DOKTORANDSKÉ DNY 2018

sborník workshopu doktorandů FJFI
oboru Matematické inženýrství

16. a 23. listopadu 2018

P. Ambrož, Z. Masáková (editoři)

Doktorandské dny 2018
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

Počet stran 194, Vydání 1.

Seznam příspěvků

Předmluva

Workshop Doktorandské dny probíhá v roce 2018 už po třinácté, tentokrát v datech 16. a 23. listopadu. Jedná se o setkání doktorandů oboru Matematické inženýrství na Fakultě jaderné a fyzikálně inženýrské ČVUT v Praze. Tento obor společně zajišťují katedry matematiky, fyziky a softwarového inženýrství FJFI ve spolupráci s několika pracovišti Akademie věd ČR, zejména Ústavem teorie informace a automatizace, Ústavem Informatiky a Ústavem jaderné fyziky.

Workshop doktorandské dny slouží k předvedení práce našich studentů za poslední rok. Tématika přednášek pokrývá velkou šíři oblastí moderní aplikované matematiky, informatiky a matematické fyziky. Příspěvky v tomto sborníku jsou buď plným textem studentských přednášek, nebo abstrakty s odkazy na články publikované ve sbornících významných konferencí, či v odborných časopisech.

Workshop i letos finančně podporuje grant SVK 25/18/F4.

Editoři

Compensation of Inaccurate Meteorological Data in Source Term Determination Problem Using Bayesian Methods

Alkomiet Belal

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

advisor: Václav Šmídl, Department of Adaptive Systems
Institute of Information Theory and Automation, AS CR

Abstract. A common formulation of the source term determination problem is based on linear inverse problem y = Mx+e, where y is the the vector of observations, M is the source-receptor-sensitivity (SRS) matrix, x is the unknown source term, and e is the model residue. The matrix M is computed using a selected atmospheric transport model fed by a selected meteorological data. The inverse problem is typically ill-conditioned which requires the use of regularization terms. However, the matrix M is typically assumed to be known exactly. The mismatch between the true underlying SRS matrix and that from the numerical model is hard to assess. It is assumed that the residue is a combination of the observation error and the SRS error. Distinction between these errors is typically made explicitly as in methods based on outlier detection, or implicitly in methods based on estimation of the residue covariance matrix. In this contribution, we propose two probabilistic models of the residue statistics. The first model is based on the assumption of non-Gaussian statistics of the residue. In particular, we choose Student-t model of the residues and design a Variational Bayesian estimation procedure for the source term. The second model is based on Gaussian model of the residues with non-diagonal covariance matrix. Since estimation of the full covariance matrix is not possible due to limited data, we propose several variants of the covariance parametrization using restricted parametrization. We estimate the parameters of the residue model jointly with the source term. The proposed models will be tested on two data sets: the ETEX experiment for which a ground truth source term is known, and the data from recent detection of ruthenium over Europe and Siberia in late September and early October 2017. We will show that the proposed models improve estimation of the source term for the ETEX dataset. Since the origin of the Ruthenium release is still unclear, we will show the effect of the proposed models on estimated maps of the source location. The source location maps are also computed using Bayesian model selection approach.

Keywords: Bayesian inference, atmospheric transport model, inverse modeling

Full paper: O. Tichý, V. Šmídl, M. Hýža, K. Šindelářová, L. Ulrych, A. Belal. *Compensation of inaccurate meteorological data in source term determination problem using Bayesian methods.* Geophysical Research Abstracts **20**, EGU2018-15902, 2018.

References

[1] V. Šmídl and A. Quinn. *A The Variational Bayes Method in Signal Processing.* In SpringerVerlag (Berlin/Heidelberg, 2006).

[2] O. Tichý , V. Šmídl, R. Hofman, and A. Stohl. *LS-APC v1.0: a tuning-free method for the linear inverse problem and its application to source-term determination.* Springer (2016).

[3] C. Bishop. *Pattern recognition and machine learning.* Springer(2006).

Kernel and Divergence Techniques in High Energy Physics Separations*

Petr Bouř

3rd year of PGS, email: petr.bour@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. Binary decision trees under the Bayesian decision technique are used for supervised classification of high-dimensional data. We present a great potential of adaptive kernel density estimation as the nested separation method of the supervised binary divergence decision tree. Also, we provide a proof of alternative computing approach for kernel estimates utilizing Fourier transform. Further, we apply our method to Monte Carlo data set from the particle accelerator Tevatron at DØ experiment in Fermilab and provide final top-antitop signal separation results. We have achieved up to 82 % AUC while using the restricted feature selection entering the signal separation procedure.

Keywords: binary classification, decision trees, Fourier transform, kernel density estimation, machine learning, phi-divergence, top quark

Abstrakt. Binární rozhodovací stromy jsou vhodným nástrojem pro použití při statistické klasifikaci vícedimenzionálních dat. Pro uzly binárního rozhodovacího stromu navrhujeme implementaci adaptivního jádrového odhadu hustoty pravděpodobnosti, který diskriminuje třídy pozorování v klasifikační úloze. Pro detailnější vhled do problematiky jádrových odhadů navíc prokážeme alternativní možnost jejich výpočtu skrze Fourierovu transformaci. Naši komplexní metodu natrénujeme na Monte Carlo datasetu pro klasifikační úlohu separace rozpadu top-antitop kvarků na částicovém urychlovači Tevatron při experimentu DØ ve Fermilabu. Při finální separaci extrémně znečištěného signálu dosahujeme úspěšnosti klasifikace až 82 % AUC pro omezenou volbu fyzikálních příznaků vstupujících jako prediktory do klasifikační úlohy.

Klíčová slova: binární klasifikace, rozhodovací stromy, Fourierova transformace, jádrové odhady hustoty pravděpodobnosti, strojové učení, f-divergence, top kvark

Full paper: P. Bouř et al. *Kernel and divergence techniques in high energy physics separations*. J. Phys.: Conf. Ser. **898 (7)** (2017), 072004.

*This work has been supported by the grants LM2015068 (MYES), LG15047 (MYES), GA16-09848S (GACR) and SGS15/214/OHK4/3T/14 (CTU).

Use of GPU for Molecular Simulations of Nucleation and Metastable State

David Celný

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

advisors:

Jiří Kolafa, Department of Physical Chemistry, UCT in Prague

Roland Span, Department of Thermodynamics, Ruhr-Universität Bochum

Tomáš Oberhuber, Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. Theoretical investigation of nucleation processes has a long tradition in thermodynamics. The research possibilities in this field were further enhanced by the simulation tools such as Molecular Dynamics (MD) enabling more direct observation of nucleation processes compared to observation in nature. The simulation ambitions are faced with a computability restriction on size of the problem and therefore experimental results, available simulation and theoretical models align only qualitatively with several orders of magnitude quantitative discrepancies. The aim is to extend the possibilities of water nucleation simulation by employing parallelization techniques of general-purpose programming on graphics processing units (GPGPU) to improve execution speed of calculation. Realization of this aim would enable larger system calculation with the advantage of improving simulation predictions of nucleation.

Keywords: homogeneous nucleation, GPGPU, Molecular dynamics,

Abstrakt. Výzkum nukleace má na poli termodynamiky dlouhou tradici. Kromě experimentu a teorietického výzkumu jsou k dispozici také nástroje molekulární dynamiky(MD), které umožňují mikroskopický pohled na nukleační procesy. V současné době jsou však srovnání výsledků z experimentů, teorie a simulací zatížené řádovou kvantitativním chybou. Tato chyba v případě simulacích je způsobena nejspíše příliš nízkou velikostí referenčního simulovaného systému. V porovnáním s eperimentálně zkoumanými vzorky jsou simulované systémy řádově menší. Cílem této práce je urychlení výpočtu simulace nukleace pomocí použití grafických výpočetních jednotek (GPU) a technik použití GPU pro negrafické výpočty (GPGPU). Urychlení výpočtu umožní výpočet nuklace u větších systémů a tímto způsobem zlepšit předpověď nukleační rychlosti.

Klíčová slova: homogenní nukleace, GPGPU, Molekulární dynamika,

1 Introduction

The process of nucleation is a prevalent cornerstone of physics of condensed matter from eighteen century when the phenomenon of nucleation was first indirectly observed in supercooled water[4]. Nucleation was afterwards experimentally and theoretically indirectly investigated [16, 6] followed in year 1926 by the theoretical work on nucleation

published by Volmer and Weber [24] about classical nucleation theory (CNT). Lately, a number of of books was published [9, 22, 8, 10] due to the wide occurrence of the nucleation phenomenon. The wide applicability is a consequence of nucleation processes close connection with thermophysical properties, atmospheric processes, uses in industry or the interdisciplinary similarities with other processes occurring in nature. Industrial applications include equations of state, connection with CCS technologies[28] for lowering the carbon dioxide in atmosphere or methods of preparation platinum nanoparticles for Polymer electrolyte membrane (PEM) fuel cells[25, 3].

2 Theoretical background

The theoretical background of the presented work consist of three main areas namely nucleation, molecular simulation and general purpose computation on graphical processing units(GPGPU). The brief overview of nucleation and classical nucleation theory(CNT) is provided first. The modelling area of research is presented in the second section. The last section consist of programming segment dealing with the parallel programming paradigm on Nvidia graphics processing units(GPU) with the CUDA C/C++ language extension.

2.1 Nucleation

One of the accepted views of nucleation is as the process of first order phase transition. It is usually, but not restricted to, the vapour <–> liquid transition. An example could be the formation of water droplets in clouds or formation of carbon dioxide bubbles in carbonated drinks. The initial description of the nucleation and CNT, according to the Kalikmanov [8], was provided in first half of twentieth century by Volmer & Weber, Becker & Döring and Zeldovich [24, 2, 27].

It is important to note that basic CNT models describe homogeneous nucleation. The term homogeneous imply that no external forces or system impurities(external agents) are present. The homogeneous nucleation is almost non-existent in nature because of the requirement of chemically pure system. The examples introduced in the initial description are prime examples of the fact: impurities in clouds or drinks. The external agents lower the energy barrier of the translation, because they are the initiating nucleation centres. This situation is called heterogeneous nucleation and it is beyond the scope of this study. For brevity of the following description the transition from vapour to liquid is considered, therefore the nucleation process described here concerns the formation of droplets.

2.1.1 Classical nucleation theory

Before the CNT can be described it has to be mentioned that the process of nucleation is restricted to a special thermodynamic state only. This state is called the metastable state and can be defined as a region between spinodal and binodal curve restricting the system thermodynamic variables. For detail about these curves see Fig 5.a) in Kelton [10]. The more general definition is that the system in such a state favours the local first order phase transition. The system then transfers into the state with a lower free energy by crossing an energy barrier illustrated in Figure 1. The general idea of nucleation is that

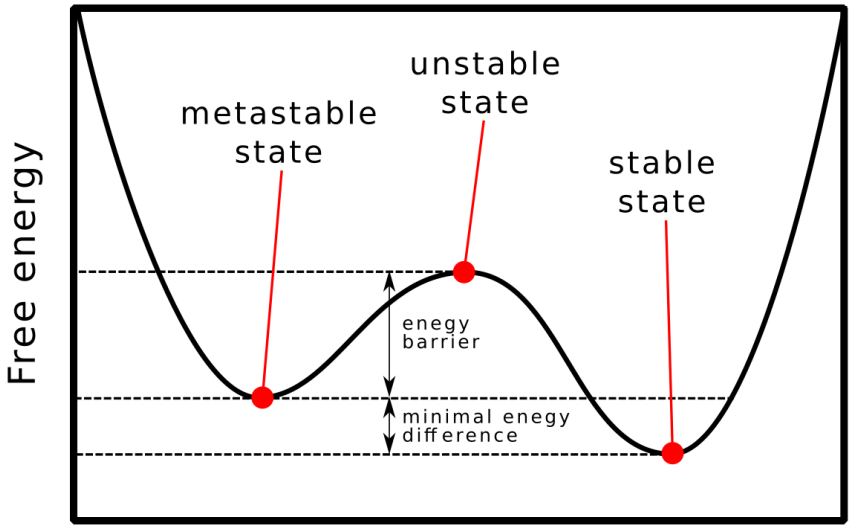


Figure 1: Illustration of free energy function depicting the free energy corresponding to the system stability. This picture was based upon Fig. 3.1 in [8].

while energy requirement for the transition as a whole system at once are too high the localized transition driven by system's fluctuation is a feasible. This means that system particles random collisions create dimmers, trimers which can grow into clusters until critical cluster is reached.

The general form of CNT presents a way how to calculate the nucleation rate of a metastable system. The nucleation rate is the amount of clusters formed in a unit of space during a unit of time. The cluster where probabilities of cluster growth and cluster shrinkage are equal is called critical cluster. The critical cluster is directly connected to the nucleation rate as an input parameter of nucleation rate showing how fast nucleation is, based on the characteristics of critical cluster. It is also an important property for physical and even technical application such as turbine development and construction where the rate and size of formed droplet determines turbine blade corrosion [12] or in the PEM fuel cells catalytic layer preparation [3]

The problem of nucleation in its native form is quite complicated, therefore CNT adopts the capillarity assumption which simplifies the description of very small clusters and considers these cluster to behave as macroscopic. The assumption neglects the effects of curvature on surface tension and assumes a bulk density within the centre of each cluster. Additionally a steady state of nucleation is assumed meaning that the flow of particles is independent on the size and previous history and therefore can be described as an exchange of a single particle. This can be understood as the nucleation is Markov process schematically illustrated by Figure 2 with omitted dependence on number of particles n.

Under these assumptions the CNT formula can be derived according to [8, 22] for a

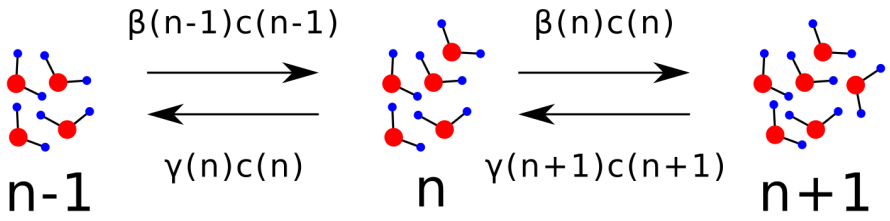


Figure 2: Steady state nucleation kinetics illustration for water molecules as system particles. The growth probabilities are dependent on concentrations c and condensation coefficients β, γ. For steady state the dependency on n should be omitted.

steady state metastable system that states the nucleation rate J as:

$$J = J_0 \exp\left(\frac{-\Delta G^*}{k_\mathrm{B} T}\right) \tag{1}$$

Where J is an exponential function of the energy barrier ΔG^* with flux pre-factor $J_0 = J_0(n_c)$ dependent on the number of particles in critical cluster particle count n_c. While the flux J is considered at steady state it is independent on the size but the observation corresponds to the value of critical cluster $n = n_c$. The factor J_0 gets form of

$$J_0 = \mathcal{Z} \nu A(n_c) \rho^V \tag{2}$$

The term $\nu A(n_c)$ evaluates the rate at which particles attach themselves to the cluster relative to the number density ν. ρ^V is the density of the supersaturated vapour. Additionally, the Zeldovich [27] factor $\mathcal{Z}$ represents the probability of critical cluster crossover of energy barrier. This factor is expressed in terms of critical cluster count as:

$$\mathcal{Z} = \frac{1}{3}\sqrt{\frac{\theta}{\pi}}\, n_c^{-2/3} \tag{3}$$

Where θ is the dimensionless surface tension evaluated according to the Capillarity assumption from planar surface tension $\sigma = \sigma_\infty$ as:

$$\theta = \frac{\sigma \sqrt[3]{36\pi}}{\nu_L^{2/3} k_\mathrm{B} T} \tag{4}$$

Additionally to the flux pre-factor the energy barrier ΔG^* can be expressed as:

$$\Delta G^* = \frac{16\pi}{3} \frac{\nu_\mathrm{L}^2 \sigma^3}{\Delta \mu^2} \tag{5}$$

For the formula of critical cluster the Kelvin equation [21]

$$p^V = p^{\mathrm{sat}} \exp\left(\frac{2\sigma_\infty \nu^L}{k_\mathrm{B} T r_t}\right), \tag{6}$$

can be employed to express the vapour pressure p^V. Where r_t is the droplet radius at surface of tension radius and p_{sat} is the saturation pressure. The size of the critical cluster is therefore written as:

$$n_c = \left(\frac{2\theta}{3 \ln(p^V / p^{\text{sat}})} \right)^3 \tag{7}$$

where the argument of the logarithm is usually denoted as saturation which can be used as a descriptor of how supersaturated system is i.e. how far into the metastable region from the saturation line the system is placed.

2.2 Molecular simulations

An alternative approach to experiment are molecular simulation tools performing a pseudo-experiment in computer memory [18]. This offers a benefit of a reproducible system evolution which can be performed in experimentally unfeasible setting or measure experimentally unreachable properties. This advantage is also a partial drawback as the experiment reflects the nature while a simulation can be incorrectly setup and provide unphysical results. It is therefore very important to verify the simulations with experiments and already verified models.

The property obtained from simulation tool is in the form of a mean value. When different thermodynamic ensembles are considered one can obtain the mean energy of the system, pressure and also nucleation speed. Simulation tools can be distinguished into two main categories depending on the type of mean value produced. The time mean values produced during system evolution are ascribed to molecular dynamics. On the other hand the ergodic theorem can be employed to produce samples mean. This mean is evaluated on different system samples generated by Monte Carlo simulation. This study is focused on the molecular dynamics presented below.

2.2.1 Molecular dynamics

Molecular dynamics considers a Newtonian description of a mechanical system with the prescribed potential field. In this setup, the modelled system gets its mechanical behaviour from the Newtonian physics although methods taking into account quantum behaviour are also available. More detailed explanation can be found in [5, 18, 1].

The general idea of molecular simulation is that for a provided initial configuration the model produce time series of snapshots representing a state of system in simulated time. These time snapshots can be used for post-processing, or quantities of interest can be computed during the simulation. For example the immediate energy of the system can be computed online. Analysis of these output data can yield a desired mean property value. The analysis of time snapshots produces also structural information about the system such as the radial distribution function or cluster distribution.

Expanding further on the nature of the Newtonian problem are the interactions. The simple form of potential of non-bonded (denoted as U_{nb}) system (system containing only one type of unbounded atoms such as Ar) is

$$U_{nb}(r^N) = \sum_{i=1}^{N} u^E(r_i) + \sum_{i=1}^{N} \sum_{j>i}^{N} u^P(r_i, r_j) + \sum_{i=1}^{N} \sum_{j>i}^{N} \sum_{k>j}^{N} u^T(r_i, r_j, r_k) + \dots \tag{8}$$

where the first sum represents external effecting potential u^N, the second sum stands for two-body interaction and following sums similarly represents the k-body interactions. In many cases the external potential is zero. Additionally, considering the ratio of two-body, three-body and further interactions, the three of more interaction terms are to be neglected without causing significant decrease of model accuracy. The two-body potential can be reduced from two position vector to a function of interparticle separation, $u(r_{ij})$. The corresponding force per particle is obtained as negative differentiation of the potential

$$U(r^N) = \sum_{i,j} u(r_{ij}) \tag{9a}$$

$$f_i = -\frac{\partial}{\partial r_i} U(r^N) = -\sum_{i \neq j} u'(r_{ij}) \frac{r_{ij}}{|r_{ij}|}. \tag{9b}$$

The force formula is evaluated directly instead of run-time differentiation. The force per particle is then used to calculate atom velocities and new atom positions for a time advanced by a constant time-step. In the case of bare calculation the computational time is then incremented leading to the next iteration.

This setup would suffice for the calculation of the mentioned argon system. For the molecules, the molecule bonds has to be considered with corresponding changes to the potential computation. This is done in algorithm that takes into account that bonds represents a constraints leading to the solution of a set of equations with implied constrains on molecular level, such as the Lagrange multipliers methods. With the prime focus of this work being water we can select simpler but efficient solver method in the form of SHAKE [19] algorithm. SHAKE algorithm solves the equation set by the Gauss-Seidel iteration leading to the linear order of convergence.

For the water simulation the particle charges have to be taken into account. Electrostatic part of the potential field has a theoretical difference compared to the so far considered potential, which has short range effect. The previously considered potential can be safely truncated after so called cutoff distance while the electrostatic potentials have much larger range. The range effect arise from the analysis of how Culomb's law decay with distance. This consideration leads to the special summation methods, as e.g. Ewald summation. In this stage of work we have considered larger cutoff distance for the potential to bypass the error caused by truncation. The used potential is smoothed and shifted to remove jumps in potential spline.

For the complete overview the problem complexity key-points are noted. The greatest time requirement is imposed by the potential/force calculation as could be inferred from the form of equation (9b), which is a $\mathcal{O}(n^2)$ problem. The remainder of calculation is in form of $c\,\mathcal{O}(n)$ where the constant c represent the additional overhead in position updates and inclusion of the bond constrain solver of no greater complexity than $\mathcal{O}(n)$.

2.3 General purpose GPU programming

The term GPGPU denotes an initiative to perform operation not related to drawing graphics on the GPU hardware. These initial attempts required to bend the problem formulation to fit into graphics drivers primitives. Situation changed in the year 2007 as

Figure 3: Die schema of Fermi microarchitecture consisting of 16 streaming multiprocessors (SM) coalesced into four graphics processing clusters (GPC)

CUDA [13] was officially introduced enabling simplified development of GPGPU applications. The definition adopted from CUDA handbook [26] specifies CUDA as "A proprietary toolchain from NVIDIA that enables C programmers to write parallel code in form of language extension". Other option how to develop for GPU is through the use of a general platform for heterogeneous computation named Open CL, or other tools. Employed technologies rely on the underlying hardware and because of this the compliant tools of CUDA toolkit [14] is used on the available NVIDIA GPU chips.

2.3.1 Hardware properties and limitations

Since 2001 of initial release a lot of things have changed and Nvidia hardware has traversed five generations of microarchitectures enabling simpler and more efficient executing of general purpose tasks. Extension ranging from the double precisions units, dedicated special function units, cache sizes and warp scheduling to double precision intrinsics, dynamic parallelism and artificial intelligence features. Together with these improvement of the peak performance of the arithmetic units, memory sizes, bandwidth and streaming multiprocessors (SM) count the capabilities of this devices were elevated into everyday computation coprocessors.

While the hardware has improved rapidly it is still not as general purpose as a CPU. The main differences arise form the structure of the chip and its construction. In example situation of increasing scaling up a simple problem such as vector multiplication the hardware limiting factor can be compared. In case of CPU it is a clock when CPU does not operate fast enough to process any more data at given time frame. The GPU on the

other hand is limited by bandwidth where the units are usually not served fast enough by the chip memory. GPU model also has to consider the processing units utilization limit and general characteristic of algorithm which may not be suited for parallel execution. Three issues into a comprehensible rules how to develop programs for GPU called the programming paradigm.

2.4 Programming paradigm

Considering the hardware limitations the paradigm is summarised into three categories: memory hierarchy, thread hierarchy and computation intensity [26, 20].

Memory hierarchy is concerned with three different memory spaces available on GPU chip: global memory, shared memory and local memory. These blue memory spaces in Figure 3 are not equal and has significant access latency. The penalty for accessing *local* –> *shared* –> *global* memory increase by factor of ≈ 7 and ≈ 10 respectively. Consequentially the program has to utilize the spaces efficiently to achieve highest utilization with available bandwidth. An example is preloading data into shared memory and operate on top of it before returning result.

The thread hierarchy regards the overall efficiency of program. Considering the grouping of the individual threads into blocks and blocks into grid the utilization of the chip resources vary. While it is a native constraint that a group of 32 threads is called warp which is usually (depending on the microarchitecture) executed simultaneously on one SM (shown in Figure 3) the overall composition of threads into blocks and grids is left for the programmer. The correct grouping of the threads is also influenced by the number of registers used by each thread which is connected to memory hierarchy. The desired output is achieving high occupancy and therefore high utilisation of GPU resources to mask latencies of warps waiting for memory. The example technique is kernel unification and thread per block variation.

The third category is the computation intensity reflecting the amount of calculation done per memory transfer. With generally costly memory access the aim is to utilise the processing units (green fields in Figure 3) enough that leaves enough time for new memory fetch. The intensity is dictated by the problem but correct instruction order can improve the computational intensity. The techniques of computation intensity includes instruction level parallelism (ILP) when one thread is assigned more work cell.

3 Solution approach

The problem of homogeneous nucleation introduced in theory overview leads to a MD setup that exhibits greater system heterogeneity, i.e. the system is composed of clusters of high density within low density surroundings. This situation disrupts the assumption made in most MD packages about computationally homogeneous system. During the conducted research this issue is addressed proposing a MD model that can avoid heterogeneity while preserving a good level of parallel execution for speed benefits.

3.1 Problem formulation

To observe nucleation, the initial configuration must be in metastable state. This requires a delicate manipulation of the system energy and box size which is beyond the scope of this brief introduction.

The simulation then proceeds as micro canonical ensemble employing the NVE type of simulation where the number of particles N, volume V and energy E are constant. The model uses the Verlet [23] algorithm with Leap-Frog scheme [7] for the step iteration. For the constant particle count the periodic boundary conditions are used. The volume of the system is fixed for the testing purposes and the energy fluctuation is used as the testing metric. While the total energy of a system should remain constant the system energy fluctuates around this mean value. In case the total energy of the system exhibit decay the numerical errors of used arithmetic and model should be considered. This requires to use the double precision(DP) instead of single precision(SP) floating-point arithmetic (FPA).

3.2 Algorithm development & parallelization

In this section the most time consuming part of simulation is considered: the force evaluation. Our first algorithm proposal for force evaluation was centred around idea of implementing heterogeneous domain decomposition algorithm that would reflect the cluster formation. The initial formulation led to an algorithm that would be more suitable for CPU implementation because of the control mechanisms complexity. For the GPU hardware the algorithm of changing domains in the proposed form were unfortunately inefficient. This also led to the abandonment of top-down implementation approach.

The analysis and trial implementations have shown that it is best to adopt the bottom-up approach. The first considered algorithm was the brute-force approach called “naive" with full force evaluation. This led to evaluation of the $N * N$ matrix of interactions shown schematically in figure 4 a). Even with the potential cutoff truncation the problem exposes enough parallelism to consider more sophisticated methods. The example method named “strip" was inspired by brute force approach adopted from the gravitational potential example in [15] is illustrated in Figure 4 b). The analysis of this type of implemented algorithms revealed the workload efficiency issues leading to improved form of “2D" methods with ILP in form of workload shown in Figure 4 c). This fact is illustrated by a thread (black dot) crossing multiple cells (i, j interaction). In schematic case of c) example the workload = 2.

It is important to note that in i index or row index all employed threads have to somehow communicate the incremented partial results. The best situation is achieved in case b) with fewest communication. On the other hand the computational strain has to be also considered which led to the c) type improvements.

The methods are compared in more detail with the sequential algorithm without optimization. The algorithm was only executed taking the symmetry of the interaction into account resulting in the halved size of interaction number. The test was calculated on Nvidia GeForce 930M (930M) during development to observe which methods yield best performance as depicted in Table 1 showing the undertaken optimization direction.

The full force implementation is easily scalable but the overall amount of work drag

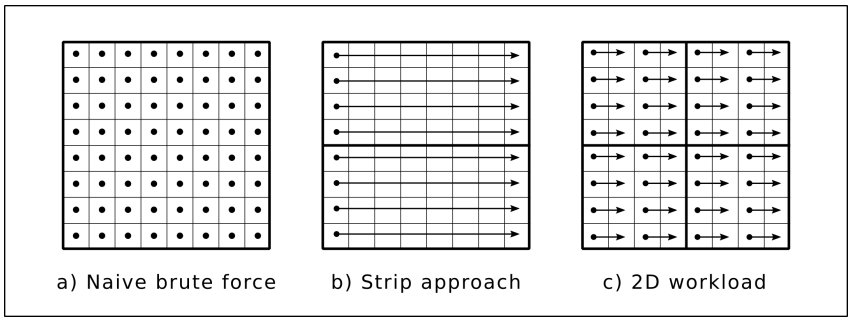


Figure 4: The illustration of the three algorithm approaches.

alg. code	type	unit	features	time per it. [ms]	speed up
sequential		CPU	half interaction	26.30	1
naive	a)	GPU	multi-kernel	2.0212	13.0153
naive II	a)	GPU	multi-kernel, preload	2.0073	13.1051
strip	b)	GPU	strips eval.	1.1486	22.9034
2D	c)	GPU	grid eval, workload	0.7630	34.4745
2D II	c)	GPU	grid eval, preload position, workload	0.7215	36.4605
2D sh	c)	GPU	grid eval, preload data, workload	0.7156	36.7598
2D sh II	c)	GPU	grid eval, preload data, workload	0.7151	36.7859

Table 1: Comparison of the developed algorithm variants for force evaluation in SP arithmetic with sequential CPU implementation. The features listed illustrates the improvement made. The tested configuration is 1024 Ar particles in periodic box $20\text{Å} \cdot 20\text{Å} \cdot 20\text{Å}$ with same scaled potential, scaled temperature $T = 1$ and cutoff= 3Å. The unit used were Intel i5-6200U and 930M (Maxwell microarchitecture).

the execution down especially when considering DP. As the solution to this issue the pre-emptive truncation of unuseful interaction has to take place. The first most used method is domain decomposition [17] partitioning the simulation space which complicates the control scheme. The second method is the Verlet list [23] a replication methods that hold multiple copies/pointers of the same particle for evaluation. The Verlet list presently used is constructed in such a way that only molecules within specified distance are included as neighbours. The initial N^2 number of required force evaluation is effectively shrunk to $< n_v * N$ where n_v is maximum number of neighbours. The neighbour list has to be reconstructed with $\mathcal{O}(N^2)$ complexity but the reconstruction is performed once after several steps lowering the interaction evaluation total count.

The last modification of present algorithm is in the form of centre of mass method of evaluation. The motion of the centre of mass is separated from the relative motion of atoms with the centre of mass. This leads to a more precise method as numerical rounding errors are lowered while during distance and velocity calculations numbers of similar order are manipulated (in molecular or atomistic layer). This approach also has implication in the SHAKE method and force evaluation where individual interaction are now performed on molecular level. In case of SPCE water model used presently for testing each molecule contain three atoms that together with another molecule result in 9 atom-atom interaction evaluation coupled into single molecule-molecule interaction. The molecule level interaction is handled by each thread resulting in increased ILP. The result of the current version of algorithm are presented in next section.

4 Intermediate results

Continuing the development of the proposed algorithm, significant speed up times were achieved. In this section the benchmark results are presented and discussed. Additionally, the graphs of the energy behaviour are shown to illustrate the computation method accuracy level.

The results of the present version with Verlet list and SHAKE method are presented in Table 2 comparing the sequential full sum(FS) algorithm and fastest link cell list(LCL) algorithms of the MACSIMUS[11] program developed by J. Kolafa. The results were obtained using two available Nvidia GPU deployed for testing. These chips are consumer GPU aimed at graphics rendering and therefore do not exhibit high DP throughputs. The nominal DP throughput of 930M is 18.4 GFLOPS while for Nvidia GeForce GTX 1070 (GTX1070) it is 180.7 GFLOPS. The GTX1070 should therefore be up to ten times faster than 930M, which is shown in case of argon in Table 2. For nitrogen and water in Table 2 the ratios are lower pointing to a possible space for optimization of the molecule bonds calculation in SHAKE procedure. The comparison of present implementation with the CPU show a great promise already outperforming the best available algorithm in MACSIMUS. It can be expected that on chips with higher DP throughput the speed-up should be better and it is our present goal to investigate this further.

Energy evolution of the same SPCE water system is shown in figure 5. We can see the evolution of the system from initial configuration with zero near potential due to the molecules not yet interacting. The following steps depict the propagation of interaction and corresponding kinetic energy compensation leading to comparatively constant total

algorithm	unit	subs	T[K]	N	time per it. [ms]	speed up FS	speed up LCL
full sum	CPU	Ar	1	4096	104	1	0.0183
link-cell list	CPU	Ar	1	4096	1.91	54.4502	1
CoM verlet	930M	Ar	1	4096	3.4271	30.3457	0.5573
CoM verlet	GTX1070	Ar	1	4096	0.3920	265.2767	4.8719
full sum	CPU	N_2	200	2048	11	1	0.1
link-cell list	CPU	N_2	200	2048	1.1	10	1
CoM verlet	930M	N_2	200	2048	2.6	4.23065	0.42306
CoM verlet	GTX1070	N_2	200	2048	0.4904	22.42754	2.24275
full sum	CPU	H_2O	350	1024	45.5	1	0.0703
link-cell list	CPU	H_2O	350	1024	3.2	14.2187	1
CoM verlet	930M	H_2O	350	1024	10.416188	4.3682	0.3072
CoM verlet	GTX1070	H_2O	350	1024	1.713735	26.5501	1.8672

Table 2: Benchmark testing the developed algorithm against CPU implementation used by MACSIMUS[11]. The tested configuration were aligned into same sized periodic box $100\text{Å} \cdot 100\text{Å} \cdot 100\text{Å}$, Verlet list cutoff vary accordingly with used substance: $tof_{\text{Ar}} = 3\text{Å}$, $cuttof_{\text{N}_2} = 12.404250\text{Å}$, $cuttof_{\text{H}_2\text{O}} = 19.031250\text{Å}$. The processing units used were Intel Duo E8600, 930M (Maxwell microarchitecture) and GTX1070 (Pascal microarchitecture).

energy.

The second Figure 6 is zoomed total energy calculation obtained from two GPU chips starting with same initial condition. This illustrates the fact that for identical initial configurations system behaviour differs. The figure shows energy graphs discrepancies intensification at higher time step numbers. Figure 6 also illustrates the energy fluctuations magnitudes. The initial total energy is elevated and after relaxation time of $\approx$7000 steps it fluctuates around mean total energy value. Notice that the standard deviation after relaxation time is no more than 100 energy computational units, four orders of magnitude lower compared to the absolute value of total energy. This is reasonable value considering the simulation settings and precision settings of the SHAKE algorithm. All simulated data presented in Table 2 and Figures 5 and 6 were computed as NVE simulation. With same block setting of 32 thread per block and higher testing ILP of 64 elements per element.

5 Conclusion

The aim of this work is the development of a fast molecular dynamics algorithm for the nucleation process calculation. This research is connected with the theme of Diploma thesis where it explores the simulation part of the nucleation phenomenon in more detail. The development of algorithm is preceded by classical nucleation theory overview, basics of Molecular dynamics and GPGPU. Theoretical findings are consequentially applied in the molecular simulation algorithm implemented for GPU with overview of development process.

This initial aim was partially fulfilled with the current version of the algorithm that is already competitive compared to the most efficient CPU implementation. The pre-

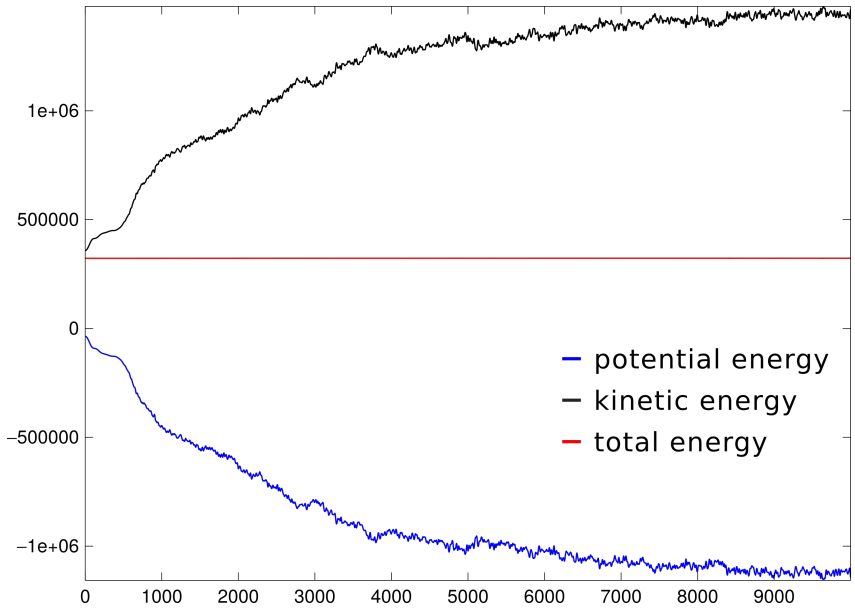


Figure 5: The energy evolution of 1024 molecules SPCE water system during first 10000 iteration steps at 350 K. The x-axis represents the time step number and y-axis is in computational energy units.

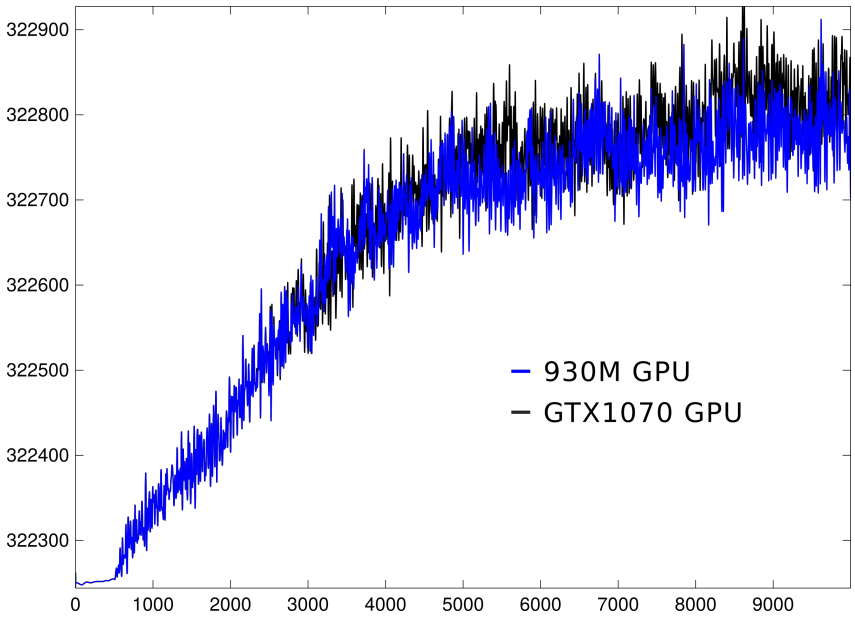


Figure 6: The total energy (red line of Figure 5 computed by 930M and GTX1070. The x-axis represent time step number and y-axis are computational energy units.

liminary energy behaviour also shows the algorithm function correctly, opting for the improvement of the model on more powerful GPU hardware. The algorithm will be further extended to account for advanced water and metal nucleation. Additional algorithm optimizations are planned with the extension into multi GPU environment.

References

[1] M. P. Allen and D. J. Tildesley. *Computer Simulation of Liquids.* Oxford University Press, 2 edition, (2017).

[2] R. Becker and W. Döring. *Kinetic treatment of the formation of nuclei in oversaturated steam.* Ann Phys **5** (1935), 719–752.

[3] P. Brault, S. Chuon, and J.-M. Bauchire. *Molecular dynamics simulations of platinum plasma sputtering: A comparative case study.* FRONTIERS IN PHYSICS **4** (MAY 10 2016).

[4] D. G. Fahrenheit. *Experimenta & observationes de congelatione aquae in vacuo factae a dg fahrenheit, rss.* Philosophical Transactions (1683-1775) **33** (1724), 78–84.

[5] D. Frenkel and B. Smit. *Understanding Molecular Simulation, Second Edition: From Algorithms to Applications (Computational Science Series 1).* 2 edition, (2001).

[6] J. W. Gibbs. *On the equilibrium of heterogeneous substances.* Trans. Connecticut Akad. (1878).

[7] R. W. Hockney. *The potential calculation and some applications.* Methods Comput. Phys. **9** (1970), 136.

[8] V. Kalikmanov. *Nucleation theory*, volume 860. Springer, (2012).

[9] D. Kashchiev. *Nucleation.* Elsevier, (2000).

[10] K. Kelton and A. L. Greer. *Nucleation in condensed matter: applications in materials and biology*, volume 15. Elsevier, (2010).

[11] J. Kolafa. Macsimus. http://old.vscht.cz/fch/software/macsimus/. Accessed: 08.09.2018.

[12] M. Kolovratník, J. Hrubý, V. Ždímal, O. Bartoš, I. Jiříček, P. Moravec, and N. Zíková. *Nanoparticles found in superheated steam: a quantitative analysis of possible heterogeneous condensation nuclei.* Proceedings of the Institution of Mechanical Engineers, Part A: Journal of Power and Energy **228** (2014), 186–193.

[13] J. Nickolls, I. Buck, M. Garland, and K. Skadron. *Scalable parallel programming with cuda.* In 'ACM SIGGRAPH 2008 classes', 16. ACM, (2008).

[14] NVIDIA®. Cuda toolkit. https://developer.nvidia.com/cuda-toolkit. Accessed: 08.09.2018.

[15] L. Nyland, M. Harris, and J. Prins. *Gpu gems 3.* Fast N-Body Simulation with CUDA (2007).

[16] W. Ostwald. *Studien über die bildung und umwandlung fester körper.* Zeitschrift für physikalische Chemie **22** (1897), 289–330.

[17] M. Pinches, D. Tildesley, and W. Smith. *Large scale molecular dynamics on parallel computers using the link-cell algorithm.* Molecular Simulation **6** (1991), 51–87.

[18] D. C. Rapaport. *The art of molecular dynamics simulation.* Cambridge university press, (2004).

[19] J.-P. Ryckaert, G. Ciccotti, and H. J. Berendsen. *Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes.* Journal of Computational Physics **23** (1977), 327–341.

[20] J. Sanders and E. Kandrot. *CUDA by example: an introduction to general-purpose GPU programming.* Addison-Wesley Professional, (2010).

[21] W. Thomson. *4. on the equilibrium of vapour at a curved surface of liquid.* Proceedings of the Royal Society of Edinburgh **7** (1872), 63–68.

[22] H. Vehkamäki. *Classical Nucleation Theory in Multicomponent Systems.* Springer-Verlag Berlin Heidelberg, 1 edition, (2006).

[23] L. Verlet. *Computer" experiments" on classical fluids. i. thermodynamical properties of lennard-jones molecules.* Physical review **159** (1967), 98.

[24] M. Volmer and A. Weber. *Nuclei formation in supersaturated states.* Z. Physik. Chem., ll9 (1925), 277–301.

[25] K. Wegner, P. Piseri, H. V. Tafreshi, and P. Milani. *Cluster beam deposition: a tool for nanoscale science and technology.* JOURNAL OF PHYSICS D-APPLIED PHYSICS **39** (NOV 21 2006), R439–R459.

[26] N. Wilt. *The cuda handbook: A comprehensive guide to gpu programming.* Pearson Education, (2013).

[27] Y. B. Zeldovich. *On the theory of new phase formation: cavitation.* Acta Physicochem., USSR **18** (1943), 1.

[28] M. Čenský, J. Hrubý, V. Vinš, J. Hykl, and B. Šmíd. *Investigation of droplet nucleation in ccs relevant systems–design and testing of the expansion chamber.* In 'EPJ Web of Conferences', volume 180, 02015. EDP Sciences, (2018).

Translation and Rotation Invariant Method of Renyi Dimension Estimation*

Martin Dlask

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

advisors:

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

Pavel Sovka, Department of Circuit Theory
Faculty of Electrical Engineering, CTU in Prague

Abstract. A fractal dimension is a non-integer characteristic that measures the space filling of an arbitrary Lebesgue-measurable set. The conventional methods usually provide a biased estimation of the fractal dimension, and therefore it is necessary to develop more complex methods for its estimation. A new characteristic based on the Parzen estimate formula is presented, and for the analysis of correlation dimension, a novel approach that employs the log-linear dependence of a modified Renyi entropy is used. The new formula for the Renyi entropy has been investigated both theoretically and experimentally on selected fractal sets.

Keywords: Parzen estimate, Renyi entropy, Monte Carlo, Renyi dimension

Abstrakt. Fraktální dimenze je neceločíselná charakteristika, která je definovaná pro každou Lebegueovsky měřitelnou množinu a udává, jaké množství prostoru vyplňuje. Tradiční přístupy jejího odhadu vedou často pouze k vychýlenému odhadu, a proto je zapotřebí vyvíjet nové komplexnější metody, které by odhad fraktální dimenze zpřesnily. Nová metoda představená ve článku je založena na Parzenově odhadu a slouží k výpočtu korelační dimenze při využití logaritmicko-lineární závislosti na Rényho entropii. Nový předpis pro odhad Rényho entropie byl teoreticky dokázán a následně byla jeho užitečnost experimentálně ověřena na fraktálních množinách se známou dimenzí.

Klíčová slova: Parzenův odhad, Rényho entropie, Monte Carlo, Rényho dimenze

Full paper: M. Dlask, J. Kukal. *Translation and rotation invariant method of Renyi dimension estimation.* Chaos, Solitons & Fractals, Elsevier BV. **114** (2018), 536–541.

*This work has been supported by the grant SGS14/208/OHK4/3T/14.

Numerical Analysis of Immersed Boundary – Lattice Boltzmann Method for Fluid-Structure Interaction*

Pavel Eichler

1st year of PGS, email: eichlpa1@fjfi.cvut.cz
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Radek Fučík, Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. In this article, we deal with the numerical immersed boundary-lattice Boltzmann method for simulation of fluid-structure interaction in 2D. We consider an interaction of incompressible, Newtonian fluid in an isothermal system with an elastic fiber, which represents the immersed body boundary. First, a short introduction to the lattice Boltzmann and immersed boundary method is presented and the combination of these two methods is briefly discussed. Next, the choice of the smooth approximation of the Dirac delta function and the discretization of the immersed body is discussed. One of the major drawbacks of the immersed boundary method is the penetrative flow through the immersed boundary that should be impermeable and thus the effect of the immersed body boundary discretization is analyzed on the deformation of an elastic fiber tightened in the domain with a cavity flow. The results show that the restrictions placed on the discretization in literature are not necessary. In the last part, the deformation of an elastic fiber with one fixed end behind a cylindrical obstacle is studied.

Keywords: Cascaded lattice Boltzmann method, Immersed boundary method, Penalty Immersed boundary method, Computational study, Lagrangian point spacing.

Abstrakt. V tomto příspěvku se zabýváme numerickou metodou vnořené hranice a metodou lattice Boltzmann pro simulaci interakce tekutiny s překážkou ve 2D. Předpokládáme interakci nestlačitelné, newtonovské kapaliny v izotermálním systému s elastickým vláknem, které představuje hranici vnořeného tělesa. Nejprve je uveden krátký úvod do metody lattice Boltzmann, metody vnořené hranice a krátce je uvedena jejich kombinace. Dále je diskutována volba aproximace Diracovy delta funkce a diskretizace vnořeného tělesa. Jedna z hlavních nevýhod metody vnořené hranice je penetrativní tok skrz hranici, která je nepropustná, a proto je analyzován vliv diskretizace vnořeného tělesa na úloze deformace elastického vlákna nataženého v oblasti s kavitačním prouděním. Výsledky ukazují, že omezení kladená na diskretizaci v literatuře nejsou nutná. V poslední části je diskutována deformace elastického vlákna s jedním pevným koncem za kruhovou překážkou.

Klíčová slova: Kaskádová lattice Boltzmannova metoda, Metoda vnořené hranice, Penalizovaná metoda vnořené hranice, Výpočetní studie, Vzdálenost lagrangeovských bodů.

*This work has been supported by the Czech Science Foundation project No. 18-09539S, by the grant No. SGS17/194/OHK4/3T/14 of the Grant Agency of the Czech Technical University in Prague and by the project No. 15-27178A of the Ministry of Health of the Czech Republic.

Full paper: P. Eichler, R. Fučík, R. Straka. *Computational study of immersed boundary – lattice Boltzmann method for fluid interaction with an elastic body.* Submitted to Discrete & Continuous Dynamical Systems Series S (2018).

Area-Level Gamma Mixed Model*

Ondřej Faltys

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

advisor: Tomáš Hobza, Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. In practise we can encounter many problems where is useful to employ small area estimation (SAE) methods to obtain reliable estimates of characteristics of interest (e.g. means, totals). The contribution deals with an area-level gamma mixed model where responses have the gamma distribution. To obtain estimates of regression parameters and predictors of random effects, the ML Laplace approximation algorithm is introduced. Subsequently, an algorithm for calculating empirical best predictors (EBP) is derived. Simulation experiments are conducted to check the behaviour of EBP and Plug-in predictor.

Keywords: area-level model, Laplace approximation algorithm, EBP, simulations

Abstrakt. V praxi lze narazit na řadu problémů, kdy je užitečné použít metody odhadování v malých oblastech, abychom získali odhady charakteristik, které nás zajímají (např. středních hodnot, celkového součtu). Tento článek pojednává o statistickém modelu na úrovni oblastí, kde odezvy mají gamma rozdělení. Pomocí Laplaceova aproximačního algoritmu získáme odhady regresních parametrů a predikce náhodných efektů. Následně navrhneme algoritmus pro výpočet EBP prediktoru. Simulačními experimenty poté zkoumáme vlastnosti EBP prediktoru a plug-in prediktoru.

Klíčová slova: model na úrovni oblastí, Laplaceův aproximační algoritmus, EBP, simulace

1 Introduction

There are two kinds of models that are distinguished in Small Area Estimation (SAE): area-level models and unit-level models. Under a term, small area, we can imagine a real geographical area or a socio-economic group. The number of areas is finite and we denote it as D. Let N_d be the total of individuals in area d, $d = 1, \ldots, D$. We can suppose that a sample of length $n_d < N_d$ was obtained, e.g. by a survey, from each area. If the data are available for each individual (i.e. responses and associated auxiliary variables), we refer to a unit-level model. If not, we can use the collected data to compute a direct estimate (e.g. mean) that represents a response for a whole area, i.e. there is one response for each area. Auxiliary data must also have an area-level interpretation. In this case we refer to area-level model.

One of the basic area-level models is the Fay-Herriot model that can be expressed as (see [1])

$$y_d = \mathbf{x}_d^T \boldsymbol{\beta} + v_d + e_d, \quad d = 1, \ldots, D, \tag{1}$$

*This work has arisen in cooperation with Domingo Morales: the author used some parts from a still not published article dealing with this topic.

where $\boldsymbol{\beta}$ is a vector of regression parameters, $e_d \sim N(0, \sigma_d^2)$ are independent sampling errors and $v_d \sim N(0, \sigma_v^2)$ are independent random effects. It is assumed that the random effects are independent on the samplings errors and the variances $\sigma_1^2, \ldots, \sigma_D^2$ are known. The model has $p+1$ unknown parameters: $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ and σ_v^2. The vector $\mathbf{x}_d$ denotes the auxiliary data. The task is then to estimate the quantity $\mu_d = \mathbf{x}_d^T\boldsymbol{\beta} + v_d$ to improve the direct estimates y_d $(d = 1, \ldots, D)$. That is equivalent to minimize the sampling errors e_d. In this article we suppose that the responses y_d have the gamma distribution and the structure of the model is the same as in (1).

2 Model

We consider a set of random effects $\{v_d : d = 1, \ldots, D\}$ such that $v_d \overset{\text{iid}}{\sim} N(0, 1)$. In matrix notation we have $\mathbf{v} = (v_1, \ldots, v_D)^T \sim N_D(\mathbf{0}, \mathbf{I}_D)$, i.e.

$$f_{\mathbf{v}}(\mathbf{v}) = \frac{1}{(2\pi)^{D/2}} \exp\left\{-\frac{1}{2}\mathbf{v}^T\mathbf{v}\right\}.$$

The conditional distribution of the target variable y_d given v_d is

$$y_d|v_d \sim \text{Gamma}\left(\nu_d, a_d = \frac{\nu_d}{\mu_d}\right), \quad d = 1, \ldots, D$$

and the density follows

$$f(y_d|v_d) = \frac{a_d^{\nu_d}}{\Gamma(\nu_d)} y_d^{\nu_d - 1} \exp\{-a_d y_d\} I_{(0,\infty)}(y_d) = \left(\frac{\nu_d}{\mu_d}\right)^{\nu_d} \frac{y_d^{\nu_d-1}}{\Gamma(\nu_d)} \exp\left\{-\frac{\nu_d}{\mu_d} y_d\right\} I_{(0,\infty)}(y_d).$$

The expectation and variance of the conditional random variable y_d given v_d are

$$E[y_d|v_d] = \frac{\nu_d}{a_d} = \mu_d, \quad \text{var}[y_d|v_d] = \frac{v_d}{a_d^2} = \frac{\mu_d^2}{\nu_d}.$$

The canonical link for the gamma distribution (see [2]) is the inverse link, $g(x) = \frac{1}{x}$, then we model the conditional expectation μ_d as

$$g(\mu_d) = \frac{1}{\mu_d} = \mathbf{x}_d^T\boldsymbol{\beta} + \phi v_d, \quad d = 1, \ldots, D, \tag{2}$$

where $\beta = (\beta_1, \ldots, \beta_p)^T$ and $\mathbf{x}_d^T = (x_{d1}, \ldots, x_{dp})$. Considering the data $\mathbf{y} = (y_1, \ldots, y_D)^T$ satisfy the assumptions of GLMM (see [3]) the random variables $y_d|v_d$, $i = 1, \ldots, D$, are independent, i.e. $f(\mathbf{y}|\mathbf{v}) = \prod_{i=1}^D f(y_d|v_d)$. Finally, we get

$$f(\mathbf{y}) = \int_{\mathbb{R}^D} f(\mathbf{y}|\mathbf{v}) f_{\mathbf{v}}(\mathbf{v}) \mathrm{d}\mathbf{v} = \int_{\mathbb{R}^D} \psi(\mathbf{y}, \mathbf{v}) \mathrm{d}\mathbf{v}, \tag{3}$$

where

$$\begin{aligned}\psi(\mathbf{y},\mathbf{v}) &= (2\pi)^{-D/2}\exp\left\{-\frac{\mathbf{v}^T\mathbf{v}}{2}\right\}\prod_{d=1}^{D}\left(\frac{\nu_d}{\mu_d}\right)^{\nu_d}\frac{y_d^{\nu_d-1}}{\Gamma(\nu_d)}\exp\left\{-\frac{\nu_d}{\mu_d}y_d\right\}\\ &= (2\pi)^{-D/2}\exp\left\{-\frac{\mathbf{v}^T\mathbf{v}}{2}\right\}\left(\prod_{d=1}^{D}\frac{\nu_d^{\nu_d}y_d^{\nu_d-1}}{\Gamma(\nu_d)}\right)\exp\left\{\sum_{d=1}^{D}\nu_d\log(\mathbf{x}_d^T\boldsymbol{\beta}+\phi v_d)\right\}\times\\ &\times\exp\left\{-\sum_{k=1}^{p}\left(\sum_{d=1}^{D}\nu_d y_d x_{dk}\right)\beta_k-\phi\sum_{d=1}^{D}\nu_d y_d v_d\right\}.\end{aligned}$$

The partial derivatives of $\mu_d=\frac{1}{\mathbf{x}_d^T\boldsymbol{\beta}+\phi v_d}$ are

$$\frac{\partial\mu_d}{\partial\beta_r}=-\frac{x_{dr}}{(\mathbf{x}_d^T\boldsymbol{\beta}+\phi v_d)^2}=-x_{dr}\mu_d^2,\quad \frac{\partial\mu_d}{\partial\phi}=-\frac{v_d}{(\mathbf{x}_d^T\boldsymbol{\beta}+\phi v_d)^2}=-v_d\mu_d^2.$$

There are $p+1$ unknown parameters in this model: $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_p)^T$ and ϕ. Due to the fact that the integral in (3) cannot be calculated explicitly we employ the ML Laplace approximation algorithm to obtain estimates of these parameters.

Remark: In practise, y_d is a direct estimate of a domain total or mean with estimated design-based variance $\sigma_d^2=\mathrm{var}_\pi(y_d)$. By equating $\mathrm{var}(y_d|v_d)$ to σ_d^2 and substituting μ_d by y_d, we get $\sigma_d^2=\frac{y_d^2}{\nu_d}$.

3 ML Laplace approximation algorithm

3.1 Laplace approximation to the likelihood

Let $h:\mathbb{R}\mapsto\mathbb{R}$ be a twice continuously differentiable function with a global maximum at x_0, i.e. $\dot{h}(x_0)=0$ and $\ddot{h}(x_0)<0$. Taylor's series expansion of $h(x)$ around x_0 yields to

$$h(x)=h(x_0)+\frac{1}{2}\ddot{h}(x_0)(x-x_0)^2+o(|x-x_0|^2)\approx h(x_0)+\frac{1}{2}\ddot{h}(x_0)(x-x_0)^2.$$

The univariate Laplace approximation is

$$\begin{aligned}\int_{-\infty}^{\infty}\mathrm{e}^{h(x)} &\approx \int_{-\infty}^{\infty}\mathrm{e}^{h(x_0)}\exp\left\{-\frac{1}{2}(-\ddot{h}(x_0))(x-x_0)^2\right\}\mathrm{d}x\\ &= (2\pi)^{1/2}(-\ddot{h}(x_0))^{-1/2}\mathrm{e}^{h(x_0)}\int_{-\infty}^{\infty}\frac{\exp\left\{-\frac{1}{2}\left(\frac{x-x_0}{(-\ddot{h}(x_0))^{-1/2}}\right)^2\right\}}{(2\pi)^{1/2}(-\ddot{h}(x_0))^{-1/2}}\mathrm{d}x\\ &= (2\pi)^{1/2}(-\ddot{h}(x_0))^{-1/2}\mathrm{e}^{h(x_0)}.\end{aligned}\tag{4}$$

Recalling assumptions, $v_1,\ldots,v_d\sim N(0,1)$ are independent and

$$y_d|v_d\overset{\text{ind}}{\sim}\text{Gamma}\left(\nu_d,\frac{\nu_d}{\mu_d}\right),\quad \mu_d=\mu_d(v_d)=(\mathbf{x}_d^T\boldsymbol{\beta}+\phi v_d)^{-1},\quad d=1,\ldots,D.$$

The marginal density of y_d can be expressed as

$$\begin{aligned} f(y_d) &= \int_{-\infty}^{\infty} f(y_d|v_d) f(v_d) \mathrm{d}v_d \\ &= \int_{-\infty}^{\infty} \frac{\nu_d^{\nu_d} y_d^{\nu_d - 1}}{(2\pi)^{1/2}\Gamma(\nu_d)} \exp\{\nu_d \log(\mathbf{x}_d^T\boldsymbol{\beta} + \phi v_d) - \nu_d y_d(\mathbf{x}_d^T\boldsymbol{\beta} + \phi v_d)\} \exp\left\{-\frac{1}{2}v_d^2\right\} \mathrm{d}v_d \\ &= \frac{\nu_d^{\nu_d} y_d^{\nu_d - 1}}{(2\pi)^{1/2}\Gamma(\nu_d)} \int_{-\infty}^{\infty} \exp\left\{-\frac{v_d^2}{2} + \nu_d \log(\mathbf{x}_d^T\boldsymbol{\beta} + \phi v_d) - \nu_d y_d(\mathbf{x}_d^T\boldsymbol{\beta} + \phi v_d)\right\} \mathrm{d}v_d \\ &= \frac{\nu_d^{\nu_d} y_d^{\nu_d - 1}}{(2\pi)^{1/2}\Gamma(\nu_d)} \int_{-\infty}^{\infty} \exp\{h(v_d)\} \mathrm{d}v_d, \end{aligned}$$

where

$$\begin{aligned} h(v_d) &= -\frac{v_d^2}{2} + \nu_d \log(\mathbf{x}_d^T\boldsymbol{\beta} + \phi v_d) - \nu_d y_d(\mathbf{x}_d^T\boldsymbol{\beta} + \phi v_d), \qquad (5) \\ \dot{h}(v_d) &= -v_d + \frac{\nu_d \phi}{\mathbf{x}_d^T\boldsymbol{\beta} + \phi v_d} - \phi\nu_d y_d = -v_d + \phi\nu_d\mu_d(v_d) - \phi\nu_d y_d, \\ \ddot{h}(v_d) &= -\left(1 + \frac{\phi^2\nu_d}{(\mathbf{x}_d^T\boldsymbol{\beta} + \phi v_d)^2}\right) = -(1 + \phi^2\nu_d\mu_d^2(v_d)). \end{aligned}$$

Let v_{0d} denote the global maximum of h then $\dot{h}(v_{0d}) = 0$ and $\ddot{h}(v_{0d}) < 0$. By applying (4) in $v_d = v_{0d}$, we get

$$\begin{aligned} f(y_d) \approx{}& \frac{\nu_d^{\nu_d} y_d^{\nu_d - 1}}{\Gamma(\nu_d)} (1 + \phi^2\nu_d\mu_d^2(v_{0d}))^{-1/2} \times \\ & \times \exp\left\{-\frac{v_{0d}^2}{2} + \nu_d \log(\mathbf{x}_d^T\boldsymbol{\beta} + \phi v_{0d}) - \nu_d y_d(\mathbf{x}_d^T\boldsymbol{\beta} + \phi v_{0d})\right\}. \end{aligned}$$

It holds that $y_1, \ldots, y_D$ are unconditionally independent and then the likelihood has the form $L(\boldsymbol{\beta}, \phi) = \prod_{i=1}^{D} f(y_i)$. The log-likelihood is $l(\boldsymbol{\beta}, \phi) = \sum_{d=1}^{D} l_d$, where

$$\begin{aligned} l_d = \log f(y_d) \approx l_{0d} = \log \frac{\nu_d^{\nu_d} y_d^{\nu_d - 1}}{\Gamma(\nu_d)} - \frac{1}{2}\log\xi_{0d} - \frac{v_{0d}^2}{2} + \nu_d \log(\mathbf{x}_d^T\boldsymbol{\beta} + \phi v_{0d}) \\ - \nu_d y_d(\mathbf{x}_d^T\boldsymbol{\beta} + \phi v_{0d}), \end{aligned}$$

where $\xi_{0d} = 1 + \phi^2\nu_d\mu_{0d}^2$ and $\mu_{0d} = \mu_d(v_{0d})$. The first derivatives of μ_{0d} and ξ_{0d} are

$$\begin{aligned} \frac{\partial\mu_{0d}}{\partial\beta_r} &= -x_{dr}\mu_{0d}^2, \quad \eta_{0dr} = \frac{\partial\xi_{0d}}{\partial\beta_r} = -2\phi^2\nu_d x_{dr}\mu_{0d}^3, \\ \frac{\partial\mu_{0d}}{\partial\phi} &= -v_{0d}\mu_{0d}^2, \quad \eta_{0d} = \frac{\partial\xi_{0d}}{\partial\phi} = 2\phi\nu_d\mu_{0d}^2 - 2\phi^2\nu_d v_{0d}\mu_{0d}^3. \end{aligned}$$

The first derivatives of l_{0d} with respect to β_r and ϕ are

$$\frac{\partial l_{0d}}{\partial\beta_r} = -\frac{1}{2}\frac{\eta_{0dr}}{\xi_{0d}} + \nu_d x_{dr}\mu_{0d} - \nu_d x_{dr} y_d, \quad \frac{\partial l_{0d}}{\partial\phi} = -\frac{1}{2}\frac{\eta_{0d}}{\xi_{0d}} + \nu_d v_{0d}\mu_{0d} - \nu_d v_{0d} y_d.$$

It holds that

$$\frac{\partial \eta_{0dr}}{\partial \beta_s} = 6\phi^2 \nu_d x_{dr} x_{ds} \mu_{0d}^4, \quad \frac{\partial \eta_{0dr}}{\partial \phi} = -4\phi \nu_d x_{dr} \mu_{0d}^3 + 6\phi^2 \nu_d x_{dr} v_{0d} \mu_{0d}^4,$$
$$\frac{\partial \eta_{0d}}{\partial \beta_r} = -4\phi \nu_d x_{dr} \mu_{0d}^3 + 6\phi^2 \nu_d v_{0d} x_{dr} \mu_{0d}^4, \quad \frac{\partial \eta_{0d}}{\partial \phi} = 2\nu_d \mu_{0d}^2 - 8\phi \nu_d v_{0d} \mu_{0d}^3 + 6\phi^2 \nu_d v_{0d}^2 \mu_{0d}^4.$$

The second partial derivatives of l_d are

$$\frac{\partial^2 l_{0d}}{\partial \beta_s \partial \beta_r} = -\frac{1}{2} \frac{\frac{\partial \eta_{0dr}}{\partial \beta_s} \xi_{0d} - \eta_{0dr} \eta_{0ds}}{\xi_{0d}^2} - \nu_d x_{dr} x_{ds} \mu_{0d}^2,$$
$$\frac{\partial^2 l_{0d}}{\partial \phi \partial \beta_r} = -\frac{1}{2} \frac{\frac{\partial \eta_{0dr}}{\partial \phi} \xi_{0d} - \eta_{0dr} \eta_{0d}}{\xi_{0d}^2} - \nu_d v_{0d} x_{dr} \mu_{0d}^2,$$
$$\frac{\partial^2 l_{0d}}{\partial \phi^2} = -\frac{1}{2} \frac{\frac{\partial \eta_{0d}}{\partial \phi} \xi_{0d} - \eta_{0d}^2}{\xi_{0d}^2} - \nu_d v_{0d}^2 \mu_{0d}^2.$$

For $r, s = 1, \ldots, p+1$, the components of the score vector and the Hessian matrix are

$$U_{0r} = \sum_{d=1}^{D} \frac{\partial l_{0d}}{\partial \beta_r}, \quad U_{0p+1} = \sum_{d=1}^{D} \frac{\partial l_{0d}}{\partial \phi},$$
$$H_{0rs} = H_{0sr} = \sum_{d=1}^{D} \frac{\partial^2 l_{0d}}{\partial \beta_s \partial \beta_r}, H_{0rp+1} = H_{0p+1r} = \sum_{d=1}^{D} \frac{\partial^2 l_{0d}}{\partial \phi \partial \beta_r}, H_{0p+1p+1} = \sum_{d=1}^{D} \frac{\partial^2 l_{0d}}{\partial \phi^2}.$$

In matrix form we have $\mathbf{U}_0 = \mathbf{U}_0(\boldsymbol{\theta}) = (U_{01}, \ldots, U_{0p+1})^T$ and $\mathbf{H}_0 = \mathbf{H}_0(\boldsymbol{\theta}) = (H_{0rs})_{r,s=1,\ldots,p+1}$, where $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \phi)^T$. The Newton-Raphson algorithm maximizes $l_0(\boldsymbol{\theta})$, with fixed $v_d = v_{0d}$, $d = 1, \ldots, D$. The updating equation is

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \mathbf{H}_0^{-1}(\boldsymbol{\theta}^{(k)}) \mathbf{U}_0(\boldsymbol{\theta}^{(k)}). \tag{6}$$

For $d = 1, \ldots, D$, the Newton-Raphson algorithm maximizes $h(v_d) = h(v_d, \boldsymbol{\theta})$, defined in (5), with $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ fixed. The updating equation is

$$v_d^{(k+1)} = v_d^{(k)} - \frac{\dot{h}(v_d^{(k)}, \boldsymbol{\theta}_0)}{\ddot{h}(v_d^{(k)}, \boldsymbol{\theta}_0)}. \tag{7}$$

3.2 Algorithm

The ML Laplace approximation algorithm is

1. Set the initial values $k = 0$, $\boldsymbol{\theta}^{(0)}$, $\boldsymbol{\theta}^{(-1)} = \boldsymbol{\theta}^{(0)} + \mathbf{1}_{p+1}$, $v_d^{(0)} = 0$, $v_d^{(-1)} = 1$, $d = 1, \ldots, D$.
2. Until $||\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^{(k-1)}|| < \varepsilon_1$, $|v_d^{(k)} - v_d^{(k-1)}| < \varepsilon_2$, $d = 1, \ldots, D$, do
   (a) Apply algorithm (7) with seeds $v_d^{(k)}$, $d = 1, \ldots, D$, convergence tolerance ε_2 and $\boldsymbol{\theta} = \boldsymbol{\theta}^{(k)}$ fixed. Output: $v_d^{(k+1)}$, $d = 1, \ldots, D$.

