagine a situation as figure 1 shows, where we have a noise-free etalon (model set), and we want another set to compare to it. The other set is identical with the etalon but has a random dot far away from it in a distance d. The result of comparison of such sets using standard Hausdorff distance will be highly disappointing. Instead of zero, or at least distance very close to zero, we will receive distance d.

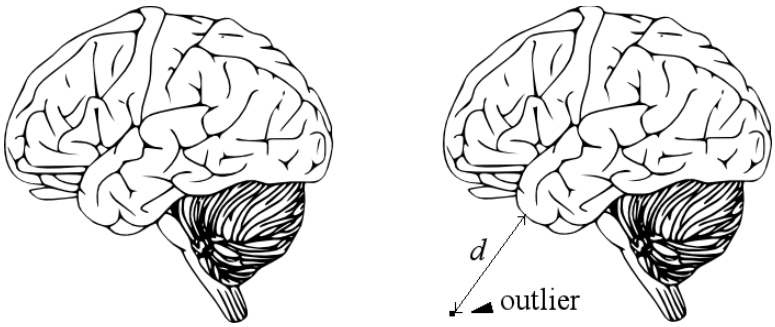


Figure 1: Two sets with Hausdorff distance d

In the real world, data sets received from various sensors are always disturbed by noise which is random in most cases. As we have just shown, the Hausdorff distance in its standard form is generally not suitable as dissimilarity measure of such noisy sets. Therefore, for the real world shape matching it is necessary to modify the Hausdorff distance in a way to make it more robust and stable to noise and outliers.

Some modified Hausdorff distances have already been proposed to achieve the goal. However, modifications of Hausdorff distance can cause violation of certain axioms of metric. Consequently, the modified Hausdorff distance may no longer be a metric. In support of more robust measure this is not a big trouble while the new modified distance corresponds with an intuitive notion of shape resemblance. In the next sections we will describe these modifications and discuss their properties.

3 Modifications of Hausdorff Distance

3.1 Partial Hausdorff distance

Partial Hausdorff distance was proposed by Huttenlocher et al. in [4]. He extended the definition of Hausdorff distance to enable comparison of objects that are partially hidden from the view. It has been shown by many experiments that the partial Hausdorff distance yields good results also for matching binary sets disturbed by impulse noise.

Consider the partial Hausdorff distance $h(A,B)$ from set A to B where p,q are numbers of elements of the sets and $dist(\vec{a},B)$ is defined by eq. (3). The computation of such distance rank all points from set A by a distance to the nearest point of set B. Partial Hausdorff distance as defined by eq. (4) determines the distance by the largest ranked element of set A. When instead of taking the largest ranked element, we take the K^{th} ranked element of A, where $1 \leq K \leq p$, to determine the distance from A to B, we receive the definition of *partial directed Hausdorff distance* as Huttenlocher *et al.* proposed it:

$$h_K(A,B) := K^{th}_{\vec{a}\in A} dist(\vec{a},B) \tag{6}$$

That is, for each element of A, the distance to the nearest element of B is computed and then the elements of A are ranked according to the respective values of this distance. The K^{th} ranked distance d expresses that K of the elements of set A are each within a distance d of some point of B. When we put $K = p$, we get the standard directed Hausdorff distance $h(A,B)$.

The K value is an input parameter for the computation. The range of possible values is dependent upon the number of set elements. To generalize this input argument it is better to specify some fraction x, where $0 \leq x \leq 1$. Subsequently, K can be computed by $K = \lfloor xp \rfloor$. The notation of directed distance using generalized input argument in a form of fraction or percentage can look like:

$$h_x(A,B) := {}^{x}K^{th}_{\vec{a}\in A} dist(\vec{a},B) \tag{7}$$

Directed partial Hausdorff distance as defined by eq.(6) or (7) has a nice property of automatically selecting the best matching points of A. Thus, it is not required to specify which part of the set A is to be compared to the set B. Computation of $h(A,B)$ determines the distances from elements of A to the nearest points from B and after ranking, the K nearest elements are taken under consideration.

Undirected partial Hausdorff distance is naturally defined as:

$$H_{K,L}(A,B) := \max\{h_K(A,B), h_L(B,A)\} \tag{8}$$

3.2 Modified Hausdorff distance

Modified Hausdorff distance (MHD) was proposed by Dubuisson and Jain in [5]. They compared 24 different undirected distance measures based on Hausdorff distance and tested them for matching two objects based on their edge points. Tested measures were created by combining 6 different directed distances $h_i(A,B)$ with 4 functions $f_j(A,B)$ providing undirected resulting distance. Dubuisson and Jain conclude that the proposed MHD yields the best results in object matching among all tested measures. Moreover, it has the following desirable properties:

- The value of measure increases as the degree of mismatch is growing.
- It is robust to outliers that might result from segmentation errors.

MHD proposed in their paper uses the following directed distance

$$h(A,B) := \frac{1}{p}\sum_{\vec{a}\in A} dist(\vec{a},B) \tag{9}$$

and the function that combines two directed distances in order to receive undirected distance measure

$$f(A,B) := \max\{h(A,B), h(B,A)\} \tag{10}$$

3.3 Windowed HD

In January 2007 in a preprint and later published in [6] a new approach was proposed for determining dissimilarity between two sets using Hausdorff-like distances. While previous and other modifications of Hausdorff distances were global and except the classical Hausdorff distance and the MHD required some input arguments, the windowed Hausdorff distance operates locally and does not require any input parameters. Furthermore, while the global ones produce only one number that expresses the dissimilarity between two sets, the windowed Hausdorff distance produce a dissimilarity map where local mismatches can be examined. The main idea of windowed Hausdorff distance is to define some window, moving it over two sets, counting the Hausdorff distance within the window and record the results into dissimilarity map. Thus, the windowed Hausdorff distance could be naively defined by applying the global Hausdorff distance within a window:

$$H_w(A,B) := \max\{h_w(A,B), h_w(B,A)\} \tag{11}$$

where

$$h_w(A,B) := \max_{\vec{a}\in A\cap W}\left(\min_{\vec{b}\in B\cap W} \varrho\left(\vec{a},\vec{b}\right)\right) \tag{12}$$

This definition is naive, because it does not take into account the possibility that there could be no elements or elements of only one of two sets in the window W. What is the distance in those cases? Furthermore, it is necessary to assure that the distance be coherent when the window is moved or resized, which is not the case in this naive definition. To eliminate these negative facts, the naive windowed Hausdorff distance definition should be modified so that it respects the following principles:

- The distance value should not decrease if the window size is enlarged.
- The distance values obtained in different cases (representative of one set in the window, representatives of two sets in the window...) shall be consistent so as to have smooth transition when the window is modified

The improved definition, which takes into account the previous principles, involves three different directed Hausdorff distances, which supplies three possible cases of presence

of set points in the window. It makes use of the distance to the frontier $Fr(W)$ of the window W. In this discrete case we consider that the frontier $Fr(W)$ is between the elements. For example the frontier of the ball $B(x,n)$ is the line between $B(x,n)$ and $B(x,n+1) \setminus B(x,n)$. The distance of a point $x \in B(x,n)$ to the frontier is equal to the distance to the elements just behind the frontier.

Definition: Let A, B be two bounded sets of $\mathbb{R}^r$. $H_w(A,B) = \max\{h_w(A,B), h_w(B,A)\}$ where:

- If $A \cap W \neq \emptyset \wedge B \cap W \neq \emptyset$

$$h_w(A,B) := \max_{\vec{a} \in A \cap W} \left[\min_{\vec{b} \in B \cap W} \varrho\left(\vec{a}, \vec{b}\right), \min_{\vec{w} \in Fr(W)} \varrho(\vec{a}, \vec{w}) \right] \tag{13}$$

- If $A \cap W \neq \emptyset \wedge B \cap W = \emptyset$

$$h_w(A,B) := \max_{\vec{a} \in A \cap W} \left[\min_{\vec{w} \in Fr(W)} \varrho(\vec{a}, \vec{w}) \right] \tag{14}$$

- If $A \cap W = \emptyset$

$$h_w(A,B) := 0 \tag{15}$$

Note that according to this new definition it is possible to measure the windowed Hausdorff distance even if the window contains no element or elements of only one set. In case there is no point of A nor of B in W, both directed distances $h_w(A,B)$ and $h_w(B,A)$ are equal to zero, and therefore, the global distance $H_w(A,B)$ is zero too. In case there is exactly one set without point in W, one of two directed distances is equal to zero and the expression of the other one takes into account the distance to the border of W.

3.3.1 Properties of windowed Hausdorff distance

- *Symmetry* and *non-negativity* (by definition)
- *Identity*

  Let A, B be a bounded sets of points of $\mathbb{R}^r$. Let W be a convex closed subset of $\mathbb{R}^r$. $H_w(A,B) = 0 \Leftrightarrow A \cap W = B \cap W$
- *Boundary*

  Let $\vec{x} \in \mathbb{R}^r$ and $r > 0$. Let define $W = B(\vec{x}, r)$ then $H_w(A,B) \leq H(A,B)$.
- *Growth*

  Let $V = B(\vec{x_v}, r_v)$ and $W = B(\vec{x_w}, r_w)$ be two closed balls such as $V \subset W$. Then $H_v(A,B) \leq H_w(A,B)$

These properties ensure that the value measured in the window does not decrease when the window is enlarged. As the window W slides all over two sets, the values in the produced dissimilarity map will remain between 0 and $H(A,B)$. The properties of boundary and growth give a frame to an optimum window-size-criterion definition. In [6] it is shown that the window $W = B(\vec{x}, r)$ gives a local measure when the Hausdorff distance in this window is maximum i.e. $H_{B(\vec{x},r)}(A,B) = r > 0$. This maximum value is reached only if exactly one element of A is in the center of the window and no B elements are within it. The maximum local measure for fixed $\vec{x}$ is the biggest possible $r > 0$ where $H_{B(\vec{x},r)}(A,B) = r$. The maximum local measure is the distance from the central point of window for instance $\vec{a} \in A$ to the nearest point of the other set: $r_m = dist(\vec{a}, B)$. The maximum local measure centered in general point $\vec{x}$ could be computed and recorded to the local distance map by the following technique.

3.3.2 Local distance map

Definition: Let A and B be two non-empty finite sets of points of $\mathbb{R}^r$ and let $\vec{x} \in \mathbb{R}^r$, the local distance map $LDMap(\vec{x})$ is defined by:

$$LDMap(\vec{x}) = |\mathbb{I}_A(\vec{x}) - \mathbb{I}_B(\vec{x})| \max\{dist(\vec{x}, A), dist(\vec{x}, B)\} \tag{16}$$

where $\mathbb{I}_A(\vec{x})$ is equal to 1 if $\vec{x} \in A$ and 0 otherwise.

The maximum value in the LDMap is the Hausdorff distance $H(A,B)$. This value is present in the map at least once. Figure 2 shows 2D local distance map of two brains slices cut at the same level.

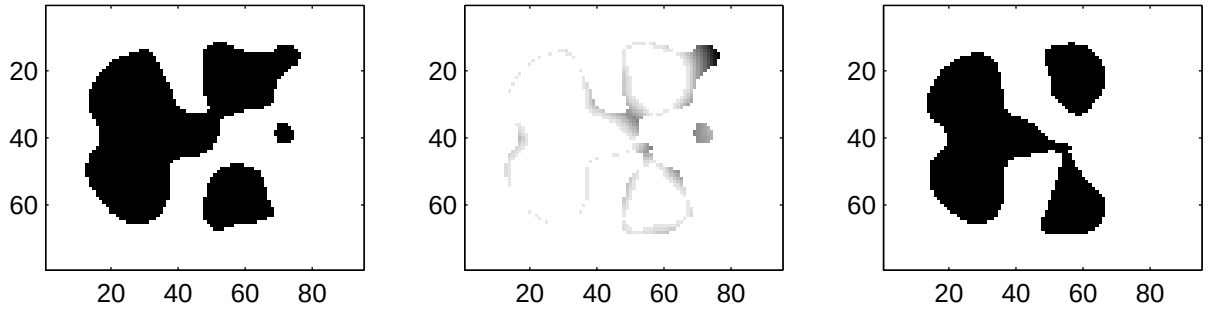


Figure 2: Slices of two brains (left, right) and slice of corresponding 3D LDMap (center)

Local distance map can be visualized also for 3D images as shown in figure 3.

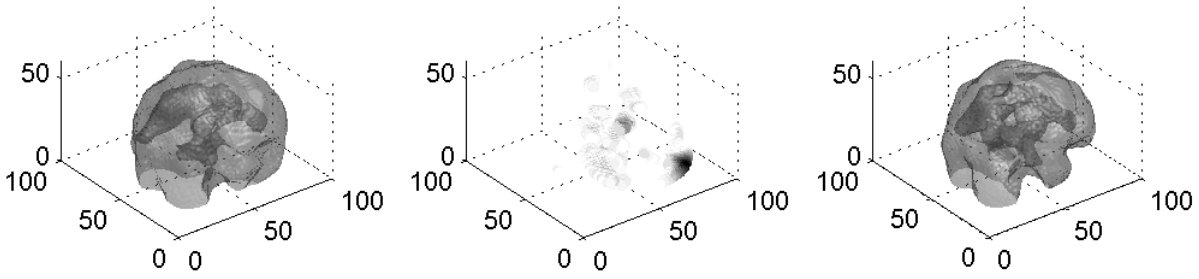


Figure 3: Brain 3D SPECT images and their LDMap

4 Conclusion

The methods mentioned in this paper have been examined deeply over 2D images. It has been proved by many experiments that they are relatively robust and stable to noise and outliers. Also, effective algorithms have been developed, which reduces the computation time and save memory. We extended the last mentioned method WHD to third dimension and we generated a 3D LDMap showing the usability of Windowed Hausdorff distance in medicine. The 3D LDMap highlights local dissimilarities between two human organs, which can be helpful for doctors when trying to detect abnormalities. In the further research, these LDMaps will be examined in more detail and utilized for automatic identification of Alzheimer disease.

Acknowledgment: The support of grant OHK4-027/10 CTU in Prague is gratefully acknowledged. The Authors would also like to thank Helena Trojanova and Renata Pichova from Clinique of Nuclear Medicine FNKV in Prague for providing the image data.

References

[1] A. Papadopoulos. *Metric spaces, convexity and nonpositive curvature.* European mathematical society, 2005, pp 105-110, ISBN 3-03719-010-8.

[2] A. M. J. Skulimowski. *Mathematical bases for the numerical evaluation of the Hausdorff distance.* Preprints of the 9th IMACS World Congress, Oslo, August 5-9, 1985; Vol. 5, pp.343-346.

[3] H. Alt, B. Behrends, J. Blomer. *Measuring the resemblance of polygonal shapes.* in Proc. Seventh ACM Symp. Comput. Geometry, 1991.

[4] D. P. Huttenlocher, G. A. Klanderman, W. J. Rucklidge. *Comparing images using the Hausdorff distance.* IEEE Transactions on pattern analysis and machine intelligence, vol. 15, no. 9, 1993, pp. 850-863.

[5] M. P. Dubuisson, A. K. Jain. *A modified Hausdorff distance for object matching*. In: Proc. 12th Internat. Conf. on Pattern Recognition, Jerusalem, Israel, October 1994, pp. 566-568

[6] E. Baudrier, F. Nicolier , G. Millon, R. Su. *Binary-image comparison with local dissimilarity quantification*. Pattern Recognition, vol. 41, issue 5, May 2008, pp. 1461-1478.

Parallel Algorithms for Numerical Solution of Laser Plasma Hydrodynamics

Ľuboš Bednárik

3rd year of PGS, email: lbs@centrum.sk
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Richard Liska, Department of Physical Electronics, Faculty for Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. Laser plasma simulations are well modelled by Lagrangian hydrodynamical equations with heat conductivity and laser absorption included. Pure Lagrangian simulation, however, may suffer from severe mesh distortion which can cause the failure of the computation. This difficulty is overcomed with the use of an Arbitrary Lagrangian Eulerian method, treating the hyperbolic hydrodynamical part of the model. The parabolic heat conductivity part is treated by splitting and mimetic method. A simulation of laser and material interaction demonstrates the usefullness of the method.

Keywords:

Abstrakt. Simulácia laserovej plazmy je typicky modelovaná prostredníctvom Lagrangeovských hydrodynamických rovníc s dodatočnými členmi pre tepelnú vodivosť a absorpciu laseru. Čisto Lagrangeovská simulácia včak môže trpieť vážnymi poruchami sieťky, čo dokáže zapríčiniť zlyhanie výpočtu. Tento problém sa dá prekonať použitím ALE metódy pre výpočet hyperbolickej hydrodynamickej časti modelu. Parabolická časť pre tepelnú vodivosť je riešená prostredníctvom splitting metódy a mimetickej metódy konečných diferencií. Použiteľnosť metódy je ilustrovaná na simulácii interakcie laserového zväzku s cieľovým materiálom.

Kľúčové slová:

1 Introduction

The Arbitrary Lagrangian Eulerian methods are a popular group of methods for simulation of continuum mechanics problems where the laser plasma simulation belongs. Compressible laser plasma typically includes regions of high compression and large expansion which require treatment by Lagrangian hydrodynamics, with heat conductivity and laser absorption included, allowing large scale changes of the computational domain. The computational mesh with the boundaries and boundary condition is fixed to the fluid and moves with the fluid. In some cases, e. g. in problems solving shear flows, however, the moving mesh can degenerate and become invalid with inverted cells when some node crosses the opposite edge of the same cell.

Therefore, the hyperbolic part of the model (hydrodynamics) is treated by the Arbitrary Lagrangian Eulerian (ALE) method which avoids moving mesh distortion and parallelized by means of the OpenMP library. The ALE method is a combination of Lagrangian and Eulerian methods and consists of three phases, the standard Lagrangian

computation, rezoning, and remaping. The rezoning is a simple mesh modification in order to repair different local degenerations, while the remapping corresponds to the Eulerian part of the ALE method and allows the mass flux between cells.

The parabolic part of the model (heat conductivity) is treated by splitting by an implicit mimetic finite difference method. Heat conductivity is presented in classical Spitzer-Harm form and the heat conductivity coefficient is a non-linear function of the temperature. The mimetic method works well on bad quality meshes, appearing in the Lagrangian simulations where non-linear heat conductivity effects like heat waves or discontinous diffusion coefficient can be observed. Methods used to solve the parabolic part of the model are parallelized by means of the LAPACK library which gives the possibility to perform linear algebra calculations in parallel, especially to obtain the solution of a tridiagonal matrix in parallel.

For the laser absorption there is presented a model where the laser is absorbed on critical surface. Under the term critical suface there is understand a surface where the critical density is reached.

2 ALE Method

For compressible fluid flow with heat conductivity and laser absorption the Euler equations in Lagrangian coordinates can be written in form

$$\frac{d\eta}{dt} = \vec{v}_S, \tag{1}$$

$$\frac{d\vec{v}}{dt} = -p_S, \tag{2}$$

$$\frac{d\varepsilon}{dt} = -p\vec{v}_S - W_S - L_S, \tag{3}$$

where t stands for time, $\eta = 1/\rho$, ρ is density, v velocity, p pressure, ε specific internal density, W is heat flux, and finally, L is energy flux density of laser radiation. Equations express consecutively law of conservation of mass, law of conservation of momentum, and law of conservation of energy.

Additionaly, the system is supplemented with the equation for mesh movement

$$\frac{d\vec{x}}{dt} = \vec{v}, \tag{4}$$

where $\vec{x}$ represents the position vector.

Finally, the system is completed with the equations of state (EOS) as functions $p = p(\varepsilon, \rho)$, $T = T(\varepsilon, \rho)$. For the ideal gas they have the form

$$p = \varepsilon\rho(\gamma - 1) \tag{5}$$

$$T = \frac{A}{Z+1}\frac{p}{c_p\rho},\ c_p = \frac{k_B}{m_u} \tag{6}$$

where γ is heat capacity ratio, Z degree of ionization, A atomic (proton) number, k_B Boltzmann constant and $m_u = 1,6605.10^{-24}g$ atomic mass unit.

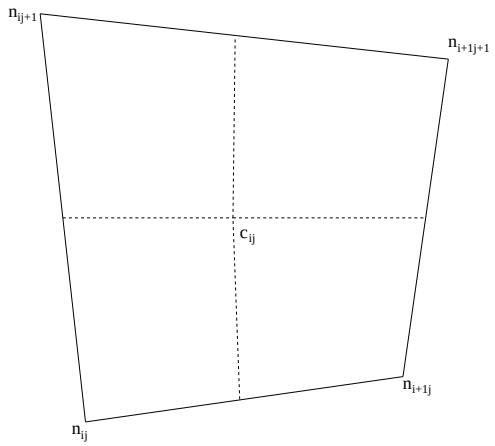


Figure 1: Cell geometry. A quadrilaterall cell divided into 4 subzones by connecting the cell center with the edge centers.

The system of equations (1), (2), (3) is split into hyperbolic part containing Lagrangian hydrodynamical equations and laser absorption

$$\frac{d\eta}{dt} = \vec{v}_S \tag{7}$$

$$\frac{d\vec{v}}{dt} = -p_S \tag{8}$$

$$\frac{d\varepsilon}{dt} = -p\vec{v}_S - L_S \tag{9}$$

and into parabolic part containing equation for heat conductivity

$$\frac{d\varepsilon}{dt} = -W_S - L_S. \tag{10}$$

2.1 Lagrange step

The Lagrangian hydrodynamical equations are numerically treated in two-dimensional domain on quadrilateral, logically rectangular computational mesh by means of compatible staggered discretization [4]. This discretization places scalar quantities (ρ, ε, p, T) into the mesh cells and vector quantities ($\vec{v}$,$\vec{x}$) into mesh nodes. Each quadrilateral cell, zone, is divided into 4 subzones by connecting the cell center with the edge centers as shown in Figure 1.

As it was mentioned, the mesh movement satisfies the equation (4) which can be discretized in the form

$$\frac{d\vec{x}_n}{dt} = \vec{v}_n \tag{11}$$

for each node n. After the mesh movement the density can be calculated in standard way as

$$\rho_c = \frac{m_c}{V_c}, \rho_s = \frac{m_s}{V_s} \tag{12}$$

for each cell c and subzone s, and seeing the Lagrangian assumption that the mass does not flow through the mesh and submesh edges, the mass of each cell and subzone remains constant.

The discrete law of conservation of momentum (2) at node n can be written as

$$m_n \frac{d\vec{v}_n}{dt} = \vec{F}_n = \sum_{c(n)} \vec{F}_{c(n),n} \tag{13}$$

where $\vec{F}_{c(n),n}$ is the force from the cell c (neighboring the node n) to the node n, and $\vec{F}_n$ is the total force from all four cells in the neighborhood of the node n affecting the node n. Forces $\vec{F}_{c(n),n}$ are composed of zonal pressure force $\vec{F}^p_{c(n),n}$, subzonal pressure force $\vec{F}^{dp}_{c(n),n}$, and artificial viscosity force $\vec{F}^q_{c(n),n}$

$$\vec{F}_{c(n),n} = \vec{F}^p_{c(n),n} + \vec{F}^{dp}_{c(n),n} + \vec{F}^q_{c(n),n}. \tag{14}$$

The forces $\vec{F}_{c(n),n}$ are used also in discretization of law of conservation of energy (3)

$$m_c \frac{d\varepsilon_c}{dt} = -\sum_{n(c)} \vec{F}_{n(c),c}\vec{v}_{n(c)}, \tag{15}$$

where $n(c)$ represents a node in the neighborhood of the cell c, and $\vec{F}_{n(c),c}$ is the force from node n (neighboring the cell c) affecting the cell c. This guarantees conservation of the total energy [4].

The viscosity force term $\vec{F}^q_{c(n),n}$ in (14) can be expressed by means of several viscosity types. One of the simplest artifical viscosity is a simple bulk viscosity based on Kuropatenko formula [5, 11]. Another forms of viscosities that can be used are edge viscosity [5] and tensor viscosity [3].

The system of ordinary differential equations (11), (13), (15), where the mass is conserved due to Lagrangian assumptions, represents the spatial discretization of the system of hydrodynamical equations. In addition, the system is also discretized by means of a predictor-corrector second order method for all nodes and all cells in time.

2.2 Laser absorption

In the energy equation (9) there is a term L representing the energy transfered to the system because of the absorption of the laser radiation. It is assumed in the form

$$L = div\vec{I},$$

where $\vec{I}$ represents the laser intensity. The laser is absorpted only at a critical surface which is the isosurface with a critical density

$$\rho_c = 1,86 \cdot 10^{-3} \frac{A}{Z\lambda_\mu^2},$$

where A is atomic (proton) number, Z degree of ionization, and λ_μ laser wavelength in μm.

The intensity in places of material with subcritical density is given by laser radial and temporal Gaussian profiles. Behind the critical surface the laser intensity is zero. The (x, y) components of laser intensity are projected on the edge normals at the edge midpoints and divergence of the intensity approximated by the standard formula derived from Green's theorem.

2.3 Heat conductivity

The parabolic part of the system, the equation for heat conductivity, is assumed in the form

$$\frac{dT}{dt} = \nabla.(\kappa \nabla T), \tag{16}$$

where κ stands for heat conductivity, and the term $\nabla.(\kappa\nabla T)$ represents the heat flux. The equation is treated after each Lagrangian step by a scheme fully implicit in time which allows the timestep to be equal to the timestep of the hyperbolic system. Operators of divergence and gradient are discretized by a mimetic finite difference method [12], leading to a system with a symmetric positive definite matrix which is solved by conjugate gradient method.

2.4 Rezoning and remapping

Allowing the mesh to move, it can become strongly deformed, and these deformations need to be improved. The rezoning phase of the ALE algorithm, which covers mesh smoothing and untangling, is a way how to repair these mesh distortions. However, because of remapping, it is necessary to move only the vertexes which have to be moved, and as little as possible. There exist several methods for rezoning, the combination of a feasible set method and global optimization [14] or reference Jacobian method [9]. The algorithm used in our ALE computation is the Winslow smoothing method [16].

In the remapping phase of the ALE method there is performed a conservative interpolation of conserved quantities from the original, the old Lagrangian mesh to the new, smoother one. It is required this procedure conserves the mass, each component of the momentum, and the total energy. Furthermore, the monotonicity, or at least the local bounds, for density, velocity and specific energy have to be preserved, and also the remapping should be as accurate as possible. All these can be achieved by a method which, first, performs a piecewiese linear reconstruction [1] of conserved quantities, then integrates the reconstruction [10, 6] over regions swept by edges as the edges move from old mesh to the new one, and finally, corrects (repairs [10, 13]) possibly created new extrema, which does not preserve the local bounds, by redistribution to the neighbouring cells.

Other techniques almost always need to have satisfied all the imposed requirements. They can combine a low-order intercell fluxes (which preserve local bounds by default) with a higher-order (generally unconstrained) fluxes. An example can be seen in [15].

3 Results

As the numerical example there is presented a laser and material interaction simulation from laser plasma physics performed at Prague Asterix Laser System (PALS) facility. A massive alluminium target is irradiated from the top by intensive laser beam pulse. The laser beam is operating at the energy $115\,J$ with the wavelength $\lambda = 438\,nm$ and pulse length $300\,ps$. The simulation starts at the moment of impact and continues till the simulation time $t = 10\,ps$ as can be seen on Figure 2.

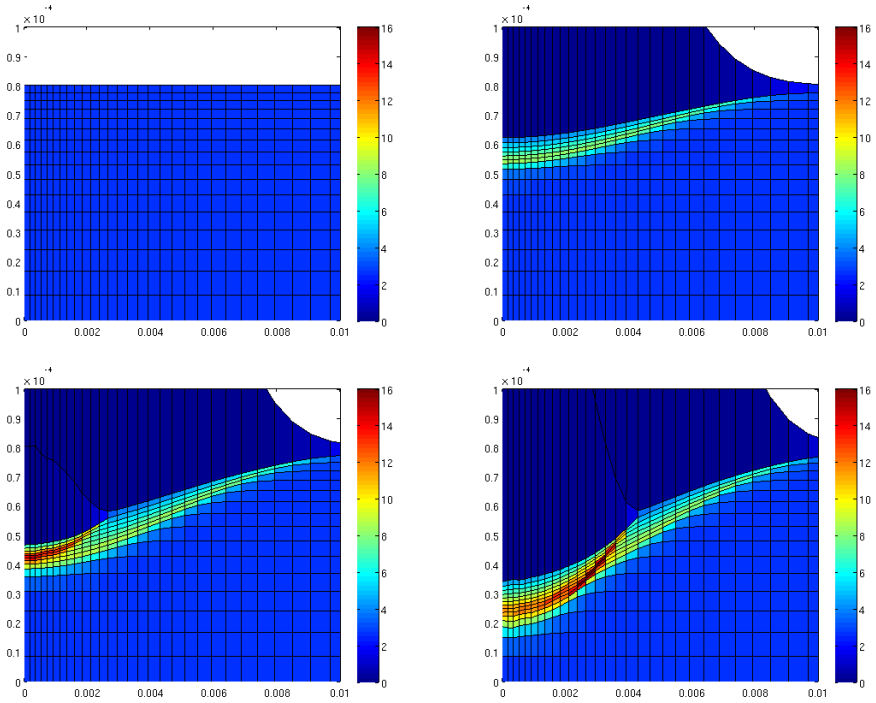


Figure 2: Simulation of material irradiation by intensive laser beam pulse. On the top-left picture there can be seen density of the massive target just before the impact $t = 0$, on the top-right at time $t = 4\,ps$, then at time $t = 7\,ps$ on the bottom-left, and finally, at time $t = 10\,ps$ on the bottom-right picture. The colormap displays the density range from 0 to $16\,g.cm^{-3}$.

In pure Lagrangian simulation the computation fails very soon due to fatal mesh distortion, however, with the ALE algorithm, the mesh smoothness is preserved, and the computation can continue till the final time. The smoothing was performed with weighted-average rezoning and remapping with piece-wise linear interpolation with re-pairing.

After the laser impact, the massive target starts to compress itself which leads to its temperature increase and material evaporation. Part of the material becomes ionized and forms an expanding plasma corona. A crater is formed in the massive target, representing the interface between liquid and gas phase. Later in time, the crater stops in further movement, while a shockwave formed at the solid-liquid phase interface is still propagating deeper in the target material.

4 Conclusions

An arbitrary Lagrangian Eulerian algorithm has been developed in order to simulate different laser plasma problems. The algorithm is developed on logically orthogonal 2D grid, and includes pure Lagrangian calculation with heat conductivity and laser absorption followed by mesh rezoning and remapping procedure. It has been applied to a simulation, laser and material interaction problem, inspired by the experiment performed at the PALS facility. The Lagrangian calculation without the ALE extension was unable to calculate the solution due to severe mesh distortions, while the complete ALE simulation provides reasonable results what demonstrates the usefullness of the algorithm.

Acknowledgements

This research has been supported by the Czech Ministry of Education grant MSM6840770022. The author thanks R. Liska for leading support, and P. Vachal and M. Kucharik for constructive comments and discussion.

References

[1] T. J. Barth. *Numerical methods for gasdynamic systems on unstructured meshes*, in An introduction to Recent Developments in Theory and Numerics for Conservation Laws, C. R. D. Kroner, M. Ohlberger, ed., Berlin, 1997, Lecture Notes in Computational Science and Engineering, Springer, 195-284.

[2] S. Borodziuk, A. Kasperczuk, T. Pisarczyk, K. Rohlena, J. Ullschmied, M. Kalal, J. Limpouch, and P. Pisarczyk. *Application of laser simulation method for the analysis of crater formation experiment on PALS laser*, Czechoslovak Journal of Physics **53** (2003) 799-810.

[3] J. C. Campbell and M. J. Shashkov. *A tensor articial viscosity using a mimetic finite difference algorithm*, Journal of Computational Physics, **172**(2) (2001) 739-765.

[4] J. Caramana, D. E. Burton, M. J. Shashkov, and P. P. Whalen. *The construction of compatible hydrodynamics algorithms utilizing conservation of total energy*, Journal of Computational Physics **146** (1998) 227-262.

[5] E. J. Caramana, M. J. Shashkov, and P. P. Whalen. *Formulations of artifcial viscosity for muti-dimensional shock wave computations*, Journal of Computational Physics, **144**(2) (1998) 70-97.

[6] R. Garimella, M. Kuchařík, and M. Shashkov. *Efficient algorithm for local-bound-preserving remapping in ALE methods*, in Numerical Mathematics and Advanced Applications, M. Feistauer, V. Dolejsi, P. Knobloch, and K. Najzar, eds., Springer-Verlag Berlin Heidelberk New York, 2004, 358-367.

[7] T. Kapin, M. Kuchařík, J. Limpouch, R. Liska, P. Váchal. *Arbitrary Lagrangian Eulerian method for laser plasma simulations*, International Journal of Numerical Methods in Fluids **56** (2008) 1337-1342.

[8] T. Kapin, M. Kuchařík, J. Limpouch, R. Liska. *Hydrodynamic simulations of laser interactions with low-density foams*, Czechoslovak journal of Physics **56** (2006) B493-B499.

[9] P. Knupp, L. Margolin, and M. Shashkov. *Reference Jacobian optimization-based rezone strategies for arbitrary Lagrangian Eulerian methods*, Journal of Computational Physics **176** (2002) 93-128.

[10] M. Kuchařík, M. Shashkov, and B. Wendroff. *An efficient linearity-and-bound-preserving remapping method*, Journal of Computational Physics **188** (2003) 462-471.

[11] V. F. Kuropatenko. *Difference methods for solutions of problems of mathematical physics*, volume 1, page 116, American Mathematical Society, Providence, 1967.

[12] M. Shashkov and S. Steinberg. *Solving diffusion equations wit rough coefficients in rough grids*, Journal of Computational Physics **129** (1996) 383-405.

[13] M. Shashkov and B. Wendroff. *The repair paradigm and application to conservation laws*, Journal of Computational Physics **198** (2004) 265-277.

[14] P. Váchal, R. Garimella, and M. Shashkov. *Untangling of 2D meshes in ALE simulations*, Journal of Computational Physics **196** (2004) 627-644.

[15] P. Váchal and R. Liska. *Sequential Flux-Corrected Remapping for ALE Method*, in Numerical Mathematics and Advanced Applications. ENUMATH 2005, Springer, 2006, 671-679.

[16] A. M. Winslow. *Equipotential zoning of two-dimensional meshes*, Technical Report UCRL-7312, Lawrence Livermore National Laboratory, 1963.

Morphological Analysis of 3D SPECT Images via Nilpotent t-Norms in Diagnosis of Alzheimer's Disease

Tomáš Bělíček

2nd year of PGS, email: belictom@fjfi.cvut.cz
Department of Software Engineering in Economy
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Jaromír Kukal, Department of Software Engineering in Economy,
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. This article describes a fuzzy image processing method which can serve as a potential diagnostic tool for Alzheimer's disease. We have set up a sequence of several image-processing processes based on morphological fuzzy edge detection followed by watershed segmentation. Then we undertook an analysis of nilpotent t-norms forming the edge detectors and the parameter of Gaussian filter, and in the end we carried out an statistical evaluation of segments of human brains images. Our goal was to demonstrate the usefulness of the Łukasiewicz BL-algebra in feature extraction of 3D biomedical images by using enhanced methods of image morphology.

Keywords: image processing, fuzzy logic, watershed transformation, Alzheimer's disease

Abstrakt. Tento článek popisuje metodu fuzzy zpracování obrazu, která má potenciál sloužit jako diagnostický nástroj pro Alzheimrovu chorobu. Navrhli jsme poslouponost několika procesů založených na fuzzy detekci hran pomocí morfologických operátorů a segmentaci obrazu. Poté jsme analyzovali nilpotentní normy, které tvoří základ pro hranové detektory a parametr Gaussova filtru. Nakonec jsme statisticky vyhodnotili počty získaných segmentů z obrazů lidských mozků. Našim cílem bylo demonstrovat užitečnost Łukasiewiczovy BL-algebry při extrakci rysů 3D biomedicínských obrazů za použití rozšířených morfologických metod.

Klíčová slova: zpracování obrazu, fuzzy logika, transformace pomocí rozvodí, Alzheimerova choroba

1 Introduction

Alzheimer's disease (AD) is the most frequent degenerative dementia. Despite important progress in the field of neurology and neurosciences during the last years, its diagnosis remains, however, based essentially on clinical appreciation. Current diagnostic criteria reach sensitivity and specificity about maximally 80-87 %.

The principal idea of our research takes advantage of the medical finding that affected brains are usually characterized by a different structure of gray and white matter. Hence, we predicted that we would get a different number of segments after applying some convenient segmentation methods. To achieve this goal, we had to set up an image-processing

procedure consisting of appropriate functions. The edge detection was based on morphological fuzzy operators in coordination with transformations of the image contrast, while operators were taken from the Lukasiewicz BL-algebra.

2 Edges and nilpotent t-norms

As we have mentioned, the edge detectors in our process make use of fuzzy logic. Our approach is based on t-norms [6] and [1], but we differ in results via generator theory. The novel decomposition of edge decomposition is based on increasing bijection transform as intensity preprocessing and the absence of inverse transform. Speaking more accurately, the detectors are built of fuzzy operators from BL-algebras generated by nilpotent t-norms. Every standard BL-algebra is a standard residuated lattice [3] and [2], with two more axioms:

$$x \wedge y = x \otimes (x \to y) \tag{1}$$

$$(x \to y) \vee (y \to x) = 1 \tag{2}$$

for all $x, y, z \in [0, 1]$ and it is formed by $\mathfrak{L} = \langle [0, 1], \vee, \wedge, \otimes, \to, 0, 1 \rangle$ where $\otimes$ represents a t-norm and $\to$ corresponds to residuum. The word 'standard' means that we operate over BL-algebras having a support in the real interval $[0; 1]$. Similar utilization can be found in [8],[9] and [10].

We chose the BL-algebra because it is suitable for our gray-scale images where intensities flow between 0 and 1 and its logic operations produce the desired behavior, concretely speaking they contribute by suppressing and heightening specific edges and areas in images.

Every t-norm defines its BL-algebra and its derived operators such as addition, subtraction or biresiduum, causes different behavior on systems where it is used. We focused on a subset of nilpotent t-norms in our application of edge detector and we figured out several detectors. Inspirative results were taken from [14] or [13]. The main goal was to prove that every formula from a general BL-algebra generated by a nilpotent t-norm is equivalent to the same formula expressed in the Łukasiewicz BL-algebra and the bijection ϕ. Let us say that every function $\mathrm{F} : [0, 1]^n \to [0, 1]$ is a formula realized in some BL-algebra if it is combined exclusively by operators from the same BL-algebra or by their derived operators. We can say that every formula realized in some BL-algebra generated by a nilpotent t-norm can be expressed by the same formula realized in the Łukasiewicz BL-algebra and a bijection ϕ.

$\mathcal{L}_1$ is a standard BL-algebra generated by a nilpotent continuous t-norm. $\mathcal{L}_2$ is a standard Łukasiewicz BL-algebra. Let $\varphi : [0, 1] \to [0, 1]$ be an increasing bijection, $\mathrm{F} : [0, 1]^n \to [0, 1]$ and $\vec{x} = (x_1, \ldots, x_n)$, where $n \in \mathbf{N}$. If F_1 is F realized in $\mathcal{L}_1$ then

$$\mathrm{F}_1(\vec{x}) = \varphi^{-1}(\mathrm{F}_2(\varphi(x_1), \ldots, \varphi(x_n))), \tag{3}$$

where F_2 is F realized in $\mathcal{L}_2$.

For the purposes of BL-algebra application in image processing, we always work with some volume element (voxel) and its neighborhood in a 3D image matrix. We define a function built of BL-algebra operators, which transform the intensity of a voxel with regard to intensities of neighboring voxels. We define the voxel neighborhood of a given central voxel $V_{u,v,w}$ from a 3D image $\mathfrak{I}$ as

$$\mathcal{N}_R(V_{u,v,w}) = \{V_{i,j,k} \in \mathfrak{I} : |u-i| \leq R \wedge |v-j| \leq R \wedge |w-k| \leq R\},$$

where $R \in \mathbf{N}_0$. Generally, a function transforming intensities of image voxels is a local operation based on subtracting minimum from maximum in a neighborhood of a central element; hence, it heightens edges in a global aspect. Such function can be modified in some way, but it must be still built of BL-algebra operators. We substitute the subtraction by the distance operator in two cases of the following examples of edge-heightening functions. The function has three parameters $y = \mathrm{f3}(c, R)$ where c represents a central voxel and R is a radius of the neighborhood $\mathcal{N}_R(c)$ of a central element c. Here are several fuzzy edge detecting functions where index i runs through the neighborhood of a central element c:

- *fuzzy Minkowski sausage:* $\mathrm{f}_1 = \max(x_i) \ominus \min(x_i)$
- *fuzzy morphological edge detector* $\mathrm{f}_2 = (\max(x_i) \ominus c) \wedge (c \ominus \min(x_i))$
- *modified fuzzy morphological edge detector* $\mathrm{f}_3 = (\max(x_i) \ominus c) \oplus (c \ominus \min(x_i))$

Our approach to the functions is similar to [4], [5] and [7]. The presented formulas of fuzzy morphological operators in BL-algebra aim to alternate classical gradient methods based on derivation. Thus, each element of an output image of some edge-detecting operator represents the fuzzy measure of being the edge with regard to intensities of other elements. It means that the measure of being a part of an edge must be computed in respect to the maximum of an image. The introduced edge detecting formulas can be expressed in the Łukasiewicz BL-algebra according to the theorem. After applying this relation we get successively

$$\mathrm{f}_1 = \varphi^{-1}[\varphi(\max(x_i)) \ominus_{\mathbf{Ł}} \varphi(\min(x_i))] \tag{4}$$

$$\mathrm{f}_2 = \varphi^{-1}[\varphi(\max(x_i)) \wedge_{\mathbf{Ł}} \varphi(\min(x_i))] \tag{5}$$

$$\mathrm{f}_3 = \varphi^{-1}[\varphi(\max(x_i)) \ominus_{\mathbf{Ł}} \varphi(c)) \oplus_{\mathbf{Ł}} (\varphi(c) \ominus_{\mathbf{Ł}} \varphi(\min(x_i)))] \tag{6}$$

If we focus on the increasing bijection $\varphi : [0,1] \to [0,1]$, we can observe that it changes contrast and brightness of an image. In the later experiment we design this function by polynomial $\varphi(x) = ax + bx^2 + cx^3$ where parameters a, b, c are counted upon given values of contrast and brightness and the assumption that brightness and contrast are defined as

$$bri = \int_0^1 \varphi(x)dx = \frac{1}{2}a + \frac{1}{4}b + \frac{1}{8}c$$

$$contr = \varphi'(\frac{1}{2}) = a + b + \frac{3}{4}c,$$

where $\varphi(0) = 0$ and $\varphi(1) = 1$ (thus $a + b + c = 1$). We choose only a solution that fulfils the condition of increasing polynomial on the real interval $[0, 1]$. The operation φ^{-1} is not finally executed owing to the zero-effect on the edge detection, which is based only on the value comparison.

3 Patients and Control Groups

We enrolled SPECT data from 17 available adult patients (10 males, 7 females) with definite Alzheimer's disease (AD) confirmed by post mortem brain autopsy. As controls (CN) we used SPECT data from randomly chosen 10 patients (7 males, 3 females) with amyotrophic lateral sclerosis (ALS), a neurodegenerative disorder affecting predominantly upper and lower motor neurons. These patients underwent SPECT and detailed cognitive evaluation as a part of a research protocol (submitted data).

4 Methodology

Let the initial input image matrix be denoted $\mathbf{A}$. The method of whole-image segmentation consists of several steps. At first, we carry out the image filtration using the kernel given by the relation

$$\mathbf{F}(\vec{v}) = e^{-\frac{\|\vec{v}\|^2}{2\sigma^2}} \tag{7}$$

where $\vec{v}$ is the space coordinate vector and σ represents the filter radius. The filtration is provided by convolution

$$\mathbf{B} = \mathbf{A} * \mathbf{F}. \tag{8}$$

At the second step, voxel intensities are transformed into the real interval $[0; 1]$ to get a matrix applicable for fuzzy edge detectors. This step must be carried out owing to the noise that arises from acquiring a digital record on the SPECT device

$$\mathbf{C} = \frac{\mathbf{B}}{\max(\mathbf{B})}. \tag{9}$$

Then the normalization is followed by soft thresholding which cuts off the image values below θ_1. This is made feasible by subtracting θ_1 in the Łukasiewicz BL-algebra from each element of an image matrix

$$\mathbf{D} = \max(\mathbf{C} - \theta_1, 0). \tag{10}$$

The next step consists of fuzzy edge detection which was comprehensively described above. Firstly, we apply a φ-function, counted from variables *contr* (contrast) and *bri* (brightness), on an image, and finally we continue with application of one of three fuzzy edge detector function f_i using voxel neighborhood with radius R

$$\mathbf{G} = \text{detect}(\mathbf{D}, \text{f}_i, R, contr, bri). \tag{11}$$

The second soft thresholding with θ_2-value comes right after the fuzzy edge detection and serves as a tool for cutting off image values representing edges that may be illusory or too narrow. This is provided in the same way as in the first thresholding.

$$\mathbf{H} = \max(\mathbf{G} - \theta_2, 0) \tag{12}$$

Keeping these edges on an image may cause an undesirable impact on evaluated features such as the final number of totalled segments. The final step is represented by the watershed transformation. It returns an image containing separate regions marked by integers that have been created by flooding an image from the global minimum to the global maximum. The function for watershed transform wshed computes a label matrix identifying the watershed regions with the parameter of connectivity system. This parameter identifying the neighborhood system of the watershed transform was fixed upon the 6-connectivity system in 3D space as

$$\mathbf{J} = \text{wshed}(\mathbf{H}). \tag{13}$$

The watershed method including a computational algorithm is described in [15]. Such transformed images have served for evaluating particular features. They were processed by means of several experiments based on evaluating different features and comparing both classes of images. The previously presented sequence of image processes was applied to every 3D input image and then it was used for evaluating a selected feature.

If we summarize the designed method, we need to analyze the following parameters to find their optimal intervals for the finest results of our application:

- σ - the filter radius from $(0; +\text{inf})$
- θ_1 - the first threshold parameter from $[0, 1]$
- *contr* - the value of contrast from $[0.2, 1.4]$
- *bri* - the value of brightness was fixed upon 0.5
- $\text{f}_1, \text{f}_2, \text{f}_3$ - the fuzzy edge detecting function
- θ_2 - the second threshold parameter from $[0, 1]$
- R - the radius of the edge detector's neighborhood from $\mathbb{N}$

5 Results

We have chosen the observed feature as a total segment number of an image processed by the designed method. The values of all images from both groups were used for evaluating Student's two-sample t-test. The output p-value represents the probability that the null hypothesis is true, where the null hypothesis claims that mean values of two random

Table 1: The system parameters for the optimum $p - value = 4.48 \times 10^{-2}$ evaluated on entire brains.

Edge	R	*contr*	σ	θ_1	θ_2	*nhood*
f_2	4	0.8676	1.509	0.4900	0.3548	6

Table 2: The system parameters for the optimum $p - value = 3.816 \times 10^{-7}$ evaluated on pariatal area.

Edge	R	*contr*	σ	θ_1	θ_2	*nhood*
f_2	4	1.0148	0.2763	0.8997	0.3272	6

variables are identical.

As far as the phase of the edge detection is concerned, the scale of contrast had an unambiguously significant influence on the final number of segments. The change of brightness had no influence on the final segmentation due to the interprocess of fuzzy edge detectors, therefore we did not take this parameter into account while evaluating our feature. Another important parameter was found in the initial threshold θ_1. Since it sets elements with values below the θ_1 to zero, it suppresses edges which are false due to noise or they are completely unimportant with respect to other high edges in the image.

The first aim of the evaluation was to find optimal configuration of our system, i.e. values of parameters σ, θ_1, θ_2, *contrast*, edge detecting function f and radius R which lead to the minimum p-value of the Student's two-sample test. We reached our optima by means of a heuristic optimization method based on the differential evolution described in [16], [17] and [18].

The first image processing and evaluation was made over entire area of a human brain. The table 5 shows the final values of input parameters and the optimum t-value (respectively p-value) computed by using differential evolution algorithm.

Next part of this article presents results of the evaluation made only on the part of a brain called pariatal area, where a brain matter is usually mostly affected by the disease. From the medical point of view, structure of brains in these two symmetric areas is more disrupted for AD than for CN class. The following table 5 contains final values of the evaluation at pariatal area and the boxplot graph for each brain with number of segments in y-axis.

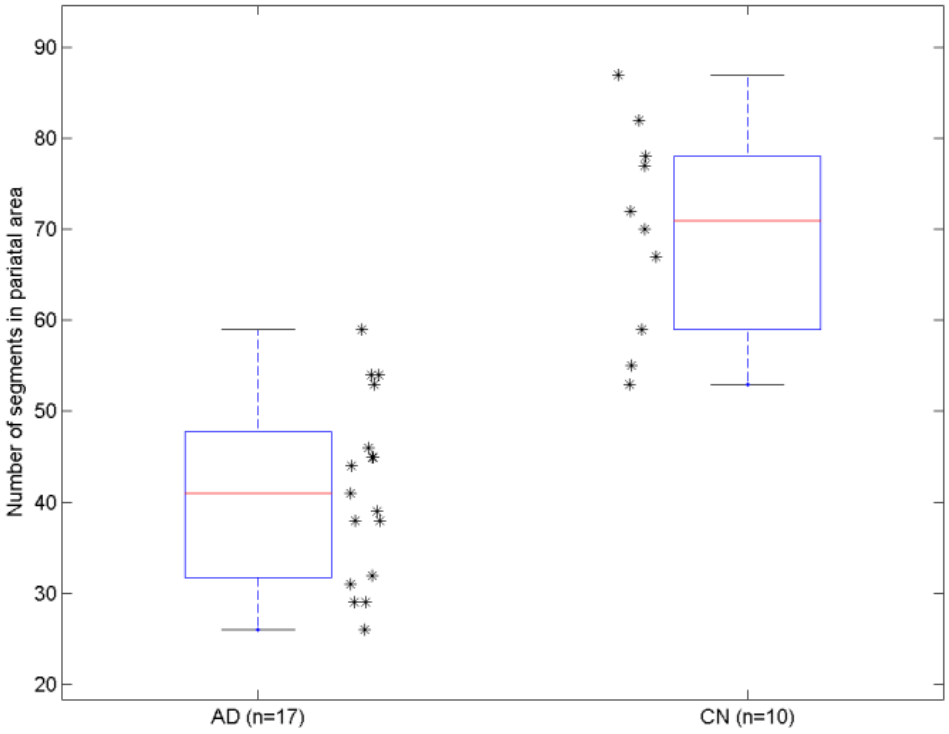


Figure 1: Boxpot graph with number of segements for each class evaluated with parameters given from pariatal areas

Such results still need to be verified on a larger number of patients with Alzheimer's and healthy controls and compared with results from patients with other dementia subtypes. This research will be also supported by another approach based on fuzzy edge detection in dodecahedral grid system.

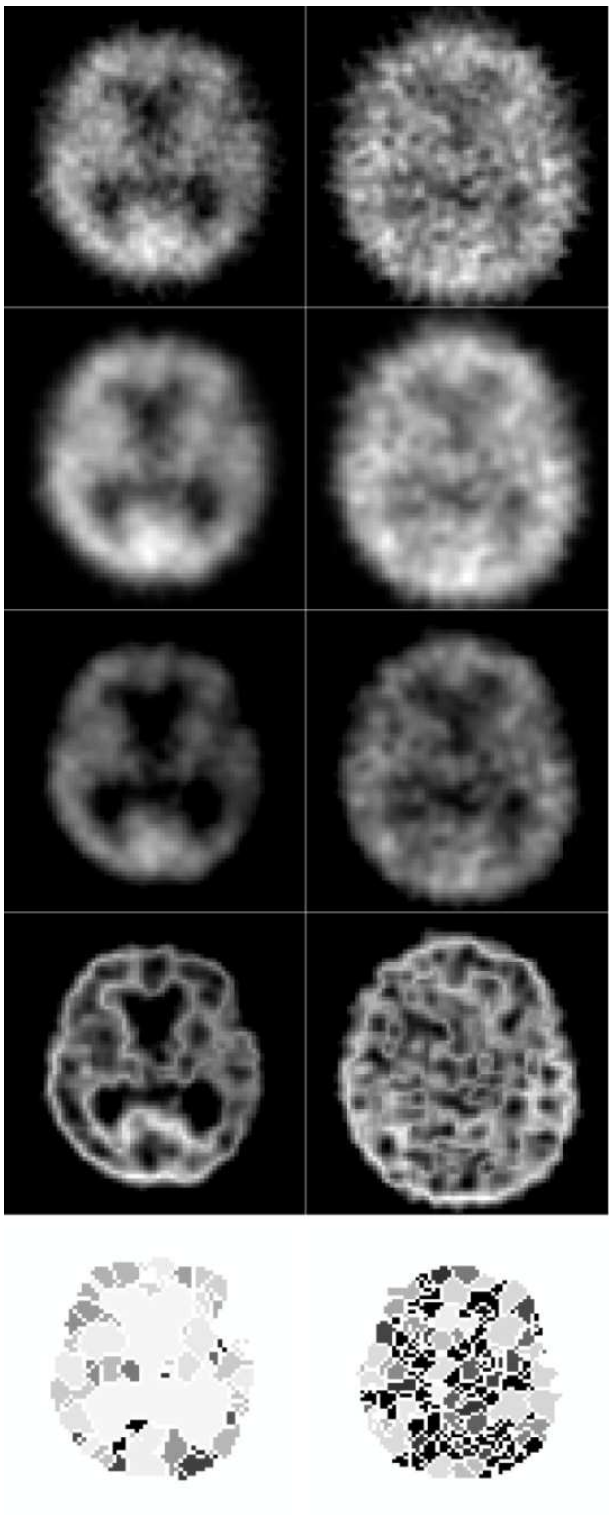


Figure 2: The processing of input images AD (left) and CN (right) from the beginning (top) to the end (bottom).

References

[1] M. Gonzalez. M. Hidalgo. *Edge-Images using a uniform-based fuzzy mathematical morphology: opening and closing.* Advances in computational vision and medical image processing, (2009)

[2] S. Gotwalld. P. Hájek. *Triangular norm-based mathematical fuzzy logics.* Elsevier, Amsterdam, (2005)

[3] P. Hájek. *Basic fuzzy logic and BL-algebras.* Soft Computing, Springer, (1998)

[4] J. Kukal, D. Majerová. *Multicriteria approach to 2D image de-noising by means of Łukasiewicz algebra with square root.* Neural Network World, Czech Republic, (2002)

[5] J. Kukal, D. Majerová, V. Musoko, A. Pavelka, A. Procházka. *Nonlinear Filtering of 3D MRI in Matlab* CTU, Prague, (2004)

[6] C. Lopez, Molina et al. *A t-norm based approach to edge detection* Computational and Ambient Intelligence, (2009)

[7] D. Majerová. *Image Processing By Means of Łukasiewicz Algebra with Square Root*, PhD thesis at CTU, Prague, (2004)

[8] P. Margos, V. Tzouvas, G. Stamou. *Lattice Fuzzy Signal Morphology Operators and Generalized Image Gradients.* National Technical University of Athens, (2003)

[9] P. Margos, V. Tzouvas, G. Stamou. *Synthesis and Applications of Lattice Image Morphology operators Based on Fuzzy Norms.* National Technical University of Athens, (2001)

[10] P. Margos, V. Tzouvas, G. Stamou. *A Lattice Control Model of Fuzzy Dynamical Systems in State-Space.* National Technical University of Athens, (2000)

[11] M. Navara, P. Olšák. *Principles of Fuzzy Sets* CTU, Prague, (2002)

[12] V. Novák. *Fuzzy Sets and Their Applications* SNTL, Prague, (1990)

[13] E. P. Klement, M. Mesiar. *Logical, Algebraic, and Probabilistic Aspects of Triangular Norms* Elsevier, Amsterdam, (2005)

[14] E. P. Klement, M. Mesiar. *Logical, Algebraic, Analytic and Probabilistic Aspect of Triangular Norms.* Elsevier, (2005)

[15] J. B. T. M. Roerdink, A. Meijster. *The Watershed Transform: Definitions, Algorithms and Parallelization Strategies.* IOS Press, Groningen, (2001)

[16] J. Tvrdík. *Differential Evolution with Competitive Setting of its Control Parameters, TASK Quarterly.* vol. 11, page 169-179, ISSN 1428-6394, (2007)

[17] J. Tvrdík, I. Křivý, L. Mišík. *Adaptive Population-based Search: Application to Estimation of Nonlinear Regression Parameters.* COMPUT STAT DATA AN., vol. 52, page 713-724., ISSN 0167-9473, (2007)

[18] J. Tvrdík. *Adaptive Differential Evolution: Application to Nonlinear Regression.* Proceedings of the International Multiconference on Computer Science and Information Technology. Wisla: PTI, page 193-202, (2007)

Transport of Colloids Through Heterogeneous Porous Media

Pavel Beneš

2nd year of PGS, email: benespa1@fjfi.cvut.cz
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Jiří Mikyška, Department of Mathematics, the Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. The goal of this contribution is to describe the transport of colloids in heterogeneous porous media. This work includes equations describing the flow field, transport of colloids, and deposition of colloids in porous media. Then we describe a numerical discretization of the system of equations describing the colloid transport by means of operator splitting finite volume method. We present some numerical results at the end of the contribution.

Keywords: transport of colloidal particles, heterogeneous porous medium, finite volume method

Abstrakt. Hlavním cílem tohto příspěvku je popis transportu koloidů v heterogenním porézním prostředí. Tato práce obsahuje rovnice popisující proudové pole, transport koloidů a jejich ukládání v porézním prostředí. Dále je v práci obsažena numerická diskretizace tohoto systému rovnic popisujícího transport koloidů za použití rozkladu operátoru a metody konečných objemů. Na závěr příspěvku jsou uvedeny některé dosažené výsledky.

Klíčová slova: transport koloidních částic, heterogenni porézní médium, metoda konečných objemů

1 Introduction

Colloids are small particles with at least one dimension smaller than 100 nm. Because of their size, colloid particles are strongly attracted to the pore surfaces. On the other hand, colloids, like nanoiron particles, can be strongly reactive and can be used in remediation of contamined sites. To plan a suitable remediation strategy, one has to understand mechanisms of colloid transport and their deposition in the subsurface. This understanding can be obtained by means of numerical models. This paper contains equations describing colloidal transport in porous media. Then introduce semi-explicit scheme and present some results of numerical experiments.

2 The Physical Model

This section presents equations describing the colloidal transport in porous media [1].

2.1 Flow Field Equation

The following equation describes a distribution of pressure in a porous media

$$\frac{\partial(\Phi\varrho)}{\partial t} - \operatorname{div}\left(\frac{k\varrho}{\mu}(\nabla p - \varrho\mathbf{g})\right) = \varrho(s_+ - s_-), \tag{1}$$

where k is the permeability, μ the dynamic viscosity, $\mathbf{g}$ the gravity, ϱ is density, s_+ and s_- are the sources and sinks and $p\ [Pa]$ is the unknown fluid pressure. When the pressure distribution is known the Darcy velocity can be computed using from the Darcy law

$$\mathbf{q} = -\frac{k}{\mu}(\nabla p - \varrho\mathbf{g}). \tag{2}$$

The flow field will be necessary for description of the colloidal transport.

2.2 Colloid Transport Equation

The colloid transport equation can be derived from the mass balance of colloids over the REV (representative element volume). There are three main mechanisms controlling the colloidal transport: hydrodynamic dispersion, advection and colloid deposition and release. This can be described by the generalized advection dispersion equation, where the unknown is the particle number concentration n

$$\frac{\partial n}{\partial t} = \nabla\cdot(D\nabla n) - \nabla\cdot(\mathbf{V}\cdot n) - \frac{f}{\pi a_p^2}\frac{\partial\theta}{\partial t}, \tag{3}$$

where θ is the specific surface coverage, defined as

$$\theta = \frac{\text{total cross-section area of deposited colloids}}{\text{interstitial surface area of the porous media solid matrix}},$$

f is specific surface area

$$f = \frac{\text{interstitial surface area}}{\text{porous medium pore volume}},$$

a_p is the radius of colloidal particles, D is the particle hydrodynamic dispersion tensor and $\mathbf{V}$ is the particle velocity vector. It is possible to write the particle hydrodynamic dispersion tensor as

$$D_{ij} = \alpha_T\bar{V}\delta_{ij} + (\alpha_L - \alpha_T)\frac{\bar{V}_i\bar{V}_j}{\bar{V}} + D_dT\delta_{ij},$$

where D_d is the Stokes-Einstein diffusivity, $\bar{V}_i$, $\bar{V}_j$ are components of the interstitial velocity, α_L is the longitudinal dispersivity, α_T is the transverse dispersivity and T is the tortuosity of the porous medium.

2.3 Colloid Deposition and Release

Let λ be the percentage part of the solid matrix with favorable conditions for colloid deposition. This can be for example areas with iron oxides on its surface. These surfaces are typically positively charged and colloids are typically negatively charged. Deposition on the surfaces is usually irreversible. On the rest $(1-\lambda)$ of the solid matrix surface are unfavorable conditions for the colloidal deposition. Deposition takes place on both parts, but difference in rates can be huge. For particle surface coverage rate we can adopt this patchwise model

$$\frac{\partial\theta}{\partial t} = \lambda\frac{\partial\theta_f}{\partial t} + (1-\lambda)\frac{\partial\theta_u}{\partial t}, \tag{4}$$

where θ_f is the favorable surface fraction and θ_u is the unfavorable surface fraction. The rates are described by the following partial differential equations

$$\frac{\partial\theta_f}{\partial t} = \pi a_p^2 k_{dep,f} n B(\theta_f) - k_{det,f}\theta_f R(\theta_f), \tag{5}$$

$$\frac{\partial\theta_u}{\partial t} = \pi a_p^2 k_{dep,u} n B(\theta_u) - k_{det,u}\theta_u R(\theta_u), \tag{6}$$

where k_{dep} is the colloid deposition rate constant, k_{det} is the colloid release rate constant, $B(\theta)$ is the dynamic blocking function and $R(\theta)$ is the dynamic release function. The colloid deposition rate coefficient k_{dep} can be expressed by means of a single collector efficiency η

$$k_{dep} = \frac{\eta\varepsilon V}{4} = \frac{\alpha\eta_0\varepsilon V}{4}, \tag{7}$$

where V is the fluid advection velocity, ε is porosity and η_0 is the favorable single collector removal efficiency.

2.4 Dynamic Blocking and Release Functions $B(\theta)$, $R(\theta)$

The dynamic blocking functions characterize the particle deposition [4]. When the collector is particle free at the beginning, blocking function has value $B(\theta) = 1$. As the deposited particles block the surface more and more, $B(\theta)$ decreases. At the maximum attainable surface coverage $\theta = \theta_{max}$ (jamming limit), $B(\theta) = 0$.

2.4.1 RSA Dynamic Blocking Function

For colloidal particles depositing on the oppositely charged collector surface, these conditions for use of RSA model are valid [4]:

- attachment is irreversible as long as conditions do not change
- surface diffusion is negligible
- particle-particle contact is prohibited

For low and moderate surface coverage, the function $B(\theta)$ has this form

$$B(\theta) = 1 - 4\theta_\infty\frac{\theta}{\theta_{max}} + \frac{6\sqrt{3}}{\pi}\left(\theta_\infty\frac{\theta}{\theta_{max}}\right)^2 + \left(\frac{40}{\sqrt{3}\pi} - \frac{176}{3\pi^2}\right)\left(\theta_\infty\frac{\theta}{\theta_{max}}\right)^3,$$

where θ_∞ is the hard sphere jamming limit.

2.4.2 Dynamic Release Function

The dynamic release function describes the probability of colloid release from the porous media surface covered by retained colloids [1]. This function should in general depend on the colloid residence time and the retained colloid concentration. Because the colloid release is not well understood, we will use $R(\theta) = 1$.

3 Mathematical Model

This section shows solved equations, initial and boundary conditions. By substituting equations describing the colloid deposition and release (4), (5) and (6) into (3), we obtain the following expression

$$\frac{\partial n}{\partial t} = \nabla \cdot (D\nabla n) - \nabla \cdot (\mathbf{V} \cdot n) - \frac{f}{\pi a_p^2}((\lambda \pi a_p^2 k_{dep,f} B(\theta_f) + (1-\lambda)\pi a_p^2 k_{dep,u} B(\theta_u))n - ((\lambda \pi k_{det,f}\theta_f R(\theta_f) + (1-\lambda)k_{dep,u}\theta_u R(\theta_u)). \quad (8)$$

We assume that $K(\theta) = 1$ (first-order kinetics release mechanism) and use the following notations

$$\begin{aligned} \gamma &= \tfrac{f}{\pi a_p^2}, \\ K_a(\theta_f, \theta_u) &= \pi a_p^2[\lambda k_{dep,f} B(\theta_f) + (1-\lambda)k_{dep,u} B(\theta_u)], \\ K_r(\theta_f, \theta_u) &= \lambda \pi k_{det,f}\theta_f + (1-\lambda)k_{dep,u}\theta_u. \end{aligned} \quad (9)$$

Under these assumptions, the following equation is obtained

$$\frac{\partial n}{\partial t} = \nabla \cdot (D\nabla n) - \nabla \cdot (\mathbf{V} \cdot n) - \frac{K_a(\theta_f, \theta_u)}{\gamma} n + \frac{K_r(\theta_f, \theta_u)}{\gamma}. \quad (10)$$

In (10), V is a known velocity field given by a flow model. We complete this equation with (5) and (6)

$$\frac{\partial \theta_f}{\partial t} = \pi a_p^2 k_{dep,f} n B(\theta_f) - k_{det,f}\theta_f, \quad (11)$$

$$\frac{\partial \theta_u}{\partial t} = \pi a_p^2 k_{dep,u} n B(\theta_u) - k_{det,u}\theta_u. \quad (12)$$

To solve this system, we will need boundary conditions for equation (10) and initial conditions for each equation (10), (11) and (12). Let us consider a rectangular domain Ω with boundary Γ, where lower boundary is denoted Γ_1, right Γ_2, upper Γ_3 and left Γ_4 (Fig. 1).

For concentration equation (10), we will prescribe an initial condition

$$n(\mathbf{x}, 0) = n_0(\mathbf{x}) \text{ for } \mathbf{x} \in \Omega, \quad (13)$$

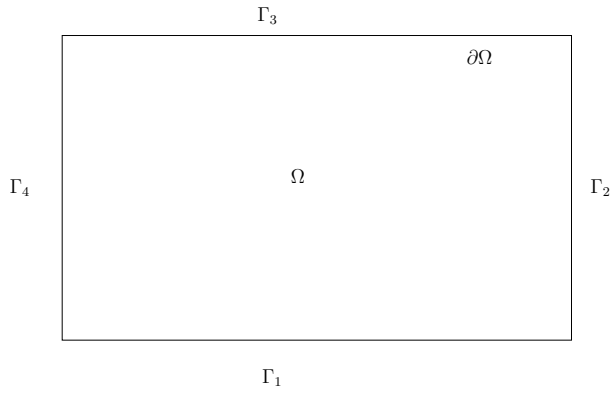


Figure 1: The domain Ω.

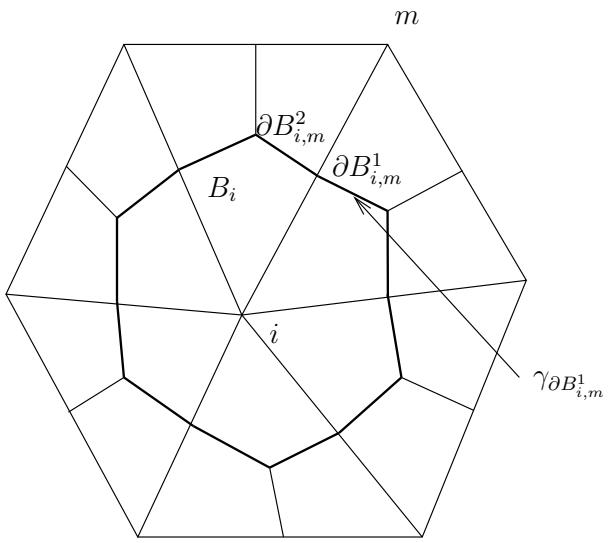


Figure 2: The exclusive subdomain for node i.

and boundary conditions describing concentration of colloids on Γ

$$n(\mathbf{x},t) = n_i(\mathbf{x},t) \text{ for } \mathbf{x} \in \Gamma_i,\ i \in 1,\ldots,4\ . \tag{14}$$

For equations (11) and (12) we need to prescribe initial conditions for θ_f and θ_u. As there are initially no deposited colloids,

$$\theta_f(\mathbf{x},0) = \theta_u(\mathbf{x},0) = 0 \text{ for } \mathbf{x} \in \Omega. \tag{15}$$

4 Numerical Solution

We discuss the discretization methods for solving (10), (11) and (12). Although our numerical solution is computed on a rectangular grid, we develop the scheme for the transport problem for a more general case of an unstructured mesh in two dimensions composed of triangles and quadrangles of the domain Ω, which is called the primary mesh. We construct a dual mesh by connecting barycentres of each element with midpoints of all its sides in each element from the primary grid. In this way we obtain a polygon around each node from the primary mesh (on the boundary of the domain $\partial\Omega$, polygons are incomplete). For a primary mesh node i, we call this polygon B_i, the exclusive subdomain of node i. ∂B_i consists of several abscissae and each of abscissa belongs to one abscissa connecting node i with his neighbor m. For each couple i, m, there are two abscissae, we denote them $\partial B^l_{i,m}$. The midpoint of the abscissa $\partial B^l_{i,m}$ is denoted $\gamma_{\partial B^l_{i,m}}$ (Fig. 2). The time level is denoted by superscript k. The length of abcissa $\partial B^l_{i,m}$ is denoted $|\partial B^l_{i,m}|$. The same numerical grid is used for solving the flow field and the transport equation.

4.1 The flow field

The flow problem (1) is discretized using the finite volume method. To solve the linear equations system in the scheme is used the PETSc library for programming language C.

4.2 The transport equation

Explicit scheme has the disadvantage that the time steps has to be limited due to CFL condition [6]. For this reason we implemented semi-implicit numerical scheme [5], which will enable us to use larger time steps compared to the explicit scheme.

Equation (10)

$$\frac{\partial n}{\partial t} = \nabla \cdot (D \nabla n) - \nabla \cdot (\mathbf{V} \cdot n) - \frac{K_a(\theta_f, \theta_u)}{\gamma} n + \frac{K_r(\theta_f, \theta_u)}{\gamma} \tag{16}$$

is solved using the operator splitting technique. At first we solve explicitly convection and reaction parts of the equation

$$\frac{\partial n}{\partial t} = -\nabla \cdot (\mathbf{V} \cdot n) - \frac{K_a(\theta_f, \theta_u)}{\gamma} n + \frac{K_r(\theta_f, \theta_u)}{\gamma} \tag{17}$$

obtained from (10) by setting $D = 0$. We discretize (17) as follows

$$\left[\frac{n_i^{k+\frac{1}{2}} - n_i^k}{\Delta t} + \frac{K_a(\theta_{f,i}^k, \theta_{u,i}^k)}{\gamma} n_i^k - \frac{K_r(\theta_{f,i}^k, \theta_{u,i}^k)}{\gamma} \right] |B_i| + \sum_{m,l} \int_{\partial B_{i,m}^l} (\mathbf{V}(\gamma_{\partial B_{i,m}^l}) \cdot n_{i,m,l}^\star) \cdot \mathbf{n}_{\partial B_{i,m}^l} |\partial B_{i,m}^l| = 0, \tag{18}$$

where the upwind value is given as

$$n_{i,m,l}^\star = \begin{cases} n_i^k & \text{for} \quad \mathbf{n}_{\partial B_{i,m}^l} \cdot \mathbf{V}(\gamma_{\partial B_{i,m}^l}) > 0, \\ n_m^k & \text{for} \quad \mathbf{n}_{\partial B_{i,m}^l} \cdot \mathbf{V}(\gamma_{\partial B_{i,m}^l}) \leq 0. \end{cases} \tag{19}$$

The value of $n_i^{k+\frac{1}{2}}$ is used as an initial condition and (14) as boundary conditions for solving the diffusion equation

$$\frac{\partial n}{\partial t} = \nabla \cdot (D \nabla n), \tag{20}$$

which is solved using the backward Euler scheme

$$\left[\frac{n_i^{k+1} - n_i^{k+\frac{1}{2}}}{\Delta t} \right] |B_i| = \sum_{m,l} \left[(D(\gamma_{\partial B_{i,m}^l})(\nabla n)^{k+1}(\gamma_{\partial B_{i,m}^l})) \cdot \mathbf{n}_{\partial B_{i,m}^l} |\partial B_{i,m}^l| \right]. \tag{21}$$

We denote number of nodes in one row of our numerical grid n_r. On a rectangular grid with grid sizes Δx, Δy, equation (21) reads as

$$\left[\frac{n_i^{k+1}-n_i^{k+\frac{1}{2}}}{\Delta t}\right]|B_i| - \Delta x D_{yy}(\gamma_{\partial B_{i,i+n_r}})\left(\frac{n_{i+n_r}^{k+1}-n_i^{k+1}}{\Delta y}\right) +$$
$$\Delta y D_{xx}(\gamma_{\partial B_{i,i-1}})\left(\frac{n_i^{k+1}-n_{i-1}^{k+1}}{\Delta x}\right) + \Delta x D_{yy}(\gamma_{\partial B_{i,i-n_r}})\left(\frac{n_i^{k+1}-n_{i-n_r}^{k+1}}{\Delta y}\right) -$$
$$\Delta y D_{xx}(\gamma_{\partial B_{i,i+1}})\left(\frac{n_{i+1}^{k+1}-n_i^{k+1}}{\Delta x}\right) = 0. \qquad (22)$$

In equation (22) the terms containing boundary values can be eliminated into the right hand side. In every time step we need to solve the system $An^{k+1} = b$, where

$$\begin{aligned}
A_{i,i-n_r} &= -\frac{\Delta x D_{yy}(\gamma_{\partial B_{i,i-n_r}})}{\Delta y} \\
A_{i,i-1} &= -\frac{\Delta y D_{xx}(\gamma_{\partial B_{i,i-1}})}{\Delta x} \\
A_{i,i} &= \frac{|B_i|}{\Delta t} + \frac{\Delta x D_{yy}(\gamma_{\partial B_{i,i+n_r}})}{\Delta y} + \frac{\Delta y D_{xx}(\gamma_{\partial B_{i,i-1}})}{\Delta x} + \frac{\Delta x D_{yy}(\gamma_{\partial B_{i,i-n_r}})}{\Delta y} + \frac{\Delta y D_{xx}(\gamma_{\partial B_{i,i+1}})}{\Delta x} \\
A_{i,i+1} &= -\frac{\Delta y D_{xx}(\gamma_{\partial B_{i,i+1}})}{\Delta x} \\
A_{i,i+n_r} &= -\frac{\Delta x D_{yy}(\gamma_{\partial B_{i,i+n_r}})}{\Delta y}
\end{aligned}$$

and $A_{i,j} = 0$ elsewhere. The right hand side of the solved system b reads as

$$b_i = \frac{|B_i|}{\Delta t} n_i^k, \qquad (23)$$

where index i goes through all nodes. The boundary terms can be eliminated into the b_i.

5 Results

In this section we present results describing transport of colloids in a heterogeneous porous media. We are given a square domain Ω of size 3×3m with two heterogenities. The first one is a square $[0.5, 0.75] \times [1.25, 1.5]$ with higher λ and k_{dep} and affects only the transport of colloids. The second heterogeneity is a circle in the middle of Ω with diameter of 1 where is ten times smaller permeability then in the rest of Ω.

For pressure we have no flow boundary conditions on Γ_1 and Γ_3 and Dirichlet's boundary conditions (hydrostatic pressure) $p(\mathbf{x}) = 10.0^5 - y\varrho g$ on Γ_2 and $p(\mathbf{x}) = 1.1 \cdot 10^5 - y\varrho g$ on Γ_4. In the beginning no colloidal particles are present in the area. We prescribe a boundary condition

$$n(\mathbf{x}, t) = \begin{cases} 10^{14}[m^{-3}] & \text{for} \quad t \leq 0.5\text{day} \\ 0 & \text{for} \quad 1.0 \geq t > 0.5\text{day} \end{cases} \qquad (24)$$

for $y \in [1, 2]$ and 0 elsewhere on Γ_4 and $n(\mathbf{x}, t) = 0$ for $x \in \Gamma_1,\ \Gamma_2,\ \Gamma_3$ and $t \in [0, 1\text{day}]$ We are interested in the distribution of colloids in domain Ω in time of one day. Our results are showing the number concentration of colloidal particles contained in water in pores n divided by 10^{14}, so that the resulting values are rescaled between 0 and 1. The

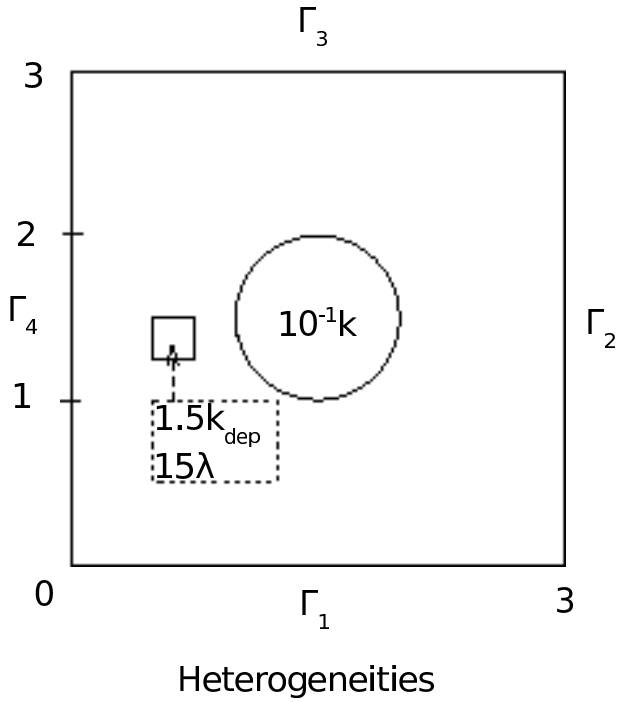


Figure 3: Compuatation domain.

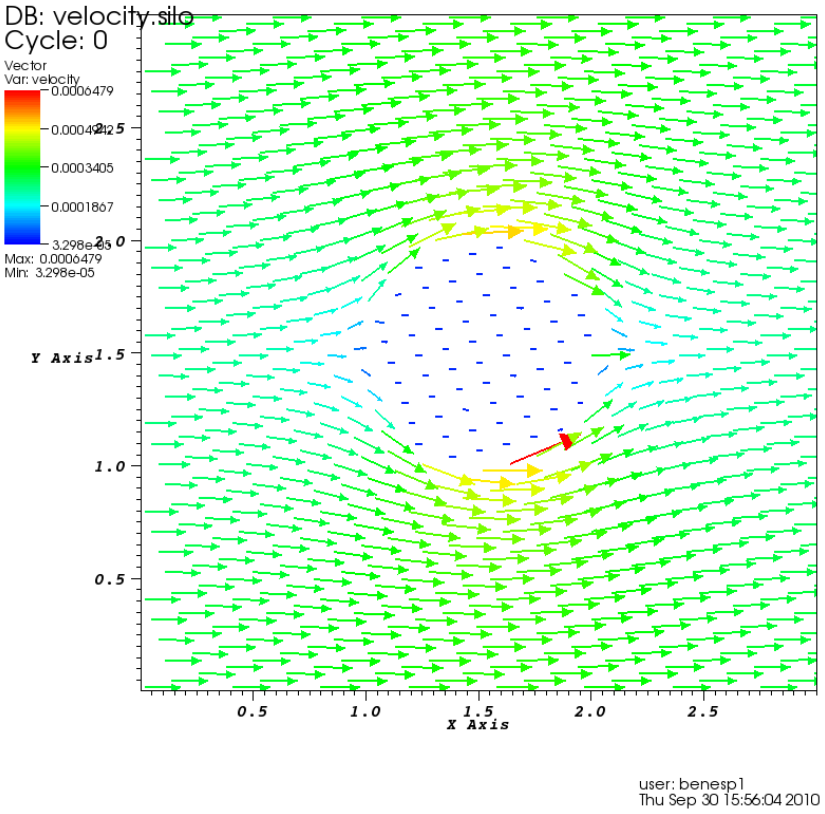


Figure 4: The flow field.

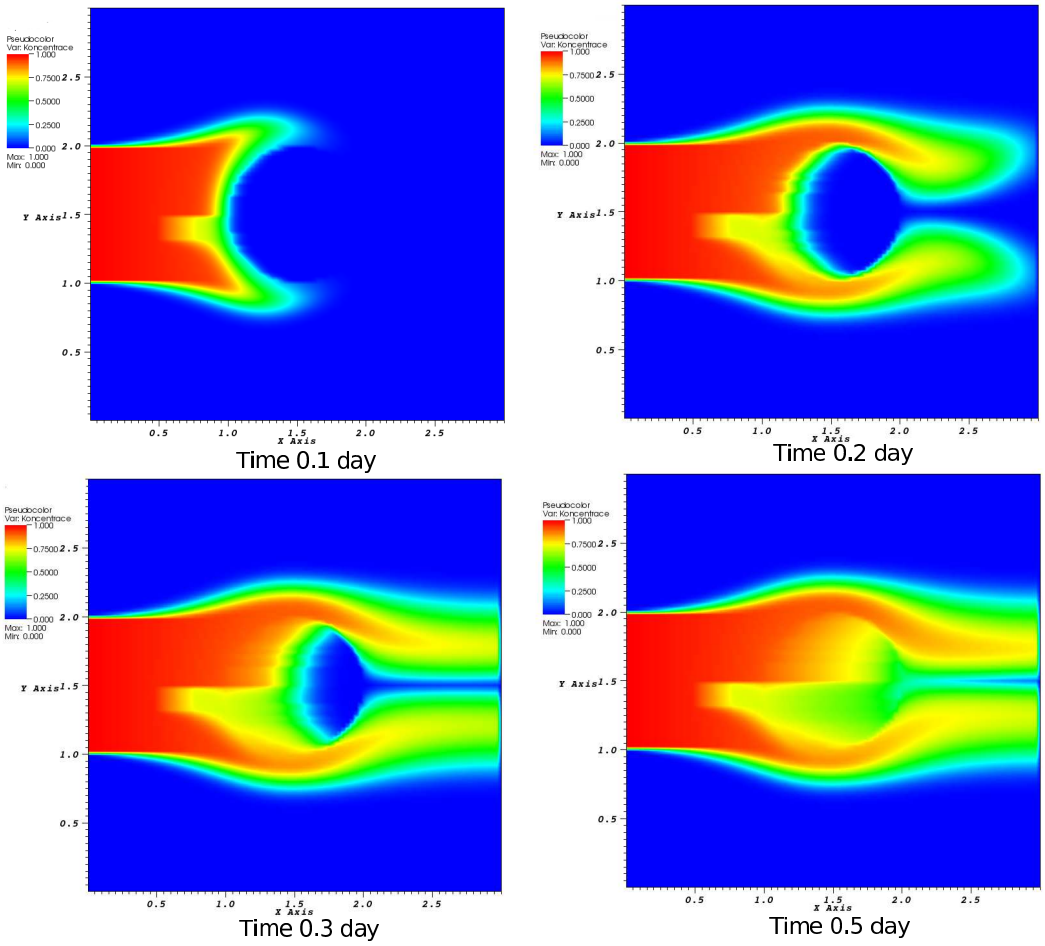


Figure 5: Evolution of the number concentration n divided by 10^{14} in time; semi-implicit scheme.

rectangular numerical grid with 100×100 nodes was used for computations. Results computed by the semi-implicit numerical scheme are shown in Figure 5.

Figure 4 depicts the flow field. The heterogeneity in permeability made most of water to flow around the circle lower permeability heterogeneity in the middle of the area. Second figure 5 shows the time evolution of normed concentration of colloids during transport with higher deposition rate in the square heterogeneity with higher deposition constants. Transport is of course influenced by means of the permeability heterogeneity due to flow field.

6 Conclusion

In this contribution a summary of equations describing the colloid transport was presented. The equations were discretized by means of the semi-implicit scheme based on the operator splitting technique using first order upwind (19) for approximation of the convection term. Numerical results show the expected behavior. Most of the colloids go around low permeability heterogeneity and on the heterogeneity with higher colloid deposition quantities colloids deposits with much higher rate.

7 Acknowledgment

This work was supported by the project "Mathematical Modeling of Multi-Phase Porous Media Flow", project No. 201/08/P567 of the Grant Agency of the Czech Republic, 2008-2010 (principal investigator J. Mikyška)

References

[1] N. Sun, M. Elimelech, N.-Z. Sun *A novel two-dimensional model for colloid transport in physically and geochemically heterogeneous porous media.* Journal of Contaminant Hydrology 49, (2001), 173–199.

[2] N.-Z. Sun *Mathematical Modeling of Groundwater Pollution.* Springer-Verlag, New York.

[3] N.-Z. Sun W.W.-G, Yeh *A proposed upstream weight numerical method for simulating pollutant transport in groundwater.* Water Resour. Res. 19 (1983), 1489–1500.

[4] J.N. Ryan, M. Elimelech *Review Colloid mobilization and transport in groundwater.* Colloids and Surfaces A: Physicochemical and Engineering Aspects 107 (1996), 1–56.

[5] R.J. LeVeque, J. Oliger *Numerical methods based on additive splittings for hyperbolic partial differential equations.* Math. Comp. 40 162 (1983), 469–497.

