technische universität
dortmund

SFB 876

Technical report for Collaborative Research Center SFB 876

Providing Information by Resource-Constrained Data Analysis

Dezember 2015

Part of the work on this report has been supported by Deutsche Forschungsgemeinschaft (DFG) within the Collaborative Research Center SFB 876 "Providing Information by Resource-Constrained Analysis"

Speaker: Prof. Dr. Katharina Morik
Address: Technische Universität Dortmund
Fachbereich Informatik
Lehrstuhl für Künstliche Intelligenz, LS VIII
D-44221 Dortmund

Inhaltsverzeichnis

SFB 876 Verfügbarkeit von
Information durch Analyse unter
Ressourcenbeschränkung

Subproject A1
Data Mining for Ubiquitous System Software

Katharina Morik Olaf Spinczyk

Spatio-Temporal Reparametrization of the Multivariate Gaussian

Nico Piatkowski
Lehrstuhl für Künstliche Intelligenz
Technische Universität Dortmund
nico.piatkowski@tu-dortmund.de

In this report, it is shown how our reparametrization and regularization of spatio-temporal discrete state space models, can be carried over to Gaussian graphical models. We discuss the pros and cons of performing the reparametrization either on the standard form or on the exponential family form of the multivariate Gaussian.

1 Introduction

Large probabilistic graphical models [2, 5] may be parametrized by millions of variables. Approaches to penalize parameter vectors with many non-zero weights are available (e.g., the LASSO [1,4]). However, setting model parameters to zero implies a change of the underlying conditional independence structure [2]. In some cases, this might not be desired. To overcome this issue for spatio-temporal data, a combination of reparametrization and regularization is proposed, called STRF [3], which allows the regularization of large graphical models while keeping the conditional independence structure intact. While the model in the aforementioned work is presented entirely for discrete data, the underlying concept can be extended to continuous data as well. Here, this idea is discussed for multivariate Gaussian data where the conditional independence structure is encoded by the entries of the inverse covariance matrix [2]. Let $\boldsymbol{x} \in \mathbb{R}^n$. The most common representation of the multivariate Gaussian probability density is the *standard form*,

$$p_{\boldsymbol{\mu},\Sigma}(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^n|\Sigma|}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \tag{1}$$

with mean vector $\boldsymbol{\mu} \in \mathbb{R}^n$ and covariance matrix $\Sigma \in \mathbb{R}^{n\times n}$. Here, $|\Sigma|$ denotes the determinant of Σ. The reparametrization from [3] is formulated for models in *exponential family form*

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \exp\left(\langle \boldsymbol{\theta}, \phi(\boldsymbol{x})\rangle - \ln Z(\boldsymbol{\theta})\right) \tag{2}$$

with log partition function $\ln Z(\boldsymbol{\theta}) = \ln \int \exp\left(\langle \boldsymbol{\theta}, \phi(\boldsymbol{x})\rangle\right) \nu(\mathrm{d}^n\boldsymbol{x})$ and base measure ν. In general, the base measure ν might be the counting measure, in which case p is a probability mass function; alternatively, for a continuous random vector, the base measure could be the ordinary Lebesgue measure on $\mathbb{R}$ [5]. Because of this, (2) is valid for discrete and continuous random variables without any changes in notation. Now, we show how to rewrite the multivariate Gaussian in exponential family form. Let $\mathrm{vec}(\cdot) : \mathbb{R}^{n\times n} \to \mathbb{R}^{n^2}$ be a function that maps the elements of a $n \times n$ matrix to a n^2-dimensional vector in an arbitrary but fixed order. If the model parameters $\boldsymbol{\theta} \in \mathbb{R}^d$ and sufficient statistic $\phi : \mathbb{R}^n \to \mathbb{R}^d$ are chosen according to

$$\boldsymbol{\theta} = \begin{bmatrix} -\frac{1}{2}\,\mathrm{vec}(\Sigma^{-1}) \\ \Sigma^{-1}\boldsymbol{\mu} \end{bmatrix} \quad \text{and} \quad \phi(\boldsymbol{x}) = \begin{bmatrix} \mathrm{vec}(\boldsymbol{x}\boldsymbol{x}^\top) \\ \boldsymbol{x} \end{bmatrix}$$

respectively, equations (1) and (2) are equal, i.e., $p_{\boldsymbol{\mu},\Sigma}(\boldsymbol{x}) = p_{\boldsymbol{\theta}}(\boldsymbol{x}), \forall \boldsymbol{x} \in \mathbb{R}^n$. To see this, we apply the n-dimensional Gaussian integral to derive the closed form of $\ln Z(\boldsymbol{\theta})$:

$$\begin{aligned} \ln Z(\boldsymbol{\theta}) &= \ln \int \exp\left(\langle \boldsymbol{\theta}, \phi(\boldsymbol{x})\rangle\right) \nu(\mathrm{d}^n\boldsymbol{x}) = \ln \int \exp\left(-\frac{1}{2}\boldsymbol{x}^\top\Sigma^{-1}\boldsymbol{x} \ + \ \boldsymbol{x}^\top\Sigma^{-1}\boldsymbol{\mu}\right) \nu(\mathrm{d}^n\boldsymbol{x}) \\ &= \ln\left(\sqrt{(2\pi)^n|\Sigma|}\exp\left(\frac{1}{2}\boldsymbol{\mu}^\top\Sigma^{-1}\boldsymbol{\mu}\right)\right). \end{aligned}$$

Plugging this into Eq. (2) and rearranging, one arrives at the multivariate Gaussian density

$$\begin{aligned} p_{\boldsymbol{\theta}}(\boldsymbol{x}) &= \exp\left(\langle \boldsymbol{\theta}, \phi(\boldsymbol{x})\rangle - \ln Z(\boldsymbol{\theta})\right) \\ &= \frac{1}{\sqrt{(2\pi)^n|\Sigma|}}\exp\left(-\frac{1}{2}\boldsymbol{x}^\top\Sigma^{-1}\boldsymbol{x} \ + \ \boldsymbol{x}^\top\Sigma^{-1}\boldsymbol{\mu} \ - \frac{1}{2}\boldsymbol{\mu}^\top\Sigma^{-1}\boldsymbol{\mu}\right) \\ &= \frac{1}{\sqrt{(2\pi)^n|\Sigma|}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) = p_{\boldsymbol{\mu},\Sigma}(\boldsymbol{x}). \end{aligned}$$

2 Spatio-temporal models

Data which is generated simultaneously by multiple sources at consecutive points in time may be described by the affine first-order vector-autoregressive stochastic process

$$\boldsymbol{x}^{t+1} = \boldsymbol{A}^t\boldsymbol{x}^t + \boldsymbol{\varepsilon}^t \quad \text{and} \quad \boldsymbol{x}^0 = \boldsymbol{\varepsilon}^0 \tag{3}$$

with data[1] $\boldsymbol{x}^t, \boldsymbol{x}^{t+1} \in \mathbb{R}^n$, transition matrix $\boldsymbol{A}^t \in \mathbb{R}^{n \times n}$ and Gaussian noise $\boldsymbol{\varepsilon}^t \sim \mathcal{N}(0, \Sigma^t)$. Since $\boldsymbol{\varepsilon}^t$ is a random variable, $\boldsymbol{x}^{t+1}$ is a random variable as well, with $\boldsymbol{x}^{t+1} \sim \mathcal{N}(\boldsymbol{A}^t\boldsymbol{x}^t, \Sigma^t)$. Notice that Σ^t may be interpreted as *spatial* structure between variables which are measured simultaneously at time t. Likewise, $\boldsymbol{A}^t$ is the *temporal* structure that describes how measurements at time $t+1$ arise from measurements at time t. Without any additional assumption the spatio-temporal structure $(\boldsymbol{A}^t, \Sigma^t)$ could change at any point in time. This causes serious problems in parameter estimation since the underlying model would have an infinite number of parameters, namely $\{(\boldsymbol{A}^t, \Sigma^t)\}_{0 \leq t \leq \infty}$. We assume that the process is periodic with period T, i.e., $\boldsymbol{A}^t = \boldsymbol{A}^{t+T}$ and $\Sigma^t = \Sigma^{t+T}$, to overcome this issue. Moreover, to match the setting in [3], we make two additional assumptions. First, the conditional independence structure is constant over time and given, i.e., $\Sigma^t = \Sigma, \forall 0 \leq t \leq \infty$. This may also be interpreted as strong homoscedasticity assumption which is quite mandatory in most regression settings. However, a weaker homoscedasticity assumption, namely $\Sigma^t = \Sigma^{t+T}$, would suffice. Second, any variable $\boldsymbol{x}_i^t$ may only have temporal influence on variables $\boldsymbol{x}_j^{t+1}$ on which it has spatial influence, i.e., $\Sigma_{i,j}^{-1} = 0 \Rightarrow \boldsymbol{A}_{ij}^t = 0$. Both assumptions are mainly for ease of notation and can be relaxed or even fully removed. Finally, the task is to estimate the entries of the matrices $\Sigma, \boldsymbol{A}^0, \boldsymbol{A}^1, \ldots, \boldsymbol{A}^{T-1}$ where the non-zero pattern in Σ^{-1} (and hence in all $\boldsymbol{A}^t$) is known. Let $\boldsymbol{X} = (\boldsymbol{x}^0, \boldsymbol{x}^1, \ldots, \boldsymbol{x}^{T-1})$ be one complete period. Furthermore, let $\mathcal{D} = (\boldsymbol{X}^1, \ldots, \boldsymbol{X}^N)$ be a series that consists of N consecutive periods with a total of NT data points. Our definition of the stochastic process (3) implies that $p(\mathcal{D}) = p(\boldsymbol{x}^0)\prod_{j=1}^{NT} p(\boldsymbol{x}^j \mid \boldsymbol{x}^{j-1})$. Each non-zero entry of any $\boldsymbol{A}^t$ may then be consistently estimated by the maximum-likelihood (ML) method, which is equivalent to solving nT least-squares regression problems of the form $\min_{\boldsymbol{A}_i^t} \sum_{(\boldsymbol{x}^t, \boldsymbol{x}^{t+1}) \in \mathcal{D}} (\boldsymbol{x}_i^{t+1} - \langle \boldsymbol{A}_i^t, \boldsymbol{x}^t \rangle)^2$ where $\boldsymbol{A}_i^t$ is the i-th row of matrix $\boldsymbol{A}^t$. The non-zero elements of Σ may also be estimated by the ML method—afterwards or jointly together with the $\boldsymbol{A}^t$ matrices.

3 Reparametrization

While models like (3) are easy to estimate, they suffer from the fact that multiple consecutive matrices $\boldsymbol{A}^t, \boldsymbol{A}^{t+1}$ might be very similar. This issue arises quite naturally when sensors are sampled periodically and independent of the data. As a result, the model has too many degrees of freedom and might be prone to overfitting the data. In [3], we propose to reparametrize exponential families over spatio-temporal data such that $\boldsymbol{\theta}^t = \sum_{i=0}^{t} \frac{1}{t-i+1}\Delta^i$. If we investigate $p(\mathcal{D})$ (as defined above) in exponential family form, the definition of conditional probability implies for the factors $p(\boldsymbol{x}^t \mid \boldsymbol{x}^{t-1})$ that

$$\boldsymbol{\theta}^t = \begin{bmatrix} -\frac{1}{2}\operatorname{vec}(\Sigma^{-1}) \\ \operatorname{vec}(\Sigma^{-1}\boldsymbol{A}^t) \end{bmatrix} \quad \text{and} \quad \phi(\boldsymbol{x}^t, \boldsymbol{x}^{t-1}) = \begin{bmatrix} \operatorname{vec}(\boldsymbol{x}^t\boldsymbol{x}^{t\top}) \\ \operatorname{vec}(\boldsymbol{x}^t\boldsymbol{x}^{t-1\top}) \end{bmatrix}.$$

[1] Superscripts like $\boldsymbol{x}^t$ denote object $\boldsymbol{x}$ at *time* t and not the t-th power of $\boldsymbol{x}$.

Multiple issues show up if the aforementioned reparametrization is applied to these parameters, because of the coupling $\Sigma^{-1}\boldsymbol{A}^t$ in $\boldsymbol{\theta}^t$. First, $\boldsymbol{A}^t$ and Σ can no longer be estimated separately. Second, Σ is a covariance matrix and must hence be positive definite. It is unclear how positive definiteness of Σ can be assured in this setting. Third, it is assumed that $\Sigma^{-1}_{i,j} = 0 \Rightarrow \boldsymbol{A}^t_{ij} = 0$, a constraint that is hard to implement with the above reparametrization. Finally, it would be desirable to read off the spatial and temporal influence directly from the parameters. But here, one matrix inversion and one matrix multiplication have to be carried out in order to compute $\boldsymbol{A}^t = \Sigma\Sigma^{-1}\boldsymbol{A}^t$.

As an alternative, we propose to apply the reparametrization directly to the matrices $\boldsymbol{A}^t$ instead, i.e., $\boldsymbol{A}^t = \sum_{i=0}^{t} \frac{1}{t-i+1}\Delta^i$. This addresses all of the above issues but still comes with some restrictions. The estimation of the transition matrices is indeed independent from the estimation of Σ. However, in contrast to the basic model from Sec. 2, the reparametrized transition matrices Δ^t can no longer be estimated in separate least-squares problems, since the Δ^i appear in the definition of all $\boldsymbol{A}^t$ with $t \geq i$. Regarding the second problem, maintaining the positive definiteness of Σ can now be handled by the same methods which apply to the non-reparametrized model [1]. The same holds for the third problem. The entries of $\boldsymbol{A}^t$ and Σ^{-1} which should be 0 can be initialized to 0 and not be updated during the estimation procedure. The last problem is obviously solved, since the $\boldsymbol{A}^t$ and Σ are no longer coupled. Finally, ℓ_1-regularization [1, 4] is applied to the new parameters Δ^t which pins their entries down to zero, whenever the difference between parameters of consecutive time steps is small. Preliminary results show, that the proposed procedure delivers sparse models without changing the conditional independence structure.

References

[1] Jerome Friedman, Trevor. Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[2] Steffen L. Lauritzen. *Graphical Models*. Oxford University Press, Oxford, UK, 1996.

[3] Nico Piatkowski, Sangkyun Lee, and Katharina Morik. Spatio-temporal random fields: compressible representation and distributed estimation. *Machine learning*, 93(1):115–139, 2013.

[4] Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society Series B*, 67(1):91–108, 2005.

[5] Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008.

kCQL: A Novel Paradigm for the Acquisition of OS Data

Jochen Streicher
Lehrstuhl für Informatik 12
Technische Universität Dortmund
jochen.streicher@tu-dortmund.de

Declarative acquisition of both OS events and state was not possible before. With kCQL, it is. Such a highly abstract interface to low-level OS data is powerful in terms of expressiveness and ease of use, and the prototype's overhead is on a reasonable scale. Nevertheless, there are still unanswered questions regarding its efficient implementation and the choice of abstractions in the data model.

1 Introduction

Dynamic instrumentation or tracing frameworks, like *SystemTap, DTrace* [4], or *Fay* [5], allow one to extract internal data from operating systems tasks at a high level of abstraction, or even, partially, in a declarative way. The more recent *PiCO QL [6]* enables actual (non-modifying) SQL queries over a relational representation of the Unix kernel state, something which is not possible using existing event-based data acquisition tools. However, PiCO QL does not allow tracing events, and is thus a complementary approach to existing tracing tools. An earlier protoype that aimed to combine the best of both worlds was based on the *Aurora* stream query language [1]. The experiences from that approach led to kCQL.

```
Packets: RSTREAM (
  SELECT packet.datalen AS len, packet.direction AS dir, process.pid AS
   pid, process.name AS pname
  FROM packet [now], socket, process
  WHERE packet.sid = socket.sid AND socket.pid = process.pid
);
```

Listing 1: *Packets*: Assigns network packets to processes.

```
Files: SELECT DISTINCT P.name, F.inode_name, F.inode_mode & 0400,
    F.inode_mode & 040, F.inode_mode & 4
FROM process AS P, file AS F, process_group AS PG
WHERE P.pid = F.pid AND P.pid = PG.pid AND F.mode & 1 = 1 AND (F.inode_uid
    != P.cred_fsuid OR F.inode_mode & 0400 = 0) AND (P.cred_fsgid !=
    PG.gid OR F.inode_mode & 040 = 0) AND F.inode_mode & 4 = 0;
```

Listing 2: *Files*: A continuously updated relation containing files opened with currently insufficient permissions.

2 kCQL

Data model and language of kCQL are based on the Stanford Continuous Query Language (CQL). [3] The data model is based on streams and time-varying relations. Contrary to Aurora [1], relational operations are allowed only on relations. This way, CQL stays very close to SQL, extending it simply by a few operators to convert between streams and relations. Listing 1 shows a query that produces a data stream by capturing network packets and assigns them to the corresponding process, whereas the query in Listing 2 tracks files opened for reading by processes that do not have the necessary permissions. Note, that the latter may look like a typical relational database query, but it *is* a continious query, which continuously produces results.

The overall architecture (Figure 1 Left) is comparable to the earlier prototype: Relation and event sources offered by different address spaces are visible and can be used in queries from any (other) address space, leading to a distributed query plan with a DSMS engine instance in every participating address space.

The DSMS engine is based on Stanford STREAM [2], which represents relations as *update streams*. That is different from the earlier protoype, which used *direct access* to OS data structures when stream tuples are joined to them. While this representation leads to consistent joins, kCQL has to track every update to time-varying relations and thus requires instrumentation for any event that updates these. The technical details of its integration into kCQL, in particular the adaptation to asynchronous data sources, are described in the paper. [7]

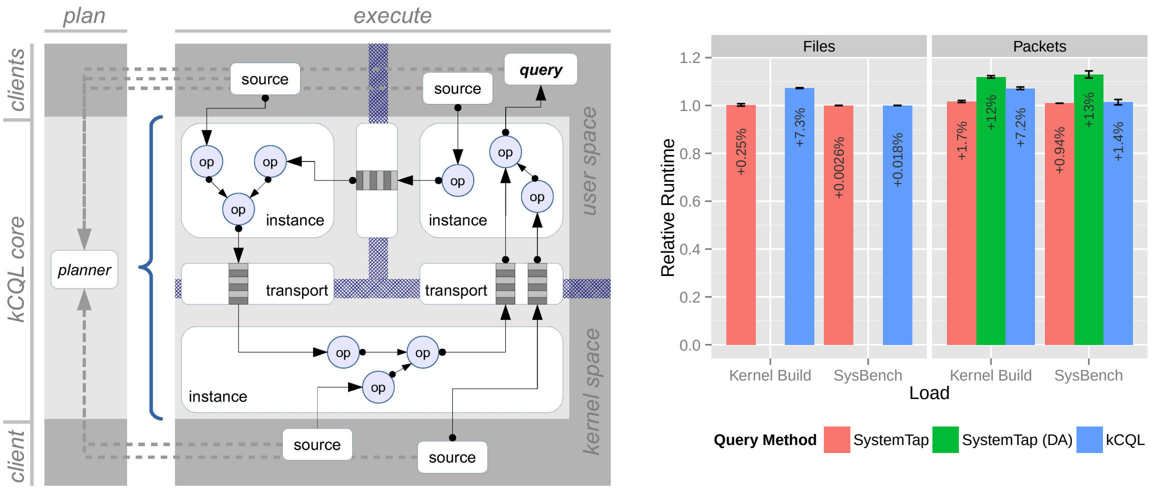


Figure 1: Left: Architecture of kCQL. Right: Comparison of average relative runtime compared to clean runs of kernel build and SysBench.

3 Results

The declarative approach was evaluated regarding the overhead of query execution using the *Packets* (Listing 1) and *Files* (Listing 2) queries and corresponding SystemTap variants thereof. For the *Packets* query, two SystemTap implementations were used: One that uses update streams for relations, resembling kCQL's mode of operation, and one that uses direct access for the process list.

Two different load scenarios were used, both without (baseline) and with the described data acquisition scenarios. The load scenarios consist of SysBench[1] (CPU and user mode only) and a full Linux kernel build (same as above, CPU and I/O activity). For the first scenario, a second machine connected with a direct Gigabit Ethernet connection sent TCP packets at full speed. Figure 1 (right) shows the relative runtime for these loads compared to the baseline.

It is obvious that direct access to kernel state (DA) is rather unfavorable for the *Packets* query. However, using incrementally updated copies of the kernel state, the painstakingly optimized SystemTap scripts naturally generate less overhead than kCQL. While kCQL does not yet excel here, the overhead is already reasonable. The SystemTap implementation of the *Files* query can make use of an invariant: The situation we track in the query can only occur after a call to *setuid*, which is not apparent from kCQL's level of abstraction.

[1] http://sysbench.sourceforge.net

4 Conclusion and Future Work

While resulting in consistent joins, the *update stream* representation is not generally superior to *direct access* with regard to performance. The latter obviously leads to less overhead for joins between streams and relations when the stream emits tuples at low frequency or when updates to the relation are frequent. As already became apparent with the earlier prototype, direct access gives rise to further performance gains, namely the exploitation of *natural indices* and *explicit weak consistency*. Consequently, those variants will have to be evaluated within kCQL in order to come up with an answer.

While the shortcut over *setuid* used by SystemTap variant of the *Files* query would be possible with kCQL, using a stream of setuid events, this would mean leaving the convenient level of abstraction based on relations. In fact, kernel data contains multiple related levels of abstraction: We can write a query that counts packets, although the kernel does that already on its own. Explicitly modeling these relations could be used for query optimization, ultimately allowing one to automatically use the setuid-shortcut, even if the query is based solely on relations on a higher level of abstraction. This, as well the demand for rapid development of new data providers suggests a domain specific language to capture the relations between the abstract data model and the low-level OS code.

The already distributed nature of kCQL as well as the easy extensibility of the data model to multiple instances of data sources enables natural extension to multiple devices, for example, a distributed ubiquitous system.

References

[1] Daniel J. Abadi, Don Carney, Ugur Çetintemel, Mitch Cherniack, Christian Convey, Sangdon Lee, Michael Stonebraker, Nesime Tatbul, and Stan Zdonik. Aurora: A new model and architecture for data stream management. *The VLDB Journal – The International Journal on Very Large Data Bases*, 12(2):120–139, August 2003.

[2] Arvind Arasu, Brian Babcock, Shivnath Babu, John Cieslewicz, Mayur Datar, Keith Ito, Rajeev Motwani, Utkarsh Srivastava, and Jennifer Widom. STREAM: The Stanford data stream management system. Technical Report 2004-20, Stanford InfoLab, 2004.

[3] Arvind Arasu, Shivnath Babu, and Jennifer Widom. The CQL continuous query language: Semantic foundations and query execution. *The VLDB Journal – The International Journal on Very Large Data Bases*, 15(2):121–142, June 2006.

[4] Bryan M. Cantrill, Michael W. Shapiro, and Adam H. Leventhal. Dynamic instrumentation of production systems. June/July 2004.

[5] Úlfar Erlingsson, Marcus Peinado, Simon Peter, Mihai Budiu, and Gloria Mainar-Ruiz. Fay: Extensible distributed tracing from kernels to clusters. *ACM Transactions on Computer Systems*, 30(4):13:1–13:35, November 2012.

[6] Marios Fragkoulis, Diomidis Spinellis, Panos Louridas, and Angelos Bilas. Relational access to Unix kernel data structures. pages 12:1–12:14, 2014.

[7] Jochen Streicher, Alexander Lochmann, and Olaf Spinczyk. kCQL: Declarative stream-based acquisition and processing of diagnostic OS data. In *Proceedings of the Conference on Timely Results in Operating Systems (TRIOS)*. ACM, 2015.

SFB 876 Verfügbarkeit von
Information durch Analyse unter
Ressourcenbeschränkung

Subproject A2
Algorithmic aspects of learning methods in embedded systems

Christian Sohler Jens Teubner

Curve Simplification Approach to Clustering Time Series Under the Fréchet Distance

Amer Krivošija
Lehrstuhl für effiziente Algorithmen und Komplexitätstheorie
Technische Universität Dortmund
amer.krivosija@tu-dortmund.de

For a given problem of clustering of n unidimensional time series of complexity m under the Fréchet distance, it is a curve simplification approach using *signatures* that enables us to extract the key information on the input time series as well to find the constant-sized candidates set for cluster centers. We explore the properties of signatures and ask if such approach would be possible in the multidimensional case.

Introduction

Let the set $T = \{\tau_1, \ldots, \tau_n\}$ be given, where τ_i is a time series of complexity m, for $1 \leq i \leq n$, and every entry of time series is an ordered pair (w_j, t_j), where $w_j \in \mathbb{R}$ and t_j is a time stamp, for $1 \leq j \leq m$. The problem of clustering of time series from T, called curves, under the Fréchet distance was first studied in [6]. The $(1+\varepsilon)$-approximation algorithms with near-linear running times in terms of the input size were found for k-center and k-median problems, under the condition that the complexity of the cluster centers is bounded by the constant ℓ. We assume that k, ℓ and ε are input constants. These problems were considered for the case that the values w_j in the input time series are from $\mathbb{R}$ (i.e. unidimensional case), and it is shown that even then these problems are **NP**-hard.

For these problems to be solved it is worth to consider the curve simplifications technique. They reduce the complexities of the input time series by ignoring the measurements –

vertices of the input curve that do not increase the Fréchet distance to the original curve beyond some limit, therefore being “not important”. The simplification approach addresses therefore the problem of noisy measurement as well.

Formally, a simplification of a curve is a curve which is lower in complexity (it has fewer vertices) than the original curve and which is similar (being at the small Fréchet distance) to the original curve. This is captured by the following standard definitions.

Definition 1. *We call a curve π a minimum-error ℓ-simplification of τ if for any curve π' of at most ℓ vertices, it holds that $d_F(\pi', \tau) \geq d_F(\pi, \tau)$.*

Definition 2. *We call a curve π a minimum-size ε-simplification of τ if $d_F(\pi, \tau) \leq \varepsilon$ and for any curve π' such that $d_F(\pi', \tau) \leq \varepsilon$, it holds that the complexity of π' is at least as much as the complexity of π.*

The simplification problem has been studied under different names for multidimensional curves and under various error measures, in domains, such as cartography, computational geometry, and pattern recognition. Often, the simplified curve is restricted to vertices of the input curve and the endpoints are kept. However, in the clustering setting under Fréchet distance, we need to use the more general problem definitions stated above.

Historically, the first minimal-size curve simplification algorithm was a heuristic algorithm independently suggested in the 1970’s by Ramer and Douglas and Peucker [4, 10] and it remains popular in the area of geographic information science until today. It uses the Hausdorff error measure and has running time $O(m^2)$ (where m denotes the complexity of the input curve), but does not offer a bound to the size of the simplified curve. Recently, worst-case and average-case lower bounds on the number of vertices obtained by this algorithm were proven by Daskalakis *et al.* [3]. Imai and Iri [8] solved both the minimum-error and minimum-size simplification problem under the Hausdorff distance by modeling it as a shortest path problem in directed acyclic graphs.

Curve simplification using the Fréchet distance was first proposed by Godau [7]. The current state-of-the-art approximation algorithm for simplification under the Fréchet distance was suggested by Agarwal *et al.* [1]. This algorithm computes a 2-approximate minimal-size simplification in time $O(m \log m)$.

For time series, a concept similar to simplification called segmentation has been extensively studied in the area of data mining. The standard approach for computing exact segmentations is to use dynamic programming which yields a running time of $O(m^2)$.

Our contribution

To capture the shape of the input time series and its critical points, therefore reducing its complexity while describing the input well, we introduce the concept of *signatures* as a special kind of simplification. Our definition aligns with the work on computing important minima and maxima in the context of time series compression. Such technique is similar to the idea of shortcutting used for partial curve matching in [2, 5]. Intuitively, the signatures provide us with the "shape" of a time series at multiple scales.

The formal definition of δ-signature (for a given $\delta > 0$) is long and already given in [9]. For a given time series τ, observed as polygonal curve $\tau : [0, 1] \to \mathbb{R}$, its δ-signature is a curve $\sigma = v_1, \ldots, v_\ell$, defined by a subset of vertices of τ.

The signatures always exist. We can calculate them efficiently for a given error δ in time $O(m)$, as well as for the given goal complexity ℓ in time $O(m \log m)$, using the idea introduced by Driemel and Har-Peled [5]. They suggested the concept of a vertex permutation with the aim of preprocessing a curve for curve simplification. The idea is that any prefix of the permutation represents a bicriteria approximation to the minimal-error curve simplification. Our signatures have the strong hierarchical structure and enable us to calculate the 2-approximation to the minimum-error ℓ-simplification and later, the efficient clustering algorithms.

Our main technical result, that allows the reduction of the input curves and the constant-sized candidates set for the cluster centers is the following theorem.

Theorem 3. *Let $\sigma = v_1, \ldots, v_\ell$ be a δ-signature of $\tau = w_1, \ldots, w_m$. Let $r_j = [v_j - \delta, v_j + \delta]$ be ranges centered at the vertices of σ ordered along σ, where $r_1 = [v_1 - 4\delta, v_1 + 4\delta]$ and $r_\ell = [v_\ell - 4\delta, v_\ell + 4\delta]$. Let π be a curve with $d_F(\tau, \pi) \leq \delta$ and let π' be a curve obtained by removing some vertex $u_i = \pi(p_i)$ from π with $u_i \notin \bigcup_{1 \leq j \leq \ell} r_j$. It holds that $d_F(\tau, \pi') \leq \delta$.*

We describe the proof idea. We consider the witness matching f from π to τ which maps each point on π to a point on τ within distance δ. When u_i is removed, we consider the new curve π', which differs from π in the subcurve $\pi[p^-, p^+]$. We want to construct the witness Fréchet matching f' based on f, such that the distance is kept bounded by δ. We need to show that the subcurve $\tau[f(p^-), f(p^+)]$ can be matched to some subcurve on π', while respecting the monotonicity of the matching.

To do so, we have to perform a case analysis, based on the structure of the two curves. We assume w.l.o.g. that $\pi(p_i)$ is a local minimum on π and that it holds for the signature vertices that $\tau(s_j) < \tau(s_{j+1})$, where $\tau[s_j, s_{j+1}]$ is a curve segment that contains the minimal value on $\tau[f(p^-), f(p^+)]$. Both assumptions can be removed by mirroring or reversal traversing of the curves π and τ. We conclude that the only signature vertex that can be in the subcurve $\tau[f(p^-), f(p^+)]$ is $\tau(s_{j+1})$.

We divide the proof into subcases depending whether the signature vertex $\tau(s_{j+1})$ is outside or inside of $\tau[f(p^-), f(p^+)]$. It appears that the latter case is harder to be repaired, since it can induce iterative matching scheme. We consider the curve $\pi[p^+, 1]$ and observe whether it further ascends over or descends below some value or stays within these bounds. Our construction ensures that the extended subcurves cover the subcurve $\tau[f(p^-), f(p^+)]$ and that the subcurves line up appropriately with the subcurves of π before and after matched subcurve.

It is a natural question whether the approach to curve simplification using signatures is possible in multidimensional space? The answer appears to be (probably) negative, since the properties of the signatures heavily use the fact that the input time series values lie on one axis, but there is no proof yet for such a claim.