(b) Apply algorithm (6) with seed $\boldsymbol{\theta}^{(k)}$, convergence tolerance ε_1 and $v_{0d} = v_d^{(k+1)}$ fixed, $d = 1, \ldots, D$. Output: $\boldsymbol{\theta}^{(k+1)}$.

(c) $k \leftarrow k + 1$

3. Output: $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^{(k)}$, $\hat{v}_d = v_d^{(k)}$, $d = 1, \ldots, D$.

4 Empirical best predictors

We introduce the empirical best predictors (EBP) and derive a corresponding algorithm for numerical simulations. As already mentioned, the conditional distribution of $\mathbf{y}$, given $\mathbf{v}$, is

$$f(\mathbf{y}|\mathbf{v}) = \prod_{d=1}^{D} f(y_d|v_d)\,,$$

where

$$f(y_d|v_d) = \exp\left\{\log\frac{\nu_d^{\nu_d} y_d^{\nu_d - 1}}{\Gamma(\nu_d)} + \nu_d \log(\mathbf{x}_d^T\boldsymbol{\beta} + \phi v_d) - \nu_d y_d(\mathbf{x}_d^T\boldsymbol{\beta} + \phi v_d)\right\}$$

and the p.d.f. of $\mathbf{v}$ is

$$f(\mathbf{v}) = \prod_{d=1}^{D} f(v_d), \quad f(v_d) = (2\pi)^{-1/2}\exp\left\{-\frac{1}{2}v_d^2\right\}.$$

The best predictor of $\mu_d = \mu_d(\boldsymbol{\theta}, v_d)$ is $\hat{\mu}_d(\boldsymbol{\theta}) = E_\theta[\mu_d|\mathbf{y}]$. Considering the area-level model, we have got $E_\theta[\mu_d|\mathbf{y}] = E_\theta[\mu_d|y_d]$ and

$$E_\theta[\mu_d|y_d] = \frac{\int_R (\mathbf{x}_d^T\boldsymbol{\beta} + \phi v_d)^{-1} f(y_d|v_d) f(v_d)\mathrm{d}v_d}{\int_R f(y_d|v_d) f(v_d)\mathrm{d}v_d} = \frac{\mathcal{N}_d(y_d, \boldsymbol{\theta})}{\mathcal{D}_d(y_d, \boldsymbol{\theta})} = \frac{N_d(y_d, \boldsymbol{\theta})}{D_d(y_d, \boldsymbol{\theta})}\,,$$

where $\mathcal{N}_d = \mathcal{N}_d(y_d, \boldsymbol{\theta})$, $\mathcal{D}_d = \mathcal{D}_d(y_d, \boldsymbol{\theta})$, $N_d = N_d(y_d, \boldsymbol{\theta})$ and $D_d = D_d(y_d, \boldsymbol{\theta})$ are

$$\mathcal{N}_d = \int_R (\mathbf{x}_d^T\boldsymbol{\beta} + \phi v_d)^{-1} \exp\left\{\log\frac{\nu_d^{\nu_d} y_d^{\nu_d - 1}}{\Gamma(\nu_d)} + \nu_d \log(\mathbf{x}_d^T\boldsymbol{\beta} + \phi v_d) - \nu_d y_d(\mathbf{x}_d^T\boldsymbol{\beta} + \phi v_d)\right\} f(v_d)\mathrm{d}v_d\,,$$

$$\mathcal{D}_d = \int_R \exp\left\{\log\frac{\nu_d^{\nu_d} y_d^{\nu_d - 1}}{\Gamma(\nu_d)} + \nu_d \log(\mathbf{x}_d^T\boldsymbol{\beta} + \phi v_d) - \nu_d y_d(\mathbf{x}_d^T\boldsymbol{\beta} + \phi v_d)\right\} f(v_d)\mathrm{d}v_d\,,$$

$$N_d = \int_R (\mathbf{x}_d^T\boldsymbol{\beta} + \phi v_d)^{-1} \exp\left\{\nu_d \log(\mathbf{x}_d^T\boldsymbol{\beta} + \phi v_d) - \nu_d y_d(\mathbf{x}_d^T\boldsymbol{\beta} + \phi v_d)\right\} f(v_d)\mathrm{d}v_d\,,$$

$$D_d = \int_R \exp\left\{\nu_d \log(\mathbf{x}_d^T\boldsymbol{\beta} + \phi v_d) - \nu_d y_d(\mathbf{x}_d^T\boldsymbol{\beta} + \phi v_d)\right\} f(v_d)\mathrm{d}v_d\,.$$

The best predictor of v_d is

$$\hat{v}_d(\boldsymbol{\theta}) = E_\theta[v_d|y_d] = \frac{\int_R v_d f(y_d|v_d) f(v_d)\mathrm{d}v_d}{\int_R f(y_d|v_d) f(v_d)\mathrm{d}v_d} = \frac{N_{v,d}(y_d, \boldsymbol{\theta})}{D_d(y_d, \boldsymbol{\theta})}\,,$$

where

$$N_{v,d}(y_d, \boldsymbol{\theta}) = \int_R v_d \exp\left\{\nu_d \log(\mathbf{x}_d^T\boldsymbol{\beta} + \phi v_d) - \nu_d y_d(\mathbf{x}_d^T\boldsymbol{\beta} + \phi v_d)\right\} f(v_d)\mathrm{d}v_d\,.$$

4.1 Algorithm

The EBP of μ_d is $\hat{\mu}_d = \hat{\mu}_d(\hat{\boldsymbol{\theta}})$ and the EBP of v_d is $\hat{v}_d = \hat{v}_d(\hat{\boldsymbol{\theta}})$. Both predictors can be approximated by the following algorithm:

1. Estimate $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}^T, \hat{\phi})^T$.
2. For $s = 1, \ldots, S$: generate $v_d^{(s)} \overset{iid}{\sim} N(0,1)$ and put $v_d^{(S+s)} = -v_d^{(s)}$.
3. Calculate $\hat{\mu}_d(\hat{\boldsymbol{\theta}}) = \hat{N}_d / \hat{D}_d$ where

$$\hat{N}_d = \frac{1}{2S} \sum_{s=1}^{2S} (\mathbf{x}_d^T \hat{\boldsymbol{\beta}} + \hat{\phi} v_d^{(s)})^{-1} \exp\left\{ \nu_d \log(\mathbf{x}_d^T \hat{\boldsymbol{\beta}} + \hat{\phi} v_d^{(s)}) - \nu_d y_d (\mathbf{x}_d^T \hat{\boldsymbol{\beta}} + \hat{\phi} v_d^{(s)}) \right\},$$

$$\hat{D}_d = \frac{1}{2S} \sum_{s=1}^{2S} \exp\left\{ \nu_d \log(\mathbf{x}_d^T \hat{\boldsymbol{\beta}} + \hat{\phi} v_d^{(s)}) - \nu_d y_d (\mathbf{x}_d^T \hat{\boldsymbol{\beta}} + \hat{\phi} v_d^{(s)}) \right\}.$$

4. Calculate $\hat{v}_d(\hat{\boldsymbol{\theta}}) = \hat{N}_{v,d} / \hat{D}_d$, where

$$\hat{N}_{v,d} = \frac{1}{2S} \sum_{s=1}^{2S} v_d^{(s)} \exp\{ \nu_d \log(\mathbf{x}_d^T \hat{\boldsymbol{\beta}} + \hat{\phi} v_d^{(s)}) - \nu_d y_d (\mathbf{x}_d^T \hat{\boldsymbol{\beta}} + \hat{\phi} v_d^{(s)}) \}.$$

Remark: The plug-in estimator of μ_d is $\tilde{\mu}_d = (\mathbf{x}_d^T \hat{\boldsymbol{\beta}} + \hat{\phi} \hat{v}_d)^{-1}$.

4.2 Bootstrap estimation of MSE

The following procedure calculates a parametric bootstrap estimator of $MSE(\hat{\mu}_d)$.

1. Fit the model to the sample and calculate the estimator $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}^T, \hat{\phi})^T$.
2. Repeat B times ($b = 1, \ldots, B$):
   (a) Generate $v_d^* \sim N(0,1)$, $y_d^* \sim \text{Gamma}\left(\nu_d, \frac{\nu_d}{\mu_d^*}\right)$, $\mu_d^* = (\mathbf{x}_d^T \hat{\boldsymbol{\beta}} + \hat{\phi} v_d^*)^{-1}$, $d = 1, \ldots, D$.
   (b) For each bootstrap sample, calculate the estimator $\hat{\boldsymbol{\theta}}^*$ and the EBP $\hat{\mu}_d^* = \hat{\mu}_d^*(\hat{\boldsymbol{\theta}}^*)$.
3. Output: $mse^*(\hat{\mu}_d) = \frac{1}{B} \sum_{b=1}^{B} (\hat{\mu}_d^{*(b)} - \mu_d^{*(b)})^2$

5 Simulation experiments

The target of simulations is to check the behaviour of the Laplace fitting algorithm. We set the true values of the regression parameters to $\beta_0 = 0.05$, $\beta_1 = 0.1$ and $\phi = 0.01$, i.e. $p = 2$. Next we suppose that $D = 50, 100, 150, 200$ is the number of domains. For $d = 1, \ldots, D$, we generate $\nu_d = 100$, $x_d = \frac{d+D}{D}$, $v_d \sim N(0,1)$ and

$$y_d | v_d \sim \text{Gamma}\left(\nu_d, \frac{\nu_d}{\mu_d}\right), \quad \mu_d = (\beta_0 + \beta_1 x_d + \phi v_d)^{-1}.$$

5.1 Simulation 1

The behaviour of the EBP and the plug-in estimator of the expectation μ_d is investigated in this simulation. The steps are

1. Repeat $K = 100$ times $(k = 1, \ldots, K)$

   1.1. Generate a sample of size D.

   1.2. Estimate $\hat{\beta}_0^{(k)}$, $\hat{\beta}_1^{(k)}$, $\hat{\phi}^{(k)}$, calculate EBP $\hat{\mu}(\hat{\boldsymbol{\theta}}^{(k)})$ and plug-in estimator $\tilde{\mu}_d^{(k)} = (\hat{\beta}_0^{(k)} + x_d\hat{\beta}_1^{(k)} + \hat{\phi}^{(k)}\hat{v}_d^{(k)})^{-1}$.

2. Output: For $\hat{\mu}_d^{(k)} \in \{\hat{\mu}(\hat{\boldsymbol{\theta}}^{(k)}), \tilde{\mu}_d^{(k)}\}$, $d = 1, \ldots, D$, calculate

$$\bar{\mu}_d = \frac{1}{K}\sum_{k=1}^{K}\mu_d^{(k)},\ RB_d = \frac{\sum_{k=1}^{K}(\hat{\mu}_d^{(k)} - \mu_d^{(k)})}{K|\bar{\mu}_d|},\ RE_d = \frac{\sqrt{\frac{1}{K}\sum_{k=1}^{K}(\hat{\mu}_d^{(k)} - \mu_d^{(k)})^2}}{|\bar{\mu}_d|},$$

$$B_d = \frac{1}{K}\sum_{k=1}^{K}(\hat{\mu}_d^{(k)} - \mu_d^{(k)}),\ E_d = \frac{1}{K}\sum_{k=1}^{K}(\hat{\mu}_d^{(k)} - \mu_d^{(k)})^2,$$

$$B = \frac{1}{D}\sum_{d=1}^{D}B_d,\ RB = \frac{1}{D}\sum_{d=1}^{D}RB_d,\ E = \frac{1}{D}\sum_{d=1}^{D}E_d,\ RE = \frac{1}{D}\sum_{d=1}^{D}RE_d.$$

	$D = 50$	$D = 100$	$D = 150$	$D = 200$
EBP	5.67e-03	-6.39e-03	-2.61e-03	-2.97e-03
PLUG	-7.75e-04	-1.34e-02	-9.98e-03	-1.08e-02

Table 5.1.1. B.

	$D = 50$	$D = 100$	$D = 150$	$D = 200$
EBP	1.39e-03	-1.07e-03	-3.56e-04	-6.87e-04
PLUG	1.46e-04	-2.41e-03	-1.77e-03	-2.17e-03

Table 5.1.2. RB (in %).

	$D = 50$	$D = 100$	$D = 150$	$D = 200$
EBP	7.72e-02	7e-02	6.71e-02	6.45e-02
PLUG	7.71e-02	7e-02	6.72e-02	6.45e-02

Table 5.1.3. E.

	$D = 50$	$D = 100$	$D = 150$	$D = 200$
EBP	5.23e-02	4.86e-02	4.79e-02	4.69e-02
PLUG	5.23e-02	4.86e-02	4.79e-02	4.69e-02

Table 5.1.4. RE (in %).

5.2 Simulation 2

The behaviour of the bootstrap MSE estimator $mse^*(\hat{\mu}_d)$ is investigated in this simulation. The steps are

1. Repeat $K = 100$ times $(k = 1, \ldots, K)$

   1.1. Generate a sample of size D.

   1.2. Estimate $\hat{\boldsymbol{\theta}}^{(k)} = (\hat{\beta}_0^{(k)}, \hat{\beta}_1^{(k)}, \hat{\phi}^{(k)})$, calculate $mse_d^{*(k)} = mse_d^*(\hat{\mu}_d(\hat{\boldsymbol{\theta}}^{(k)}))$, $d = 1, \ldots, D$, and take E_d from Simulation 1.

2. Output:

$$Rb_d = \frac{\frac{1}{K}\sum_{k=1}^{K}(mse_d^{*(k)} - E_d)}{|E_d|},\ Re_d = \frac{\sqrt{\frac{1}{K}\sum_{k=1}^{K}(mse_d^{*(k)} - E_d)^2}}{|E_d|},$$

$$b_d = \frac{1}{K}\sum_{k=1}^{K}(mse_d^{*(k)} - E_d),\ e_d = \frac{1}{K}\sum_{k=1}^{K}(mse_d^{*(k)} - E_d)^2,$$

$$b = \frac{1}{D}\sum_{d=1}^{D} b_d,\ Rb = \frac{1}{D}\sum_{d=1}^{D} Rb_d,\ e = \frac{1}{D}\sum_{d=1}^{D} e_d,\ Re = \frac{1}{D}\sum_{d=1}^{D} Re_d.$$

	$D = 50$	$D = 100$	$D = 150$	$D = 200$
EBP	-1.13e-02	-9.13e-03	-4.57e-03	-3.33e-03
PLUG	-1.14e-02	-9.41e-03	-4.72e-03	-3.46e-03

Table 5.2.1. b.

	$D = 50$	$D = 100$	$D = 150$	$D = 200$
EBP	-0.15	-8.87e-02	-3.96e-02	-2.80e-02
PLUG	-0.15	-9.15e-02	-4.08e-02	-3.01e-02

Table 5.2.2. Rb (in %).

	$D = 50$	$D = 100$	$D = 150$	$D = 200$
EBP	2.07e-03	1.41e-03	1.13e-03	8.53e-04
PLUG	2.06e-03	1.41e-03	1.13e-03	8.52e-04

Table 5.2.3. e.

	$D = 50$	$D = 100$	$D = 150$	$D = 200$
EBP	5.4e-01	5e-01	4.6e-01	4.12e-01
PLUG	5.4e-01	5e-01	4.6e-01	4.12e-01

Table 5.2.4. Re (in %).

The results of the first simulation show us that both predictors (EBP, Plug) behave in a very similar way. The average variance E, table 5.1.3., is almost the same for both predictors and falls slightly as D grows. This fact is even more obvious in table 5.1.4. showing the relative average variance RE. The relative error is about 5% for both predictors as can be seen from this table. Tables 5.1.1. and 5.1.2. represents the average deviation and the relative average deviation of both predictors.

In the second simulation we try to estimate the variance E_d $(d = 1, \ldots, D)$ from the first simulation by means of the bootstrap algorithm. Statistics b_d and Rb_d are the average deviation and the relative average deviation of mse_d^* from E_d. Statistics e_d and Re_d then represents the mean square error and the relative mean square error for the domain d. Unfortunately, Re from table 5.2.4. gives the considerable relative error, ca. 50%, for both predictors.

6 Conclusion

Area-level gamma mixed model (2) was introduced and formulae for estimation of the true regression parameters (6) and for prediction of random effects (7) were derived. Subsequently, a corresponding numerical algorithm for obtaining these estimates and predictions was supplied. An algorithm for prediction EBP of the expectation μ_d denoting a characteristic of interest for the domain d was given.

The crucial points of this article are simulations experiments investigating the properties of EBP and Plug-in predictor. The results of the first simulation, tables 5.1.1. to 5.1.4., are satisfactory. In table 5.1.4. we can see that the relative error is about 5% for both predictors. The second simulation, however, points to a considerable relative error, ca. 50%, for both predictors, as can be seen from table 5.2.4.. That possibly makes both predictors useless. This issue must be resolved by further research.

References

[1] T. Hobza. *Model-based methods for small area estimation.* Habilitation Thesis (2017), 23–26.

[2] Ch. E. McCulloch, S. R. Searle. *Generalized, Linear, and Mixed Models.* Wiley Series in Probability and Statistics (2001), 135–142.

[3] M. Kulich. *Advanced Regression Models.* `https://www.karlin.mff.cuni.cz/~kulich/vyuka/pokreg/doc/advreg\_notes\_180220.pdf` (2018), 96–97.

[4] N. E. Breslow, D. G. Clayton. *Approximate Inference in Generalized Linear Mixed Models.* Journal of the American Statistical Association, **Vol. 88**, **No. 421** (1993), 9–25.

Density-Approximating Neural Network Models for Anomaly Detection

Martin Flusser

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

advisors:

Petr Somol, Cognitive Research at Cisco Systems

Vladimír Jarý, Department of Software Engineering
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. The anomaly detection is sub field of artificial intelligence the aim of which is identifying data that are somehow different from an expected pattern. Anomaly detection is also known as one-class classification because it is a similar task to the classification with the only difference: The training set contains the only class. This makes the task difficult because the character of the anomalous data is unknown when the model is trained.

We propose an alternative use of neural models in anomaly detection. Traditionally, in anomaly detection context the common use of neural models is in form of auto-encoders. Through the use of auto-encoders the true anomality is proxied by reconstruction error. Auto-encoders often perform well but do not guarantee to perform as expected in all cases. A popular more direct way of modeling anomality distribution is through k-Nearest Neighbor models. Although kNN can perform better than auto-encoders in some cases, their applicability can be seriously impaired by their space and time complexity especially with high-dimensional large-scale data. The alternative we propose is to model the distribution imposed by kNN using neural networks. We show that such neural models are capable of achieving comparable accuracy to kNN while reducing computational complexity by orders of magnitude. The de-noising effect of a neural model with limited number of neurons and layers is shown to lead to accuracy improvements in some cases. We evaluate the proposed idea against standard kNN and auto-encoders on a large set of benchmark data and show that in majority of cases it is possible to improve on accuracy or computational cost.

Keywords: Anomaly detection, neural network, auto-encoder, nearest neighbor

Abstrakt. Detekce anomálií je podoborem umělé inteligence a zabývá se nalezením anomálních prvků. Jako anomální se dají považovat data (pozorování), která jsou rozdílná, buď od vzorových dat, nebo od očekávaného vzoru. Tato úloha se někdy nazývá jako jednotřídní klasifikace, protože pro trénování modelu jsou k dispozici pouze data z jedné konkrétní třídy. Avšak detekce anomálií je mnohem složitější a obtížnější úkol než klasifikace, protože při detekci anomálii není předem znám charakter anomálních dat a je nutné rozhodovat, jak velké výchylky musí data dosáhnout, aby byla detekována jako anomální.

V článku představujeme alternativní použití neuronových sítí pro detekci anomálií. Neuronové sítě se v kontextu detekce anomálií používají zejména ve formě auto-encoderu, kde se míra anomálie získá jako rekonstrukční chyba. Auto-encodery často fungují dobře, ale v některých případech mohou selhat, nebo se chovat jinak než je očekáváno. Mnohem přímější způsob,

jak modelovat anomálii, je s použitím metod nejbližších sousedů. Přestože kNN nabízí v detekci anomálií často lepší výsledky, jeho použitelnost je limitována časovou a prostorovou složitostí, zejména pak pro velká a vysokodimenzionální data. Zde publikovaná alternativa je modelovat kNN distribuci anomálií pomocí neuronové sítě. Ukazujeme, že taková neuronová sít dosahuje srovnatelné přesnosti s kNN, ale s řádově lepší složitostí. Díky zahlazení neuronové sítě dochází k odšumování. Díky tomu je pak v mnoha případech přesnější než kNN. Přesnost navrhované metody je experimentálně porovnána se standartním kNN a také auto-encoderem za použití nejvyspělejší testovací metodologie a testovacích dat. Ukazuje se, že ve většině případů dokáže navrhovaná metoda zlepšit přesnost nebo složitost.

Klíčová slova: Detekce anomálií, auto-encoder, neuronové sítě, metoda nejbližšího souseda

Full paper: M. Flusser, T. Pevný, and P. Somol. *Density-approximating neural network models for anomaly detection.* ACM SIGKDD workshop on outlier detection deconstructed (8 2018). London, United Kingdom. `https://www.andrew.cmu.edu/user/lakoglu/odd/accepted_papers/ODD_v50_paper_19.pdf` or: `goo.gl/73yvmG`.

Analýza krevních vzorků pacientů podstupujících transplantaci krvetvorných buněk ohrožených sinusoidálním obstrukčním syndromem*

Kateřina Henclová†

3. ročník PGS, email: katerina.henclova@fjfi.cvut.cz
Katedra matematiky
Fakulta jaderná a fyzikálně inženýrská, ČVUT v Praze

školitel: Václav Šmídl, Oddělení adaptivních systémů
Ústav teorie informace a automatizace, AV ČR

Abstract. Sinusoidal obstruction syndrome (hepatic veno-occlusive disease) is a relatively frequent, serious and often fatal complication of hematopoietic stem cell transplantation. It is caused directly by high-dose chemotherapy, yet little is known about its biochemistry. Hence, in order to have more precise diagnostic criteria, early syndrome diagnosis or prediction and its effective treatment, the ultimate long-term goal of the corresponding medical research is to discover the syndrome's biomarkers. Study [1] is first to explore possibilities of mathematical approach and analysis of the patients' data collected by the Institute of Hematology and Blood Transfusion.

The author is provided with a series of 4 blood samples from each of 21 patients. Each sample is analyzed with gel electrophoresis and is to be handled in the form of a 581-dimensional, noisy vector. All patients, selected to be otherwise medically comparable, are labeled "positive" or "negative" based on whether they experienced the syndrome or not. However, no further information about their condition is given. The goal is to find such 1-5 dimensions (features) that are related just and only to the syndrome, i.e. to find features representing its biomarkers. To do so, the data is preprocessed in a fitting manner, analyzed using LASSO and further manually combined in order to produce a satisfactory result: 3 features to be forwarded for detailed chemical analysis.

Keywords: data analysis, big data in medicine, sinusoidal obstruction syndrome

Abstrakt. Sinusoidální obstrukční syndrom (veno-okluzivní choroba jater) je relativně častou, vážnou a mnohdy fatální komplikací vyskytující se u pacientů po transplantaci krvetvorných buněk. Je přímo zapříčiněn vysokou toxicitou chemoterapie, ale přesné probíhající biochemické procesy nejsou známé. Objev biomarkerů je nutný pro stanovení přesných diagnostických kritérií, včasné diagnózy nebo i předvídání rozvinutí syndromu a také pro umožnění jeho efektivní léčby. Toto je dlouhodobým cílem příslušného medicínského výzkumu. Studie [1] je první, která zkoumá možnosti matematického přístupu a analýzy dat pacientů Ústavu hematologie a krevní transfuze.

*Tato práce byla podpořena Fondem Neuron pro podporu vědy

†Poděkování patří doc. RNDr. Janu Vybíralovi, Ph.D. a doc. Ing. Václavu Šmídlovi, Ph.D. za konzultace a Ústavu hematologie a krevní transfuze za poskytnutá data.

Od každého ze 21 pacientů má autorka k dispozici sadu 4 krevních vzorků. Každý byl zpracován metodou gelové elektroforézy, a tedy je matematicky uchopitelný jakožto zašuměný 581-dimenzionální vektor. Pacienti, navybíraní tak, aby v jiných ohledech byl jejich stav srovnatelný, jsou označeni jako "pozitivní" nebo "negativní" podle toho, zda syndromem trpěli. Žádná další informace o jejich zdravotním stavu však není k dispozici. Cílem je najít 1-5 dimenzí (příznaků), které souvisí jen a pouze s výskytem syndromu, a tedy reprezentují hledané biomarkery. Za tím účelem jsou data vhodně předpřipravena, analyzována pomocí metody LASSO a nakonec zpracována i manuálně. Výsledkem je nalezení 3 slibných příznaků, které budou dále podrobně chemicky zkoumány.

Klíčová slova: analýza dat, big data v medicíně, sinusoidální obstrukční syndrom

Literatura

[1] K. Henclová. *Little Data Analysis of Bone Marrow Transplant Patients.* Submitted to Proceedings of Stochastic and Physical Monitoring Systems 2018, (2018).

Diffusive and Kohonen Learning Strategy: Stock Market*

Radek Hřebík

3rd year of PGS, email: Radek.Hrebik@seznam.cz
Department of Software Engineering
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisors:

Josef Jablonský, Department of Econometrics
Faculty of Informatics and Statistics, University of Economics, Prague

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

Abstract. Traditional self organised map (SOM) is learned by Kohonen paradigm. Novel model of self organisation is based on diffusion modelling in continuous space which is a good approximation of endorphins propagation in real brain. Therefore the structure of system is described by neuron coordinates instead of neighborhood relationship in traditional SOM. Neuron activation using diffusion process is discussed. Novel diffusive learning algorithm is based on this activation mentioned above. Paper is focused on market analysis via Kohonen and diffusion learning. Kohonen SOM is used as referential method. Using logarithmic differences of stock market data is preferred as data preprocessing.

Keywords: self organization, Kohonen map, diffusion learning, SOM, stock market

Abstrakt. Tradiční úloha samoorganizace (SOM) je založena na Kohonenově učení. Nový model samoorganizace je založen na difuzním modelování v kontinuálním prostoru, což představuje aproximaci šíření endorfinů v mozku. Struktura systému je popsána neuronovými souřadnicemi namísto sousedství v tradičním SOM. Příspěvek popisuje aktivaci neuronu za použití difúzního procesu. Nový algoritmus difuzního učení je založen na této aktivaci. Použití algoritmu je zaměřeno na analýzu trhu prostřednictvím srovnání Kohonenova a difuzního učení. Kohonenovo učení slouží jako referenční metoda. Data burzovních indexů jsou předzpracována s využitím logaritmických diferencí.

Klíčová slova: samoorganizace, Kohonenova mapa, difuzní učení, SOM, akciový trh

1 Introduction

There are many approaches how to perform modelling of self organisation. They can be directly inspired by anatomy and physiology of neuronal system or rather by other ideas which are easy to realize. Our research is inspired by pudding model of atom in physics [11, 9], where the nucleus of atoms are supposed as points (raisins) in the electron continuum (pudding). In the case of self organisation we will place individual neurons instead of atom nucleus into the continuum which would transfer the information in the

*This work has been supported by the grant SGS 17/196/OHK4/3T/14

system. The second inspiration is strongly connected with brain physiology study about slow signal propagation in central nervous system.
The SOM can be directly applied to the analysis of time series of stock prices. The logarithmic differences i.e. daily yields are evaluated for given stock first. Then we can define the state of the stock as its history in previous days. These states are real vectors of x-dimensions which are in mutual relationships and the relationship structural is useful for the study of stock market evaluation. This approach will be applied both to individual stocks and to their system.

2 Pudding Model of SOM

Our model of self-organised map is based on specific assumptions:

- Finite number of neurons is placed in fix positions like raisins in a pudding.
- The neurons are surrounded by unconstrained continuum as analogy of the pudding base.
- Neuron interconnections are omitted.
- The pattern set stays outside the pudding and only sequentially activates individual neurons.
- Neuron activities generate concentration profile of substrate in the pudding.
- Substrate concentration influences the learning rates of individual neurons.
- The learning process changes weights as neuron properties.

The *pudding model* description begins with remembering of basic facts. Let $m, n, H \in \mathbb{N}$ be number of patterns, pattern dimensionality and number of SOM neurons [3]. The individual patterns are $\mathbf{x}_j \in \mathbb{R}^n$, where $j = 1, ..., m$ and form the pattern set $\mathcal{S} = \{\mathbf{x}_1, ..., \mathbf{x}_m\}$. The fixed positions of individual neurons in continuum are $\mathbf{p}_i \in \mathbb{R}^N$ for $i = 1, ..., H$ and reflects the topology of SOM [8] which is subject of network design. The diffusion process in continuum can be easily expressed using matrix $\mathbf{D} \in (\mathbb{R}_0^+)^{H \times H}$ of mutual distances $d_{i,j} = \|\mathbf{p}_i - \mathbf{p}_j\|_2$. Therefore the resulting SOM is invariant to translation and rotation of its structure. Let $\Delta t > 0$ be learning period and the diffusion in continuum will be studied only in discrete time $t_k = k \cdot \Delta t$, where $k \in \mathbb{N}_0$. The result of SOM learning is the system of weights [10] $\mathbf{w}_i \in \mathbb{R}^n$, where $i = 1, ..., H$ of course. We begin with random weights setting $\mathbf{w}_i(0)$. The weights evolve during learning process and their values in time t_q are denoted as $\mathbf{w}_i(q)$, where $q \in \mathbb{N}_0$. The *pudding model* is based on substrate concentrations in neurons and given time. Being prepared to SOM learning we have to study the concentration profile first.

2.1 Single Activation

The pudding SOM learning is based on the activation of single neuron. We will study j-th neuron which is supposed to be active in time t_k. Therefore, formally $j = \varphi_k$. The concentration profile in $\mathbb{R}^N$ is depicted on the left part of figure 1 for $N = 2$. But it is not necessary to study the substrate concentration in any point. The learning is based only on the concentration in neuron points. The concentration in time t_q is

$$\mathrm{c}(\mathbf{y}, \mathbf{p}_j, t_q) = \frac{1}{(4\pi D(t_q - t_k))^{N/2}} \cdot \exp\left(-\frac{\|\mathbf{y} - \mathbf{p}_j\|_2^2}{4D(t_q - t_k)}\right) \cdot \exp\left(-\lambda(t_q - t_k)\right) \qquad (1)$$

for $q > k$. The formula can be simplified to

$$\mathrm{c}(\mathbf{p}_i, \mathbf{p}_j, t_q) = \frac{1}{(4\pi D(q-k)\Delta t)^{N/2}} \cdot \exp\left(-\frac{d_{i,j}^2}{4D(q-k)\Delta t}\right) \cdot \exp\left(-\lambda(q-k)\Delta t\right). \qquad (2)$$

After the substitution $a = 4D\Delta t > 0$, $b = \lambda\Delta t > 0$ we obtain resulting activation formula

$$\mathrm{c}(\mathbf{p}_i, \mathbf{p}_j, t_q) = (\pi a(q-k))^{-N/2} \cdot \exp\left(-\frac{d_{i,j}^2}{a(q-k)} - b(q-k)\right). \qquad (3)$$

When $\min(d_{i,j} \geq 1)$, then we suggest to use $a = 1, b = 1/10$ for the first experiments as will be demonstrated in next sections.

2.2 Complete Activation

The SOM learning is based on the substrate concentrations in q-th step in time t_q. This concentration represents the result of previous activation sequence $\varphi_1, \varphi_2, ...\varphi_{q-1}$ using single activation model (3). Due to linearity we can use the additivity principle and directly calculate the cumulative concentration in i-th neuron and step q

$$c_{i,q} = \sum_{k=1}^{q-1} c(\mathbf{p}_i, \mathbf{p}_{\varphi_k}, t_q - t_k) = \frac{1}{(\pi a)^{N/2}} \cdot \sum_{k=1}^{q-1} \frac{\exp\left(-\frac{d_{i,\varphi_k}^2}{a(q-k)} - b(q-k)\right)}{(q-k)^{N/2}}. \qquad (4)$$

Resulting formula consists of all concentration information which are necessary for the SOM learning. Therefore the concentration $c_{i,q}$ is only a function of activation history, SOM topology and parameters a, b. But the history is result of learning which will be studied in next section.
The full concentration profile in $2D$ pudding after 99 random activation steps ($q = 100$) is depicted on the right part of figure 1.

2.3 Diffusive Learning of SOM

Novel learning algorithm is completely devoted to Kohonen learning rules [7] as follows. The weight of i-th neuron is changed in q-th step by rule

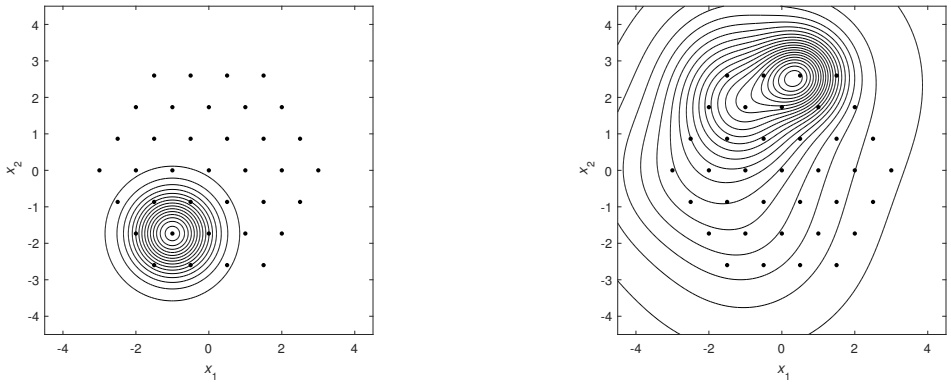


Figure 1: Concentration profile after single (left) and complete (right) activation ($N = 2, a = 1, b = 1/10, H = 37, q = 100$)

$$\mathbf{w}_i(q) = \mathbf{w}_i(q-1) + \alpha(q) \cdot c_{i,q} \cdot (\mathbf{x}_q - \mathbf{w}_i(q-1)) \tag{5}$$

for $i = 1, ..., H$, $\mathbf{x}_q \sim \mathrm{U}(\mathcal{S})$ is uniformly selected pattern from $\mathcal{S}$, $c_{i,q}$ is substrate concentration according to (4) and $\alpha(q) > 0$ is ageing function which is supposed to be non-increasing. The winner is also selected according to Kohonen rule [7] as

$$\varphi_q \in \arg\min_{k=1,...,H} \|\mathbf{x}_q - \mathbf{w}_k\|_2. \tag{6}$$

The main difference between the traditional SOM learning [1] and our approach [6] is in the application of diffusive equation which generates the concentration profile (4). The learning feedback is driven by winner index φ_q from (6) which is used in next step of concentration calculations (4).
As in traditional SOM learning we have to initialize the weights [1] and use appropriate ageing strategy. We recommend generate the initial weights from the multivariate Gaussian distribution as

$$\mathbf{w}_i(0) \sim \mathrm{N}(\mathrm{E}\mathbf{X}, \mathrm{var}\mathbf{X}/100) \tag{7}$$

for $i = 1, ..., H$. The ageing function $\alpha(q)$ can be constant in the first experiments, but satisfying $\alpha(q) \cdot c_{i,q} \leq 1$ to avoid learning instability.

3 Stock Market States

In our previous studies [4, 5] we investigated the states of country economics using annual change of macroeconomical indicators and traditional principal component analysis. In this study we focused on the states of individual stocks [2] and their changes which are easy to analyse and visualise using Pudding SOM algorithm. The analysis of single stock is based on price time series

$$\{a_k\}_{k=0}^{D-1} \tag{8}$$

in period of D days where $a_k \geq 0$ is the stock price in k-th day. First, we have to calculate daily yield as

$$z_k = \Delta \ln a_k = \ln(a_{k+1}/a_k) \tag{9}$$

The hypothesis behind the study is that the state reconstruction brings more information than the individual daily price or yield. The stock state is defined using slighting window of length $w \in N$ as vector

$$s_k = (z_k, ..., z_{k+w-1}) \in \mathbb{R}^w \tag{10}$$

To avoid the effect of dimensionality curse we will study the states for $w <= 30$ respecting the habit of 5 market days per week we will study the market states for $w = 5, 10, 15, 20, 25, 30$ in fixed hexagonal topology of 19 nodes SOM.

3.1 Case Study: Stock Market in Period 2000 - 2018

As input data we use daily prices of 10 main indices index SP500 (USA), index Dow Jones 30 (USA), index Nasdaq (USA), index Russell 2000 (USA), index DAX (Germany), index CAC40 (France), index BEL20 (Belgium), index EURONEXT100 (Europe), index NIKKEI (Japan) and Hang Seng index (Hong Kong). We have used data from March 2000 to March 2018. Therefore the SOM consists of 43070 states from 4312 days. The diffusive learning and traditional Kohonen approach have been compared for $w \in \{5, 10, 15, 20, 25, 30\}$ which corresponds to working week period. Selected results are demonstrated in figure 2 as number of states in SOM nodes. The results for $w >= 10$ are very similar and we decided to analyse individual stocks only for diffusive learning and $w = 5$. As seen the highest concentration are in the central nodes which represent the average behaviour. Meanwhile the extreme case in perimeter nodes are not so frequent in all cases.
Using the result of learning we can trace the states of selected stock in selected years. The situation in the year 2008 i.e. the beginning of crisis is demonstrated in figure 3. The anomalous market behaviour caused higher frequency of extreme states in perimeter nodes as opposite to normal behaviour. The example of standard behaviour is captured in figure 4.

4 Conclusion

The new method of diffusive learning has been applied to stock market data and its results are similar to Kohonen SOM. The state reconstruction has been performed using one week history of individual stocks. There is significant difference between stocks states during crisis period and normal stock behaviour. The crisis can be characterized by high frequency of extreme states at the perimeter of SOM. So that the last year can be characterized by normal stock behaviour.

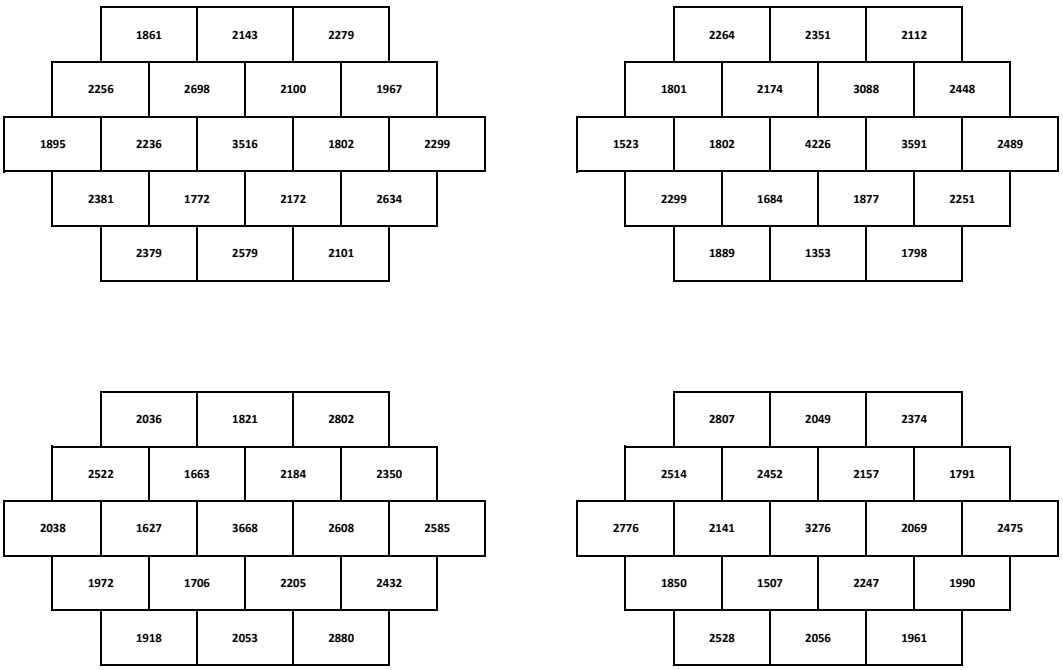


Figure 2: Resulting SOMs in case of diffusive learning with $w = 5$ (top left), $w = 10$ (top right) and Kohonen learning with $w = 5$ (bottom left), $w = 10$ (bottom right)

References

[1] E. Alonso. *Computational Neuroscience for Advancing Artificial Intelligence: Models, Methods and Applications: Models, Methods and Applications*. Premier reference source. Medical Information Science Reference, 2010.

[2] M. Dlask and J. Kukal. Correlation dimension as a measure of stock market variability. In *Mathematical Methods in Economics*, pages 119–124, Hradec Kralove, 2017. University of Hradec Kralove.

[3] D. Graupe. *Deep Learning Neural Networks: Design and Case Studies*. 2016.

[4] R. Hrebik and J.• Kukal. Multivarietal data whitening of main trends in economic development. In *Mathematical Methods in Economics*, pages 279–284, Plzeň, 2015. University of West Bohemia.

[5] Radek Hrebik and Jaromir Kukal. The economics and data whitening: Data visualisation. In *Federated Conference on Software Development and Object Technologies*, pages 91–101. Springer, 2015.

[6] Radek Hrebik and Jaromir Kukal. Self-organization via diffusion modelling. *IEEE Transactions on Neural Networks and Learning Systems*, page submitted, 2018.

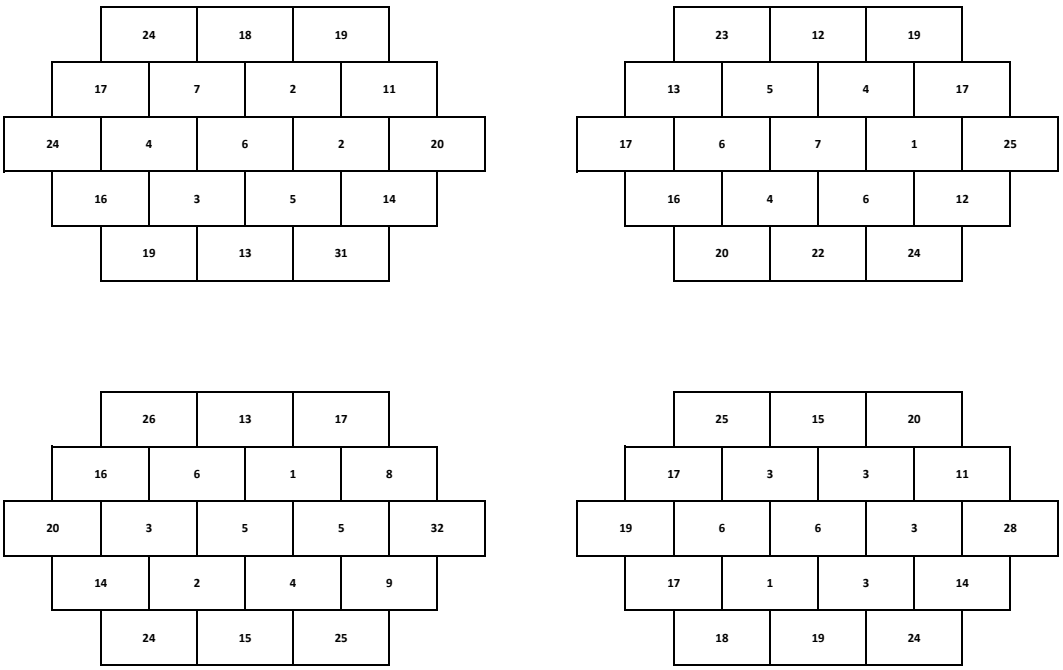


Figure 3: Resulting SOMs for SP 500 (left top), DAX (right top), NIKKEI (left bottom) and Hang Seng (right bottom) for diffusive learning with $w = 5$ in 2008

[7] T. Kohonen. *Self-Organizing Maps*. Springer Series in Information Sciences. Springer Berlin Heidelberg, 2012.

[8] E. Oja and S. Kaski. *Kohonen Maps*. Elsevier Science, 1999.

[9] Andrew Philippides, Phil Husbands, and Michael O'Shea. Four-dimensional neuronal signaling by nitric oxide: a computational analysis. *Journal of Neuroscience*, 20(3):1199–1207, 2000.

[10] A. Rettberg, M.C. Zanella, M. Amann, M. Keckeisen, and F.J. Rammig. *Analysis, Architectures and Modelling of Embedded Systems: Third IFIP TC 10 International Embedded Systems Symposium, IESS 2009, Langenargen, Germany, September 14-16, 2009, Proceedings*. IFIP Advances in Information and Communication Technology. Springer Berlin Heidelberg, 2009.

[11] J. J. Thomson. On the structure of the atom: an investigation of the stability and periods of oscillation of a number of corpuscles arranged at equal intervals around the circumference of a circle; with application of the results to the theory of atomic structure. *Philosophical Magazine Series 6*, 7(39):237–265, 1904.

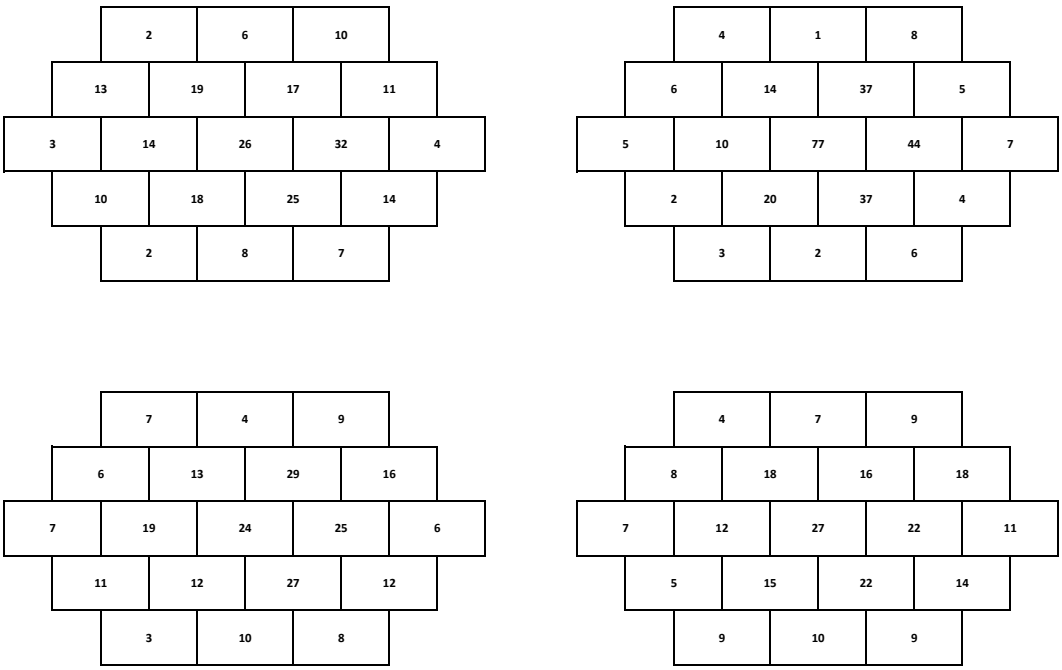


Figure 4: Resulting SOMs for SP 500 (left top), DAX (right top), NIKKEI (left bottom) and Hang Seng (right bottom) for diffusive learning with $w = 5$ in 2017

Mathematical Model of Signal Propagation in Excitable Media*

Jakub Kantner

1st year of PGS, email: jakub.kantner@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. This contribution deals with a model of signal propagation in excitable media based on a system of reaction-diffusion equations. Such media have the ability to exhibit a large response in reaction to a small deviation from the rest state. An example of such media is the nerve tissue or the heart tissue. The first part of the contribution briefly describes the origin and the propagation of the cardiac action potential in the heart. In the second part, the mathematical properties of the model are discussed. Next, the contribution studies a numerical solution of the problem based on the finite-difference method. Finally, a numerical study is performed in both homogeneous and heterogeneous medium with focus on interactions of propagating signals with obstacles in the medium.

Keywords: excitable media, reaction-diffusion equations, FitzHugh-Nagumo model, cardiac action potential, signal propagation, invariant regions

Abstrakt. Příspěvek se věnuje modelu šíření signálu v excitovatelném prostředí založeném na soustavě reakčně-difuzních rovnic. Taková prostředí mají vlastnost reagovat velkou odezvou na malou výchylku za stabilní polohy. Jejich příkladem je nervová nebo srdeční tkáň. První část příspěvku krátce popisuje vznik a šíření srdečního akčního potenciálu v srdeční tkáni. Ve druhé části jsou diskutovány matematické vlastnosti modelu. Příspěvek se dále zabývá numerickým řešením problému pomocí metody konečných diferencí. V poslední části jsou provedeny výpočetní studie analyzující chování modelu v homogenním i heterogenním prostředí se zvláštním zřetelem na interakci šířících se signálů s překážkami v prostředí.

Klíčová slova: excitovatelné prostředí, reakčně-difuzní rovnice, FitzHugh-Nagumův model, srdeční akční potenciál, šíření signálu, invariantni regiony

Full paper: J. Kantner, M. Beneš. *Mathematical Model of Signal Propagation in Excitable Media.* Full text is in preparation for the submission to the journal Discrete & Continuous Dynamical Systems – Series S, 2018.

*This work has been supported by the grant No. SGS17/194/OHK4/3T/14 of the Grant Agency of the Czech Technical University in Prague and by the project No. 15-27178A "Quantitative Mapping of Myocard and of Flow Dynamics by Means of MR Imaging for Patients with Nonischemic Cardiomyopathy Development of Methodology" of the Ministry of Health of the Czech Republic the project No. 15-27178A of the Ministry of Health of the Czech Republic.

Cramér-Rao Induced Bound for Complex Independent Component Extraction*

Václav Kautský

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

advisors:

Zbyněk Koldovský, Institute of Information Technology and Electronics
Faculty of Mechatronics, Informatics, and Interdisciplinary Studies, TUL

Petr Tichavský, Department of Stochastic Informatics
Institute of Information Theory and Automation, CAS

Abstract. Cramér-Rao Lower Bound (CRLB) for the complex-valued blind source extraction problem based on the assumption the target signal is independent of the other signals is derived. A recent approach called Independent Component Extraction is used where, compared to Independent Component Analysis (ICA), the mixing model contains minimum number of parameters needed for the extraction problem. The target signal is assumed to be non-Gaussian or noncircular Gaussian while the other signals, which are not separated from each other, are modeled as a circular Gaussian mixture. A CRLB-induced Bound (CRIB) for Interference-to-Signal Ratio (ISR) is derived and compared with similar bound for ICA. Numerical simulations with selected BSE algorithms show the correspondence between empirical results and the theory.

Keywords: Blind Source Extraction, Cramér-Rao Lower Bound, Independent Component Analysis, Independent Component Extraction

Abstrakt. Tento článek se zabývá odvozením Crámerovy-Raovy dolní meze pro slepou separaci komplexních signálů, která je založena na statistické nezávislosti cílového signálu a ostatních signálů. Využívá se nového přístupu, tzv. ICE, který vychází z Analýzy nezávislých komponent (ICA), avšak využívá minimální počet parametrů potřebných pro extrakci. Cílový signál je požadován negaussovský nebo necirkulární gaussovský. Ostatní signály, které nejsou předmětem separace, jsou modelovány jako cirkulární gaussovská směs. Dále je odvozena Crámerova-Raova dolní mez pro Interference-to-Signal Ratio (ISR). Tato mez je následně porovnána se stejnou mezí pro ICA. Numerické simulace využívající vybrané algoritmy pro slepou separaci potvrzují dobrou shodu mezi empirickými výsledky a teorií.

Klíčová slova: Analýza nezávislých komponent, Crámerova-Raova dolní mez, ICE, Slepá separace signálu

*This work was supported by The Czech Science Foundation through Project No. 17-00902S, by the United States Department of the Navy, Office of Naval Research Global, through Project No. N62909-18-1-2040 and by the Grant SGS18/188/OHK4/3T/14 provided by the Ministry of Education, Youth, and Sports of the Czech Republic (MŠMT ČR).

1 Introduction

In Independent Component Analysis (ICA), the instantaneous linear mixing model

$$\mathbf{x} = \mathbf{A}\mathbf{u} \tag{1}$$

is studied, where $\mathbf{x}$ is a $d \times 1$ vector of d mixed signals, $\mathbf{A}$ is a $d \times d$ non-singular mixing matrix, and $\mathbf{u}$ is a $d \times 1$ vector of source signals that are assumed to be *mutually independent* [1]. The jth signal u_j (the jth element of $\mathbf{u}$) is modeled as a random variable with the probability density function (pdf) $p_j(\cdot)$. The goal is to estimate $\mathbf{A}^{-1}$ from $\mathbf{x}$ through finding a square de-mixing matrix $\mathbf{W}$ such that $\mathbf{y} = \mathbf{W}\mathbf{x}$ are as independent as possible. In this paper, complex-valued signals and parameters will be considered.

It was shown that $\mathbf{A}^{-1}$ can be identified up to the order and scales of its rows if it holds that at most one source signal has the complex Gaussian pdf or that no two complex Gaussian source signals have the same circularity coefficient [2]. Then, $\mathbf{W}$ can be estimated as such that $\mathbf{G} = \mathbf{W}\mathbf{A} \approx \mathbf{P}\mathbf{\Lambda}$, where $\mathbf{P}$ and $\mathbf{\Lambda}$ is a permutation and diagonal matrix, respectively.

$\mathbf{G}$ reflects the separation accuracy as its ijth element, G_{ij}, determines the presence of u_j in the ith separated signal y_i. The Cramér-Rao Lower Bound (CRLB) provides an algorithm-independent lower bound on the variance of the estimation. For the complex-valued ICA problem, it was derived using the CRLB that

$$\mathrm{E}[G_{ij}^2] \geq \frac{1}{N} \frac{\kappa_j}{\kappa_i \kappa_j - 1}, \qquad i \neq j, \tag{2}$$

where N is the number of i.i.d. samples; $\kappa_i = \mathrm{E}[|\psi_i|^2]$ where $\psi_i(x) = -\partial/\partial x \log p_i(x)$ is the score function related to p_i. For normalized variables it holds that $\kappa_i \geq 1$, and $\kappa_i = 1$ if and only if the ith pdf is circular Gaussian [6].

This paper addresses the Blind Source Extraction (BSE) problem where the goal is to extract only one signal of interest (SOI) from the other signals in $\mathbf{x}$ (which will be called "background"). This has been solved, in the context of ICA, through extracting a signal by minimizing its entropy [1]. The information theory-based approach, however, does not allow us to directly compute the CRLB for the extraction problem. Recently, we have revised the BSE in approach called Independent Component Extraction (ICE) [5]. Here, the mixing model (1) is re-parameterized so that it contains a minimum number of parameters that are necessary for the extraction of the SOI. The statistical model is based on the assumption that the SOI is independent from the background, similarly to ICA, and the background is assumed to be Gaussian. Then, the computation of the CRLB is straightforward using the likelihood function.

This way, we have computed the bound for the real-valued case in [4]. It was shown that the bound coincides with that for ICA if the background is Gaussian and is also in a good agreement with asymptotic performance analyses of several BSE algorithms [10, 3]. In this paper, we generalize this result for the complex-valued case where the SOI is assumed to be non-Gaussian or non-circular Gaussian, while the background is modeled as circular Gaussian.

The ICE problem is described in Section II. Section III is devoted to the computation of the Fisher Information Matrix, which is used in Section IV to derive the Cramér-

Rao Induced Bound (CRIB). Section V is devoted to simulations and comparisons, and Section VI concludes the article.

Technicalities: Plain letters denote scalars, bold letters denote vectors, and bold capital letters denote matrices. Upper index $\cdot^T$, $\cdot^H$, or $\cdot^*$ denotes, respectively, transposition, conjugate transpose, or complex conjugate. The Matlab convention for matrix/vector concatenation and indexing will be used, e.g., $[1;\, \mathbf{g}] = [1,\, \mathbf{g}^T]^T$, and $(\mathbf{A})_{j,:}$ is the jth row of $\mathbf{A}$. A complex random vector $\mathbf{x}$ is called circular if its pseudo-covariance is $\mathtt{pcov}(\mathbf{x}) = \mathrm{E}\big[(\mathbf{x}-\mathrm{E}[\mathbf{x}])\,(\mathbf{x}-\mathrm{E}[\mathbf{x}])^T\big] = \mathbf{0}$, otherwise, $\mathbf{x}$ is non-circular; $\mathrm{E}[\cdot]$ stands for the expectation operator. The second-order circularity coefficient γ of a complex-valued random variable x with zero mean, see [2], is defined as in [8] $\gamma = \big|\mathrm{E}\big[x^2\big]\big| \,/ \mathrm{E}\big[|x|^2\big]$. Thus, $\gamma \in [0,1]$ and $\gamma = 0$ holds for circular random variable.

2 Problem Statement

2.1 Algebraic Model

Without any loss of generality, let the SOI be $s = u_1$. Let the mixing matrix $\mathbf{A}$ be partitioned as $\mathbf{A} = [\mathbf{a},\, \mathbf{A}_2]$, and let $\mathbf{x}$ be written as $\mathbf{x} = \mathbf{A}\mathbf{u} = \mathbf{a}s + \mathbf{y}$, where $\mathbf{y} = \mathbf{A}_2\mathbf{u}_2$ and $\mathbf{u}_2 = [u_2, \ldots, u_d]^T$. The identification of $\mathbf{A}_2$ is not needed for the extraction of s, so it can be replaced by an arbitrary $\mathbf{Q}$ whose columns span the same subspace as those of $\mathbf{A}_2$; the new mixing matrix is $\mathbf{A}_{\mathrm{ICE}} = [\mathbf{a},\, \mathbf{Q}]$. To remove the uncertainty of $\mathbf{Q}$, we focus on the inverse matrix $\mathbf{A}_{\mathrm{ICE}}^{-1} = \mathbf{W}_{\mathrm{ICE}}$.

Let $\mathbf{a}$ and $\mathbf{W}_{\mathrm{ICE}}$ be partitioned, respectively, as $\mathbf{a} = [\gamma;\, \mathbf{g}]$ and $\mathbf{W}_{\mathrm{ICE}} = [\mathbf{w}^H;\, \mathbf{B}]$. To separate s from the other signals, $\mathbf{B}$ must be orthogonal to $\mathbf{a}$. A straightforward selection is $\mathbf{B} = [\mathbf{g} \quad -\gamma\mathbf{I}_{d-1}]$ where $\mathbf{I}_d$ denotes the $d \times d$ identity matrix. Let $\mathbf{w} = [\beta;\, \mathbf{h}]$. Then from $\mathbf{A}_{\mathrm{ICE}}^{-1} = \mathbf{W}_{\mathrm{ICE}}$ it follows that

$$\mathbf{W}_{\mathrm{ICE}} = \begin{pmatrix} \mathbf{w}^H \\ \mathbf{B} \end{pmatrix} = \begin{pmatrix} \beta^* & \mathbf{h}^H \\ \mathbf{g} & -\gamma\mathbf{I}_{d-1} \end{pmatrix}, \tag{3}$$

and

$$\mathbf{A}_{\mathrm{ICE}} = [\mathbf{a}, \quad \mathbf{Q}] = \begin{pmatrix} \gamma & \mathbf{h}^H \\ \mathbf{g} & \frac{1}{\gamma}\left(\mathbf{g}\mathbf{h}^H - \mathbf{I}_{d-1}\right) \end{pmatrix}, \tag{4}$$

where β and γ are linked through the condition $\beta\gamma = 1 - \mathbf{h}^H\mathbf{g}$.

Now, the ICE mixing model can be written as

$$\mathbf{x} = \mathbf{A}_{\mathrm{ICE}}\mathbf{v}, \tag{5}$$

where $\mathbf{v} = [s; \mathbf{z}]$, and $\mathbf{z} = \mathbf{B}\mathbf{x}$. Since the scales of s and of $\mathbf{a}$ are ambiguous (s and $\mathbf{a}$ can be substituted, respectively, by αs and $\alpha^{-1}\mathbf{a}$ with any $\alpha \neq 0$), we can fix the uncertainty by putting $\gamma = 1$. Then, the only free parameters of the ICE model are $\mathbf{g}$ and $\mathbf{h}$.

2.2 Statistical Model

The fundamental assumption of ICA/ICE states that s and $\mathbf{z}$ are independent, which means that their joint pdf can be factorized as the product of marginal pdfs. Furthermore,

we will assume that s has a non-Gaussian pdf denoted as $p_s(s)$ while $\mathbf{z}$ has multivariate circular Gaussian pdf with covariance $\mathbf{C}_\mathbf{z}$. Hence, from (5), the pdf of $\mathbf{x}$ is

$$p_\mathbf{x}(\mathbf{x}) = p_s(\mathbf{w}^H\mathbf{x})p_\mathbf{z}(\mathbf{B}\mathbf{x})|\det(\mathbf{W}_{\text{ICE}})|^2, \tag{6}$$

where $\mathbf{W}_{\text{ICE}}$, $\mathbf{w}$, and $\mathbf{B}$ depend on $\mathbf{g}$ and $\mathbf{h}$ as described by (3), and $p_\mathbf{z}$ is the pdf of the circular Gaussian distribution with zero mean and covariance $\mathbf{C}_\mathbf{z}$, denoted by $\mathcal{CN}(\mathbf{0}, \mathbf{C}_\mathbf{z})$.

A straightforward calculus, not shown here to save space, can show that for $\gamma = 1$, $|\det(\mathbf{W}_{\text{ICE}})| = 1$; see [5]. The log-likelihood function for one signal sample is thus equal to

$$\mathcal{L}(\mathbf{x}|\boldsymbol{\theta}, \boldsymbol{\xi}) = \log p_s(\mathbf{w}^H\mathbf{x}) - \mathbf{x}^H\mathbf{B}^H\mathbf{C}_\mathbf{z}^{-1}\mathbf{B}\mathbf{x} - \log\left(|\mathbf{C}_\mathbf{z}|\right) - (d-1)\log(2\pi), \tag{7}$$

where $\boldsymbol{\theta} = [\mathbf{g}; \mathbf{h}; \mathbf{r}]$ denotes the parameter vector, in which $\mathbf{r}$ stands for the $d(d-1)/2 \times 1$ vector stacking the off-diagonal elements of $\mathbf{C}_z^{-1}$, and $\boldsymbol{\xi}$ is the $d \times 1$ vector stacking the real-valued diagonal elements of $\mathbf{C}_z^{-1}$; $|\mathbf{C}_\mathbf{z}|$ denotes the determinant of $\mathbf{C}_\mathbf{z}$; $\mathbf{r}$ and $\boldsymbol{\xi}$ are nuisance parameters.