[6] R.J. LeVeque *Finite-Volume Methods for Hyperbolic Problems.* Cambridge Press, (2002)

Robustified total least squares

Jiří Franc

1st year of PGS, email: francji1@fjfi.cvut.cz
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Jan Ámos Víšek, Department of Macroeconomics and Econometrics,
Faculty of Social Sciences, Institute of Economic Studies, Charles University

Abstract. Classical regression estimators, such as the ordinary least squares (LS), are sensitive to occurrence of outliers and are not consistent when the orthogonality condition fails. There have been several robust estimators that can cope with this problem. The development of instrumental weighted variables (IWV), the robust version of instrumental variables methods, is reviewed. The alternative approach in regression methods when orthogonality condition is breaking and both independent and dependent variables are considered to be measured with errors is called total least squares. The existence and uniqueness of the solution is discussed and different approaches of calculation are described. The robustified version of TLS based on the idea of downweighting the influential points is presented and its properties are discussed. Finally the generalization of TLS to mixed LS-TLS and its robustified version is mentioned.

Keywords: robust regression analysis, instrumental weighted variables, robustified total least squares

Abstrakt. Klasické regresní odhady, jako metoda nejmenších čtverců, jsou citlivé na výskyt odlehlých pozorování a nejsou konzistentní když nezávislé proměnné jsou měřeny s chybou. Jedna z možností jak se s tímto případem vypořádat je použít metodu instrumentálnách vážených proměnných. Další možností je robustifikovat totální nemenší čtverce (TLS). Základní vlastnosti, existence a jednoznačnost řešení i způsoby výpočtu klasických totálních čtverců jsou diskutovány. Dále jsou představeny možnosti robustifikace TLS pomocí penalizace vlivných bodů. Na závěr je uvedeno zobecnění TLS na metodu smíšených nejmenších čtverců-totálních nejmenších čtverců a robustifikace této metody.

Klíčová slova: robustní regresní analýza, instrumentální vážené proměnné, robustifikované totální nejmenší čtverce

1 Introduction

Let us consider the multiple linear regression model

$$Y_i = X_{i,1}\beta_1^0 + X_{i,2}\beta_2^0 + \cdots + X_{i,p}\beta_p^0 - \varepsilon_i = X_i^T\beta^0 - \varepsilon_i \qquad i = 1 \dots n,$$

or in the matrix notation

$$\mathbf{Y} = \mathbf{X}\beta^0 - \varepsilon,$$

where $\mathbf{Y} \in \mathbb{R}^{n\times 1}$ is a vector of response (dependent) variable, $\mathbf{X} \in \mathbb{R}^{n\times p}$ is a matrix of predictors (independent variables), $\beta^0 \in \mathbb{R}^{p\times 1}$ is a vector of unknown regression coefficients

and $\varepsilon \in \mathbb{R}^{n\times 1}$ a vector of unknow error terms (vector of disturbances). The objective is to estimate the unknown regression coefficients and express the dependent variable as a linear function of the independent variables. We assume that $\left\{\left(X_i^T, \varepsilon_i\right)^T\right\}_{i=1}^{+\infty}$ is a sequence of *iid* $(p+1)$-dimensional random variables with a absolutely continuous distribution function, but the explanatory variables X_i's can be correlated with the error terms ε_i's. Our goal is to find a robust estimator of the regression coefficients, which resists well to a violation of theoretical conditions and to a contamination of the data. Many regression estimators such as the ordinary least squares are not robust, i.e. they are very sensitive to outliers. One of the best known and the most used robust estimator is the least trimmed squares (LTS), which was proposed by Rousseeuw in 1984 (see [3]). It minimizes the sum of the h smallest squared residuals, where the j-th residual is defined as $r_j(\beta) = r_j = Y_j - X_j^T\beta$. The TLS estimator is defined as follows

$$\hat{\beta}^{(LTS,h,n)} = \arg\min_{\beta\in\mathbb{R}^p} \sum_{i=1}^{h} r_{(i)}^2(\beta),$$

where h is an optional parameter satisfying $\frac{n}{2} \leq h \leq n$ and $r_{(i)}^2$ is the i-th least squared residual, i.e. for any $\beta \in \mathbb{R}^p$

$$r_{(1)}^2(\beta) \leq r_{(2)}^2(\beta) \leq \ldots \leq r_{(n)}^2(\beta).$$

For $h = \left\lfloor \frac{n}{2} \right\rfloor + \left\lfloor \frac{p+1}{2} \right\rfloor$ the LTS reaches the maximum possible value of the breakdown point equal to $\frac{n-p+2}{2n}$. The existence of the LTS estimator is given by the existence of the LS estimator for all subsamples of size h. While it is very difficult to find exact value of LTS estimate for larger sets of observations, the approximative algorithms are usually used. The oldest one defined by Rousseeuw does the approximation in the following way. Let select randomly an $(p+1)$ observations and apply the least squares method on them. For the estimated parameter β evaluate residuals for all n observations. Then select h observations with the smallest squared residuals and compute again the least squares estimation. Repeat last two steps until convergence. Repeat the procedure number of times with different initial estimate to get more candidates. The candidate which has the smallest value of the objective function (sum of squared residuals) is taken as the LTS estimate. Due to the high computational complexity of the LTS estimator there have been appeared a lot of algorithms in the literature for last years, but the mentioned one is sufficient for our purpose. One of the main disadvantages of the LTS is its infinite local sensitivity, because the "change of weights" between trimmed and not trimmed points is too sharp. We can cope with this problem by the definition of some continuous weighting function and multiply the residuals by a weights from $\langle 0, 1\rangle$. This is exactly the way how the least weighted squares estimator is defined. According to [4]

$$\hat{\beta}^{(LWS,w,n)} = \arg\min_{\beta\in\mathbb{R}^p} \sum_{i=1}^{n} w_i r_{(i)}^2(\beta) = \arg\min_{\beta\in\mathbb{R}^p} \sum_{i=1}^{n} w\left(\frac{i-1}{n}\right) r_{(i)}^2(\beta),$$

where weights w_i are defined by the weight function $w : \langle 0, 1\rangle \to \langle 0, 1\rangle$, which is absolutely continuous, $w(0) = 1$ and non-increasing with the derivative $w'(t)$ bounded from below

by a constant $(-L)$, where $L \geq 0$. Furthermore we can rewrite the previous definition of LWS and instead of ordering the residuals we can reorder the weights ($w_i := w_i(\beta)$). For any $i \in \{1, \ldots, n\}$ let us denote by $\pi(\beta, i)$ the random rank of the i-th residual as

$$\pi(\beta, i) = j \in \{1, \ldots, n\} \qquad \Leftrightarrow \qquad r_i^2(\beta) = r_{(j)}^2(\beta)$$

Then we have

$$\hat{\beta}^{(LWS,w,n)} = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} w\left(\frac{\pi(\beta, i) - 1}{n}\right) r_i^2(\beta)$$

and the least weighted squares are solution to the normal equations

$$\mathrm{NE}_{Y,X,n}(\beta) = \sum_{i=1}^{n} w\left(\frac{\pi(\beta, i) - 1}{n}\right) X_i \left(Y_i - X_i^T \beta\right) = 0. \tag{1}$$

The solution to LWS estimate exists, because it is equal to the solution to classical weighted least squares, where weights have one certain permutation. Consequently our problem, how to find the least weighted squares estimator, is equal to the problem, how to find "the best" weighted least squares among $n!$ possibilities. Since n is not usually "small" number, we can not find a deterministic solution to this extremal problem and we need to use some approximative algorithm again. One of the most simplest, but sufficient one, is based on the same procedure that we described in LTS section.

In econometrics, the explanatory variables are frequently assumed to be correlated with the random error ε, that is, $\mathbf{X}$ is supposed to be correlated with ε such that $p\lim\left(\frac{1}{n}\mathbf{X}^T\varepsilon\right) \neq \mathbf{0}$. If we now apply LS, LTS or LWS estimators, we get an inconsistent estimate of β^0. One of the best known example of the situation, when the orthogonality condition fails (i.e. $E\left[X_i\varepsilon_i\right] \neq 0$), is the model in which the explanatory variables are measured with a random error. We consider an overdetermined set of n linear equations $\mathbf{Y} \approx \mathbf{X}\beta$, where $\mathbf{Y} \in \mathbb{R}^{n\times 1}$, $\mathbf{X} \in \mathbb{R}^{n\times p}$ and $\beta \in \mathbb{R}^{p\times 1}$ is a parameter of interest. We suppose that

$$Y_i = Y_{0i} - \varepsilon_i \qquad X_i = X_{0i} - \theta_i$$

and that there exists $\beta^0 \in \mathbb{R}^{p\times 1}$ such that

$$Y_{0i} = X_{0i}\beta^0,$$

i.e.

$$Y_i + \varepsilon_i = (X_i + \theta_i)\beta^0 \qquad i = 1 \ldots n. \tag{2}$$

Assuming usually that $\mathbb{E}\left[\varepsilon_i\right] = 0$, $\mathbb{E}\left[\varepsilon_i^2\right] = \sigma^2 \in (0, \infty)$ and $\mathbb{E}\left[\theta_i\right] = 0$, $\mathbb{E}\left[\theta_i\theta_i^T\right] = \Sigma_\theta$ nonsingular and $\mathbb{E}\left[\theta_i\varepsilon_i\right] = 0$. If we consider now classical regression model

$$Y_i = X_{0i}\beta^0 - \varepsilon_i = (X_i + \theta_i)\beta^0 - \varepsilon_i = X_i\beta^0 + \theta_i\beta^0 - \varepsilon_i = X_i\beta^0 + e_i,$$

we can easily find out that orthogonality condition is broken.

$$\mathbb{E}\left[X_i e_i\right] = \mathbb{E}\left[(X_{0i} - \theta_i) \cdot \left(\theta_i\beta^0 - \varepsilon_i\right)\right] = -\Sigma_\theta\beta^0.$$

If $\beta^0 \neq 0$ then $\Sigma_\theta\beta^0 \neq 0$ and not only the LS estimate of the regression coefficients is inconsistent. In the following sections we describe two possibilities how to cope with such a cases when the orthogonality condition is broken and the data set contains outliers.

2 Instrumental Weighted Variables

Suppose there is an $n \times p$ matrix of some variables $\mathbf{Z}$, called instruments. If instrumental variables are uncorrelated with errors and the matrix of correlations between the variables in $\mathbf{X}$ and the variables in $\mathbf{Z}$ is of maximum possible rank (equal to p) then we call the instruments proper. Let $\{Z_i\}_{i=1}^n$ be any sequence of p-dimensional proper instrumental variables, then the instrumental variable estimator $\hat{\beta}^{(IV,n)}$ of β^0 is defined by

$$\hat{\beta}^{(IV,n)} = \left(\mathbf{Z}^T\mathbf{X}\right)^{-1}\mathbf{Z}^T\mathbf{Y}.$$

We can easily obtain "classical" weighted instrumental variables by adding weights into the definition of instrumental variables

$$\hat{\beta}^{(WIV,n,W)} = \left(\mathbf{Z}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{Z}^T\mathbf{W}\mathbf{Y},$$

where $\mathbf{W}$ is in our case an $n \times n$ diagonal matrix of weights. For weighted instrumental variables the weights are assigned to the observation a priori, according to an external rule or some previous knowledge of the problem. The weighted instrumental variables (WIV) estimation $\hat{\beta}^{(WIV,n,W)}$ can by also defined as the solution to the following normal equations

$$\mathbf{Z}^T\mathbf{W}\left(\mathbf{Y} - \mathbf{X}\beta\right) = \mathbf{0}.$$

Since $\hat{\beta}^{(WIV,n,W)}$ is not robust with respect to outliers and leverage points, we are going to use the idea of implicit weighting of the squared residuals from the LWS and define the robust version of the instrumental variables estimator.

Let $\{Z_i\}_{i=1}^{+\infty}$ be any sequence of p-dimensional proper instrumental variables. Then the solution of the normal equations

$$NE_{Y,X,Z,n}(\beta) = \sum_{i=1}^{n} w\left(\frac{\pi(\beta,i)-1}{n}\right) Z_i\left(Y_i - X_i^T\beta\right) = 0 \tag{3}$$

will be called the instrumental weighted variables estimation (IWV) of β^0 and denoted by $\hat{\beta}^{(IWV,n)}$. As we described in previous text the relation between classical weighted least squares and the robust LWS, we can find similar relation for classical WIV and newly defined IWV. Suppose we have some weighting function w and set of proper instruments $\mathbf{Z}$. Then we can permutate the set of weights

$$\left\{w\left(\frac{1-1}{n}\right), w\left(\frac{2-1}{n}\right), \ldots, w\left(\frac{n-1}{n}\right)\right\}$$

according to some permutation $\pi \in \mathcal{P}$, where $\mathcal{P}$ denotes the set of all permutations of the indices $\{1, 2, \ldots, n\}$ for any $n \in \mathbb{N}$. Let compute for all $\pi \in \mathcal{P}_n$

$$\hat{\beta}^{(WIV,n,W(\pi))} = (\mathbf{Z}^T\mathbf{W}(\pi)\mathbf{X})^{-1}\mathbf{Z}^T\mathbf{W}(\pi)\mathbf{Y}.$$

Recall that weights in instrumental weighted variables satisfy the condition that the smallest residual obtains the largest weight, the second smallest residual obtains the

second largest weight, etc. till the largest residual obtains the smallest weight. Hence, we can minimize again the sum of weighted residuals. We have $n!$ possibilities how to permutate weights and hence $n!$ solutions of classical weighted instrumental variables. We compute the classical WIV for all $\pi \in \mathcal{P}$ and find the certain permutation $\pi_{best} \in \mathcal{P}$ which minimizes the the sum of weighted residuals among all $\hat{\beta}^{(WIV,n,W(\pi))}$. Then IWV estimator can be defined as

$$\hat{\beta}^{(IWV,n)} = \hat{\beta}^{(WIV,n,W(\pi_{best}))} = (\mathbf{Z}^T\mathbf{W}(\pi_{best})\mathbf{X})^{-1}\mathbf{Z}^T\mathbf{W}(\pi_{best})\mathbf{Y}.$$

In [5] is shown that all solutions of the corresponding normal equations (3) are bounded in probability and the weak consistency of the IWV is proved. Another technical approach of the proof of consistency is in [2], but it still using the idea of bounding the solutions of (3) with some probability and the strengthened Glivenko-Cantelli theorem $\sup_{\beta\in R^p} \sup_{r\in R} \sqrt{n} \left| F_\beta^{(n)}(r) - F_\beta(r) \right| = O_p(1)$. In [2], among others, are described some approximative algorithms that compute the IWV estimator $\hat{\beta}^{(IWV,n)}$ of a given linear regression problem. The first type of algorithms is based on the idea of iterative reweighting which was described in the introduction section. The $(j+1)$th iteration of the IWV estimator is obtained as:

$$\hat{\beta}_{(j+1)}^{\left(IWV,n,\mathbf{W}\left(\hat{\beta}_{(j)}^{(IWV,n)}\right)\right)} = (\mathbf{Z}^T\mathbf{W}\left(\hat{\beta}_{(j)}^{(IWV,n)}\right)\mathbf{X})^{-1}\mathbf{Z}^T\mathbf{W}\left(\hat{\beta}_{(j)}^{(IWV,n)}\right)\mathbf{Y},$$

where as the initial estimate $\hat{\beta}_{(0)}^{(IWV,n)}$ we can consider the simple LS estimator of $(p+1)$ randomly picked different observations and

$$W(\beta) = \operatorname{diag}\{w_1, w_2, \ldots, w_n\} \quad \text{s} \quad w_i = w\left(\frac{\pi(\beta,i)-1}{n}\right).$$

The second type of algorithms is based on on theory of simulated annealing and use Metropolis-Hastings algorithm for Markov Chain - Monte Carlo and anther one use genetic algorithms. All of them have been tested for several different simulations and verified not only on data sets where outliers and leverage points occur but also on data sets where regressors are correlated with error terms. The first one is much faster and for larger dataset ($n > 50$) gives significantly better results, but two remaining probabilistic algorithms give still sufficiently good estimation compared with LS, IV or LWS.

The disadvantage of the IWV method is that the credibility of the estimates hinges on the selection of suitable instruments. To find such instrumental variables that are not correlated with the error terms and that are highly correlated with the explanatory variables can be hard. That is why we introduce another approach that is much better especially in such a cases when both response variable and explanatory variables are measured with a random error (the model (2)). This model is sometimes called errors-in-variables model and the approach how to estimate unknown parameter β^0 is known as orthogonal regression or total least squares.

3 Total Least Squares

The total least squares method is viewed as a tool for deriving approximate linear models and its systematic investigation was started by Golub and Van Loan paper in 1980 [6]. Assume again the overdetermined system of n linear equations

$$\mathbf{Y} \approx \mathbf{X}\beta, \quad \mathbf{Y} \in \mathbb{R}^n,\ \mathbf{X} \in \mathbb{R}^{n\times p},\ n > p,$$

$$\mathbf{Y} \approx \mathbf{X}\beta \Leftrightarrow \mathbf{X}\beta \approx \mathbf{Y}|_{\mathcal{R}(\mathbf{X})} + \mathbf{Y}|_{\mathcal{N}(\mathbf{X}^T)}\,,$$

$$\mathbf{X}\beta = \mathbf{Y}|_{\mathcal{R}(\mathbf{X})}\,,\ \mathbb{R}^p = \mathcal{N}(\mathbf{X}) \oplus \mathcal{R}(\mathbf{X}^T) \Rightarrow \beta \in \mathcal{R}(\mathbf{X}^T).$$

Since the exact solution need not exist, we try to find some approximation, which is best in some sense. The idea is to modify all data points in such a way that some norm of the modification is minimized subject to the constraint that the modified vectors satisfy some linear relation. There are many possible way how to define the approximation, but the most frequent ones are ordinary least squares, data least squares and total least squares approach. Given an overdetermined set of n linear equations $\mathbf{Y} \approx \mathbf{X}\beta$ in p unknowns β then

- the ordinary least squares problem seeks to

$$\hat{\beta}^{(OLS,n)} = \min_{\beta\in\mathbb{R}^p, \varepsilon\in\mathbb{R}^n} \|\varepsilon\|_2 \quad \text{subject to} \quad \mathbf{Y} + \varepsilon = \mathbf{X}\beta. \tag{4}$$

  $\hat{\beta}^{(OLS,n)}$ is called a OLS solution to the problem (4) and ε is called the corresponding OLS correction.

- the data least squares problem seeks to

$$\hat{\beta}^{(DLS,n)} = \min_{\beta\in\mathbb{R}^p, \mathbf{\Theta}\in\mathbb{R}^{n\times(p)}} \|\mathbf{\Theta}\|_F \quad \text{subject to} \quad \mathbf{Y} = (\mathbf{X} + \mathbf{\Theta})\beta. \tag{5}$$

  $\hat{\beta}^{(DLS,n)}$ is called a DLS solution to the problem (5) and $\mathbf{\Theta}$ is called the corresponding DLS correction.

- the total least squares problem seeks to

$$\hat{\beta}^{(TLS,n)} = \min_{\beta\in\mathbb{R}^p, [\varepsilon,\mathbf{\Theta}]\in\mathbb{R}^{n\times(p+1)}} \|[\varepsilon, \mathbf{\Theta}]\|_F \quad \text{subject to} \quad \mathbf{Y} + \varepsilon = (\mathbf{X} + \mathbf{\Theta})\beta. \tag{6}$$

  $\hat{\beta}^{(TLS,n)}$ is called a TLS solution to the problem (6) and $[\varepsilon, \mathbf{\Theta}]$ is called the corresponding TLS correction.

The suitable norm used in previous definitions of the DLS and the TLS problem is called the Frobenius norm and for the matrix $\mathbf{X}$ is defined as follows

$$\|\mathbf{X}\|_F = \sqrt{\sum_{i=1}^{n}\sum_{i=1}^{p} x_{ij}^2} = \sqrt{\text{trace}(\mathbf{X}^T\mathbf{X})} = \sqrt{\sum_{i=1}^{\min\{n,p\}} \sigma_i^2} = \sqrt{\sum_{i=1}^{\text{rank}(\mathbf{X})} \sigma_i^2}, \tag{7}$$

where σ_i's are the singular values of the matrix $\mathbf{X}$.

While in the OLS approach we assume that we know all data points from matrix $\mathbf{X}$ exactly and we measure with errors only the response variable $\mathbf{Y}$ in the TLS approach we assume that both the response variable and the predictors are perturbed. Let us consider an n-element point set $P \in \mathbb{R}^{p+1}$, whose ith point is $p_i = (X_{i,1}, \ldots, X_{i,p}, Y_i)^T$. We will denote by the term hyperplane a p-dimensional hyperplane, which is nonvertical (i.e. the last coordinate of its normal vector is nonzero). A model parameter vector β corresponds to a hyperplane, which we will denote by $\rho(\beta)$ or simply ρ. The residual $d_i(\rho)$ is defined to be the signed orthogonal distance from ρ to p_i. In this formulation the total least squares problem is equivalent to computing the hyperplane that minimizes the sum of the squared orthogonal distances, while OLS minimizes the sum of the squared vertical distances from the data points p_i to the fitting hyperplane ρ. The normal vector of the hyperplane ρ is $\nu = \left[\beta^T, -1\right]^T$. Then the formula for the orthogonal distance of point $p_i \in \mathbb{R}^{p+1}$ from the hyperplane ρ is

$$\frac{\left|\nu^T(A - p_i)\right|}{\|\nu\|},$$

where $A \in \rho$ is arbitrary point. Then we can formulate the total least square problem as

$$\begin{aligned}
\hat{\beta}^{(TLS,n)} &= \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \frac{\left|\nu^T(A - p_i)\right|^2}{\|\nu\|^2} = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \frac{\left|\left[\beta^T, -1\right] \begin{bmatrix} X_i \\ Y_i \end{bmatrix}\right|^2}{\|[\beta^T, -1]\|^2} \\
&= \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{1 + \|\beta\|^2} \sum_{i=1}^{n} |Y_i - X_i\beta|^2 = \arg\min_{\beta \in \mathbb{R}^p} \frac{\|\mathbf{Y} - \mathbf{X}\beta\|}{\sqrt{1 + \|\beta\|^2}}. \qquad (8)
\end{aligned}$$

Before we give conditions for uniqueness and existence of a TLS solution, we introduce some important tools such as singular value decomposition.

Singular Value Decomposition Theorem

The singular value decomposition (SVD) of the matrix $[\mathbf{X}, \mathbf{Y}] \in \mathbb{R}^{n \times (p+1)}$ is defined by

$$[\mathbf{X}, \mathbf{Y}] = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T, \qquad (9)$$

where $\mathbb{U} = [u_1, \ldots, u_n] \in \mathbb{R}^{n \times n}$ and $\mathbf{V} = [v_1, \ldots, v_{p+1}] \in \mathbb{R}^{(p+1) \times (p+1)}$ are orthonormal matrices that contain the left and the right singular vectors, respectively. $\Sigma = \operatorname{diag}\{\sigma_1, \ldots, \sigma_r\} \in \mathbb{R}^{r \times r}$, $r = \min\{n, p+1\}$, $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r \geq 0$ are the singular values of the matrix $[\mathbf{X}, \mathbf{Y}]$ in decreasing order of magnitude. The triplet (u_i, σ_i, v_i) is called a singular value triplet. If we assume that $rank([\mathbf{X}, \mathbf{Y}]) = r$ then the dyadic decomposition of the matrix $[\mathbf{X}, \mathbf{Y}]$ is following

$$[\mathbf{X}, \mathbf{Y}] = \sum_{i=1}^{r} \sigma_i u_i v_i^T \qquad (10)$$

and decompose the matrix $[\mathbf{X}, \mathbf{Y}]$ of rank r in a sum of r matrices of rank one. Note that numbers σ_i's are square roots of nonzero eigenvalues of the symmetric and nonnegative

definite matrices $[\mathbf{X},\mathbf{Y}]^T[\mathbf{X},\mathbf{Y}]$ and $[\mathbf{X},\mathbf{Y}][\mathbf{X},\mathbf{Y}]^T$ related to eigenvectors $\{u_1,\ldots,u_r\}$ and $\{v_1,\ldots,v_r\}$.
To solve the problem (6) with the TLS, bring the set into form $[\mathbf{X},\mathbf{Y}][\beta,-1]^T \approx 0$. We want to find „true" values $[\mathbf{X_0},\mathbf{Y_0}]$ such that $\|[\mathbf{X},\mathbf{Y}]-[\mathbf{X_0},\mathbf{Y_0}]\|_F$ is minimal and $[\mathbf{X_0},\mathbf{Y_0}][\beta,-1]^T = 0$ for some β. If $\sigma_{p+1} \neq 0$ then $\mathrm{rank}([\mathbf{X},\mathbf{Y}]) = p+1$. There is no nonzero vector in the orthogonal complement of the space generated by the rows of $[\mathbf{X},\mathbf{Y}]$. In order to obtain a solution, the rank of $[\mathbf{X},\mathbf{Y}]$ must be reduced to p, i.e. $\mathrm{rank}([\mathbf{X_0},\mathbf{Y_0}]) = p$. Let SVD of $[\mathbf{X},\mathbf{Y}]$ be given by (9) then for $[\mathbf{X_0},\mathbf{Y_0}] = \sum_{i=1}^{p}\sigma_i u_i v_i^T$ Eckart-Young-Mirsky theorem tells us that

$$\min_{\mathrm{rank}(\mathbf{A})=p}\|[\mathbf{X},\mathbf{Y}]-\mathbf{A}]\|_F = \|[\mathbf{X},\mathbf{Y}]-[\mathbf{X_0},\mathbf{Y_0}]\|_F = \sqrt{\sum_{i=p+1}^{p+1}\sigma_i^2} = \sigma_{p+1}. \tag{11}$$

By another words, the best rank p TLS approximation $[\mathbf{X_0},\mathbf{Y_0}]$ of $[\mathbf{X},\mathbf{Y}]$ is obtained by settings the smallest singular value σ_{p+1} to zero. This was firstly investigate by Golub and Van Loan (see [6]) and the following theorem gives conditions for uniqueness and existence of a TLS solution.

Solution to the TLS problem

Let the SVD of $[\mathbf{X},\mathbf{Y}] = \sum_{i=1}^{r}\sigma_i u_i v_i^T$ and $\sigma_{min}(\mathbf{X})$ be the smallest singular value of $\mathbf{X}$. If $\sigma_{min}(\mathbf{X}) > \sigma_{p+1}$, then the TLS solution

$$\hat{\beta}^{(TLS,n)} = -\frac{1}{v_{p+1,p+1}}[v_{1,p+1},\ldots,v_{p,p+1}]^T \tag{12}$$

exists and is the unique solution to $\mathbf{Y_0} = \mathbf{X_0}\beta$ and the corresponding TLS correction matrix is given by

$$[\varepsilon,\mathbf{\Theta}] = \sigma_{p+1}u_{p+1}v_{p+1}^T. \tag{13}$$

The condition $\sigma_{min}(X) > \sigma_{p+1}$ ensure the uniques of TLS solution. If we suppose that $[\mathbf{X},\mathbf{Y}]$ has full column rank, this condition is generically satisfied. For general case see [1], where the situation when $\sigma_{min}(X) = \sigma_{p+1}$ is analyzed. Since singular vectors v_i's are eigenvectors of the matrix $[\mathbf{X},\mathbf{Y}]^T[\mathbf{X},\mathbf{Y}]$, $\hat{\beta}^{(TLS,n)}$ satisfies the following eigenvector equations

$$[\mathbf{X},\mathbf{Y}]^T[\mathbf{X},\mathbf{Y}]\begin{bmatrix}\hat{\beta}^{(TLS,n)}\\ -1\end{bmatrix} = \begin{bmatrix}\mathbf{X}^T\mathbf{X} & \mathbf{X}^T\mathbf{Y}\\ \mathbf{Y}^T\mathbf{X} & \mathbf{Y}^T\mathbf{Y}\end{bmatrix}\begin{bmatrix}\hat{\beta}^{(TLS,n)}\\ -1\end{bmatrix} = \sigma_{p+1}^2\begin{bmatrix}\hat{\beta}^{(TLS,n)}\\ -1\end{bmatrix}$$

and we can write the closed-form expression of the TLS solution

$$\hat{\beta}^{(TLS,n)} = (\mathbf{X}^T\mathbf{X}-\sigma_{p+1}^2\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y}.$$

The previous formula tells us that the TLS solution is more ill-conditioned than the LS solution. It can be dangerous to compute TLS estimate by this way and that is why we will evaluate TLS solution by the help of SVD and equation (12). The computational stability and speed can by further improved by using the Golub-Kahan bidiagonalization

(GKB) to the matrix $[\mathbf{X}, \mathbf{Y}]$ and use the connection among TLS problem, GKB and Krylov subspaces. This concept is called core problem and has been developed by Paige and Strakoš (see [7]). The idea is to find by the help of GKB two orthonormal matrices $\mathbf{P}, \mathbf{Q}$ such that

$$\mathbf{P}^T [\mathbf{Y}, \mathbf{XQ}] = \begin{bmatrix} b_1 & \mathbf{A}_{11} & 0 \\ 0 & 0 & \mathbf{A}_{22} \end{bmatrix}$$

where the block $\mathbf{A}_{11}$ is lower bidiagonal with nonzero bidiagonal elements. Moreover, the matrix $\mathbf{A}_{11}$ has full column rank and its singular values are simple. The matrix $\mathbf{A}_{11}$ has minimal dimensions, and $\mathbf{A}_{22}$ has maximal dimensions and the first elements of all left singular vectors of $\mathbf{A}_{11}$, are nonzero. These properties guarantee that the subproblem $b_1 \approx \mathbf{A}_{11}\beta_{11}$ has minimal dimensions and contains all necessary and sufficient information for solving the original problem $\mathbf{Y} \approx \mathbf{X}\beta$. All irrelevant and redundant information is contained in $\mathbf{A}_{22}$. The asymptotical behaviour of the TLS estimator such as consistency or asymptotic normality is shown and proved by in [9], among others. If the errors in the observations are independent random variables with zero mean and equal variance, TLS gives better estimate than does LS. The problem arises when outliers are present then accuracy of the TLS estimate deteriorates considerably, because classical TLS estimation is not robust estimator.

4 Robustified Total Least Squares

The goal of this section is to propose a robustified version of TLS estimator that is based on the idea of downweighting the influential points. We want to find such a estimators that will combine the advantages of both TLS and respectively LTS and LWS. Firstly let us define the total least trimmed squares.

Total Least Trimmed Squares (TLTS)

TLTS minimizes the sum of the h smallest squared orthogonal distances of data points p_i's from the pth dimensional hyperplane $\rho(\beta)$. The j-th orthogonal distances is denoted by d_j and defined by

$$d_j \quad = \quad \frac{|Y_j - X_j\beta|^2}{1 + \|\beta\|^2}. \tag{14}$$

The TLTS estimator is defined as follows

$$\hat{\beta}^{(TLTS,n)} \quad = \quad \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{h} d^2_{(i)}, \tag{15}$$

where h is an optional parameter satisfying $\frac{n}{2} \leq h \leq n$ and $d^2_{(i)}$ is the i-th least squared orthogonal distance, i.e. for any $\beta \in \mathbb{R}^p$

$$d^2_{(1)}(\beta) \leq d^2_{(2)}(\beta) \leq \ldots \leq d^2_{(n)}(\beta).$$

TLTS estimator has similar properties as LTS estimator. The existence of TLTS is given by the existence of the TLS for subsamples of size h. The computational complexity to find the exact solution is given again by examination of all $\binom{n}{h}$ possible subsamples and

the candidate with the smallest value of the objective function is TLTS estimate. The approximative algorithm to evaluate TLTS is following:
For k=1 to number of iteration do

1. Pick randomly $(p+1)$ data points and compute TLS estimate $\hat{\beta}^{(TLS,p+1)}$ by the help of SVD.
2. Compute the orthogonal distance for all n data points from the pth dimensional hyperplane $\rho(\hat{\beta}^{(TLS)})$.
3. Select the h data points with the smallest squared orthogonal distances d_i's.
4. Compute TLS estimate $\hat{\beta}^{(TLS,h)}$ by the help of SVD for selected data points.
5. Repeat steps 2-4 until convergence.
6. If the value of the objective function is the smallest one among the values, that have been reached up to this moment, store the appropriate estimation as a TLTS.

This algorithm is very fast and generally gives satisfactory results. We also tried to implement algorithms based on theory of annealing or on genetics algorithms, results of these algorithms are sufficient, but there is still some pending work. The larger simulation study is the question of present work. The disadvantages of the TLTS is its infinite local sensitivity, hence we modify the estimator by adding some continuous weighting function and multiply the distances by a weights from $\langle 0, 1\rangle$.
Total Least Weighted Squares

$$\hat{\beta}^{(TLWS,w,n)} = \arg\min_{\beta\in\mathbb{R}^p} \sum_{i=1}^{n} w\left(\frac{i-1}{n}\right) d^2_{(i)}(\beta) = \arg\min_{\beta\in\mathbb{R}^p} \sum_{i=1}^{n} w\left(\frac{\pi(\beta,i)-1}{n}\right) d^2_i(\beta),$$

where weights w_i are defined by the weight function $w : \langle 0, 1\rangle \to \langle 0, 1\rangle$, which is absolutely continuous, $w(0) = 1$ and non-increasing with the derivative $w'(t)$ bounded from below by a constant $(-L)$, where $L \geq 0$ and $\pi(\beta, i)$ is the random rank of the i-th residual as previously. The evaluation of this estimator, algorithms and the large sample properties are under research.

5 Mixed Ordinary Least Squares Total Least Squares

Sometimes the linear modeling problem $\mathbf{Y} \approx \mathbf{X}\beta$ contains the intercept (i.e. $X_{i1} = 1,\ i = 1, \ldots n$) or some columns of $\mathbf{X}$ may be known exactly. In this cases the TLS solution cannot give the accurate estimation of parameters β. It is natural to require that the corresponding columns of the data matrix $\mathbf{X}$ be unperturbed since they are known exactly. The generalization of the TLS approach is called mixed least squares - total least squares problem (mixed LS-TLS). Let us suppose the overdetermined system of n linear equations

$$\mathbf{Y} \approx \mathbf{X}\beta, \quad \mathbf{Y} \in \mathbb{R}^n,\ \mathbf{X} \in \mathbb{R}^{n\times p},\ n > p,$$

$$\text{partition} \quad \begin{array}{ll} \mathbf{X} = \left[\mathbf{X}^{(1)}, \mathbf{X}^{(2)}\right] & \mathbf{X}^{(1)} \in \mathbb{R}^{n\times p_1},\ \mathbf{X}^{(2)} \in \mathbb{R}^{n\times p_2} \\ \beta^T = \left[{\beta^{(1)}}^T, {\beta^{(2)}}^T\right] & \beta^{(1)} \in \mathbb{R}^{p_1},\ \beta^{(2)} \in \mathbb{R}^{p_2} \end{array}$$

and assume that the columns of $\mathbf{X}^{(1)}$ are error free and $p_1 + p_2 = p$. Then the mixed LS-TLS problem seeks to

$$\hat{\beta}^{(LS-TLS,n)} = \min_{\beta\in\mathbb{R}^p, [\varepsilon,\mathbf{\Theta}]\in\mathbb{R}^{n\times(p_2+1)}} \|[\varepsilon, \mathbf{\Theta}]\|_F \quad \text{subject to} \quad \mathbf{Y}+\varepsilon = \mathbf{X}^{(1)}\beta^{(1)}+(\mathbf{X}^{(2)}+\mathbf{\Theta})\beta^{(2)}. \tag{16}$$

$\hat{\beta}^{(LS-TLS,n)}$ is called a mixed LS-TLS solution to the problem (16) and $[\varepsilon, \mathbf{\Theta}]$ is the corresponding LS-TLS correction. By varying p_1 from zero to p, the mixed LS-TLS problem can handle also with any ordinary LS or ordinary TLS problem. To solve the mixed LS-TLS problem, due to Golub, we us QR factorization, solve ordinary TLS problem of reduced dimension and after that we compute the first p_1 components of $\hat{\beta}^{(LS-TLS,n)}$. Let a matrix $\left[\mathbf{X}^{(1)}, \mathbf{X}^{(2)}\right]$ be given, have full column rank and columns of $\mathbf{X}^{(1)}$ are error free. Suppose that $0 < p_1 < p$ then compute the QR factorization

$$\left[\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \mathbf{Y}\right] = \mathbf{Q}\begin{bmatrix} \mathbf{R_{11}} & \mathbf{R_{12}} & \mathbf{R_{Y_1}} \\ 0 & \mathbf{R_{22}} & \mathbf{R_{Y_2}} \end{bmatrix},$$

where $\mathbf{Q}$ is orthogonal, $\mathbf{R_{11}} \in \mathbb{R}^{p_1\times p_1}$ and $\mathbf{R_{22}} \in \mathbb{R}^{n-p_1\times p_2+1}$ are upper triangular. Then compute the ordinary TLS solution $\hat{\beta}^{(TLS,n-p_1)}$ of $\mathbf{R_{Y_2}} \approx \mathbf{R_{22}}\beta$ which gives us the the last p_2 components of $\hat{\beta}^{(LS-TLS,n)}$. The first p_1 components we obtain from the solution of following equation

$$\mathbf{R_{11}}\hat{\beta}^{(LS,p_1)} = \mathbf{R_{Y_1}} - \mathbf{R_{12}}\hat{\beta}^{(TLS,n-p_1)}.$$

The mixed LS-TLS solution is $\hat{\beta}^{(LS-TLS,n)} = \left[\hat{\beta}^{(LS,p_1)}, \hat{\beta}^{(TLS,n-p_1)}\right]$. Unfortunately this universal estimator is not robust and gives misleading results when outliers occur.

6 Robustified Mixed Least Squares Total Least Squares

The robustification of mixed LS-TLS estimator is not straightforward as in ordinary total least squares. Let denote by ρ_1 the $p_1 + 1$ dimensional hyperplane given by the normal vector $\nu_1 = \left[\hat{\beta}_1^{(LS-TLS,n)}, \ldots, \hat{\beta}_{p_1}^{(LS-TLS,n)}, -1\right]^T$ and by ρ_2 the $p_2 + 1$ dimensional hyperplane given by the normal vector $\nu_2 = \left[\hat{\beta}_{p_1+1}^{(LS-TLS,n)}, \ldots, \hat{\beta}_{p}^{(LS-TLS,n)}, -1\right]^T$. Then we can compute the squared vertical distance of each data point $[X_{i1}, \ldots, X_{ip_1}, Y_i]$ from the hyperplane ρ_1 and the orthogonal distance of each data point $[X_{ip_1+1}, \ldots, X_{ip}, Y_i]$ from the hyperplane ρ_2. Now we need to take some reasonable combination of these two distances, identify the influential points and downweight them. Another possibility is to identify the influential points separately and instead of discarding s outermost points from the $p+1$ dimensional hyperplane ρ given by the normal vector $\left[\hat{\beta}^{(LS-TLS,n)}, -1\right]^T$ we discard $s/2$ points from the first part by the help of LTS and $s/2$ points from the second part by the help of TLTS. However, algorithms and properties of robustified mixed LS-TLS estimator are still under research and large simulation study is under preparation.

7 Conclusions

In this paper we reviewed the development and extensions of estimation of parameters in linear regression model when outliers occur and the orthogonality condition fails. The most frequent example of this problem is a case when the explanatory variables are measured with a random errors. In algebraic point of view is the problem of overdetermined system $\mathbf{Y} \approx \mathbf{X}\beta$. We propose and described two approaches how to solve this kind of problem. The first approach is based on theory of instrumental variables and the second one on theory of total least squares. We described how to determine the solution to the basic TLS problem from the SVD. We propose the robustified versions of IV and of TLS with outlines of the algorithms for computations of the solutions of the instrumental weighted variables or total least trimmed squares. Furthermore we generalized the TLS to mixed LS-TLS method and propose the robustified version. To all mentioned estimators and methods we ran small simulation study and both results and MATLAB codes of algorithms are available on the request. Since the error-in-variables model corresponds to TLS, this field of mathematics connects the algebraic and numerical mathematics with statistics. For further reading we can recommend [1] or [8].

References

[1] Van Huffel, S. and Vandewalle, J. *The Total Least Squares Problem: Computational Aspects and Analysis*. SIAM Philadelphia (1991).

[2] Franc, J. *Robustified instrumental variables, Master thesis*. FNSPE, Czech Technical University in Prague (2009).

[3] Rousseeuw, P. J. *Least Median of Squares Regression* . Journal of the American Statistical Association 79, (1984), 871-–880 .

[4] Víšek, J. Á. *Regression with high breakdown point*. Robust 2000 ,(2000), 324–356 .

[5] Víšek, J. Á. *Consistency of the instrumental weighted variables*. Annals of the Institute of Statistical Mathematics 61, number 3, (2009), 543–578.

[6] Golub, G. and Van Loan, C. *An analysis of the total least squares problem*. SIAM J. Numerical Analysis 17, (1980), 883–893.

[7] Paige, C. C. and Strakoš, Z. *Core problems in linear algebraic systems*. SIAM Journal on Matrix Analysis and Applications 27, (2006), 861–875.

[8] Markovsky, I. and Van Huffel, S. *Overview of total least squares methods*. Signal Processing 87, number 10, (2007), 2283–2302.

[9] Glesser, L. J. *Estimation in a multivariate errors-in-variables regression model: Large sample results*. Annals of Statistics 9, (1981), 24–44.

Performances of Modified Power Divergence Estimators in Normal Models*

Iva Frýdlová

7th year of PGS, email: ivafrydlova@quick.cz
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Václav Kůs, Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. Point estimators based on minimization of information-theoretic divergences between empirical and hypothetical distribution induce a problem when working with continuous families which are measure-theoretically orthogonal with the family of empirical distributions. In this case the ϕ-divergence is always equal to its upper bound and the minimum ϕ-divergence estimates are trivial. Broniatowski and Vajda in [2] proposed several modifications of the minimum divergence rule to provide a solution to the above mentioned problem. We examine these modifications in practical use.

Keywords: divergences, minimum ϕ-divergence estimation, maximum subdivergence estimators, minimum superdivergence estimators, simulations

Abstrakt. Bodové odhady založené na minimalizaci ϕ-divergencí mezi empirickou a hypotetickou distribucí přináší problém, pokud pracujeme se spojitými rodinami hustot, které jsou ortogonální (vzhledem k dominující σ-konečné míře) s rodinami empirických distribucí. V takovém případě je ϕ-divergence vždy rovna své horní mezi a odhad s minimální ϕ-divergencí je tudíž triviální. Broniatowski a Vajda v [2] navrhli několik modifikací tohoto typu odhadu a poskytli tak řešení zmiňovaného problému. V tomto příspěvku se věnujeme praktickému využití těchto modifikovaných odhadů.

Klíčová slova: divergence, odhad s minimální ϕ-divergencí, odhad s maximální subdivergencí, odhad s minimální superdivergencí, simulace

1 Introduction

As was already mentioned in many publications, the well known information-theoretic measures of divergence of probability measures introduced in the 60ties by A. Renyi and I. Csiszar cannot be directly applied in statistical estimation, since the divergence between the theoretical absolutely continuous probability measure and the discrete empirical probability measure is always equal to its upper bound and often takes on infinite values. In 2008 - 2009 Broniatowski & Vajda ([2]) studied and extended two different modifications of divergences proposed independently in 2006 ([5], [1]). They altered the traditional ϕ-divergences into *subdivergences* and *superdivergences* and defined *maximum subdivergence estimators with escort parameter* θ and *minimum superdivergence estimators*. We

*This work has been supported by grants GA CR 102/07/1131, GA CR P202/10/0618, SGS OHK4-007/10 and MSMT 1M0572.

shall present the key results of extensive simulation study of these types of point estimators. The main interest of our research ([4]) was to examine these modifications in practical use as to the consistency, robustness and efficiency of the estimators. We focus on the well known family of power divergences parametrized by α in the normal distribution model. We run a comparative computer simulation for several randomly selected contaminated and uncontaminated data sets, and we study the behavior of estimators for different sample sizes and different ϕ-divergence parameters.

2 ϕ-divergences and Minimum ϕ-divergence Estimators

This chapter introduces the ϕ-divergences, their basic characteristics, and the concept of minimum divergence estimation. We mention several problems encountered when working with these estimators, and we suggest some possiblilities to bypass them.

Let $(\mathcal{X}, \mathcal{A})$ be a measurable space and let $\mathcal{P}$ be a set of all probability measures on $(\mathcal{X}, \mathcal{A})$. If $P \in \mathcal{P}$ is dominated by a σ-finite measure λ on $(\mathcal{X}, \mathcal{A})$, then $p = dP/d\lambda$ is a Radon-Nikodym density of P with respect to measure λ.

Definition 1.. Let $P, Q \in \mathcal{P}$, $\{P, Q\} \ll \lambda$, $p = dP/d\lambda$ and $q = dQ/d\lambda$. A *ϕ-divergence of distributions P a Q* is a function $D_\phi : \mathcal{P} \times \mathcal{P} \to [0, \infty]$ defined by

$$D_\phi(P, Q) = \int_{\mathcal{X}} \phi\left(\frac{p}{q}\right)\, dQ = \int_{\mathcal{X}} q\, \phi\left(\frac{p}{q}\right)\, d\lambda\ , \tag{1}$$

where $\phi : (0, \infty) \to \mathbb{R}$ is a convex function.

For this formula to be well defined, we put

$$q\, \phi\left(\frac{p}{q}\right) = \begin{cases} q\, \phi(0) & if\ p = 0 \\ p\, \phi(\infty)/\infty & if\ q = 0, \end{cases}$$

where $\phi(0) := \lim_{t \to 0_+} \phi(t)$ and $\phi(\infty)/\infty := \lim_{t \to \infty} \frac{\phi(t)}{t}$, while "$0 \cdot \infty = 0$".

For ϕ-divergences it holds the following theorem.

Theorem 1.. *For each generating function ϕ it holds $\phi(1) \le D_\phi(P, Q) \le \phi(0) + \phi(\infty)/\infty$ for every $P, Q \in \mathcal{P}$, where the left equality takes place if $P = Q$ and the right equality takes place if $P \perp Q$, i.e. P, Q are singular.*

From now on, we shall consider only ϕ which are twice differentiable, strictly convex generating functions with $\phi(1) = 0$ and continuous extension to $t = 0_+$ denoted by $\phi(0)$. We let $\mathbf{\Phi}$ be the class of all such functions. As to the probability measures, we will deal with P and Q which are either measure-theoretically equivalent, $P \equiv Q$ (i.e. $pq > 0$ λ-a.s.), or measure-theoretically orthogonal, $P \perp Q$ (i.e. $pq = 0$ λ-a.s.).

In the sequel, we shall use the power divergences

$$D_\alpha(P, Q) := D_{\phi_\alpha}(P, Q), \quad \alpha \in \mathbb{R}, \tag{2}$$

where

$$\phi_\alpha(t) = \frac{t^\alpha - \alpha(t-1) - 1}{\alpha(\alpha - 1)} \quad \alpha \neq 0, \alpha \neq 1 \tag{3}$$

with the limiting cases

$$\phi_0(t) = -\ln t + t - 1 \quad \text{and} \quad \phi_1(t) = t \ln t - t + 1. \tag{4}$$

For these functions it holds as a result of Theorem 1

$$0 \leq D_\alpha(P,Q) \leq \begin{cases} \frac{1}{\alpha(1-\alpha)} & \text{if} \quad 0 < \alpha < 1 \\ \infty & \text{otherwise.} \end{cases} \tag{5}$$

The left equality takes place if and only if $P = Q$. If $0 < \alpha < 1$ then the right equality takes place if and only if $P \perp Q$. Otherwise it takes place if $\alpha \leq 0$ and $Q \not\ll P$, i.e. $Q(\mathcal{X} - S(P)) > 0$, or if $\alpha \geq 1$ and $P \not\ll Q$, i.e. $P(\mathcal{X} - S(Q)) > 0$.

Now, let $X_1, X_2, ..., X_n$ be independent and identically distributed observations governed by $P_{\theta_0} \in \mathcal{P}$, where $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is a family of probability measures on $(\mathcal{X}, \mathcal{A})$, $\Theta \subset \mathbb{R}^d$ is a parameter space, and we assume that for every $\theta, \theta_0 \in \Theta$, $\theta \neq \theta_0$ holds $P_\theta \neq P_{\theta_0}$ and $P_\theta \equiv P_{\theta_0}$. Moreover, we assume the family $\mathcal{P}$ to be nonatomic (continous), i.e. for all $\theta \in \Theta$ and $x \in \mathcal{X}$ we require $P_\theta(\{x\}) = 0$. We also let the data $X_1, X_2, ..., X_n$ to be represented by an empirical probability measure $P_n = \frac{1}{n}\sum_{i=1}^n P_{X_i}$, where P_x is the Dirac probability measure with all mass concenrated at the point $x \in \mathcal{X}$.

Definition 2.. Let $\phi \in \mathbf{\Phi}$. We say that an estimator $\widehat{\theta}_n : \mathcal{X}^n \to \Theta$ of a true parameter $\theta_0 \in \Theta$ is a *minimum ϕ-divergence estimator* if for the corresponding D_ϕ it holds that

$$\widehat{\theta}_n = \text{argmin}_\theta D_\phi(P_\theta, P_n).$$

The problem we encounter with these estimators is that the continuous family $\mathcal{P}$ and the family of empirical distributions $\mathcal{P}_{emp}$ are measure-theoretically orthogonal, i.e. $P_\theta \perp P_n$ for every $P_\theta \in \mathcal{P}$ and $P_n \in \mathcal{P}_{emp}$. This implies that for every $P_\theta \in \mathcal{P}$ and $P_n \in \mathcal{P}_{emp}$

$$D_\phi(P_\theta, P_n) = \phi_\alpha(0) + \phi_\alpha(\infty)/\infty$$

and the above defined estimates are trivial. To face this problem, it is possible to use some prior smoothing of the data or another nonparametric density estimation like we did by implementing histogram in [3], but these methods bring another unpleasant obstructions such as bandwidth selection. In the next section, we present several modifications of the minimum divergence rule studied by Broniatowski & Vajda ([2]) avoiding these complications as well.

3 Power subdivergence and superdivergence estimators

We shall regard the probability measures $P \in \mathcal{P}$ and $Q \in \mathcal{Q}$ for $\mathcal{Q} = \mathcal{P} \cup \mathcal{P}_{emp}$.

Consider the family of finite expectations

$$\underline{\mathrm{D}}_{\phi,\tilde{\theta}}(P_\theta, Q) = \int \phi'(p_\theta/p_{\tilde{\theta}})\, \mathrm{d}P_\theta + \int \phi^{\#}(p_\theta/p_{\tilde{\theta}})\, \mathrm{d}Q, \quad (P_\theta,\ Q) \in \mathcal{P} \otimes \mathcal{Q} \tag{6}$$

parametrized by $(\phi, \tilde{\theta}) \in \mathbf{\Phi} \otimes \mathbf{\Theta}$, where

$$\phi^{\#}(t) = \phi(t) - t\phi'(t) \quad \text{for every} \quad \phi \in \mathbf{\Phi}$$

and ϕ' denotes the derivative of ϕ. For (6) to be correctly defined, we assume that the integrals exist and have a finite value.

Now, the *maximum subdivergence estimators* with escort parameter $\theta \in \mathbf{\Theta}$ (briefly, the $\max\underline{\mathrm{D}}_\phi$-estimators) are defined as

$$\tilde{\theta}_{\phi,\theta,n} = \mathrm{argmax}_{\tilde{\theta}}\, \underline{\mathrm{D}}_{\phi,\tilde{\theta}}(P_\theta, P_n) = \mathrm{argmax}_{\tilde{\theta}} \left[\int \phi'\left(\frac{p_\theta}{p_{\tilde{\theta}}}\right) \mathrm{d}P_\theta + \frac{1}{n}\sum_{i=1}^{n} \phi^{\#}\left(\frac{p_\theta(X_i)}{p_{\tilde{\theta}}(X_i)}\right)\right]$$

and the *minimum superdivergence estimators* (briefly, the $\min\bar{\mathrm{D}}_\phi$-estimators) as

$$\theta_{\phi,n} = \mathrm{argmin}_{\theta}\, \sup_{\tilde{\theta}} \underline{\mathrm{D}}_{\phi,\tilde{\theta}}(P_\theta, P_n) = \mathrm{argmin}_{\theta}\, \sup_{\tilde{\theta}} \left[\int \phi'\left(\frac{p_\theta}{p_{\tilde{\theta}}}\right) \mathrm{d}P_\theta + \frac{1}{n}\sum_{i=1}^{n} \phi^{\#}\left(\frac{p_\theta(X_i)}{p_{\tilde{\theta}}(X_i)}\right)\right].$$

If we restrict ourselves to a subclass of these estimators determined by the power divergences given in (2), by employing the power functions ϕ_α from (3) and (4) we receive for $\alpha > 0$ formulas

$$\tilde{\theta}_{\alpha,\theta,n} = \mathrm{argmin}_{\tilde{\theta}}\, M_{\alpha,\theta}(P_n, \tilde{\theta}) \tag{7}$$

and

$$\theta_{\alpha,n} = \mathrm{argmax}_{\theta} \mathrm{inf}_{\tilde{\theta}}\, M_{\alpha,\theta}(P_n, \tilde{\theta}) \quad \text{or} \quad \theta_{\alpha,n} = \mathrm{argmax}_{\theta}\, M_{\alpha,\theta}(P_n, \tilde{\theta}_{\alpha,\theta,n}) \tag{8}$$

where

$$\begin{aligned} M_{\alpha,\theta}(P_n, \tilde{\theta}) &= \frac{1}{1-\alpha}\int \left(\frac{p_\theta}{p_{\tilde{\theta}}}\right)^{\alpha} \mathrm{d}P_{\tilde{\theta}} + \frac{1}{\alpha n}\sum_{i=1}^{n}\left(\frac{p_\theta(X_i)}{p_{\tilde{\theta}}(X_i)}\right)^{\alpha} \quad \text{if } \alpha > 0,\ \alpha \neq 1 \\ &= -\int \ln \frac{p_\theta}{p_{\tilde{\theta}}}\, \mathrm{d}P_\theta + \frac{1}{n}\sum_{i=1}^{n} \frac{p_\theta(X_i)}{p_{\tilde{\theta}}(X_i)} \qquad \text{if } \alpha = 1 \end{aligned} \tag{9}$$

for all $Q \in \mathcal{Q} = \mathcal{P} \cup \mathcal{Q}$ and

$$\tilde{\theta}_{0,\theta,n} = \mathrm{argmax}_{\tilde{\theta}}\, \Sigma_{i=1}^{n} \ln p_{\tilde{\theta}}(X_i) \quad \text{and} \quad \theta_{0,n} = \mathrm{argmax}_{\theta}\, \Sigma_{i=1}^{n} \ln p_\theta(X_i) \tag{10}$$

for $\alpha = 0$. It is obvious that in this case the estimators coincide with MLE's, hence the classes of $\max\underline{\mathrm{D}}_\phi$ -estimators and of $\min\bar{\mathrm{D}}_\phi$-estimators are extensions of the MLE.

3.1 Power subdivergence estimators and power superdivergence estimators in the normal distribution model

Let the observation space $(\mathcal{X}, \mathcal{A})$ be $(\mathbb{R}, \mathcal{B})$ and $\mathcal{P} = \{P_{\mu,\sigma} : \mu \in \mathbb{R},\ \sigma > 0\}$ be the normal family with parameters of location μ and scale σ (i.e. variances σ^2). We are interested in the $\min\bar{\mathrm{D}}_\alpha$-estimates $(\mu_{\alpha,n}, \sigma_{\alpha,n})$ and the $\max\underline{\mathrm{D}}_\alpha$-estimates $(\tilde{\mu}_{\alpha,\mu,\sigma,n}, \tilde{\sigma}_{\alpha,\mu,\sigma,n})$ with power parameters $\alpha \geq 0$ and escort parameters $(\mu, \sigma) \in \mathbb{R} \otimes (0, \infty)$.

If $\alpha = 0$ then these estimators reduce to

$$(\mu_{0,n}, \sigma_{0,n}) = (\tilde{\mu}_{0,\mu,\sigma,n}, \tilde{\sigma}_{0,\mu,\sigma,n}) = \left(\frac{1}{n}\sum_{i=1}^{n} X_i,\ \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \tilde{\mu}_{0,\mu,\sigma,n})^2}\right) \tag{11}$$

which is a maximum likelihood estimate in the family of normal distributions.

For $\alpha > 0$, $\alpha \neq 1$ the function (9) becomes

$$M_{\alpha,\mu,\sigma}(P_n, \tilde{\mu}, \tilde{\sigma}) = \frac{1}{1-\alpha}\int \left(\frac{p_{\mu,\sigma}}{p_{\tilde{\mu},\tilde{\sigma}}}\right)^{\alpha} \mathrm{d}P_{\tilde{\mu},\tilde{\sigma}} + \frac{1}{\alpha n}\sum_{i=1}^{n}\left(\frac{p_{\mu,\sigma}(X_i)}{p_{\tilde{\mu},\tilde{\sigma}}(X_i)}\right)^{\alpha} \tag{12}$$

where

$$\left(\frac{p_{\mu,\sigma}(x)}{p_{\tilde{\mu},\tilde{\sigma}}(x)}\right)^{\alpha} = \left(\frac{\tilde{\sigma}}{\sigma}\right)^{\alpha} \exp\left\{\frac{\alpha\,(x-\tilde{\mu})^2}{2\tilde{\sigma}^2} - \frac{\alpha\,(x-\mu)^2}{2\sigma^2}\right\},$$

and

$$\int \left(\frac{p_{\mu,\sigma}}{p_{\tilde{\mu},\tilde{\sigma}}}\right)^{\alpha} \mathrm{d}P_{\tilde{\mu},\tilde{\sigma}} = \exp\left\{\frac{-\alpha(1-\alpha)(\mu-\tilde{\mu})^2}{2[\alpha\tilde{\sigma}^2+(1-\alpha)\sigma^2]} - \ln\frac{\sqrt{\alpha\tilde{\sigma}^2+(1-\alpha)\sigma^2}}{\tilde{\sigma}^{\alpha}\sigma^{1-\alpha}}\right\}.$$

For $\alpha = 1$

$$\begin{aligned} M_{1,\mu,\sigma}(P_n,\tilde{\mu},\tilde{\sigma}) &= \lim_{\alpha\to 1} M_{\alpha,\mu,\sigma}(P_n,\tilde{\mu},\tilde{\sigma}) \\ &= \frac{-(\mu-\tilde{\mu})^2}{2\tilde{\sigma}^2} - \frac{1}{2}\left[-\ln\left(\frac{\sigma}{\tilde{\sigma}}\right)^2 + \left(\frac{\sigma}{\tilde{\sigma}}\right)^2 - 1\right] \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\tilde{\sigma}}{\sigma}\right)\exp\left\{\frac{(X_i-\tilde{\mu})^2}{2\tilde{\sigma}^2} - \frac{(X_i-\mu)^2}{2\sigma^2}\right\}. \end{aligned} \tag{13}$$

In [2], Vajda shows that the max$\underline{\mathrm{D}}_{\alpha}$-estimators of location are Fisher consistent in the normal family $\mathcal{P}_{\sigma} = \{P_{\mu,\sigma} = N(\mu,\sigma^2) : \mu \in \mathbb{R}\}$ with $\sigma > 0$ fixed if and only if $\sigma = 1$, which suggests an easy loss of consistency of these estimators. We shall inspect this property by simulations in the next chapter. We shall also examine whether the max$\underline{\mathrm{D}}_{\alpha}$-estimators escorted by MLE $\tau_n = \tilde{\mu}_{0,\mu,n}$, i.e. $\tilde{\mu}_{\alpha,\tau_n,n}$, are Fischer consistent under all hypothetical models $P_{\mu,\sigma} = N(\mu,\sigma^2)$, $\sigma > 0$, and possibly consistent and robust under the contaminated versions of these models.

4 Computer Simulations

This chapter is to present the results obtained by applying the methods of Broniatowski & Vajda introduced in the previous chapters. We target the study at the power subdivergence and power superdivergence estimators of location given by

$$\mu_{0,n} = \tilde{\mu}_{0,\mu,n} = \bar{\boldsymbol{X}}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

and $\mu_{\alpha,n} = \mathrm{argmax}_{\mu}\mathrm{inf}_{\tilde{\mu}}\, M_{\alpha,\mu}(P_n,\tilde{\mu})$ $\tilde{\mu}_{\alpha,\mu,n} = \mathrm{argmin}_{\tilde{\mu}}\, M_{\alpha,\mu}(P_n,\tilde{\mu})$

for $\alpha > 0$ with $M_{\alpha,\mu}(P_n,\tilde{\mu})$ given by (12) with parameter $\sigma = 1$, and the power subdivergence and power superdivergence estimators of scale given by

$$\sigma_{0,n} = \tilde{\sigma}_{0,\sigma,n} = \boldsymbol{S}_n = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

and $\sigma_{\alpha,n} = \mathrm{argmax}_{\sigma}\mathrm{inf}_{\tilde{\sigma}}\, M_{\alpha,\sigma}(P_n,\tilde{\sigma})$ $\tilde{\sigma}_{\alpha,\sigma,n} = \mathrm{argmin}_{\tilde{\sigma}}\, M_{\alpha,\sigma}(P_n,\tilde{\sigma})$

for $\alpha > 0$ with $M_{\alpha,\sigma}(P_n,\tilde{\sigma})$ given by (12) with parameter $\mu = 0$. Here $X_1,...,X_n$ are observations on the convex mixtures $P_{\varepsilon} = (1-\varepsilon)P + \varepsilon Q$, P is a standard normal model with location $\mu = 0$ and scale $\sigma = 1$, further denoted by $N(0,1)$, and Q is successively normal ($N(0,9), N(0,100)$), logistic ($Lo(0,1)$), and Cauchy ($C(0,1)$) distribution. The contamination we use is 0, 1, 5, 10, 20, 30 percent respectively, i.e. ε takes on the values

0, 0.01, 0.05, 0.1, 0.2, and 0.3. The sample size n is considered successively 20, 50, 100, 200, 500.

In case of $\min\bar{\mathrm{D}}_\alpha$-estimators $\mu_{\alpha,n}$, $\sigma_{\alpha,n}$ we take into account only power parameters 0, 0.01, 0.05, 0.1, 0.2, and 0.5. In the case of $\max\underline{\mathrm{D}}_\alpha$-estimators $\tilde{\mu}_{\alpha,\mu,n}$ we consider the same values of power parameter and in addition to that we select the escort parameters $\mu =$ 0, 0.1, 0.2, 0.5, 1 and finally $\mu = \bar{\boldsymbol{X}}_n$ (MLE). For $\max\underline{\mathrm{D}}_\alpha$-estimators $\tilde{\sigma}_{\alpha,\sigma,n}$ we consider the same values of power parameter and the escort parameters $\sigma = 0.5$, 1, 1.2, 1.5, 2 and finally $\sigma = \boldsymbol{S}_n$ (MLE) and $\sigma = 1.483\ \mathrm{med}_j(|X_j|)$ (MAD estimate of scale for known location parameter equal to 0, otherwise MAD $= 1.483\ \mathrm{med}_j(|X_j - \mathrm{med}_i(X_i)|)$).

To evaluate the behavior of power superdivergence (or power subdivergence) estimators we generate K different data samples (K=100 or K=1000) to gain K different estimates (further indexed by $^{(k)}$) and we compute means and standard deviations

$$m(\mu) = \frac{1}{K}\sum_{k=1}^{K}\mu_{\alpha,n}^{(k)} \qquad s(\mu) = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(\mu_{\alpha,n}^{(k)} - m(\mu))^2}$$

$$m(\sigma) = \frac{1}{K}\sum_{k=1}^{K}\sigma_{\alpha,n}^{(k)} \qquad s(\sigma) = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(\sigma_{\alpha,n}^{(k)} - m(\sigma))^2}$$

of the $\min\bar{\mathrm{D}}_\alpha$-estimators (or $\max\underline{\mathrm{D}}_\alpha$-estimators) and maximum likelihood estimators $\bar{\boldsymbol{X}}_n^{(k)}$ and $\boldsymbol{S}_n^{(k)}$. Making use of these we receive the relative empirical efficiencies

$$\mathrm{eref}(\mu) = \frac{\frac{1}{K}\sum_{k=1}^{K}(\bar{\boldsymbol{X}}_n^{(k)})^2}{\frac{1}{K}\sum_{k=1}^{K}(\mu_{\alpha,n}^{(k)})^2} \qquad \mathrm{eref}(\sigma) = \frac{\frac{1}{K}\sum_{k=1}^{K}(\boldsymbol{S}_n^{(k)} - 1)^2}{\frac{1}{K}\sum_{k=1}^{K}(\sigma_{\alpha,n}^{(k)} - 1)^2}.$$

4.1 Results for power subdivergence estimators of location

First we inspect the development of consistency, efficiency and robustness in case of mixture $(1-\varepsilon)N(0,1)+\varepsilon N(0,9)$ for moving value of escorting parameter $\mu = 0$, 0.1, 0.2, 0.5, 1 and contamination parameter $\varepsilon = 0$, 0.01, 0.05, 0.1, 0.2, and 0.3.

For power parameter $\alpha = 0$ we can conclude that the estimates coincide with MLE, i.e. $\mathrm{eref}(\tilde{\mu}) = 1$, as was expected. In case of escort parameter $\mu = 0$, the $\max\underline{\mathrm{D}}_\alpha$-estimators for the uncontaminated data still more or less copy the behavior of MLE even for values of $\alpha > 0$, but as the contamination grows, we observe that the mean and standard deviation of $\max\underline{\mathrm{D}}_\alpha$-estimator move apart from MLE taking on lower values than maximum likelihood estimate of the contaminated data. In case of $m(\tilde{\mu})$ the difference is only slight (yet favourable), but in case of $s(\tilde{\mu})$ is the difference apparent (cf. Figure 1) and causes a fair increase in empirical relative efficiency. Figure 2 displays the development of $\mathrm{eref}(\tilde{\mu})$ for different values of power parameter α showing us that the robustness tendency is growing stronger with α increasing. Since the dependence on sample size n is almost constant for $n > 50$, we present in Figure 3 the value of $\mathrm{eref}(\tilde{\mu})$ only for $n = 500$ as a function of contamination parameter ε for different levels of α. This shows the rising efficiency of $\max\underline{\mathrm{D}}_\alpha$-estimator (compared to MLE with $\alpha = 0$) with increasing contamination.

All that was stated above holds for $\mu = 0$. However, the situation changes to the worse for the parameter μ moving to 1. The consistency, efficiency, even the robustness

tendencies slowly vanish, and we see that apart from the case of $\mu = 0$ the $\max\underline{D}_\alpha$-estimators do not possess the useful properties we would desire.

The previously described behavior can be seen also for the other mixtures, i.e. contamination by $N(0, 100)$, $Lo(0, 1)$ and $C(0, 1)$. It was only observed to grow stronger as the outliers get farther away, as is the case of contamination by Cauchy distribution. Especially the robustness of the estimator escorted by $\mu = 0$ is rather stunning compared to MLE. Unfortunately also the loss of consistency for μ moving to 1 is faster.

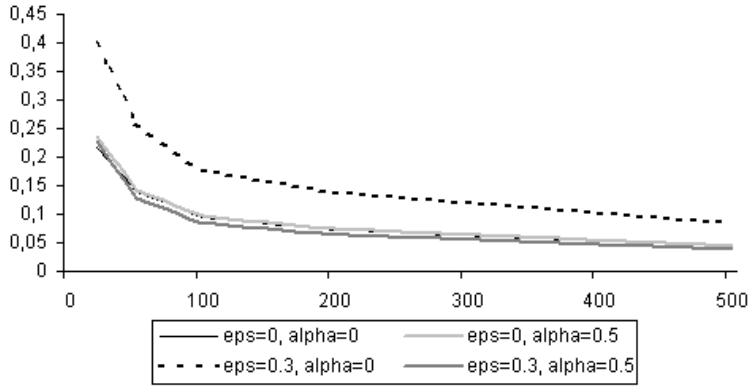


Figure 1 : Dependency of standard deviation of the $\max\underline{D}_\alpha$-estimators with escort parameter $\mu = 0$ on sample size n for data distributed by $(1 - \varepsilon)N(0, 1) + \varepsilon N(0, 9)$

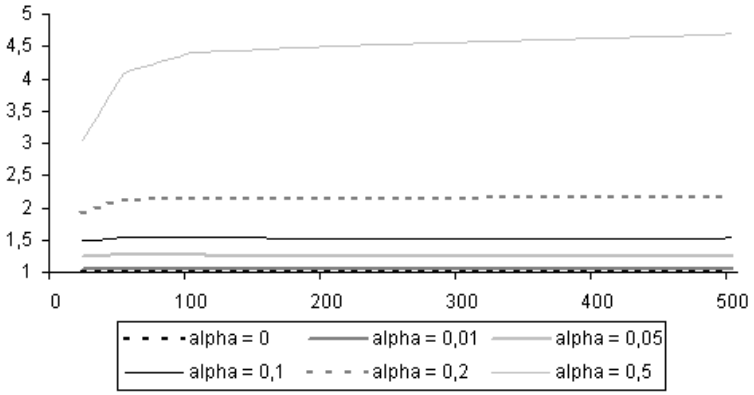


Figure 2 : Dependency of empirical relative efficiency of the $\max\underline{D}_\alpha$-estimators with escort parameter $\mu = 0$ on sample size n for data distributed by $0.7N(0, 1) + 0.3N(0, 9)$

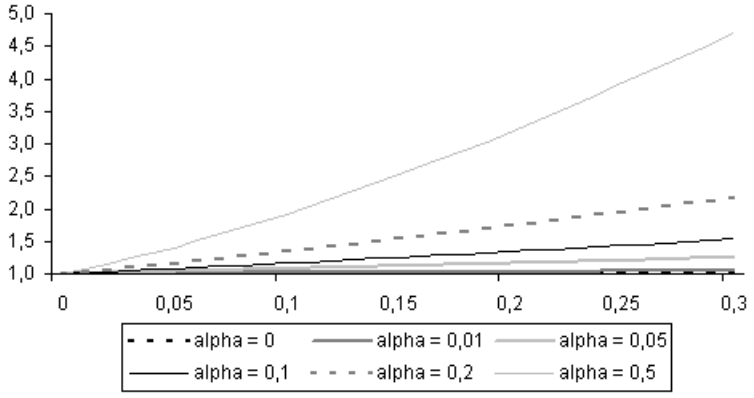


Figure 3 : Dependency of empirical relative efficiency of the $\max\underline{D}_\alpha$-estimators with escort parameter $\mu = 0$ on contam. parameter ε for data distributed by $(1-\varepsilon)N(0, 1)+\varepsilon N(0, 9)$

In accordance with the fact that the best results we obtained were for $\mu = 0$ which is the true parameter of the estimated data, some very good results were received for the value of the escort parameter $\mu = \bar{\boldsymbol{X}}_n$ as was already indicated by theory in [2].

For contamination by $N(0, 9)$, $N(0, 100)$ and $Lo(0, 1)$ we received perfect match with MLE for all values of ε. Nevertheless, an outstanding behavior was noticed in case of contamination by Cauchy distribution, where the power subdivergence estimator shows a significant resistance to distant outliers (cf. Figure 4). In this situation, the standard deviation $s(\tilde{\mu})$ of the maximum likelihood estimator with great volatility copies the occurrence of extreme outliers, while the standard deviation of MLE-escorted subdivergence estimator retains low values and steady convergence to 0. This, clearly, results also in huge empirical relative efficiency.

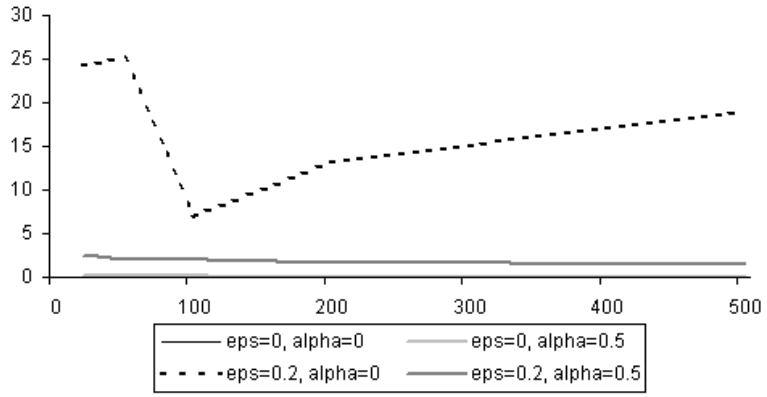


Figure 4 : Dependency of standard deviation of the max$\underline{\mathrm{D}}_\alpha$-estimators with escort parameter $\mu = \bar{\boldsymbol{X}}_n$ on sample size n for data distributed by $(1-\varepsilon)N(0,1) + \varepsilon C(0,1)$

4.2 Results for power subdivergence estimators of scale

Lets again first inspect the development of consistency, efficiency and robustness in case of mixture $(1-\varepsilon)N(0,1)+\varepsilon N(0,9)$ for the values of escorting parameter $\sigma = 0.5$, 1, 1.2, 1.5, 2 and contamination parameter $\varepsilon = 0$, 0.01, 0.05, 0.1, 0.2, and 0.3.

As expected, for $\alpha = 0$ we get the exact MLE, hence eref$(\tilde{\sigma})$ is always equal to 1. For $\varepsilon = 0$, i.e. the uncontaminated data, the subdivergence estimators more or less correspond with the maximum likelihood estimators, but they do not outperform them. With rising value of parameter α also the standard deviation $s(\tilde{\sigma})$ rises a little, which causes a certain loss of efficiency. For $\varepsilon > 0$ and escort parameters $\sigma = 1$, 1.2, 1.5, and 2 we observe a loss of consistency, however the MLE loses its consistency too, and with rising contamination we see that the max$\underline{\mathrm{D}}_\alpha$-estimators possess lower values of means and standard deviations then MLE and their performance is therefore better. The best results were received for escort parameter $\sigma = 0.5$. Here, the estimates retained the consistency even for highly contaminated data, showed substantially lower values of $m(\tilde{\sigma})$ and $s(\tilde{\sigma})$, which resulted in high empirical relative efficiency (cf. Table 1).

The max$\underline{\mathrm{D}}_\alpha$-estimators for the other mixtures behave very much the same, the described behavior only gets stronger with contamination by distant outliers. For data contaminated by $N(0, 100)$ and $C(0, 1)$, the subdivergence estimators perform better then MLE even for very small level of contamination $\varepsilon = 0.01$ and all values of escort parameter σ.

As in the location case, we tried to escort the subdivergence estimator with the MLE $\sigma = \boldsymbol{S}_n$. However, we received only a perfect match with maximum likelihood estimator, showing no robusness whatsoever. This motivated us to plug in a simple and robust estimate of scale called median absolute deviation (MAD), which showed up to be a better choice. For this type of estimators we observe that while the power parameter α moves away from 0, the values of $m(\tilde{\sigma})$ decrease and the values of $s(\tilde{\sigma})$ increase. This causes the efficiency to rise at first and than fall down slowly (cf. Table 2). However, by direct comparison with Table 1 we see that the performances of these estimators are not as good as those of the estimators escorted by $\sigma = 0.5$.

α/n	50			100			200			500		
	$m(\tilde{\sigma})$	$s(\tilde{\sigma})$	$eref(\tilde{\sigma})$	$m(\tilde{\sigma})$	$s(\tilde{\sigma})$	$eref(\tilde{\sigma})$	$m(\tilde{\sigma})$	$s(\tilde{\sigma})$	$eref(\tilde{\sigma})$	$m(\tilde{\sigma})$	$s(\tilde{\sigma})$	$eref(\tilde{\sigma})$
0.00	1.336	0.256	1.000	1.326	0.175	1.000	1.336	0.125	1.000	1.336	0.079	1.000
0.01	1.297	0.214	1.326	1.291	0.147	1.288	1.296	0.104	1.301	1.297	0.065	1.288
0.05	1.207	0.147	2.757	1.206	0.103	2.580	1.205	0.073	2.705	1.207	0.044	2.671
0.10	1.156	0.125	4.450	1.155	0.088	4.321	1.152	0.062	4.749	1.153	0.038	4.789
0.20	1.114	0.121	6.446	1.110	0.083	7.210	1.105	0.059	8.763	1.106	0.036	9.454
0.50	1.089	0.143	6.302	1.079	0.097	8.780	1.072	0.069	12.97	1.072	0.043	16.94

Table 1: The evaluation characteristics of $\max\underline{\mathrm{D}}_\alpha$-estimators for mixture $0.9N(0,1) + 0.1N(0,9)$ and escort parameter $\sigma = 0.5$

α/n	50			100			200			500		
	$m(\tilde{\sigma})$	$s(\tilde{\sigma})$	$eref(\tilde{\sigma})$	$m(\tilde{\sigma})$	$s(\tilde{\sigma})$	$eref(\tilde{\sigma})$	$m(\tilde{\sigma})$	$s(\tilde{\sigma})$	$eref(\tilde{\sigma})$	$m(\tilde{\sigma})$	$s(\tilde{\sigma})$	$eref(\tilde{\sigma})$
0.00	1.336	0.256	1.000	1.326	0.175	1.000	1.336	0.125	1.000	1.336	0.079	1.000
0.01	1.176	0.204	2.464	1.132	0.140	3.711	1.106	0.103	5.911	1.085	0.073	9.559
0.05	1.108	0.191	3.688	1.082	0.147	4.806	1.072	0.113	7.093	1.072	0.091	8.853
0.10	1.088	0.199	3.776	1.071	0.162	4.381	1.067	0.130	6.005	1.071	0.106	7.283
0.20	1.073	0.215	3.450	1.062	0.180	3.761	1.065	0.154	4.618	1.070	0.126	5.750
0.50	1.064	0.252	2.638	1.053	0.211	2.895	1.062	0.191	3.181	1.070	0.158	3.992

Table 2: The evaluation characteristics of $\max\underline{\mathrm{D}}_\alpha$-estimators for mixture $0.9N(0,1) + 0.1N(0,9)$ and escort parameter $\sigma = MAD$

4.3 Results for power superdivergence estimators

For the power superdivergence estimators of location, as well as in case of $\max\underline{\mathrm{D}}_\alpha$-estimators of location escorted by $\mu = \bar{\boldsymbol{X}}_n$, we received a perfect match with maximum likelihood estimatior for all mixtures except for the mixture $(1-\varepsilon)N(0,1) + \varepsilon C(0,1)$. In this case again, with higher contamination the $\min\bar{\mathrm{D}}_\alpha$-estimators show favourable robusness and rising efficiency. Yet, these robustness tendencies are not as strong as with power subdivergence estimators escorted by $\mu = 0$ or $\mu = \bar{\boldsymbol{X}}_n$.

When estimating the scale parameter, we also received estimates that very well coincided with the MLE, even in the case of contamination with Cauchy distribution. These estimates showed no robustness at all.

Another demotivating feature of superdivergence estimators computation is extremely high computing time caused by double optimization. This price is too high to pay for the above mentioned robustness, and it strongly discourages the users from further utilization.

5 Conclusion

To resume, the power subdivergence estimators of location do not possess the required properties except for the case with escorting parameter $\mu = 0$ or $\mu = \bar{\boldsymbol{X}}_n$. These estimates show consistency copying maximum likelihood estimates, and they exhibit high empirical relative efficiency and considerable robustness. This resistance to distant outliers together with consistency and efficiency is a key result of our simulation, and it motivates us to explore the max$\underline{\mathrm{D}}_\alpha$-estimators further.

The power subdivergence estimators of scale also suffer by the loss of consistency but their performance is usually better than that of the maximum likelihood estimator, especially when the data are contaminated by distant outliers. The best results were received for the cases with escorting parameter $\sigma = 0.5$ where the consistency was retained and the efficiency was high due to the decrease of means and standard deviations of the subdivergence estimators. Some very good results were obtained also for escort parameter equal to median absolute deviation, but the performances surprisingly were not better those of the subdivergence estimators escorted by $\sigma = 0.5$. This also brings up many questions and inspires us for the future research.

The power superdivergence estimators of location are equivalent to standard maximum likelihood estimator, apart from the cases of high contamination by heavy-tailed distribution. Here it displays certain robustness which, however, does not overly impress and which does not compensate the extraordinary computational demands. The power superdivergence estimators of scale also very well correspond with the standard maximum likelihood estimator.

References

[1] M. Broniatowski and A. Keziou. *Minimization of ϕ-divergences on sets of signed measures*. Studia Scientiarum Mathematica Hungarica, **vol. 43**, (2006), 403–442.

[2] M. Broniatowski and I. Vajda. *Several Applications of Divegence Criteria in Continuous Families*. Research Report No. 2257, Institute of Information Theory and Automation, Prague, (2009).

[3] I. Frýdlová. *Minimum Kolmogorov distance estimators*. Thesis, Czech Technical University, Prague, 2004.

[4] I. Frýdlová. *Modified Power Divergence Estimators: Performances in Location Models*. Research Report No. 2258, Institute of Information Theory and Automation, Prague, (2009).

[5] F. Liese and I. Vajda. *On divergences and informations in statistics and information theory*. IEEE Transactions on Information Theory, **vol. 52**, No. 10, (2006), 4394–4412.

Minimum Distance Estimate*

Jitka Hanousková

3rd year of PGS, email: kj@email.cz
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Václav Kůs, Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. Minimum distance density estimates (MDE) are considered. Via numerical simulation, robustness and consistency of many types of MDE are examined. We consider Kolmogorov, Lévy, discrepancy, and Cramer–von Misses distances. For all but last distances we have proven consistency of the order $n^{-1/2}$ in L_1–norm if the sample is non–contaminated. Graphs for contaminated case are presented and discussed. Further, new type of MDE are introduced, namely, with generalized Cramer–von Mises (GCM) and Kolmogorov-Cramer ($KC_{\alpha,m}$) distance. Various types of GCM estimates are simulated and results are presented and discussed. As results of simulation show, the new defined estimates possess some robustness and consistency even for heavily contaminated distributions (35% contamination).

Keywords: minimum distance estimate, consistency, Cramer-von Mises estimate

Abstrakt. Zkoumáme odhady s minimální vzdáleností (MDE). Pomocí numerické simulace je zkoumána konzistence a robustnost těchto odhadu. Uvažujeme odhady s minimální Kolmogorovskou, Lévyho, diskrepanční a Cramer-von Mises vzdáleností. Pro všechny až na poslední zmíněný máme teoreticky dokázanou konzistenci a řád konzistence $n^{-1/2}$ v L_1–normě a střední hodnotě L_1–normy pro neznečištěnou distribuci. Grafy pro znečištěný případ jsou prezentovány a diskutovány. Dále jsou zavedeny dva nové odhady s minimální vzdáleností, jmenovitě se zobecněnou Cramer-von Mises vzdáleností a Kolmogorov-Crame vzdáleností. Výsledky simulací ukazují, že oba nově zavedené odhady vykazují robustnost a konzistenci i pro velká znečištění.

Klíčová slova: odhady s minimální vzdáleností, konzistence, Cramer-von Mises odhad

1 Introduction

This paper focuses on the minimum distance density estimates. Consistency and robustness of many type of the minimum distance estimates are explored. For non contaminated case Kus [10] has proven conditions for consistency of the order $n^{-1/2}$ in (expected) L_1-norm for Kolmogorov estimate. Hanousková [8] has weakened this conditions and extended them on Lévy and discrepancy estimate. Further, Cramer-von Mises and newly defined generalized Cramer-von Mises and Kolmogorov-Cramer estimates are explored via simulation. We consider non contaminated and contaminated distributions and explore consistency and robustness of all above mentioned estimates.

*This work has been supported by the grant SGS OHK4–007/10.

2 Basic Concepts

We introduce the notation used in the following text. Let λ be a σ-finite measure on $(\mathbb{R}, \mathcal{B})$, where $\mathcal{B}$ is a borel σ-field on $\mathbb{R}$. Let $\mathcal{F}_\lambda$ be the set of distributions on $(\mathbb{R}, \mathcal{B})$ which are absolutely continuous with respect to the measure λ. Let us denote by $\mathcal{D}_\lambda$ the set in Banach space $\mathrm{L}_1(\mathbb{R}, \mathrm{d}\lambda)$ containing densities corresponding to distribution functions in $\mathcal{F}_\lambda$, by $\mathcal{D}$ arbitrary nonvoid subset of $\mathcal{D}_\lambda$, and by $\mathcal{F}$ a respective subset of $\mathcal{F}_\lambda$. Further, $\mathbb{X}_n = (X_1, \ldots, X_n)$ denotes a random vector with independent components distributed by a density f. Next, $F_n(x)$ represents the empirical distribution function based on $\mathbb{X}_n$, similarly $\nu_n(B)$ stands for the empirical measure

$$F_n(x) = \frac{1}{n} \sum_{j=1}^{n} \mathbf{I}_{\{X_j \le x\}}\,,\ x \in \mathbb{R}, \qquad \nu_n(B) = \frac{1}{n} \sum_{j=1}^{n} \mathbf{I}_{\{X_j \in B\}}\,,\ B \in \mathcal{B}\,, \tag{1}$$

where $\mathbf{I}_{\{X_j \le x\}}$ and $\mathbf{I}_{\{X_j \in B\}}$ stand for indicators of the corresponding events.

Definition 1. Let d be a distance on a given set of probability measures $\mathcal{P}$. An estimate $\widehat{P}_n$ of a measure $P \in \mathcal{P}$ (for nonparametric model), or an estimate $\widehat{\theta}_n \in \Theta$ of a parameter θ (for parametric model) is called minimum d-distance estimate if it holds that

$$\widehat{P}_n = \arg\min_{P \in \mathcal{P}} d(P, \nu_n) \quad \text{a. s.}\,, \tag{2}$$

$$\widehat{\theta}_n = \arg\min_{\theta \in \Theta} d(P_\theta, \nu_n) \quad \text{a. s.}\,, \tag{3}$$

and $\widehat{P}_n$, $\widehat{\theta}_n$ exist. If the measures from $\mathcal{P}$ are absolutely continuous with respect to the measure λ, then the density $\widehat{f}_n$ corresponding to the measure $\widehat{P}_n$ ($\widehat{f}_n = \mathrm{d}\widehat{P}_n/\mathrm{d}\lambda$) provides the minimum d distance estimate of probability density f corresponding to the measure $\mathcal{P}$.