References

[1] P. K. Agarwal, S. Har-Peled, N. H. Mustafa, and Y. Wang. Near-linear time approximation algorithms for curve simplification. *Algorithmica*, 42:203–219, 2005.

[2] K. Buchin, M. Buchin, and Y. Wang. Exact algorithms for partial curve matching via the Fréchet distance. In *Proceedings of the 20th ACM-SIAM Symposium on Discrete Algorithms*, pages 645–654, 2009.

[3] C. Daskalakis, I. Diakonikolas, and M. Yannakakis. How good is the chord algorithm? *SIAM Journal of Computing*, page to appear, 2015. http://arxiv.org/abs/1309.7084.

[4] D. H. Douglas and T. K. Peucker. *Algorithms for the Reduction of the Number of Points Required to Represent a Digitized Line or its Caricature*, pages 15–28. John Wiley & Sons, Ltd, 2011.

[5] A. Driemel and S. Har-Peled. Jaywalking your dog – computing the Fréchet distance with shortcuts. *SIAM Journal of Computing*, 42(5):1830–1866, 2013.

[6] A. Driemel, A. Krivošija, and C. Sohler. Clustering time series under the Fréchet distance. *Proceedings of the 26th ACM-SIAM Symposium on Discrete Algorithms*, to appear, 2016.

[7] M. Godau. A natural metric for curves —- computing the distance for polygonal chains and approximation algorithms. In *Proceedings of the 8th Annual Symposium on Theoretical Aspects of Computer Science*, pages 127–136. Springer, 1991.

[8] H. Imai and M. Iri. Polygonal approximations of a curve – formulations and algorithms. In G. Toussaint, editor, *Computational Morphology*, pages 71–86. North-Holland, Amsterdam, 1988.

[9] K. Morik and W. E. Rhode. Technical report for collaborative research center sfb 876 - graduate school. Technical Report 10, TU Dortmund, 2014.

[10] U. Ramer. An iterative procedure for the polygonal approximation of plane curves. *Computer Graphics and Image Processing*, 1(3):244–256, 1972.

Optimal Algorithms for Diameter in Sliding Windows

Chris Schwiegelshohn

Lehrstuhl für effiziente Algorithmen und Komplexitätstheorie
Technische Universität Dortmund
chris.schwiegelshohn@tu-dortmund.de

In this paper, we investigate small space summaries for the diameter, i. e., the maximum distance between two input points in sliding windows. We give an algorithm maintaining a $3 + \varepsilon$ approximate diameter in arbitrary metric spaces with nearly optimal space requirements. Our lower bounds separate the space requirements of sliding windows from those found in insertion only streams. To our knowledge, no such separation was previously known.

1 Introduction

Analyzing big data sets from streams is a topic that has received considerable theoretical and practical attention. In the normal streaming setting, we constrain our algorithms to use as little space as possible while computing high-quality solutions. The complexity of diameter is well understood for insertion-only streams where input points arrive one by one. The more general settings like dynamic streams and the sliding window model have also received some attention. Both models aim to incorporate dynamic behavior; in dynamic streams input points are removed via a dedicated delete operation and in the sliding window model older input expires as new elements arrive. In this paper, we consider a fixed window size consisting of N points, which is the most studied variant of this model in theoretical computer science, but our algorithms also work in the case that the window considers a dynamic number of points, for instance given in some time frame.

1.1 Related Work

In the metric distance oracle model there exists a folklore 2 approximation that maintains the first point p and the point with maximum distance from p. Guha [5] showed this algorithm to be essentially optimal, as no algorithm storing less than $\Omega(n)$ points can achieve a ratio better than $2 - \varepsilon$ for any $\varepsilon > 0$. In Euclidean spaces, Agarwal et al. [1] proposed a $(1 + \varepsilon)$-approximation using $O(\varepsilon^{-(d-1)/2})$ points. The best streaming algorithm with no exponential dependency on d is due to Agarwal and Sharathkumar [2] with an almost tight approximation ratio of $\sqrt{2} + \varepsilon$ in $O(d\varepsilon^{-3}\log(1/\varepsilon))$ space. For dynamic streams, Indyk [6] gave a sketching scheme with approximation factor $c > \sqrt{2}$ and space $O(dn^{1/(c^2-1)}\log n)$. The first work to consider the diameter in the sliding window model was done by Feigenbaum et al. [4]. Their algorithm uses $O(\frac{1}{\varepsilon}\log^3 N(\log\alpha + \log\log N + \log\frac{1}{\varepsilon})$ bits of space in one dimension and $O(\left(\frac{1}{\varepsilon}\right)^{(d+1)/2}\log^3 N(\log\alpha + \log\log N + \frac{1}{\varepsilon}))$ bits of space in d dimensions, where $\alpha = \frac{\max ||p-q||}{\min ||p-q||}$ is the ratio of largest and smallest possible distance by the points. The also give a lower bound of $\Omega(\frac{1}{\varepsilon}\log N \log\alpha)$ for a $(1+\varepsilon)$ approximation factor in one dimension and, implicitly, a $\Omega(\log R)$ space bound for any multiplicative approximation factor. This lower bound was later matched by Chan and Sadjad [3], who also gave an improved space bound of $O(\left(\frac{1}{\varepsilon}\right)^{(d+1)/2}\log\frac{\alpha}{\varepsilon})$ points for higher dimensions. For more general metric spaces, they obtain a 6 approximation using $O(\sqrt{N}\log\alpha)$ space.

2 Our Contribution and Techniques

We give a $(3 + \varepsilon)$ approximation for the metric diameter problem using $O(\varepsilon^{-1}\log\alpha)$ points. In addition, given a window size N, we provide a lower bound of $\Omega(\sqrt[3]{N})$ points for any algorithm with an approximation factor better than 3 in the metric oracle distance model. Together with the $\Omega(\log\alpha)$ lower bound for any multiplicative approximation by Feigenbaum et al. [4], this shows that our algorithm is almost tight. Our algorithm aims to find for each estimate γ of the diameter two certificate points with distance greater than γ, while maintaining the two most recent points close to the two points forming the certificate. With every additional input point, we check whether we are able to update the certificate to a more recent timestamp. The main technical lemma and algorithm are presented in Section 3.

Theorem 1. *Given a set of points A with aspect ratio α and a window of size N, there exists an algorithm computing a $3(1 + \epsilon)$-approximate solution for the metric diameter problem storing at most $8/\epsilon \cdot \ln\alpha$ points.*

For space restrictions, we will not provide any details on the lower bound, but simply state the result.

Theorem 2. *Any randomized sliding window algorithm outputting two points at distance greater than $\frac{1}{3} \cdot diam(A)$ with probability bounded away from $\frac{1}{2}$ for the distance oracle metric diameter problem with constant aspect ratio requires $\Omega(\sqrt[3]{N})$ memory.*

3 Algorithm and Analysis

Algorithm 1 Sliding Window Algorithm for $(\gamma, 3 \cdot \gamma)$-gap Diameter

```
$c_{old}, c_{new} \leftarrow$ null
$r_{old}, r_{new} \leftarrow$ null
for all element $p$ of the stream do
    if certificate point $c_{old}$ expires then
        if $c_{new} \neq$ null then
            $c_{old} \leftarrow r_{old}$; $r_{old} \leftarrow r_{new}$; $c_{new}, r_{new} =$ null;
        else
            $c_{old} \leftarrow r_{old}$;
        end if
    end if
    INSERT($p$)
end for
procedure Insert($p$)
    if ($c_{old} =$ null $\wedge$ $c_{new} =$ null) then
        $c_{old}, r_{old} = p$;
    else if (dist($p, c_{old}$) $< \gamma \wedge c_{new} =$ null) then
        $r_{old} = p$
    else if dist($p, r_{new}$) $> \gamma$ then
        $c_{old}, r_{old} \leftarrow r_{new}$; $c_{new}, r_{new} \leftarrow p$;
    else if dist($p, c_{new}$) $> \gamma$ then
        $c_{old} \leftarrow c_{new}$; $c_{new} \leftarrow p$; $r_{old} \leftarrow r_{new}$; $r_{new} \leftarrow p$;
    else if (dist($p, r_{old}$) $> \gamma \wedge r_{old} \neq c_{old}$) then
        $c_{old} \leftarrow r_{old}$; $c_{new} \leftarrow p$; $r_{old} \leftarrow r_{new}$; $r_{new} = p$;
    else if (dist($p, c_{old}$) $> \gamma \wedge c_{new} =$ null) then
        $c_{new} \leftarrow p$;
    else
        $r_{new} = p$;
    end if
end procedure
```

Lemma 1. *At any given time t, the following two statements hold:*

1. *If $c_{old}(t) \neq r_{old}(t)$ and $c_{new}(t) \neq$ **null**, then dist$(r_{old}(t), c_{new}(t)) \leq \gamma$ and dist$(r_{old}(t), r_{new}(t)) \leq \gamma$.*
2. *If $c_{new}(t) \neq$ **null**, then $T(c_{old}(t)) \geq T(r_{old}(t)) = 1 + T(c_{new}(t)) \geq T(r_{new}(t))$.*

Lemma 2. *If at time t $c_{new}(t) =$ **null**, then for any two points p and p' in the current window, we have*

- *If $T(p), T(p') \geq T(c_{old}(t))$, then dist$(p, p') \leq 2\gamma$ and*
- *If $T(p) < T(c_{old}(t))$, then dist$(p, c_{old}(t)) \leq \gamma$.*

Proof. We proceed by induction over arriving points. The base case of our induction consists of only one point and is therefore trivial. For the inductive step, we first assume that at time $t-1$ we also had only one center $c_{old}(t-1)$. If $c_{old}(t-1)$ expired, then $c_{old}(t) = r_{old}(t-1)$ and for all points q submitted after $c_{old}(t-1)$ we have by inductive hypothesis $\mathrm{dist}(q, c_{old}(t-1)) \leq \gamma$ and hence $\mathrm{dist}(q, c_{old}(t)) \leq 2\gamma$. In addition, the newest point p did not produce a certificate, so $\mathrm{dist}(p, c_{old}(t)) \leq \gamma$. If $c_{old}(t-1)$ did not expire, then the claim holds for all points except the newest point p by inductive hypothesis. For p we have $\mathrm{dist}(p, c_{old}(t-1)) \leq \gamma$, since we did not produce a new certificate.

What remains is the case where we had two certificate points and $c_{old}(t-1)$ expired. Any point submitted prior to $c_{old}(t-1)$ is expired. In this case $c_{old}(t) = r_{old}(t-1)$. Every pair of points p and p' submitted after $c_{old}(t-1)$ and prior to $r_{old}(t-1) = c_{old}(t)$ did not produce a certificate with $c_{old}(t-1)$ and therefore we have $\mathrm{dist}(p, c_{old}(t-1)) \leq \gamma$ and $\mathrm{dist}(p, p') \leq 2\gamma$. Note that in particular $\mathrm{dist}(p, c_{old}(t)) \leq 2\gamma$, which proves item 1.

For item 2, we first observe $r_{old}(t-1) = r_{old}(t-T(c_{new}(t-1))$ due to Lemma 1.2. Every point p submitted after $r_{old}(t-1)$ at some time $T(c_{new}(t-1)) \leq t' \leq t-1$ became $r_{new}(t')$, and by Lemma 1.1 we then have $\mathrm{dist}(p, c_{old}(t)) = \mathrm{dist}(p, r_{old}(t-1)) \leq \gamma$. □

References

[1] Pankaj K. Agarwal, Jirí Matousek, and Subhash Suri. Farthest neighbors, maximum spanning trees and related problems in higher dimensions. *Comput. Geom.*, 1:189–201, 1991.

[2] Pankaj K. Agarwal and R. Sharathkumar. Streaming algorithms for extent problems in high dimensions. *Algorithmica*, 72(1):83–98, 2015.

[3] Timothy M. Chan and Bashir S. Sadjad. Geometric optimization problems over sliding windows. *Int. J. Comput. Geometry Appl.*, 16(2-3):145–158, 2006.

[4] Joan Feigenbaum, Sampath Kannan, and Jian Zhang. Computing diameter in the streaming and sliding-window models. *Algorithmica*, 41(1):25–41, 2004.

[5] Sudipto Guha. Tight results for clustering and summarizing data streams. In *Database Theory - ICDT 2009, 12th International Conference, St. Petersburg, Russia, March 23-25, 2009, Proceedings*, pages 268–275, 2009.

[6] Piotr Indyk. Better algorithms for high-dimensional proximity problems via asymmetric embeddings. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms, January 12-14, 2003, Baltimore, Maryland, USA.*, pages 539–545, 2003.

Subproject A3
Methods for Efficient Resource Utilization in Machine Learning Algorithms

Jörg Rahnenführer   Peter Marwedel

Ressource-Efficient Parallel Machine Learning Applications

Helena Kotthaus
Computer Science 12
TU Dortmund University
helena.kotthaus@tu-dortmund.de

Our goal is to provide a resource-aware scheduling strategy for parallel machine learning R programs. To develop such a strategy we need to analyze the performance of those applications. Our tool traceR [1] allows the user to profile resource usage of an application [2]. This profiling functionality was previously limited to sequential R applications. Parallel computing, however, is becoming more and more popular since R is increasingly used to process large data sets. Here, a vast amount of resources is needed. We therefore have improved traceR to allow for profiling parallel applications also. In the following, we will show how traceR is applied to analyze parallel R applications, and how it could be integrated to guide scheduling decisions in our resource-aware model based optimization framework.

1 Parallel Performance Analysis

TraceR can be used for common cases, like parallelization on multiple cores or parallelization on multiple machines [3, 4, 5]. It produces profiles for CPU and memory utilizations to support the development of parallel applications. TraceR can also serve as a constraint to determine the degree of parallelism. The gain from parallel execution can be negated if the memory requirements of the processes exceed the capacity of the RAM. By triggering to many parallel jobs the OS starts to swap out memory which leads to inefficient resource utilization. An example profile produced by traceR illustrating this case is shown in Figure 1. Here, a program that evaluates different parameter configurations of a SVM classification task is analyzed. The x-axis represents the runtime in seconds. For each

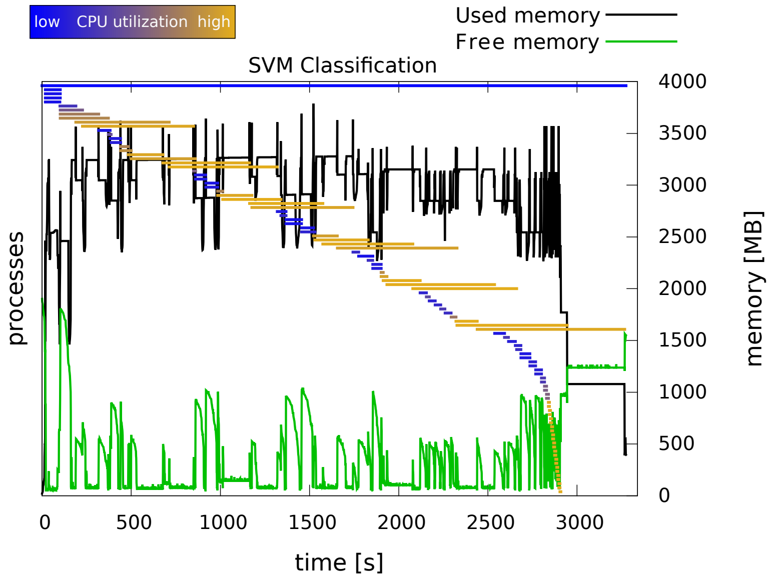


Figure 1: Relative CPU utilization and memory utilization of an R program evaluating different parameter settings of a SVM classification task using load balancing on 4 cores on a single machine with 2 GB of RAM.

parallel job runtime and CPU utilization are illustrated by a horizontal line. The length of a line represents the runtime, while its color indicates the relative CPU utilization and varies from blue via purple to orange. CPU utilization is calculated over the entire runtime of a process. The master process is shown on top while the worker processes are shown below, sorted by starting time. The y-axis on the right side of Figure 1 shows memory utilization. Memory behavior is indicated by two curves. The green curve shows the amount of free main memory during program execution, reported by the Linux kernel, and the black curve shows the amount of memory allocated by all parallel processes over time. All jobs are processed on 4 cores on a single machine. Due to the R load balancing mechanism, those jobs are dynamically allocated to 4 worker processes. Each job evaluates the performance of an SVM for a given parameter setting. As can be seen (Figure 1) their is a high variance of computation times between parallel jobs, even when load balancing is applied. This indicates that R's load balancing mechanism is not sufficient for our SVM task. Additionally their are several blue lines representing low CPU utilizations and thus inefficient resource utilization. This is caused by jobs waiting for processing their swapped out data. It is also represented by the green curve that indicates free memory which is most of the time close to zero. Here, one solution for reducing memory utilization and for producing a high cpu utilization is to use only 2 worker processes instead of 4.

This example shows that the development of an optimized scheduling strategy for parallel applications also has to consider memory requirements. Furthermore, an efficient load balancing mechanism is needed to guaranty efficient resource utilization.

2 Efficient Scheduling for Model Based Optimization

Our parallel performance analysis with TraceR can steer the development of parallel R programs aimed at overcoming inefficient resource utilizations we are currently facing. Our future goal is to realize an optimized scheduling strategy for parallel machine learning applications used within model- based optimization (MBO). The MBO approach, which is in general a black-box optimization method, can be used to find an optimal machine learning algorithm configuration for a given dataset in an efficient way. In a wider sense it can also be applied to find the best algorithm with its best configuration. To do so, an enormous amount of different parameter sets of each algorithm needs to be evaluated. While, due to the model-based approach, this does not comprise every possible combination of those, this process still includes computation-intensive evaluation processes that need to be executed in parallel to reduce runtime. However, running these in a naive parallel way leads to inefficient resource utilization and thus to unacceptably long runtimes. The reason for this is the high variance in runtime and memory usage that is induced by different parameter configurations. In cooperation with Jakob Richter, we therefore plan to develop a resource-aware model based optimization framework, which is shown in Figure 2. All purple boxes represent possible settings for our framework and therefore possible experiments to improve the efficiency of our method. The yellow boxes include static information needed for optimization steps. And the blue boxes are illustrating the 4 main steps of our optimization cycle.

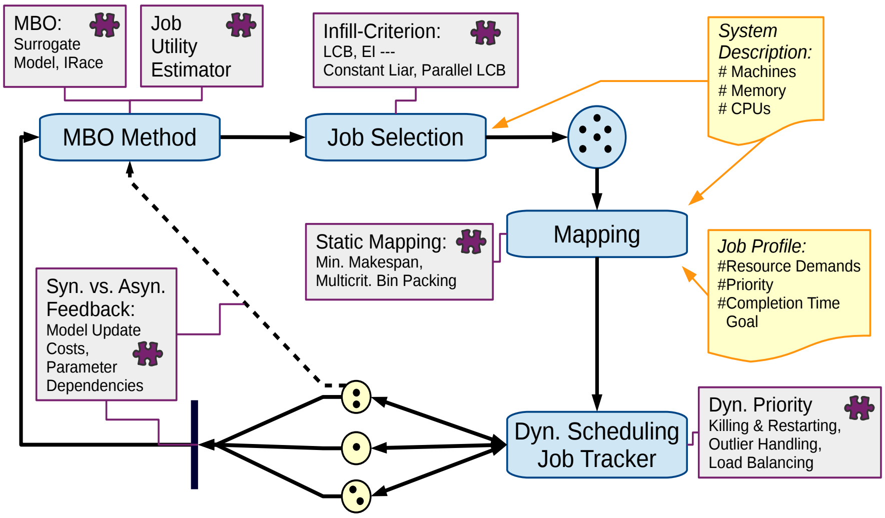

MBO: Surrogate Model, IRace
Job Utility Estimator
Infill-Criterion: LCB, EI --- Constant Liar, Parallel LCB
System Description: # Machines # Memory # CPUs
MBO Method
Job Selection
Static Mapping: Min. Makespan, Multicrit. Bin Packing
Mapping
Job Profile: #Resource Demands #Priority #Completion Time Goal
Syn. vs. Asyn. Feedback: Model Update Costs, Parameter Dependencies
Dyn. Scheduling Job Tracker
Dyn. Priority Killing & Restarting, Outlier Handling, Load Balancing

Figure 2: Resource-Aware Model Based Optimization Framework

The first step is the *MBO Method*. Here, a strategy, like the surrogate regression model, is applied to estimate the quality of machine learning algorithms that we want to compare to build a model for the given data. Each configuration or so called job has different resource demands that are dependent on the algorithm itself and its parameters. For resource- aware scheduling it is necessary to know the resource demands of each job before execution. Such job profiles are produced though a *Job Utility Estimator* on the basis of previous runs. Here our traceR tool can be applied to gather profiling information during runtime. As a next main step the *Job Selection* chooses a group of jobs to be executed. This selection depends on the priority of a specific algorithm and parameter setting, its job profile and on the system description (e.g. number of CPUs). To determine the priority, a so called *Infill-Criterion* is applied, e.g. the LCB (lower confident bound). The third main step is the *Mapping* that allocates the selected jobs to machines and cores. Therefore it can apply different algorithms, like bin packing, to guaranty an efficient resource utilization. Here, the system description and job profiles serve as inputs. As a last main step jobs are executed by a *Dynamic Scheduling* method and are being monitored by the *Job Tracker*. Since job profiles are only estimated on the basis of previous runs an underestimation, of e.g. runtime values, is possible. In such a case, a job needs to be rescheduled or stopped to guarantee efficient resource utilization. After a job has finished its computation, there are two possibilities to update the model. One way is the synchronous *Feedback*, where the results of all jobs from one iteration are gathered before the model is updated. The other way is to update the model each time with just the results of the job that has finished its computation. Whether asynchronous or synchronous feedback is chosen depends also on the costs of a model update.

References

[1] TraceR: A Profiling Tool for R. TU Dortmund University. 2015. URL https://github.com/allr/traceR-installer

[2] Kotthaus, H., Korb, I., Engel, M., Marwedel, P., Dynamic Page Sharing Optimization for the R Language. Proceedings of the 10th Symposium on Dynamic Languages (DLS'14). Portland, USA. pp. 79-90. 2014.

[3] Kotthaus, H., Korb, I., Marwedel, P., Performance Analysis for Parallel R Programs: Towards Efficient Resource Utilization. Abstract Booklet of the International R User Conference (UseR!). Aalborg, Denmark, pp. 66. 2015.

[4] Kotthaus, H., Korb, I., Marwedel, P., Distributed Performance Analysis for R. R Implementation, Optimization and Tooling Workshop (RIOT). Prag, Czech, 2015.

[5] Kotthaus, H., Korb, I., Marwedel, P., Performance Analysis for Parallel R Programs: Towards Efficient Resource Utilization. Technical Report 01/2015, Department of Computer Science 12, TU Dortmund University, July 2015, SFB876 Project A3

Virus Detection on Embedded Systems and Approximate Communication

Olaf Neugebauer
Computer Science 12
TU Dortmund University
olaf.neugebauer@tu-dortmund.de

Plasmon-based Virus Detection on Heterogeneous Embedded Systems

Future application scenarios of embedded systems, like mobile medical diagnosis, face significant challenges to provide sufficient performance while operating on a constrained energy budget. For example, complex biological virus detection in combination with the PAMONO sensor, usually, due to its demands on calculation power, is executed on PCs or laptops. To enable a mobile battery-driven virus detection, embedded systems must be used. In the future, embedded systems are expected to provide significant amounts of computing power to process large data volumes. But are todays embedded systems capable of performing such tasks?

Modern embedded MPSoC platforms use heterogeneous CPU and GPU cores providing a large number of optimization parameters. This allows to find useful trade-offs between energy consumption and performance for a given application. In [4], we describe how the complex data processing required for PAMONO can efficiently be implemented on a state-of-the-art heterogeneous MPSoC platform. An additional optimization dimension explored is the achieved quality of service. Reducing the virus detection accuracy enables additional optimizations not achievable by modifying hardware or software parameters alone.

Instead of relying on often inaccurate simulation models, our design space exploration employs a hardware-in-the-loop approach to evaluate the performance and energy consumption on the embedded target platform. Trade-offs between performance, energy and accuracy are controlled by a genetic algorithm running on a PC control system which deploys the evaluation tasks to a number of connected embedded boards.

To explore the solution space, we extended an existing GPU design space exploration [1] to consider platform specific features and real hardware. Thus, the extended version supports measurement of GPU, CPU and RAM energy consumptions, run time executed on real systems. Further, the internals of the existing approach were improved as well. This approach was used to investigate the trade-offs between run time, energy consumption and accuracy of the virus detection application. To evaluate the abilities, we analyzed three different evaluation scenarios: A design space exploration of the software parameters only (cf. Figure 1(a)), a design space exploration of the hardware configuration only and a combined hardware/software design space exploration (cf. Figure 1(b)).

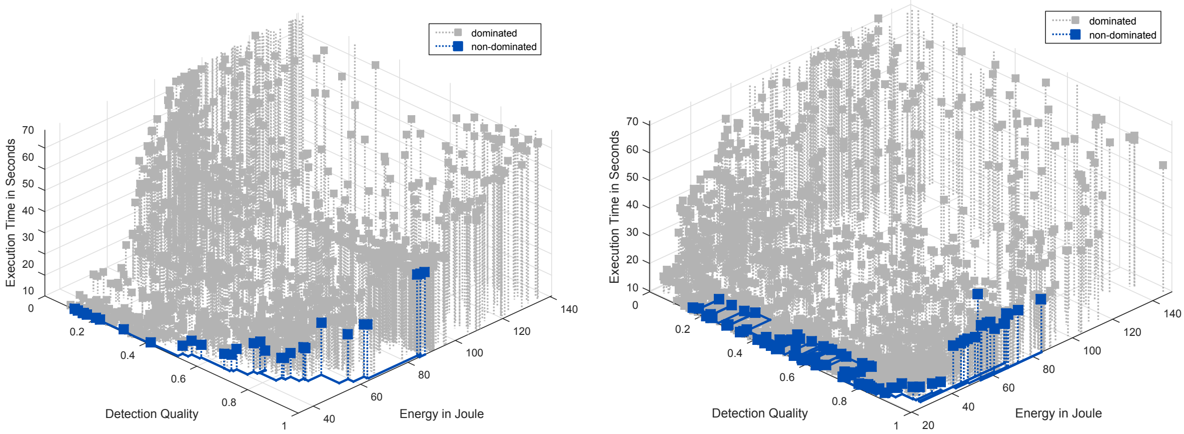
A surprising result is that the frame rate could be increased from 7.5 fps to 30.7 fps (speedup of 4.1) without losing accuracy in the detection quality. This enables the embedded system to process images with the best possible detection quality live from camera. At the same time, the energy consumption could also decreased by 84%, from 370 Joule down to 57.5 Joule. By taking a reduction of accuracy into account, energy consumption could be reduced by 93% with a reasonable detection quality of 96.9% and a frame rate of 60.8 fps (speedup of 8.1). These positive results have exceeded our expectations and lay the foundation for a mobile usage of this virus detection.

In times were increasing processor's clock frequency is not an option due to temperature and power issues, new solutions need to be found. The results of the conducted experiments show that accepting inaccurate solutions opens a new optimization dimension which must be further investigated to increase performance even more in the future.

Approximate Communication

Future embedded systems will be composed of dozens of different computation units connected through a common communication infrastructure. This infrastructure is known to be *the* performance bottleneck of parallel programs. Thus, efficient utilization of the underlying hardware and implementations in combination with application knowledge is mandatory. We investigate how to apply ideas from approximate computation in order to remove or at least widen this bottleneck. Due to typical memory restrictions of embedded systems, providing different approximate implementations utilizing alternative communication patterns is typically not possible.

We propose *approximate communication* to improve the communication efficiency of parallel applications utilizing shared channels. The basic idea behind approximate communication is to provide each communication channel with default values for communicated parameters provided either by a developer or by static analysis. Thus, program parts waiting for data to be received can be executed even if not all data is readily available. Careful



(a) Experiment Exp2$_{sw}$

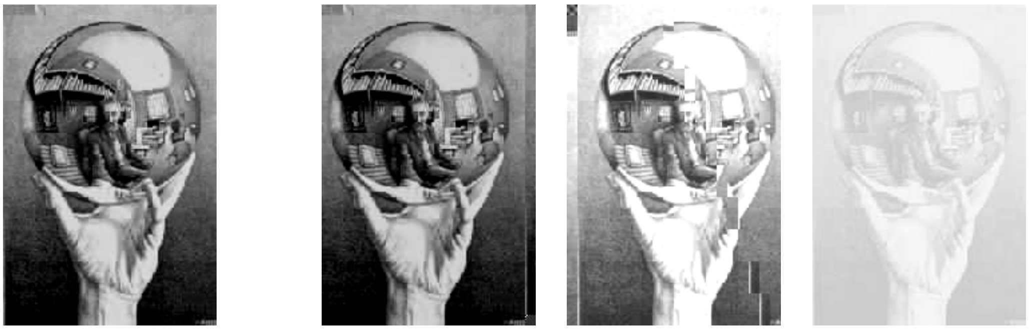
(b) Experiment Exp3$_{hw\&sw}$

Figure 1: Pareto fronts and dominated points for two of the three experiments. The non-dominated points are plotted on top of the dominated points for better visualization. For execution time and energy consumption lower values are better and for detection quality higher values are better.

analysis and selection of parameters and values ensure that *sufficiently good* results are obtained.

We expect this functionality to be especially useful in situations where, e.g. meeting a soft deadlines or saving energy is more important than obtaining a precise result. By only changing the communication, no modification to the original application is necessary. In addition, the behavior of the approximate communication can be adapted during runtime to deal with environmental changes. For example, under light communication load, all data can be communicated and processed, resulting in *precise* results. In case of high communication load, approximate communication will transfer less data by using default values or a previously saved parameter. Assuming careful selection of the omittable data elements, a *sufficiently good* result will then be obtained. Approximate communication will also be useful in scenarios where the producer of data is delayed, e.g., when its execution is preempted by higher priority tasks.

In first preliminary experiments, conducted on real hardware, we observed potential benefits of approximate communication. Here, we introduced random artificially delay into the communication of parallel applications using our PICO framework [3]. A timeout during the receiving of data was used to proceed with approximate data. Run time and energy consumption could be improved compared to the execution without a timeout. However, we noticed that the influence of this approximation technique is highly application-dependent. Figure 2 shows results for the JPEG encoding experiment, which picture counts as acceptable highly depends of the environment this application is used. Thus, application knowledge extracted from the source code or annotated by experts is



(a) (b) (c) (d)

Figure 2: Figure a) shows the optimal result and Figures b)-c) show impact of approximate communication on JPEG encoding.

required to achieve the desired performance improvements with respect to an acceptable result's quality.

Current work focuses on extending our parallelization and communication optimization framework PICO [2] towards approximate computation like approximate communication. Multiple quality metrics are investigated and implemented, leading into an automatic evaluation framework for various approximation techniques.

References

[1] Pascal Libuschewski, Peter Marwedel, Dominic Siedhoff, and Heinrich Müller. Multi-objective energy-aware GPGPU design space exploration for medical or industrial applications. In *Proceedings of Signal-Image Technology and Internet-Based Systems, SITIS*, 2014.