3 Fisher Information Matrix

To compute the CRLB, we use the approach for the mixed case with real and complex-valued parameters described in [7]. Let $\tilde{\boldsymbol{\theta}} = (\boldsymbol{\theta}, \boldsymbol{\theta}^*, \boldsymbol{\xi})$. According to [7], for any unbiased estimator of $\tilde{\boldsymbol{\theta}}$, it holds that

$$\texttt{cov}\left(\tilde{\boldsymbol{\theta}}\right) \succeq \mathcal{J}^{-1}\left(\tilde{\boldsymbol{\theta}}\right) = \texttt{CRLB}\left(\tilde{\boldsymbol{\theta}}\right), \tag{8}$$

where $\mathcal{J}\left(\tilde{\boldsymbol{\theta}}\right)$ is the Fisher information matrix (FIM), and $\mathbf{C} \succeq \mathbf{D}$ means that $\mathbf{C} - \mathbf{D}$ is a positive semi-definite matrix. The FIM can be partitioned as

$$\mathcal{J}\left(\tilde{\boldsymbol{\theta}}\right) = \begin{pmatrix} \mathbf{F} & \mathbf{P} & \mathbf{R} \\ \mathbf{P}^* & \mathbf{F}^* & \mathbf{R}^* \\ \mathbf{R}^H & \mathbf{R}^T & \mathbf{F}_{\boldsymbol{\xi}} \end{pmatrix}, \tag{9}$$

where

$$\mathbf{F} = \mathrm{E}\left[\frac{\partial\mathcal{L}}{\partial\boldsymbol{\theta}^*}\left(\frac{\partial\mathcal{L}}{\partial\boldsymbol{\theta}^*}\right)^H\right], \qquad \mathbf{P} = \mathrm{E}\left[\frac{\partial\mathcal{L}}{\partial\boldsymbol{\theta}^*}\left(\frac{\partial\mathcal{L}}{\partial\boldsymbol{\theta}^*}\right)^T\right],$$
$$\mathbf{R} = \mathrm{E}\left[\frac{\partial\mathcal{L}}{\partial\boldsymbol{\theta}^*}\left(\frac{\partial\mathcal{L}}{\partial\boldsymbol{\xi}}\right)^T\right], \qquad \mathbf{F}_{\boldsymbol{\xi}} = \mathrm{E}\left[\frac{\partial\mathcal{L}}{\partial\boldsymbol{\xi}}\left(\frac{\partial\mathcal{L}}{\partial\boldsymbol{\xi}}\right)^T\right],$$

where the derivatives with respect to complex-valued parameters are defined according to the Wirtinger calculus.

The derivatives of the log-likelihood function (7) are as follows.

$$\frac{\partial\mathcal{L}(\mathbf{x}|\boldsymbol{\theta}, \boldsymbol{\xi})}{\partial\mathbf{g}^*} = \left(\frac{\partial\mathcal{L}(\mathbf{x}|\boldsymbol{\theta}, \boldsymbol{\xi})}{\partial\mathbf{g}}\right)^* = \psi^*(s)\mathbf{h}x_1^* - x_1^*\mathbf{C}_\mathbf{z}^{-1}\mathbf{z}, \tag{10}$$

$$\frac{\partial\mathcal{L}(\mathbf{x}|\boldsymbol{\theta}, \boldsymbol{\xi})}{\partial\mathbf{h}^*} = \left(\frac{\partial\mathcal{L}(\mathbf{x}|\boldsymbol{\theta}, \boldsymbol{\xi})}{\partial\mathbf{h}}\right)^* = \psi(s)\mathbf{z}, \tag{11}$$

$$\frac{\partial\mathcal{L}(\mathbf{x}|\boldsymbol{\theta}, \boldsymbol{\xi})}{\partial r_{i,j}^*} = \left(\frac{\partial\mathcal{L}(\mathbf{x}|\boldsymbol{\theta}, \boldsymbol{\xi})}{\partial r_{i,j}}\right)^* = -z_i z_j^* + (\mathbf{C}_\mathbf{z})_{i,j}, \tag{12}$$

where $r_{i,j}$ is the ijth element of of $\mathbf{C}_\mathbf{z}^{-1}$; $\psi(s) = -\frac{\partial \ln p_s(s,s^*)}{\partial s}$. Next, for the real-valued parameter it holds

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\xi}} = \left(\frac{\partial \mathcal{L}(\mathbf{x}|\boldsymbol{\theta},\boldsymbol{\xi})}{\partial r_{i,i}}\right) = -z_i z_i^* + (\mathbf{C}_\mathbf{z})_{i,i}\,. \tag{13}$$

Let $\mathbf{F}$ be partitioned as

$$\mathbf{F} = \begin{pmatrix} \mathbf{F}_{\mathbf{g},\mathbf{g}} & \mathbf{F}_{\mathbf{g},\mathbf{h}} & \mathbf{F}_{\mathbf{g},\mathbf{r}} \\ \mathbf{F}_{\mathbf{h},\mathbf{g}} & \mathbf{F}_{\mathbf{h},\mathbf{h}} & \mathbf{F}_{\mathbf{h},\mathbf{r}} \\ \mathbf{F}_{\mathbf{r},\mathbf{g}} & \mathbf{F}_{\mathbf{r},\mathbf{h}} & \mathbf{F}_{\mathbf{r},\mathbf{r}} \end{pmatrix}. \tag{14}$$

Using (10)-(13), the blocks of $\mathbf{F}$ are as follows.

$$\mathbf{F}_{\mathbf{g},\mathbf{g}} = \eta \mathbf{h}\mathbf{h}^H + \kappa(\mathbf{h}^H\mathbf{C}_\mathbf{z}\mathbf{h})\mathbf{h}\mathbf{h}^H - \mathbf{h}\mathbf{h}^H + \sigma_s^2\mathbf{C}_\mathbf{z}^{-1} + (\mathbf{h}^H\mathbf{C}_\mathbf{z}\mathbf{h})\mathbf{C}_\mathbf{z}^{-1}, \tag{15}$$

where

$$\kappa = \mathrm{E}[|\psi(s)|^2], \tag{16}$$
$$\eta = \mathrm{E}[|\psi(s)|^2|s|^2], \tag{17}$$
$$\sigma_s^2 = \mathrm{E}[|s|^2], \tag{18}$$

and where the following identities have been used.

$$\mathrm{E}[\psi(s)s] = \mathrm{E}[\psi^*(s)s^*] = 1, \tag{19}$$
$$y_1 = \mathbf{h}^H\mathbf{z}, \tag{20}$$
$$\mathrm{E}\left[y_1^2\mathbf{y}\mathbf{y}^H\right] = \mathrm{E}\left[y_1^2\right]\mathrm{E}\left[\mathbf{y}\mathbf{y}^H\right] + \mathrm{E}\left[y_1^*\mathbf{y}\right]\mathrm{E}\left[y_1\mathbf{y}^H\right] + \mathrm{E}\left[y_1\mathbf{y}\right]\mathrm{E}\left[y_1^*\mathbf{y}^H\right], \tag{21}$$
$$\mathbf{P}_z = \mathrm{E}[\mathbf{z}\mathbf{z}^T] = \mathbf{O}, \tag{22}$$

where (21) holds for complex Gaussian variables, (22) follows from the circularity of $\mathbf{z}$, and $\mathbf{O}$ denotes the zero matrix of the corresponding dimension. Next,

$$\mathbf{F}_{\mathbf{h},\mathbf{h}} = \mathrm{E}\left[|\psi(s)|^2\mathbf{z}\mathbf{z}^H\right] = \kappa\mathbf{C}_\mathbf{z}, \tag{23}$$
$$(\mathbf{F}_{\mathbf{r},\mathbf{r}})_{m,n} = (\mathbf{C}_\mathbf{z})_{i,k}\,(\mathbf{C}_\mathbf{z})_{l,j}\,, \tag{24}$$
$$\mathbf{F}_{\mathbf{g},\mathbf{h}} = \beta\mathbf{h}\mathbf{h}^T\mathbf{P}_\mathbf{z}^* - \mathbf{I}_{d-1} = -\mathbf{I}_{d-1}, \tag{25}$$
$$\mathbf{F}_{\mathbf{h},\mathbf{r}} = \mathbf{O}, \tag{26}$$

where $m = j + \frac{i-1}{2}(2d-2-i)$, $n = l + \frac{k-1}{2}(2d-2-k)$ for $i = 1,\ldots,d-1$, $j = 1,\ldots,i$ and $k = 1,\ldots,d-1$, $l = 1,\ldots,k$; $\beta = \mathrm{E}[(\psi^*(s))^2]$. Then,

$$(\mathbf{F}_{\mathbf{g},\mathbf{r}})_{:,k} = \mathbf{C}_\mathbf{z}^{-1}\,(\mathbf{C}_\mathbf{z})_{:,i}\,(\mathbf{C}_\mathbf{z})_{j,:}\,\mathbf{h}, \tag{27}$$

where the index of k-th column reads $k = j + \frac{i-1}{2}(2d-2-i)$ for $i = 1,\ldots,d-1$, $j = 1,\ldots,i$. The other blocks of (14) follow from the hermiticity of $\mathbf{F}_{\boldsymbol{\theta}}$.

Now, we describe the computation of $\mathbf{P}$ in (9). Let $\mathbf{P}$ be partitioned as

$$\mathbf{P} = \begin{pmatrix} \mathbf{P}_{\mathbf{g},\mathbf{g}} & \mathbf{P}_{\mathbf{g},\mathbf{h}} & \mathbf{P}_{\mathbf{g},\mathbf{r}} \\ \mathbf{P}_{\mathbf{h},\mathbf{g}} & \mathbf{P}_{\mathbf{h},\mathbf{h}} & \mathbf{P}_{\mathbf{h},\mathbf{r}} \\ \mathbf{P}_{\mathbf{r},\mathbf{g}} & \mathbf{P}_{\mathbf{r},\mathbf{h}} & \mathbf{P}_{\mathbf{r},\mathbf{r}} \end{pmatrix}. \tag{28}$$

Then,

$$\mathbf{P}_{\mathbf{g},\mathbf{g}} = \rho \mathbf{h}\mathbf{h}^T - 2\mathbf{h}\mathbf{h}^T, \tag{29}$$

where

$$\rho = \mathrm{E}[(\psi^*(s))^2 (s^*)^2], \tag{30}$$
$$\sigma_{s^*}^2 = \mathrm{E}[(s^*)^2]. \tag{31}$$

Next,

$$\mathbf{P}_{\mathbf{h},\mathbf{h}} = \mathrm{E}\left[\psi^2(s)\right] \mathbf{P}_{\mathbf{z}} = \mathbf{O}, \tag{32}$$
$$\left(\mathbf{P}_{\mathbf{r},\mathbf{r}}\right)_{m,n} = \left(\mathbf{C}_{\mathbf{z}}\right)_{i,l} \left(\mathbf{C}_{\mathbf{z}}\right)_{k,j}, \tag{33}$$

where $m = j + \frac{i-1}{2}(2d-2-i)$, $n = l + \frac{k-1}{2}(2d-2-k)$ for $i = 1, \ldots, d-1$, $j = 1, \ldots, i$ and $k = 1, \ldots, d-1$, $l = 1, \ldots, k$. The other blocks read

$$\mathbf{P}_{\mathbf{g},\mathbf{h}} = \kappa \mathbf{h}\mathbf{h}^T \mathbf{C}_{\mathbf{z}}^*, \tag{34}$$
$$\left(\mathbf{P}_{\mathbf{g},\mathbf{r}}\right)_{:,k} = \mathbf{C}_{\mathbf{z}}^{-1} \left(\mathbf{C}_{\mathbf{z}}\right)_{:,j} \left(\mathbf{C}_{\mathbf{z}}\right)_{i,:} \mathbf{h}, \tag{35}$$
$$\mathbf{P}_{\mathbf{h},\mathbf{r}} = \mathbf{O}. \tag{36}$$

Finally, since it holds that

$$\mathbf{R} = \left(\mathbf{F}_{\mathbf{g},\mathbf{c}}^{\mathbf{T}}, \mathbf{F}_{\mathbf{h},\mathbf{c}}^{\mathbf{T}}, \mathbf{F}_{\mathbf{c},\mathbf{c}}^{\mathbf{T}}\right)^T, \tag{37}$$

the elements of $\mathbf{R}$ follows from (24) and (27) when considering $i = j$. Similarly, $\mathbf{F}_{\boldsymbol{\xi}}$ is obtained from (24) when $i = j \wedge k = l$.

4 CRLB-Induced Bound for ISR

A lower bound for the achievable mean Interference-to-Signal Ratio (ISR) can be derived using the CRLB. Let $\widehat{\mathbf{w}}$ be an estimated vector that separates s from $\mathbf{x}$, and let $\mathbf{A}_{\mathrm{ICE}}$ be the true mixing matrix in (5). Then, the ISR of the extracted signal $\widehat{s} = \widehat{\mathbf{w}}^H \mathbf{x}$ is

$$\mathtt{ISR} = \frac{\mathrm{E}[|\widehat{\mathbf{w}}^H \mathbf{y}|^2]}{\mathrm{E}[|\widehat{\mathbf{w}}^H \mathbf{a} s|^2]} = \frac{\mathbf{q}_2^H \mathbf{C}_{\mathbf{z}} \mathbf{q}_2}{|q_1|^2 \sigma_s^2} \approx \frac{1}{\sigma_s^2} \mathbf{q}_2^H \mathbf{C}_{\mathbf{z}} \mathbf{q}_2, \tag{38}$$

where $\mathbf{q}^H = [q_1, \mathbf{q}_2^H] = \widehat{\mathbf{w}}^H \mathbf{A}_{\mathrm{ICE}} = [\widehat{\mathbf{w}}^H \mathbf{a}, \widehat{\mathbf{w}}^H \mathbf{Q}]$. The last approximation is valid for "small" estimation error in $\widehat{\mathbf{w}}$, i.e., $|q_1|^2 \approx 1$ and $\mathbf{q} \approx \mathbf{e}_1$ (the unit vector). Then, the mean ISR value reads

$$\mathrm{E}\left[\mathtt{ISR}\right] \approx \frac{1}{\sigma_s^2} \mathrm{E}\left[\mathbf{q}_2^H \mathbf{C}_{\mathbf{z}} \mathbf{q}_2\right] = \frac{1}{\sigma_s^2} \mathtt{tr}\left(\mathbf{C}_{\mathbf{z}} \mathrm{E}\left[\mathbf{q}_2 \mathbf{q}_2^H\right]\right). \tag{39}$$

Finally, (39) reads

$$\mathrm{E}\left[\mathtt{ISR}\right] \approx \frac{1}{\sigma_s^2} \mathtt{tr}\left(\mathbf{C}_{\mathbf{z}} \mathtt{cov}\left(\mathbf{q}_2\right)\right). \tag{40}$$

The equivariance property of the BSE problem [10] enables us to consider the special case when $\mathbf{g} = \mathbf{h} = \mathbf{0}$, without any loss on generality. Then, $\mathbf{q}_2 = \widehat{\mathbf{h}}$, where $\widehat{\mathbf{h}} = \widehat{\mathbf{w}}_{2:d}$, and

$$\mathrm{E}\left[\mathtt{ISR}\right] \approx \frac{1}{\sigma_s^2}\mathtt{tr}(\mathbf{C}_\mathbf{z}\mathtt{cov}(\widehat{\mathbf{h}})) \geq \frac{1}{\sigma_s^2}\mathtt{tr}\left(\mathbf{C}_\mathbf{z}\mathtt{CRLB}\left(\mathbf{h}\right)\right), \tag{41}$$

where $\mathtt{CRLB}(\mathbf{h})$ denotes the diagonal block of $\mathcal{J}^{-1}\left(\tilde{\boldsymbol{\theta}}\right)$ corresponding to the parameter vector $\mathbf{h}$. For $\mathbf{h} = \mathbf{g} = \mathbf{0}$, the FIM simplifies to

$$\mathcal{J}\left(\tilde{\boldsymbol{\theta}}\right) = \begin{pmatrix} \sigma_s^2\mathbf{C}_\mathbf{z}^{-1} & -\mathbf{I}_{d-1} & \mathbf{O} \\ -\mathbf{I}_{d-1} & \kappa\mathbf{C}_\mathbf{z} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} & \mathbf{D} \end{pmatrix}, \tag{42}$$

where $\mathbf{D}$ is not the target of interest to derive in this paper. Since $\mathcal{J}\left(\tilde{\boldsymbol{\theta}}\right)$ is a block diagonal matrix, the block for $\mathtt{CRLB}(\mathbf{h})$ reads

$$\mathtt{CRLB}(\mathbf{h}) = \left(\kappa\mathbf{C}_\mathbf{z} - \frac{1}{\sigma_s^2}\mathbf{C}_\mathbf{z}\right)^{-1} = \frac{\sigma_s^2}{\kappa\sigma_s^2 - 1}\mathbf{C}_\mathbf{z}^{-1}. \tag{43}$$

Using (41) and (43), and by considering N observations, the CRLB-induced bound for ISR is obtained as

$$\mathrm{E}\left[\mathtt{ISR}\right] \geq \frac{1}{N}\frac{d-1}{\kappa\sigma_s^2 - 1}. \tag{44}$$

This result is in a good agreement with (2) in the following sense: Assume that $u_1, \ldots, u_d$ have all unit variance and that $u_2, \ldots, u_d$ have all circular Gaussian pdf, which means that $\kappa_j = 1$ for $j = 2, \ldots, d$. In that special case, the bound for the achievable ISR through ICA induced by (2) coincides with (44). Since, $\kappa\sigma_s^2 \approx \kappa_{norm}$, where $\kappa_{norm} = \mathrm{E}[|\psi(s)|^2]$ when s is normalized to the unit variance, note, that the bound (44) does not depend on the variances of signals.

5 Simulations

In simulations, we compare the bound for ICE with empirical mean ISR achieved by two methods: non-circular FastICA (NC-FastICA) for complex-valued signals from [9] designed for components belonging to the complex Generalized Gaussian Distribution (GGD) family [8] and OGICE (Orthogonally Constrained ICE) from [5]. OGICE is derived as an ICE algorithm based on maximum likelihood principle, so it might achieve asymptotic efficiency provided that it is initialized in the region of convergence to the SOI and the true score function is used as the internal nonlinear function.

In a trial, $d = 5$ independent complex-valued signals are generated. The target signal is drawn from a complex GGD with zero mean, unit variance, a shape parameter $\alpha \in (0, +\infty)$, and a circularity coefficient $\gamma \in [0, 1]$. The corresponding pdf is [6]

$$p(s, s^*) = \frac{\alpha\rho\exp\left(-\left[\frac{\rho/2}{\gamma^2-1}\left(\gamma s^2 + \gamma(s^*)^2 - 2ss^*\right)\right]^\alpha\right)}{\pi\Gamma(1/\alpha)(1-\gamma^2)^{\frac{1}{2}}}, \tag{45}$$

where $\rho = \frac{\Gamma(2/\alpha)}{\Gamma(1/\alpha)}$, and $\Gamma(\cdot)$ is the Gamma function. The other signals are circular Gaussian, which corresponds to $\alpha = 1$ and $\gamma = 0$. All signals are mixed by a random mixing matrix $\mathbf{A}$ according to (1).

OGICE is initialized by a randomly perturbed first column of $\mathbf{A}$, while the initialization of NC-FastICA is random in full (the output channel containing the separated SOI is determined based on the known ISR). In OGICE, the nonlinearity is the same as the true score function derived from (45) , that is,

$$\psi(s, s^*) = \frac{2\alpha(\rho/2)^\alpha}{(\gamma^2 - 1)^\alpha} \left(\gamma s^2 + \gamma (s^*)^2 - 2ss^*\right)^{\alpha - 1} (\gamma s - s^*), \tag{46}$$

where the parameters are set to their true values. NC-FastICA is endowed by the nonlinearity proposed in [9].

Figs. 1, 2, 3 and 4, show average ISR achieved by the algorithms in 1000 trials, respectively, for varying N, α, and γ. The average ISRs achieved by OGICE are very close to the bound (44), where, for the GGD pdf (45), $\kappa\sigma_s^2 = \kappa_{norm}$ is equal to [6]

$$\kappa_{norm} = \mathrm{E}\left[|\psi(s)|^2\right] = \frac{\alpha^2 \Gamma(2/\alpha)}{(1 - \gamma^2)\Gamma^2(1/\alpha)}. \tag{47}$$

The performance of NC-FastICA appears to be limited due to the choice of the nonlinearity, which is equal to the true score function of the SOI only when $\alpha = \frac{3}{2}$.

In Figs. 2 and 3, the ISR for sub-gaussian ($\alpha > 1$) and super-Gaussian ($\alpha < 1$) SOI is shown, respectively. For $\alpha = 1$, all signals, including the SOI, are circular Gaussian, in which case the mixing coefficients are not identifiable. Therefore, the ISRs approach 0 dB, which means no separation.

In Fig. 4, the non-circular Gaussian SOI with varying circularity is considered. The performance of NC-FastICA does not show any dependence on γ, which agrees with [9]. The ISR achieved by OGICE approaches the CRIB, which confirms that a non-circular Gaussian signal can be extracted from the other Gaussian signals when their circularity coefficient is different. This condition becomes violated as γ approaches 0, which corresponds with the decaying ISR.

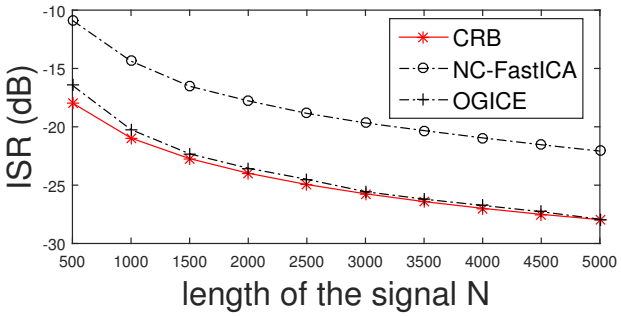


Figure 1: Average ISR for $d = 5$, $\alpha = \frac{1}{2}$, and varying N.

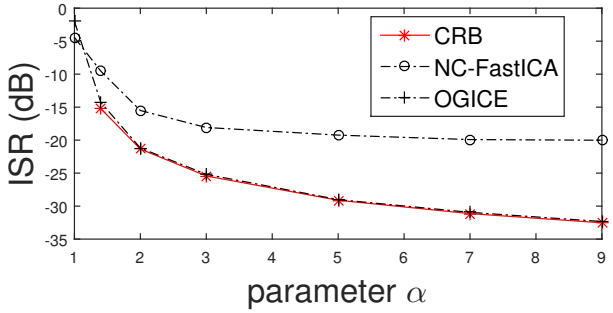


Figure 2: Average ISR for $d = 5$, $N = 2500$ and varying α.

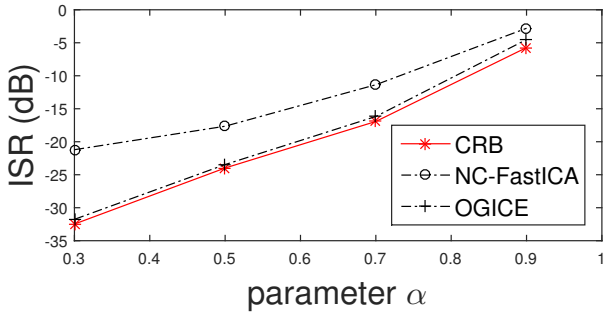


Figure 3: Average ISR for $d = 5$, $N = 2500$ and varying α.

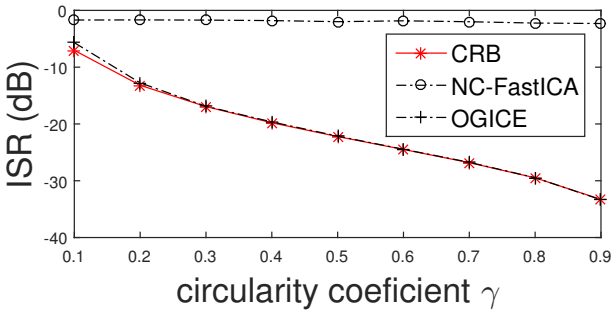


Figure 4: Average ISR for $d = 5$, $N = 2500$, $\alpha = 1$ and varying circularity coeficient γ.

6 Conclusions

The lower bound for achievable ISR through ICE model, the CRIB, is valid for both, circular and non-circular sources. The CRIB was shown to be attainable by OGICE, when the target signal is non-Gaussian or non-circular Gaussian. The derived bound depends on the target signal distribution and the length of data and it coincides with that for ICA when all but the target signals are circular Gaussian.

References

[1] P. Comon and C. Jutten. *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. Independent Component Analysis and Applications Series. Elsevier Science, (2010).

[2] J. Eriksson and V. Koivunen. *Complex random vectors and ICA models: Identifiability, uniqueness, and separability*. In 'IEEE Trans. Information Theory', volume 52, 1017–1029, (March 2006).

[3] A. Hyvärinen. *One-unit contrast functions for independent component analysis: a statistical analysis*. In 'Neural Networks for Signal Processing VII. Proceedings of the 1997 IEEE Signal Processing Society Workshop', 388–397, (Sep 1997).

[4] V. Kautský, Z. Koldovský, and P. Tichavský. *Cramér-Rao-induced bound for interference-to-signal ratio achievable through non-gaussian independent component extraction*. In '2017 IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)', 94–97, (Dec 2017).

[5] Z. Koldovský, P. Tichavský, and V. Kautský. *Orthogonally constrained independent component extraction: Blind MPDR beamforming*. 1195–1199, (September 2017).

[6] B. Loesch and B. Yang. *Cramér-Rao bound for circular and noncircular complex independent component analysis*. In 'IEEE Trans. Signal Processing', volume 61, 365–379, (Jan 2013).

[7] T. Menni, E. Chaumette, P. Larzabal, and J. P. Barbot. *New results on deterministic Cramér–Rao bounds for real and complex parameters*. In 'IEEE Trans. Signal Processing', volume 60, 1032–1049, (March 2012).

[8] M. Novey, T. Adali, and A. Roy. *A complex generalized gaussian distribution–characterization, generation, and estimation*. In 'IEEE Trans. Signal Processing', volume 58, 1427–1433, (March 2010).

[9] M. Novey and T. Adali. *On extending the complex FastICA algorithm to noncircular sources*. In 'IEEE Trans. Signal Processing', volume 56, 2148–2154, (May 2008).

[10] P. Tichavský, Z. Koldovský, and E. Oja. *Performance analysis of the FastICA algorithm and Cramér-Rao bounds for linear independent component analysis*. **54** (April 2006), 1189–1203.

Statically Configurable Representation of Conforming Unstructured Homogeneous Meshes for High-Performance Computing*

Jakub Klinkovský

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

advisors:

Tomáš Oberhuber, Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Radek Fučík, Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. A general data structure for an efficient representation of *conforming unstructured homogeneous* meshes for scientific computations on CPU and GPU-based systems is presented. Compared to existing open-source C++ libraries, such as OpenMesh [4], ViennaGrid [12], DUNE [3, 2], deal.II [1], OpenFoam [6], or libMesh [7], our implementation does not provide advanced features such as adaptive mesh refinement or non-conforming hierarchical meshes, but it focuses on efficient computations on the GPU accelerators.

We propose a data structure for an efficient representation of a class of unstructured meshes on GPUs as well as traditional CPU-based systems. The data structure presented is implemented as part of the TNL library [1] using the C++ language and the CUDA framework [9] for GPU utilization. The most notable features we aim to provide are high configurability via templates of the C++ language and efficient internal memory layout for CPU and GPU-based systems. The use of static configuration avoids the storage of unnecessary dynamic data, which minimizes the size of the data structures and increases efficiency of the cache and registers usage. The internal memory layout is based on state–of–the–art sparse matrix formats [8, 11, 10] for the storage of incidence matrices. Consequently, the coalesced memory access pattern [9] can be utilized during the parallel mesh traversal on GPU and the size of the data structure representing mesh entities can be kept constant, depending only on the entity shape but not on the number of its neighbours.

The efficiency of the implemented data structure is verified using several benchmark problems and its applicability to advanced numerical methods is demonstrated using the mixed-hybrid finite element method (MHFEM) applied to the two-phase flow in porous media [5]. We show speed-ups that rise above 32 in 2D and 59 in 3D when compared to sequential CPU computations, and above 5 in 2D and 11 in 3D when compared to a ten-threaded CPU computations.

Keywords: unstructured mesh, data structure, GPGPU, performance evaluation

*The work was supported by OPVVV project no. CZ.02.1.01/0.0/0.0/16_019/0000765: *Research Center for Informatics*, project no. 15-27178A of Ministry of Health of the Czech Republic, and the Student Grant Agency of the Czech Technical University in Prague project no. SGS17/194/OHK4/3T/14.

[1] http://www.tnl-project.org

Abstrakt. Článek se věnuje obecné datové struktuře pro efektivní reprezentaci *konformních nestrukturovaných homogenních* sítí pro vědecké výpočty na CPU a GPU. Ve srovnání s existujícími open-source knihovnami v jazyce C++, jako například OpenMesh [4], ViennaGrid [12], DUNE [3, 2], deal.II [1], OpenFoam [6] nebo libMesh [7], naše implementace neposkytuje pokročilé funkce, jako například adaptivní zjemnění nebo nekonformní hierarchické sítě, ale místo toho se zaměřuje na efektivní využití výpočetních akcelerátorů GPU.

V článku popisujeme datovou strukturu pro efektivní reprezentaci třídy nestrukturovaných sítí na GPU i tradičních systémech využívajících jen CPU. Tato datová struktura je implementovaná v rámci knihovny TNL[1] pomocí jazyka C++ a frameworku CUDA [9] pro využití GPU. Poskytované funkce této datové struktury zahrnují zejména vysokou konfigurovatelnost pomocí šablon jazyka C++ a efektivní interní rozložení dat v paměti. Pomocí statické konfigurace se lze vyhnout ukládání nepotřebných dat, což minimalizuje velikost datové struktury a zvyšuje efektivitu využití cache a registrů. Interní organizace dat v paměti počítače je založena na použití moderních formátů pro řídké matice [8, 11, 10] pro uložení incidenčních matic, což umožňuje využít sloučené přístupy do paměti na GPU [9].

Efektivita implementované datové struktury je ověřena pomocí několika testovacích úloh a vhodnost pro použití v pokročilých numerických metodách je prokázána pomocí numerického schématu založeného na hybridní metodě smíšených konečných prvků pro řešení dvoufázového proudění v porézním prostředí [5]. Dosažené výsledky při použití GPU vykazují urychlení více než 32 ve 2D a 59 ve 3D oproti sekvenčnímu výpočtu na CPU a urychlení více než 5 ve 2D a 11 ve 3D oproti desetivláknovému výpočtu na CPU.

Klíčová slova: nestrukturovaná síť, datová struktura, GPGPU, porovnání výkonu

Full paper: J. Klinkovský, T. Oberhuber, R. Fučík, V. Žabka. *Statically configurable representation of conforming unstructured homogeneous meshes for high-performance computing.* Submitted to Computer Physics Communications (2018).

References

[1] W. Bangerth, R. Hartmann, and G. Kanschat. *deal.II – a general purpose object oriented finite element library.* ACM Trans. Math. Softw. **33** (2007), 24/1–24/27.

[2] P. Bastian, M. Blatt, A. Dedner, C. Engwer, R. Klöfkorn, R. Kornhuber, M. Ohlberger, and O. Sander. *A generic grid interface for parallel and adaptive scientific computing. Part II: implementation and tests in DUNE.* Computing **82** (2008), 121–138.

[3] P. Bastian, M. Blatt, A. Dedner, C. Engwer, R. Klöfkorn, M. Ohlberger, and O. Sander. *A generic grid interface for parallel and adaptive scientific computing. Part I: abstract framework.* Computing **82** (2008), 103–119.

[4] M. Botsch, S. Steinberg, S. Bischoff, and L. Kobbelt. *Openmesh – a generic and efficient polygon mesh data structure.* In '1st OpenSG Symposium', (2002).

[5] R. Fučík, J. Klinkovský, J. Solovský, T. Oberhuber, and J. Mikyška. *Multidimensional mixed–hybrid finite element method for compositional two–phase flow in heterogeneous porous media and its massive parallel implementation on GPU.* Submitted to Computer Physics Communications (2017).

[6] H. Jasak, A. Jemcov, Z. Tukovic, et al. *OpenFOAM: A C++ library for complex physics simulations.* In 'International workshop on coupled methods in numerical dynamics', volume 1000, 1–20. IUC Dubrovnik, Croatia, (2007).

[7] B. S. Kirk, J. W. Peterson, R. H. Stogner, and G. F. Carey. *libMesh: A C++ Library for Parallel Adaptive Mesh Refinement/Coarsening Simulations.* Engineering with Computers **22** (2006), 237–254.

[8] A. Monakov, A. Lokhmotov, and A. Avetisyan. *Automatically tuning sparse matrix-vector multiplication for GPU architectures.* In 'High Performance Embedded Architectures and Compilers', Springer (2010), 111–125.

[9] NVIDIA. *CUDA Toolkit Documentation, version 9.0.* Nvidia, (2017). URL: `https://docs.nvidia.com/cuda/archive/9.0/`.

[10] T. Oberhuber and M. Heller. *Improved row-grouped CSR format for storing of sparse matrices on GPU.* In 'Proceedings of Algoritmy', 282–290, (2012).

[11] T. Oberhuber, J. Vacata, and A. Suzuki. *New row-grouped CSR format for storing the sparse matrices on GPU with implementation in CUDA.* Acta Technica CSAV **56** (2011), 447–466.

[12] F. Rudolf, K. Rupp, and J. Weinbub. Viennagrid 2.1.0 – user manual, (2014).

Rigidity of Spectra in Damped Unitary Ensembles of Hyperbolic Kind

Ondřej Kollert

3rd year of PGS, email: ondra.kollert@gmail.com
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Milan Krbálek, Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. One-dimensional systems with repulsively interacting elements can possess various forms of dependency structures described for instance by the spacings between the elements. For the simplest structure, where the spacings are independent, the standard methods of the counting process theory can be conveniently used. Nevertheless, this is not the case for most of real application systems where the elements can be represented for example by the eigenvalues of a self-adjoint random matrix, times of event occurrences or spatial locations of agents. Namely, such systems show extraordinarily complex dependencies where even farther spacings might be correlated. In that situation, the analytic solution is difficult, rather impossible to obtain. Some of the preliminary observations for the system of Gaussian random matrices and also the so called hyperbolic damped matrices are provided in this text. Its substantial part will be devoted to the comparison of rigidity with and without the spacing correlation.

Keywords: Rigidity, Hyperbolic Damped Matrix, Spacing Correlation, Repulsive Interaction

Abstrakt. Jedno-dimenzionalní systémy s repulsivně interagujícími elementy mohou vykazovat různé typy závislostních struktur, jež mohou být popsány např. odstupy sousedních elementů. Pro nejjednodušší případ, kde jsou všechny odstupy nezávislé, mohou být použity standardní metody teorie čítacích procesů. V reálných systémech se však většinou s tímto případem nesetkáváme. Může se jednat například o systémy, kde jsou příslušné elementy reprezentovány vlastními čísly samo-sdružených náhodných matic, časy příchodů jistých událostí, nebo poloh objektů v jedno-dimenzionálním prostoru. Takové systémy vykazují výjimečně složité závislosti kde se i vzdálenější odstupy daných elementů ovlivňují. V těchto případech jsou analytická odvození obtížná, ne-li nemožná získat. Některé z prvotních výsledků pro gaussovské a také takzvané hyperbolické tlumené matice jsou poskytnuty v tomto textu. Významná část textu se bude zabývat porovnáním rigidity s a bez závislosti mezi odstupy příslušných vlastních čísel.

Klíčová slova: rigidita, hyperbolická tlumená matice, korelace odstupů, odpudivá interakce

Introductory Talk

The random matrix theory is well known for its applications in describing various agent systems. In particular, eigenvalues of certain types of random matrices possess a similar interaction dependency structure as one-dimensional random variables characterizing certain agent systems. These random variables represent for instance arrival times of events or the locations of some objects in space while the interaction among them has

mostly a repulsive character. In other words, the probability of two neighboring random variables being very close to each other is very small for such systems.

In the work [1] and the corresponding paper [2], the theory of counting processes is used to mathematically describe the interaction dependency structure. The theory deals with the quantities such as level spacing or rigidity which are useful for comparing and classifying various types of interaction dependencies. For these quantities, many powerful theoretical results were obtained. However, the assumption of independence between neighboring distances of the ordered one-dimensional random variables (arrival times, object locations, random matrix eigenvalues,...) had to be satisfied. This assumption is very strong and it is not always met. As matter of fact, most of the studied systems do not fulfill such a requirement. For example, the system of random eigenvalues possess much more complicated interaction dependency structure.

The goal of this work is to provide some ways of revealing deeper dependency structure withing different types of random matrix spectra. In particular, the Wigner random matrices and the damped unitary matrices will be dealt with. For both of these types of random matrices, the neighboring spacings of the consecutive ordered eigenvalues are not independent. As a consequence, the theoretical results from the work [1] do not apply for system of these eigenvalues. Its actually extremely hard, rather impossible to derive any analytic results of that kind for the ensemble of random variables with such a complicated dependency structure. Thats why most of the approaches introduced in this work will be rather statistical and decriptive. Nevertheless, their significance and originality is undoubtedly unquestionable.

1 Gaussian and Damped Random Matrices

As mentioned, this work focuses on the properties of spectra of Wigner and damped random matrices. The definition of the first mentioned type of random matrices is provided below.

Definition 1. *Matrix* $\mathbf{W}_n = (W_{ij})_{i,j=1}^n$ *is said to be a Wigner matrix if* $W_{ij} = W_{ij}^*$ *and* W_{ij} *are independent identically distributed random variables with* $\mathrm{E}(W_{ij}) = 0$ *and* $\mathrm{E}|W_{ij}|^2 < +\infty$ *for* $i \leq j$.

The notion of Winger matrix is still quite general to be dealt with practically. Thus, the distribution of its elements will be considered to be Gaussian in this work. That yields the so called Gaussian orthogonal ensembles (GOE) and Gaussian unitary ensembles (GUE) in case the corresponding elements are distributed according to the normal distributions $\mathrm{N}(0,1)_{\mathbb{R}}$ and $\mathrm{N}(0,1)_{\mathbb{C}}$ respectively. These particular types of random matrices are actually the most famous ones and there are already many results known about them some of which can be found for instance in [8].

The second types of random matrices we will focus on here are the so called damped random matrices. They were first dealth with in the paper [7] as the numerical implementation of the so called Calogero-Moser random matrices. However, the notation 'damped' was introduced in the paper [4]. The general definition of this notion is given below.

Definition 2. *The random matrix* $\mathbf{D} = \left(D_{ij}\right)_{i,j=1}^{n}$ *is said to be a damped random matrix if* $(D_{ii})_{i=1}^{n}$ *are i.i.d. and* $D_{ij} = \mathrm{i}g/f_n(i-j)$ *a.s. for* $i \neq j$ *where* g *is positive. The function* f_n *is required to be continuous, odd and to satisfy* $|\lim_{t\to\infty} f_n(t)| = \infty$.

Again, the notion of damped matrices is quite general to deal with. Hence, let us assume the distribution of the diagonal elements to be $\mathrm{N}(0,1)_{\mathbb{R}}$ and the damping function f_n to satisfy

$$\frac{1}{f_n(t)} = \frac{2\pi}{n\sinh(2\pi t/n)}$$

for $t \in \mathbb{R}$. The matrices with such a choice of the function f_n are said to come from damped unitary ensembles of hyperbolic type ($\mathrm{DUE_h}$). They were chosen to be studied in this work because their eigenvalues describe the locations of the vehicles located in a one lane road quite well. Studying this correspondence is actually the next step in the research of the author.

2 Level Spacing and Rigidity

First of all, let us have a look at the level spacing of the spectra for the random matrices defined in the previous section. This fundamental quantity is used the most when it comes to describing the system of repulsively interacting elements. In our case, level spacing represents the spacing between neighboring ordered eigenvalues.

To actually investigate this quantity, it is crucial to apply the so called unfolding to the spectra first. That is a procedure where the individual eigenvalues are transformed so that the resulting neighboring eigenvalue spacings have the same scale. In fact, such a transformation is just the distribution function of the mixture of all the eigenvalues. As a result, the distribution of all the transformed eigenvalues is uniform. The eigenvalue

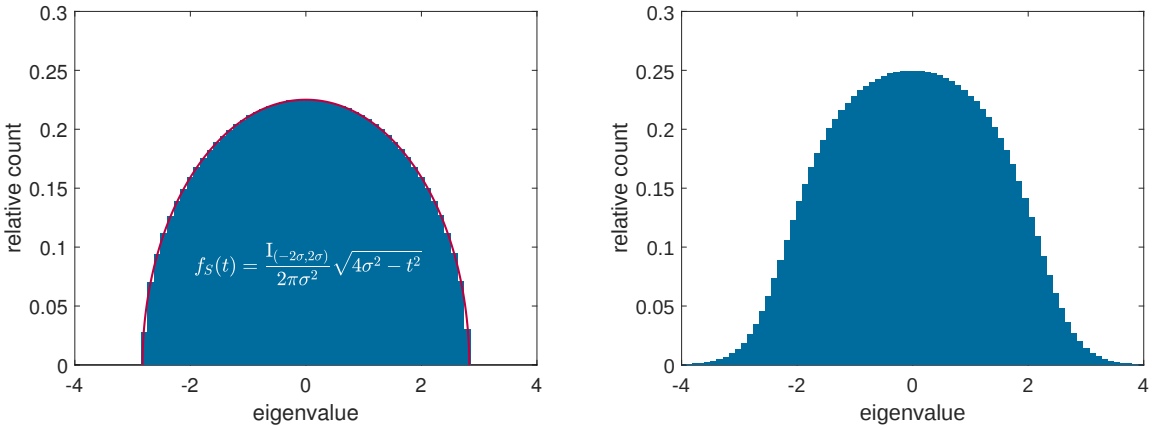


Figure 1: Histograms of eigenvalue densities for GUE (on the left) and $\mathrm{DUE_h}$ for $g = 0.5$ (on the right).

distributions (mixture distributions of all the eigenvalues) for some of the considered types of random matrices before unfolding are depicted in the figure 1. On the picture on the left, the asymptotic formula for the eigenvalue density is provided. This result is

also known as the Wigner semicircle law. The analytical version of the distribution on the second picture is not known yet. However, it is being intensely studied these days. The unfolding procedure is theoretically and practically dealt with in the text [3].

Even though the eigenvalue distributions in the figure 1 are not that different from each other, there is a remarkable distinction in the behavior of the individual eigenvalues in the spectra. In the work [3], it was found out that after the unfolding is applied, the nearest-neighbor spacings are identically distributed accross the whole spectra of Gaussian Matrices. In the case of hyperbolic damped matrices, the spacing distribution strongly depends on the position of the corresponding eigenvalues in the spectra. Nevertheless, it seems that the distribution is quite stable in the its middle part as can be seen from the estimations of the parameters α and β in the figure 2. That is also why only the

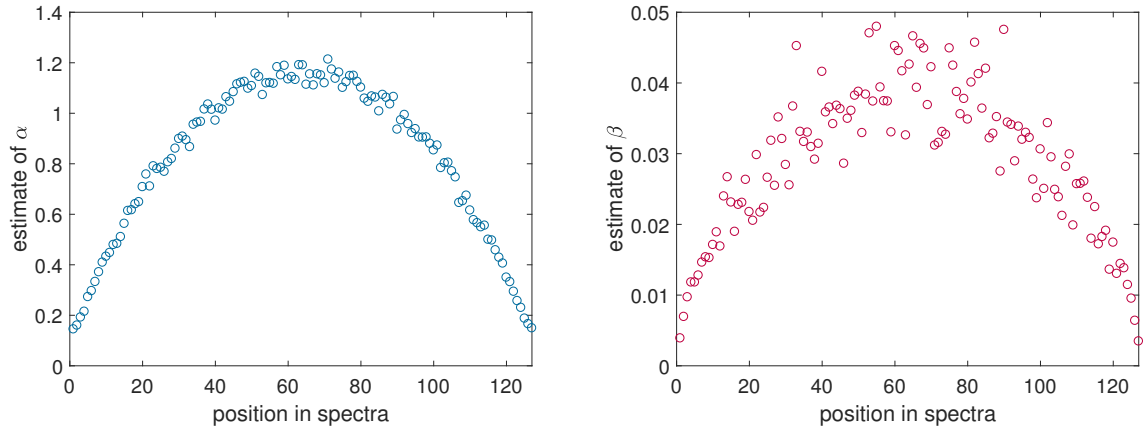


Figure 2: The estimates of the parameters α and β of the GIG density for the spacings of the nearest eigenvalues from $\mathrm{DUE_h}$ for the choice $g = 0.25$.

realizations of the spacings from that part of the spectra were considered for the case of $\mathrm{DUE_h}$. Particularly, the eigenvalues on the positions from 400 to 600 out of total 1000 eigenvalues generated and used throughout this whole text.

The histograms of the nearest-neighbor spacing distributions are depicted in the figure 3 for both Gaussian and hyporbolic damped random matrices. To be able to compare various spacing densities, the respective expected values are always required to be constant. In this work, we will consider that constant to be simply equal to one. From the figure 3, we can see that the spacing distribution of eigenvalues from GUE seems to have a lower variance than that of GOE. In case of $\mathrm{DUE_h}$, the variance decreases when the parameter increases. Comparing the two types of random matrices, the variance of Gaussian matrices is lower than that of hyperbolic damped matrices (for the values of g high enough). We point out the variance here since it naturally expresses the repulsive interaction between two consecutive eigenvalues.

The lines in the figure 3 represent the theoretical estimations of the distributions using the Wigner-type surmises

$$f_{\mathrm{GOE,GUE}}(t) = ct^{\gamma}\mathrm{e}^{-dt^2}, \quad f_{\mathrm{DUE_h}}(t) = ct^{\alpha}\mathrm{e}^{-\beta/t-dt} \tag{1}$$

for $t > 0$ where c, d are chosen so that the distributions have the expected value one. The parameter γ is equal to one (Weibull density) for GOE and in the case of GUE, it has the

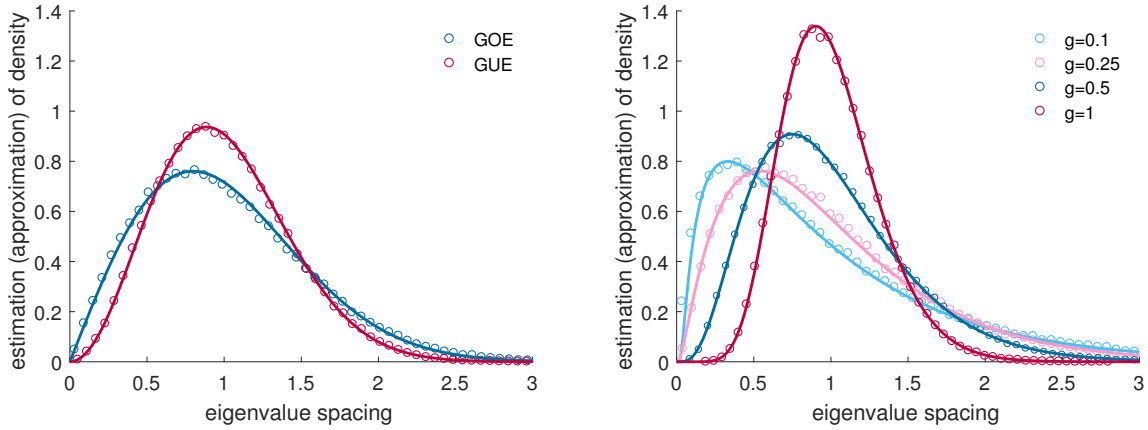


Figure 3: Histograms and fits of level spacing densities between two nearest eigenvalues from GOE, GUE (on the left) and $\mathrm{DUE_h}$ (on the right) for various values of the parameter g.

value two. The right-hand side function represents the density of the generalized inverse Gaussian (GIG) distribution. Its parameters α and β are the increasing functions of the matrix parameter g as can be seen in the figure (4). The trends of $\alpha(g)$ and $\beta(g)$ were studied in [3] more in detail.

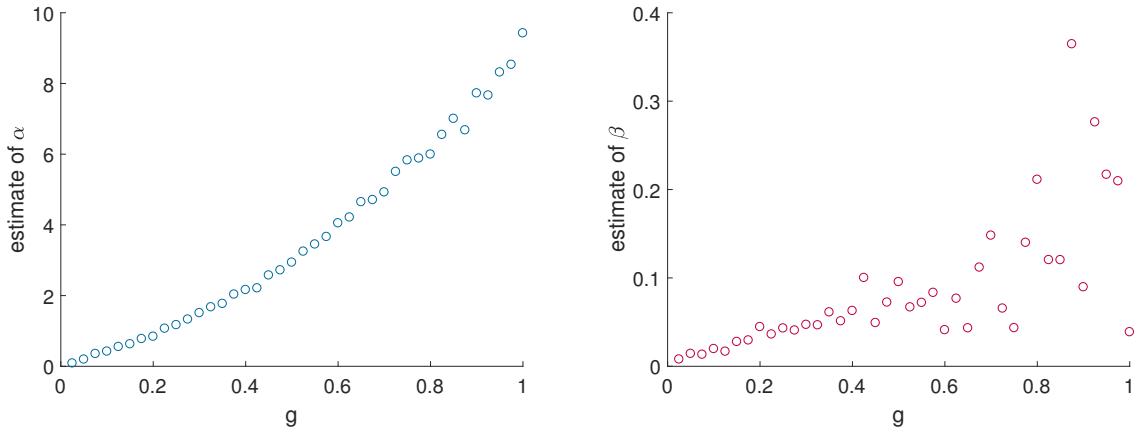


Figure 4: The estimates of the parameters α and β of the GIG density for the spacings of the nearest eigenvalues from $\mathrm{DUE_h}$ depending on the parameter g.

The part of the densities (1) containing the parameters α, β and γ corresponds to the behavior of the functions around zero. Thus, these parameters are directly related to the repulsive interaction between consecutive eigenvalues. To be more specific, the higher the values of α, β and γ are, the lower the variances of the distributions are and thus, the more the eigenvalues repulse each other. Comparing the behavior of the densities (1) around zero, the eigenvalues of DUE has much stronger type of repulsion than those of Guassian matrices.

The level spacing quantity is quite simple to obtain and thus, it is uselful for the first glimpse of the behavior of the spectra. Nevertheless, it expresses only the very local

interaction of eigenvalues. To include the higher-order kind of interaction, the quantity called rigity is more suitable to use. It is defined as the variance of the number of eigenvalues on the interval of certain length after the unfolding is applied to the spectra. Indeed, it really includes the interaction dependency among more distant eigenvalues. The estimates of this quanity for various interval lengths are depicted in the figure 5.

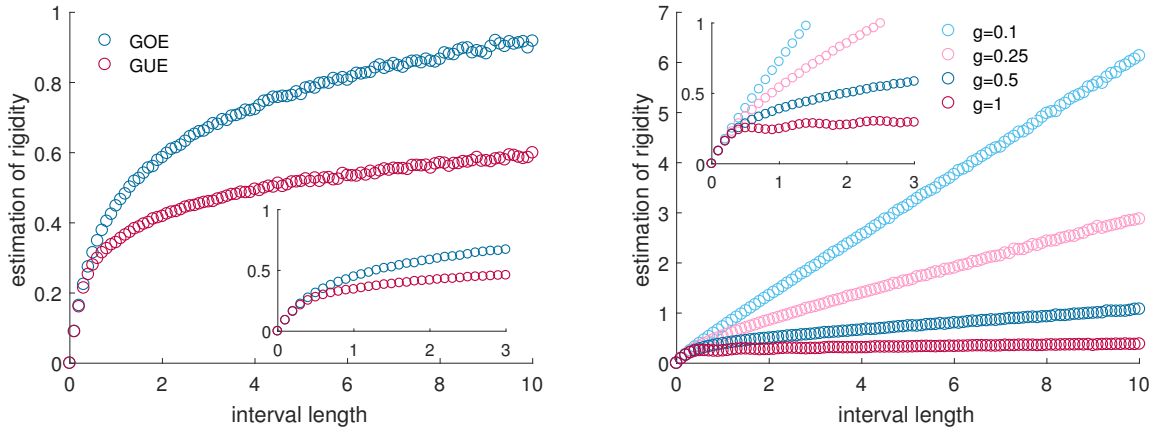


Figure 5: The estimations of the rigidity for GOE, GUE (on the left) and $\mathrm{DUE_h}$ (on the right) for various values of the parameter g.

According to the picture, the tail of rigidity is lower in the case of GUE than in that of GOE and the tail for $\mathrm{DUE_h}$ decreases when parameter g increases. In other words, the behavior of rigidity for higher values of its argument seems to represent the repulsive interaction between the elements as well as it was previously deduced with the variance of the level spacing distribution. But what is astonishing here is that even though the neighboring eigenvalues of $\mathrm{DUE_h}$ (for the values of g high enough) repulse each other more than those of GOE or GUE, the asymptotic trend of the rigidity increases less significantly in the case of the Gaussian matrices. In particular, the functions depicted on the left of the figure 5 have logaritmic trends while the ones on the righ-hand side have simply linear trend. The reason for such a discrepancy is that the dependency between distinct eigenvalue spacings in GOE and GUE is stronger than that in $\mathrm{DUE_h}$. More thorough investigation of this phenomena is the main goal of this work and it will be dealt with in the next section.

Before we proceed to the next section, let us compare the variance functions for the spectra of both studied kinds of random matrices under the assumption of independece of distinct eigenvalue spacings. This scenario can be obtained by random shuffling of the nearest-neighbor spacings. The figure 6 shows the resulting rigidities compared to previously shown ones. As can be seen, all the depicted functions have a linear asymptotic trend now. This was actually thoroughly studied in the work [1] and presented within the paper [2] using the theory of counting processes. It was shown that the asymptotic behavior of rigidity is linear under the assumed independence of distinct spacings. Particularly, it was proved that the asympotic expansion of ridigity (denoted as Δ) is

$$\Delta(t) = \frac{\mu_2 - \mu_1^2}{\mu_1^3} t + \frac{3\mu_2^2 - 2\mu_1\mu_3}{6\mu_1^4} + \mathrm{o}(1) \quad t \to \infty \tag{2}$$

where μ_k is the k-th moment of the corresponding level spacing. From this formula, it is apparent that the assymptotic behavior of rigidity depends only on the first three moments of the level spacing distribution. That is not surprising since the corresponding

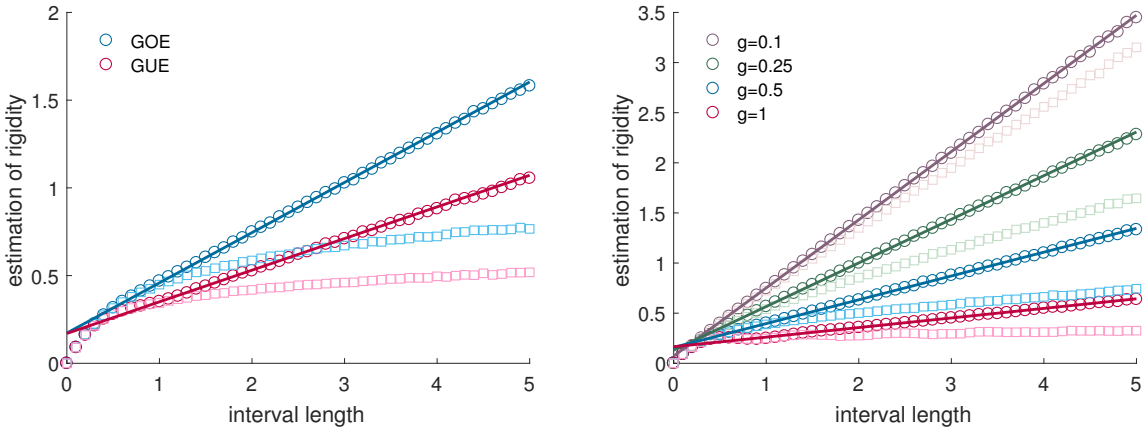


Figure 6: The estimations of the rigidity for GOE, GUE (on the left) and $\mathrm{DUE_h}$ (on the right) for various values of the parameter g after the shuffling of the eigenvalue spacings. The circles represent the case of the shuffled eigenvalue spacing and the squares the originally generated spacings. The lines are the estimates of the asymptote using the result (2).

counting process is entirely defined only through that distribution since the neighboring spacings are independent. As a consequence, the slopes of the estimated asymptotes depicted in the figure 6 are strictly higher in the case of GOE and GUE than those of $\mathrm{DUE_h}$ (for the values of g high enough). That is already consistent to the comparison of the variances of the densities (1) discussed previously since according to (2), the slope of rigidity is equal to the variance of level spacing.

The slope and the intercept in (2) can be thus estimated from the available data using the estimates of the first three moments of the spacing distribution. The estimates of the asymptotic lines for all the considered random matrix cases are also presented in the figure 6.

The simple form of the expansion (2) allows one not only to simply estimate the asymptote from data, but compute it analytically as well provided that the corresponding spacing distribution is known. For the eigenvalues of the considered random matrices, we know only approximations (1). In the case of the Gaussian matrices, the approximative expansions take the forms

$$\Delta_{\mathrm{GOE}}(t) \approx \frac{4-\pi}{\pi}t + \frac{8-2\pi}{\pi^2} + \mathrm{o}(1) \quad t \to \infty\,,$$

$$\Delta_{\mathrm{GUE}}(t) \approx \frac{3\pi-8}{8}t + \frac{27\pi^2-64\pi}{384} + \mathrm{o}(1) \quad t \to \infty$$

where all the values were calculated directly. Considering the GIG density in (1), the calculation of the corresponding moments is not that straightforward.

Let us now present the method of finding an approximative expansion of the rigidity for $\mathrm{DUE_h}$. The required moments of the GIG distribution satisfy the well known relation

$$\mu_k = \left(\frac{\beta}{d}\right)^{k/2} \frac{K_{\alpha+k+1}\left(2\sqrt{\beta d}\right)}{K_{\alpha+1}\left(2\sqrt{\beta d}\right)} \tag{3}$$

where K_α is the Macdonald function of the order α. Plugging the relation (3) into the expansion (2) would give us the desired analytical approximation. However, its expression would be a bit cumbersome. Using the famous equation

$$K_{\alpha-1}(t) - K_{\alpha+1}(t) = -\frac{2\alpha}{t} K_\alpha(t)$$

in the combination with the relation (3) yields the possibility to express the moments of the GIG distribution in a very slick form only through the parameters α, β and the constant d. However, the specific expressions of the moments have to be calculated through the recurrent formula

$$\mu_k = \frac{\alpha + k}{d}\mu_{k-1} + \frac{\beta}{d}\mu_{k-2} \tag{4}$$

for $k \in \{2, 3, ...\}$ where μ_1 is considered to be one as was mentioned before. The relation above represents the homogeneous difference equation with non-constant coefficients. Its explicit solution is the aim of the further research of the author.

The relation (4) now suffices to easily compute the second and third moment of the GIG distribution. Using the expression (2), the analytical approximation of the slope a and the intercept b of the rigidity $\Delta_{\mathrm{DUE_h}}$ can be written in the form

$$\begin{aligned} a &\approx \frac{\alpha + \beta + 2}{d} - 1 \\ b &\approx \frac{(\alpha + \beta + 2)(\alpha + 3\beta)}{6d^2} - \frac{\beta}{3d} \end{aligned} \tag{5}$$

It is important to mention that the approximations of such type were first obtained in the work [5]. Their little downside is that they still depend on the variable d which cannot be explicitly expressed through the parameters α and β. Nevertheless, it can be numerically calculated using the equation (3) for $k = 1$ since the expected value μ_1 is assumed to be constant. The actual values of the variable d are for various values of α and β depicted in the figure 7. As shown on the pictures, the behavior of the function d is not very wild. It is linear when β is kept constant while for α being constant, the function is asymptotically linear where the asymptotes are visualized in the figure too. From both graphs, it is also possible to deduce the asymptotic expansions

$$\begin{aligned} d(\alpha, \beta) &= \alpha + \beta + 1 + \mathrm{o}(1) \quad \beta \to 0_+ \\ d(\alpha, \beta) &= \alpha + \beta + \frac{3}{2} + \mathrm{o}(1) \quad \beta \to \infty \end{aligned} \tag{6}$$

The first relation is implied by the value of d while $\beta = 0$ for which the respective density in (1) is that of gamma distribution. The second asymptotic relation above can be

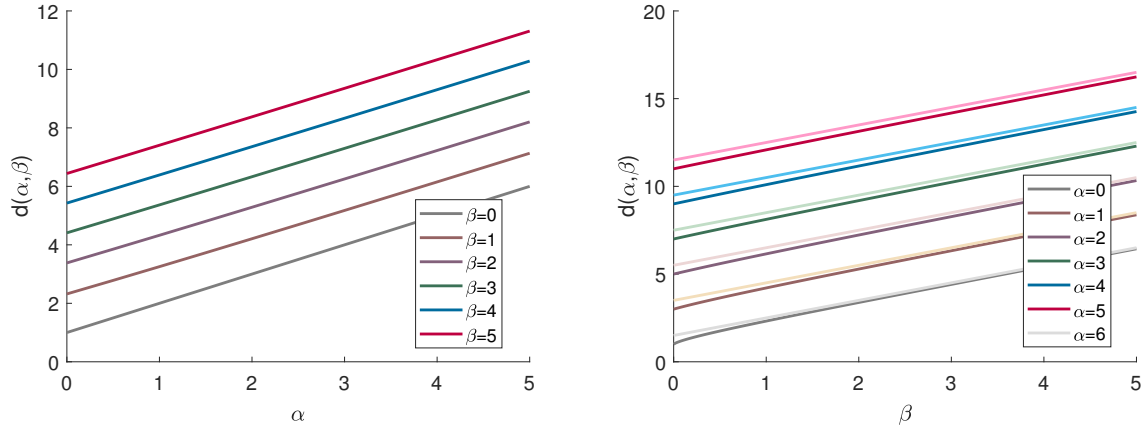


Figure 7: The function $d(\alpha, \beta)$ for various values of the parameter α (on the left) and various values of the parameter β (on the right).

obtained analytically through the solution of approximative Bessel differential equation. That is thoroughly dealt with for example in the work [6]. Another observation based on the figure 7 is that the convergence of d in β to its asymptote depends on the value of the parameter α. To be more specific, the higher the value of α is, the slower the function d is becoming linear when $\beta \to \infty$.

Before we finally go to the last section of this work, let us provide the comparison of the estimations of the asymptote in (2) based on the generated data and its theoretical approximations we just derived. As can be seen from the figure 2, the GIG distributions

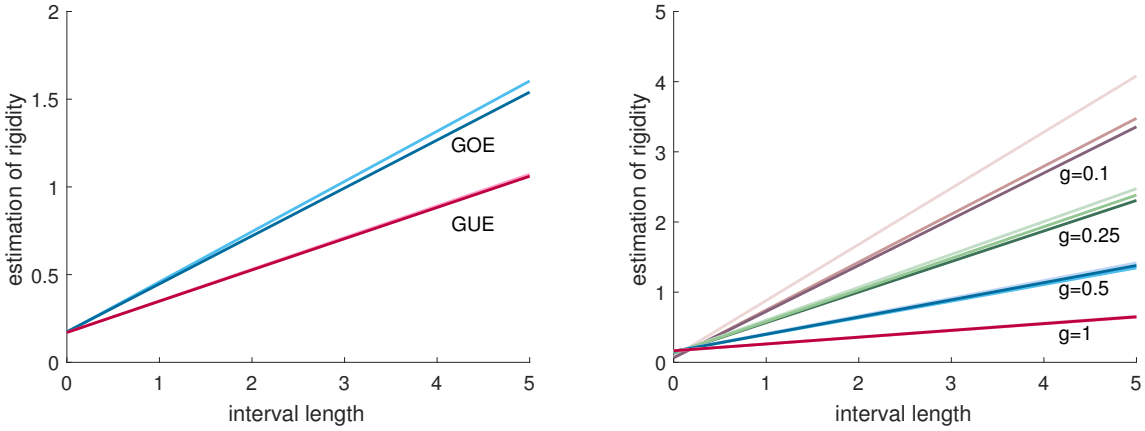


Figure 8: The comparisons between estimations (lower lines) and approximations of the asymtptote of the rigidity for the Gaussian (on the left) and the hyperbolic damped unitary matrices (on the right). In the latter case, the middle lines represent the approximations with the precise value of d and upper ones the approximations with approximated value of d.

we work here with have quite low values of the parameter β. Let us then use the relation (6) to deduce the approximation $d(\alpha, \beta) \approx \alpha+\beta+1$ and plug it in the expressions (5). The respective comparisons are given in the figure 8. The estimates based on the data should

be very precise so the comparions basically show how good the theoretical approximations are. Apparently, the stronger the repulsive interaction between neighboring eigenvalues is, the more precise the analytically obtained lines are. At last, the approximation $d(\alpha, \beta) \approx \alpha+\beta+1$ seems to work very well in all presented values of the parameter g, maybe except for the case $g = 0.1$. Let us now move to the final section where the the dependencies between distinct eigenvalue spacings will be investigated.

3 Autocorrelation Function for Sequence of Spacings

In the previous section, it was pointed out that the repulsive character of the neighboring eigenvalues does not have to be necessarily connected to the assymptotic behavior of the corresponding rigidity function. In the following text, this little mystery will be finally solved.

First of all, it is crucial to realise that the variance of the number of eigenvalues in a certain interval depends on the distribution of all eigenvalues in the spectra. Thus, rigidity function involves the information about the whole spectra, i.e. the interaction among not only two neighboring eigenvalues, but also the ones further from each other. As a consequence, let us now deal with the distribution of the spacings of the ordered eigenvalues between which there are several other eigenvalues. Such a quantity is said to be a multifold spacing. As well as in the case of the nearest neighbor spacing, it is first necessary to transform the spectra via unfolding.