In the following text ρ_d denotes a distance between two probability densities $f, g \in \mathcal{D}_\lambda$ defined by $\rho_d(f, g) = d(P, Q)$, where $P, Q \in \mathcal{P}$ are the corresponding probability measures. If the distance d is a metric, ρ_d need not be a metric. We use the concept of equivalence classes ($f \sim g$ iff $\rho_d(f, g) = 0$) in order to convert ρ_d into a metric.

We deal with Kolmogorov (ρ_K), total variation (ρ_V), discrepancy (ρ_D), Cramer-von Mises (ρ_{C-M}), and Lévy (ρ_L) distances. And further, we define generalized Cramer-von

Mises distance (ρ_{GCM}).

$$\rho_K(f,g) = \sup_{x\in\mathbb{R}} |F(x) - G(x)|, \tag{4}$$

$$\rho_V(f,g) = \int_{\mathbb{R}} |f - g| \mathrm{d}\lambda, \tag{5}$$

$$\rho_D(f,g) = \sup_{B\in\mathbf{B}} |P(B) - Q(B)|, \tag{6}$$

$$\rho_{C-M}(f,g) = \int_{\mathbb{R}} (F(x) - G(x))^2 \mathrm{d}F(x), \tag{7}$$

$$\rho_L(f,g) = \inf\{\varepsilon > 0 : G(x-\varepsilon) - \varepsilon \leq F(x) \leq G(x+\varepsilon) + \varepsilon\}, \tag{8}$$

$$\rho_{GCM}(f,g) = \int_{\mathbb{R}} |F(x) - G(x)|^\alpha \mathrm{d}F(x), \tag{9}$$

where $\mathbf{B}$ is the set of all closed balls, and P, Q are probability measures with distribution functions F, G corresponding to the densities f, g. And α is a non-negative real parameter of generalized Cramer-von Mises distance. Obviously, for the choice of parametr $\alpha = 2$ the generalized Cramer-von Mises distance converts to the original Cramer-von Mises distance.

Further, we define new type of minimum distance estimator. We define so called Kolmogorov-Cramer distance with parameter α, m ($d_{KC_{\alpha,m}}$) between empirical distribution function and arbitrary distribution function in the following way

Definition 2. Let $(x_1, ..., x_n)$ be a realization of random vector $\mathbb{X}_n$, α real parameter, and F arbitrary distribution function then a sequence $(G_i)_1^{2n}$ is defined as

$$G_i = |F_n(x_i) - F(x_i)|^\alpha \text{ pro } i = 1, ..., n \tag{10}$$

$$G_{2n+1-i} = |F_{n-}(x_i) - F_-(x_i)|^\alpha \text{ pro } i = 1, ..., n\,. \tag{11}$$

Where $F_{n-}(x_i) = \lim_{x\to x_i-} F_{n-}(x)$ and simillar $F_-(x_i) = \lim_{x\to x_i-} F_-(x)$. Then Kolmogorov-Cramer distance is defined

$$d_{KC_{\alpha,m}}(F_n, F) = \frac{1}{m}\sum_{i=1}^{m} G_{(i)}\,, \tag{12}$$

where $G_{(i)}$ denotes arranging in order of size.

In other words, we can say that the Kolmogorov-Cramer distance is one over m times sum of m largest of the G_i, $i = 1, ..., 2n$. We call the distance Kolmogorov-Cramer, because for $m = 1$ it converts to Kolmogorov distance to the pover α and the shape is inspired by Cramer-von Mises distance.

Definition 3. We say that an estimate $\widehat{f_n}$ of a density f is consistent in the given ρ_d distance (in the expected ρ_d distance) iff $\rho_d(\widehat{f_n}, f) \to 0$ a.s. ($E\rho_d(\widehat{f_n}, f) \to 0$). We say that the estimate $\widehat{f_n}$ is consistent of the order $r_n \to 0$ in the distance ρ_d, (in the expected ρ_d distance) iff $\rho_d(\widehat{f_n}, f) = O_p(r_n)$, ($E\rho_d(\widehat{f_n}, f) = O(r_n)$).

3 Consistency in L_1–norm

Kus [10] presents conditions for consistency and for consistency of the order $n^{-1/2}$ of Kolmogorov estimates. Conditions are based on domination relation between Kolmogorov distance and total variation distance. Furthermore sufficient condition for the domination relation is proven ibid. All previous results could be summarized in the following way. If the degree of variation of family $\mathcal{D}$ is finite, then all Kolmogorov estimates of densities from $\mathcal{D}$ are consistent of the order $n^{-1/2}$ in the (expected) L_1–norm. Further, generalization of this theory is given in Hanousková [8]. Namely, new (weaker) conditions for consistency and for consistency of the order $n^{-1/2}$ of Kolmogorov estimates are proven. For this aim an asymptotic domination and a partial degree of variation were defined. The main result could be summarized by following statement. If the partial degree of variation of family $\mathcal{D}$ is finite, then all Kolmogorov estimates of densities from $\mathcal{D}$ are consistent in the (expected) L_1–norm. And if, moreover, some additional, but not as restrictive as finiteness of the degree of variation, assumptions hold, then we gain the order $n^{-1/2}$ of consistency in the L_1–norm and in the expected L_1–norm. For details see Hanousková [8].

Now we explore the consistency and the order of consistency of our minimum distance estimates for the case of distances differing from the Kolmogorov distance. Let us suppose that the asymptotic domination is fulfilled for a family $\mathcal{D} \subset \mathcal{D}_\lambda$. Further, assume that for a given distance d it holds both inequalities $\rho_d \le h_1(\rho_K)$ and $\rho_K \le h_2(\rho_d)$ with two real functions h_1, h_2 continuous at zero point with zero value in zero argument. If there exist positive constants K_i such that $h_i(x) \le K_i x$, $i = 1, 2$, in a neighborhood of zero, then the minimum d distance estimates of densities from $\mathcal{D}$ are consistent of the order $n^{-1/2}$ in L_1–norm and the expected L_1–norm. (See Hanousková [8] for details). The above stated conditions are satisfied for example for Lévy and discrepancy distances, i.e.

$$\rho_K(f,g) \le \rho_D(f,g) \le 2\rho_K(f,g)\,, \tag{13}$$

$$\rho_L(f,g) \le \rho_K(f,g) \le (1+\sup|G'|)\rho_L(f,g)\,. \tag{14}$$

For specific inequalities derived in spaces of probability densities, see Gibbs & Su [5].

Moreover, we were able to prove consistency and $n^{-1/2}$ order of consistency for Kolmogorov-Cramer estimate even though the upper mentioned inequalities do not hold. For Cramer-von Mises and generalized Cramer- von Mises distances we failed to show the desired inequalities leading to $n^{-1/2}$ consistency. Nevertheless, we hoped for consistency of this two estimates, so we produced numerical simulation to ascertain. Via simulation we study robustness of all upper mentioned estimates.

4 Numerical simulation

Further, we considered normal distribution $N(0,1)$ contaminated by the normal distribution $N(0,100)$ and explore consistency of all above mentioned minimum distance estimators. For contaminated case we have not theoretical results guaranteeing the consistency. Thus, this case is explored only via simulation. Normal distributions with 5%, 10%, 15% contamination were considered and the consistency results (and related robustness

as well) are presented in Figure 1 for Kolmogorov, Lévy, discrepancy, and Cramer-von Mises estimate and Figures 2, 3 present consistency results for generalized Cramer-von Mises estimates with parameter $\alpha = 0.25, 0.5, 0.75, 1.0, 1.2, 1.4, 1.6, 1.8$.

The consistency is now getting worse with increasing contamination for the Kolmogorov, Lévy, and discrepancy metrics. Under 10% and 15% contamination all estimates, except for Cramer–von Mises and generalized Cramer-von Mises, achieve the best L_1–error for $n = 50$ or $n = 100$ observations and for greater sample sizes the L_1–norm slightly increases possibly to a limiting value. Nevertheless, it can be seen from Figure 1 that the Cramer–von Mises and all examined types of generalized Cramer-von Mises estimates seem to preserve good consistency even under 5%,10%, and 15% contamination. Thus, Cramer–von Mises and generalized Cramer-von Mises estimates show some robust behavior.

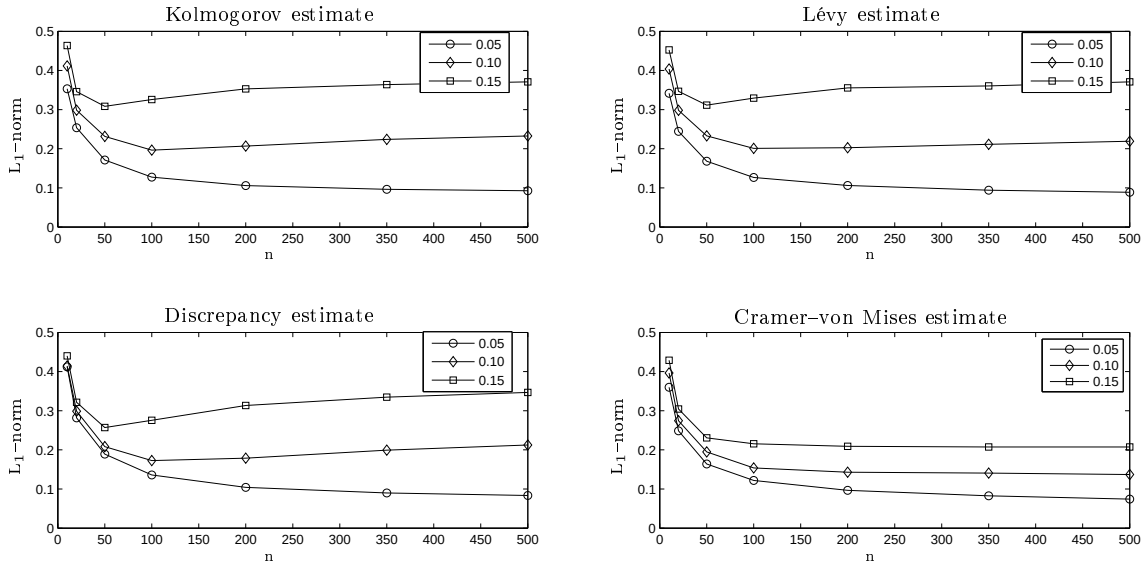


Figure 1: L_1–error of MD estimates for Normal distribution with parameters $\mu_0 = 0$, $\sigma_0^2 = 1$ contaminated by Normal distribution with parameters $\mu = 0$, $\sigma^2 = 100$.

Regarding the robustness, it can be seen directly from the Figures 1, 2, 3 that Cramer–von Mises type of estimate has the most robust behavior of all for sample sizes greater than 200. Further, we want to determine which choice of parameter α is the best (in sence of robustness and consistency, too). However, the shapes of graphs for Cramer-von Mises type of estimates are very similar and it is not easy to determine which choice of α is the best. Therefor Figure 4 presents average absolute error of estimated parameter (σ) with respect to the true value of parameter (σ_0) for 5%,10%, 15%, and 35% contamination. As it can be seen the best result for 5% contamination is achieved for parameter $\alpha = 0.5, 0.75, 1.0$, for 10%,15%, and 35% contamination is the best result achieved for $\alpha = 0.25$. For the 5% contamination case the trend is not monotonic. The line is serrated with local extremes (minims). For 10% contamination case the average absolute error is monotonically decreasing for $\alpha \leq 1$. For all but 5% and 10% contamination cases

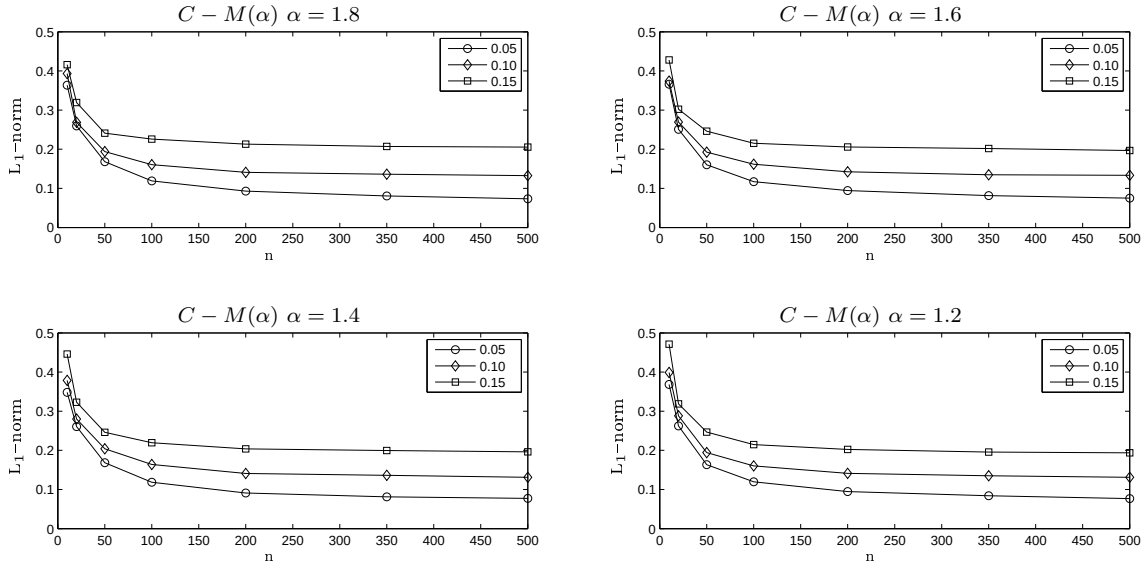


Figure 2: L_1–error of generalized Cramer-von Mises estimates for Normal distribution with parameters $\mu_0 = 0$, $\sigma_0^2 = 1$ contaminated by Normal distribution with parameters $\mu = 0$, $\sigma^2 = 100$.

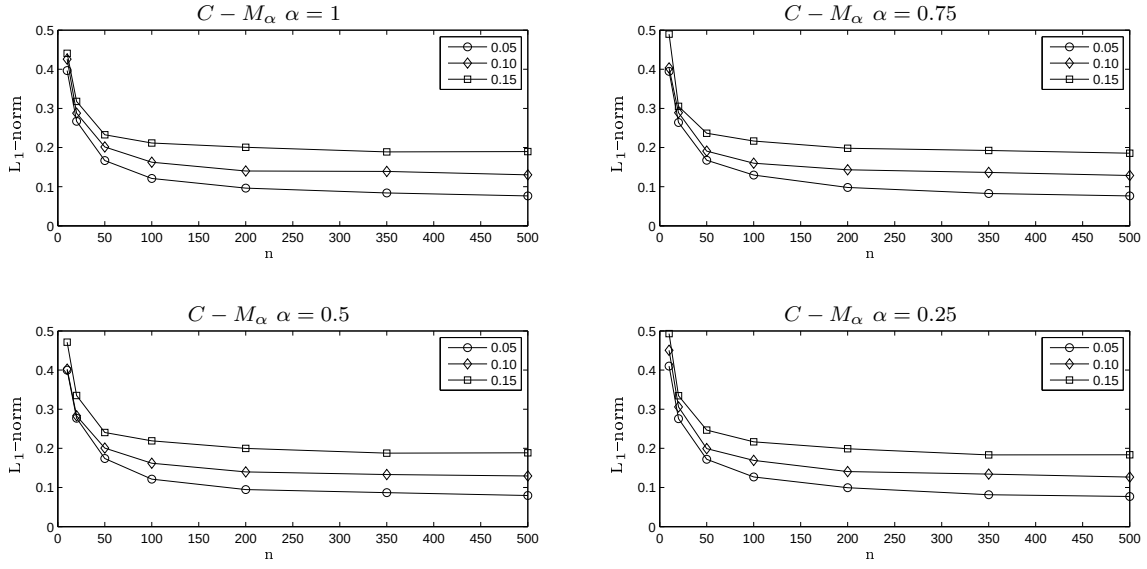


Figure 3: L_1–error of generalized Cramer-von Mises estimates for Normal distribution with parameters $\mu_0 = 0$, $\sigma_0^2 = 1$ contaminated by Normal distribution with parameters $\mu = 0$, $\sigma^2 = 100$.

decreases the average absolute error monotonically if the α tends to zero. It could lead us to the idea that the smaller parameter α we choose the more robust estimator we

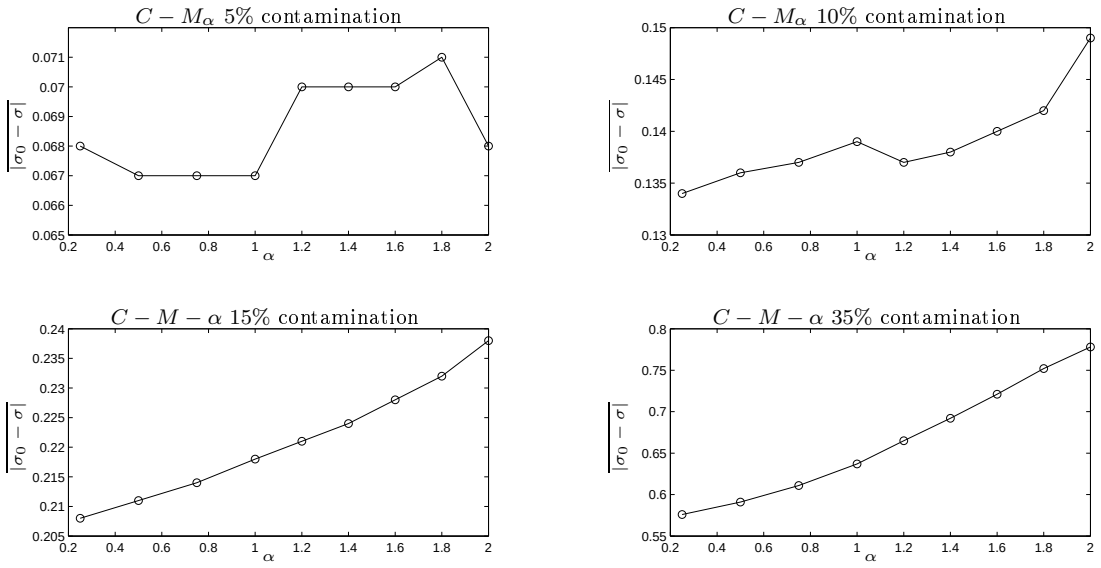


Figure 4: Average absolute error of estimated parameter with respect to the true value of parameter $\overline{|\sigma_0 - \sigma|}$ of various types of generalized Cramer-von Mises estimates.

gain. However, parameter α could not be taken as small as possible. Naturally, there is a threat of loosing efficiency due to gaining robustness. For more accurately determining of the best choice of parameter α to obtain robust and efficient estimate a comprehensive simulation study would be beneficial.

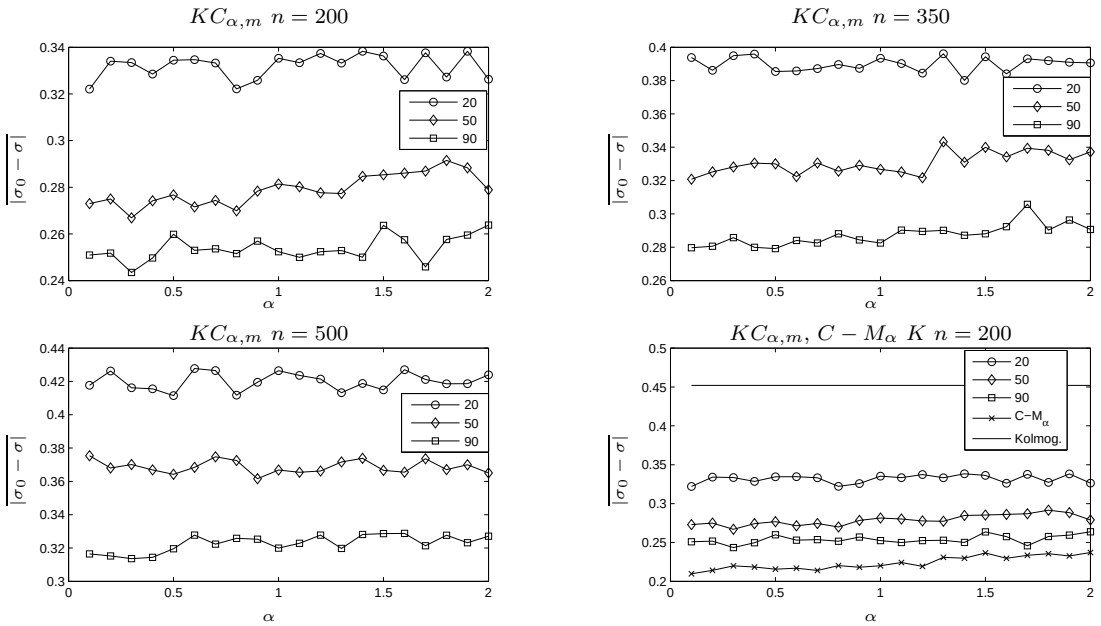


Figure 5: Average absolute error of estimated parameter with respect to the true value of parameter $\overline{|\sigma_0 - \sigma|}$ of Kolmogorov-Cramer estimates for various sample sizes.

Further we examine the robustness property for newly defined Kolmogorov-Cramer estimate. Theoretical proof guarantee us the consistency and order of consistency of this estimate. For robustness we have no theoretical results; our investigation is based on simulation. Figure 5 presents average absolute error of estimated parameter (σ) with respect to the true value of parameter (σ_0) for 15% contamination and various sample sizes ($n = 200, 350, 500$). As can be seen from the Figure 5 the bigger the parameter m is the smaller absolute error we gain for all examined sample sizes. And as can be seen from the last graph in Figure 5 the Kolmogorov estimate has the biggest absolute error, and the smallest absolute error has the Cramer-von Mises estimate with parameter α. The Kolmogorov-Cramer estimate's absolute errors lie between absolut error of Kolmogorov and Cramer-von Mises estimate. The result of this simulation study is following, the bigger parametr m we choose the smaller absolute error we gain. It means that the Kolmogorov estimate is the less robust and Cramer-von Mises is the most robust of our examined estimator. Newly defined Kolmogorov-Cramer estimator create transition from Kolmogorov to Cramer-von Mises estimate. However, for Cramer-vonmises estimate we have not proven the consistency (and order of consistency) and for Kolmogorov and Kolmogorov-Cramer estimate we have theoretical results for non contaminated distribution.

5 Conclusion

Via numerical simulation we explored consistency of Cramer-von Mises and generalized Cramer-von Mises estimates. Further, consistency on contaminated distribution was explored. Kolmogorov, Lévy, and discrepancy estimate loosed their consistency on contaminated sample. On the other hand Cramer-von Mises type of estimate preserve some consistency even under heavier contamination. Moreover, we determine the best choice of parameter α of generalized Cramer-von Mises estimate in sense of robustness and consistency too. As the best choice was determined case $\alpha = 0.25$, the smallest of explored choices (Choice of parameter α inside interval $< 0.25, 2 >$) were explored). However the possible change of efficiency was not explored. Thus, more detail simulation study will be beneficial. In the end the robustness of newly defined Kolmogorov-Cramer estimate was explored via simulation with this results. The bigger parameter m we choose the more robust the estimate is, i.e. the most robust of Kolmogorov-Cramer estimate is the Cramer-von Mises estimate. However, for Cramer-von Mises estimate we have no theoretical results about consistency for non contaminated distribution. More detailed numerical study and would be beneficial. Moreover, there is a chance to extend the proof of the order of consistency of Kolmogorov-Cramer estimate to case when the parametr m depends on sample size n.

References

[1] L. Devroye, L. Györfi. *Nonparametric Densiti Estimation: The L_1-View.* Wiley, New York, (1985).

[2] L. Devroye, L. Györfi, G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer, New York, (1996).

[3] L. Friedrich, I. Vajda. On Divergences and Informations in Statistics and Information Theory. *IEEE Transactions on Information Theory*, **52**, (2006), 4394-4412.

[4] I. Frýdlova. Odhady pravděpodobnostních hustot s minimální Kolmogorovskou vzdáleností. *FJFI ČVUT Praha*, (2004).

[5] A. L.Gibbs, F. E. Su. On choosing and bounding probability metrics. *International Statistical Review*, **70**, (2002), 419–435.

[6] L. Györfi, I. Vajda,E. C. van der Meulen. Family of point estmates yielded by L_1-consistent density estimate. *L1-Statistical Analysis and Related Methods*, (1992), 415-430.

[7] L. Györfi, I. Vajda,E. C. van der Meulen. Minimum Kolmogorov Distance Estimates of Parameters and Parametrized Distributions. *Metrika*, **43**, (1996), 237-255.

[8] J. Hanousková. Asymptotické vlastnosti odhadu s minimální Kolmogorovskou vzdáleností. *FJFI ČVUT Praha*, (2009).

[9] A. J. Izenman. Recent Developments in Nonparametric Density Estimation. *Journal of the American Statistical Association*, **86**, (1991), 205-224.

[10] V. Kus. Nonparametric density estimates consistent of the order of $n^{-1/2}$ in the L_1-norm. *Metrika*, **60**, (2004), 1–14.

[11] L. Pardoi. *Statistical Interference Based on Divergence Measures.* Taylor and Francis Group, LLC, Boca Raton, (2006).

Phase-Field Approach to Crystal Growth

Hung Hoang Dieu

3rd year of PGS, email: hoangdieu@fjfi.cvut.cz
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Michal Beneš, Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. The phase-field method has appeared in the context of diffuse interfaces. It has been applied to the three major materials processes: solidification, solid-state phase transformation, and grain growth and coarsening. Very recently, a number of new phase-field models have been developed for modelling thin films and surfaces (see [4]). The first part of this contribution is concerned with the phase-field model of spiral crystal growth [7] described by the Burton-Cabrera-Frank theory [2]. Here, we investigate the influence of numerical parameters on the growth patterns. We then present computational studies related to the pattern formation and to the dependence on model parameters. The second part is concerned with the phase-field model [9, 10] of heteroepitaxial growth. Finally, we present our latest results.

Keywords: phase-field method, spiral growth, heteroepitaxial growth, ATG instability, FDM, FEM

Abstrakt. Metoda phase-field se objevila v souvislosti s difuzními rozhraními a byla aplikována na tři hlavní procesy v materiálech: tuhnutí, fázový přechod pevné látky a růst zrn. V současné době řada nových modelů phase-field byla vyvinuta pro modelování povrchů a tenkých vrstev (viz [4]). První část tohoto příspěvku se týká modelu phase-field pro spirálový růst krystalů [7], popsaný Burton-Cabrera-Frankovou teorií [2]. Zde zkoumáme vliv numerických parametrů na růst krystalů, výsledky jsou pak prezentovány. Druhá část se zabývá modelem phase-field pro heteroepitaxní růst, založeným na [9] a [10]. Nakonec prezentujeme naše nejnovější výsledky.

Klíčová slova: metoda phase-field, spirálový růst, heteroepitaxní růst, ATG nestabilita, metoda sítí, metoda konečných prvků

1 Spiral Crystal Growth

1.1 The Model

Crystallization is the process where solid crystals are formed from melt, solution, or vapour phase. There are two major stages involved in the crystallization process – *nucleation* and *crystal growth*. Nucleation is the stage where crystal forming units (atoms, ions or molecules) gather into clusters which are unstable until they reach a critical size. Stable clusters are called nuclei. After nuclei are created, crystal growth begins. It is the stage where new crystal forming units are incorporated into the crystal lattice.

Real crystals contain dislocations which are crystallographic defects in the structure of the crystal lattice. The presence of dislocations influences the mechanism of crystal growth. If a screw dislocation is present in the crystal lattice of the substrate, a step with

a zero height at the dislocation core is created. This step winds around the dislocation and produce a spiral.

Classically epitaxial crystal growth is modelled using Burton-Cabrera-Frank (BCF) theory (see [2]). According to that theory, atoms are first adsorbed to the crystalline surface. Such atoms are called adatoms. They then diffuse freely along the surface and they can either desorb from the surface with a probability $1/\tau_S$ per unit time, or they are incorporated into the crystal at one of the three sites: ledge site, step site or kink site. Incorporation at a kink site will be the most energetically favourable.

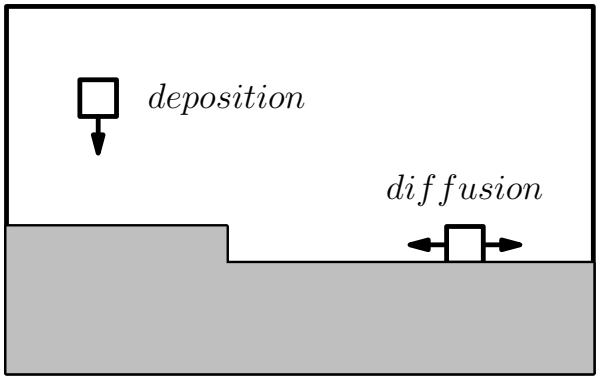


Figure 1: Burton-Cabrera-Frank model.

The basic equations in the phase-field formulation [7] of BCF model are

$$\partial_t c = D\Delta c - \frac{c}{\tau_S} + F - \Omega^{-1}\partial_t \Phi, \tag{1}$$

$$\alpha\partial_t \Phi = \xi^2\Delta\Phi + sin(2\pi(\Phi - \Phi_S)) + \lambda c(1 + cos(2\pi(\Phi - \Phi_S))), \tag{2}$$

where c is the adatom density, D is the surface diffusion coefficient, τ_S is the mean time for the desorption of adatoms from the surface, F is the deposition rate, Φ is the surface height in units of atoms, α is the time relaxation parameter, ξ is the width of steps between terraces, Φ_S is the height of the initial substrate surface and λ is the coupling constant.

The boundary conditions are given by

$$\frac{\partial c}{\partial n}(t, \mathbf{x}) = \frac{\partial \Phi}{\partial n}(t, \mathbf{x}) = 0, t \in (0, T). \tag{3}$$

The initial conditions are given by

$$c(0, \mathbf{x}) = 0, \tag{4}$$

$$\Phi(0, \mathbf{x}) = \Phi_S(\mathbf{x}). \tag{5}$$

1.2 Numerical scheme

We use an explicit scheme of the finite difference method to solve the free boundary problem of spiral crystal growth. The first step in the discretization is to divide the

computational domain into a two-dimensional grid and then derivatives are replaced with equivalent finite differences.

We consider the computational domain S to be a rectangle $(0, L_1) \times (0, L_2)$ which is to be discretized. We partition the domain S using a grid of internal nodes $\omega_h = \{(ih_1, jh_2) | i = 1, ..., N_1 - 1, j = 1, ..., N_2 - 1\}$, where $h_1 = \frac{L_1}{N_1}, h_2 = \frac{L_2}{N_2}$ are the mesh sizes in S. We discretize the time interval using a mesh $[0, T] : T_\tau = \{k\tau | k = 0, ..., N_T\}$, where $\tau = \frac{T}{N_T}$ is a time step. Then we can consider a grid function $u : T_\tau \times \omega_h \to \mathbb{R}$ for which $u^k_{ij} = u(ih_1, jh_2, k\tau)$.

The time derivative is approximated by forward difference

$$\partial_t u^k_{ij} \approx \frac{u^{k+1}_{ij} - u^k_{ij}}{\tau},$$

and the space derivatives are approximated by second-order central differences:

$$\partial^2_x u^k_{ij} \approx \frac{u^k_{i+1,j} - 2u^k_{ij} + u^k_{i-1,j}}{h_1^2},$$
$$\partial^2_y u^k_{ij} \approx \frac{u^k_{i,j+1} - 2u^k_{ij} + u^k_{i,j-1}}{h_2^2}.$$

Then the Laplace operator in two dimensions is given by $\Delta_h u^k{}_{ij} = \partial^2_x u^k_{ij} + \partial^2_y u^k_{ij}$.

The explicit scheme has the form

$$\begin{aligned} \alpha \frac{\Phi^{k+1}_{ij} - \Phi^k_{ij}}{\tau} &= \xi^2 \Delta_h \Phi^k_{ij} + sin(2\pi(\Phi^k_{ij} - \Phi^k_{S_{ij}})) \\ &\quad + \lambda c^k_{ij}(1 + cos(2\pi(\Phi^k_{ij} - \Phi^k_{S_{ij}}))) \qquad (6) \\ \frac{c^{k+1}_{ij} - c^k_{ij}}{\tau} &= D\Delta_h c^k_{ij} - \frac{c^k_{ij}}{\tau_S} + F - \Omega^{-1} \frac{\Phi^{k+1}_{ij} - \Phi^k_{ij}}{\tau} \qquad (7) \end{aligned}$$

for $i = 1, ..., N_1 - 1, j = 1, ..., N_2 - 1, k = 0, ..., N_T$.

Discretization of the epitaxial crystal growth problem leads to a system of equations

$$\begin{aligned} \Phi^{k+1}_{ij} &= \Phi^k_{ij} + \frac{\tau\xi^2}{\alpha} \frac{\Phi^k_{i+1,j} + \Phi^k_{i,j+1} - 4\Phi^k_{ij} + \Phi^k_{i,j-1} + \Phi^k_{i-1,j}}{h^2} \\ &\quad + \frac{\tau}{\alpha} sin(2\pi(\Phi^k_{ij} - \Phi^k_{S_{ij}})) \\ &\quad + \frac{\tau\lambda}{\alpha} c^k_{ij}(1 + cos(2\pi(\Phi^k_{ij} - \Phi^k_{S_{ij}}))) \qquad (8) \\ c^{k+1}_{ij} &= c^k_{ij} + \tau D \frac{c^k_{i+1,j} + c^k_{i,j+1} - 4c^k_{ij} + c^k_{i,j-1} + c^k_{i-1,j}}{h^2} \\ &\quad - \frac{\tau}{\tau_S} c^k_{ij} + \tau F - \frac{\Phi^{k+1}_{ij} - \Phi^k_{ij}}{\Omega} \qquad (9) \end{aligned}$$

for $i = 1, ..., N_1 - 1, j = 1, ..., N_2 - 1, k = 0, ..., N_T$. That means we can obtain the values at time $k + 1$ from the corresponding ones at time k.

For $h = h_1 = h_2$ this explicit method is known to be numerically stable and convergent whenever $\frac{\xi^2\tau}{\alpha h^2} \leq \frac{1}{4}$ and $\tau(\frac{4D_S}{h^2} + \frac{1}{\tau_S}) \leq 1$.

The boundary conditions are treated by mirroring the values in the inner nodes across the boundary.

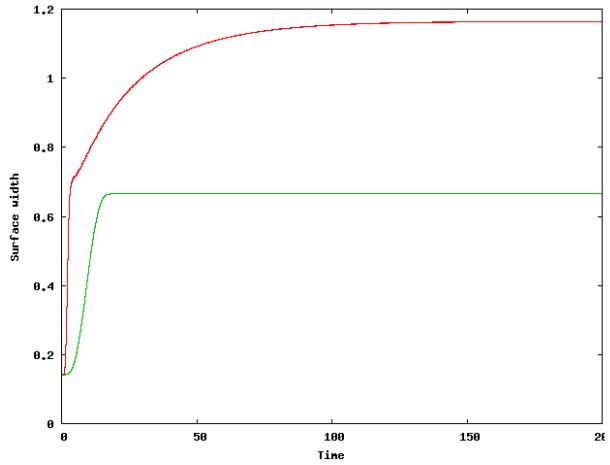


Figure 2: Comparison of transient dynamics for different desorption times. Green (bottom) line: $\tau_S = 0.1$, the surface width quickly levels off and remains constant. Red (top) line: large τ_S, the surface width changes slowly in time.

1.3 Numerical Results

In the numerical experiments, we investigated the influence of the parameter τ_S to the spiral growth. First, transient dynamics is quantified by defining the so called surface width $w(t)$ which represents the mean fluctuation of the surface height

$$w(t) = \frac{1}{2}\langle \Phi(x,t)^2 - \langle \Phi(x,t)\rangle^2\rangle^{1/2},$$

where $\langle f\rangle = L^{-2}\int_S f dx$. $(L = h(N-1) = 50)$ (see Fig. 2).

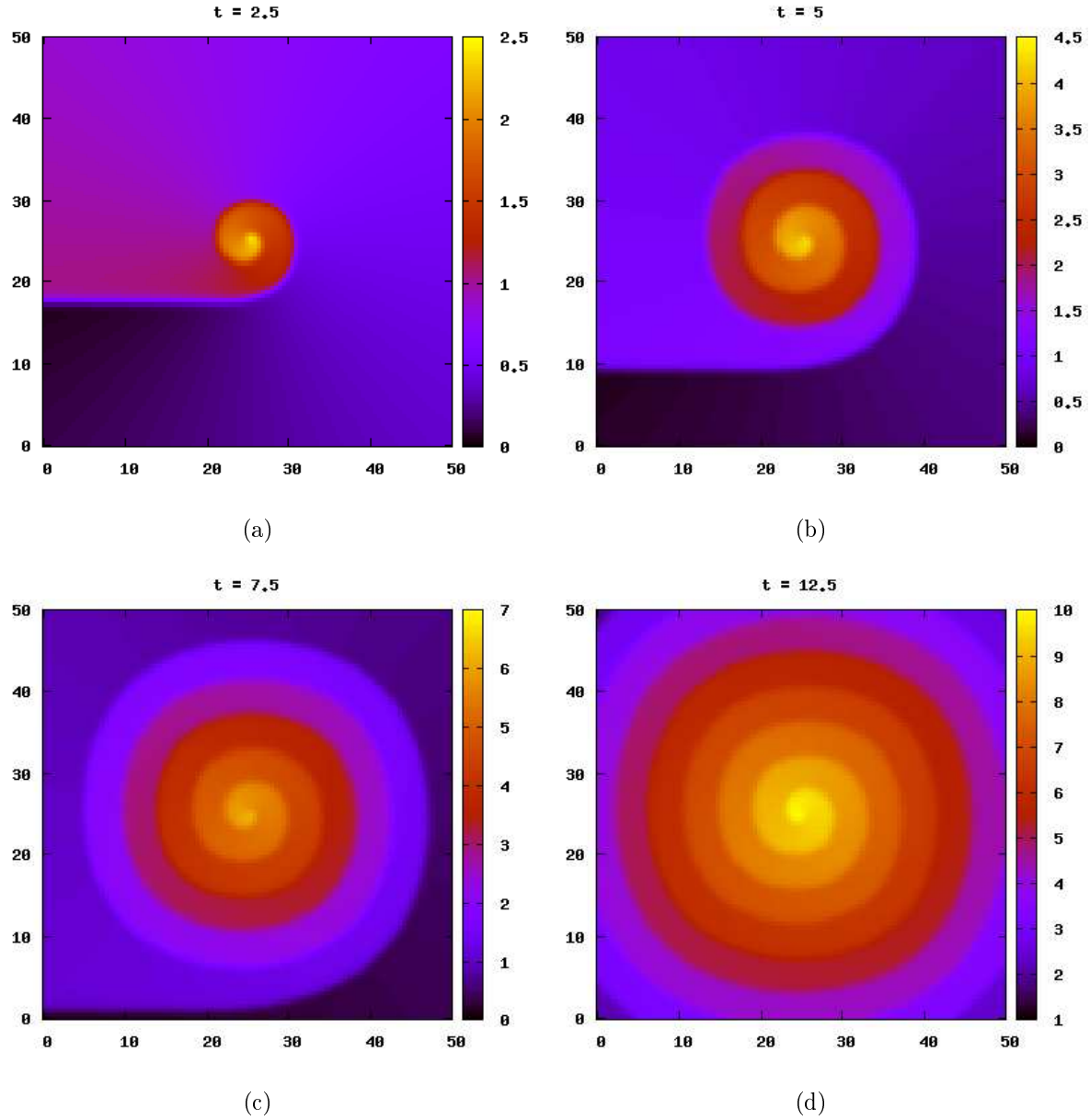
Then, the parameters are set up as follows: $\Omega = 2.0$, $\alpha = 1.0$, $\xi = 1.0$, $\lambda = 10.0$, $D_S = 2.0$, $F = 3.0$, $\tau = 0.00025$, $N_T = 100000$, so that $T = 25$. The dimensions of ω_h are 100×100 and the spatial step size is set to $50/99$. The initial height of the substrate Φ_S is formed by $\frac{arctan(y/x)}{2\pi}$ for the dislocation. We observed two distinguished growth regimes. As can be seen in Fig. 3 for small τ_S, the spiral finds its final step spacing l essentially after a single rotation. In contrast, for very large τ_S the transient spiral ridge evolves slowly towards a spiral with a constant l. This surface evolution is demonstrated in Fig. 4.

From these numerical simulations we conclude that step spacing is dependent on desorption time. The larger desorption time is, the smaller the step spacing is.

2 Heteroepitaxial Growth

Epitaxy refers to the oriented growth of crystalline material onto the single crystal surface. The orientation is determined by the underlying crystal. In general, we distinguish two cases:

- Homoepitaxy – the growth layers of the material and the substrate are of the same chemical composition.



(a) (b)

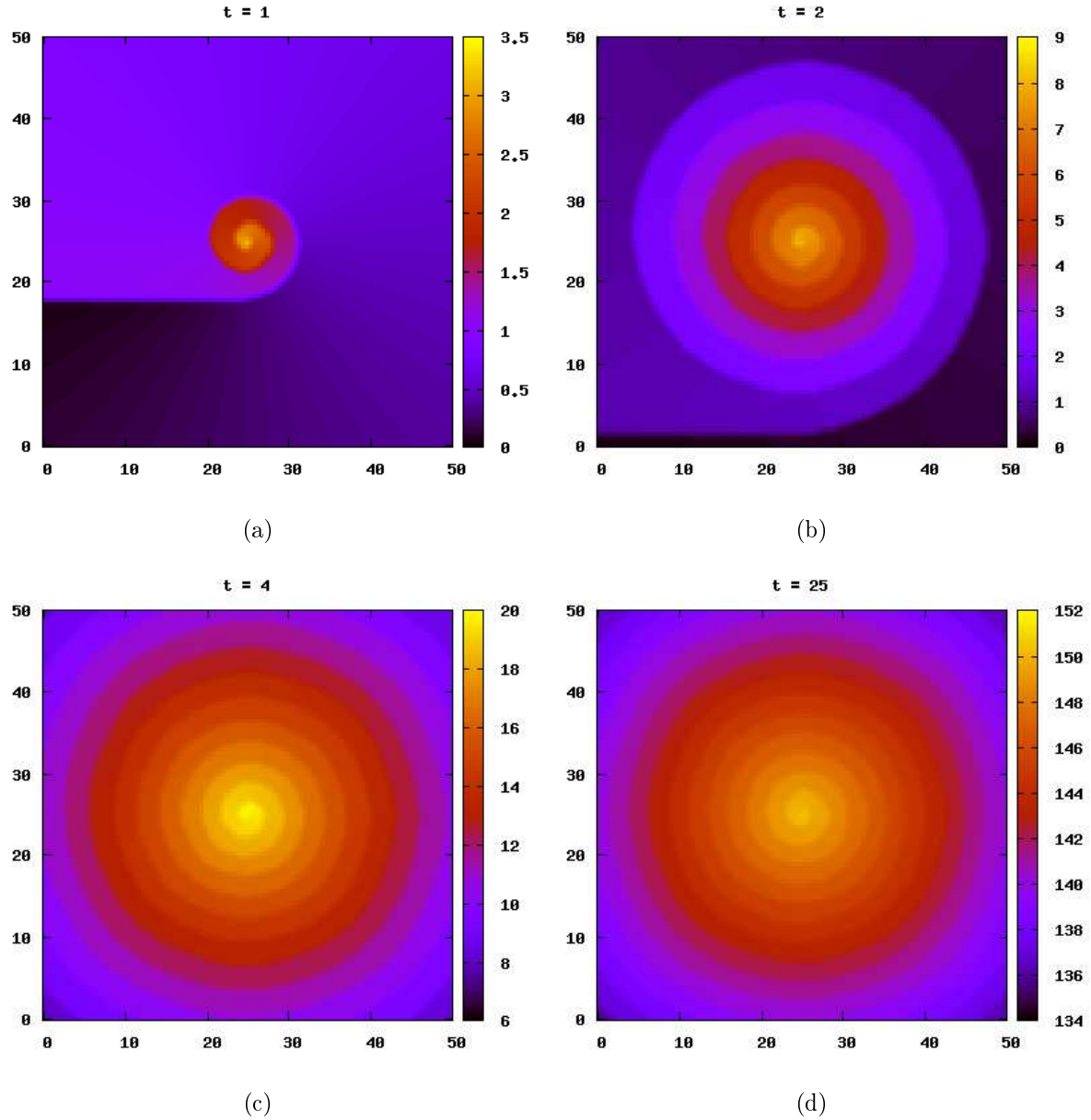
(c) (d)

Figure 3: Spiral ridge at different times t for $\tau_S = 0.1$. Colour palette represents the surface height.

- Heteroepitaxy – the growth layers of the material and the substrate are of the different chemical compositions.

Our aim is to study heteroepitaxial growth which is under misfit stress. This leads to morphological instability (known as Asaro-Tiller-Grinfeld instability).

We consider a system Ω consisting of two regions – a solid epitaxial film $\Omega^e(t)$ and vapour phase $\Omega^v(t)$. The solid-vapour interface is denoted $\Gamma(t)$, which is a function of



(a) (b)

(c) (d)

Figure 4: Spiral ridge at different times t for large τ_S. Colour palette represents the surface height.

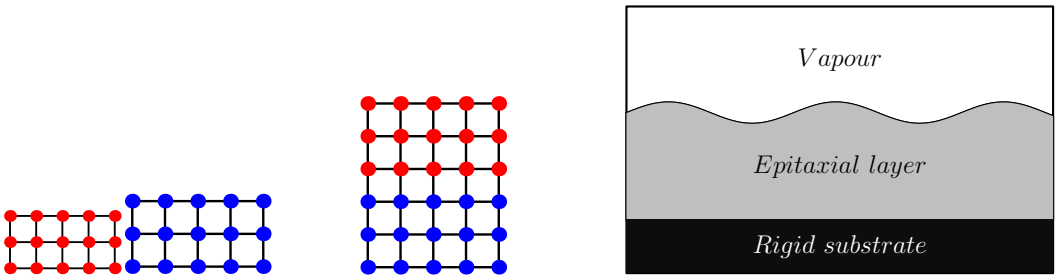
time t (see Fig. 5c). We introduce a non-conserved order parameter

$$\Phi(t, \mathbf{x}) = \begin{cases} 0 & \mathbf{x} \in \Omega^v \\ 1 & \mathbf{x} \in \Omega^e \end{cases} .$$

Here, the linear elastic theory is used. The stress tensor $\sigma_{ij}^{(v)}$ in the vapour is given by Hooke's law

$$\sigma_{ij}^{(v)} = 2\mu^{(v)}\epsilon_{ij} + \lambda^{(v)}\epsilon_{kk}\delta_{ij},$$

where einstein summation convention is implied.



(a) Relaxed state. (b) Strained state. (c) Model description.

Figure 5: Heteroepitaxial growth.

Following [14] the stress tensor $\sigma_{ij}^{(e)}$ in the epitaxial film is given by

$$\sigma_{ij}^{(e)} = 2\mu^{(e)}\epsilon_{ij} + \lambda^{(e)}\epsilon_{kk}\delta_{ij} - \epsilon^m\{\tfrac{1+\nu^{(e)}}{1-2\nu^{(e)}}\}\delta_{ij},$$

where $\mu^{(*)}, \lambda^{(*)}$ are Lamé constants, $\nu^{(*)}$ is Poisson's ratio, where $* \in \{e, v\}$. $\epsilon^m = \frac{a_e - a_s}{a_s}$ is the misfit strain, where a_e, a_s are lattice constants of epitaxial film or substrate. The strain tensor is given by $\epsilon_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right)$, where u_i is the ith component of the displacement vector.

The stress tensor in the system is determined from

$$0 = \frac{\partial}{\partial x_j}\{h(\Phi)\sigma_{ij}^{(e)} - [1 - h(\Phi)]\sigma_{ij}^{(v)}\}, \tag{10}$$

where $h(\Phi) = \Phi^2(3 - 2\Phi)$ is the weight function for the epitaxial layer.

The equation of motion is

$$\begin{aligned} \xi\partial_t\Phi &= A\xi\Delta\Phi + \frac{B}{\xi}g'(\Phi) \\ &\quad + Ch'(\Phi)\{(\mu^{(e)} - \mu^{(v)})\epsilon_{ij}\epsilon_{ij} + \frac{\lambda^{(e)} - \lambda^{(v)}}{2}(\epsilon_{ii})^2 \\ &\quad - \frac{1+\nu^{(e)}}{1-2\nu^{(e)}}(\epsilon^m)^2\}, \end{aligned} \tag{11}$$

where $\Phi = 1$ represents the solid phase, $\Phi = 0$ represents the liquid phase, $0 < \Phi < 1$ represents the diffuse interface, ξ is the width of the transition region, A, B, C are constants, $g'(\Phi) = 2\Phi(1 - \Phi)(1 - 2\Phi)$, and $h'(\Phi) = 6\Phi(1 - \Phi)$.

2.1 Numerical results

We implemented the model using the explicit scheme based on FDM for the phase-field equation (11). For the elastic problem, we used FreeFem++ based on FEM. Computations of stress field were very time consuming.

In the numerical experiments, we set up initial surface of heteroepitaxial film to be rectangular and material parameters of silicon are taken. We observed that both tops and valleys of the surface profile deepen but the valleys deepen at higher velocity (see

Fig. 6). It is not obvious from the experiments whether this can lead to fracture. We found that numerical noise avoid us to simulate the problem in longer time. Therefore, our aim in the future is to develop better numerical schemes suitable for the model of heteroepitaxial growth.

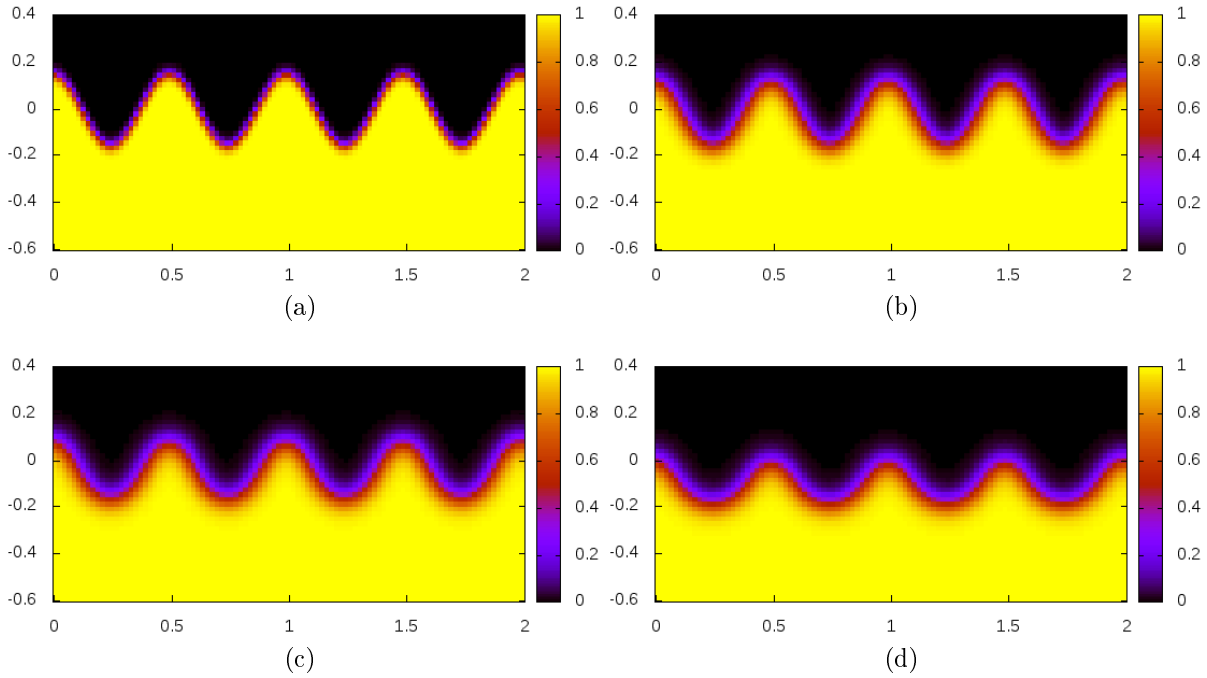


Figure 6: Evolution of heteroepitaxy at different times.

References

[1] A. Baskaran. *Modeling and Simulation of Heteroepitaxial Growth.* University of Michigan (2009).

[2] W. K. Burton, N. Cabrera, and F. C. Frank, *The Growth of Crystals and the Equilibrium Structure of their Surfaces*, Phil. Trans. Roy. Soc. 243 (1951), 299.

[3] V. Chalupecký, H. Emmerich, *Numerical scheme for two-scale model of liquid phase epitaxy*, In: Beneš M., Kimura M. and Nakaki T., Eds. *Proceedings of Czech Japanese Seminar in Applied Mathematics 2006*, in COE Lecture Note, Vol. 6, Faculty of Mathematics, Kyushu University Fukuoka, 2007, ISSN 1881-4042, pp. 50–61.

[4] L. Q. Chen. *Phase-Field Models for Microstructure Evolution.* Annual Review of Materials Research, Vol. 32 (2002), 113–140.

[5] D. H. Hoang. *Epitaxial Crystal Growth.* Proceedings of ALGORITMY (2009).

[6] D. H. Hoang. *Numerical Simulation of Spiral Growth by Phase-Field Method*. Doktoandské dny (2009).

[7] A. Karma, and M. Plapp, *Spiral Surface Growth without Desorption*, Phys. Rev. Lett. 81 no. 20 (1998), pp. 4444–4447.

[8] K. Kassner, C. Misbah. *Phase field for stress-induced surface instabilities*. Europhys. Lett. 46 (1999), 217.

[9] K. Kassner, C. Misbah, J. Müller, J. Kappey, and P. Kohlert. *Phase-field for elastic surface instabilities*. Phys. Rev. E 63 (2001), 036117.

[10] H. Emmerich. *Modeling elastic effects in epitaxial growth*. Continuum Mechanics and Thermodynamics, Volume 15, Issue 2 (2003), 197-215.

[11] I. V. Markov. *Crystal growth for beginners: fundamentals of nucleation, crystal growth and epitaxy*. World Scientific (2003).

[12] J. Müller, M. Grant. *Model of Surface Instabilities Induced by Stress*. Phys. Rev. Lett. 82 (1999), 1736–1739.

[13] I. Steinbach. *Phase-field models in materials science*. Modelling Simul. Mater. Sci. Eng. 17 (2009), 073001.

[14] B. J. Spencer et al.. *Morphological instability in epitaxially strained dislocation-free solid films*. Phys. Rev. Lett. 67 (1991), 3696–3699.

Porovnání existujících NoSQL DBMS z hlediska škálovatelnosti

František Jahoda

1. ročník PGS, email: jahoda@cs.cas.cz
Katedra matematiky
Fakulta jaderná a fyzikálně inženýrská, ČVUT v Praze

školitel: Július Štuller, Ústav informatiky, AVČR

Abstract. In recent years, the so called NoSQL DBMS starts profiling against a widely used and proven relational DBMS. These database systems are characterized by an effort to improve scalability at the expense of features of the database system. They are designed for specific applications and thus are different from each other. The article compares the three most popular CouchDB, MongoDB and Google Bigtable with emphasis on scalability.

Keywords: scalability, noSQL, MongoDB, CouchDB, Google BigTable

Abstrakt. V poslední době se vůči široce používaným a osvědčeným relačním DBMS začínají profilovat tzv. NoSQL DBMS. Tyto databázové systémy se vyznačují snahou o lepší škálovatelnost za cenu snížení požadavků kladených na databázový systém. Tyto systémy jsou konstruovány pro specifické oblasti použití a jsou tedy od sebe odlišné. V článku srovnávám tři nejpopulárnější CouchDB, MongoDB a Google BigTable s důrazem na škálovatelnost.

Klíčová slova: škálovatelnost, noSQL, MongoDB, CouchDB, Google BigTable

1 NoSQL DBMS (Structured data storages)

je souhrný název pro různé databázové systémy (*DBMS*), jež se snaží umožnit škálovatelnost aplikací na nich postavených. Mezi představitele těchto systému patří CoucheDB, MongoDB, Google BigTable a další. Tyto systémy nemají jednotnou funkčnost stanovenou standardem. V zásadě je lze rozdělit na systémy reprezentující data jako dokumenty, grafy, provázané objekty, slovníky (hash tabulky) a XML databáze. Cíleně relaxují některé podmínky ACID (atomičnost transakcí, konzistenci, izolaci změn v rámci transakcí a trvanlivost dat). Dotazování těchto systému je též omezené, často neexistuje podpora pro podobné operace jako mají relační databázové systémy: např. operace databázového spojení (join), agregační funkce. V těchto systémech se počítá s nasazením na clusteru a je tedy kladen důraz na efektivitu síťové komunikace mezi uzly clusteru. Tyto DBMS se nedotazují pomocí SQL, ale většinou pomocí vlastního proprietárního jazyka s omezenou funkčností.

Společným rysem těchto systémů je, že nevyžadují definovat schéma databáze a umožňují vložit tzv. řídká data. Každý záznam tedy může mít např. vyplněné jiné sloupce. Nevýhodou tohoto přístupu je, že konzistenci databáze si musí hlídat uživatel sám. Výhodou naopak je, že změny požadavku na databázi nevedou ke změně databázového schématu, která může být časově náročná a v produkčním prostředí lze tedy např. omezit odstávky

databáze. V databázi totiž mohou existovat záznamy se starým i novým formátem vedle sebe.

2 Škálovatelnost

Škálovatelnost je v softwarovém inženýrství žádaná vlastnost, která vyjadřuje schopnost zvládat přijatelným způsobem zvětšující se množství práce. Škálovatelností v kontextu distribuovaných databází rozumíme schopnost systému vyhovovat zadaným požadavkům při změně následujích proměnných: množství dat, množství operací zápisu a množství operací čtení. V praxi se ovšem jen zřídka setkáme při růstu aplikace pouze s jedním druhem škálovatelnosti a tak má smysl řešit reakci na všechny předchozí druhy škálovatelnosti zároveň.

Hrubě řečeno, abychom označili systém za dobře škálovatelný, očekáváme, že reakce databáze by se neměla prodlužovat při růstu předchozích proměnných. Toho samozřejmě nelze dosáhnout na stejném počítačovém vybavení a tak se předpokládá, že narůst jednotlivých proměnných je úměrně kompenzován nárůstem počtu strojů, které jsou schopny požadavky paralelně zpracovávat.

Ukazuje se, že velkou překážkou škálovatelnosti bývají náročné operace, které komerční relační databáze běžně zajišťují: synchronizace transakcí při práci na společných datech, replikace databáze a zajištění konzistence databáze v síťovém prostředí.

Správa zámků může tvořit podstatnou část režije DBMS. Režie plánování transakcí, tak aby nebyla ohrožena integrita databáze, může při vzrůstajícím zatížení vést k prudkému propadu výkonu.

Další podstatnou částí režije DBMS se stává síťová komunikace mezi jednotlivými stroji a mechanismus udržování aktuální verze dat na všech strojích. To je často realizováno pomocí časově náročného dvoufázového potvrzení (two phase commit).

NoSQL DBMS se problémům s růstem zátěže snaží čelit omezením délky transakcí na jeden záznam, rozdělením souvisejících dat na stejný stroj (shardování), duplikací dat (replikace) a omezením podmínky na konzistenci databáze při zátěži. Replikace se často používá i k zajištění větší spolehlivosti databáze uložené na více strojích v síťovém prostředí. Ne všechny noSQL DBMS však tento problém řeší stejně a proto má smysl prozkoumat v čem se v této oblasti liší a tím potenciálním uživatelům umožnit vybrat DBMS, jež se jim pro jejich aplikaci bude hodit nejlépe.

3 Google BigTable

Google BigTable [5] je používána pro Google App Engine Datastore [2]. Jedná se o DBMS komerční, který společnost Google navrhla pro své cloud služby. Je *silně konzistentní*, tedy každé čtení vrácí výsledek po poslední operaci zápisu. Toho je docíleno tím, že pro každý záznam je stanoveno, kdo drží jeho hlavní verzi. Silnou konzistenci lze relaxovat na *eventuální konzistenci*. Při eventuální konzistenci mohou být vrácena starší data ze záložní kopie. Dotazování tohoto DMBS je záměrně omezeno a např. operaci databázové spojení je třeba implementovat v aplikační vrstvě.

Databáze je organizovaná a srovnaná podle klíčů, ke každému klíči náleží jeden dokument, který se skládá z několika sloupců s hodnotami. Každá hodnota má časovou značku a v jedné buňce (určené klíčem a sloupcem) může tedy být několika hodnot. Konzistence se tedy zajišťuje pomocí Multiversion Concurency Control (*MVCC*), tento systém kontroly konzistence zajišťuje, že každá operace čtení pracuje nad obrazem databáze, jak databáze vypadala při spuštění této operace. Zároveň lze nastavit, kolik verzí hodnot se v daném sloupci má uchovávat. Databázový stroj pak starší verze automaticky uvolňuje.

Dokumenty databáze jsou seskupeny do bloků tzv. tablets. Tyto bloky jsou vytvářeny dynamicky a různé bloky mohou být umístěny na různých strojích clusteru. (Celá databáze je navržena pro provoz v clusteru).

I když je v API možné operace nad databází seskupovat, atomicita operace je zaručena pouze v rámci změny jediného dokumentu. Kontrola konzistence databáze je zanechána na uživateli a v databázi neexistuje mechanismus integritních podmínek jako jsou cizí kliče nebo unikátní hodnoty. Rozhraní Google app engine umožňuje definovat transakce nad více dokumenty, ale je nutné specifikovat, že se tyto dokumenty mohou účastnit společné transakce už v době jejich vytvoření. Konzistence databáze je zajištěna vůči selhání uzlu clusteru i výpadku spojení mezi clustery.

S popisu je tedy zřejmé, že vysoké škálovatelnosti je dosaženo pomocí rozdělení databáze do více bloků a zamezením vazeb mezi jednotlivými bloky.

4 CouchDB

CouchDB [1][4] je open source DBMS s Apache licencí, který je postaven na jazyku Erlang a jehož rozhraní je realizováno pomocí HTTP protokolu s využitím pravidel REST a JSON formátu záznamů. Stejně jako Google BigTable se pro zajištění konzistence databáze používá MVCC model pro změnu hodnot. Pokud v průběhu změny záznamu dojde ke změně jiným procesem, DBMS ohlásí konflikt při zápisu a aplikace se může pokusit o nový zápis. Databáze je bezschémová, každý záznam se skládá z klíčů, které jsou typované.

Je možné provozovat paralelně několik instancí databáze s tím, že změny si instance posílají mezi sebou. V této situaci se může stát, že dojde k zápisu do stejného záznamu do obou instancí zároveň a tedy ke konfliktu. Tento konflikt je vyřešen automaticky tak, že za aktuální verzi se zvolí posledně uložený zápis instance a starší verze se uloží pro pozdější řešení konfliktu. Řešení konfliktu je zanecháno na aplikaci, která by si měla zvolit možnost řešení, která je pro daný záznam vhodná (nic neměnit, vrátit se ke starší verzi, nebo nějakým způsobem záznamy sjednotit).

5 MongoDB

MongoDB [3][6] je též open source DBMS s AGPL licencí (licence podobná GPL 3.0), která se ale vztahuje jen na samotný DBMS a ne již na aplikace jej využívající. DBMS je postaven na jazyku C++. Přestože je v mnohém podobný CouchDB v některých ohledech se liší. Místo MVCC modelu pro updatování záznamů používá "in place update", a proto se nedoporučuje používat v konfiguraci master-master. DBMS je v konfiguraci master-slave silně konzistentní. Operací "in place update" je myšlena záměna záznamu přímo na

	Google BigTable	CouchDB	MongoDB
horizontální rozdělení	ano	ano	ano
replikace	tablet servers	master-master	master-slave
atomicita	nutno určit při zadávání dat	jeden dokument	jeden dokument
kontrola souběhu	MVCC	MVCC	update in place
model konzistence	silná i eventuální	silná i eventuální	eventuální

Tabulka 1: Srovnání vlastností NoSQL databází

disku. Toto se provede jen pokud se délka záznamu o moc nezvětší, v opačném případě je nutno záznam uložit jinde na disku. Jelikož pro každý záznam není uchovávána jeho historie a specificky konflikty při měnění záznamů, není možné řešit konflikty při eventuální konzistenci. Tato nevýhoda je však vyvážena lepším využitím operační paměti a diskového místa.

MongoDB se od CouchDB liší i větším důrazem na velikost přenášených dat a s DBMS není nutné komunikovat jen pomocí REST rozhraní, ale je tak možno činit i pomocí proprietárních konektorů.

6 Celkové srovnání

V porovnání databází jsem se záměrně zaměřil na vlastnosti spojené s ACID charakteristikami transakcí a provozem v síti a vynechal jsem jiné rozdíly, které však při výběru databáze mohou též hrát významnou roli. Celkové srovnání je shrnuto v tabulce 1.

7 Závěr

NoSQL databáze jsou schopné škálování i v situacích, kdy toho běžné komerční relační databáze schopné nejsou. Zároveň díky omezení povolených operací vedou k aplikacím, které nevyužívají špatně škálovatelných vlastností a umožňují snadněji předvídat, jakým způsobem porostou nároky systému. NoSQL DBMS vytváří alternativu pro aplikace, u kterých lze očekávat výrazný nárůst požadavků a potřebu provozovat databázi v clusteru, za cenu podstatného omezení z funkcionality poskytované klasickými relačními DBMS.

Literatura

[1] Couchdb online description.
http://couchdb.apache.org/docs/overview.html.

[2] Google app engine online documentation.
http://code.google.com/appengine/articles/storage_breakdown.html.

[3] Mongodb homepage.
http://www.mongodb.org.

[4] J. C. Anderson, J. Lehnardt, and N. Slater. *CouchDB: The Definitive Guide.*

[5] F. Chang, J. Dean, S. Ghemawat, W. C. Hsieh, D. A. Wallach, M. Burrows, T. Chandra, A. Fikes, and R. E. Gruber. Bigtable: A distributed storage system for structured data. In 'OSDI'06: Seventh Symposium on Operating System Design and Implementation'.

[6] K. Chodorow and M. Dirolf. *MongoDB: The Definitive Guide.*

COMPASS Database Upgrade

Vladimír Jarý

2nd year of PGS, email: jaryvlad@kmlinux.fjfi.cvut.cz
Department of Software Engineering in Economy
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Miroslav Virius, Departement of Software Engineering in Economy,
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. Modern experiments in the particle physics depend on computer systems which are used to store and analyze large quantities of data. This paper describes the data acquisition system of the COMPASS experiment at CERN and focuses on the role of databases in this system. At first, the existing database architecture is analyzed, then the new architecture is proposed. The proposal includes replication, monitoring, and back ups to achieve high reliability and availability of the database service. Finally, implementation of the proposal is reviewed. This paper also briefly covers several database optimization techniques which were required to reduce the server load. Some possible future improvements are also discussed.

Keywords: data acquisition, database, high availability, COMPASS

Abstrakt. Moderní experimenty ve fyzice elementárních částic používají počítače pro simulace, pro řízení a pro sběr a analýzu dat. Tento článek popisuje systém pro sběr dat používaný experimentem COMPASS v Evropské organizaci pro jaderný výzkum (CERN) a zaměřuje se na roli databází v tomto systému. Nejprve je popsána stávající databázová architektura, poté je představen návrh nové architektury. Pro zajištění vysoké dostupnosti a spolehlivosti databázové služby se v návrhu počítá s replikací, s pravidelným zálohováním a s dohledovým systémem. Následně článek popisuje řešení několika problémů, které se vyskytly během implementace návrhu. Článek také stručně zmiňuje prostředky optimalizace databázové struktury a databázových dotazů, které byly použity pro snížení zátěže serverů. V závěru jsou vyjmenována další možná vylepšení aktualizované architektury.

Klíčová slova: sběr dat, databáze, vysoká dostupnost, COMPASS

1 Introduction

Today, computers participate in every phase of experiments in the particle physics: They are used for detector simulations, for the experiment control, for the data acquisition, and for the data analysis. In this paper, the database system used by the COMPASS experiment is described. At first, the experiment is briefly introduced and the data acquisition system is presented in extent that is necessary for the following discussion. In the following section, the existing database architecture is reviewed and our proposal of the upgrade is presented. In the last section, we analyze the solution of several problems that occurred during the upgrade.

2 The COMPASS experiment

COMPASS, which is an abbreviation for the *Common Muon and Proton Apparatus for Structure and Spectroscopy*, is a fixed target experiment operating on the SPS (*Super Proton Synchrotron*) particle accelerator at CERN [1]. Scientific program of the COMPASS experiment was approved by CERN in 1997; it consists of the muon and hadron programs. Data taking started in 2002. Today, the proposal of the second phase (*COMPASS–II*) has been submitted to the CERN scientific council [2]; this proposal includes 3 programs which would run at least until 2015, if approved by CERN.

2.1 Data acquisition system

The beam of the accelerated particles (muons or hadrons, depending on the program) provided by the SPS accelerator hits the polarized target. When the beam particles interact with the target, secondary particles are produced. The particles passing through the spectrometer are registered by the system of detectors. The beam is not continuous, it consists of the co called spills. Typical spill contains 10^8 particles. The flight and decay of a particle in the spectrometer is known as an *event*. Each event can be described by roughly 35 kB of data. Total year production attacks the value of 500 TB (508078 GB in 2004, [5]).

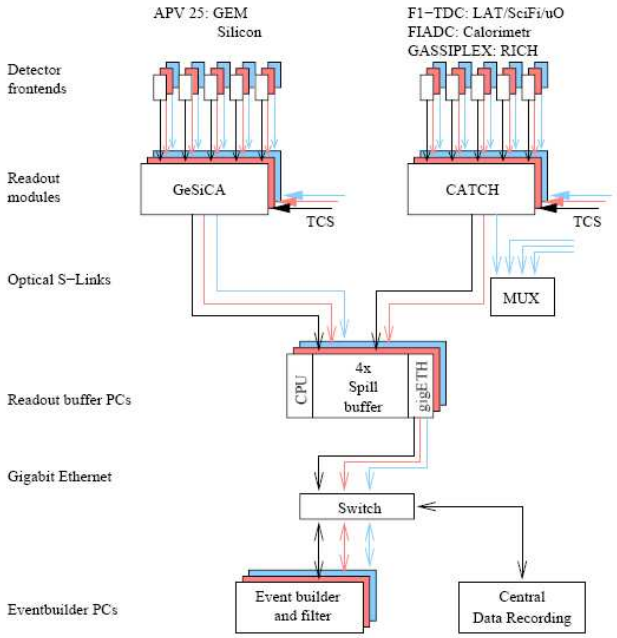


Figure 1: Layers of the DAQ system according to [4]

System for the *data acquisition* (*DAQ*) consists of several layers (see Fig. 1). First, the front-end electronics, which is part of detectors, digitizes the analog signal generated by the detected particle. Front-end electronics also handle the delay caused by the time of flight of the particle (the length of the spectrometer is about 50 m). The raw data

produced by front-end electronics is collected by the front-end boards called *CATCH* and *GeSiCA*. In this layer, event header is appended to data. This header will be used by the following layers to reconstruct events from blocks of data coming from different detectors. Data files are transferred to the following layer – the *ROB* computers – by the *S-LINK* interface. *S-LINK* is a high speed bus developed in CERN for the ATLAS experiment. *ROB* (*Read out buffer*) computers act as a cache for data. Data packets are received during spill and are send to the last layer, to the *event builder* computers (*EVB*). As their name suggests, EVB computers are used to reconstruct events. Data files representing events are sent using the *CDR* (*Central data recording*) facility into the permanent storage *CASTOR* (*CERN Advanced storage*) after some delay.

Software for the DAQ is based on the *DATE* system which has been developed for the ALICE experiment at the Large Hadron Collider experiment. The COMPASS DAQ system combines industry standard equipment (such as Gigabit Ethernet) with prototypes (e.g., the *CATCH* front-end board was developed in Freiburg). There is currently a proposal to replace *ROB* and *EVB* computers by the architecture based on *FPGA* (*Field-programmable gate array*) circuits and to reuse existing read out buffers and event builders for online filtering and analysis. More information about the COMPASS DAQ can be found in [8].

3 Existing database architecture

It was previously stated that the year data production exceeds 500 TB. These data are not stored in the database, they are saved on tapes in the *CASTOR*. Databases in COMPASS manage meta-information about the run of the experiment. These meta-information include detector configuration, beam parameters, or software logs.

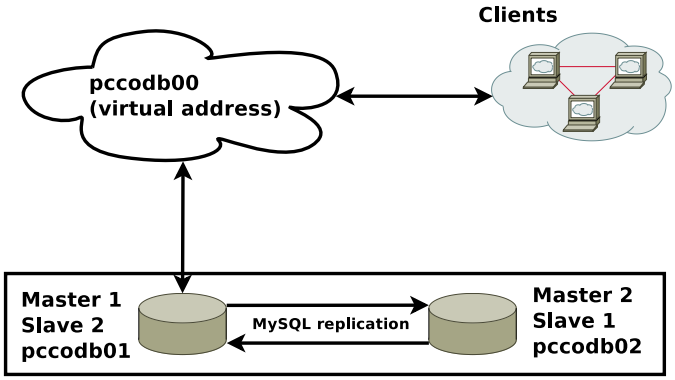


Figure 2: Existing database architecture

COMPASS uses two physical database servers named *pccodb01* and *pccodb02*. These servers are powered by 32-bit operating system *Scientific Linux CERN* which is rebranded *Red Hat Enterprise Linux*. As a database software, the MySQL server has been selected because of its performance. Two physical servers are synchronized by the *master–master* replication [10]. Server *pccodb01* acts as a master of slave server *pccodb02*. At the same time, server *pccodb02* acts as a master of slave server *pccodb01*. This means that both servers contain the same data. This mechanism helps to achieve high availability of

the database service. Furthermore, during the replication, all queries are written to the binary log. This log can be considered as an incremental back up. Clients connect to database through the virtual address *pccodb00* which normally points to *pccodb01*. In case the *pccodb01* experiences problems (crash, overload), the virtual address is reconfigured to point to *pccodb02*. This process is transparent to clients. When the *pccodb01* recovers, it resynchronizes itself with the *pccodb02* by means of the replication.

Each MySQL server contains roughly 20 logical databases. Two largest databases *beamdb2009* and *DATE2009_log* hold about 15 GB of data. The *beamdb2009* stores information about beam parameters, the *DATE2009_log* stores error logs of the *DATE* software. Another large database *runlb* stores logbook entries provided by shift crew, database *DATE2009* stores configuration of event builders and read out buffers.

4 Proposal of database upgrade

The existing architecture became overloaded several times during last year. For the 2010 run, it was expected that the load would increase as a consequence of higher intensity of the beam. We were asked to design a new database architecture that would sustain increased load.

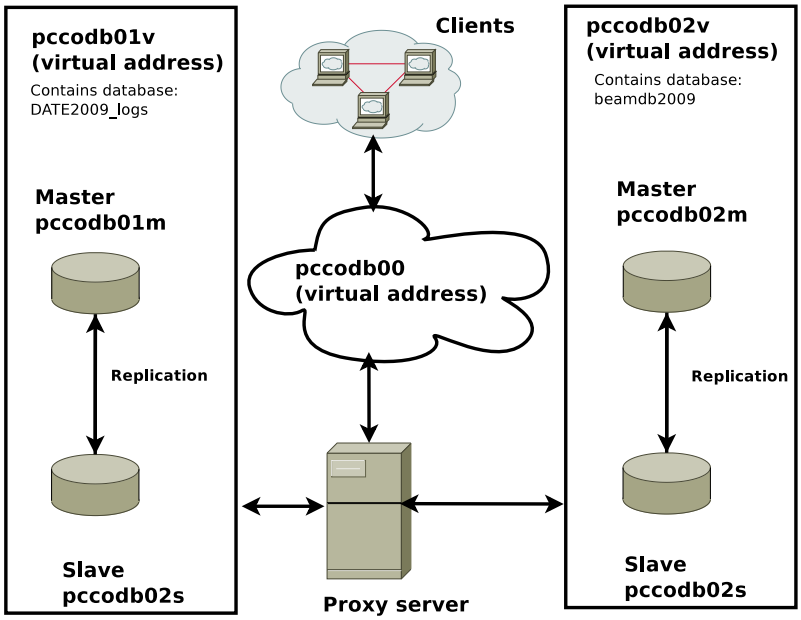


Figure 3: Proposed upgrade of the database architecture

The main idea of the new architecture lies in splitting two of the largest databases on separate servers. The proposal also counts with master–master replication to guarantee high availability. Thus, the first pair of physical servers manages the *DATE2009_log* database, the second pair manages the *beamdb2009* database. Remaining smaller databases are distributed evenly among the pairs of servers. Fifth (physical) server acts as a proxy server. It connects clients to the corresponding server according to the requested database. This server is also used for additional tasks: monitoring, back ups, and HTTP

server. Clients would connect to the proxy through virtual address *pccodb00*. This means that migration to the new architecture should be transparent.

The proposal has been presented on the meeting of the front-end electronics group and with some modification it has been approved. Unfortunately, only three physical servers were provided. On the other hand, the hardware configuration of new servers is much better in comparison to the old servers (see table 1). Thus, the new architecture combines features of the old architecture with features of the proposal.

As in the original architecture, all databases are stored on two servers (named *pccodb11* and *pccodb12*) which are synchronized using the replication. Third server (*pccodb10*) is used as a proxy which is accessible via virtual address *pccodb00*, so there is no need to reconfigure clients. These servers are part of the COMPASS internal network and are located directly in the experimental hall. To increase safety of the data, server *pccodb11* is replicated to the server *compass02* which runs in the CERN computing center and acts as a gateway to the Internet. The *compass02* is replicated into the computer centers of participating institutes. This configuration is known as a *chain replication*.

	Old server	*New server*
Memory	3 GB	16 GB
Processor	2 cores at 3 GHz (Xeon)	8 cores at 2.5 GHz (Xeon)
OS	32b SLC 4.7	64b SLC 5.4
Linux	2.6.9	2.6.18
Server	MySQL 4.1.22	MySQL 5.1.45

Table 1: Configuration of old and new servers

According to the proposal, the proxy server is also running the *HTTP* service. This service includes database management tool *phpMyAdmin*, run logbook application, and web interface of the monitoring software Nagios. Nagios has been selected because of its modular design: it is relatively easy to configure it to meet specific needs. Its functionality can be extended by plug-ins, there are many plug-ins available in the standard Nagios installation. In addition, we developed a custom plug-in which monitors a temperature of CPU cores. It was decided to monitor at least the following quantities: the uptime and load average of servers, the state of MySQL processes, the state of replication, the state of the *cron* daemon, and the core temperatures. The *Nagios Remote Plug-in Executor (NRPE)* agent is installed on database servers to intermediate communication between Nagios and plug-ins. NRPE receives requests sent by Nagios, executes them, and sends back the result. Nagios displays the state of the monitored servers using the web interface. If the Nagios detects a problem, it notifies administrator by an e-mail. In certain circumstances, it can also attempt to fix the problem. For example, the proxy server normally connects all clients to the *pccodb11* server. If the Nagios detects that this server is down, it reconfigures the proxy to redirect all clients to the *pccodb12* server.

4.1 Implementing the proposal

The migration started with the installation of the operating system (OS) on the new servers. The support of the Scientific Linux CERN release 4 ends later this year, thus it

has been decided to use newer release (5.4) of this OS. At the same time, the architecture has been switched from 32 b to 64 b. The first problem appeared during the installation: The servers did not contain optical drive, so bootable flash disk had to be prepared. The rest of the installation was completed without any other issues; we have chosen the *ext3* as a file system for MySQL data directory.

Also the MySQL server software was upgraded to the latest stable version (5.1.45). Precompiled package for the SLC 5 was not available, the installation was done by compiling the source codes. In the *configure* stage of installation, several features were enabled (e.g. support for very large tables, storage engine *InnoDB*). In the next step, additional customization was achieved by editing the server configuration file. As a base, the *my.cnf.huge* template was used. This template is designed for heavily loaded database servers, yet it appeared that some parameters were not sufficient: The limit on the number of simultaneously opened files had to be increased. Moreover, binary logging was turned on to enable replication. We also enabled logging of *slow queries*. Knowledge of problematic queries is very important for optimization.

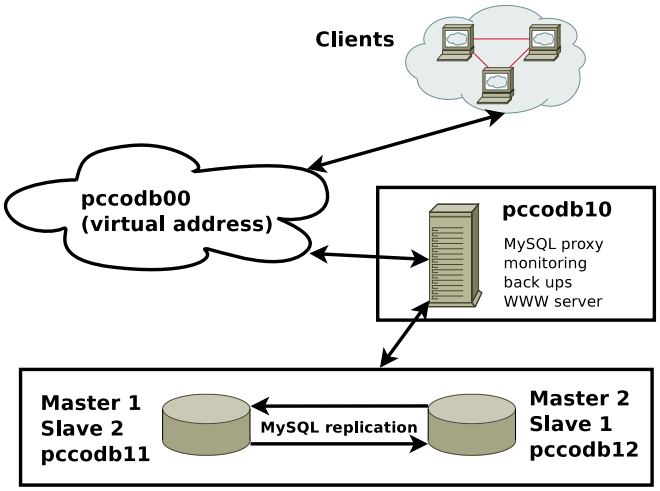


Figure 4: New database architecture

The migration continued by dumping all data from old servers. The tool *mysqldump* was used for this purpose. It saves dumps as a script with SQL commands (*CREATE TABLE* followed by *INSERT* statements) which restore dumped data. The scripts were copied using the *scp* tool to the new servers. The data was imported into the *pccodb11* server and the *pccodb12* was automatically synchronized due to the replication. It is imperative to verify that data were imported correctly. The idea was to dump data from new server and compare the dumped files with dumps from old server. Due to the size of dumps (several gigabytes for largest databases), fast method of the file comparison was necessary. The comparison method is based on the *md5sum* tool which calculates *md5 hash* of given file [7]. If the hashes of dumps from old and new servers differ, it indicates that there was problem with migration. The *md5* is not an injective function, i.e. different dumps can have the same hash but the probability of the hash collision is very small and can be ignored.

This method proved that the majority of databases was imported successfully. Only 3 databases had produced different hashes. It was required to determine the cause of the

failure. The *md5* function can only decide whether the dumps differ or not. On the other hand, another tool *diff* compares files line by line and writes the differences. However, it needs more time and resources (CPU, memory) to do its job. The output of the *diff* helped with identification of the problem: The definition of the data type *DECIMAL(m, n)* has changed in recent version of the MySQL server [11]. This type stores rational numbers, the M parameter represents the significant digits, the N parameter represents the number of the digits following the decimal point. Old version of MySQL stored this data type as a character string, each digit was saved as a character as well as the sign. In this representation, the range of the positive numbers could be extended by one magnitude. E.g., the type *DECIMAL(5, 2)* represented the rational numbers from the interval $[-999.99, 9999.99]$. In the version 5.0.3 of the MySQL server, the definition changed in order to comply with the SQL standard. According to the standard, the data type *DECIMAL(M, N)* represents rational numbers with up to $M - N$ digits before the decimal point and up to N digits after the decimal point. Thus, the type *DECIMAL(5, 2)* represents interval $[-999.99, 999.99]$. To fix the problem, it was necessary to modify the *CREATE TABLE* SQL command in dumps from old server before importing it into new server.

Another problem appeared at the end of the migration, after the old servers had been disconnected. Some clients were not allowed to connect to new servers. Inspection of the system table with privileges (*'mysql'.'user'*) revealed that the affected clients were permitted to login only from address *pccodb00*. This was not problem on the old architecture, because the virtual address pointed directly on the database server. On the other hand, in new architecture, the virtual address points to the proxy server *pccodb10* (see Fig. 4). The problem was solved by changing the address in system table from *pccodb00* to *pccodb10*.

To enhance safety of the data, regular back ups were scheduled. There are three types of back ups: daily, hourly, and incremental. Daily back ups contain all data from all databases, hourly back ups contain all data from only smaller databases (i.e. all databases except *beamdb2009* and *DATE2009_log*). Back ups are created by shell script that is automatically invoked by the scheduler *cron*. Script uses *mysqldump* program to retrieve data and *gzip* to compress dumps; it takes approximately 30 minutes to dump all data. Back ups are purged after two days to save disk space. During the replication, all queries that modify data are written into binary log which is read by replication slave. This log can be also used as an incremental back up. In case of accident (disk failure, dropped database), it is possible to recover all data.

4.2 Database optimizations

When working with larger tables, it is important to optimize table structure and queries. As already mentioned, we have enabled logging of slow queries, i.e. queries that take long time to execute. MySQL provides useful command called *EXPLAIN* which explains how is the query evaluated. The command informs which table indices are used (if any), how many rows must be examined, whether temporary tables need to be created, or if the query can be split into subqueries. This output should be used to improve both schema (by adding indices) and query (by splitting it to subqueries, using LIMIT keyword to

reduce number of returned records, etc).

Additional improvements can be achieved by reducing amount of data in tables, thus less demanding data types should be used wherever possible. MySQL server in version 5.1 brings support for a new feature: the *partitioned tables*. Such a table is divided into several *partitions* according to some function defined over a set of table columns. Each partition can be treated as a separate table. Execution of some queries can be greatly improved by using partitioning. In these cases, only partition which can contain the desired rows are searched. This technique is known as a partition pruning. In addition, each partition can be stored on different disk. This can improve speed of the *SUM* operation by parallel processing of multiple partitions.

Finally, it is also possible to improve performance of the database by choosing appropriate storage engine for tables. In [3], we have compared two most popular engines: *MyISAM* and *InnoDB*. *MyISAM* is much faster than *InnoDB* when comparing speed of row insertion. On the other hand, *MyISAM* engine has stricter limitation on length of index. If longer index is needed, *InnoDB* should be used. *InnoDB* also supports fully *ACID*[1] compliant transactions. COMPASS uses the *MyISAM* engine. Table in the *MyISAM* engine can be packed by the *myisampack* utility. This tool reduces table size to $40\% - 70\%$ depending on contained data but also makes it read-only [11]. This feature could be used to compress old databases (e.g. *beamdb2006*, *DATE2006_log*).