[2] Olaf Neugebauer, Michael Engel, and Peter Marwedel. Multi-objective aware communication optimization for resource-restricted embedded systems. In *Proceedings of Architecture of Computing Systems. Proceedings, ARCS*, 2015.

[3] Olaf Neugebauer, Michael Engel, and Peter Marwedel. A Parallelization Approach for Resource Restricted Embedded Heterogeneous MPSoCs Inspired by OpenMP. *Journal of Systems and Software - Special Issue on Software Engineering for Parallel Systems*, in press - available Fall 2015.

[4] Olaf Neugebauer, Pascal Libuschewski, Michael Engel, Heinrich Müller, and Peter Marwedel. Plasmon-based virus detection on heterogeneous embedded systems. In *Proceesings of Workshop on Software & Compilers for Embedded Systems, SCOPES*, 2015.

Faster Model Based Optimization through Resource Aware Scheduling

Jakob Richter
Faculty of Statistics
TU Dortmund University
jakob.richter@tu-dortmund.de

In my ongoing work in cooperation with Helena Kotthaus we develop a scheduled model based optimization (MBO) approach that leads to a more efficient utilization of parallel computer architectures. In the field of hyperparameter optimization efficient black-box optimization is necessary to find a well-performing algorithm for a specific problem. Recently, with the MBO approach an efficient framework for black-box optimization was the topic of many publications [2, 4, 6, 7]. It fits a regression model on the set of known evaluations to prevent evaluations of unpromising configurations and thus saves resources. However, its sequential approach disables it from using parallel resources efficiently until now, as only the most promising configuration is evaluated on the black-box per iteration. To solve that problem we are working on an extension of MBO which allows for multiple promising configurations to be evaluated on multiple processors at the same time. Problems to solve are: What are the most promising configurations given the regression model? How to schedule the configurations given different priorities and noisy estimations of runtime and memory demands?

1 The synchronous approach

The synchronous approach is closely related to the classical MBO approach. Whereas in the classical approach only the one most promising candidate configuration is evaluated on the black-box per iteration, in the parallel MBO approach a set of k candidate configurations is evaluated in parallel. Seen as promising are those configurations that are predicted to be a global minimum and/or those that lead to a decrease of model uncertainty. In equation 1 the definition of the *Lower Confidence Bound* (LCB) [5] is given, which is one criterion for the promisingness.

$$LCB(\boldsymbol{x}) := \hat{y}(\boldsymbol{x}) - \lambda \cdot \hat{s}^2(\boldsymbol{x}), \tag{1}$$

whereas $\hat{y}(\boldsymbol{x})$ denotes the predicted outcome of the black-box for the candidate configuration $\boldsymbol{x}$ and $\hat{s}^2(\boldsymbol{x})$ denotes uncertainty of the regression model. Proposals on how to choose k candidates simultaneously have been made in [1]. Usually k is chosen equivalent to the number of available cores. In this way k candidates are evaluated in every iteration. When all evaluations are finished, the model will be updated with the results and the next k candidates will be proposed according to the updated model. However, this can lead to high amounts of unused CPU-time as our research has shown.

Depending on the black-box runtimes can vary with the values of configuration. This is made obvious if we highlight one of our major use cases of black-box optimization: Hyperparameter optimization for machine learning methods. In this scenario there are different cases in which runtime is determined by the size of the data set, which is always the same for one tuning problem, and the training process of a chosen machine learning method.

One simple case is hyperparameter tuning of a Support Vector Machine (SVM) on a fairly large data set. It is a property of SVMs that training runtime is highly dependent on cost parameter C and e.g. on parameter γ, when using the RBF-Kernel function. That runtime can hugely differ becomes even more obvious in a more complex case where hyperparameter tuning includes a tuning over different machine learning algorithms, which all have different training runtimes. Given the fact that the runtime highly depends on the hyperparameter configuration, it becomes clear that when running k black-box instances on k CPUs some instances will finish early and the continuation of the MBO iteration has to wait. This leaves frames of CPU-time open which we can use to evaluate instances which we believe will have a short runtime. To estimate the runtime we apply the same method that is used to estimate the outcome of the black-box. This enables us to have a larger amount $l > k$ instances that we can schedule on k CPUs.

Figure 1 shows the general framework for our approach, including a selection of possible settings. Starting from the *Optimization Method* we are technically free to use MBO or other optimization methods as e.g. IRace. The *Job Utility Estimator* provides estimated runtimes and memory demands for every candidate configuration. In connection with the optimization method the *Infill-Criterion* generates a set of candidates including a prioritization. In the *Mapping*-Stage, together with estimated resources, candidates are scheduled based on available resources and their priority. The *Job Tracker* takes care of the distribution and execution of the instances as well as killing possible outliers. Finally in the synchronous approach the iteration starts over when all instances are finished, and the optimizer can use the results to generate new instances.

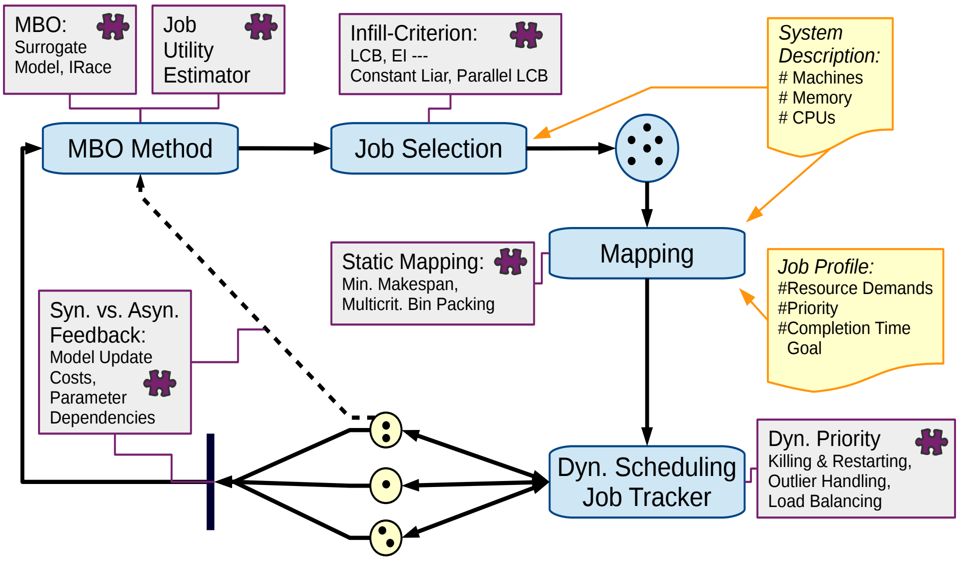


Figure 1: Outline of the Resource Aware Model Based Optimization (RAMBO) Framework, highlighting the key components of interest in the research.

2 The asynchronus approach

The asynchronous approach does not differ from the classical MBO approach heavily either. Its differences lay deeper in the technical sense which makes it harder to implement. However, we think it is a promising way as it has the potential to minimize the unused CPU resources to theoretically zero.

In Figure 1 the difference to the synchronous approach is shown by the dashed diagonal line from a single black-box evaluation result in the bottom to the *Optimization Method* Box. Whereas in the synchronous approach the algorithm always waits until all instances are finished, in the asynchronous approach it continues as soon as one instance finishes. The result will be used to update the regression model. Still running instances will be taken into account by assuming that their outcome is exactly as predicted – this is called the *constant liar* method. In contrast to the synchronous approach this method can handle unreliable estimations of utilization better and it can incorporate new knowledge about the black box faster.

3 Multi Fidelity Model Based Optimization

Further research is done in advancing on the results of my master thesis. Often the black-box to optimize is expensive to evaluate and there are further black-boxes which approximate the target black-box and are cheaper. In the machine learning scenario the black-box optimization is the hyperparameter tuning for a specific dataset and machine learning algorithm. Cheaper black-boxes can be created by subsetting the original dataset.

Initial ideas on how to conduct such multi-fidelity optimization using Kriging as a regression method are covered in [3]. In my work I extended this idea in such ways that multi-fidelity optimization is not dependent on a specific regression method anymore. This enables tuning over mixed hyperparameter spaces (e.g. combined algorithm selection and tuning) by using e.g. random forest regression.

The potential of that work to further improve the efficiency of the previously introduced parallel model based optimization is very high. Possible gaps of unused CPU-time could be easily filled with instances of cheap black-box evaluations, which will also potentially increase accuracy of the regression.

References

[1] B. Bischl et al. "MOI-MBO: Multiobjective Infill for Parallel Model-Based Optimization". In: *Learning and Intelligent Optimization*. Ed. by Panos M. Pardalos et al. Lecture Notes in Computer Science. Springer, 2014, pp. 173–186. ISBN: 978-3-319-09583-7. DOI: 10.1007/978-3-319-09584-4_17. URL: http://www.caopt.com/LION8/Papers/Bischl.pdf.

[2] Alexander IJ Forrester and Andy J Keane. "Recent advances in surrogate-based optimization". In: *Progress in Aerospace Sciences* 45.1 (2009), pp. 50–79.

[3] Deng Huang et al. "Sequential kriging optimization using multiple-fidelity evaluations". In: *Structural and Multidisciplinary Optimization* 32.5 (2006), pp. 369–382.

[4] Frank Hutter, Holger H Hoos, and Kevin Leyton-Brown. "Sequential model-based optimization for general algorithm configuration". In: *Learning and Intelligent Optimization*. Springer, 2011, pp. 507–523.

[5] Donald R Jones. "A taxonomy of global optimization methods based on response surfaces". In: *Journal of global optimization* 21.4 (2001), pp. 345–383.

[6] Donald R Jones, Matthias Schonlau, and William J Welch. "Efficient global optimization of expensive black-box functions". In: *Journal of Global optimization* 13.4 (1998), pp. 455–492.

[7] Chris Thornton et al. "Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms". In: *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2013, pp. 847–855.

Subproject A4
Resource efficient and distributed platforms for integrative data analysis

Christian Wietfeld   Michael ten Hompel   Olaf Spinczyk

Reducing the Resource Consumption of a Fault-Tolerant Operating System by Static Analysis

Christoph Borchert
Department of Computer Science 12
TU Dortmund
christoph.borchert@tu-dortmund.de

This reports describes current advances in the AspectC++ compiler with respect to static whole-program analysis. Such analyses are used to optimize the resource consumption of operating-system fault-tolerance implementations. In a case study on the L4/Fiasco.OC microkernel operating system, these analyses lead to a significant reduction of CPU utilization and a slight increase in fault tolerance.

1 Introduction

Fault tolerance of operating systems can be implemented by several software mechanisms, such as *Generic Object Protection (GOP)* [2] and *Virtual-Function Pointer (VPtr) Protection* [1]. Both approaches add redundancy to kernel data structures (*objects*) and carry out runtime checks to detect and correct errors in such data structures. For instance, the GOP mechanism introduces runtime checks before every member-function call, before returning to a member function, and before every data-member access. Although effective, such an approach leads to heavy computational load. This report describes an optimization based on static whole-program analysis to reduce the demand on the CPU resource by identifying source-code locations suitable for omitting runtime checks while still retaining the same degree of fault tolerance (Section 2). Section 3 quantifies the benefits of the proposed whole-program analysis by taking the example of the *L4/Fiasco.OC* microkernel [4].

2 Whole-Program Analysis and Optimization

The following optimization for the aforementioned GOP mechanism is based on the insight that, in general, it is not beneficial to repeatedly check an object within a short time frame. Thus, I address two efficiency problems:

1. **Short-running functions:** After returning from a function call, the active object gets checked for errors that accumulated during the function call. For functions that return quickly, for example, inline getter or setter functions, it is better to leave the object unprotected for few instructions.

2. **Call sequences on the same object:** The following excerpt from the L4/Fiasco.OC source code shows two consecutive invocations on the same object:

   ```
   Sched_context *sc = sched_context();
   /* some lines of code omitted here ... */
   sc->set(p);
   sc->replenish();
   ```

   Instead of checking the object again on the second call (sc->replenish()), it would be more efficient to verify the code only once before the entire call sequence.

Both efficiency problems can be solved by leaving out check operations on recently used objects. I propose static source-code analysis to automatically identify respective source-code locations. Therefore, I extended the AspectC++ [5] compiler by static control-flow and data-flow analyses. Internally, the AspectC++ compiler builds a *project model* that aggregates all results from the static analyses and thereby abstracts from the syntax tree. Figure 1 shows an excerpt from the project metamodel. Each depicted class gets instantiated according to the source code being compiled, e.g., an object from the class `Call` corresponds to one function call present in the source code. The callee, an instance of the class `Function`, is linked via the target association. Taking the example of a function call, the base class `Access` holds the following additional results:

- **lid:** A local identifier describing the order of each function call.
- **target_object_lid:** A local object identifier that is assigned to each object used as a call target. When two function calls have an identical target_object_lid, then the same object is used with certainty. Different IDs indicate either different objects or that the static pointer analysis cannot prove their identity.
- **cfg_block_lid:** A local identifier describing the basic block a call belongs to. A basic block is a linear sequence of program statements that are always executed in order.

The AspectC++ compiler currently processes one translation unit at a time, and, thus, the project model only contains partial information on all files *during* compilation. However, the partial project model from each translation unit is serialized into a shared XML file (the *project repository*). Thus, after all translation units have been compiled once, the project repository contains the complete model for *whole-program analysis*. In this report, I use the *XQuery* language to extract the whole-program information from the XML project repository. Therewith, I generate a new header file that describes the source-code locations where the object checking can be optimized out. Using this new header file, I compile each translation unit once more to apply the optimization.

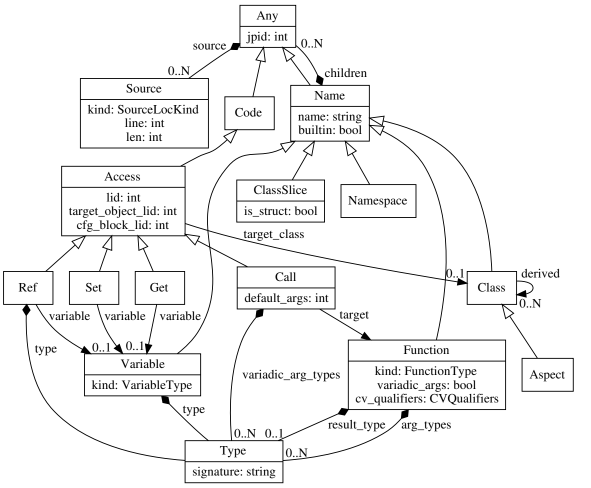


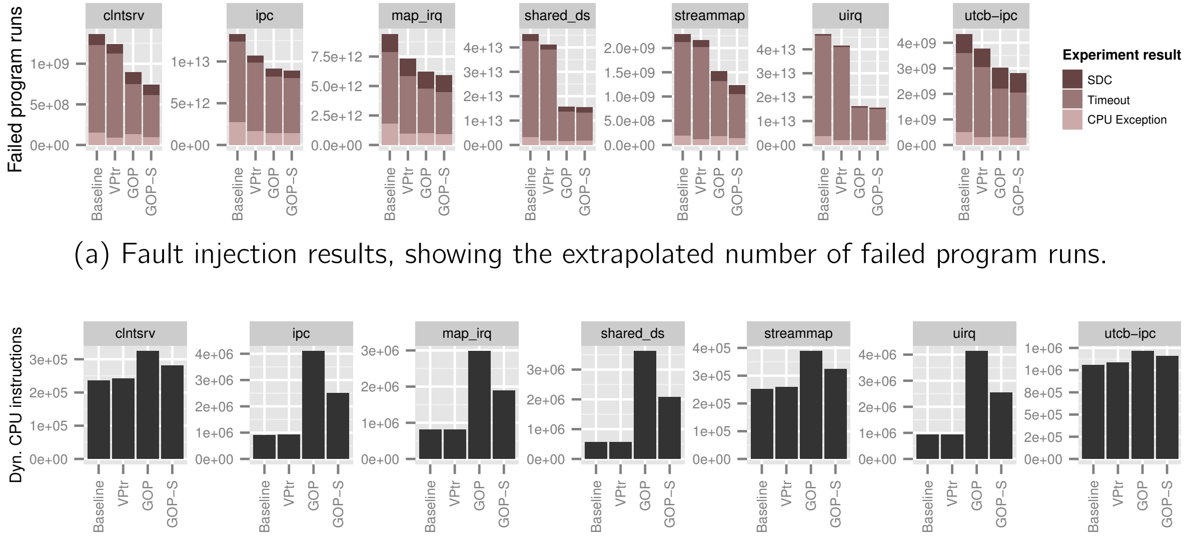
Figure 1: Excerpt from the AspectC++ project metamodel [3]

3 Evaluation

In this section, the static optimization is applied to the L4/Fiasco.OC microkernel in combination with GOP and VPtr protection. Figure 2a shows the susceptibility of four kernel variants, each running seven benchmark programs, to random memory errors: *Baseline* denotes an unprotected kernel, *VPtr* refers to protection by replicated virtual function pointers, *GOP* describes the initial GOP mechanism, and *GOP-S* adds the static whole-program optimization. Likewise, Figure 2b shows the number of dynamically executed CPU instructions at runtime. The static optimization reduces the CPU utilization notably, and even the fault tolerance of the statically optimized variant is slightly better. Because the optimized kernel is so much faster, the probability of being hit by a random memory error is drastically reduced, while keeping the essential checks after long periods of time.

4 Conclusions

This report demonstrates that, by applying static whole-program analysis with the AspectC++ compiler, the resource costs of operating-system fault tolerance can be drastically reduced. Furthermore, optimizations based on that analysis do not necessarily reduce the fault tolerance of the operating system, but can even slightly improve its dependability.



(a) Fault injection results, showing the extrapolated number of failed program runs.

(b) Dynamic CPU instructions executed at runtime.

Figure 2: Fault tolerance and runtime comparision of four microkernel variants [3].

References

[1] Christoph Borchert, Horst Schirmeier, and Olaf Spinczyk. Protecting the dynamic dispatch in C++ by dependability aspects. In *1st GI W'shop on SW-Based Methods for Robust Embedded Sys. (SOBRES '12)*, Lecture Notes in Informatics, pages 521–535. German Society of Informatics, September 2012.

[2] Christoph Borchert, Horst Schirmeier, and Olaf Spinczyk. Generative software-based memory error detection and correction for operating system data structures. In *43rd IEEE/IFIP Int. Conf. on Dep. Sys. & Netw. (DSN '13)*. IEEE, June 2013.

[3] Christoph Borchert and Olaf Spinczyk. Hardening an L4 microkernel against soft errors by aspect-oriented programming and whole-program analysis. In *8th W'shop on Progr. Lang. and OSes (PLOS '15)*, pages 1–7, New York, NY, USA, October 2015. ACM.

[4] Adam Lackorzynski and Alexander Warg. Taming subsystems: Capabilities as universal resource access control in L4. In *Proceedings of the Second Workshop on Isolation and Integration in Embedded Systems*, IIES '09, pages 25–30, New York, NY, USA, 2009. ACM.

[5] Olaf Spinczyk and Daniel Lohmann. The design and implementation of AspectC++. *Knowledge-Based Systems, Special Issue on Techniques to Produce Intelligent Secure Software*, 20(7):636–651, 2007.

Towards the Awareness of Energy Contexts in Embedded Operating Systems

Markus Buschhoff
Department of Computer Science 12
Technische Universität Dortmund
markus.buschhoff@tu-dortmund.de

To deploy deeply embedded systems that are adaptive regarding their energy consumption, it is necessary to achieve awareness of the energetic situation ("context") a deployed system is in. Within this report we will put a spotlight on the simulation of some context aspects, namely the long-time harvest of energy using solar cells, and the simulation and scheduling of system tasks according to the state of an energy buffer to increase the buffer's efficiency by exploiting its charging and stress asymmetries.

1 Simulation of tasks and energy buffers during long time outdoor solar harvest

The interaction of energy harvesting power supplies with adaptive systems shows a high complexity due to the many of different components influencing the energy harvest and energy consumption.

On the side of the energy supply, there is a high environmental dynamic that heavily influences the outcome of energy. This is moderated by energy buffers, which store energy during phases of high gain to keep up supply through phases of negative energetic balance.

On the consumption side, there are a multiple of tasks in a system that are often able to scale their behavior for the price of decreasing service quality. As an example, take a sensor node that uses an operating system task to measure data and store it into memory. The frequency of this task determines the granularity of measurements and thus the system's service quality. Additionally, a second task might be implemented to transfer the collected data using a radio link. The frequency of this task is determined by different factors: The targeted service quality (actuality of data sent), the amount of memory available, and the available energy for sending.

Within the operating system we want mechanisms to balance the supply and the consumption of energy in a way that optimizes the service quality. The most promising approaches are predictive algorithms used in environments that allow for predictability.

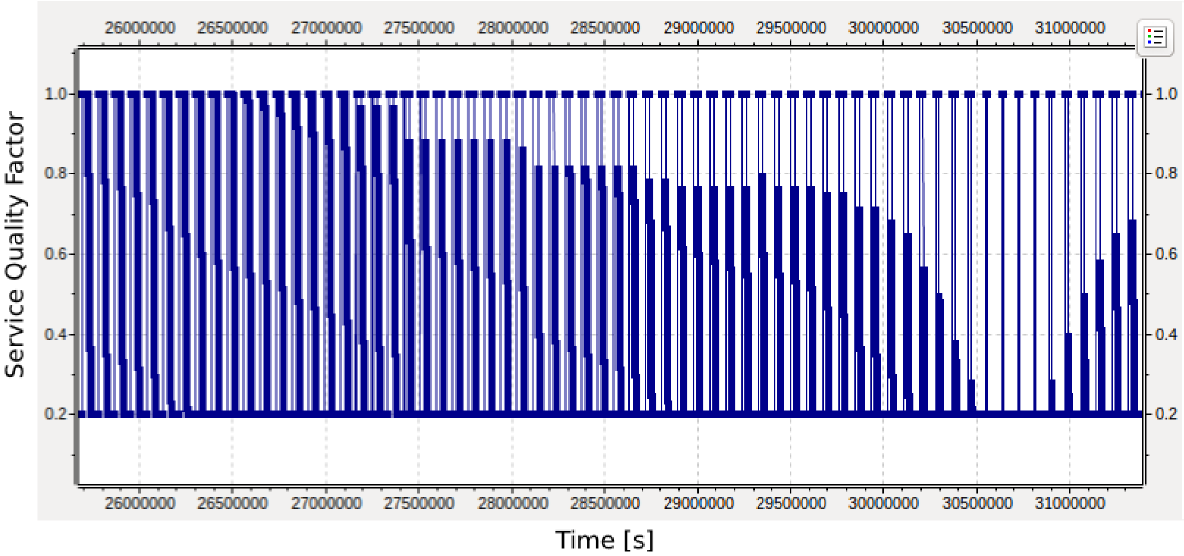


Figure 1: Adaptivity plot by Mobisim over the last quarter of a year with changing sunlight for a solar harvester. The service quality is scaled down to the defined bounds during night times and in winter.

To evaluate the efficiency of such adaptive algorithms in a typical embedded system setup, we enhanced Mobisim [1] to enable the modeling of solar potentials over the year for a given latitude, thus including night and day cycles as well as summer and winter daytime durations. Additionally, Mobisim can now calculate basic solar cell models (linear voltage and current models) to simulate the solar harvest. Together with energy buffers and a task model with scaling service qualities, Mobisim is now able to graph the adaptiveness of energy based task schedulers (Figure 1).

2 Scheduling for the exploit of non-linear effects in energy buffers

There are several energy buffering technologies available to store an excess of harvested energy for later use. Among them are rechargeable batteries and capacitors with extreme electrical characteristics ("supercaps"). All energy buffers have an unideal efficiency, thus loosing energy during recharging or stress. Since chemical processes within the buffer often do not occur immediately after an event, the efficiency of a buffer is also dependent on recovery periods.

This leads to the assumption, that assuming a given model of energy loss and efficiency for a buffer, these changes in efficiency over time can be exploited to create a "buffer-aware" schedule of operating system tasks that is more energy efficient than other (naive?) schedules.

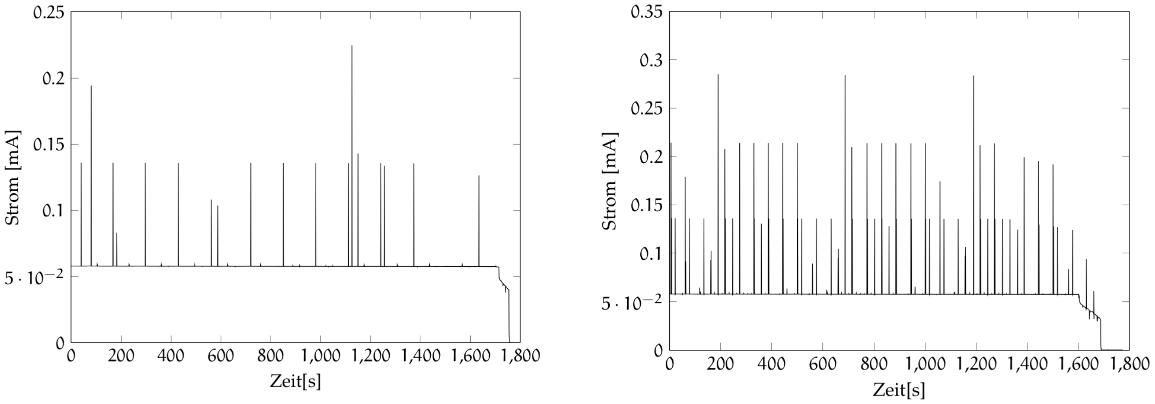
To prove this thesis, a supercap was measured and modeled utilizing the RC Ladder Model [3] - a modeling approach that describes the single capacitor in its behavior as series of parallel capacitors, each with different capacitance and series resistance. This model describes quite well, how self-discharge and the internal transshipping of charge spreads the available energy over time.

Based on this model, a simulator was developed to calculate the available energy for a given point in time. This simulator could then be used to drive an evolutionary algorithm to optimize the available energy for a given set of cyclic operating system system tasks.

The tasks where modeled by assuming a cyclic run-to-completion pattern: A task is started and runs to its end, requiring a constant amount of energy. After completion, the operating system may pause the task within given time boundaries before restarting it. Generally, it is considered desirable to run the task as often as possible within these boundaries. But, the evolutionary algorithm may choose to reduce the task frequency to optimize towards a given energy target.

The algorithmic results shown in the master thesis of Michael Hesse [2] claim, that under the constraint of using all available energy during a given time of exactly 30 minutes, the task frequency can diverge from 17 runs during that time (Figure 2a) to a maximum of 33 runs with an optimized schedule (Figure 2b). Experiments with real hardware show, that in both cases the used supercap reaches the low threshold voltage close to 30 minutes (>28 minutes), just as requested, using these very different schedules for the same tasks, thus proving the assumption of practically exploitable, non-linear effects to be correct.

Within the next year we plan to intensify research on this topic by examining further energy buffers and by evaluating the accuracy and the possible energy gain of our approach. Eventually, we want to find simple rules that apply for a given energy buffer to



(a) Unoptimized tasks. Task run 17 times. [2]

(b) Optimized scenario, up to 33 runs. [2]

Figure 2: Exploiting the recovery and self-discharge cycles of a supercap.

deploy online scheduling decisions as an alternative to the static evolutionary algorithm approach.

References

[1] Markus Buschhoff, Jochen Streicher, Björn Dusza, Christian Wietfeld, and Olaf Spinczyk. MobiSIM: A simulation library for resource prediction of smartphones and wireless sensor networks. In *Proceedings of the 46th Annual Simulation Symposium*, ANSS '13, Society for Computer Simulation International, 2013. Society for Computer Simulation International.

[2] Michael Hesse. Berücksichtigung nicht-linearer energiequelleneffekte im eingebetteten betriebssystem kratos. Diplomarbeit, Universität Dortmund, Germany, September 2015.

[3] L. Zubieta and Richard Bonert. Characterization of double-layer capacitors for power electronics applications. *Industry Applications, IEEE Transactions on*, 36(1):199–205, Jan 2000.

Client-Based Connectivity Estimation for Resource-Efficient Communications in LTE

Robert Falkenberg
Lehrstuhl für Kommunikationsnetze
Technische Universität Dortmund
robert.falkenberg@tu-dortmund.de

This report gives a brief overview on the connectivity estimation of LTE mobile network links on the basis of available information and data at the mobile-equipment site. That rating is required to choose the most efficient link in case the device provides multiple communication paths through different technologies at the same time. Especially in vehicular applications, which are affected by continuously changing radio conditions, resource-efficient communication can improve the transmission of telemetry, emergency calls, and other data for further assistance systems.

1 Introduction

The multiplicity of simultaneously available communication technologies, e.g., LTE and WiFi, challenges the mobile equipment to choose the *best* link for their data transfers. What is meant by *best* link depends on device type and working context: In an energy harvesting environment, this is most often the path with the lowest energy costs while an emergency-reporting system typically heads for a low latency. The choice remains with the particular device arising the need for proper indicators to allow that decision.

Heading for that goal, this report focuses on the utilization of accessible measurement values within mobile equipment in LTE networks. From these values we aim to derive an estimated transmission time when using that link under the current radio conditions.

At its core, this research consists of the identification of suitable measurement values, the evaluation of the reliability, and the synthesis of an applicable estimation model.

1.1 Performance of a Wireless Communication Link

In order to estimate the connectivity of an LTE link in terms of data rate or transmission time, it is important to know about the influencing factors. Figure 1 shows a generic

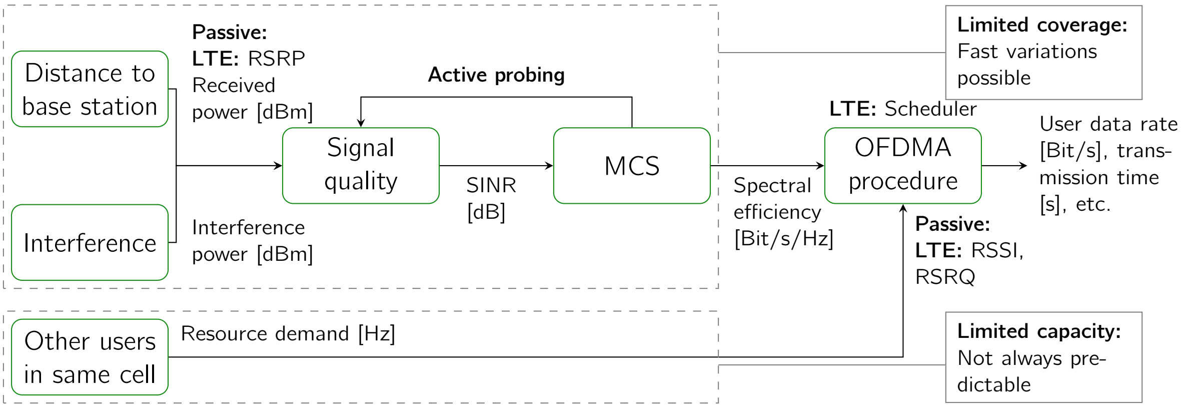


Figure 1: Performance of wireless communication systems dependent on signal quality and network load.

overview of the link performance in a wireless communication system applied to the implementation in LTE. The main three influences are the distance to the base station, the intensity of interference and the amount of other active users attached to the same cell.