Instead of dealing with the distributions of the multifold spacings as a whole, let us have a look at their variances. Their estimates for spectra of the considered random matrices are depicted in the figure 9. As can be seen, the variance does not increase

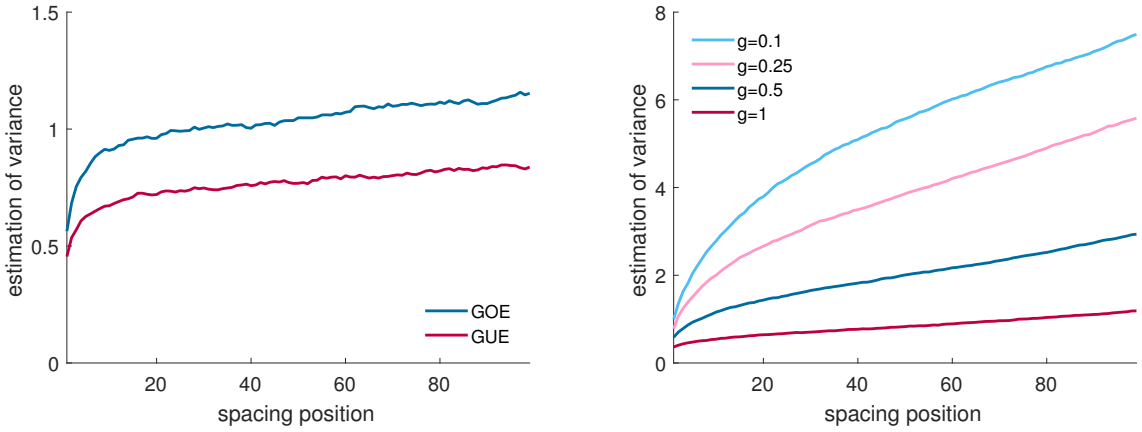


Figure 9: The estimates of the variances of the multifold spacing between eigenvalues of GOE and GUE (on the left) and of DUE_h (on the right).

rapidly with the difference netween the positions of the considered eigenvalues. The increase of the function is even less significant when the spacings of farther eigenvalues are considered. That is partly caused by the restriction that the eigenvalue distribution is uniform after the unfolding is applied. Considering this restriction is the same for all the studied spectra, the variance function in the case of Gaussian matrices increases

much slower than that of hyperbolic damped matrices. That directly implies that the variance of the eigenvalues within $\mathrm{DUE_h}$ is much larger than that of the eigenvalues from Gaussian matrices. In conclusion, even though the eigenvalues of the damped matrices repulse each other more (for the value of g high enough), their positions in spectra have much more variability than those of Gaussian matrices. This deduction is supported

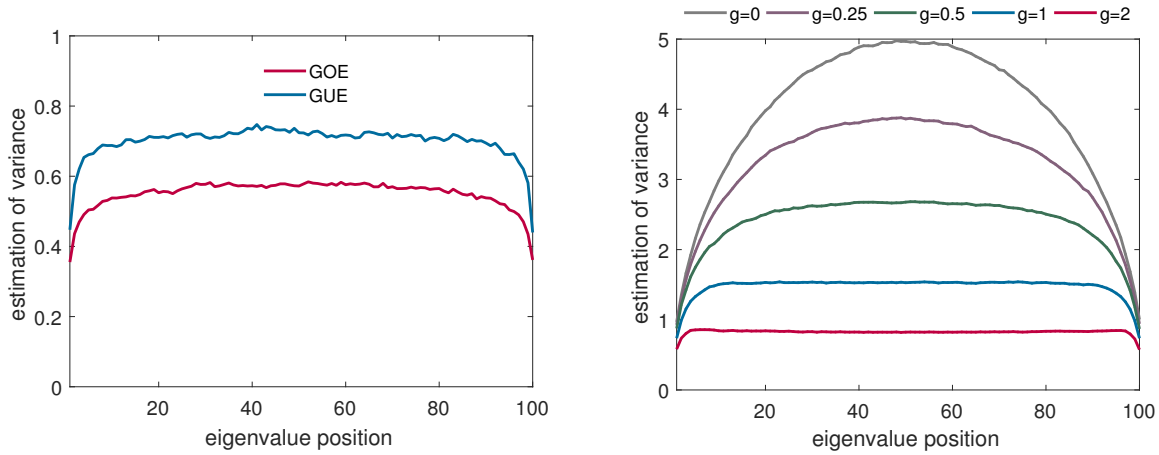


Figure 10: The estimates of the variances of the eigenvalues from GOE and GUE (on the left) and from $\mathrm{DUE_h}$ (on the right).

by the figure 10 in which the variances of the eigenvalues for both types of matrices are depicted. Additionally, the variance of the independent ordered random variables from uniform distribution (case $g = 0$) is depicted in the figure too. According to it, the influence of the restriction for the variables to be within the interval of the uniform density is actually quite significant.

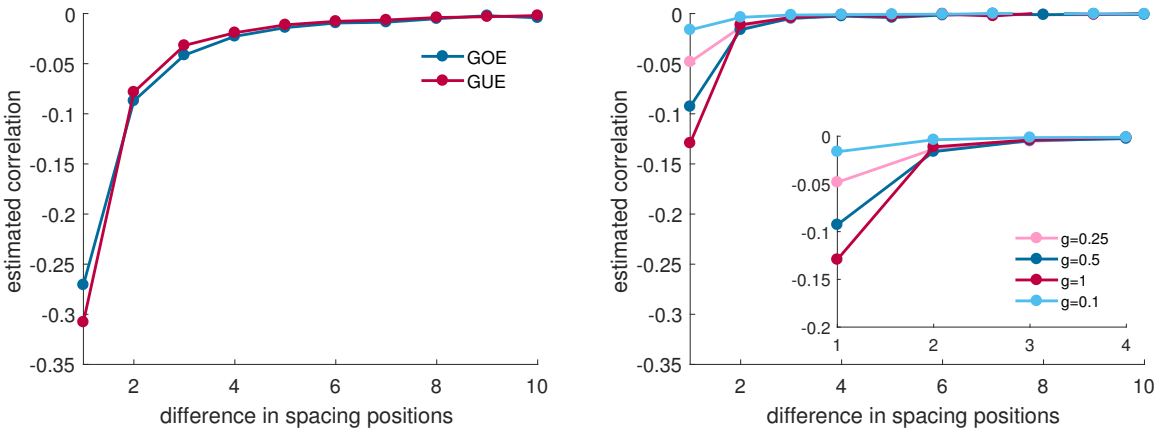


Figure 11: The estimates of the correlations of the eigenvalue spacings within GOE and GUE (on the left) and within $\mathrm{DUE_h}$ (on the right).

The conclusions just made explain why the rigidity of Gaussian matrices increase much lower than that of the damped matrices. But what is the reason for the difference in the variances of multifold spacing distributions when the two types of the random matrices are compared, that is still left to be clarified. As a matter of fact, it is due

to much stronger dependency structure of the distinct nearest-neighbor spacings of the eigenvalues in Gaussian matrices than it is in the case of the damped matrices.

In the figure 11, the estimation of the autocorrelation function of the sequence of the nearest-neighbor spacings is visualized. The correlation between neighboring eigenvalue spacings is stronger in GUE than in GOE. However, in both of these cases, the correlation is overall fairly stronger than in all the presented hyperbolic damped matrices. Not only the individual values of the correlations are higher in the case of Gaussian matrices, but the range of interaction of the eigenvalues from GOE and GUE is significantly higher as well. Another important observation is that all the estimates obtained have a negative sign. The negativity of the correlations actually says that if there is a spacing with large or low value, the following spacing will more likely have lower or larger value respectively.

To tackle the problem a bit more mathematically, let us denote the n-th ordered eigenvalue as Λ_n and the sequence of nearest-neighbor spacings as $S_i = \Lambda_i - \Lambda_{i-1}$ where $S_1 := \Lambda_1$. Than we can write

$$\mathrm{Var}(\Lambda_n) = \mathrm{Var}\left(\sum_{i=1}^{n} S_i\right) = \sum_{i=1}^{n} \mathrm{Var}(S_i) + \sum_{i \neq j} \mathrm{Cov}(S_i, S_j)\,.$$

The left-hand sum corresponds to the case when the spacings are completely independent. As was discussed, that allows one to derive many useful results for the related counting process. In a general case and also in the spectra of most random matrices, the right-hand sum is non-zero. In addition, its value is negative since the covariances (corrrelations) of all the spacings are negative as can be seen also in the figure 11. That explains why the variances of the eigenvalues in GOE and GUE are much lower than those in $\mathrm{DUE_h}$.

References

[1] O. Kollert, *Analysis of Random Matrix Spectra Using Counting Process Theory (diploma thesis)*, ČVUT v Praze, Fakulta jaderná a fyzikálně inženýrská, Praha (2015)

[2] O. Kollert, M. Krbalek, T. Hobza, *Level Counting Process Theory*, not published yet

[3] O. Kollert, *Unfolding of Spectra in Damped Unitary Ensembles of Hyperbolic Kind*, proceedings of Doktorandské dny 2017, ČVUT, FJFI

[4] T. Hobza, M. Krbalek, *Inner Structure of Vehicular Ensembles and Random Matrix Theory*, Physics Letters A 380 (2016)

[5] J. Vacková, *Perturbation Theory for Statistical Rigidity in Particle Systems(diplomová práce)*, ČVUT v Praze, Fakulta jaderná a fyzikálně inženýrská, Praha (2017)

[6] J. Vacková, *Multi-headway Statistics in Ensembles with Composite Potentials (research project)*, ČVUT v Praze, Fakulta jaderná a fyzikálně inženýrská, Praha (2016)

[7] E. Bogomolny, O. Giraud, C. Schmit, *Integrable random matrix ensembles*, Nonlinearity 24 (2011)

[8] M.L. Mehta *Random Matrices*, Academic Press, New York (2004)

Causal Network Discovery by Iterative Conditioning: Comparison of Algorithms*

Jakub Kořenek

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

advisor: Jaroslav Hlinka, Department of Complex Systems
Institute of Computer Science, CAS

Abstract. Estimating causal interactions in complex networks is an important problem encountered in many fields of current science. While a theoretical solution for detecting the graph of causal interactions has been previously formulated in the framework of prediction improvement, it generally requires the computation of high-dimensional information functionals – a situation invoking the curse of dimensionality with increasing network size. Recently, several methods have been proposed to alleviate this problem, based on iterative procedures for assessment of conditional (in)dependences. In the current work, we bring a comparison of several such prominent approaches. This is done both by theoretical comparison of the algorithms using a formulation in a common framework, and by numerical simulations including realistic complex coupling patterns. The theoretical analysis shows that the algorithms are strongly related; including one algorithm being in particular situations equivalent to the first phase of another. Numerical simulations suggest that the accuracy of most of the algorithms is under suitable parameter choice almost indistinguishable. However, particularly for large networks there are substantial differences in their computational demands, suggesting some of the algorithms are relatively more efficient in the case of sparse networks, while other perform better in the case of dense networks. The most recent variant of the algorithm by Runge et al. then provides a promising speedup particularly for large sparse networks, albeit appears to lead to a substantial decrease in accuracy in some scenarios. Based on the analysis of the reviewed algorithms, we propose a hybrid approach that provides competitive results both concerning computational efficiency and accuracy.

Keywords: Causality, Complex systems, VAR(1) process

Abstrakt. Detekce kauzálních vtahů mezi prvky komplexního systému je problém diskutovaný napříč vědeckými disciplínami. Jednou z metod navržených k této detekci je takzvaná transferní entropie, která je založená na výpočtu podmíněné vzájemné informace. V praxi však tato metoda naráží na problém odhadu podmíněné vzájemné informace vysoké dimenze při znalosti časových řad omezené délky. V poslední době tedy bylo navrženo několik algoritmů, které řeší problém vysoké dimenze úlohy pomocí iterativního podmiňování, v této práci přinášíme srovnání přesnosti a výpočetní složitosti těchto algoritmů. Numerické simulace naznačují, že přesnost studovaných algoritmů je při vhodné volbě vstupních parametrů téměř totožná, nicméně výpočetní složitost se u jednotlivých algoritmů výrazné liší, a to v závislosti na hustotě a velikosti odhadované sítě. Zatímco Rungeho algoritmus (a jeho modifikace) se zdají být velice efektivní

*This work was supported by the Czech Health Research Council Projects No. NV15-29835A, No. NV15-33250A, and No. NV17-28427A; and by project Nr. LO1611 with a financial support from the MEYS under the NPU I program.

při odhadu řídkých sítí, při odhadu velkých hustých sítí jeho výpočetní složitost dramaticky narůstá. Na základě studia těchto algoritmů jsme dále navrhli modifikaci Sunova algoritmu, která se zdá být vysoce efektivní při srovnatelné přesnosti s dříve navrženými algoritmy.

Klíčová slova: Kauzalita, Komplexní systémy, VAR(1) proces

Full paper: Hlinka, J.; Korenek, J., Causal network discovery by iterative conditioning: comparison of algorithms, https://arxiv.org/abs/1804.08173, 2018

Affine Moment Invariants of Vector Fields*

Jitka Kostková

4th year of PGS, email: kostkjit@fjfi.cvut.cz
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Jan Flusser, Department of Image Processing
Institute of Information Theory and Automation, CAS

Abstract. Vector fields are a special kind of multidimensional data, which are in a certain sense similar to digital color images, but are distinct from them in several aspects. A 2D vector field $\mathbf{f}(\mathbf{x})$ can be mathematically described as a pair of scalar fields (images) $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}))$. At each point $\mathbf{x} = (x, y)$, the value of $\mathbf{f}(\mathbf{x})$ shows the direction and the magnitude of the quantity which has been measured.

A common task in the analysis of vector fields is to detect patterns of interest. This includes detection of singularities such as sinks, vortices and saddle points, and other patterns similar to those stored in the database. The detection of patterns is realized by template matching. The matching algorithm must be in the first place invariant with respect to all possible deformations of the field. To detect the patterns of interest in the field, we can not use the methods for scalar images, but special matching methods and algorithms must be developed.

In this paper, we propose a method for the description and matching of vector field patterns under an unknown affine transformation of the field. Unlike digital images, transformations of vector fields act not only on the spatial coordinates but also on the field values, which makes the detection different from the image case. The transformation $\mathbf{f}'(\mathbf{x}) = B\mathbf{f}(A^{-1}\mathbf{x})$, where A and B be regular matrices, is called *independent total affine transformation* of the field $\mathbf{f}$.

To measure the similarity between the template and the field patch, we propose original invariants with respect to total affine transformation. They are designed from the vector field geometric moments. The invariance is reached by integration of *cross-products* $C_{ij} = x_i y_j - x_j y_i$ and *component cross-products* $F_{kl} = f_1(x_k, y_k) f_2(x_l, y_l) - f_1(x_l, y_l) f_2(x_k, y_k)$. The integral

$$V(\mathbf{f}) = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \prod_{k,j=1}^{r} C_{kj}^{n_{kj}} \cdot F_{kj}^{v_{kj}} \cdot \prod_{i=1}^{r} \mathrm{d}x_i \, \mathrm{d}y_i$$

is a relative invariant

$$V(\mathbf{f}') = J_B^v J_A^w |J_A|^r V(\mathbf{f})$$

and the subsequent normalization is needed.

It is demonstrated by experiments on real data from fluid mechanics (a vortex detection in a fluid flow), that they perform significantly better than potential competitors.

Keywords: Vector field, total transformation, affine invariants, template matching, vector field moments

Abstrakt. Vektorová pole jsou speciálním typem vícerozměrných dat, která jsou sice v jistém smyslu podobná barevným obrázkům, ale liší se významně v několika ohledech. Na dvourozměrné

*This work was supported by the Czech Science Foundation (Grant No. GA18-07247S) and by the Grant Agency of the Czech Technical University (Grant No. SGS18/188/OHK4/3T/14).

vektorové pole $\mathbf{f}(\mathbf{x})$ můžeme z matematického hlediska nahlížet jako na dvojici skalárních polí (obrázků) $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}))$. V každém bodě $\mathbf{x} = (x, y)$ vektor $\mathbf{f}(\mathbf{x})$ popisuje velikost a směr pozorované veličiny.

Rozšířenou úlohou analýzy vektorových polí je vyhledávání významných struktur, což zahrnuje detekci singularit - zřídel, sedlových bodů, vírů, a dalších specifických vzorů. Při detekci se často využívá databázového vyhledávání. Vyhledávací algoritmus musí být především nezávislý na všech přípustných deformacích vektorového pole. Abychom mohli v daném vektorovém poli vyhledávat specifické vzory, je potřeba vyvinout pro tento typ dat speciální metody a algoritmy.

V tomto článku navrhujeme metodu pro popis a vyhledávání vzorů ve vektorových polích při neznámé afinní transformaci pole. Na rozdíl od digitálních obrázků nepůsobí transformace vektorového pole pouze na prostorové souřadnice, ale mění také hodnoty pole, což je největší rozdíl mezi vektorovými poli a skalárními obrázky. Zobrazení $\mathbf{f}'(\mathbf{x}) = B\mathbf{f}(A^{-1}\mathbf{x})$, kde A a B jsou regulární matice, nazýváme nezávislou totální afinní transformací pole $\mathbf{f}$.

Aby bylo možné měřit podobnost mezi vzory a částmi polí, navrhujeme originální invarianty vůči totální afinní transformaci, které jsou tvořeny geometrickými momenty. Invarianty konstruujeme jako integrál ze součinu členů obsahujících kombinaci souřadnic $C_{ij} = x_i y_j - x_j y_i$ a komponent vektorového pole $F_{kl} = f_1(x_k, y_k) f_2(x_l, y_l) - f_1(x_l, y_l) f_2(x_k, y_k)$. Takovýto integrál

$$V(\mathbf{f}) = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \prod_{k,j=1}^{r} C_{kj}^{n_{kj}} \cdot F_{kj}^{v_{kj}} \cdot \prod_{i=1}^{r} \mathrm{d}x_i \, \mathrm{d}y_i$$

je ale pouze relativním invariantem

$$V(\mathbf{f}') = J_B^v J_A^w |J_A|^r V(\mathbf{f})$$

a je třeba jej následně normalizovat.

Na experimentech s reálnými daty z mechaniky kapalin (detekci vírů v proudící kapalině) ukazujeme, že navrhované invarianty si vedou výrazně lépe než jejich potenciální konkurenční metody.

Klíčová slova: Vektorové pole, totální transfromace, afinní invarianty, vyhledávání vzorů, momenty vektorových polí

Původní článek:

- J. Kostková, J. Flusser and T. Suk (in press). *Affine Invariants of Vector Fields*, Proceedings of the 2018 IEEE International Conference on Image Processing (ICIP 2018), Athens, 7–10 October 2018.

Accuracy at Top in Intrusion Detection

Václav Mácha

2nd year of PGS, email: machava2@fjfi.cvut.cz
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisors:
Václav Šmídl, Department of Adaptive Systems
Institute of Information Theory and Automation, CAS
Lukáš Adam,
Southern University of Science and Technology, Shenzhen, China

Abstract. Intrusion detection systems are frequently implemented as an ensemble of heterogeneous yet relatively simple detectors, which allows combining general-purpose and highly-specialized detectors. Aggregation of the output of these detectors is a challenging task especially in the considered domain where the data are highly imbalanced and the requirement of very low positive rate is essential. It has been recently shown that approach known as Accuracy at top is very suitable for this task. The complexity of the original optimal algorithm motivated the search for a computationally affordable alternative. We review previous work and the best performing heuristic algorithm. We propose two novel algorithms for the solution of the optimization task based on the surrogate approximation of the true quantile. On dataset from HTTP proxy logs, we show that one of the proposed algorithms systematically outperforms the algorithm.

Keywords: Anomaly detection, Accuracy at top, Intrusion detection

Abstrakt. Systémy pro detekci narušení jsou často konstruovány jako soubory heterogenních a přesto poměrně jednoduchých detektorů, které umožňují kombinovat univerzální detektory s vysoce specializovanými detektory. Hledání vhodné kombinace výstupů těchto detektorů je velmi složitý problém a to obzvláště v uvažované doméně, kde jsou vstupní data velmi nevyvážená a obsahují pouze malé procento potenciálně škodlivých (pozitivních) vzorků. Jako vhodný přístup k této problematice se ukazuje úloha zvaná Accuracy at top. Nicméně náročnost deterministického algoritmu řešícího tuto úlohu motivuje k hledání méně náročných alternativ. V této práci poskytujeme stručný popis nejlepšího heuristického algoritmu pro danou úlohu. Dále navrhujeme dva nové algoritmy založené na aproximaci skutečného kvantilu pomocí ztrátové funkce. Na datovém souboru z protokolů HTTP proxy ukazujeme, že jeden z navržených algoritmů systematicky dosahuje lepších výsledků než porovnávaný heuristický algoritmus.

Klíčová slova: Detekce anomálií, Accuracy at top, detekce narušení

1 Introduction

Accuracy at top (Acc@Top) is a binary decision problem, where a classifier returns τ fraction of the total of n samples such that the number of positive[1] ones among returned

[1] Without loss of generality it is assumed that samples of user's interest have a positive label.

is maximized. A good classifier should be therefore precise for samples with the highest score, while the performance on the rest of them is not important. Such classifiers found many applications in information retrieval systems, where for a given query the most relevant documents should be returned. Furthermore, they are useful in domains, where a large number of samples needs to be quickly screened and select those, which are possibly interesting for further evaluation. A prime example of such a domain is a computer security, where the intrusion detection system needs to make a decision over millions of events every five minutes [9]. Notice that a common property of domains, where Acc@Top is important is a high-class imbalance between objects of interest and remaining "background" objects.

Thinking about Acc@Top as a binary decision problem with an imbalance in class priors and costs of error, it should be possible to be solved by appropriately penalizing differently false positives and false negatives, as [2], yet these approaches are rarely used in practice due to difficulty in finding the right values of costs. Many works reformulate the problem as a ranking problem, where interesting samples should be ranked higher than non-interesting ones. An example is RankBoost [7] maximizing area under ROC curve, which is equivalent to optimizing ranking. Since the only accuracy in the top τ-quantile is important, Infinite Push [13] and Top Push [12] concentrate on the higher-ranked negatives and try to push them down. Both algorithms are also computationally efficient. Ref. [11] optimizes convex upper bound on the number of errors among the top k items. The algorithms solve a convex problem with exponentially many constraints by repeatedly optimizing a problem with the smaller number of constraints and adding the most violating ones after each iteration. A large number of constraints make the method computationally expensive.

None of the above formulations optimize the problem of interest, but rather sidestep it since samples outside of the quantile contribute to the error. Ref. [3] presented an algorithm finding the optimal solution of the Acc@Top problem, but the algorithm suffers from a complexity of $O(n^4)$ where n is the number of samples. Such complexity prevents the algorithm to be used on problems of decent size. This has been addressed by [10] by proposing a simple variant of the gradient descent, where after each step of the gradient descent method (GD) the solution is projected on the quantile constraint by alternating the threshold (not optimized by GD). Albeit losing the optimality of [3], large-scale experiments with different types of noise and showed that this heuristic algorithm exceeds all prior art. Therefore this algorithm is taken in this work as a baseline.

2 Problem Definition(s)

We assume that samples $\mathbf{x}$ come from an unknown mixture of distributions $P = \pi \cdot P^+ + (1-\pi) \cdot P^-$, where P^+ is an m-dimensional distribution of positive samples, P^- is an m-dimensional distribution of negative samples and $\pi \in (0,1)$ is the fraction of positive samples. The goal is to find function $f \in \mathcal{F}$ (where $\mathcal{F}$ is a class of functions $\mathbb{R}^m \to \mathbb{R}$) minimizing the number of false-positive samples in the top τ-quantile of distribution $f(\mathbf{x})$,

$\mathbf{x} \sim P$. This can be mathematically written as

$$\arg\min_{f \in \mathcal{F}} \quad \mathbb{E}_{\mathbf{x} \sim P^-} \left[f(\mathbf{x}) \geq t \right], \tag{1a}$$

$$\text{s.t.} \quad t \text{ is the top } \tau\text{-quantile of } f(\mathbf{x}), \tag{1b}$$

where $[\cdot]$ is the Iverson bracket which is equal to one if the argument is true and to zero otherwise, t is the top τ-quantile of the distribution $f(\mathbf{x})$, $\mathbf{x} \sim P$ defined as

$$t = \arg\max_{t} \left\{ \mathbb{E}_{\mathbf{x} \sim P} \left[f(\mathbf{x}) \geq t \right] \geq \tau \right\}. \tag{2}$$

Since the true probability distributions P^+ and P^- are unknown, the expectations are replaced by empirical estimates

$$\mathbb{E}_{x \sim P^-} \left[f(x) \geq t \right] \approx \frac{1}{n^-} \sum_{\mathbf{x} \in \mathcal{X}^-} \left[f(\mathbf{x}) \geq t \right], \quad \mathbb{E}_{x \sim P^+} \left[f(x) < t \right] \approx \frac{1}{n^+} \sum_{\mathbf{x} \in \mathcal{X}^+} \left[f(\mathbf{x}) < t \right], \tag{3}$$

where $\mathcal{X}/\mathcal{X}^+/\mathcal{X}^-$ denote the sets of all/positive/negative samples and $n/n^+/n^-$ their respective sizes. Furthermore, by $\hat{t}$ we denote the empirical estimate of the τ-quantile of the distribution $f(\mathbf{x})$, $\mathbf{x} \sim P$ defined by

$$\hat{t} = \arg\max_{t \in \mathbb{R}} \left\{ \sum_{\mathbf{x} \in \mathcal{X}} \left[f(\mathbf{x}) \geq t \right] \geq n \cdot \tau \right\}. \tag{4}$$

To simplify the above equations, we will use true-positive counts tp, false-negative counts fn, true-negative counts tn, and false-positive counts fp defined by

$$\mathrm{tp}(f,t) = \sum_{\mathbf{x} \in \mathcal{X}^+} \left[f(\mathbf{x}) \geq t \right], \qquad \mathrm{fn}(f,t) = \sum_{\mathbf{x} \in \mathcal{X}^+} \left[f(\mathbf{x}) < t \right],$$

$$\mathrm{tn}(f,t) = \sum_{\mathbf{x} \in \mathcal{X}^-} \left[f(\mathbf{x}) < t \right], \qquad \mathrm{fp}(f,t) = \sum_{\mathbf{x} \in \mathcal{X}^-} \left[f(\mathbf{x}) \geq t \right].$$

Using this notation, the empirical quantile (4) can be rewritten as

$$\hat{t} = \arg\max_{t \in \mathbb{R}} \left\{ \mathrm{tp}(f,t) + \mathrm{fp}(f,t) \geq n \cdot \tau \right\}.$$

Provided that $n \cdot \tau$ is an integer, the previous relation may be equivalently written as

$$\mathrm{tp}(f,\hat{t}) + \mathrm{fp}(f,\hat{t}) = n \cdot \tau. \tag{5}$$

In the opposite case, without any loss of generality, we may increase the value of τ. Note that this is not needed anymore when we pass to a continuous approximation later in (11). Then problem (1) can be approximated by

$$\arg\min_{f \in \mathcal{F}} \quad \frac{1}{n^-} \mathrm{fp}(f,t), \tag{6a}$$

$$\text{s.t.} \quad \mathrm{tp}(f,t) + \mathrm{fp}(f,t) = n \cdot \tau. \tag{6b}$$

For the analysis of this problem, the next theorem will play a crucial part.

Theorem 1. *For the true quantile $\hat{t}$ we have*

$$\mathrm{fp}(f,\hat{t}) = \mathrm{fn}(f,\hat{t}) + n\cdot\tau - n^{+}.$$

Proof. Denoting $\mathrm{p}(f,t) = \mathrm{tp}(f,t) + \mathrm{fp}(f,t)$ to be the count of all positive samples, we observe that $\mathrm{p}(f,\hat{t}) = n\cdot\tau$. Then we have

$$\mathrm{fp}(f,\hat{t}) = \mathrm{p}(f,\hat{t}) - \mathrm{tp}(f,\hat{t}) = n\cdot\tau - \mathrm{tp}(f,\hat{t}) = n\cdot\tau - n^{+} + \mathrm{fn}(f,\hat{t}).$$

□

Note that the previous theorem does not hold if the true quantile $\hat{t}$ is replaced by an arbitrary t. Since shifting the objective by a constant or multiplying it by a positive scalar does not change the optimal solution and since the only feasible point of (6b) is $\hat{t}$, Theorem 1 tells us that (6) is for any $\alpha\in[0,1]$ equivalent to

$$\begin{aligned} \underset{f\in\mathcal{F}}{\arg\min}\quad & \alpha\cdot\mathrm{fp}(f,t) + (1-\alpha)\cdot\mathrm{fn}(f,t), && \text{(7a)}\\ \text{s.t.}\quad & \mathrm{tp}(f,t) + \mathrm{fp}(f,t) = n\cdot\tau. && \text{(7b)} \end{aligned}$$

By using relation

$$\mathrm{tp}(f,t) + \mathrm{fn}(f,t) + \mathrm{fp}(f,t) + \mathrm{tp}(f,t) = n, \tag{8}$$

the τ-quantile can be expressed using the true-positives and the false-positives and we arrive at another equivalent formulation

$$\begin{aligned} \underset{f\in\mathcal{F}}{\arg\min}\quad & \alpha\cdot\mathrm{fp}(f,t) + (1-\alpha)\cdot\mathrm{fn}(f,t), && \text{(9a)}\\ \text{s.t.}\quad & \mathrm{tn}(f,t) + \mathrm{fn}(f,t) = n\cdot(1-\tau). && \text{(9b)} \end{aligned}$$

To obtain a numerically tractable problems, we have to further approximate (7) and (9) by their surrogate counterparts. Even though problems (6), (7) and (9) are equivalent to each other for any $\alpha\in[0,1]$ and form a discrete approximation of the continuous model (1), this is no longer true when we apply surrogate approximations. In the next sections, we derive different approximations of these equivalent problems and comment on their differences.

2.1 Surrogate Models

Since the empirical loss function (9a) is neither convex nor continuous, finding the optimal solution is an NP-complete problem [10]. A usual approach is to replace the Iverson bracket $[\cdot]$ with a convex surrogate function. The classical choices are the exponential, hinge, and logistic loss functions

$$\begin{aligned} l_{\exp}(f(\mathbf{x}),t,y) &= \exp\left\{-y\cdot(t-f(\mathbf{x}))\right\},\\ l_{h}(f(\mathbf{x}),t,y) &= \max\left\{0,1-y\cdot(t-f(\mathbf{x}))\right\},\\ l_{\log}(f(\mathbf{x}),t,y) &= \frac{1}{\ln 2}\ln\left(1+\exp\left\{-y\cdot(t-f(\mathbf{x}))\right\}\right), \end{aligned} \tag{10}$$

where $y \in \{-1, 1\}$ indicates which samples should be counted, i.e., $y = +1$ indicates samples above the threshold t and $y = -1$ indicates samples below the threshold t. In the text below the symbol $l\left(f(\mathbf{x}), t, y\right)$ denotes any of these surrogate functions.

Using the surrogate function, the true-positive, the false-negative, the false-positive and the true-negative counts can be bounded from above as in [5] by

$$\overline{\mathrm{tp}}(f,t) = \sum_{\mathbf{x}\in\mathcal{X}^+} l\left(f(\mathbf{x}), t, +1\right) \geq \mathrm{tp}(f,t), \quad \overline{\mathrm{fn}}(f,t) = \sum_{\mathbf{x}\in\mathcal{X}^+} l\left(f(\mathbf{x}), t, -1\right) \geq \mathrm{fn}(f,t), \tag{11a}$$

$$\overline{\mathrm{fp}}(f,t) = \sum_{\mathbf{x}\in\mathcal{X}^-} l\left(f(\mathbf{x}), t, +1\right) \geq \mathrm{fp}(f,t), \quad \overline{\mathrm{tn}}(f,t) = \sum_{\mathbf{x}\in\mathcal{X}^-} l\left(f(\mathbf{x}), t, -1\right) \geq \mathrm{tn}(f,t). \tag{11b}$$

Then it is natural to approximate (7) and (9) by

$$\arg\min_{f\in\mathcal{F}} \quad \alpha \cdot \overline{\mathrm{fp}}(f,t) + (1-\alpha) \cdot \overline{\mathrm{fn}}(f,t), \tag{12a}$$

$$\text{s.t.} \quad \overline{\mathrm{tp}}(f,t) + \overline{\mathrm{fp}}(f,t) = n \cdot \tau, \tag{12b}$$

and

$$\arg\min_{f\in\mathcal{F}} \quad \alpha \cdot \overline{\mathrm{fp}}(f,t) + (1-\alpha) \cdot \overline{\mathrm{fn}}(f,t), \tag{13a}$$

$$\text{s.t.} \quad \overline{\mathrm{tn}}(f,t) + \overline{\mathrm{fn}}(f,t) = n \cdot (1-\tau), \tag{13b}$$

respectively. To derive bounds, we first realize that due to (5), all feasible solutions of (12) satisfy

$$\mathrm{tp}(f,t) + \mathrm{fp}(f,t) \leq n \cdot \tau$$

and all feasible solutions of (13) satisfy

$$\mathrm{tn}(f,t) + \mathrm{fn}(f,t) \leq n \cdot (1-\tau).$$

Then due to (11) solving (12) provides an upper bound on the true quantile $\hat{t}$ while solving (13) provides a lower bound on the quantile.

3 Systematization of Estimation Algorithms

Our aim in this section is to derive and compare algorithms which can be used to solve problems (12), (13). We consider only linear classifiers: the class of functions $\mathcal{F}$, therefore, equals to

$$\mathcal{F} = \left\{ f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} \mid \mathbf{w} \in \mathbb{R}^m \right\}.$$

3.1 Gradient method with projection

The algorithm proposed in [10] solves a problem where the surrogate is applied only to the objective but not to the constraints, thus

$$\arg\min_{\mathbf{w}\in\mathbb{R}^m, t\in\mathbb{R}} \quad \alpha \cdot \overline{\mathrm{fp}}(\mathbf{w}^\top\mathbf{x}, t) + (1-\alpha) \cdot \overline{\mathrm{fn}}(\mathbf{w}^\top\mathbf{x}, t), \tag{14a}$$

$$\text{s.t.} \quad \mathrm{tp}(\mathbf{w}^\top\mathbf{x}, t) + \mathrm{fp}(\mathbf{w}^\top\mathbf{x}, t) = n \cdot \tau, \tag{14b}$$

$$\sum_{i=1}^{m} w_i = 1, \tag{14c}$$

by alternating between the gradient step with respect to weights $\mathbf{w}$ and the projection on the quantile constraint. Notice that the optimization task above contains constraint of the space of linear classifiers $\mathcal{F}$. We assume that this constraint was necessary due to exponential loss function used in [10]. Although the exponential loss in our experiments is not used, we use the proposed algorithm unchanged and to satisfy the second constraint we use L_2 normalization. The major contribution with respect to [3] is its computational efficiency. The algorithm is summarized in Algorithm 1.

Algorithm 1 The *grill* algorithm proposed in [10] used to solve (14).

1: **input:** dataset $\mathcal{X}$ with binary labels y, confidence level τ
2: **output:** classifier weights $\mathbf{w}$, exact quantile t
3: **repeat**
4: set t_{k+1} to be the exact quantile from (14b)
5: compute the gradient $\mathbf{g}_k$ of the objective (14a)
6: set $\mathbf{w}_{k+1} \leftarrow \mathbf{w}_k - \eta_k \cdot \mathbf{g}_k(\mathbf{w}_k, t_{k+1})$
7: compute L_2 normalization of the vector $\mathbf{w}_{k+1}$
8: set $k \leftarrow k + 1$
9: **until** a termination criterion is satisfied

3.2 Smooth optimization of quantile surrogate

3.2.1 Lagrange multipliers

When passing to the surrogates, the optimization problem becomes smooth and thus can be solved by methods of smooth optimization. Here, we concentrate mainly on problem (12). Its Lagrangian reads

$$\mathcal{L}(\mathbf{w}, t; \lambda) = \alpha \cdot \mathrm{fp}(\mathbf{w}^\top \mathbf{x}, t) + (1 - \alpha) \cdot \overline{\mathrm{fn}}(\mathbf{w}^\top \mathbf{x}, t) + \\ + \lambda \cdot \left(\overline{\mathrm{tp}}(\mathbf{w}^\top \mathbf{x}, t) + \overline{\mathrm{fp}}(\mathbf{w}^\top \mathbf{x}, t) - n \cdot \tau \right), \quad (15)$$

where λ is a Lagrange multiplier. Then (12) amount to

$$\operatorname*{arg\,min}_{\mathbf{w} \in \mathbb{R}^m, t \in \mathbb{R}} \max_{\lambda \in \mathbb{R}} \quad \mathcal{L}(\mathbf{w}, t; \lambda). \quad (16)$$

We propose the most basic solution to this saddle problem in Algorithm 2, where we alternate between minimizing the Lagrangian with respect to the primal variables and maximizing it with respect to the dual variable.

3.2.2 Alternating direction method of multipliers

First, we define the augmented Lagrangian as

$$\mathcal{L}_\rho(\mathbf{w}, t; \lambda) = \alpha \cdot \mathrm{fp}(\mathbf{w}^\top \mathbf{x}, t) + (1 - \alpha) \cdot \overline{\mathrm{fn}}(\mathbf{w}^\top \mathbf{x}, t) + \\ + \frac{\rho}{2} \left(\frac{1}{n} \cdot \overline{\mathrm{tp}}(\mathbf{w}^\top \mathbf{x}, t) + \frac{1}{n} \cdot \overline{\mathrm{fp}}(\mathbf{w}^\top \mathbf{x}, t) - \tau + \lambda \right)^2, \quad (17)$$

Then the basic solution using alternating direction method of multipliers (ADMM) can be described as in Algorithm 3.

Algorithm 2 Proposed *surrq* algorithm based on the Lagrange multipliers method.

input: dataset $\mathcal{X}$ with binary labels y, confidence level τ
output: classifier weights $\mathbf{w}$, exact quantile t
repeat
    compute the gradient $\mathbf{g}_k$ of the Lagrangian (15) with respect to w, t
    set $(\mathbf{w}_{k+1}, t_{k+1}) \leftarrow (\mathbf{w}_k, t_k) - \eta_k \cdot \mathbf{g}_k(\mathbf{w}_k, t_k, \lambda_k)$
    compute the gradient $\mathbf{q}_k$ of the Lagrangian (15) with respect to λ
    set $\lambda_{k+1} \leftarrow \lambda_k + \eta_k \cdot \mathbf{q}_k(\mathbf{w}_{k+1}, t_{k+1}, \lambda_k)$
    set $k \leftarrow k + 1$
until a termination criterion is satisfied

Algorithm 3 Proposed *surrq-admm* algorithm based on the ADMM method.

input: dataset $\mathcal{X}$ with binary labels y, confidence level τ, $\lambda_0 = 0$
output: classifier weights $\mathbf{w}$, exact quantile t
repeat
    use gradient based algorithms (as in Algorithm 2) to solve optimization task

$$(\mathbf{w}_{k+1}, t_{k+1}) \leftarrow \underset{\mathbf{w}\in\mathbb{R}^m, t\in\mathbb{R}}{\arg\min} \quad \mathcal{L}_\rho(\mathbf{w}, t, \lambda_k)$$

    set $\lambda_{k+1} \leftarrow \lambda_k + q(\mathbf{w}_{k+1}^\top \mathbf{x}, t_{k+1}) - \tau$
    set $k \leftarrow k + 1$
until a termination criterion is satisfied

4 Experiments with Intrusion Detection Data

4.1 Evaluation dataset

In the experiments we used dataset created by the NetFlow anomaly detection engine [8, 14] which processes NetFlow [1] records exported by routers and other network traffic shaping devices. This dataset was manually labeled by Cisco analysts. More precise description of the used dataset and the data generating and labeling process is in [10].

The use of machine learning methods in security is frequently hindered by the lack of fully labeled datasets. While samples labeled as malicious are usually connected to a malicious behavior, due to human error, samples labeled as background are frequently actually malicious. We introduced three types of noise to model this phenomenon:

- None: No noise is present.
- ALN: All malicious activity types are present. Half of each malicious activity type is mislabeled.
- MAT: Half of the malicious activity types were removed from the training data but kept in the testing data. Half of the present malicious activity types are mislabeled.
- MLT: All malicious activity types are present. Half of the malicious activity types is completely mislabeled while the other half is only partially mislabeled.

The testing set is always noiseless.

4.2 Experimental settings

We have already described the numerical methods in Section 3. Concerning the parameters, we used the step length set to $\eta = 0.01$ and we stopped each algorithm after 1000 iterations. As the surrogate loss function, we used the hinge surrogate (10) and for all three algorithms we set $\alpha = 0.5$. The initial weights were uniform, thus $\mathbf{w}_0 = \frac{1}{m}$, where m is the number of anomaly detectors. As an initial value of the Lagrange multiplier in Algorithm 2 we used $\lambda_0 = 1.$ and in Algorithm 3 we used parameter $\rho = 1$.

4.3 Results

To evaluate the algorithm performance, one of the most popular approaches is the area under the Receiver Operating Characteristic curve (AUCROC) [6]. However, since we have highly unbalanced data and since our region of interest is the top τ-quantile, the preferred option is the area under the Precision-Recall curve (AUCPR) [4]. Assuming that malware samples have positive labels, precision and recall is defined by

$$\text{Precision} = \frac{\text{tp}}{\text{tp} + \text{fp}}, \quad \text{Recall} = \frac{\text{tp}}{\text{tp} + \text{fn}}.$$

To get a point on the PR curve, we fixed τ, run the optimization algorithm and computed the precision and recall on the top τ-quantile on the testing set. The whole PR curve is then the collection of all points for different values of τ.

Experimental results are summarized in Fig. 1 via PR-curves of all considered algorithms. Each graph corresponds to one type of noise introduced in Section 4.1. As can be seen, the *surrq-admm* algorithm outperforms other algorithms in almost all cases however for the MLT noise (d) this algorithm did not work well. Moreover, this algorithm is the most computationally expensive since it has to solve an auxiliary optimization task in each iteration. On the other hand, the *surrq* and *grill* algorithms are computationally cheaper but much worse than *surrq-admm* in most cases. In addition, for the *surrq-admm* algorithm, there is plenty of room to improve the algorithm, because we used a really naive implementation.

5 Conclusion

We have presented systematization of the methods used to solve Acc@Top problem in the security domain and proposed a new methods based on the surrogate approximation of the quantile. The proposed *surrq-admm* method consistently outperforms the prior art in almost all realization of the noise that is common in network intrusion detection. The main drawback is time complexity, which could be partially mitigated by using more advanced versions of SGD (such as Adam or RMSProp).

References

[1] E. B. Claise. Cisco Systems NetFlow Services export Version 9. (October 2004).

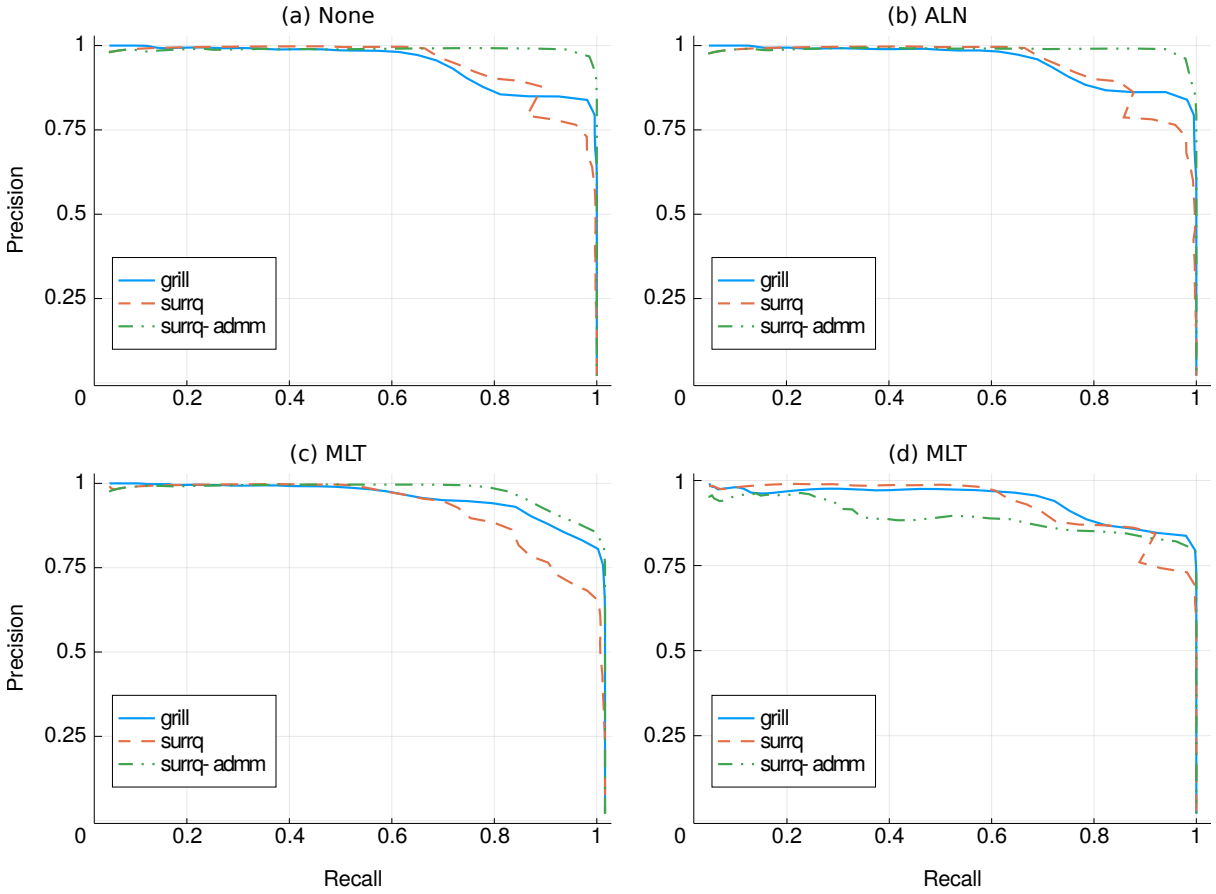


Figure 1: PR curves for all algorithms described in Section 3 and for all types of noises.

[2] F. R. Bach, D. Heckerman, and E. Horvitz. *Considering cost asymmetry in learning classifiers.* Journal of Machine Learing Research **7** (2006), 1713–1741.

[3] S. Boyd, C. Cortes, M. Mohri, and A. Radovanovic. Accuracy at the top. In 'Advances in neural information processing systems', 953–961, (2012).

[4] J. Davis and M. Goadrich. The relationship between precision-recall and roc curves. In 'Proceedings of the 23rd international conference on Machine learning', 233–240. ACM, (2006).

[5] E. Eban, M. Schain, A. Mackey, A. Gordon, R. Rifkin, and G. Elidan. Scalable learning of non-decomposable objectives. In 'Artificial Intelligence and Statistics', 832–840, (2017).

[6] T. Fawcett. *An introduction to roc analysis.* Pattern recognition letters **27** (2006), 861–874.

[7] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer. *An efficient boosting algorithm for combining preferences.* The Journal of machine learning research **4** (2003), 933–969.

[8] S. Garcia, M. Grill, J. Stiborek, and A. Zunino. *An empirical comparison of botnet detection methods.* Computers & Security **45** (2014), 100–123.

[9] M. Grill. *Automatic intrusion detection in computer networks.* Bachelor's Thesis (2008).

[10] M. Grill and T. Pevný. *Learning combination of anomaly detectors for security domain.* Computer Networks **107** (2016), 55–63.

[11] T. Joachims. A support vector method for multivariate performance measures. In 'Proceedings of the 22Nd International Conference on Machine Learning', ICML '05, 377–384, New York, NY, USA, (2005). ACM.

[12] N. Li, R. Jin, and Z.-H. Zhou. Top rank optimization in linear time. In 'Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1', NIPS'14, 1502–1510, Cambridge, MA, USA, (2014). MIT Press.

[13] A. Rakotomamonjy. *Sparse support vector infinite push.* arXiv preprint arXiv:1206.6432 (2012).

[14] M. Rehák, M. Pěchouček, M. Grill, J. Stiborek, K. Bartoš, and P. Čeleda. *Adaptive multiagent system for network traffic monitoring.* IEEE Intelligent Systems (2009), 16–25.

Modifications to the Fractal Patterns in Quantum Purification*

Martin Malachov

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

advisor: Igor Jex, Department of Physics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. Recent research has found that certain protocol designed for entanglement purification during quantum communication induces chaotic dynamic in pure input states. Additionally, it was proven that the induced fractal pattern is preserved in general mixed state dynamics; in other words, the chaos is robust to external noise. The fractal structure of sensitive states undergoes sudden changes resembling phase transition when the purity of the initial states decreases, i.e. the purity plays the role of the temperature.

We recapitulate these results and show that they are valid also for other protocols. We present a whole class of modifications to the original protocol using so called twirling operators. Studying these modifications we find new fractal structures and dynamical regimes. The chaotic dynamics of the quantum systems can be very rich. We show evidence that the phase transitions are also present in the new protocols but the quality of the phases and the transition temperatures change.

We present not only the mathematical results but also the methods we have developed to characterise the new dynamical regimes.

Keywords: qubit, chaos, phase transition

Abstrakt. Nedávné výzkumy prokázaly, že jistý protokol vyvinutý pro purifiikaci kvantového provázání v kvantové komunikaci, vyvolává chotickou dynamiku v čistých vstupních stavech. Dále bylo dokázáno, že vyvolané fraktální vzory se zachovávají i ve stavech smíšených; jinými slovy, chaos je odolný vůči vnějšímu šumu. Tyto fraktální struktury stavů citlivých na perturbaci procházejí fázovým přechodem, když se snižuje čistota počátečních stavů, tj. čistota hraje roli teploty.

Stručně shrneme nejnovější výsledky a ukážeme, že jsou platné i pro jiné protokoly. Uvedeme celou třídu modifikací původního protokolů užitím nových tzv. twirlingových operátorů. Pomocí těchto modifikací nalézáme nové frakktální struktury a dynamické režimy. Chaotická dynamika kvantových systémů tudíž může být překvapivě bohatá. Na příkladech ukážeme, že k fázovým přechodům dochází i v modifikovaných protokolech, mění se ovšem kvalita fází a také teploty přechodů.

Neprezentujeme pouze matematické závěry, ale také metody, které jsme vyvinuli k charakterizaci dynamických režimů.

Klíčová slova: qubit, chaos, fázový přechod

*This work has been supported by the Grant Agency of the Czech Technical University in Prague, grant No. SGS16/241/OHK/3T/14.

1 Introduction

The original protocol [1, 2] was designed to purify entanglement of the physical system consisting of two qubits. However, the protocol can be easily modified to act on a single qubit only. The main idea of both variations is to employ CNOT gate and measurement based selection. The measurement is responsible for nonlinear (irreversible) evolution at the cost of a copy of the original system. When the protocol is iterated the nonlinearity can result into chaos meaning sensitivity to the initial states - the slightest perturbation can dramatically change the asymptotic evolution. The chaotic states typically form fractal shaped structures or so called strange attractors.

In our case, the chaotic structure induced by the protocol [2] for initially pure states can be approached using theory of complex functions [3], see [2, 4]. The chaotic states form so called Julia set of the evolution function

$$f_0(z) = \frac{1 - z^2}{1 + z^2} \tag{1}$$

but for mixed states there is no applicable theory because the protocol manifests as three connected rational polynomial functions of three variables. For the purpose of describing the evolution of general mixed state we use following parameterisation of density matrices in computational basis, it is also known as Fano representation:

$$\rho = \frac{1}{2}\begin{pmatrix} 1 + w & u - iv \\ u + iv & 1 - w \end{pmatrix} \tag{2}$$

The evolution $\rho \to \rho'$ is then realised:

$$u' = \frac{2w}{1 + w^2}\,, \quad v' = \frac{-2uv}{1 + w^2}\,, \quad w' = \frac{u^2 - v^2}{1 + w^2} \tag{3}$$

When studying the action of the protocol on the mixed states it is useful to take into account all initial states with the same purity $P = \frac{1+u^2+v^2+w^2}{2}$. Such states form a sphere in the Euclidean space of coordinates (u, v, w). The evolution numerically calculated on these spheres reveals that the fractal structure present on the Bloch sphere $P = 1$ is preserved when P decreases. New dynamical regime is accessible - as soon as the purity is $P < 1$, there are states converging to the maximally mixed state $\rho = \frac{1}{2}\mathbb{1}$. There is a purity threshold - if the state has lower purity, it certainly converges to this attractor, i.e. cannot be purified in the sense of the original protocol.

In total, there are four possible asymptotic regimes: 1)+2) state converges to the pure cycle $|0\rangle \leftrightarrow (|0\rangle + |1\rangle)/\sqrt{2}$; 3) state converges to the mixed attractor $\frac{1}{2}\mathbb{1}$; 4) state is unstable and can exhibit chaotic or quasi-stable evolution. These states are formed by borders of convergence regions of previous attractors. Analytically, the chaos is proven only in pure states.

In the next text we present research analogous to [5] for more general protocols. We use the same approach regarding parameterisation and numerical analysis.

2 Protocol modification

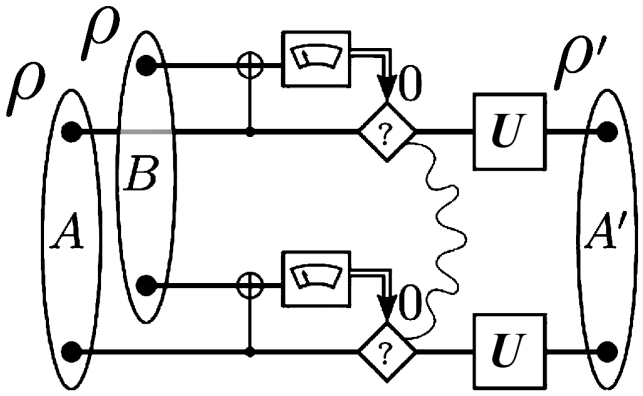


Figure 1: Protocol scheme. Main ingredient is the CNOT gate followed by measurement. The measurement based selection is responsible for nonlinearity of the protocol. Additional modification can be implemented by twirling gate U. In [5] Hadamard gate was used.

A natural question arises: what happens when the original protocol is modified? We now consider protocol with scheme according to figure 2. The modification lies in the use of general twirling operators U. For the purposes of this paper we choose a single-parameter family of operators

$$U_\vartheta = \begin{pmatrix} 1 & e^{i\vartheta} \\ -e^{-i\vartheta} & 1 \end{pmatrix} \tag{4}$$

This choice reduces for $\vartheta = 0$ to a case previously studied, see [6]. The gate is conjugated to the Hadamard gate used in the original protocol in the sense that two applications of both protocols on the same input state yield the same results in both cases.

First, we evaluate the evolution of pure states, we use the same approach as in [2, 5, 7]. We project the Bloch sphere onto complex plane, i.e. we parameterise each pure state uniquely with a complex number. The evolution is given by:

$$|\psi\rangle = \frac{1}{1+|z|^2}\begin{pmatrix}1\\ z\end{pmatrix}; \quad |\psi\rangle \to |\psi\rangle' \sim U_\vartheta \begin{pmatrix}1\\ z^2\end{pmatrix} \tag{5}$$

The evolution of such state is given by complex function:

$$z' = f_\vartheta(z) = \frac{z^2 - e^{-i\vartheta}}{e^{i\vartheta}z^2 + 1} \tag{6}$$

these functions are complex rational polynomial in variable z so they can be studied using theory [3]. The evolution of a general mixed state is driven by three rational polynomial functions depending on the parameter ϑ:

$$u' = \frac{-2w\cos\vartheta + (u^2-v^2)\sin^2\vartheta + 2uv\frac{\sin 2\vartheta}{2}}{1+w^2}, \; v' = \frac{2w\sin\vartheta + (u^2-v^2)\frac{\sin 2\vartheta}{2} + 2uv\cos^2\vartheta}{1+w^2}, \\ w' = \frac{(u^2-v^2)\cos\vartheta - 2uv\sin\vartheta}{1+w^2} \tag{7}$$

The mathematical background for such maps is insufficient (literature like [8, 9, 10] usually focuses on strictly defined problem and offers no general results) and we rely on numerical approach. This approach must be executed carefully because of computational precision.

3 Methods

In case of studying evolution of pure states one can employ the theory [3] on the evolution function. It has been shown that the lengths of critical cycles can vary in a complicated pattern to the chosen twirling operator parameters, [7]. The critical cycles are found by evaluating the derivative

$$\frac{d}{dz} f_\vartheta(z) \stackrel{!}{=} 0 \tag{8}$$

solving this equation one finds critical points and calculating their evolution one finds the ciritical cycles. The repellent cycles help to find the Julia set of the function.

We now follow slightly different approach. Instead of counting length of the critical cycles we decide to characterise the dynamics of the protocol by calculating the dimension of Julia set which typically is a fractal. The set itself is estimated as the border of basins of attraction which are obtained by considering (a grid approximating) all initial states and evaluating their evolution. States with the same asymptotic regime belong to the same basin of attraction and are assigned the same colour in the images we call attractor maps. These maps capture exactly the asymptotic dynamics as described and the borders of coloured form the sought Julia set.

The box counting dimension is a numerical concept approximating the dimension of an object. The studied object is covered by boxes of small size $\sim \varepsilon$ and the number N_ε of the boxes needed is compared to it in following way:

$$\mathcal{D} = \lim_{\varepsilon \to 0} \frac{\log N_\varepsilon}{\log 1/\varepsilon} \tag{9}$$

The exact formulation of the principle, its independence on the boxes shape, size and covering scheme can be verified in [11]. We perform the box counting estimate in following manner: after obtaining sufficiently many points $\vec{x}$ of the desired set (in the complex space or the threedimensional u, v, w space with respect to the situation) we calculate the number of boxes as follows: the boxes are determined by flooring the plane coordinates of the points:

$$\vec{x} \to \frac{1}{k} \lfloor k\vec{x} \rfloor \tag{10}$$

This procedure identifies points that are close to each other within scale $\frac{1}{k}$. The number of covering boxes is easily the number of unique points left after this flooring method is applied to all points of the considered set. Polishing the scale with increasing k we can simulate the limit (with restrictions given by the resolution of the numerical calculations) and estimate the dimension given by equation 9.

We can also face another problem: When the protocol would be generalised, the period of the critical cycles can be very large, the Julia set can be very complicated and fill the complex plane very densely. Than the box counting method can fail due to the numerical precision. Additionally, in the threedimensional dynamics of mixed states there is no guarantee that the chaotic states do not form a small set with nonempty interior (i.e. the box counting method would yield $\mathcal{D} = 2$) - such thing is impossible for pure states as the Julia set has no interior point or fills the whole complex space. To distinguish such regimes we suggest using another approach based on 'volume' of chaotic states. This

approach would compare number of states reaching certain asymptotic regimes after sufficient number of iterations.

To visualise the asymptotic dynamics we choose following method. We choose a two-dimensional surface - (single qubit) states with fixed initial purity. These initial states are parameterised in spherical coordinates:

$$u = \sqrt{2P-1}\sin\theta\cos\varphi\,,\quad v = \sqrt{2P-1}\sin\theta\sin\varphi\,,\quad w = \sqrt{2P-1}\cos\theta \tag{11}$$

where $P = (1+u^2+v^2+w^2)/2$ is the purity of the state. The coordinate space $\varphi \in [0,2\pi)\times\theta \in [0,\pi]$ is a rectangle and each considered state is a pixel (point) in this rectangle up to the poles which are projected as lines. Each state is uniquely parameterised with the triplet (P,φ,θ). When considering the attractor map, we fix the value P, in the rectangle of angles we assign each pixel a colour that codes the asymptotical regime of the state. We call the coloured rectangle an attractor map. The dominant feature of these map are coloured islands - slices through basins of attractions (formally, these objects are defined only in the dynamics of complex functions but we use the same term because current situation is straightforward analogy). The borders of coloured islands are formed of sensitive states which can be driven to different attractor depending on perturbation given to the initial state prior to evolution. In case of pure states, this border is the Julia set of the evolution function.

4 New dynamical regimes

The dynamical regimes for changing ϑ can be first approached analytically by the theory [3] when considering only pure states as the input states. We find critical points (we work on the Riemann sphere including $z=\infty$):

$$f'_\vartheta(z) = \frac{4z}{(1+e^{i\vartheta}z^2)^2} = 0 \;\Rightarrow\; z = 0 \vee z = \infty \tag{12}$$

When evaluating the evolution of these points in order to determine orbits of critical points we find that $f_\vartheta(\infty) = e^{-i\vartheta} = -f_\vartheta(0)$ which implies $f^2_\vartheta(0) = f^2_\vartheta(\infty)$. Both critical orbits coincide for any parameter ϑ. However, it is impossible to track whole the orbit for general ϑ - we present general evolution of point 0:

$$0 \to -e^{-i\vartheta} \to \frac{e^{-i\vartheta}-1}{e^{i\vartheta}+1} \to -e^{-i\vartheta}\frac{e^{3i\vartheta}+e^{2i\vartheta}+3e^{i\vartheta}-1}{e^{3i\vartheta}+3e^{2i\vartheta}-e^{i\vartheta}+1} \to \tag{13}$$

$$\to -e^{-i\vartheta}\frac{e^{7i\vartheta}+5e^{6i\vartheta}+5e^{5i\vartheta}-11e^{4i\vartheta}+3e^{3i\vartheta}-9e^{2i\vartheta}+7e^{i\vartheta}-1}{e^{7i\vartheta}+7e^{6i\vartheta}+9e^{5i\vartheta}+3e^{4i\vartheta}+11e^{3i\vartheta}+5e^{2i\vartheta}-5e^{i\vartheta}+1} \to \dots$$

The polynomials become more involved as more iterations are performed. Generally based on these, we find that there exist no parameters ϑ such that 0 would be critical fixed point or form critical cycle (periodic orbit of the critical point) with period 3. The only parameter ϑ for which 0 forms a 2-cycle or 4-cycle is $\vartheta = 0$. Although the orbit does not have to be generally periodical (be a cycle), the theory [3] states that the critical orbit converges to attractive or parabolic cycle. In order to find the Julia set, one can

use this fact together with backward iterations which make an attractive cycle repelling. The inverse function f_ϑ^{-1} has two branches for any ϑ as f_ϑ combines z^2:

$$f_\vartheta^{-1}(z) = \pm\sqrt{\frac{z + e^{-i\vartheta}}{1 - ze^{i\vartheta}}} \tag{14}$$

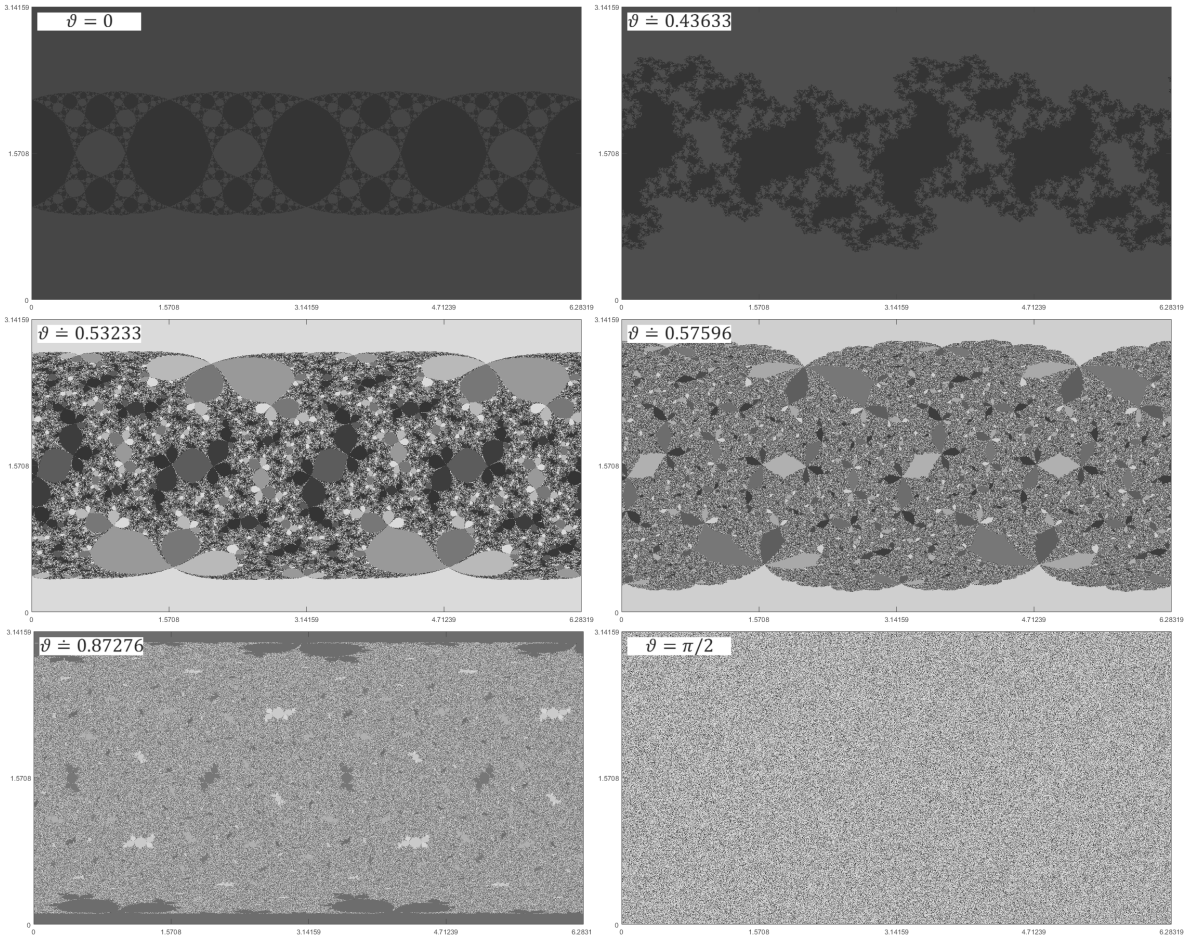


Figure 2: Pure states dynamics of protocols modified by twirling operator U_ϑ. The attractor map with $\varphi \in [0, 2\pi]$ on x-axis, $\phi \in [0, \pi]$ on y-axis codes the asymptotic dynamical regime of a state (point in the map) with colour (grey shades). Generally but with exceptions, with increasing ϑ the fractal grows through the complex plane.