Currently, MySQL Proxy running on the *pccodb10* redirects all queries to one database server (usually to the *pccodb11*). But the proxy has much greater potential: For example, it can be used to implement load balancing. In the load balancing mode, all queries that modify the data (*INSERT*, *UPDATE*, *DELETE*) are sent to one server, queries that retrieve data (*SELECT*) are distributed across multiple servers. The proxy server contains a script written in the *Lua* language that checks availability of database servers, their load, and the replication lag to decide which server executes the query. At the end of the year, the load balancing will be tested in the current database architecture. If the tests succeed, the load balancing will be implemented before the start of the 2011 run.

5 Conclusion

Data acquisition system of the COMPASS experiment has been described. We have analyzed existing database architecture of the experiment and prepared proposal of new architecture. This proposal has been approved with some modifications and implemented before the start of the 2010 data taking. New databases servers are now running without problems for several months, old servers have already been used as additional event builders.

For the near future, there are plans to use some advanced MySQL features to improve performance of servers. These plans include the partitioning of large tables, the change of the storage engine of several older tables, and also using MySQL proxy as a load balancer. We have been also asked to use Nagios to monitor other machines that participate in data taking (event builders, read out buffers).

[1] Atomicity, Consistency, Isolation, Durability

Acknowledgment

This work has been supported by grants MŠMT LA08015 and SGS 10/094.

References

[1] P. Abbon et al. (the COMPASS collaboration): *The COMPASS experiment at CERN*. In: Nucl. Instrum. Methods Phys. Res., A 577, 3 (2007) pp. 455–518

[2] Ch. Adolph et al. (the COMPASS collaboration): *COMPASS-II proposal*. CERN-SPSC-2010-014; SPSC-P-340 (May 2010)

[3] L. Fleková, V. Jarý, T. Liška, M. Virius: *Využití databází v rámci fyzikálního experimentu COMPASS*. In: Konference Tvorba softwaru 2010, Ostrava: VŠB - Technická univerzita Ostrava, 2010, ISBN 978-80-248-2225-9 pp. 68–75.

[4] E. Gurzhiy: *Řízení distribuovaných výpočtů při zpracování experimentálních dat*. Praha: CTU, 2007.

[5] A. Král, T. Liška, M. Virius: *Experiment COMPASS a počítače*. In Československý časopis pro fyziku 5. Prague, Czech Republic, 05/2005. pp. 472.

[6] T. Liška: *The Cluster and GRID Computing on Mass Data Production Systems for High Energy Physics Experiments*. Praha: CTU, February 2009.

[7] R. Rivest: *The MD5 Message–Digest Algorithm* [Internet RFC 1321]. April 1992. Available at: `http://tools.ietf.org/html/rfc1321`

[8] L. Schmitt et al.: *The DAQ of the COMPASS experiment*. In: 13th IEEE-NPSS Real Time Conference 2003, Montreal, Canada, 18–23 May 2003, pp. 439–444

[9] *COMPASS page* [online]. 2010. Available at: `http://wwwcompass.cern.ch`

[10] *MySQL Master–Master Replication* [online]. 2010. Available at: `http://www.howtoforge.com/mysql_master_master_replication`

[11] *MySQL 5.1 Reference Manual* [online]. 2010. Available at: `http://dev.mysql.com/doc/refman/5.1/en/`

Výskyt makroskopických fenoménů v modifikovaném termodynamickém dopravním modelu

Katarína Kittanová

2. ročník PGS, email: kittakat@fjfi.cvut.cz
Katedra matematiky
Fakulta jaderná a fyzikálně inženýrská, ČVUT v Praze

školitel: Milan Krbálek, Katedra matematiky, Fakulta jaderná a fyzikálně inženýrská, ČVUT

Abstract. This article deals with traffic modeling. Introduction shows the difference between analyzing traffic on microscopic and macroscopic level. Then, a possible way to gain an universal model, powerful in both analysis, is introduced. It is based on thermodynamical traffic model, which is proper for exploring the microscopical structure. After modification, some simulations have been made and the macroscopic phenomena have been observed. Also the microscopic structure in new model has been confirmed.

Keywords: traffic modeling, traffic congestion, modified metropolis algorithm, thermodynamical gas model

Abstrakt. Tento článek se zaobírá modelováním dopravy. V úvodu je krátce vysvětlen rozdíl mezi mikroskopickou a makroskopickou analýzou dopravy. Pak je nastíněn možný postup získání univerzálního modelu. Východiskem je termodynamický dopravní model, používaný pro zkoumání mikroskopické struktury, v kterém je provedena modifikace. V upraveném modelu jsou provedeny první simulace indikující výskyt makroskopických jevů a porovnání ověřující zachování mikroskopické struktury.

Klíčová slova: modelování dopravy, dopravní zácpa, modofikovaný metropolisův algortimus, model termodynamického plynu

1 Úvod

Zájem o zkoumání a modelování dopravy v posledních desetiletích neustále rostl. Bylo použito mnoho přístupů a analýz dopravních vzorků na makroskopické a mikroskopické úrovni. Jedním z nejvýznamnějších cílu dopravního modelování je představení universálního modelu generujícího mikroskopickou strukturu zodpovídající skutečnosti, ve kterém se navíc při volbě vhodných parametrů vyskytují makroskopické jevy.

1.1 Macroscopická struktura

Makroskopický přístup zkoumá dopravu na globální úrovni. Mezi nejvýznamnější veličiny popisující makroskopickou strukturu dopravního systému patří dopravní tok a hustota

provozu. Dopravní tok udává počet vozidel zaznamenaných na pevně určeném úseku vozovky za jednotku času.

$$\Phi = \frac{N}{t}.$$

Hustota provozu zas představuje počet vozidel nacházejících se v jednom okamihu na jednotce délky vozovky.

$$\rho = \frac{N}{l}.$$

Při zkoumání dopravy je zájem kladen především na vztah uvedených veličin $\Phi = \Phi(\rho)$. Grafická reprezentace téhle závislosti se nazývá fundamentální diagram.
Aktuální pozice dopravní vzorky ve fundamentálním diagramu udává dopravní režim, ve kterém se tenhle vzorek nachází. Rozlišují se dva resp. tři hlavní dopravní režimy. Režim volné dopravy je charakterizován relativně vysokou rychlostí, kterou můžou vozidla dosáhnout a nízkou mírou vlivu ostatních vozidel na trajektorii. Takové stavy jsou zaznamenány v levé části fundamentálního diagramu, kde je závislost dopravního toku na hustotě provozu téměř lineární. V reálné dopravě zodpovídá režim volné dopravy např. pohybu na dálnici vysokou rychlostí. Příkladem synchronizované dopravy je naopak dopravní zácpa, kdy je řidič nucen pohybovat se v závislosti na okolních vozidlech, aby předešel nárazu. Ve fundamentálním diagramu jsou tyhle případy zachyceny v pravé části, kde dochází ke zvyšování hustoty provozu, ale dopravní tok klesá. Někdy je pak extrémní případ dopravní zácpy, kdy vozidla střídavě stojí a pohybují se ve vlnách, brán jako samostatný režim.

1.2 Microscopická struktura

Mikroskopická analýza dopravní vzorky zkoumá jednotlivé vozidla. Detektory zachytávají rychlost $i$tého vozidla v_i a časové okamihy t_i a τ_i kdy přední resp. zadní nárazník vozidla mine detektor. Ze získaných údajů se dopočítává délka vozidla l_i a vzdálenost od $(i{+}1)$ního vozidla r_i pomocí vztahů

$$l_i = v_i(\tau_i - t_i),$$
$$r_i = v_i(t_i - \tau_{i-1}).$$

Pro analýzu dopravního vzorku jsou pak důležité především pravděpodobnostní rozdělení rychlostí $q(v)$ a vzdáleností mezi sousedními vozidly $p(r)$.

2 Termodynamický buněčný dopravní model

Hlavním cílem je pokusit se nastínit univerzální model použitelný na analýzu mikroskopické i makroskopické struktury dopravního vzorku. Jako inspirace nám poslouží Nagelův-Schreckenbergův buněčný model, který dokáže produkovat makroskopické jevy, a dopravní modely založené na termodynamice plynu, ve kterých se tyhle jevy nevyskytují. Termodynamický přístup však narozdíl od prvního zmiňovaného modelu přináší mikroskopickou strukturu odpovídající struktuře reálních dopravných dat. Nový model by měl

tedy ideálně kombinovat výhody obou.
Východiskem pro něj bude model založen na jednodimenzionálním krátkodosahovém termodynamickém plynu, kterého částice se budou pohybovat pouze v diskrétních oblastech, jak tomu je u Nag.-Schreck. modelu. N částic uvažovaného plynu je umístněno na kružnici s obvodem délky N rozdělené na m buněk stejné velikosti $\frac{N}{m}$. Aby měl model smysl, musí platit $m > N$. Částice se pohybují pouze skokově mezi buňkami. Pozice $l$té částice se bude značit x_l a hodnota této veličiny bude udávat pořadové číslo buňky, ve které se $l$tá částice aktuálně nachází. Vzdálenost dvou následujících částice se bude značit r_l a dopočítávat vztahem

$$r_l = (x_{l+1} - x_l)(\frac{N}{m}).$$

S ohledem na rovnost

$$\sum_{l=1}^{N} r_l = N$$

bude její střední hodnota rovna 1, co se dá brát taky jako $\frac{m}{N}$ buněk. Rychlost částice je charakterizována délkou skoku. Omezení maximální rychlostí se tedy promítne do omezení maximální délky skoku, udávané v počtu buněk, která se bude značit w.
Uvažovaný soubor částic je navíc umístněn v teplotní lázni s termodynamickou teplotou T. Z praktických důvodů se bude místo T používat inverzní termodynamická teplota β zavedená vztahem

$$\beta = \frac{1}{kT},$$

kde k reprezentuje Boltzmannovu konstantu.
Jako výchozí konfiguraci částic se může použít ekvidistantní nebo náhodné rozmístnění. Pro oba případy systém směruje k termální rovnováze nezávislé od počáteční konfigurace a závislé od hodnoty β. Termální rovnováha je charakterizována stabilní hodnotou potenciální energie U, hodnota které se pro konkrétní konfiguraci částic vypočítá pomocí vztahu

$$U = \sum_{l=1}^{N} \frac{1}{r_l}. \tag{1}$$

Tedy dosažení termální rovnováhy odpovídá stabilizaci hodnoty této veličiny. Inverzní termodynamická teplota β ovlivňuje hodnotu U po dosažení uvedené rovnováhy. β může být v dopravním modelu prezentována jako parametr udávající stres řidiče. Ten je zapříčiněn hlavně snahou předejít kolizi s vozidlem před ním, dále dopravními předpisy a značkami, stavem silnice, povětrnostními podmínkami a taky aktuálním psychickým stavem řidiče. Hodnotu parametru β v reální dopravě nelze změřit. Víme však, že velikost tohohle dopravního stresu roste s rostoucí hustotou provozu a maximum dosahuje v zácpách. Závisí tedy na aktuálním dopravním režimu a poloze dopravního vzorku ve fundamentálním diagramu.
Modely využívající termodynamický přístup většinou simulují dopravnou situaci s konstantní hodnotou β. Tedy pro vysoké hodnoty tohohle parametru, kdy se předpokládá

dopravní zácpa, se berou v potaz pouze vozidla vytvářející tuhle zácpu. Pro synchronizovaný dopravní režim jsou charakteristické krátké a téměř stejné rozestupy. V termodynamickém modelu se však vzdálenosti přeškálují, aby jejich střední hodnota byla rovna 1. Proto nedochází ke tvorbě dopravních zácep jak je známe z Nag.-Schreck. modelu a fakt, že se jedná o vozidla v synchronizovaném dopravním režimu může být zjištěn jenom analyzováním pravděpodobnostního rozdělení vzdáleností. K pozorování zácpy je potřeba rozlišovat dvě oblasti: oblast samotné dopravní zácpy a oblast, kde k zácpě nedochází. Ty logicky musí mít odlišnou hodnotu parametru β.
Proto bude kružnice v uvažovaném modelu rozdělena na dvě části, přičemž každé části bude přiřazena jiná hodnota inverzní termodynamické teploty β_1 resp. β_2. Tahle modifikace by při vhodném nastavení měla vynutit vznik zácpy. Ještě zbývá vyřešit jak oblasti vymezit. Jelikož jsou disjunktní a souhrnně pokrývají celou kružnici, spokojíme se s vymezením jedné z nich. Nejjednodušším řešením je pevně ji definovat jako část kružnice. To ale příliš neodpovídá reální situaci, kde můžeme pozorovat pohyb zácpy, proto se přistoupilo k definici pomocí částic. Pro potřeby tohoto článku byl požit model, kde je jedna oblast definována jako okolí pevně zvolené částice. Konkrétně šlo o osminu délky kruhu kolem sté částice. Nalezení ideální varianty rozmístnění obou oblastí však zůstává otevřenou otázkou a může sloužit jako motivace pro další výzkum.
Pro simulaci procesu ustalování termodynamické rovnováhy byl použit modifikovaný metropolisův algoritmus obsahující následující kroky:

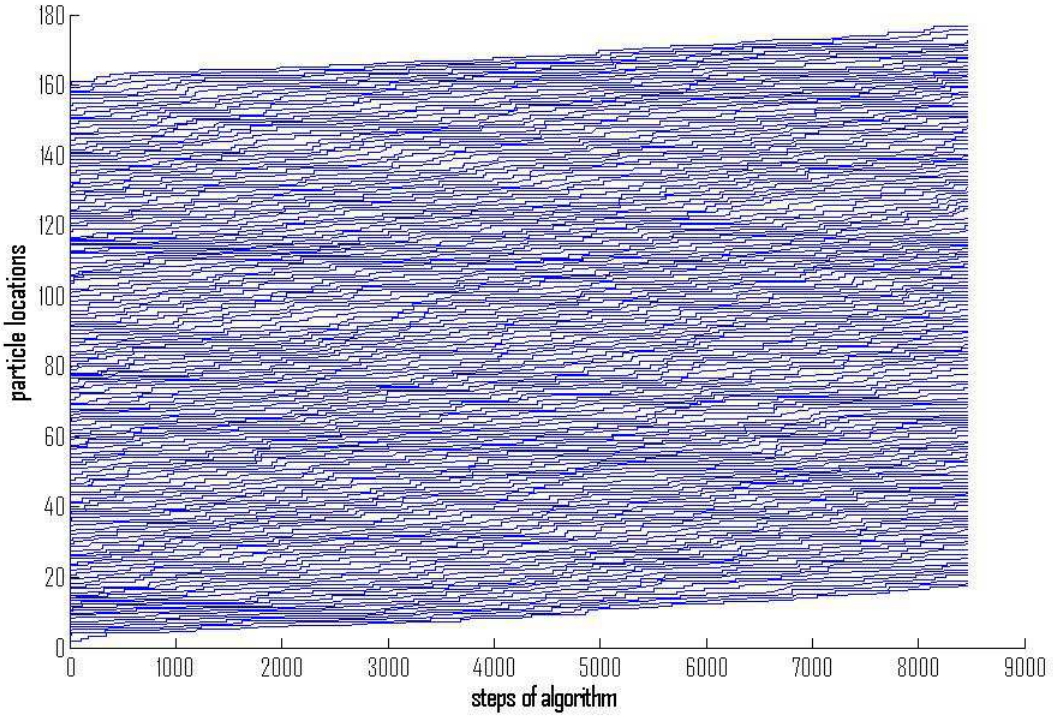
- Je vypočítána hodnota potenciální energie pro aktuální konfiguraci částic pomocí vztahu (1)
- Náhodně je vybrán index l částice která se pokusí o skok.
- Je vygenerováno číslo δ jako realizace náhodné veličiny rovnoměrně rozdělené na intervalu $(0, 1)$ reprezentující náhodný faktor v délce skoku.
- Délka skoku je získána pomocí vztahu $\overline{w} = w\delta$ a následné diskretizace $\tilde{w} = \lceil \overline{w}\delta \rceil$. Horní celá část je použita pro zajištění nenulového výsledku.
- Nová předpokládaná pozice $\acute{x}_l$ $l$té částice je dána vztahem $\acute{x}_l = x_l + \tilde{w}$. V uvažovaném modelu není povolena změna pořadí částic, proto může být nová pozice $\acute{x}_l$ akceptována pouze v případě, že je splněna nerovnost $\acute{x}_l < x_{l+1}$.
- Je vypočítána nová hodnota potenciální energie $\acute{U}$ pro konfiguraci obsahující $\acute{x}_l$.
- Obě hodnoty potenciální energie se porovnají. V případě, že $\acute{U} < U_0$, skok z x_l do $\acute{x}_l$ je přijat a konfigurace změněna. Jinak je vypočten Boltzmannův faktor q jako

  $$q = exp^{-\beta_l \Delta U} = exp^{-(\acute{U} - U_0)}$$

  resp.

  $$q = exp^{-\beta_2 \Delta U} = exp^{-(\acute{U} - U_0)},$$

  přičemž je vybrána hodnota β_l resp. β_2 v závislosti na tom, ve které oblasti se $l$tá částice aktuálně nachází. Pak je vybráno náhodné číslo g rovnoměrně rozděleno



Obrázek 1: Příklad trajektorií částic pro konstantní hodnotu parametru β

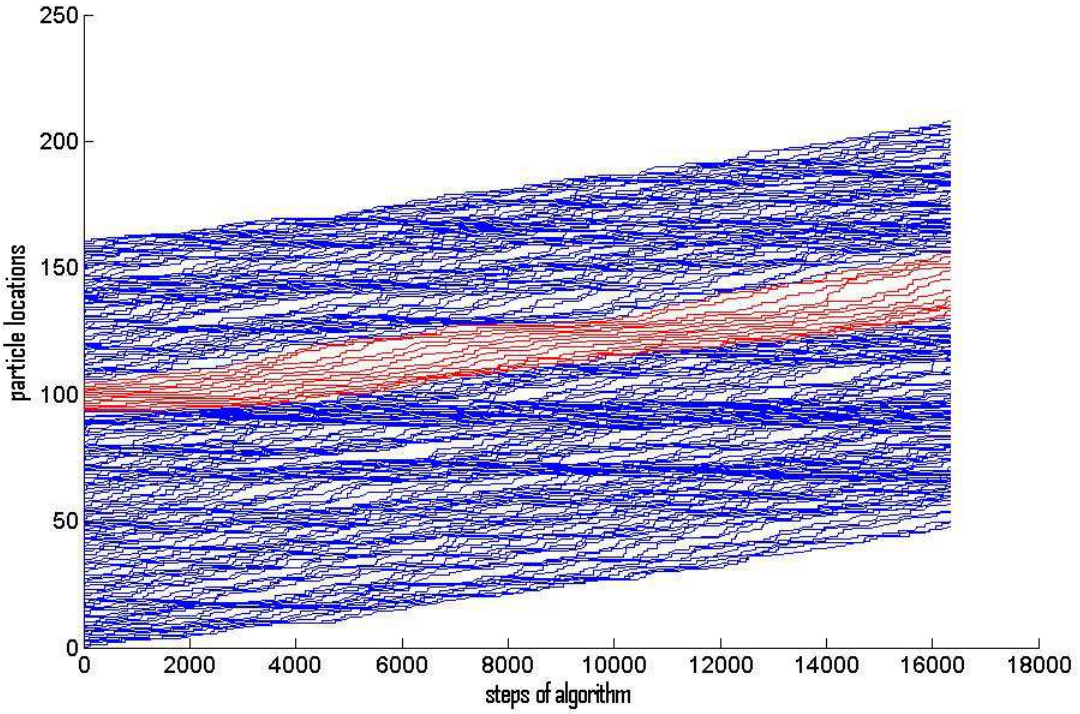
na intervalu $(0, 1)$ a porovnáno s Boltzmannovým faktorem. Při splnění nerovnosti $q > g$ je skok přijat, jinak zůstává konfigurace nezměněna.

3 Výsledky

3.1 Trajektorie částic

Aplikováním předchozího algoritmu získáváme simulaci pohybu vozidel. Sledování změn v pozici konkrétní částice udává její trajektorii. Ty lze po vykonání dostatečného počtu kroků metropolisova algoritmu zobrazit formou diagramu a tak získat představu o rozmístnění částic za daných podmínek. Tahle simulace se může provádět pro rozličné nastavení hodnot parametrů β_l a β_2. Nejprve se však zaměříme na triviální případ $\beta_l = \beta_2$. Model s konstantní hodnotou inverzní termodynamické teploty pro celou kružnici je již dobře znám. Pro názornost si zobrazíme trajektorie částic na grafu (1). Jediné pozorovatelné zhuštění je způsobeno počáteční náhodnou konfigurací a je brzy rozpuštěno.

Nyní již přistoupíme k simulaci pro různé hodnoty β_l a β_2. Jeden z výsledků dané simulace je zobrazen na grafu (2). Odlišná barva představuje oblast s jinou hodnotou parametru β, menší červená oblast ji má vyšší než modrý zbytek. Lze si povšimnout, že na spodním okraji červené oblasti vzniká zhuštění, které se proti směru pohybu šíří modrou oblastí. Naopak na druhé straně se zácpa rozpouští a v horní části modré oblasti se již částice pohybují volně, bez zhuštění. Porovnáním grafů (1) a (2) lze dojít k závěru,



Obrázek 2: Příklad trajektorií částic pro dvě rozdílne hodnoty β_1 a β_2

že předpokládané zácpy jako zástupci makroskopických jevů se vyskytují pouze na grafu (2). A tedy modifikace uvedená v předchozí kapitole pravděpodobně vyvolává jejich vznik. Ještě zbývá ověřit jestli se zachovala mikroskopická struktura, která u modelů založených na termodynamickém plynu koresponduje se strukturou reálných dopravných vzorků.

3.2 Rozdělení vzdáleností

Pro analyzování mikroskopické struktury bude důležitý především tvar pravděpodobnostního rozdělení vzdáleností dvou následujících částic. Ten byl v termodynamickém dopravním modelu odvozen [1] z Hamiltoniánu daného systému

$$H = \frac{m}{2}\sum_{i=1}^{N}(v_i - \overline{v})^2 - C\sum_{i=1}^{N}\frac{1}{r_i},$$

kde m označuje hmotnost částice a $\overline{v}$ průměrnou rychlost v souboru. Výsledkem je pak funkce

$$P(r) = \Theta(r)Aexp[-\frac{\beta}{r} - Br], \tag{2}$$

kde $\Theta(r)$ označuje Heavisidovu funkci, β představuje již zmíněnou inverzní termodynamickou teplotu a A a B jsou normalizační konstanty. Jejich hodnoty se získávají z dvou normalizačních rovnic.

$$\int_0^\infty Aexp[-\frac{\beta}{r} - Br]\mathrm{d}r = 1,$$

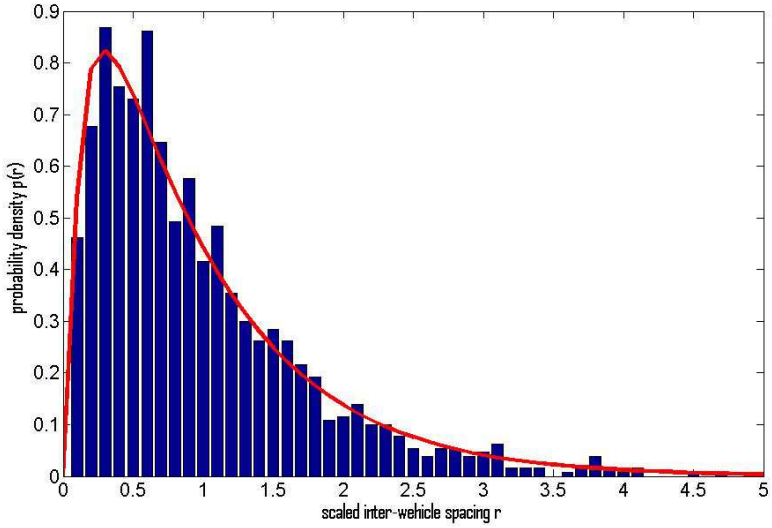
$$\langle r\rangle = \int_0^\infty rAexp[-\frac{\beta}{r} - Br]\mathrm{d}r = 1.$$

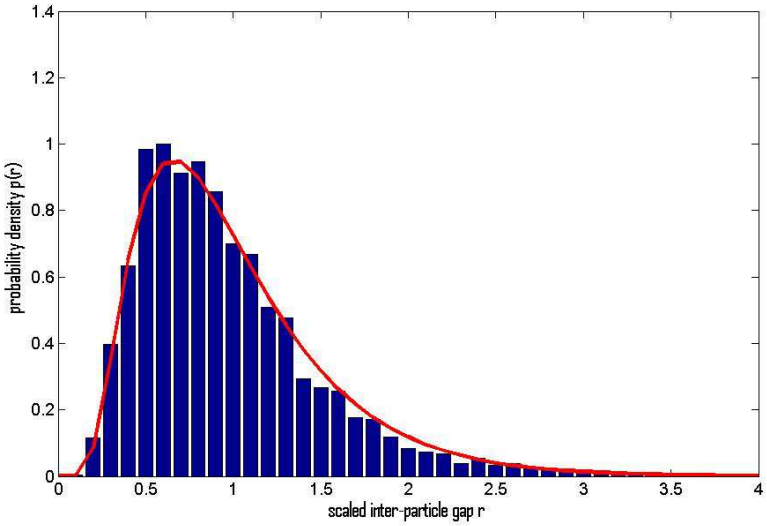
První vztah popisuje vlastnost každé hustoty pravděpodobnosti, druhý zajišťuje střední hodnotu r rovnou 1. Jejich aproximativní vyjádření v závislosti na hodnotě parametru β je následující

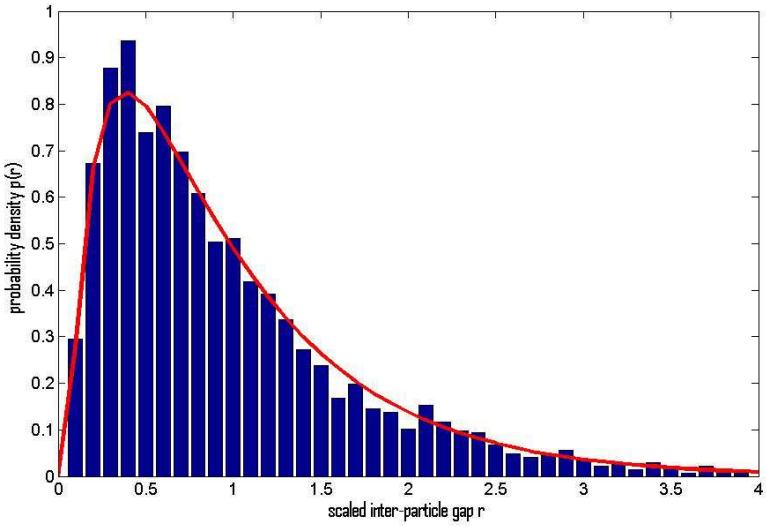
$$B = \beta + \frac{3 - exp[sqrt\beta]}{2},$$

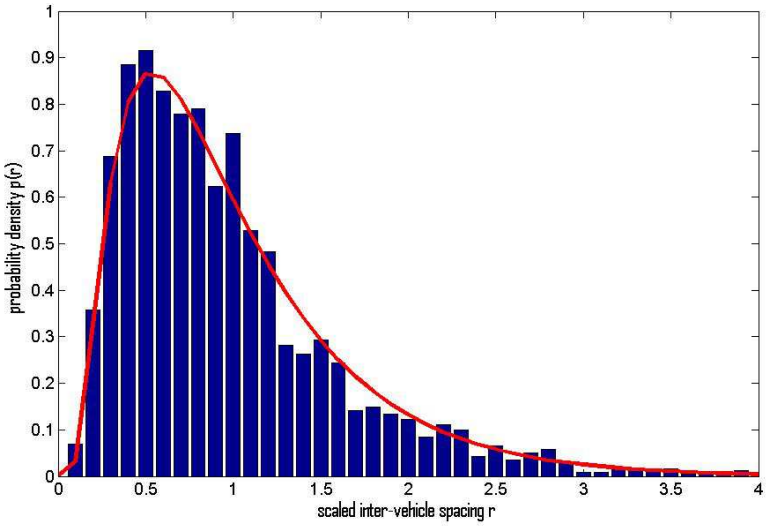
$$A^{-1} = 2\sqrt{\frac{\beta}{B}}K_{1(2\sqrt{\beta B})},$$

kde $K_1(x)$ označuje modifikovanou Besselovou funkci druhého druhu.
Získaný tvar pravděpodobnostního rozdělení vzdáleností odpovídá struktuře reálných dopravných dat, jak již bylo dokázáno. Tato funkce je vhodnou aproximací jak u vzorků v režimu volné dopravy, tak u těch, kde dochází ke zhuštění, rozdíl je v hodnotě parametru β. Přibližně platí, že režim volné dopravy koresponduje s nižšími hodnotami inverzní termodynamické teploty, zatímco při zácpách hodnota tohohle parametru stoupá. Pro extrémní případ $\beta = 0$ jde o Poissonovo rozdělení, které se používá pro nezávisle se pohybující částice.
Pro získání pravděpodobnostního rozdělení ve zkoumaném modelu se použijí data ze simulací. Aby se zamezilo vlivu počátečního rozmístnění, budou se analyzovat pouze konečné konfigurace po ustálení rovnováhy. Taky je potřeba provést více běhů metropolisova algoritmu se stejnými parametry pro dostatek dat. Kružnice v modelu byla rozdělena na dvě části s rozdílnými hodnotami parametru β, čemu odpovídají také odlišné hustoty pravděpodobnosti vzdáleností. Proto musí být i částice získané ze simulace rozdělené do dvou souborů podle toho, ve které oblasti se nacházejí. Uvážíme-li fakt, že na okrajích obou oblastí dochází k deformacím, musíme ještě ze souborů vyloučit částice z okrajů oblastí. Tím ale zůstane dost dat pro analýzu poskytující smysluplné výsledky pouze ve větší z oblastí a proto dále pracujeme jenom s ní. Samotnou funkci hustoty pravděpodobnosti vzdáleností získáme jako histogram vzdáleností částic z analyzovaných údajů. Protože hodnota inverzní termodynamické teploty β pro analyzovaný soubor je známá, je dán i požadovaný tvar pravděpodobnostního rozdělení a to (2)


probability density p(r)
scaled inter-vehicle spacing r


probability density p(r)
scaled inter-particle gap r


probability density p(r)
scaled inter-particle gap r



Histogram znázorňuje funkci hustoty pravděpodobnosti získanou ze simulací a křivka reprezentuje její požadovaný tvar.

Na obrázcích (3.2) jsou zobrazeny porovnání hustoty pravděpodobnosti ze zkoumaného modelu oproti předpokládanému tvaru pro různé hodnoty parametru β. Vizuálním srovnáním je potvrzeno, že výsledné histogramy zodpovídají požadovanému tvaru funkce. Drobné odchylky můžou být způsobené nedostatkem dat v souborech.

4 Závěr

Výsledky simulací a prvních analýz naznačují, že model představen v tomhle článku může být krokem správným směrem k vyvinutí univerzálního dopravního modelu spojující lokální a globální pohled na dopravu. Ten by pak kvantitativně i kvalitativně vysvětlil všechny jevy a zákonitosti známe ze silnic a dálnic. Prvním pozitivním impulzem je výskyt dopravní zácpy jako makroskopického fenoménu v modelu, který zachovává požadovanou mikroskopickou strukturu. Mělo by to sloužit jako motivace pro podrobnější analýzu právě představeného modelu, jako i zkoumání dalších možností modifikace. Klíčem ke globálnějšímu pohledu na dopravu je jak se zdá upuštění od předpokladu konstantní inverzní termodynamické teploty β v celém zkoumaném vzorku.

Literatura

[1] M. Krbálek. *Equilibrum distributions in thermodynamical traffic gas.* J. Phys. A: Math. Theor **40** (2007).

[2] Nagel, K. and M. Scgreckenberg. *A cellular automaton model for freeway traffic.* Journal de Physique I, France, **2** (1992).

[3] Krbálek, M. and D. Helbing *Determination of interaction potentials in freeway traffic from steady-state statistics.* Physica A, **333** (2004).

[4] Krbálek, M. *Inter-vehicle gap statistics on signal-controlled crossroads*. J. Phys. A: Math. Theor **42** (2009).

Nonparametric Predictive Control via Lazy Learning and Stochastic Optimization

Karel Macek

4th year of PGS, email: karel.macek@fjfi.cvut.cz
Department of Software Engineering in Economy
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Jaromír Kukal, Department of Software Engineering in Economics, FNSPE, CTU in Prague

Abstract. This paper provides a novel method for predictive control of a stochastic dynamic system. The prediction is implemented by local regression which makes no assumptions about the model's structure and parameters. The method works on the principle of dynamic programming form the horizon to actual time instant. The paper focuses on risk-averse optimization related to the distribution of total loss. Hence, the uncertainity is sustained by additional data points.

Keywords: local regression, dynamic decision making, risk management, stochastic programming

Abstrakt. Tento příspěvek předkládá novou metodu prediktivního řízení pro stochastické dynamické systémy. Predikce je realizována pomocí lokální regrese, která neklade žádné předpoklady na strukturu modelu a jeho parametry. Metoda pracuje na principu dynamického programování. Přípěvek věnuje optimalizací za averze vůči riziku. Proto je informace o neurčitosti udržována pomocí přídatných datových bodů.

Klíčová slova: lokální regrese, dynamické rozhodování, řízení rizik, stochastická programování

1 Introduction

Two directions have been distinguished in advanced control, namely model oriented and data oriented [7] approaches. Model oriented systems require some explicit prior knowledge based on first principles. However, for some systems, no explicit parameterized model is known since (i) the science have not already provided relevant first-principles, (ii) modeling of the system by first principles would be extremely complicated, or (iii) the process is very non-stacionary and it is impossible to capture it by one parameterized model, even the parameters would be addaptive. This situation occurs typically when the operation of the system is impacted by human factors, e.g. in Internet activities, stock exchange or also HVAC systems, especially in buildings large complexes.

Absence of an explicit model make eventual control difficult. Some authors solve the problem by some approximation, very often by neural networks [2, 10], speaking about model free control. In fact, they use a model, even it is a black box model.

This paper provides a novel combination of data-centric regression and stochastic optimization for risk-dirigible predicitve control. It is organized as follows: Section 2 provides used building blocks, i.e. lazy learning and stochastic optimization. Next, Section 3 formulates the control algorithm and discusses some of its properties. Then,

Section 4 demonstrates the function of the algorithm on a simple examples. Finally, Section 5 summarizes the paper.

2 Building Blocks

2.1 Value at risk optimization

Before we will introduce the system and its control, it is necesarry to define the optimization paradigm as such. Since we are going to optimize a stochastic function in terms of value at risk, we have to define these terms:

Definition 1 (Value At Risk). *Let Z be a random variable with cdf H_Z and let $\alpha \in [0,1]$, value of Z at risk α is then defined as:*

$$\begin{aligned} \mathrm{V@R}_\alpha(Z) &= H_Z^{-1}(1-\alpha) && (1)\\ &= \inf\{t:\ \boldsymbol{P}(Z \le t) \ge 1-\alpha\} && (2)\\ &= \inf\{t:\ \boldsymbol{P}(Z > t) \le \alpha\} && (3) \end{aligned}$$

Definition 2 (Random mapping). *$\mathbb{R}^m \to \mathbb{R}^n$ is a mapping where to each $\mathbb{x} \in \mathbb{R}^m$ is assigned only one distribution over $\mathbb{R}^m$. We will use the notation $(Y|X=x)$ for this where x is called decision.*

Note that if we consider X to be a random vector and a conditioned random vector is given $Y|X$. Random mapping, in this case assinges to the distributions X and $Y|X$ the marginalization $Y:\ f(y) = \int f(y|x)f(x)dx$.

The radnom mapping is therefore defined for both point and probabilistic decisions which is more general and enables eventual fully probabilistic design. It has to be mentioned that in both cases the decision returns a probability distribution of Y.

Now, we are able to formulate

Definition 3 (Value At Risk Optimization Problem). *Let*

1. *$Y|X$ be the objective function with the risk level α_o*
2. *$E|X$ be equality constraints with the risk levels $\alpha^{(e)}$*
3. *$G|X$ be unequality constraints with the risk levels $\alpha^{(g)}$*

The goal of the value-at-risk optimization problem is to find such (random) vector $X0$ so it holds:

- *$\forall i : \boldsymbol{P}(E_i \neq 0|X0) \le \alpha_e$ and*
- *$\forall i : \boldsymbol{P}(G_i \ge 0|X0) \le \alpha_e$ and*
- *$\mathrm{V@R}_{\alpha_o}(Y|X0)$ is minimal.*

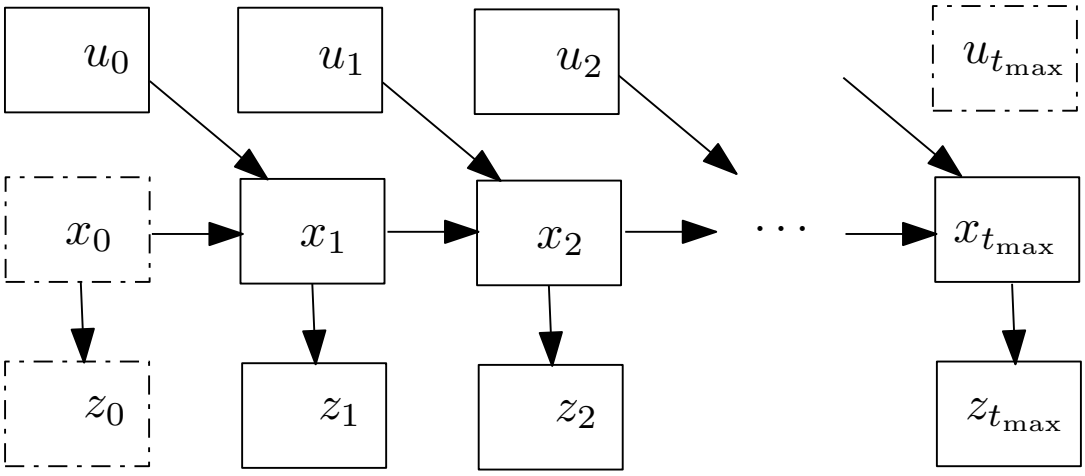


Figure 1: System dynamics for predictive horizon $t_{\max}$ where dashed boxes are not considered for the optimal control.

where $\alpha = 0.05$, typically.

It have to be mentioned here that this problem may have no solution. The reason can be that the constraints are in conflict. This is a typical situation for all constrained optimization problems. However, there might be also an issue with the probabilities. Let us consider $E_1 = x - 1 + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Even the $\sigma^2 > 0$ is arbitrary small and $x = 1$, the $\boldsymbol{P}(E_1 = 0) = 0$ almost surely since the distribution is continuous. Hence, the constraints in form of equation require a tie at zero or they have to be replaced by intervals around zero.

2.2 $V@R$ control

We will consider a Markovian system of order one (the generalization to higher orders is straightforward) with action and observation models as probability distribution functions:

$$F(X(t)|X(t-1), U(t-1)) \tag{4}$$
$$G(Y(t)|X(t), U(t)) \tag{5}$$

The system dynamics is presented in Figure 1 where the arrows correspond to the stochastic mappings F, G. Mention that we use $G(Y(t)|X(t)$ instead of $G(Y(t)|X(t), U(t))$ since the loss depending on $U(t)$ can be considered as a part of the state $U(t+1)$. Let F and G have densities functions f and g. The loss function is defined as a sum of partial losses $z = \sum_{t=1}^{t_{\max}} z(t)$. All state variables are measured, this approach does not use any unobserved variables. Consider random mapping $Z|U_0, U_1 \ldots, U_{t_{\max}-1}, X_0$ with following

cdf:

$$\underbrace{\iiint}_{t_{\max}} \prod_{t=2}^{t_{\max}} \left\{ \underbrace{\iiint}_{t-1} f(z(t)|x(t)) \prod_{i=0}^{t} g(x(t)|x(t-1),u(t-1)) dx(t_{\max}-1)\dots dx(1) \right\} \cdot$$

$$\cdot f(s - \sum_{t=2}^{t_{\max}} z_t | x(1)) dz(1)\dots dz(t_{\max})$$

Then the problem is to minimize z via changing the control strategy $u(0), u(1)\dots u(t_{\max}-1)$ so the expression equals given $\alpha \in [0,1]$. We make no assumption on g and f, we only assume a data set of (x,u,z) measurements from the past. Note that

$$\prod_{i=1}^{t} g(x(t)|x(t-1),u(t-1)) \tag{6}$$

is application of the chain rule.

There is already intensive work on risk-averse control of stochastic systems. The risk aversion can be expressed in form of the utility function [12]. In this case, usual application of mean value operator for each step of the dynamic programming is possible, for more detail on favorable properties of the mean value in stochastic dynamic programming see [14].

Usage of true risk measures requires alternative approach. The risk measures are adjusted so they can be pushed back in the dynamic programming process. This claim requires more assumptions like convexity, subadditivity and homogenity of the risk measures [1, 13] which do not hold for $V@R$. and are focused mostly no discrete systems. This paper however provides a method which is able to work with wider class of risk measures, including the $V@R$.

2.3 Local Regresion and Conficence Sets

Local regression is an essencial topic in nonparametric statistics. Let us consider following dependency:

$$Y = f(x) + \epsilon(x) \tag{7}$$

where $Y \in \mathbb{R}$ and $x \in M$, M is a metric space with metric $d : M \times M$ and $\epsilon(x)$ is a random variable with zero mean.

Local regression based on kernels provide methods to construct the estimate of $\hat{r} \approx r$ and provide eventually some information about the error $\epsilon(x)$.

The estimator is defined as follows:

$$\hat{r}(x) = \sum_{i=1}^{n} l_i(x) Y_i \tag{8}$$

where

$$l_i(x) = \frac{K\left(\frac{x-x_i}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{x-x_j}{h}\right)} \tag{9}$$

and K is a kernel function[1].

The local regression makes no assumptions on model if sufficient amount of data is available. The local regression is also applicable in the dynamic systems and the nonstacionarity can be covered easily [8]. For our purpose of risk-dirigible control, we have to treat the risk. There are some alternatives where the most recent is based directly on considered quantile regression [4]. In this work we will nevertheless assume the errors $\epsilon(x)$ to be distributed normally. In this case we can adopt the concept of confidence bands [18]. First of all, the variance have to be estimated:

1. Estimate the $\hat{r}$
2. Define $Z_i = \log(Y_i - \hat{r}(x_i))^2$
3. Estimate the $\hat{q}$ as local regression of the Z_i depending on x_i
4. Obtain $\hat{\sigma}^2(x) = e^{\hat{q}(x)}$

This holds only for points from the training set while for the other the distance has some impact on the variance. The adjustement is perfomred as follows:

$$s(x) = \sqrt{\sum_{i=1}^{n} \hat{\sigma}^2(x) l_i^2(x)} \tag{10}$$

The confidence band can be expressed in form $[\hat{r}(x) - cs(x), \hat{r}(x) - cs(x)]$. The constant $c > 0$ can be find for one dimensional in an explicit sophisticated way [16]. For our purposes we adopted - however - direct quantile prediction [3] which consist in minimization:

$$\frac{K(x - x_i)}{h} \rho\left(Y_i - \theta\right) \tag{11}$$

where $\rho_\alpha(x) = |x| + (2\alpha - 1)x$ and θ the desired α-percentile.

2.4 Stochastic Optimization

In this work, we have to minimize the $\mathrm{V@R}_\alpha(Y|x)$ with respect to x. Using fixed α, the problem becomes deterministic. The objective function is - however - very general and usual suitable properties like linearity or convexity cannot be ensured. Therefore, methods of stochastic optimization can be applied. Both population based or single point methods can be used, e.g. the differential evolution [15], particle swarm optimization [5], threshold accepting [19], simulated annealing etc. Furthemore, it is possible to use some kind of advanced metaheuristics like memetic algorithms [6], DEBR18 [17] or OSOOM [11].

If we want to treat constraints, it is possible to adopt lexicographic comparison mechanism [9]:

[1] In multidimensional case, the it is possible to use usuall kernell function with a metric, e.g Eucidian [18]. Alternatively, contributions in dimensions can be multiplied.

- If x_1 and x_2 satisfy both the stochastic constraints at least at required level α_1 then, they are compared by their $\mathrm{V@R}_{\alpha_2}(x_1)$ and $\mathrm{V@R}_{\alpha_2}(x_1)$ values.
- Otherwise, x_1 and x_2 are compared by their probabilities they will both satisfy the constraints.

Finally, it have to be added that deterministic constraints are a special case of the stochastic ones.

3 Proposed Algorithm of Predictive Control

Predictive control attempts to find a control strategy for a predictive horizon. Proposed algorithm works on principle of dynamic programming:

1. **Initialization:** $t = t_{\max} + 1$, $C_t = 0$
2. **Update:** $C_{t-1} = z_i(X) + C_t$
3. **Regression:** Having data (X, U, C_t), regress intented percentil α of $C_t|U, X$. However for the next step, it is necessary to regress also other representative percentiles or parameters of $C_t|U, X$
4. **Optimization:** For all X from the original set find $U^* = \arg\max_U \mathrm{V@R}_\alpha(C_t|U, X)$. For the step t, related data should be used with higher weights, e.g. for $t = t_{\max}$, last training data will be prefered. This might help to avoid the impact of potential nonstacionarity. These weights can be used as additional input for weighting in the local regression.
5. **Particle Filtering:** For all pairs (X, U) optimizing the $\mathrm{V@R}_\alpha(C_t)$ sample some values from $C_t|X, U$ and use it as a data set for the next iteration.
6. **Next step** Go to 2)

4 Application

This work is intended to be applied in two applied problems where the risk aversion plays an important role and parameterized models are not available. The first application is in supervisory control of HVAC[2] systems. The goal is to ensure required comfort and minimize operational costs. Nevertheless, the demands of zones in a building complex on heat and fresh air is impacted by human factor, changing operation. Some dynamics can be covered by first-principle modeling like heat accumulation in internal mass, air flows etc. However, there are usualy not enough sensors required for this approach.

The other application area is the portfolio management where the risk is an important topic. Also in this case, parameterized models are missing. Hence, the data-centric risk averse control seems to be a promising.

[2]Heating, ventilation, air conditioning.

Finally, this topic relates to author's PhD thesis which attempts to control a optimization procedure composed of various optimization methods. The author adopts in the thesis proposal also the risk averse approach that should lead to reliable results in the optimization.

5 Conclustion

Proposed approach has proved high level flexibility and genericity. The major disadvantages consists however in the computational complexity since each evaluation of the objective function has to process the whole data set and this makes the optimization very slow. This is also a key point of possible improvements. The data set could be reduced to a reasonable reference data set (like in case of Support Vector Machines), nevetheless, this reduction might worsen the estimates of the confidence bands. Alternatively, the process of query for the objective function calculation could be speeded up by an advanced data structure.

Other topic to be improved is the confidence interval estimation which are typically dependent, at least for several subsequenting steps.

References

[1] P. Artzner, F. Delbaen, J. M. Eber, and D. Heath. Coherent measures of risk. *Mathematical Finance*, 9(3), 1999.

[2] Leandro dos Santos Coelho, Marcelo Wicthoff Pessôa, Rodrigo Rodrigues Sumar, and Antonio Augusto Rodrigues Coelho. Model-free adaptive control design using evolutionary-neural compensator. *Expert Systems with Applications*, 37(1):499 – 508, 2010.

[3] Ali Gannoun, Jérôme Saracco, and Keming Yu. Nonparametric prediction by conditional median and quantiles. *Journal of Statistical Planning and Inference*, 117(2):207 – 223, 2003.

[4] Wolfgang Karl Härdle, Ritov Ya'acov, and Song Song. Partial linear quantile regression and bootstrap confidence bands. SFB 649 Discussion Papers SFB649DP2010-002, Humboldt University, Collaborative Research Center 649, 2010.

[5] James Kennedy and Russell C. Eberhart. Particle swarm optimization. page 1942–1948, Piscataway, NJ, 1995. IEEE Int. Conf. on Neural Networks.

[6] Natalio Krasnogor and Steven Gustafson. Toward truly "memetic" memetic algorithms: discussion and proofs of concept. In *Advances in Nature-Inspired Computation: The PPSN VII Workshops. PEDAL (Parallel, Emergent and Distributed Architectures Lab). University of Reading. ISBN 0-9543481-0-9. icalp.tex; 9/12/2003; 16:52; p.21 22 Natalio Krasnogor, Steven Gustafson*, pages 0–9543481, 2002.

[7] Rudolf Kulhavý. Data-centric decision support. In *Proceedings of the American Control Conference, 2002*, volume 4, pages 3395 – 3400 vol.4, 2002.

[8] Rudolf Kulhavý. Bayesian analysis of stochastic system dynamics. In *Proceedings of the 25th International Conference of the System Dynamics Society*, 2007.

[9] Tomáš Kunt. Methods of nonlinear optimization (in czech). Bachelor project, Faculty of Nuclear Sciences and Physical Engineering, CTU, 2010.

[10] Ivana Lukáčová, Ján Pitel, and Tomáš Saloky. Model-free adaptive heating process control. In *XXXIV. Seminar ASR '2009 "Instruments and Control"*, 2009.

[11] Karel Macek, Josef Boštík, and Jaromír Kukal. Reinforcement learning in global optimization heuristics. In *Mendel, 16ts International Conference on Soft Computing*, 2010.

[12] Zuzana Macova and Daniel Sevcovic. Weakly nonlinear analysis of the hamilton-jacobi-bellman equation arising from pension savings management. Quantitative Finance Papers 0905.0155, arXiv.org, May 2009.

[13] André Mundt. *Dynamic risk management with Markov decision processes.* PhD thesis, Universität Karlsruhe, 2007.

[14] Ivan Nagy, Lenka Pavelková, Evgenia Suzdaleva, Jitka Homolková, and Miroslav Kárný. Bayesian decision making. Technical report, UTIA AVČR, 2005.

[15] Reiner Storn and Ken Price. Differential evolution - a simple and efficient heuristic for global optimization. *Journal for Global Optimization*, 11(341-359), 1997.

[16] Jiayang Sun, Clive Loader, and William P. Mccormick. Confidence bands in generalized linear models. *Ann. Statist.28*, 2:429–460, 1994.

[17] Josef Tvrdík. Adaptation in differential evolution: A numerical comparison. *Applied Soft Computing*, 9(3):1149 – 1155, 2009.

[18] Larry Wasserman. *All of Nonparametric Statistics.* Springer Berlin / Heidelberg, 2006.

[19] Peter Winker. *Optimization Heuristics in Econometrics.* Willey Series in Probability and Statistics. Willey, 2001.

Využití hexagonální topologie 2D obrazu k diagnostice Alzheimerovy demence

Jakub Nerad

1. ročník PGS, email: jakubnerad@gmail.com
Katedra softwarového inženýrství v ekonomii
Fakulta jaderná a fyzikálně inženýrská, ČVUT v Praze

školitel: Jaromír Kukal, Katedra softwarové inženýrství v ekonomii, Fakulta jaderná a fyzikálně inženýrská, ČVUT v Praze

Abstract. This paper deals with image processing methods in the hexagonal topology. The first part describes the implementation of the hexagonal grid. In the second and third part deals with the various operations of the hexagonal image (noise reduction, edge detection and morphological operations). The last section is devoted to the use of hexagonal topology in real biomedical data for diagnosing Alzheimer's disease.

Keywords: image processing, hexagonal topology, Alzheimer's Disease

Abstrakt. V této práci se zabývám metodami zpracování obrazu v hexagonální topologii. V první části práce se věnuji implementaci hexagonální mřížky. Ve druhé a třetí části se věnuji jednotlivým operacím nad hexagonálním obrazem (odstranění šumu, detekce hran a morfologické operace). V poslední části se věnuji využití hexagonální topologie na reálných biomedicínských datech k diagnostice Alzheimerovy choroby.

Klíčová slova: zpracování obrazu, hexagonální topologie, Alzheimerova nemoc

1 Úvod

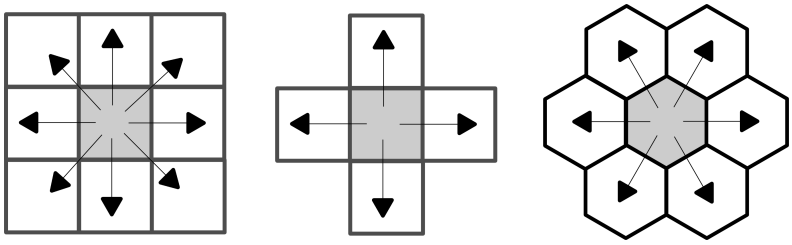
V oblasti zpracování obrazu je často zmiňováno využití hexagonální mřížky jako alternativa ke konvenční čtvercové. Ovšem nebývá toto téma příliš rozvinuto. Cílem tohoto textu je tuto problematiku popsat podrobněji a ukázat její výhody na reálných datech.

Hexagonální zpracování obrazu se od tetragonálního se liší nejčastěji v tvaru konvolučních masek, které musí odpovídat návrhu vzorkovací mřížky. Tu je vhodné tedy implementovat tak, aby převod známých konvolučních masek byl co nejsnazší a nejpřesnější. Vytvoření hexgonální mřížky je stěžejní bod, od kterého se odvíjí další postupy při zpracování obrazu.

2 Výhody hexagonálního zpracování obrazu

Hexagonální obraz vznikne vzorkováním signálu do mřížky, která je tvořená pravidelnými šestiúhelníky. Z geometrických vlastností šestiúhelníku vyplývají základní vlastnosti této topologie:

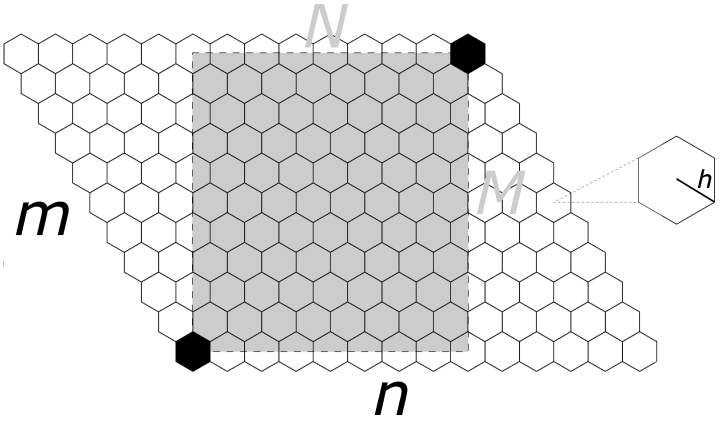
- Šestiúhelník zahrnuje více prostoru než čtverec, čímž lépe aproximuje kruh. To znamená, že hustota vzorkování hexagonální mřížkou je větší než u čtvercové mřížky.
- Každý obrazový element má šest stejně vzdálených sousedů, kteří sdílí hranu. Z toho vyplývá, že křivky mohou být v hexagonální mřížce mnohem lépe zastoupeny než ve čtvercové.
- Existuje jednoznačné okolí bodu, na rozdíl od čtvercové mřížky, kde okolí bodu může být definováno jako *4-sousedství* (sousedé jsou jen ti, kteří sdílí hranu) nebo *8-sousedství* (sousedé jsou i ti, kteří sdílí roh).



Obrázek 1: Okolí bodu ve čtvercové mřížce (8-sousedství a 4-sousedství) a hexagonální mřížce.

3 Implementace hexagonální mřížky

Není mi známo, že by se v praxi vyskytovaly přístroje, které zaznamenávají obraz přímo do hexagonální mřížky. K dispozici mám pouze běžné obrazy zachycené do čtvercové mřížky, které se musí do hexagonální interpolovat.



Obrázek 2: Konverze mezi čtvercovou a hexagonální topologií.

Rozměry vstupního obrazu I označíme M a N, další známou hodnotou pro výpočet je poloměr kružnice h opsané šestiúhelníku, který volíme. Tyto údaje stačí k výpočtu rozměrů m, n matice H. Dále potřebujeme zjistit počáteční body x_0 a y_0.

Pro souřadnice levého dolního rohu platí $x \geq N$ a $y \geq 1$ a pro pravý horní roh $x \leq 1$ a $y \geq M$. Z toho vyplývají následující vztahy:

$$x_0 + h(m-1) + \frac{h}{2}(m-m) \geq N \Rightarrow x_0 \geq N - h(n-1) \tag{1}$$

$$y_0 + \frac{h}{2}\sqrt{3}(m-m) \geq 1 \Rightarrow y_0 \geq 1 \tag{2}$$

$$x_0 + h(1-1) + \frac{h}{2}(m-1) \geq 1 \tag{3}$$

$$1 + \frac{h}{2}\sqrt{3}(m-1) \geq M \Rightarrow 1 \tag{4}$$

Z výrazů plyne

$$m = \left\lceil \frac{2(M-1)}{h\sqrt{3}} \right\rceil + 1, \qquad n = \left\lceil \frac{(m-1)}{2} + \frac{(N-1)}{h} \right\rceil + 1 \ . \tag{5}$$

Hodnoty souřadnic obrazových elementů jsou

$$x_{i,j} = x_0 + h(j-1) + \frac{h}{2}(m-i), \qquad y_{i,j} = y_0 + \frac{h}{2}\sqrt{3}(m-i),$$
$$\text{kde } i = 1, \ldots, m; j = 1, \ldots, n. \tag{6}$$

4 Zpracování obrazu

Důležitou operací při zpracování obrazu je *konvoluce*. Konvoluce dvourozměrných funkcí f a h je definována integrálem

$$\mathrm{f}(x,y) * \mathrm{h}(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathrm{f}(x-a, y-b)h(a,b)\mathrm{d}a\mathrm{d}b \ ,$$

kde $h(x,y)$ je konvoluční jádro.

V digitální zpracování obrazu se využívá diskrétní konvoluce, diskretizace předchozího integrálu zní

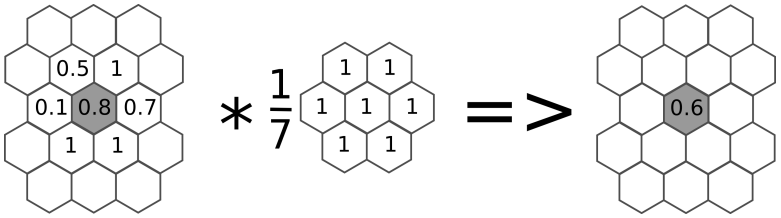
$$\mathrm{I}(x,y) * \mathrm{h}(x,y) = \sum_{j=-k}^{k}\sum_{j=-k}^{k} \mathrm{I}(x-i, y-j)\mathrm{h}(i,j) \ ,$$

kde I je diskrétní obraz a $h(x,y)$ jádro konvoluce.

4.1 Nízkofrekvenční filtry

Nízkofrekvenční filtry slouží k vyhlazování obrazu, jejich cílem je odstranění nežádoucího šumu, který se v obraze nachází.

Pokud máme více obrazů stejné předlohy, můžeme použít k odstranění šumu postup, kdy zprůměrujeme hodnoty obrazových bodů o stejných souřadnicích. Výhodou tohoto postupu je, že nedochází k rozmazání obrazu. Ovšem ve většině případů při zpracování obrazu máme k dispozici pouze jednu předlohu a musíme tedy použít jiné metody.



Obrázek 3: Ukázka konvoluce v hexagonálním obrazu.

Metody pracující pouze s jedním obrazem spoléhají na nadbytečné informace v obraze. Sousední obrazové body mívají většinou stejnou nebo velmi podobnou úroveň jasu. Zašuměné obrazy lze tak na základně provedení analýzy okolních bodů opravit. Hodnota vybraného bodu je nahrazena hodnotou typickou v jeho okolí. Typickou hodnotou mám na mysli např. průměr, výběrový průměr nebo medián.

Základní metodou vyhlazování obrazu je průměrování. Jedná se o lineární vyhlazování. Každému bodu je přiřazen nový jas, který je aritmetickým průměrem původních jasů ve zvoleném okolí. Obraz je zpracován konvoluční maskou, která popisuje chování funkce $\mathrm{h}(i,j)$ s tím, že střed masky má souřadnice $(0,0)$. Může mít například tvar

$$h = \frac{1}{7}\begin{bmatrix} 1 & 1 & * \\ 1 & 1 & 1 \\ * & 1 & 1 \end{bmatrix} . \tag{7}$$

Tento neobvyklý tvar masky (matice) je dán implementací hexagonální mřížky. Hvězdičky představují prázdnou (neexistující) hodnotu. Například v programovacím prostředí Matlab jsou implementovány jako `NaN` (Not-a-Number).

Můžeme samozřejmě volit i masky jiné velikosti, dokonce masky o rozměru 5×5 se v hexagonální topologii osvědčily. Ve své praktické části používám i zajímavé masky dvou „poloměrů“. Vnější poloměr R udává celkovou velikost masky a vnitřní r je poloměr okolí, na které nebude při zpracování brán zřetel. Platí, že $R > r \geq 0$.

Příkladem může být maska

$$h = \frac{1}{12}\begin{bmatrix} 1 & 1 & 1 & * & * \\ 1 & 0 & 0 & 1 & * \\ 1 & 0 & 0 & 0 & 1 \\ * & 1 & 0 & 0 & 1 \\ * & * & 1 & 1 & 1 \end{bmatrix} , \tag{8}$$

kde $R = 2$ a $r = 1$.

Ovlivnit proces zpracování lze i zvýhodněním respektive znevýhodněním, některé oblasti v konvoluční masce. Typicky to může být zvětšením váhy bodů blízkých středu masky. Například jako

$$h = \frac{1}{10}\begin{bmatrix} 1 & 1 & * \\ 1 & 4 & 1 \\ * & 1 & 1 \end{bmatrix} \text{ nebo } h = \frac{1}{27}\begin{bmatrix} 1 & 1 & 1 & * & * \\ 1 & 2 & 2 & 1 & * \\ 1 & 2 & 3 & 2 & 1 \\ * & 1 & 2 & 2 & 1 \\ * & * & 1 & 1 & 1 \end{bmatrix} . \tag{9}$$

U těchto masek můžeme místo obyčejného průměru volit i *vážený průměr*.

Pro odstranění šumu lze využít i masky, jejichž váhy odpovídají hodnotám *Gaussovy funkce*.

Gaussovo rozdělení pro 2D je definováno jako

$$\mathrm{G}(x,y) = \frac{1}{\sqrt{2\pi\sigma}} \mathrm{e}^{-\frac{x^2+y^2}{2\sigma^2}} \ .$$

Diskrétní aproximací získáme konvoluční masku.

Hlavní nevýhodou lineárních metod vyhlazování je rozmazávání hran. Toto lze řešit využít nelineárních metod jako je použití mediánu nebo rotující masky.

5 Vysokofrekvenční filtry

Mezi typické aplikace vysokofrekvenční filtru patří zejména detekce hran.

Detekce hran je důležitá operace v biologickém i v počítačovém vidění. Podle hran člověk dokáže rozlišovat objekty. K základním předpokladům nalezení hrany v obraze je náhle se měnící hodnota jasu.

V počítačovém zpracování obrazu patří detekce hran k základní operaci, na kterou navazují další aplikace (např. rozpoznání objektů).

Hrana je určena tím, jak náhle se změní obrazová funkce $\mathrm{f}(x,y)$. Nástrojem na zachycení změn funkce dvou proměnných jsou parciální derivace – změnu udává její gradient, určující směr největšího růstu a strmost. Obrazové body s největším gradientem se nazývají hranami.

Jednou z metod nalezení hrany je využití tzv. operátorů. Těch je několik a liší se podle toho zda jsou závislé na rotaci či nikoliv. Představitel operátoru, který je nezávislý na rotaci a udává pouze sílu hrany, je **Laplaceův operátor**.

Ukázka masky Laplaceova operátoru

$$h = \begin{bmatrix} 1 & 1 & * \\ 1 & -6 & 1 \\ * & 1 & 1 \end{bmatrix} . \tag{10}$$

Při praktických testech se ukázalo, že u Laplaceova operátoru a hexagonální topologii funguje lépe větší maska, např.

$$h = \begin{bmatrix} 1 & 1 & 1 & * & * \\ 1 & 0 & 0 & 1 & * \\ 1 & 0 & -12 & 0 & 1 \\ * & 1 & 0 & 0 & 1 \\ * & * & 1 & 1 & 1 \end{bmatrix} . \tag{11}$$

Dalším z operátorů je **operátor Prewittové**, který je závislý na rotaci. Využívá více masek a jejich počet je závislý na počtu možných rotací. Pro masku 3×3 je těchto směrů šest, ukázka 12.

$$h_1 = \begin{bmatrix} 1 & 1 & * \\ 0 & 0 & 0 \\ * & -1 & -1 \end{bmatrix}, \quad h_2 = \begin{bmatrix} 0 & 1 & * \\ -1 & 0 & 1 \\ * & -1 & 0 \end{bmatrix}, \quad h_3 = \begin{bmatrix} -1 & 0 & * \\ -1 & 0 & 1 \\ * & 0 & 1 \end{bmatrix},$$

$$h_4 = \begin{bmatrix} -1 & -1 & * \\ 0 & 0 & 0 \\ * & 1 & 1 \end{bmatrix}, \quad h_5 = \begin{bmatrix} -1 & -1 & * \\ 0 & 0 & 0 \\ * & 1 & 1 \end{bmatrix}, \quad h_6 = \begin{bmatrix} 0 & -1 & * \\ 1 & 0 & -1 \\ * & 1 & 0 \end{bmatrix} \tag{12}$$

Po zpracování obrazu všemi maskami je vybrána například hodnota s největší hodnotou gradientu.

6 Morfologické operace

Morfologické operace patří mezi nelineární operátory, využívající masky, která se zde nazývá *strukturní element*.

Morfologické operace patří mezi nelineární operátory, využívající masky, která se zde nazývá *strukturní element*. Nejdříve než se dostanu k jednotlivým operacím, je potřeba zavést pojmy *vršek* a *stín* množiny a *Minkowského operace*.

6.1 Minkowského operace, vršek množiny a stín množiny

V [2] jsou pojmy vršek a stín definovány následovně. Vršek množiny A je funkce definovaná na $(n-1)$–rozměrném definičním oboru. Pro každou $(n-1)$-tici je vršek nejvyšší hodnota zbylé poslední souřadnice množiny A. Pro euklidovský prostor má nejvyšší hodnota význam suprema.

Nec $\mathbf{A} \subseteq \mathcal{E}^n$ a nechť definiční obor

$$\mathbf{F} = \{x \in \mathcal{E}^{n-1} \text{ pro některá } y \in \mathcal{E}, (x,y) \in \mathbf{A}\ \}\ .$$

Vršek množiny $\mathbf{A}$, označovaný T[$\mathbf{A}$], je zobrazením $\mathrm{F} \to \mathcal{E}$ definovaným jako

$$\mathrm{T}[\mathbf{A}](x) = \max\{y, (x,y) \in \mathbf{A}\ \}.$$

Stínem funkce f je množina sestávající se z vršku f a celého prostoru pod ním.

Nechť $\mathbf{F} \subseteq \mathcal{E}^{n-1}$ a $\mathrm{f} : \mathbf{F} \to \mathcal{E}$. **Stín** funkce f se označuje U[f], $U[\mathrm{f}] \subseteq \mathbf{F} \times \mathcal{E}$

$$\mathrm{U}[\mathrm{f}] = \{(x,y) \in \mathbf{F} \times \mathcal{E}, y \subseteq \mathrm{f}(x)\}.$$

Definice. *Nechť M je základní množina a $\mathbf{A}$, $\mathbf{B}$ jsou objekty na M. Minkowského součet a rozdíl definujeme pomocí základních operací*

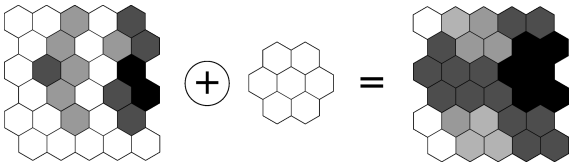
- *součet*

$$\mathbf{A}\ \oplus \mathbf{B} = \bigcup_{\beta \in \mathbf{B}} (\mathbf{A}\ + \beta)$$

- *rozdíl*

$$\mathbf{A}\ \ominus \mathbf{B} = \bigcap_{\beta \in \mathbf{B}} (\mathbf{A}\ + \beta),$$

kde $\mathbf{A} + \beta$ představuje posun množiny $\mathbf{A}$ ve směru vektoru β .



Obrázek 4: Dilatace

6.2 Dilatace

Jelikož šedotónová morfologie je zobecnění morfologie binární, před zavedením příslušné definice dilatace (později i eroze) pro šedotónový obraz uvedu nejprve definici této operace pro binární morfologii.

Definice. *Dilatace* D *je v binární morfologii definována pomocí Minkowského součtu jako*

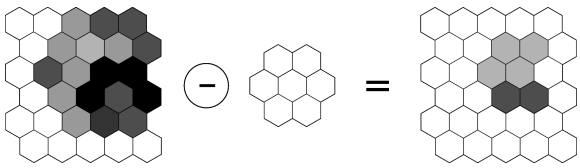
$$D(\mathbf{A},\mathbf{B}) = \mathbf{A} \oplus_B \mathbf{B} = \bigcup_{\beta \in \mathbf{B}} (\mathbf{A} \ + \beta).$$

Definice. *Nechť* $\mathbf{F}$,$\mathbf{K} \subseteq \mathcal{E}^{n-1}$, $f : \mathbf{F} \to \mathcal{E}, k : \mathbf{K} \subseteq \mathcal{E}$.
Šedotónová dilatace $\oplus$ *funkce* f *s funkcí* k*, definována jako*

$$f \oplus k = \mathrm{T}\{\mathrm{U}[f] \oplus_B \mathrm{U}[k]\} \ .$$

Tato definice není příliš vhodná pro algoritmizaci, proto se zavádí postup přes maximum součtů v množině

$$(\mathrm{f} \oplus \mathrm{k})(x) = \max\{\mathrm{f}(x-z) + \mathrm{k}(z), z \ \in \mathbf{K}, x-z \ \in \mathbf{F}\}.$$



Obrázek 5: Eroze

6.3 Eroze

Tak jako v předchozí sekci věnované dilataci je nutné nejdříve začít definicí eroze binárního obrazu.

Definice. *Eroze* E *je v binární morfologii definována pomocí Minkowského rozdílu jako*

$$E(\mathbf{A},\mathbf{B}) = \mathbf{A} \ominus_B \mathbf{B} = \bigcap_{\beta \in \mathbf{B}} (\mathbf{A} + \beta).$$

nyní je možno přejít k definici eroze v šedotónovém obrazu.

Definice. *Nechť* $\mathbf{F},\mathbf{K} \subseteq \mathcal{E}^{n-1}$, $f : \mathbf{F} \to \mathcal{E}$, $k : \mathbf{K} \subseteq \mathcal{E}$. *Šedotónová eroze* $\ominus$ *množiny* f *množinou* k, *definována jako*

$$f \ominus k = \mathrm{T}\{\mathrm{U}[f] \ominus_B \mathrm{U}[k]\}.$$

(*Poznámka:* $\ominus$ na pravé straně je erozí binárních obrazů)
Opět skutečný výpočet eroze v praxi probíhá jinak

$$(\mathrm{f} \ominus \mathrm{k})(x) = \min_{z \in \mathbf{K}} \{\mathrm{f}(x+z) - \mathrm{k}(z)\}.$$

Další z morfologických operací jsou operace tref či miň, otevření, uzavření či ztenčování. Všechny tyto operace využívají kombinace dilatace a eroze.

7 Využití hexagonální topologie k diagnostice Alzheimerovy choroby

Obsahem této kapitoly je zjistit, zda znalosti uvedené v předchozích kapitolách povedou k odhalení *Alzheimerovy demence* (AD).

8 Alzheimerova demence

Alzheimerova demence (někdy označována i jako Alzheimerova choroba) je neurodegenerativní onemocnění mozku, při kterém dochází k postupné demenci [4]. Nemoc se projevuje poruchou tzv. kognitivních funkcí - myšlení, paměti a úsudku.

Choroba je v současné době nevyléčitelná. V roce 2008 trpělov touto nemocí v České republice přibližně 120 tisíc lidí [4].

8.1 Předzpracovnání dat

Nejdříve je nutné snímky předzpracovat (tzv. preprocessing). Jelikož data jsem měl ve formě trojrozměrné matice $[m, n, h] = [79, 95, 69]$, bylo nejdříve nutné provést řez v určité hladině. Rozhodl jsem se řez provést na hladině $h = 30$, která subjektivně poskytuje největší množství informací.

Poté následovalo zbavení obrazu nežádoucího šumu pomocí nízkofrekvenčních filtrů, detekovat hrany, převést obraz do binární podoby a najít hranice.

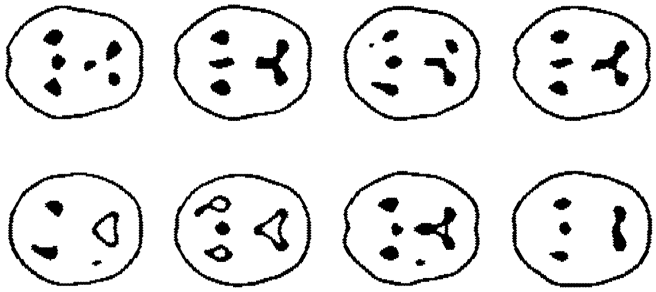
Až poté byly provedeny statitické operace a rozhodnutí zda uvedený postup vede k detekci AD.

8.2 Vytvoření etalonu zdravého pacienta

Při pozorování zpracovaných obrazů jsem si všiml, že obrazy zdravých pacientů jsou si velmi podobné, na rozdíl od nemocných pacientů, kde struktura snímků byla odlišná. Demonstrováno na obr. 6.

Tohoto poznatku jsem využil k vytvoření etalonu zdravého pacienta, který slouží k rozlišení zdravého pacienta od nemocného.

Etalony jsem vytvořil dva $\mathbf{E}_1$ a $\mathbf{E}_2$. Základem prvního byla Gaussova filtrace a druhého dilatace.



Obrázek 6: V horní řadě jsou zpracované obrazy zdravých pacientů. V dolní řadě pacienti trpící AD.

8.3 Samotná detekce AD

Předzpracované obrazy podle postupu v 8.1 jsou nyní porovnány s vytvořenými etalony. Kritériem pro rozpoznání snímků zdravých a nemocných pacientů je počet obrazových elementů pacienta $\mathbf{I}$ ležící mimo etalon $\mathbf{E}$. Tomu odpovídá veličina

$$c = \sharp(\mathbf{I} \setminus \mathbf{E}) \ .$$

Ze souboru nemocných byl sestaven vektor hodnot $\overline{c}_{AD} = (c_1, ..., c_{20})$. Analogicky byl ze souboru zdravých sestaven vektor $\overline{c}_{CN} = (c_1^{'}, ..., c_{26}^{'})$. K testování hypotézy H_0 o schodě průměrů byl použit t-test na hladině významnosti $p = 0,05$. Cílem bylo prokázat, že při vhodné volbě parametrů zpracování a etalonu je statisticky významný rozdíl mezi hodnotami c pro nemocné a zdravé pacienty.

V tabulkách 1 a 2 jsou výsledky testování s různými hodnotami parametrů R (poloměr masky při vyhlazování), p (práh při detekci hran) a p_{bin} (práh při tvorbě binárního obrazu).

Tabulka 1: Testování s $\mathbf{E}_1$

Test č.	1	2	3	4
R	5	3	3	5
p	0,38	0,38	0,48	0,48
p_{bin}	0,4	0,4	0,4	0,4
$p - hodnota$	0,0167	0,00035	0,0001	0,007

Ve výsledcích lze vidět, že přes různé nastavení parametrů se udržuje hladina významnosti pod hodnotou $0,05$.

Tabulka 2: Testování s $\mathbf{E}_2$

Test č.	1	2	3	4
R	5	3	3	5
p	0,38	0,38	0,48	0,48
p_{bin}	0,4	0,4	0,4	0,4
$p - hodnota$	0,0155	0,00034	0.0135	0,0002

9 Závěr

Představil jsem základní možnosti práce s hexagonálním obrazem a ukázku reálného použití při detekci Alzheimerovy choroby. I když výsledky ukazují dobré hodnoty, jsem opatrný v jejich prezentaci, protože testovacích dat (snímky pacientů) nebylo mnoho. Ale byl udělán první krok k potvrzení, že využití hexagonální topologie je dobrá cesta a má smysl se jí dále zabývat. V současné době již mám k dispozici větší a kvalitnější soubor dat a předpokládám, že pomocí nich budu schopen lépe demonstrovat výhody této topologie.

Literatura

[1] MIDDLETON, Lee, SIVASWAMY, Jayanthi. *Hexagonal Image Processing : A Practical Approach.* London : Springer-Verlag, 2005. 259 s. ISBN 1-85233-914-4.

[2] HLAVÁČ, Václav. *Zpracování signálů a obrazů.* Praha : Vydavatelství ČVUT, 2002. 220 s. Skriptum. ČVUT, Fakulta elektrotechnická. ISBN 80-01-02114-9.

[3] ŠEBEST, Miroslav. *Digitálna morfológia v hexagonálnej mriežke v Matlabe.* Praha, 2006. 68 s. ČVUT, FJFI, Katedra softwarového inženýrství v ekonomii. Vedoucí dimplové práce doc. Ing. Jaromír Kukal, Ph.D.

[4] Wikipedie: Otevřená encyklopedie: *Alzheimerova choroba* [online]. c2010 [citováno 3. 5. 2010]. Dostupný z WWW: http://cs.wikipedia.org/w/index.php?title=Alzheimerova_choroba&oldid=5275102

Spectral Analysis of Predictive Error in Alzheimer's Disease Diagnostics

Olga Orlova

1st year of PGS, email: orlovolg@fjfi.cvut.cz
Department of Software Engineering in Economy
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Jaromír Kukal, Department of Software Engineering in Economy,
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. This paper describes the new method which is based on non-linear one-step predictor, which is designed as MLP neural network. It is a kind of low-pass non-linear filter. The difference between raw EEG and the ANN output is then a subject of band spectral analysis. The differences in this power spectrum between Alzheimer diseased and control patient group are statistically significant.

Keywords: EEG, Alzheimer's disease, ANN, model error, spectral analysis

Abstrakt. Tento článek popisuje novou metodu, která je založená na nelineárním jednokrokovém prediktoru, navrženém jako neuronová síť MLP. Je to druh nízko-úrovňového nelineárního filtru. Rozdíl mezi hrubým EEG a výstupem neuronové sítě je pak předmětem pásmové frekvenční analýzy. Rozdíly ve výkonovém spektru mezi skupiny zdraých pacientů a pacientů trpících Alzheimerovou demenci jsou statisticky významné.

Klíčová slova: EEG, Alzheimerova demence, neuronové sítě, chyba predikce, spektrální analýza

1 Introduction

Alzheimer's disease (AD) is the most common dementia. This disease affects approximately 7 % of people older than 65 years and 40 % of people older than 80 years [7]. Dementia is characterized by memory decline and others neurophysiological changes that occur in the elderly and the risk of disease increases exponentially with age.

EEG signals reflect the bioelectrical activity of the brain. Electroencephalographic records are one of the tools for diagnosis of neurological diseases. Traditional analysis relies mainly on detection of spectral changes: performs the analysis of selected frequency bands, then calculate the corresponding spectral powers, whose changes may indicate dysfunction of the nervous system.

Analysis of the power spectrum of a healthy active brain suggests [2] that there are four main frequency bands: δ $(0.5-4\,\mathrm{Hz})$, θ $(5-8\,\mathrm{Hz})$, α $(9-12\,\mathrm{Hz})$, β $(13-20\,\mathrm{Hz})$. In the frequency domain, there are established following differences in the EEG records of healthy patients and patients with Alzheimer's dementia: an increase in theta and delta rhythms, decline in beta rhythm and slowing of alpha rhythm.

Many works [3], [4], [5] also show the possibility of using artifical intelligence to solve the problem of identification of Alzheimer's disease. Among them those employing neural network to address this problem. Dominating amount of the works has many common

features: perform artefacts cleaning of EEG record, perform the analysis of time series decomposition to frequency bands, perform further processing of these bands. After all these operations neural network is used for classification purposes.

The possibility of classification of healthy people and people with Alzheimer's disease, in this study, is explored using the following assumptions:
- don't use EEG signals adjusted with artefacts filtering;
- use neural network to detect patters in the EEG signal;
- prediction error is used for classification purposes, this prediction error was obtained from the use of neural network, trained on human health and subsequently used for a person with Alzheimer's disease;
- Don't use an artifical neural network for classification purpose.

2 Alzheimer's disease diagnosis via EEG

Electroencefalography is a continuous multi-channel recording of electrical potential difference. This recording was measured by electrodes placed on the scalp in some way. EEG was first introduced and described by Berger [6] in connection with the study of sleep. Data which are collected from a typical EEG experiment are a sequence of time points sampled at $128 - 1024\,\mathrm{Hz}$ in general.

EEG recordings were obtained from 16 healthy people and 16 people with Alzheimer's disease. All patients sat in a chair in a darkened room, they were at a quiet state and had their eyes closed. Measurements were performed using 21 active electrodes placed on the surface of the head in line with the international 10/20 system. Sampling frequency was $200\,\mathrm{Hz}$.

3 ANN as intelligent filter of EEG

3.1 Signal description

One of the main points of this work is to construct a model that would allow to assess the general patterns of EEG recording in healthy people and would create preconditions for the decision rules required to classify EEG recordings of healthy people and people with Alzheimer's disease.

Formulation of the problem is largely procedural in nature, taking into account only one factor, the one that EEG record reflects changes in brain bioelectrical activity, and among these changes are those that carry for us important informations.

In the most general case, the behavior of EEG signal can be described as a superposition of a function z which describes important informations for us, and some random component e. The estimation $\hat{z}$ of the signal function z will be implemented by using neural network, in this paper. Random component $\hat{e}$, which arises as a result of this assessment, will serve as the calculation of the noise component e.

3.2 Neural network

Let $n \in \mathbb{N}$ be number of inputs, $N \in \mathbb{N}$ be number of outputs and $H \in \mathbb{N}$ be number of neurons in the hidden layer. Let $\mathrm{x} \in \mathbb{R}^n$ be input vector, $\mathrm{y} \in \mathbb{R}^N$ be output vector and $\mathrm{h} \in \mathbb{R}^H$ be signal vector in the hidden layer. The three layer ANN − multi-layer perceptron (MLP) operates according to equations

$$\mathrm{h} = \mathsf{f}(\mathbb{W}\mathrm{x} + \mathrm{w}_0) \tag{1}$$

$$\hat{\mathrm{z}} = \mathbb{V}\mathrm{h} + \mathrm{v}_0 \tag{2}$$

where
$\mathbb{W} \in \mathbb{R}^{H\times n}$, $\mathbb{V} \in \mathbb{R}^{N\times H}$ are weight matrices, $\mathrm{w}_0 \in \mathbb{R}^H$, $\mathrm{v}_0 \in \mathbb{R}^N$ are biases and f is a non-polynomial function. After the decompositions:

$$\mathbb{W} = \begin{pmatrix} \mathrm{w}_1 \\ \vdots \\ \mathrm{w}_H \end{pmatrix}, \ \mathbb{V} = \begin{pmatrix} \mathrm{v}_1 \\ \vdots \\ \mathrm{v}_N \end{pmatrix}$$

We can establish the vector of ANN parameters

$$\mathrm{p} = (\mathrm{w}_0^T, \mathrm{w}_1^T, \ldots, \mathrm{w}_H^T, \mathrm{v}_0^T, \mathrm{v}_1^T, \ldots, \mathrm{v}_N^T)$$

consisting of $M = (n+1)H + (H+1)N$ real coordinates. The resulting MLP as ANN can be formally rewritten as

$$\mathrm{y} = \mathsf{ANN}(\mathrm{x}, \mathrm{p}) \tag{3}$$

3.3 Learning strategy

Let $m \in \mathbb{N}$ be number of patterns for ANN learning. The pattern set can be represented via matrices

$$\mathbb{X} = \begin{pmatrix} \mathrm{x}_1 \\ \vdots \\ \mathrm{x}_m \end{pmatrix} \in \mathbb{R}^{m\times n}, \quad \mathbb{Y} = \begin{pmatrix} \mathrm{y}_1 \\ \vdots \\ \mathrm{y}_m \end{pmatrix} \in \mathbb{R}^{m\times N}$$

The method of least squares was used to ANN learning. Adequate objective function

$$\mathsf{F}(\mathrm{p}) = \sum_{k=1}^{m} \|\mathrm{y}_k - \mathsf{ANN}(\mathrm{x}_k, \mathrm{p})\|^2 \tag{4}$$

is thus subject of minimization.

Due to multi-modality of $\mathsf{F}(\mathrm{p})$ we applied Fast Simulated Annealing (FSA) method. The algorithm of FSA produces a parameter sequence $\mathrm{p}_0, \mathrm{p}_1, \ldots, \mathrm{p}_k \in \mathbb{R}^M$ beginning with initial vector p_0 according to the rule:

$$\begin{aligned} \mathrm{p}_{k+1} &= \mathrm{p}_k^* \quad \text{when} \quad \mathsf{F}(\mathrm{p}_k^*) < \mathsf{F}(\mathrm{p}_k) + T_k \tan\rho_k \\ \mathrm{p}_{k+1} &= \mathrm{p}_k \quad \text{when} \quad \mathsf{F}(\mathrm{p}_k^*) \geq \mathsf{F}(\mathrm{p}_k) + T_k \tan\rho_k \end{aligned} \tag{5}$$

where $\mathrm{p}_k^* = \mathrm{p}_k + gT_k \cdot \tan \mathrm{r}_k$, $\rho_k, r_{k,j} \sim U(-\frac{\pi}{2}, +\frac{\pi}{2})$ are independent uniformly distributed random variables, $g > 0$ is scaling factor and $T_k > 0$ is dimensionless temperature. The strategy of FSA cooling is

$$T_k = \frac{T_0}{1 + \lfloor \frac{k}{r} \rfloor} \tag{6}$$

where $r \in \mathbb{N}$ is a repeating period.