The first two factors limit the coverage of the cell due to path loss and the raising disturbances by neighboring cells. This leads to a certain signal quality in terms of Signal to Interference Ratio (SINR) from which the network chooses a suitable Modulation and Coding Scheme (MCS). The higher the SINR, the more bits can be transferred within the same bandwidth and time interval. This ratio of *bits per second and hertz* is called spectral efficiency.

Since a cellular communication system is a multiple access system, the radio resources need to be shared among all active participants attached to the cell which is the third limiting factor. In LTE this is applied with Orthogonal Frequency Division Multiple Access (OFDMA), which is well known for its robustness, flexibility, and efficiency. The base station – in LTE called eNodeB – allocates the available radio resources to the participants according to their MCS resulting in an individual data rate of each attendant.

1.2 Available Indicators from Application Layer

To evaluate the channel quality in an LTE User Equipment (UE) several channel indicators are defined by 3rd Generation Partnership Project (3GPP) [1]. A subset of these indicators is accessible at UE's application layer:

Reference Signal Received Power (RSRP**) in** dBm : This value describes the average received power of the resource elements carrying cell-specific reference signals over the utilized bandwidth. As the eNodeB sends these signals with a constant power, this provides information to the UE about the position within the cell.

Receive Signal Strength Indicator (RSSI**) in** dBm : In contrast to RSRP, this value

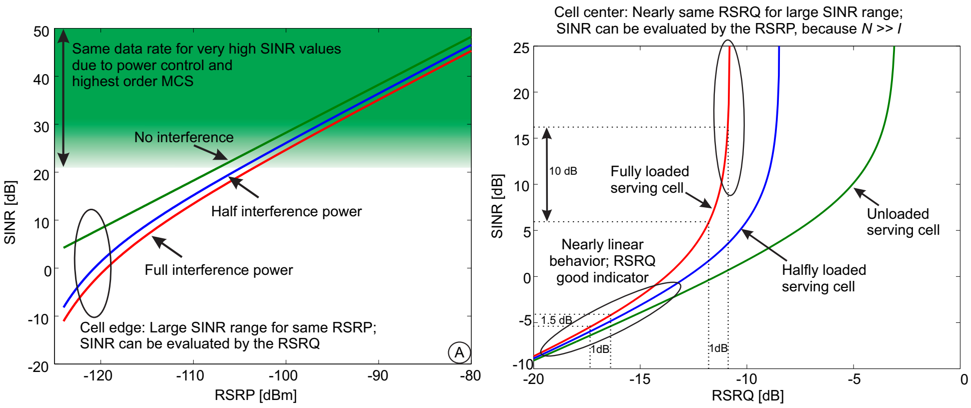


Figure 2: Relationship between RSRP (left), RSRQ (right) and SINR in an LTE cell under different levels of interference and load.

reflects the average received power S over all resource elements within the used bandwidth including noise N and interference I:

$$\mathrm{RSSI} = S + N + I. \tag{1}$$

Reference Signal Received Quality (RSRQ**)** : Subtracting RSRP and RSSI together with a normalization to the number of occupied Physical Resource Blocks (PRBs) gives a channel qualitiy indicator which is nearly independent from path loss:

$$\mathrm{RSRQ} = 10\log_{10}\left(N_{\mathrm{PRB}}\right) + \mathrm{RSRP} - \mathrm{RSSI}. \tag{2}$$

1.3 Relationship of RSRP, RSRQ and SINR

The achievable data rate derives from the MCS which depends on the Signal to Interference Ratio (SINR) at the receiver. That ratio can be put in correlation with the value RSRP and RSRQ in the following manner:

$$\mathrm{SINR} = \frac{\mathrm{RSRP}}{N+I} \qquad \text{and} \qquad \mathrm{SINR} = \frac{1}{\frac{1}{12\cdot\mathrm{RSRQ}} - p} \tag{3}$$

where $p \in [0, 1]$ denotes the current cell's load. The first equation is plotted left in Figure 2 and shows the influence of interference to that relaionship. At high RSRP values the resulting SINR hardly depends on disturbances, leading to a reliable indicator for estimating the achievable data rate. Unfortunately, at low RSRP values the uncertainty grows spanning a wide range of possible SINR for a specific RSRP value.

This usually happens towards the edge of a cell where the neighboring cell's signal power raises. As the UE does not measure the power level of neighboring payload channels, this indicator can not give reliable estimations in these cases.

In contrast, the relationship between RSRQ and SINR has an inverted behavior as shown in Figure 2 on the right-hand side. Low RSRQ values lead to an exact indication of SINR whereas higher RSRQ values highly scatter over a huge SINR range. Hence, at least both, RSRQ and RSRP, must be taken into account for a proper link estimation.

2 Method

In order to investigate the behavior of these values on physical devices in a real environment we performed measurements in a dedicated LTE network as shown in Figure 3. The setup contains two concurring eNodeB having several UEs attached to them which are utilized as Smart Traffic Generators (STG) [3]. A Device Under Test (DUT) was placed in the cell center and at the edge towards the neighboring cell trying to upload 100 kB of data. In each case all permutations of full/low cell load within the own an neighboring cell were tested capturing RSRP, RSRQ, and transmission time.

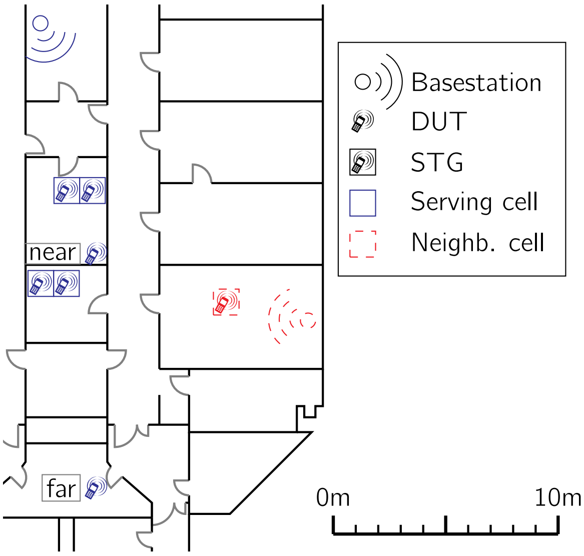


Figure 3: Measurement setup.

3 First Results

Our results of repeated measurements show a reliable outcome under certain restrictions. A high RSRP value always leads to a short transmission time which increases slightly due to resource sharing under full cell load. On the other hand, at low RSRP values the upload takes significantly longer having the RSRQ value as amplificating factor: The smaller that value, the greater the variance in upload time. Having that said, the RSRP reliably identifies connectivity hot spots with excellent radio conditions resulting in a low transmission time together with a low variance. However at high distances and under weaker radio conditions, duration and variance highly depend on the inner and neighbor cell load. In this case interference is the dominating parameter which must be derived from RSRQ. Measurements in a public LTE network showed similar results.

4 Conclusion and Further Research

The current state of the research shows that RSRP is suitable for estimating SINR in good channel conditions. At distances with low signal and high interference, RSRQ has turned out as a suitable indicator. In order to improve the reliability of the estimations in the transient area further indicators will be extracted at physical layer utilizing Open Air Interface [2]. This topic is currently under progress.

References

[1] 3GPP TS 36.214 - Physical Layer Measurements, V 12.1.0, Dec. 2014.

[2] EURECOM Mobile Communications Department. OpenAir5G LAB. `http://openairinterface.eurecom.fr`, November 2015.

[3] D. Kaulbars, F. Schweikowski, and C. Wietfeld. Spatially distributed traffic generation for stress testing the robustness of mission-critical smart grid communication. In *IEEE GLOBECOM 2015 Workshop on SmartGrid Resilience*, San Diego, USA, December 2015. Accepted for presentation.

Resource aware Flow Management for Secure and Reliable Cloud Access in Heterogeneous Network Environment

Maike Kuhnert
Lehrstuhl für Kommunikationsnetze
Technische Universität Dortmund
maike.kuhnert@tu-dortmund.de

Currently, the development of 5G networks considers the increasing demand for mobility aware high capacity networks. In parallel Cloud services and applications become more and more important for daily network usage (e.g., Cloud storage, collaboration platforms). The main bottlenecks and barriers for efficient use are lack of capacity and coverage as well as the significant demand of flexibility of wireless access networks and in addition security concerns during data transmission. Client based flow management relying on network coding can handle the problem of security concerns and enables the combined usage of network access technologies based on their quality indicators (e.g., PER, Throughput) that are estimated at the clients' devices directly.

1 Solution Overview

One of the main challenge in 5G development is the handling of flexible network infrastructures that rises the need of an efficient and dynamic evaluation of wireless access networks in range of a mobile terminal. Due to varying network environments, the received signal strength at the mobile terminal is not a suitable performance criteria anymore for reliable and secure network access. Specific client based quality indicators, like models for throughput estimation under varying conditions or energy models relying on different behavior like CoPoMo [1], are needed to select the best suitable network. Figure 1 depicts the holistic solution for secure and reliable transmission in wireless networks.

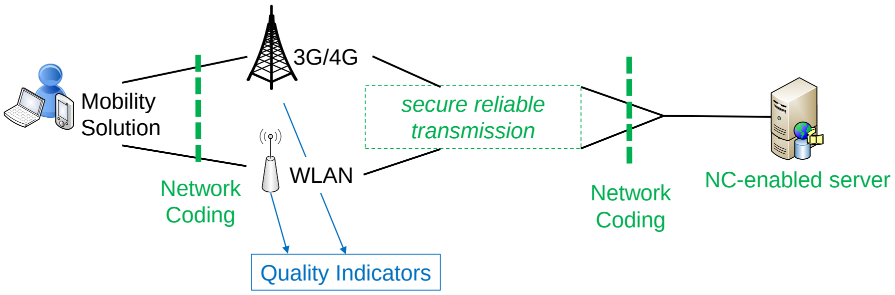


Figure 1: secure and reliable transmission using network coding relying on client based quality indicators of wireless networks

Network evaluation in terms of developing quality indicators and its models, network selection in terms of selection algorithms and flow management are the three main mechanisms that are combined for a holistic client based mobility solution and can be seen as an extension of mobile IP [2], MPTCP [3], client based solution CSHMU [4] and existing network coding implementations [5]. The later will be evaluated in this technical report.

2 Flow Management based on Network Coding

Client based network coding starts with the adjustment of preferences for data transmission. Fast means combined usage of links to achieve a high throughput as possible, whereas secure means high reliability and high security against eavesdropping (see Figure 2. Using only two links does not allow to use a redundant link so for successful transmission of all packets both links have to transfer there coded packets successfully.

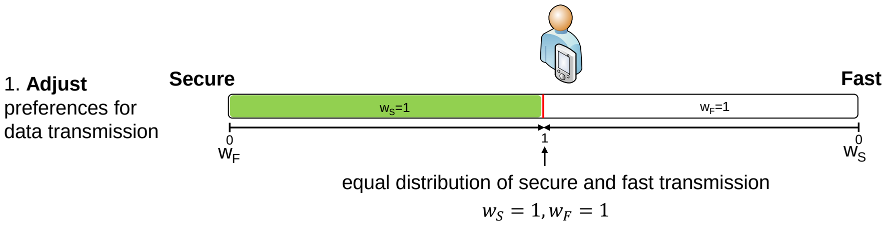


Figure 2: adjustment of preferences for secure and fast transmission

Based on the so defined weights setting for w_F and w_S the distribution P_{L_i} for the specific links L_i is calculated by:

$$P_{L_i} = \max(w_s \cdot C_{PER}, w_F \cdot C_{Thr})\forall i \tag{1}$$

The distribution at the boundary cases C_{PER} and C_{Thr} is derived by the adaption of the link quality relations to 100% packets that should be transmitted using network coding.

The evaluation of quality indicators based on clients knowledge is based on specific models and depends on the performance criteria (e.g. throughput, energy). Table 1 shows an example of packet allocation to links depending on the values of the quality indicator packet error rate (PER) and throughput.

Link L_i	PER [%]	Throughput [MBit/s]	C_{PER}	C_{Thr}
L_1	2	5	58.82	14.28
L_2	4	10	29.41	28.57
L_3	10	20	11.76	57.14

Table 1: quality indicators of available networks in range and distribution of packets to links in boundary cases: $w_S = 1, w_F = 0$ and $w_S = 0, w_F = 1$

Formula 1 enables the allocation of packets to the link. Specific weight settings lead to a distribution of packets over 100%, therefore some packets will be duplicated and enhance the reliability.

3 Proof of Concept

The above section describes the concept behind the application of network coding in the network access selection and management process in a 5G environment. Using network simulator NS–3, it can be proven that the usage of network coding has an benefit against traditional mechanisms (e.g. routing). Figure 3 displays the results of 50 simulation run using random linear network coding (RLNC) with different settings and no coding in a basis mesh network environment using 5 hops. The results in the case of 10% PER show that there is a benefit using network coding in nearly 80% of the simulation runs.

This basic implementation of network coding will now be adapted to user controlled parameterization of network coding to handle the user requirements as well as the dynamic environment reflected by specific quality indicators.

4 Conclusion and Outlook

Network coding is an efficient method to combine traditional processes with efficient coding of data to enhance network reliability and security. In future work the integration of network coding into existing mobility solutions will be one part of the holistic client based mobility solution within 5G networks. In parallel, the derivation of quality indicators and the correctness of their models for estimating the current network situation is of high value and will be further investigated.

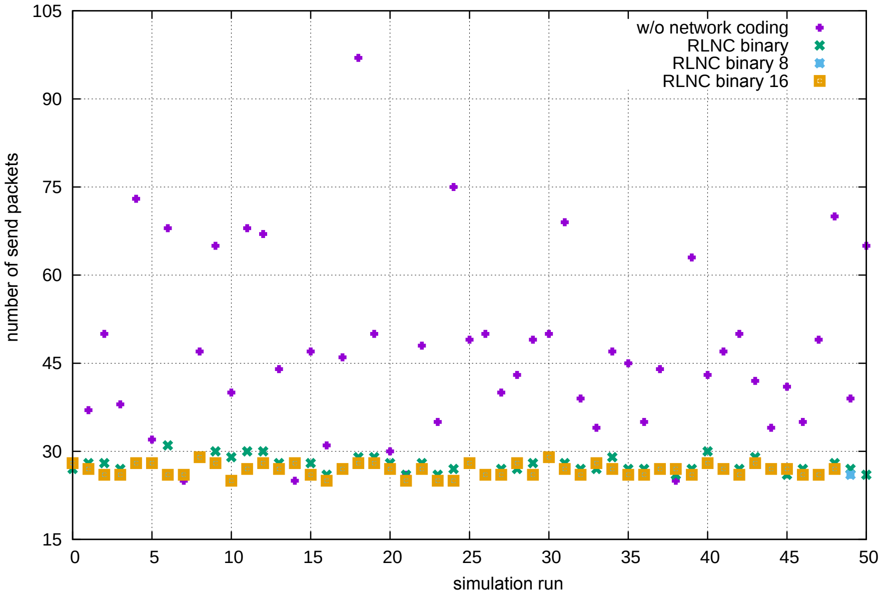


Figure 3: number of packets needed for the transmission of 20 symbols, comparing traditional routing approach with random linear network coding (RLNC)

References

[1] Dusza B., Ide C., Cheng L. and Wietfeld C., CoPoMo: A Context-Aware Power Consumption Model for LTE User Equipment, Transactions on Emerging Telecommunications Technologies (ETT), Wiley. 2013.

[2] Devarapalli, V., Wakikawa, R., Petrescu, A., Thubert, P.: Network Mobility (NEMO) Basic Support Protocol, Request for Comments (Draft Standard) 3963, Internet Engineering Task Force, 2005.

[3] Gonzalez, M., Higashino, T., Okada, M.: Radio access considerations for data offloading with multipath TCP in cellular/WiFi networks. 2013 International Conference on Information Networking (ICOIN) 2013

[4] Tran, T., Kuhnert, M., Wietfeld, C., Energy-efficient Handoff Decision Algorithms for CSH-MU Mobility Solution, 21st IEEE International Conference on Computer Communication Networks (ICCCN), 2nd International Workshop on Context-aware QoS Provisioning and Management for Emerging Networks, Applications and Services, IEEE, Munich, Germany, 2012.

[5] Hundeboll, M.; Pahlevani, P.; Lucani, D.E.; Fitzek, F.H.P., Throughput vs. Delay in Lossy Wireless Mesh Networks with Random Linear Network Coding, 20th European Wireless Conference, 2014.

A Measurement Platform for Photovoltaic Energy Harvesting in Indoor Low Light Environment

Mojtaba Masoudinejad
Lehrstuhl für Förder- und Lagerwesen
Technische Universität Dortmund
mojtaba.masoudinejad@tu-dortmund.de

This report presents a platform specially designed for high-accuracy measurement of indoor artificial lighting and environmental information for the evaluation of indoor photovoltaic energy harvesting when used in environments with ultra-low harvesting potentials with less than 1 W/m^2 irradiance.

1 Introduction

Flexibility and modularity of systems have always played a major role in the field of materials handling [1]. Smart objects with ability to understand and react on their environment are a key solution for achieving this flexibility [2]. The autonomous and intelligent load carrier embodied as *Intelligent Bin* (inBin) is developed [1] according to the IoT device requirements to fulfill the current growing demand [2]. It is a container with the ability of storing data about its contents' attributes, showing data on its display, communicating through a wireless network, in addition to interaction with the operator.

Same as any other IoT entity, energy plays a major role on the feasibility and functionality of the inBin. The inBin should have minimum mobility restrictions and all functions should be fulfilled in an energy neutral manner. It has been shown that among light, vibration and thermal energy, the PhotoVoltaic (PV) generated energy is more promising and efficient within this application [2].

Despite maturity of outdoor PV applications, few researches have been undertaken on the PV cells behavior under indoor artificial lighting [3]. Available analysis are mainly

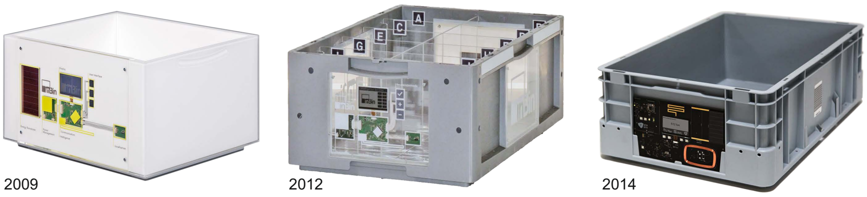


Figure 1: Three generation of the inBin smart object

focused on the applications under Standard Testing Condition (STC) with an arbitrary maximum terrestrial intensity, one sun spectra, 1000 W/m^2 perpendicular to the cell panel in 25 $^{\circ C}$ [4]. In materials handling applications, artificial illumination is mostly the only available lighting. Consequently, an exclusive intensity evaluation and spectrum analysis of this lighting without any outdoor effect is required. A setup to recreate the indoor controlled environment and provide valid PV measurement data is explained here.

2 Measurement Platform

Three types of data are collected within the platform: the lighting and PV cell specification, the environmental condition and set up data of the measurement platform.

Spectrometry is the most reliable way of light measurement. A 2-inch integration sphere from *StellarSphere IC2* with a 180° field of view and wavelength range of 200-1700 nm samples the light and transfers it through a fiber optic connection to a BLACK-Comet spectrometer measuring wavelengths in the range of 200-1100 nm.

In addition to the integration sphere, three PV cells, temperature sensor, infra red sensor and RGB measurement sensors are mounted on a board which enables measurement of all required parameters within a small area with homogeneous light quality. The homogeneity of the light assures a reliable measurement from all sensors.

An interfacing board connects the control PC to the measurement board. In addition to the conversion and transfer of data, this interface selects the active PV cell on the measuring board according to the command from the operator.

To measure four quadrants VI curve characteristics, the Agilent (Keysight) B2902A as a two channel precision Source/Measure Unit (SMU) with resolution of 100 fA and 100 nV is used. It connects to the active PV cell with a 4-wire connection setup as shown in Fig. 2 to eliminate the voltage error caused by the test leads residual resistance.

Four light sources with different types (LED, halogen and florescent) are used in the measurement platform. All used sources are dimmable and their outcoming light is modified through a dimming board based on the operator's command.

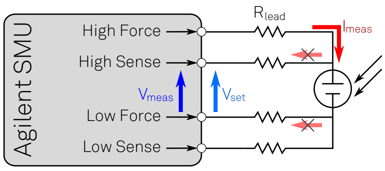


Figure 2: A 4-wire connection setup for measuring IPV cells four quadrants characteristics

The overall signaling structure of the measurement platform, connecting all electrical and electro-optical components is shown in Fig. 3.

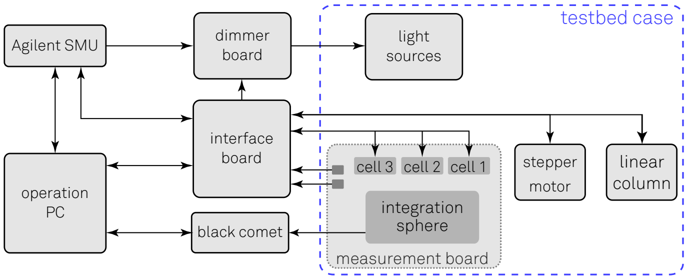


Figure 3: Schematic electrical signaling of all electrical and optical components

All these components and devices are mounted within a black colored hard wooden case with the size of 1700×700×650 *mm*. All interior walls of this case are covered with a rippled black sponge layer to minimize any reflection inside the platform. The overall built platform can be seen in Fig. 4.

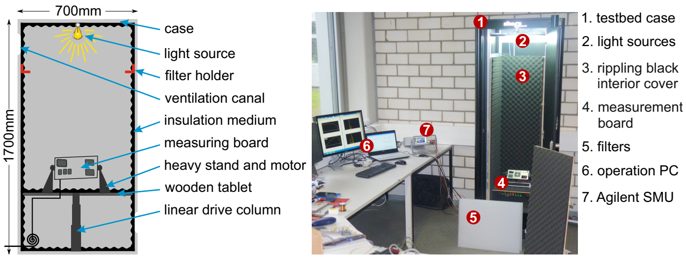


Figure 4: Left: a schematic structure of the platform, Right: the built platform

3 Light Modification

To replicate different indoor light conditions, the measurement platform is able to modify the light intensity with four techniques as below:

Distance alteration: based on the inverse square law, light will lose energy when travels further, leading to a reduced intensity. Therefore, the variation of the physical distance between light source and PV cell represents the most basic modification.

Dimming: manipulation of the input power to the light source changes the illuminance of the light and can be used as a light intensity control parameter.

Filtering: optical filters are able to absorb and reflect some portion of the incident light. Therefore, the energy of the passed light beam would be less.

Incident angle: changing the incident angle between the source and cell reduces the active surface of the IPV cell that leads into a smaller value of the incident light.

4 Conclusion and Future Works

The light quality and characteristics of indoor environments are key factors to consider during the design of smart objects powered by PV cells. The requirements of a measurement platform for evaluation of the indoor light condition is presented here. In addition to the general specification for measurements, a suggestion for components, mechanical structure and required signaling is presented. To replicate possible indoor conditions, four light modification techniques are provided.

Installing an active temperature control system is required as the first future task. Moreover, a better interior cover with less reflection would improve the measurement quality.

References

[1] J. Emmerich, M. Roidl, T. Bich and M. ten Hompel; *Entwicklung von energieautarken, intelligenten Ladehilfsmitteln am Beispiel des inBin*; Logistics Journal, Vol. 2012. (urn:nbn:de:0009-14-34309)

[2] A. Kamagaew, T. Kirks and M. ten Hompel; *Energy potential detection for autarkic Smart Object design in facility logistics*; IEEE International Conference on Control System, Computing and Engineering (ICCSCE), pp. 285-290; 2011.

[3] M. Müller, J. Wienold, W.D. Walker and L.M. Reindl; *Characterization of indoor photovoltaic devices and light*; 34th IEEE Photovoltaic Specialists Conference (PVSC), pp. 738-743; 2009.

[4] J.F. Randall and J. Jacot; *Is AM1.5 applicable in practice? Modelling eight photovoltaic materials with respect to light intensity and two spectra*; Renewable Energy, Vol. 28, Issue 12, ISSN 0960-1481, pp. 1851-1864; 2003.

SFB 876 Verfügbarkeit von
Information durch Analyse unter
Ressourcenbeschränkung

Subproject A6
Ressourceneffiziente Analyse von Graphen

Christian Sohler Petra Mutzel Kristian Kersting

Detection of q-Unique DNA Sequences in Genomes

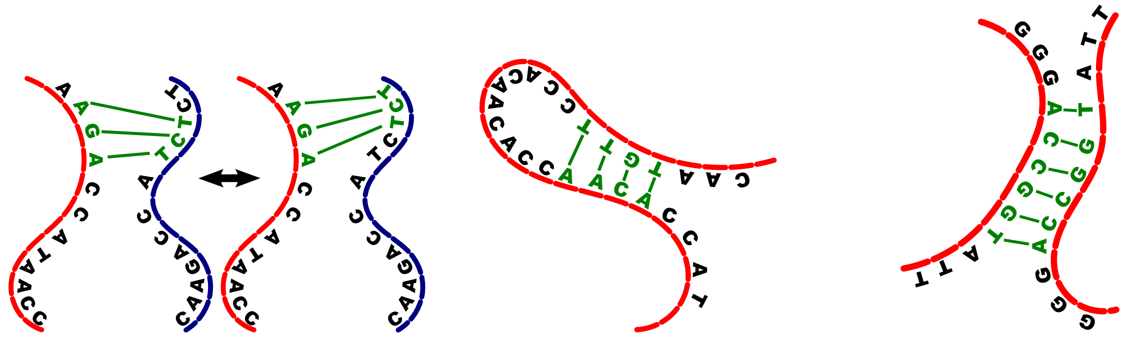
Marianna D'Addario
Department of Computer Science 11
Technische Universität Dortmund
marianna.daddario@tu-dortmund.de

A DNA sequence is q-unique if every q-gram occurs at most once and no reverse complementary q-gram appears. Designing such sequences is already addressed in a former paper [1]. Now we are interested in finding q-unique sequences in real genomes. In this report four viruses genomes were inspected. We found that for two of the genomes the distribution of q-unique sequences is not random, while for the other two it appears to be.

This report aims to show how q-unique sequences are distributed in a real genome. In the field of nanotechnology DNA is used to construct specific geometrical and topological targets [5]. The clue of the construction from DNA is the self-assembly of structures. Combining a set of designed DNA sequences, the structure will build up only due to directed hybridization of the sequences [3], [4]. For this purpose every single DNA sequence should be designed so that only one hybridization state is possible. Especially complementary parts within the same sequence result in formation of secondary structures. Designing q-unique sequences solves this problem. Now we are interested in finding those q-unique sequences in real genomes. Section 1 gives a short introduction into the most important terms. A description of first results is given in Section 2. Finally, Section 3 concludes the report and determines future work.

1 Definition of q-Unique DNA Sequences

The hybridization between two DNA sequences will be formed if one sequence is the reverse complement of the other. Every character c from the DNA alphabet $\Sigma = \{\text{A,C,T,G}\}$ has a complement $\overline{c}$, that is $\overline{\text{A}} = \text{T}$, $\overline{\text{T}} = \text{A}$, $\overline{\text{C}} = \text{G}$ and $\overline{\text{G}} = \text{C}$. For a sequence



(a) The q-gram TCT shows up twice

(b) Sequence containing complementary q-gram

(c) Self-complementary q-gram, only possible for even q

Figure 1: Problems with not q-unique sequences

$s = \sigma_1\sigma_2 \cdots \sigma_n$, where $\sigma \in \Sigma$ the *reverse complement* $\overline{s}$ is defined as $\overline{s} := \overline{\sigma}_n\overline{\sigma}_{n-1} \cdots \overline{\sigma}_1$. A string of length q is called a *q-gram*. A q-gram that is a substring of a given sequence s is called a *q-gram of s*. In the following, for a given sequence s and a given q, we consider the sequence of overlapping q-grams of s. For example $s =$ GATTACA and $q = 4$, we obtain the q-gram sequences (GATT, ATTA, TTAC, TACA).

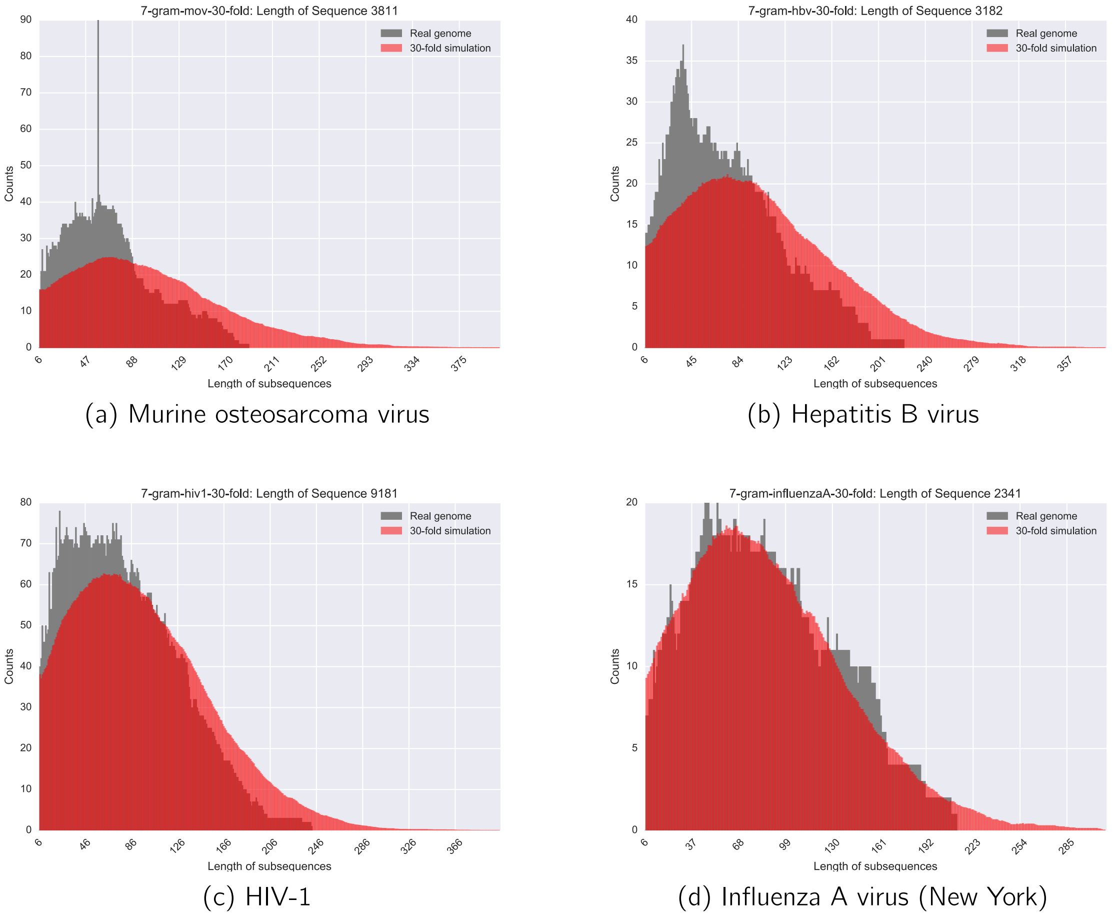
A sequence s is said to be *q-unique* [2] if the following requirements are fulfilled:

1. Every q-gram occurs at most once in s.
2. If a q-gram occurs in s, then its reverse complement does not.

Figure 1 shows problems for the self-assembly of specific targets that derive from not q-unique sequences. The effect of having the same q-gram twice is shown in Figure 1a. The hybridization is not stable due to multiple binding possibilities. In the presence of its reverse complement, the q-gram tends to bind as shown in Figure 1b. This leads to a secondary structure of the sequence and will obstacle the formation of designed target. Note that the second requirement implies that (for even q) no self-complementary q-gram may occur in s. Self-complementary q-grams would lead to bindings between two equal strands, like in Figure 1c. For odd q, self-complementary q-grams do not exist.