We now investigate another feature present in case $\vartheta = 0$ and discussed in [4, 5]. There is a period-2 cycle which is found by solving $f_\vartheta^2(z) = z$. This equation gives five solutions. Two of them form a cycle:

$$z_1 = \frac{-e^{i\vartheta} - \sqrt{e^{i\vartheta} + e^{2i\vartheta} - e^{3i\vartheta}}}{e^{i\vartheta} + e^{2i\vartheta}} \overset{f_\vartheta}{\leftrightarrow} z_2 = \frac{-e^{i\vartheta} + \sqrt{e^{i\vartheta} + e^{2i\vartheta} - e^{3i\vartheta}}}{e^{i\vartheta} + e^{2i\vartheta}} \tag{15}$$

Note that $z_{1,2}(\vartheta) = \overline{z_{1,2}(2\pi - \vartheta)}$. Remaining three solutions $z_{3,4,5}$ are fixed states with very complicated analytical form. Their stability is investigated via derivative of the evolution function 6 yielding $|f'_\vartheta(z_{3,4,5})| \in (1, 2]$ which means these fixed states are

repelling (they belong to the Julia set) for all parameters ϑ. The stability of the attractive cycle is given by:

$$\left|\left(f_\vartheta^2\right)'(z_1)\right| = \left|\left(f_\vartheta^2\right)'(z_2)\right| = 2|\sin\vartheta| \tag{16}$$

where the power 2 of the function stands for two iterations of the function. This result is surprisingly elegant and distinguishes three cases (since the ϑ is an angle, we only consider $\vartheta \in [0, 2\pi]$):

1. $(0 \leq)\, \vartheta < \frac{\pi}{6} \vee \frac{5\pi}{6} < \vartheta < \frac{7\pi}{6} \vee \frac{11\pi}{6} < \vartheta\, (\leq 2\pi)$: the cycle $z_1 \leftrightarrow z_2$ is attracting.
2. $\vartheta \in \left\{\frac{\pi}{6}, \frac{5\pi}{6}, \frac{7\pi}{6}, \frac{11\pi}{6}\right\}$: the cycle $z_1 \leftrightarrow z_2$ is parabolic.
3. $\frac{\pi}{6} < \vartheta < \frac{5\pi}{6} \vee \frac{7\pi}{6} < \vartheta < \frac{11\pi}{6}$: the cycle $z_1 \leftrightarrow z_2$ is repelling.

This fact can be used when searching for points of Julia set. This cycle is part of it in cases 2. and 3. However, it is more convenient to use the backward iterations 14.

5 Phase transition

We now describe what happens to the fractal when the ϑ is increasing from 0. The fractal structure performs a torsion and suddenly proliferates through the complex plane. Although the fractal seems to gradually dominate the complex plane, it does not identify with it. Compare images of different fractal patterns in figure 2 for various values ϑ. Even though the image may seem to exhibit random noise, the fractal has zero volume and separates immensely high number of immediate basins of attraction.

The global action of the protocol on generally mixed states can be approached only numerically. However, there is an important piece of analytic result. The maximally mixed state is a fixed state of the protocol for any ϑ. Numerically we find following general property of the mixed dynamics: *The asymptotic dynamics found analytically for pure states is enriched with only one additional attractive asymptotical regime - the maximally mixed state $\rho = \frac{1}{2}$. The fractal pattern decays with lower purity.* This decaying is meant in the sense that the fractal pattern shrinks in size but keeps its fractalness. This effect is in detail explained in [5] (case analogous to $\vartheta = 0$) and we find it present for all values of ϑ. In figure 3 we show an example of the fractal structure changing when noise is added to the initial states.

The quality of the fractal structure is understood via its dimension. An acceptable way to estimate the dimension lies in the box counting method, see section 3 or consult with bibliography, e.g. [11]. When the box counting dimension is evaluated on attractor maps obtained for fixed value ϑ and various values of P, i.e. we check the structure of sensitive states depending on their initial purity, we find just like in [5] that the dimension changes suddenly. The discrete jumps when P is lowered is interpreted as phase transition. We claim that the transitions are not unique or $\vartheta = 0$ but are a rather general property. Few examples can be seen in figure 4. The temperatures of the transitions are not determined precisely due to high requirements on computational time and memory. Current results suggest that the fractal dimension has sudden jumps also when P is constant and ϑ is changed. This allows us to state a hypothesis that a phase diagram in ϑ, P space could be drawn. It would contain clear islands of constant phase separated by transition curves.

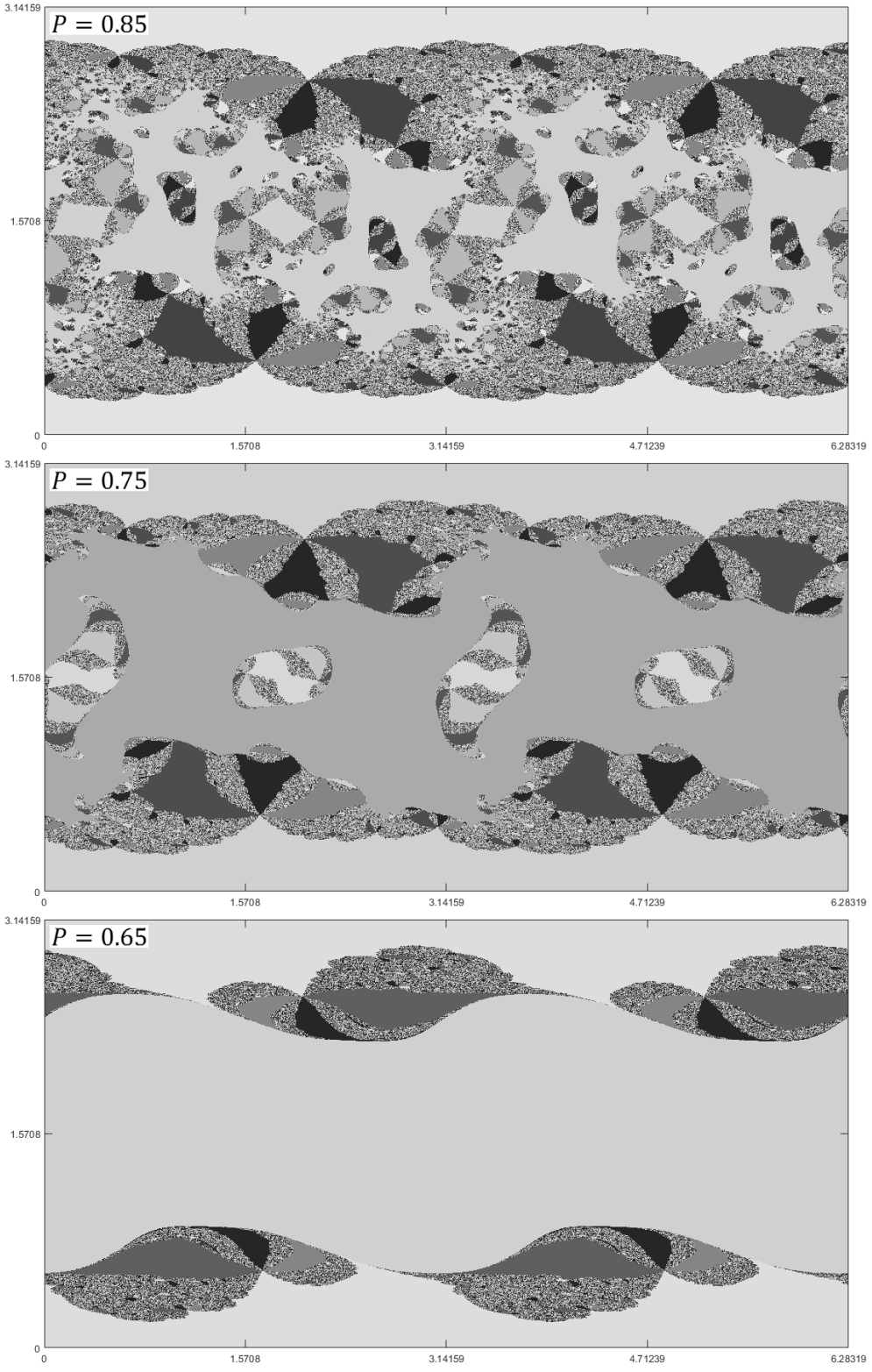


Figure 3: The fractal pattern of sensitive states decays when considering states with lower initial purity. Attractor maps for protocol with parameter chosen $\vartheta = 33^{\circ} = \frac{11}{60}\pi \doteq 0.576$.

An extremal case occurs for $\vartheta = \frac{\pi}{2}$ as each pure state is chaotic (Julia set of the evolution function 6 is the whole complex plane, see [12]) and each mixed state converges to the maximally mixed state.

The fractal pattern appearing in the mixed states consists of states that are sensitive to initial conditions. The sensitiveness comes from the fact that the fractal border separates regions of convergence to two different asymptotic regimes and the slightest perturbation can change the convergence. Therefore, these states cannot be in principle handled numerically - the finite precision of the computer can result into wrong estimation of asymptotic dynamics. In consequence, the other condition necessary to claim that the states are chaotic - topological transitivity - cannot be proven. An analytic approach is not known now. However, we can speak of quasi-chaotic behaviour as in [5].

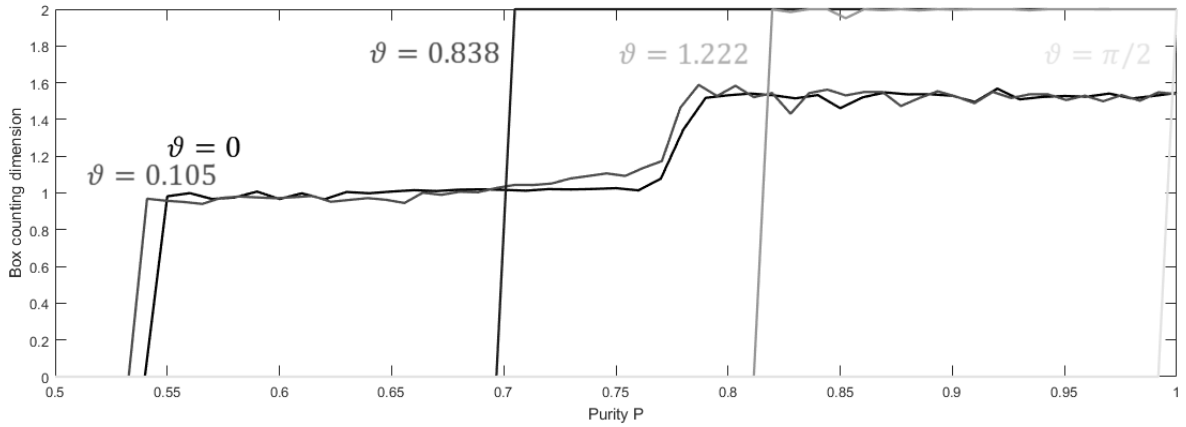


Figure 4: Dimension of the structure of sensitive states depending on the purity of the initial states. Transition is not unique for $\vartheta = 0$, examples of protocols with various ϑ demonstrate that the phase and the transition temperatures change with the twirling operator.

6 Conclusions

We have presented generalisations to a protocol that implements nonlinear evolution using measurement-based selection and modification. The protocol generalisation lies in using more general twirling gate. We have chosen a single parametric family of operator depending on angle ϑ.

Analytically, we have shown that the critical orbits start in $0, \infty$ and collide for all values ϑ. We have also found fixed states which are always repelling and period-2 cycle which may be attracting, parabolic or repelling depending on ϑ. The critical values where the nature of the cycle changes can be also traced form observing attractor maps, the fractal suddenly stretches through the complex plane.

Numerical methods suggest that the dynamics of mixed states is (compared to the pure states) enriched with additional attractor, the maximally mixed state $\rho = \frac{1}{2}\mathbb{1}$. This state attracts the more states the lower purity of the input states is. The fractal structure of sensitive states undergoes phase transition - on few examples we have shown that the dimension of the fractal changes suddenly when the control parameter - the purity -

is lowered. Due to the computational requirements we have not yet calculated phase diagram. This would probably visualise the relation of the transition temperatures and the actual phase (fractal dimension) depending on ϑ.

The phase diagram can be the goal for the next research as well as studying other generalisations of the protocol, namely hen other rotation matrices are used as twirling operators.

References

[1] H. Bechmann-Pasquinucci et. al. *Phys. Lett. A* **242** (1998), 198–204

[2] G. Alber, A. Delgado, N. Gisin, I. Jex. *J. Phys. A: Math. Gen.* **34** (2001), 8821

[3] J.W. Milnor. *Dynamics in One Complex Variable*, 3^{rd} edition (Princeton University Press) (2000)

[4] Kiss T., Vymětal S., Tóth L.D., Gábris A., Jex I., Alber G. *Phys. Rev. Lett.* **107**, 100501 (2011)

[5] M. Malachov M., I. Jex, O. Kálmán, T. Kiss. *ArXiv e-prints* 1809.00140 (2018)

[6] M. Malachov. *Chaotic Dynamics of Purification Protocols*, Masters' thesis (FJFI ČVUT v Praze) (2015)

[7] T. Kiss, I. Jex, G. Alber, S. Vymětal. *Phys. Rev. A* **74** (2006), 040301(R)

[8] J.E. Fornæss, B. Stensønes. *Lectures on Counterexamples in Several Complex Variables*, (AMS Chelsea Publishing, 2007).

[9] J.E. Fornæs. *Dynamics in Several Complex Variables* (Amer. Math. Soc.) (1994)

[10] S. Morosawa, Y. Nishimura, M. Taniguchi, T. Ueda. *Holomorphic Dynamics* (Cambridge University Press) (2000)

[11] K. Falconer. *Fractal Geometry; Mathematical Foundations and Applications*, 2^{nd} ed. (Willey & Sons) (2003)

[12] A. Gilyén et al. *Sci. Rep.***6** (2016), 20076

Fixed Points of Sturmian Morphisms and Their Derivated Words*

Kateřina Medková

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

advisors:

Edita Pelantová, Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Karel Klouda, Department of Applied Mathematics
Faculty of Information Technology, CTU in Prague

Abstract. Sturmian words are aperiodic infinite words over a binary alphabet having the least factor complexity possible, i.e. $\mathcal{C}(n) = n+1$ for each $n \in \mathbb{N}$. They are probably the most studied objects in combinatorics on words. In this article we study the derivated words to Sturmian words fixed by a primitive morphism.

Let $\mathbf{u} = u_0u_1u_2\cdots$ be a binary infinite word with $u_i \in \{0,1\}$ and let $w = u_iu_{i+1}\cdots u_{i+n-1}$ be its factor. The integer i is called an occurrence of the factor w in $\mathbf{u}$. A return word to a factor w is a word $u_iu_{i+1}\cdots u_{j-1}$ with $i < j$ being two consecutive occurrences of w in $\mathbf{u}$. Any infinite uniformly recurrent word $\mathbf{u}$ can be written as a concatenation of a finite number of return words to a chosen prefix w of $\mathbf{u}$. The ordering of the return words in this concatenation is coded by the derivated word to $\mathbf{u}$ and its prefix w. In 1998, Durand proved that a fixed point $\mathbf{u}$ of a primitive morphism has only finitely many distinct derivated words and each derivated word is fixed by a primitive morphism as well. Our aim is to follow this result and describe in detail the set of derivated words to Sturmian words.

Let $\mathbf{u}$ be a Sturmian word which is a fixed point of a Sturmian morphism ψ. Any primitive Sturmian morphism may be decomposed using the elementary Sturmian morphisms. Using this decomposition we derive the main result of our article, which is an exact description of the morphisms fixing the derivated words of $\mathbf{u}$. More precisely, we provide an algorithm which to a given Sturmian morphism ψ lists the morphisms fixing the derivated words of the fixed point of ψ. We continue our study by counting the number of derivated words. We provide the exact number of derivated words for two specific classes of Sturmian morphisms. For a general Sturmian morphism ψ, we give a sharp upper bound on their number.

Keywords: derivated word, return word, sturmian morphism, sturmian word

Abstrakt. Sturmovská slova jsou aperiodická nekonečná slova nad binární abecedou, která mají nejnižší možnou faktorovou komplexitu, tedy komplexitu $\mathcal{C}(n) = n + 1$ pro každé $n \in \mathbb{N}$. V kombinatorice na slovech patří mezi vůbec nejstudovanější objekty. V tomto článku se věnujeme studiu derivovaných slov ke sturmovským slovům, která jsou pevnými body primitivních morfismů.

*This work has been supported by the Ministry of Education, Youth and Sports of the Czech Republic project no. CZ.02.1.01/0.0/0.0/16_019/0000778 and by the Czech Technical University in Prague grant SGS17/193/OHK4/3T/14.

Nechť $\mathbf{u} = u_0u_1u_2\cdots$ je binární nekonečné slovo s písmeny $u_i \in \{0,1\}$ a nechť $w = u_iu_{i+1}\cdots u_{i+n-1}$ je jeho faktor. Přirozené číslo i nazýváme výskytem faktoru w v $\mathbf{u}$. Návratové slovo k faktoru w je slovo $u_iu_{i+1}\cdots u_{j-1}$, kde $i < j$ jsou dva po sobě jsoucí výskyty w v $\mathbf{u}$. Každé uniformně rekurentní nekonečné slovo $\mathbf{u}$ můžeme zapsat jako zřetězení návratových slov ke zvolenému prefixu w slova $\mathbf{u}$. Pořadí jednotlivých návratových slov v tomto zřetězení je kódováno derivovaným slovem ke slovu $\mathbf{u}$ a jeho prefixu w. V roce 1998 Durand ukázal, že pevný bod $\mathbf{u}$ primitivního morfismu má vždy pouze konečný počet různých derivovaných slov a že tato derivovaná slova jsou také pevnými body primitivních morfismů. Naším cílem je navázat na tento výsledek a detailněji popsat množinu derivovaných slov ke sturmovským slovům.

Nechť $\mathbf{u}$ je sturmovské slovo, které je pevným bodem sturmovského morfismu ψ. Každý primitivní sturmovský morfismus můžeme rozložit na tzv. elementární sturmovské morfismy. Vhodné využití tohoto rozkladu je klíčovým krokem k odvození hlavního výsledku tohoto článku, což je přesný popis morfismů fixujících derivovaná slova k $\mathbf{u}$. Konstruujeme algoritmus, který k danému sturmovskému morfismu ψ najde morfismy, které fixují derivovaná slova k pevnému bodu morfismu ψ. Dále pokračujeme vyčíslením počtu derivovaných slov. Pro dvě speciální třídy sturmovských morfismů určujeme přesný počet derivovaných slov. Pro obecný sturmovský morfismus pak stanovujeme horní odhad jejich počtu a ukazujeme, že se v některých případech tohoto odhadu nabývá.

Klíčová slova: derivované slovo, návratové slovo, sturmovský morfismus, sturmovské slovo

Full paper: K. Klouda, K. Medková, E. Pelantová and Š. Starosta. *Fixed points of Sturmian morphisms and their derivated words.* Theoret. Comput. Sci. **743** (2018), 23–37.

Properties of Curvature Flow in Codimension Two*

Jiří Minarčík

1st year of PGS, email: minarji2@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 aim of this work is to analyze the differences between the curvature flow of planar and space curves. First, we investigate the curvature driven flow in the direction of the principal normal vector. We compare theoretical properties of this motion to the classical two-dimensional case and analyze the motion of special classes of space curves. We present several new theoretical results including the Shrinking ball theorem which states that curves under the curvature flow cannot cross certain family of shrinking spheres. We also studied the binormal flow and its application in vortex dynamics. Finally, we examine numerical solution to the geometric flow problems and present selected computational results.

Keywords: curvature flow, binormal flow, parametric approach, vortex dynamics

Abstrakt. Cílem této práce je srovnání pohybu křivek v rovině a třírozměrném prostoru. Nejprve je zkoumán pohyb ve směru normálového vektoru s rychlostí rovnou křivosti v daném bodě. Vlastnosti takto zadaného pohybu jsou porovnány se standardní verzí tohoto problému, ve které se křivky pohybují v rovině. Práce obsahuje konkrétní příklady demonstrující důležité rozdíly a novou třídu prostorových křivek, které se svým chováním podobají křivkám v rovině. Dále jsou představeny nové teoretické výsledky umožňující pohyb křivek prostorově omezit. Dalším zkoumaných problémem je pohyb ve směru binormály a jeho použití v oblasti dynamiky vírových smyček. Práce je zakončena rozborem numerického řešení zmíněných problémů a přehledem vybraných výpočetních výsledků.

Klíčová slova: prostorové křivky, parametrická formulace, geometrický tok, pohyb vírů

References

[1] J. Minarčík, M. Kimura, M. Beneš. *Comparing Motion of Curves and Hypersurfaces in* $\mathbb{R}^m$. Submitted to Discrete and Continuous Dynamical Systems Series B, 2018.

*This work has been supported by the grant Application of advanced supercomputing methods for mathematical modeling of natural processes, project of the Student Grant Agency of the Czech Technical University in Prague No. SGS17/194/OHK4/3T/14 2017-19

Surrogate Model Selection in Combination with the CMA-ES*

Zbyněk Pitra†

5th year of PGS, email: z.pitra@gmail.com
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Martin Holeňa, Department of Machine Learning
Institute of Computer Science, CAS

Abstract. Many real-world problems belong to the area of continuous black-box optimization, where evolutionary optimizers have become very popular inspite of the fact that such optimizers require a great amount of real-world fitness function evaluations, which can be very expensive or time-consuming. Hence, regression surrogate models are often utilized to evaluate some points instead of the fitness function. The Doubly Trained Surrogate Covariance Matrix Adaptation Evolution Strategy (DTS-CMA-ES) [4] is a surrogate-assisted version of the state-of-the-art continuous black-box optimizer CMA-ES [2] using Gausssian processes (GP) as a surrogate model to predict the whole distribution of the fitness function. In [5], the DTS-CMA-ES is studied in connection with the boosted regression forest, another regression model capable to estimate the distribution. The reported results suggest that convenient choice of the surrogate model plays an important role. Therefore, in [6], the surrogate model selection using previously obtained knowledge through Exploratory landscape analysis [3] is studied on random forests and GP. In addition, selection of a covariance function defining the main character of GP in connection with the CMA-ES was investigated [7, 8]. The main results of research into reducing the number of fitness evaluations of the CMA-ES using GP were summarized in [1].

Keywords: black-box optimization, surrogate model, Gaussian process, random forest, metalearning

Abstrakt. Mnoho praktických úloh spadá do oblasti spojité black-box optimalizace, kde se evoluční algoritmy staly velmi populární navzdory faktu, že vyžadují velké množství vyhodnocení skutečné fitness funkce, což může být velice drahé nebo časové náročné. Z tohoto důvodu jsou regresní náhradní modely často využívány k ohodnocení některých bodů namísto původní fitness funkce. Doubly Trained Surrogate Covariance Matrix Adaptation Evolution Strategy (DTS-CMA-ES) [4] je verzí v současnosti nejlepšího algoritmu pro spojitou optimalizaci CMA-ES [2], která využívá gaussovské procesy jako náhradní model k odhadům rozdělení celé fitness funkce. Ve článku [5] je DTS-CMA-ES zkoumán v kombinaci s boostovanými regresními lesy, které jsou dalším regresním modelem schopným odhadovat rozdělení funkce. Uvedené výsledky naznačují, že vhodná volba náhradního modelu hraje důležitou roli. Z tohoto důvodu je ve článku [6] zkoumán problém výběru náhradního modelu na základě předchozích znalostí pomocí metody Exploratory landscape analysis [3] na náhodných lesech a gaussovských procesech. Dále byl výzkum zaměřen na výběr kovarianční funkce gaussovského procesu, která jej ve velké míře charakterizuje, ve spojení s algoritmem CMA-ES [7, 8]. Nejdůležitější výsledky výzkumu v

*The reported research was supported by the Czech Science Foundation grant No. 17-01251, by the Grant Agency of the Czech Technical University in Prague with its grant No. SGS17/193/OHK4/3T/14.

†This study has been provided in cooperation with Lukáš Bajer and Jakub Repický.

oblasti snížení počtu vyhodnocení fitness funkce algoritmem CMA-ES pomocí gaussovských procesů byli shrnuty v [1].

Klíčová slova: black-box optimalizace, náhradní modelování, gaussovské procesy, náhodné lesy, metaučení

References

[1] L. Bajer, Z. Pitra, J. Repický, and M. Holeňa. *Gaussian process surrogate models for the CMA evolution strategy.* Evolutionary Computation (to be released), 30.

[2] N. Hansen. *The CMA Evolution Strategy: A Comparing Review.* In 'Towards a New Evolutionary Computation', J. A. Lozano, P. Larrañaga, I. Inza, and E. Bengoetxea, (eds.), number 192 in Studies in Fuzziness and Soft Computing, Springer Berlin Heidelberg (January 2006), 75–102.

[3] O. Mersmann, B. Bischl, H. Trautmann, M. Preuss, C. Weihs, and G. Rudolph. Exploratory landscape analysis. GECCO '11, 829–836. ACM, (2011).

[4] Z. Pitra, L. Bajer, and M. Holeňa. *Doubly Trained Evolution Control for the Surrogate CMA-ES*, 59–68. Springer International Publishing, Cham, (2016).

[5] Z. Pitra, J. Repický, and M. Holeňa. Boosted regression forest for the doubly trained surrogate covariance matrix adaptation evolution strategy. In 'ITAT 2018 Proceedings', S. Krajči, (ed.), volume 2203, 72–79. CEUR Workshop Proceedings, (September 2018).

[6] Z. Pitra, L. Bajer, J. Repický, and M. Holeňa. Transfer of knowledge for surrogate model selection in cost-aware optimization. In 'Workshop on Interactive Adaptive Learning Proceedings', G. Krempl, (ed.), ECML PKDD 2018, 89–94, (2018).

[7] J. Repický, Z. Pitra, and M. Holeňa. Automated selection of covariance function for gaussian process surrogate models. In 'ITAT 2018 Proceedings', S. Krajči, (ed.), volume 2203, 64–71. CEUR Workshop Proceedings, (September 2018).

[8] J. Repický, Z. Pitra, and M. Holeňa. Adaptive selection of gaussian process model for active learning in expensive optimization. In 'Workshop on Interactive Adaptive Learning Proceedings', G. Krempl, (ed.), ECML PKDD 2018, 80–84, (2018).

Revisiting Transitions between Superstatistics*

Martin Prokš

2nd year of PGS, email: proksma6@fjfi.cvut.cz
Department of Physics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Petr Jizba, Department of Physics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. This work aims to provide an accurate method for a detection of a transition between Superstatistics. A slight improvement over the currently published method is achieved. Superstatistics framework is briefly recalled and a rather new concept of transition of Superstatistics, introduced by Beck and Xu in 2016, is reexamined. In addition, an original synthetic model for Superstatistical transition suggested by Beck is discussed. It is shown that its modified version which takes into account a stochastic nature of the transition better reflects empirically observed transitions.

Keywords: Superstatistics, Transition of Superstatistics, Monte Carlo simulation, time series

Abstrakt. Tato práce má za cíl představit přesnou metodu pro detekci přechodu mezi Superstatistikami. Podařilo se docílit zlepšení oproti nedávno publikované metodě. Superstatistika je krátce připomenuta a následně je přezkoumán relativně nový pojem transmutace mezi Superstatistikami. Tento jev byl představen Ch. Beckem and Xu roku 2016. Navíc původní model, navržený Beckem, zachycující přechod mezi Superstatistikami je podrobněji diskutován. Je ukázáno, že jeho modifikovaná verze, která bere v potaz stochastickou povahu přechodu, lépe reflektuje empiricky pozorované přechody.

Klíčová slova: Superstatistika, Přechod mezi Superstatistikami, Monte Carlo simulace, časové řady

1 Introduction

Superstatistics is a well known term in a field of non-equilibrium statistical physics. It describes a system in a local thermodynamic equilibrium. However, only recently a new spin in a Superstatistical paradigm has been introduced. Namely a transition of Superstatistics [4]. Its application is predominantly in time series analysis. The basic premise is that the Superstatistical smearing distribution may change on different time scales. The first experimental evidence for this phenomenon was introduced by Beck and Xu in [4]. The pioneering paper was followed by our paper [5] which introduced a more reliable method for detecting the alleged transition. The method was based on leveraging statistical distances in order to decide a favorable probability distribution on various time scales. Nevertheless, there were doubts about a significance level. Therefore, in this paper we revisit the procedure and by using Monte Carlo method, we provide a probability of

*This work has been supported by the Grant Agency of the Czech Technical University in Prague, grant No. SGS16/239/OHK4/3T/14 and by Czech Science Foundation Grant No. 17-33812L.

successful determination of a correct smearing distribution. By doing so, we can asses a significance level to the transition.

2 From Superstatistic to Transition between Superstatistics

Superstatistics is a concept devised by Beck for systems with fluctuating intensive parameter, e.g. temperature, which are therefore in non-equilibrium state. The idea first appeared in the paper [1], and the term Superstatistics was coined later in the successive paper [3].

The assumption is that the system is in non-equilibrium steady state and is composed of many cells which are locally in equilibrium but with different values of intensive parameter, e.g. temperature. This intensive parameter in each cell changes on a long time scale much larger than a relaxation time of the cell.

As a Superstatistics is meant the generalized Boltzmann factor which describes the whole system composed of small subsystems in local equilibrium.

$$B(E) = \int_0^{+\infty} f(\beta) e^{-\beta E} \mathrm{d}\beta, \tag{1}$$

where $f(\beta)$ is a smearing distribution.

In general, the only restriction on $f(\beta)$ is to allow only positive values for β. However, experince showed that three distributions fit various empirical data especially well. Therefore, Beck in [2] suggested so called universality classes. It contains *Gamma distribution*

$$f(\beta) = \frac{1}{\Gamma(\frac{n}{2})} \left(\frac{n}{2\beta_0}\right)^{\frac{n}{2}} \beta^{\frac{n}{2}-1} \exp\left(-\frac{n\beta}{2\beta_0}\right), \tag{2}$$

Log-normal distribution

$$f(\beta) = \frac{1}{\beta\sqrt{2\pi s^2}} \exp\left(-\frac{(\log\beta - \log\mu)^2}{2s^2}\right) \tag{3}$$

and *Inverse* χ^2 *distribution*. The last distribution is disregarded here because it was shown in [5] that it gives poor results for date at our disposal.

Superstatistics is definitely a great idea which has justifiable motivation, and furthermore, it has been successfully fitted on empirical data. Therefore, there is little doubt about usefulness of Superstatistics, however, a new broader model was recently introduced by father of Superstatistics in [4]. It is claimed that transition of Superstatistics is possible when we look at the time series at different time scales.

By transition between Superstatistics is then meant a change from one smearing distribution to another. For example, in a financial time series log-returns may be well described by Log-normal distribution at a small (minute) time scale, however, if one moves to a higher time scales (hours) a better description may be using a Gamma distribution. This kind of Superstatistical transition is examined in the next section.

3 Transition confirmation

Originally in the pioneering paper [4], the transition was assessed only by looking at histogram of β at two time scales (minutes and days). Unfortunately, it is impossible to really detect the transition merely from histogram. Therefore, in [5] a method based on statistical distances was employed. It allowed to see in a quantitative way a change from one Superstatistics to another. However, this still lacked a level of significance because the fact that statistical distance is smaller for one distribution than for another may be just a manifestation of a random error. A slight improvement described in the followings attempts to address this issue.

The main difference from the method in [5] is that we try to assign a probability distribution to each time scale according to all three statistical distances. Kolmogorov-Smirnov distance

$$D_n = \sup_x |F_n(x) - F(x)|, \tag{4}$$

Cramér-von Mises distance

$$C_n = n \int_{-\infty}^{+\infty} (F_n(x) - F(x))^2 \mathrm{d}F(x) \tag{5}$$

and Anderson–Darling distance

$$A_n = n \int_{-\infty}^{+\infty} \frac{(F_n(x) - F(x))^2}{F(x)(1 - F(x))} \mathrm{d}F(x), \tag{6}$$

where $F(x)$ is fully specified distribution function and $F_n(x)$ is empirical distribution function

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(u_i \leq x). \tag{7}$$

As a correct distribution at a given time scale is considered a distribution which is favorable by at least two distances. Then, the ideal procedure would be to obtain a probability of successfully discriminating between the two distributions for each statistical distances and hence assigning a value of significance for chosen distribution at each time scale. This would allow us to claim that at time scale e.g. 20 *min* the smearing distribution is highly probably better described by Log-normal distribution while at time scale 300 *min* it is very likely a Gamma distribution. Hence the transition may be considered as a significant.

Unfortunately, statistical properties of the distances cannot be obtained if parameters are estimated, the only possibility is to use Monte Carlo simulations to determine a probability of successfully detected transition. The same procedure is used in a Liliford test which uses Kolmogorov-Smirnov distance, see [6].

Appart from Lilliefors test, the method of recognizing probability distribution was inspired by [7], where only Kolmogorov-Smirnov distance was used to discriminate between two-parametric distribution families. It was shown that the distance measure provides a reliable discriminating criteria.

The Monte Carlo simulation is conducted as follows:

1. Estimate parameters for Gamma and Log-normal distribution from data by Maximum-Likelihood method.

2. Generate a random sample of the size available for a particular time scale and company from Gamma or Log-normal distribution (using parameters from the previous step) depending on a distribution favorable by probability distances.

3. For both Gamma and Log-normal distribution calculate probability distances from the empirical distribution (estimated from the generated sample).

4. Use the decision criteria for choosing Gamma or Log-normal distribution and mark the trial as successful if the distribution matches the one generated in the step 2.

5. Repeat the steps $2-4$ 10^5 times and estimate the probability of successfully selecting the probability distribution by relative frequencies.

The dataset used for testing this method is the same as in [4], i.e. stock prices of seven US companies from different sectors recorded on the minute-tick basis during a period from the $2nd$ Jan, 1998, to $22nd$ May, 2013. The output of the simulation is depicted in figure 4. It confirms the conclusion from [5] about a transition for company Alcola Inc. and Wal-Mart Stories Inc. Moreover, it is seen that the time series for the company Bank of America indeed do not exhibit transition of Superstatistics. The key point to notice is a relatively high probability of successfully discriminating the two distributions. Therefore, it may be concluded that the transition, especially for AA company, is not smooth but oscillates between the two distributions around the transition point. This statistically significant observation is examined in the next section.

4 Beck's Transition Model

In the original mention of Superstatistical transition [4], Beck and Xu suggested a so called *synthetic model*

$$\beta_\tau = \kappa_\tau L_{\tau_0} + \big(1 - \kappa_\tau\big) G_{\tau_\infty}, \tag{8}$$

where L_{τ_0} and G_{τ_∞} are two random variables with Log-normal and Gamma distribution respectively. The suffixes τ_0 and τ_∞ denotes small and large time scales respectively where the distribution of β is Log-normal and Gamma. For data at hand $\tau_0 = 20\ min$ and $\tau_\infty \approx 500\ min$. L_{τ_0} and G_{τ_∞} may be thought of as an asymptotic distributions. The parameter $\kappa \in \langle 0, 1\rangle$ is a function of a time scale τ and is responsible for a smooth transition from a region dominated by Log-normal distribution to one with Gamma distribution on larger time scales. A reasonable functional form for κ which may reproduce the observed transition is

$$\kappa(\tau) = \frac{1}{2}\Big(tanh\big(a(\tau - b)\big) + 1\Big). \tag{9}$$

The parameter a controls sharpness of a transition and b selects a time scale at which the transition occurs. See figure 1 for demonstration of the sharpness parameter.

However, as Monte Carlo simulations shows a deterministic increasing function of a time scale does not reflect the observed rough evolution of the transition. (Compare

figure 2 where a transition generated from the synthetic model 8 with κ given by eq. 9 is depicted and figure 4 showing the observed transitions for companies Alcola Inc. (AA) and Wal-Mart Stories Inc. (WMT)).

The explanation for this discrepancy is rather simple. Since the distance measures between probability distributions used for discriminating the two regions have a significant statistical power (due to a large sample size especially at small time scales), it will at a certain level of κ (likely $\kappa \approx \frac{1}{2}$) flip from Log-normal distribution to Gamma distribution and never oscillate between those two states as seen in figure 2.

In this paper we propose a model which better captures an observed behavior of real transitions. As can be seen from fig 4, the transitions posses a stochastic nature. For example, time series for Wal-Mart Stores Inc. the transition from Log-normal to Gamma region occurs around a time scale of 60 *min*. Nevertheless, an quick unpredictable transitions happens much sooner and also on higher time scale occasional flip to Log-normal region is observed. The stochastic nature is more pronounced for Alcola Inc. company, where the transition again occurs around a time scale 60 *min*, but unlike for WMT, it is very slow (corresponds to a small sharpness parameter in 9). Even at $\tau \geq 300$ *min* an occasional flip back to Log-normal distribution is observed.

The suggested modification incorporating a random element into the model is to consider κ as a random variable. (Strictly speaking a stochastic process since κ is parametrized by a time scale τ.) κ is a parameter in $\langle 0, 1 \rangle$, therefore it is necessary to take a probability distribution with a compact support. A well known distribution with this property is a Beta distribution

$$p(x) = \frac{x^{\gamma-1}(1-x)^{\delta-1}}{B(\gamma,\delta)}, \qquad \gamma, \delta > 0. \tag{10}$$

The original parametrization is not the best choice, therefore, an alternative one is used which contains the mode of the distribution μ and so called concentration ν

$$\begin{aligned} \gamma &= \mu(\nu-2)+1, \\ \delta &= (1-\mu)(\nu-2)+1. \end{aligned}$$

The concentration together with sharpness parameter in 9 controls how centered the transition is and should complement each other, i.e. they should act as a one degree of freedom. The mode parameter μ is then time scale dependent and its functional form is given by eq. 9. See figure 3 showing a shape of the probability distribution for various time scales. For $\tau \approx \tau_0$ there is a low probability for κ to leave a Log-normal region while for $\tau \approx \tau_\infty$ κ predominantly stays in a Gamma region.

It should be noted that even though the model reflects well the empirical transition, so far there have not been found a suitable estimator for corresponding parameters in the model.

5 Conclusion

The Beck's synthetic model for Superstatistical transition was revisited. It was shown that its modified version, which involves a random element, is able to correctly reproduce

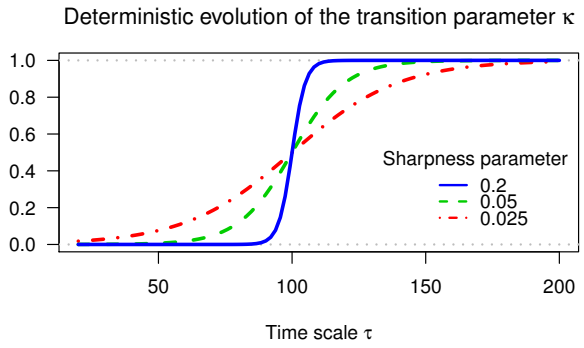


Figure 1: Deterministic evolution of the transition parameter

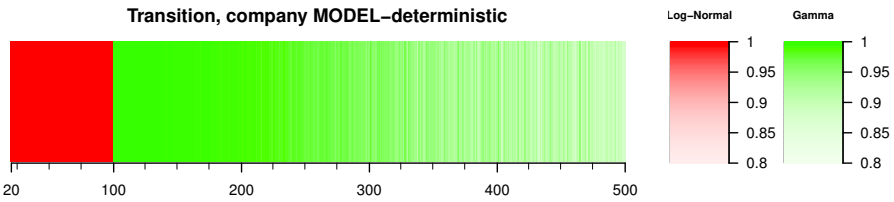


Figure 2: Transition for the deterministic model

observed transitions and therefore may serve as a suitable model. The modification is done by incorporating a random element into the transition parameter κ. Namely, κ is considered to be a stochastic process (parametrized by time scale) with Beta distribution. Moreover, a better method for assessing transition of Superstatistics was provided which assigns a probability of successfully discriminating between two Superstatistical regions. Those probabilities needs to be obtained by Monte Carlo simulations.

References

[1] C. Beck. Dynamical foundatiouns of nonextensive statistical mechanics. *Phys. Rev. Lett.*, 87, 2001.

[2] C. Beck. Recent developments in superstatistics. *Brazilian Journal of Physics*, Vol. 39, no. 2A, 2009.

[3] C. Beck and E.G.D. Cohen. Superstatistics. *Physica A*, 322:267, 2003.

[4] C. Beck and D. Xu. Transition from lognormal to χ^2 -superstatistics for financial time series. *Physica A*, 453:173–183, 2016.

[5] P. Jizba, J. Korbel, H. Lavička, M. Prokš, V. Svoboda, and C. Beck. Transitions between superstatistical regimes: Validity, breakdown and applications. *Physica A*, 493:29–46, 2018.

[6] H. Lilliefors. On the kolmogorov–smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, Vol. 62:399–402, 1967.

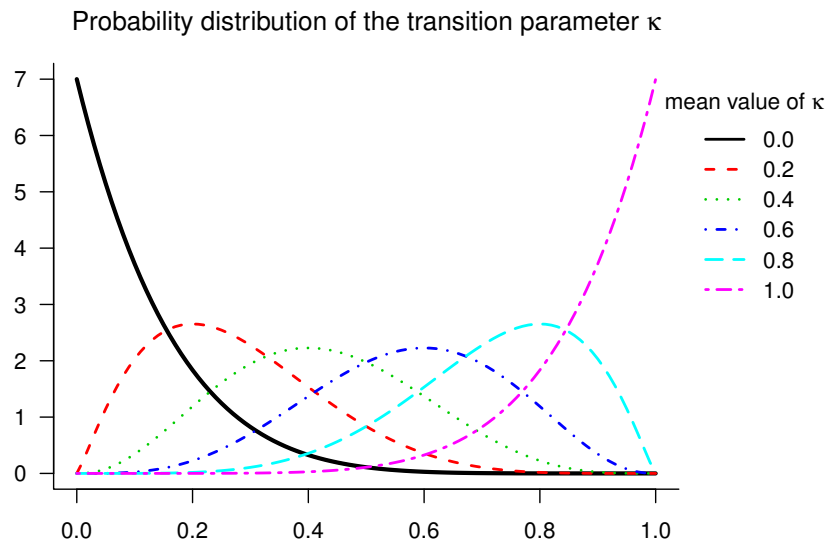


Figure 3: Probability distribution of the transition parameter

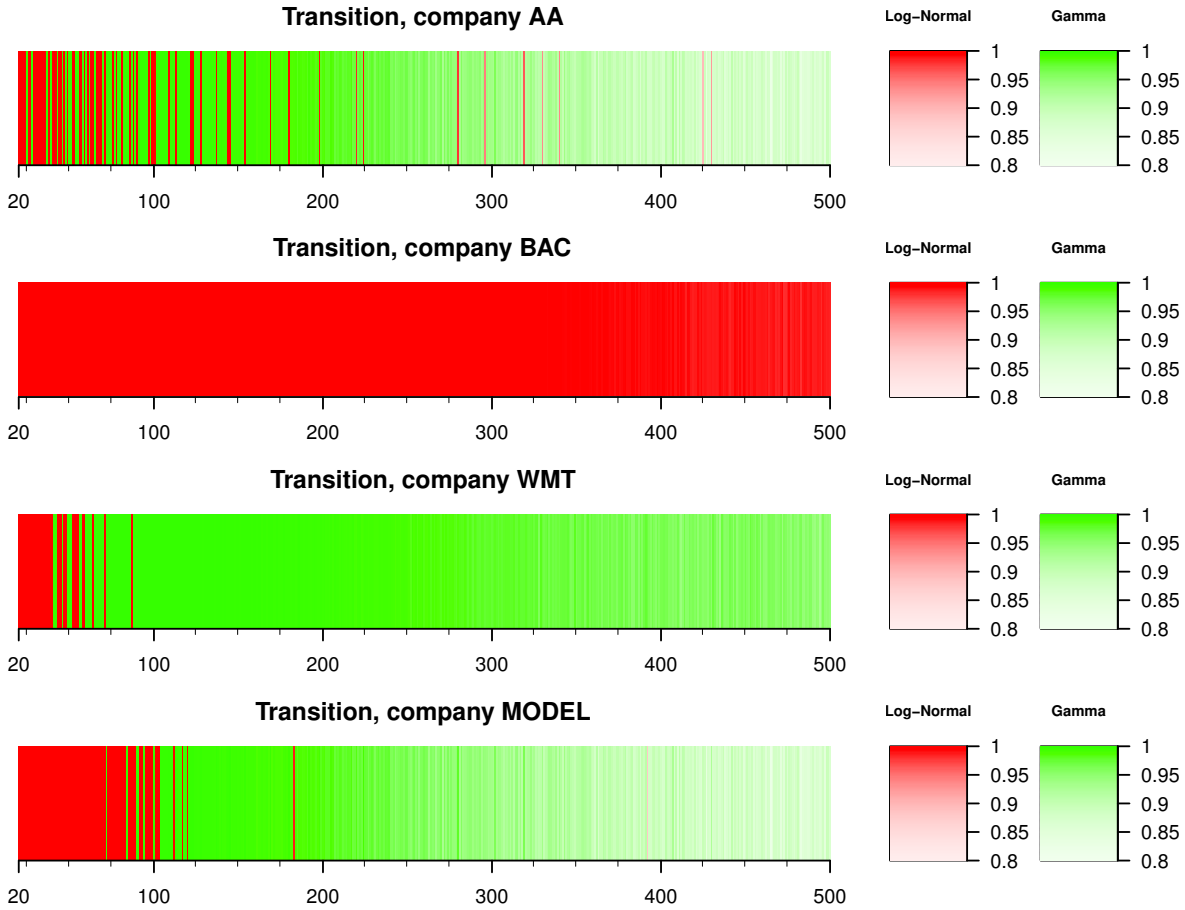


Figure 4: Transitions for companies Alcola Inc. (AA), Bank of America Corporation (BAC), Wal-Mart Stories Inc. (WMT) and the synthetic model incorporating randomness.

[7] A. W. Marshall, J.C. Meza, and I. Olkin. Can data recognize its parent distribution? *Journal of Computational and Graphical Statistics*, 10:555–580, 2001.

Heart Attack Mortality Prediction: An Application of Machine Learning Methods*

Issam Salman

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

advisor: Jiří Vomlel, Department of Decision-Making Theory
Institute of Information Theory and Automation, CAS

Abstract. The heart is an important organ in the human body, and Acute Myocardial Infarction (AMI) is the leading cause of death in most countries. Researchers are diverting a lot of data analysis work to assist doctors in predicting the heart problem. An analysis of the data related to different health problems and its functions can help in predicting with a degree of certainty the wellness of this organ. Our research reported in this paper is twofold. In the first part of the paper, we compare different predictive models of hospital mortality for patients with AMI. All results presented in this part are based on real data of about 603 patients from a hospital in Czechia and about 184 patients from two hospitals in Syria. Although the learned models may be specific to the data, we also draw more general conclusions that we believe are generally valid. In the second part of the paper, because the data is incomplete and imbalanced we develop the Chow-Liu and tree-augmented naive Bayesian (TAN) to deal with that data in better conditions, and measure the quality of these algorithms with other algorithms.

Abstrakt. Srdce je důležitým orgánem v lidském těle, a akutní infarkt myokardu (AMI) je hlavní příčinou úmrtí ve většině zemí. Výzkumníci směřují spoustu práce s analýzou dat pomáhat lékařům při předvídání problému se srdcem. Analýza dat souvisejících s různými zdravotní problémy a jeho funkce mohou pomáhat předvídat s jistotou wellness tohoto orgánu. Náš výzkum uvedený v tomto článku je dvojí. V první části práce porovnáváme různé prediktivní modely nemocniční úmrtnosti u pacientů s AMI. Všechny výsledky uvedené v této části jsou založeny na reálných datech přibližně 603 pacientů z nemocnice v ČR a přibližně 184 pacientů ze dvou nemocnic v Sýrii. I když naučené modely mohou být specifické pro data, vyvozujeme také obecnější závěry, o kterých se domníváme, že jsou obecně platné. V druhé části práce, protože data jsou neúplná a nevyrovnaná, vyvíjíme Chow-Liu a naivně bayesovský (TAN) augmentovaný strom, abychom se s těmito daty lépe zabývali a měřili kvalitu těchto algoritmů s jinými algoritmy.

Keywords: Machine Learning, Data mining, Classification, Bayesian networks, Acute Myocardial Infarction

*This work has been supported by the SGS grant CTU SGS16/253/OHK3/3T/14.

1 Introduction

An enormous amount of data is being generated every day. Analysing big data sets is impossible without the help of automated procedures. Machine learning [1] provides these procedures. The most commonly used form of machine learning is supervised classification [2]. Its goal is to learn a mapping from the descriptive features of an object to the set of possible classes, given a set of features-class pairs.

Probabilities play a central role in modern machine learning [12]. Probabilistic graphical models (PGMs) [15] have emerged as a general framework for describing and applying probabilistic models. A PGM allows us to efficiently encode a joint distribution over some random variables by making assumptions of conditional independence.

A Bayesian network classifier (BNC) [7] is a Bayesian network applied to the classification task. BNCs have many strengths, including: good interpretability, the possibility of including prior knowledge about a domain, and competitive predictive performance. They have been successfully applied in practice, e.g. [13], [8], and [11].

Acute myocardial infarction (AMI) is commonly known as a heart attack. A heart attack occurs when an artery leading to the heart becomes completely blocked and the heart doesn't get enough blood or oxygen. Without oxygen, cells in that area of the heart die. AMI is responsible for more than a half of deaths in most countries worldwide. Its treatment has a significant socioeconomic impact.

One of the main objectives of our research is to design, analyze, and verify a predictive model of hospital mortality based on clinical data about patients. A model that predicts mortality well can be used, for example, for the evaluation of medical care in different hospitals. The evaluation based on mere mortality would not be fair to hospitals that treat complicated cases often. It seems better to measure the quality of the health care using the difference between predicted and observed mortality.

A related work was published by [3], the authors analyze the mortality data in U.S. hospitals using the logistic regression model. In another work by [4], the authors design and verify a predictive model of hospital mortality in ST Elevation Myocardial Infarction (STEMI). In other work by [14], the authors analyze the medical records of patients suffering Myocardial infarction from a third world country - Syria - and a developed country - Czechia - and present an idea of how to deal with incomplete and imbalanced data for Tree Augmented NB (TAN).

2 Data

Our data-set contains data from 787 patients from 2 different countries (603 patients from Czechia and 184 are from Syria) characterized by 24 variables. The attributes are listed in the Table 1. Most records contain missing values, i.e., for most patients only some attribute values are available, and some attributes are not available for Syrian patients, i.e. the data is incomplete. The thirty-days mortality is recorded for all patients. There are 89% of the patients survived, i.e. the data is imbalanced.

In Czechia, the results of blood tests are reported in millimoles per liter of blood. In Syria some of the measurements are reported in milligrams per liter and some in millimoles per liter. We standardized all measurements to the millimoles per liter scale.

Attribute	**Code**	**type**	**value range in data**	**Country**
Age	AGE	real	[23, 94]	SYR, CZ
Height	HT	real	[145, 205]	CZ
Weight	WT	real	[35, 150]	CZ
Body Mass Index	BMI	real	[16.65, 48.98]	CZ
Gender	SEX	nominal	{male, female}	SYR, CZ
Nationality	NAT	nominal	{Czech, Syrian}	SYR, CZ
STEMI Location	STEMI	nominal	{inferior, anterior, lateral}	SYR, CZ
Hospital	Hospital	nominal	{CZ, SYR1, SYR2}	SYR, CZ
Kalium	K	real	[2.25, 7.07]	CZ
Urea	UR	real	[1.6, 61]	SYR, CZ
Kreatinin	KREA	real	[17, 525]	SYR, CZ
Uric acid	KM	real	[97, 935]	SYR, CZ
Albumin	ALB	real	[16, 60]	SYR, CZ
HDL Cholesterol	HDLC	real	[0.38, 2.92]	SYR, CZ
Cholesterol	CH	real	[1.8, 9.9]	SYR, CZ
Triacylglycerol	TAG	real	[0.31, 11.9]	SYR, CZ
LDL Cholesterol	LDLC	real	[0.261, 7.79]	SYR, CZ
Glucose	GLU	real	[2.77, 25.7]	SYR, CZ
C-reactive protein	CRP	real	[0.3, 359]	SYR, CZ
Cystatin C	CYSC	real	[0.2, 5.22]	SYR, CZ
N-terminal prohormone of brain natriuretic peptide	NTBNP	real	[22.2, 35000]	CZ
Troponin	TRPT	real	[0, 25]	CZ
Glomerular filtration rate (based on MDRD)	GFMD	real	[0.13, 7.31]	CZ
Glomerular filtration rate (based on Cystatin C)	GFCD	real	[0.09, 7.17]	CZ

Table 1: Attributes

3 Machine Learning Methods

Since the explanatory variables may combine their influence and the influence of a variable may be mediated by another variable, it is worth studying the relations of variables altogether. We will do it in two steps: (1) since the mortality prediction is of our primary interest, we will compare how different classifiers are able to predict mortality, (2) to get an overall picture of the relations between all variables, we will learn some of Bayesian network models from the collected data, (3)to handle incomplete and imbalanced data, we provide an idea of how to develop the Chow-Liu [17] and tree-augmented naive Bayesian (TAN) algorithms [7] to be able to process this data.

We will work with different versions of data. They depend on how we treat variables that have more than two states: (1) real valued ordinal variables, (2) discrete valued variables (with at most five states), and (3) binary variables. We will discuss the values' transformation in more detail in the next sections.

3.1 Ordinal attributes

In our data, we have several categorical variables (sometimes also called nominal variables). These are variables that have two or more categories. For example, gender is a categorical variable having two categories (male and female). However, for some machine learning methods we need ordinal attributes which are attributes whose values have an ordering of values that is natural for the quantification of their impact on the class. This is satisfied by all attributes that can take only two values – even if they are nominal, e.g. by gender (0 for male, 1 for female), mortality (0 for survived, 1 for died). In our data it seems that the ordinality can be assumed for most real valued attributes, but note that there might also exist laboratory tests whose values deviate from a normal range in both directions (i.e. both lower and higher values) may both increase the mortality. We will refer to the ordinal data as D.ORD.

3.2 Discrete attributes

A discrete variable is a variable that can take values from a finite set. It is common to work with discrete variables since that simplifies calculations and some classification methods require them. we will used NA in the results of that classification methods. To get statistically reliable estimates of model parameters it is advisable to keep the number of values as low as possible while still being able to express the significant relations. We performed discretization of all real-valued attributes. It is not easy to find the optimum number and the values of split points in discretization. Fortunately, there exists the Czech National Code Book that classifies numeric laboratory results, with respect to age and gender, into nine groups $1, 2, \ldots, 9$. Group 5 corresponds to standard values in the standard population. We further reduced the number of states to 5 by joining some groups together. We will refer to data in this form as D.DISCR.

3.3 Binary attributes

Binary data are data whose variables can take on only two possible states, traditionally termed 0 and 1 in accordance with the binary numeral system and Boolean algebra. In our case, all laboratory tests are encoded using two binary attributes. The first attribute takes a value of 0 for the standard values of the test and a value of 1 if the values are decreased. The second attribute takes a value of 0 for the standard values of the test and value of 1 if the values are increased. The attributes Age, Height, and Weight are removed. From the demographic group of attributes only Gender and the Body Mass Index (BMI) were kept with BMI being encoded using two binary attributes BMI high and BMI low where the BMI is greater than the mean takes a value of 1, otherwise it takes a value of 0. We will refer to data in this form as D.BIN.

3.4 Attribute Selection

Before learning a model, we preprocess the data. Usually, one of the most useful parts of preprocessing is the attribute selection, where irrelevant attributes are removed. Attribute selection is a process by which we automatically search for the best subset of

attributes in our dataset. The notion of "best" is relative to the problem we are trying to solve, but typically means the highest accuracy. Three key benefits of performing attribute selection on our data are:

- Reduces Overfitting: Less redundant data means less opportunity to make decisions based on a noise.
- Improves Accuracy: Less misleading data means that modeling accuracy improves.
- Reduces Training Time: Less data means that algorithms train faster.

The CfsSubsetEval method of Weka [6] selects the subsets of attributes that are highly correlated with the class while having low inter-correlation. We searched the space of all subsets by a greedy best first search with backtracking. Data D after the application of this attribute selection method will be suffixed as D.AS.

3.5 Tested classifiers

For tests, we used a large subset of classifiers implemented in Weka. Classifiers that performed best in the preliminary tests qualified for the final tests. In the final tests we compared the following classifiers:

- **Decision tree C4.5** [19].
- **Logistic regression** [20].
- **Naive Bayes (NB) classifier** [21] assume that the value of a particular explanatory variable (attribute) is independent of the value of any other attribute given the class variable.
- **NB-Tree** generating a decision tree with naive Bayes classifiers at the leaves [22]
- **Bayesian network (BN) classifiers** (1) learned by K2 algorithm [23] - referred as BN.K and (2) Tree Augmented Naive Bayes classifier refer as BN.TAN [7].

We use the leave-one-out cross validation as the model evaluation method. It means that N separate times, the classifier is trained on all the data except for one point and a prediction is made for that point. After that the average error is computed and used to evaluate the model.

3.6 Results of experiments

For each data record classified by a classifier there are possible classification results. Either the classifier got a positive example labeled as positive (in our data the positive example is the patient survived) or it made a mistake and marked it as negative. Conversely, a negative example may have been mislabeled as a positive one, or correctly marked as negative. This defines the following metrics:

- **True Positives (TP):** number of positive examples, labeled as such.

- **False Positives (FP):** number of negative examples, labeled as positive.
- **True Negatives (TN):** number of negative examples, labeled as such.
- **False Negatives (FN):** number of positive examples, labeled as negative.

Our results are summarized in Table 2 using the following measures of the prediction quality:

- **Accuracy** measures how often the classifier makes the correct prediction. It is the ratio between the number of correct predictions and the total number of predictions.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Recall** is also known as sensitivity. It is the fraction of positive instances that are correctly classified as positive (rate of true positives).

$$REC = \frac{TP}{TP + FN}$$

- **Precision** is the fraction of true positives over the number of all reported positives.

$$PRE = \frac{TP}{TP + FP}$$

- **F-measure** is the harmonic mean of the precision and the recall

$$F = 2 \cdot \frac{PRE \cdot REC}{PRE + REC}$$

- **Specificity** is the fraction of true negatives over the number of all negatives.

$$SPE = \frac{TN}{FP + TN}$$

- **Area under the ROC curve (AUC)**. The ROC curve shows how the classifier can sacrifice the true positive rate (recall or sensitivity) for the false positive rate (1-specificity) by plotting the TP rate to the FP rate. In other words, it shows you how many correct positive classifications can be gained as you allow for more and more false positives. As an example, in Figure 2 we report the ROC curve for the Naive Bayes classifier with the ordinal attributes. Its area under the curve is 0.782.

In Table 2 we compare the results of different classifiers on different versions of data. The C4.5 classifier with D.DISCR has the highest accuracy of 0.942, its recall and precision are also among the best achieved. But its area under the ROC curve is very low, only 0.371, which suggests that this classifier can not be satisfactorily tuned if we want to sacrifice precision to recall or vice versa.

In Figure 1, we present the tree structure of the C4.5 learned from the discrete data. It has achieved the highest accuracy from all tested classifiers. Its structure is surprisingly

Table 2: Results of experiments

Classifier	Criteria	D.ORD	D.ORD.AS	D.DISCR	D.DISCR.AS	D.BIN	D.BIN.AS
Naive Bayes	ACC	0.855	0.925	0.860	0.914	0.875	0.911
	AUC	**0.782**	0.722	0.744	0.781	0.695	0.717
	Recall	0.439	0.158	0.351	0.368	0.246	0.140
	Prec.	0.234	0.450	0.215	0.396	0.203	0.276
	F-measure	0.305	0.234	0.267	0.382	0.222	0.186
C4.5	ACC	0.935	0.933	**0.942**	0.921	0.926	0.927
	AUC	0.527	0.621	0.371	0.627	0.528	0.273
	Recall	0.263	0.105	0.246	0.123	0.070	0.035
	Prec.	0.625	0.750	0.875	0.368	0.444	0.333
	F-measure	0.370	0.185	0.384	0.184	0.121	0.063
LOG.REG	ACC	0.930	0.925	0.907	0.919	0.926	0.919
	AUC	0.746	0.755	0.622	0.746	0.675	0.746
	Recall	0.140	0.018	0.193	0.140	0.070	0.140
	Prec.	0.571	0.250	0.289	0.364	0.364	0.364
	F-measure	0.225	0.033	0.232	0.203	0.118	0.203
NB-Tree	ACC	0.932	0.936	0.914	0.920	0.913	0.920
	AUC	0.658	0.480	0.701	0.726	0.701	0.726
	Recall	0.211	0.228	0.228	0.088	0.070	0.088
	Prec.	0.600	0.684	0.310	0.313	0.211	0.313
	F-measure	0.312	0.342	0.263	0.137	0.105	0.137
BN.K2	ACC	NA	NA	0.886	0.918	0.900	0.926
	AUC	NA	NA	0.750	0.775	0.687	0.671
	Recall	NA	NA	0.316	0.368	0.193	0.105
	Prec.	NA	NA	0.265	0.429	0.256	0.462
	F-measure	NA	NA	0.288	**0.396**	0.220	0.171
BN.TAN	ACC	NA	NA	0.908	0.925	0.904	0.927
	AUC	NA	NA	0.721	0.768	0.653	0.642
	Recall	NA	NA	0.193	0.228	0.088	0.053
	Prec.	NA	NA	0.297	0,464	0.179	0.333
	F-measure	NA	NA	0.234	0.306	0.118	0.091

simple. If the patient is Czech then it is predicted to survive if the patient is Syrian then the LDL cholesterol value should be checked. If it is below 4.78 then the patient is predicted to survive, otherwise, if LDL cholesterol value is between 4.78 and 6.28 then it depends on the Syrian hospital in which he/she is treated. If he/she is treated in the public hospital (SYR1) then he/she dies; if he/she is treated in the private one (SYR2) then he/she survives. If his/her LDL cholesterol values are higher than 6.28 then he/she dies (no matter what Syrian hospital he/she is treated in). The simplicity of the C4.5 classifier is in line with the general recommendation that in order to avoid the over-fitting of training data the models should be as simple as possible.

The highest area under the ROC curve (AUC) was achieved by Naive Bayes classifier with the ordinal attributes. The highest value of F-measure was achieved by BN.K2 with discrete attributes selected by the method CfsSubsetEval method of Weka [6]. The learned BN model is actually also a Naive Bayes model – see Figure 3. We can conclude that there is no single winner – a classifier that would be the best in all considered criteria. Also, the classifiers differ in what variables they consider to be important for AMI mortality prediction. We believe that it is worth learning diverse classifiers since it may help medical specialists to get a deeper insight into the modeled problem.

```
Hospital <= 0: 0 (603.0/36.0)
Hospital > 0
|   LDLC <= 4.78: 0 (157.86/6.0)
|   4.78 < LDLC <= 6.28
|   |   Hospital <= 1: 1 (12.95/2.95)
|   |   Hospital > 1: 0 (9.0/1.0)
|   LDLC > 6.28: 1 (4.18/0.18)
```

Figure 1: Decision tree C4.5 learned from D.DISCR has the highest accuracy 0.943 of all tested models.

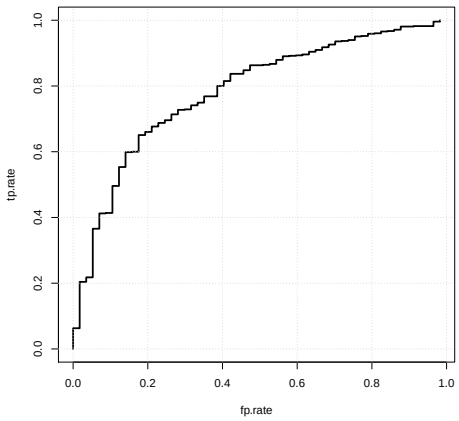


Figure 2: ROC for the Naive Bayes classifier with ordinal attributes

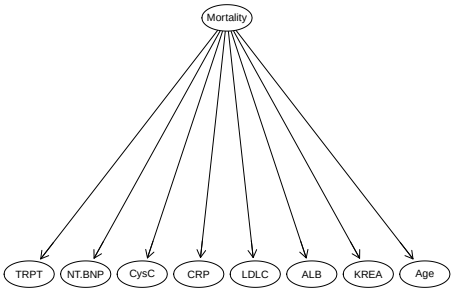


Figure 3: BN learned by BN.K2

4 Dealing with incomplete and imbalanced data

As we can see from the section Data, our data-set contains incomplete and imbalanced data. In [14] we have presented an idea to develop TAN [7] to handle incomplete and imbalanced data (Algoritm 1, and Algoritm 2), where the Conditional Mutual Information "CMI" is defined as:

$$I(X,Y|Z) = \sum_{\mathbf{x,y,z}} f(\mathbf{x,y,z}) \log \frac{f(\mathbf{z})f(\mathbf{x,y,z})}{f(\mathbf{x,z})f(\mathbf{y,z})}$$

where the sum is only over $\mathbf{x,y,z}$ such that $f(\mathbf{x,z}) > 0$ and $f(\mathbf{y,z}) > 0$.