4 Spectral analysis of EEG signal

Power spectrum describes the energy distribution of the frequencies of the dynamic system. The dynamic development of simple systems can usually be described by a certain frequency range. An opposite situation is typical for complex systems: cannot be selected any particular frequency band. Frequency components were processed in each of the four frequency bands using the following relationship.

$$r_{band} = \frac{\sum_{band} |\mathrm{fft}|^2}{\sum |\mathrm{fft}|^2} \cdot 100\,\% \tag{7}$$

where fft is a result of application of the fast Fourier transform to analyse EEG signal, *band* is one of the four main frequency bands: δ, θ, α or β. Due to the large non-linearity of EEG signal, Fourier transformation was applied not only to the data itself but also on the resulting prediction error obtained by using neural networks. It is natural to expect that the non-linearity error is smaller than the non-linearity of the original signal.

5 Results

To eliminate noise in the EEG data, EEG signals of all patients were analysed in the range of indices from 20000 to 50000. Identify patterns in the EEG recording was performed by one-step prediction using MLP neural network. The used MLP network consisted of one hidden layer with four hidden neurons. Hyperbolic tangent function was used as an activation function of MLP network. As a standard healthy person has been chosen one patient (pivot) whose EEG signal had the average statistical characteristics regarding the set of healthy people. The neural network was trained on the EEG signal of this patient, generalization abilities of used neural network were tested on the EEG signals of the remaining healthy patients. Subsequently, the neural network was applied to the EEG signals of patients with Alzheimer's disease. All electrodes were used for the prediction of EEG signals using MLP network in the corresponding EEG signals of pivot patient. Training all neural network were performed by FSA algorithm, containing 300 interior and 300 exterior cycles. Classification of healthy patients and patients with Alzheimer's disease was based on a review of one-step prediction error signal EEG using MLP neural network learning on EEG signal of a healthy pivot patient, and then used for prediction EEG signals others patients. In accordance with the above definitions, we have: $n = 1$, $H = 4$, $N = 1$, $T_0 = 0.001$, $g = 1$, $r = 300$, $k_{max} = 300$.

Results of band spectral analysis of individual channels for raw EEG data are collected in the Tab. 1. Individual p-values are results of two-sample two-sided t-test of hypothesis

H_0 that relative power (for given frequency band and channel) is the same for AD and CN group of patients. Adequate ROC diagram is depicted on the Fig. 1 for four bands and first channel. Tab. 2 shows results of the t-test of significance for each band, in the power spectrum of EEG signal, originating from the first channel, at a significance level of 5 %.

Table 1: P-values for significant differences (AD × CN) in the case of raw EEG (t-test)

channel	δ	θ	α	β
1	$9,4 \times 10^{-2}$	$1,6 \times 10^{-2}$	$1,1 \times 10^{-1}$	$2,6 \times 10^{-2}$
2	$3,2 \times 10^{-1}$	$1,9 \times 10^{-2}$	$2,4 \times 10^{-2}$	$1,1 \times 10^{-1}$
3	$3,6 \times 10^{-1}$	$3,3 \times 10^{-2}$	$\mathbf{1,6 \times 10^{-4}}$	$1,4 \times 10^{-1}$
4	$1,4 \times 10^{-1}$	$\mathbf{8,1 \times 10^{-3}}$	$\mathbf{2,4 \times 10^{-3}}$	$3,8 \times 10^{-1}$
5	$2,1 \times 10^{-1}$	$\mathbf{6,8 \times 10^{-3}}$	$1,7 \times 10^{-2}$	$8,0 \times 10^{-2}$
6	$3,7 \times 10^{-1}$	$\mathbf{5,2 \times 10^{-3}}$	$\mathbf{7,0 \times 10^{-3}}$	$1,3 \times 10^{-1}$
7	$8,8 \times 10^{-1}$	$8,1 \times 10^{-2}$	$\mathbf{3,9 \times 10^{-3}}$	$2,7 \times 10^{-1}$
8	$5,2 \times 10^{-1}$	$9,7 \times 10^{-2}$	$\mathbf{3,0 \times 10^{-5}}$	$6,2 \times 10^{-1}$
9	$7,8 \times 10^{-1}$	$3,2 \times 10^{-1}$	$\mathbf{3,9 \times 10^{-4}}$	$2,9 \times 10^{-2}$
10	$9,8 \times 10^{-2}$	$\mathbf{3,2 \times 10^{-3}}$	$\mathbf{3,5 \times 10^{-4}}$	$7,2 \times 10^{-2}$
11	$3,7 \times 10^{-1}$	$9,2 \times 10^{-2}$	$\mathbf{9,0 \times 10^{-4}}$	$1,6 \times 10^{-1}$
12	$5,8 \times 10^{-1}$	$6,9 \times 10^{-2}$	$\mathbf{9,8 \times 10^{-4}}$	$8,0 \times 10^{-1}$
13	$6,5 \times 10^{-1}$	$5,4 \times 10^{-1}$	$\mathbf{2,0 \times 10^{-4}}$	$6,9 \times 10^{-2}$
14	$8,4 \times 10^{-1}$	$2,8 \times 10^{-1}$	$\mathbf{8,0 \times 10^{-5}}$	$1,2 \times 10^{-2}$
15	$3,0 \times 10^{-1}$	$8,3 \times 10^{-1}$	$\mathbf{7,6 \times 10^{-4}}$	$\mathbf{7,2 \times 10^{-4}}$
16	$2,5 \times 10^{-1}$	$7,2 \times 10^{-2}$	$\mathbf{3,0 \times 10^{-5}}$	$1,1 \times 10^{-1}$
17	$5,5 \times 10^{-1}$	$3,9 \times 10^{-2}$	$\mathbf{2,0 \times 10^{-3}}$	$5,0 \times 10^{-2}$
18	$7,7 \times 10^{-1}$	$7,0 \times 10^{-1}$	$\mathbf{9,0 \times 10^{-3}}$	$\mathbf{1,3 \times 10^{-3}}$
19	$7,8 \times 10^{-1}$	$3,0 \times 10^{-1}$	$\mathbf{9,9 \times 10^{-4}}$	$3,1 \times 10^{-2}$

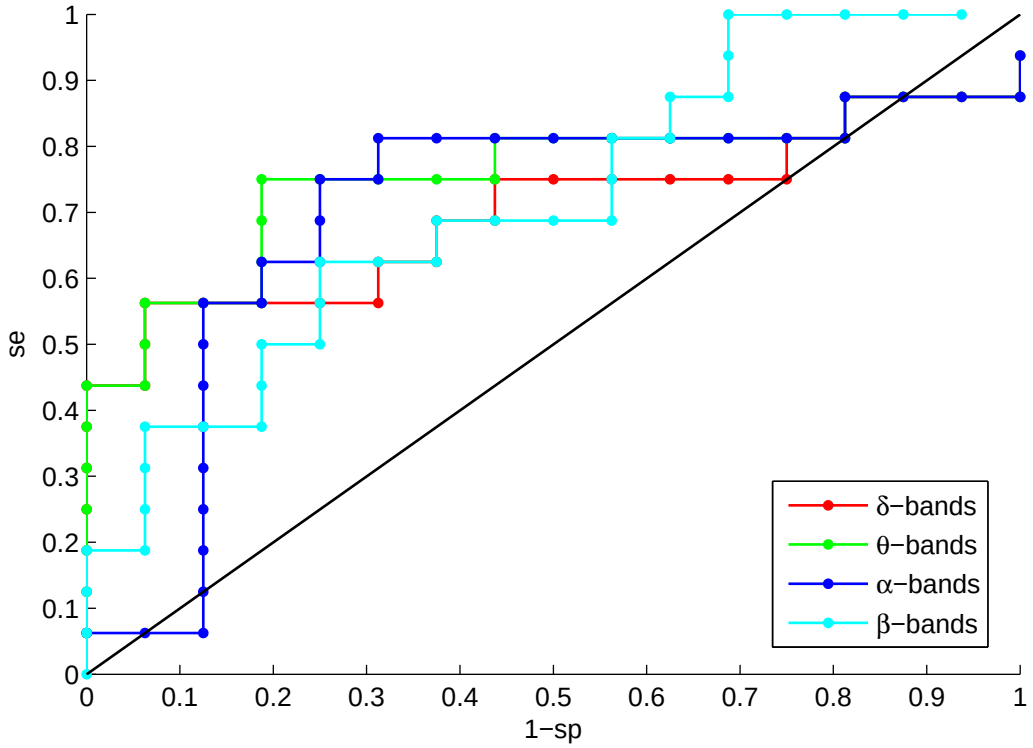


Figure 1: ROC for the 1st channel and δ, θ, α, β bands in the case of raw EEG data

Table 2: The significance of differences in the power spectrum of EEG signal (t-test)

	h	p	t	df	s
$\delta-$band	0	$0,0944$	$-1,7274$	30	$7,4846$
$\theta-$band	1	$0,0164$	$-2,5439$	30	$2,8343$
$\alpha-$band	0	$0,1074$	$-1,6596$	30	$1,5903$
$\beta-$band	1	$0,0261$	$2,3399$	30	$2,7321$

The differences between raw EEG and the output of ANN is called here as prediction error. The best are the results of band spectral analysis of the prediction error. Relative power of prediction in given channel and band was also subject of statistical testing. Results are involved in the Tab. 3. Adequate ROC diagram is depicted on the Fig. 2 for the four bands and first channel. Table 4 shows results of the t-test of significance for each band, in the power spectrum of prediction error, originating from the first channel.

Table 3: P-values for significant differences (AD × CN) in the case of model error (t-test)

channel	δ	θ	α	β
1	$2,4 \times 10^{-2}$	$\mathbf{6,4 \times 10^{-3}}$	$\mathbf{7,2 \times 10^{-3}}$	$\mathbf{1,7 \times 10^{-3}}$
2	$3,1 \times 10^{-2}$	$2,3 \times 10^{-2}$	$2,2 \times 10^{-2}$	$\mathbf{1,7 \times 10^{-3}}$
3	$6,7 \times 10^{-1}$	$4,9 \times 10^{-1}$	$\mathbf{2,8 \times 10^{-3}}$	$1,7 \times 10^{-2}$
4	$4,0 \times 10^{-1}$	$7,7 \times 10^{-2}$	$\mathbf{4,8 \times 10^{-3}}$	$2,1 \times 10^{-2}$
5	$2,1 \times 10^{-1}$	$5,5 \times 10^{-2}$	$1,8 \times 10^{-2}$	$\mathbf{5,0 \times 10^{-3}}$
6	$6,3 \times 10^{-1}$	$1,9 \times 10^{-1}$	$\mathbf{7,7 \times 10^{-3}}$	$\mathbf{6,6 \times 10^{-3}}$
7	$7,2 \times 10^{-1}$	$4,4 \times 10^{-1}$	$\mathbf{3,6 \times 10^{-3}}$	$2,0 \times 10^{-2}$
8	$8,1 \times 10^{-1}$	$9,0 \times 10^{-1}$	$\mathbf{5,1 \times 10^{-4}}$	$1,9 \times 10^{-1}$
9	$5,6 \times 10^{-1}$	$3,0 \times 10^{-1}$	$\mathbf{1,2 \times 10^{-3}}$	$2,0 \times 10^{-2}$
10	$2,1 \times 10^{-1}$	$5,2 \times 10^{-2}$	$\mathbf{1,7 \times 10^{-3}}$	$\mathbf{3,6 \times 10^{-3}}$
11	$1,6 \times 10^{-1}$	$7,6 \times 10^{-2}$	$\mathbf{7,1 \times 10^{-4}}$	$1,6 \times 10^{-2}$
12	$9,0 \times 10^{-1}$	$4,0 \times 10^{-1}$	$2,1 \times 10^{-2}$	$5,8 \times 10^{-2}$
13	$2,7 \times 10^{-1}$	$3,7 \times 10^{-1}$	$\mathbf{2,8 \times 10^{-3}}$	$\mathbf{2,0 \times 10^{-3}}$
14	$7,6 \times 10^{-1}$	$6,1 \times 10^{-1}$	$\mathbf{4,1 \times 10^{-4}}$	$\mathbf{1,0 \times 10^{-3}}$
15	$9,9 \times 10^{-1}$	$3,9 \times 10^{-1}$	$\mathbf{9,4 \times 10^{-4}}$	$\mathbf{6,6 \times 10^{-4}}$
16	$4,3 \times 10^{-1}$	$5,6 \times 10^{-1}$	$\mathbf{4,5 \times 10^{-4}}$	$\mathbf{1,4 \times 10^{-3}}$
17	$6,1 \times 10^{-1}$	$3,9 \times 10^{-1}$	$\mathbf{1,3 \times 10^{-3}}$	$\mathbf{1,2 \times 10^{-3}}$
18	$9,8 \times 10^{-1}$	$9,1 \times 10^{-1}$	$\mathbf{9,5 \times 10^{-3}}$	$\mathbf{1,6 \times 10^{-4}}$
19	$3,7 \times 10^{-1}$	$4,4 \times 10^{-1}$	$\mathbf{5,0 \times 10^{-3}}$	$\mathbf{3,0 \times 10^{-4}}$

Table 4: The significance of differences in the power spectrum of the prediction error (t-test)

	h	p	t	df	s
$\delta-$band	1	0.0238	−2.3815	30	1.5916
$\theta-$band	1	0,0064	−2.9289	30	1.5619
$\alpha-$band	1	0.0072	−2.8850	30	2.0161
$\beta-$band	1	0.0017	3.4542	30	3.9116

6 Conclusions

Band-power spectrum of raw EEG is efficient tools for the classification of Alzheimer diseased patients against control normal patients. Bound-power spectral analysis of prediction error come to statistically significant results. Namely β-band relative power in the case channels $1, 2, 5, 6, 10, 13-19$ has p-value < 0.001 in the case of two-sample two-sided t-test. The relative power of ANN prediction error is significantly lower in the case of Alzheimer's disease. It corresponds with hypothesis of diseased β-activity in the right frontal domain of the human brain in the case of given dementia. Classification purposes, as β-band well as α-band, enable more stable results in the case of channels $13-19$.

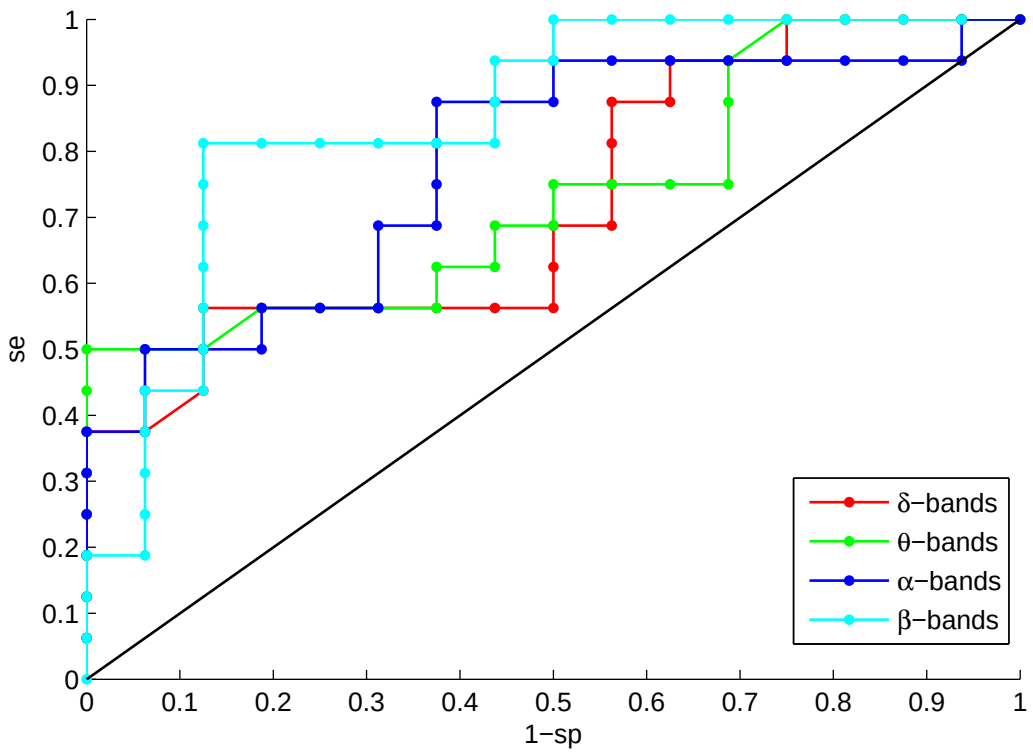


Figure 2: ROC for the 1st channel and δ, θ, α, β bands in the case of model error

References

[1] C. Bishop. *Neural Networks for Pattern Recognition*, ISBN 0-19-853849-9, Oxford Press

[2] K. Revett. *On the use of rough sets for artefact extraction from EEG datasets*, IEEE 0-7695-2999-2/07

[3] C. Castellaro, et. al.. *An artifical intelligence approach to classify and analyse EEG trace*, Neurophysiol Clin 2002; 32:193-214, Elsevier, S0987705302003027/FLA

[4] A. C. Tsoi, et. al.. *Classification of Electroencephalogram using Artifical Neural Networks*, University of Queensland St. Lucia, Queensland 4072, Australia 1993

[5] A. A. Petrosian, et. al.. *Recurrent neural network-based approach for early recognition of Alzheimer's disease in EEG*, S1388-2457(01)00579-X, Elsevier 2001

[6] J. H. Siegel. *The Neural Control of Sleep and Waking*, ISBN 0387954929, Springer

[7] C. Lehmann, et. al.. *Application and comparison of classification algorithms for recognition of Alzheimer's disease in electrical brain activity (EEG)*, Journal of Neuroscience Methods 161 (2007) 342-350, Elsevier 0165-0270

MDA and Agile – Choose or Combine?

Marek Rosa

1st year of PGS, email: rosamare@fjfi.cvut.cz
Department of Software Engineering in Economy
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Vojtěch Merunka, Department of Software Engineering in Economy,
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. This paper discusses whether MDA and Agile, two software design and development approaches, are exclusive or can be combined. After a brief indtroduction of both of them, the paper defines a set of observed features and traces the presence in one or the other or both. We see a great potential in the combination of these two approaches with a lot of supplemental features in regard to each other. However, model based evolution will still have to wait before the majority of developers together with tools vendors come along to this approach as a new standard.

Keywords: MDA, Model Driven Architecture, Agile, Software Development Methodology

Abstrakt. Tento článek se snaží zhodnoti Agilní a MDA přístup k návrhu a vývoji software z pohledu jejich možného zkombinovaní. Po stručném představení obou metodik následuje vlba požadovaných vlastností k porovnání a zmapování jejich přítomnsti v té či oné. V kombinaci obou přístupů vidíme velký potenciál, jak se mohou vzájemně doplnit a přispět ke zvýšení výsledné efektivity. Rozvoj modelovacího přístupu bude muset nicméně ještě počkat než majorita vývojářů spolu s dodavateli vývojových nástrojů přijmou tento způsob za nový standard.

Klíčová slova: MDA, Modelem řízená architektura, Agilní vývoj, Metodika vývoje software

1 Introduction

During the last 9 years, there has been a lot of buzz around two catch-phrases in the software development world — Model Driven Architecture[11] and Agile[4]. Both of these describe a different way how to approach software development, focusing, among others, on minimizing the failure rate of information systems. Over these years, many debates took place on whether the modelling (the foundation of MDA) should drive the whole development process or should be used rather informally only for temporary artifacts (the way Agile treats it).

Model Driven Architecture (MDA), or the modelling with UML itself, proposes an approach to software engineering with most of the work being done at a higher level of abstraction, working with platform-independent models and striving for reusability.

Agile, on the other hand, draws attention by involving customer and application experts in the overall development life-cycle. Its main focus is primarily on the code and the distinction between design and implementation tasks is put aside. Extreme programming [7], for example, encourages developers to select the most simple solution as opposed to designing for reusability.

The following two sections briefly re-introduce both approaches, highlighting the key attributes of each one. In section 4, individual attributes are evaluated for comparison and intercompatibility and presented in the form of a table.

2 Model Driven Development/Architecture

Model Driven Architecture was introduced as a set of guidelines for software specifications, design, and development, defined by the Object Management Group[3]. These guidelines describe the development life-cycle which is not all that different from the traditional approaches. The difference is in the artifacts being created during the process. In case of MDA, the main artifacts are models. Specifically, a set of models which comprises three core models on three different levels of abstraction and distance from the target platform. Fully automatic both-way transformations are defined between these three models[15].

The first of the top-level models is a Platform Independent Model (PIM), independent of the technology of implementation and target platform. This model is mostly used in the analysis phase, but when following an iterative development methodology and thanks to the automated transformations, it can also be refined continuously.

PIM transforms into a Platform Specific Model (PSM) which consists of sub-models interdependent with the specific platform (e.g., EJB), see Figure 1.

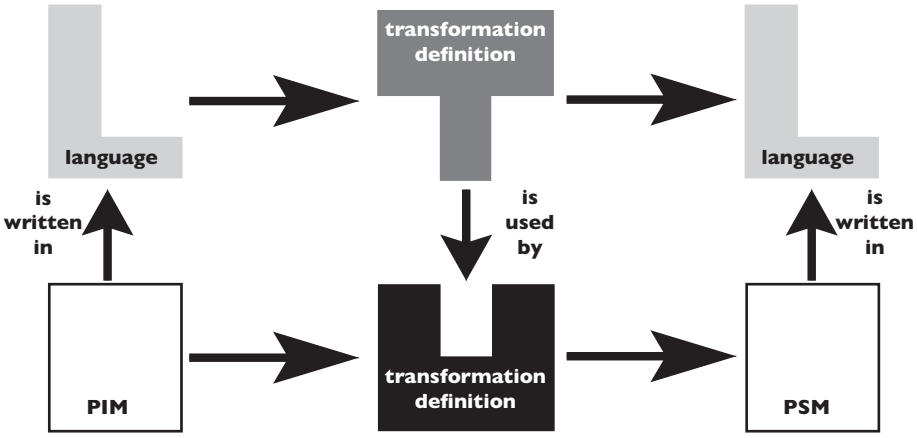


Figure 1: The MDA Framework [15]

And finally, the third model is the final, fully generated and executable code.

As has been already said, MDA raises the levels of abstraction and reuse. This is achieved by combining the phase of design and specification with the phase of manual implementation following the outputs of design phase. History shows nice examples of similarly significant paradigm shifts.

In the 1980s, the software development field went through a shift from manually producing assembly code to programming in high level programming languages and having the compiler producing assembly code. No doubt that the level of abstraction and platform independence has increased. As for the reuse, it was a long way through the decades of functions, procedural paradigm, object oriented paradigm, components and frameworks until the very present domain models [9].

Another advantage is the interoperability [11]. Interoperability represents MDA feature of making multiple PSMs generated from one PIM able to communicate with each other. This is achieved by the concept of bridges — automatically generated connectors between individual PSMs. This feature can save a lot of time during development.

It is also worth noting that the transformations used are often also a subject of the development because the generic transformation is often not enough. At least not yet. This can lead to either prolonging the total delivery time or the need of utilizing additional resources.

3 Agile Approach

Agile movement started around 2001. It was a reaction to a much-greater-than-necessary number of experience of slipping schedules, growing budgets, unsatisfied customers, ineffective practices, etc. Fear of failing again and again led to more and more constraining processes demanding a growing number of documents and reports [7].

By this motivation, the Agile Alliance was formed and it outlined the basic values and principles. Agile methods, formerly known as light-weight processes, promote disciplined but flexible project management process, run in short iterations and continuously refining the estimates. Using a big-scale method like PERT which does not fit smaller projects (i.e., most of the software projects, because PERT does not scale well to the man-day level), or a method like CPM which does not correlate its concurrency in tasks with the real-world developers' capabilities, is suppressed in favor of relative estimates based on the nearest experience of the last iteration.

Estimates expressed in virtual points represent amount or portion of tasks/features that will be implemented in the iteration, first one is more or less guessed [6]. Based on these points is calculated the effectivity, measured as a number of points delivered in the last iteration. It is called velocity.

The approach draws high attention to the customer. He is involved as much as possible in the whole process, he is allowed and expected to prioritize the features during planning sessions, even change the requirements and collaborate by all means.

Developers need to adopt a set of engineering best practices focused on quality, delivery time, and maintainability. These include: test-driven development, pair programming, continuous integration, etc. Working code is at the core of all and any developer can work at any part of it at any time — a practice called collective ownership. And most importantly, at the end of every iteration there has to be a working program.

Changes to requirements are harnessed with confidence backed up with automated test suite guarding regressions. Acceptance tests written in a scripting language verify the expected behaviour.

Agile suppresses production of comprehensive documents unless really and immediately needed. The design is evolved gradually rather than prescribed by big design up front [13] — incremental development.

4 Putting It Together

This section will try to give an overview of the selected features and properties of both approaches, thus mapping the presence in one or the other or both, every attribute being completed with our statement. Some of the features are heavily dependent on the maturity and availability of the tools to support them; when this is the case for at least one of the subjects, the attribute is prefixed with **T:**.

Attribute	Agile	MDA	Statement / Comment
Reuse	By default, does not explicitly promote reuse. Starts with the simplest thing that could possibly work and the design goes through the evolution before the final state.	Support reuse by having main focus on the more abstract and general level - the model. Some of the MDA methodologies even explicitly prescribe techniques to create potentially reusable artifacts [1].	MDA's higher level of abstraction, visual artifacts and model execution speaks for the advantage of MDA in focus on reusability.
Risk Management	Agile methods are designed to mitigate risks like changing requirements and schedules [14].	Existing methodologies do not provide coverage for the activity [1].	It would seem that Agile is better prepared for risk management; however, the general techniques of risk management can be applied and used with both of the approaches.

Continued on next page

Attribute	Agile	MDA	Statement / Comment
Adoption Prospects	For agile development there are quite mature tools supporting Agile practices and existing methodologies are also proved by time. The troubles with adoption are mainly in the shift of involving customers, learning new practices and habits, and bringing management and developers on the same page.	Broad adoption is waiting for sound tools or integration with major tools being used [12]. It also requires developers and analysts to extend their portfolio of software engineering techniques with modelling skills [13]. One of key impediments to adoption is the a priori assumption that model driven code cannot possibly work [10]. Building a deliverable system early in Agile style allays the fears and brings team sceptics on board.	Agile being definitely more prevalent has a better adoption starting point. However, it is very likely that MDA will begin to attract more and more attention both of the developer and tool vendors.
Tools support	The two most popular IDEs - Eclipse and Visual Studio - and most of the tools improvements are directed towards programming activities and Agile practices support [12].	Neither of the two most popular IDEs - Eclipse and Visual Studio - has yet paid enough attention to attract wider audience to the MDA approach, the first being arguably further.	As the most development tools focus on the code, it is Agile practices which has a better support at this time, MDA tools are either not mature enough or specific for a certain area (Netsilon [5], BridgePoint for the Shlaer-Mellor method), or too expensive.

Continued on next page

Attribute	Agile	MDA	Statement / Comment
T: Coping with legacy systems	It is usually difficult to apply Agile practices on the legacy code which was not designed in that manner. Especially lack of test suite increases the risk of defects from refactoring and changing the original behaviour.	MDA with a tool with full round-trip engineering support can easily reconstruct the model from legacy code and start rebuilding from there.	Legacy systems are a very sound argument for MDA approach as they are able (with appropriate tool) to import the old system and modify it on the model level.
T: Prototyping	Directly aims at continuously involving customer and application experts via frequent prototypes and the "test first" paradigm.	With proper tool with support of an executable version of UML and PSMs, the model can be used for rapid prototyping.	Prototyping is an important activity in nowadays business software development and it is achievable with both approaches; however, MDA might still suffer from immature tools.
T: Test Covering	Given the test-first paradigm, the project started with agile approach usually exhibit a good test coverage.	Same as previous attribute, the executable UML would provide support for the activity [8]. Another approach proposes usage of visual contracts to define test cases [2].	Basically the same statement as the one above applies here with the notion of that model testing requires different techniques.

Table 1: Features and Properties of MDA and Agile

5 Conclusions

Both MDA and Agile claim an increased productivity as one of the benefits they can provide. For MDA, this comes out from high potential of the fact that the main artifact is the model which can be easier to understand for customers and stakeholders than the code, test execution outputs, or bare business logic prototypes. In Agile context, productivity is achieved by short iterations with working software at the end of each, automated test suite created with the test-first approach, working closely with the customer to review the results and decisions immediately.

In that, there might be an opportunity for even further productivity increase — having customer on site but working on the model level, while the model is the executable

test suite (or it can be at least automatically generated). Customer's presence and tasks prioritization (together with Agile promotion of simple design and solutions) might also help preventing modeling a too-large system (horizontal scope creep) or too general (premature generalization - vertical scope creep) [10].

There is a great potential in the combination of these two approaches with a lot of supplemental features in regard to each other. However, model based evolution will still have to wait before it is very well supported by the major tools being used [13].

References

[1] Mohsen Asadi and Raman Ramsin. Mda-based methodologies: An analytical survey. In Ina Schieferdecker and Alan Hartman, editors, *Model Driven Architecture – Foundations and Applications*, volume 5095 of *Lecture Notes in Computer Science*, pages 419–431. Springer Berlin / Heidelberg, 2010.

[2] Gregor Engels, Baris Güldali, and Marc Lohmann. Towards model-driven unit testing. In Thomas Kühne, editor, *Models in Software Engineering*, volume 4364 of *Lecture Notes in Computer Science*, pages 182–192. Springer Berlin / Heidelberg, 2007.

[3] Object Management Group. Model driven architecture (mda). OMG Document ormsc/2001- 07-01, 2001.

[4] Jim Highsmith and Martin Fowler. The agile manifesto. *Software Development Magazine*, 9(8):29–30, 2001.

[5] William Kaim, Philippe Studer, and Pierre-Alain Muller. Model driven architecture for agile web information system engineering. In *Object-Oriented Information Systems*, volume 2817 of *Lecture Notes in Computer Science*, pages 299–303. Springer Berlin / Heidelberg, 2003.

[6] Robert C. Martin. Pert, cmp and agile project management, October 2003.

[7] Robert C. Martin and Micah Martin. *Agile Principles, Patterns, and Practices in C# (Robert C. Martin)*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 2006.

[8] Stephen J. Mellor. Agile mda. Technical report, 2005.

[9] Stephen J. Mellor, Scott Kendall, Axel Uhl, and Dirk Weise. *MDA Distilled*. Addison Wesley Longman Publishing Co., Inc., Redwood City, CA, USA, 2004.

[10] Stephen J. Mellor and Leon Starr. Six lessons learned using mda. In Nuno Nunes, Bran Selic, Alberto Rodrigues da Silva, and Ambrosio Toval Alvarez, editors, *UML Modeling Languages and Applications*, volume 3297 of *Lecture Notes in Computer Science*, pages 198–202. Springer Berlin / Heidelberg, 2005.

[11] OMG, `http://www.omg.org/mda/`.

[12] Óscar Pastor. From extreme programming to extreme non-programming: Is it the right time for model transformation technologies? In Stephane Bressan, Josef Küng, and Roland Wagner, editors, *Database and Expert Systems Applications*, volume 4080 of *Lecture Notes in Computer Science*, pages 64–72. Springer Berlin / Heidelberg, 2006.

[13] Bernhard Rumpe. Agile modeling with the uml. In Martin Wirsing, Alexander Knapp, and Simonetta Balsamo, editors, *Radical Innovations of Software and Systems Engineering in the Future*, volume 2941 of *Lecture Notes in Computer Science*, pages 59–65. Springer Berlin / Heidelberg, 2004.

[14] Richard Turner and Apurva Jain. Agile meets cmmi: Culture clash or common cause? In Don Wells and Laurie Williams, editors, *Extreme Programming and Agile Methods — XP/Agile Universe 2002*, volume 2418 of *Lecture Notes in Computer Science*, pages 153–165. Springer Berlin / Heidelberg, 2002.

[15] Jos Warmer and Anneke Kleppe. *The Object Constraint Language: Getting Your Models Ready for MDA*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2003.

Irregular $\mathcal{PT}$-symmetric Point Interactions*

Petr Siegl

3rd year of PGS, email: siegl@ujf.cas.cz
Department of Physics, Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague
Nuclear Physics Institute ASCR
Laboratoire Astroparticule et Cosmologie, Université Paris 7

advisor: Miloslav Znojil, Nuclear Physics Institute ASCR
Jean-Pierre Gazeau, Laboratoire Astroparticule et Cosmologie, Université Paris 7

Abstract. We study certain classes of $\mathcal{PT}$-symmetric extensions of symmetric second derivative operators. The operators are similar to self-adjoint ones except particular irregular cases. Spectral properties of the latter are very far from those of self-adjoint extensions: operators with empty resolvent set and empty spectrum are present.

Keywords: $\mathcal{PT}$-symmetry, point interactions, irregular boundary conditions

Abstrakt. Studujeme třídy $\mathcal{PT}$-symmetrických rozšíření symetrických diferenciálních operátorů druhého řádu. Rozšíření jsou podobná samosdruženým operátorům s výjimkou speciálních případů, jejichž spektrální vlastnosti jsou velmi odlišné od samosdružených rozšíření: existují rozšíření s prázdnou resolventní množinou a s prázdným spektrem.

Klíčová slova: $\mathcal{PT}$-symetrie, bodové interakce, iregulární okrajové podmínky

1 Introduction

$\mathcal{PT}$-symmetric operators, a special case of operators with antilinear symmetry, have been intensively studied in both physical and mathematical context as a result of the observation that the spectrum of such operators may be real and discrete [5]. Although it is known that some $\mathcal{PT}$-symmetric operators are special case of quasi-Hermitian ones [7], or equivalently, they can be mapped by similarity transformation to the self-adjoint ones, see *e.g.* [3, 11, 16] for examples, the spectrum of $\mathcal{PT}$-symmetric operators may be also complex, *e.g.* complex conjugated eigenvalues may appear. The complex conjugated pairs of eigenvalues instead of the real ones are actually the simplest possible deviation of the spectrum from the self-adjoint case. In fact, the class of operators with antilinear symmetry is much larger. The residual spectrum of operators (even bounded) with antilinear symmetry may be non-empty and the point spectrum of such operator may be uncountable [17], *i.e.* operators may be non-spectral [8].

We consider $\mathcal{PT}$-symmetric point interactions on a line described in general in [1]. It has been established that the spectrum of $\mathcal{PT}$-symmetric point interaction on a line

*This work has been supported by the Grant Agency of the Czech Technical University in Prague, grant No. SGS OHK4-010/10.

can include up to two real or complex conjugated eigenvalues in addition to the continuous part [1]. However, it has been noticed in a more recent work [3] that an irregular case having uncountable point spectrum is present among previously studied extensions. Starting from this observation we proceed with an analysis of the analogous models on a finite interval showing that the spectrum of the corresponding operators can be empty or entire complex plane depending on boundary conditions imposed at the endpoints. All these examples show that $\mathcal{PT}$-symmetry together with pseudo-Hermiticity and J-self-adjointness may be a very weak requirement allowing not only complexification of some (or all) discrete eigenvalues.

In the physical framework of $\mathcal{PT}$-symmetric Quantum Mechanics [4, 15], the fact that the point interactions can completely and dramatically change the spectrum was not expected. Nonetheless, considering operators being not even similar to normal ones brings expected unusual spectral effects. We remark that these examples illustrate the necessity of the non-empty residual set assumption in [10, III, Corollary 6.34], claiming that the extension of a finite order has a compact resolvent if and only if some other extension of the same operator has a compact resolvent.

We recall a definition of $\mathcal{PT}$-symmetric point interactions in the first section and we also formulate slightly more precisely the claim of [1] concerning the $\mathcal{PT}$-self-adjointness of the operators. In the next section, we consider a particular $\mathcal{PT}$-symmetric point interaction for the model defined on a line, we summarize results on the spectrum and indicate a connection to the collapse of quasi-Hermiticity in the irregular case. Models defined on the finite interval $(-l, l)$ are studied in the last section. The dependence of the spectrum on the boundary conditions at $\pm l$ is described in details.

The interesting spectral effects caused by certain $\mathcal{PT}$-symmetric point interactions can be expected when considering general classification of boundary conditions, *cf.* [8], since the studied extensions correspond to irregular boundary conditions. The recent work [14] dealing with pseudo-Hermitian extensions with empty resolvent set shows that their presence is essential for existence of an additional fundamental symmetry that can be used for explicit construction of $\mathcal{C}$ operators.

Operators $\mathcal{P}$ and $\mathcal{T}$ are defined in $L^2(\mathbb{R})$ space in the following way, the parity $\mathcal{P}$ acts as $(\mathcal{P}\psi)(x) = \psi(-x)$ and the time reversal symmetry $\mathcal{T}$ is the complex conjugation $(\mathcal{T}\psi)(x) = \overline{\psi}(x)$. We say that an operator A is $\mathcal{PT}$-symmetric if $(\mathcal{PT})A \subset A(\mathcal{PT})$.

2 $\mathcal{PT}$-symmetric point interactions

A family of $\mathcal{PT}$-symmetric point interaction at the origin was determined in [1] by the two types of boundary conditions: connected and separated. Differential operator L, corresponding to the point interaction,

$$L = -\frac{\mathrm{d}^2}{\mathrm{d}x^2} \tag{1}$$

is defined on the domain $\mathrm{Dom}\,(L)$ consisting of functions ψ from $W^{2,2}(\mathbb{R}\setminus\{0\})$ satisfying boundary conditions described by parameters $b, c, \psi, \theta, h_0, h_1$ in the following way

i) connected case:

$$\begin{pmatrix}\psi(0+)\\ \psi'(0+)\end{pmatrix} = B\begin{pmatrix}\psi(0-)\\ \psi'(0-)\end{pmatrix}, \tag{2}$$

with the matrix B equal to

$$B = e^{\mathrm{i}\theta}\begin{pmatrix}\sqrt{1+bc}\,e^{\mathrm{i}\phi} & b\\ c & \sqrt{1+bc}\,e^{-\mathrm{i}\phi}\end{pmatrix}, \tag{3}$$

with real parameters $b \geq 0$, $c \geq -1/b$, $\theta, \phi \in (-\pi, \pi]$.

ii) separated case:

$$\begin{aligned} h_0\psi'(0+) &= h_1 e^{\mathrm{i}\theta}\psi(0+),\\ h_0\psi'(0-) &= -h_1 e^{-\mathrm{i}\theta}\psi(0-), \end{aligned} \tag{4}$$

with the real phase parameter $\theta \in [0, 2\pi)$ and with the parameter $\mathbf{h} = (h_0, h_1)$ taken from the real projective space $\mathbf{P}^1$.

The operator L is an extension of a symmetric densely defined operator $L_0 = -\mathrm{d}^2/\mathrm{d}x^2$ with the domain $\mathrm{Dom}\,(L_0) = C_0^\infty(\mathbb{R}\setminus\{0\})$. L can be also viewed as a restriction of $L_{max} = L_0^* = -\mathrm{d}^2/\mathrm{d}x^2$ with the domain $\mathrm{Dom}\,(L_{max}) = W^{2,2}(\mathbb{R}\setminus\{0\})$.

The separated $\mathcal{PT}$-symmetric boundary conditions have been studied in several works [11, 6, 13, 12] and we will not consider this case further.

We would like to remark that the claim of [1] that all operators L satisfy the property $L^* = \mathcal{P}L\mathcal{P}$ is not entirely accurate for the connected case. If we express explicitly the boundary conditions corresponding to the adjoint operator L^*, we conclude that $L^* = \mathcal{P}L\mathcal{P}$ holds if and only if $\theta = 0$. Nevertheless, none of the other claims of [1] is affected by this fact because of the unitary equivalence of the operators corresponding to the different choices of θ. Further, we will consider $\theta = 0$ only.

We summarize symmetry properties of L. The proof of the following proposition is straightforward application of boundary conditions for L, L^* and actions of operators $\mathcal{P}$ and $\mathcal{T}$.

Proposition 1. *Let L be the second derivative operator corresponding to the connected $\mathcal{PT}$-symmetric point interaction at the origin* (1)–(3) *with the choice $\theta = 0$ in the boundary conditions. Then*
i) $L^* = \mathcal{P}L\mathcal{P}$,
ii) $(\mathcal{PT})L \subset L(\mathcal{PT})$,
iii) $L^* = \mathcal{T}L\mathcal{T}$.

The first symmetry is referred to as the $\mathcal{P}$-pseudo-Hermiticity or $\mathcal{PT}$-self-adjointness, the second one is the $\mathcal{PT}$-symmetry in its original sense and the third one is the $\mathcal{T}$-self-adjointness, the special case of J-self-adjointness, where J is an antilinear isometric involution, *i.e.* $J^2 = I$ and $\langle Jx, Jy\rangle = \langle y, x\rangle$ for all $x, y \in \mathcal{H}$. The importance of $\mathcal{T}$-self-adjointness for $\mathcal{PT}$-symmetric models was stressed in [6], one of the reasons is that the residual spectrum of J-self-adjoint operators is empty [9, Lem. III.5.4].

We remark that the property *i)* of the Proposition 1 guarantees that the operator L is closed. To this end take into the consideration closedness of every adjoint operator, the relation *i)*, and $\mathcal{P} = \mathcal{P}^{-1} \in \mathscr{B}(\mathcal{H})$. The closedness of the considered extensions can be alternatively shown with help of [10, III, Problem 5.11] as well.

3 Model on a line

The spectrum of this model has been investigated firstly in [1, Thm.2, Prop.1], it basically consists of the branch of continuous spectrum $[0,\infty)$ and up to two real or complex conjugated eigenvalues. However, it was observed in [3] that there is one "exceptional" case for particular choice of parameters for which the resolvent set of corresponding operator is empty. Our aim is to investigate this particular operator into more details.

Let us study the connected case with $\theta = b = c = 0$, *i.e.* the boundary conditions for L_ϕ read

$$\psi(0+) = e^{\mathrm{i}\phi}\psi(0-), \quad \psi'(0+) = e^{-\mathrm{i}\phi}\psi'(0-), \tag{5}$$

where $\phi \in (-\pi,\pi]$. The adjoint operator L_ϕ^* can be found explicitly, $L_\phi^* = L_{-\phi}$ and the case $\phi = \pi$ corresponds to the self-adjoint operator.

Spectral properties of L_ϕ for $\phi \neq \pm\frac{\pi}{2}$ are very simple, the spectrum is continuous without any eigenvalues,

$$\sigma(L_\phi) = \sigma_c(L_\phi) = [0,\infty), \quad \phi \neq \pm\frac{\pi}{2}. \tag{6}$$

It is possible to find an invertible positive bounded operator Θ with bounded inverse satisfying

$$L_\phi^*\Theta_\phi = \Theta_\phi L_\phi, \quad \phi \neq \pm\frac{\pi}{2}, \tag{7}$$

in other words, to show that L_ϕ is quasi-Hermitian [7] or, equivalently, that L_ϕ is similar to a self-adjoint operator. The explicit formula for the operator Θ and its square root was obtained by different approaches in [3, 16, 2],

$$\Theta_\phi = I - \mathrm{i}\sin\phi P_{\mathrm{sign}}\mathcal{P}, \tag{8}$$

where the operator P_{sign} acts as a multiplication by the function $\operatorname{sign} x$. The spectrum of Θ_ϕ consists of two eigenvalues $1 \pm \sin\phi$ of infinite multiplicities,

$$\sigma(\Theta_\phi) = \sigma_p(\Theta_\phi) = \{1 \pm \sin\phi\}. \tag{9}$$

We denote $L_\pm, \Theta_\pm$ the operators corresponding to $\phi = \pm\frac{\pi}{2}$. The relation (7) is still valid for $\phi = \pm\frac{\pi}{2}$, however, operators $\Theta_\pm$ are no longer invertible. Moreover, we can see that formula for the resolvent [1, eq.(17)] collapses because the expression [1, eq.(18)] appearing in the denominator is identically zero. These facts are reflected in unusual spectral properties of $L_\pm$ being far from those of self-adjoint operators.

Proposition 2. *Spectra of the operators $L_\pm$ include all complex numbers, interval $[0,\infty)$ is the continuous part and every $\lambda \in \mathbb{C}\setminus[0,\infty)$ belongs to the point spectrum.*

$$\sigma_p(L_\pm) = \mathbb{C}\setminus[0,\infty), \quad \sigma_c(L_\pm) = [0,\infty). \tag{10}$$

Proof. The eigenfunctions corresponding to eigenvalues from $\mathbb{C}\setminus[0,\infty)$ can be found explicitly, see [3] for the details. □

4 Models on a finite interval

We consider a finite interval $(-l, l)$ and a second derivative operator L_ϕ corresponding to the $\mathcal{PT}$-symmetric interaction at origin of the type (5). The domain of L_ϕ consists of functions ψ belonging to the Sobolev space $W^{2,2}((-l,0)\cup(0,l))$ and satisfying boundary conditions (5) at origin and some other boundary conditions at $\pm l$ being specified later. Our aim is to study the spectrum of such differential operators, particularly if $\phi = \pm\pi/2$ where the choice of the boundary conditions at $\pm l$ plays an essential role.

We distinguish two classes of boundary conditions being imposed at $\pm l$: symmetric and $\mathcal{PT}$-symmetric ones. The symmetric boundary conditions are determined by a unitary matrix U entering well known relation

$$(U - I)\Psi(l) + \mathrm{i}\,(U + I)\Psi'(l) = 0, \tag{11}$$

where

$$\Psi(l) = \begin{pmatrix} \psi(l) \\ \psi(-l) \end{pmatrix}, \qquad \Psi'(l) = \begin{pmatrix} \psi'(l) \\ -\psi'(-l) \end{pmatrix}. \tag{12}$$

The $\mathcal{PT}$-symmetric boundary conditions are defined by relations (2)-(4).

We summarize spectral properties of L_ϕ in following propositions. As we may expect, the cases $\phi = \pm\pi/2$ exhibit unusual features.

Proposition 3. *Let L_ϕ be the second derivative operator in $L^2((-l,l))$ corresponding to the $\mathcal{PT}$-symmetric point interaction (5) at origin with symmetric boundary conditions* (11)–(12) *at $\pm l$.*

If $\phi \neq \pm\pi/2$, then the spectrum of L_ϕ is discrete and its eigenvalues $\lambda = k^2$ are solutions of the equation

$$\begin{aligned} \cos\phi\Big(P_1(U) - 2\mathrm{i}kP_2(U)\cos 2kl + k^2P_3(U)\sin 2kl\Big) + \\ +2\mathrm{i}k\Big(u_{12} + u_{21} + \mathrm{i}(u_{11} - u_{22})\sin\phi\Big) = 0, \end{aligned} \tag{13}$$

where u_{ij} are elements of the unitary matrix U and

$$\begin{aligned} P_1(U) &= 1 - u_{11} - u_{12}u_{21} - u_{22} + u_{11}u_{22}, \\ P_2(U) &= 1 + u_{12}u_{21} - u_{11}u_{22}, \\ P_3(U) &= 1 + u_{11} - u_{12}u_{21} + u_{22} + u_{11}u_{22}. \end{aligned} \tag{14}$$

If $\phi = \pm\pi/2$, then the point spectrum of $L_\pm$ is either empty or entire $\mathbb{C}$. The latter case occurs if and only if

$$u_{12} + u_{21} \pm \mathrm{i}(u_{11} - u_{22}) = 0. \tag{15}$$

If we take into consideration usual Dirichlet ($U = -I$) and Neumann ($U = I$) boundary conditions at $\pm l$, then the condition (15) is fulfilled, thus the spectrum of $L_\pm$ is the entire complex plane.

Next, we apply both connected and separated $\mathcal{PT}$-symmetric boundary conditions at $\pm l$. It may be expected for connected case that the second point interaction (parameters are denoted by the subscript 2) of the type $b_2 = 0, c_2 = 0, \phi_2 = \pm\pi/2$ produces analogous interesting spectral effects.

Proposition 4. *Let L_ϕ be the second derivative operator in $L^2((-l,l))$ corresponding to the $\mathcal{PT}$-symmetric point interaction (5) at origin with connected $\mathcal{PT}$-symmetric boundary conditions (2) at $\pm l$.*

If $\phi \neq \pm\pi/2, \phi_2 \neq \pm\pi/2$ or $\phi \neq \pm\pi/2, \phi_2 = \pm\pi/2$ and $b_2 \neq 0$ or $c_2 \neq 0$, then the spectrum of L_ϕ is discrete and its eigenvalues $\lambda = k^2$ are solutions of the equation

$$\cos\phi\left(\left(b_2k^2 - c_2\right)\sin 2kl + 2k\sqrt{1+b_2c_2}\cos\phi_2\cos 2kl\right) + \\ +2k\left(\sqrt{1+b_2c_2}\sin\phi\sin\phi_2 - 1\right) = 0. \tag{16}$$

If $\phi = \pm\pi/2$, then the point spectrum of $L_\pm$ is either empty or entire $\mathbb{C}$. The latter case occurs if and only if

$$\sqrt{1+b_2c_2}\sin\phi_2 - 1 = 0. \tag{17}$$

If $b_2 = 0, c_2 = 0, \phi_2 = \pm\pi/2$, then the point spectrum of $L_\pm$ is either empty or entire $\mathbb{C}$. The latter case occurs if and only if $\phi = \pm\pi/2$.

Proposition 5. *Let L_ϕ be second derivative operator in $L^2((-l,l))$ corresponding to the $\mathcal{PT}$-symmetric point interaction (5) at origin with separated $\mathcal{PT}$-symmetric boundary conditions (4) at $\pm l$.*

If $\phi \neq \pm\pi/2$ and $\theta \neq 0, \pi$, then the spectrum of L_ϕ is discrete and its eigenvalues $\lambda = k^2$ are solutions of equation

$$\cos\phi\left(2h_0h_1k\cos 2kl\cos\theta + \left(h_0^2k^2 - h_1^2\right)\sin 2kl\right) - 2h_0h_1k\sin\theta\sin\phi = 0. \tag{18}$$

If $\phi = \pm\pi/2$, then the point spectrum of $L_\pm$ is either empty or entire $\mathbb{C}$. The latter case occurs if and only if $\theta = 0, \pi$.

Remark 1. *The case of empty point spectrum actually means that the whole spectrum is empty because the resolvent is compact in this case.*

Proof. We solve the differential equation $L_\phi\psi = \lambda\psi$ together with both boundary conditions. We search for a non-zero eigenfunction and this yields the secular equations (13), (16), (18). If we insert $\phi = \pm\pi/2$ or other assumptions on the rest of the parameters into the equations, we obtain the assertions concerning the empty and entire $\mathbb{C}$ point spectrum.

In order to prove the claim of the non-empty discrete spectrum and of the remark above we show that the resolvent is compact in these cases. We calculate the resolvent explicitly for the operator L_+ in Proposition 5. The remaining resolvents can be obtain by analogous procedure. At first, using standard Green function approach, we calculate the resolvent corresponding to the $L_1 = -\mathrm{d}^2/\mathrm{d}x^2$ on $(-l,l)$ with separated $\mathcal{PT}$-symmetric conditions (4) at $\pm l$.

$$\left(R_{L_1}(\lambda)g\right)(x) = \int_{-l}^{l} G(x,y)g(y)\mathrm{d}y, \tag{19}$$

where $g \in L_2(\mathbb{R})$, $\lambda = k^2$, and

$$G(x,y) = \frac{1}{W(k)}\begin{cases} u_-(x)u_+(y), & x \leq y \\ u_-(y)u_+(x), & x \geq y, \end{cases} \tag{20}$$

$$
\begin{aligned}
W(k) &= -k\left(\left(k^2-\frac{h_1^2}{h_0^2}\right)\sin 2kl + 2k\frac{h_1^2}{h_0^2}\cos\theta\cos 2kl\right),\\
u_\pm(x) &= A_\pm\cos kx + B_\pm\sin kx,\\
A_- &= -k\cos kl + e^{-\mathrm{i}\theta}\frac{h_1}{h_0}\sin kl, \quad B_- = e^{-\mathrm{i}\theta}\frac{h_1}{h_0}\cos kl + k\sin kl,\\
A_+ &= k\cos kl - e^{\mathrm{i}\theta}\frac{h_1}{h_0}\sin kl, \qquad B_+ = e^{\mathrm{i}\theta}\frac{h_1}{h_0}\cos kl + k\sin kl.
\end{aligned}
\tag{21}
$$

We may easily check that functions $u_\pm$ satisfy appropriate boundary condition (4) at $\pm l$.

We define operators L_{min} and L_{max} both acting as $-\mathrm{d}^2/\mathrm{d}x^2$, the domain of L_{min} consists of $\psi\in W^{2,2}((-l,l))$ satisfying $\psi(0)=\psi'(0)=0$ and the separated $\mathcal{PT}$-symmetric boundary conditions (4) at $\pm l$, while the domain of L_{max} are $\psi\in W^{2,2}((-l,0)\cup(0,l))$ satisfying the separated $\mathcal{PT}$-symmetric boundary conditions (4) at $\pm l$. Both L_1 and L_ϕ are extensions of L_{min} and restrictions of L_{max} and therefore the resolvent of L_ϕ can be written in the form

$$
\left(R_{L_\phi}(k^2)g\right)(x) = \left(R_{L_1}(k^2)g\right)(x) + C_-(k)e_-(x) + C_+(k)e_+(x), \tag{22}
$$

with

$$
e_\pm(x) = \vartheta(\pm x)u_\pm(x), \tag{23}
$$

where $\vartheta(x)$ is the Heaviside step function, and $C_\pm(k)$ are to be determine. We require $R_{L_\phi}(k^2)g\in\operatorname{Dom}L_\phi$, thus it must satisfy boundary conditions (5). This leads to the system of linear equations for $C_\pm(k)$

$$
\begin{pmatrix} e^{\mathrm{i}\phi}A_+ & -A_- \\ -e^{-\mathrm{i}\phi}kB_+ & kB_- \end{pmatrix}\begin{pmatrix} C_-(k) \\ C_+(k)\end{pmatrix} = \begin{pmatrix} \left(e^{\mathrm{i}\phi}-1\right)F_1(0) \\ \left(e^{-\mathrm{i}\phi}-1\right)F_1'(0)\end{pmatrix}, \tag{24}
$$

where

$$
F_1(x) \;=\; \left(R_{L_1}(k^2)g\right)(x), \quad F_1'(x) = \frac{\mathrm{d}}{\mathrm{d}x}F_1(x). \tag{25}
$$

The solution exists if determinant of the matrix on the l.h.s. of (25) denoted M further is non-zero. On the other hand the condition $\det M=0$ yields eigenvalue equation (18).

Solutions $C_\pm(k)$ have the following form:

$$
\begin{aligned}
C_-(k) &= \frac{e^{\mathrm{i}\phi}-1}{\det M}\left(B_-F(0) - \frac{e^{-\mathrm{i}\phi}}{k}A_-F'(0)\right),\\
C_+(k) &= \frac{e^{\mathrm{i}\phi}-1}{\det M}\left(e^{-\mathrm{i}\phi}B_+F(0) - \frac{1}{k}A_+F'(0)\right).
\end{aligned}
\tag{26}
$$

If we consider k for which $\det M\neq 0$ and $W(k)\neq 0$, then $C_\pm(k)$ are bounded and estimates

$$
|C_\pm(k)| \le C(k)\|g\| \tag{27}
$$

are valid for a constant $C(k)$ depending on k. $R_{L_1}(k^2)$ is a compact operator and if we add rank one, *i.e.* also compact, operators $C_\pm(k)e_\pm$ we get $R_{L_\phi}(k^2)$ which is then also

compact for the fixed k. Whence, by the resolvent identity, $R_{L_\phi}(k^2)$ is compact for all $k^2 \in \varrho(L_\phi)$.

This claim remains true also for $\phi = \pi/2$ and $\theta \neq 0, \pi$ because

$$\det M = -2k^2 \frac{h_1}{h_0} \sin\theta. \tag{28}$$

We can alternatively finish the proof by using [10, III, Corollary 6.34]. In order to prove that resolvent is compact it suffices to show that the resolvent set is non-empty, *i.e.* to find some k for which $R_{L_\phi}(k^2) \in \mathscr{B}(\mathcal{H})$. □

References

[1] S. Albeverio, S. M. Fei, and P. Kurasov. *Point Interactions: PT-Hermiticity and Reality of the Spectrum.* Letters in Mathematical Physics **59** (2002), 227–242.

[2] S. Albeverio, U. Gunther, and S. Kuzhel. *J-self-adjoint operators with $\mathcal{C}$-symmetries: an extension theory approach.* Journal of Physics A: Mathematical and Theoretical **42** (2009), 105205 (22pp).

[3] S. Albeverio and S. Kuzhel. *One-dimensional Schrödinger operators with $\mathcal{P}$-symmetric zero-range potentials.* Journal of Physics A: Mathematical and General **38** (2005), 4975–4988.

[4] C. M. Bender. *Making sense of non-Hermitian Hamiltonians.* Reports on Progress in Physics **70** (2007), 947–1018.

[5] C. M. Bender and S. Boettcher. *Real spectra in non-hermitian hamiltonians having PT symmetry.* Physical Review Letters **80** (Jun 1998), 5243–5246.

[6] D. Borisov and D. Krejčiřík. *PT-symmetric waveguides.* Integral Equations Operator Theory **62** (2008), 489–515.

[7] J. Dieudonné. *Quasi-Hermitian operators.* Proceedings Of The International Symposium on Linear Spaces (July 1961), 115–123.

[8] N. Dunford and J. T. Schwartz. *Linear Operators, Part 3, Spectral Operators.* Wiley-Interscience, (1971).

[9] D. E. Edmunds and W. D. Evans. *Spectral Theory and Differential Operators (Oxford Mathematical Monographs).* Oxford University Press, USA, (1987).

[10] T. Kato. *Perturbation theory for linear operators.* Springer-Verlag, (1966).

[11] D. Krejčiřík, H. Bíla, and M. Znojil. *Closed formula for the metric in the Hilbert space of a $\mathcal{PT}$-symmetric model.* Journal of Physics A: Mathematical and General **39** (2006), 10143–10153.

[12] D. Krejčiřík and P. Siegl. $\mathcal{PT}$-symmetric Models in Curved Manifolds. To appear in Journal of Physics A: Mathematical and Theoretical,(2010).

[13] D. Krejčiřík and M. Tater. *Non-Hermitian spectral effects in a $\mathcal{PT}$-symmetric waveguide.* Journal of Physics A: Mathematical and Theoretical **41** (2008), 244013 (14pp).

[14] S. Kuzhel and C. Trunk. On a class of J-self-adjoint operators with empty resolvent set. arXiv:1009.0873, (2010).

[15] A. Mostafazadeh. *Pseudo-Hermitian Representation of Quantum Mechanics.* International Journal of Geometric Methods in Modern Physics (to appear).

[16] P. Siegl. *Supersymmetric quasi-Hermitian Hamiltonians with point interactions on a loop.* Journal of Physics A: Mathematical and Theoretical **41** (2008), 244025 (11pp).

[17] P. Siegl. *The non-equivalence of pseudo-Hermiticity and presence of antilinear symmetry.* Pramana - Journal of Physics **73** (2009), 279–287.

Example of an Infinite Word with Specific Properties*

Štěpán Starosta

2nd year of PGS, email: sstarosta@seznam.cz
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Edita Pelantová, Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. We will exhibit an example of an infinite word, whose language is closed under reversal, and the word itself is rich in palindromes and contains infinitely many non-palindromic bispecial factors. We will prove the mentioned properties. We will also mention the motivation for such an example in the context of generalizations of Sturmian words.

Keywords: palindrome, rich word, bispecial factor

Abstrakt. Uvedeme příklad nekonečného slova, které má jazyk uzavřený na reverzi, je plně saturováno palindromy a obsahuje nekonečně mnoho nepalindromických bispeciálních faktorů. Zmíněné vlastnosti dokážeme a zmíníme i motivaci pro hledání takového slova v kontextu zobecnění Sturmovských slov.

Klíčová slova: palindrom, slovo bohaté na palindromy, bispeciální faktor

1 Introduction

Combinatorics on words deals mainly with infinite words - infinite sequences of letters. It is a relatively new domain, it dates to the beginning of 20^{th} century. Since then, its growth is accelerating until today. It is intimately connected with other mathematical domains. One of the closest connection is to symbolic dynamics where the state of the system is represented by an infinite word. A lot of combinatorial properties of an infinite word have their dynamical equivalent in a symbolic dynamical system (see for instance [4] or [5]).

We will deal with some specific combinatorial properties of infinite words. The situation is usually simpler on a binary alphabet. Some binary infinite words are quite well explored and their properties have been generalized to larger alphabet. In [2] a well-known class of binary words, Sturmian words, served as an inspiration for exploring the generalized properties and their relations. Sturmian words are an interesting object as they can be defined in many ways and they appear in very different situations in the world around us. Some of their combinatorial characterizations rely on the notion of

*This work has been supported by the Czech Science Foundation grant no. 201/09/0584, by the grant no. MSM6840770039 and LC06002 of the Ministry of Education, Youth, and Sports of the Czech Republic and by the grant no. SGS10/085OHK4/1T/14 of the Grant Agency of the Czech Technical University in Prague

palindrome - a word read the same from the left as from the right. One of the properties of Sturmian words is that they are fully saturated by palindromes, i.e., a Sturmian word cannot contain more palindromes. Some of their generalizations also fulfill that property (which is referred to as richness or fullness).

In this report we will give one example of an infinite word used in [2] in the context of generalizations of Sturmian words. We will require this example to be also saturated by palindromes and contain non-palindromic factors (subwords that occur without gaps in the infinite word) - see later for precise definition. We will give proof of its properties.

Section 2 gives some necessary notions and definition from combinatorics on words while Section 3 contains the example and proofs.

2 Preliminaries

An **alphabet** $\mathcal{A}$ is a finite set of symbols called **letters**. A **finite word** is a finite sequence of letters. By a **language** we mean a set of finite words. The set of all finite words over the alphabet $\mathcal{A}$ is denoted by $\mathcal{A}^*$ and includes the empty word ε. When equipped with the operation of concatenation, $\mathcal{A}^*$ is a monoid.

An **infinite word** is an infinite sequence of letters. For an infinite word $\mathbf{u} = (u_i)_{i=0}^{+\infty}$, where $u_i \in \mathcal{A}$ for all i, we say that a finite word w is a **factor** of $\mathbf{u}$ if there exists an integer $k \geq 0$ such that $w = u_k u_{k+1} \dots u_{k+n-1}$. The integer k is said to be an **occurrence** of w in $\mathbf{u}$. The integer n is the length of the word w, denoted $|w|$.

An infinite word is **recurrent** if every factor occurs infinitely many times, i.e., has infinitely many occurrences. An infinite word is **uniformly recurrent** if the gaps between consecutive occurrences of every factor are bounded.

The set of all factors of an infinite word $\mathbf{u}$, including the empty word ε, is denoted $\mathcal{L}(\mathbf{u})$. This set is said to be the **language of u**. We say that a factor $v \in \mathcal{L}(\mathbf{u})$ is a **right extension** of a factor $w \in \mathcal{L}(\mathbf{u})$ if there exists a letter $x \in \mathcal{A}$ such that $v = wx$. If a factor w has more than one right extension, we say it is **right special** (RS). Note that in a language of an infinite word, every factor has at least one right extension. The definition of **left extension** and **left special** (LS) factor is analogous. If an infinite word is recurrent, then also every factor has at least one left extension. If a factor is right and left special, we say it is **bispecial** BS.

Factor complexity $\mathcal{C}(n)$ is a mapping associated to an infinite word $\mathbf{u}$ which to an integer n associated the number of distinct factor of length n, i.e.,

$$\mathcal{C}(n) = \#\left\{w \in \mathcal{L}(\mathbf{u}) \mid |w| = n\right\}.$$

The **bilateral order** of a factor $w \in \mathcal{L}(\mathbf{u})$ was introduced in [3] as the number $\mathrm{b(w)} = \#\{\mathrm{awb} \mid \mathrm{awb} \in \mathcal{L}(\mathbf{u}), \mathrm{a}, \mathrm{b} \in \mathcal{A}\} - \#\{\mathrm{aw} \mid \mathrm{aw} \in \mathcal{L}(\mathbf{u}), \mathrm{a} \in \mathcal{A}\} - \#\{\mathrm{wb} \mid \mathrm{wb} \in \mathcal{L}(\mathbf{u}), \mathrm{b} \in \mathcal{A}\} + 1$. Factors can be classified according to their bilateral order. If $\mathrm{b(w)} = 0$, we say the factor is **ordinary**. If $\mathrm{b(w)} > 0$, it is said to be **strong**. Otherwise the factor is **weak**. It can be readily seen from the definitions that if w is not BS, then it is ordinary. The importance of bilateral orders stems from the fact that the second difference of the factor complexity can be expressed in terms of bilateral orders of factors. Since only bispecial factors can have non-zero bilateral order, they play important role while studying the language of an infinite word.

The **mirror image** or reversal of a word $w = w_0 w_1 \dots w_n$ is defined as

$$\tilde{w} = w_n w_{n-1} \dots w_0.$$

If a language contains with every factor w also its reversal $\tilde{w}$, we say it is **closed under reversal**.

A finite word such that $w = \tilde{w}$ is a **palindrome**. Given a palindrome $w \in \mathcal{L}(\mathbf{u})$, we say that xwx, $x \in \mathcal{A}$, is a **palindromic extension** of w if $xwx \in \mathcal{L}(\mathbf{u})$.

More basic notions and theorems can be found for instance in [6] or [4].

In [2] more properties of infinite words and relations between them are given. We will mention some of them. We say that an infinite word $\mathbf{u}$ satisfies

- Property $\mathcal{C}$ if its factor complexity is $\mathcal{C}(n) = (\#\mathcal{A} - 1)n + 1$ for all $n \geq 0$;
- Property $\mathcal{PE}$ if every its palindrome of $\mathcal{L}(\mathbf{u})$ has a unique palindromic extension;
- Property $\mathcal{BO}$ if every its factor is ordinary.

Note that most of the mentioned definitions can be rewritten in a more general way for any language. As we will deal only with languages of infinite words, we keep the notions for the language of an infinite word $\mathcal{L}(\mathbf{u})$.

We will use the following relation taken from [2].

Theorem 1 ([2] Theorem 22)**.** *Let* $\mathbf{u}$ *be an infinite word with language closed under reversal satisfying Property* $\mathcal{C}$*. Then Properties* $\mathcal{PE}$ *and* $\mathcal{BO}$ *are equivalent.*

This is an example of a relation between generalized properties of Sturmian words. On a binary alphabet, if we add aperiodicity to the Property $\mathcal{BO}$, the 3 mentioned Properties are equivalent even if we relax to condition of being closed under reversal on the language to be closed under reversal.

The next claim is a reformulation of Corollary 27 from the same paper. It refers to an interesting class of words called **rich** words. A finite word w is rich if it contains $|w| + 1$ palindromic factors (including the empty word). It can be shown that this is the maximum number of palindromic factors that a word can contain, thus the name rich. An infinite word is rich if all its factors are rich. In other words, rich infinite words are fully saturated by palindromic factors.

Claim 2. *Let* $\mathbf{u}$ *be an infinite word with language closed under reversal satisfying Property* $\mathcal{C}$*. If the property* $\mathcal{PE}$ *is satisfied, then the word is rich.*

We will need another theorem.

Theorem 3 (folklore)**.** *If* $\mathbf{u}$ *is a uniformly recurrent word that contains infinitely many distinct palindromes, then its language* $\mathcal{L}(\mathbf{u})$ *is closed under reversal.*

Proof. Suppose for contradiction that there exists a factor $w \in \mathcal{L}(\mathbf{u})$ such that $\tilde{w} \notin \mathcal{L}(\mathbf{u})$. Since $\mathbf{u}$ is uniformly recurrent, there exists an integer K, such that every factor of length at least K contains every factor of length $|w|$. As there is infinitely many palindromes, we can find a palindrome $p \in \mathcal{L}(\mathbf{u})$ such that $|p| \geq K$. As w occurs in p, $\tilde{w}$ occurs also in p and therefore in $\mathcal{L}(\mathbf{u})$ – contradiction. □

3 Example of an infinite word with desired properties

The goal is to construct a rich infinite word having its language closed reversal and containing non-palindromic BS factors. The motivation is that all known examples contain only palindromic BS factors and such example would serve as a counterexample. For larger context see [2].

Example 4. A ternary word with such properties is $\mathbf{v} = \pi(\mathbf{u})$, where $\mathbf{u} = \varphi^2(\mathbf{u})$ (i.e., $\mathbf{u}$ is a fixed point of a morphism φ^2) and $\varphi : \{A, B, C, D\}^* \to \{A, B, C, D\}^*$ and $\pi : \{A, B, C, D\}^* \to \{a, b, c\}^*$ are the following morphisms

$$\varphi : A \to CAC,\ B \to CACBD,\ C \to BDBCA,\ D \to BDB,$$

$$\pi : A \to ba,\ B \to b,\ C \to a,\ D \to abc.$$

The idea of the construction is very different from the actual proof and is inspired by [7]. To prove its properties we will use notions and claims from the previous section. However the proof is rather technical and is separated into 5 lemmas.

The first two lemmas concern the word $\mathbf{u}$.

Lemma 5. *Let $x \in \{A, B, C, D\}$ and $n \geq 1$. Then the word $\varphi^{2n}(x)$ equals py_0y_1 where p is a palindrome and y_0 and y_1 are letters.*

Proof. The proof will be done by induction on n. We will prove the following. For any letter x and any integer n, $\varphi^{2n}(x)$ without last two letters is a palindrome and words $\varphi^{2n}(CA)$ and $\varphi^{2n}(DB)$ differ only on their last two letters.

Before proceeding, note that the suffix of length 2 of $\varphi^{2n}(x)$ can be determined for any n. They are enumerated in Table 1.

x	suffix of $\varphi^{2n}(x)$
A	CA
B	DB
C	AC
D	BD

Table 1: Suffixes of length 2 of words $\varphi^{2n}(x)$

For $n = 1$, the claim can be directly verified.

Suppose now the claim is true for n. Let first $x = A$. We want to prove that

$$\varphi^{2n+2}(A) = \varphi^{2n}\left(BDBCACACBDBCA\right)$$

without last two letters is a palindrome. Let p_y denote the palindrome such that $\varphi^{2n}(y) = p_y y_0 y_1$ with y_0 and y_1 being letters. Using this notation, we can rewrite $\varphi^{2n+2}(A)$ as

$$p_B DB p_D BD p_B DB p_C AC p_A CA p_C AC p_A CA p_C AC p_B DB p_D BD p_B DB p_C AC p_A CA.$$

The fact that $\varphi^{2n}(CA)$ and $\varphi^{2n}(DB)$ differ only on their last two letters can be denoted as

$$p_C AC p_A = p_D BD p_B. \tag{1}$$

Putting these two equalities together we get $\varphi^{2n+2}(A) =$

$$p_BDBp_DBDp_BDBp_DBDp_BCAp_CACp_ACAp_CACp_BDBp_DBDp_BDBp_DBDp_BCA.$$

It is easy to verify that if we cut the suffix CA, we get a palindrome.

For $x = B$ the proof is analogous.

For the remaining two letters C and D we can consider a morphism exchanging the letters σ: $A \leftrightarrow D$ and $B \leftrightarrow C$. It is clear that $\varphi^{2n+2}(C) = \sigma(\varphi^{2n+2}(B))$ and $\varphi^{2n+2}(D) = \sigma(\varphi^{2n+2}(A))$. The first part of the claim is proved.

To prove the remaining relation, one can see that

$$\varphi^{2n+2}(CA) = \varphi^{2n}(BDBCACAC) = p_BDBp_DBDp_BDBp_CACp_ACAp_CACp_ACAp_CAC$$

and

$$\varphi^{2n+2}(DB) = \varphi^{2n}(BDBCACBD) = p_BDBp_DBDp_BDBp_CACp_ACAp_CACp_BDBp_DCA.$$

Comparing the two words, one can see that only their suffix differs and we in fact want to prove that the word p_ACAp_C equals p_BDBp_D. This is true since we can apply σ to the relation (1). On the left-hand side we get

$$\sigma(p_CACp_A) = p_BDBp_D$$

and on the right-hand we have

$$\sigma(p_DBDp_B) = p_ACAp_C.$$

The equality $\sigma(p_x) = p_{\sigma(x)}$ for any letter x is due to the fact that $\sigma(\varphi(x)) = \varphi(\sigma(x))$.

□

The last lemma implies that $\mathcal{L}(\mathbf{u})$ contains infinitely many palindromes. Theorem 3 then implies that $\mathcal{L}(\mathbf{u})$ is closed under reversal.

Lemma 6. *Every LS factor of* $\mathbf{u}$ *is a prefix of* $\varphi^{2n}(B)$ *or* $\varphi^{2n}(C)$ *for some* $n \in \mathbb{N}$.

Proof. It is readily seen that B is LS factor, with DB and CB its left extensions. The same holds for C, whose left extensions are AC and BC.

>From the definition of φ^2, we can see that all short left special factors are prefixes of $\varphi^2(B)$ or $\varphi^2(B)$.

On the other hand, one can see that both $\varphi^{2n}(B)$ and $\varphi^{2n}(C)$ are left special for all n.

Suppose now that $w \in \mathcal{L}(\mathbf{u})$, $|w| > |\varphi^2(B)|$, is left special. Such factor w must be a factor of an image by φ^2 of a shorter LS factor v. Since for any letter x, the word $\varphi^2(x)$ ends by x, we can easily identify the preimage v to be a prefix of $\varphi^{2n}(B)$ or $\varphi^{2n}(C)$ for some n big enough.

□

The following lemmas concern the infinite word $\mathbf{v}$.

Lemma 7. *Let $p \in \mathcal{L}(\mathbf{u})$ be a non-empty palindrome different from B and C. Then $\pi(p) = xyp'$ where x and y are letters and p' is a palindrome.*

Proof. The proof is done by induction on $|p|$. For short palindromes, the claim can be verified easily.

Suppose the claim holds for a palindrome p, $|p| > 3$. Let $\pi(p) = xyp'$ where p' is a palindrome.

We will now deal apart with the four possible cases ApA, BpB, CpC and DpD - palindromes of length $|p| + 2$. Let us recall the factors of $\mathbf{u}$ of length 2

$$\mathcal{L}_2(\mathbf{u}) = \{AC, CA, CB, BC, CD, DB\}.$$

It will serve us to determine xy in each case.

1. ApA: Since $ApA \in \mathcal{L}(\mathbf{v})$, it is clear that either CA or CB is a prefix of p. In both cases this implies that $xy = ab$. Altogether we have $\pi(ApA) = baabp'ba$.

2. BpB : All 3 possible prefixes of p, namely D, CA and CB, imply that $xy = ab$. One can see that $\pi(BpB) = babp'b$.

3. CpC : 3 possible prefixes of p are BD, BC and A. Thus $xy = ba$. Therefore $\pi(CpC) = abap'a$.

4. DpD : 2 possible prefixes of p are BD and BC. As $xy = ba$, we have $\pi(DpD) = abcbap'abc$.

□

Lemma 8. *Every LS factor of $\mathbf{v}$ is a prefix of $\pi(\varphi^{2n}(B))$ or $\pi(\varphi^{2n}(C))$ for some $n \in \mathbb{N}$.*

Proof. Let $w \in \mathcal{L}(\mathbf{v})$ be a LS factor. Let xw and yw, where x and y are letters, be its left extensions. It is clear that there exists a left special factor $W \in \mathcal{L}(\mathbf{u})$ such that xw is a factor of XW, where X is a letter. Analogously, there exists an extension YW. Since the pair (X, Y) is either (A, B) or (C, D), and words in pairs $(\pi(A), \pi(B))$ and $(\pi(C), \pi(D))$ end in different letters, one can see that w is a prefix of $\pi(W)$. The claim then follows from Lemma 6. □

Lemma 9. *The infinite word $\mathbf{v}$ contains infinitely many non-palindromic BS factors.*

Proof. Consider $r \in \mathcal{L}(\mathbf{u})$ to a BS factor starting by the letter B and ending by the letter C. We will firstly show that $\pi(r)ba$ is a non-palindromic BS factor of $\mathcal{L}(\mathbf{v})$. In the second part we will show that such a factor r exists and there are infinitely many factors with such properties.

It is clear that r is extended in $\mathcal{L}(\mathbf{u})$ by the letters B and C to the left and to the right. Since $\pi(B)$ and $\pi(C)$ end in different letter, we can see that $\pi(r)$ is LS. Since the last letter of r is C, the factor r can be extended to the right by one of the following words: $\{BD, BCA, BCB, AC\}$. The longest common prefix of the words $\{\pi(BD), \pi(BCA), \pi(BCB), \pi(AC)\}$ is ba and is not equal to neither of them. Therefore,

$\pi(r)ba$ is right special. As the first letter of r is B, the first letter of $\pi(r)ba$ is b. Thus $\pi(r)ba$ is a non-palindromic BS factor of $\mathbf{v}$.

It remains to show that $\mathcal{L}(\mathbf{u})$ contains infinitely many BS factors starting by the letter B and ending by the letter C. To prove that take an arbitrary integer n and a non-empty prefix w of the word $\varphi^{2n}(B)$. According to Lemma 6 w is a LS factor. As $\mathbf{u}$ is uniformly recurrent, we can extend w to the right in a unique manner until we have a factor $s \in \mathcal{L}(\mathbf{u})$ which is also right special.

Since $\mathbf{u}$ is closed under reversal, there are two possibilities for the last letter of s. It is either B or C.

If the last letter is C, then s is followed by one of 3 factors: AC, BD or BC. The longest common prefix of factors $\{\varphi^2(AC), \varphi^2(BD), \varphi^2(BC)\}$ is the word $\varphi^2(B)$. Moreover, the longest common prefix is shorter than any of these 3 words. Thus the word $\varphi^2(s)\varphi^2(B)$ is right special with two right extending letters - B and C. Note that $\varphi^2(sB)$ ends by the letter B. According to Lemma 6, $\varphi^2(sB)$ is also left special since it is an image by φ^2 of a LS factor.

If the last letter of s is B, the situation is analogous. We can construct a longer BS factor $\varphi^2(sC)$, this time ending by the letter C.

Altogether we can find an infinite sequence of non-palindromic BS factors of $\mathcal{L}(\mathbf{u})$ starting by B and ending by C. □

Proof of properties of $\mathbf{v}$. As $\mathbf{v}$ contains infinitely many distinct palindromes (Lemma 7) and is a morphic image of a uniformly recurrent word, thus uniformly recurrent, according to Theorem 3 the language $\mathcal{L}(\mathbf{v})$ is closed under reversal. Using Lemma 8 we see that $\mathcal{L}(\mathbf{v})$ has 2 LS factors of each positive length. This implies that for $n \geq 0$ we have

$$\mathcal{C}(n+1) - \mathcal{C}(n) = 2$$

as any other non-special factor has only one left extension. Therefore, $\mathbf{v}$ has complexity $\mathcal{C}(n) = 2n + 1$, i.e., Property $\mathcal{C}$ holds. Since there are two infinite words whose prefixes are exactly LS factors, with 2 left extensions each time, it can be deduced from the definition of the bilateral order and the fact that the language is closed under reversal that $\mathbf{v}$ contains only ordinary BS factors. Applying Theorem 1, Property $\mathcal{PE}$ holds as well. Finally, as Lemma 9 states, $\mathcal{L}(\mathbf{v})$ contains infinitely many non-palindromic BS factors. □

The word $\mathbf{u}$ is ternary. One might ask if there exists a binary word with such properties. The answer is positive, we can use $\mathbf{v}$ to construct such word. Let us define a morphism ψ as follows:

$$\psi : a \to 01, b \to 010, c \to 01011.$$

In fact, this morphism is from a class denoted P_{ret} treated in [1]. We will not give details of this class, however the form of ψ, Property $\mathcal{PE}$ and richness of $\mathbf{v}$ imply that $\psi(\mathbf{v})$ is rich and has language closed under reversal.

In the proof of Lemma 9, it is in fact stated that $\mathbf{v}$ contains infinitely many BS factors starting by b, ending by a, whose right and left extending letters are a and b. Take r to be a such BS factor. It is readily seen that $\psi(r)010$ is a BS factor of $\psi(\mathbf{v})$. Since r starts by b and ends by a, it follows that $\psi(r)010$ has a prefix 0100 and a suffix 1010, thus it is

not a palindrome. Thus, $\psi(\mathbf{u})$ is a binary word having the same properties as the word $\mathbf{u}$.

4 Final remark

In the proof of the properties of $\mathbf{u}$ we have used a common technique used to analyse a language of an infinite word which is produced by morphisms. Such technique may not be sufficient for some special cases of morphisms.

References

[1] L. Balková, E. Pelantová, and Š. Starosta. *Infinite words with finite defect.* http://arxiv.org/abs/1009.5105

[2] L. Balková, E. Pelantová, and Š. Starosta. *Sturmian jungle (or garden?) on multiliteral alphabets.* http://arxiv.org/abs/1003.1224, to appear in RAIRO-Theor. Inf. Appl.

[3] J. Cassaigne. *Complexity and special factors.* Bull. Belg. Math. Soc. Simon Stevin 4 **1** (1997), 67–88.

[4] N. P. Fogg. *Substitutions in Arithmetics, Dynamics and Combinatorics.* Springer, 1st edition, (2002).

[5] P. Kůrka. *Topological and Symbolic Dynamics.* Societé Mathématique de France, (2004).

[6] M.Lothaire. *Algebraic combinatorics on words.* Number 90 in Encyclopedia of Mathematics and its Applications. Cambridge University Press, (2002).

[7] G. Rote. *Sequences with subword complexity* $2n$. J. Number Th. **46** (1993), 196–213.

A Quantitative Criterion for Artificial Diffusion in Numerical Schemes

Pavel Strachota

3rd year of PGS, email: pavel.strachota@fjfi.cvut.cz
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Michal Beneš, Deptartment of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. In the course of tuning the developed numerical algorithm for MR-DTI data visualization, it was necessary to introduce a measurement technique capable of quantitatively assessing the artificial isotropic diffusion in numerical schemes for PDE. Based on such assessment, a qualified choice of the numerical scheme can be made. This contribution describes the proposed measurement technique based on total variation evaluation. The procedure is applied to several numerical discretizations of the anisotropic diffusion model based on the Allen-Cahn equation and the obtained results are presented.

Keywords: Allen-Cahn equation, artificial numerical diffusion, finite volume method, multipoint flux approximation, total variation

Abstrakt. V průběhu ladění numerického algoritmu pro vizualizace dat z MR-DTI bylo nutné vyvinout metodu schopnou kvantitativně posoudit míru umělé numerické difuze, která je přítomna v každém numerickém schématu pro diskretizaci PRD. Na základě takového měření lze pak kvalifikovaně rozhodnout o volbě numerického schématu. Příspěvěk popisuje techniku měření založenou na vyhodnocování totální variace. Tato procedura je použita na několik numerických diskretizací modelu anizotropní difuze založeného na Allenově-Cahnově rovnici a jsou prezentovány dosažené výsledky.

Klíčová slova: Allenova-Cahnova rovnice, umělá numerická difuze, metoda konečných objemů, vícebodová aproximace toku, totální variace

1 Introduction

The Allen-Cahn equation having its origin in phase modeling in physics [1] has since found its application in other fields, including image processing and mathematical visualization [2, 9]. In particular, in order to visualize the streamlines of a given tensor field in 3D, an initial boundary value problem for the modified Allen-Cahn equation with incorporated anisotropy can be used (see [9, 11] and [7]), giving similar results to the LIC method[3, 5]. We begin with the problem formulation and describe its numerical solution using several flux approximation schemes on a rectangular grid. The schemes suffer from an undesired numerical dissipation effect which demonstrates itself as an additional isotropic diffusion of the solution. Hence, we proceed with the development of a measurement technique that would provide for assessing the amount of the numerical diffusion produced by the schemes. A quantitative scheme comparison criterion is thereby created.

2 Problem for the Allen-Cahn equation with anisotropy

Formulation. Assume there is a symmetric positive definite tensor field $\boldsymbol{D} : \bar{\Omega} \mapsto \mathbb{R}^{3\times 3}$ where $\Omega \subset \mathbb{R}^3$ is a block shaped domain. On the time interval $\mathcal{J} = (0, T)$, the initial boundary value problem for the anisotropic Allen-Cahn equation reads

$$\xi \frac{\partial p}{\partial t} = \xi \nabla \cdot \boldsymbol{D} \nabla p + \frac{1}{\xi} f_0(p) \qquad \text{in } \mathcal{J} \times \Omega, \tag{2.1}$$

$$\left.\frac{\partial p}{\partial n}\right|_{\partial\Omega} = 0 \qquad \text{on } \bar{\mathcal{J}} \times \partial\Omega, \tag{2.2}$$

$$p|_{t=0} = I \qquad \text{in } \Omega \tag{2.3}$$

where

$$f_0(p) = p(1-p)\left(p - \frac{1}{2}\right).$$

Let $x \in \Omega$. Thanks to $\boldsymbol{D}(x)$ in the diffusion term on the right hand side of (2.1), the diffusion of p at x is focused into the direction of the principal eigenvector of $\boldsymbol{D}(x)$, or more precisely, with the directional distribution described by the ellipsoid

$$\left\{ \boldsymbol{\eta} \in \mathbb{R}^3 \middle| \boldsymbol{\eta}^{\mathrm{T}} \boldsymbol{D}(x)^{-1} \boldsymbol{\eta} = 1 \right\}.$$

In terms of tensor field visualization, we choose the initial condition I in (2.3) as a noisy texture, preferably an impulse noise. Due to the anisotropic diffusion process carried out by solving (2.1-2.3), the solution p changes in time from noise to an organized structure. Streamlines of the field of principal eigenvectors of $\boldsymbol{D}$ can be recognized there as parts with locally similar value of p. The term f_0 efficiently increases contrast of the resulting 3D image provided that the parameter ξ and the final time T are chosen appropriately (in our case by experiment). In order to actually view the resulting 3D image $p(\cdot, T)$, 2D slices through Ω can be helpful.