2 Finding q-Unique Sequences in Genomes

We pose the question how many q-unique subsequences exist within a genome and how their lengths are distributed. Let G be the sequence of a genome (or a chromosome) and s_i the q-unique subsequence that begins at position i of G with length u_i. Find the set S_G containing all subsequences of G that are q-unique. In particular the set S_G contains for every position i of G a q-unique sequence that can be a subsequence of the prior found subsequence. Consider $|G|$ the length of G, then we define $S_G = \{s_i : u_i \geq u_{i-1} - 1, \; \forall i \in$



(a) Murine osteosarcoma virus

(b) Hepatitis B virus

(c) HIV-1

(d) Influenza A virus (New York)

Figure 2: Comparison between lengths of 7-unique subsequences in genomes and random texts. A set 30 random texts have been generated for every genome. The overlaying (red) histogram shows the average number for every length of 7-unique subsequences.

$|G|\}$. Consider, e. g. $G =$ ACCTTGGCAAT and $q = 2$, then we find $S_G = \{$ACCTTG, CCTTG, CTTGGC, TTGGC, TGGC, GGCAAT$\}$.

In a first approach, we calculate all q-unique subsequences with $q = 7$ for a set of viruses. We chose four genomes: Murine osteosacroma, Hepatitis B, HIV-1, Influenza A (A/New York/392/2004(H3N2)). We downloaded all reference genomes from the NCBI GenBank (`http://www.ncbi.nlm.nih.gov/genbank/`). Figure 2 shows a histogram over the length of all found 7-unique subsequences.

Additionally, for every genome 30 random texts have been generated and a histogram over the averaged number of q-unique subsequences has been added. This overlaying histogram (in red) represents the distribution of the length of all q-unique subsequences within a random genome. The random texts are based on a Markov model of order 0, that considers for a given g the concentration of all g-grams within the real genome. For the simulation of the genomes MoSDi (`https://bitbucket.org/tobiasmarschall/mosdi`) was used.

3 Conclusion and Future Work

Based on the histograms shown in Figure 2a and 2b we assume that for those viruses the lengths of q-unique subsequences are not randomly distributed. The difference between the histograms for real genome and random texts is noticeable. The Murine osteosacroma and the Hepatitis B viruses have a higher number of short q-unique subsequences than expected. For HIV-1 and Influenza A virus the shapes of the histograms are rather similar. This indicates a random distribution of lengths of 7-unique subsequences within those viruses.

A next step of this work is to analyse more genomes. For a better overview it is necessary to compute sets of q-unique subsequences for all reasonable values of q. Also we have to consider genomes of various groups, like bacteria, viruses and fungi separately. In addition, the simulation of the genomes has to be done with every reasonable value of g for the estimation of the underlying text model.

References

[1] Marianna D'Addario, Nils Kriege, and Sven Rahmann. Designing q-Unique DNA Sequences with Integer Linear Programs and Euler Tours in De Bruijn Graphs. In *GCB*, pages 82–92, 2012.

[2] U. Feldkamp, H. Rauhe, and W. Banzhaf. Software tools for DNA sequence design. *Genetic Programming and Evolvable Machines*, 4(2):153–171, 2003.

[3] Lauren Hakker, Alexandria N Marchi, Kimberly A Harris, Thomas H LaBean, and Paul F Agris. Structural and thermodynamic analysis of modified nucleosides in self-assembled DNA cross-tiles. *Journal of Biomolecular Structure and Dynamics*, 32(2):319–329, 2014.

[4] Barbara Saccà and Christof M Niemeyer. DNA origami: the art of folding DNA. *Angewandte Chemie International Edition*, 51(1):58–66, 2012.

[5] Nadrian C Seeman. DNA nanotechnology: novel DNA constructions. *Annual review of biophysics and biomolecular structure*, 27(1):225–248, 1998.

Computing and Enumerating Maximum Common Subgraph Isomorphisms in restricted graph classes

Andre Droschinsky
Chair of Algorithm Engineering (LS11)
Technische Universität Dortmund
andre.droschinsky@tu-dortmund.de

The maximum common subgraph isomorphism (MCS) problem asks for an isomorphism of maximum size between subgraphs of two given input graphs. This problem is well known to be NP-hard in general graphs. However, restricting both the input graphs and the common subgraph to trees renders polynomial time algorithms. For more general graph classes like outerplanar graphs or series parallel graphs polynomial time algorithms are available, if blocks and bridges are preserved (BBP). Since a maximum common subgraph isomorphism is not unique in general, it is of interest to find and list all solutions. In this report we describe our newly developed algorithm to enumerate all maximum common subtree isomorphisms (MCSTs) [2]. We further discuss improvements regarding the time complexity of that algorithm as well as its applicability on other graph classes. The results are in submission to an international conference [3].

Introduction. In many application areas such as pattern recognition [1], or chem- and bioinformatics [7, 10], it is an important task to elucidate similarities between structured objects like proteins or small molecules. A widely-used and successful approach is to model objects as graphs and to identify their MCSs. As a MCS apparently is not unique, it is of interest to find the set of *all* solutions. Since the number of solutions may be superexponential in the input size, the running time cannot be expected to be polynomial in this case. For this reason, enumeration algorithms are said to have *polynomial-delay* if the running time between the output of two solutions (including initialization and halting) is polynomially bounded in the input size [5].

Unfortunately, the MCS problem is known to be NP-hard and consequently a polynomial-time algorithm is not even known for finding a single maximum solution. However, few tractable variants of the MCS problem are known. Edmonds was reported [9] to have proposed a polynomial time algorithm for solving the maximum common subtree isomorphism (MCST) problem, where the input graphs and the subgraphs are trees, by means of maximum weight bipartite matching. Related problems like the *maximum agreement subtree* of rooted trees are well-studied and are, for example, considered in [6] using similar ideas to those of Edmonds [9]. Polynomial time algorithms were presented to find a connected MCS in outerplanar graphs under the additional requirement that blocks, i.e., maximal biconnected subgraphs, and bridges of the input graphs are preserved [11]. However, our enumeration algorithm [2] to compute all maximum common subtree isomorphisms of two given trees was the first proposed polynomial-delay algorithm for any of these efficiently solvable variants.

First, we present the ideas of that algorithm. Next, we discuss our recent ideas leading to a large improvement of the running time. Finally we evaluate strategies to employ the algorithm to other graph classes.

Computing all maximum common subtrees isomorphisms of two given trees. The basic approach to enumerate all MCSTs is separated into two steps. The first step is to compute the order of an MCST between the input trees G and H. Edmonds algorithm solves this problem in time $\mathcal{O}(n^5)$, where n is the order of the larger tree. This is a dynamic programming approach, where MCSTs on pairs of rooted subtrees of G and H are computed. For these pairs the roots must be mapped to each other. Further vertex mappings between these rooted subtrees are determined by maximum weight matchings (MWMs) in certain bipartite graphs. We store these solutions and reuse them in the second step, the enumeration part. After the first step we know which initial mapped edges allow to be expanded into an MCST between G and H. We enumerate all these pairs of edges and recursively add further edges based on the stored MWMs. However, a MWM is not unique in general, therefore we have to consider all the different MWMs in the bipartite graphs mentioned above. To do this we have adapted a polynomial-delay enumeration algorithm from Uno [14] for listing perfect matchings.

In this basic approach each MCST φ is enumerated as many times as there are edges mapped by φ. This is based on the fact, that each initially selected edge-pair, which is mapped by φ, can be expanded into φ. We solved this problem by subsequently deleting edges $e \in E(G)$ after all MCSTs including e have been enumerated. Then we proved that we obtained a polynomial-delay algorithm with time bound $\mathcal{O}(n^6 + Mn^2)$, where M is the total number of different MCSTs.

Improving the running time. Recently [2], we investigated options to improve the running time. One successful alteration was a vertex based approach, i.e., starting from a pair of mapped vertices and add further pairs of vertices, again by means of MWMs. In other words, we investigated the case of rooted trees and considered every possible pair of roots to obtain an MCST between the unrooted trees G and H. Unfortunately, this alteration alone did not change the worst case running time. However, we proved that we still obtain an MCST between G and H if we consider only a certain set of subtrees of G. This allowed to reduce the time bound to $\mathcal{O}(n^4)$. We also proved new upper time bounds dependent on different sizes of G and H as well as the degree of the trees. Edmonds algorithms as well as the node based approach are based on computing MWMs. Therefore we analyzed the bipartite graph instances on which the MWMs are computed. We discovered certain dependencies between them. Exploiting this allowed a cubic worst case running time. We also made some improvements on the enumeration of the MWMs and ultimately on the enumeration on MCSTs, based on a variant of the enumeration algorithm for perfect matchings presented in [4]. These new results, including an experimental evaluation, are currently in submission to an international conference [3].

Employing the algorithm to other graph classes. One might ask if it is possible to reuse the ideas for computing an MCST between trees on other graph classes. Unfortunately even computing a subgraph isomorphism is NP-hard when the smaller graph is a tree and the other is outerplanar [13]. For this reason, a problem variation was introduced, where the common subgraph is required to be block and bridge preserving (BBP). This restriction renders algorithms in outerplanar graphs possible, which are efficient in both, theory and practice. BBP-MCSs yield meaningful results for molecular graphs and even compare favorably to ordinary MCSs in empirical studies [12]. In that paper an approach for the MCS problem which lead to a running time of $\mathcal{O}(n^4)$ in unrooted trees was suggested[1]. Their algorithm also is based to some extend on Edmonds algorithm. Therefore it is possible to apply our new techniques for computing an MCST into their BBP-MCS algorithm. A first experimental evaluation showed a notable improvement, especially for larger graphs. A more careful analysis as well as providing an algorithm to enumerate all BBP-MCSs between outerplanar graphs is one of our current research topics.

References

[1] D. Conte, P. Foggia, C. Sansone, and M. Vento. Thirty years of graph matching in pattern recognition. *Int. J. Pattern. Recognit. Artif. Intell.*, 18(3):265–298, 2004.

[2] Andre Droschinsky, Bernhard Heinemann, Nils Kriege, and Petra Mutzel. Enumeration of maximum common subtree isomorphisms with polynomial-delay. In Hee-Kap Ahn and Chan-Su Shin, editors, *Algorithms and Computation (ISAAC)*, LNCS, pages 81–93. Springer, 2014.

[1] The paper claims a running time of $\mathcal{O}(n^{2.5})$, but the analysis appears to be flawed, see [8, Sec. 3.3.3] for details. Our experimental findings support the time of $\mathcal{O}(n^4)$.

[3] Andre Droschinsky, Petra Mutzel, and Nils Kriege. Maximum common subtree isomorphism in cubic time. Currently in submission to the 18th Conference on Integer Programming and Combinatorial Optimization (IPCO 2016).

[4] Peter Eades and Tadao Takaoka, editors. *Algorithms and Computation, 12th International Symposium, ISAAC 2001, Christchurch, New Zealand, December 19-21, 2001, Proceedings*, volume 2223 of *LNCS*. Springer, 2001.

[5] David S. Johnson, Mihalis Yannakakis, and Christos H. Papadimitriou. On generating all maximal independent sets. *Information Processing Letters*, 27(3):119–123, 1988.

[6] Ming-Yang Kao, Tak-Wah Lam, Wing-Kin Sung, and Hing-Fung Ting. An even faster and more unifying algorithm for comparing trees via unbalanced bipartite matchings. *Journal of Algorithms*, 40(2):212–233, 2001.

[7] Ina Koch. Enumerating all connected maximal common subgraphs in two graphs. *Theoretical Computer Science*, 250(1–2):1–30, 2001.

[8] Nils Morton Kriege. *Comparing Graphs: Algorithms & Applications.* PhD thesis, TU Dortmund, 2015.

[9] David W. Matula. Subtree isomorphism in $O(n^{5/2})$. In P. Hell B. Alspach and D.J. Miller, editors, *Algorithmic Aspects of Combinatorics*, volume 2 of *Annals of Discrete Mathematics*, pages 91–106. Elsevier, 1978.

[10] John W. Raymond and Peter Willett. Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *J. Comput. Aided Mol. Des.*, 16(7):521–533, 2002.

[11] Leander Schietgat, Jan Ramon, and Maurice Bruynooghe. A polynomial-time metric for outerplanar graphs. In Paolo Frasconi, Kristian Kersting, and Koji Tsuda, editors, *Mining and Learning with Graphs*, 2007.

[12] Leander Schietgat, Jan Ramon, and Maurice Bruynooghe. A polynomial-time maximum common subgraph algorithm for outerplanar graphs and its application to chemoinformatics. *Annals of Mathematics and Artificial Intelligence*, 69(4):343–376, December 2013.

[13] Maciej M. Sysło. The subgraph isomorphism problem for outerplanar graphs. *Theoretical Computer Science*, 17(1):91–97, 1982.

[14] Takeaki Uno. Algorithms for enumerating all perfect, maximum and maximal matchings in bipartite graphs. In *Algorithms and Computation (ISAAC)*, volume 1350 of *LNCS*, pages 92–101. Springer, 1997.

Scalable Graph Kernels

Christopher Morris
Chair of Algorithm Engineering (LS11)
TU Dortmund University
christopher.morris@tu-dortmund.de

Real-world graph data is often annotated with continuous node or edge attributes, e.g., physical measurements in the case of chemical molecules, such as atomic mass or bond energy. Unfortunately, state-of-the-art graph kernels for graphs with continuous attributes tend to be slow in comparision with graph kernels that are restricted to discrete labels. In the following, we report our progress towards the development of a scalable graph kernel for graphs with continuous attributes. Moreover, we provide directions for further research for extending the field of application of graph kernels.

The various graph kernels proposed in recent years, e.g., cf. [11], can be divided into approaches that either compute feature maps (i) explicitly, or (ii) implicitly. Explicit computation schemes have been shown to be scalable, e.g., cf. [9], and allow the use of fast linear kernel methods, cf. [4], while implicit ones are typically slow, cf. [2, 6, 7].

Alternatively, we may divide graph kernels according to their ability to handle annotations of vertices and edges, e.g., real-valued vectors. The proposed graph kernels are either (i) restricted to discrete labels, or (ii) compare attributes by user-specified kernels. Remarkably, these two subsets of graph kernels largely coincide, i.e., graphs with discrete labels can be compared efficiently by graph kernels based on explicit feature maps, whereas graph kernels supporting complex annotations use implicit computation and tend to be slow.

Hence, we see the following as a promosing direction for future research: First, we want to show how to obtain explicit feature maps for recently proposed graph kernels employing implicit computation. More specifically, we want to derive approximative explicit mappings of state-of-the-art graph kernels for graphs with continuous attributes such as the recently proposed *GraphHopper graph kernel*, cf. [2], and provide theoretical error bounds. Based on the seminal results of Rahimi and Recht, cf. [8], which prove exponential tails bounds for the approximation ratio of widely-used kernels like the Gaussian RBF kernel, our approach promises high efficiency.

Secondly, we are investigating a method to employ graph kernels based on explicit feature maps, which were originally designed for graphs with discrete labels, to graphs with continuous attributes. That is, in each iteration we first map the set of graphs with continuous attributes to graphs with discrete labels using *probalistic binning functions*. These functions act as an unbiased estimator of an implicit basis kernel, e.g., cf. [1, 3]. For each such resulting graph, we employ a graph kernel for graphs with discrete labels to compute a feature map. Finally, the feature maps of each graph of each iteration are concatinated into a single feature map. From a theory viewpoint, we want to prove that the resulting kernel approximates complex kernels on annotations with high probability, i.e., we want to provide exponential tail bounds. Hence, we extend the work of Rahimi and Recht to more complex structures such as graph data. Finally, we want to backup our theoretical results in an extensive experimental comparison using known data sets from the literature as well as new, larger data sets steming from rational drug design. Additionally, on could reduce the dimensionality of the feature maps by applying state-of-the-art dimension reduction techniques, such as [5, 10, 12].

Thirdly, we are interested in extending the field of application of graph kernels to data not stemming from chemistry or biology. In the past, graph kernels have been almost solely used for the classification of chemical molecules, e.g., protein structures. An interesting direction for further research would be to use graph kernels for *text classification* or *document classification*. Text classification is of particular interest because the amount of data is very large and hence acts as a benchmark for scalable classification algorithms. Consequently, we want to develop algorithms to transform text documents into graphs which represent the local structure of a document. Subsequently, we want to develop efficient graph kernels to classify text documents via text graphs. Since the resulting graphs are very large, we have to resort to graph kernels based on sampling aproaches.

Therefore, we want to develop sampling approaches to speed up existing fast graph kernels, e.g., [9]. Moreover, we want to show tight theoretical bounds for these approaches. Additionally, we want to conduct an extensive experimental study to validate our theoretical findings as well as compare our approach to state-of-the-art text classifiers.

Fourthly, we want to investigate the influence of different *kernel methods*, e.g., support vector machines or gaussian processes, on the classification accurracy of graph kernels. In the past, only support vector machines have been employed for this task. Hence, the question arises if the employed kernel method has a significant influence on the classification accuracy of graph kernels.

References

[1] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni. “Locality-sensitive hashing scheme based on p-stable distributions”. In: *Proceedings of the Twentieth Annual ACM Symposium on Computational Geometry*. Ed. by J. Snoeyink and J.-D. Boissonnat. New York: ACM, 2004, pp. 253–262.

[2] A. Feragen, N. Kasenburg, J. Petersen, M. D. Bruijne, and Borgwardt K. M. “Scalable kernels for graphs with continuous attributes”. In: *Advances in Neural Information Processing System*. Ed. by C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger. Red Hook, NY: Curran Associates, 2013, pp. 216–224.

[3] P. Indyk and R. Motwani. “Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality”. In: *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing*. Ed. by J. S. Vitter. New York: ACM, 1998, pp. 604–613.

[4] T. Joachims. “Training Linear SVMs in Linear Time”. In: *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: ACM, 2006, pp. 217–226.

[5] P. Kar and H. Karnick. “Random Feature Maps for Dot Product Kernels”. In: *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*. Ed. by N. D Lawrence and M. A. Girolami. Vol. 22. 2012, pp. 583–591.

[6] N. M. Kriege and P. Mutzel. "Subgraph Matching Kernels for Attributed Graphs". In: *Proceedings of the Twenty-sixth Annual International Conference on Machine Learning*. Ed. by J. Langford and J Pineau. Madison, WI: Omnipress, 2012, pp. 1113–1120.

[7] F. Orsini, P. Frasconi, and L. De Raedt. "Graph Invariant Kernels". In: *Proceedings of the Twenty-fourth International Joint Conference on Artificial Intelligence*. Ed. by Q. Yang and M. Wooldridge. Palo Alto CA: AAAI Press, 2015, pp. 3756–3762.

[8] A. Rahimi and B. Recht. "Random features for large-scale kernel machines". In: *Advances in Neural Information Processing Systems*. Ed. by D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou. Red Hook, NY: Curran Associates, 2008, pp. 1177–1184.

[9] N. Shervashidze, P. Schweitzer, E. J. van Leeuwen, K. Mehlhorn, and K. M. Borgwardt. "Weisfeiler-Lehman Graph Kernels". In: *Journal of Machine Learning Research* 12 (2011), pp. 2539–2561.

[10] A. Vedaldi and A. Zisserman. "Sparse kernel approximations for efficient classification and detection". In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*. Washington, DC: IEEE, 2012, pp. 2320–2327.

[11] S. V. N. Vishwanathan, N. N. Schraudolph, R. Kondor, and K. M. Borgwardt. "Graph Kernels". In: *Journal of Machine Learning Research* 11 (2010), pp. 1201–1242.

[12] K. Weinberger, A. Dasgupta, J. Langford, A. J. Smola, and J. Attenberg. "Feature Hashing for Large Scale Multitask Learning". In: *Proceedings of the twenty-sixth Annual International Conference on Machine Learning*. Ed. by L. Bottou and M. Littman. New York, NY: ACM, 2009, pp. 1113–1120.

SFB 876 Verfügbarkeit von
Information durch Analyse unter
Ressourcenbeschränkung

Subproject B1
Analysis of Spectrometry Data with Restricted Resources

Jörg Rahnenführer

Jörg Ingo Baumbach

A comparison of analysis processes in MCC-IMS

Salome Horsch
Statistical Methods in Genetics and Chemometrics
Technische Universität Dortmund
salome.horsch@tu-dortmund.de

Diagnosing diseases via breath sample analysis is a budding field of research since it has many advantages compared to traditional diagnosting tools like the analysis of blood. Multi-capillary-column-Ion-mobility-spectrometry (MCC-IMS) is a breath analyzing technique which is fast, cheap and doesn't require a vacuum. It is an imaging technique so the first step towards statistical classification of diseases is feature extraction in the form of peak detection followed by peak alignment among several measurements. Afterwards classification algorithms are applied. In order to evaluate the quality of different methods for these three kinds of problems we consider 156 possible combinations and compare their respective classification results.

The analysis of human breath becomes feasible by MCC-IMS. This thechnology allows the detection of volatile organic compounds (VOCs) in a breath sample at ambient pressure and temperature within approximately ten minutes. It showed promising results in many different applications (see [1] for an overview). It is an imaging technique and therefore requires preprocessing. There are different methods available but it is unclear, which ones work best. Currently this preprocessing is accomplished manually. For broad application of MCC-IMS automated algorithms are necessary. We seek to answer the question if algorithms exist that can at least keep up with the manual approach. After feature definition statistical classification algorithms are applied, also leading to the question which ones work best in this scenario.

Data The used data stems from the Department of Pulmonology, Ruhrlandklinik, University Hospital of Essen, Germany containing 67 observations. The airways of 37 probands were infected or colonized by Pseudomonas aeruginosa. 30 probands were healthy non-smoker controls.

1 Methods

The process of MCC-IMS consists of two steps resulting in three variables per metabolite. The combination of the first two variables is specific for the correspondent metabolite in the air. The third variable carries information about the signal intensity, the amount of that compound in the breath sample. The three variables are displayed in a heat map with colors representing the signal intensity.

1. Peak picking The combination of the three variables results in one image per measurement with visible peaks that need to be identified and quantified in order to recognize the ingredients of the breath sample. Different methods were applied (VisualNow (VN) manual and automatic, Local Maxima (LM), Peak Model Estimation (PME), Peak Detection by Slope Analysis (PDSA), Savitzky-Golay Laplace-operator filtering thresholding Regions (SGLTR) and Online Peak Model Estimation (OPME)). Each method includes its own filtering steps to reduce noise in the image.

2. Peak clustering After the peak picking step we have a list for each measurement with information about the found peaks, usually the location and an intensity value for each peak. To obtain a data table with observations and well defined features the peaks have to be aligned among the measurements to decide which peaks are the same. We call the alignment of peak locations peak clustering. Different methods were applied (VisualNow (VN), Grid Square (GS), Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Cluster Editing (CE) and EM Clustering (EM)).

3. Classification After these two steps our dataset contains observations from two classes (here diseased and healthy) and intensities of some peaks. Several statistical algorithms (Support Vector Machine (SVM) with linear and rbf-kernels, Classification Tree (CT), Random Forest (RF), Generalized Boosted Models (GBM) and K-Nearest-Neighbor (KNN)) are used, trying to distiguish the two groups. We use 10-fold cross-validation (CV) and 10-fold nested CV where parameter optimization was necessary to ensure internal validity. Since CV results always depend on the random split, we repeat each CV run 50 times to observe the variability.

More information and further references on the methods mentioned in this section can be found in [2].

2 Results

To evaluate which peak picking, peak clustering and classification algorithms are best in this scenario, we combine them (where technically feasible) resulting in 156 constellations. We use accuracy and AUC values as performance measures.
The results published in [2] are visualized as boxplots in Figure 1. Each panel contains the

boxplots of one classification algorithm. The boxes separated by vertical lines belong to one peak picking method and the colors correspond to a certain peak clustering method. The boxes result from the 50 CV runs. A line is drawn at an accuracy of 0.8 for orientation.

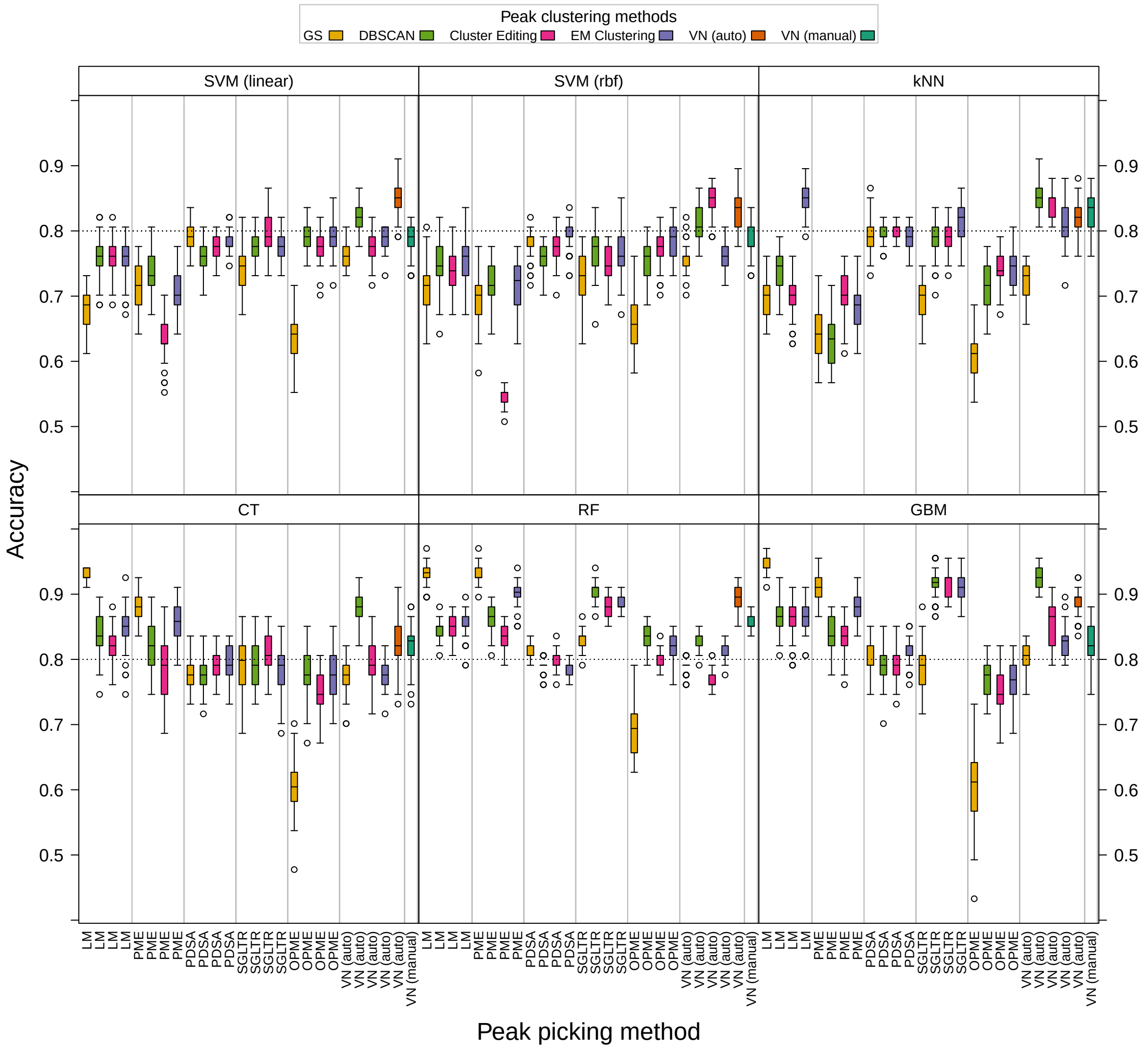


Figure 1: Boxplots for the Accuracies achieved by 156 combinations of peak picking, peak clustering and classification in a 50 times repeated CV.

Most of the boxes of the k-Nearest-Neighbor algorithm and the SVMs (for the linear as well as for the rbf kernel) lie below the line at 0.8. In the second row the results of the three tree based classification algorithms are displayed. The Classification Tree achieves better results than the methods mentioned before but the highest accuracies are reached by the *extensions* of the CT, namely the Random Forest and the GBM. Most of their boxes lie above the line.

Picking	Clustering	Class	Accuracy	AUC
LM	GS	GBM	0.940	0.960
LM	GS	RF	0.933	0.977
LM	GS	CT	0.925	0.927
PME	GS	RF	0.925	0.985
SGLTR	CE	GBM	0.925	0.974
VN (auto)	DB	GBM	0.925	0.921
SGLTR	DB	GBM	0.918	0.979
PME	GS	GBM	0.910	0.930
SGLTR	DB	RF	0.910	0.980
SGLTR	EM	GBM	0.910	0.973
PME	EM	RF	0.903	0.918
VN (manual)	VN (manual)	RF	0.851	0.923

Table 1: Combinations of peak picking, peak clustering and classification algorithms with a median accuracy above 0.9.

It is not easy to determine which peak picking and peak clustering methods work best. The accuracy depends on certain combinations. To get a quick overview Table 1 contains all combinations with a median accuracy above 0.9. The correspondent median AUC values are also displayed. All of these combinations used a tree based classification algorithm. One used Classification Trees, four Random Forests and six the boosting approach. The best combination of peak picking and peak clustering methods is Local Maxima peak picking with Grid Squares peak clustering with median Accuracies of 0.925 to 0.940 depending on the used classification algorithm. But also other combinations lead to good results, e.g. combinations containing SGLTR or PME.

Outlook: Our current research extends theses analyses particularly with regard to the examination of further datasets. We hope to find combinations of peak picking, peak clustering and classification algorithms that lead to good and stable results on most datasets in order to accomplish standard procedures for these tasks in MCC-IMS.

References

[1] T. Fink, J. I. Baumbach, and S. Kreuer. Ion mobility spectrometry in breath research. *J. Breath Res.*, 2, 2014.