In a similar way, we can create a procedure that enables the Chow-Liu algorithm to deal with incomplete data, where a normal Chow-Liu algorithm [17] just deals with complete data. The procedure is shown in Algorithm 3. where the Mutual Information "MI" is defined as:

$$I(X,Y) = \sum_{\mathbf{x,y}} f(\mathbf{x,y}) \log \frac{f(\mathbf{x,y})}{f(\mathbf{x})f(\mathbf{y})}$$

Algorithm 1 TAN For Incomplete Data

1: Read $D = \{\mathbf{u}_1, \ldots, \mathbf{u}_N\}, \mathbf{u}_m = (a_1, \ldots, a_n, c), m \in \{1, \ldots, N\}$
2: **procedure** CMI$(A_i, A_j, C\})$ ▷ // Conditional Mutual Information
3: $\overline{D} = \{\overline{\mathbf{u}}_1, \ldots, \overline{\mathbf{u}}_N\}, \overline{\mathbf{u}}_m = (a_i, a_j, c), m \in \{1, \ldots, N\}$, such that $\mathbf{u}_m = (a_1, \ldots, a_n, c) \in D$
4: **Foreach** $\overline{\mathbf{u}}_m \in \overline{D}$
5: **If**$(a_i == NA | a_j == NA)$
6: Delete $\overline{\mathbf{u}}_m$ from $\overline{D}$
7: **endfor**
8: Compute $I_p = I(A_i, A_j|C)$ from $\overline{D}$
9: **return** I_p
10: **Endprocedure**
11: Compute $I_p = I(A_i, A_j|C))$ between each pair of attributes, $i \neq j$, using the Procedure CMI.
12: Build a complete undirected graph in which the vertices are the attributes $A_1, A_2, \ldots, A_n$. Annotate the weight of an edge connecting A_i to A_j by $I_p = I(A_i, A_j|C))$.
13: Build a maximum weighted spanning tree.
14: Transform the resulting undirected tree to a directed one by choosing a root variable and setting the direction of all edges to be outward from it.
15: Construct a TAN model by adding a vertex labeled by C and adding edges from C to all other nodes in the graph.

where the sum is only over **x,y** such that $f(\mathbf{x}) > 0$ and $f(\mathbf{y}) > 0$.

The idea behind Algorithms 1 and 3 is that we believe if we use more data then the estimates of mutual information and conditional mutual information are more reliable.

4.1 Results

We will refer to TAN, and Chow-Liu which deal with incomplete and imbalanced data as (TANI and CLI). The results are summarized in Table 3. We compare the results of our methods with TAN in bnclassify [16], Chow-Liu [17], EM algorithm [9] for Chow-Liu using Hugin [18], normal TAN [7], and [10] (this algorithm deals with TAN based on the EM principle, where they have proposed an adaptation of the learning process of Tree Augmented Naive Bayes classifier from incomplete data, where, any variable can has missing values in the dataset), we will refer to it as (FL), and SMOTE algorithm [5] for TAN, on two versions of dataset (binary and discrete attributes). For measures of the prediction quality, we use log-likelihood (LL) and AUC. Also, We use the 10 fold cross validation as the model evaluation method. Algorithm TANI with D.BIN has achieved the highest area under the ROC curve (AUC) (ROC = 0.953) and the highest LL with -2744.4279. The results of Algorithm 1 is better than the normal TAN algorithm in the both datasets D.DISCR and D.Bin. But SMOTE algorithm with TAN has achieved the second highest LL with D.DISCR (LL= -6043.0785) but the area under the ROC curve (AUC) is (ROC = 0.802), also its ROC is better than the ROC(s) of Algorithm 1 with D.DISCR and Algorithm 3 with the both data-sets. We can conclude that the TANI is

Algorithm 2 Procedure for WeightMatrix computation with incomplete and imbalance data

var
  M The number of samples for the majority class
  N The number of samples for the minority class
  D_T All instances of the majority class, $D_T \subset D$
  D_F All instances of the minority class, $D_F \subset D$
integer division $L = M/N$
Divide D_T to L parts, $D_{T_k}, k \in \{1, \ldots, L\}$
Foreach D_{T_k}
  $D_k = D_{T_k} \cup D_F$
EndForeach
Compute WeightMatrix $\mathbb{I}_{p_k}$ foreach D_k
$\hat{\mathbb{I}}_p$ = the average of $\mathbb{I}_{p_k}, k \in 1, \ldots, L$ ▷ // $\hat{\mathbb{I}}_p$ is the final WeightMatrix

a single winner with D.Bin.

5 Quality of classifiers tested on artificial data

The data which we have is not big enough to have a very good result. Where TAN [7] is a reliable model and has been tested on many data-sets, we decided to use the model BN.TAN [7]"its result presented in Table 2" to generate a sequence of datasets with those sizes (3000 - 5000 - 7000 - 10000) and 10% missing data, to test the Algorithms (Algo 1, TANI, and FL [10]). The result show's in the Figures(4 and 5). We can see that our Algorithm 1 is better than other algorithms, also TANI seems not good with the big binary datasets.

6 Conclusions

We used medical data on patients with AIM to compare the results of (a) classification models and (b) Bayesian networks modeling the relations found in data. Although the conclusions might seem to be specific only for the used data here, we also report general observations.

In principle, the BN learning algorithms are able to discover the mediated correlation, since they test not only pairwise independence but also the conditional independence given values of other variables.

Bayesian networks are a tool of choice for reasoning in uncertainty, with incomplete data. However, often, Bayesian network structural learning only deals with complete data. We have proposed here an adaptation of the learning process of the Chow-Liu and Tree Augmented Naive Bayes classifier (TAN) from incomplete and imbalanced datasets. These methods have been successfully tested on our dataset. We have seen that the TANI algorithm is a single winner with D.Bin.

Algorithm 3 Procedure Chow-Liu for incomplete data

Read $D = \{\mathbf{u}_1, \ldots, \mathbf{u}_N\}, \mathbf{u}_m = (a_1, \ldots, a_n), m \in \{1, \ldots, N\}$
procedure MI(A_i, A_j) ▷ // Mutual Information
  $\overline{D} = \{\overline{\mathbf{u}}_1, \ldots, \overline{\mathbf{u}}_N\}, \overline{\mathbf{u}}_m = (a_i, a_j), m \in \{1, \ldots, N\}, a_i, a_j \in \mathbf{u}_m$, such that $\mathbf{u}_m = (a_1, \ldots, a_n) \in D$
  Foreach $\overline{\mathbf{u}}_m \in \overline{D}$
    If$(a_i == NA | a_j == NA)$
      Delete $\overline{\mathbf{u}}_m$ from $\overline{D}$
  endfor
  Compute $I_p = I(X, Y)$ from $\overline{D}$
  return I_p
Endprocedure
Compute $I_p = I(A_i, A_j)$ between each pair of attributes, $i \neq j$, using the Procedure MI.
Build a complete undirected graph in which the vertices are the attributes $A_1, A_2, \ldots, A_n$. Annotate the weight of an edge connecting A_i to A_j by $I_p = I(A_i, A_j)$.
Build a maximum weighted spanning tree.
Transform the resulting undirected tree to a directed one by choosing a root variable and setting the direction of all edges to be outward from it.

Table 3: BN Results

		D.DISCR	D.Bin
TAN_bnclassify	AUC	0.804	0.448
	LL	-7340.9414	-7497.461
FL	AUC	0.77081	0.871
	LL	-11319.6	-6368.38
SMOTE.TAN	AUC	0.802	0.818
	LL	**-6043.0785**	-7168.8239
Chow-Liu	AUC	0.723	0.69
	LL	-12763.4	-6396.2673
EM-Chow-Liu	AUC	0.6917	0.71
	LL	-11508.2	-8869.63
BN.TAN	AUC	0.62	0.67
	LL	-11321.406	-6368.453
Algo1	AUC	0.77	0.93
	LL	-19914.4937	-2819.3032
Algo3	AUC	0.75	0.73
	LL	-6145.0196	-2755.7778
TANI	AUC	**0.82**	**0.953**
	LL	-9393.4688	**-2744.4279**
CLI	AUC	0.476	0.8956
	LL	-6317.81655	-2953.3373

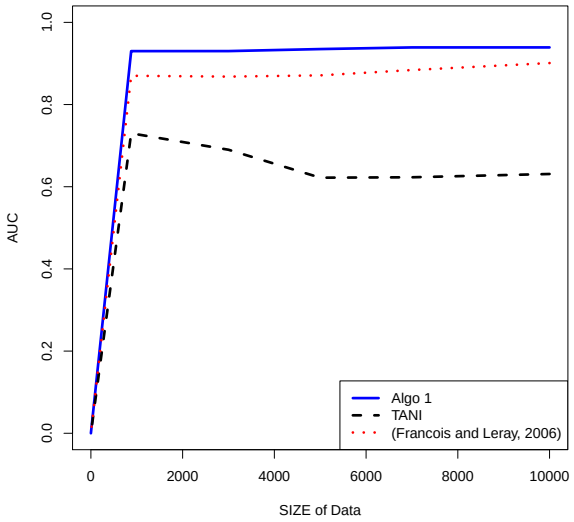


Figure 4: AUC quality of classifiers (D.Bin)

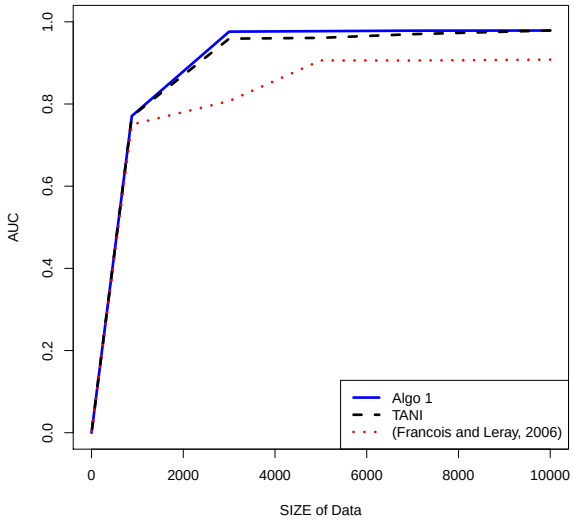


Figure 5: AUC quality of classifiers (D.DISCR)

References

[1] K. P. Murphy, *Machine learning: a probabilistic perspective*, The MIT Press, 2012.

[2] R. Duda and P. Hart and D.G. Stork, *Pattern Classification*, Wiley and Sons, 2001.

[3] H. M. Krumholz, S.-L. T. Normand, D. H. Galusha, J. A. Mattera, A. S. Rich, Y. Wang and Y. Wang, *Risk-Adjustment Models for AMI and HF 30-Day Mortality, Methodology*, Harvard Medical School, Department of Health Care Policy, (2007).

[4] J. Vomlel and H. Kružík and P. Tůma and J. Přeček, and M. Hutyra, *Machine Learning Methods for Mortality Prediction in Patients with ST Elevation Myocardial Infarction*, In the Proceedings of The Nineth Workshop on Uncertainty Processing WUPES'12, Czech Republic, 204–213, (2012).

[5] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, *Synthetic Minority Over-sampling Technique*, Journal of Artificial Intelligence Research, 11, 321–357, (2002).

[6] M. Hall and E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and H. Witten, *The WEKA Data Mining Software: an Update*, In 'ACM SIGKDD Exploration ACM SIGKDD Explorations', 11 (2009), 10–18.

[7] N. Friedman, D. Geiger, and M. Goldszmidt, Bayesian network classifiers, Machine Learning Journal, 29 (1997), 131–163.

[8] Feature selection in Bayesian classifiers for the prognosis of survival of cirrhotic patients treated with TIPS, R. Blanco and I. Inza and P. L. naga , Journal of Biomedical Informatics 8, 507–543, 2005.

[9] I. Cohen and F. Cozman and N. Sebe and M. C. Cirelo and T. S. Huang, *Semi-supervised learning of classifiers: theory, algorithms and their application to human-computer interaction*, IEEE Transactions on Pattern Analysis and Machine Intelligence 26, 1553–1568, (2004).

[10] O. C. H. Francois and P. Leray, *Learning the Tree Augmented Naive Bayes Classifier from incomplete datasets*, Third European Workshop on Probabilistic Graphical Models, 91–98, (2006).

[11] D. Heckerman and E. Horvitz and B. Nathwani, Toward normative expert systems, 1: The PATHFINDER project, Methods of Information in Medicine, 31, 90–105, 1992.

[12] T. Hastie and R. Tibshirani and J. Friedman, *The Elements of Statistical Learning*, Springer, 2009

[13] A Bayesian Network Model for Diagnosis of Liver Disorders, A. Onisko and M. Druzdzel and H. Wasyluk, Proceedings of the Eleventh Conference on Biocybernetics and Biomedical Engineering, 842–846, 1999.

[14] A machine learning method for incomplete and imbalanced medical data, I. Salman and J. Vomlel, PROCEEDINGS OF THE 20TH CZECH-JAPAN SEMINAR ON DATA ANALYSIS AND DECISION MAKING UNDER UNCERTAINTY, Pardubice, CZECH REPUBLIC, 188–195, (2017).

[15] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques*, MIT press, 2009

[16] B. Mihaljevic and C. Bielza and P. Larranaga, =Comments on bnclassify package runtimes, URl: http://127.0.0.1:15009/library/bnclassify/doc/runtimes.pdf, 2015

[17] C. Chow and C. Liu, Approximating discrete probability distributions with dependence trees, IEEE Trans, on Info, Theory 14, 462–467, (1968).

[18] Hugin, Hugin Expert A/S. Hugin Expert, `http://www.hugin.com`, 2010

[19] C4.5: Programs for Machine Learning, R. Quinlan and M. Kaufmann, Machine Learning , 29, 131–163, 1993

[20] Ridge estimators in logistic regression, S. le Cessie and J.C. van Houwelingen, University of Leiden, The Netherlands, 1992.

[21] Pattern Classification and Scene Analysis, R. O. Duda and P. E. Hart, Wiley-Interscience, Oxford, 30, 106–110, 1973.

[22] Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid, R. Kohavi, In Proceedings: Second International Conference on Knowledge Discovery and Data Mining, 202–207, 1996.

[23] A Bayesian Method for the Induction of Probabilistic Networks from Data, G. F. Cooper, Machine Learning, 9, 309–347, 1992.

Quantum Square Well with Logarithmic Central Spike*

Iveta Semorádová

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

advisor: Miloslav Znojil, Department of Theoretical Physics
Nuclear Physics Institute, CAS

Abstract. Singular repulsive barrier $V(x) = -g\ln(|x|)$ inside a square well is interpreted and studied as a linear analogue of the state-dependent interaction $\mathcal{L}_{eff}(x) = -g\ln[\psi^*(x)\psi(x)]$ in nonlinear Schrödinger equation. In the linearized case, Rayleigh-Schrödinger perturbation theory is shown to provide a closed-form spectrum at the sufficiently small g or after an amendment of the unperturbed Hamiltonian. At any spike-strength g, the model remains solvable numerically, by the matching of wave functions. Analytically, the singularity is shown regularized via the change of variables $x = \exp y$ which interchanges the roles of the asymptotic and central boundary conditions.

Keywords: state-dependence of interactions, effective Hamiltonians, logarithmic nonlinearities, linearized quantum toy model.

Abstrakt. Singulární repulsivní bariéra $V(x) = -g\log(|x|)$ uvnitř čtvercové jámy je interpretována a studována jako lineární analog interakce závislé na stavech $\mathcal{L}_{eff}(x) = -g\log[\psi^*(x)\psi(x)]$ v nelineární Schrödingerově rovnici. V linearizovaném případě je ukázáno, že Rayleigh-Schrödingerova poruchová teorie poskytuje spektrum v uzavřeném tvaru pro dostatečně malé g nebo po pozměnění neporušeného Hamiltoniánu. Pro jakoukoliv sílu bariéry g, model zůstává numericky řesitelný spojením vlnových funkcí. Analyticky je singularita regularizována pomocí záměny proměnných $x = \exp y$, která vymění role asymptotických a centrálních hraničních podmínek.

Klíčová slova: interakce závislé na stavech, efektivní Hamiltoniány, logaritmické nelinearity, linearizovaný kvantový zjednoduěný model.

Full paper: M. Znojil, I. Semorádová. *Quantum square well with logarithmic central spike.* Modern Physics Letters A **33** (2018), 1850009.

*This work was supported by the Grant Agency of the Czech Technical University in Prague, grant No. SGS16/239/OHK4/3T/14

Perfect State Transfer by Means of Discrete-Time Quantum Walks on Complete Bipartite Graph*

Stanislav Skoupý

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

advisor: Martin Štefaňák, Department of Physics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. We consider the state transfer algorithm based on discrete-time quantum walks on complete bipartite graph between two vertices, the sender and the receiver. We show in our previous paper that the perfect state transfer is not achieved when the sender and the receiver are in the opposite parts of the graph and when those parts have different size. We prove in this article that the use of Grover coin followed by phase shift by π on the marked vertices achieves the perfect state transfer on all complete bipartite graphs.

Keywords: quantum walk, state transfer, complete bipartite graph

Abstrakt. Uvažujeme model algorytmu přenosu stavu založeného na kvantových procházkách v diskrétním čase na kompletním bipartitním grafu mezi dvěma vrcholy, odesilatelem a příjemcem. V našem předchozím článku jsme ukázali, že k úplnému přenosu stavu nedochází v případě, kdy odesilatel a příjemce jsou v různých částech grafu a tyto části mají různou velikost. V tomto článku dokážeme, že při použití Groverovy mince následované fázovým posunem o π na označené vrcholy dochází k úplnému přenosu stavu na všech kompletních bipartitních grafech.

Klíčová slova: kvantová procházka, přenos stavu, úplný bipartitní graf

1 Introduction

Quantum walks are important part of quantum computer science. State transfer algorithm is a one type of algorithm that can be implemented by means of quantum walks by modifying the quantum walk search algorithm. In this article we use the model of discrete-time quantum walk with coins and we focus on the state transfer algorithm on complete bipartitite graph. In [5] we show that perfect state transfer is achieved for the star graph and complete graph. We define the perfect state transfer as the quantum walker moving from one vertex to another vertex with probability close to 1. Algorithm that we use does not allow transfer of inner state of the walker. In [6] we prove that perfect state transfer is also possible on complete bipartite graph but not for all cases. In this article we show how with different setting in the algorithm we can achieve perfect state transfer on all complete bipartite graphs.

*This work was supported from Czech Technical University in Prague under project Centrum pokročilých aplikovaných přírodních věd (CZ.02.1.01/0.0/0.0/16_019/0000778).

In the first section we describe the general scheme of the state transfer algorithm. Then we show the numerical simulation that suggests that perfect state transfer can be achieved for all complete bipartite graphs and in the last section we present a proof that this is truly the case.

2 General scheme

In this section we present the general scheme of the state transfer algorithm based on the discrete-time quantum walks search algorithm on the complete bipartite graph $K_{N,M}$. This algorithm is based on work in [1],[2] and [3]. The complete bipartite graph $K_{N,M}$ is graph composed of 2 parts where there are no edges between vertices in the same part and vertex form one part is connected to all vertices in the other part. We consider state transfer between 2 vertices, sender and receiver, and the sender vertex will always be in the part of the graph with N vertices. At first in this section we construct the Hilbert space of the walk and then we introduce evolution operator of the state transfer algorithm. Last but not least we describe the state transfer algorithm itself.

Let us first describe the Hilbert space of the walk at complete bipartite graph. Because the graph is bipartite we can write the Hilbert space $\mathcal{H}$ as a direct sum

$$\mathcal{H} = \mathcal{H}^1 \oplus \mathcal{H}^2 \tag{1}$$

where each space $\mathcal{H}^i$ corresponds to different part of the graph. We consider that the sender is in the space $\mathcal{H}^1$. Each Hilbert space has $\mathcal{H}^i$ a form of tensor product of position space $\mathcal{H}_P$ and so called coin $\mathcal{H}_C$ space which directs the evolution of the walk. Hence we get

$$\mathcal{H}^i = \mathcal{H}^i_P \otimes \mathcal{H}^i_C. \tag{2}$$

The states in the position space describe the position of the walker and states in the coin space describe the direction where walker will move. To clarify the notation we use latin letters to describe the position in the first part of the graph and greek letters to describe the position in the second part of the graph. Thus having the graph $K_{N,M}$ the position space of the first part of the graph $\mathcal{H}^1_P$ is spanned by states $|j\rangle_P$ where j goes from 1 to N. The coin space $\mathcal{H}^1_C$ is spanned by states $|\alpha\rangle_C$ where α goes from 1 to M. The position space of the second part of the graph $\mathcal{H}^2_P$ is spanned by contrast by states $|\alpha\rangle_P$ where α goes from 1 to M and the coin space $\mathcal{H}^2_C$ is spanned by states $|j\rangle_C$ where j goes from 1 to N. So the basis states of $\mathcal{H}^1$ has the form $|j,\alpha\rangle = |j\rangle_P \otimes |\alpha\rangle_C$, $\mathcal{H}^2$ with basis states $|\alpha, j\rangle = |\alpha\rangle_P \otimes |j\rangle_C$. Since the sender lies in the $\mathcal{H}^1$ we denote its position by $|s\rangle_P$.

The evolution operator $\hat{U}$ of each step of the discrete-time quantum walk with coins is composed of 2 operators, shift operator $\hat{S}$ and coin operator $\hat{C}$. The shift operator moves the walker from one position to the other and in our basis it has a following form

$$\hat{S} = \sum_{j=1}^{N}\sum_{\alpha=1}^{M} \left(|j,\alpha\rangle\langle\alpha,j| + |\alpha,j\rangle\langle j,\alpha|\right). \tag{3}$$

The coin operator $\hat{C}$ acts locally at each vertex and main idea of the algorithm is to use one local coin operator on the sender and the receiver vertices and different coin operator

on the rest vertices of the graph. The local coin operator that we use on non-marked vertices is so called Grover operator [4]

$$\hat{G}_v = -\hat{I} + 2|\psi_v\rangle\langle\psi_v| \tag{4}$$

where $|\psi_v\rangle$ is equal superposition of all direction of the coin space at vertex v. At the sender and the receiver we can use either phase shift by π or the Grover coin operator followed by phase shift by π. In our paper on state transfer on the complete bipartite graph [6] we use the phase shift by π at marked vertices which is not the best choice. In this article we show that the Grover coin followed by phase shift by π achieves better results. We discuss these differences in the following section. Finally each step of the algorithm is given by evolution operator given by $\hat{U} = \hat{S}\hat{C}$.

At last the steps of the state transfer algorithm based on discrete-time quantum walks are following:

1. Initialize the system in the superposition of all directions at the sender vertex

$$|init\rangle = \frac{1}{\sqrt{M}} \sum_{\alpha=1}^{M} |s, \alpha\rangle. \tag{5}$$

2. Apply the evolution operator $\hat{U}$ T-times.

3. Measure the position of the walker.

The number of steps T generally depends on the choice of the coin and on the size of the graph. However ss we show later the number of steps for correct choice of the coin depends only on the size of the part of the graph which contains the receiver vertex. We not yet write the whole coin operator, since it depends on the relative position of the sender and the receiver vertices. We have 2 cases, first when both marked vertices are in the same part of the graph and second when they are in the opposite parts of the graph. We discuss those cases in the following section.

3 Different coins

In this section we examine the different choices of the local coin operator applied on the sender and receiver vertex. We start with the case where both marked vertices are in the same part of the graph, then we move to the case where they lie in the opposite parts of the graph.

At first we present the case when the sender and the receiver are in the same part of the graph. We label the state corresponding to position of the receiver vertex by $|r\rangle_P$. In this case the coin operator $\hat{C}$ reads

$$\hat{C} = \sum_{j\neq s,r}^{N} |j\rangle\langle j|_P \otimes \hat{G}_j + \sum_{\alpha=1}^{M} |\alpha\rangle\langle\alpha|_P \otimes \hat{G}_\alpha + (|s\rangle\langle s|_P + |r\rangle\langle r|_P) \otimes \hat{C}_L \tag{6}$$

where $|j\rangle\langle j|_P$ is projector operator at position space and $\hat{C}_L$ is local coin of our choice that is applied at each step at marked vertices. In our paper [6] we choose as the coin

operator phase shift by π. In this case where both the sender and the receiver are in the same part of the graph the other choice, the Grover coin followed by phase shift by π, does not change the evolution of the system. Both choices achieve the perfect state transfer in the number of steps which scales $O\left(\sqrt{N}\right)$. The size of other part of the graph does not play a role in fidelity of the state transfer or the number of steps. The evolution of fidelity in this case is displayed in figure (1).

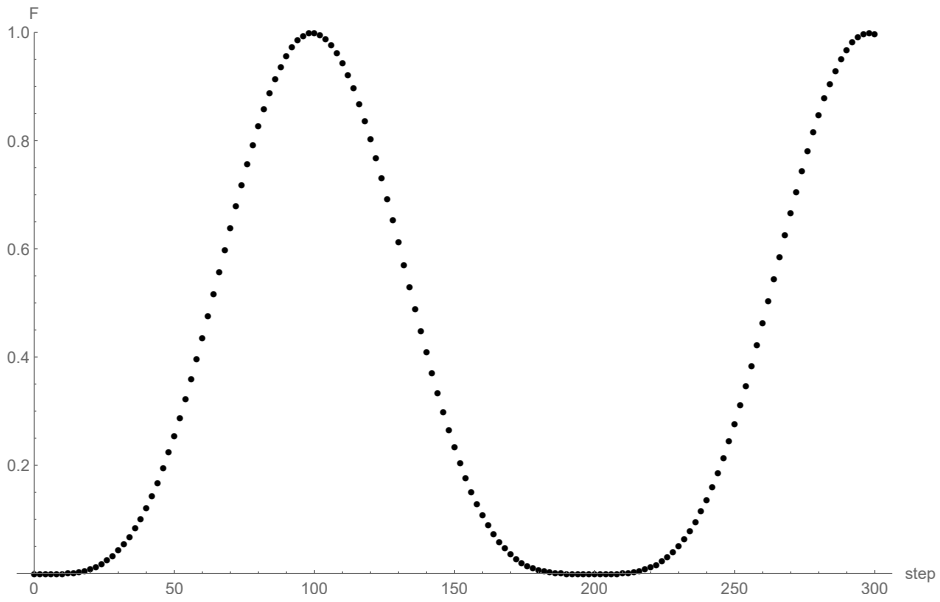


Figure 1: There is evolution of the fidelity $\mathcal{F}$ of the state transfer during 300 steps in the graph when both the sender and the receiver are in the same part of graph. This part of the graph has 2000 vertices. Only even steps are plotted because at odd steps fidelity is equal to 0. The coin operator at the marked vertices is the Grover coin followed by phase shift by π.

In the second case the sender and receiver are in the opposite parts of the graph. Hence we label the position of the receiver by state $|\rho\rangle_P$. For this case the coin operator reads

$$\hat{C} = \sum_{j\neq s}^{N} |j\rangle\langle j|_P \otimes \hat{G}_j + \sum_{\alpha\neq\rho}^{M} |\alpha\rangle\langle\alpha|_P \otimes \hat{G}_\alpha + |s\rangle\langle s|_P \otimes \hat{C}_{L1} + |\rho\rangle\langle\rho|_P \otimes \hat{C}_{L2} \qquad (7)$$

where $\hat{C}_{L1}$ and $\hat{C}_{L2}$ are the local coins operators at the sender and the receiver. Unlike the case where both marked vertices are in the same part the choice of the local coins operator changes the evolution of the system. We examine phase shift by π as the local coin operator in the paper [6] and we find that the perfect state transfer is achieved only in the case where both parts have the same size, otherwise fidelity of the state transfer does not go to one. See the figure (2) where there is displayed dependence of the fidelity on the size of the part with the sender and the receiver vertices. The maximal fidelity is

in this case given by

$$\mathcal{F}_{max}(N,M) = \left(\frac{\sqrt{NM} + \sqrt{(N-1)(M-1)}}{N+M-1} \right)^2 .. \tag{8}$$

However, the choice of the Grover coin followed by phase shift of π achieves better results. As we prove in following section this choice always achieves the perfect state transfer for all sizes of the part with the sender vertex. In the figure (3) we see the comparison of both choices of the coin operators.

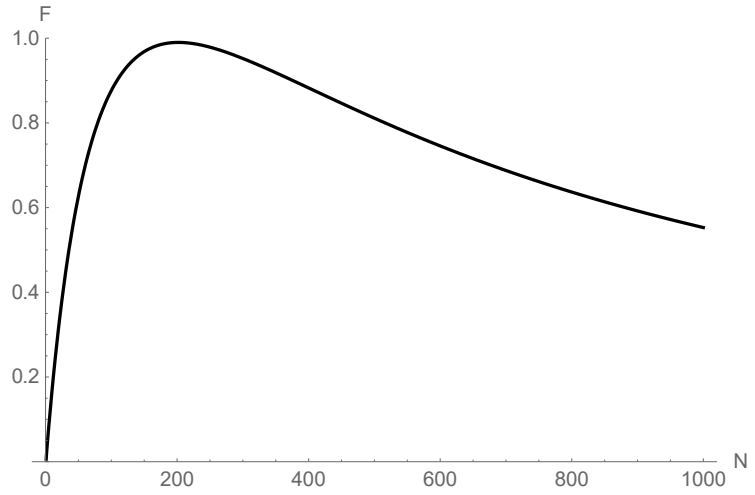


Figure 2: The maximum of fidelity $\mathcal{F}_{max}$ depending on the size of the part with the sender given by (8) with choice of the local coin being phase shift by π. The part with the receiver vertex has $M = 200$ vertices. The perfect state transfer is reached for $N = 200$.

4 Proof of perfect state transfer

In this section we calculate the fidelity of the state transfer for general graph $K_{N,M}$ for the case when the sender is in the part of the graph with N vertices and the receiver is in the opposite part with size M. We consider case where we use the Grover coin operator followed by phase shift by π on the receiver and the sender. We use the same methods for simplifying the calculation as we use in the papers [5] and [6]. Since the graph is bipartite and we are interested in the probability of the state transfer to the receiver vertex we use in the calculation the square of evolution operator $\hat{U}^2$ instead of the evolution operator $\hat{U}$, i.e we make 2 steps at once. We also find the invariant subspace of the walk which preserves all information of the walk but it has much smaller dimension than the Hilbert space of the walk $\mathcal{H}$. It is easy to verify that the subspace $\mathcal{H}_{eff}$ of the Hilbert space $\mathcal{H}_2$

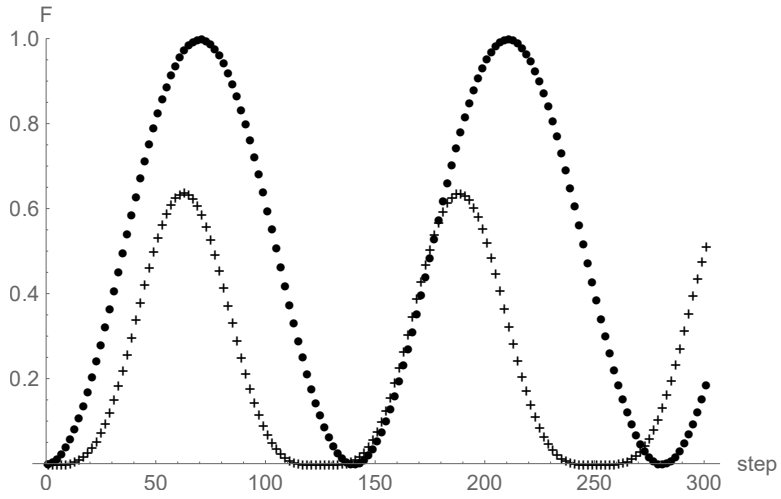


Figure 3: The evolution of the fidelity of the state transfer on the graph $K_{500,2000}$ where the receiver is in the part with 2000 vertices and the sender is the opposite part. Only odd steps are plotted. Dots correspond to the choice the Grover coin followed by phase shift by π. We see that the fidelity goes to 1 and the perfect state transfer is achieved. "+" correspond to phase shift by π and the success of the transfer does not exceed 0.7.

which is spanned by following states

$$\begin{aligned} |\nu_1\rangle &= |\rho, s\rangle \\ |\nu_2\rangle &= \frac{1}{\sqrt{N-1}} \sum_{j\neq s}^{N} |\rho, j\rangle \\ |\nu_3\rangle &= \frac{1}{\sqrt{M-1}} \sum_{\alpha\neq \rho}^{M} |\alpha, s\rangle \\ |\nu_4\rangle &= \frac{1}{\sqrt{(M-1)(N-1)}} \sum_{\alpha\neq\rho}^{M} \sum_{j\neq s}^{N} |\alpha, j\rangle \end{aligned} , \tag{9}$$

is invariant with respect to the $\hat{U}^2$. Also the initial state of the walk after one step, i.e. one application of the evolution operator, lies in the subspace $\mathcal{H}_{eff}$. It reads

$$|init'\rangle = \hat{U}|init\rangle = -\frac{1}{\sqrt{M}}|\nu_1\rangle - \sqrt{\frac{M-1}{M}}|\nu_3\rangle. \tag{10}$$

Further more if we write the state of the system after $2t+1$ steps in basis of subspace $\mathcal{H}_{eff}$ as follows

$$|\psi(t)\rangle = U^{2t}|init'\rangle = \alpha(t)|\nu_1\rangle + \beta(t)|\nu_2\rangle + \gamma(t)|\nu_3\rangle + \delta(t)|\nu_4\rangle \tag{11}$$

we see that the fidelity of the state transfer reads

$$\begin{aligned} \mathcal{F}(t) &= \sum_{j=1}^{N} |\langle \rho, j|\psi(t)\rangle|^2 = |\alpha(t)|^2 |\langle \rho, s|\nu_1\rangle|^2 + |\beta(t)|^2 \sum_{j\neq s} |\langle \rho, j|\nu_2\rangle|^2 = \\ &= |\alpha(t)|^2 + |\beta(t)|^2 \end{aligned} . \tag{12}$$

Hence, the calculation can be reduce to calculation in 4-dimensional subspace.

We introduce effective evolution operator $\hat{U}_{eff}$ as the square of evolution operator $\hat{U}$ in subspace $\mathcal{H}_{eff}$. It is given by following unitary matrix

$$U_{eff} = \begin{pmatrix} \frac{(M-2)(N-2)}{MN} & -\frac{2(M-2)\sqrt{N-1}}{MN} & \frac{2(N-2)\sqrt{M-1}}{MN} & -\frac{4\sqrt{(M-1)(N-1)}}{MN} \\ \frac{2(M-2)\sqrt{N-1}}{MN} & \frac{(M-2)(N-2)}{MN} & \frac{4\sqrt{(M-1)(N-1)}}{MN} & \frac{2(N-2)\sqrt{M-1}}{MN} \\ -\frac{2(N-2)\sqrt{M-1}}{MN} & \frac{4\sqrt{(M-1)(N-1)}}{MN} & \frac{(M-2)(N-2)}{MN} & -\frac{2(M-2)\sqrt{N-1}}{MN} \\ -\frac{4\sqrt{(M-1)(N-1)}}{MN} & -\frac{2(N-2)\sqrt{M-1}}{MN} & \frac{2(M-2)\sqrt{N-1}}{MN} & \frac{(M-2)(N-2)}{MN} \end{pmatrix}. \tag{13}$$

The spectrum of U_{eff} is composed of two pairs of conjugate eigenvalues which are given by

$$\lambda_{\pm 1,2} = e^{\pm i\omega_{1,2}}, \tag{14}$$

where the eigenphases have a following form

$$\begin{aligned} \omega_1 &= \arccos\left(\frac{(M-2)(N-2) - 4\sqrt{(M-1)(N-1)}}{MN}\right) \\ \omega_2 &= \arccos\left(\frac{(M-2)(N-2) + 4\sqrt{(M-1)(N-1)}}{MN}\right) \end{aligned}. \tag{15}$$

We see that the second eigenphase ω_2 for the case $K_{N,N}$ is equal to 0 and second pair of eigenvalues goes to $\lambda_2 = 1$. This case we threat separately later, first we look at cases there both parts of the graph has different number of vertices. In this case the corresponding eigenvectors read

$$\begin{aligned} |\lambda_{\pm 1}\rangle &= \pm\frac{i}{2}\left(|\nu_1\rangle + |\nu_4\rangle\right) + \frac{1}{2}\left(|\nu_2\rangle - |\nu_3\rangle\right) \\ |\lambda_{\pm 2}\rangle &= \pm\frac{i}{2}\left(|\nu_1\rangle - |\nu_4\rangle\right) + \frac{\mathrm{sgn}(M-N)}{2}\left(|\nu_2\rangle + |\nu_3\rangle\right) \end{aligned}. \tag{16}$$

We rewrite the state of the evolution (11) in the eigenbasis of $\hat{U}_{eff}$ and we get

$$\begin{aligned} |\psi(t)\rangle = {} & e^{i\omega_1 t}\langle\lambda_{+1}|init'\rangle|\lambda_{+1}\rangle + e^{-i\omega_1 t}\langle\lambda_{-1}|init'\rangle|\lambda_{-1}\rangle \\ & + e^{i\omega_2 t}\langle\lambda_{+2}|init'\rangle|\lambda_{+2}\rangle + e^{-i\omega_2 t}\langle\lambda_{-2}|init'\rangle|\lambda_{-2}\rangle \end{aligned}. \tag{17}$$

From (17) we calculate the $\alpha(t)$ as follows

$$\begin{aligned} \alpha(t) &= \langle\nu_1|\psi(t)\rangle = \\ &= e^{i\omega_1 t}\langle\lambda_{+1}|init'\rangle\langle\nu_1|\lambda_{+1}\rangle + e^{-i\omega_1 t}\langle\lambda_{-1}|init'\rangle\langle\nu_1|\lambda_{-1}\rangle \\ &+ e^{i\omega_2 t}\langle\lambda_{+2}|init'\rangle\langle\nu_1|\lambda_{+2}\rangle + e^{-i\omega_2 t}\langle\lambda_{-2}|init'\rangle\langle\nu_1|\lambda_{-2}\rangle = \\ &= \frac{-1+i\sqrt{M-1}}{4\sqrt{M}}e^{i\omega_1 t} + \frac{-1-i\sqrt{M-1}}{4\sqrt{M}}e^{-i\omega_1 t} + \\ &+ \frac{-1-i\sqrt{M-1}\mathrm{sgn}(M-N)}{4\sqrt{M}}e^{i\omega_2 t} + \frac{-1+i\sqrt{M-1}\mathrm{sgn}(M-N)}{4\sqrt{M}}e^{-i\omega_2 t} = \\ &= \frac{-\cos(\omega_1 t) - \cos(\omega_2 t) - \sqrt{M-1}\sin(\omega_1 t) + \mathrm{sgn}(M-N)\sqrt{M-1}\sin(\omega_2 t)}{2\sqrt{M}} \end{aligned} \tag{18}$$

For $\beta(t)$ we do the same calculation and we get

$$\begin{aligned}\beta(t) &= \langle \nu_2 | \psi(t) \rangle = \\ &= \frac{\sqrt{M-1}\cos(\omega_1 t) - \sqrt{M-1}\cos(\omega_2 t) - \sin(\omega_1 t) - \mathrm{sgn}(M-N)\sin(\omega_2 t)}{2\sqrt{M}}\end{aligned} \tag{19}$$

Using (18) and (19) in (12) we get following form of the fidelity

$$\mathcal{F}(t) = \frac{1}{2} - \frac{(M-2)\cos(t\varphi) - \sqrt{M-1}\sin(t\varphi)}{2M} \tag{20}$$

where φ depends on relative size of the parts of $K_{N,M}$ as follows

$$\begin{aligned}\varphi_{M<N} &= \omega_1 + \omega_2 \\ \varphi_{M>N} &= \omega_1 - \omega_2\end{aligned}. \tag{21}$$

t in (20) corresponds to the number of application of $\hat{U}_{eff}$ and the number of steps of the state transfer algorithm is given by $T = 2t + 1$. In order to calculate φ we use following properties of arccos function

$$\begin{aligned}\arccos(x) + \arccos(y) &= \arccos\left(xy - \sqrt{1-x^2}\sqrt{1-y^2}\right) \text{ for } x + y \geq 0 \\ \arccos(x) - \arccos(y) &= \arccos\left(xy + \sqrt{1-x^2}\sqrt{1-y^2}\right) \text{ for } x - y < 0\end{aligned}. \tag{22}$$

In the case where the part with the sender is bigger, i.e. $N > M$, we get

$$\varphi_{N>M} = \omega_1 + \omega_2 = \arccos\left(\frac{M^2 - 8M + 8}{M^2}\right) \text{ when } \frac{2(M-2)(N-2)}{MN} \geq 0 \tag{23}$$

and since the condition hold for $N > 2$ and $M > 2$ we can use the formula (22). In the second case $N < M$ φ has a form

$$\varphi_{N<M} = \omega_1 - \omega_2 = \arccos\left(\frac{M^2 - 8M + 8}{M^2}\right) \text{ when } -\frac{8\sqrt{(M-1)(N-1)}}{MN} < 0. \tag{24}$$

After calculating the cases where both parts have different size we return to the special case of $M = N$. The spectrum is composed of one pair of conjugate eigenvalues $\lambda_{\pm 1}$, that are given by (14), and degenerate eigenvalue $\lambda_2 = 1$. The eigenphase of $\lambda_{\pm 1}$ reads

$$\omega = \arccos\left(\frac{M^2 - 8M + 8}{M^2}\right) \tag{25}$$

and the corresponding eigenvectors do not change from the general case and are given by (16). The eigenstates corresponding to λ_2 read

$$\begin{aligned}|\lambda_{2,1}\rangle &= \frac{1}{\sqrt{2M}}\left(|\nu_1\rangle - |\nu_4\rangle\right) + \sqrt{\frac{M-1}{2M}}\left(|\nu_2\rangle + |\nu_3\rangle\right) \\ |\lambda_{2,2}\rangle &= \sqrt{\frac{M-1}{2M}}\left(|\nu_4\rangle - |\nu_1\rangle\right) + \frac{1}{\sqrt{2M}}\left(|\nu_2\rangle + |\nu_3\rangle\right).\end{aligned} \tag{26}$$

We chose (26) so that $|\lambda_{2,2}\rangle$ is orthogonal to the initial state (10). We again calculate $\alpha(t)$ and $\beta(t)$ and get

$$\begin{aligned}\alpha(t) = \langle\nu_1|\psi(t)\rangle &= e^{i\omega t}\langle\lambda_{+1}|init'\rangle\langle\nu_1|\lambda_{+1}\rangle + e^{-i\omega t}\langle\lambda_{-1}|init'\rangle\langle\nu_1|\lambda_{-1}\rangle + \langle\lambda_{2,1}|init'\rangle\langle\nu_1|\lambda_{2.1}\rangle = \\ &= \frac{-1-\cos(\omega t)-\sqrt{M-1}\sin(\omega t)}{2\sqrt{M}} \\ \beta(t) = \langle\nu_2|\psi(t)\rangle &= \frac{-\sqrt{M-1}-\sin(\omega t)+\sqrt{M-1}\cos(\omega t)}{2\sqrt{M}}\end{aligned} \tag{27}$$

The fidelity is given by the same expression (20) where φ is now equal to eigenphase (25). Since expression the fidelity is the same for all cases of the relative size of the parts we find the condition for fidelity to reach its maximal value. This condition reads

$$\varphi t = \arccos\left(\frac{2-M}{M}\right). \tag{28}$$

For this condition fidelity (20) reaches 1, thus perfect state transfer is achieved for all cases od $K_{N,M}$. Last but not least the number of steps T to reach the maximal fidelity of the state transfer algorithm is the closest odd integer to the number given by

$$T = 2t+1 = \frac{2\arccos\left(\frac{2-M}{M}\right)}{\arccos\left(\frac{M^2-8M+8}{M^2}\right)} + 1 \doteq \frac{\pi}{2}\sqrt{M} + O\left(\frac{1}{\sqrt{M}}\right) \tag{29}$$

where we use the Taylor expansion for large M. Form (29) we see that number of steps scales as square root of the size of the part with the receiver vertex. The numerical simulation of the evolution of the fidelity of the state transfer algorithm is displayed in figure (4) and is compared with (20).

5 Conclusion

We show that choise of the local coin operator on the marked vertices in [6] is not optimal in the case when the sender and the receiver are in the different parts of the graph. We improve our results from [6] and prove that perfect state transfer is achieved on all complete bipartite graphs with the Grover coin operator followed by phase shift by π acting an the marked vertices. The number of steps of the state transfer algorithm depends only on the size of the part which contains the receiver vertex and scales as square root of that size, which is the same behavior as in other cases of perfect state transfer on highly symmetric graphs in [5].

References

[1] N. Shenvi, J. Kempe, K. B. Whaley. *A quantum random walk search algorithm*, Phys. Rev. A 67, 052307 (2003)

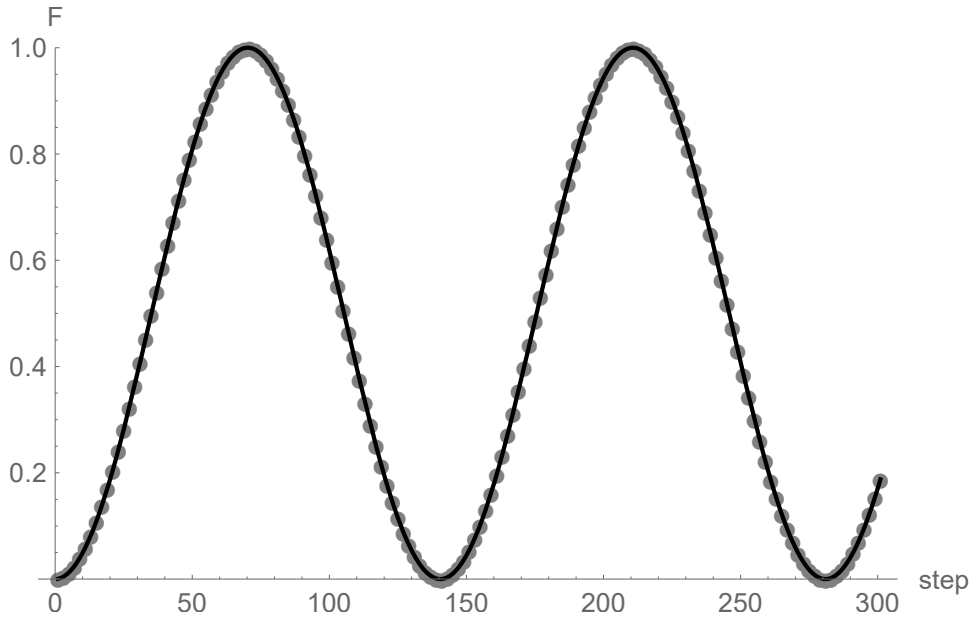


Figure 4: The evolution of the fidelity of the state transfer algorithm during 300 steps on the complete bipartite graph $K_{500,2000}$ where the receiver is in the part with 2000 vertices. Only odd steps are plotted. Dots correspond to the numerical simulation. Line represents the predicted fidelity (20).

[2] A. Ambainis, J. Kempe, A. Rivoch. *Coins make quantum walks faster*, Proceedings of the 16th Annual ACM-SIAM Symposium on Discrete Algorithms (2005), p. 1099

[3] B. Hein, G. Tanner. *Wave communication across regular lattice*, Phys. Rev. Lett. 103, 260501 (2009)

[4] Lov K. Grover. *Quantum Mechanic helps in searching for a needle in a haystack*, Phys. Rev Lett. 78, 325 (1997)

[5] M. Štefaňák, S. Skoupý. *Perfect state transfer by means of discrete-time quantum walk search algorithms on highly symmetric graphs*, Phys. Rev. A 94, 022301 (2016)

[6] M. Štefaňák, S. Skoupý. *Perfect state transfer by means of discrete-time quantum walk on complete bipartite graphs*, Quantum Inf. Process. 16, 72 (2017)

Unified Presentation and Comparison of Various Formulations of the Phase Stability and Phase Equilibrium Calculation Problems

Tomáš Smejkal*

2nd year of PGS, email: smejkto5@fjfi.cvut.cz
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

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

Abstract. In this paper we present general formulations of the phase-equilibrium and phase-stability problems for multicomponent mixtures and verify that these formulations generalize the problems of phase-equilibrium and phase-stability at constant volume, temperature, and mole numbers (VTN-flash), at constant internal energy, volume, and mole numbers (UVN-flash), and at constant pressure, temperature, and mole numbers (PTN-flash). Furthermore, we develop a numerical method for solving the general formulation of phase-equilibrium problems. This algorithm is based on the direct minimization of the objective function with respect to the constraints. The algorithm uses a modified Newton-Raphson method, along with a modified Cholesky decomposition of the Hessian matrix to generate a sequence of states with decreasing values of the objective function. The algorithm was implemented in C++ and using generic programing we have a single, portable solver for all three flash formulations. Properties of the algorithm are shown on phase-equilibria problems of multicomponent mixtures in different specifications and with different levels of difficulty. Complexities and numerical performance of the individual flash formulations are discussed.

Keywords: C++ templates, general formulation, generic programing, phase equilibrium calculation, phase stability testing, modified Cholesky decomposition, multicomponent mixtures, Newton-Raphson method, optimization, VTN-flash, PTN-flash, UVN-flash

Abstrakt. V tomto článku představíme obecné formulace problémů fázové rovnováhy a fázové stability vícesložkových směsí a ověříme, že tyto formulace zobecňují problémy fázové rovnováhy a stability při kontantním objemu, teplotě a látkovém množství (VTN-specifikace), při konstantní vnitřní energii, objemu a látkovém množství (UVN-specifikace) a při konstantní teplotě, tlaku a látkovém množství (PTN-specifikace). Mimoto jsme vytvořili novou numerickou metodu pro řešení obecké formulace problému fázové rovnováhy a fázové stability. Tento algoritmus je založen na přímé minimalizaci účelové funkce vzhledem k omezujícím podmínkám. Algoritmus využívá modifikanou Newtonovu metodu a modifikovanou Choleského dekompozici Hessovy matice, které generují posloupnost stavů s klesající hodnotou účelové funce. Tento algorimus byl implementován v jazyce C++ a s využitím generického programovaní jsme vytvořili jeden řešič pro všechny tři specifikace. Vlastnosti algoritmu jsou ukázány na úlohách fázové

*Tato práce byla podpořena grantem Investigation of shallow subsurface flow with phase transitions, projekt č. 17–06759S Grantové agentury České republiky, 2017-2019.

rovnováhy vícesložkových směsí s různou obtížností. Složitost jednotlivých specifikací je diskutována.

Klíčová slova: C++ šablony, obecná formulace, generické programování, výpočet rovnovážných stavů, testování fázové rovnováhy, modifikovaná Choleského dekompozice, vícesložkové směsi, Newtonova-Raphsonova metoda, optimalizace, VTN-flash, PTN-flash, UVN-flash

Plná verze: T. Smejkal, J. Mikyška. Unified presentation and comparison of various formulations of the phase stability and phase equilibrium calculation problems. Fluid Phase Equilibria **476**, Part B (2018), 61–88.

Optical Flow-Based Non-Rigid Registration of Cardiac MRI Images*

Kateřina Solovská

1st year of PGS, email: katerina.solovska@gmail.com
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Tomáš Oberhuber, Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. This work deals with non-rigid registration of cardiac MRI images [7, 1], specifically the MOLLI sequences [5]. MOLLI sequence consists of 11 heart images acquired over 17 cardiac cycles. The images of MOLLI sequence are used for pixel-wise estimation of T_1 relaxation time values. The acquisition is ECG triggered and the images are therefore acquired at the same cardiac phase. However, registration is necessary to correct the deformations that occur because of the patient's imperfect breath-holding during the acquisition. The main characteristic of the MOLLI sequence is the evolving intensity of the tissues and also large variation of the image contrast. This characteristics of the sequence make the registration process challenging and make the use of intensity-based registration method impossible.

For this purpose, we propose a method based on optical flow [4, 3], using information obtained by image segmentation. The first step of the registration process, is segmentation of the regions of interest, using the level set method [2, 8]. The segmented objects are represented by distance maps. The transformation between original images is determined by applying the optical flow method to the distance maps. The registration process is independent on the varying intensity and takes into account only the shape and position of the segmented areas, such as the myocardium or the ventricles. We also propose an approach of combining the method of optical flow on distance maps for first part of the MOLLI sequence, where the changes of intensity and contrast are most significant, and optical flow on original images for the second part of this sequence, where the changes of intensity are relatively small.

The implementation of the proposed method is described and the method is tested on several MOLLI sequences. The results are compared to results of the method based on maximisation of mutual information [6]. For images with significant changes in intensity, the better results are obtained by the proposed method.

Keywords: distance function, image registration, image segmentation, optical flow

Abstrakt. Článek se zabývá nerigidní registrací snímků srdce [7, 1] z magnetické rezonance. Konkrétně je řešen problém registrace MOLLI sekvencí [5]. MOLLI sekvence se skládá z 11 snímků srdce pořízených během 17 srdečních cyklů. Snímky MOLLI sekvence jsou využívány pro výpočet hodnoty T_1 v jednotlivých pixelech. Sekvence jsou pořizovány za pomoci EKG triggeringu, což umožňuje pořízení všech snímku ve stejné fázi srdečního cyklu. Pro odstranění deformací způsobených nedostatečných zadržením dechu pacienta v průběhu pořizování sekvence

*This work has been supported by the grant no.NV15-27178A: Quantitative mapping of myocard and of flow dynamics by means of MR imaging for patients with nonischemic cardiomyopathy - development of methodology.

je ovšem potřeba dodatečná registrace. Hlavní charakteristikou MOLLI sekvence je vývoj intenzity jednotlivých tkání a výrazné změny kontrastu na snímcích. Tato vlastnost sekvence znemožňuje použití registračních metod založených na shodě intenzity.

V této práci je navržena metoda založená na výpočtu optického toku [4, 3], využívající segmentaci. Prvním krokem při zpracování sekvence je segmentace klíčových oblastí pomocí vrstevnicové metody [2, 8]. Transformace mezi snímky je poté určena aplikací metody optického toku na distanční mapy reprezentující segmentované objekty. Proces registrace není závislý na změnách intenzity a zohledňuje pouze tvar a polohu segmentovaných objektů, jako je například myokard nebo srdeční komory.

Dále je navržen postup využívající výpočet optického toku mezi distančními mapami pro první část sekvence, kde jsou změny intenzity a kontrastu nejvýraznější, a výpočet optického toku mezi původními snímky pro druhou část sekvence, kde jsou změny intenzity relativně malé.

Je popsám způsob implementace navrhované metody a metoda je otestována na řadě reálných MOLLI sekvencí. Výsledky registrace pomocí navrhované metody jsou srovnány s výsledky metody založené na maximalizaci mutual information [6]. V případě snímků s výrazně rozdílnými intenzitami bylo při použití metody optického toku dosaženo lepších výsledků.

Klíčová slova: distanční funkce, optický tok, registrace obrazu, segmentace obrazu,

Full paper: K. Solovská , T. Oberhuber, J. Tintěra, R. Chabiniok *Optical flow-based non-rigid registration of cardiac MRI images.* Submitted to Discrete and Continuous Dynamical Systems Series S (2018).

References

[1] J. S. Duncan and N. Ayache. *Medical image analysis: Progress over two decades and the challenges ahead.* IEEE transactions on pattern analysis and machine intelligence **22** (2000), 85–106.

[2] L. C. Evans, J. Spruck, et al. *Motion of level sets by mean curvature I.* Journal of Differential Geometry **33** (1991), 635–681.

[3] M. Heinrich, J. Schnabel, F. Gleeson, M. Brady, and M. Jenkinson. *Non-rigid multimodal medical image registration using optical flow and gradient orientation.* Proc. Medical Image Analysis and Understanding (2010), 141–145.

[4] B. K. Horn and B. G. Schunck. *Determining optical flow.* Artificial intelligence **17** (1981), 185–203.

[5] P. Kellman, J. R. Wilson, H. Xue, M. Ugander, and A. E. Arai. *Extracellular volume fraction mapping in the myocardium, Part 1: Evaluation of an automated method.* Journal of Cardiovascular Magnetic Resonance **14** (2012), 63.

[6] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens. *Multimodality image registration by maximization of mutual information.* IEEE transactions on Medical Imaging **16** (1997), 187–198.

[7] T. Makela, P. Clarysse, O. Sipila, N. Pauna, Q. C. Pham, T. Katila, and I. E. Magnin. *A review of cardiac image registration methods.* IEEE Transactions on medical imaging **21** (2002), 1011–1021.

[8] S. Osher and J. A. Sethian. *Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton-Jacobi formulations.* Journal of computational physics **79** (1988), 12–49.

MHFEM with BDDC for Two-Phase Flow in Porous Media in 2D and 3D*

Jakub Solovský

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

advisor: Radek Fučík, Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. This work deals with the application of the domain decomposition method [4] with the BDDC preconditioner [2, 3] for two-phase flow problems in porous media. We briefly describe the spatial discretization of the problem that is based on the mixed-hybrid finite element method (MHFEM) [1] and semi-implicit time discretization. Then in detail, we describe the domain decomposition method with the BDDC preconditioner, discuss the differences between the 2D and 3D cases, and present necessary modifications of the algorithm to improve its efficiency for the more complicated 3D case. We describe the parallel implementation of the method using `MPI` and highlight the critical steps of the algorithm for the performance and scalability. The parallel implementation is then tested on benchmark problems in 2D and 3D and its efficiency is investigated for various meshes. The numerical results on the computational cluster Galileo at CINECA show that the method preserves high efficiency for increasing numbers of processes and, therefore, allows solving problems on very fine meshes.

Keywords: BDDC, domain decomposition, porous media, mixed-hybrid finite element method, MPI, two-phase flow

Abstrakt. Článek se věnuje aplikaci metody domain decomposition [4] s BDDC předpodmíněním [2, 3] na úlohy dvoufázového proudění v porézním prostředí. Nejprve stručně popíšeme prostorovou diskretizaci založenou na smíšené hybridní metodě konečných prvků (MHFEM) [1] a semi-implicitní časovou diskretizaci úlohy. Poté se podrobně věnujeme metodě domain decomposition s BDDC předpodmíněním a rozebereme rozdíly mezi situací ve 2D a 3D a nutné modifikace algoritmu pro zachování efektivity i v komplikovanějším případě ve 3D. Podrobně popíšeme paralelní implementaci využívající `MPI` a popíšeme kritické části tohoto algoritmu, na kterých závisí výkon a škálovatelnost. Paralelní implementace je poté otestována na vybraných úlohách ve 2D a 3D, na kterých měříme efektivitu pro různé numerické sítě. Výsledky na výpočetním klastru Galileo v CINECA ukazují, že metoda zachovává vysokou efektivitu i pro větší počet procesů a umožňuje řešit úlohy na velmi jemných sítích.

Klíčová slova: BDDC, domain decomposition, porézní prostředí, smíšená hybridní metoda konečných prvků, MPI, dvoufázové proudění

Full paper: J. Solovský, R. Fučík, J. Šístek. *MHFEM with BDDC for two-phase flow in*

*The work was supported by the Czech Science Foundation project no. 17-06759S: Investigation of shallow subsurface flow with phase transitions, by grant No. SGS17/194/OHK4/3T/14 of the Grant Agency of the Czech Technical University in Prague, and by the Project HPC-EUROPA3 (INFRAIA-2016-1-730897), with the support of the European Community - Research Infrastructure Action under the H2020.

porous media in 2D and 3D. Submitted to Discrete and Continuous Dynamical Systems Series S (2018).

References

[1] F. Brezzi and M. Fortin. *Mixed and Hybrid Finite Element Methods*. Springer-Verlag, (1991).

[2] C. R. Dohrmann. *A Preconditioner for Sustructuring Based on Constrained Energy Minimization*. SIAM Journal of Scientific Computing **25** (2003), 246–258.

[3] M. Hanek, J. Šístek, and P. Burda. *An application of the BDDC method to the Navier-Stokes equations in 3-D cavity*. Programs and Algorithms of Numerical Mathematics, Proceedings of Seminar. Dolní Maxov, June 8-13, 2014 (2015), 77–85.

[4] A. Toselli and O. Windlund. *Domain Decomposition Methods – Algorithms and Theory*. Springer-Verlag, (2005).

Hurst Index and P-Variation*

Václav Svoboda

3rd year of PGS, email: svobova5@fjfi.cvut.cz
Department of Physics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisor: Petr Jizba, Department of Physics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. The concept of scaling and self-similarity has been widely studied in number of fields. Closely related is a notion of Hurst exponent, there is a considerable amount of literature regarding theoretical properties of Hurst exponent, its role in phase transitions and its various generalizations [11].
The theoretical definition of Hurst exponent depends on the properties of the underlying process on infinitesimal scale. For practical purposes we however need to estimate the Hurst exponent from empirical discrete data. Number of methods has been proposed, we in particular look at few them.
We examine three methods. First discussed method is DMA, which is considered to be stable, computationally effective and one of the most efficient methods existing today. The second one is method based on p-variation which is derived from properties of fractional processes. Third one is simple method based on direct measurement of diffusion parameter D. We introduce and compare these methods by applying them on artificially generated data.
The methods we present can be applied for any type of data, however we aim to apply these to financial time series in the future. Therefore we will focus on relatively short time series (tens of thousands data points), so the results are relevant to standardly analyzed data in finance. We will evaluate the Hurst index on rolling window subset of original data and study the evolution of Hurst exponent when scaling properties of underlying process change.

Abstrakt. Koncepty samopodobnosti a škálování jsou velmi době prostudovány. Úzce s nimi souvisí koncept Hurstova exponentu, jehož teoretické vlastnosti jsou také velmi dobře známy.
Teoreticky je Hurstuv exponent definován na základě infinitisimálních vlastností studovaného procesu. Pro reálné aplikace musíme ovšem být schopni určit Hurstuv exponent na základě empirických, diskrétních dat.
Podíváme se na tři metody pro odhad Hurstova exponentu. Prvni je DMA, metoda považovaná za jednu z nejspolehlivějších v současnosti. Druhá metoda využívá teoreticke vlastnosti p-variace frakčních procesu, poslední metoda přímočaře odhaduje difusni parametr D.
Tyto metody následně aplikujeme na uměle vygenerovaná data a sledujeme vývoj Hurstova exponentu na rolling-window podmnožine vygenerovaných dat.

Keywords: Hurst exponent, DMA, p-variation, Fractional processes, Transformation of statistics, Finacial time series

*This work was supported by the Grant Agency of the Czech Technical University in Prague, grant No. SGS16/239/OHK4/3T/14 and by Czech Science Foundation Grant No. 17-33812L.

1 Introduction

Mandelbrot originally defined Hurst index as scaling of the variance of stochastic process

$$\sigma_t^2 = \int_{\mathbb{R}} x^2 p(x,t) dx = ct^{2H} \tag{1}$$

where $p(x,t)$ is density and H is Hurst exponent, $H = 1/2$ corresponds to Brownian motion while $H > 1/2$ leads to super-diffusive behaviour, $H < 1/2$ to sub-diffusion. In terms of financial markets $H = 1/2$ corresponds to effective markets.
This implies

$$p(t,x) = \frac{1}{t^H} F(\frac{x}{t^H}) \tag{2}$$

If we assume underling process to be Ito diffusion in form

$$dX_t = \sqrt{D(x,t)} dB_t \tag{3}$$

then using Ito calculus [15] we get

$$\sigma_t^2 = \int_0^t ds \int_{\mathbb{R}} p(x,s) D(x,s) dx \tag{4}$$

which implies

$$D(x,t) = t^{2H-1} D(\frac{x}{t^H}) \tag{5}$$

If the underlying process has stationary increments ie. $E(X(t+\Delta t) - X(t)) = c\Delta t^{2H}$ we get the following for correlation function

$$C(0,\Delta t) = \frac{E(X(t)X(t+\Delta t))}{EX^2(t)} = 2^{2H-1} - 1 \tag{6}$$

This implies that for every diffusive self-similar process driven by BM and with stationary increments either $H = 1/2$ or the process has long-distance correlations.
Case $H = 1/2$ is the BM, the case $H \neq 1/2$ corresponds to FBM, in this case the diffusion kernel $D(x,t)$ depends also on the past values of x leading to auto correlations and non-Markovian processes. Interestingly if we omit the condition of stationary increments class of Markov diffusion processes can be constructed with $H \neq 1/2$, so to use empirical value of H as proof of autocorrelation for Ito diffusion processes is possible only if condition of stationary increments is satisfied.
More generally we can define the self-similarity as

$$X_t \stackrel{d}{=} t^H X_1 \tag{7}$$

which is equivalent to condition (2) which is equivalent to condition (1) only if the variance is finite.
This is the second mechanism leading to self-similarity, when the driving noise is not BM but heavy tailed *alpha*-stable process. Stable processes are self-similar subclass of Levy processes (ie. processes with independent and stationary increments) with asymptotic behaviour given by

$$p_\alpha(x) \sim \frac{1}{|x|^{\alpha+1}} \quad |x| \to \infty \quad \alpha \in (0,2) \tag{8}$$

which implies infinite variance.