Numerical solution. For numerical solution, the *method of lines* [8] is utilized. Applying a finite volume discretization scheme in space, the problem (2.1-2.3) is converted to a system of ODE in the general form

$$\frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}t} = \boldsymbol{f}(t, \boldsymbol{p}). \tag{2.4}$$

Thereafter, we employ the 4th order Runge-Kutta-Merson solver with adaptive time stepping to solve (2.4).

Describing the finite volume scheme, (2.4) can also be referred to as the semidiscrete scheme and written in the form

$$\xi \frac{\mathrm{d}}{\mathrm{d}t} p_K(t) = \xi \sum_{\sigma \in \mathcal{E}_K} F_{K,\sigma}(t) + \frac{1}{\xi} f_{0,K}(t) \qquad \forall K \in \mathcal{T} \tag{2.5}$$

where $\mathcal{T}$ is an admissible finite volume mesh [4], $K \in \mathcal{T}$ is one particular control volume (cell) and $\mathcal{E}_K$ is the set of all faces of the cell K. $F_{K,\sigma}(t)$ represent the respective numerical

fluxes at the time t, which contain difference quotients approximating the derivatives $\partial_x p$, $\partial_y p$, $\partial_z p$ at the center of the face σ.

Artificial dissipation and finite volume scheme design. As indicated in the introduction, all schemes introduce a certain amount of *artificial (numerical) isotropic diffusion* in the solution. However, its strength depends on the exact form of $F_{K,\sigma}$. This phenomenon needs to be suppressed as much as possible as it may significantly deteriorate the visual quality of the result due to *blurring*. Its cause lies in the occurrence of high frequency structures in the solution: both the initial noise and the forming streamlines. To be treated correctly, they require the difference operators used in $F_{K,\sigma}$ to be of an appropriate order [10, 6].

We have assembled and investigated numerical schemes using the following approximations of the derivatives in the flux term:

- second order central difference approximation with linear interpolation of the missing points in the difference stencil;
- fourth order *multipoint flux approximation* (MPFA) central difference scheme with linear interpolation;
- fourth order MPFA central difference scheme with *cubic* interpolation.

Thereto, a classical forward-backward first order finite difference (FD) scheme has been added. For more details on the design of the MPFA differences, see [11].

3 Artificial diffusion measurement

Having the results available obtained by using different schemes but based on identical input settings, one can try to compare them visually to decide on the scheme with the least artificial diffusion. In Figure 3.1, an example of such comparison is demonstrated on a real-data MR-DTI neural tract visualization. In the center part of the images, a major neural tract in the shape of U is displayed in the form of streamlines. It can be observed that the FD scheme produces undesired isotropic diffusion greatly dependent on the prescribed direction of diffusion. This is related to the asymmetry of the difference stencil. The 2nd order central difference flux approximation used in the FV scheme is already symmetric. However, it is clearly outperformed by the scheme based on MPFA which causes significantly weaker blurring.

Scheme assessment by total variation. In this part we introduce a quantitative measure of the artificial diffusion in the schemes. For this purpose, the total variation of the numerical solution $p^h = p^h(t)$ finds its rather unusual application. It is defined as

$$TV\left(p^h\right) = \sum_{K \in \mathcal{T}} \left|\nabla_h p_K^h\right| m(K) \tag{3.1}$$

where $\nabla^h p_K^h$ represents the discrete approximation of the gradient and $m(K)$ is the measure of the cell K. From the image processing point of view, the value of TV is

FD FV 2nd order MPFA cubic 4th order

Figure 3.1: Artificial diffusion in different numerical schemes. Crops from colorized MR-DTI visualizations based on real data, transverse plane slice.

proportional to both the number of edges in the image p^h and its contrast. Both these quantities assume their maxima for the noisy initial condition and change in time along with the diffuse evolution of the numerical solution. Performing two computations with identical settings except for the choice of the numerical scheme, it is possible to directly compare the TV values of the results. The scheme producing an image with a greater value of TV exhibits less artificial diffusion as it maintains more edges, more contrast, or both.

Scheme comparison methodology. We have performed extensive testing with phantom input tensor fields to investigate the behavior of the schemes depending on the prescribed direction of diffusion. For each triple of spherical coordinates $(r = 1, \varphi, \theta)$ where $\varphi \in [0, 360°]$, $\theta \in [-90°, 90°]$, let a unit vector

$$\boldsymbol{v}_1(\varphi, \theta) = (\cos\varphi\cos\theta, \sin\varphi\cos\theta, \sin\theta)$$

represent the principal eigenvector of a uniform tensor field $\boldsymbol{D}(\varphi, \theta)$, corresponding to the eigenvalue $\lambda_1 = 100$. The remaining eigenvalues are $\lambda_2 = \lambda_3 = 1$ and the eigenvectors $\boldsymbol{v}_2, \boldsymbol{v}_3$ complete the orthonormal basis of R^3. Afterwards, a computation is carried out using $\boldsymbol{D}(\varphi, \theta)$ as input data and subsequently, TV is evaluated from the resulting datasets. The TV values alone are not of particular interest since they depend on both the grid dimensions and the size of the domain Ω. However, the relative differences of TV between schemes provide the desired information.

The results of the procedure described above performed for all the four schemes in several time levels are shown in Figures 3.2-3.6. In all the graphs, TV is normalized so that the maximum in each chart is 1. Settings of all important computation parameters can be found in the figure captions. In Figure 3.2, the latitude θ is fixed to 0 and the longitude φ traverses the angles from 0° to 350° with the step 10°. The same is true for Figure 3.3 which only differs from Figure 3.2 in the setting of parameter ξ. Figure 3.4 depicts the "diagonal" cut through the space (φ, θ) in the range from 0° to 90°, including the worst situation for all schemes where $\varphi = \theta = 45°$. Finally, Figures 3.5 and 3.6 contain surface plots of all measured combinations of φ, θ for the FD scheme and the MPFA FV scheme with cubic interpolation, respectively.

Observations from Figures 3.2-3.6 can be summed up as follows:

- Artificial diffusion clearly depends on $\boldsymbol{v}_1$ and occurs least when the direction $\boldsymbol{v}_1$ is aligned with coordinate axes. For the FD scheme, a straightforward explanation

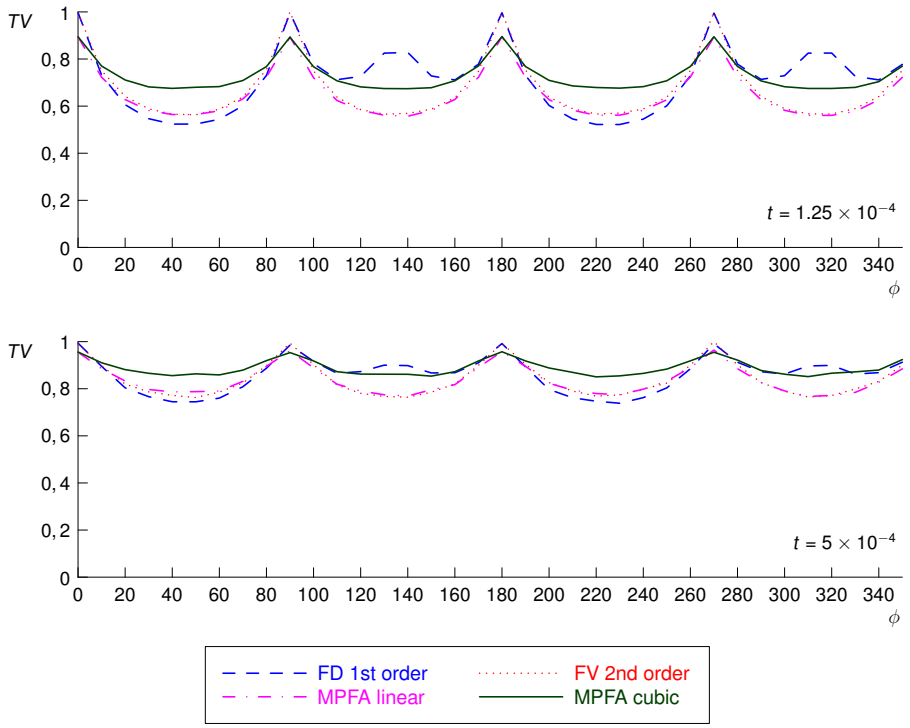


Figure 3.2: Comparison of numerical schemes based on TV, $\xi = 5 \times 10^{-3}$, $\theta = 0$, $\varphi \in [0°, 350°]$.

can be given: In the degenerate case $\lambda_2 = \lambda_3 \to 0$, the equation systems for different rows of grid nodes along $\boldsymbol{v}_1$ become independent.

- The performance of all schemes improves (i.e. TV rises) with growing time. This is obvious as the ongoing diffusion gradually limits the frequency spectrum of the solution. At the beginning, the infinite spectrum of the initial condition can not be handled properly by any difference operator.
- The performance of the schemes improves with decreasing ξ (compare Figures 3.2 and 3.3).
- The FD scheme exhibits a highly asymmetric behavior (see Figures 3.3 and 3.5).
- All FV schemes are symmetric.
- The FV scheme with MPFA and cubic interpolation outperforms all other schemes in the comparison except for the FD scheme when $\boldsymbol{v}_1$ is aligned with some coordinate axis.

4 Conclusion

We have developed an approach for measuring the amount of artificial isotropic diffusion in numerical schemes. Thorough computational studies based on phantom input data

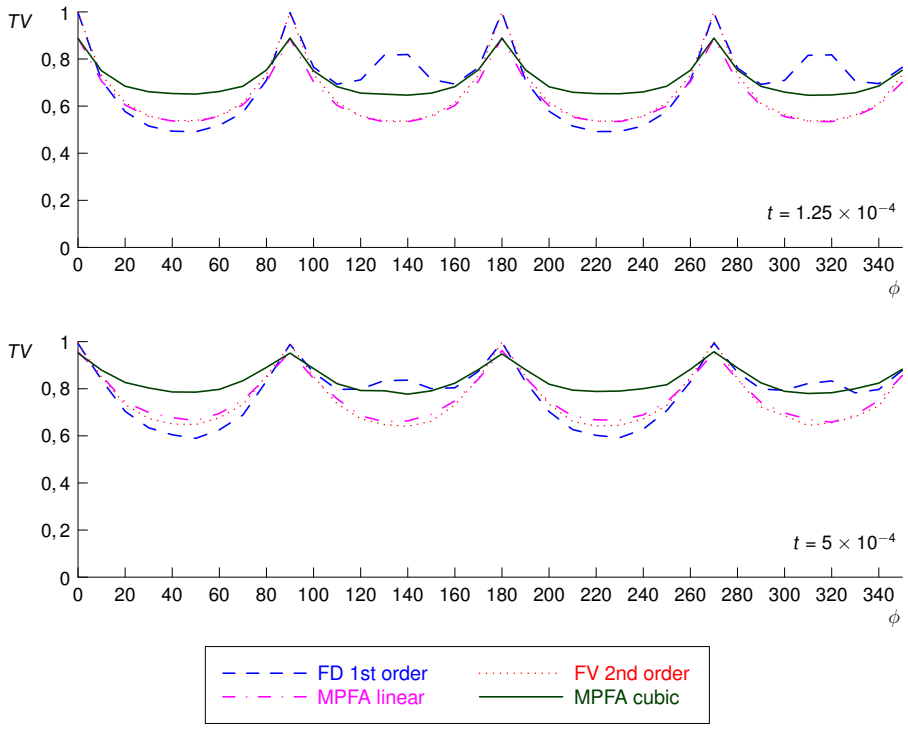


Figure 3.3: Comparison of numerical schemes based on TV, $\xi = 10^{-2}$, $\theta = 0$, $\varphi \in [0°, 350°]$.

confirm that this technique fulfills the given objective and produces results in agreement with an intuitive notion of blurring observable in images obtained by solving (2.5). Introducing a suitable threshold in (3.1), the measurement can also be applied to computations with real MR-DTI input data.

Acknowledgments: This work was carried out under the HPC-EUROPA++ project (project number: 211437), with the support of the European Community - Research Infrastructure Action of the FP7 "Coordination and support action" Program. This work was supported by the Grant Agency of the Czech Technical University in Prague, grant No. SGS10/086/OHK4/1T/14. Partial support of the project "Jindřich Nečas Center for Mathematical Modeling", No. LC06052. Special thanks to the colleagues at the Institute for Clinical and Experimental Medicine (IKEM) in Prague for providing input datasets, consultations, and support.

References

[1] S. Allen and J. W. Cahn. *A microscopic theory for antiphase boundary motion and its application to antiphase domain coarsening.* Acta Metall. **27** (1979), 1084–1095.

[2] M. Beneš, V. Chalupecký, and K. Mikula. *Geometrical image segmentation by the Allen-Cahn equation.* Appl. Numer. Math. **51** (2004), 187–205.

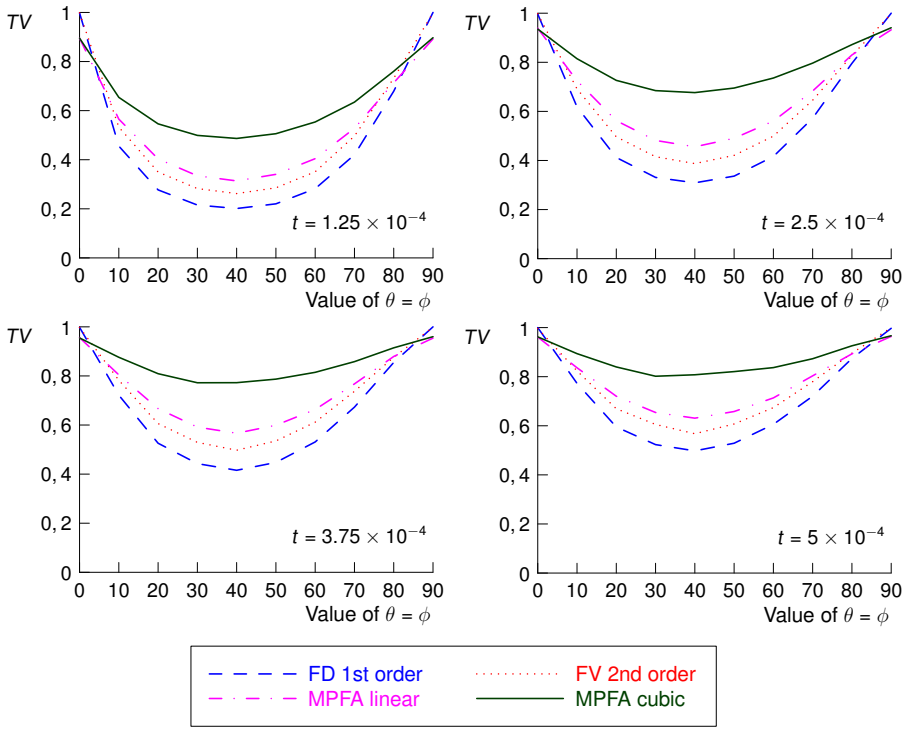


Figure 3.4: Comparison of numerical schemes based on TV in different time levels, $\xi = 10^{-2}$, $\theta \in [0°, 90°]$, $\varphi = \theta$.

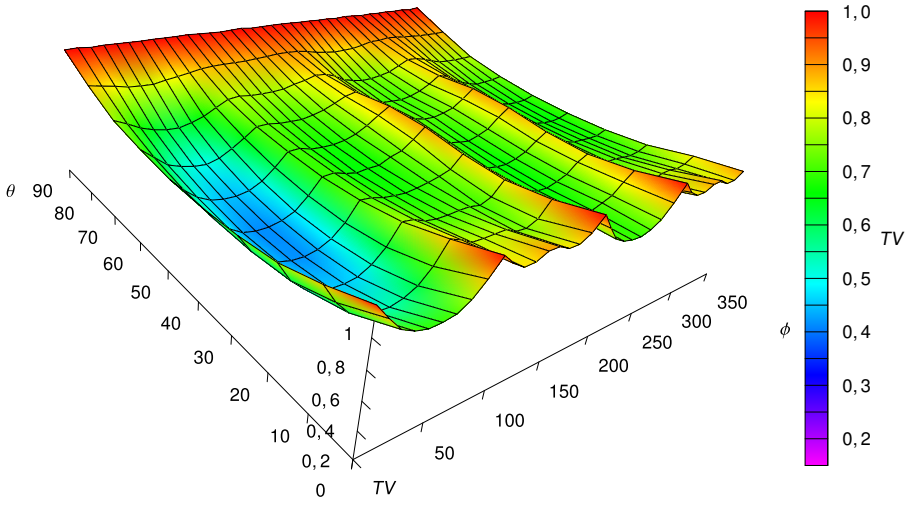


Figure 3.5: TV graph for the FD scheme, $t = 3.75 \times 10^{-4}$, $\xi = 5 \times 10^{-3}$, $\theta \in [0°, 90°]$, $\varphi \in [0°, 350°]$.

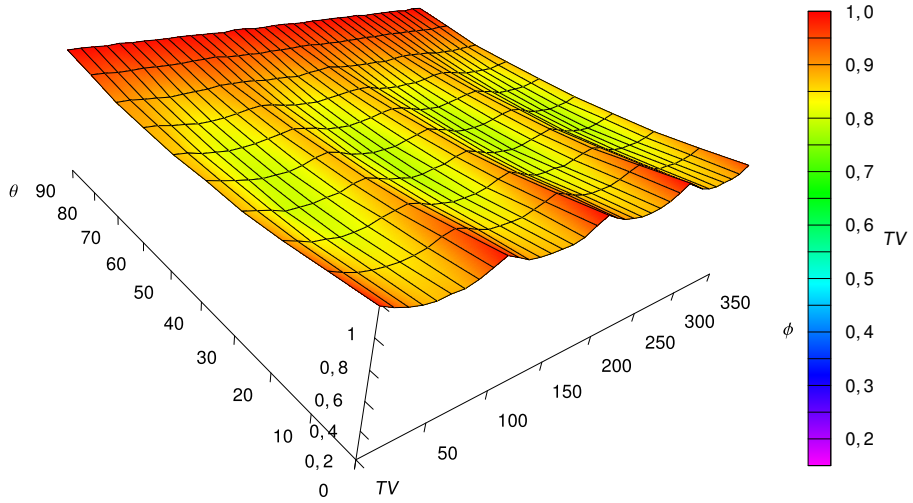


Figure 3.6: TV graph for the FV scheme with MPFA and cubic interpolation, $t = 3.75 \times 10^{-4}$, $\xi = 5 \times 10^{-3}$, $\theta \in [0°, 90°]$, $\varphi \in [0°, 350°]$.

[3] B. Cabral and L. C. Leedom. Imaging vector fields using line integral convolution. In 'SIGGRAPH '93: Proceedings of the 20th annual conference on Computer graphics and interactive techniques', 263–270, New York, NY, USA, (1993). ACM.

[4] R. Eymard, T. Gallouët, and R. Herbin. *Finite volume methods.* In 'Handbook of Numerical Analysis', P. G. Ciarlet and J. L. Lions, (eds.), volume 7, Elsevier (2000), 715–1022.

[5] E. Hsu. Generalized line integral convolution rendering of diffusion tensor fields. In 'Proc. Intl. Soc. Mag. Reson. Med', volume 9, 790, (2001).

[6] H. Lomax, T. H. Pulliam, and D. W. Zingg. *Fundamentals of Computational Fluid Dynamics.* Springer, (2001).

[7] T. Preußer and M. Rumpf. Anisotropic nonlinear diffusion in flow visualization. In 'Proceedings of IEEE Visualization 1999', 325–332, (1999).

[8] W. E. Schiesser. *The Numerical Method of Lines: Integration of Partial Differential Equations.* Academic Press, San Diego, (1991).

[9] P. Strachota. Vector field visualization by means of anisotropic diffusion. In 'Proceedings of Czech Japanese Seminar in Applied Mathematics 2006', M. Beneš, M. Kimura, and T. Nakaki, (eds.), volume 6 of *COE Lecture Note*, 193–205. Faculty of Mathematics, Kyushu University Fukuoka, (2007).

[10] P. Strachota. Antidissipative numerical schemes for the anisotropic diffusion operator in problems for the Allen-Cahn equation. In 'ALGORITMY 2009 - Proceedings of contributed lectures and posters', A. Handlovičová, P. Frolkovič, K. Mikula, and D. Ševčovič, (eds.), volume 18, 134–142. Slovak University of Technology in Bratislava, (2009).

[11] P. Strachota. *Implementation of the MR tractography visualization kit based on the anisotropic Allen-Cahn equation.* Kybernetika **45** (2009), 657–669.

The Characteristic Function for a Particular Class of Infinite Jacobi Matrices

František Štampach

1st year of PGS, email: stampfra@fjfi.cvut.cz
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Pavel Šťovíček, Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU

Abstract. Spectral properties of semi-infinite symmetric Jacobi matrices with diagonal created by a real and strictly increasing sequence such that the reciprocal sequence belongs to ℓ^2 is studied. Parallels to the diagonal are composed of a positive and bounded sequence. It is shown the spectrum of these matrices is simple and discrete. Further functions $\mathfrak{E}$ and $\mathfrak{F}$ with simple and nice algebraic properties are defined on a subset of the space of complex sequences. Some special functions are expressible in terms of these functions, first of all the Bessel functions of first kind. The main result of this work is finding a function which is expressed with the aid of $\mathfrak{F}$ applied to a certain sequence with a property that zeros of this function coincide with eigenvalues of the Jacobi matrix under investigation. At the end general results are applied to a simple example. It is demonstrated the spectrum of the semi-infinite Jacobi matrix with linear diagonal and constant parallels coincides with zeros of the Bessel function of the first kind considered as the function of its order.

Keywords: tridiagonal matrix, Jacobi matrix, eigenvalue problem, characteristic function

Abstrakt. V tomto příspěvku jsou studovány vlastnosti spektra polo-nekonečných symetrických Jacobiho matic, jejichž diagonála je tvořena reálnou a ryze rostoucí posloupností s převrácenou hodnotou v ℓ^2. Vedlejší diagonály tvoří pozitivní a omezená posloupnost. Je ukázáno, že spektrum těchto matic je jednoduché a diskrétní. Dále jsou zevedeny funkce $\mathfrak{E}$ a $\mathfrak{F}$ na podmnožině prostoru komplexních posloupností. Tyto funkce mají jendoduché a pěkné algebraické vlastnosti. Některé speciální funkce lze vyjádřit pomocí funkcí $\mathfrak{E}$ a $\mathfrak{F}$, například Besselovy funkce prvního druhu. Hlavním výsledkem práce je nalezení funkce vyjádřené pomocí $\mathfrak{F}$ aplikované na určitou posloupnost, která má tu vlastnost, že její nuly odpovídají vlastním hodnotám studované Jacobiho matice. Na příkladu je ukázáno, že spektrum polo-nekonečné Jacobiho matice, jejíž diagonála závisí lineárně na indexu a vedlejší diagonály tvoří kladný parametr, se shoduje s množinou všech nul Besselovy funkce prvního druhu jako funkce jejího řádu.

Klíčová slova: tridiagonální matice, Jacobiho matice, vlastní čísla, charakteristická funkce

1 Introduction

In the whole paper I assume real strictly insreasing sequence $\{\lambda_n\}_{n=1}^{\infty}$ satisfying condition

$$\sum_{n=1}^{\infty}\frac{1}{\lambda_n^2}<\infty, \tag{1}$$

and positive and bounded sequence $\{w_n\}_{n=1}^{\infty}$ to be given. Note the assumptions imply

$$\lim_{n\to\infty}\lambda_n = +\infty.$$

Next let me denote J a seminfinite symmetric Jacobi matrix whose diagonal is created by sequence $\{\lambda_n\}_{n=1}^{\infty}$ and parallels to the diagonal are consist of sequence $\{w_n\}_{n=1}^{\infty}$, i.e.

$$J := \begin{pmatrix} \lambda_1 & w_1 & & & \\ w_1 & \lambda_2 & w_2 & & \\ & w_2 & \lambda_3 & w_3 & \\ & & \ddots & \ddots & \ddots \end{pmatrix}. \tag{2}$$

Further J_n denotes the truncation of J, i.e.

$$J_n = \begin{pmatrix} \lambda_1 & w_1 & & & \\ w_1 & \lambda_2 & w_2 & & \\ & \ddots & \ddots & \ddots & \\ & & w_{n-2} & \lambda_{n-1} & w_{n-1} \\ & & & w_{n-1} & \lambda_n \end{pmatrix}. \tag{3}$$

There exists a sequence of polynomials $q_k(\xi)$, $k = 1, 2, \ldots$, such that the degree of $q_k(\xi)$ equals to $k-1$, the coefficients of $q_k(\xi)$ are rational functions of $\lambda_1, \ldots, \lambda_{k-1}$ and $w_1, \ldots, w_{k-1}$, and whenever the sequence $\{x_k\}_{k=1}^{\infty}$ solves the eigenvalue equation

$$\begin{aligned} &\lambda_1 x_1 + w_1 x_2 = \xi x_1, \\ &w_{k-1}x_{k-1} + \lambda_k x_k + w_k x_{k+1} = \xi x_k, \quad k = 2, 3, \ldots \end{aligned}$$

with an eigenvalue $\xi \in \mathbb{C}$ then it holds

$$x_k = q_k(\xi)x_1, \quad k = 1, 2, \ldots$$

Actually, one sets $q_1(\xi) := 1$ and the sequence of polynomials is unambiguously defined by the reccurent relation

$$q_{k+1}(\xi) = \frac{\xi - \lambda_k}{w_k}q_k(\xi) - \frac{w_{k-1}}{w_k}q_{k-1}(\xi), \quad k = 1, 2, \ldots$$

(where one has to set $w_0 := 0$).

Corollary 1. *Any eigenvalue of J regarded as an operator in $\ell^2(\mathbb{N})$ is simple.*

Proposition 2. *The spectrum of J is discrete.*

Proof. Operator J can be decomposed as

$$J = \Lambda + W + W^*$$

where $\Lambda := \operatorname{diag}(\lambda_1, \lambda_2, \dots)$ and

$$W := \begin{pmatrix} 0 & w_1 & 0 & 0 & \\ 0 & 0 & w_2 & 0 & \\ 0 & 0 & 0 & w_3 & \\ & & \ddots & \ddots & \ddots \end{pmatrix}.$$

Since sequence $\{w_n\}_{n=1}^{\infty}$ is bounded W is bounded, $\|W\| = \sqrt{\|WW^*\|} = \sup w_n$. Obviously the essential spectrum of Λ is empty, hence Λ has a compact resolvent. Consequently, $W + W^*$ is Λ-compact self-adjoint perturbation ($(W + W^*)(\Lambda - i)^{-1}$ is compact) and, by the Weyl criterion, the essential spectrum of J is empty. □

2 Function $\mathfrak{E}$ and function $\mathfrak{F}$

Definition 3. Define $\mathfrak{E} : D \to \mathbb{C}$, $\mathfrak{F} : D \to \mathbb{C}$,

$$\mathfrak{E}(x) = 1 + \sum_{m=1}^{\infty} \sum_{k_1=1}^{\infty} \sum_{k_2=k_1+2}^{\infty} \dots \sum_{k_m=k_{m-1}+2}^{\infty} x_{k_1} x_{k_1+1} x_{k_2} x_{k_2+1} \dots x_{k_m} x_{k_m+1}$$

and

$$\mathfrak{F}(x) = 1 + \sum_{m=1}^{\infty} (-1)^m \sum_{k_1=1}^{\infty} \sum_{k_2=k_1+2}^{\infty} \dots \sum_{k_m=k_{m-1}+2}^{\infty} x_{k_1} x_{k_1+1} x_{k_2} x_{k_2+1} \dots x_{k_m} x_{k_m+1}$$

where

$$D = \left\{ \{x_k\}_{k=1}^{\infty} \subset \mathbb{C};\ \sum_{k=1}^{\infty} |x_k x_{k+1}| < \infty \right\}.$$

For a finite number of complex variables we identify $\mathfrak{F}(x_1, x_2, \dots, x_n)$ with $\mathfrak{F}(x)$ where $x = (x_1, x_2, \dots, x_n, 0, 0, 0, \dots)$ and similarly for $\mathfrak{E}$. By convention, we also put $\mathfrak{E}(\emptyset) = \mathfrak{F}(\emptyset) = 1$ where $\emptyset$ is the empty sequence.

Remark 4. Note that the domain D is not a linear space. One has, however, $\ell^2(\mathbb{N}) \subset D$.

The function $\mathfrak{F}$ is continuous functional on $\ell^2(\mathbb{N})$ which satisfies the following identities. First, the recurrent relation

$$\mathfrak{F}(x) = \mathfrak{F}(Tx) - x_1 x_2\, \mathfrak{F}(T^2 x) \tag{4}$$

holds for all $x \in D$ (T stands for the shift operator). This identity admits a generalization

$$\mathfrak{F}(x) = \mathfrak{F}(x_1, \dots, x_k)\, \mathfrak{F}(T^k x) - \mathfrak{F}(x_1, \dots, x_{k-1}) x_k x_{k+1} \mathfrak{F}(T^{k+1} x), \quad k = 1, 2, \dots. \tag{5}$$

Next, the Bessel function of the first kind can be expressed with the aid of $\mathfrak{F}$. More precisely, for $\nu \notin -\mathbb{N}$, one has

$$J_\nu(2w) = \frac{w^\nu}{\Gamma(\nu + 1)} \mathfrak{F}\left(\left\{ \frac{w}{\nu + k} \right\}_{k=1}^{\infty} \right). \tag{6}$$

The respective proofs of these identitites are work out in [4].

The function $\mathfrak{E}$ has very similar properties as $\mathfrak{F}$. $\mathfrak{E}$ is continuous functional on $\ell^2(\mathbb{N})$ and it holds

$$\mathfrak{E}(x) = \mathfrak{E}(Tx) + x_1 x_2\, \mathfrak{E}(T^2 x), \tag{7}$$

$$\mathfrak{E}(x) = \mathfrak{E}(x_1,\dots,x_k)\, \mathfrak{E}(T^k x) + \mathfrak{E}(x_1,\dots,x_{k-1}) x_k x_{k+1} \mathfrak{E}(T^{k+1} x), \quad k = 1,2,\dots, \tag{8}$$

and

$$I_\nu(2w) = \frac{w^\nu}{\Gamma(\nu+1)} \mathfrak{E}\left(\left\{ \frac{w}{\nu+k} \right\}_{k=1}^{\infty} \right) \tag{9}$$

where $x \in D$, $\nu \notin -\mathbb{N}$ and I stands for the modified Bessel function of the first kind. All the proofs of (7), (8) and (9) can be done by the same way as the proofs of (4), (5) and (6) with only slight modifications.

Finnaly, an obvious inequality

$$|\mathfrak{F}(x)| \leq \mathfrak{E}(|x|) \tag{10}$$

holds for any $x \in D, |x| = (|x_1|, |x_2|, \dots)$.

Application of $\mathfrak{F}$ to a finite complex sequence can be unambiguously defined with the aid of a determinant of a Jacobi matrix.

Proposition 5. *For $d \in \mathbb{N}$ and $\{x_j\}_{j=1}^d \subset \mathbb{C}$, one has*

$$\mathfrak{F}(x_1,x_2,\dots,x_d) = \left| \begin{pmatrix} 1 & x_1 & & & & \\ x_2 & 1 & x_2 & & & \\ & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & & & x_{d-1} & 1 & x_{d-1} \\ & & & & x_d & 1 \end{pmatrix} \right|. \tag{11}$$

Proof. The case $d = 1, 2$ is easy to verify. Denote the RHS of (11) by $D(x_1, x_2, \dots, x_d)$. By expanding $D(x_1, x_2, \dots, x_d)$ along the first row one finds out the recurrence rule

$$D(x_1,x_2,\dots,x_d) = D(x_2,x_3,\dots,x_d) - x_1 x_2 D(x_3,x_4,\dots,x_d)$$

holds. Now, it suffices to apply the induction hypothesis and (4). □

Remark 6. The Jacobi matrix J_n can be decomposed into the product

$$J_n = G_n \tilde{J}_n G_n \tag{12}$$

where $G_n = \operatorname{diag}(\gamma_1, \gamma_2, \dots, \gamma_n)$ is a diagonal matrix and $\tilde{J}_n$ is a Jacobi matrix with all units on the neighboring parallels to the diagonal,

$$\tilde{J}_n = \begin{pmatrix} \tilde{\lambda}_1 & 1 & & & & \\ 1 & \tilde{\lambda}_2 & 1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & \tilde{\lambda}_{n-1} & 1 \\ & & & & 1 & \tilde{\lambda}_n \end{pmatrix}.$$

One can put

$$\gamma_{2k-1} = \prod_{j=1}^{k-1} \frac{w_{2j}}{w_{2j-1}}\,,\ \gamma_{2k} = w_1 \prod_{j=1}^{k-1} \frac{w_{2j+1}}{w_{2j}}\,,\ k = 1, 2, 3, \ldots. \tag{13}$$

Alternatively, the sequence $\{\gamma_k\}_{k=1}^n$ is defined recursively by $\gamma_1 = 1$, $\gamma_{k+1} = w_k/\gamma_k$. Furthermore, $\tilde{\lambda}_k = \lambda_k/\gamma_k^2$. With this choice, (12) is clearly true.

Consequently, the characteristic function of finite symmetric Jacobi matrix J_n can be expressed with the aid of $\mathfrak{F}$.

Proposition 7. *Let $d \in \mathbb{N}$ and $\{\gamma_k\}_{k=1}^d$ be the sequence defined in (13). Then it holds*

$$\det(J_d - zI_d) = \left(\prod_{k=1}^{d}(\lambda_k - z)\right) \mathfrak{F}\left(\frac{\gamma_1^2}{\lambda_1 - z}, \frac{\gamma_2^2}{\lambda_2 - z}, \ldots, \frac{\gamma_d^2}{\lambda_d - z}\right) \tag{14}$$

for all $z \in \mathbb{C}$.

Proof. In view of Remark 6 and Proposition 5, one finds out $\det(J_d - zI_d)$ is equal to

$$\det(G_d) \det \begin{pmatrix} \frac{\lambda_1 - z}{\gamma_1^2} & 1 & & & \\ 1 & \frac{\lambda_2 - z}{\gamma_2^2} & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & \frac{\lambda_{d-1} - z}{\gamma_{d-1}^2} & 1 \\ & & & 1 & \frac{\lambda_d - z}{\gamma_d^2} \end{pmatrix} \det(G_d)$$
$$= \left(\prod_{k=1}^{d}(\lambda_k - z)\right) \mathfrak{F}\left(\frac{\gamma_1^2}{\lambda_1 - z}, \frac{\gamma_2^2}{\lambda_2 - z}, \ldots, \frac{\gamma_d^2}{\lambda_d - z}\right).$$

□

3 Preliminaries

In this section I introduce several preliminary propositions which are necessary for deriving main results of this paper. Respective proofs of these propositions are omitted because they are too extensive and/or too technical.

First, sequence of functions

$$\mathfrak{F}\left(\left\{\frac{\gamma_k^2}{\lambda_k - z}\right\}_{k=1}^{n}\right)$$

converges to function

$$\mathfrak{F}\left(\left\{\frac{\gamma_k^2}{\lambda_k - z}\right\}_{k=1}^{\infty}\right)$$

loccaly uniformly in $z \in \mathbb{C} \setminus \{\lambda_k\}_{k=1}^{\infty}$ with $n \to \infty$. However, one can claim more. The same proposition holds for the first derivative of these functions with respect to z. Hence, it holds

$$\lim_{n\to\infty} \frac{d}{dz}\mathfrak{F}\left(\left\{\frac{\gamma_k^2}{\lambda_k - z}\right\}_{k=1}^{n}\right) = \frac{d}{dz}\mathfrak{F}\left(\left\{\frac{\gamma_k^2}{\lambda_k - z}\right\}_{k=1}^{\infty}\right)$$

and the convergence is local uniform in $z \in \mathbb{C} \setminus \{\lambda_k\}_{k=1}^{\infty}$.

Since

$$\mathfrak{F}\left(\left\{\frac{\gamma_k^2}{\lambda_k - z}\right\}_{k=1}^{n}\right)$$

is the analytic function in z on set $\mathbb{C} \setminus \{\lambda_k\}_{k=1}^{n}$ the limit function

$$\mathfrak{F}\left(\left\{\frac{\gamma_k^2}{\lambda_k - z}\right\}_{k=1}^{\infty}\right)$$

is also the analytic function in z on set $\mathbb{C}\setminus\{\lambda_k\}_{k=1}^{\infty}$. This proposition is based on Theorem 8.8 stated in [5] (where the assumption of uniform convergence is replaced by local uniform convergence). In other words the limit function is meromorphic on $\mathbb{C}$ with poles in the points of $\{\lambda_k\}_{k=1}^{\infty}$. Moreover these poles are of order 1.

Definition 8. Let me denote

$$f_{s,k}^{(n)}(z) := \begin{cases} (z-\lambda_s)\mathfrak{F}\left(\frac{\gamma_k^2}{\lambda_k - z}, \frac{\gamma_{k+1}^2}{\lambda_{k+1}-z}, \ldots, \frac{\gamma_n^2}{\lambda_n - z}\right), & \text{if } z \in H_\epsilon(\lambda_s) \setminus \{\lambda_s\} \\ \lim_{z\to\lambda_s}(z-\lambda_s)\mathfrak{F}\left(\frac{\gamma_k^2}{\lambda_k - z}, \frac{\gamma_{k+1}^2}{\lambda_{k+1}-z}, \ldots, \frac{\gamma_n^2}{\lambda_n - z}\right), & \text{if } z = \lambda_s \end{cases} \tag{15}$$

and

$$f_{s,k}(z) := \begin{cases} (z-\lambda_s)\mathfrak{F}\left(\frac{\gamma_k^2}{\lambda_k - z}, \frac{\gamma_{k+1}^2}{\lambda_{k+1}-z}, \ldots\right), & \text{if } z \in H_\epsilon(\lambda_s) \setminus \{\lambda_s\} \\ \lim_{z\to\lambda_s}(z-\lambda_s)\mathfrak{F}\left(\frac{\gamma_k^2}{\lambda_k - z}, \frac{\gamma_{k+1}^2}{\lambda_{k+1}-z}, \ldots\right), & \text{if } z = \lambda_s \end{cases} \tag{16}$$

where $s, k, n \in \mathbb{N}$, $k \leq n+1$ and $H_\epsilon(\lambda_s) = \{z \in \mathbb{C} : |z - \lambda_s| < \epsilon\}$.

Remark 9. First, it is not important how small ϵ in the previous definition is. One can take for instance $\epsilon = \min\{\lambda_s - \lambda_{s-1}, \lambda_{s+1} - \lambda_s\}$ if $s > 1$ and $\epsilon = \lambda_2 - \lambda_1$ if $s = 1$. I need the functions to be defined only locally. Second, the limit in the previous definition exists and it is finite (use (5) and strict increase of sequence $\{\lambda_k\}$).

Let $s, k \in \mathbb{N}$ then sequence of functions $f_{s,k}^{(n)}(z)$ converges to $f_{s,k}(z)$ uniformly on a neighbourhood of λ_s with $n \to \infty$. However, the same proposition holds for the first derivative with respect to z. Hence, sequence of functions $\frac{d}{dz}f_{s,k}^{(n)}(z)$ converges to $\frac{d}{dz}f_{s,k}(z)$ uniformly on a neighbourhood of λ_s with $n \to \infty$.

3.1 Christoffel-Darboux-like identities

Identities derived below for the function $\mathfrak{F}$ are analogies of the Christoffel-Darboux formula (see [1, Chp. 1] for details).

Definition 10. Let me denote

$$R_k^{(n)}(z) \equiv R_k(z) := \prod_{l=k+1}^{n} \left(\frac{z-\lambda_l}{w_{l-1}} \right) \mathfrak{F}\left(\left\{ \frac{\gamma_l^2}{\lambda_l - z} \right\}_{l=k+1}^{n} \right) \tag{17}$$

where $k, n \in \mathbb{N}$, $0 \leq k \leq n$ (set $w_0 := 1$). Note $R_k^{(n)}$ is a polynomial in z of degree $n-k$.

Recurrence rule (4) rewritten in terms of R_k has a form

$$w_{k-1}R_{k-1}(z) + (\lambda_k - z)R_k(z) + w_k R_{k+1}(z) = 0 \tag{18}$$

which holds for $1 \leq k \leq n-1$. Further the identity

$$(\mu - \lambda)\sum_{k=l}^{n} R_k(\mu)R_k(\lambda) = w_{l-1}(R_l(\lambda)R_{l-1}(\mu) - R_l(\mu)R_{l-1}(\lambda)) \tag{19}$$

holds for all $n \in \mathbb{N}$, $l \in \{1, 2, \ldots, n\}$ and $\lambda, \mu \in \mathbb{C}$. To prove (19) one can proceed by induction in $l = n, n-1, \ldots, 1$. By putting $l = 1$ in (19) and makeing limit $\mu \to \lambda$ one arrives at the identity

$$\sum_{k=1}^{n} (R_k(\lambda))^2 = R_1(\lambda)R_0'(\lambda) - R_0(\lambda)R_1'(\lambda) \tag{20}$$

which holds for all $n \in \mathbb{N}$ and $\lambda \in \mathbb{C}$. Finnaly one can use definititon relation (17) and identity (20) to find out the equation

$$\begin{aligned} &\textstyle\sum_{k=1}^{n} \left(\prod_{l=2}^{k} \left(\frac{z-\lambda_l}{w_{l-1}} \right) \mathfrak{F}\left(\left\{ \frac{\gamma_l^2}{\lambda_l - z} \right\}_{l=k+1}^{n} \right) \right)^2 = \mathfrak{F}\left(\left\{ \frac{\gamma_l^2}{\lambda_l - z} \right\}_{l=1}^{n} \right) \mathfrak{F}\left(\left\{ \frac{\gamma_l^2}{\lambda_l - z} \right\}_{l=2}^{n} \right) \\ &+(z-\lambda_1)\left[\mathfrak{F}\left(\left\{ \frac{\gamma_l^2}{\lambda_l - z} \right\}_{l=2}^{n} \right) \frac{d}{dz}\mathfrak{F}\left(\left\{ \frac{\gamma_l^2}{\lambda_l - z} \right\}_{l=1}^{n} \right) - \mathfrak{F}\left(\left\{ \frac{\gamma_l^2}{\lambda_l - z} \right\}_{l=1}^{n} \right) \frac{d}{dz}\mathfrak{F}\left(\left\{ \frac{\gamma_l^2}{\lambda_l - z} \right\}_{l=2}^{n} \right) \right] \end{aligned} \tag{21}$$

holds for all $n \in \mathbb{N}$ and $z \in \mathbb{C} \setminus \{\lambda_k\}_{k=1}^{n}$. A similar formula can be derived for functions $f_{s,k}^{(n)}$. Thus the identity

$$\begin{aligned} &\sum_{k=1}^{n} \left(\prod_{l=2}^{k} \left(\frac{z-\lambda_l}{w_{l-1}} \right) f_{s,k+1}^{(n)}(z) \right)^2 = f_{s,1}^{(n)}(z) f_{s,2}^{(n)}(z) \\ &+(z-\lambda_1)\left[f_{s,2}^{(n)}(z)\frac{d}{dz}f_{s,1}^{(n)}(z) - f_{s,1}^{(n)}(z)\frac{d}{dz}f_{s,2}^{(n)}(z) \right] \end{aligned} \tag{22}$$

holds for all $s, n \in \mathbb{N}$ and z from a neighbourhood of λ_s.

3.2 Simple zeros

The identities derived in the previous subsection let me prove a proposition concerning a multiplicity of zeros of the function

$$\mathfrak{F}\left(\left\{ \frac{\gamma_k^2}{\lambda_k - z} \right\}_{k=1}^{\infty} \right)$$

considered as a function of variable z. The assumption of strict increase of sequence $\{\lambda_n\}_{n=1}^{\infty}$ is essential. Proofs are only indicated in this subsection.

Proposition 11. *Let $z_0 \in \mathbb{C} \setminus \{\lambda_k\}$ such that*

$$\mathfrak{F}\left(\left\{\frac{\gamma_k^2}{\lambda_k - z_0}\right\}_{k=1}^{\infty}\right) = 0$$

then z_0 is a simple zero.

Proof. One can prove the statement by contradiction. If $z_0 \in \mathbb{C} \setminus \{\lambda_k\}$ is a multiple zero one can send n to infinity in (21) and put $z = z_0$. Consequently, by switching $\lim_{n\to\infty}$ and $\frac{d}{dz}$ (correct due to local uniform convergence discussed at the beginning of section 3), the RHS of (21) vanishes. Then one can arrive at a contradiction. □

A similar property holds for function $f_{s,1}$.

Proposition 12. *Let $s \in \mathbb{N}$ and $f_{s,1}(\lambda_s) = 0$ then λ_s is a simple zero.*

Proof. Similarly as in the previous proof one can prove the statement by contradiction using (22) instead of (21). □

4 Main results

In this section it will be shown the eigenvalues of J coincides with zeroes of a function which will be expressed with the aid of the function $\mathfrak{F}$.

At the beginning let me recall the main result demonstated in [2] where it is prooved the spectrum of J is equal to the set of all limit points of sequences of eigenvalues of truncated finite-dimensional matrices J_n. Thus the equivalence

$$\lambda \in \operatorname{spec}(J) \Leftrightarrow (\exists\{k_n\} \subset \mathbb{N}, k_n < k_{n+1})(\exists\{\tilde{\lambda_n}\} \subset \mathbb{R}, \tilde{\lambda_n} \in \operatorname{spec}(J_{k_n}))(\lim_{n\to\infty} \tilde{\lambda_n} = \lambda) \quad (23)$$

holds.

Proposition 13. *The implication*

$$(\lambda \in \operatorname{spec}(J) \wedge \lambda \notin \{\lambda_k\}_{k=1}^{\infty}) \quad \Longrightarrow \quad \mathfrak{F}\left(\left\{\frac{\gamma_k^2}{\lambda_k - \lambda}\right\}_{k=1}^{\infty}\right) = 0$$

holds.

Proof. Let $\lambda \in \operatorname{spec}(J)$ and $\lambda \notin \{\lambda_k\}_{k=1}^{\infty}$. According to (23) one has a real sequence $\{\tilde{\lambda_n}\}$ such that $\lim_{n\to\infty} \tilde{\lambda_n} = \lambda$ and, by equality (14), it holds

$$\det(J_{k_n} - \tilde{\lambda_n} I_{k_n}) = \prod_{k=1}^{k_n} (\lambda_k - \tilde{\lambda_n}) \mathfrak{F}\left(\left\{\frac{\gamma_k^2}{\lambda_k - \tilde{\lambda_n}}\right\}_{k=1}^{k_n}\right) = 0$$

for all $n \in \mathbb{N}$. Without loss of generality one can assume $\lambda_k \neq \tilde{\lambda_n}$ for all $k, n \in \mathbb{N}$. Hence one gets

$$\mathfrak{F}\left(\left\{\frac{\gamma_k^2}{\lambda_k - \tilde{\lambda_n}}\right\}_{k=1}^{k_n}\right) = 0, \qquad \text{for all } n \in \mathbb{N}.$$

Finnaly, the last equality together with the argument concernig local uniform convergence and analiticity discussed at the beginning of section 3 imply

$$\mathfrak{F}\left(\left\{\frac{\gamma_k^2}{\lambda_k-\lambda}\right\}_{k=1}^{\infty}\right)=0.$$

□

Proposition 14. *Let* $z_0\in\mathbb{C}\setminus\{\lambda_k\}$ *such that*

$$\mathfrak{F}\left(\left\{\frac{\gamma_l^2}{\lambda_l-z_0}\right\}_{l=1}^{\infty}\right)=0$$

then $z_0\in\operatorname{spec}(J)$.

Proof. Since, according to Proposition 11, z_0 is a simple zero, $\mathfrak{F}\left(\left\{\frac{\gamma_l^2}{\lambda_l-z}\right\}_{l=1}^{\infty}\right)$ is continuous in z_0 and $\mathfrak{F}\left(\left\{\frac{\gamma_l^2}{\lambda_l-z}\right\}_{l=1}^{\infty}\right)$ is locally uniformly aproximated by sequence $\mathfrak{F}\left(\left\{\frac{\gamma_l^2}{\lambda_l-z}\right\}_{l=1}^{n}\right)$ in $z=z_0$ there exists a real sequence $\{z_n\}$ such that $\lim_{n\to\infty}z_n=z_0$ and

$$\mathfrak{F}\left(\left\{\frac{\gamma_l^2}{\lambda_l-z_n}\right\}_{l=1}^{n}\right)=0$$

for all n (sufficiently large). Hence $z_n\in\operatorname{spec}(J_n)$ due to identity (14). Finally, equivalence (23) implies $z_0\in\operatorname{spec}(J)$. □

Thus, it was shown that

$$spec(J)\setminus\{\lambda_k\}=\left\{z\in\mathbb{C};\mathfrak{F}\left(\left\{\frac{\gamma_l^2}{\lambda_l-z}\right\}_{l=1}^{\infty}\right)=0\right\}.$$

But what about the points of sequence $\{\lambda_k\}_{k=1}^{\infty}$? These points can be also in the spectrum of J. Further it will be derived a similar condition for them to be in the spectrum.

Proposition 15. *Let* $s\in\mathbb{N}$ *and* $\lambda_s\in\operatorname{spec}(J)$ *then* $f_{s,1}(\lambda_s)=0$.

Proof. According (23) there exists a real sequence $\{\lambda_n^s\}_{n=1}^{\infty}$ such that $\lim_{n\to\infty}\lambda_n^s=\lambda_s$ and $\lambda_n^s\in\operatorname{spec}(J_{k_n})$. Without loss of generality one can assume $\lambda_n^s\neq\lambda_k$ for all $n,k\in\mathbb{N}$, $k\neq s$. By using (14) one has

$$0=\det(J_{k_n}-\lambda_n^sI_{k_n})=\prod_{k=1,k\neq s}^{k_n}(\lambda_k-\lambda_n^s)f_{s,1}^{(k_n)}(\lambda_n^s)$$

for all $n\in\mathbb{N}$. Hence

$$f_{s,1}^{(k_n)}(\lambda_n^s)=0,\qquad\text{for all } n\in\mathbb{N}.$$

Finnaly, by taking into account the local uniform convergence of sequence $f_{s,1}^{(n)}$ (discussed at the beginning of section 3), it suffices to send n to infinity in the last equation to get

$$f_{s,1}(\lambda_s)=0.$$

□

Proposition 16. *Let $s \in \mathbb{N}$ and $f_{s,1}(\lambda_s) = 0$ then $\lambda_s \in \operatorname{spec}(J)$.*

Proof. Since, according to Proposition 12, λ_s is a simple zero, $f_{s,1}$ is continuous in λ_s and $f_{s,1}$ is locally uniformly aproximated by sequence $f_{s,1}^{(n)}$ in λ_s there exists a real sequence $\{z_n^{(s)}\}$ such that $\lim_{n\to\infty} z_n^{(s)} = \lambda_s$ and

$$f_{s,1}^{(n)}(z_n^{(s)}) = 0$$

for all n (sufficiently large). Then $z_n^{(s)} \in \operatorname{spec}(J_n)$ because, by identity (14), one has

$$\det(J_n - z_n^{(s)} I_n) = \prod_{k=1, k\neq s}^{n} (\lambda_k - z_n^{(s)}) f_{s,1}^{(n)}(z_n^{(s)}) = 0.$$

Finally, equivalence (23) implies $\lambda_s \in \operatorname{spec}(J)$. □

5 Summary and example

Let me summarize the main results.

Let $\{w_n\}_{n=1}^{\infty}$ be positive and bouded sequence and $\{\lambda_n\}_{n=1}^{\infty}$ be real and strictly increasing sequence satisfying the condition

$$\sum_{n=1}^{\infty} \frac{1}{\lambda_n^2} < \infty.$$

Then it holds

$$z \in \operatorname{spec}(J) \setminus \{\lambda_n\}_{n=1}^{\infty} \Longleftrightarrow \mathfrak{F}\left(\left\{\frac{\gamma_k^2}{\lambda_k - z}\right\}_{k=1}^{\infty}\right) = 0 \tag{24}$$

and

$$\lambda_s \in \operatorname{spec}(J) \Longleftrightarrow \lim_{z\to\lambda_s} (\lambda_s - z)\mathfrak{F}\left(\left\{\frac{\gamma_k^2}{\lambda_k - z}\right\}_{k=1}^{\infty}\right) = 0. \tag{25}$$

Example 17. Let $\lambda_n = n\lambda$, $\lambda > 0$ and $w_n = w > 0$ for all $n \in \mathbb{N}$. Then

$$J(\lambda, w) = \begin{pmatrix} \lambda & w & & & \\ w & 2\lambda & w & & \\ & w & 3\lambda & w & \\ & & \ddots & \ddots & \ddots \end{pmatrix}.$$

Since $J(\lambda, w) = \lambda J(1, w/\lambda)$ it is sufficient to investigate spectrum of $J \equiv J(1, w)$ without loss of generality. With this choice it holds

$$\gamma_n = \begin{cases} 1, & \text{if } n \text{ is odd} \\ w, & \text{if } n \text{ is even.} \end{cases}$$

Next, it is easy to check

$$\mathfrak{F}\left(\left\{\frac{\gamma_k^2}{k - z}\right\}_{k=1}^{\infty}\right) = \mathfrak{F}\left(\left\{\frac{w}{k - z}\right\}_{k=1}^{\infty}\right).$$

According to (6) one has

$$\mathfrak{F}\left(\left\{\frac{w}{k-z}\right\}_{k=1}^{\infty}\right) = w^z\Gamma(1-z)J_{-z}(2w).$$

Since term $w^z\Gamma(1-z)$ does not effect the zeros of function $\mathfrak{F}\left(\left\{\frac{w}{k-z}\right\}_{k=1}^{\infty}\right)$ and moreover the term $\Gamma(1-z)$ causes the singularities in $z = 1, 2, \ldots$ of function $\mathfrak{F}\left(\left\{\frac{w}{k-z}\right\}_{k=1}^{\infty}\right)$ one can put (24) and (25) together and one arrives at the statement

$$z \in \mathrm{spec}(J) \iff J_{-z}(2w) = 0 \tag{26}$$

or equivalently

$$\mathrm{spec}(J) = \{z \in \mathbb{C};\ J_{-z}(2w) = 0\}. \tag{27}$$

Finnaly one has

$$\mathrm{spec}(J(\lambda, w)) = \left\{\lambda z \in \mathbb{C};\ J_{-z}\left(\frac{2w}{\lambda}\right) = 0\right\}. \tag{28}$$

References

[1] N. I. Akhiezer. *The classical moment problem and some related questions in analysis.* Oliver & Boyd, Edinburgh, (1965).

[2] E. K. Ifantis, C. G. Kokologiannaki, E. Petropoulou. *Limit points of eigenvalues of truncated unbounded tridiagonal operators.* Centr. Europ. J. Math. (2007) 335-344.

[3] V. B. Kuznetsov, E. K. Sklyanin. *Eigenproblem for Jacobi matrices: hypergeometric series solution.* arXiv:math.CO/0509298 v2, (2006)

[4] F. Štampach. *On the eigenvalue problem for a particular class of Jacobi matrices.* In 'SVOČ competetion work (Ostrava, 2010)'.

[5] L. Vrána. *Matematická analýza III - diferenciální počet.* CTU, Prague, (1990).

Clustering via the Distribution Mixtures

Jan Tláskal

2nd year of PGS, email: tlaskjan@fjfi.cvut.cz
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Václav Kůs, Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. The finite distribution mixtures present a wide class of probability distributions. Apart from the obvious application as the distribution estimator of the population with more than one underlying independent phenomena, the mixtures are successfully applied in the model based clustering. If we constraint the members of the mixture to arise from one specific family or type of parametric distributions, each cluster would refer to one component of the mixture. The membership of the observed sample to a cluster is given simply as the maximum probability on the components of the mixture, i.e. by the Mahalanobis distance, and weighted by the weights of the mixture. This approach is feasible even for overlapping clusters and strongly uneven numbers of the members of the clusters, where standard methods of cluster analysis fall short. We provide with an introduction to the distribution mixtures, focusing on the problem of fitting the mixture to observed sample using the maximum likelihood approach and the EM algorithm, as well as the assessment of the optimal number of components.

Keywords: distribution mixtures, model based cluster analysis, order of a parametric model

Abstrakt. Distribuční směsi představují nesmírně širokou třídu distribučních funkcí. Použití distribuční směsi pro odhad neznámých multimodálních rozdělení pravděpodobnosti je přirozené, směsi ale mohou být využity i ve shlukové analýze. Omezímeli komponenty směsi na jeden typ parametrického rozdělení pravděpodobnosti, každý shluk v naměřených datech bude odpovídat jedné komponentě směsi. Příslušnost naměřených hodnot ke komponentě směsi získáme přes maximální věrohodný odhad. Distribuční směs je vhodný model i tam, kde běžné metody shlukové analýzy selhávají, pro překrývající se shluky a pro shluky s významně různými počty prvků.

Klíčová slova: distribuční směs, shluková analýza, řád parametrického modelu

1 Introduction

We understand under the term (finite) distribution mixture a convex combination of distribution functions. This seems like a trivial concept. Nevertheless, if we constraint components of the mixture to be from parametric families of distributions we get a quite general set of distribution functions which retain some of the convenient properties of the parametric models, e.g. that the mixture is fully described with set of scalar parameters.

The distribution mixture is usually employed whenever there is strong evidence of multimodality in the data. Our scope will be the possible use of the distribution mixture as a model based method in cluster analysis, where the clusters are considered to be generated from the distribution mixture components with normal densities. Since each

component of the distribution mixture should ideally refer to one cluster in the data, the problem of choosing an optimal number of components of the mixture becomes quite urgent. As we will see later, there is no definite answer to the problem of optimal number of components and the approach taken depends strongly on the judgement of the experimenter.

2 Clustering via the distribution mixture

In our narrowed scope, we denote the *distribution mixture* any convex combination in a form of

$$p(x) = \sum_{j=1}^{M} \alpha_j p_j(x), \quad \sum_{j=1}^{M} \alpha_j = 1, \;\; \alpha_j \geq 0, \tag{1}$$

where $p_j(x)$ are probability density functions on $\mathbb{R}^D$, α_j are the weight factors, $x \in \mathbb{R}^D$, M is the number of the components. Further, if the components are from the same parametric density family, we denote

$$p(x|\Theta) = \sum_{j=1}^{M} \alpha_j p_j(x|\theta_j), \quad \sum_{j=1}^{M} \alpha_j = 1, \tag{2}$$

where $\theta_j \in \mathbb{R}^s$ is the vector denoting the j-th component of the mixture.

Let $\mathbf{x} = (x_1, \ldots, x_N)$ be sequence of i.i.d. observation of a random variable X which has the density $p(x|\Theta)$. We expect the data to form M clusters according to the components of the mixture. Then, we prescribe to which component of the mixture the observation x_i belongs with vector t_i

$$(t_i)_k = \frac{\alpha_k p_k(x_i|\theta_k)}{\sum_{m=1}^{M} \alpha_m p_m(x_i|\theta_m)}, \quad k \in (1, \ldots, M). \tag{3}$$

The k-th element of t_i evaluates the probability of x_i belonging to the k-th component of the distribution mixture. If we prefer having only a scalar indicator, we define t_i simply as

$$t_i = \arg\max_{k \in M} \alpha_k p_k(x_i|\theta_k). \tag{4}$$

3 Fitting the distribution mixture

Let there be a i.i.d. sequence $\mathbf{x} = (x_1, \ldots, x_N)$ drawn from range of a random variable X having the distribution mixture density $p(x|\Theta)$, $\Theta = (\alpha_1, \ldots, \alpha_{M-1}, \theta_1, \ldots, \theta_M)$. As the distribution mixture is fully described by the vector of parameters Θ, fitting the distribution mixture to the data $\mathbf{x}$ means finding an estimate Θ^* of the real parameter Θ. Our preferred method is the *maximum likelihood method.*

We denote the *log-likelihood function* as $l(\Theta|\mathbf{x}) = \ln p(\mathbf{x}|\Theta)$. The *maximum likelihood estimate* is then a vector Θ^* that satisfies

$$\Theta^* = \arg\max_{\Theta\in\Delta} l(\Theta|\mathbf{x}), \tag{5}$$

where Δ is a domain from which we take the possible candidates for Θ^*. In the case of the same types of components we have

$$l(\Theta|\mathbf{x}) = \ln p(\mathbf{x}|\Theta) \overset{iid}{=} \ln \prod_{i=1}^{N} p(x_i|\Theta) = \sum_{i=1}^{N} \ln \sum_{j=1}^{M} \alpha_j p_j(x_i|\theta_j). \tag{6}$$

In most practical applications, the probability density of the mixture is a differentiable function. To get he candidates for Θ^* we take the partial derivatives of $l(\Theta|\mathbf{x})$ with respect to all elements of Θ and solve the system of equations.

The maximum likelihood method applied on the distribution mixture bear some specific problems:

- The mixture is not *identifiable*. The identifiability property means that for every distinct Θ_1, $\Theta_2 \in \Delta$, $p(x|\Theta_1) \neq p(x|\Theta_2)$ a.s. Since the identifiability property does not hold for the distribution mixture, the sequence of ML estimators does not have to be consistent. Some identifiability problems can be resolved easily, e.g. the component permutation invariance can be solved by ascending sorting of the components, but some cannot be resolved at all. The [2] presents examples when the estimator is stuck within a connected subset of Δ, where the singular Fisher's information matrix is singular.

- The loglikelihood function is not *bounded* even for the simplest mixture, since every observation x_i gives rise to a singularity of the loglikelihood function. Consider a mixture of two heteroscedastic normal densities. If we prescribe the mean of the second component to be exactly equal to the i-th observation and we force the variation of the second component to go to zero, the loglikelihood function will grow beyond all bounds. In the context of the cluster analysis, the unwelcome behavior of the mixture implies that one component fits to a single observation. These solutions are the so called "spurious clusters" solutions and they are disregarded. However, the presence of spurious clusters mean that that the global maximizer of the loglikelihood function does not exist, we have to cope with local maximum solutions. The practitioner usually looks for a sequence of solutions of (5) that seems to be consistent with growing number of observations.

- The analytical work with the term (6) proves to be difficult. Fortunately, there is a simple iterative algorithm which provides the local maximizers of the likelihood function which is called the *Expectation Maximization algorithm*.

4 EM algorithm

Before we can proceed to the definition of the steps of the EM algorithm we need to take one step back and formulate the problem (3), i.e. from which component of the

mixture the i-th observation origins, via the missing information principle. We introduce the random variable $Z = \binom{X}{Y}$, where X are the in the experiment observable data , Y is an unobservable random variable that provides the information from which component of the mixture was the observed X drawn. We refer to Z as to the complete data, X as the observed data, and Y as the missing data. Let $\mathbf{x} = (x_1, \ldots, x_N)$ a i.i.d. random sequence drawn from X, which was observed, and sequence $\mathbf{y} = (y_1, \ldots, y_N)$ from Y which is missing. If $\mathbf{x}$ is i.i.d. then $\mathbf{y}$ is i.i.d. We denote

$$\mathbf{z} = \left((x_1, y_1)^T, \ldots, (x_N, y_N)^T\right), \tag{7}$$

the complete data. The joint distribution density $P(\mathbf{x}|\Theta)$ and the loglikelihood function of the complete data $l_c(\Theta|\mathbf{z})$ can be then expressed in the form

$$P(\mathbf{x}|\Theta) = P(\mathbf{z}|\Theta)/P(\mathbf{y}|\mathbf{x}, \Theta) \tag{8}$$

$$l(\Theta|\mathbf{x}) = l_c(\Theta|\mathbf{z}) - \ln P(\mathbf{y}|\mathbf{x}, \Theta) \tag{9}$$

By taking the conditional expectation of $l_c(\Theta|\mathbf{z})$ with respect to $\mathbf{x}$, $\Phi \in \Delta$ we get

$$E[l(\Theta|\mathbf{x})|\mathbf{x}, \Phi] = E[l_c(\Theta|\mathbf{z})|\mathbf{x}, \Phi] - E[\ln P(\mathbf{y}|\mathbf{x}, \Theta)|\mathbf{x}, \Phi], \tag{10}$$

$$l(\Theta|\mathbf{x}) = E[l_c(\Theta|\mathbf{z})|\mathbf{x}, \Phi] - E[\ln P(\mathbf{y}|\mathbf{x}, \Theta)|\mathbf{x}, \Phi]. \tag{11}$$

We denote the conditional expectation of the complete data loglikelihood function as

$$q(\Theta, \Phi) = E[l_c(\Theta|\mathbf{z})|\mathbf{x}, \Phi]. \tag{12}$$

The k-th interation of the EM algorithm consist of two steps, the expectation and the maximization step:
E-step: Calculate $q(\Theta, \Theta^k)$.
M-step: Maximize $q(\Theta, \Theta^k)$ with respect to first argument Θ,

$$\Theta^{k+1} = \arg\max_{\Theta \in \Delta} q(\Theta, \Theta^k). \tag{13}$$

The iteration of EM algorithm is repeated until convergence. It can be proved [4] that the iteration of the EM algorithm either increases the value of the loglikelihood function $l(\Theta^k|\mathbf{x})$ or Θ^k is already a stationary point $\hat{\Theta}$ of $l(\Theta|\mathbf{x})$, albeit it does not have to be a local minimum.

The advantage of the EM algorithm is that for a few distributions, among which belongs the mixture of normal components, the iteration of the EM algorithm can was solved analytically and the formula to calculate Θ^{k+1} was derived in closed form.

We skip the lengthy derivation of both steps of the EM algorithm for the mixture of normal components and provide only with the final formula to illustrate the how simple the iteration of EM algorithm gets (detailed derivation can be found in [3]). For k-th iteration, $\Theta^k = (\alpha_1^k, \ldots, \alpha_{M-1}^k, \theta_1^k, \ldots, \theta_M^k)$, $\Theta^{k+1} = (\alpha_1^{k+1}, \ldots, \alpha_{M-1}^{k+1}, \theta_1^{k+1}, \ldots, \theta_M^{k+1})$, $\theta_j^k = (\mu_l, \mathbb{C}_l)$, $l = 1, \ldots, M$, we obtain

$$\alpha_l^{k+1} = 1/N \sum_{i=1}^{N} p(l|x_i, \Theta^k), \tag{14}$$

$$\mu_l^{k+1} = \frac{\sum_{i=1}^{N} p(l|x_i, \Theta^k) x_i}{\sum_{i=1}^{N} p(l|x_i, \Theta^k)} \tag{15}$$

$$\mathbb{C}_l^{k+1} = \frac{1}{\sum_{i=1}^{N} p(l|x_i, \Theta^k)} \sum_{i=1}^{N} p(l|x_i, \Theta^k) \mathbb{B}_{i,l} \tag{16}$$

where,

$$p_l(x_i|\theta_l) = \frac{1}{(2\pi)^{D/2}|\mathbb{C}_l|^{1/2}} e^{-\frac{1}{2}(x_i - \mu_l)^T \mathbb{C}_l^{-1}(x_i - \mu_l)},$$
$$\mathbb{B}_{i,l} = (x_i - \mu_l^{k+1})(x_i - \mu_l^{k+1})^T.$$

We now list the most important properties of the EM algorithm:

1. The EM algorithm converges to a stationary point $\hat{\Theta}$ of the loglikelihood $l(\Theta|\mathbf{x})$.
2. The EM algorithm naturally suppresses superfluous components by forcing their weights to zero.
3. The local maximum found by the EM algorithm strongly depends on the initial point of the algorithm Θ^0, since the EM algorithm cannot escape a stationary point once it is reached. We usually initialize the EM algorithm either randomly or with some traditional cluster analysis method and run the EM algorithm several times.
4. The convergence rate gets slow in the later iterations which are close the to local maximum. We usually stop the iterations once $l(\Theta^{k+1}|\mathbf{x})/l(\Theta^k|\mathbf{x}) - 1$ descends under certain $\varepsilon > 0$.

5 The number of components of the mixture

Up until now, we did silently assume that the number M of the components of the distribution mixture is fixed and known, although in many practical problems the number of clusters in the data is unknown. The assessment of the optimal number of components

of the distribution mixture pertains to a more general problem called *the optimal order of a parametric model*.

There is a wide variety of different methods and criteria available in the literature, because there is no definite answer what number of parameters of the model is optimal. The opinions vary in different experiment setups and from author to author. Further, in the context of cluster analysis with the distribution mixture, the optimal number of components that will explain the observed data with sufficient level of precision will probably differ from the optimal number of components needed for discerning the clusters.

Out of all possible methods we consider the penalization criteria of the loglikelihood function the most feasible. All loglikelihood penalization criteria share the same structure. For each possible number of components M, the maximum likelihood estimate Θ_M^* (MLE) is calculated, and then the loglikelihood $l(\Theta_M^*|\mathbf{x})$ is penalized for the complexity of the model. The penalization term is necessary, since the loglikelihood at MLE will rise with increasing M as we maximize over a larger parametric space Δ_M. The optimal number of components M_0 is chosen so it would minimize the loglikelihood at the MLE minus the penalization term, which can be written as

$$M_0 = \arg\min_{M\in\mathbb{N}}[-2l(\Theta_M^*|\mathbf{x}) + 2\xi(N, M, \Theta_M^*, d_M(\Theta_M^*)], \tag{17}$$

where M is the number of components, N is the number of observations, ξ is the penalization term, d_M is the number of free parameters in the mixture described by Θ_M^*.

A different choice of a penalization term ξ yields a different penalization criteria.

We will discuss the four most important criteria used in clustering applications.

1. Akaike's information criterion (AIK)
2. Bayes information criterion (BIC)
3. Entropy criterion (ENC)
4. Integrated classification likelihood (ICL)

Akaike's information criterion (AIC)

The AIC, originally named Another information criterion has become widely used in the regression analysis. The optimal number of parameters of the model, in our case the number of components M_0 is chosen as

$$M_0 = \arg\min_{M\in\mathbb{N}}[-2l(\Theta_M^*|\mathbf{x}) + d_M(\Theta_M^*)]. \tag{18}$$

Bayes information criterion (BIC)

The BIC was published by professor Schwartz in 1979, [5]. The optimum number of components is chosen as

$$M_0 = \arg\min_{M\in\mathbb{N}}[-2l(\Theta_M^*|\mathbf{x}) + d_M(\Theta_M^*)\ln N]. \tag{19}$$

The derivation of the BIC is done by Bayes methods. The proof of consistency of the criterion provided in [5], i.e. that with $N \to \infty$ the BIC chooses asymptotically

the true number of parameters in the model, holds only for the regular exponential family of distributions. Some of the necessary assumptions are broken by the distribution mixtures, thus the BIC is not guaranteed to be a consistent estimator of the true number of components of the mixture. We note that this did not stop the practitioners to use the BIC frequently and the BIC is now a standard criterion applied on distribution mixtures and other criteria are frequently benchmarked to the BIC.

The Entropy criterion (ENC)
As both AIC, BIC criterions are theoretical criteria for assessment of the optimal number of model parameters, the other two we mention, the ENC and ICL are both criteria that was derived with the sole purpose of finding the optimal number of components of the mixture applied in cluster analysis.

We recall the notation (7) of the complete data $\mathbf{z} = ((x_1, y_1)^T, \ldots, (x_N, y_N)^T)$. The component indicators y_i are assumed to be vectors from $\mathbb{R}^M$ with k-th element equal to 1 if the observation x_i comes from the k-th component of the mixture, 0 otherwise. We may write the loglikelihood of the complete data in the form analogical to (9)

$$\ln k(\mathbf{z}|\mathbf{x}, \Theta_M) = l_c(\Theta_M|\mathbf{z}) - l(\Theta_M|\mathbf{x}), \tag{20}$$

$$l(\Theta_M|\mathbf{x}) = \sum_{i=1}^{N} \ln \sum_{k=1}^{M} \alpha_k p_k(x_i|\theta_k), \tag{21}$$

$$l_c(\Theta_M|\mathbf{z}) = \sum_{i=1}^{N} \sum_{k=1}^{M} y_{i,k} \ln \alpha_k p_k(x_i|\theta_k). \tag{22}$$

Further, we get from (21) and z (22)

$$\ln k(\mathbf{z}|\mathbf{x}, \Theta_M) = \sum_{i=1}^{N} \ln \sum_{k=1}^{M} y_{i,k} \ln t_{i,k}, \tag{23}$$

$$t_{i,k} = \frac{\alpha_k p_k(x_i|\theta_k)}{\sum_{j=1}^{M} \alpha_j p_k(x_i|\theta_j)}. \tag{24}$$

The value of $\ln k(\mathbf{z}|\mathbf{x}, \Theta_M)$ is unknown, but we can calculate its negatively taken expectation with respect to the observation $\mathbf{x}$, which we denote $EN(\Theta_M, \mathbf{x})$,

$$EN(\Theta_M, \mathbf{x}) = -E[\ln k(\mathbf{z}|\mathbf{x}, \Theta_M)|\mathbf{x}, \Theta_M] = -\sum_{i=1}^{N} \sum_{k=1}^{M} t_{i,k} \ln t_{i,k}. \tag{25}$$

The term $EN(\Theta_M, \mathbf{x})$ can be interpreted as the entropy of the fuzzy classification matrix with elements $t_{i,k}$, hence the name of the criteria.

Finally we approximate (21) with its expectation and use the MLE Θ_M^* for Θ_M,

$$l_c(\Theta_M^*|\mathbf{z}) \approx l(\Theta_M^*|\mathbf{x}) - EN(\Theta_M^*, \mathbf{x}). \tag{26}$$

The EN criterion is defined to maximize the complete likelihood $l_c(\Theta_M^*|\mathbf{x})$,

$$M_0 = \arg\min_{M\in\mathbb{N}}[-2l(\Theta_M^*|\mathbf{x}) + 2EN(\Theta_M^*, \mathbf{x})]. \tag{27}$$

The entropy $EN(\Theta_M^*, \mathbf{x})$ serves as measure of cluster separation. If the clusters are well separated, the terms $t_{i,k}$ would have one element almost equal 1 and the other would be 0, which leads to almost zero value of $EN(\Theta_M^*, \mathbf{x})$. On the other hand, the entropy reaches a high value for strongly overlapping clusters.

Integrated classification likelihood (ICL)
Although the ENC criterion can be used as it is, it has one significant shortcoming, since it does not penalize the model for complexity. The ICL criterion attempts to overcome the lack of penalization terms, the optimal number of components M_0 satisfies

$$M_0 = \arg\min_{M\in\mathbb{N}}[ICL(\Theta_M^*, \mathbf{x})], \tag{28}$$

$$\begin{aligned} ICL(\Theta_M^*, \mathbf{x}) &= -2l(\Theta_M^*|\mathbf{x}) + 2EN(\Theta_M^*, \mathbf{x}) + \tilde{d}_M \ln N \\ &+ 2N\sum_{k=1}^{M}\alpha_k^* \ln\alpha_k^* - 2\sum_{k=1}^{M}\ln\Gamma(n_k + \frac{1}{2}) \\ &+ 2M\ln\Gamma(\frac{1}{2}) + 2\ln\Gamma(N + \frac{M}{2}) - 2\ln\Gamma(\frac{M}{2}), \end{aligned} \tag{29}$$

where $\tilde{d}_M = d_M - M + 1$, $n_k = \sum_{i=1}^{N} y_{i,k}$, Γ is the gamma function. Naturally, we approximate the gamma function with Stirling formula

$$\ln\Gamma(u) = (u - \frac{1}{2})\ln u - u + \frac{1}{2}\ln 2\pi. \tag{30}$$

The main idea behind the ICL criterion is straightforward, the ICL is taken as the posterior mode of the complete loglikelihood $l_c(\Theta_M^*|\mathbf{x})$, where the prior $p(\Theta_M)$ is assumed to be from the Dirichlet distribution family. The derivation itself is quite complex and beyond the scope of this paper, the reader is referred to [6] for the detailed derivation.

6 Discussion

The distribution mixtures are a powerful tool in statistics but seemingly underused. They are employed the most successfully whenever there is a strong evidence that the observed data have multiple modes, where the parametric models are bound to fall short. They are considered a compromise between simple parametric and general nonparametric methods. For more applications of the distribution mixture see [1].

The estimate of the optimal number number of components of the mixture is crucial in the cluster analysis. We provided four likelihood penalization based criteria, the AIC, BIC, NEC and ICL. Our experiments with the data from the acoustic emission confirm the general opinion within the model based clustering community that the AIC and BIC

seem to overestimate the number of clusters, whereas the ICL usually chooses the best model.

There are other methods frequently used to assess the right number of the components of the mixture, for example the likelihood ratio test (LRT) between each two possible candidates for M_0. The LRT is generally applicable, since the compared models are nested, i.e. that if $M_0 < \widetilde{M_0}$, the corresponding parametric spaces satisfy $\Delta_M \subset \Delta_{\widetilde{M}}$. The problem we see with the LRT is that since we do not have the asymptotic normality of the MLE for the mixtures, therefore the asymptotic distribution of likelihood ratio is not guaranteed to be chi-squared and it has to be sampled via bootstrapping methods.

References

[1] G. McLachlan, D. Peel. *Finite Mixture Models*. J. Wiley, New York (2000).

[2] G. McLachlan, T. Krishnan. *The EM algorithm and Extensions*. J. Wiley, New York (1997).

[3] J. Tláskal. *Statistical Classification Methods in Accoustic Emission*. Diploma thesis (MSc) at FNSPE CTU in Prague, (2009).

[4] A.P. Dempster, N.M. Laird, D.B. Rubin. *Maximum likelihood from incomplete data via the EM algorithm (with discussion).* Journal of the Royal Statistical Society, **B 39** (1995), 1-38.

[5] G. Schwartz. *Estimating the dimension of a model.* Annals of statistics, **6** (1978), 461-466.

[6] C. Biernacki, G. Celeux, G. Covaert. *Assesing a mixture model for clustering with the integrated classification likelihood.* Technical Report No 3521. Rhône-Alpes: INRIA, (1998).

Využití neuronových sítí k modelování úrokových měr stanovených ČNB*

Tran Van Quang

2. ročník PGS, email: tran@vse.cz
Katedra softwarového inženýrství v ekonomii
Fakulta jaderná a fyzikálně inženýrská, ČVUT v Praze

školitel: Jaromír Kukal, Katedra softwarového inženýrství v ekonomii, Fakulta jaderná a fyzikálně inženýrská, ČVUT v Praze

Abstract. Knowledgeability about interest rates set by a central bank is very important for all participants in an economy. In this paper we have used publicly available data to model the behavior of Czech National Bank in terms of how they manipulated interest rates when conducted the monetary policy recently. Using artificial neural networks with different architecture, namely the multilayer perceptron and radial basis function types, and with the different number of hidden neurons, we modeled the discount rates of CNB on a set of macroeconomic variables including inflation rate (CPI), exchange rate of CZK to Euro, interest rate PRIBOR 3M, unemployment rate, the economic growth rate, the growth rate of money aggregate M2 and interest rate set by European Central Bank. Having also compared the results obtained from modeling by these artificial neural networks with the results from a linear model run on the same set of regressors, we found that in terms of predicted errors, the neural network modeling gives superior results over the ones from the linear model.

Keywords: linear model, artificial neural network, MLP and RBF networks, CNB short-term interest rate

Abstrakt. Znalost úrokových měr stanovených centrální bankou je velmi důležitá pro všechny účastníky dané ekonomiky. V tomto příspěvku jsem používal veřejně dostupná data k modelování chování České národní banky při výkonu měnové politiky prostřednictvím diskontních úrokových sazeb. Byly použity neuronové sítě s různou architekturou, konkrétně vícevrstvé sítě perceptron (MLP) a sítě s radiální bází (RBF), a s různým počtem skrytých neuronů a diskontní sazby ČNB byly modelovány jako funkce makroekonomických proměnných jako míry inflace, měnového kurzu české koruny vůči Euru, úrokové míry PRIBOR 3M, míry nezaměstnanosti tempa růstu HDP, míry růstu peněžního agregátu M2 a úrokové míry stanovené Evropské centrální banky. Výsledky modelování úrokových měr ČNB neuronovými sítěmi byly potom porovnány s výsledky získanými lineárním modelem. Ukázalo se, že neuronové sítě poskytují mnohem přesnější výsledky měřené střední chybou modelu.

Klíčová slova: lineární model, umělé neuronové sítě, MLP a RBF sítě, diskontní sazba ČNB

*Tato práce byla podpořena grantem číslo SGS10/092/OHK4/1T/14 Ministerstva školství, mládeže a tělovýchovy České republiky.

1 Úvod

V posledních několika desetiletích se objevil nový nástroj pro zpracování dat zvaný umělá neuronová síť. Jak už jeho názvu vyplývá, neuronové sítě se snaží napodobovat chování skutečných biologických neuronů v lidském mozku a skládají se ze sítí neuronů propojených synapsemi předávajícími výstupy z jednoho neuronu na druhý s výjimkou vstupních a výstupních neuronů. Díky své struktuře jsou neuronové sítě velmi flexibilní a jsou schopné zachytit i složité nelineární závislosti mezi vysvětlujícími a vysvětlovanými proměnnými, což jedna z předností neuronových sítí při modelování dat. Další výhodou oproti tradičním modelům je to, že výsledky modelů založených na umělých neuronových sítích (ANN) jsou ovlivněny hierarchickou strukturou těchto sítí. Proto je zajímavé studovat vliv typu neuronové sítě, počtu skrytých neuronů a počtu identifikovaných parametrů (vah ANN) na dosaženou přesnost modelování. Tyto přednosti neuronových sítí budou ověřeny při modelování chování České národní banky.
Centrální banky ve vyspělých tržních ekonomikách v poslední době stále nejčastěji používají krátkodobou úrokovou sazbu pro realizaci měnové politiky. Snaží se prokázat svou důvěruhodnost transparentností, svoje rozhodnutí o úrokových sazbách zakládá na objektivních indikátorech, jejichž vliv na ekonomické prostředí je vnímán různými subjekty v ekonomice. Protože změny úrokových sazeb ovlivňují všechny aktivity domácího bankovního sektoru a stejně tak i veškeré operace na kapitálovém trhu včetně toku kapitálu ze zahraničí, zmíněné subjekty se snaží odhadnout s předstihem, jak bude centrální banka zacházet se svými úrokovými nástroji. Znalost výše úrokových měr stanovených centrální bankou je velmi důležitá pro všechny subjekty v dané ekonomice.
Uvedená skutečnost je dostatečně silný motiv k modelování chování České národní banky (ČNB) na základě veřejně přístupných informací. K tomu účelu slouží různé typy neuronových sítí, jejich výstupy lze následně využít k jednokrokové predikci její změny úrokové sazby. Jde tedy o externí modelování chování ČNB v diskrétním čase, jehož výstupem je měsíční relativní změna úrokové sazby. Dále bude zkoumán vliv architektury námi zvolených neuronových sítí na přesnost našeho modelu. Tento příspěvek je strukturován následovně: v následující části budu charakterizovat, jak ČNB provádí svou měnovou politiku. Pak vymezím modely, které budou použity pro modelování chování ČNB. V další části podrobně analyzuji kritéria pro hodnocení kvality jednotlivých modelů. Výsledky modelování chování ČNB budou uvedeny v navazující části článku a v závěru bude celkové vyhodnocení.

2 Umělá neuronová síť jako nástroj modelování

Umělá neuronová síť (ANN) je jedním z nástrojů umělé inteligence (AI), o kterém lze rovněž tvrdit, že je z pohledu matematické statistiky též modelem, jehož parametry je třeba odhadnout například metodami nelineární regrese (Haykin, 1999, Šnorek a Jiřina, 1997). Z hlediska teorie systémů ve stejné souvislosti hovoříme opět o modelu, jehož parametry se adaptují, jsou předmětem optimalizace nebo je třeba je identifikovat. V analogii s lineárním modelem je vhodné problematiku ANN zúžit pouze na neuronové sítě s jedním výstupem (predikovaná relativní změna) několika vstupy (relativní změny ekonomických ukazatelů) a s dopředným šířením signálů. Tím se vyhneme rekurentním výpočtům a

cyklům v topologii sítě. Každá taková ANN je reprezentována acyklickým orientovaným grafem, jehož vrcholy lze formálně očíslovat tak, aby signál postupoval ve směru vzrůstajícího indexu vrcholu. V případě ANN vrchol nazýváme umělým neuronem, což není nic jiného než procesor pro zpracování informace z do něj vstupujících hran. Každá vstupní hrana nese dvě zásadní informace: svou váhu a hodnotu signálu z předchozího vrcholu (neuronu). Vrcholy, do kterých nevede žádná hrana, nazýváme vstupní neurony neboli vstupy systému. Vrcholy, ze kterých nevede žádná hrana, nazýváme výstupní neurony neboli výstupy systému. V našem případě máme pouze jeden výstupní neuron. Zbylé vrcholy slouží jako mezičlánky přenosu informace a nazýváme je skrytými neurony. K transformaci hodnot signálů dochází tedy pouze ve skrytých as výstupních neuronech. Každý takový umělý neuron má svůj model, tj. funkci, která zpracuje hodnoty všech vah a signálů do něj vstupujících a vytvoří z nich nový signál. Problematika ANN se tedy týká pouze tří aspektů: topologie grafu sítě, modelů jednotlivých neuronů a nastavení (určení, učení, odhad, adaptace, optimalizace, identifikace) vah jednotlivých hran orientovaného grafu.