[2] S. Horsch, D. Kopczynski, J. I. Baumbach, J. Rahnenführer, and S. Rahmann. From raw ion mobility measurements to disease classification: a comparison of analysis processes. *PeerJ PrePrints*, 3, 2015.

Subproject B2
Resource optimizing real time analysis of artifactious image sequences for the detection of nano objects

Heinrich Müller Roland Hergenröder Jian Jia Chen

Utilization Based Schedulability and Optimization Analysis for Non-Preemptive Static Priority Scheduling

Georg von der Brüggen
Computer Science 12
Technische Universität Dortmund
georg.von-der-brueggen@tu-dortmund.de

Especially for many embedded systems it is important that tasks can be executed with real time guarantees. This means that the execution of a task does not only provide a correct value but that this value is provided within a certain amount of time, commonly known the tasks deadline. In scheduling analysis preemption is often considered to be to be very important to ensure the schedulability of real time systems as it allows to allocate the processor to high-priority tasks nearly immediately. However, for practical real time systems preemption is often not allowed due to the resulting preemption overhead. This overhead is often neglected in theoretical analysis but can be in the same magnitude as the actual execution times in practical cases. In our paper presented in the Euromicro Conference on Real-Time Systems (ECRTS) 2015 [4] we provided the first safe sufficient schedulability test in hyperbolic form that can verify the schedulability of static priority non-preemptive scheduling algorithms. For many important scheduling schemes like Deadline Monotonic and Rate Monotonic scheduling this test runs in linear time. We used this test to provide a better upper bound on the processor speedup factors for non-preemptive fixed priority scheduling of constrained and implicit deadline tasks sets. We also provided the first general utilization bound for non-preemptive Rate Monotonic scheduling that is based on task utilization and the task sets blocking factor $\gamma > 0$.

Schedulability Test For real time task sets $\tau = \{t_1, \ \ldots, \ \tau_n\}$ the sporadic task model is widely adopted, where each task τ_i is described by its relative deadline D_i, its Worst Case Execution Time (WCET) C_i, and its minimum inter arrival time or period T_i. It assumes that each tasks releases an infinite number of instances called jobs where the release times of consecutive jobs are at least separated by the period. The task utilization is $U_i = \frac{C_i}{T_i}$. To guarantee the schedulability of a task set it is necessary that $R_i \leq D_i \ \forall \tau_i$ where R_i is the Worst Case Response Time of τ_i.

For static priority scheduling algorithms all tasks and the related jobs are assigned with a fixed priority. We focused on constrained ($D_i \leq T_i$) and implicit deadline ($D_i = T_i$) task sets. In this case we can use the critical instant theorem to test the schedulability of a task τ_k, i.e., release the task together with all higher priority tasks and release all subsequent jobs of higher priority tasks as early as possible. In the preemptive case the schedulability of a task set can be determined by calculating the R_i of all tasks using Time Demand Analysis (TDA) [3]. A task τ_k is schedulable if the following equation holds [3]

$$\exists t \text{ with } 0 < t \leq D_k \ \text{ and } \ C_k + \sum_{\tau_i \in hp(\tau_k)} \left\lceil \frac{t}{T_i} \right\rceil C_i \leq t$$

where $hp(\tau_k)$ is the subset of tasks in τ with priority higher then τ_k. If this holds for all $\tau_k \ \in \ \tau$ the task set is schedulable by the given scheduling scheme.

For non-preemptive scheduling the blocking time B_k of each task, i.e., the maximum time the tasks can be blocked by lower priority tasks running in non-preemptive mode, has to be taken into account as well. A *strict upper bound* of the maximum blocking time of a task τ_k can be determined as $B_k \ = \ \max_{\tau_i \ \in \ lp(\tau_k)} \{C_i\}$ where $lp(\tau_k)$ is the subset of tasks with priority lower then τ_k. The blocking factor γ_k of task τ_k is defined as B_k/C_k. Due to this small overestimation of the blocking time we could use a revised form of TDA to determine the schedulability of τ_k for non-preemptive fixed priority scheduling:

$$\exists t \text{ with } 0 < t \leq D_k \quad \text{and} \ \ C_k + B_k + \sum_{\tau_i \in hp(\tau_k)} \left\lceil \frac{t}{T_i} \right\rceil C_i \leq t$$

It is sufficient to only test the $t \ \in [0; \ D_k]$ where a task with higher then τ_k arrives. This leads to a sudo-polynomial runtime of TDA. We sacrificed some presition to restrict this further by only testing D_k and the last release of higher priority tasks before D_k the test runs in linear time. We referenced to this last release of τ_i before D_k as t_i and used $t_k = D_k$ for notation brevity. This allowed us to rewrite the formulation in a way that the test is based on the utilization of the task instead of WCET and period:

$$\exists t_j \in \{t_1, \ \ldots, \ t_k\} \text{ and } C_k + B_k + \sum_{i=1}^{k-1} t_i U_i + \sum_{i=1}^{j-1} t_i U_i \leq t_j$$

From this equation we derived the following schedulability test for a task τ_k in a non-preemptive task set under static priority scheduling:

A task τ_k in a non-preemptive sporadic task system with constrained deadlines can be feasibly scheduled by a fixed-priority scheduling algorithm, if the schedulability for all higher priority tasks has already been ensured and the following condition holds:

$$\left(\frac{C_k + B_k}{D_k} + 1\right) \prod_{\tau_j \in hp(\tau_k)} (U_j + 1) \leq 2$$

If the $\tau_i \in \tau$ are tested in increasing priority order this test can be used to verify the schedulability of the entire task set in linear time as $\prod\limits_{\tau_j \in hp(\tau_k)} (U_j + 1)$ for τ_k can be calculated directly from the value used to test τ_{k-1} and the utilization of τ_{k-1}.

Speedup Factor Static priority scheduling is widely adopted in real life embedded systems. This is due to its easy implementation, thus leading to a smaller overhead during runtime compared to dynamic priority scheduling. Nevertheless, dynamic priority scheduling has better response time guarantees and is, in general, able to successfully schedule task sets with higher utilizations. This loss can be determined using a processor speedup factor for an scheduling algorithm A, defined in [2] as $f^A = \max\limits_{\forall \tau} \left\{ \frac{f^A(\tau)}{f^{opt}(\tau)} \right\}$, where $f^{opt}(\tau)$ is the speed an optimal scheduling algorithm needs to successfully schedule τ and $f^A(\tau)$ is the necessary speed if algorithm A is used. Previously in the non-preemptive case for the speedup factor of static priority algorithms in comparison to the optimal work conserving dynamic priority scheduling algorithm earliest deadline first a lower bound of ≈ 1.76322 and an upper bound of 2 was known for both implicit and constrained deadline task sets. Using the presented linear time schedulability test we showed the upper bound of the speedup factor is ≈ 1.76322 as well, thus closing the gap.

Utilization Bounds Due to the worst case setting of blocking a task τ_k by a lower priority tasks with very large WCET and period compared to the deadline of τ_k the general utilization bound of non-preemptive scheduling for implicit deadline task sets is 0 for both dynamic and static priority scheduling. However, if the relation between deadline and blocking time is bounded, schedulability guarantees based on task utilizations can still be provided. The task sets blocking factor is defined as $\gamma = \max_{\tau_k} \left\{ \max\limits_{\tau_i \in lp(\tau_k)} \left\{ \frac{C_i}{C_k} \right\} \right\} = \max_{\tau_k} \{\gamma_k\}$.

Previously only a utilization bound of 25% for the CAN-Bus was provided by Andersson and Tovar [1]. They show that the utilization bound is $\frac{1}{1+\gamma}$ for $\gamma \geq 2$. In Theorem 4 in [4] we showed that $\frac{1}{1+\gamma}$ also holds for $0 < \gamma < 2$ and thus for the first time provided a general utilization bound for non-preemptive scheduling based on the blocking factor. We improved this result in Theorem 9 in [4]. The related curves are presented in Figure 1.

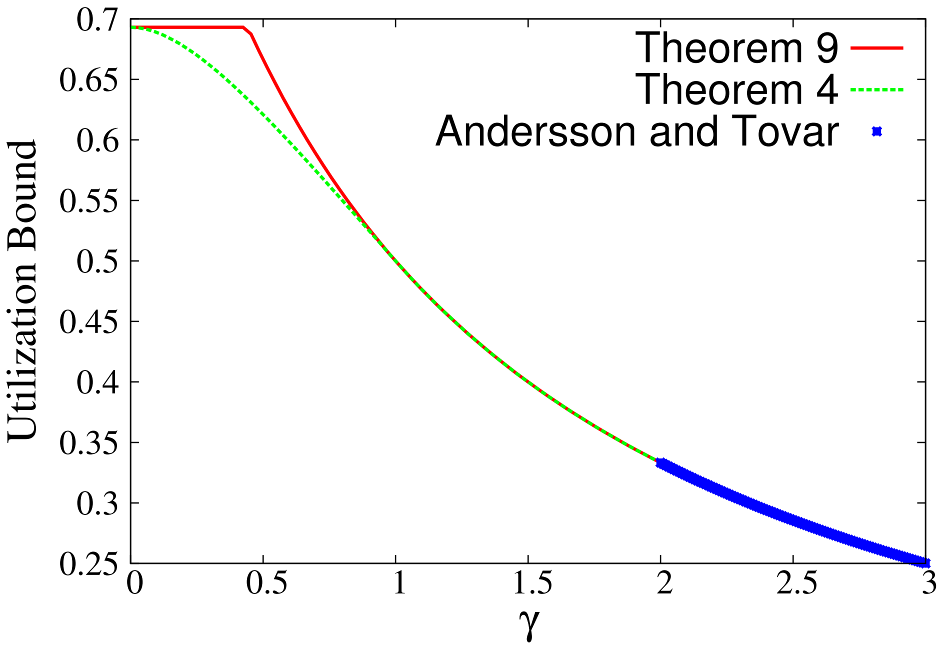


Figure 1: Comparison of the total utilization bound of RM-NP with respect to $\gamma = \max_{\tau_k}\left\{\max_{\tau_i \in lp(\tau_k)}\left\{\frac{C_i}{C_k}\right\}\right\}$ provided by Andersson and Tovar [1], Theorem 4 in [4], and Theorem 9 in [4].

References

[1] Björn Andersson and Eduardo Tovar. The utilization bound of non-preemptive rate-monotonic scheduling in controller area networks is 25%. In *IEEE Fourth International Symposium on Industrial Embedded Systems - SIES*, pages 11–18, 2009.

[2] Robert I. Davis, Thomas Rothvoß, Sanjoy K. Baruah, and Alan Burns. Exact quantification of the suboptimality of uniprocessor fixed priority pre-emptive scheduling. *Real-Time Systems*, 43, 2009.

[3] John P. Lehoczky, Lui Sha, and Y. Ding. The rate monotonic scheduling algorithm: Exact characterization and average case behavior. In *IEEE Real-Time Systems Symposium'89*, pages 166–171, 1989.

[4] Georg von der Bruggen, Jian-Jia Chen, and Wen-Hung Huang. Schedulability and optimization analysis for non-preemptive static priority scheduling based on task utilization and blocking factors. In *27th Euromicro Conference on Real-Time Systems, ECRTS*, pages 90–101, 2015.

Scheduling Self-Suspension Tasks in Mobile Cloud Computing

Wen-Hung Kevin Huang
SFB 876, Project B2
Computer Science XII, TU Dortmund
wen-hung.huang@tu-dortmund.de

Self-suspension is becoming a prominent characteristic in real-time systems such as: (i) I/O-intensive systems (ii) multi-core processors, and (iii) computation offloading systems with coprocessors, like Graphics Processing Units (GPUs) or mobile cloud computing. In this report, we study self-suspension systems under fixed-priority (FP) fixed-relative-deadline (FRD) algorithm by using release enforcement to control self-suspension tasks' behavior. Specifically, we use equal-deadline assignment (EDA) to assign the release phases of computations and suspensions.

1 Introduction

Over the last decade, mobile devices are becoming more and more pervasive: applications with ever richer functionalities are provided anytime, anywhere such as (*i*) global positioning systems (GPS) (*ii*) real-time language translation (*iii*) social networking. However, the resources on mobile systems are limited in terms of batter life, network bandwidth, storage capacity, and processing ability. Such resource scarceness poses a big challenge of the design and implementation of a mobile system that exploits its full potential of mobile computing. Mobile cloud computing is a natural solution that addresses this problem by executing mobile applications on shared resource providers external to the mobile system. As shown in Figure 1, some execution of a process is offloaded to an external machine, a.k.a. cloud, as so to optimize execution time and energy use.

Before making the offloading decision, the offloadable parts of a program have to be identified. This is usually done by manually program annotating . A typical approach represents the program as a graph: the vertices represent the computational components (such as functions) and the edges represent the communication between them. Figure 2 shows an example of dividing a program using graph partition. The program takes input

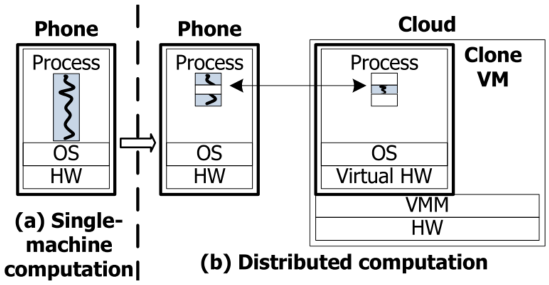


Figure 1: (a) The classic execution on a mobile system and (b) the execution with cloud computing, as shown in [2]

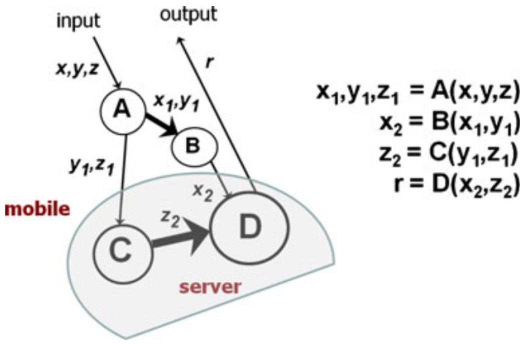


Figure 2: A program is expressed as a graph and partitioning the graph between a mobile system and a server, as shown in [4]

x, y, z and gives output r. In the figure, the computation consists of four functions A, B, C, D. Each of these functions is a possible candidate for offloading. The objective is to decide which of these functions to offload, depending on their requirements of energy consumption and execution time, as illustrated in Figure 2.

In this report, we assume that the offloading decision is static. In other words, the program is partitioned during development. After static program partitioning, each task becomes a multi-segment self-suspension task. A multi-segment self-suspension task τ_i is characterized by 3 tuples:

$$\tau_i = \left((C_i^0, S_i^0, C_i^1, S_i^1, \ldots, S_i^{m_i-2}, C_i^{m_i-1}), T_i, D_i\right) \tag{1}$$

where T_i denotes the minimum inter-arrival time of τ_i, D_i denotes the relative deadline of task τ_i; C_i^j denotes the upper bound on the execution time of the $(j+1)$th- computation segment; and S_i^j denotes the upper bound on the suspension time of the $(j+1)$th-suspension interval. In this work, we aim at answering the question of how to schedule these self-suspension tasks such that all the tasks are schedulable: every task meets its relative deadline. In this report, we focus on fixed-priority scheduling, in which each task is associated with a unique priority level. Specifically, we prioritize tasks according to the suspension-laxity-monotonic (SLM) priority assignment: the smaller the suspension laxity $D_i - S_i$, the higher the priority level.

2 Fixed-Relative-Deadline (FRD) Scheduling and Generalized Multiframe (GMF) Tasks

The Fixed-Relative-Deadline (FRD) scheduler works based on a release (scheduling) enforcement that assigns each computation segment C_i^j with a relative deadline D_i^j:

- The release times between two consecutive computation segments C_i^j, C_i^{j+1} are separated by *exactly* D_i^j time units plus the upper bound time on the suspension interval, i.e. $D_i^j + S_i^j$.
- Each suspension interval S_i^{j+1} is released at the absolute deadline of the previous computation segment. That is, if a computation segment C_i^j is released at time t and is set an absolute deadline $t + D_i^j$ under FRD scheduling, then the follow-up suspension interval is released at $t + D_i^j$.
- The sum of the assigned relative deadlines of task τ_i's computation segments cannot exceed $D_i - S_i$, i.e., $\sum_{j=0}^{m_i-1} D_i^j \leq D_i - S_i$.

Equal-Deadline Assignment (EDA) assigns equal relative deadlines to computation segments, considering from a total slack $D_i - S_i$:

$$D_i^0 = D_i^1 = \cdots = D_i^{m_i-1} = \frac{D_i - S_i}{m_i} \tag{2}$$

The generalized multiframe (GMF) task model was first introduced by Baruah et al. [1]. A GMF task ψ_i consisting of m_i frames is characterized by the 3-tuple $(\vec{C}_i, \vec{D}_i, \vec{T}_i)$, where $\vec{C}_i$,$\vec{D}_i$, and $\vec{T}_i$ are m_i-ary vectors $(C_i^0, C_i^1, ..., C_i^{m_i-1})$ of execution requirements, $(D_i^0, D_i^1, ..., D_i^{m_i-1})$ of relative deadlines, $(T_i^0, T_i^1, ..., T_i^{m_i-1})$ of minimum inter-arrival times, respectively. In fact, tasks under the FRD scheduler can be equivalently transformed to the well-known generalized multiframe (GMF) [1] tasks. The details can be found in [3]. As a result, the schedulability test proposed in [5] can be thereafter adopted.

3 Experimental Results

We conduct experiments using synthesized task sets for evaluating the schedulability tests as follows: the polynomial-time test for dynamic self-suspension tasks under RM scheduling, denoted by *Idv-Burst-RM*; the pesudo-polynomial-time algorithm and analysis for handling the general dynamic self-suspension system under fixed-priority scheduling, denoted by *PASS-OPA*; the pseudopolynomial-time analysis considering jitter, under SLM priority assignment, denoted by PH-SLM; and the proposed method using SLM priority assignment in this report, denoted by EDAGMF-SLM.

In Figure 3, we show the result for the performance by these tests above in terms of the acceptance ratio, from different suspension types and lengths. The performance by Idv-Burst-RM is inferior to all the others in all the cases. The acceptance ratios by PH-SLM and PASS-OPA are almost identical. The proposed EDAGMF-SLM is far more effective than PASS-OPA and PH-SLM in the case of the rare suspension type (Figure 3a, 3b, and 3c). In the case of the frequent suspension type (Figure 3d, 3e, and 3f), the acceptance by EDAGMF-SLM drops down close to that by PASS-OPA and PH-SLM. Generally speaking, our proposed approach, by using enforcement, is impractical to be adopted for task sets with many suspension intervals (≥ 5); however, it is highly useful for task sets with small numbers of suspension intervals (< 5).

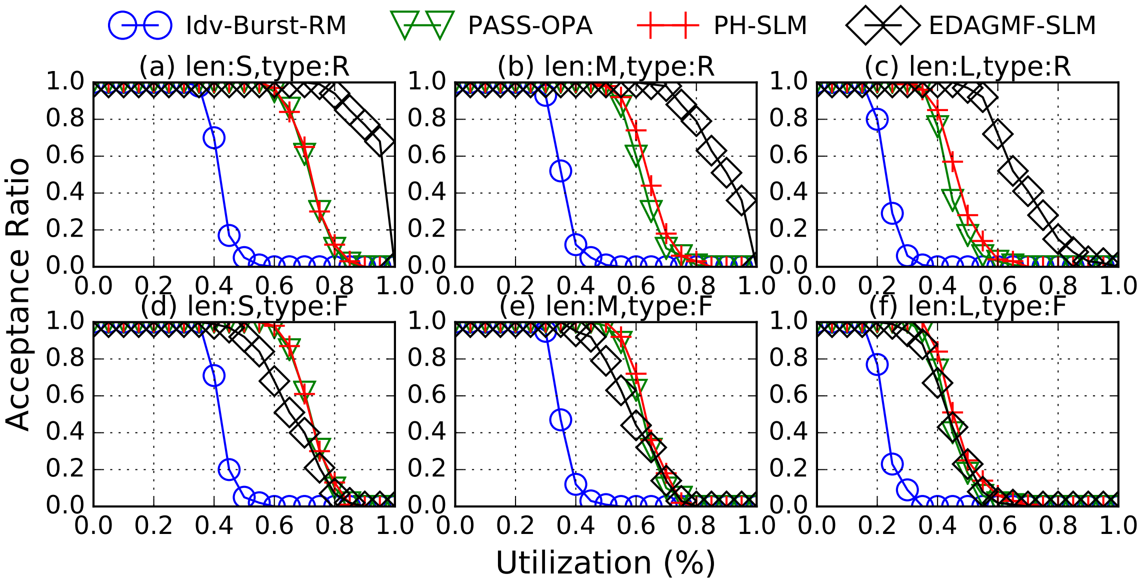


Figure 3: Comparison with different types of suspension lengths (leng) and different types of suspension frequency (type).

4 Conclusion

In this report, we propose to adopt SLM priority assignment and the FRD scheduler using EDA for scheduling multi-segment self-suspending tasks. We empirically show that our proposed approach is highly effective if the number of suspending intervals is small (< 5), in terms of the number of task sets that are deemed schedulable, despite that it is impractical for task sets with many suspension intervals (≥ 5). As a result, we recommend that the constraint on the maximum number of segments in the solver for finding the partitioning strategy be imposed so as to avoid such unfavorable cases.

References

[1] Sanjoy Baruah, Deji Chen, Sergey Gorinsky, and Aloysius Mok. Generalized multiframe tasks. *Real-Time Systems*, 17(1):5–22, 1999.

[2] Byung-Gon Chun, Sunghwan Ihm, Petros Maniatis, Mayur Naik, and Ashwin Patti. Clonecloud: elastic execution between mobile device and cloud. In *European Conference on Computer Systems, Proceedings of the Sixth European conference on Computer systems, EuroSys 2011, Salzburg, Austria, April 10-13, 2011*, pages 301–314, 2011.

[3] Wen-Hung Huang and Jian-Jia Chen. Self-suspension real-time tasks under fixed-relative-deadline fixed-priority scheduling. In *Design, Automation and Test in Europe (DATE)*, Dresden, Germany, 14 -18th Mar 2016.

[4] Karthik Kumar, Jibang Liu, Yung-Hsiang Lu, and Bharat K. Bhargava. A survey of computation offloading for mobile systems. *MONET*, 18(1):129–140, 2013.

[5] Hiroaki Takada and Ken Sakamura. Schedulability of generalized multiframe task sets under static priority assignment. In *RTCSA*, pages 80–86, 1997.

Crest lines of real functions for image analysis and point-set regression

Thomas Kehrt
Lehrstuhl VII Graphische Systeme
Technische Universität Dortmund
thomas.kehrt@tu-dortmund.de

The report is devoted to aspects of ridges, or crests, of real functions of two variables. It starts with various definitions of ridges or crests lines know in literature. Next, it presents several areas of application of the concept in image analysis and point-set regression. Then the common basic algorithmic approach to the calculation of crest lines is outlined and an own adaptation is described. Finally topics of possible future research are compiled.

1 Crest lines

Ridge or crest lines of real functions of two variables are some sort of maximum besides the "'tops of mountains"' (zero gradient) which usually are considered as maxima in mathematical optimization. Crest lines are a set of curves whose points are local maxima of the function in at least one dimension. There are several possiblities to formalize this intuitive definition [6].

One possibilty of defining ridge or crest lines is to employ curvature concepts of differential geometry [2]. In this case crest lines are the locus of points on a surface whose absolute value of largest curvature is locally maximal in the associated principal direction [4].

A second possiblity is to use concepts of vector field topology. The vector field under consideration is the vector field of gradients of the given function. In this case, crest line points are points with minimal gradient length in direction of the tangent of the iso-line traversing the point under consideration, cf. chapter 3. As a dual view, the level set topology may be considered [5]. Alternatively, crest lines may be defined as separatrices in vector fields [9].

2 Applications

Crest lines are of interest in several applications, e.g. image analysis, point set clustering, and regression in point sets, as outlined in the following.

A gray-level image can be considered as a function in two variables. In images showing tube-like structures, crest lines induce a reduction to curve structures. The curve structures may be employed as features for image registration. An example is registration of individual brains with a brain atlas [8]. Futhermore, the crest line may serve as some sort of skeleton, similar to the medial axis, which may be used for image content recognition.

Many applications can be reduced to the analysis of big sets of finite points. Useful smooth functions related to point sets are so-called kernel density functions. A kernel density function $D(x)$ is a mean sum of kernel functions $K_i(x)$, i.e. $D(x) = \sum_{i=0}^{N} K_i(x)$, where each of the N point contributes one kernel function. Typical examples of kernel functions are Gaussians which are centered at the given points. Kernel density functions may be used for point set clustering. A prominent example is mean shift clustering [1]. In this case clusters are induced by maxima of the kernel density function. Considering crest lines in addition could give more insight into the structure of feature spaces. This could lead to refined classifications of the events related to the features.

An example of an application of the analysis of big point sets is boundary reconstruction in noisy point data of a 3D-scanner. Figure 1, left, shows registered slices of 3D-surface scans of a human head as an example. The aim of reconstruction is a continuous curve (in 2D) or surface (in 3D) which approximates the given point set. Such curves or surfaces, respectively, might be defined by crest lines. A general view is to consider those curves or surfaces as continuous regressions of the data set.

A special application is the reconstruction of trajectories of small moving objects in noisy image sequences. In this case curve structures approximating the trajectories are fitted. This task occurs in the PAMONO sensor of sub-project B2 of the SFB 876 for so-called rolling viruses.

A further possibility of asigning continuous fuctions to a point set are distance functions. The distance function d of a point set is the infimum taken over distance kernels $d_p(x) = ||x - p||$, where p denotes an input point. In this case "‘valleys"’ are considered which are "‘ridges"’ of the negative distance function. A variant of distance function is the signed distance function, Signed distance functions are defined with respect to a closed, oriented curve (2D) or surface (3D) which induce inside and outside regions. Then the distance kernel is defined as $d_{\text{signed}}(x) = d(x)$ if x is outside, and $d_{\text{signed}}(x) = -d(x)$ if x is inside, where d(x) is the unsigned distance function. For finite point sets with normals, the normals may be used to define the signature.

3 Algorithms

Crest lines are usually calculated by so-called marching algorithms. Thirion et al. [7, 8] have developed a marching algorithm especially for crest lines. Based on the observation that the second derivative is maximal on crest-lines, they define an extremality equation based on third-order derivation to identify crest lines.

In the special case of signed distance functions, the well-known marching-cubes algorithm (MC algorithm, [3]) or one of its more recent variants may be employed. The great advantage of the MC approach is its robustness for explicitly extracting separating surfaces. The crucical problem is the requirement of a signature whose definition on finite point sets may be non-trivial, if possible at all.

In the following an algorithm of crest line calculation based on marching along extremal gradients is presented, i.e. the gradient vector field view is applied (chap. 1). The algorithm is demonstated for the example of reconstructing a boundary curve of a region represented by the scan data of Fig. 1, left. To recover the boundary, a density function D for the given points as desrcibed above is used which employs a kernel with low-pass characteristic. The second and the fourth image from the left of Fig. 1 visualize such density functions for kernels of different width, where every pixel represents a function value of an evaluation on a grid, using a gray scale mapping of the value.

The extremal gradients required are determined from isolines using the observation that the length of the gradient is locally minimal along the isoline at intersection points of isolines with crest lines. The fourth image from the left of Fig. 1 shows several gray levels whose boundaries are isolines (please enlarge in the pdf-viewer). Since the gradient at a point x is orthogonal to the tangent t of the isoline traversing x, a computationally easier condition is that the length of the gardient is minimum in direction orthogonal to the gradient, i.e.

$$\lim_{\epsilon \to 0} \frac{\|\nabla D(x + \epsilon t)\| - \|\nabla D(x)\|}{\|\epsilon t\|} = 0.$$

The images of Fig. 1 in the middle and on the right show the gray-coded values of this expression for the two density functions. The thin black curve following the ridge approximates the crest line. It may be explicitly reported by tracking the black pixels along the grid, or by grid-free incremental numerical tracking.

4 Topics of future research

An issue of existing marching algorithms different from MC is the numerical stability, in particular in the case of non-sharp crests. A further topic is the generalization of crest lines to functions of higher dimension, and the desgin of related stable and efficient algorithms. Here, the vector field view could be helpful. The usefulness of crests for

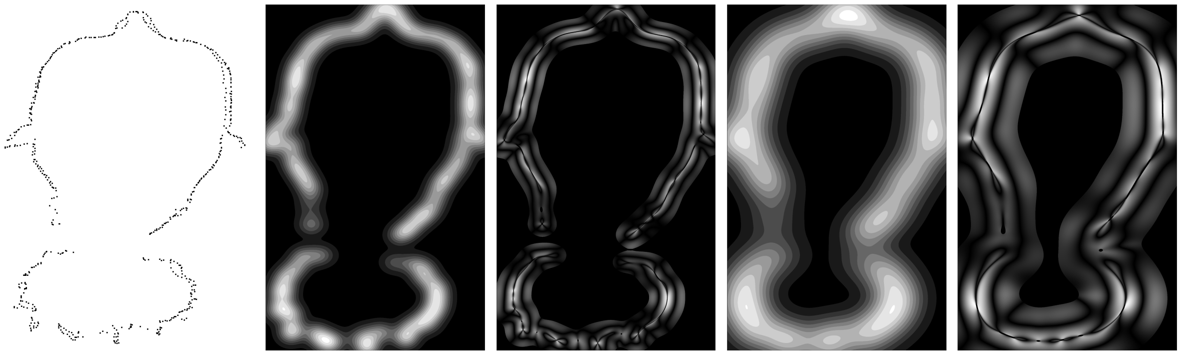


Figure 1: Crest-line criterion applied to point set (left) with isolines for comparison. Center pair shows smaller kernel diameter then right pair. Enlarge in your favourite pdf-Viewer.

applications has to be further explored, in particular for surface reconstruction, regression, and classification.

References

[1] Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(5):603–619, May 2002.

[2] M.P. do Carmo. *Differential Geometry of Curves and Surfaces.* Prentice Hall, Englewood Cliffs, New Jersey, 1976.

[3] William E. Lorensen and Harvey E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. *SIGGRAPH Comput. Graph.*, 21(4):163–169, August 1987.

[4] Olivier Monga and Serge Benayoun. Using partial derivatives of 3d images to extract typical surface features. *Comput. Vis. Image Underst.*, 61(2):171–189, March 1995.

[5] V. Pascucci and K. Cole-McLaughlin. Efficient computation of the topology of level sets. In *Proceedings of the Conference on Visualization '02*, VIS '02, pages 187–194, Washington, DC, USA, 2002. IEEE Computer Society.

[6] G. Subsol. Crest lines for curve based warping. In A. W. Toga, editor, *Brain Warping*, chapter 13, pages 225–246. Academic Press, 1998.

[7] Jean-Philippe Thirion and Alexis Gourdon. The 3D marching lines algorithm and its application to crest lines extraction. Research Report RR-1672, INRIA, 1992.

[8] Jean-Philippe Thirion and Alexis Gourdon. The marching lines algorithm : new results and proofs. Technical Report RR-1881, INRIA, 1993.