2 Hurst exponent, fractional processes and p-variation

The fractional processes are processes used to describe anomalous diffusion, ie. diffusion where standard Brownian scaling $\langle x^2(t)\rangle \sim \sigma t$ doesn't work.
There are essentially two main approaches to fractional diffusion. The first one focuses on PDF of underlying problem as process is defined using using fractional Fokker-Planck equations, for example [1]

$$\frac{\partial W}{\partial t} = {}_0D_t^{1-\gamma} K_\alpha \frac{\partial^2}{\partial x^2} W(x,t) \tag{9}$$

The other approach that we will follow focuses more on trajectories of process itself, the process is defined via stochastic equations, this approach has certain advantages when we want to focus on path properties of the process (Hurst exponent, fractal dimension etc.). For example the FBM motion we discussed can be defined as integral

$$B_H(t) = \int_{\mathbb{R}} \left((t-x)_+^d - (-x)_+^d\right) dB(x) \tag{10}$$

where $d = H - 1/2$ is a diffusion parameter determining the type of diffusive behaviour and $(x)_+ = max(x, 0)$. It can be easily shown that FBM is Gaussian H-self similar process with stationary increments and its increments are positively correlated for $H > 1/2$ and negatively for $H < 1/2$. Notice that condition (5) is satisfied for FBM.
More generally we can assume stable fractional processes as

$$L_H^\alpha(t) = \int_{\mathbb{R}} \left((t-x)_+^d - (-x)_+^d\right) dL_\alpha(x) \tag{11}$$

where L_α is α-stable symmetric process, $d = H - 1/\alpha$ with $H \in (0, 1)$.
Fractional stable processes are H-self similar processes with stationary increments. They can be described via their characteristic function [4]

$$\varphi_t^{H,\alpha}(z) = e^{-(ct^H|z|)^\alpha} \tag{12}$$

To be able to estimate empirically H of fractional process we will introduce autoregressive fractionally integrated moving average model (ARFIMA) [5,13], which is discrete time analogue of fractional processes. It generalizes the standard linear ARMA model in two ways. The general form of ARFIMA model is

$$\mathcal{A}_p(B)X_t = \mathcal{B}_q(B)(1-B)^{-d} Z_t \tag{13}$$

where B is a lag operator, $\mathcal{A}, \mathcal{B}$ are polynomials of order p respectively q and Z_t are iid α-stable variables with $\alpha > 1$ representing random noise.
The term $(1-B)^{-d}$ is defined via Taylor expansion as

$$(1-B)^{-d} Z_t = \sum_{i=0}^{\infty} \frac{\Gamma(i+d)}{\Gamma(i)\Gamma(d+1)} Z_{t-i} \tag{14}$$

We denote the above defined model as $ARFIMA(p, d, q, \alpha)$, for it to be correctly specified (converge a.s.) the following must hold [6]

$$0 < H = d + 1/\alpha < 1 \tag{15}$$

Furthermore if roots of polynomial $\mathcal{A}_p$ lie outside of unit circle the ARFIMA process is stationary.
Stationary ARFIMA process is asymptotically H self-similar with $H = d + 1/\alpha$. The most important result for us is the following limiting relation, let X be ARFIMA process then

$$N^{-H} \sum_{i=1}^{\lfloor Nt \rfloor} X_i \xrightarrow{\mathcal{D}} L_H^{\alpha}(t) \quad N \to \infty \tag{16}$$

So ARFIMA model can be considered as discrete time version of Levy fractional processes.
The case $d = 0$ leads to ARMA processes (with α-stable noise) and exponentially decaying autocorrelation functions. The case $d > 0$ is similar to the case of fractional Brownian motion and leads to long range dependence.

P-variation of fractional process

We will introduce notion of p-variation and asymptotic results for behaviour of p-variation of ARFIMA process in this section.
First we define general sample p-variation of process $X_{t \in <0,T>}$, $t_k = T * k2^m,\ k = 0, .., 2^m$

$$V_m^p = \sum_{i=1}^{k} |X(t_i) - X(t_{i-1})|^p \tag{17}$$

The p-variation is then defined as

$$V^p = \lim_{m \to \infty} V_m^p \tag{18}$$

The p-variation measures process fluctuation so its connection with Hurst index is not surprising.
For empirical discrete data we define sample p-variation of process $X_{i \in \{1..N\}}$ of lag m as [16]

$$V_m^p = \sum_{i=0}^{N/m-1} |X_{(i+1)m} - X_{im}|^p \tag{19}$$

Let us assume that X is cumulative sum process of stationary ARFIMA process, then for sufficiently large N/m it holds [6]

1. if $\alpha = 2$ or if $1 < \alpha < 2$ and $d \geq 0$

$$V_m^p \sim m^{Hp-1} \tag{20}$$

2. if $1 < \alpha < 2$ and $d < 0$

$$V_m^p \sim m^{Hp-p/\alpha} \tag{21}$$

It worth noticing that in the first case variation increases with growing p but it decreases in the second case. Also in the first case the variation V_p converges to finite number for $p = 1/H$ and diverges only for $p > 1/H$ in the second case V_p diverges always. This is caused by the fact that in the second case the process has unbounded trajectories.

3 Empirical estimation of H

We introduce three methods for empirical estimation of Hurst index in this section. The performance of these methods will be discussed in next section when we apply them to artificial data.

3.1 DMA

The most common methods like RS-estimate or DFA start by dividing data into bins first. In the case of RS method we look at range of the process inside these bins, DFA performs detrending inside the bins and measures fluctuations of remaining noise.
The idea behind DMA [17] is similar but it seems to work bit better then these methods, we define

$$\sigma^2_{DMA} = \frac{1}{N-n}\sum_{t=n}^{N}(x(t) - \widetilde{x}_n(t))^2 \tag{22}$$

where N is length of the whole time series and

$$x_n(t) = \frac{1}{n}\sum_{k=0}^{n-1} x(t-k) \tag{23}$$

So in comparison with DFA we remove 'trend' given by moving average instead of determining the trend via regression.
We calculate the σ_{DMA} for multiple values of n, the following relation should hold

$$\sigma^2_{DMA} \sim n^H \tag{24}$$

So we can obtain H by running regression $\log \sigma^2_{DMA} \sim \log n$. The range of n we use to estimate σ_{DMA} is quite important and can change the result significantly, after some testing we decided to take $n = 10, .., 100$ for samples containing thousands of data points.

3.2 P-variation based estimation

If the fractional process satisfies the following:
$\alpha = 2$ or if $1 < \alpha < 2$ and $d \geq 0$ then

$$V^p_m \sim m^{Hp-1} \tag{25}$$

In this case we propose the following method for estimation of H

1. Estimate V^p_m for $p = 1/\{0.01, 0.02..1\}$

2. For fixed m find p that minimizes $(\frac{V_m^p - V_1^p}{V_1^p})^2$

3. estimate $H = 1/p$

The appropriate choice of m has to done based on sample size, generally it is better to choose larger m as long as N/m remains sufficiently large, we used default value $m = 3$ and it seems to work reasonably.
The second case $1 < \alpha < 2$ and $d < 0$ can be theoretically transformed into the first case by using concept of surrogate data, which means essentially reshuffling the (stationary increment) data. That should break the correlation structure within the data and essentially set $d = 0$, however it is not really needed because in this case the process has unbounded trajectories and other properties undesirable for modeling real world time series.

3.3 Diffusion parameter based estimation

Another very simple approach is to directly estimate parameter d using sample variance

$$M_N(t) = \frac{1}{N - t - 1} \sum_{i=0}^{N-t} (X_{i+t} - X_i)^2 \tag{26}$$

If the process X_t is cumulative sum process of stationary ARFIMA process (with $\alpha > 1$) then the following asymptotic relation holds

$$M_N(t) \sim t^{2d+1} \tag{27}$$

Notice the case $d < 0$ with non-Gaussian noise, in this case the large jumps produced by stable noise are compensated by large jumps of opposite sign and on average the diffusion of the process is slower that in the standard Brownian case.
The proposed method of estimation of parameter d is the following:

1. Estimate $M_N(t)$ for $t = 1, 2, .., 10$

2. Run regression $\ln M_N(t) \sim \ln t$

3. Take the calculated slope δ and calculate $d = \frac{\delta - 1}{2}$

The proposed estimator is consistent, it has been tested and seems to produce reliable results.
To estimate $H = d + 1/\alpha$ we still need to estimate the stability index α. One approach would be to assume $\alpha = 1/2$, for example methods like RS, DFA or DMA assume this implicitly. However we decided to estimate α directly using Koutrouvelis regression method for estimation of parameters of stable distributions.This method essentially fits the parameters by running regression on empirical characteristic function. The method seems to be stable and computationally quite efficient compared to MC based methods.

4 Application to artificial data

In this section we will apply the above introduced methods to artificial data. In particular we test the methods on BM, FBM with $H = 0.3$ and $H = 0.7$ and on α-stable process with $\alpha = 1.8$ implying $H \simeq 0.56$. The last method estimating d and α separately should be taken in perspective because it uses parametric method to estimate stability index which obviously works well on artificial data but might be misleading on real data.

Brownian motion

First we generated trajectory of BM with 50000 data points. We set the $m = 3$ for the p-variation method and $n = 10, .., 100$ for DMA. We estimated Hurst index on rolling window subset of size 5000 data points and in each iteration moved the window by 100 data points. We got the following time evolution of Hurst index.

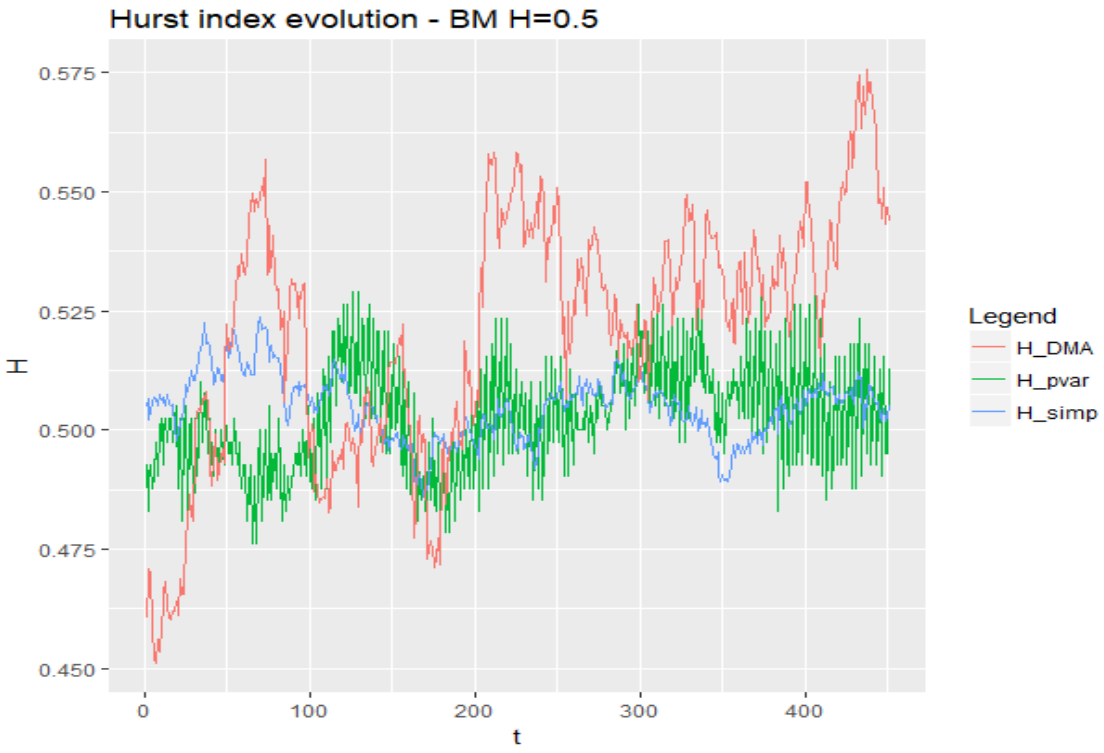


Figure 1: Hurst index evolution BM

The green line in Figure 1 corresponds to method based on p-variation, red one to DMA method and blue one to the method estimating d and α separately.
As we can seen the p-variation method is more accurate in measuring true Hurst exponent then DMA but at the same time it has much higher fluctuations. The time evolution of H would be obviously smoother for longer series however we purposely study shorter time series which are more useful for applications to financial markets.

Fractional Brownian motion

We performed same analysis for FBM both super-diffusive and sub-diffusive case. In both cases we obtained similar results, again method based on p-variation is more accurate but oscillates quite a lot. The results for super-diffusive case are depicted in Figure 2.

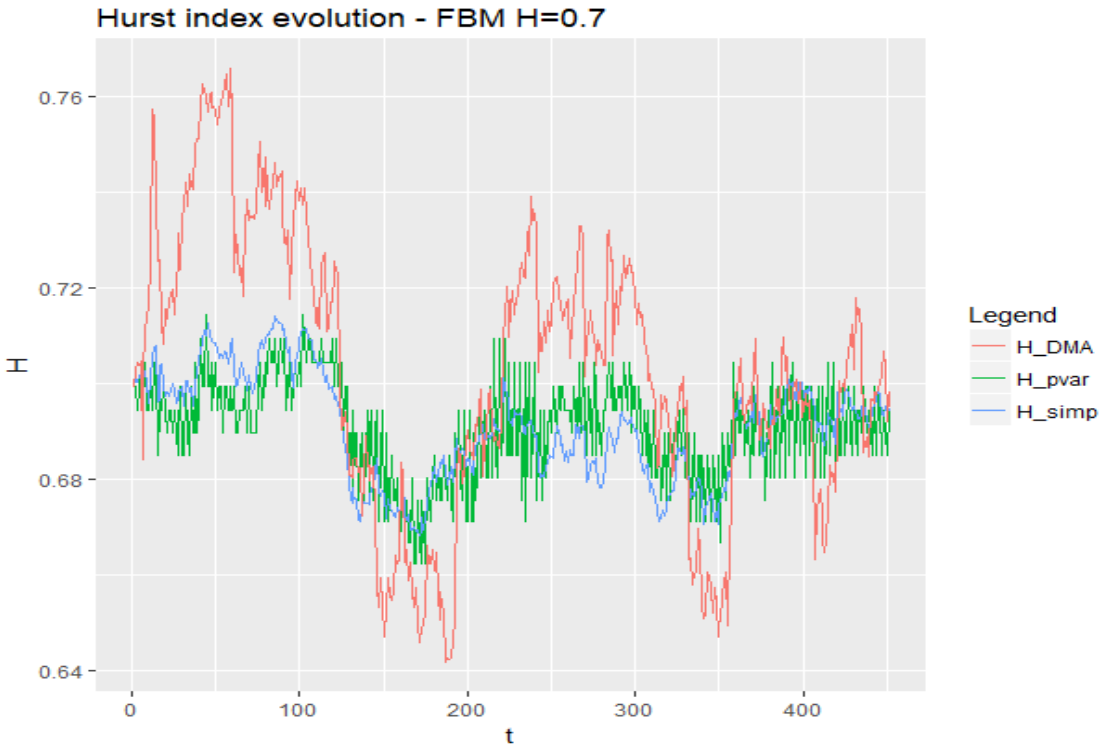


Figure 2: Hurst index evolution FBM

Stable process

Situation for stable processes is much more tricky, it is complicated to measure fractionality caused by infinite variance on finite samples which leads to problems.
We generated again 50K data points of stable process with $\alpha = 1.8$. We got the following time evolution of Hurst index.

As you can see p-variation method fails but also DMA method gives some strange results. This is caused by extreme values that can be generated from stable distribution, especially p-variation method is (on shorter datasets) extremely effected by them.
If we remove these 'extreme observations' from dataset we get 'normal' evolution of Hurst index, for example if we damp 1% of extreme observation (Figure 5.) then the p-variation method gives Brownian like result and p-variation fluctuates only little bit more then in Brownian case.

Change of regime

We will calculate the evolution of Hurst exponent for series that changes its dynamic and observe how these transitions affect evolution of Hurst exponent.
In particular first we applied the methods to series that starts as BM but then quickly

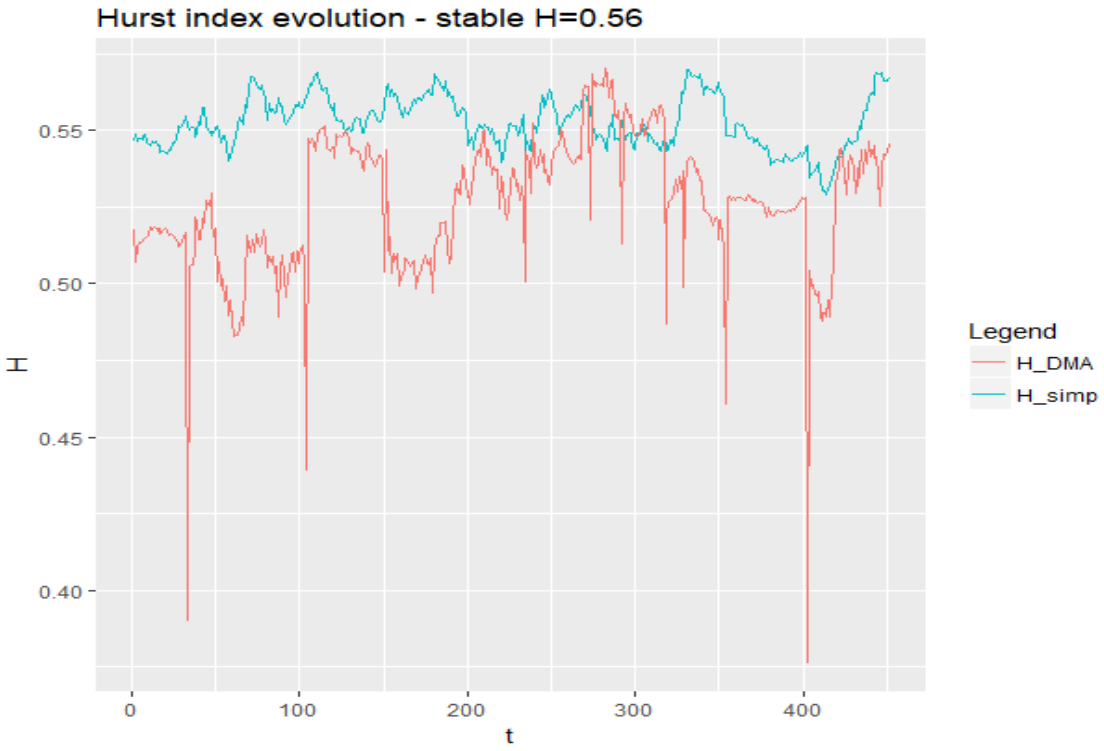


Figure 3: Hurst index evolution stable process

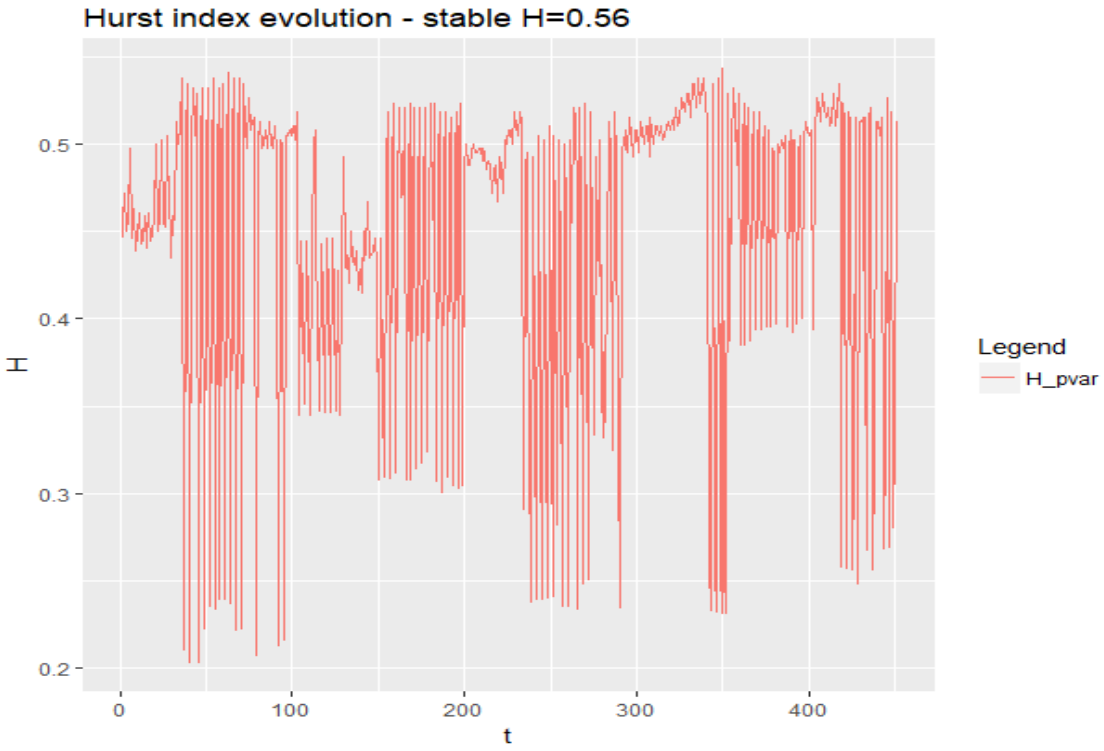


Figure 4: Hurst index evolution stable process

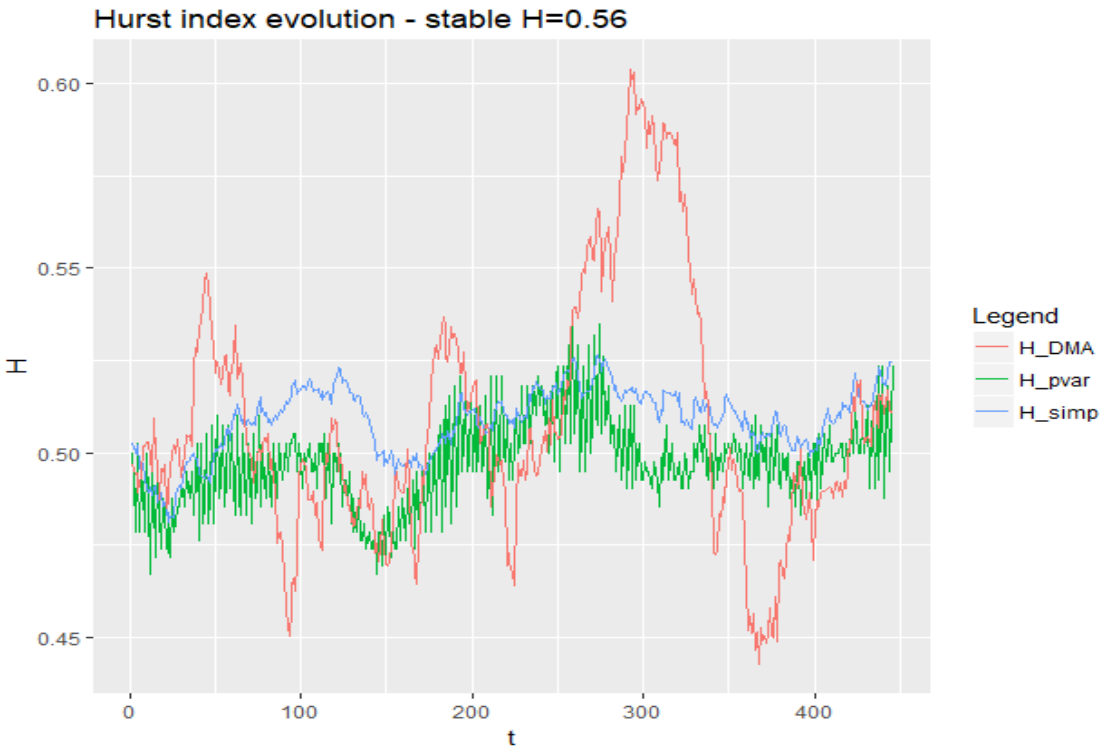


Figure 5: Hurst index evolution stable damped process

switches to sub-diffusive regime with $H = 0.3$. All methods reliably captured this (Figure 6.).

More interesting is a switch from BM to stable process with $\alpha = 1.8$. The results in Figure 7. show that DMA captures the regime switch but if we did not know that the regime change occurs there it wouldn't be clear if the change is relevant. On the other hand while p-variation based method fails to give accurate estimate of H its evolution changes significantly and it clearly indicates regime change.

Conclusion

The main goal of this paper was to investigate empirical performance of proposed method for estimation of Hurst exponent based on p-variation and compare it to well known DMA method and simple method based on directly measuring variance scaling.
For FBM is the p-variation method very precise and outperforms DMA. However it is quite sensitive to extreme values which leads to complete breakdown when applied to stable processes, however also DMA gives not so great results for stable processes and in general all methods using path properties only don't capture fractionality caused by infinite variance very well.
The main upside of DMA however is that it can handle processes with non-stationary increments, the other methods are not very well suited for that. For example when applied on moving average of BM, DMA gives values around $H = 0.5$ while the other two methods give unrealistically high H.

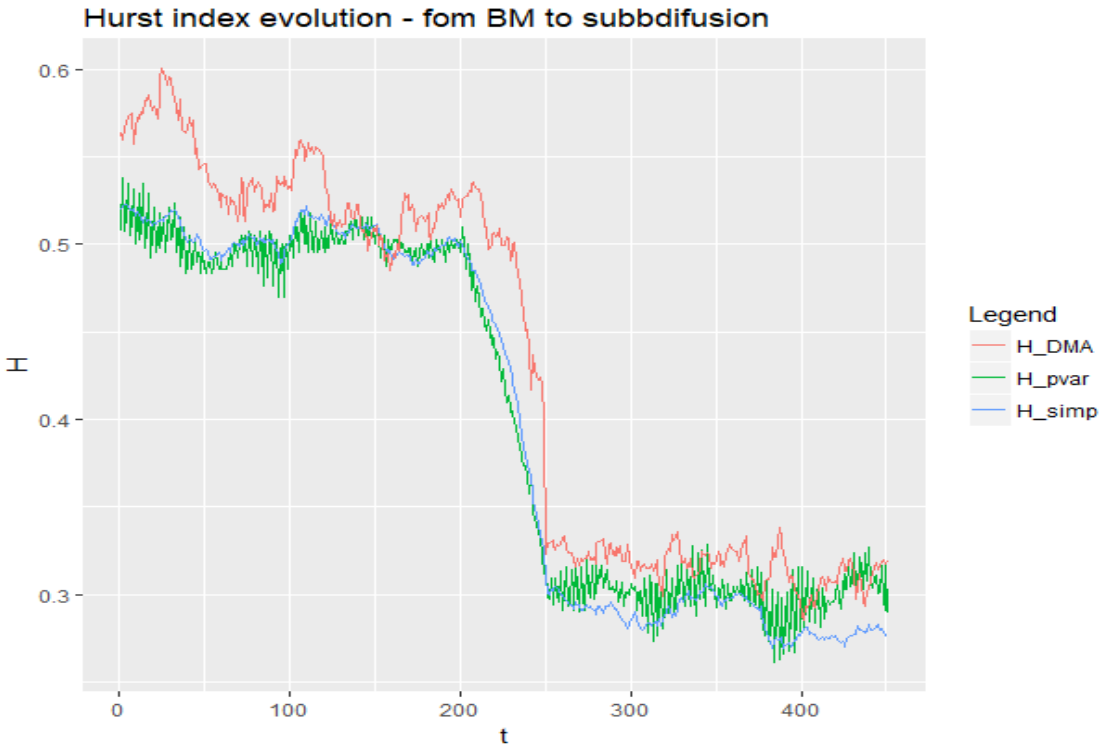


Figure 6: Hurst index evolution regime switch

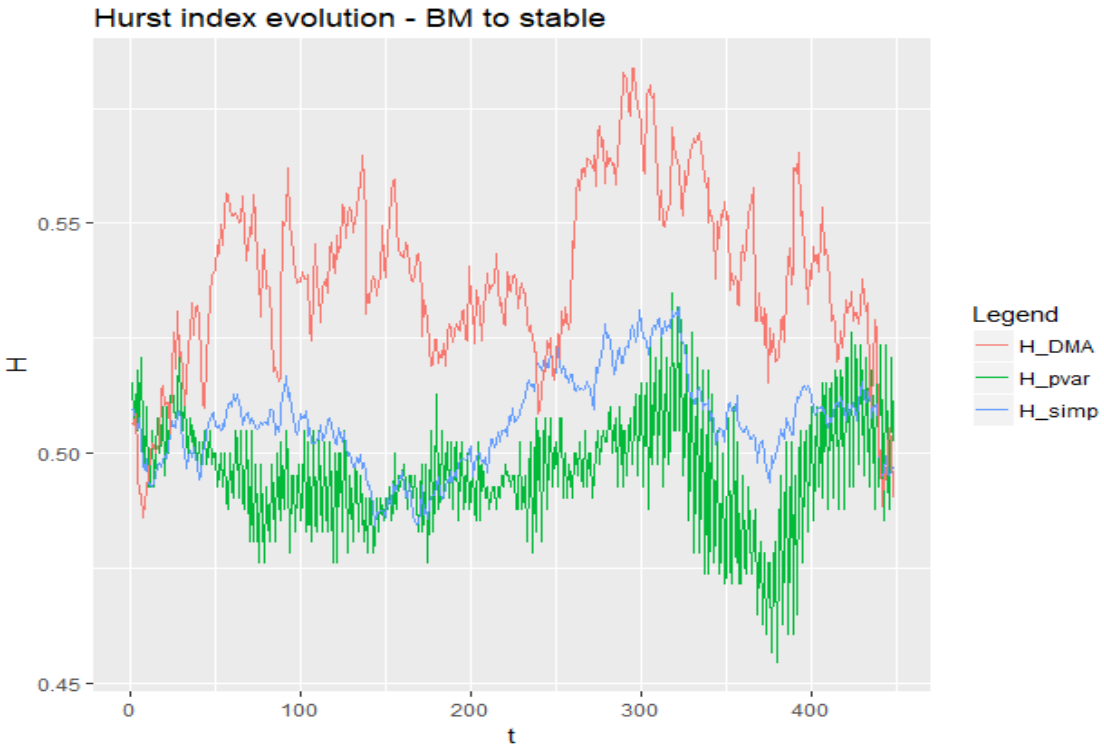


Figure 7: Hurst index evolution regime switch

So overall the DMA method is more useful but is less accurate for FBM. Also the p-variation method has higher sensitivity to outliers and other anomalies which is often undesirable but can be possibly used to detect anomalies and regime changes in time series.

References

[1] R. Metzler, J. Klafter. *The random walks's guide to anomalous diffusion: A fractional dynamics approach.* Physics reports 339 1-77, 2000.

[2] P. Tankov. *Financial modelling with jump processes.* Chapman-Hall/CRC, 2003.

[3] K. S. Miller, B. Ross. *An introduction to the fractional calculus and fractional differential equations.* Wiley, 1993.

[4] M. Teuerle, A. Wylomanska, G. Sykora *Modelling anomalous diffusion by subordinated Levy stable processes.* Journal of statistical mechanics: Theory and experiment. 10.1088/1742-5468/2013/05/P05016.

[5] Granger, C. W. J.; Joyeux, R. . *An introduction to long-memory time series models and fractional differencing.* Journal of Time Series Analysis (1980). 1: 15–30.

[6] K. Burnecki, A. Weron. *Algorithms for testing of fractional dynamics: a practical guide to ARFIMA modelling.* Journal of statistical mechanics: Theory and experiment. 10.1088/1742-5468/2014/10/P10036.

[7] T. Marquardt. *Fractional Levy processes with an application to long memory moving average process.* Bernoulli 12(6), 2006, 1099-1126.

[8] K. Falconer. *Fractal geometry, Mathematical foundations and applications.* Wiley, 1989.

[9] R. N. Mantegna, H. E. Stanley. *An introduction to econophysics.* CUP, Cambridge, 2000.

[10] B. V. Gnedenko, A. N. Kolmogorov. *Limit distributions for sums of independent random variables.* Adison-Wesley, 1968.

[11] B.B. Mandelbrot. *Fractals and scaling in finance.* SELECTA VOLUME E, 1996.

[12] W.Paul, J. Baschnagel. *Stochastic Processes: From physics to finance.* Springer, 2000.

[13] Hosking, J. R. M. *Fractional differencing.* Biometrika. 68 (1): 165–176, 1981.

[14] M. Matsui, A. Takemura. *Goodness of fit tests for symmetric stable distributions.* http://www.e.u-tokyo.ac.jp/cirje/research/03research02dp.html.

[15] B.Oksendal. *Stochastic differential equations.* Springer, 1994.

[16] R. Norvaise, D.Salopek. *Estimating the p-Variation Index of a Sample Function: An Application to Financial Data Set.* Methodology And Computing In Applied Probability, March 2002, Volume 4, Issue 1, pp 27–53.

[17] A. Carbonea, G. Castellia, H.E. Stanley. *Time-dependent Hurst exponent in financial time series.* Physica A 344 (2004) 267–271.

[18] Kevin E. Bassler, Gemunu H. Gunaratne, Joseph L. McCauley. *Markov Processes, Hurst Exponents, and Nonlinear Diffusion Equations.* Physics Department, University of Houston.

[19] Marcin Magdziarz, Jakub Karol Slezak and Justyna Wójcik. *Estimation and testing of the Hurst parameter using p-variation.* Journal of Physics A Mathematical and Theoretical, July 2013.

Independent Component Extraction Using Nonstationarity and Multiplicative Update*

Ondřej Šembera

4th year of PGS, email: sembera@utia.cas.cz
Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

advisors:

Petr Tichavský, Department of Stochastic Informatics
Institute of Information Theory and Automation, CAS

Zbyněk Koldovský, Institute of Information Technology and Electronics
Faculty of Mechatronics, Informatics, and Interdisciplinary Studies, TUL

Abstract. This paper deals with the problem of extracting one independent nonstationary source from an instantaneous linear mixture. We propose a parameterization of the de-mixing matrix and show that other parameterizations do not lead to better results. We design four algorithms for the estimation of the de-mixing matrix based on the maximum likelihood principle and nonstationarity modeling of the sources. In the experiments, we test the robustness of the algorithms against inaccurate initialization. The two proposed algorithms show promising results.

Keywords: blind source extraction, independent component analysis, nonstationarity modeling

Abstrakt. Tento článek se zabývá extrakcí jednoho nestacionárního zdroje z okamžité lineární směsi. Navrhujeme parametrizaci demixující matice a ukazujeme, že neexistují lepší parametrizace. Dále odvozujeme dvě metody extrakce založené na principu maximální věrohodnosti a modelování nestacionarity zdrojů. V experimentální části testujeme robustnost těchto metod vůči nepřesné inicializaci.

Klíčová slova: Slepá extrakce signálu, analýza nezávislých komponent, modelování nestacionarity

1 Introduction

Blind source separation (BSS) and blind source extraction (BSE) represent a wide class of algorithms and techniques that have found their applications in many areas [1]. The main task is to separate independent signals or signal subspaces from their linear mixture with a little or no information about the separated sources and the mixing parameters.

In the BSE, the goal is to separate only one desired signal or signal subspace from the mixture. Some BSS methods reside in performing as many BSE tasks as is the number of signals to be extracted, while constraining the BSE units so that each one extracts a different signal [2, 4]. The BSS and BSE techniques differ from each other in principles and statistical models of the desired signals of interest that enable us to

*This work has been supported by The Czech Science Foundation through Project No. 17-00902S

achieve the separation/extraction. There are three main principles: non-Gaussianity, spectral diversity, and nonstationarity [8, 11]. The BSS algorithms for the latter two models are mostly based on an approximate joint diagonalization of covariance matrices of the mixed signals [8, 10]. They mainly differ in speed and accuracy, but only a few approaches based on these principles exist for the BSE [7]. One goal of this paper is to explore the BSE problem based on the signals' independence and nonstationarity.

Performance of the BSE methods can be compared to that of the BSS methods provided that the desired signal obtained through BSE is one of the signals separated by BSS. The BSS methods may, in general, be more computationally complex because they estimate more parameters than BSE. On the other hand, in BSE, it may happen that the method would separate another independent source in different from the desired one. Therefore, one important performance criterion of the BSE algorithms is their stability, which specifies the probability of separation of the desired component, or vice versa, the probability of a failure due to estimating an unwanted signal instead. Also, BSE methods can be less accurate as they model the observed mixture with a lower number of parameters than BSS methods [11, 5].

In this paper, we solve the BSE problem through assumed nonstationarity of the sources, as this is a strong property of some real-world signals such as speech. Our approach is similar to that used in [6]; it will be referred to as Independent Component Extraction (ICE). The de-mixing matrix model is structured using the minimum number of parameters that are necessary for the signal extraction and two BSE algorithms are proposed. The proposed algorithms are compared with BGSEP algorithm [10] and an algorithm of Shalvi and Weinstein [9], which was originally designed for the blind identification of one non-stationary signal from its mixture with stationary interference and has been very popular in audio separation problems; see, e.g., [3]. For an extended version of the paper see [12].

The rest of the paper is organized as follows: In Section II. we introduce the problem of independent component extraction and explore the ambiguity of the solution to the problem. Then we describe a non-stationary Gaussian model of the independent sources and derive a contrast function using the maximum likelihood principle. We conclude the Section II. by introducing specific parameterization of the demixing matrix. In Section III. we describe the proposed algorithms. First, we derive a method using a multiplicative update and nonstationarity modeling of both the signal of interest and the interference. Next, we discuss the possible simplification introduced by assuming stationary interference. Finally, in Section IV. we test the robustness of the algorithms to an inaccurate initialization.

We use italic lowercase letters to denote scalars, bold lowercase letters for vectors and bold uppercase ones for matrices. The elements of a vector are denoted by italic lowercase letters with lower subscripts. Thus the i-th element of a vector $\mathbf{x}$ would be denoted by x_i. For submatrices, we will use bold uppercase letters with two lower subscripts. We will use the Matlab notation to express the range of indices. For example, if $\mathbf{A} \in \mathbb{R}^{I \times J}$ is a matrix and $1 \leq i_1 < i_2 \leq I$, $1 \leq j_1 < j_2 \leq J$, then $\mathbf{A}_{i_1:i_2,j_1:j_2}$ is a submatrix of $\mathbf{A}$ with rows in the range from i_1 to i_2, and columns in the range fromj_1 to j_2. The i_1-th row of $\mathbf{A}$ would be denoted as $\mathbf{A}_{i_1,:}$ the i,j-th element of A as $\mathbf{A}_{i,j}$. We will use a trace operator $\mathrm{tr}(\,\cdot\,)$ defined for a square matrix as a sum of its diagonal elements.

2 Independent component extraction

2.1 Problem of ICE

In this paper, we deal with determined linear instantaneous mixtures of independent sources, i.e., the observed data $\mathbf{X} \in \mathbb{R}^{d \times N}$ follow the generating model

$$\mathbf{X} = \mathbf{A}\mathbf{U}, \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{d \times d}$ is an unknown nonsingular mixing matrix, later called the mixing matrix, $\mathbf{U} \in \mathbb{R}^{d \times N}$ is the matrix of the sources. The columns $\mathbf{x}_t$ of $\mathbf{X}$ and $\mathbf{u}_t$ of $\mathbf{U}$ correspond to the samples of the observed mixtures and the samples of the sources, respectively. The sources, i.e., the rows of $\mathbf{U}$, are assumed to be independent.

In contrast to independent component analysis, which aims at estimating all the sources $\mathbf{U}_{1,:} \dots \mathbf{U}_{d,:}$, the independent component extraction aims at estimating just one of the sources, from now on called the signal of interest or SOI, and to extract it from the mixtures. Without any loss of generality let us assume that $\mathbf{s} = \mathbf{U}_{1,:}$ is the signal of interest and $\mathbf{Z} = \mathbf{U}_{2:d,:}$ is the interference. Let $\mathbf{W} \in \mathbb{R}^{d \times d}$ be a nonsingular matrix and let

$$\begin{pmatrix} \tilde{\mathbf{s}} \\ \tilde{\mathbf{Z}} \end{pmatrix} = \mathbf{W}\mathbf{X}. \tag{2}$$

The goal of ICE is to find a nonsingular matrix $\mathbf{W}$ such that the following extraction conditions hold:

1. $\tilde{\mathbf{s}}$ is equal to SOI up to the scaling by a nonzero factor,
2. $\tilde{\mathbf{Z}}$ is independent of $\tilde{\mathbf{s}}$.

The matrix $\mathbf{W}$ is then called a demixing matrix. It can be shown that, if the extraction conditions hold, then $\tilde{\mathbf{Z}}$ is equal to $\mathbf{Z}$ up to a multiplication by a nonsingular matrix, i.e., there is a nonsingular matrix $\mathbf{B} \in \mathbb{R}^{(d-1)\times(d-1)}$ such that $\tilde{\mathbf{Z}} = \mathbf{B}\mathbf{Z}$ holds. Performing ICE thus gives us an estimate of the signal of interest and a $d-1$ - dimensional mixture of the interference sources $\tilde{\mathbf{Z}}$ (which we will refer to as interference from now on). The ICE can then again be used on the mixture $\tilde{\mathbf{Z}}$ to yield another source estimate, say, $\mathbf{U}_{2,:}$. This can be repeated until the full separation is achieved, i.e., all the sources $\mathbf{U}_{1,:}, \dots, \mathbf{U}_{d,:}$ have been estimated. However, the full ICA algorithms are better suited to this task, and the main purpose of the ICE algorithm proposed bellow is the separation of just one of the components.

A demixing matrix exists since, for example, $\mathbf{W} = \mathbf{A}^{-1}$ meets the extraction conditions 1. and 2., and it is not unique. Given any demixing matrix $\mathbf{W}_1$, any nonzero scalar α, and any nonsingular matrix $\mathbf{B} \in \mathbb{R}^{(d-1)\times(d-1)}$, matrix

$$\mathbf{W}_2 = \begin{pmatrix} \alpha & \\ & \mathbf{B} \end{pmatrix} \mathbf{W}_1 \tag{3}$$

is also a demixing matrix. The sought demixing matrix is thus unique only up to an action of a $(d^2 - 2d + 2)$-parametric group of transformations, which leaves us with only $2d - 2$ parameters to estimate. The section 2.4 deals with the parameterization in detail.

2.2 Modelling the source signals

We place the following assumptions on the signal of interest $\tilde{s}$ and the interference $\tilde{\mathbf{z}}$: the time domain can be divided into M blocks of the same length T such that, in the m-th block, the signal of interest is an i.i.d. sequence of Gaussian variables with the zero mean and a variance σ_m^2, and the interference is an i.i.d. sequence with joint Gaussian distribution with the zero mean and a covariance matrix $\mathbf{C}_m$. Hence, for $t = (m-1)T + 1 \dots mT$, we get $\tilde{s}_t \sim N(0, \sigma_m^2)$, $\tilde{\mathbf{z}}_t \sim N_{d-1}(0, \mathbf{C}_m)$ and the probability densities of the SOI and the interference take on the form

$$\begin{aligned} p_s(\tilde{s}_t|\sigma_m^2) =& \frac{1}{\sqrt{2\pi\sigma_m^2}} \exp\Big[-\frac{\tilde{s}_t^2}{2\sigma_m^2}\Big], \\ p_{\mathbf{z}}(\tilde{\mathbf{z}}_t|\mathbf{C}_m) =& \frac{1}{\sqrt{2\pi}^{d-1}\sqrt{\det \mathbf{C}_m}} \exp\Big[-\frac{1}{2}\tilde{\mathbf{z}}_t^T \mathbf{C}_m^{-1} \tilde{\mathbf{z}}_t\Big]. \end{aligned} \tag{4}$$

The successful extraction of SOI depends on its nonstationarity, i.e., there have to be at least two blocks indexed by m and n such that $\sigma_m^2 \neq \sigma_n^2$. On the other hand, the interference can be assumed to be stationary, which occurs when the covariance matrix of interference is constant, $\mathbf{C}_m = \mathbf{C}$. While this stationarity assumption may lead to a worse approximation of the true interference distribution, and thus to worse separation, it also yields a smaller number of nuisance parameters and lower computational costs.

2.3 Contrast function

The demixing matrix is sought as the minimizer of the so-called contrast function $J(\mathbf{W})$. A properly chosen contrast should be minimized when $\mathbf{W}$ is a demixing matrix and it should be invariant with respect to the transformation described in (3), since such a transformation does not improve or worsen the separation. The contrast here is derived via the maximum likelihood principle. The negative logarithmic likelihood of the demixing matrix $\mathbf{W}$ and the nuisance parameters σ_m^2, $\mathbf{C}_m$ given N samples of the observed mixtures is

$$\begin{aligned} J =& \frac{T}{2} \sum_{m=1}^{M} \Bigg\{ \ln \sigma_m^2 + \ln\det \mathbf{C}_m + \frac{1}{\sigma_m^2} \mathbf{W}_{1,:} \tilde{\mathbf{R}}_m \mathbf{W}_{1,:}^T \\ & + \operatorname{tr}\Big[\mathbf{C}_m^{-1} \mathbf{W}_{2:d,:} \tilde{\mathbf{R}}_m \mathbf{W}_{2:d,:}^T\Big] \Bigg\} - N \ln\det \mathbf{W}, \end{aligned} \tag{5}$$

where

$$\tilde{\mathbf{R}}_m = \frac{1}{T} \sum_{t=(m-1)T+1}^{mT} \mathbf{x}_t \mathbf{x}_t^T \tag{6}$$

is the sample covariance matrix of the mixtures in the m-th block. The maximum likelihood estimate of the demixing matrix is found by minimizing (5) with respect to the matrix $\mathbf{W}$, variances σ_m^2 and covariances $\mathbf{C}_m$. Note that the function in Formula (5) is not a function of only the data and the demixing matrix and it is not invariant with

respect to the symmetry transformation of $\mathbf{W}$ with σ_m^2 and $\mathbf{C}_m$ being fixed. Thus it does not meet our requirements for the contrast. However, minimization of (5) with respect to $\mathbf{W}$, σ_m^2 and $\mathbf{C}_m$ is equivalent to minimization of

$$\tilde{J}(\mathbf{W}) = \min_{\sigma_m^2, \mathbf{C}_m} J(\mathbf{W}, \sigma_m^2, \mathbf{C}_m) \tag{7}$$

with respect to $\mathbf{W}$. Now $\tilde{J}(\mathbf{W})$ is clearly a function of only $\mathbf{W}$ and the data. In addition, the minimum of (5) with respect to σ_m^2 and $\mathbf{C}_m$ can be easily computed by setting the partial derivatives of (5) with respect to $\sigma_m^2, \mathbf{C}_m$ to zero which leads to $\sigma_m^2 = \mathbf{W}_{1,:}\tilde{\mathbf{R}}_m\mathbf{W}_{1,:}^T$ and $\mathbf{C}_m = \mathbf{W}_{2:d,:}\tilde{\mathbf{R}}_m\mathbf{W}_{2:d,:}^T$. Substituting these into (5) yields

$$\tilde{J}(\mathbf{W}) = \frac{T}{2}\sum_{m=1}^{M}\left\{\ln \mathbf{W}_{1,:}\tilde{\mathbf{R}}_m\mathbf{W}_{1,:}^T + \ln\det \mathbf{W}_{2:d,:}\tilde{\mathbf{R}}_m\mathbf{W}_{2:d,:}^T\right\} - N\,\ln\det\mathbf{W} + \frac{Nd}{2}. \tag{8}$$

Now it is easily seen that $\tilde{J}(\mathbf{W})$ is invariant with respect to the transformation (3). The demixing matrix can be estimated either by minimizing $\tilde{J}(\mathbf{W})$ with respect to $\mathbf{W}$ or by minimizing $J(\mathbf{W}, \sigma_m^2, \mathbf{C}_m)$ with respect to $\mathbf{W}$ and the nuisance parameters. In the following, we choose the latter option.

2.4 Parameterization of the demixing matrix

The demixing matrix in ICE is determined only up to a transformation described in (3). Any two matrices $\mathbf{W}_1, \mathbf{W}_2$ that are related through (3) give a separation of the same quality and are equivalent from the viewpoint of ICE. To get rid of this ambiguity, we choose one representative of each class of equivalence and seek the demixing matrix among these representatives.

Let $\mathbf{A}$ be the true mixing matrix, and let $\mathbf{B}$ be a submatrix of $\mathbf{A}^{-1}$ created by removing the first row and the first column, $\mathbf{B} = (\mathbf{A}^{-1})_{2:d,2:d}$, and let $\alpha = (\mathbf{A}^{-1})_{1,1}$. Finally, let us assume that α is nonzero and $\mathbf{B}$ is nonsingular. Multiplying the mixing equation $\mathbf{U} = \mathbf{A}^{-1}\mathbf{X}$ by a block diagonal matrix $\begin{pmatrix}\alpha^{-1} & \\ & \mathbf{B}^{-1}\end{pmatrix}$ yields

$$\begin{pmatrix}\tilde{\mathbf{s}} \\ \tilde{\mathbf{Z}}\end{pmatrix} = \begin{pmatrix}1 & \mathbf{h}^T \\ \mathbf{g} & \mathbf{I}\end{pmatrix}\mathbf{X}. \tag{9}$$

Since $\tilde{\mathbf{s}}$ is a scaled signal of interest, and $\tilde{\mathbf{Z}}$ is independent of $\tilde{\mathbf{s}}$, the matrix

$$\mathbf{W}(\mathbf{g}, \mathbf{h}) = \begin{pmatrix}1 & \mathbf{h}^T \\ \mathbf{g} & \mathbf{I}\end{pmatrix} \tag{10}$$

is a demixing matrix. Thus we can seek the demixing matrix in the form (10), provided the regularity conditions $\alpha \neq 0$, $\det\mathbf{B} \neq 0$ hold.

A natural question arises, whether this parameterization is optimal. As a result of the invariance property, the contrast function is constant on each of the classes of equivalence. Consequently, choosing a different parameterization and minimizing the contrast in this parameterization would yield exactly the same equivalence class, i.e., the separation would not be improved.

3 Estimation algorithm

3.1 Multiplicative step method

The demixing matrix is estimated using a multiplicative update. The main idea is the following: Given an estimate $\mathbf{W}_i$ of the demixing matrix, let us seek a better estimate in the form

$$\mathbf{W}_{i+1} = \mathbf{W}_{res}\mathbf{W}_i, \tag{11}$$

where the so called residual demixing matrix $\mathbf{W}_{res}$ is chosen appropriately. As was shown in the previous section, the residual demixing matrix $\mathbf{W}_{res}$ can be assumed to take on the parametric form

$$\mathbf{W}_{res} = \begin{pmatrix} 1 & \mathbf{h}^T \\ \mathbf{g} & \mathbf{I} \end{pmatrix}. \tag{12}$$

The parameters $\mathbf{g}, \mathbf{h}$ are sought by minimizing the contrast $J(\mathbf{W}_{res}\mathbf{W}_i)$, which can be rewritten as

$$\begin{aligned} J(\mathbf{W}_{res}\mathbf{W}_i) =& \frac{T}{2}\sum_{m=1}^{M}\Bigg\{ \ln\sigma_m^2 + \ln\det\mathbf{C}_m + \frac{1}{\sigma_m^2}(\rho_m^{(i)} + 2\mathbf{h}^T\mathbf{r}_m^{(i)} + \mathbf{h}^T\mathbf{R}_m^{(i)}\mathbf{h}) \\ &+ \operatorname{tr}\left[\mathbf{C}_m^{-1}(\mathbf{R}_m^{(i)} + \mathbf{g}(\mathbf{r}_m^{(i)})^T + \mathbf{r}_m^{(i)}\mathbf{g}^T + \rho_m^{(i)}\mathbf{g}\mathbf{g}^T)\right]\Bigg\} - N\ln(1-\mathbf{h}^T\mathbf{g}). \end{aligned} \tag{13}$$

Here $\rho_m^{(i)}, \mathbf{r}_m^{(i)}, \mathbf{R}_m^{(i)}$ are blocks of the sample covariance matrix of the partially demixed data $\mathbf{x}^{(i)} = \mathbf{W}_i\mathbf{x}$,

$$\begin{pmatrix} \rho_m^{(i)} & (\mathbf{r}_m^{(i)})^T \\ \mathbf{r}_m^{(i)} & \mathbf{R}_m^{(i)} \end{pmatrix} = \frac{1}{T}\sum_{t=(m-1)*T+1}^{mT}\mathbf{x}_t^{(i)}(\mathbf{x}_t^{(i)})^T. \tag{14}$$

The contrast (13) is minimized with respect to the parameters $\mathbf{g}$, $\mathbf{h}$ and the nuisance parameters σ_m^2 and $\mathbf{C}_m$. The minimization method consists of three distinctive steps.

Initialization: Firstly, the parameters of the residual demixing matrix are initialized as $\mathbf{g} = \mathbf{h} = \mathbf{0}$ which simply means that the previous estimate $\mathbf{W}_i$ is used as the initialization for $\mathbf{W}_{i+1}$. The nuisance parameters σ_m^2 and $\mathbf{C}_m$ does not need to be initialized as they can be estimated in a closed form.

Update of σ_m^2 and $\mathbf{C}_m$: Setting partial derivatives of the contrast (13) with respect to σ_m^2 and $\mathbf{C}_m$ to zero and setting $\mathbf{g} = \mathbf{h} = \mathbf{0}$ yields a necessary condition for σ_m^2 and $\mathbf{C}_m$ to be minimizers of (13) in a form

$$\sigma_m^2 = \rho_m^{(i)}, \tag{15}$$

$$\mathbf{C}_m = \mathbf{R}_m^{(i)}. \tag{16}$$

Update of g and h: With σ_m^2 and $\mathbf{C}_m$ fixed to (15)(16), we update the parameters $\mathbf{g}$ and $\mathbf{h}$ as

$$\begin{pmatrix} \mathbf{g} \\ \mathbf{h} \end{pmatrix} = -\mathbf{H}\nabla J, \tag{17}$$

where ∇J and $\mathbf{H}$ are the gradient and the mean Hessian of the contrast with respect to $\mathbf{g}$ and $\mathbf{h}$ computed at $\mathbf{g} = \mathbf{h} = \mathbf{0}$. Performing the necessary computations yields

$$\begin{pmatrix} \mathbf{g} \\ \mathbf{h} \end{pmatrix} = - \begin{pmatrix} T \sum_{m=1}^{M} \mathbf{C}_m^{-1} \sigma_m^2 & N\mathbf{I} \\ N\mathbf{I} & T \sum_{m=1}^{M} \frac{1}{\sigma_m^2} \mathbf{C}_m \end{pmatrix}^{-1} \begin{pmatrix} T \sum_{m=1}^{M} \mathbf{C}_m^{-1} \mathbf{r}_m \\ T \sum_{m=1}^{M} \frac{1}{\sigma_m^2} \mathbf{r}_m \end{pmatrix}. \tag{18}$$

Having computed $\mathbf{g}, \mathbf{h}$ from (18), we set $\mathbf{W}_{res}$ as in (12) and update the demixing matrix as in (11). These steps are repeated until convergence is achieved.

3.2 Covariance inversion

The inversion of the covariance matrix $\mathbf{C}_m$ needed for(18) can be avoided by utilizing the special structure of the demixing matrix. Note that the matrices $\mathbf{R}_m^{(i)}$ and $\mathbf{R}_m^{(i+1)}$, which estimate $\mathbf{C}_m$ in the i-th step and the $i+1$-th step, differ only by an additive factor of rank two, that is

$$\mathbf{R}_m^{(i+1)} = \mathbf{R}_m^{(i)} + \begin{pmatrix} \mathbf{g} & \mathbf{r}_m^{(i)} \end{pmatrix} \begin{pmatrix} (\mathbf{r}_m^{(i)})^T + \rho_m^{(i)} \mathbf{g}^T \\ \mathbf{g}^T \end{pmatrix}.$$

The Woodbury matrix lemma then states that the inverse of $\mathbf{R}_m^{(i+1)}$ is equal to the inverse of $\mathbf{R}_m^{(i)}$ plus rank-2 matrix update, computation of which requires only the inversion of a 2×2 matrix. Thus instead of inverting the $(d-1) \times (d-1)$ matrix $\mathbf{R}_m^{(i)}$ in each step of the algorithm, we just need to invert it once at the initialization of the algorithm and then to invert a 2×2 matrix in each step. The complexity of the algorithm is then dominated by the complexity of the inversion of the mean Hessian, which is $\mathcal{O}(d^3)$.

3.3 Stationary interference

Another simplification can be made by assuming stationarity of the interference. This may worsen the achieved separation if the true interference is nonstationary. On the other hand, it greatly reduces the computational cost since it enables us to avoid the inversion of the Hessian at each step.

Let us assume taht the interference is stationary Gaussian with the zero mean and covariance $\mathbf{C}$. This corresponds to $\mathbf{C}_m = \mathbf{C}$ for each m. The estimate of the covariance $\mathbf{C}$ is then computed as in (16) but with the right-hand side averaged over m, i.e., $\mathbf{C} = \frac{1}{M} \sum_{m=1}^{M} \mathbf{R}_m^{(i)}$. The mean Hessian from (17) takes on the same form as in (18) with $\mathbf{C}$ substituted in place of $\mathbf{C}_m$ and it can be inverted using the block matrix inversion scheme, yielding

$$\mathbf{H}^{-1} = \begin{pmatrix} (\frac{1}{N}\gamma + \beta\gamma^2)\mathbf{C} & -\beta\gamma\mathbf{I} \\ -\beta\gamma\mathbf{I} & \beta\mathbf{C}^{-1} \end{pmatrix}, \tag{19}$$

where $\gamma = (\frac{1}{M} \sum_{m=1}^{M} \sigma_m^2)^{-1}$ and $\beta = \frac{1}{N}(\frac{1}{M} \sum_{m=1}^{M} \frac{1}{\sigma_m^2} - \gamma)^{-1}$.

4 Numerical experiments

4.1 Robustness to initial error

In the following experiment we study the robustness of BGICE to a small error in initialization. We distinguish two cases: nonstationary interference and stationary interference. In each of the trials, we simulate $d = 6$ sources on $M = 10$ blocks, each with length $T = 1000$ samples. On each block, the signal of interest and the interference are sampled from a Gaussian distribution with zero mean and the following variances resp. covariances. For $i = 1, \ldots, d$ and $m = 1, \ldots, M$ we set

$$\sigma_{im}^2 = \sin(\frac{\pi m i}{M}) + 1.1, \tag{20}$$

$$\mathbf{C}_m = \mathrm{diag}(\sigma_{2m}^2 \ldots \sigma_{6m}^2) \tag{21}$$

Here σ_{1m}^2 is taken as the variance of the SOI in the m-th block and either $\mathbf{C}_m$ or $\mathbf{C} = \frac{1}{M}\sum_{m=1}^{M}\mathbf{C}_m$ as the covariance of the interference in the m-th block. The SOI and the interference are then mixed by a mixing matrix $\mathbf{A}$ generated by drawing each of it's elements from the uniform distribution on the interval $[1, 2]$. The pdBGICE algorithm is initialized with $\mathbf{g}, \mathbf{h}$ computed from the true demixing matrix $\mathbf{W} = \mathbf{A}^{-1}$ as

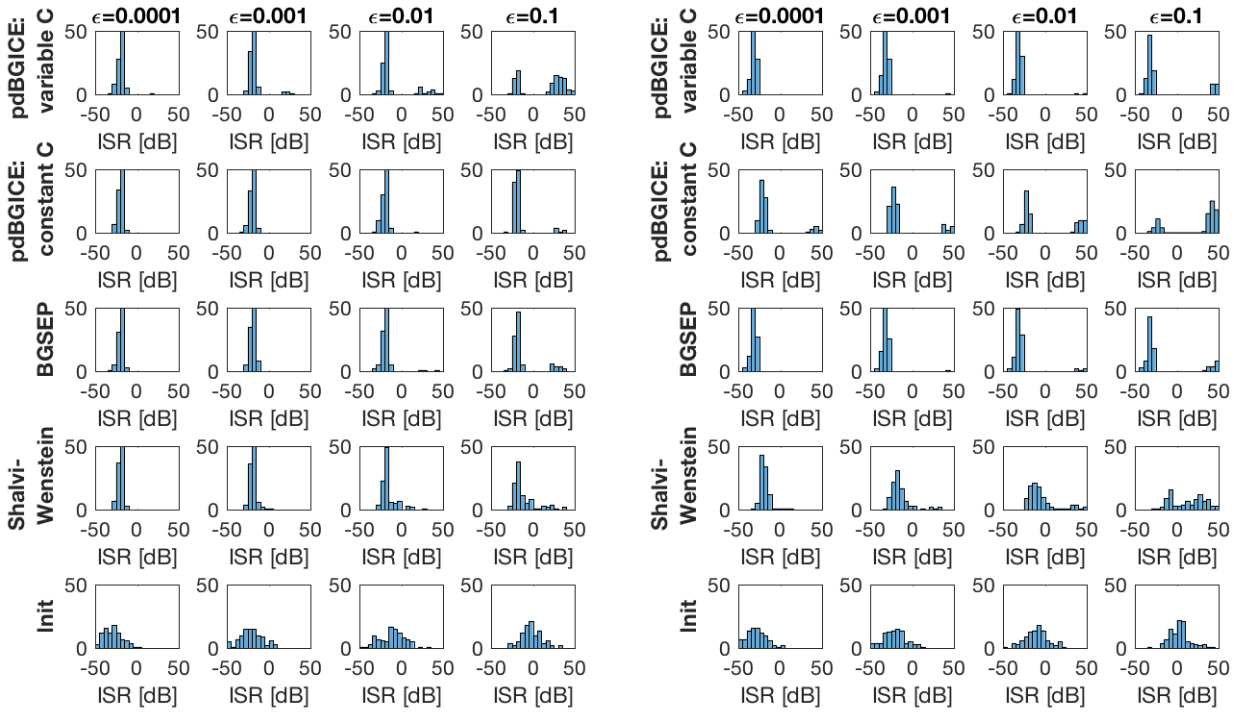
$$\mathbf{g} = \mathbf{W}_{2:d,2:d}^{-1}\mathbf{W}_{2:d,1} + \epsilon_g, \tag{22}$$

$$\mathbf{h} = \mathbf{W}_{11}^{-1}\mathbf{W}_{1,2:d}^{T} + \epsilon_h. \tag{23}$$

where ϵ_g and ϵ_h are randomly generated vectors with $\|\epsilon_g\| = \|\epsilon_h\| = \epsilon$. The parameter ϵ determining the size of the initialization error ranges from 10^{-4} to 10^{-1}. For each value of ϵ, we run 200 trials. Given the estimated demixing matrix, we compute the ISR in the estimated SOI. The values of ISR from all the trials are shown in histograms in Fig. 1. The values around -30 dB correspond to perfect separation, values around 30 dB to a convergence to the wrong component. Values around 0 dB indicate the algorithm did not converge. The method Init corresponds to using just the initialization without any further separation.

In Fig. 1a, we plot the results for the stationary interference case. The algorithm pdBGICE: variable C always achieves a separation, although for $\epsilon = 0.1$, the wrong component is separated in a majority of the trials. pdBGICE: constant C gives the best results. It rarely separates the wrong components even for $\epsilon = 0.1$ and never fails to converge. The BGSEP algorithm yields a good separation for $\epsilon \in \{10^{-4}, 10^{-3}, 10^{-2}\}$. For $\epsilon = 0.1$, the algorithm separates the wrong component in some trials. The method of Shalvi and Weinstein achieves a good ISR for $\epsilon \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$.

In Fig. 1b we plot the results for the non-stationary interference case. pdBGICE: variable C yields the best results, rarely separating the wrong component even for $\epsilon = 0.1$ and never failing completely. Similarly, pdBGICE: constant C never fails completely but the rate of separation of the wrong component is substantially higher. The BGSEP algorithm yields an accurate separation for $\epsilon \in \{10^{-4}, 10^{-3}, 10^{-2}\}$. For $\epsilon = 0.1$, the algorithm separates the wrong component in some trials. The method of Shalvi and Weinstein achieves only a low separation, no separation at all or separation of the wrong component for $\epsilon > 10^{-4}$. Note that this method is specifically designed for the case



(a) (b)

Figure 1: Histograms of the ISR obtained through the pdBGICE, BGSEP and Shalvi-Weinstein method for different values of the initialization error in the cases of (a) stationary interference, (b) nonstationary interference.

of stationary independence and cannot work properly with non-stationary interference. The relatively good low ISR achieved for $\epsilon = 10^{-4}$ can be attributed to the accurate initialization.

5 Conclusion

In this paper, we examined the problem of blind source extraction. First, we have described the problem of the independent component extraction and proposed a suitable parameterization of the sought de-mixing matrix. We described a natural equivalence relation on the set of possible de-mixing matrices corresponding to the problem of ICE, and have used this relation to show that there are no better parameterizations. Second, we have proposed two separation algorithms using the maximum likelihood principle and nonstationarity modeling of the sources. These algorithms either model the interference as nonstationary, leading to lower CRLB but higher complexity, or as stationary.

In the experimental part, we have tested the algorithms on artificial data following the proposed model. We have explored the robustness of the algorithms to an inaccurate initialization. The two proposed algorithms yield promising results, achieving an accurate separation in a high percentage of the trials, showing a low rate of wrong component

separation and outperforming the competing BGSEP method, as well as the method of Shalvi-Weinstein.