Nejčastější je hierarchické uspořádání ANN do jednotlivých vrstev. Vrstva ANN je skupina neuronů (vrcholů), které spolu přímo ani nepřímo nekomunikují. Komunikace je dále omezena tak, že do neuronů dané vrstvy vedou signály pouze z předchozí vrstvy. Z toho okamžitě plyne, že pořadí vrstev, které musí být respektováno. Pak hovoříme o první (vstupní) vrstvě tvořené všemi vstupními neurony, druhé, třetí skryté vrstvě a konečně o poslední (výstupní) vrstvě) obsahující všechny výstupní neurony. Jednoduché úlohy lze řešit pomocí dvouvrstvé topologie ANN, kdy zcela chybí skryté vrstvy. Většinu aplikací lze zvládnout pomocí třívrstvé topologie s optimálním počtem skrytých neuronů v jediné skryté vrstvě. Z hlediska modelů jednotlivých umělých neuronů již dávno nejde o napodobování chování jednotlivých nervových buněk živých organizmů. Jde o modely, které se v oblasti AI pragmaticky osvědčily při konstrukci univerzálních aproximací spojitých funkcí. Do první skupiny patří modely založené na skalárním součinu vektoru vah a vektoru vstupujících hodnot. Takto vzniklý vážený součet (lineární kombinace signálů) je většinou modifikován vhodnou nelinearitou. Oblíbená nelinearita tohoto typu má tvar sigmoidální funkce (hyperbolické tangenty). Do druhé skupiny patří modely založené na rozdílu vektoru vah a vektoru vstupujících hodnot a jeho normě (délka rozdílového vektoru). Pokud vektor vah chápeme jako popis etalonu (vzorové situace), pak tímto postupem určíme vzdálenost vstupujícího vektoru (reality) od etalonu (prototypu). Výsledná vzdálenost je pak opět zpracována vhodnou nelinearitou. Výstupní signál takového neuronu obvykle chápeme jako informaci o podobnosti reality a etalonu. Vzdálenost většinou měříme euklidovskou normou a nelinearita má obvykle tvar Gaussovy zvonovité funkce. Do třetí skupiny obvykle řadíme modely umělého neuronu produkované s využitím operátorů fuzzy logiky. V této souvislosti se vyskytuje pojem fuzzy-neuronová síť, což není nic jiného, než ANN, ve které existuje neuron z uvedené třetí skupiny.

V drtivé většině reálných aplikací se setkáváme s ANN obsahující pouze jednu homogenní skrytou vrstvu a jednu homogenní výstupní vrstvu. Pokud jsou ve výstupní vrstvě pouze lineární modely neuronu, pak to má výhody při nastavování vah mezi skrytou a výstupní vrstvou metodami lineární algebry. Pokud jsou ve skryté vrstvě pouze neurony prvního typu se sigmoidální charakteristikou, pak hovoříme o třívrstvém či obecně vícevrstvém perceptronu (MLP). Pokud jsou ve skryté vrstvě pouze neurony druhého typu s eukli-

dovskou normou a výstupní vrstva je lineární, pak hovoříme o síti s radiální bází (RBF). Samostatnou kapitolu tvoří metody učení ANN, což z hlediska matematické statistiky není nic jiného než určení bodového odhadu parametrů celého modelu ANN na základě reálných dat. Nejčastěji se setkáváme s učením MLP či RBF gradientními metodami (metoda konjugovaných gradientů, stochastická gradientní metoda, zpětné šíření neboli backpropagation. Při něm vycházíme z náhodného odhadu vah modelu a iteračně získáme jeden z lokálních extrémů součtu čtverců odchylek chování modelu ANN od požadovaných hodnot. K tomuto účelu slouží známá gradientní metoda vycházejicí z různých počátečních hodnot vah ANN. Alternativou k takovému postupu je využití heuristik pro hledání globálního minima součtu čtverců odchylek, např. simulované žíhání (SA, FSA, ASA), modely chování kolonií (ACO), modely migrace jedinců (SOMA), modely chování hejn (PSO), genetické algoritmy optimalizace (GO), evoluční vyhledávání (ES) a diferenciální evoluce (DE).

3 Použité modely ANN

V tomto příspěvku jsou využity dva modely s jednou skrytou vrstvou a jedním výstupním neuronem (třívrstvá ANN). První z nich je třívrstvá perceptronová síť s lineárním výstupním neuronem, o které je známo, že je univerzálním aproximátorem na třídě spojitých omezených funkcí. To znamená, že s rostoucím počtem skrytých neuronů klesá chyba aproximace k nule, nikoli však ve statistickém slova smyslu. Výstup uvedené sítě je dán vztahem

$$y = v_0 + \sum_{k=1}^{H} v_k \tanh(w_{k,0} + \sum_{j=1}^{n} w_{k,j} x_j) \tag{1}$$

kde H je zadaný počet skrytých neuronů, $w_{1,0}, ..., w_{H,n}$ jsou neznámé váhy jednotlivých vstupů a $w_0, ..., w_H$ jsou neznámé výstupní váhy. Celkový počet parametrů uvedeného nelineárního modelu je roven $np = H(n+2)+1$. Pokud preferujeme sítě s radiální bází, pak dáme přednost třívrstvé neuronové síti RBF, která je rovněž univerzálním aproximátorem spojitých ohraničených funkcí. Výstup uvedené sítě je dán vztahem

$$y = v_0 + \sum_{k=1}^{H} v_k \exp(-\frac{\sum_{j=1}^{n}(x_j - w_{j,k})^2}{2w_{k,0}^2}) \tag{2}$$

Celkový počet parametrů uvedeného nelineárního modelu je opět roven $np = H(n+2)+1$.

Kritéria hodnocení kvality modelu

Chceme-li vzájemně porovnávat lineární model a nelineární modely ANN, musíme stanovit srovnatelné podmínky pro oba typy modelů. Vzhledem k tomu, že lineární regrese je standardním nástrojem odhady parametrů lineárního modelu, je přirozené odhadovat váhy (neznámé parametry) sítí MLP a RBF stejnou metodou, tedy minimalizací součtu čtverců odchylek, což je mnohem náročnější úloha vzhledem k nelinearitě příslušných modelů. Příslušná účelová funkce, jejíž minimum je hledáno, je

$$SSQ_{rel} = \sum_{i=1}^{m} (\rho_i - \rho_i^{ANN})^2 \tag{3}$$

kde ρ_i, ρ_i^{ANN} je relativní přírůstek úrokové sazby a jeho predikce na výstupu ANN. V případě lineární ANN je známo explicitní řešení s využitím lineární algebry. Pokud je úloha nelineární, osvědčily se k jejímu řešení kompetitivní heuristiky diferenciální evoluce. V tom případě je třeba si uvědomit, že není zajištěno nalezení globálního optima SSQ_{rel}, neboť učení ANN je NP-úplný problém, takže se nezávisle na metodě hledání neznámých vah je nutné uchýlit k opakovaným numerickým experimentům s danou heuristikou a tak získatjistotu přesnějších výsledků. Jednotícím kritériem kvality modelu pro lineární model i ANN je střední chyba predikce relativní diference úrokové sazby ČNB daná vztahem

$$S.E. = \sqrt{\frac{SSQ_{rel}}{m - np}} \tag{4}$$

Vzhledem k možnosti permutace neuronů ve skryté vrstvě ANN není možné posuzovat statistickou významnost jednotlivých vah ANN respektive stanovovat jejich směrodatnou odchylku.

4 Měnová politika České národní banky

V tržním hospodářství hraje centrální banka nezastupitelnou úlohu v bankovním systému. Kromě jiných funkcí se jí výhradně svěřuje výkon měnové politiky státu. Měnová politika je proces, kdy centrální banka využívá různé nástroje k dosažení určitých cílů, jejich realizace přispívá k stabilizaci a růstu národní ekonomiky. Tyto cíle jsou cenová stabilita, vysoká zaměstnanost, ekonomický růst, stabilita úrokové míry a měnového kurzu a v neposlední řadě i stabilita finančního trhu. Je patrné, že tyto cíle nejsou od sebe dělitelné a dosažení jednoho z nich může být realizováno na úkor druhých. K dosažení těchto cílů má centrální banka několik nástrojů, které jsou tržní i netržní, tedy regulatorní povahy. Protože regulační podmínky a opatření ve vyspělých ekonomikách jsou standardní, v současné době centrální banky ve vyspělých tržních ekonomikách inklinují k použití tržních nástrojů jako operace na volném trhu nebo úrokových nástrojů k dosažení cílů své měnové politiky. Česká národní banka (ČNB) podle Ústavy České republiky a zákona o České národní bance (zákon č. 6/1993 Sb.) je centrální bankou České republiky. Kromě emisních pravomocí a dohledu nad finančním sektorem operujícím na českém území Česká národní banka vykonává měnovou politiku, jejímž hlavním cílem je zabezpečit cenovou stabilitu v České republice. ČNB také podporuje jiné cíle hospodářské politiky vlády České republiky jako ekonomický růst, nízkou nezaměstnanost a vnitřní a vnější stabilitu ekonomiky, pokud tyto cíle nejsou v konfliktu s hlavním cílem ČNB. Jako cenová stabilita se rozumí stabilitou spotřebitelských cen měřenou indexem spotřebitelských cen (CPI) poskytovaným Českým statistickým úřadem. Je třeba podotknout, že jako cenovou stabilitu ČNB nepočítá s absolutní stabilitou, tedy nulovou inflaci, nýbrž kalkuluje s mírně kladnou inflací, jež mimo jiné zohledňuje pozitivní změny v kvalitě nových zboží a služeb poskytovaných spotřebitelům, které docházejí průběžně v ekonomické realitě. K dosažení svých cílů používá ČNB v současné době tzv. režim cílování inflace, který centrální banky vyspělých tržních ekonomik začaly rozšířeně aplikovat od konce devadesátých let minulého století (Mandel a Tomšík, 2008). Cílování inflace je takový proces, při němž centrální banka odhaduje a zveřejňuje cílované hodnoty míry inflace v budoucnu a provádí takovou

měnovou politiku prostřednictvím změn úrokových sazeb a jiných měnových nástrojů, aby se skutečná inflace v budoucnu co nejvíce přiblížila k cílované hodnotě. Např. centrální banka si stanovuje cíl, že inflace v ekonomice bude 2% v budoucnu. Při plnění svého cíle centrální banka v tuto chvíli pozoruje různé, které mohou potenciálně vyvinout silný tlak na zvýšení inflace v budoucnu, a to jak ze strany nabídky, tak i ze strany poptávky. Chce-li centrální banka toto nebezpečí eliminovat a svůj inflační cíl splnit, tak musí zvýšit nominální úrokovou míru. Nicméně zvýšení úrokové míry vede ke zdražení peněz v ekonomice, což negativně ovlivňuje ekonomickou aktivitu různých subjektů v ekonomice v podobě zpomalení ekonomického růstu nebo zvýšení nezaměstnanosti. Výhoda měnové politiky v režimu cílování inflace spočívá v tom, že pokud centrální banka (v našem případě ČNB) je dostatečně kredibilní a dovede si svůj inflační cíl splnit správnou měnou politikou, získá tím důvěru veřejnosti. Ta na oplátku zabuduje do svého očekávání cílovanou hodnotu inflace deklarovanou centrální bankou a tím se vytvoří inflačně stabilní prostředí v ekonomice. Podmínkou k prosazení takové měnové politiky je nezávislost centrální banky na vládě a ČNB ze zákona tuto nezávislost má. V tuto chvíli ČNB vyhlašuje inflační cíl ve výši 2% s tím, že se skutečná inflace nebude lišit od této hodnoty více než 1%.

5 Modely chování ČNB

Chování ČNB je zkoumána v diskrétním čase. Navíc pro modelování a predikce není je třeba vhodná selekce ukazatelů a a jejich transformace. Z tohoto důvodu je modelována relativní změna úrokové sazby pomocí relativních změn makroekonomických ukazatelů v předchozím období. Jelikož Česká národní banka mění úrokové sazby nepravidelně, bylo také nutné tyto nepravidelné změny převedeny na data s měsíční periodou. Ostatní makroekonomické ukazatele je dostupné již v měsíčních periodách. Pouze data o ekonomickém růstu jsou přepočítány na tuto periodu lineární interpolací. Pro všechny studované veličiny jsou tak k dispozici časové řady s periodou jeden měsíc. Pokud budeme studovat časovou řadu $\{\xi\}_{k=1}^{n}$ tvořenou kladnými hodnotami, pak při modelování můžeme využít absolutní diference $\delta_k = \xi_{k-1} - \xi_k$ nebo relativní diference $\rho_k = 100\frac{\xi_{k-1}-\xi_k}{\xi_k}$. Absolutní diference se užívají v souvislosti s lineárními modely. Pro snadnou aplikaci ANN je třeba použít relativní změny (v %) s tím, že některé makroekonomické ukazatele již uvedený relativní tvar mají. Úlohu o jednokrokové predikci v časové řadě relativních diferencí úrokové sazby je třeba konvertovat na úlohu o lineární nebo nelineární regresi. Zavedeme-li dvě úrovně indexování relativních diferencí tak, že první index odpovídá časovému kroku a druhý index sledované veličině (úroková sazba ČNB má index roven jedné), pak jednokroková predikce je dána vztahem $\rho_{k+1,1} = f(\rho_{k,1}, \rho_{k,n})$, kde n je počet sledovaných veličin a $y = f(x)$ je příslušný model. Z časových řad délky M tak dostaneme řady relativních diferencí délky $M - 1$, které umožní realizaci statistického výběru o rozsahu $m = M - 2$. Nyní se můžeme věnovat jednotlivým modelům. V prvé řadě byl studován lineární model ve tvaru

$$y = w_0 + \sum_{k=1}^{n} w_k x_k \tag{5}$$

kde jsou neznámé váhy jednotlivých vstupů. S využitím metodiky lineární regrese snadno určíme nejen bodové odhady parametrů (vah) a chybu modelu, ale též jejich směrodatné odchylky parametrů a příslušné p-hodnoty (testování hypotéz o nulovosti parametrů). Uvedený lineární model můžeme chápat jako referenční ANN bez skryté vrstvy s jedním lineárním neuronem ve výstupní vrstvě. Celkový počet parametrů uvedeného modelu je roven $np = n + 1$. Cílem studie je porovnat chování takového lineárního systému s nelineárními modely reprezentovanými rovněž ANN.

6 Použitá data a výsledky odhadů parametrů modelů

Pro modelování chování České národní banky jsou vybrány tyto řady: jako úroková sazba je zvolena diskontní sazba ČNB. Měřítkem inflace je index spotřebitelské ceny. Indikátorem vnější stability české měny je průměrný měsíční kurz koruny vůči společné evropské měně Euru. Dále jsou to průměrná měsíční úroková sazba na mezibankovním trhu na dobu tří měsíců Pribor3M, měsíční tempo růstu peněžního agregátu M2, měsíční míra nezaměstnanosti, čtvrtletní tempo růstu HDP v České republice a diskontní sazba Evropské centrální banky, a to v období od 1.1.1999 do 1.9.2008. Zdrojem údajů o diskontní sazbě, kurzu české koruny vůči euru, mezibankovní úrokové sazby a měsíčního tempa růstu peněžního agregátu M2 je statistika České národní banky. Údaje o inflaci a tempu růstu HDP jsou získány ze statistiky Statistického úřadu ČR. Údaje o nezaměstnanosti jsou získány ze statistiky Ministerstva práce a sociálních věcí. Údaje o diskontní sazbě ECB jsou z její statistiky. Protože všechny proměnné zahrnuté v modelu nemají stejného měřítko ani stejnou povahu, je třeba přistoupit k transformaci některých z nich. Údaje o inflaci (CPI), tempu růstu HDP a tempu růstu měnového agregátu M2 mají již požadovaný charakter, mohou vstoupit do modelu beze změny. Ostatní proměnné, i když některé z nich jsou udávané v procentech jako úrokové sazby nebo míra nezaměstnanosti, jsou převedeny na nejdříve na přírůstkovou formu, pak i na relativně přírůstkový tvar. Je patrné, že nejlepší transformace je ta, která převádí všechny proměnné jsou ve stejných měřítkách a stejné povahy, které navíc jsou symetrické kolem jejich průměrů. Touto úpravou se počet pozorování snižuje z 117 na 116. Relativní přírůstky diskontní úrokové sazby ČNB jsou pak modelovány na hodnotách vybraných proměnných zpožděných o jedno období včetně samotných relativních přírůstků diskontní úrokové sazby. Závislost relativních přírůstků diskontní úrokové míry České národní banky na vybraných proměnných je nejdříve modelována metodou nejmenších čtverců a výsledné odhady vlivů jednotlivých proměnných jsou uvedeny v tabulce 1, kde kromě samotných odhadů jsou uvedeny také standardní chyby jednotlivých odhadů a jejich p-value, tj. hladina, od které se zamítá nulová hypotéza, že hodnota odhadovaného parametru je nula. Podle p-value jsou statisticky významné odhady zpožděné diskontní úrokové sazby ČNB, zpožděné úrokové sazby na mezibankovním trhu Pribor 3M a zpožděné úrokové sazby Evropské centrální banky. Významnost těchto proměnných je odůvodnitelná, protože rozhodnutí ČNB o změně úrokové sazby musí vycházet z jejích minulých uzávěrů a mezibankovní úroková sazba je odvozena od úrokové sazby stanovené ČNB a ČNB při svém rozhodování musí brát v úvahu situaci v nejbližším a také nejdůležitějším okolí jejího měnového prostoru. Odhady parametrů ostatních proměnných obsažených v modelu jsou statisticky nevýznamné a nelze tedy odmítnout nulovou hypotézu, že hodnoty těchto parametrů

Tabulka 1: Výsledky odhadu lineárního modelu chování ČNB

regresor	w	S.E.	p-value
Konstanta	-4,2310	2,4901	0,0922
Úroková sazba ČNB	-0,3359	0,1105	0,0030
Inflace ČR (CPI)	0,0274	0,4548	0,9521
Průměrný měsíční kurz EUR/CZ	-0,0660	0,6275	0,9165
IR Pribor 3M	1,0911	0,2116	0,0000
Tempo růstu M2	0,2815	0,2594	0,2804
Měsíční míra nezaměstnanosti	0,0646	0,2181	0,7676
Úroková sazba ECB	0,2706	0,0944	0,0050
Tempo růstu HDP ČR (čtvrtletní)	0,2277	0,4573	0,6197

Tabulka 2: Vliv počtu neuronů na chybu MPL a RBF modelu

počet neuronů	2	3	4	5
MPL	6,60162	5,85499	6,34161	6,71695
RBF	6,56605	6,06271	5,34662	5,68335

mohou být nulové. Tato skutečnost může být způsobena tím, že tyto proměnné mohou být silně korelované. Například míra nezaměstnanosti může být ovlivněna tempem růstu HDP a naopak při poklesu tempa růstu HDP může dojít k růstu nezaměstnanosti. Chyba lineárního modelu je se = 7,4206. Porovnáváme-li ji se standardní odchylkou řady relativních přírůstků diskontní úrokové míry České národní banky (s = 8,6550), je patrné, že jen poměrně malá část její variability je vysvětlena tímto modelem. Závislost relativních přírůstků diskontní úrokové míry ČNB na vybraných proměnných je také modelována vícevrstvou neuronovou sítí (MLP – multilayer perceptron). Tato síť se skládá z jedné vrstvy vstupních neuronů, pak ji následuje skrytá vrstva a na konci této sítě je výstupní vrstva, ze které vychází výstup jako lineární kombinace výstupů ze skryté vrstvy. Námi zkoumaná závislost je modelována s různým počtem neuronů ve skryté vrstvě. V tabulce 2 jsou uvedeny výsledky modelování závislosti relativních přírůstků diskontní úrokové míry České národní banky na vybraných proměnných prezentované chybami modelu v závislosti na počtu skrytých neuronů. Počet neuronů v poslední vrstvě je vždy o jeden neuron více než ve skryté vrstvě. Je vidět, že nejmenší chybu (5,855) poskytuje neuronová síť se třemi skrytými neurony. Pro tuto síť jsou uvedeny optimální váhy jednotlivých vstupních proměnných ve skryté vrstvě (tabulka 3) a váhy výstupů ze skrytých neuronů do lineární výstupní vrstvy (tabulka 4). Porovnáváme-li chybu modelu neuronové sítě s chybou lineárního modelu, zjistíme, že variabilita relativních přírůstků diskontní úrokové míry České národní banky na vybraných proměnných je vysvětlena mnohem více prostřednictvím vícevrstvé neuronové sítě než lineárním modelem. Závislost relativních přírůstků diskontní úrokové míry České národní banky na vybraných proměnných je dále modelována sítí RBF neuronů. Typologie této sítě je podobná vícevrstvé síti. Skládá se z vrstvy vstupních neuronů, dále je vrstva RBF neuronů a poslední je výstupní vrstva, ze které vychází výstup jako lineární kombinace výstupů ze skryté vrstvy. Závislost je také modelována s různým počtem RBF neuronů ve skryté vrstvě. V tabulce 2 jsou uvedeny výsledky

Tabulka 3: Optimální váhy jednotlivých vstupů ve skryté vrstvě MLP sítě

váha	1	2	3
Konstanta	0,927236	-0,787776	-0,937465
Sazba ČNB	-0,095104	0,082983	0,096009
Inflace ČR (CPI)	-0,058158	0,020444	0,061885
Průměrný měsíční kurz EUR/CZ	-0,050555	-0,099868	0,069388
IR Pribor 3M	0,026756	-0,074350	-0,018224
Tempo růstu M2	-0,080787	0,096671	0,078897
Měsíční míra nezaměstnanosti	-0,084985	0,046552	0,089975
Sazba ECB	0,068586	-0,050390	-0,070029
Tempo růstu HDP ČR (čtvrtletní)	-0,075676	-0,028200	0,087045

Tabulka 4: Váhy ve výstupní vrstvě optimální MLP a RBF sítě

	váha	
č. neuronů	MPL	RBF
0	-0,00180	0,000000000002
1	7,49334	-0,000000000181
2	0,75171	-0,000000071741
3	6,79324	1,606886617377
4	-	0,000000000002

modelování závislosti relativních přírůstků diskontní úrokové míry České národní banky na vybraných proměnných prezentované chybami modelu v závislosti na počtu skrytých RBF neuronů. Je vidět, že nejmenší chybu (5,347) poskytuje RBF neuronová síť se čtyřmi skrytými neurony. Pro tuto síť jsou také uvedeny optimální váhy jednotlivých vstupních proměnných ve skryté vrstvě (tabulka 5) a váhy výstupů ze skrytých neuronů do lineární výstupní vrstvy (tabulka 4). Porovnáváme-li chybu modelu neuronové sítě s chybou lineárního modelu, zjistíme, že variabilita relativních přírůstků diskontní úrokové míry České národní banky na vybraných proměnných je také lépe vysvětlena RBF neuronovou sítí než lineárním modelem. Porovnáváme-li nejmenší chybu modelu RBF sítě s nejmenší chybou modelu MLP sítě, zjistíme, že optimální RBF síť poskytuje menší chybu než optimální MLP síť. V obou případech je však variabilita relativních přírůstků diskontní úrokové sazby ČNB vysvětlena méně než na 50 %.

7 Závěr

V této práci byly použity dva typy neuronových sítí k modelování chování České národní banky na základě veřejně dostupných makroekonomických dat. Byla modelována je relativní změna diskontní úrokové sazby ČNB. Za předpokladu, že Česká národní banka pracuje s adaptivním očekáváním, byly použity zpožděné makroekonomické veličiny jako prediktory pro modelování chování ČNB s výjimkou diskontní sazby, kdy se spíše než adaptivní očekávání předpokládá určitá kontinuita v chování centrální banky. Bylo prokázáno, že některé makroekonomické veličiny umožňují jednokrokovou predikci relativní

Tabulka 5: Optimální váhy jednotlivých proměnných ve skryté vrstvě RBF sítě

váha	1	2	3	4
Poloměr RBF	4,676286	6,485502	1,711815	3,842197
Sazba ČNB	0,428009	-22,757768	-6,064228	-22,639839
Inflace ČR (CPI)	5,514883	-0,253952	5,934267	1,515577
Průměrný měsíční kurz EUR/CZ	-3,288189	3,502609	-0,155341	-2,069613
IR Pribor 3M	-12,000000	15,148288	11,658247	12,575753
Tempo růstu M2	7,086276	3,202215	5,404994	12,754242
Měsíční míra nezaměstnanosti	0,067192	-3,156935	0,884224	-0,845098
Sazba ECB	5,634724	-14,858715	5,975916	16,641846
Tempo růstu HDP ČR (čtvrtletní)	2,854650	5,297753	4,822723	2,301018

změny diskontní úrokové sazby ČNB. V případě lineárního modelu jsou statisticky významné vlivy zpožděné diskontní úrokové sazby ČNB, zpožděné úrokové sazby na mezibankovním trhu Pribor 3M a zpožděné úrokové sazby Evropské centrální banky. Standardní odchylka řady relativních přírůstků diskontní úrokové míry České národní banky je 8,655 % a chyba predikce lineárním modelem je 7,421 %. Tento malý rozdíl vypovídá o obtížné predikovatelnosti chování ČNB lineárním modelem.
Jako nelineární alternativa k lineárnímu modelu byly vybrány umělé neuronové sítě typu MLP a RBF sítě a ve snaze najít optimální počet skrytých neuronů zajišťující co nejnižší chybu predikce bylo zjištěno, že pro neuronovou síť typu MLP se pro tři skryté neurony podařilo snížit chybu predikce na 5,855 % při topologii ANN 8-3-1. Neuronová síť typu RBF docílila pro čtyři skryté neurony chybu predikce 5,347 % s topologií ANN 8-4-1, což je nejlepší dosažený výsledek. Rozdíl v chybě jednotlivých modelů sice není řádový, ale modely neuronových sítí mohou v oblasti predikce chování ČNB přinést podstatné zlepšení v porovnání s lineárním modelem. Je třeba si povšimnout i řádových rozdílů v hodnotách vah mezi skrytou a výstupní vrstvou. V síti MLP se převážně uplatňuje pouze první a třetí neuron, avšak druhý neuron nelze zcela eliminovat, neboť model se dvěma skrytými neurony má větší chybu predikce. Obdobná situace je u sítě RBF, kde se nejvíce uplatnil pouze třetí neuron a zbylé tři neurony hrají symbolickou roli. Ačkoliv oba nelineární prediktory neumožňují jednoznačnou interpretaci vah ANN, čímž je jejich analytický význam mírně snížen, modelování úrokových měr s využitím neuronových sítí je užitečná alternativa ve srovnání s tradičním lineárním modelem.

Literatura

[1] S. Haykin. *Neural Networks: A comprehensive foundations, 2nd edition.* Upper Saddle River, New Jersey, (1999).

[2] I. Křivý, J. Tvrdík, R. Krpec. *Stochastic Algorithms in Nonlinear Regression.* In 'Computational Statistics and Data Analysis'. 33(3), (2000) 277-290.

[3] M. Mandel, V. Tomšík. *Monetární ekonomie v malé otevřené ekonomice.* Management Press, Praha, (2008).

[4] Z. Michalewicz, D. B. Fogel. *How to Solve It: Modern Heuristics, 2nd edition.* Springer,(2004).

[5] M. Šnorek, M. Jiřina. *Neuronové sítě a neuropočítače.* Skripta ČVUT, Praha, (1997).

[6] J. Tvrdík, I. Křivý. *Simple Evolutionary Heuristics for Global Optimization.* In Computational Statistics and Data Analysis. 30(3), (1999) 345-352.

[7] Zákon o České národní bance, Zákon č. 6/1993 Sbírka zákonů

Business Process Modeling and Business to IT Transformation Revisited

Barış Unal, Josef Myslín

1st year of PGS, email: brsunalmsc@gmail.com
Department of Software Engineering in Economy
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Vojtech Merunka, Deptartment of Software Engineering in Economy,
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. Business process modeling is one of the essential parts of system development and served as foundation for subsequent activities like analysis, design and implementation. BPM techniques should be clear and easy that everybody who plays a role for description of the system, could understand it and it should use different artifacts to provide essential properties of the system at high abstraction level that later development efforts could use these models effectively. It also plays an important role for communication between all stakeholders and requirement analysis. However, semantic gap between business process modeling and subsequent information system development activities is the main problem. In this paper, different business process modeling techniques and their contribution to corresponding information system development are discussed.

Keywords: business process modeling, conceptual gap, BORM, model driven engineering, information system development

Abstrakt. Business modelování je jedna ze základních částí procesu vývoje a slouží jako základ posloupnosti aktivit jako jsou analýza, návrh a implementace. Techniky business modelování by měly být jasné a snadno pochopitelné pro každého, kdo se podílí na popisu systému, případně kdo by měl tento popsi pochopit a kdo by měl poskytnout základní parametr systému na vysoké úrovni abstrakce tak, aby při pozdějším vývoji bylo možno tyto modely efektivně použít. Hrají také významnou roli při komunikaci všech účastníků procesu vývoje a analytikem požadavků. Bohužel, velkým problémem je sméantická mezera mezi procesem business modelování a návaznými aktivitami vývoje systémů. V tomto článku jsou diskutovány techniky byznys modelování a jejich konsekvence vzhledem k dalším korespondujícím aktivitám.

Klíčová slova: modelování business procesů, koncepční rozdíly, Borm, model řízeného inženýrství, informační systém vývoje

1 Introduction

A business process model is an abstraction of the real world complex business systems. Main purpose of a business model is to create a medium where all stakeholders including domain experts, customers, end-users, system analyst and software system developers communicate. In addition to that, business process models are also the foundation of subsequent system modeling activities. That is the reason that business process models should be transformed to the IT world without loss of information in a consistent manner

which enables complete and satisfactory end-user products. Main question of system development comes from the semantic gap between business process description and the corresponding information system built on that business specifications. Even though there are so many system development processes, this conceptual difference is still main concern among these approaches. In this paper, different BPM techniques and their way of bridging this conceptual gap are illustrated.

Since full system development process life cycle is comprehensive and comprises various aspects and artifacts, the scope of this paper is delimited with different BPM techniques with the main concern of semantic gap between business domain and system IT.

In the following sections, with the excursion into different BPM techniques, required properties, which should be satisfied to minimize conceptual gap, are discussed. As a kind of BPM techniques, BORM [9]and its way of handling this problem are emphasized.

2 Main Problem - Business Domain to IT Transformation

As mentioned in the preceding section, main source of incoherency between business and IT world, could be specified as *conceptual gap*. More formally Goldberg [1] uses different terms to define this problem. *Concept space* describes the system under development from the user/expert view point. *The articulation space* is used to define communication medium between user/expert and system analyst. The constructed model as a feedback to that user/ expert's expectation is called as *analyst space*. Basically, difference between this concept and corresponding analyst's space is called as the *conceptual gap,* which is addressed in this paper.

The cost of any kinds of requirements misconception among participants of system development increases exponentially as it gets detected later on the following phases of development, especially after product delivery to end-user. The cost of bugs which are detected later in the development process, is much more than the cost of any bugs detected earlier during initial phases. So, it gives a big responsibility to the business process modeling in order to get an unambiguous, exact, concrete and consistent representation of the system under development. The second step is to transform this business models into corresponding software system development environments without loss of information and in a consistent manner. BPM and software development are different disciplines and have their own artifacts. Moreover, users and domain experts have totally different expertise and knowledge level with respect to software developers who deal with system's technical and implementation aspects. BPM techniques should provide participants of both worlds with a common ground for easy communication to create concrete software system specifications through business to IT transformation.

3 Business Process Modeling Approaches

Even though there are lots of system development processes in the literature, problem of business to IT transformation is still problem. Menes et al.[8] define two dimensions for transformation. *A horizontal transformation* is a transformation at the same level of

abstraction whereas a *vertical transformation* renders an input into more detailed output through refinements. Stein et al.[14] propose a framework for business to IT transformation. In this framework, it is stated that business process models should be platform independent and the platform specific IT solution, nevertheless must be derived from these process models through vertical transformation, not horizontal. According to [14], horizontal transformation guarantees totally independent business and IT environments, however, vertical transformation requires a business process model which should have IT related perspectives to be configured by domain experts at high level abstraction. Vertical transformation is accepted as a more reasonable approach compared to the horizontal transformation to bridge the gap between business and IT.

In this section, as a special and important discipline of system development processes, BPM techniques are illustrated with the concentration on their support for subsequent software system development efforts in order to bridge the cultural gap.

3.1 RUP and BPM

Business process modeling is one of the Rational Unified Process disciplines. [4] It is performed during *inception phase* of system development life-cycle. Similar approaches and artifacts as in software development are suggested for BPM in RUP to create an strong inter-dependency between BPM and rest of the system development. As in subsequent software development effort, use case modeling technique and business object model with system actors are used during this phase. UML (Unified Modeling Language) is used for description language as in the rest of software engineering disciplines like requirement, analysis and design. Although UML is reach enough for software engineering and provides different modeling techniques from different points of view of the system under development, it is claimed that using the same techniques for business description might not be comprehensible and appropriate for domain experts and users of the system, who do not have technical knowledge as system and software engineers have. Domain experts and users, nevertheless should contribute to the system description actively, especially in the early and during maintenance phases.

As discussed in the following sections, more user centric methodologies with different specification techniques behind them, like EPC, Petri Nets or FSM (Fine State Machines) have been developed. In stead of using RUP's own business process modeling techniques, business process descriptions specified by these methodologies could be transformed into the RUP software development environment to exploit UML for subsequent software engineering disciplines.

3.2 ARIS

ARIS (Architecture of Integrated Information Systems) framework is one of methodologies for modeling of business processes [11]. Today, this methodology is widely used in the world of business modeling. The author of this framework is Professor August-Wilhelm Scheer; its research has started in 1990s. Today ARIS is complex methodology which offers for analyzing processes and it takes a holistic view of process design, management, work flow, and application processing with a powerful tool for modeling. The main modeling technique used in ARIS is Event-driven process chain (EPC). This tool is directed

graph which follows the process from begin to end. The basic of idea lies in the fact that processes consist of events and our reactions. In other words – our activities are affected by events. EPC diagram is the sub-sequence of events and functions, where two events are linked by a function. Then a function represents our reaction on some events. This concept expects that every event has some reaction and every reaction is conditioned by some event. We never do something without reason and without suggestion and no event is without response. But the basic version of EPC is not able to express more difficult and complex processes. So it has been extended, which is called Extended EPC(eEPC). eEPC chart is able to show not only events and functions, but also inputs, outputs, roles and organization units. The most important property in eEPC is the possibility of flow control. Now we can use logical connectors which allow analyst to model branches and conditions. So this is the way how we can model business process in language, which is natural to workers in business. They feel process in the way of activities that they have to accomplish as a result of some events. Analyst has to define (in cooperation with stakeholders from the business where processes are analyzed) the initial event, which starts whole process and then he has to find sub-sequence of reactions and events. The chart as result is understandable for analyst as well as for active participants from business.

3.3 BPMN(Business Process Modeling Notation)

Business Process Modeling Notation (BPMN) is a graphical notation (the set of graphical objects and rules, which determine the possibility of connections of objects) that system analyst can use for business process modeling [15]. The primary goal of the research activity is the same as the goal of this paper – bridge the semantic gap between business modeling and subsequent activities. Now this method is one of the standards for business process modeling and it is widely used (with many of modeling tools). The second goal is to make the method easy as well as usable for modeling of complex processes. Important aspect is the way how to convert model to IT implementation for running processes. Mostly, BPMN is used for automatic model transformation towards IT domain. (BPMN has so many features that it is not easy to make full automatic conversion. Some tools restrict some features and then they can do this automatic conversion. For detailed techniques for model transformation using BPMN, you can check.[10]) The result of modeling is one diagram, Business Process Diagram (BPD), which is network of graphical objects and flows between them. Graph can also contain lanes, which can represent roles, departments etc. For details of notation, you can find in [15]. As an advantage, research of this notation is oriented to cooperation between analyst and participants from business.

3.4 *XMDD* (Extreme Model Driven Design)

In *XMDD*[5], an holistic approach is proposed to bridge the gap between business driven requirement and IT-based realization. A well defined (jABC) framework[6] is specified to support business process modeling and model transformation to IT system.

XMDD combines different system and software development paradigms to provide an agile process. The combination of eXtreme programming, model driven design and process modeling forms the basics of *XMDD* process. In addition to that, service oriented paradigm is followed during process modeling. It is stated in [12]that

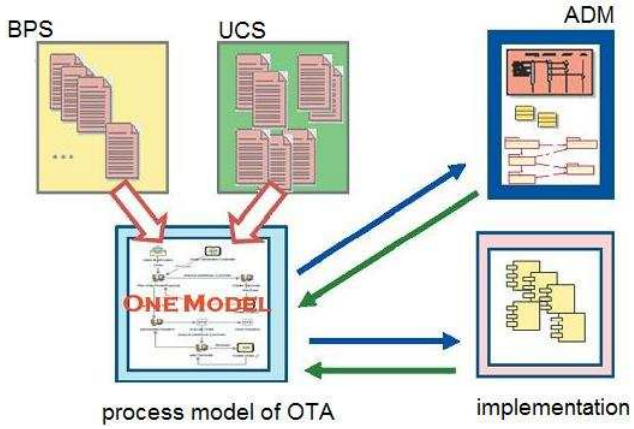


Figure 1: OTA of XMDD applied to heterogeneous landscape of the Rational Unified Process

> "A very highlevel kind of programming, in terms of orchestration, coordinates and harmonizes application-level "things" that are provided as services."

Implementation of services is regarded as a distinct task and any kind of programming paradigm could be used for service implementation, which is beyond the scope of this paper.

Most important philosophy used by *XMDD* is "One-Thing Approach" (*OTA*). It is a cooperative and hierarchical development process, which is organized by building and refining "one thing artifact" during whole process, which is called "Service Logic Model". Model refinements could occur in different forms like adding details, defining roles, responsibilities, any kind of constraints or performance requirements along the way down to the implementation in both vertical and horizontal dimensions. One thing approach guarantees consistency between model refinement efforts while keeping the loss of information minimum between consecutive process stages compared to other "many thing approaches". Model refinement continues until a sufficient level of detail is reached, where user and application functions could be implemented as elementary services by IT domain experts. So, cultural gap between business and IT domain just becomes service-oriented realization of requirements.

As shown in figure 1, business process and use case specifications in UML are combined together to build consistent one thing in *XMDD* for further refinements. Application of *OTA* for a project which was first implemented by RUP within IKEA IT group, is detailed in [2].

Moreover, domain experts and users take the control during process modeling. They participate actively in system analysis, model verification and refinement. They also experience and monitor the developed system at any phase by document browsing, simulation, full execution, or mixtures of them. So *XMDD* could also be called as "User Centric".

Briefly, one thing philosophy with empowering domain experts during business description, model refinements and verification provides an agile, consistent and comprehensible process in order to bridge the cultural gap between business and IT.

For further information about *XMDD*, you could also check references[2, 7, 13].

In the next section, BORM methodology regarding business process modeling aspect with its properties mentioned in preceding sections is discussed.

4 BORM - Our Approach

In this section, as another business process modeling methodology, Business Object Relation Modeling[9] is discussed. In contrast with XMDD discussed in the preceding section, it uses object oriented modeling paradigm for all phases of system development.

BORM is a end-to-end, full life cycle supporting system development methodology, which has been progressed since 1993. It has been used to capture knowledge of typical business systems, like business processes, business data and all related problems associated with business systems. BORM methodology has been used in development of different systems with variable sizes. Details of recent projects with BORM could be seen in [9].With these examples, BORM has proved to be effective in process of describing business systems and introduction of new requirements. The effectiveness of BORM comes from its simplicity and usage of unified method to model systems in different abstraction levels.

One of the main problems addressed in BORM is business to IT transformation. As discussed in the previous sections, active participation of stakeholders is necessary for requirement specification and creation of conceptual model of the systems to ensure that systems under development are verified easily. This requirement is only achievable through an understandable and easy modeling technique as the primary concerns for business process representation. It is widely accepted that "use case" modeling for the initial stages of development is not enough to capture all necessary aspects of the systems, and it should be supported by other modeling techniques like sequence and activity diagrams in UML.

In BORM, single diagram shows all necessary aspect of the systems at certain abstraction levels, in stead of using multiple diagrams for different views. Any change in business description is reflected apparently down to the implementation to guarantee consistent model generations. So we could say that BORM uses "*one thing approach*" like in XMDD discussed in the preceding section. Diagrams in BORM reflect the nature of object oriented paradigm, including business process description. It is fully object-oriented development process. BORM has several concepts associated with different stages of system development and some rules considering model transformation between these concepts.

In business process diagram, main concepts are business objects, their behaviors (functions), data links and communication between objects. Every object is represented as finite state machine(FSM), where transitions between object's states are initiated by events. Three dimensions of object oriented paradigm, which are data, behavior and history of objects are represented in this model. It combines three diagrams of UML together, state, communication and sequence diagrams. In figure 2, an invoice business process modeled in BORM is shown as an example.

As it could be seen, rectangles and ovals represent objects states and behaviors (function) respectively. Arrows between objects are used to model communication and data transfer between them. Concepts used for business process description is simple, understandable and comprehensive, which is necessary to minimize conceptual gap. It allows

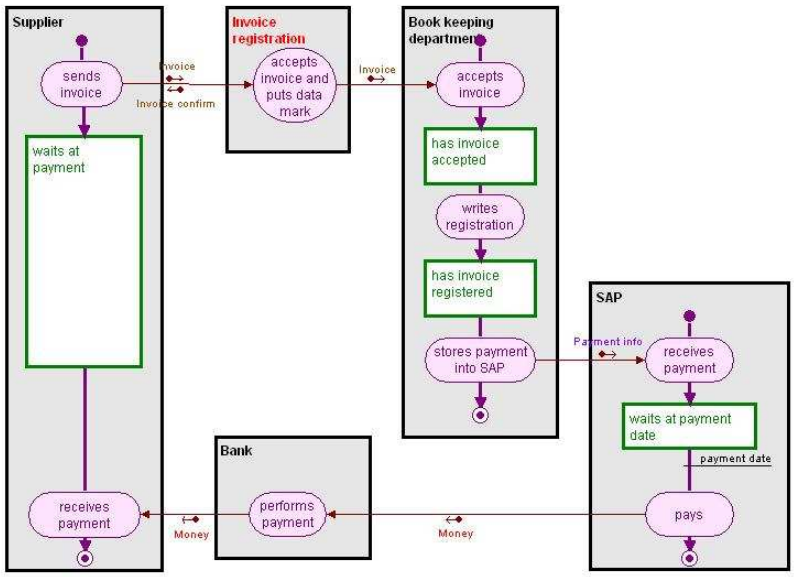


Figure 2: Invoice Business Process in BORM

domain experts and users of the system to participate in business process description actively to specify system requirements concretely and unambiguously. Domain experts and users could also test, verify and validate the system under development without any relation to software realization. So it could be said that BORM borrows "test first" paradigm with active participation of domain experts and end-users at the modeling level. *BORM is a user centric approach.*

Business process model built by all stakeholders is step-by-step transformed towards to the final model of the system for software implementation. BORM uses specific rules for model transformation to keep consistent set of models. Models in BORM are organized hierarchically via model decomposition. An object or behavior of a model must be linked to another object or behavior in an upper or lower level of hierarchy. Two new concepts for vertical and horizontal transformations as mentioned in section 3, are used to define this hierarchy mechanism. *Model aggregation/refinement* uses IS-A or HAS-A relation to add details to models across different abstraction layers whereas *model filtration* is used for simplification, encapsulation or hiding unnecessary parts of models at the same aggregation/refinement levels. Model aggregation/refinement and filtration mechanism is shown in figure 3. In this example, a library process model is transformed by model aggregation/refinement and filtration mechanism. This mechanism allows model transformation both in vertical and horizontal dimensions. Any change in a model at some certain level of hierarchy is reflected through model hierarchy in a consistent manner.

Moreover, BORM exploits fast prototyping through use of totally dynamic and pure object oriented implementation in Smalltalk, which is necessary to validate and verify system under development at early stages.

For the detailed description of all concepts, models and transformation rules used in BORM, you could see [3, 9].

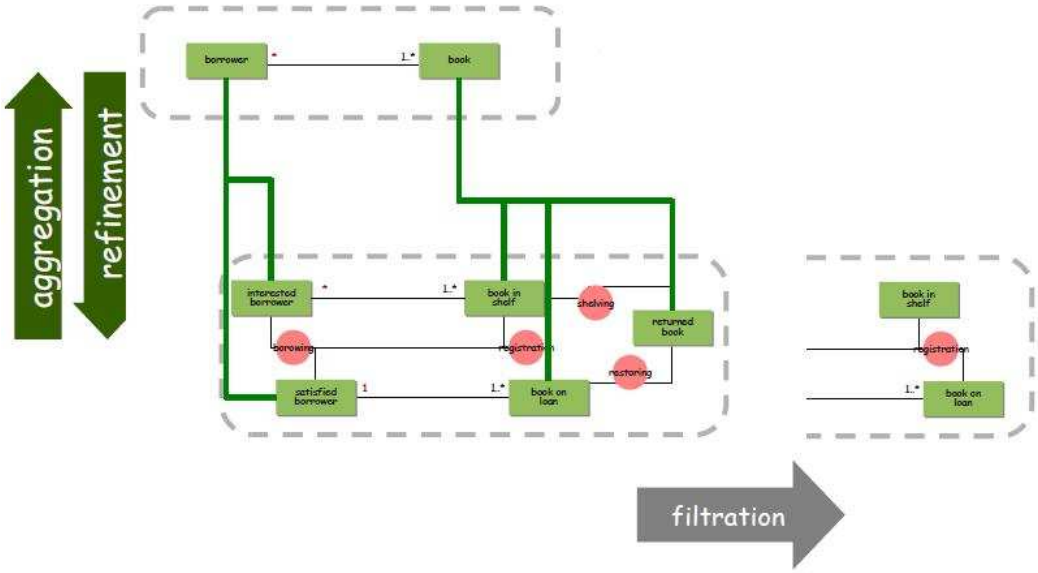


Figure 3: Model aggregation/refinement and filtration

5 Conclusion

As a conclusion, business process modeling constitutes initial stages of system development and its main purpose should provide a medium for communication between all stakeholders in a project to bridge the gap between business and IT. Throughout the paper, it has been emphasized that any approach to build common understanding of the business derived requirements should have certain characteristics.

Business process models must be simple, understandable and comprehensive enough to represent requirements of the system under development concretely and unambiguously. By doing so, domain experts and users who do not have technical knowledge and experience, could be empowered and have direct control on the process during early stages of development. We call it "user centric" approach. To keep the model hierarchy consistent and simple, one artifact which combines different aspects of the process together in one model, should be built and transformation toward to realization of final product should be based on step by step refinements of this unique model, which is called "one thing approach".

In addition to characteristics summarized above, in order to minimize the misconceptions among stakeholders, business process approaches must provides tools to test, verify and experience the system whenever it is necessary during development process.

Regarding the business process modeling characteristics discussed so far, we could say that BORM provides an object oriented and agile methodology through use of "user centric" and OTA with verification and validation supports during development process. Moreover, model aggregation/refinement and filtration concepts are defined to build a consistent model hierarchy in BORM. By doing so, step by step transformation towards software implementation is achieved. From the practical perspective, BORM has been used in different projects for 10 years. This experience proves that our clients prefer to understand and simulate each important relationship between materials, finance, resources,

and information in simple and concrete model, like in BORM business process model, which shows the importance and power of *user centric*, *OTA* in practice as discussed so far.

References

[1] Goldberg, A. and Rubin, K. S. (1995) *Succeeding with Objects* – Decision Frameworks for Project Management, Addison Wesley, ISBN 0-201-62878-3.

[2] Hormann, M., Margaria, T., Mender, T., Nagel, R., Steffen and B., Trinh, H.(2008) *The jABC Approach to Rigorous Collaborative Development of SCM Applications*, Proc. ISoLA 08, Springer Verlag, 2008, pp. 724-737

[3] Knott, R. P., Merunka, V. and Polak J.(2003) The BORM methodology: a third-generation fully object-oriented methodology. In: Knowledge-Based Systems Elsevier Science International, New York, 2003

[4] Kruchten, P. (2003) *Rational Unified Process, The: An Introduction*, Third Edition, Addison Wesley, ISBN : 0-321-19770-4.

[5] Margaria, T. and Steffen, B.(2008) *Agile IT: Thinking in User-Centric Models*, Proc. ISoLA 2008, Springer, pp. 493-505.

[6] Margaria, T. and Steffen, B.(2009) *Business Process Modelling in the jABC: The One-Thing-Approach*, Handbook of Research on Business Process Modeling, IGI Global, ISBN: 978-1-60566-288-6

[7] Margaria, T. and Steffen, B. (2006) *Service Engineering: Linking Business and IT*, Computer, Oct. 2006, pp. 45-55.

[8] Mens, T., Czarnecki, K., Van Gorp, P.(2005) *A taxonomy of model transformations.* In Bezivin, J., Heckel, R., eds.: *Language Engineering for Model-Driven Software Development.* Number 04101 in Dagstuhl Seminar Proceedings, Germany

[9] Merunka, V., Polak, J., Brožek, J. and Šebek M. (2009) *BORM - Business Object Relation Modeling*, Proceedings of the Fifteenth Americas Conference on Information Systems, San Francisco, California-2009.

[10] Pautasso, C. and Koehler, J.(2008), *Preliminary Proceedings of 1st International Workshop on Model-Driven Engineering For Business Process Management,* Milano, Italy, 2008.

[11] Scheer, A. W.,(1992) *Architecture of Integrated Information System*, Springer Verlag.

[12] Steffen,B. and Narayan, P. (2007) *Full Life-Cycle Support for End-to-End Processes*, Computer, Nov. 2007, pp. 64-73

[13] Steffen, B., Margaria, T., Nagel R., Jorges, S. and Kubczak C.(2006) *Model-Driven Development with the jABC*, Proc. Haifa Verification Conf., Springer, LNCS 4383, pp. 92-101

[14] Stein, S., Kuhne, S. and Ivanov, K. (2008) *Business to IT Transformations Revisited,* 1st International Workshop on Model-Driven Engineering For Business Process Management, Milano, Italy-2008

[15] *Business Process Modeling Notations* www.bpmn.org

Built-up Structure Criticality*

Daniel Vašata

4th year of PGS, email: daniel.vasata@gmail.com
Department of Physics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Petr Šeba, University of Hradec Králové & Doppler Institute for Mathematical Physics and Applied Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. We analyse the structure of built-up land in the centre of big cities in the Czech Republic using the framework of statistical physics. To do this, both the variance of the built-up area and the number variance of built-up landed plots in spheres are calculated. In both cases the variance as a function of a circle radius follows a power law. The obtained value of the exponents are comparable to values typical for critical systems. The study is based on cadastral data from the Czech Republic.

Keywords: urban structure, critical systems, self-organized criticality

Abstrakt. V naší práci se zabýváme studiem struktury zastavěných ploch z pohledu statistické fyziky v centrech velkých měst v České republice. Za tímto účelem počítáme rozptyl zastavěné plochy a rozptyl počtu zastavěných pozemků ve sférách. V obou případech vykazuje tento rozptyl jako funkce poloměru mocninnou závislost. Získané hodnoty exponentů jsou srovnatelné s hodnotamy pro kritické systémy. Studie je založena na datech z katastru nemovitostí České republiky.

Klíčová slova: struktura městské zástavby, kritické systémy, samo-organizovaná kritikalita

1 Introduction

Urban land represents an important part of overall landscape where most of people live. The structure of cities is influence by cultural, sociological economic, political and other processes.

As was shown, despite the apparent complexity, some simple universal properties and rules can be found. The classical work discuss the size distribution of the cities [11]. If the cities are ranked by their number of inhabitants, then the rank-size distribution follows a power law. From physical point of view, it is interesting to study the spatial properties of urban pattern. Complex spatial features associated with urban systems are often being described within fractal self-similarity concept [1, 2, 8, 9].

Our aim is to study the principles of the urban structure on smaller scales. The main topic of this article is the analysis of the built-up pattern in the inner urban area based on the data from Czech Cadastre. The use of land parcels allows us to study the built-up structure over the smallest possible scale. We are going to show the connection of urban

*This work has been supported by the grant No. 202/08/H072 of the Grant Agency of the Czech Republic and by the grant No. SGS10/211/OHK4/2T/14 of the Czech Technical University in Prague.

land to so called critical systems from thermodynamics. Particularly that some spatial properties of built-up land pattern are the same as for critical systems.

Built-up land in the centre of cities is chosen, because one can there expect the uniform density of built-up land. By uniform, we mean the same probability for a point to be built-up thorough the history or possible future of the city, not the present state (concrete realization). Together with the previously mentioned fractal properties of urban systems and the fact that from economic point of view, the built-up land represents the new phase between other types of land, the effort to study critical properties seems natural.

2 Critical phenomena

Let us now recall some specific features of this so called critical systems [5, 6, 7]. Phase transitions within thermodynamic description of a system occurs in points where thermodynamic potential becomes non-analytic. Such non-analyticity may be discontinuous (first order phase transition) or continuous (second order). Properties of different thermodynamic systems near the continuous phase transition show specific universalities for which the term critical phenomena is used. In the past decades similar behaviour was also found in various systems from natural and social sciences.

If one is interested in the static spacial structure then certainly the most important property is the scaling invariance connected with the change of quantities under a change of length scale. In simple terms, if a part of a system is magnified to the same size as the original system, it is not possible to distinguish between the magnified part and the original system. In other words, near the critical point there exists only one characteristic length of the system, the correlation length ξ, which is solely responsible for singular contributions to thermodynamic quantities. At the critical point correlation length diverges thus no characteristic length is presented and the system is invariant under scale transformations.

In order to explicitly describe these features one needs to define an order parameter M as a thermodynamic quantity that distinguish between the two phases and approaches zero at the critical point as the phases become identical there. Well studied examples of order parameter are density difference between gas and liquid phase near critical point, shear modulus in liquid-solid phase transition or magnetization in ferromagnet-paramagnet transition. The conjugate field H is defined by the relation

$$dW = -HdM, \tag{1}$$

where dW denotes the work done on the system when the order parameter changes by dM. Canonical partition function can be written in the form

$$\mathcal{Z}(T,H) = \sum e^{-\beta(\mathscr{H}-H\tilde{M})}, \qquad \beta = \frac{1}{k_B T}, \tag{2}$$

where $\mathscr{H}$ and $\tilde{M}$ are values of the Hamiltonian and the order parameter, respectively, for a concrete realization. The sum is as usual taken over the whole ensemble. Order parameter can be from the partition sum obtained through Gibbs free energy given by

$G(T,H) = -k_B T \ln \mathcal{Z}$:

$$M = -\left(\frac{\partial G}{\partial H}\right)_T = \frac{1}{\mathcal{Z}} \sum \tilde{M} e^{-\beta(H - H\tilde{M})}. \tag{3}$$

It is also useful to define a local order parameter value $m(\boldsymbol{r})$ by the relation

$$M = \int_V \langle m(\boldsymbol{r}) \rangle d\boldsymbol{r}, \qquad V \subset \mathbb{R}^d, \tag{4}$$

where d is the dimension of the space and $\langle ... \rangle$ stands for ensemble average. If the system is homogeneous and isotropic then

$$\langle m(\boldsymbol{r}) \rangle = \langle m(0) \rangle = m = \frac{M}{V}, \qquad \forall \boldsymbol{r} \in V. \tag{5}$$

The susceptibility is given by the derivative of the order parameter density with respect to its conjugate field

$$\chi(T,H) = \left(\frac{\partial m}{\partial H}\right)_T = \frac{\beta}{V}\left[\frac{1}{\mathcal{Z}} \sum \tilde{M}^2 e^{-\beta(H - H\tilde{M})} - \left(\frac{1}{\mathcal{Z}} \sum \tilde{M} e^{-\beta(H - H\tilde{M})}\right)^2\right]. \tag{6}$$

Together with the definition of the order parameter density one gets

$$\chi = \frac{1}{k_B T V} \int_V \int_V G(\boldsymbol{r}_1, \boldsymbol{r}_2, T) d\boldsymbol{r}_1 d\boldsymbol{r}_2, \tag{7}$$

where $G(\boldsymbol{r}_1, \boldsymbol{r}_2, T)$ stands for two-point correlation function

$$G(\boldsymbol{r}_1, \boldsymbol{r}_2) = \langle (m(\boldsymbol{r}_1) - \langle m(\boldsymbol{r}_1) \rangle)(m(\boldsymbol{r}_2) - \langle m(\boldsymbol{r}_2) \rangle) \rangle . \tag{8}$$

The correlation function $G(\boldsymbol{r}_1, \boldsymbol{r}_2)$ describes the fluctuations of the order parameter and under homogeneity and isotropy can be simplified to

$$G(r) = \langle m(\boldsymbol{r}) m(0) \rangle - m^2, \tag{9}$$

where $r = |\boldsymbol{r}|$. Thus the final relation for susceptibility known as the *fluctuation-dissipation theorem* is

$$\chi(T) = \frac{1}{k_B T} \int_V G(r,T) d\boldsymbol{r}. \tag{10}$$

As we stated before, the correlation length diverge when approaching the critical point. Suppose the critical point occurs at the point T_c, H_c in the parameter space and the spacial size is infinite i.e. thermodynamic limit $V \to +\infty$. Introducing the reduced temperature $t = (T - T_c)/T_c$ the correlation length ξ is assumed to diverge as

$$\xi(t) \propto |t|^{-\nu}, \qquad \nu > 0. \tag{11}$$

The scaling assumption for correlation function can be written in the following form:

$$G(r) = \frac{\Psi_{\pm}(r/\xi)}{r^{d-2+\eta}}, \tag{12}$$

where subscripts $\pm$ denotes two different functions $\Psi_+(x)$ for $t > 0$, and $\Psi_-(x)$ for $t < 0$ as these directions can be generally different. Index η appearing in the exponent of power law part of $G(r)$ is called the anomalous dimension. Inserting this relation into equation (10) leads to

$$\chi = \frac{1}{k_B T} \int \frac{\Psi_\pm(r/\xi)}{r^{d-2+\eta}} d\boldsymbol{r}. \tag{13}$$

The final scaling form for susceptibility is

$$\chi = \frac{\xi^{2-\eta}}{k_B T} \int \frac{\Psi_\pm(x)}{x^{d-2+\eta}} d\boldsymbol{x} = K\xi^{2-\eta} \propto |t|^{\nu(2-\eta)}\,. \tag{14}$$

Similar relations hold also for particle density in the grand canonical ensemble [7, 4]. To obtain them we should replace the order parameter by the mean total number of particles $M \to \langle N \rangle$, conjugate field by the chemical potential $H \to \mu$, canonical partition function by the grand canonical one $\mathcal{Z}(T, H) \to \Xi(T, \mu, V)$ and the Gibbs free energy by the Grand potential $G \to \Omega$. Local order parameter is then just the density of particles $\langle \rho(\boldsymbol{r}) \rangle = \rho_0 = \langle N \rangle / V$. The role of susceptibility takes here the isothermal compressibility κ_T. Under such replacement it is easy to see the following relations:

$$\langle N \rangle = -\left(\frac{\partial \Omega}{\partial \mu}\right)_{T,V} = k_B T \left(\frac{\partial \ln \Xi}{\partial \mu}\right)_{T,V}, \tag{15}$$

where $\Xi(T, \mu, V) = \sum e^{-\beta(\mathscr{H} - \mu N)}$ and

$$\langle N^2 \rangle = \frac{1}{\beta^2 \Xi} \left(\frac{\partial^2 \ln \Xi}{\partial \mu^2}\right)_{T,V} = -\frac{1}{\beta} \left(\frac{\partial^2 \Omega}{\partial \mu^2}\right)_{T,V} + \langle N \rangle^2 \tag{16}$$

Fluctuations of the number of particles are thus given by

$$\langle N^2 \rangle - \langle N \rangle^2 = k_B T \left(\frac{\partial \langle N \rangle}{\partial \mu}\right)_{T,V}. \tag{17}$$

After a little play with Jacobians, Maxwell relation, the fact that Gibbs free energy is linear in N and the definition of compressibility

$$\kappa_T = -\frac{1}{V} \left(\frac{\partial V}{\partial p}\right)_T, \tag{18}$$

one obtains

$$\langle N^2 \rangle - \langle N \rangle^2 = k_B T \rho_0^2 V \kappa_T. \tag{19}$$

Defining the two-point correlation function $G(r)$ for $\rho(\boldsymbol{r})$ analogously to (8), the fluctuation-dissipation theorem can be written in the form

$$\kappa_T = \frac{1}{k_B T \rho_0^2} \int G(r) d\boldsymbol{r}. \tag{20}$$

Scaling forms for correlation length and correlation function are again described by the functional forms (11) and (12), respectively. The fluctuation-dissipation theorem gives analogously to (14)

$$\kappa_T = K\xi^{2-\eta} \propto |t|^{\nu(2-\eta)}\,. \tag{21}$$

The only qualitative difference between order parameter and particle density is in the singular structure of a concrete realization. Density function for point particles located at points $\boldsymbol{r}_1,\ \boldsymbol{r}_2,\ \boldsymbol{r}_3,\ \boldsymbol{r}_4,\ ... \in \mathbb{R}^d$ is given by

$$\rho(\boldsymbol{r}) = \sum_{i=1}^{\infty} \delta(\boldsymbol{r} - \boldsymbol{r}_i). \tag{22}$$

The correlation function can be written [7] in the form

$$G(\boldsymbol{r}_2 - \boldsymbol{r}_1) = \rho_0 \delta(\boldsymbol{r}_2 - \boldsymbol{r}_1) + \mathcal{G}(\boldsymbol{r}_2 - \boldsymbol{r}_1), \tag{23}$$

where $\mathcal{G}(\boldsymbol{r}_2 - \boldsymbol{r}_1)$ is the non-diagonal part meaningful only for $r = |\boldsymbol{r}_2 - \boldsymbol{r}_1| > 0$. Thus the difference of correlation of the order parameter and particle density is only in the diagonal δ therm which of course doesn't influence the character of the divergence near the critical point.

2.1 General parameter

From scaling laws it is analogously possible to show that the character of fluctuations given by the power-law divergence of two-point correlation function and general susceptibility holds also for the extensive thermodynamic variables that do not approach zero at the critical point. One such example is the density discussed above. The reason for working with the order parameter near the critical point lies in the possibility to expand the free energy in the powers of m and its derivatives and use this expansion in analytical derivation of the properties (with or without renormalization theory).

Since our attention now is not in the scale field modelling we are not limited to the assumption of the respective parameter be 0 at the critical point. In the next we will thus work with general parameter of the system. Under this parameter we also understand the density of particles if needed. In the previous section we showed that in such case the only difference is the presence of the diagonal part in the relation for correlation function (23).

2.2 Parameter variance in spheres

The useful tool to analyse experimental data is the variance of the parameter in spheres. For the parameter $m(\boldsymbol{r})$ with homogeneous and isotropic distribution $\langle m(\boldsymbol{r})\rangle = m$ the cumulative value of the parameter in the sphere of radius R is given by

$$M(R) = \int_{S(R)} m(\boldsymbol{r}) d\boldsymbol{r}, \tag{24}$$

where the sphere is the set $S(R) \equiv \{\boldsymbol{r} \in \mathbb{R}^d | \, |\boldsymbol{r}| < R\}$ with a volume $|S(R)|$. The centres of the spheres are not important because of the homogeneity of the parameter distribution. The parameter variance is defined [3] as

$$\sigma^2(R) = \langle M(R)^2\rangle - \langle M(R)\rangle^2, \tag{25}$$

where

$$\langle M(R)\rangle = \int_{S(R)} \langle m(\boldsymbol{r})\rangle d\boldsymbol{r} = m\,|S(R)|\,, \tag{26}$$

and

$$\langle M(R)^2\rangle = \int_{S(R)}\int_{S(R)} \langle m(\boldsymbol{r}_1)m(\boldsymbol{r}_2)\rangle d\boldsymbol{r}_1 d\boldsymbol{r}_2. \tag{27}$$

Using the definition (8) of the two-point correlation function, $\sigma^2(R)$ can be expressed by

$$\sigma^2(R) = \int_{S(R)}\int_{S(R)} G(\boldsymbol{r}_1 - \boldsymbol{r}_2) d\boldsymbol{r}_1 d\boldsymbol{r}_2. \tag{28}$$

It is reasonable to change the variables and obtain the relation

$$\sigma^2(R) = |S(R)| \int_{S(2R)} G(r) d\boldsymbol{r}. \tag{29}$$

If we compare this relation with the fluctuation dissipation theorem (10) then in the thermodynamic limit $V \to +\infty$ outside the critical point, where the susceptibility is finite, the following limit holds

$$\lim_{R\to+\infty} \sigma^2(R) = k_B T \chi(T) \lim_{R\to+\infty} |S(R)|\,. \tag{30}$$

This leads to

$$\sigma^2(R) \propto R^d \sim \langle M(R)\rangle, \qquad R \gg 1. \tag{31}$$

Outside the critical point is the parameter distribution sometimes called substantially Poissonian, since for the Poissonian point process (particle distribution in the ideal gas) is the last equation valid for all R. That can be easily derived from the fact, that the particles are non-interacting and therefore independent. The correlation function for density (23) has only the diagonal part

$$G(r) = \rho_0 \delta(r). \tag{32}$$

The density variance is therefore

$$\sigma^2(R) = \rho_0\,|S(R)| = \langle M(R)\rangle. \tag{33}$$

Different situation arise when the system is approaching the critical point. There both the susceptibility (compressibility) and correlation length ξ diverge. Spatial correlations in this region are long-ranged and the correlation function is dominated by the power-law decay (12). In the region $R \gg 1$ and $R \ll \xi$ we obtain

$$\sigma^2(R) \propto |S(R)| \int_{S(2R)} \frac{1}{r^{d-2+\eta}} d\boldsymbol{r} = C\,|S(R)| \int_0^R \frac{r^{d-1}}{r^{d-2+\eta}} dr = C\,|S(R)|\,R^{2-\eta}. \tag{34}$$

Hence the fluctuations are proportional to

$$\sigma^2(R) \propto R^{d+2-\eta} \sim \langle M(R)\rangle^{\frac{d+2-\eta}{d}}. \tag{35}$$

3 Data analysis

In this section we show that built-up land patterns have the same fluctuation properties as critical systems. As we work with the surface data, the dimension of the space in previous formulas is now $d = 2$.

Our data contain all cadastral records form the Czech Republic. Every landed plot i is characterised by its definition point $\boldsymbol{r}_i$, size (acreage) λ_i, type of land and the ownership data. Since our interest is in the built-up structure we restrict our attention only to built-up landed plots. We don't know the exact parcel shape, thus the most straightforward analysis of cadastral data is to use the point pattern given by definition points $\boldsymbol{r}_i$ of the parcels. This is in the latter text called "point" representation. For this representation the order parameter is represented by the singular point density (22). Parameter variance $\sigma^2(R)$ is than the number variance in sphere.

Another possibility is to approximate unknown parcel shapes by circles with the same acreage. The built-up land is represented as a subset Z of two dimensional surface $\mathbb{R}^2$. Order parameter in this case is just the indicator of such subset: $m(\boldsymbol{r}) = 1$ if there is a building at $\boldsymbol{r}$, $m(\boldsymbol{r}) = 0$ otherwise. Such approximation leads to errors. Fortunately the approach of estimating parameter variance in spheres is much less sensitive to them than direct estimation of correlation function. During the estimation of built-up area contained inside one concrete sphere $S(R)$ the intersection area of the sphere with every parcel represented by circle is added to cumulative result:

$$\tilde{M(R)} = \sum_{i \in I} \lambda \left(S_i \bigcap S(R) \right), \tag{36}$$

where $\lambda(.)$ denotes the Lebesgue measure on $\mathbb{R}^2$, I is the set of all built-up parcels and S_i is the circle positioned at the definition point of the i-th parcel having the same size $\lambda(S_i) = \lambda_i$. For R much larger than typical parcel perimeter this approach produce errors only in the vicinity of the sphere boundary. The effective error will therefore decrease as

$$\frac{M(R) - \tilde{M(R)}}{M(R)} \sim R^{-1}. \tag{37}$$

For "set" representation the built-up area variance in spheres is calculated.

All estimations are based on the assumption of self-averaging property [10]. It means that sufficiently large sample is a good representative of the whole ensemble. In our case however, the size of sample is limited to the area around the city centre where we can expected uniform density. For typical large Czech city the perimeter of such area is about 4 km. This size puts limitation on the perimeter of spheres in order to obtain reasonable statistics. Together with the fact that the power law dependence, if presented, is valid for $R >> 1$, one is usually restricted to work in the region 400 km $\lesssim R \lesssim$ 1000 km. Related to this, mean values in formula (25) for fixed perimeter R are estimated in the following way. Inside the studied part of the city $A \subset \mathbb{R}^2$ centres $\boldsymbol{o}_j$ of N spheres are uniformly randomly chosen so that every sphere is inherited in A, $S(\boldsymbol{o}_j, R) \subset A$. For every sphere $S(\boldsymbol{o}_j, R)$ the cumulative parameter value (number of points or built-up area) $M_j(R)$ inside is calculated. The mean value is than estimated by

$$\langle M(R) \rangle = \frac{1}{N} \sum_j M_j(R) \tag{38}$$

and the variance by

$$\sigma^2(R) = \frac{1}{N-1} \sum_j \left(M_j(R) - \langle M(R) \rangle\right)^2 . \tag{39}$$

Note the same notation for the definition (25) and for the estimator (39). It is always clear from context what does the symbol mean.

3.1 Results

We analysed 6 largest cities in the Czech Republic. For each city we calculate both number variance in spheres (point representation) and built-up area variance in spheres (set representation). The dependences of $\sigma^2(R)$ on $\langle M(R) \rangle$ in the case of set representation are plot in the log-log scale in figure 1.

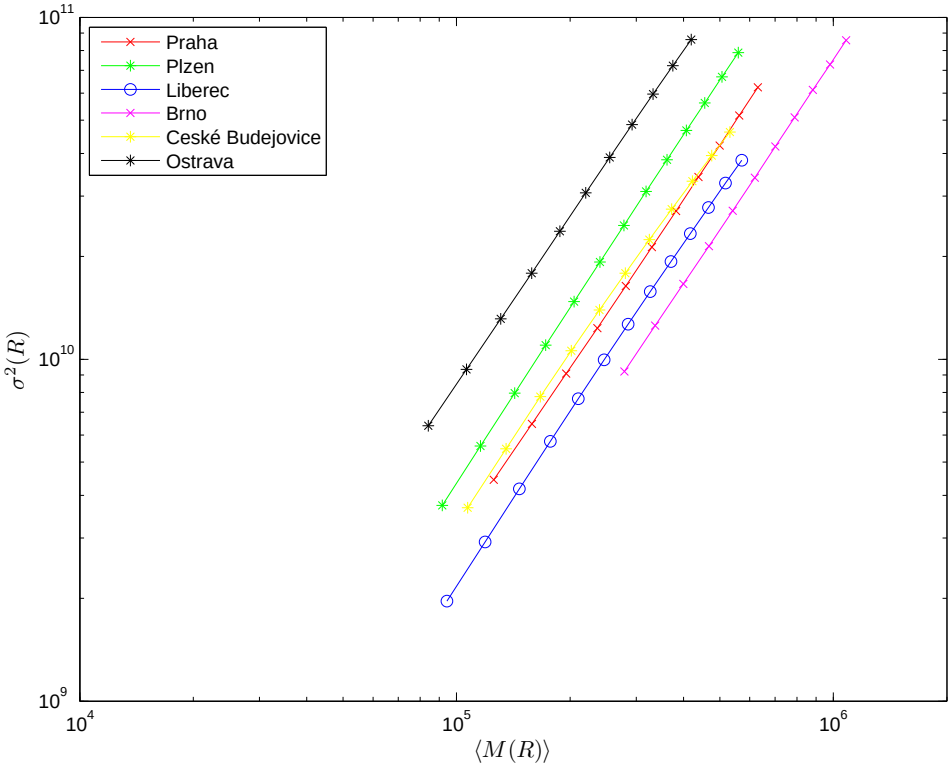


Figure 1: Dependencies of $\sigma^2(R)$ on $\langle M(R) \rangle$ in log-log scale for different cities.

It is clearly visible that the dependence for all the cities follows a power law. It can be therefore fit by the strait line (in log-log scale). From this fit we can easily determine the exponent of power-law. The summary of resulting exponents α for studied cities according to the relation

$$\sigma^2(R) \propto \langle M(R) \rangle^{\alpha} \tag{40}$$

is presented in table 1.

As was shown by (31), exponent $\alpha = 1$ express the system that is outside of the critical region, e.g. randomly positioned particles. One can see that this is not the case of built-up pattern.

Table 1: Exponents α according to power law dependence (40) of $\sigma^2(R)$ on $\langle M(R)\rangle$.

City	Point representation	Set representation
Praha	1.47	1.64
Plzeň	1.61	1.69
Liberec	1.54	1.65
Brno	1.40	1.65
České Budějovice	1.50	1.58
Ostrava	1.54	1.62

4 Conclusion

We study the built-up land pattern in the centres of 6 largest cities in the Czech Republic. Our analysis is based on cadastral data. For every parcel we know the location of the definition point, size, type of land, that uniquely determines the built-up land and other properties. For the purpose of analysis the built-up land is represented in 2 different ways - points and subset.

Because the data do not contain information about exact shape of parcels, it is useful to study the fluctuations of built-up area in spheres (circles). This leads, especially for set representation, to effective error that decreases with increasing perimeter R of the spheres.

The computations show, that for both representations the dependence of fluctuations on the mean value of the parameter follows a power law. Moreover the set representation, as can be expected, seems to be more universal. The values of exponent α in the relation $\sigma^2(R) \sim \langle M(R)\rangle^\alpha$, for different cities in the Czech Republic are very close to the value $\alpha = 1.64$.

We can conclude that the inner urban area structure is correlated with a long ranged power-law dependence. This shows the connection between critical systems and the urban system. The power-law exponent seems to be independent of the concrete city, being therefore determined only by the fact that it represents an inner urban structure. Such observation is very interesting and the connection between urban area and critical systems may be useful to development and verification of further urban models. The probable explanation may be inherited in the connection of built-up land to various networks, e.g. transportation, water supply, sewerage, electricity.

References

[1] M. Batty and P. Longley. *Fractal Cities: A Geometry of Form and Function.* Academic Press, first edition, (1994).

[2] P. Frankhauser and R. Sadler. *Fractal analysis of urban structures.* Natural Structures - Principles, Strategies and Models in Architecture and Nature. Proc. Int. Symp. SFB 230 **4** (1992), 57–65.

[3] A. Gabrielli, M. Joyce, and F. Sylos Labini. *Glass-like universe: Real-space correlation properties of standard cosmological models.* Phys. Rev. D **65** (Apr 2002), 083523.

[4] N. Goldenfeld. *Lectures on Phase Transitions And the Renormalization Group.* Perseus Books, (1992).

[5] I. Herbut. *A Modern Approach to Critical Phenomena.* Cambridge University Press, Cambridge, U.K., (2007).

[6] K. Huang. *Statistical Mechanics.* John Wiley & Sons, 2nd edition, (1987).

[7] L. D. Landau and E. M. Lifschitz. *Statistical Physics.* Pergamon Press, 3rd edition, (1980).

[8] H. A. Makse, J. S. Andrade, M. Batty, S. Havlin, and H. E. Stanley. *Modeling urban growth patterns with correlated percolation.* Phys. Rev. E **58** (Dec 1998), 7054–7062.

[9] F. Schweitzer. *Brownian Agents and Active Particles.* Springer, first edition, (2003).

[10] D. Sornette. *Critical Phenomena in Natural Sciences.* Springer, (2000).

[11] G. K. Zipf. *Human Behavior and the Principle of Least Effort.* Addison-Wesley, (1949).

Backward Stochastic Differential Equations and its Application to Stochastic Control*

Petr Veverka

2nd year of PGS, email: veverka@utia.cas.cz
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Miloslav Vošvrda, Institute of Information Theory and Automation, ASCR

Abstract. In this article, we introduce the concept of Backward Stochastic Differential Equations (BSDE), provide fundamental theorems of existence and uniqueness of the solution for some essential cases and we show by example its important connections to financial mathematics. Finally, we focus on vast applications of BSDE to stochastic control via Pontryagin's maximum principle.

Keywords: backward stochastic differential equations, stochastic control, stochastic maximum principle

Abstrakt. V tomto článku představíme koncept zpětných stochastických diferenciálních rovnic (BSDE) a vyslovíme zásadní věty o existenci a jednoznačnosti řešení takových rovnic v obecném případě. Na příkladu dále ilustrujeme jedno z jejich možných využití v oblasti řízení finančního portfolia. Poslední část je věnována uplatnění zpětných rovnic v teorii stochastického řízení užitím Pontrjaginova principu maxima.

Klíčová slova: zpětné stochastické diferenciální rovnice, stochastické řízení, stochastický princip maxima

1 Introduction

The domain of BSDE, in its full generality, was first studied in 1990 by Pardoux and Peng who formulated the general problem of BSDE and proved some fundamental theorems including the central one - the existence and uniqueness of the solution, see [3]. Since then, BSDE have found a variety of applications in finance, in physics but also in even more theoretical fields such as stochastic control, theory of random processes probability distributions, probabilistic representation of elliptic and parabolic-type deterministic PDE's, numerical methods for PDE's and many other.
The first section of the article gives an introduction to BSDE - we start by the theorem of Pardoux and Peng for finite time horizon BSDE and then we proceed to infinite time horizon case considering, in addition, Lévy driven stochastic noise. We refer to [5], [7] and [9] for an overview on generalizations of this type. Further, to present an example of a practical model using the BDSE theory. We show how the theory can be applied to

*This work has been supported by grants no. 402/09/H045 and no. P402/10/1610 of the Czech Science Foundation.

the European Call Option hedging problem. In the second section, we formulate the task of stochastic control and associated maximum stochastic maximum principle and discuss some other extension of the model.

2 Backward stochastic differential equations (BSDE)

2.1 Finite time horizon case

The main motivation for introducing the BSDE is the need for solving problems with terminal condition of the following type

$$\begin{aligned} -dY_t &= f(t, Y_t, Z_t)dt - Z_t dW_t, \quad \forall t \in [0, T), \text{ a.s.} \\ Y_T &= \xi, \text{ a.s.}, \end{aligned} \tag{1}$$

where $0 < T < +\infty$ is a finite time horizon, $\big(\Omega, \mathcal{F}, \mathbf{P}\big)$ is a standard probability space equipped by a standard $\mathbb{R}^d$-valued Wiener process $\big(W_t\big)_{t\in[0,T]}$. Let $\big(\mathcal{F}_t^W\big)_{t\in[0,T]}$ be the canonical filtration of W_t, i.e. $\mathcal{F}_t^W = \sigma\big(W_s; s \le t\big)$ and $\big(\mathcal{F}_t\big)_{t\in[0,T]}$ be its completion. The function f (called *drift*) and the random variable ξ (*terminal condition*) are, in fact, the only inputs of the equation.

Definition 1: *The couple (f, ξ) is called standard parameters of the equation (1) if it holds*

- $\xi \in \mathbf{L}^2(\mathcal{F}_T; \mathbb{R}^n)$, i.e. ξ is an $\mathcal{F}_T$-measurable r.v., $\mathbb{R}^n$-valued, satisfying $\mathbf{E}||\xi||^2 < +\infty$
- $f : \Omega \times [0, T] \times \mathbb{R}^n \times \mathbb{R}^{n\times d} \to \mathbb{R}$, i.e. $(\omega, t, y, z) \mapsto f(\omega, t, y, z) \in \mathbb{R}$
- f is an application $\mathcal{F} \otimes \mathcal{B}(\mathbb{R}) \otimes \mathcal{B}(\mathbb{R}^n)$ - progressively measurable
- $\forall t \in [0, T] : f(\cdot, t, 0, 0) \in \mathcal{H}^2(\mathbb{R})$, i.e. $f(\cdot, t, 0, 0)$ is $\mathcal{F}_t$ -progressive with $\mathbf{E}\int_0^T f^2(\cdot, t, 0, 0)dt < +\infty$
- f is uniformly Lipschitz in y and z, i.e. $\exists C > 0$ that $|f(\omega, t, y_1, z_1) - f(\omega, t, y_2, z_2)| \le C(|y_1 - y_2| + |z_1 - z_2|)$ $\forall y_1, y_2 \in \mathbb{R}, \ \forall z_1, z_2 \in \mathbb{R}^n, \ d\mathbf{P} \otimes dt$ a.s.

Generally, we denote as $\mathcal{H}^2(\mathcal{X})$ the set of stochastic processes $(\varphi_t)_{t\in[0,T]}$, $\mathcal{F}_t$ - progressive, with values in Banach space $\mathcal{X}$, satisfying $\mathbf{E}\int_0^T ||\varphi_t||_{\mathcal{X}}^2 dt < +\infty$.

The properties of standard parameters are sufficient conditions for the existence and uniqueness of the solution which is an assertion of the following theorem proved by Pardoux and Peng in [3].

Theorem 1: *Let (f, ξ) be standard parameters. Then the BSDE (1) has a unique solution $(Y_t, Z_t)_{t\in[0,T]} \in \mathcal{H}^2(\mathbb{R}^n) \times \mathcal{H}^2(\mathbb{R}^{n\times d})$.*

Idea of the proof: We define an application $\Phi : \mathcal{H}^2(\mathbb{R}^n) \times \mathcal{H}^2(\mathbb{R}^{n\times d}) \to \mathcal{H}^2(\mathbb{R}^n) \times \mathcal{H}^2(\mathbb{R}^{n\times d})$ so that $\Phi(U,V) = (Y,Z)$ where

$$\begin{aligned} -dY_t &= f(t,U_t,V_t)dt - Z_t dW_t, \quad \forall t \in [0,T) \text{ a.s.} \\ Y_T &= \xi, \text{ a.s.} \end{aligned} \tag{2}$$

To have Φ defined correctly, one must show that there exists a unique solution to (2) belonging to the product space $\mathcal{H}^2(\mathbb{R}^n) \times \mathcal{H}^2(\mathbb{R}^{n\times d})$. Note that in (2), the driver f does not depend on Y_t and Z_t.

Further, we realize that (Y,Z) solves (1) iff $\Phi(Y,Z) = (Y,Z)$ therefore, (Y,Z) is a fixed point of Φ (on a Banach space $\mathcal{H}^2(\mathbb{R}^n) \times \mathcal{H}^2(\mathbb{R}^{n\times d})$). It is possible to show that Φ is a contraction on $\mathcal{H}^2(\mathbb{R}^n) \times \mathcal{H}^2(\mathbb{R}^{n\times d})$ for the norm $||\cdot||_\beta$ where $\beta > 0$ is chosen properly and

$$||(Y,Z)||_\beta^2 = \mathbf{E}\int_0^T e^{\beta s}||Y_s||^2 ds + \mathbf{E}\int_0^T e^{\beta s}||Z_s||^2 ds.$$

Then the solution to BSDE (1) exists uniquely by the fixed point theorem.

Remark 1: 1) The process $(Z_t)_{t\in[0,T]}$, introduced by Theorem 1, ensures the adaptability of the process $(Y_t)_{t\in[0,T]}$.

2) The uniqueness of the solution means that if (Y_t, Z_t) and $(\tilde{Y}_t, \tilde{Z}_t)$ are two solutions to (1) then $\mathbf{E}\int_0^T ||Y_t - \tilde{Y}_t||^2 dt = \mathbf{E}\int_0^T ||Z_t - \tilde{Z}_t||^2 dt = 0$.

3) Since the process $(Y_t)_{t\in[0,T]}$ has continuous trajectories a.s., the space $\mathcal{H}^2(\mathbb{R}^n)$ in Definition 1 can be replaced with the space $\mathcal{S}^2(\mathbb{R}^n)$ which is a set of $\mathcal{F}_t$-adapted processes $(Y_t)_{t\in[0,T]}$ with $\mathbf{E}\big[\sup_{0\le t\le T} ||Y_t||^2\big] < +\infty$.

Theorem 1, in general, says nothing about the form of the solution even if it exists. Nevertheless, it is possible to express and compute it in some special cases. One such a case is a linear model, i.e. $f(t,Y_t,Z_t) = \beta_t Y_t + \gamma_t' Z_t + \varphi_t$ where $(\beta_t)_{t\in[0,T]}$ and $(\gamma_t)_{t\in[0,T]}$ are two processes $\mathcal{F}_t$ - progressively measurable, bounded, with values in $\mathbb{R}$ and $\mathbb{R}^n$, respectively. $(\varphi_t)_{t\in[0,T]}$ is a $\mathcal{F}_t$ - progressively measurable, $\mathbb{R}$-valued process, square-integrable. We suppose that Y_t and Z_t have corresponding dimensions, i.e. they are $\mathbb{R}$ and $\mathbb{R}^n$ - valued, respectively. Then we have, due to Pardoux and Peng [3],

Theorem 2: *The linear BSDE*

$$\begin{aligned} -dY_t &= (\beta_t Y_t + \gamma_t' Z_t + \varphi_t)dt - Z_t' dW_t, \quad \forall t \in [0,T) \text{ a.s.} \\ Y_T &= \xi \quad \text{a.s.} \end{aligned} \tag{3}$$

has a unique solution $Y_t = \mathbf{E}\big[H_T\xi + \int_t^T H_s\varphi_s ds|\mathcal{F}_t\big]$, $\forall t \in [0,T]$ a.s.,
where the process $(H_t)_{t\in[0,T]}$ *is a solution to the following SDE*

$$dH_t = H_t(\beta_t dt + \gamma_t' dW_t); H_0 = 1.$$

Remark 2: 1) The second solution process $(Z_t)_{t\in[0,T]}$ is obtained by applying the integral representation theorem for square-integrable continuous martingales (see e.g. [1]) to the martingale $M_t = Y_t + \int_0^t H_s\varphi_s ds$.

2.2 Example

To see one possible application of BSDE, we give a classical example. It concerns the hedging task for a European Call Option in a complete market.

We consider a financial market model with $n+1$ assets $(S^0, S^1, ..., S^n)$ whose price dynamics is given by the following SDE's

- $dS_t^0 = S_t^0 r_t dt$ (one non-risky asset)
- $dS_t^i = S_t^i\big(b_t^i dt + \sigma_t^i dW_t\big),\ i = 1, ..., n$ (n risky assets) where

$(r_t)_{t\in[0,T]}$, $(b_t)_{t\in[0,T]}$ and $(\sigma_t)_{t\in[0,T]}$ are $\mathbb{R}$, $\mathbb{R}^n$, $\mathbb{R}^{n,n}$ - valued bounded processes, $\mathcal{F}_t$ - progressive. Moreover, we assume that there is a bounded process $(\theta_t)_{t\in[0,T]}$ with values in $\mathbb{R}^n$. $(\theta_t)_{t\in[0,T]}$ is called *market price of risk* and it ensures the absence of arbitrage in the market.