Hunting for Biological Viruses

Dominic Siedhoff
Lehrstuhl für Graphische Systeme
Technische Universität Dortmund
dominic.siedhoff@tu-dortmund.de

1 Extended Analysis Software: SynOpSis

SynOpSis (Synthesis/Optimization/Analysis) is a method for automatic detection and classification of objects in time series of images. It has been developed in the context of the *PAMONO* sensor for biological virus detection. While the full method is described in the upcoming thesis [Sie], this report focuses on the most recent advances: *SynOpSis* (cf. Figure 1) optimizes the parameters of an object detector and a classifier using synthetic data. Besides their sequential optimization using single aggregate objectives for detector and classifier, an option for global optimization has been added, optimizing both elements simultaneously. The NSGA-II algorithm [DPA+02] is used for multiobjective optimization. For the detector, Recall is maximized, and the rate of multiple responses to a single object ('M-Rate') is minimized. Precision is not optimized because the detector aims at high sensitivity, and false positives are sorted out by the classifier. Consequently, for the classifier, Precision and Recall are optimized, summing to four objectives in total. This variant of optimization will be denoted 'global 4' in the following. As Jain and Deb state [JD13], four objectives may overburden NSGA-II. The number of objectives can be reduced by multiplying the Recall functions, thus ignoring whether an object is missed by the detector or by the classifier. This variant will be denoted 'global 3', while 'sequential' will denote non-global optimization of two objectives each, for detector and classifier.

In order to focus the search in the practically relevant part of the Pareto-front, desirability functions (DFs) [TM09] are used. A DF assigns a degree of desirability in $[0, 1]$ to an objective value. Given DFs for all objectives, the desirability index (DI) [TM09] is a way of scalarizing multiple objectives, while reflecting user preferences encoded in the DFs. The desirability approach can be used in three ways: 1. The DI can identify the most desirable individual on a Pareto front, found in an optimization not using desirabilities.

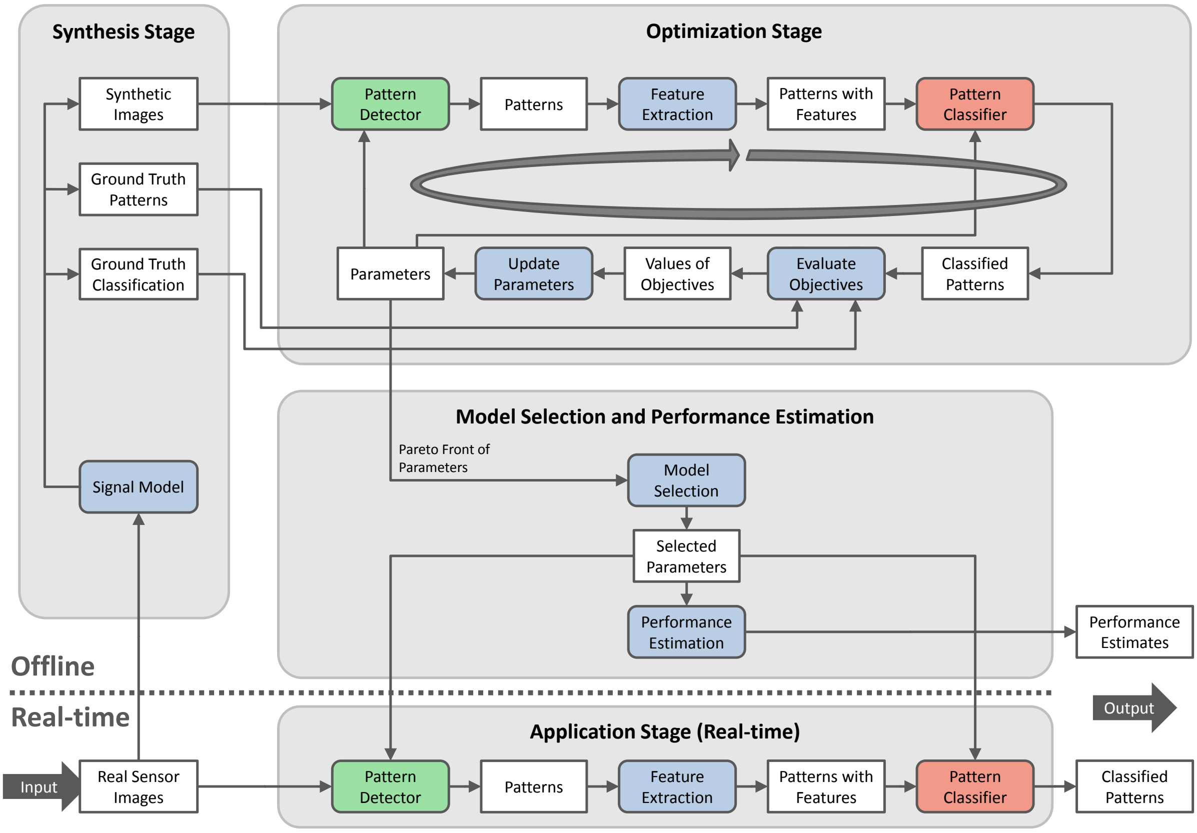


Figure 1: Scheme of the *SynOpSis* approach (Synthesis/Optimization/Analysis).

This variant is denoted 'disabled' in the following plots. 2. DFs of multiple objectives can be optimized instead of the original objectives, focusing the search in the desirable part of the front. This variant is denoted 'multiobjective'. 3. The DI of all objectives can be computed and undergo singleobjective optimization. This variant is denoted 'scalarize'.

2 Validation and Results

Figure 2(a) explores all combinations of optimization variants (sequential, global 4, global 3) with desirability variants (disabled, multiobjective, scalarize) by plotting medians and quartiles of objectives and further quality measures, as attained in the analysis of real data ('Application Stage' in Figure 1). Detector Precision is not optimized and serves to indicate the ratio of false positives the classifier has to cope with. Classifier D-Rate is not optimized and indicates the deviation between the actual number of objects and that reported by the classifier. For example D-Rate $= -0.2$ means that the number of objects was understimated by 20%. Computing the three-objective-variant and optimizing its DI ('global 3 / scalarize') performs best in terms of median D-Rate over all experiments and cross validation folds.

Figures 2(b) and (c) resolve the results obtained for ‘global 3 / scalarize’ to the experiment level, aggregating solely over cross validation folds. As expected, quality measures and their spread (as indicated by the quartiles) are better for training (b) than for real data (c), with a tendency of underestimating object counts in both. Performance is furthermore correlated with dataset difficulty: The 100 nm dataset performs worst, due to its low signal-to-noise-ratio. Performance on the 200 nm datasets is generally better, except for the low quality (LQ) sensor surface. Using a different camera (Guppy) does not affect performance much. For the medium quality (MQ) sensor surface D-Rate is paradoxically better than for the high quality (HQ) surface.

3 Conclusion and Outlook

As a summary, the performance of the *SynOpSis* approach, which optimizes parameters and learns classifying models on synthetic data, has been validated on real *PAMONO* sensor measurements. Further directions to be investigated are meta modeling [HHL11] to save optimization time and integrate optimization results of multiple experiments. This is important with regard to multiple cooperating *PAMONO* sensors, taking different measurements: A pool of optimization results stored in a meta model can be used to predict performance for candidate parameters, thus accelerating optimization. An ultimate goal is to predict parameters themselves from instance-based features. Alternatively, parameter optimization can be offloaded to powerful GPU servers via LTE.

References

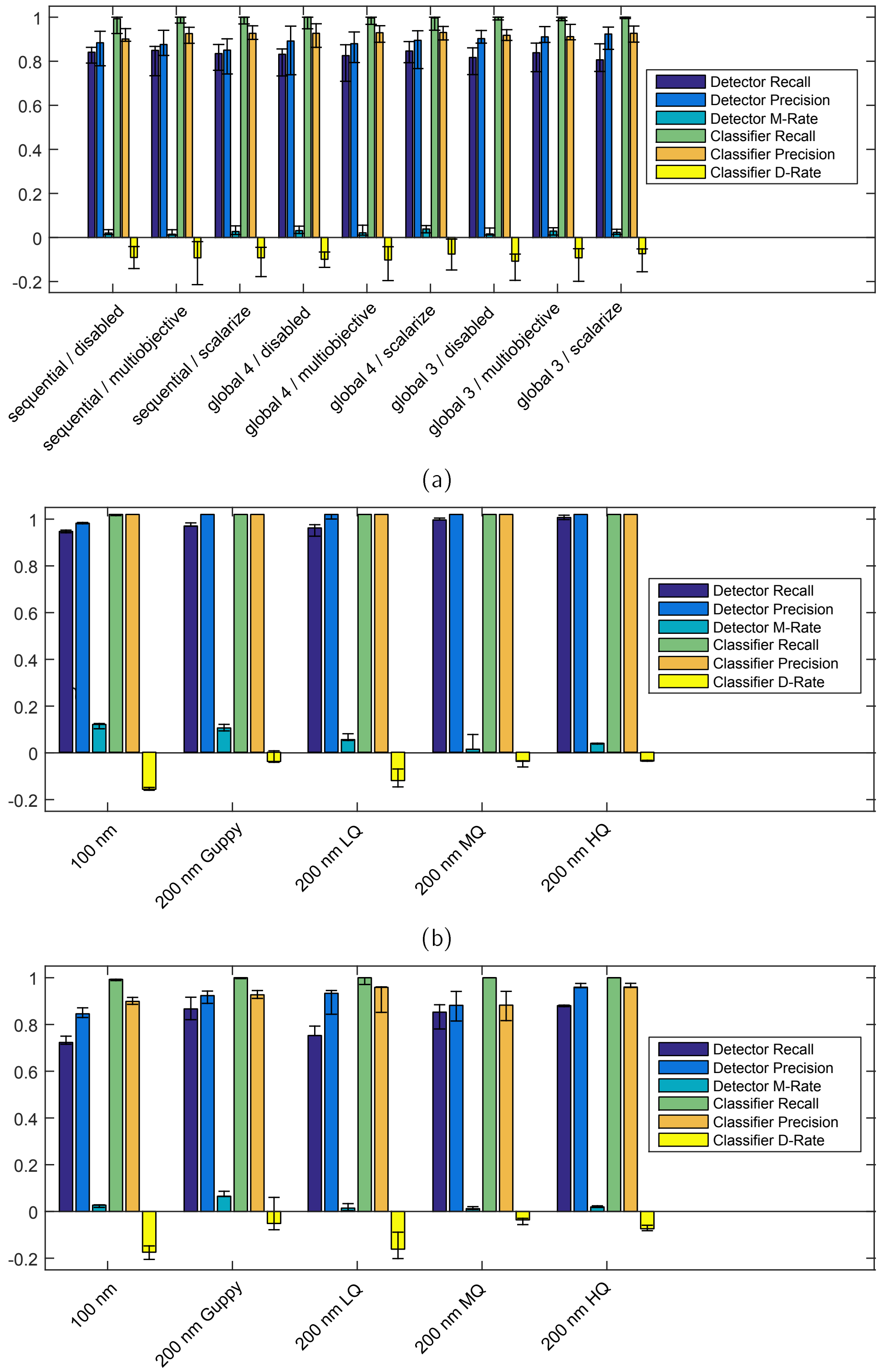
[DPA+02] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. “A fast and elitist multi-objective genetic algorithm: NSGA-II”. In: *Evolutionary Computation, IEEE Transactions on* 6.2 (2002), pp. 182–197.

[HHL11] F. Hutter, H. H. Hoos, and K. Leyton-Brown. “Sequential model-based optimization for general algorithm configuration”. In: *Learning and Intelligent Optimization*. Springer, 2011, pp. 507–523.

[JD13] H. Jain and K. Deb. “An improved adaptive approach for elitist nondominated sorting genetic algorithm for many-objective optimization”. In: *Evolutionary Multi-Criterion Optimization*. Springer. 2013, pp. 307–321.

[Sie] D. Siedhoff. *A Parameter-optimizing Model-based Approach to the Analysis of Low-SNR Image Sequences for Biological Virus Detection*. (PhD thesis in preparation).

[TM09] H. Trautmann and J. Mehnen. “Preference-based Pareto optimization in certain and noisy environments”. In: *Engineering Optimization* (2009).



(a)

(b)

(c)

Figure 2: Medians and quartiles for measures of analysis quality are shown. (a) explores all combinations of optimization and desirability variants on real sensor data. (b) and (c) show results by experiment, as obtained using the 'global 3 / scalarize' combination which performed best in (a). (b) shows results on synthetic training data, while (c) was measured on real sensor data.

SFB 876 Verfügbarkeit von
Information durch Analyse unter
Ressourcenbeschränkung

Subproject B3
Data Mining on Sensor Data of Automated Processes

Jochen Deuse Katharina Morik

Mixed-Model Assembly Line Balancing - Forming product-families to increase efficiency

Benedikt Konrad
Institut für Produktionssysteme
Professur für Arbeits- und Produktionssysteme
Technische Universität Dortmund
benedikt.konrad@ips.tu-dortmund.de

Assembly Line Balancing (ALB) approaches aim at optimally assigning tasks to workstations. While traditional ALB solution procedures lead to increasing idle times and therefore inefficiencies in high mix, low volume production settings, alternative approaches have to be designed. This paper discusses approaches of dynamic line balancing and introduces a product-resemblance driven approach to line balancing.

1 Assembly Line Balancing

The Assembly Line Balancing Problem (ALBP) describes the problem of assigning assembly operations called tasks to the stations of an assembly line. Research on ALBP can be traced back to Helgeson et al. [5] and Jackson [6] who present the first formulation of the problem setting and a first solution procedure, respectively.

In general, ALBP can be differentiated with respect to the complexity of the problem statement into Simple Assembly Line Balancing Problems (SALBP) and General Assembly Line Balancing Problems (GALBP). SALBP are focusing assembly lines on which only a single product is assembled according to a well-defined list of tasks. The corresponding assembly line consists of serially linked line station, all of which are identical, such that each task may be assigned to each station. Moreover, each assembly task's duration is deterministic and independent of the sequence of the tasks' execution. [1] If any of these conditions is relaxed, the problem is called a GALBP. Most relevant in terms of practical application of ALBP is its extension to thsoe assembly lines on which a wide variety of different product or product variants are produced. These types of GALBP are referred to as Multi Model Assembly Line Balancing (MuMoALBP) or Mixed Model Assembly Line Balancing (MiMoALBP). While in MuMoALBP assembly is organized in batches with a batch consisting of identical products, MiMoALBP allows to assembly all different products or product variants in an arbitrary order. [1]

In order to solve MiMoALBP it is common practice to reduce the problem to the Single Model Assembly Line Balancing Problem or if passible to SALBP. This is achieved by combining all different products or variants. In line balancing products are represented by their precedence graph. A precedence graph is a acyclic directed graph in which tasks are related to the vertices and precedence constraints are represented by the set of edges. Consequently, combining the different products requires deriving the joint precedence graph of all products. In order to account for each product's individual demand, tasks are weighted according to the demand. [2] By this, the variety of different products is reduced to one averaged one. Once the mixed model problem is reduced to the single model one, solution approaches of the latter are applied. A comprehensive review of these can be found in [1].

Although this approach allows for efficiently solving the problem stated, the line balance computed is only efficient if customer orders are similar to the average product. As customer demand tends to be erratic and varies significantly according to volume as well as mix, the computed line balance might be almost arbitrarily efficient or inefficient. The following approaches were developed address this shortfall.

2 Dynamic Assembly Line Balancing

Ostolaza et al. [8] as well as Gel et al. [4] present approaches for a dynamic approach to assembly line balancing. To increase worker flexibility beyond common approaches such as parallel work stations or floating workers, tasks assigned to a station are split into fixed and shared ones. Fixed tasks are imperatively conducted by the operator of the very station. In contrast, shared tasks may be executed by an operator from a neighboring station who has already completed all tasks assigned to his own station. Using this approach, overload situations can easily be accounted for.

While this approach concentrates on worker flexibility, the work of Reinhart and Proepster [10] addresses flexibility in planning and organization of assembly lines. To account for variability in customer demand and product dependent workloads, the authors propose a scenario driven approach with each predefined scenario representing a specific type-volume-mix of demand. To further increase flexibility, variant-specific work content is assigned to dedicated highly variable work stations. Consequently, changes in variant-specific tasks only affect few stations that can be adapted effortlessly.

In contrast to the approaches introduced previously, Deuse et al. [3] propose to combine MiMoALBP and MuMoALBP to add to balancing efficiency. The underlying principle is to analyze product and process related data in order to define homogenous product families for which line balancing can be computed. Instead of creating a balance for all products (MiMoALBP) family-specific balances are derived (MuMoALBP) with each family-specific being a MiMoALBP.

3 Assessing Product Resemblance in Assembly Lines

A prerequisite for this approach is to identify and assess product- and process-induced variability that leads to efficiency losses. In general, different products or product variants might require among others individual assembly tasks, different work contents of each task, precedence restrictions of tasks, station equipment, material and parts or operator qualifications. Consequently, homogeneous product families contain products with resemblance in these characteristics.

In case of required tasks, equipment or material, structural resemblance of different products is defined by the presence or absence of each characteristic, i. e. homogeneity is detected by an interpretation of binary data. On the contrary, work content is measured in time units, which adds ratio scaled variables to similarity analysis. Moreover, execution sequences for tasks is information that is included in precedence graphs and therefore cannot be evaluated using one of the afore-mentioned approaches. To access this information in family formation feature extraction approaches for graphs or trees have to be employed. Among others, tree edit distance [7] and graph histograms [9] are examples of techniques to define graph similarity which are to be evaluated.

The examples given line out, that assessing product resemblance in assembly lines is not trivial due to the variety of differently scaled characteristics to be included. Besides this, the relative importance of characteristics for product resemblance has to be considered. For instance, the presence of a subgroup of tasks may be more relevant for grouping than the remaining tasks, similar contents might be more desirable than similar tasks or presence of similar tasks might be important, while precedence restrictions might be less relevant. That is to say, a weighting approach has to be incorporated into family formation which has to be conducted in light of the individual use case. Once the data is preprocessed, families can be formed by means of unsupervised learning and are passed over as input to assembly line balancing.

4 Conclusion

Identifying homogeneous product families poses a foundation to improve line balancing in case of high mix, low volume assembly line production. It was lined out that those characteristics have to be identified which cause efficiency losses in balancing as a first step. Consequently, resemblance of product and product variants can be analyzed based on these criteria. Finally, for those product-families established ALB can be conducted.

References

[1] Ilker Baybars. A survey of exact algorithms for the simple assembly line balancing problem. *Management Science*, 32(8):909–932, 1986.

[2] Nils Boysen, Malte Fliedner, and Armin Scholl. Assembly line balancing: Joint precedence graphs under high product variety. *IIE Transactions*, 41(3):183–193, 2009.

[3] Jochen Deuse, Fabian Bohnen, and Benedikt Konrad. Renaissance der gruppentechnologie. *Zeitschrift für wirtschaftlichen Fabrikbetrieb*, 106(5):337–341, 2011.

[4] Esma S. Gel, Wallace J. Hopp, and Mark P. van Oyen. Factors affecting opportunity of worksharing as a dynamic line balancing mechanism. *IIE Transactions*, 34(10):847–863, 2002.

[5] W. B. Helgeson, M. E. Salveson, and W. W. Smith. How to balance an assembly line. *Management Report*, 7, 1954.

[6] James R. Jackson. A computing procedure for a line balancing problem. *Management Science*, 2(3):261–271, 1956.

[7] Shin-Yee Lu. A tree-to-tree distance and its application to cluster analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):219–224, 1979.

[8] Jose Ostolaza, L. Joseph Thomas, and John O. McClain. The use of dynamic (state-dependent) assembly-line balancing to improve throughput. *Cornell University, Johnson Graduate School of Management*, 89(5), 1989.

[9] A. N. Papadopoulos and Y. Manolopoulos. Structure-based similarity search with graph histograms. In *Proceedings of the 10th International Workshop on Database and Expert Systems Applications*, pages 174–178, 1999.

[10] Gunther Reinhart and Markus Pröpster. Handlungsfelder der dynamischen austaktung: Ein prozess zur reaktion auf nachfrageschwankungen im nutzfahrzeugbau. *Zeitschrift für wirtschaftlichen Fabrikbetrieb*, 107(6):404–408, 2012.

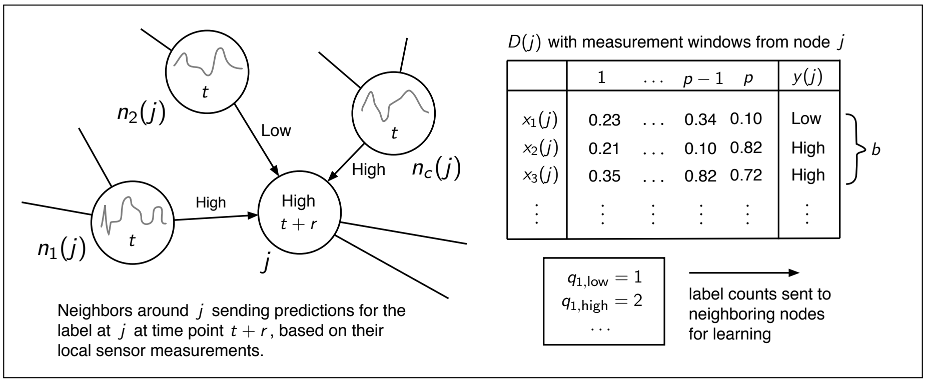
Communication-efficient distributed learning of spatio-temporal local models

Marco Stolpe
Lehrstuhl für Künstliche Intelligenz, LS 8
Technische Universität Dortmund
marco.stolpe@tu-dortmund.de

Advances in hardware technology have led to a proliferation of things getting connected to global networks. The network of all connected objects together is called the *Internet of Things (IoT)*. Today, data generated by such objects is transmitted to central cloud systems. However, experts estimate that by 2020, the IoT will consist of almost 50 billion objects. In his ground breaking ubiquitous computing white paper, Mark Weiser even spoke of hundreds connected small devices in a single room. With so many connected participants, transmitting all data to a central location is no longer scalable. This report presents a scalable privacy-preserving decentralized in-network algorithm that exchanges space-time aggregated values with a restricted number of topologically close neighboring nodes. The algorithm's performance and communication costs are discussed in terms of traffic flow prediction in the city of Dublin.

1 Introduction

We assume a network of m nodes $j = 1, \dots, m$ that each deliver an infinite series of real-valued raw measurements, assuming a constant sample rate. Each sensor has a spatial location. In many applications, decisions are based on the prediction of discrete categories. Further given is therefore a mapping $d : \mathbb{R} \rightarrow Y$ of raw measurements to categories $Y = \{Y_1, \dots, Y_l\}$. Each node j provides a set $D(j)$ for supervised offline learning,


$n_2(j)$
t
Low
t
High
$n_c(j)$
High
$n_1(j)$
t
High
$t+r$
j
Neighbors around j sending predictions for the label at j at time point $t+r$, based on their local sensor measurements.
$D(j)$ with measurement windows from node j
b
$q_{1,\text{low}} = 1$
$q_{1,\text{high}} = 2$
$\dots$
label counts sent to neighboring nodes for learning

	1	$\dots$	$p-1$	p	$y(j)$
$x_1(j)$	0.23	$\dots$	0.34	0.10	Low
$x_2(j)$	0.21	$\dots$	0.10	0.82	High
$x_3(j)$	0.35	$\dots$	0.82	0.72	High
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Figure 1: Distributed Learning of Local Models

containing n pairs $(x_i(j), y_i(j))$ of training examples $x_i(j) \in \mathbb{R}^p$ and labels $y_i(j) \in Y$. Each $x_i(j)$ is created by sliding a window of size p with step size 1 over the stream of measurements at node j. When recording from time step s, observations $x_i(j)$ are windows of measurements and labels y_i are discretized measurements r time steps ahead.

Learning is restricted to j and c neighboring nodes with indices $n_1(j), \dots, n_c(j)$. Based on the datasets at j and its neighbors, we want to learn a *local* model $f(j)$ that, given current windows $x(j), x(n_1(j)), \dots, x(n_c(j))$ of sensor readings, predicts the category $y(j)$ at node j with horizon r correctly. This situation is depicted in Fig. 1. The data is *vertically partitioned*, since each neighboring node of j only stores partial information about x, i.e. a subset of features.

Few distributed algorithms learning from vertically partitioned data are known. In the following, a decentralized in-network algorithm is presented that exchanges only aggregate label information with a restricted set of topologically close neighboring nodes.

2 Distributed Learning of Spatio-Temporal Local Models

The learning task described can be solved in at least two different ways. One way is to send measurements from neighbors to j, combine the according windows with j's labels and, based on this data, learn $f(j)$ at j. Another way is to combine windowed measurements at each neighbor with labels from j, i.e. $D_j(k) = \{(x_i(k), y_i(j))\}_{i=1,\dots,n}$, and learn models $f_j(k)$ at nodes $k = j, n_1(j), \dots, n_c(j)$ to predict $y(j)$. Model $f(j)$ could then be, for instance, a majority vote over predictions from j and its neighbors. The first approach may respect joint dependencies, but due to sending raw measurements isn't privacy-preserving. The latter approach preserves privacy by a discretization of values and saves communication by encoding the data in less bits.

Privacy and communication can be further improved. Given a partitioning of observations $x_1, \ldots, x_n$ into batches B_u, $u = 1, \ldots, h$ and label proportions π_{uv} for each batch u and class Y_v, algorithms for *learning from label proportions* learn model $f : X \rightarrow Y$ that assigns labels to individual observations. Instead of sending all labels to neighbors, only the counts of labels need to be sent. A simple partitioning of the data into b-sized batches is a division over consecutive time intervals. Node j counts how often each class occurs in each batch and sends these counts in a $h \times l$ matrix $Q(j)$, $h = \lceil n/b \rceil$, to its neighboring nodes. These transform $Q(j)$ into a label proportion matrix $\Pi(j)$, yielding the original problem of learning from label proportions at neighboring nodes of j.

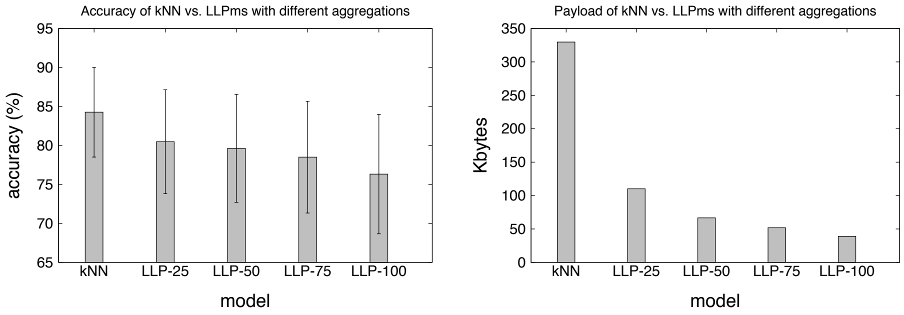
Several algorithms for learning from label proportions are referenced in [1]. The LLP algorithm [2] fits a constraint scenario and can handle multiple classes. LLP first clusters all observations ($k \geq |Y|$) and then minimizes the mean squared error (MSE) between the given label proportions and those as calculated by different label assignments to clusters. For performance reasons, attribute weights are not optimized as in the original paper. Further, the exhaustive labeling strategy can be replaced by a more efficient local search with a multistart strategy. This modified version is called LLP$_{\text{ms}}$. The distributed learning algorithm works as follows:

```
For $j$ = 1 to $m$ do        /* in parallel */
        divide $D(j)$ into batches $B_1, \ldots, B_h$
        calculate label counts for each batch and store them in $Q(j)$
        send $Q(j)$ to nodes $n_1(j), \ldots, n_c(j)$
        For $k = j, n_1(j), \ldots, n_c(j)$ do        /* in parallel */
            calculate $\Pi(j)$ from $Q(j)$
            train LLP$_{\text{ms}}$ model $f_j(k)$ at node $k$
```

Each node stores $c+1$ different models, for itself and each of its neighbors. All models are *local* in the sense that learning is restricted to local neighborhoods. Moreover, the algorithm works fully *in-network*, as no central coordinator is needed for local synchronization and learning between peer nodes.

3 Evaluation

The algorithm has been introduced and evaluated in the context of traffic flow prediction [3] at junctions in the city of Dublin. LLP$_{\text{ms}}$, kNN and STRF [4] baseline models were trained at 296 different sensor locations (junctions) across the city, with a prediction horizon of 15 minutes. kNN models outperform the STRF, achieving an average accuracy of 85.7% over all junctions, whereas the STRF yields only 78.1%. Figure 2 shows the trade-off between accuracy and payload sent for kNN and LLP$_{\text{ms}}$ with different amounts



(a) Accuracy

(b) Payload

Figure 2: Trade-off between accuracy and payload sent for kNN and $\mathrm{LLP}_{\mathrm{ms}}$

of aggregation. Error increases with higher aggregation, but communication costs decrease, by factors 3, 5 and 8.5. For $b = 75$, the accuracy is still in the order achieved by the more complex STRF model, though much less data needs to be communicated.

4 Conclusions

We have presented a privacy-preserving and communication-efficient distributed in-network algorithm for spatio-temporal learning. Future work will deal with questions related to streaming data and concept drift. Moreover, the algorithm's sensitivity to changing the number of neighbors or other parameters needs to be investigated. Finally, the approach needs to be evaluated in the context of other applications.

References

[1] G. Patrini, R. Nock, T. Caetano, and P. Rivera. (Almost) no label no cry. In *NIPS 27*, pages 190–198. Curran Associates, Inc., 2014.

[2] M. Stolpe and K. Morik. Learning from label proportions by optimizing cluster model selection. In *Proc. of ECML/PKDD*, pages 349–364. Springer, 2011.

[3] M. Stolpe, T. Liebig, and K. Morik. Communication-efficient learning of traffic flow in a network of wireless presence sensors. In *Proc. of the Workshop on Parallel and Distributed Computing for Knowledge Discovery in Data Bases (PDCKDD)*, CEUR Workshop Proceedings, page (to appear). CEUR-WS, 2015.

[4] N. Piatkowski, S. Lee, and K. Morik. Spatio-temporal random fields: compressible representation and distributed estimation. *Machine Learning*, 93(1):115–139, 2013.

Collaborative Research Center SFB 876 - Using Material Forming Simulation for Advanced Process Control

Mario Wiegand
Institut für Produktionssysteme
Technische Universität Dortmund
mario.wiegand@ips.tu-dortmund.de

This report gives a brief overview of the central aspects of using material forming simulation for the development of an advanced process control system in interlinked production processes. The simulation based on the finite element method is deployed to determine appropriate adjustments of process parameters in order to remedy product defects during operation.

1 Introduction

Within the Collaborative Research Center 876 project B3 targets the time-constrained analysis of sensor data by means of data mining. The analysis of recorded data is conducted using the example of hot rolling processes in the steel industry. The overall objective is the prediction of the final quality of steel bars on the basis of process data (e.g. rolling temperature, rotation speed) [2]. In the first period the prediction of a binary label (OK/not OK) was focused. The label represented an aggregated value of product quality as a result of an ultrasonic test at the end of the production line. The prediction model allowed for an early forecast as to whether the product currently manufactured meets quality requirements. Therefore, it could be decided at an early stage if the product in question should be ejected from process chain or if the production process should be continued. The process control decision helps to reduce waste of energy and of other resources [3].