References

[1] P. Comon and C. Jutten. *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. Independent Component Analysis and Applications Series. Elsevier Science, (2010).

[2] N. Delfosse and P. Loubaton. *Adaptive blind separation of independent sources: A deflation approach*. Signal Processing **45** (1995), 59 – 83.

[3] S. Gannot, D. Burshtein, and E. Weinstein. *Signal enhancement using beamforming and nonstationarity with applications to speech*. IEEE Transactions on Signal Processing **49** (Aug 2001), 1614–1626.

[4] A. Hyvärinen and E. Oja. *A fast fixed-point algorithm for independent component analysis*. Neural Computation **9** (July 1997), 1483–1492.

[5] V. Kautský, Z. Koldovský, and P. Tichavský. *Cramér-Rao-induced bound for interference-to-signal ratio achievable through non-Gaussian independent component extraction*. In '2017 IEEE 7th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)', 1–4, (Dec 2017).

[6] Z. Koldovský, P. Tichavský, and V. Kautský. *Orthogonally constrained independent component extraction: Blind MPDR beamforming*. In 'Proceedings of European Signal Processing Conference', 1195–1199, (September 2017).

[7] D.-T. Pham. *Blind partial extraction of instantaneous mixtures of sources using second order statistics*. In 'EUSIPCO 2008 - 16th European Signal Procesing Conference', 1–5, Lausanne, Switzerland, (August 2008). EUSIPCO.

[8] D.-T. Pham and J. F. Cardoso. *Blind separation of instantaneous mixtures of nonstationary sources*. IEEE Transactions on Signal Processing **49** (Sep 2001), 1837–1848.

[9] O. Shalvi and E. Weinstein. *System identification using nonstationary signals*. IEEE Transactions on Signal Processing **44** (Aug 1996), 2055–2063.

[10] P. Tichavský and A. Yeredor. *Fast approximate joint diagonalization incorporating weight matrices*. IEEE Transactions on Signal Processing **57** (March 2009), 878–891.

[11] P. Tichavský and Z. Koldovský. *Fast and accurate methods of independent component analysis: A survey*. Kybernetika **47** (2011), 426–438.

[12] O. Šembera, P. Tichavský, and Z. Koldovský. *Independent component extraction based on nonstationarity modeling*. IEEE Transactions on Signal Processing (submitted).

Deep Generative Models in Anomaly Detection*

Vít Škvára

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

advisor: Václav Šmídl, Department of Adaptive Systems
Institute of Information Theory and Automation, AS CR

Abstract. Many deep models have been recently proposed for anomaly detection. This paper presents comparison of selected generative deep models and classical anomaly detection methods on an extensive number of non–image benchmark datasets. We provide statistical comparison of the selected models, in many configurations, architectures and hyperparamaters. We arrive to conclusion that performance of the generative models is determined by the process of selection of their hyperparameters. Specifically, performance of the deep generative models deteriorates with decreasing amount of anomalous samples used in hyperparameter selection. In practical scenarios of anomaly detection, none of the deep generative models systematically outperforms the kNN.

Keywords: anomaly detection, generative models, neural networks

Abstrakt. Tento článek představuje stručný teoretický základ několika generativních modelů, založených na hlubokých neuronových sítích a přizpůsobených pro detekci anomálií. Kromě toho jsou představeny i některé klasické modely. Dále je popsán provedený experiment, který srovnává vybrané modely na velkém množství datasetů, pro různé nastavení hyperparametrů a architektur. Následuje statistické vyhodnocení experimentu. Hlavním získaným poznatkem je fakt, že proces ladění hyperparametrů je pro hluboké generativní modely velmi důležitý. Konkrétněji, jejich přesnost se zhoršuje se zmenšujícím se počtem anomálních vzorků použitých při tomto procesu. Závěrem je potom zjištění, že v provedených experimentech žádný z vybraných generativních modelů nedosahuje systematicky lepší úspěšnosti než algoritmus kNN.

Klíčová slova: detekce anomálií, generativní modely, neuronové sítě

1 Introduction

An anomaly is a data sample that is so different from the rest of the normal data that it was likely generated by a different underlying process. Anomalous samples are an object of interest for various reasons and methods for novelty or anomaly detection try to identify them. These methods are used in a plethora of domains, including medical, computer security or sensor data collection. An overview of anomaly detection methods is presented in [15].

This paper asks the following question – how do anomaly detection methods based on deep neural–network generative models stand in comparison to the methods based on

*This work has been supported by the GACR project 18-21409S

alternative paradigms? In the fashion of [13] or [5], we do not propose a new algorithm to prove it performs better than the existing methods. Instead, we aim to clarify whether the existing generative models actually bring a significant improvement over classical approaches. To our knowledge, such comparison of anomaly detection methods based on neural networks to some simpler methods is missing. Sadly, most papers proposing new methods (especially those based on deep neural networks) focuse on large image data from MNIST, CIFAR or some other publicly available database that pose an unsurmountable task for classical anomaly detection methods. However, image–based anomaly problems are a rather niche domain, therefore this paper focuses on evaluation on a varied range of datasets. When presenting a novel generative model, the authors usually compare to some baseline generative model against which they are trying to improve, a limited number of classical models such as PCA or OSVM, and at best they use the KDD dataset – see e.g. [1] or [22].

Furthermore, no one has ever done a thorough comparative study of generative deep models on a large number of different datasets. In the following text, a handful of selected generative models will be compared against each other and classical novelty detection methods in a statistical way on a large number of carefully crafted benchmark datasets. We do not claim to provide a complete overview of generative models applicable to anomaly detection task, but a general comparison that may simplify future assessment of novel methods. We admit that we have not tested or implemented state–of–the art methods such as [24] or [23].

Additionally, we endeavor to create a standardized, publicly available implementation of different models. Apart from the actual model implementations, we create a framework for proper training, testing and comparing the models.

Finally, we propose a number of important questions that may lead to a better understanding of individual tested algorithms and their behavior on different datasets. Using experimental data and a thorough comparison methodology, we may be able to answer some of these questions or provide valuable insight.

2 Background

Generative models are used to generate samples from some learned data distribution $p(x)$. In the case of anomaly detection, this is the distribution of normal data. Therefore, even though all following methods are unsupervised by default, we strive to train the generative model with (mostly) normal data. Even though it may be expensive to obtain labels, they are also useful for tuning of hyperparameters, as will be discussed later.

When the model has learned the normal data distribution, it can be used to compute an anomaly score function $f : \mathcal{X} \to \mathbb{R}$ for a sample $x \in \mathcal{X}$ from the data space. The convention used here is that higher the anomaly score, the more likely it is that the sample is an anomaly. For generative models, reconstruction error (in case of autoencoders), discriminator score (for adversarial models) or their combination can be used.

This section contains a brief theoretical background of the used generative models based on neural nets and two classical methods. The description of cost functions, anomaly score functions and their parameters will be given.

2.1 Autoencoding models

2.1.1 Autoencoder

An autoencoder (AE) is a cornerstone of many multi layer perceptron (MLP) models. Although not a generative model by itself, it may be used as a baseline for the variational autoencoder generative model. It is easily used for the anomaly detection task (see [16], [20] or the comprehensive review of anomaly detection methods [15]).

It consists of two MLPs - an encoder and a decoder. The encoder represents the mapping $e_\phi : \mathcal{X} \to \mathcal{Z}$ that projects a sample x from the data space $\mathcal{X}$ to code z in the latent space $\mathcal{Z}$. The decoder reconstructs the code back to the data space via mapping $d_\theta : \mathcal{Z} \to \mathcal{X}$. From here on, θ and ϕ denote hidden parameters of neural nets. Both parts of the autoencoder are trained using backpropagation by minimizing the reconstruction error

$$\mathcal{L}_r(x, \phi, \theta) = ||x - d_\theta(e_\phi(x))||_2^2. \tag{1}$$

When the dimension of the code z is smaller than that of x, the autoencoder is forced to learn an efficient sparse representation of x while being robust to noise [21].

This roughly equals to learning to reconstruct samples coming from the distribution $p(x)$ of the normal data. When the trained autoencoder is shown an anomaly, it will likely not produce a good reconstruction as it has not been trained with similar samples. Therefore, the anomaly score is given by the reconstruction error of a sample

$$f_{\text{AE}}(x) = \mathcal{L}_r(x, \bar{\phi}, \bar{\theta}), \tag{2}$$

where $\bar{\phi}, \bar{\theta}$ are the fixed learned parameters of the AE.

2.1.2 Variational autoencoder

The variational autoencoder (VAE) [11] and its modifications has seen a lot of success especially in generating realistic images. Its design is very similar to that of an ordinary autoencoder. To induce the generative property, we force the encoder to produce codes z that resemble samples from some easy–to–sample–from prior distribution $p(z)$ (e.g. $\mathcal{N}(0, 1)$). Afterwards, new samples that resemble the real data can be generated by inputing codes sampled from $p(z)$ to the decoder.

In VAE, both the encoder and decoder model parameters of conditional distributions denoted as $q_\phi(z|x)$ and $p_\theta(x|z)$. We assume that the distributions are Gaussian, therefore at the output layer of the MLPs we obtain the mean and variance of the respective distributions. During training, the aim is to minimize the reconstruction loss of x while simultaneously minimizing the Kullback-Leibler divergence $D_{KL}(q_\phi(z|x)||p(z))$, which is zero if the two distributions are equal. Also, an important part of the training is the *reparametrization trick*, which creates random decoder inputs in the following manner: $z = \mu_z + \sigma_z \epsilon$, where μ_z and σ_z are outputs of the encoder and ϵ is sampled from $p(z)$. The cost function for training a VAE is following

$$\mathcal{L}_v(x, \phi, \theta) = \mathbb{E}_{q(z|x)}\left[||x - d_\theta(z)||_2^2\right] + \lambda D_{KL}(q_\phi(z|x)||p(z)), \tag{3}$$

where $\lambda = \sigma_x^2$ is a tuning parameter equal to a known variance of data. Since both $q_\phi(z|x)$ and $p(z)$ are Gaussian, there is an analytical expression for their KL divergence.

There are numerous papers describing the use of VAE for anomaly detection – [19], [22], [6], but none make a more complete comparison with other generative and classical methods and only some use non–image datasets. In [1], the authors describe the advantages of VAE over AE – it generalizes more easily since it is working on probabilities. The anomaly score function of the VAE is the reconstruction error

$$f_{\text{VAE}}(x) = \mathbb{E}_{q(z|x)}\left[||x - d_\theta(z)||_2^2\right]. \tag{4}$$

Alternatively, the log–likelihood of the code z due to prior $p(z)$ can be used.

2.2 Adversarial models

2.2.1 GAN

The simplest adversarial generative model is the *generative adversarial network* – GAN [9]. It consists of two adversaries – a generator and a discriminator that are represented by MLPs. The generator creates samples that resemble the real data, while the discriminator is trying to recognize the fake samples from the real ones. During training, they both improve – generator creates more believable samples while the discriminator gets more proficient at recognizing fakes.

Inputs of the generator are codes $z \sim p(z)$, where $p(z)$ is e.g. standard or uniform distribution. The generator is a mapping into the data space $g_\phi : \mathcal{Z} \to \mathcal{X}$. Discriminator is a mapping $d_\theta : \mathcal{X} \to [0, 1]$, i.e. its output is a scalar representing the probability that a sample comes from the true data distribution $p(x)$. The training alternates between minimizing the logit cross–entropy for discriminator

$$\mathcal{L}_d(x, \theta) = -\mathbb{E}_{p(x)}\left[\log d_\theta(x)\right] - \mathbb{E}_{p(z)}\left[\log(1 - d_\theta(g_\phi(z)))\right] \tag{5}$$

and a simplified logit cross–entropy for the generator

$$\mathcal{L}_g(\phi) = -\mathbb{E}_{p(z)}\left[\log d_\theta(g_\phi(z))\right]. \tag{6}$$

After training, one can input samples z to the generator and it should be able to generate samples that resemble those from $p(x)$.

For anomaly detection application, $p(x)$ is the distribution of normal data. We do not need to know the true form of $p(x)$, we only need to be able to sample from it. The discriminator should ideally learn the however complicated form of $p(x)$ by backpropagation. Anomaly score of a sample x is computed as a weighted average of the discriminator output and a reconstruction error of the generated sample

$$f_{\text{GAN}}(x) = -(1 - \lambda)\log(d_{\bar{\theta}}(x)) + \lambda||x - g_{\bar{\phi}}(z)||_2, \tag{7}$$

where $z \sim p(z)$ and λ is a scalar scaling parameter.

2.2.2 Feature-matching GAN

While GAN enjoys success in generation of realistic images, it is famously difficult to train. The authors of [17] proposed a number of modifications to the original simple training

process. One of them is the addition of the *feature–matching* loss to the cost function. It is based on the idea that backpropagation based on a loss computed somewhere else than the final scalar output of the discriminator may provide improved gradients for the generator, thus enabling a more stable training procedure.

The feature–matching GAN (fmGAN) uses an augmented generator cost function

$$\mathcal{L}_f(x, \phi) = \alpha \mathcal{L}_g(\phi) + \mathbb{E}_{p(x),p(z)} \left[||h_\theta(x) - h_\theta(g_\phi(z))||_2 \right], \tag{8}$$

where h_θ is the output of some intermediate (e.g. the penultimate) layer of the discriminator and α is a scalar scaling parameter.

The fmGAN has been successfully used for anomaly detection in [18]. The anomaly score function is the same as in the case of the GAN model.

2.3 Classical outlier detection methods

2.3.1 kNN

The k–nearest neighbours anomaly detection algorithm [2] is a simple yet powerful model. It is based on the assumption that normal data are grouped in the data space and anomalies are distant from them and therefore can be detected by measuring their distance from the rest of the data. It is relatively easy to implement, well described in literature and quite well performing. A large study [5] concluded that no classical anomaly detection algorithm provides a comprehensive improvement over kNN.

For faster distance computation, the training data are encoded in a KDTree structure [3]. This significantly improves the prediction times on large datasets but requires additional overhead in construction of the tree. The only hyperparameter of the algorithm is k. The version of kNN used for experiments in this paper computes the anomaly score as the average distance to k–nearest neighbours in the training dataset.

2.3.2 Isolation forest

Isolation forest [12] is also one of the more widely used anomaly detection algorithms. During training, the algorithm randomly and recursively partitions the data and stores this partitioning in a tree structure. This partitioning is done numerous times – thus a forest is constructed. Anomalies should be ideally reached on a shorter path in the tree as they require a smaller number of partitionings to be separated from the rest of the data. The hyperparameter to be tuned is the number of trees n_t. The anomaly score is computed as the negative of the average path length of a sample over all trees in the forest.

3 Experiments

This section describes the setup of experiment that was conducted on a large number of benchmark datasets. Datasets, algorithm and experiment implementation in the Julia language is made publicly available in a GitHub repository at
`https://github.com/smidl/AnomalyDetection.jl`.

algorithm	hyperparameters
kNN	$k \in \{1, \frac{1}{2}\sqrt{N}, \sqrt{N}, \frac{3}{2}\sqrt{N}, 2\sqrt{N}\}$
Isolation Forest	$n_t \in \{100, 500\}$
all deep models	$\dim(\mathcal{Z}) \in \{0.1\tilde{M}, 0.2\tilde{M}, \ldots, 0.5\tilde{M}\}$,
	$\tilde{M} = \min(200, M)$
	$n_h \in \{1, 2, 3\}$
VAE	$\lambda \in \{10^{-4}, 10^{-3}, \ldots, 1\}$
GAN	$\lambda \in \{0, 0.2, \ldots, 1.0\}$
fmGAN	$\alpha \in \{10^{-4}, 10^{-2}, \ldots, 10^{4}\}$
	$\lambda \in \{0, 0.2, \ldots, 1.0\}$

Table 1: Tested hyperparameter settings. M is the dimensionality of a dataset, N is the number of training samples, n_t is the number of trees, n_h is the number of hidden layers.

3.1 Experiment setup

3.1.1 Benchmark data

In order to evaluate the selected models in the most realistic fashion, the methodology of creating appropriate anomaly detection benchmarks from [8] was adopted. A list of 35 preprocessed basic datasets from [14] was used. A more detailed overview of the basic datasets and of the preprocessing can be found in the same paper.

The basic datasets were originally created from multiclass classification or regression datasets that were split into normal and anomalous data, whitened and further processed. Afterwards, it is possible to sample normal and anomalous data according to different criteria. First, there are four levels of anomaly difficulty - *easy, medium, hard* and *very hard*. Second, a contamination rate (the percentage of anomalies in a sampled dataset) can be specified. Third, it is possible to sample anomalies that are clustered or unclustered.

For the purposes of our experiments, we have sampled from each basic dataset to obtain a non-contaminated (containing no anomalies) training dataset and a contaminated testing dataset. We have used 80% of normal data for training and 20% for testing. In each of the basic datasets, there is a different number of anomalies of a given anomaly difficulty level (see the summary table in [14]). Therefore, the anomaly difficulty levels were chosen (mostly *easy* or *medium*) so that there were enough anomalies to be sampled from with respect to the chosen contamination rate. The contamination rate of the testing dataset was 5% and we sampled for non–clustered anomalies. Random sampling was repeated 10 times for each basic dataset to obtain diverse training and testing datasets .

3.1.2 Algorithm setup

The table 1 contains an overview of tested hyperparameter values. An experiment consisting of training of a model and predicting anomaly scores was carried out for every possible combination of hyperparameters. Therefore, a different number of experiments per dataset sampling was realized for different algorithms – that is no more than 5 experiments for kNN but up to 450 experiments in the case of fmGAN. For training of the

deep models, the Adam optimizer [10] was used with a learning rate of 0.001. We trained each generative model for a total of 10000 steps with a batchsize of 256 or lower in case of datasets that are not numerous enough. Relu activation functions and dense layers were used. Size of hidden layers was given by linear interpolation between the code dimension $\dim(\mathcal{Z})$ and a proper input/output dimension. The code dimension was proportionate to the input dimension $\dim(\mathcal{X})$ but never exceeded 100.

3.2 Evaluation methodology

3.2.1 Algorithm ranking

For comparison of different algorithms on multiple datasets, we use the statistical procedure described in [7]. For a given comparison metric, e.g. AUROC (area under ROC curve), we rank algorithms according to their performance on a dataset. If there is a tie, then average rank is given to the tied algorithms. Afterwards, an average rank over datasets is computed for each algorithm. Based on the average ranks, a number of statistical hypotheses may be tested. In the case of this study, we are interested in two: a) is there enough statistical evidence to reject the hypothesis that all the algorithms perform equally and b) do two individual algorithms perform differently on a statistically significant level? The answers to these questions are given by the Friedman and Nemenyi test, respectively. For our problem size (6 algorithms and 35 datasets), the critical value for the Nemenyi test at 5% confidence level is $CD_{0.05} \approx 1.2745$. If the average rank of two algorithms is larger than that, we can say that their performance is significantly different.

A concise way of visualization are the critical difference diagrams, see e.g. Fig 1. The average algorithm ranks are marked and the statistically equally performing algorithms are connected together with a wide black line.

3.2.2 Hyperparameter selection

After training of an algorithm on the training dataset, a vector of anomaly scores was computed for the testing dataset. Also, a small sample of anomalies was added to the training dataset and a vector of anomaly scores of the resulting ensemble was computed (reasoning behind this follows). Afterwards, testing and training AUROC were computed from these two vectors and known labels. This was done for each individual experiment specified by a combination of dataset, model, resampling iteration and model hyperparameter setting.

We assumed that tuning of hyperparameters is done during training. The rankings of models were produced using the testing AUROC of best–performant model in a resampling iteration, averaged over resampling iterations on a dataset. To simulate different real–world conditions in which hyperparameters are tuned, 3 methods of their selection were used.

1. Firstly, the hyperparameters were selected by the best AUROC on the testing dataset. This gives an idea of the maximum potential of the algorithm, which may be difficult to attain in reality.

	kNN	IForest	AE	VAE	GAN	fmGAN
test auc	3.94	5.63	3.47	2.07	3.90	1.99
train auc	3.13	4.61	3.63	2.84	4.46	2.33
top 5%	2.57	4.07	3.24	2.73	4.90	3.49
top 1%	2.14	3.53	3.13	2.93	4.97	4.30

Table 2: Average ranks of algorithms for different hyperparameter selection criteria.

	kNN	IForest	AE	VAE	GAN	fmGAN
t_f [s]	0.15	0.76	276.33	1088.71	844.83	879.94
t_p [s]	5.35	0.58	2.14	3.49	6.40	6.38

Table 3: Average fit t_f and predict t_p times.

2. Secondly, the hyperparameters were selected by best performance on training dataset with added anomalies (training AUROC). This simulates a more realistic scenario of hyperparameter tuning, where all the labels are known on the training dataset.

3. Thirdly, the hyperparameters were selected using the $p\%$ most anomalous samples of the training dataset with added anomalies. However, the performance was measured using the precision (rate of detected anomalies). This is the most realistic scenario in case when the training labels are expensive to obtain. The common practice is then to manually inspect the most $p\%$ anomalous samples in the training dataset and select the hyperparameters of the most precise model. For our experiments, we chose $p \in \{1, 5\}$.

These ranking methods stimulate interesting questions.

- What is the reason of an algorithm performing well in the first method and not so well in the others? Insufficient tuning of hyperparameters? Overfitting on training? Or just very difficult applicability to real–world problems where labels are not present and such tuning is not possible?

- What is the robustness of different algorithms to variation of the training data? How robust is it to a possible presence of anomalies in the training data?

- How many samples need to be manually inspected to obtain robust optimal hyperparameter settings?

- Can we analyze on which datasets are some methods more successful than the others? Do these datasets have something in common?

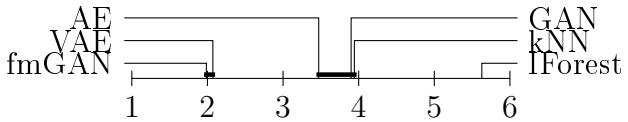


Figure 1: Critical difference diagram for the first hyperparameter selection criterion.

3.3 Experimental results

Mean ranks of algorithms across datasets for all hyperparameter selection methods are summarized in Table 2 (the complete AUC scores for all datasets and ranking criterions are not shown in this paper for the sake of brevity). Results from Table 2 shows that generative models, namely VAE and fmGAN, can produce excellent results if one has a large number of known anomalies to select right hyperparameters (methods *test auc* and *train auc*). Once the number of known anomalies decreases, as is the case of the more realistic methods *top 1%* and *top 5%*, generative models becomes inferior to methods robust to hyperparameter selection, namely k–nearest neighbours.

The Friedman test rejects the hypothesis that the models perform equally well in all four ranking criteria. Critical difference diagrams shown in Figs. 1 – 4 reveal that unless one selects hyperparameters on testing set, which is obviously cheating, not a single generative model provides a statistically significant improvement over the naive kNN anomaly detection algorithm. VAE seems to be the most robust and therefore more promising deep model, as its ranks relatively well for all the hyperparameter selection criteria.

Although the deep models do not generally outperform kNN when tuned with a limited number of labels, they can still perform well on certain datasets. These datasets may share some common characteristic which makes them suitable for application of deep models, rather than just a random fluctuation caused by the large number of experiments carried out for fmGAN and VAE. However we have not yet been able to find the connection. It is certainly not the difficulty of the anomalies, as kNN performs relatively well across all difficulty levels.

The only area in which the deep models outperform kNN is the mean prediction time on large datasets (in terms of number of training samples), see Table 3. Note that the largest benchmark dataset is *miniboone* with 93565 normal samples. In deep models, the prediction is independent on the training data size. This is however compensated for by the computational demands of their training.

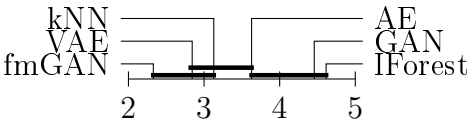


Figure 2: Critical difference diagram for the second hyperparameter selection criterion.

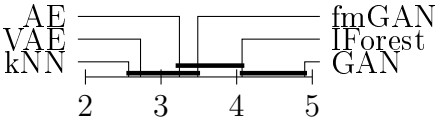


Figure 3: Critical difference diagram for the third hyperparameter selection criterion at 5% most anomalous samples.

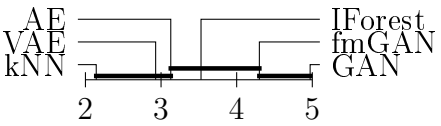


Figure 4: Critical difference diagram for the third hyperparameter selection criterion at 1% most anomalous samples.

4 Conclusions

In this paper, a selection of deep generative models adopted for anomaly detection were compared against traditional methods of k-nearest neighbours and Isolation Forests. This was done in a systematic way on a large number of benchmark datasets. Most authors of novel anomaly detection methods based on deep models do not compare to kNN, but instead compare to methods such as Local Outlier Factor [4], despite it has been shown in [14] and [5] that it does not systematically outperform kNN algorithm. We have shown that the robustness of kNN still holds even in comparison with some deep generative models.

From the conducted experiments, the main conclusion is that the major bottleneck for reliable employment of deep generative models for anomaly detection is the difficulty of hyperparameter tuning. Ideally, hyperparameter tuning should be automated and included in the training procedure, which is not the case. Especially adversarial models do not seem to be robust and their training is difficult. On the other hand, generative models showed promising potential and should be studied in a greater depth. We have studied performance of different models under a relatively simple, yet realistic, hyperparameter selection methods and criteria. An interesting question is the existence of a criterion that is more appropriate in the context of limited number of available labels.

This also leads us to questioning the appropriateness of the use of deep generative models in the unsupervised setting. Since we have shown that labels are necessary for hyperparameter tuning, maybe it is worth investigating a semi/fully–supervised deep generative models for anomaly detection.

References

[1] J. An and S. Cho. *Variational autoencoder based anomaly detection using reconstruction probability.* SNU Data Mining Center, Tech. Rep. (2015).

[2] F. Angiulli and C. Pizzuti. Fast outlier detection in high dimensional spaces. In 'European Conference on Principles of Data Mining and Knowledge Discovery', 15–27. Springer, (2002).

[3] J. L. Bentley. *Multidimensional binary search trees used for associative searching.* Communications of the ACM **18** (1975), 509–517.

[4] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. Lof: identifying density-based local outliers. In 'ACM sigmod record', volume 29, 93–104. ACM, (2000).

[5] G. O. Campos, A. Zimek, J. Sander, R. J. Campello, B. Micenková, E. Schubert, I. Assent, and M. E. Houle. *On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study.* Data Mining and Knowledge Discovery **30** (2016), 891–927.

[6] S. Clachar. *Novelty detection and cluster analysis in time series data using variational autoencoder feature maps.* The University of North Dakota, (2016).

[7] J. Demšar. *Statistical comparisons of classifiers over multiple data sets.* Journal of Machine learning research **7** (2006), 1–30.

[8] A. F. Emmott, S. Das, T. Dietterich, A. Fern, and W.-K. Wong. Systematic construction of anomaly detection benchmarks from real data. In 'Proceedings of the ACM SIGKDD workshop on outlier detection and description', 16–21. ACM, (2013).

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In 'Advances in neural information processing systems', 2672–2680, (2014).

[10] D. P. Kingma and J. Ba. *Adam: A method for stochastic optimization.* arXiv preprint arXiv:1412.6980 (2014).

[11] D. P. Kingma and M. Welling. *Auto-encoding variational bayes.* arXiv preprint arXiv:1312.6114 (2013).

[12] F. T. Liu, K. M. Ting, and Z.-H. Zhou. Isolation forest. In 'Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on', 413–422. IEEE, (2008).

[13] A. Odena, A. Oliver, C. Raffel, E. D. Cubuk, and I. Goodfellow. *Realistic evaluation of semi-supervised learning algorithms.* (2018).

[14] T. Pevný. *Loda: Lightweight on-line detector of anomalies.* Machine Learning **102** (2016), 275–304.

[15] M. A. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko. *A review of novelty detection.* Signal Processing **99** (2014), 215–249.

[16] M. Sakurada and T. Yairi. Anomaly detection using autoencoders with nonlinear dimensionality reduction. In 'Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis', 4. ACM, (2014).

[17] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In 'Advances in Neural Information Processing Systems', 2234–2242, (2016).

[18] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In 'International Conference on Information Processing in Medical Imaging', 146–157. Springer, (2017).

[19] M. Sölch, J. Bayer, M. Ludersdorfer, and P. van der Smagt. *Variational inference for on-line anomaly detection in high-dimensional time series.* arXiv preprint arXiv:1602.07109 (2016).

[20] B. B. Thompson, R. J. Marks, J. J. Choi, M. A. El-Sharkawi, M.-Y. Huang, and C. Bunje. Implicit learning in autoencoder novelty assessment. In 'Neural Networks, 2002. IJCNN'02. Proceedings of the 2002 International Joint Conference on', volume 3, 2878–2883. IEEE, (2002).

[21] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol. *Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion.* Journal of Machine Learning Research **11** (2010), 3371–3408.

[22] H. Xu, W. Chen, N. Zhao, Z. Li, J. Bu, Z. Li, Y. Liu, Y. Zhao, D. Pei, Y. Feng, et al. *Unsupervised anomaly detection via variational auto-encoder for seasonal kpis in web applications.* arXiv preprint arXiv:1802.03903 (2018).

[23] S. Zhai, Y. Cheng, W. Lu, and Z. Zhang. *Deep structured energy based models for anomaly detection.* arXiv preprint arXiv:1605.07717 (2016).

[24] B. Zong, Q. Song, M. R. Min, W. Cheng, C. Lumezanu, D. Cho, and H. Chen. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In 'International Conference on Learning Representations', (2018).

Heuristics in Blind Source Separation*

Jakub Štěch†

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

advisor: Tatiana Valentine Guy, Department of Adaptive Systems
Institute of Information Theory and Automation, CAS

Abstract. This paper deals with application of heuristic algorithms (DEBR, MCRS) in blind source separation (BSS). BSS methods focus on a separation of the (source) signal from a linear mixture. The idea of using heuristic algorithms is introduced on the independent component extraction (ICE) model. The motivation for considering heuristics is to obtain an initial guess needed by many ICE algorithms. Moreover, the comparison of this initialization, and other algorithms accuracy is performed.

Keywords: Blind Source Separation, Independent Component Extraction, Heuristics, MCRS, DEBR

Abstrakt. Tento článek se zabývá aplikací heuristických algoritmů (DEBR, MCRS) v slepé separaci zdrojů (BSS). Metody BSS se zaměřují na separaci (zdrojového) signálu z lineární směsi. Myšlenka použití heuristických algoritmů je představena na modelu ICE. Motivací pro využití heuristik je získat počáteční odhad potřebný mnoha ICE algoritmy. Navíc je provedeno srovnání těchto inicializací společně s přesnostma algoritmů.

Klíčová slova: slepá separace zdrojů, Independent Component Extraction, heurstika, MCRS, DEBR

Full paper: This paper has been submitted to the Stochastic and Physical Monitoring Systems, SPMS 2018.

*This work has been supported by the grant The Czech Science Foundation Project GA18-15970S
†Co-authors Václav Kautský

Bayesian Optimization with Heteroscedastic Gaussian Process

Lukáš Ulrych

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

advisor: Václav Šmídl, Department of Adaptive Systems
Institute of Information Theory and Automation, AS CR

Abstract. Bayesian optimization (BO) is a popular method for optimization of functions with unknown analytical form which can be evaluated only point-wise. The evaluation of these black-box functions is usually very expensive. BO proceeds by fitting the values at known points by a surrogate function (typically a Gaussian process), finding the optimum on this surrogate model and then proposing new points for evaluation. Interesting features of the method are that: i) it aims to minimize the number of necessary evaluations, and ii) it allows to optimize functions that are not evaluated exactly but with additive zero-mean noise. This noise is, however, typically assumed to be homogeneous in the whole search space. In this paper, we study the implications of the use of heteroscedastic Gaussian Process (HGP) in Bayesian optimization. The key requirement for successful use of HGP in BO is availability of a hyper-parameter estimation method that works reliably even with very low number of data. We provide a review of existing hyper-parameter estimation methods for HGP, and compare them on regression task with low number of observations. We derive the acquisition function for HGP Bayesian optimization and demonstrate its behavior on a synthetic example. We demonstrate superiority of the HGP over GP both on regression and optimization tasks.

Keywords: black-box function, Bayesian optimization, Gaussian process, hyper-parameter estimation

Abstrakt. Bayesovská optimalizace (BO) je metoda pro hledání extrémů funkcí s neznámým analytickým tvarem. Tyto tzv. black-box funkce lze typicky vyčíslit jen bodově a navíc je jejich vyčíslení velmi drahá záležitost. BO tento problém řeší představením náhradní funkce, typicky Gaussovským procesem, na kterém hledá body dalšího vyčíslení. Zajímavé vlastnosti BO: i) jejím cílem je minimalozace počtu vyčíslení neznámé funkce, ii) umožňuje optimalizovat funkce, které jsou známé s přesností na aditivní šum. Typickým předpokladem tohoto šumu je jeho gaussovské rozdělení s nulovou střední hodnotou. V tomto textu představujeme dopady použití heteroskedastického Gaussovského procesu (HGP) v BO. Klíčovým předpokladem využití, je dostupnost metod odhadů hyper-parametrů, které spolehlivě fungují i při relativně malém počtu dat. Ukážeme srovnání různých metod odhadů pro HGP a porovnáváme jejich výsledky na příkladech regrese dat. Představíme odvození akviziční funkce pro HGP a demostrujeme její chování na umělých datech.

Klíčová slova: black-box funkce, Bayesovská optimalizace, Gaussovský proces, odhad hyper-parametrů

1 Introduction

Bayesian optimization is an established method of derivative-free optimization of functions that can be evaluated only point-wise. This is typical for optimization of results of complex computations where one evaluation consumes significant amount of processing time. The Bayesian approach is based on modeling of the objective function as a realization of the stochastic process with hyper-parameters (surrogate model). Every evaluation of the objective function is treated as a new observation that is used to improve posterior distribution of the objective function. Due to availability of the full posterior distribution, it is possible to solve the decision problem where to place a new observation (i.e. where to evaluate the black-box function) to maximize the chance of finding the global extreme. This decision problem is solved by the optimization of so called acquisition function. Application of this general method requires to make choices of the model structures and evaluation methods.

The Gaussian process is the most conventional choice of the surrogate model in BO. Other models have been used such as the Student-t process that has more attractive properties especially in terms of better robustness and faster convergence to the optimum [8]. The use of non-stationary processes have been proposed in [5]. However, relatively little attention is paid to the models of the noise of observation of the objective function. It is of little importance when the function is evaluated with high accuracy, however, in problems that involve Monte Carlo simulation it is significant. This motivates our investigation of the use of heteroscedastic Gaussian process in Bayesian optimization.

Heteroscedastic GP (HGP) [4] is obtained as a hierarchical process, where the variance of observation of the objective function is modeled by another Gaussian process. Estimation of its hyper-parameters is a challenge. Various techniques have been proposed, ranging from maximum likelihood [4], Variational Bayes [11], Expectation propagation [6], to full treatment via Monte Carlo. Quality of the hyper-parameter estimation is very important in BO since full Bayesian treatment allows better space exploration and the solution is found more efficiently [9]. Therefore, we investigate suitability of existing methods for estimation of HGP for BO. Since the aim of BO is to minimize the number of function evaluations and thus the number of observations of the GP, we pay close attention to performance of the methods for relatively low number of samples. We review existing methods for HGP, test another general purpose techniques used in GP (such as adaptive importance sampling [12]), and derive our own adaptation of general purpose technique based on Look Ahead Hamiltonian Monte Carlo [10]. We will demonstrate their performance on simple synthetic data.

2 Bayesian optimization

Bayesian optimization (BO) provides probability framework for the task of the optimization of any function $f(x)$. In many real-world tasks where optimization is required it is often the case that the studied function $f(x)$ is not convex, not cheap to evaluate and even its analytical form is either unknown or outright non-existent. Such functions are called black-box functions. In these scenarios common optimization techniques, often based on convexity or derivatives, become useless and heuristic algorithms, based

on Monte Carlo approach, become unbearably slow. BO bypasses the problem of direct optimization route of these methods by introducing a surrogate model for the data. After optimizing this model to fit the data, the task at hand has shifted from optimizing the unknown, expensive to evaluate function $f(x)$ to optimizing the known, easy to evaluate surrogate model. It is not until this process is done and the possible optimum, based on the surrogate model, is found, when BO needs to evaluate the black-box function $f(x)$ to confirm or disprove the location of the optimum, thus gaining new observation. The entire process is then repeated on extended data set until some convergence requirement is met. The surrogate model BO utilizes is in the form of Gaussian process. Since finding the maximum of $f(x)$ is the same task as finding the minimum of $-f(x)$ we will follow with the assumption that the desired outcome is to find the minimum of given function.

2.1 Gaussian processes

Gaussian processes (GPs) can be viewed as an extension of a standard multivariate Gaussian distribution to the space of infinite dimension and as such define a distribution over continuous functions. GPs are fully determined by their mean and covariance functions, which depend on a set of hyper-parameters. Every finite subset of GP follows a standard multi-variate Gaussian distribution, thus working with GP is always reduced to dealing with finite multi-variate Gaussian distribution.

Here we consider a standard regression model $y = f(x) + \varepsilon$. Since our knowledge of $f(x)$ is near zero, we want to use a model that requires almost no prior information. Making an assumption about $f(x)$ to be of some given form dependent on parameters, for example linear function of x, can be too restrictive. GPs offer a solution to this problem. They represent a probability distribution over all possible functions and by letting GP to learn from the data we automatically let it to choose the most probable shape of $f(x)$.

Function $f(x)$ is modelled with GP, $f(x) \sim \mathcal{GP}(m(x), k_f(x,x))$, where $m(x)$ and $k_f(x,x)$ are monikers for mean function and covariance function respectively. Common practice is to set $m(x) = 0$ and scale the data accordingly. Choice of $k_f(x,x)$, sometimes called kernel function, is more complicated. Because always only n observations $y = (y_1, \ldots, y_n)$ in corresponding points $x = (x_1, \ldots, x_n)$ are available, GP reduces to Gaussian distribution $\mathcal{N}(\mathbf{0}_{n\times 1}, K_f)$ where $(K_f)_{i,j} = k_f(x_i, x_j)$, $i, j \in \{1, \ldots, n\}$. Gaussian noise ε is set to be zero mean with unknown covariance matrix Σ. It is straightforward to show that the joint density of y and f is

$$p(y, f) = p(y|f)\, p(f) = \mathcal{N}(f, \Sigma)\, \mathcal{N}(\mathbf{0}_{n\times 1}, K_f). \tag{1}$$

By integrating out the process f we find the distribution of y to be Gaussian with zero mean and covariance matrix $K_y = K_f + \Sigma$, $y \sim \mathcal{N}(\mathbf{0}_{n\times 1}, K_y)$. This notation is, however, incorrect, because we omit the dependence on hyperparameters, together denoted as θ, which are present in k_f and Σ. Ideally, we would want to eliminate the dependency on θ by integrating it out. That is, however, not possible, due to the intractability of the corresponding integral

$$p(y) = \int p(y|\theta)\, p(\theta)\, \mathrm{d}\theta. \tag{2}$$

Therefore, an approximation must be made. Finding probable values of θ is done in a Bayesian way by optimizing or sampling from the posterior distribution $p(\theta|y)$ which follows from Bayes's theorem $p(\theta|y) \propto p(y|\theta)\, p(\theta)$. Depending on the chosen method, either a point estimate or a set of samples is obtained and then used for approximation of (2) and consequent calculations.

2.2 GP based prediction

After optimizing its hyper-parameters, GP represents the most probable guess of how $f(x)$ looks like given the observed data. From this approximation rises another useful feature of GPs: their prediction ability. For any set of m new points x^*, for which we want to make prediction y^*, we simply make use of the knowledge that we already have about $f(x)$ contained in $p(y|\theta)$. Since it is Gaussian, we can extend it to cover x^* and arrive at

$$p(y^*, y|\theta) = \mathcal{N}\left(\begin{array}{c} \mathbf{0}_{n\times 1} \\ \mathbf{0}_{m\times 1} \end{array} , \begin{pmatrix} K_y & K_{y,y^*} \\ K_{y,y^*}^{\mathrm{T}} & K_{y^*} \end{pmatrix} \right), \tag{3}$$

where $(K_{y,y^*})_{i,j} = k\left(x_i, x_j^*\right)$, $i \in \{1,\ldots,n\}$, $j \in \{1,\ldots,m\}$ and $(K_{y^*})_{i,j} = k\left(x_i^*, x_j^*\right) + \Sigma_{i,j}^*$, $i,j \in \{1,\ldots,m\}$ represent joint covariance matrix between x and x^* and covariance matrix of x^* plus noise covariance matrix respectively. To find the desired posterior distribution for y^* we proceed by conditioning on y

$$p(y^*|y,\theta) = \mathcal{N}\left(K_{y,y^*}^{\mathrm{T}} K_y^{-1} y, K_{y^*} - K_{y,y^*}^{\mathrm{T}} K_y^{-1} K_{y,y^*}\right). \tag{4}$$

Again, the dependency on θ should be integrated out, but since it is also intractable, an approximation of $p(y^*|y)$ must be made.

For our experiments we used the squared exponential (SE) covariance function

$$k_{SE}(x_i, x_j) = \sigma_f \exp\left(-\frac{\|x_i - x_j\|_2^2}{2l}\right). \tag{5}$$

2.3 Heteroscedastic noise assumption

In real world problems the assumption of homoscedastic noise is rarely correct. Although this approximation of reality usually works fine, there are cases when we must abandon it and resort to more complex model with heteroscedastic noise. Such model keeps the Gaussian-zero-mean assumption, but presumes that variance of noise is some unknown function of x, $r(x)$, $\varepsilon \sim \mathcal{N}(0, r(x))$. To ensure that variance is always positive, we will rather work with function $g(x) = \log(r(x))$. Since $g(x)$ is completely unknown, we will use another Gaussian process to simulate it, as first proposed in [3]. In order to avoid adding false information into the model, this second-level GP will always have unknown, but constant, mean μ_g and SE covariance function.

When using heteroscedastic model, the joint distribution of y, f and g given θ is

$$p(y, f, g|\theta) = p(y|f, g, \theta)\, p(f|\theta)\, p(g|\theta) = \mathcal{N}(f, \Sigma)\, \mathcal{N}(\mathbf{0}_{n\times 1}, K_f)\, \mathcal{N}(\mu_g \mathbf{1}_{n\times 1}, K_g). \tag{6}$$

Here K_g denotes covariance matrix of noise GP. As in homoscedsatic noise case, process f can be integrated out, which yields

$$p\left(y|g,\theta\right)p\left(g|\theta\right)=\mathcal{N}\left(\mathbf{0}_{n\times 1},K_y\right)\mathcal{N}\left(\mu_g\mathbf{1}_{n\times 1},K_g\right), \tag{7}$$

where $K_y = K_f + \mathrm{diag}\left(\exp\left(g\right)\right)$. Here θ denotes all hyperparameters included in the model and $\mathrm{diag}\left(\exp\left(g\right)\right)$ a diagonal matrix with entries $\exp\left(g\left(x_1\right)\right), \ldots, \exp\left(g\left(x_n\right)\right)$. Integrating out the g process would be a desirable course of action, the corresponding integral is, however, intractable. Instead, we must treat its values as other hyperparameters and find their posterior distribution

$$p\left(g,\theta|y\right)=\frac{p\left(y|g,\theta\right)p\left(g|\theta\right)p\left(\theta\right)}{p\left(y\right)}. \tag{8}$$

The addition of the assumption of heteroscedasticity greatly affects subsequent calculations, because the predictive posterior distribution for m new points x^*

$$p\left(y^*|y\right)=\int p\left(y^*|y,g,g^*,\theta\right)p\left(g,g^*|\theta\right)p\left(\theta\right)\mathrm{d}\left(g,g^*,\theta\right) \tag{9}$$

is no longer analytically tractable. Although we can not find $p\left(y^*|y\right)$ in closed form, we can find a reasonable approximation. Noise variance perdiction g^* can be integrated out analytically with the use of factorization $p\left(g,g^*|\theta\right)=p\left(g^*|g,\theta\right)p\left(g|\theta\right)$. From the resulting distribution $p\left(y^*|y,g,\theta\right)$ both mean and covariance can be computed, allowing us to make use of moment matching technique and approximate it with Gaussian distribution which can then be used as an approximation of $p\left(y|y^*\right)$.

2.4 Acquisition functions

Since Gaussian process serves as a probability distribution the underlying black-box function, the search for the optimum of this surrogate model is done in terms of likelihoods. Specifically, we need to chose a function, called acquisition function, that is based on the GP and that, for given point x, states, how likely it is to actually be the point of optimum. Very common choice for acquisition function, and the one we used, is called Expected improvement (EI). It is defined as

$$a\left(x\right)=\mathbb{E}\left[\max\left(0,f'-f\left(x\right)\right)\right]=\left(f'-\mu\left(x\right)\right)\Phi\left(f'\right)+\Sigma\left(x,x\right)\phi\left(f'\right). \tag{10}$$

The expectation is taken with respect to the GP, denoted as $f\left(x\right)$, with mean function $\mu\left(x\right)$ and covariance function $\Sigma\left(x,x'\right)$. f' denotes the lowest value available in the data and Φ and ϕ denote cumulative density function and probability density function for Gaussian distribution with the same parameters as GP, respectively. Since we work with noise-corrupted data, f' is replaced by μ', which denotes the lowest value the GP model predicts in the same points x as the observations we have are in.

The main feature that distinguishes HGP from GP in terms of the acquisition function is the prediction of $\Sigma\left(x,x\right)$ based on (4). Our main focus is the optimization of the black-box function which is modelled by the mean function. However, we are not interested in modelling the noisy observations. For that reason the term for covariance in (4),

specifically K_{y^*}, should not include the prediction for noise. This changes nothing for GP, because its noise level is constant, however, it is paramount for HGP which noise levels vary.

The EI function is based on the GP and as such depends on its hyper-parameters, which are present in both μ and Σ. The theoretically correct approach would be to eliminate the uncetainty arising from this by integrating them out. This, however, yields yet another intractable integral and so an approximation must be made. Depending on the method of hyper-parameter estimation this is either done by point estimate (ML, VB) or by numerical approximation based on the samples available (LAHMC, PMC, AMIS).

3 Hyper-parameter estimation methods

In this section we provide brief introduction of methods that are cosidered for hyper-parameter estimation.

Probably the most well known parameter estimation method of al is Maximum Likelihood (ML). It offers what most of the other methods lack, speed and precision. However, these traits are fully utilized only when certain requirements are met.

Variational Bayes (VB) method [11] attempts to minimize relative entropy, also known as Kullback-Leibler divergence, between true posterior distribution and some approximation based on block factorization of parameters θ.

Population Monte Carlo (PMC) [1], a method based on Importance sampling technique, attempts to bypass the issue of choosing static proposal distribution $q(\theta)$ in the problem of computing the expectation of function $h(\theta)$. PMC uses advantages of MCMC simulations by optimizing parameters of $q(\theta)$ in a sequential manner.

Adaptive Multiple Importance Sampling (AMIS) [2] is another method based on Importance Sampling. Unlike PMC, which after every iteration discards obtained samples, AMIS recursively recomputes importance weigths of past samples as well as the present ones.

By using knowledge of the state space geometry, Hamiltonian Monte Carlo (HMC) [7] significantly reduces the random behavior often observed in MCMC methods, increases the quality of mixing and trims the required computational time.

An extension of standard HMC called Look Ahead Hamiltonian Monte Carlo (LAHMC) [10] is designed to further increase the acceptance probability and decrease autocorrelation of new samples.

4 Experiments

In this Section, we provide experiments for selection of the most suitable hyper-parameter learning techniques on simple regression task and advantages of HGP for simple synthetic optimization problem. The abbreviation GP will here be used for model with homogeneous noise. For HGP we use LAHMC, ML and VB methods. PMC and AMIS didn't prove to be able to handle high number of parameters. For GP we use LAHMC, ML, PMC and AMIS methods. In Bayesian optimization, both GP and HGP were modeled with unknown constant mean function.

metric	MSE			NLPD		
# of data	120	80	40	120	80	40
HGP-LAHMC	1.15 ± 0.45	1.27 ± 0.71	1.20 ± 0.97	1.81 ± 0.21	1.87 ± 0.38	2.05 ± 0.53
HGP-ML	1.13 ± 0.47	1.13 ± 0.60	1.05 ± 0.58	1.84 ± 0.23	2.02 ± 0.29	2.28 ± 0.32
HGP-VB	1.08 ± 0.48	1.19 ± 0.69	1.15 ± 1.23	1.76 ± 0.23	1.80 ± 0.37	1.90 ± 0.63
GP-AMIS	1.05 ± 0.38	1.18 ± 0.72	1.20 ± 1.02	2.11 ± 0.20	2.17 ± 0.36	2.26 ± 0.51
GP-LAHMC	1.12 ± 0.38	1.27 ± 0.69	1.39 ± 0.99	2.21 ± 0.21	2.30 ± 0.36	2.44 ± 0.49
GP-ML	1.16 ± 0.39	1.24 ± 0.67	1.32 ± 1.05	2.19 ± 0.22	2.27 ± 0.34	2.37 ± 0.51
GP-PMC	2.04 ± 0.75	2.26 ± 1.43	2.14 ± 1.98	2.26 ± 0.36	2.37 ± 0.70	2.36 ± 0.96

Table 1: Comparison of MSE and NLPD.

4.1 Comparison of hyper-parameter estimation methods on regression task

First, we deal with regression task on the synthetic data set: function $f(x) = 2\sin(2\pi x)$ evaluated in x uniformly spaced in the interval $[0, 1]$ and corrupted by zero-mean Gaussian noise with standard deviation function $\mathrm{std}(x) = 2x$. Starting from 120, we decrease the number of starting data points by 40 till 40, to see how different models and estimation methods fare in such circumstances. In total 50 data sets were generated independently for each number of starting points. Each set was then randomly split into 80% to train the model, and 20% to validate it. We used normalized mean squared error (MSE) and average negative log-predictive distribution (NLPD) to asses performance

$$\mathrm{MSE} = \frac{1}{m}\sum_{i=1}^{m}\frac{(y_i^* - \mathbb{E}[y_i^*])^2}{\mathrm{var}(y_i^*)}, \qquad \mathrm{NLPD} = \frac{1}{m}\sum_{i=1}^{m} -\log(\pi(y_i^*|y)).$$

Results are summarized in table 1. In terms of MSE, all methods, with the exception of PMC, performed comparably, however, NLPD clearly shows dominance of HGP regardless of the number of data points. Best performing method is HGP-VB, having the lowest NLDP. Student t-test ($p = 0.05$) confirmed, that NLPD values of HGP-VB are significantly better than those of HGP-ML for 80 and 40 starting points. The same holds when comparing HGP-LAHMC and HGP-ML. Comparing HGP-VB with HGP-LAHMC doesn't show any significant differences.

4.2 Example of Bayesian optimization with heteroscedastic noise

Now we consider a Bayesian optimization task. Namely, we are interested in finding the minimum of function $f(x) = -3(x + 2\pi)\cos(2(x + 2\pi)) + x$ on interval $(-5, 5)$. To every observation, a zero-mean Gaussian noise with standard deviation function $std(x) = -3(x - 4) + 5$, was added. In total, 100 independent datasets were generated with 40 different starting points in each. An example of one such dataset is ilustrated in Figure 1 along with functions $f(x)$ and $std(x)$. This combination of function and noise was chosen,

because it suffers from three local minima at $x = -3.089$, $x = 0.264$ and $x = 3.159$. In this example, the last one is the desired point of minimum. The idea is that the model with homogeneous noise will not be able to correctly distinguish the true minimum. For this example we used ML, PMC, LAHMC for GP and ML, PMC, LAHMC and VB for HGP. AMIS wasn't used, because it fared poorly on the previous example. For every combination of model and method, in total 50 iterations of Bayesian optimization were performed.

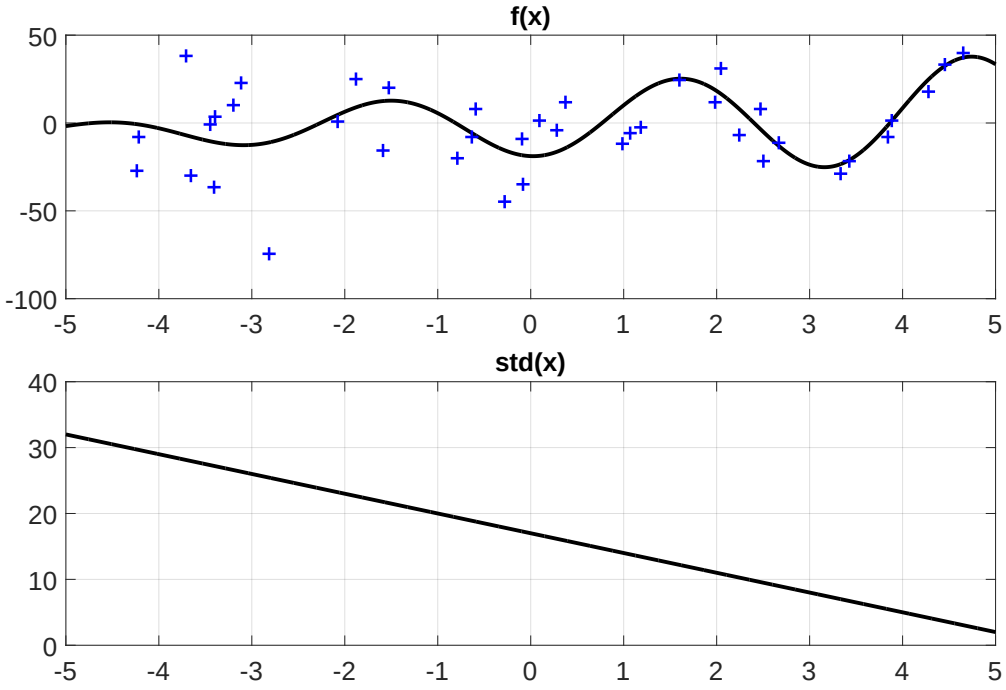


Figure 1: Example of generated dataset.

The results are presented in Figure 2. It is clear that no matter the method, the model with homogeneous noise assumption could not find the true minimum. All of them were able to find the local minimum at $x = 0.264$, but could not proceed to explore the space further. From the 90% confidence interval (the gray area) we can see, that for some data initializations this model was actually able to find the true minimum, but for different ones it was not even able to leave the local minimum around $x = -3.089$. HGP, on the other hand, paints a different picture. The only estimation method that was not able to converge to the true minimum was VB. It consistently found only the local minimum around $x = 0.264$. Every other used method (ML, PMC, LAHMC) was eventually able to find the true minimum. The only exception might be ML, because the 90% confidence interval shows a slight deviation from the correct trend, this was however subdued in the end. The other two (PMC, LAHMC) converged to the true value of minimum within 20 iterations regardless of the initial dataset.

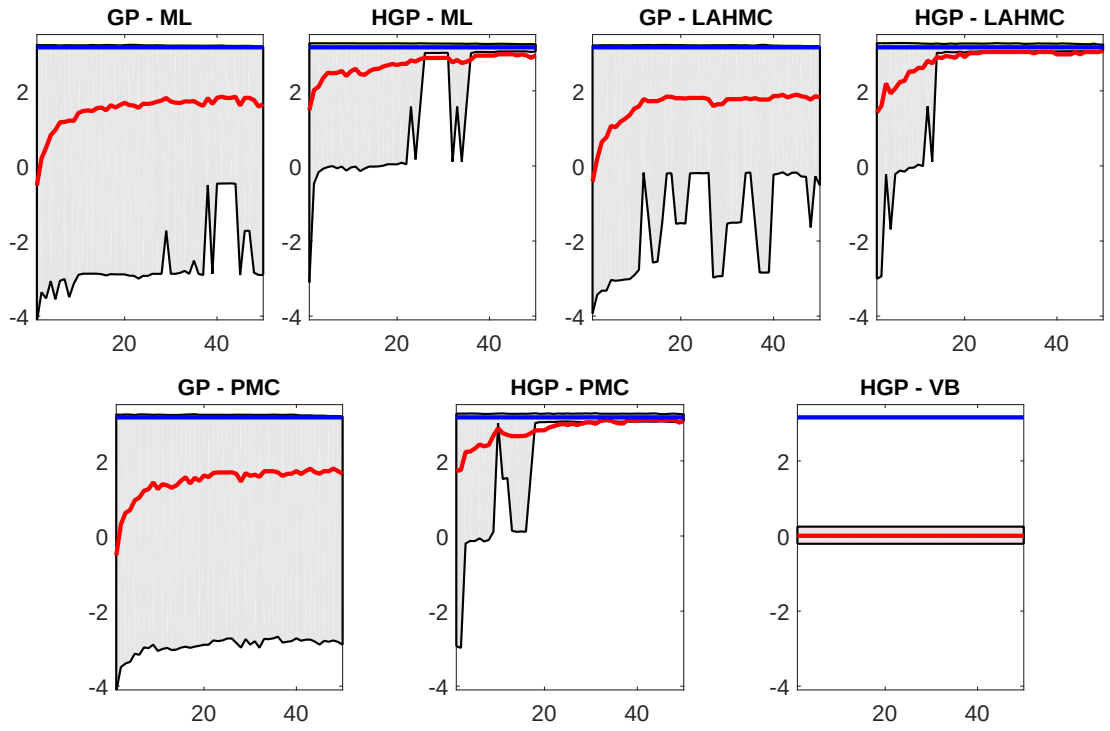


Figure 2: Optimization results. Horizontal axis represents the steps of BO, vertical axis represents the input space. The blue line (at y=3.159) is the true minimum. The red line is the estimate of minimum averaged over all datasets. The gray area is 90% confidence interval.

5 Conclusion

We introduced the concept of Bayesian optimization with Gaussian process used as a surrogate model. We expanded the standard model with homogeneous noise (GP) to include heterogeneity (HGP). For this expanded model we derived the acquisition function that is able to ignore the noise variance and focus directly on posterior mean function that models the true underlying black-box function. For hyper-parameter estimation we examined two deterministic methods (Maximum Likelihood, Variational Bayes) and three stochastic methods (Population Monte Carlo, Adaptive Multiple Importance Sampling, Look Ahead Hamiltonian Monte Carlo). First we compared the models on a simple regression example with heteroscedastic noise and demonstrated the superiority of HGP over GP mainly in scenarios when little data was available. Second we showed their performance on a Bayesian optimization task with synthetic data generated with heterogeneous noise. Again, the superiority of HGP over GP was evident. From tested estimation methods, the one that proved to be reliable in both scenarios with good results was Look Ahead Hamiltonian Monte Carlo.

References

[1] O. Cappé, A. Guillin, J.-M. Marin, and C. P. Robert. *Population monte carlo.* Journal of Computational and Graphical Statistics **13** (2004), 907 – 929.

[2] J. Cornuet, J.-M. MARIN, A. Mira, and C. P. Robert. *Adaptive multiple importance sampling.* Cornuet, Jean and MARIN, JEAN-MICHEL and Mira, Antonietta and Robert, Christian P **39** (2012), 798–812.

[3] P. W. Goldberg, C. K. Williams, and C. M. Bishop. Regression with input-dependent noise: A gaussian process treatment. In 'Advances in neural information processing systems', (1998).

[4] K. Kersting, C. Plagemann, P. Pfaff, and W. Burgard. Most likely heteroscedastic gaussian process regression. In 'Proceedings of the 24th international conference on Machine learning', 393 – 400. ACM, (2007).

[5] R. Martinez-Cantin. *Local nonstationarity for efficient bayesian optimization.* arXiv preprint arXiv:1506.02080 (2015).

[6] L. Muñoz-González, M. Lázaro-Gredilla, and A. R. Figueiras-Vidal. Heteroscedastic gaussian process regression using expectation propagation. In '2011 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)', 1 – 6. IEEE, (2011).

[7] R. M. Neal. *Handbook of Markov Chain Monte Carlo*, volume 2, chapter MCMC using Hamiltonian dynamics, 113 – 162. Chapman and Hall/CRC, (2011).

[8] A. Shah, A. G. Wilson, and Z. Ghahramani. *Student-t processes as alternatives to gaussian processes.* arXiv preprint arXiv:1402.4306 (2014).

[9] J. Snoek, H. Larochelle, and R. P. Adams. Practical bayesian optimization of machine learning algorithms. In 'Advances in neural information processing systems', 2951 – 2959, (2012).

[10] J. Sohl-Dickstein, M. Mudigonda, and M. R. DeWeese. *Hamiltonian monte carlo without detailed balance.* arXiv preprint arXiv:1409.5191 (2014).

[11] M. K. Titsias and M. Lázaro-Gredilla. Variational heteroscedastic gaussian process regression. In 'Proceedings of the 28th International Conference on Machine Learning (ICML-11)', 841 – 848, (2011).

[12] X. Xiong, V. Šmídl, and M. Filippone. *Adaptive multiple importance sampling for gaussian processes.* Journal of Statistical Computation and Simulation **87** (2017), 1644–1665.

The Microscopic Analysis of Velocity-Density Paradigm*

Jana Vacková

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

advisor: Milan Krbálek, Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering, CTU in Prague

Abstract. The relation between density and velocity is one of the most crucial aspects of pedestrian movement. Although some mechanisms are considered to be general true, our research on this topic finds some limits. Firstly, standardly used approaches for evaluating the density is unified into the general density distribution concept. This method covers point approximation, Voronoi cells approach or more sophisticated methods based on individual (kernel) distribution estimates. In this project, we assume special case that each pedestrian is considered as a source of conic density distribution and we focus on the individual density. Controlling blur parameter as the radius of the cone base and range parameter expressing the size of circular pedestrian surroundings, the benefits of short or long-range approach are discussed. When the density is set, velocity-density correlation is evaluated using the data from egress experiments realized by our research group in the last years. Within simple geometry, pedestrians passed an artificial room with controlled inflow condition that enables to observe various inner conditions. The most interesting results are measured for the state with dense crowd near the exit and free flow in the second half of the room. The rolling correlation of velocity and density is negative (as expected) only in the phase of approaching the crowd, but it turns positive when pedestrian fully entered the crowd. This surprising behavior disables to predict pedestrian velocity using density. Thus, follower-leader model using velocity of leaders to predict the motion of followers is developed. Measured by the correlation, this approach is sufficiently successful in the crowd area, and therefore it presents a good complement to velocity–density relations. Moreover, the flow conservation law provides the explanation of the positive correlation.

Keywords: Pedestrian dynamics, Egress experiment, Correlation, Individual density, Follower-leader

Abstrakt. Vztah mezi hustotou a rychlostí je jedním z klíčových aspektů pohybu chodců. Ačkoliv jsou některé mechanizmy obecně považovány za platné, náš výzkum nachází některé limity. V předložené práci nejprve sjednocujeme běžně používané přístupy pro vyhodnocování hustoty v jeden obecný koncept hustotních distribucí - tato metoda zahrnuje jak bodovou aproximaci, tak voronoiské buňky či sofistikovanější metody založené na individuálních (jádrových) distribučních odhadech. Zde uvažujeme speciální případ, kdy je každý chodec považován za zdroj kuželové hustotní distribuce, přičemž se při výpočtu zaměřujeme na jeho individuální hustotu. Zkoumáním parametru reprezentujícího rozmazání chodce (poloměr podstavy kužele) a parametru dosahu chodce (poloměr kruhového okolí chodce) diskutujeme výhody a nevýhody krátkodosahové a

*This work has been supported by the Grant SGS18/188/OHK4/3T/14 provided by the Ministry of Education, Youth, and Sports of the Czech Republic (MŠMT ČR).

dlouhodosahové varianty. S ustanovenou hustotu počítáme korelaci mezi rychlostí a hustotou používajíc data z evakuačních experimentů realizovaných naší výzkumnou skupinou v posledních letech. Chodci procházejí umělou místností, zatímco je kontrolován vstupní tok, jež umožňuje pozorovat různorodé vnitřní chování systému. Ty nejzajímavější výsledky jsou zjištěny pro stav s kongescí u východu a volným tokem v druhé části místnosti. Korelace (s pohybujícím se oknem) rychlosti a hustoty nabývá očekávaných záporných hodnot jen v případě, kdy se chodec přibližuje k davu, okamžitě se ale mění na kladnou, jakmile se chodec ocitá zcela v davu. Toto překvapivé chování nám znemožňuje predikovat rychlost chodce pouze pomocí hustoty. Proto vyvíjíme follower-leader (následník-lídr) model využívající rychlost lídra k předpovědi pohybu jeho následníka. Díky analýze korelací jejich rychlostí zjišťujeme, že je tento koncept dostatečně úspěšný v oblasti davu, a proto představuje dobrý doplněk ke vztahu rychlosti a hustoty. Navíc, zmíněnou pozitivní korelaci lze vysvětlit zákonem zachování toku.

Klíčová slova: Pohyb chodců, Evakuační experiment, Korelace, Individuální hustota, Follower-Leader

Full paper: This full paper [1] has been submitted to the *Journal of Traffic and Transportation Engineering* and is in revision process at the moment. One part of this research was presented at conference *Pedestrian and Evacuation Dynamics* [2], which was held in Lund (Sweden) in August 2018.

References

[1] Jana Vacková, Marek Bukáček. *The Microscopic Analysis of Velocity-Density Paradigm.* Journal of Traffic and Transportation Engineering. In revision process.

[2] Jana Vacková, Marek Bukáček. *Follower-Leader Concept in Microscopic Analysis of Pedestrian Movement in a Crowd.* In PED 2018 proceedings. Accepted.