The portfolio process π is an $\mathbb{R}^n$ - valued process, $\mathcal{F}_t$ - progressive whose i^{th} component π_t^i represents the amount invested into the i^{th} asset in time t. Moreover, we assume that $\mathbf{E}\int_0^T ||\sigma_t'\pi_t||^2 dt < +\infty$.

The wealth process $Y^{y_0,\pi}$, associated to the initial amount y_0 and the portfolio process π, is given as a solution to the following (forward) SDE

$$\begin{aligned} dY_t^{y_0,\pi} &= r_t Y_t^{y_0,\pi} dt + \pi_t'[b_t - r_t\mathbf{1}]dt + \pi_t'\sigma_t dW_t,\ t\in(0,T] \\ Y_0^{y_0,\pi} &= y_0,\ \text{a.s.} \end{aligned} \tag{4}$$

This approach is very intuitive for the wealth process simply expresses our wealth gained by applying our investment strategy π starting with an initial deposit y_0. What is, nevertheless, more interesting is a task of hedging a financial instrument, concretely a European Call option (EC), i.e. we look for an investment strategy π so that the terminal value Y_T^π of the corresponding wealth process would be equal to the EC pay-off which means $Y_T^\pi = (S_T - K)^+$ where K is an exercise price of the EC and S_T is the price of an underlying asset at time T. Less formally said, we can imagine EC pay-off as a random amount (contingent claim) which we will have to pay (cover) in the future (at time T). Our goal is to invest now (at $t_0 < T$) so that our wealth at time T is equal to that random amount. Formally, it means that we need to find a solution $(Y, Z) = (Y, \sigma'\pi)$ to the following BSDE

$$\begin{aligned} dY_t^\pi &= r_t Y_t^\pi dt + \pi_t'[b_t - r_t\mathbf{1}]dt + \pi_t'\sigma_t dW_t,\ t\in[0,T) \\ Y_T^\pi &= (S_T - K)^+,\ \text{a.s.} \end{aligned} \tag{5}$$

Then, if we assume,in addition, that the matrix σ is invertible, we can express our investment strategy as $\pi = \sigma^{-1}Z$.

2.3 Infinite time horizon and Lévy driven BSDE

Since the end of 1990's, there has been a huge progress in introducing jumps into BSDE models. First, just by considering an additional Poisson process but gradually, the theory was built up for general Lévy processes. The reason was, beside some specific physical tasks, that it was more and more clear that real financial asset prices do not follow normal (or better log-normal) distribution naturally obtained by using geometrical Brownian motion. Lévy-driven stochastic models were capable to improve (yet not to solve completely) the problem of heavy tails and to incorporate intuitively expected (and observed) jumps, see [4]. In this subsection we work only with $\mathbb{R}$- valued Lévy processes and we adopt the notation from [2].

Definition 2: *An adapted process $X = \big(X_t\big)_{t\geq 0}$ with $X_0 = 0$ a.s. is a Lévy process if*

1. *X has increments independent of the past, i.e. $X_t - X_s$ is independent of $\mathcal{F}_s$ for $0 \leq s < t < +\infty$; and*
2. *X has stationary increments, i.e. $X_t - X_s$ has the same distribution as X_{t-s} for $0 \leq s < t < +\infty$; and*
3. *X is continuous in probability, that is $\mathbf{P} - \lim\limits_{s\to t} X_s = X_t$.*

Remark 3: Since every Lévy process Y has a càdlàg modification X (i.e. right continuous with left limit) which is again Lévy process (see [2], Theorem 30), we will always work with this càdlàg process X.

When considering Lévy process in the model, one must specify what filtration is he or she using. In our case, we take a natural filtration of X, i.e. $\mathcal{F}_t^X = \sigma(X_s, s \leq t)$ and we proceed to completion and augmentation $\big(\mathcal{F}_t\big)_{t\geq 0}$ of the natural filtration. We lay $\mathcal{F}_\infty = \bigvee_{t\geq 0}\mathcal{F}_t \stackrel{def}{=} \sigma\big(\bigcup_{t\geq 0}\mathcal{F}_t\big)$.

Before pronouncing the existence and uniqueness theorem for infinite time horizon BSDE, we remind a crucial lemma due to Nualart and Schoutens in [8]. First, we denote as l^2 the space of real-valued sequences $(x_i)_{i\geq 1}$ such that $\sum_{i=1}^{+\infty}|x_i|^2 < +\infty$ and as $\mathcal{H}^2(l^2)$ we denote the space of l^2- valued predictable processes $\psi = \big(\psi_t\big)_{t\geq 0}$ such that

$$||\psi||^2_{\mathcal{H}^2(l^2)} = \mathbf{E}\int_0^{+\infty}\sum_{i=1}^{+\infty}|\psi_t^{(i)}|^2 dt, \tag{6}$$

Lemma 1: *Let X be a Lévy process whose associated Lévy measure ν fulfills*

1. *$\int_{\mathbb{R}}(1 \wedge z^2)\nu(dz) < +\infty$,*
2. *$\int_{(-\varepsilon,\varepsilon)^c} e^{\lambda|z|}\nu(dz) < +\infty$ for every $\varepsilon > 0$ and for some $\lambda > 0$.*

Then every square-integrable random variable $F \in L^2(\mathcal{F}_\infty)$ has a representation of the form

$$F = \mathbf{E}[F] + \int_0^{+\infty} \sum_{i=1}^{+\infty} \psi_t^{(i)} dH_t^{(i)}, \tag{7}$$

where $\left\{ \left(H_t^{(i)}\right)_{t\geq 0} \right\}_{i=1}^{+\infty}$ are strongly orthogonal martingales such that each $H^{(i)}$ is a linear combination of the Teugels martingales $Y^{(j)}$, $j = 1, ..., i$ associated to the Lévy process X.

Remark 4: See [8] and [2] for more details on this orthogonalization.

Using this representation result, it is sufficient to consider infinite time horizon BSDE of the following type

$$Y_t = \xi + \int_t^{+\infty} g(s, Y_{s_-}, Z_s) ds - \int_t^{+\infty} \sum_{i=1}^{+\infty} Z_t^{(i)} dH_t^{(i)}, \ \forall t \in [0, +\infty], \tag{8}$$

where the $\xi \in L^2(\mathcal{F}_\infty)$ and the function $g : \Omega \times [0, +\infty] \times \mathbb{R} \times l^2 \to \mathbb{R}$ fulfills

(A1): There exist two positive non-random functions $u(t) \in L^1([0, +\infty])$ and $v(t) \in L^2([0, +\infty])$ such that

$$|g(t, y_1, z_1) - g(t, y_2, z_2)| \leq v(t)|y_1 - y_2| + u(t)|z_1 - z_2| \ a.s., \quad \forall t \in [0, +\infty], (y_i, z_i) \in \mathbb{R} \times l^2, \ i = 1, 2 \tag{9}$$

(A2): $\left(g(t, y, z)\right)_{t\geq 0}$ is $\mathcal{F}_t$-progressively measurable $\forall (y, z) \in \mathbb{R} \times l^2$ with

$$\mathbf{E}\Big(\int_0^{+\infty} |g(t, 0, 0)| dt \Big)^2 < +\infty.$$

Definition 3: *A solution to BSDE (8) is a pair of processes $(Y, Z) \in \mathcal{S}^2(\mathbb{R}) \times \mathcal{H}^2(l^2)$ and satisfying (8).*

For definition of $\mathcal{S}^2(\mathbb{R})$ see Remark 1. Now we have all the tools to pronounce the existence and uniqueness theorem which is due to Zheng [7].

Theorem 3: *Let $\xi \in L^2(\mathcal{F}_\infty)$ and let g satisfy the assumptions (A1) and (A2). Then BSDE (8) has a unique solution.*

In the next section we show how BSDE naturally arise in the domain of optimal control having the meaning of conjugate variables ("generalized Lagrange multiplicators").

3 Stochastic control

3.1 Finite horizon control problem

Let $X_t^{t,x}$ be a controlled diffusion process in $\mathbb{R}^n$, i.e. $X_t^{t,x}$ is a solution to the (forward) SDE

$$dX_s^{t,x} = b(X_s^{t,x}, \alpha_s)ds + \sigma(X_s^{t,x}, \alpha_s)dW_s, \quad \forall s \in (t,T] \text{ a.s.} \tag{10}$$
$$X_t^{t,x} = x,$$

where $0 < T < +\infty$, $t \in [0,T)$, $x \in \mathbb{R}^n$, $\alpha = (\alpha_s)_{t \le s \le T}$ is an $\mathcal{F}_s$-progressively measurable A-valued control process, $A \subset \mathbb{R}^m$, $\big(W_s\big)_{s\in[t,T]}$ is an $\mathbb{R}^d$-valued standard Wiener process, $b : \mathbb{R}^n \times A \to \mathbb{R}^n$ and $\sigma : \mathbb{R}^n \times A \to \mathbb{R}^{n\times d}$ are two measurable functions satisfying a uniform Lipschitz condition in A, that means that there is a positive constant K so that

$$||b(x,a) - b(y,a)|| + ||\sigma(x,a) - \sigma(y,a)|| \le K||x-y||, \quad \forall x,y \in \mathbb{R}^n, \forall a \in A \tag{11}$$

Let us denote as $\mathcal{A}(t,x)$ the set of all admissible controls α such that

$$\mathbf{E}\Big[\int_t^T ||b(0,\alpha_s)|| + ||\sigma(0,\alpha_s)||^2 ds\Big] < +\infty \tag{12}$$

which ensures strong existence of the diffusion process X from (10).

Furthermore, let $f \in \mathcal{C}([0,T] \times \mathbb{R}^n \times A)$ and $g \in \mathcal{C}^1(\mathbb{R}^n)$ be two functions so that the following functional is meaningful (i.e. it converges)

$$J(t,x,\alpha) = \mathbf{E}\Big[\int_t^T f(s, X_s^{t,x}, \alpha_s)ds + g(X_T^{t,x})\Big], \tag{13}$$

and we define cost function $v(t,x)$ by

$$v(t,x) = \sup_{\alpha \in \mathcal{A}(t,x)} J(t,x,\alpha). \tag{14}$$

Our goal is to find such a strategy $\alpha^* \in \mathcal{A}(t,x)$ so that

$$v(t,x) = J(t,x,\alpha^*).$$

Let us define generalized Hamiltonian of the problem $\mathcal{H} : [0,T]\times\mathbb{R}^n\times A\times\mathbb{R}^n\times\mathbb{R}^{n\times d} \to \mathbb{R}$ by

$$\mathcal{H}(t,x,a,y,z) = b(x,a)'y + trace(\sigma(x,a)'z) + f(t,x,a).$$

We suppose that $\mathcal{H}$ is differentiable in x (with the gradient denoted as $\nabla_x\mathcal{H}$) and we consider the following BSDE

$$-dY_s = \nabla_x\mathcal{H}(s, X_s^{t,x}, \alpha_s, Y_s, Z_s)ds - Z_s dW_s, \quad \forall s \in [t,T) \text{ a.s.}$$
$$Y_T = \nabla_x g(X_T^{t,x}) \quad \text{a.s.} \tag{15}$$

Then we can formulate stochastic Pontryagin's maximum principle providing conditions on the optimal strategy α^*. The proof can be found in [6].

Theorem 4 (Stochastic Pontryagin's maximum principle): *Let $\hat{\alpha} \in \mathcal{A}(t,x)$ and $\hat{X}$ be the associated controlled diffusion process. Further, let us suppose that there exists a solution $(\hat{Y}, \hat{Z})$ to associated BSDE (15) such that*

1. $\mathcal{H}(t, \hat{X}_t, \hat{\alpha}, \hat{Y}, \hat{Z}) = \max_{a \in A} \mathcal{H}(t, \hat{X}_t, a, \hat{Y}, \hat{Z}), \quad \forall t \in [0,T]$ a.s.

2. $(x, a) \to \mathcal{H}(t, x, a, \hat{Y}, \hat{Z})$ *is a concave function for all* t.

Then $\hat{\alpha} = \alpha^*$, *i.e.* $\hat{\alpha}$ *is optimal control strategy to the stochastic control problem (14) which means* $v(t,x) = J(t,x,\hat{\alpha})$.

3.2 Lévy-driven stochastic control problem

The question now is if we are able to generalize the previous result to Lévy-driven stochastic control problems - both for finite and infinite time horizon. A positive answer to the first part of the question gives us the paper [5]. We note that in case of Lévy diffusion the model is

$$dX_s^{t,x} = b(s, X_s^{t,x}, \alpha_s)ds + \sigma(s, X_s^{t,x}, \alpha_s)dW_s + \int_{\mathbb{R}^n} \eta(s, X_{s_-}^{t,x}, \alpha_{s_-}, z)\bar{N}(ds, dz), \ \forall s \in (t,T] \text{ a.s.}$$
$$X_t^{t,x} = x. \tag{16}$$

The new term is an integral with respect to Poisson random measure

$$\begin{aligned} \bar{N}(ds, dz) &= \big(\bar{N}_1(ds, dz), ..., \bar{N}_l(ds, dz)\big)' \\ &= \big(N_1(ds, dz) - \chi_1(z)d\nu_1(z), ..., N_l(ds, dz) - \chi_l(z)d\nu_l(z)\big)', \end{aligned} \tag{17}$$

where $N_i(ds, dz),\ i = 1, ..., l$ are independent Poisson random measures with Lévy measures ν_i respectively, on a filtered probability space $\big(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \geq 0}, \mathbf{P}\big)$ satisfying the usual conditions. The indicator functions $\chi_i,\ i = 1, ..., l$ truncate the domain of "small and big jumps". Moreover, we assume that the control process α is predictable, left continuous with right limits. Hand in hand with these corrections, one must change the form of the generalized Hamiltonian to $\mathcal{H} : [0,T] \times \mathbb{R}^n \times A \times \mathbb{R}^n \times \mathbb{R}^{n \times d} \times \mathcal{R} \to \mathbb{R}$ so that

$$\begin{aligned} \mathcal{H}(t, x, a, y, z, r) &= b'(t,x,a)y + trace(\sigma'(t,x,a)z) + f(t,x,a) \\ &+ \int_{\mathbb{R}^n} trace\big(\eta'(t,x,a,z)r(t,z) \cdot diag(d\lambda(z))\big) \\ &+ \int_{\mathbb{R}^n} \Big[\big(\eta'(t,x,a,z)p + x'r(t,z)\big)\big(I - diag(\chi)\big)\Big]d\lambda(z), \end{aligned} \tag{18}$$

where $\mathcal{R}$ is the set of functions $r : \mathbb{R}^{n+1} \to \mathbb{R}^{n \times l}$ such that the integral in (18) converges. Again, we suppose that $\mathcal{H}$ is differentiable w.r.t. x.

Then the corresponding BSDE is of the form

$$\begin{aligned} -dY_t &= \nabla_x \mathcal{H}(t, X_t, \alpha_t, Y_t, Z_t, r(t,\cdot))dt + Z_t dW_t + \int_{\mathbb{R}^n} r(t_-, z)\bar{N}(dt, dz) \\ Y_T &= \nabla_x g(X_T). \end{aligned} \tag{19}$$

The assertion of the stochastic Pontryagin's maximum principle for this Lévy case is analogous to Theorem 4, see [5].

3.3 Infinite time horizon stochastic control problem

When considering infinite time stochastic control problem, it is useful to stress that, in fact, we are looking for a stationary optimal control α^*, that is we do not consider time dependence of functions b, σ and f.

Then the functional to maximize is

$$J(x,\alpha) = \mathbf{E}\Big[\int_0^{+\infty} e^{-\beta s} f(X_s^x, \alpha_s) ds\Big] \tag{20}$$

with the associated cost function

$$v(x) = \sup_{\alpha \in \mathcal{A}(x)} J(x,\alpha). \tag{21}$$

Again, the set of admissible controls $\mathcal{A}(x)$ is such that for all $\alpha \in \mathcal{A}(x)$ there exist a unique solution to (10) and the integral in (20) converges.

The question is, how the generalized Hamiltonian will look like when introducing also jumps in the model (by using Lévy processes) and what assumptions are needed to prove the associated Pontryagin's maximum principle. This is the goal of my current research.

The author wishes to thank to prof. Maslowski, prof. Vošvrda and to Dr. Šmíd for their help, guidance and encouragement.

References

[1] Karatzas I. and Shreve S. *Brownian motion and stochastic calculus, 2nd ed.* Springer-Verlag, 1988.

[2] Protter P. *Stochastic Integration and Differential Equations, 2nd ed.* Springer-Verlag, 1990.

[3] Pardoux E. and Peng S. *Adapted solutions of a backward stochastic differential equation. Systems and Control Letters*, **14**, 55–61, 1990.

[4] Tankov P. and Voltchkova E. *Jump-diffusion models: a practitioner's guide. Banque et Marchés*, **No. 99**, March-April 2009.

[5] Øksendal B., Sulem A. and Framstad N. C. *A sufficient stochastic maximum principle for optimal control of jump diffusions and applications to finance. J. Optimization Theory and Applications*, **121**, 77–98, 2004. Errata: *J. Optimization Theory and Applications* **124**, 511–512, 2005.

[6] Peng S. *A general stochastic maximum principle for optimal control problems. SIAM J. Control Optim.*, **28**, 966–979, 1990.

[7] Zheng S. *Infinite time interval BSDE's driven by a Lévy process.* `http://www.paper.edu.cn/index.php/default/en\_releasepaper/downPaper/201004-902`

[8] Nualart D. and Schoutens W. *Chaotic and predictable representations for Lévy processes. Stochastic Process. Appl.*, **90 (1)**, 109–122, 2000.

[9] Rong S. *On solutions of backward stochastic differential equations with jumps and applications. Stochastic Process. Appl.*, **66**, 209–236, 1997.

EEG Classification of Alzheimer's Disease Using Linear Predictive Model

Dagmar Zachová

1st year of PGS, email: zachovad@seznam.cz
Department of Software Engineering in Economy
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Jaromír Kukal, Department of Software Engineering in Economy,
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. The paper is oriented to EEG signal analysis, which is focused to quasi-stationarity hypothesis that the statistical properties of the channel signal fluctuate in time. Robust linear predictor is used for short segments of EEG as low-pass filter and the difference between the raw EEG and filter output was subject of statistical testing. Novelty is in the fluctuation measurement which enables to classify the Alzheimer's disease patients against controls.

Keywords: Alzheimer's disease, EEG, quasi-stationarity, linear predictor, robust identification

Abstrakt. Tento příspěvek je zaměřen na analýzu EEG signálu. Soustředí se na kolísání statistických vlastností v čase - hypotézu kvazistacionarity. Robustní lineární prediktor je použit jako nízkofrekvenční filtr na krátké segmenty EEG signálu. Předmětem statistického testování je rozdíl mezi skutečnými a predikovanými hodnotami EEG. Novinkou je měření této míry kolísání, což umožňuje klasifikovat pacienta s Alzheimerovou chorobou a zdravého jedince.

Klíčová slova: Alzheimerova choroba, EEG, kvazistacionarita, lineární prediktor, robustní identifikace

1 Introduction

Quasi-stationarity of EEG signal can cause difficulties in any signal processing of long sequences. If we disconnect the original series to short segments of constant length, we can use traditional methods of statistical analysis within any individual segment. Thus, the statistical properties of individual segments can be estimated correctly when the segment length is less then two seconds (in the case of EEG). But the statistical properties of segments vary in time due to the quasi-stationarity of EEG signal. The paper is oriented to statistical analysis of these fluctuations and its robust ranges.

2 Signal description

The multichannel EEG is a traditional tool for the investigation of human brain activity. The electrode signal was scrupled with constant frequency $f_s = 200$ Hz and then digitalized to the raw EEG time series X_k for $k = 1, 2, ..., L$.

The signal was partitioned to nonoverlapping segments of constant length $N \ll L$. Ideal signal should have stationarity property in the meaning that the statistical prop-

erties [6] of short segments don't vary in time. From the biomedical point of view, the stationarity of EEG is observable only for short sequences up to 2 seconds, thus for $N \ll L < 2f_s$. When the EEG scan is too long then the stationarity hypothesis falls. In this case, the EEG quasi-stationarity was subject of investigation. We used $N \ll 2f_s$ to guarantee interval stationarity of individual segments. Then the robust predictive filter was applied to every segment. The difference between the original data and the prediction was subject of statistical analysis. Various statistics of segment error sample were used and their values changed from segment to segment. Thus, the new time series of length $M = \lfloor L/N \rfloor$ of segment characteristics arisen and its members are R_k for $k = 1, 2, ..., M$. Statistical analysis of fluctuations is based on various statistical characteristics of R_k series. The process of EEG signal analysis consists of four steps:

- segmentation with X_k as result;
- within segment prediction with e_k as result;
- within segment error analysis with R_k as result;
- fluctuation analysis with Q_k as result.

3 Robust predictive model

We consider a basic linear model [4] in the form

$$Y_{k+S} = \sum_{j=1}^{H} \beta_j \varphi_j(Y_k, \ldots, Y_{k-H+1}) + \varepsilon_{k+S} \tag{1}$$

where

- N is the length of the time series segment (the number of observations);
- H is the history length of time series;
- S is the prediction step length;
- $Y_1, Y_2, \ldots, Y_N$ are observations within given segment;
- $\beta_1, \beta_2, \ldots, \beta_H$ are unknown coefficients (parameters) of the model;
- $\varphi_1(Y_k, \ldots, Y_{k-H+1})$, $\varphi_2(Y_k, \ldots, Y_{k-H+1})$, $\ldots$, $\varphi_H(Y_k, \ldots, Y_{k-H+1})$ are polynomial functions;
- ε_{k+S} is the random noise.

When we transcribe (1), we obtain an equation system that could be described in matrix form as

$$\begin{bmatrix} Y_{H+S} \\ Y_{H+1+S} \\ . \\ . \\ . \\ Y_N \end{bmatrix} = \begin{bmatrix} \varphi_{1,2,\ldots,H}(Y_H, \ldots, Y_1) \\ \varphi_{1,2,\ldots,H}(Y_{H+1}, \ldots, Y_2) \\ . \\ . \\ . \\ \varphi_{1,2,\ldots,H}(Y_{N-S}, \ldots, Y_{N-S-H+1}) \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_H \end{bmatrix} + \begin{bmatrix} \varepsilon_{H+S} \\ \varepsilon_{H+1+S} \\ . \\ . \\ . \\ \varepsilon_N \end{bmatrix}, \tag{2}$$

in other words

$$\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{3}$$

It is significant that the number of equations (degrees of freedom) must be higher than the number of estimated coefficients, i.e. $N - H - S + 1 > H$. Further, supposed that $\mathrm{E}(\boldsymbol{\varepsilon}) = \bar{0}$, where symbol E indicates the expected value. Providing this we can express estimated values Y_{k+S} (for $k = H,\ H+1,\ \ldots,\ N-S$) through the following formula

$$\mathrm{E}(Y_{k+S}) = \sum_{j=1}^{H} \beta_j \varphi_j(Y_k, \ldots, Y_{k-H+1}). \tag{4}$$

These estimated values are equal to functional values of selective regression function

$$\hat{Y}_{k+S} = \sum_{j=1}^{H} b_j \varphi_j(Y_k, \ldots, Y_{k-H+1}) \tag{5}$$

where

- b_j is the scatter estimate of unknown parameter β_j (for $j = 1, 2, ..., H$);
- $\hat{Y}_{k+S}$ is the predicted value Y_{k+S} (for $k = H,\ H+1,\ \ldots,\ N-S$).

Equation system (5) can be described in matrix form as

$$\begin{bmatrix} \hat{Y}_{H+S} \\ \hat{Y}_{H+1+S} \\ . \\ . \\ . \\ \hat{Y}_N \end{bmatrix} = \begin{bmatrix} \varphi_{1,2,\ldots,H}(Y_H, \ldots, Y_1) \\ \varphi_{1,2,\ldots,H}(Y_{H+1}, \ldots, Y_2) \\ . \\ . \\ . \\ \varphi_{1,2,\ldots,H}(Y_{N-S}, \ldots, Y_{N-S-H+1}) \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ . \\ . \\ . \\ b_H \end{bmatrix}, \tag{6}$$

in other words

$$\hat{\boldsymbol{y}} = \boldsymbol{\Phi}\boldsymbol{b}. \tag{7}$$

We can use robust methods for the coefficient estimating of model (3), i.e. for vector $\boldsymbol{b}$ specification (see 3.1).

The difference between observed and predicted value is called residue and denoted as vector $\boldsymbol{e}$. The residue in given point is equal to $e_i = \hat{Y}_i - Y_i$, therefore for the model (1) the residual vector has the form of

$$\boldsymbol{e} = \begin{bmatrix} e_{H+S} \\ e_{H+1+S} \\ . \\ . \\ . \\ e_N \end{bmatrix}. \tag{8}$$

3.1 Robust identification techniques

Robust techniques of parameter estimating represent the alternative to classical statistic methods that are very sensitive to outliers in input data. We know several types of robust estimates, namely: L-estimates, R-estimates and M-estimates. It is most suitable to apply M-estimates, the pioneer of which was Huber [2]. M-estimate of model coefficients $\boldsymbol{\beta}$ is defined via function minimization (with respect to $\boldsymbol{b}$)

$$\psi(\boldsymbol{b}) = \sum_{i=H}^{N-S} \rho\left(\frac{e_{i+S}}{\sigma}\right) = \sum_{i=H}^{N-S} \rho\left(\frac{Y_{i+S} - \boldsymbol{\Phi}_{i-H+1}^{\mathrm{T}}\boldsymbol{b}}{\sigma}\right) \tag{9}$$

where

- T is transposition symbol;
- ρ is a penalty function (see Tab. 1);
- σ is standard deviation;
- $\boldsymbol{\Phi}_i$ is i-th row of the matrix $\boldsymbol{\Phi}$.

When implementing the weight function defined as $\mathrm{w}(\xi) = \frac{\mathrm{d}\,\rho(\xi)}{\mathrm{d}\,\xi}\frac{1}{\xi}$ (see Tab. 1), satisfying $\mathrm{w}(0) = 1$ and substituting to the Taylor series of (9) we obtain a method of weighted least squares (WLS) [4]

$$\sum_{i=\mathrm{H}}^{N-S} \mathrm{w}\left(\frac{e_{i+S}}{S_N}\right) Y_{i+S}\Phi_{(i-H+1)j} = \sum_{i=H}^{N-S}\sum_{k=1}^{H} \mathrm{w}\left(\frac{e_{i+S}}{S_N}\right) \Phi_{(i-H+1)j}\Phi_{(i-H+1)k}b_k$$

where $j = 1, \ldots, H$. The method of WLS consists in implementation of the following operations:

1. initial estimate of $\boldsymbol{b}$ by means of method of least squares, iteration counter set to $l = 1$.
2. residue specification $\boldsymbol{e}$ in lth iteration;
3. calculation of weights and then $l = l + 1$;

4. specification of parameters $\boldsymbol{b}^{(l)}$ (estimate of vector $\boldsymbol{b}$ in lth iteration) and residue specification.

If the estimates $\boldsymbol{b}^{(l)}$ a $\boldsymbol{b}^{(l-1)}$ are not close enough, we repeat the steps 3 and 4. It is important when calculating the balance in step 3 that the robust estimate of standard deviation σ is not recalculated, i.e. it's specified on the basis of error residue $\boldsymbol{e}$ after the least squares method application. Such a $\boldsymbol{b}$, by which the penalty function reached the lowest value, is considered as the best estimate of parameter $\boldsymbol{\beta}$.

The question is how to get the robust estimate of standard deviation σ. There is statistics $\sigma^* = MAD_{\mathrm{E}}/0.6745$ most frequently used in practice, where MAD_{E} stands for median of E_1, E_2, ..., E_N and $E_i = \left|e_i - \widetilde{E}\right|$, $\widetilde{E}$ is median of e_1, e_2, ..., e_N.

Table 1: Robust approaches

method	$\rho(\xi)$	w (ξ)	range	constant
Tukey	$B^2\left(1-\left(1-(\xi/B)^2\right)^3\right)/6$	$\left(1-(\xi/B)^2\right)^2$	$\lvert\xi\rvert \le B$	B=4.865
	$B^2/6$	0	$\lvert\xi\rvert > B$	
Huber	$\xi^2/2$	1	$\lvert\xi\rvert \le k$	k=1.345
	$k\lvert\xi\rvert - k^2/2$	$k/\lvert\xi\rvert$	$\lvert\xi\rvert > k$	
Andrews	$A^2\left(1-\cos(\xi/A)\right)$	$(A/\xi)\sin(\xi/A)$	$\lvert\xi\rvert \le A\pi$	A=1.339
	$2A^2$	0	$\lvert\xi\rvert > A\pi$	
Welsch	$W^2\left(1-\exp\left(-(\xi/W)^2\right)\right)/2$	$\exp\left(-(\xi/W)^2\right)$	—	$W = 2.985$
Talwar	$\xi^2/2$	1	$\lvert\xi\rvert \le k$	$k = 2.795$
	$k^2/2$	0	$\lvert\xi\rvert > k$	

3.2 Statistical analysis of prediction error and time fluctuation

Let us have signal of length L, divided into segments of fixed length N and values H, S being set. Afterwards, we effect suitable robust identification of model (1), coefficient and indicate residue vector $\boldsymbol{e} = (e_1, e_2, \ldots, e_p)^{\mathrm{T}}$. Now, it's time to think of how to characterize error prediction in one segment and how best to characterize variability of error prediction of the whole signal in time. In kind of criterion featuring as total error prediction in one segment the following two characteristics can be used. The first one can be described through the relation

$$R = \left(\mathrm{E}\,|\boldsymbol{e}|^q\right)^{1/q} = \left(\frac{1}{p}\sum_{k=1}^{p}|e_k|^q\right)^{1/q} \tag{10}$$

where $q \in \langle 0, \infty)$ a $p = N - H - S + 1$. Let's identify this method as a method of root of expected value of residues (MREVR(q)).

In order to make description of the second characteristic easier let us set $a_k = |e_k|$ and let us arrange a_k in such a way that $a_{(1)} \le a_{(2)} \le \ldots \le a_{(p)}$. Afterwards, the total signal error prediction in one segment will be calculated as

$$R = a_{(\lfloor qp \rfloor)} \tag{11}$$

where parameter $q \in (0, 1\rangle$ and $p = N - H - S + 1$. This approach we call the method of quantiles of the residues (MQR(q)).

Thus, we get for each channel time series prediction errors $\{R_1, R_2, \ldots, R_M\}$, respectively structured selection $\{R_{(1)}, R_{(2)}, \ldots, R_{(M)}\}$ where $M = \lfloor L/N \rfloor$. For assess the variability of the prediction errors EEG signal in time can be used such as one of the following sample (segment) characteristics :

- maximum $R_{\max} = \max\{R_1, R_2, \ldots, R_M\}$;
- minimum $R_{\min} = \min\{R_1, R_2, \ldots, R_M\}$;
- range $R_R = R_{\max} - R_{\min}$;
- mean $\bar{R} = \frac{1}{M}\sum_{k=1}^{M} R_k$;
- standard deviation $\sigma = \sqrt{\frac{1}{M-1}\sum_{k=1}^{M}\left(R_k - \bar{R}\right)^2}$;
- median $\widetilde{R} = \frac{1}{2}(R_{(M/2)} + R_{(M/2+1)})$ for the even M, respectively $\widetilde{R} = R_{((M+1)/2)}$ for the odd M;
- median absolute deviation $MAD_Z = \widetilde{Z}$ where $\widetilde{Z}$ stands for median of Z_1, Z_2, ..., Z_M and $Z_i = \left|R_i - \widetilde{R}\right|$ for $i = 1,\ 2,\ ...,\ M$;
- 1st quartile (lower quartile) $R_{0.25} = R_{(\lfloor 0.25M \rfloor)}$;
- 3rd quartile (upper quartile) $R_{0.75} = R_{(\lfloor 0.75M \rfloor)}$;
- interquartile range $IQR = R_{0.75} - R_{0.25}$.

These aggregating characteristics will be denoted as Q in the next text.

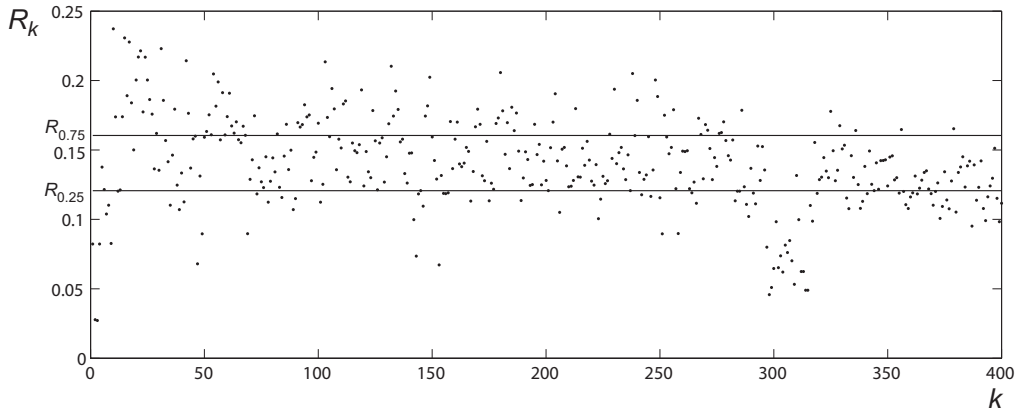


Figure 1: Fluctuation of MREVR in time for a healthy person ($IQR = R_{0.75} - R_{0.25}$)

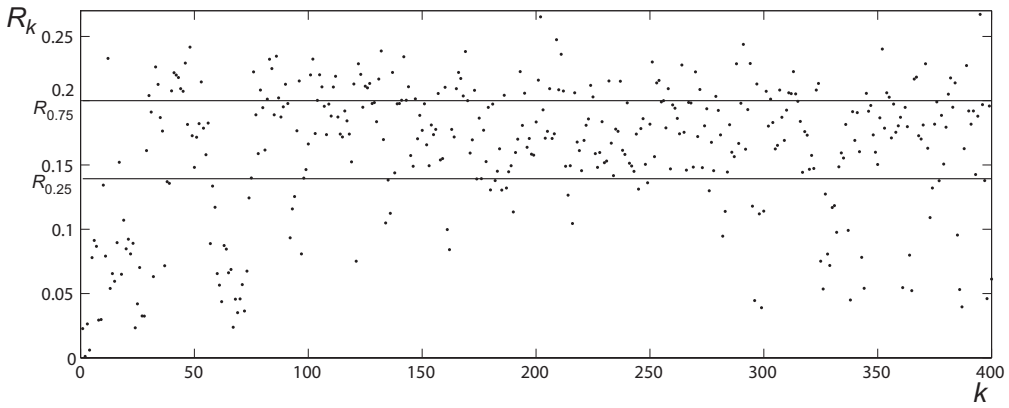


Figure 2: Fluctuation of MREVR in time for a patient with Alzheimer's disease ($IQR = R_{0.75} - R_{0.25}$)

3.3 Quality of classification

There is a direct relationship between the quality of parameter setting and the quality of classification. In our case, the optimal parameter setting has greatest differences in Q between groups AD and CN. The quality of parameter setting was driven by the apparatus of statistical hypothesis testing. Variability of the prediction error we calculated for each channel and each person. Thus, there are two samples $Q_{AD} := \{Q_1^{AD}, Q_2^{AD}, \ldots, Q_n^{AD}\}$ and $Q_{CN} := \{Q_1^{CN}, Q_2^{CN}, \ldots, Q_m^{CN}\}$ where n and m indicate the number of individuals in AD and CN groups. The null and the alternative hypothesis were constructed as follows:

H_0: expected value of random variables Q_{AD}, Q_{CN} are not different, i.e. $\mu_{AD} = \mu_{CN}$;

H_1 : expected value of random variables are different, i.e. $\mu_{AD} \neq \mu_{CN}$.

Assuming equal variances in both groups, we can use the two-sample two-sided t-test [4], where the test criterion is calculated as

$$T = \frac{\overline{Q_{AD}} - \overline{Q_{CN}}}{\sqrt{(n-1)\sigma_{AD}^2 + (m-1)\sigma_{CN}^2}} \sqrt{\frac{mn(n+m-2)}{n+m}}. \tag{12}$$

$\overline{Q_{AD}}$ and $\overline{Q_{CN}}$ denote the sample means, σ_{AD}^2 a σ_{CN}^2 are sample variances.

The criterion (12) has Student's t-distribution with $df = n+m-2$ degrees of freedom. We calculated adequate p-value for given T and df.

Another possible tool for assessing the quality of classifiers is the sensitivity and specificity. Sensitivity reflects the probability of correct classification of positive sample (AD) and specificity reflects the probability of correct classification of negative sample (CN).

Let TP (true positive) be number of samples that the classifier correctly classified into AD, let FP (false positive) be number of samples that the classifier incorrectly classified into AD, let TN (true negative) be number of samples that the classifier correctly classified into CN and let FN (false negative) be number of samples that the classifier incorrectly classified into CN. The sensitivity and the specificity can be estimated as follows:

- sensitivity (*true positive fraction*) $TPF=\frac{TP}{TP+FN}$;
- specificity (*true negative fraction*) $TNF=\frac{TN}{FP+TN}$.

The optimum, threshold for AD / CN classification is obtainable from ROC curve [1] as compromise between maximum values of TPF and TNF. We prefer to maximize $\min(TPF, TNF)$ according to minimax decision principle.

4 Results

There were 32 EEG records included in our study. The groups of AD and CD consist of 16 and 16 patients. We used international 10-20 electrode system [5]. During the measurement of electrical activity, our testers were in the bed having with closed eyes and without any stimulus. EEG data were approximately 300 seconds long with sampling frequency of 200 Hz. Electric potential was measured in millivolts.

During computer experiments, which we aimed to optimum parameter setting, we used model (1) with fixed functional base $\varphi_j(Y_k, \ldots, Y_{k-H+1}) = Y_{k-j+1}$ for $j \in \{1, 2, ..., H\}$. The following procedure was used:

- signal was divided into segments ($N = 100$);
- standardization of each segment was performed $Y_i^* = \frac{Y_i - \overline{Y}}{\sigma}$;
- default values of model parameter were used ($H = 10$, $S = 1$);
- Tukey's method was used as default robust method;
- default value for MREVR was $q = 2$;
- default value for MQR was $q = 1/2$;
- two most suitable channels were chosen on the basis of two-sample two-sided t-test at significance level of 0.05;
- with the help of p-value, optimal values for parameters N, H, S and q were found, and most suitable robust method was chosen.

Results of numerical calculations are included in the Tab. 2 using default parameters. Bold font was used for p-value below critical probability (0.05). The best in AD / CN resolution are channel 2 and 6, which were subject consequential analysis. The second aim was to study the influence of processing parameters (N, H, S) to p-value. Following parameter values were involved in the combination with Tukey's method:

- length of the segment N=100, 125, 150, 200;
- history length of time series H=6, 8, 10;
- length of the prediction step S=1, 2, 3.

The results of testing are summarized in the Tabs. 4, 5. The best results were obtained for $N = 150$, $H = 8$ or 10, $S = 1$ or 2 in the case of Turkey's method and channels 2 and 6. The parameter setting was then used for the other methods and channels. As seen in the Tab. 3, the p-values of robust methods are lower than in the squares approach (LSQ) in the case of channel 6. Similar result (except Andrew's and Huber's method) is valid in the case of channel 2 (see Tab. 3). The method MREVR is recommended for the segment error evaluation. The methods MAD and IQR are the best for the fluctuation analysis.

Table 2: Minimum p-values for the default setting

channel	p-value	characteristic	method (q)
1	**0.003301**	IQR	MQR(1/2)
2	**0.000201**	IQR	MREVR(2)
3	0.070778	$R_{\min}$	MQR(1/2)
4	0.077238	IQR	MREVR(2)
5	**0.013298**	IQR	MREVR(2)
6	**0.001757**	IQR	MREVR(2)
7	**0.002733**	MAD	MQR(1/2)
8	0.182820	$R_{\min}$	MQR(1/2)
9	0.081693	$R_{\min}$	MREVR(2)
10	0.118012	IQR	MREVR(2)
11	0.113751	IQR	MQR(1/2)
12	**0.045587**	IQR	MKR(1/2)
13	**0.012805**	R_R	MREVR(2)
14	**0.047399**	$R_{\min}$	MREVR(2)
15	0.052538	$R_{\min}$	MKR(1/2)
16	0.378503	$R_{\min}$	MREVR(2)
17	0.231664	$R_{\min}$	MREVR(2)
18	0.101802	σ	MREVR(2)
19	0.137812	$R_{\min}$	MREVR(2)

Table 3: Minimum p-values for the 6^{th} channel and different robust methods

robust method	p-value	q	method	$N - H - S$	characteristic
LSQ	0.000339	3/2	MREVR	150 - 8 - 2	IQR
Tukey	0.000326	2	MREVR	150 - 8 - 2	IQR
Andrews	0.000315	2	MREVR	150 - 8 - 2	IQR
Huber	0.000212	9/4	MREVR	150 - 8 - 2	IQR
Welsch	0.000299	2	MREVR	150 - 8 - 2	IQR
Talwar	0.000337	5/4	MREVR	150 - 8 - 2	IQR

Table 4: Minimum p-values for the 2^{nd} channel and Tukey's method

N	p-value	H	S	characteristic	method(q)
100	0.000236	10	2	$R_{\min}$	MREVR(2)
	0.000320	8	1	IQR	MQR(1/2)
125	0.000261	10	1	IQR	MREVR(2)
	0.000283	10	2	σ	MQR(1/2)
150	0.000071	10	1	IQR	MREVR(2)
	0.000225	8	1	IQR	MQR(1/2)
175	0.000127	10	1	IQR	MREVR(2)
	0.000427	10	1	IQR	MQR(1/2)
200	0.000171	10	1	MAD	MREVR(2)
	0.000495	8	1	IQR	MQR(1/2)

Table 5: Minimum p-values for the 6^{th} channel and Tukey's method

N	p-value	H	S	characteristic	method (q)
100	0.000683	8	1	IQR	MREVR(2)
	0.001112	8	1	IQR	MQR(1/2)
125	0.001047	8	1	IQR	MREVR(2)
	0.002533	10	1	σ	MQR(1/2)
150	0.000326	8	2	IQR	MREVR(2)
	0.000764	8	1	IQR	MQR(1/2)
175	0.000591	8	2	MAD	MREVR(2)
	0.002549	8	1	IQR	MQR(1/2)
200	0.001146	6	3	$R_{0.25}$	MREVR(2)
	0.004374	8	1	IQR	MQR(1/2)

Table 6: Minimum p-values for the 2^{nd} channel and different robust methods

robust method	p-value	q	method	$N-H-S$	characteristic
LSQ	6.59 X 10^{-5}	7/8	MQR	150 - 8 - 1	IQR
Tukey	6.32 X 10^{-5}	9/4	MREVR	150 - 10 - 1	IQR
Andrews	6.74 X 10^{-5}	9/4	MREVR	150 - 10 - 1	IQR
Huber	6.66 X 10^{-5}	9/4	MREVR	150 - 10 - 1	IQR
Welsch	6.11 X 10^{-5}	9/4	MREVR	150 - 10 - 1	IQR
Talwar	6.47 X 10^{-5}	7/8	MQR	150 - 8 - 1	IQR

5 Conclusion

Robust linear predictive filter was used for the characterization of signal variability within individual segments. The quasi-stationarity analysis is recommended as a tool for the classification of Alzheimer's disease against controls. The best results were obtained on EEG channel 2 with sampling period 200 Hz, segment length $N = 150$, history depth $H = 10$, step of prediction $S = 1$, Welsch's method, MREVR (method of root of expected value of residues) characteristics of EEG fluctuations. Then, the adequate optimum values are:

- p-value p-value=6.11 X 10^{-5};
- sensitivity $TPF = 81.3$;
- specificity $TNF = 87.5$.

From the biomedical point of view the novel method is comparable with the other complex methods of Alzheimer's disease diagnosis.

References

[1] T. Fawcett. *An Introduction to ROC analysis.* In 'Pattern Recognition Letters (Amsterdam, 2006)', volume 27, Elsevier Science, pp.861-876.

[2] P. J. Huber, E. M. Ronchetti *Robust Statistic.* John Wiley & Sons, Hoboken, (2009).

[3] R. G. Lehr, A. Pong *ROC Curve.* In 'Encyclopedia of Biopharmaceutical Statistics (New York, 2003)', Marcel Dekker, pp.884-891.

[4] M. Meloun, J. Militský *Statistická analýza experimentálních dat.* Academia, Praha, (1998).

[5] E. Niedermeyer, F. Lopes da Silva *Electroencephalography: Basic Principles, Clinical Applications, and Related Fields.* Lippincott Williams & Wilkins, Philadelphia, (2004).

[6] M. B. Priestley *Non-linear and Non-stationary Time Series Analysis.* Academic Press, London, (1988).

Comparison of Trading Algorithms*

Jan Zeman

4th year of PGS, email: janzeman3@seznam.cz
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: T. V. Guy, Institute of Information Theory and Automation, ASCR

Abstract. The paper continues the previous research aimed at design the automatic trading system. The paper concerns rating the quality of designed approaches. It reviews both general methods and methods specialized to trading. The proposed method is a combination of them.

Keywords: approaximative dynamic programming, Bellman function

Abstrakt. Článek navazuje na předchozí výzkum týkající se obcodování s futures. Téma je zaměřeno na hodnocení dříve navržených algoritmů. Článek reviduje hodnotící metody jak obecné tak zaměřené na problematiku obchodvání. Výsledkem je kombinovaná metoda, která je testována a hodnocena v závěrečné části.

Klíčová slova: přibližné dynamické programování, Belllmanova funkce

1 Introduction

The paper towards automatic trading system for the futures contracts. The previous research concerns the task definition and basic solution [3, 4]. The previous work proposed many approaches and we have to compare them in order to select the most suitable one. Two subtask are considered: First is how to recognize the good approach standalone, and second deals with comparison of two approaches and selecting the better one.

To recognize a good approach, a final profit can be used as the measure of a success. However in trading applications, the continuous development of the cumulative profit has higher impact than the final profit. The analyzing the cumulative is more complex due to working with the whole sequence, but can bring better insight to approach quality.

The comparison of two approaches seems to be easy, when the approaches are tested on common data set. When even more data sets are available, the comparison becomes complex, because each data set produces one dimension in results, then the comparison of multidimensional results is needed. The typical problem is: Approach A makes a total profit at five data sets $ 100000 USD, but profit was positive at only two data sets. Approach B makes a total profit only $ 50000 USD, but it makes positive profit at four of five data sets. Which approach is better? Both approaches can win, but the best should be chosen according to the preference of trader.

The paper proposes a small review of the comparison methods and applies the methods to one of the solved problems.

The paper contains two main parts. Section 2 introduces the problematics and defines the task (Sec. 2.1), defines a coefficient characterizing the quality of approach using the

*This work has been supported by the grant MŠMT 1M0572.

cumulative gain (Sec. 2.2) and introduces methods for multi-dimension comparing (Sec. 2.3). Section 3 introduces futures trading (Sec. 3.1) and coefficients used in trading (Sec. 3.2), defines algorithm of approaches rating (Sec. 3.3). The algorithm is applied and commented in Sec. 3.4.

2 Comparing methods

The section deals with definition of the solved task and given assumptions.

2.1 Task of interest

We assume a decision maker and system. Decision maker is human or machine with aims related to the system. The decision maker obtains a data y_t at the system, and design the decision u_t to reach his aims. The process is repeated each discrete time instant $t \in \{1, \ldots, T\}$. The aims of decision maker are characterized by a gain function G, which maps the system output and decisions to a real number. Higher value indicates higher success. The decision maker tries to maximize the gain function.

We focus on quality evaluation of designed decisions, hence we assume the knowledge of a whole data $y_1, \ldots, y_T$ and decision sequence $u_1, \ldots, u_T$. Moreover, we assume the knowledge of the gain function:

$$G : (y_1, \ldots, y_T, u_1, \ldots, u_T) \to \mathbb{R} \tag{1}$$

and its additive shape

$$G = \sum_{i=1}^{T} g_i, \qquad \text{where} \qquad g_i : (y_1, \ldots, y_i, u_1, \ldots, u_i) \to \mathbb{R}, \tag{2}$$

and g_i is called a one-step gain.

Let us define *cumulative gain* via:

$$G^t = \sum_{i=1}^{t} g_i. \tag{3}$$

The gain is a sum over all time instants $\{1, \ldots, T\}$, whereas cumulative gain is sum over the first t time steps $\{1, \ldots, t\}$, $t \leq T$. Hence, we use the term *final gain* for the gain from here onward. Moreover, the cumulative gain can be viewed as a sequence $G^1, \ldots, G^T$ and characterizes the approach behavior.

We assume that there are M different approaches trying to maximize the gain (2) and N testing data sets or experiment data available to compare the success of the approaches. In summary, we have $M \times N$ final gains to decide, which approach is the best. Moreover, we can obtain $M \times N \times T$ values, in order to analyze the approaches using the cumulative gains.

2.2 Cumulative gain comparison

It is disputable, whether the final gain is a good criterion for rating of the approaches. In some tasks, the good final gain can be reached only by a few last steps, hence the analysis of the cumulative gain is required. But working with a whole sequence of cumulative gain containing T values is difficult. Hence, it is needed to characterize the quality of cumulative gain by one coefficient, and this section defines such a coefficient.

The ideal cumulative gain increases, therefore the knowledge of a trend is important. To reach this knowledge, the sequence can be fitted by a linear function $y(t) = at + b$, where a, b are parameters. We assume a sequence of values $G^1, G^2, \ldots, G^T$, and we search the best values of coefficients a, b to minimize squared error $\min_{a,b} \sum_{t=1}^{T} (G^t - y(t))^2$. The obtained coefficients a_{min}, b_{min} characterize the nearest linear approximation of the original sequence. Hence, the values of a_{min}, b_{min} can be used to evaluate the success of the approach.

The coefficient a_{min} reflects a trend of cumulative gain. The positive value characterizes an increase, the negative one a decrease. The value of coefficient a is related to strength of the increase, higher value means sharper increase. Thus, it can be used as a relatively good criterion of the approach quality.

On the other hand, the linear approximation is not suitable, when the difference between original sequence and approximation $(G^t - a_{min}t - b_{min})$ is not normal distributed. This property cannot be warranted by any cumulative gain. Hence, the credibility of the coefficient a_{min} is lowered. The credibility of coefficient a_{min} is given by value of error squares $s = \sum_{t=1}^{T} (G^t - a_{min}t - b_{min})^2$, the less value of s brings better credibility of a_{min}. To obtain one characteristic coefficient, let us define *increase coefficient* c_I as follows:

$$c_I = \frac{a}{\log_{10}(s)}, \quad \text{with} \quad s = \sum_{t=1}^{T} (G^t - at - b)^2, \tag{4}$$

where a_{min}, b_{min} are coefficients of the best linear approximation of the cumulative gain sequence. The logarithm is used due to big differences in values of s for the trading task.

The higher value of c_I is rated as better result of an approach. The positive value of coefficient c_I characterizes the increase of cumulative gain, the weighting by difference s lowers the value of coefficient for bad fitted sequences. The coefficient c_I covers our requirements for working with cumulative gain, hence the further sections deals with comparing results obtained on more data sets.

2.3 Multi-dimension comparing

As was introduced, the comparison of two approach is simply, when they are tested at one data set, but when more data set is available, the decision become complex. The complexity originates from fact that the comparison has nature of multidimensional task, where each data set forms one dimension of compared vectors. Following two subsections deals with this task. Section 2.3.1 try to transform the multidimensional task to one-dimensional by weighted summing. Whereas, the Section 2.3.2 let the task multidimensional and defines comparison of vectors.

Analogical with Sec. 2.1, we assume M approaches and N testing data sets. The aim is select the best approach, hence we form M vectors $R^1, \ldots, R^M$ containing the results,

which are quality measures related to each data sets. The quality measures can be final gain, increase coefficient, or other variable characterizing the approach quality. Thus, each vector contains N values $R^i = (r_1^i, \ldots, r_N^i)$. Our aim is to chose the best approach using only this vectors.

2.3.1 Weighted sum

The first simply solution is to summarize the results and evaluate

$$\mathcal{S}^m = \sum_{n=1}^{N} r_n^m$$

for each approach $m \in \{1, \ldots, M\}$. Then each approach is characterized by one real number and it is simple to compare them.

Summing the results is simply and effective, but has a lot of disadvantages. When one of data sets produces outstanding results, the total sum is influenced by this outlayer and the results are not correct. Moreover, the maximal obtainable results must be comparable for all data sets, because the higher potential gives higher weight to given data set. The maximal and minimal possible value of results can be calculated for some special tasks and using them the following coefficient can be defined:

$$FP_n^m = \frac{r_n^m - G_n^{min}}{G_n^{max} - G_n^{min}} \times 100\%, \tag{5}$$

where G_n^{min} and G_n^{max} are minimal and maximal result values obtainable at nth data set. Let the coefficient is called *final percentage*. The final percentage express the percentage of success reached by approach according to maximal and minimal potential results reachable on the given data set. Summing FP_n^m over $n \in \{1, \ldots, N\}$ brings the equivalent results, where each experiment has the same weight independent on its potential. Instead of summing, it is better to calculate the mean value:

$$MFP^m = \frac{1}{N} \sum_{n=1}^{N} FP_n^m \tag{6}$$

the results can be interpreted as mean potential percentage of the approach m. Let coefficient MFP^m is called *mean final percentage*. The coefficient (6) is generalized weighted sum. When the minimal results potential equals zero ($G_n^{min} = 0$), then it is equivalent to weighted sum with weights: $w_n = 1/G_n^{max}$.

The coefficient MFP assigns each approach one number and the searching the best approach is transformed to sorting the number.

2.3.2 Efficient solution

Another way to compare the vectors $R^1, \ldots, R^M$ is by defining dominating and efficient solution.

The vector $R^i = (r_1^i, \ldots, r_N^i)$ is *dominated* by vector $R^j = (r_1^j, \ldots, r_N^j)$ even if following inequalities are valid:

$$\forall n \in \{1, \ldots, N\} \quad r_n^i \leq r_n^j,$$

and

$$\exists n \in \{1, \ldots, N\} \quad r_n^i < r_n^j.$$

Efficient solution is such a vector from the set $\{R^1, \ldots, R^M\}$, which is not dominated by any other vector. The term of efficient solution is taken from multiobjective optimization [1].

Taking only efficient solutions, the set of outstanding solutions can be found. The efficiency does not mix results reached on different data sets, i.e. the outstanding results on one data set cannot help the approach rating such as in poor summing the gains.

On the other hand, the efficient solutions typically forms a subset of $\{R^1, \ldots, R^M\}$. Hence, the method does not lead to one best approach, but it excludes a small set of outstanding approaches. The method cannot prefer one of efficient solutions, until the additional information about preferences is not added.

3 Example: commodity futures trading

The commodity futures trading is challenging task related to trading on stock exchanges and prices speculation. The commodity futures means an contract for delivering the commodity to given date in future. The price of contract is often object of speculation.

The speculator can speculate for following situations:

Price increase, the speculator buys the contract, it is said to open the *long* position. Then, he waits, until the price increases, and sells the contract (it is said to close the long position).

The profit is the difference of buy/sell contract price. The difference, whether speculator makes profit or loss, depends, whether the price follows his expectation. Hence, the profit from the long position is made, when the price increases, whereas the speculator loses the same value, when the price decreases.

Price decrease, the speculator sells the contact, it is said to open the *short* position. The fact, that he can sell not-owned contract, is related to principles of given exchange, the speculator can lend the contract for this operation. Then, he will buy the contract back, it is said to close the short position.

Indefinite, the speculator has no opened position. He is in so called *flat* position, or *out of market*. Speculator neither profits nor loses by this operation.

A transaction cost must be paid for each contract, which changes the position.

The period from entering the non-flat position at market to leaving the position is called *trade*. The trade is very important, because the profit in cumulative gain is only hypothetical. But at the end of the trade, the cumulative gain corresponds with the real realized profit.

3.1 Task definition

Let denote the price in time t by y_t and position held in time t by u_t. The structure of u_t is following: the absolute value $|u_t|$ sets the number of contracts in an open position;

and the signum of u_t sets the kind of position, minus for short and plus for long position. The flat position is characterized by $u_t = 0$.

For this notation the gain function is defined as:

$$G = \sum_{t=1}^{T} g_t = \sum_{t=1}^{T} \underbrace{(y_t - y_{t-1})u_{t-1} - C|u_t - u_{t-1}|}_{g_t}, \tag{7}$$

where C is the normalized transaction cost. For offline experiments, the transaction cost is artificially increased by so-called *slippages*. Slippages are required due to delay between prompting the market command and its realization, during this short time period the price can change. Second reason for slippages is that the action on market changes the price itself and this is often not included in off-line experiments. Both reasons causes that the price in real trading could be different from the value stored in data sets. To avoid this difference, the transaction cost has two parts $C = c + s$ for our task, where c is transaction cost payed to exchange provider for each contract in position, and s are slippages, which artificially make the transaction cost higher.

The slippages are estimated by an economic specialist. We use values obtained from Colosseum a.s. due our cooperation. Although the slippages makes the task more difficult, the trading system profitable at off-line data with slippages has big chance to be profitable in real trading.

3.2 Requirements to applicability

The economist have designed a lot of additional criteria to rate, whether the approach is good or bad. This criteria are closely related to the trading task. Moreover, the economist will decide, whether the approach will be applied in practice, hence is important to take this coefficients and criteria into a consideration. This section overview the main coefficients and introduces the criteria required to application of the approaches.

3.2.1 Main coefficients

Net profit is the same variable as the final gain (7).

Gross profit is the net profit calculated only over the profitable trades. The profitable trade is trade which starts with lower value of cumulative gain than finishes.

Gross loss is analogy with gross profit, but for non-profitable trades. The Gross profit is positive number, gross loss is negative number and net profit is sum of them.

Total cost is total amount of transaction cost c payed for realization of decision as was introduced in Sec. 3.1. The total cost is calculated via: $(-1)\sum_{t=1}^{T} c|u_t - u_{t-1}|$.

Total slippages is total amount of slippages s, calculated in analogy with transaction cost $(-1)\sum_{t=1}^{T} s|u_t - u_{t-1}|$. The slippages can be used for analyzing the results, because in the trading task is typical that slippages make the result negative (see [2]).

Trades is count of trades done during the experiment.

Winning/Losing trades is count of trades with positive/negative profit.

Days long/short/flat is count of time instants, when a contract was held in long/short/flat position. (The word 'days' is related to fact that we work with a day-data.)

Maximal drawdown is the biggest negative difference in cumulative gain sequence. This variable characterizes the risk related to given approach. The drawdown of bad approach is relatively same value as the final gain.

Length of drawdown characterizes the length of the maximal drawdown, i.e. how many time instants was the drawdown realized. Again, the bad approach has drawdown with comparable length as the data sequence.

3.2.2 Combinations of coefficients

The previous coefficient are raw coefficient obtainable from result. Following coefficients can be computed from the raw coefficients and give us criteria for identifying the good approach.

Percent profit gives percentage of winning trades:

$$\text{Percent profit} = \frac{\text{Winning trades}}{\text{Winning trades} + \text{Losing trades}}.$$

Profit factor is ratio of earned and lost money:

$$\text{Profit factor} = -\frac{\text{Gross profit}}{\text{Loss}}.$$

Profit per trade is average profit obtained in trade

$$\text{Profit per trade} = \frac{\text{Net profit}}{\text{Trades}}.$$

3.2.3 Criteria on good approach

There is a difference between theoretical design of approaches and its applicability in practice. Whereas, the theoretical success is each small bettering of an approach, the practical application demands significantly good results. The criteria to application of the tested approach for futures trading were designed by economic specialist from Colosseum a.s. The criteria are presented in Table 1.

3.3 Algorithm of rating

The decision, which approach is best, should be done using following rules:

1. The non-efficient approaches are excluded, the final gain is taken as measure of approach quality. This step chooses a subset of the original approaches.

Coefficient	Relation	Value
Net profit	greater than	0
Maximal drawdown	less than	1/10 net profit
Length of drawdown	less than	250 days
Percent profit	greater than	0.4
Profit factor	greater than	1.5
Profit per trade	greater than	$100 USD

Table 1: Requirements on approach to applicability in practice.

Ticker	Commodity	Exchange
CC	Cocoa	CSCE
CL	Petroleum-Crude Oil Light	NMX
FV2	5-Year U.S. Treasury Note	CBT
JY	Japanese Yen	CME
W	Wheat	CBT

Table 2: Reference markets, their tickers and exchanges.

2. The non-efficient approaches are excluded, the coefficient c_I is taken as measure of approach quality. This step chooses a subset of the original approaches.

3. The approaches are sorted by their MFP - the highest value as first.

4. The approaches are tested consequently, whether suffice the requirements on applicable approach. The proving is done over all data sets, hence each approach must satisfy $6 \times N$ conditions. The first, sufficient is rated as the best approach, because is efficient and has highest MFP.

3.4 Tuning the parameters

We have available price history from five market (see Tab. 2) and approach presented in [4], where are 2 parameters the length of regressor $l \in \{1, 2, \ldots, 10\}$ and the forgetting factor $\lambda \in \{1, 0.999, 0.99, 0.9\}$. (The explanation of the parameters is not important.) Thus, we have 40 couples of parameters and our aim is to estimate, which couple is the best. Due to availability of five data sets, the count of experiments is 200.

Table 3 reviews the results obtained by presented method (see Sec. 3.3). The values in the table were constructed by ordering the MFP coefficients (see Sec. 2.3.1), where the highest value of MFP was denoted by 1, second highest by 2 etc. And the highlighted approaches were marked as efficient in both steps 1 and 2 of algorithm from Sec. 3.3.

For last step of the algorithm, there is no approach satisfying all requirements for applicability. The nearest is the approach with the parameters $l = 1$ and $\lambda = 1$, where are satisfied 20 conditions from 30.

For the further research, the parameters couple $l = 1$ and $\lambda = 1$ will be used, although the non-applicability. The reason for this choice is that the given approach is the most

successful and moreover the analysis with respect to c_I coefficient define in Sec. 2.2 reaches also the best results (see Tab. 4).

The testing of c_I coefficient showed that approaches with value $c_I > 1.5$ have increasing cumulative gain without big drawdowns. Hence, the coefficient c_I can be used for rating the best approach in further research.

4 Conclusion

The paper concerns with the criteria of comparing approaches testing on data sets. The algorithm of the best approach choosing is designed. The algorithm is applied on the results obtained in tuning approach for futures trading task, and it chooses the best approach.

The main advantage of the designed algorithm lies in possibility to compare the approaches tested on more data sets. The algorithm combines the simply method of weighted sum with efficients solutions and applicability of approach. This combination is also great advantage.

The disadvantage of given algorithm is that the algorithm can exclude all approaches due to applicability conditions. And opposite, the efficient solution often selects big subset.

The algorithm will be tested in further research, but it make the ground idea for further algorithms in rating the approaches.

References

[1] M. Ehrgott. *Multicriteria optimization*. Lecture Notes in Economics and Mathematical Systems. Springer-Verlag, (2000).

[2] M. Kárný, J. Šindelář, Š. Pírko, and J. Zeman. Adaptively optimized trading with futures. Technical report, (2010).

[3] J. Zeman. Futures trading: Design of a strategy. In 'Proceedings of the International Conference on Operations Research and Financial Engineering 2009'. WASET, (2009).

[4] J. Zeman. A new approach to estimating the bellman function. In 'Proceedings of the 10th International PhD Workshop on Systems and Control', P. L. Hofman Radek, Šmídl Václav, (ed.). ÚTIA, AV ČR, (2009).

l	$\lambda = 1$	$\lambda = 0.999$	$\lambda = 0.99$	$\lambda = 0.9$
1	1	21	26	28
2	13	17	23	30
3	8	15	25	27
4	6	19	18	33
5	2	11	20	35
6	3	12	22	36
7	5	14	31	37
8	4	16	29	38
9	10	24	34	39
10	9	7	32	40

Table 3: Comparison of 40 approaches for Bellman function estimation, each approach is defined by couple l and λ, the efficient solutions are highlighted and the numbers in table are order of approaches by MFP.

l	$\lambda = 1$	$\lambda = 0.999$	$\lambda = 0.99$	$\lambda = 0.9$
1	1.0551	-0.3355	-1.2594	-1.7522
2	0.2444	-0.1365	-0.3234	-1.6807
3	0.6385	0.3818	-0.5466	-1.7164
4	0.4861	-0.0215	0.1084	-1.8719
5	0.6014	0.2383	-0.2796	-2.3504
6	0.6046	0.0992	-0.4663	-2.8133
7	0.5481	0.1318	-1.6669	-3.4924
8	0.5002	-0.0869	-1.2192	-3.7682
9	0.3632	-0.7274	-1.9563	-4.3346
10	0.2865	0.4667	-1.7901	-5.0938

Table 4: The mean value increase coefficient c_I calculated over available data sets.

Building Efficient Data Planner for Peta-scale Science

Michal Zerola

3rd year of PGS, email: michal.zerola@ujf.cas.cz
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Michal Šumbera, Nuclear Physics Institute, ASCR
Jérôme Lauret, Brookhaven National Laboratory, USA
Roman Barták, Faculty of Mathematics and Physics, Charles University

Abstract. Distributed computing, heavily relying on the presence of data at the proper place and time, have further raised demands for coordination of data movement on the road towards achieving high performance. Although there exist several sophisticated and efficient point-to-point data transfer tools, the lack of global planners and decision makers, answering questions such as "How and from which sources to bring the required dataset to the user?", is for most part lacking. We present our work and status of the development of an automated data planning, ensuring fairness and efficiency of data movement by focusing on the minimal time to realize data movement (delegating the data transfer itself to existing transfer tools). Its principal keystones are self-adaptation to the network/service alteration, optimal selection of transfer channels, bottlenecks avoidance and user fair-share preservation. The planning mechanism relies on Constraint Programming and Mixed Integer Programming techniques, allowing to reflect the restrictions from reality by mathematical constraints. In this paper, we will concentrate on clarifying the overall system from a software engineering point of view and present the general architecture and interconnection between centralized and distributed components of the system. The implications and benefit of our approach as well as a use case in practice made with multiple choice for sources will be presented.

Keywords: planning, data transfers, distributing computing

Abstrakt. Distribuované počítanie, ktoré závisí na dostupnosti dát v správnu dobu na správnom mieste, ešte viac zvýšilo nároky na koordináciu dátových prenosov na ceste za vysokou výkonnosťou. Hoci niekoľko sofistikovaných a výkonných 'point-to-point' prenosových nástrojov existuje, stále chýba globálny plánovač, ktorý by riešil úlohy typu "Ako a z ktorých zdrojov doručiť požadované dáta k užívateľovi?". Predstavíme prácu a stav na vývoji automatizovaného plánovacieho nástroju, zaisťujúceho efektívnosť a koordináciu dátových prenosov. Jeho hlavnými atribútami sú auto-adaptácia k zmenám služieb/sieti, optimálna selekcia prenosových kanálov, zamedzenie vzniku úzkych hrdiel a zachovanie spravodlivosti medzi užívateľmi. Plánovací mechanizmus používa techniky programovania s obmedzujúcimi podmienkami a celočíselné programovanie, čím zachytáva obmedzenia z reálneho sveta do matematických podmienok. V tomto článku sa budeme zameriavať na preverenie celkového systému z pohľadu softwarového inžinierstva a predstavíme celkovú architektúru a prepojenie jednotlivých centralizovaných a distribuovaných komponent. Ukážeme tiež dopady a výhody tohoto prístupu ako aj praktickú štúdiu založenú na požiadavkách k dátam z viacerých zdrojov.

Kľúčové slová: plánovanie, dátové prenosy, distribuované počítanie

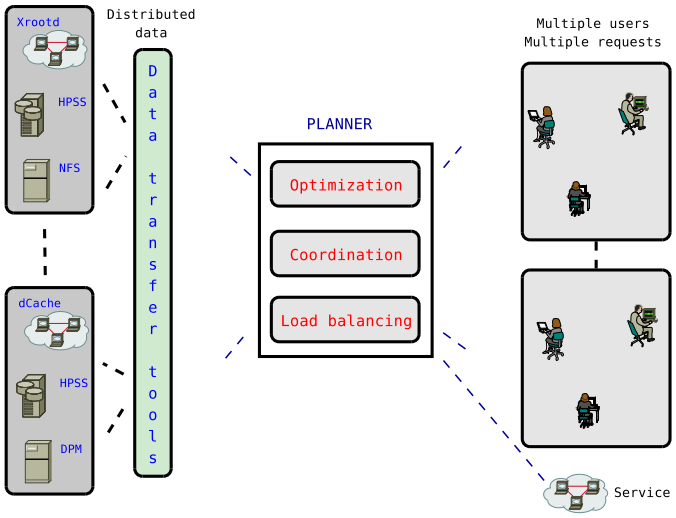


Figure 1: General view of the automated planning system. The goal is to achieve controlled and efficient utilization of the network and data services with a proper use of existing point-to-point transfer tools. At the highest level of abstraction, the planner should appear as a "box" between the user's requests and the resources.

1 Introduction

As it is widely known, distributed computing offers large harvesting potential for computing power and brings other benefits as far as it is properly exploited. On the other hand it introduces several pitfalls including concurrent access, synchronization, communications scalability as well as specific challenges such as answering key questions like "how to parallelize a task?" knowing where my data and CPU power are located. In data intensive experiments, like the one from HENP community and the STAR [1] [1] experiment, the problem is even more significant since the task usually involves processing and/or manipulation of large datasets.

This massive data processing will be hardly "fair" to users and hardly using network bandwidth efficiently unless we address and deal with planning and reasoning related to data movement and placement. In this paper we present and focus on the implementation and software engineering part of our ongoing work, while we refer to our previously published papers explaining in more depth the underlying model and theoretical background.

The purpose of our research and work is to design and develop an automated planning system acting in a multi-user and multi-service environment as shown in Fig. 1. The system acts as a "centralized" decision making component with the emphasis on **optimization**, **coordination** and **load-balancing**. The optimization guarantees the resources are not wasted and could be shared and re-used across users and sources. Coordination ensures multiple resources do not act independently so starvation or clogging do not occur, while load-balancing avoids creating bottle-necks on the resources. The intent is not to create another point-to-point data transfer point-to-point tool, but to use

[1] Solenoidal Tracker at Relativistic Heavy Ion Collider is an experiment located at the Brookhaven National Laboratory (USA). See http://www.star.bnl.gov for more information.

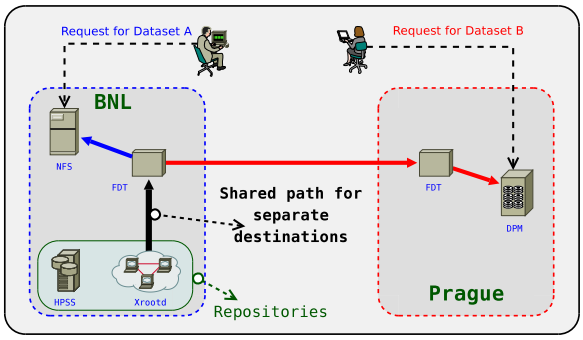


Figure 2: Optimization of the transfer paths with regards to the network structure and link bandwidth. Some network path may be re-used to satisfy multiple requests for the same data.

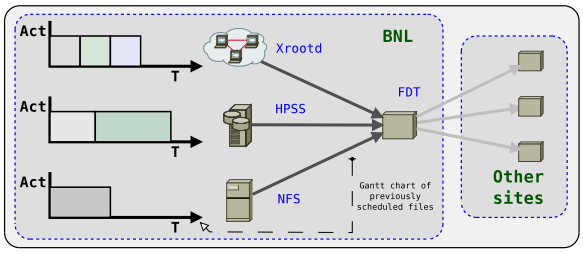


Figure 3: Optimization of the transfer paths with regards to the different data service performance/latency. Multiple sources for the same data may be naturally combined alternatively to avoid overload and service clogging.

available and practical ones in the efficient manner.

We describe the most important optimization characteristic with the help of figures Fig. 2 and 3. Let us suppose there are requests for the same (or overlapping) dataset from two users, while each of them needs the dataset to be processed at his/her specific location. The system has to reason about the possible repositories for the dataset, select the proper ones for every file (the granularity is specified by the files in our case) and produce the transfer paths for each file. The output plan should be optimal with an objective to the overall completion time of all transfers. Thus, this optimization characteristic is focusing on the network structure and respective link bandwidth. As illustrated in Figure 2, it is conceivable in our example that optimization will cause data movement to occur once on some network links while datasets will be moved to two different destinations. Moreover, the files are usually served by several data services (such as Xrootd [6], Posix file systems, Tape systems [8], ...) with different performance and latencies. Therefore, the optimization and reasoning on where to take the files available from multiple sources choice will allow making the proper selection for a file repository, respecting their intrinsic characteristic (communication and transfer speed) and scalability (Fig. 3). In other words, as soon as multiple services and sources are available, load balancing would immediately be taken into account by our planner.

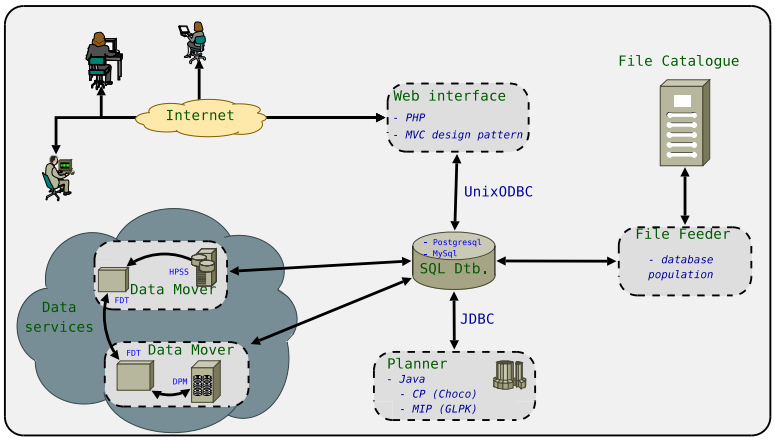


Figure 4: Architecture of the system.

2 Architecture

In this section, we will describe the architecture of the system, explaining briefly each component following the work-flow (see Fig. 4 for illustration). End users (or stand-alone services) generate requests using the web interface, written in *PHP* following the *MVC* design pattern. A request is an encapsulation of the meta-data query (as understood by STAR's File and Replica Catalogue) and the destination. The request is stored in a *SQL* database (system supports *MySQL* and *PostgreSQL*) in a Catalog agnostic manner (any Catalog should work as far as they have a LFN/PFN concept our approach relies on) with the additional information like user name, group or date of the request. Later, the component called *File Feeder* contacts the *File and Replica Catalogue* and makes the query for the requested meta-data. The output information is stored back to the database, including all possible locations for every file in a request.

The brain of the system, a component called the *Planner*, takes a subset of all requests for files to be transferred according to the preferred fair-share function. It creates the plan (transfer paths) for the selected requests and stores the plan back to the database. The individual file transfers are handled by the separate distributed component called *Data Mover*. The role of these workers is to perform a point-to-point data transfer on a particular link following the computed plan. The results and intermediate status is continuously recorded in the database and user can check the progress at any time.

We can see that the whole mechanism is a combination of **deliberative** (assuring optimality) and **reactive** planning (assuring adaptability to the changing environment). Since this is crucial to the argument, in the next section we will describe the respective two components (*Planner* and *Data Mover*) serving up as a "reasoner" and a "worker".

2.1 Planner

The *Planner* (Fig. 5-left), the brain of the system, is built on the constraint-based mathematical model. The theoretical background and our continuous progress were published in several papers ([11], [10], [12]). Therefore, we will not go into details in this pa-

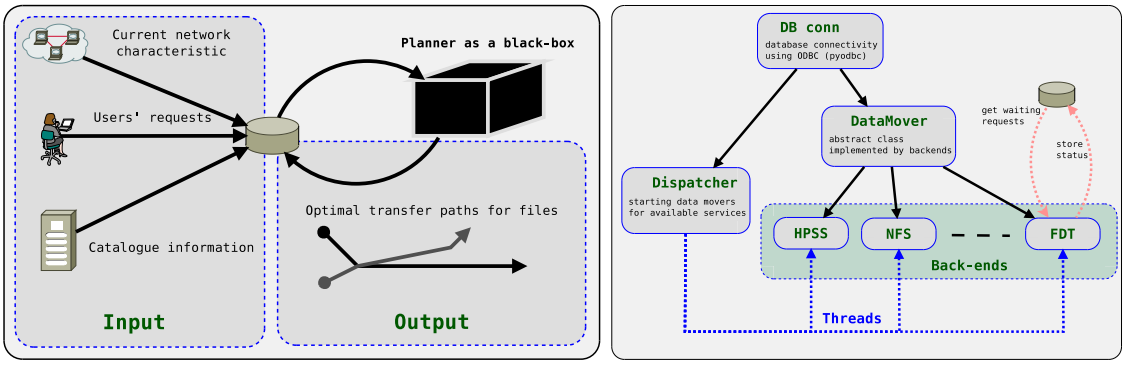


Figure 5: **Left:** Planner as a black box. **Right:** Data Mover component.

per, but only sketch out the main principles. Constraint based approach ([7]) brings a fundamental advantage in a straight forward mapping of the reality restrictions into the mathematical model. The solver uses methods from Constraint Programming and Mixed Integer Programming and the logic tries to minimize the makespan considering all possible combinations. The tree of possibilities may very well contain solutions where transferring data once on a given link lead to a minima or balancing between services lead to the fastest transfers. In all cases, the optimal solution will only be determined by the input parameters. The input consists of three parts: current characteristic of the link or network, requests to be planned (size, logical files) and information from a *File (replica) Catalogue* about possible repositories. Having all these information the solver starts a computation and stores the results directly into the database. The result is a computed transfer path (repository and oriented path to the destination) for each request. Note that multiple requests for the same files would be treated and accounted for in the plan. Our planning is also incremental - we have previously demonstrated ([9]) that a full plan or incremental planning would not make a large difference on the make span overall - the gain of an incremental approach is the ability to self-adapt based on the *Mover's* feedback.

For implementation of the solver we use **Choco** ([2]), a Java based library for constraint programming and **GLPK** ([4], [5]), a library for Mixed Integer Programming. The Java based platform allows us an easier integration with already existing tools in the STAR environment.

2.2 Data Mover

The *Data Mover* is the distributed component responsible for performing data transfers in a reactive way. Each instance is controlling data services within a given computing site and also the wide-area network connections from/to the site. It relies on the underlying data transfer tools and uses them for data movement. In our implementation, we did not address interoperability of data transfer tools (which is not the object of this work) but settled in using by the **F**ast **D**ata **T**ransfer tool (FDT [3]). The way data movers operate is reactive that is, as soon as a file appears at the source node (either at a data service or in a cache space before WAN transfer) it is marked as "ready for transfer"

and moved by the proper underlying tool. As soon as the transfer is finished another instance realizes the file is available and initiates the next move (along the computed path from the solver). Our approach is also adaptive: from the initial transfer and consequent monitoring, the real speed can be inferred and re-injected as a parameter for the next incremental plan, helping the system to converge toward realistic transfer rates rather than relying on theoretical optimum alone.

The *Data Mover* is written in *Python* language and concurrent link/service control is achieved by separate threads (Fig. 5-right).

3 Show case

To prove the validity of our planning strategy, a use case was designed and implemented. The purpose of the test was to affirm the software components work and communicate in the expected way and the quality of the computed plan is confident. The environment was for simplicity formed by two computing sites, the central **BNL** and remote **Prague**. The available data services at *BNL* were: *Xrootd*, *NFS* and *HPSS*, while in *Prague* only *NFS* was available. The wide area network (WAN) transfer was controlled by FDT. The configuration is shown in Fig. 6-left. The test hence challenges the planner in making proper decisions when multiple sources are available at the same site.

The request consisted of files available at all data services at *BNL* at the same time and the task was to bring them to the *Prague* NFS service. The test was composed of four different configurations. The planner consecutively considered:

- only Xrootd repository
- only NFS repository
- only HPSS repositories
- a combination of Xrootd, NFS and HPSS repository concurrently

The results of each configuration are shown in Fig. 6-right. As expected, while all files are located on mass storage in STAR, transfers from *HPSS* (in green) are the longest to accomplish and hence, lead to the longest delays in delivery. In our setup, the green and blue curves are near equivalent (*NFS* direct transfers are slightly faster) but it is to be noted that not all files are held on *NFS* (central storage) in STAR and pulling all files from *Xrootd* may cause significant load on a system in use primarily for batch based user analysis (hence, an additional load is not desirable). When we combined all storage sources, the makespan was equivalent to the one from *Xrootd* while the relative ratio of files transfers from the diverse sources was 19%, 38% and 43% for *HPSS*, *NFS* and *Xrootd* respectively with no load caused on any of the services. At the end, the overall bottleneck was only the WAN transfer speed - we infer our test proved the planner works as expected, since the full reasoning considering all possible repositories led to the optimum makespan. Additionally, the utilization of all services brings the advantage in the form of load-balancing and automatic use of replicas.

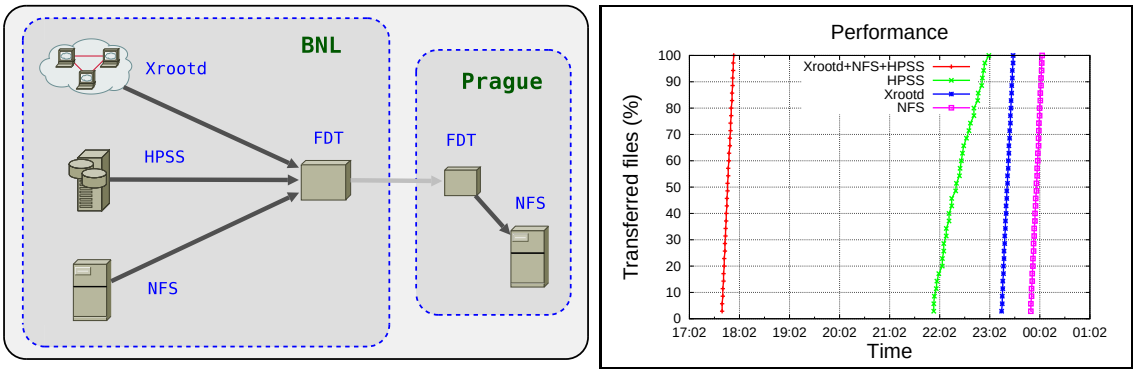


Figure 6: **Left:** The network and service configuration for the tests. **Right:** The performance of the system using 4 different configurations. On the X axis, we represent the time of transfers while Y is the percentage completion. The x-range of each curve is hence representative of the makespan.

4 Conclusions

When multiple sources for files or datasets are available along with many CPU resources in a distributed computing environment, planning is needed to ensure load balancing, efficient and fair data movement and best use of the resources. Random access to files and datasets by users could easily destroy efficiency or render sites inoperative and with this in mind, we have tackled the challenge of coordination of data transfers.

In this work, we specifically presented the architecture components and implementation of a framework, in test mode in the STAR experiment, which goal is to address the planning challenges of transfers over widely distributed resources. Based on constraint and mixed integer programming techniques, the tool was designed to incorporate elements to achieve optimization, coordination and load-balancing. Its simple yet robust architecture allows users to express their requests for files via a Web interface while a back-end planner and a set of data movers take care of the movement on the user's behalf. Within our test example of moving files to a single destination considering a dataset available from multiple-sources, we have showed that our approach lead to an optimal plan that is, producing the shortest possible makespan while causing no load on any of the storage systems by automatically load-balancing. With our model (showed to work in simulated mode [10]) and this proof of principles, we are equipped with a corner stone functional architecture and we will pursue as next steps multi-users and multi-sites transfers.

Acknowledgement

The work has been supported by the grants LC07048 and LA09013 of the Ministry of Education of the Czech Republic, the project Czech Science Foundation P202/10/1188 and by the Office of NP within the U.S. DOE Office of Science.

References

[1] S. C. J. Adams. *Experimental and theoretical challenges in the search for the quark gluon plasma: The STAR collaboration's critical assessment of the evidence from RHIC collisions.* Nuclear Physics A **757** (2005), 102.

[2] Choco. http://www.emn.fr/.

[3] FDT. http://monalisa.cern.ch/FDT.

[4] GLPK. http://www.gnu.org/software/glpk/.

[5] GLPK-java. http://glpk-java.sourceforge.net/.

[6] A. Hanushevsky, A. Dorigo, and F. Furano. The Next Generation Root File Server. In 'Proceedings of the Computing in High Energy and Nuclear Physics (CHEP) conference', 680–683, (2005).

[7] H. Simonis. *Constraint applications in networks.* In 'Handbook of Constraint Programming', F. Rossi, P. van Beek, and T. Walsh, (eds.), Elsevier (2006), chapter 25, 875–903.

[8] D. Teaff, D. Watson, and B. Coyne. The Architecture of the High Performance Storage System (HPSS). In 'Proceedings of the Goddard Conference on Mass Storage and Technologies', 28–30, (1995).

[9] M. Zerola, R. Barták, J. Lauret, and M. Šumbera. Using constraint programming to resolve the multi-source / multi-site data movement paradigm on the grid. In 'Advanced Computing and Analysis Techniques in Physics Research {PoS(ACAT08)039}', (2008).

[10] M. Zerola, R. Barták, J. Lauret, and M. Šumbera. Efficient Multi-site Data Movement in Distributed Environment. In 'Proceedings of the 10^{th} IEEE/ACM International Conference on Grid Computing (GRID)', 171–172. IEEE, (2009).

[11] M. Zerola, R. Barták, J. Lauret, and M. Šumbera. Planning Heuristics for Efficient Data Movement on the Grid. In 'Proceedings of the 4^{th} Multidisciplinary International Conference on Scheduling: Theory and Applications (MISTA)', 768–771, (2009).

[12] M. Zerola, R. Barták, J. Lauret, and M. Šumbera. Using Constraint Programming to Plan Efficient Data Movement on the Grid. In 'Proceedings of the 21^{st} International Conference on Tools with Artificial Intelligence (ICTAI)', 729–733. IEEE, (2009).