In the second period the process control system is extended allowing for the adaptation of process parameters on the basis of detailed quality predictions [1]. With it product defects can be remedied during operation by changing the parameters. In order to determine appropriate parameter adjustments experiments in the real process are not possible due to high costs. Instead material forming simulations based on the finite element method (FEM) are deployed.

In addition to the simulation of feasible parameter adaptions the FEM simulation can be used to improve the decision boundary of the prediction model. In industrial manufacturing processes the data usually has imbalanced class distributions because of the imbalance between products with high and those with low quality. Therefore, the number of positive examples greatly exceeds the number of negative examples. Thus, simulation experiments can be deployed to specifically generate negative examples and analyze critical areas of the process parameters.

2 Material forming simulation with finite elements

In general FEM is a numerical procedure for solving partial differential equations approximately. It is widely used in engineering, especially in solid and structural mechanics. In the FEM the solution region is divided into an arbitrary number of small subdivisions, called finite elements. The elements are interconnected at nodes lying on the elements boundaries [4]. This way a complex geometrical shape can be approximated by an assemblage of finite elements. The behavior of the field variable of interest (e.g. displacement, pressure) inside a finite element can be approximately computed using a simple function [4]. Thus, a complex continuous problem can be replaced by a simpler one by discretization [5].

As outlined in the introductory part, in context of hot rolling processes FEM simulations can be used to simulate the metal forming process of steel bars in order to analyze critical areas of process parameters that have not be recorded in the real process. As a result, experiments for finding feasible parameter adaptions to realize an advanced process control system can be conducted. Figure 1 gives an overview on the central elements of this control system. Besides, negative examples can be generated specifically in order to improve the decision boundary of the prediction model.

FEM simulations require relatively high computation times due to the mathematical complexity of finding an approximate solution numerically. Accordingly, it can take several hours to simulate a single processing step. That is why an integration of the simulations into the online process control system is not possible. Instead simulations have to be conducted offline.

The FEM simulation needs a variety of product information as well as details on the rolling plant as input. Therefore, we mapped geometrical information on the shape and

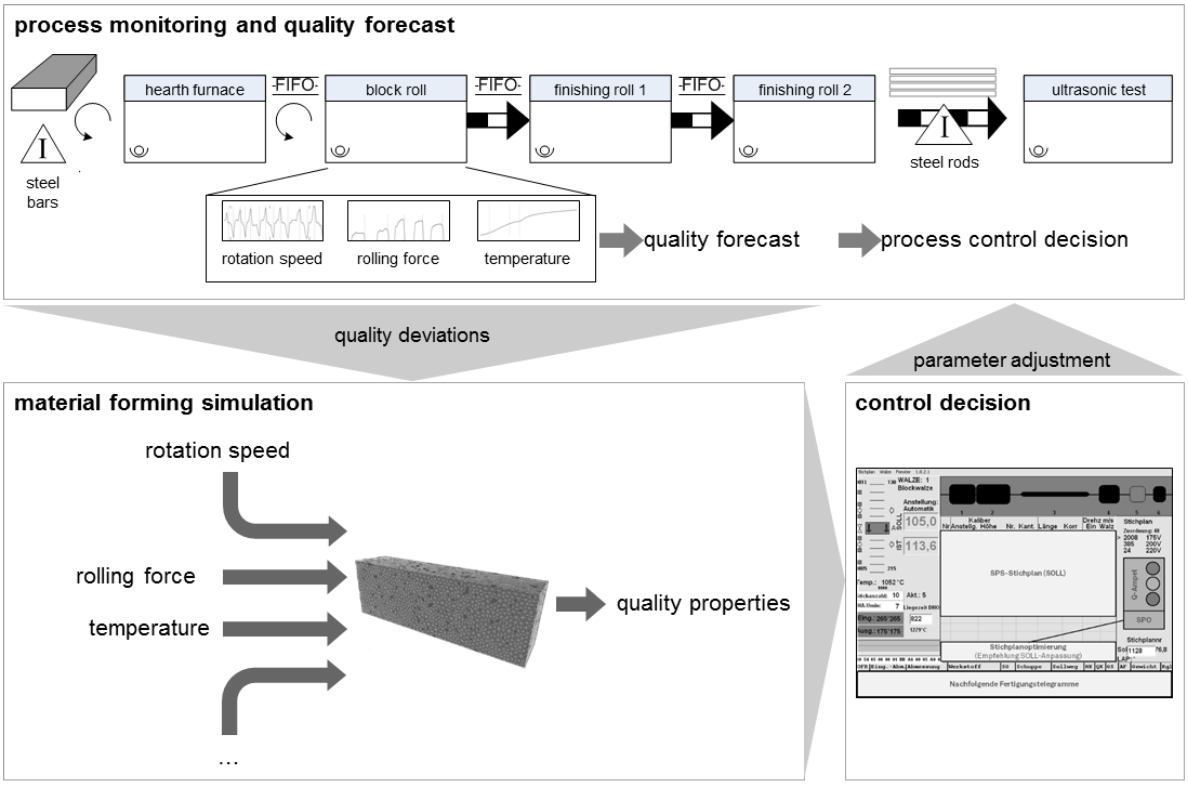


Figure 1: Using simulation for advanced process control

the dimensions of the steel bars together with three-dimensional models of the rolling plant in the FEM simulation. Furthermore, characteristics of the mechanical behavior of the material, like the elastic modulus, flow curves and isothermal transformation diagrams had to be modeled. In addition to this, a feasible representation of process parameters had to be found in order to map the recorded sensor data of the real process.

One of the most common product defects emerging in hot rolling processes are so called blowholes, voids or cavities. Depending on the concrete values of process parameters blowholes in the steel bars are reshaped and changed in size during hot rolling leading to different quality properties. Therefore, the analysis of the evolution of blowholes during hot rolling is of special interest. As output of the simulation information on the position, the size and the shape of internal material defects like blowholes can be obtained. This allows for generating a binary as well as a more detailed label. That way a more advanced forecast of product quality can be produced facilitating more complex process control decision like adjustments of process parameters to increase product quality.

Due to the computational complexity of the simulation the individual experiments can be very time consuming. Thus, only selected parameter combinations can be tested. For that reason a procedure reducing the number of simulation runs is developed.

3 Conclusion

The presented approach of using FEM simulation for testing critical parameter combinations allows for the development of an extended process control system. As a result process parameters can be adjusted online so that defects of intermediate products can be remedied during operation. Furthermore, the decision boundary can be improved by creating additional examples for particularly relevant parameter ranges.

References

[1] Jochen Deuse, Mario Wiegand, Loga Erohin, Daniel Lieber, and Ralf Klinkenberg. Big Data Analytcis in Produktion und Instandhaltung. In Hubert Biedermann, editor, *Instandhaltung im Wandel. Herausforderungen und Lösungen im Zeitalter von Industrie 4.0*, pages 33–48. TÜV Media, 2014.

[2] Benedikt Konrad, Daniel Lieber, and Jochen Deuse. Striving for zero defect production: Intelligent manufacturing control through data mining in continous rolling mill processes. In Katja Windt, editor, *Robust Manufacturing Control*, volume 1 of *Lecture Notes in Production Engineering*. Springer, 2012. Accepted for Publication.

[3] Daniel Lieber, Benedikt Konrad, Jochen Deuse, Marco Stolpe, and Katharina Morik. Sustainable interlinked manufacturing processes through real-time quality prediction. In *Leveraging Technology for a Sustainable World*, Proceedings of the 19th CIRP Conference on Life Cycle Engineering, pages 393–398. Springer, 2012.

[4] Singiresu S. Rao. The Finite Element Method in Engineering. Elsevier, 4. edition, 2005.

[5] Robert L. Taylor, Jianzhong Zhu, and Olgierd C. Zienkiewicz. The finite element method: its basis and fundamentals. Elsevier, 6. edition, 2005.

SFB 876 Verfügbarkeit von
Information durch Analyse unter
Ressourcenbeschränkung

Subproject B4
Analysis and Communication for dynamic traffic prognosis

Christian Wietfeld Michael Schreckenberg
Kristian Kersting

Prediction of Vehicular Traffic Flows with Poisson Dependency Networks

Lars Habel
Physik von Transport und Verkehr
Universität Duisburg-Essen
lars.habel@uni-due.de

This report sums up our recent research to provide better real-world input data for microscopic traffic simulations. Empirical traffic data often contains missing values. To close these data gaps, we compare a temporal approach based on exponential smoothing of historical data with a data-driven approach based on the newly proposed Poisson Dependency Networks.

1 Introduction

Microscopic road traffic simulations based on a real-world topology can be used to set up a traffic information system. Then, vehicular traffic data from real-world detectors is needed to reproduce all the recent traffic in- and outflows of the real-world system, for example at on- and off-ramps. Usually, the simulation queries new empirical data at run time and thus the permanent availability of empirical data is necessary. Unfortunately, the reliability of empirical detectors is often not good enough to ensure this requirement minute by minute. For complex topologies, this means that the simulation results then not only depend on the quality of the simulation model and the topology representation, but also on a possibly huge number of empirical traffic detectors.

Recently, we evaluated two approaches to close the resulting gaps in empirical data, one working on temporal level, the other one on level of dependencies between multiple detectors [3, 5].

2 Methods

Exponential Method The temporal approach [1] is based on exponential smoothing a set $\mathbf{j}$ of historical traffic flows. $\mathbf{j}$ comprises previously collected traffic flows from up to 30 timestamps t measured at the particular detector, which are chosen by a clustering algorithm that distinguishes between different weekdays, school holidays and public holidays. The predicted flow j_t^* is then obtained by

$$j_t^* = \alpha j_t + \alpha \sum_{i=1}^{t-1} (1-\alpha)^i j_{t-i} + (1-\alpha)^t j_0 \ , \tag{1}$$

where j_t is the most recent historical traffic flow and $\alpha = 0.8$ according to [2].

Poisson Dependency Network (PDN) For the dependency-based gap filling, we use the recently proposed Poisson Dependency Networks [4], where each node represents a single detector and each edge between nodes describes dependencies between them. Note that neighbouring detectors on the road do not have to be strongly connected in the PDN.

Here, the set $\mathbf{j}$ comprises traffic flows from other detectors, but measured at the same time. The probability function to obtain a traffic flow for detector a given all the other flows $\mathbf{j}_{\setminus a} = \mathbf{j} \setminus \mathbf{j}_a$ at that time is then denoted as

$$p(j_a|\mathbf{j}_{\setminus a}) = \frac{\lambda_a^{j_a}(\mathbf{j}_{\setminus a})}{j_a!} e^{-\lambda_a(\mathrm{j}_{\setminus a})} \ , \tag{2}$$

where $\lambda_a(\mathbf{j}_{\setminus a})$ is a function which contains all knowledge about correlations between detector a and the others. In this contribution, each λ is modelled by Poisson regression trees which have been learned by the `R`-package `rpart`.

3 Comparisons

We applied both prediction approaches to two different data sets using different learning schemes. For both comparisons, we used empirical traffic data from the Cologne orbital motorway network, which is formed by the motorways A1, A3 and A4 and is about 100 km long. Traffic data is provided by 187 detectors at 95 cross-sections.

Both approaches use historical traffic data in a certain sense. In the Exponential Method, there is usually a window of 1 week between each timestamp t. The PDN was learned

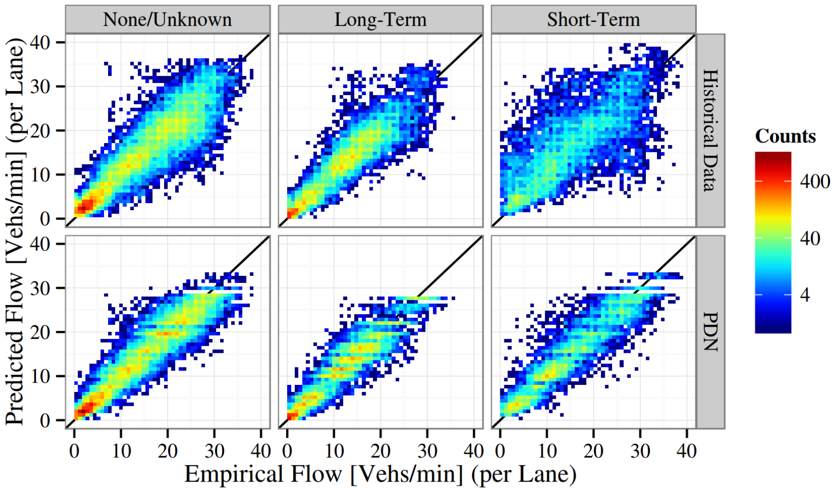


Figure 1: Comparison between empirical and predicted vehicular traffic flows (per lane) on 26/08/2014 characterised by TMC event type. First row: Exponential Method, Second row: PDN. From [5].

in two ways: In [5], a simple leave-one-out strategy was applied to traffic data from a whole day. Differences between the predictions are shown in Fig. 1.

In [3], a more realistic learning approach based on a moving 60 min-window of historical traffic data from the preceding week was used (see [3] for a detailed explanation) to predict traffic flows from the test week. The overall prediction accuracies are shown in Table 1. There, every timestamp of the whole test week is included. Both prediction methods are not faultless, because both of them basically perform a 1-week-prediction of 1 min-count data. However, positive and negative differences often balance each other out from minute to minute.

Table 1: Overall prediction accuracy. From [3].

method	RMSE [vehs/min]	NRMSE [%]
Exponential	4.93	53.2
PDN	**4.50**	**48.6**

The spatial visualisation in Fig. 2 reveals that prediction problems typically are bound to topological problems like large construction sites, which are heavily affected by congestion because of their huge bottleneck impact. They can be identified in Fig. 2 by the size of the dots, which denote the mean empirical velocity at test time. The predictions from the PDN clearly benefit from the learned dependencies, although it was trained with historical data as well.

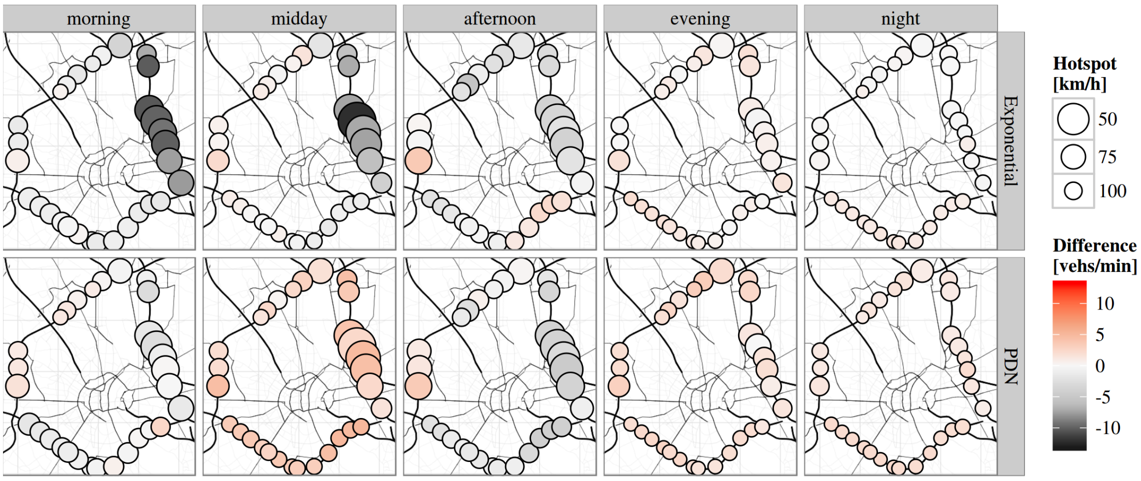


Figure 2: Spatial visualisation of differences between predicted and observed traffic flows between 21/09/2015 and 27/09/2015. Each dot represents a detector cross-section in anti-clockwise driving direction. First row: Exponential Method, Second row: PDN. From [3].

References

[1] Roland Chrobok, Oliver Kaumann, Joachim Wahle, and Michael Schreckenberg. Three categories of traffic data: Historical, current, and predictive. In E. Schnieder and Udo Becker, editors, *Proc. of 9th IFAC Symposium Control in Transportation Systems*, pages 250–255. Pergamon, 2001.

[2] Roland Chrobok, Oliver Kaumann, Joachim Wahle, and Michael Schreckenberg. Different methods of traffic forecast based on real data. *Eur. J. Oper. Res.*, 155(3):558–568, 2004.

[3] Lars Habel, Alejandro Molina, Thomas Zaksek, Kristian Kersting, and Michael Schreckenberg. Traffic Simulations With Empirical Data – How To Replace Missing Traffic Flows? In *Traffic and Granular Flow '15*. Springer, in press.

[4] Fabian Hadiji, Alejandro Molina, Sriraam Natarajan, and Kristian Kersting. Poisson Dependency Networks: Gradient Boosted Models for Multivariate Count Data. *Mach. Learn.*, 100(2-3):477–507, 2015.

[5] Christoph Ide, Fabian Hadiji, Lars Habel, Alejandro Molina, Thomas Zaksek, Michael Schreckenberg, Kristian Kersting, and Christian Wietfeld. LTE Connectivity and Vehicular Traffic Prediction based on Machine Learning Approaches. In *2015 IEEE Vehicular Technology Conference (VTC Fall)*, 2015.

Cellular Connectivity Prediction for Resource-Efficient Vehicular Applications

Christoph Ide
Lehrstuhl für Kommunikationsnetze
Technische Universität Dortmund
christoph.ide@tu-dortmund.de

Many vehicular applications in the context of Cyber-Physical Systems (CPS) can be served by cellular communication systems. The additional data traffic has to be transmitted very efficiently to minimize the interdependence with Human-to-Human (H2H) communication. In this report, a forecasting approach for cellular connectivity based machine learning methods to enable a resource-efficient communication for vehicular applications is presented. The results based on massive measurement data show that the cellular connectivity can be predicted with a probability of up to 78 %. Regarding a mobile communication system, a predictive channel-aware transmission based on machine learning methods enables a gain of 33 % concerning the spectral resource utilization of an Long Term Evolution (LTE) system.

1 Predictive Channel-Aware Transmission (pCAT) based on Data Mining

A data transmission in a LTE mobile communication system at comparatively good channel quality can be performed very efficiently due to the possibility to use high modulation and coding schemes. For this purpose, it is possible to delay sending processes of non-time-critical vehicular data to moments of favorable channel conditions of the underlying mobile communication channel as then the limited radio resources can be used more efficiently (cf. [1]). Formally expressed, particularly non-time-critical background-data are considered a sending decision that can depend on a transmission probability $p_T(t)$. It implies the current Signal-to-Interference-plus-Noise Ratio $SINR(t)$ and a connectivity forecast $\Delta SINR(t)$. After a period of no transmission $(t - t' \leq t_{min})$ continuously a transmission probability $p_T(t)$ as a function of current and future connectivity is computed. In case of success, a transmission proceeds and pCAT returns to the state of no transmission. If no transmission has proceeded until an upper boundary period $(t - t' \leq t_{max})$ is reached a transmission is proceeded immediately and pCAT returns to the state of no transmission. Details regarding the mathematical description of pCAT can be found in [2].

When applying recent past-values on the trained classification model, the mean connectivity for a future time frame is classified. This connectivity forecast based on data mining feeds the predictive component of a transmission probability in pCAT-DM [3]. A data mining process extracts knowledge from the signal course of a recent past time frame τ_d described by representative attributes (data reduction). Dependent on decision

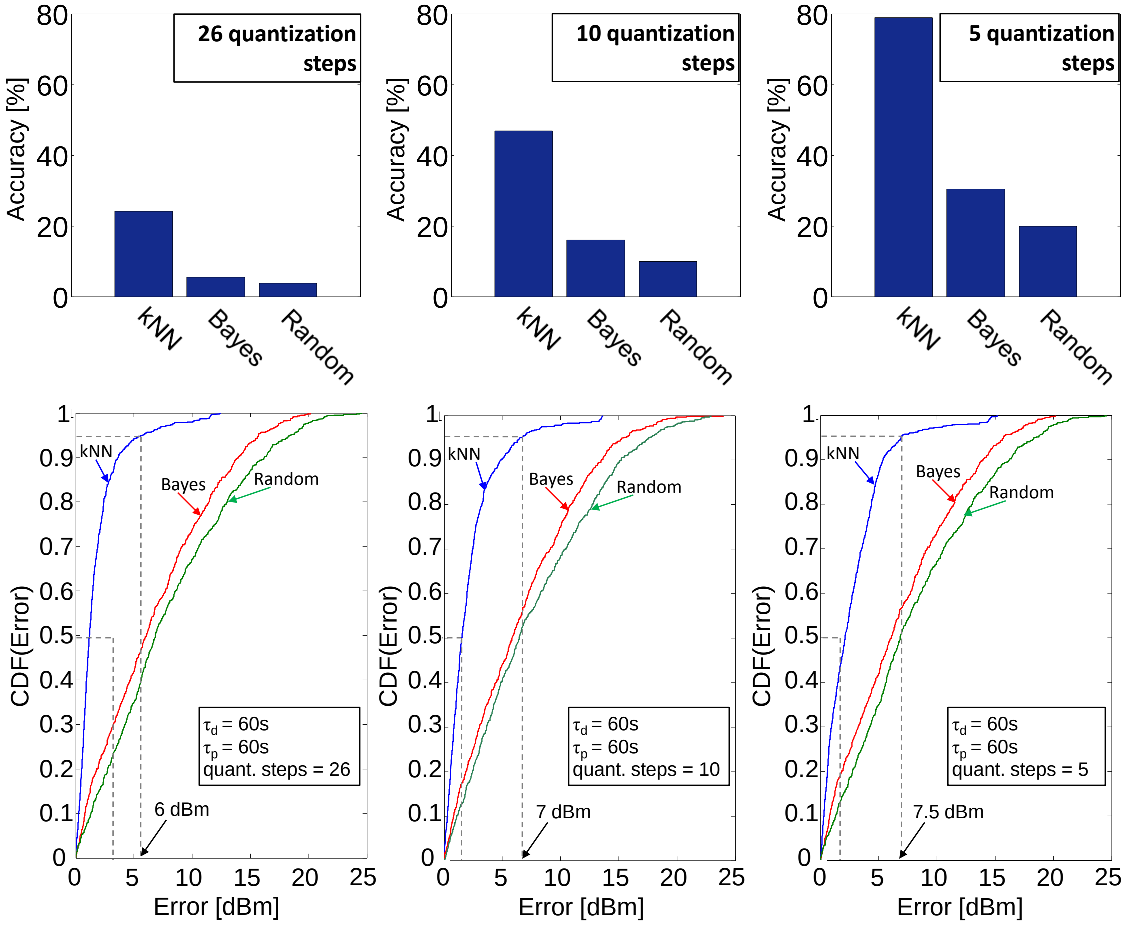
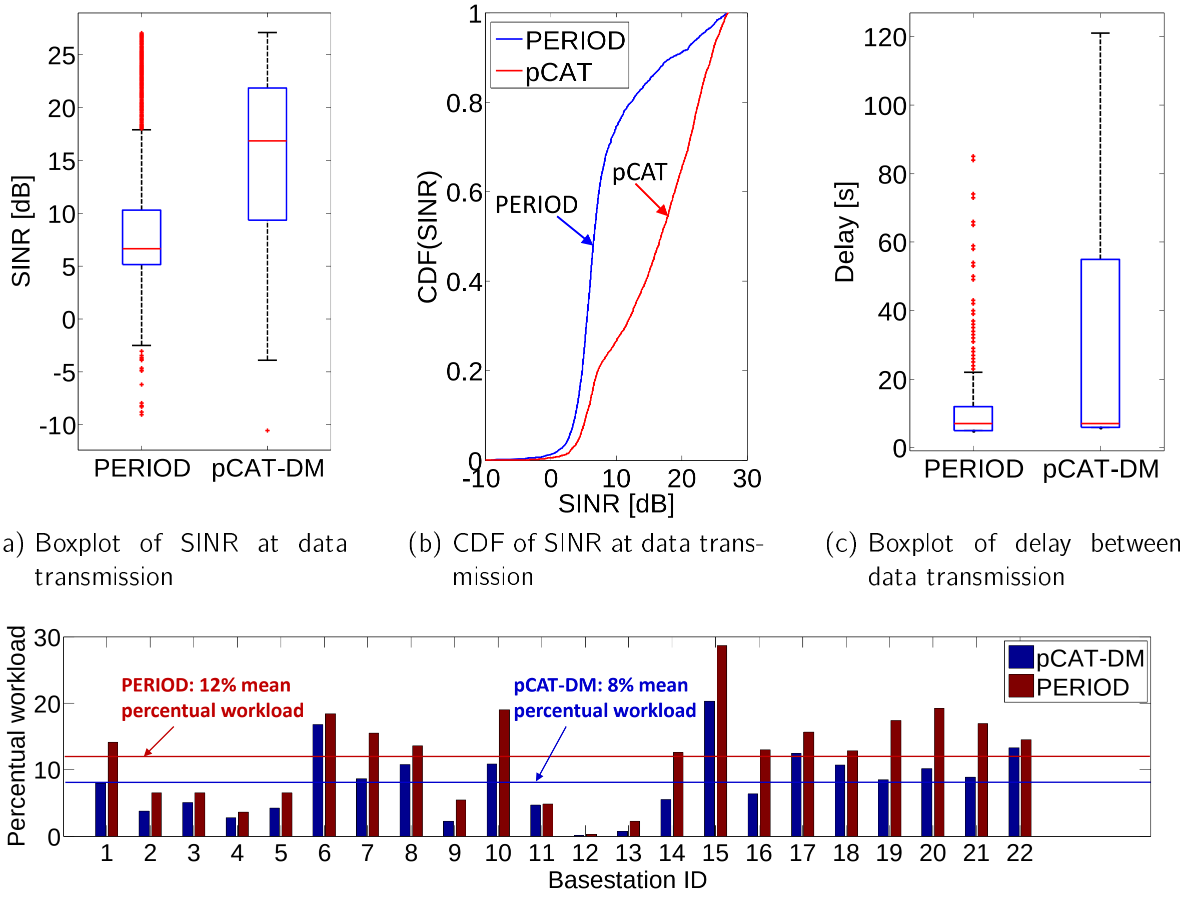


Figure 1: Classification accuracy as a function of the number of quantization steps and prediction error based on experimental data of a real UMTS network

thresholds, a k-Nearest-Neighbor (kNN) classifier predicts a mean future connectivity represented by a label in time intervals of one second. In consideration of prerequisite knowledge about any possible state of the connectivity in a represented system, e.g., influence of pathloss, shadowing, noise and interference, attributes such as current, average (mu), variance (var), min, max and gradient (mReg) values are selected. In time intervals of one second descriptive attributes are combined with a predictive label and form a data mining example. As a label represents a future change of connectivity, the computation takes a logarithmic quantization into account to converge a uniform distribution of labels, as small future changes appear exponentially more often than large changes. More details regarding the methodology for the performance evaluation can be found in [3]. In addition, in [4], general requirements for LTE-based vehicular communication and in [5], an approach for the cellular connectivity prediction based on road traffic data are presented.



(a) Boxplot of SINR at data transmission

(b) CDF of SINR at data transmission

(c) Boxplot of delay between data transmission

(d) Mean utilization of a mobile communication system reduced by 33 % in pCAT-DM

Figure 2: Comparison of SINR, delay and utilization of a mobile communication system between pCAT-DM and periodic transmission in a simulative LTE reference scenario

2 Results

To turn out the advantage of using a kNN classifier, its performance is evaluated in comparison to a pure random process and a Bayesian estimator with an a-priori probability based on the distribution of labels in a traning set. Fig. 1 shows the comparison of these classifiers dependent on the number of quantization steps of the label (SINR difference between current value and average value in the future with τ_p look ahead window). It can be seen from the figure that kNN reaches an up to four times higher prediction accuracy and a 10 dB lower prediction error in a 95% confidence interval compared to the named reference classifiers. As time-dependent connectivity forecasts are repatriated to LTE-Sim, a transmission probability taking a predictive and channel-aware component into account is computed in time steps of one second. By simulating 200 User Equipments (UE) in an LTE reference scenario, pCAT-DM turns out an immense SINR gain of 11 dB towards periodic transmission (cf. Fig. 2a, 2b). Regarding the delay of transmission time slots of 200 UEs, pCAT-DM reaches the same mean delay of periodic transmission whereas the upper quartile scatters more due to postponing transmission at increasing future connectivity (cf. Fig. 2c). On the side of a mobile radio system, the utilization can be reduced by 33 % with pCAT-DM towards a peridic transmission of the same ammount of data (cf. Fig. 2d).

References

[1] C. Ide, B. Dusza and C. Wietfeld, *Client-based Control of the Interaction between LTE MTC and Human Traffic in Vehicular Environments*, IEEE Transactions on Vehicular Technology, vol.64, no.5, pp.1856-1871, May 2015.

[2] C. Wietfeld, C. Ide and B. Dusza, *Resource-efficient Wireless Communication for Mobile Crowd Sensing*, Proc. of the 51st ACM/EDAC/IEEE Design Automation Conference (DAC), San Fransisco, USA, Jun. 2014.

[3] C. Ide, M. Nick, D. Kaulbars and C. Wietfeld, *Forecasting Cellular Connectivity for Cyber-Physical Systems: A Machine Learning Approach*, Conference on Machine Learning for Cyber Physical Systems and Industry 4.0, Lemgo, Germany, Oct. 2015.

[4] C. Wietfeld and C. Ide, *Vehicle to Infrastructure Communications, Book chapter in Vehicular Communications and Networks: Architectures, Protocols, Operation and Deployment*, Wai Chen (Ed.), Woodhead Publishing, May 2015.

[5] C. Ide, F. Hadiji, L. Habel, A. Molina, T. Zaksek, M. Schreckenberg, K. Kersting, C. Wietfeld, *LTE Connectivity and Vehicular Traffic Prediction based on Machine Learning Approaches*, IEEE Vehicular Technology Conference (VTC-Fall), Boston, USA, Sep. 2015.

Enhancing Energy Balance between LTE Base Stations and User Terminals

Dennis Kaulbars
Lehrstuhl für Kommunikationsnetze
Technische Universität Dortmund
dennis.kaulbars@tu-dortmund.de

In this report, we focus on the influence of switching-off schemes on the power consumption of terminals in Long Term Evolution (LTE) networks. In this context, we present an innovative switching-off strategy, which also takes the power consumption of mobile terminals into account for deciding whether a base station should be disabled or not, resulting in an enhanced energy balance between the infrastructure and user side.

1 Introduction

The deployment and raising availability of high-speed Long Term Evolution (LTE) networks offers many opportunities to network operators, developers and customers. Nevertheless, there are still some problems: One problem is that even new terminals still suffer from short battery lifetimes. When the number of devices in a cell increases, this problem gets even worse due to the growing competition about the limited radio resources. Another problem is that network operators are currently installing micro cells with only a short coverage range inside hotspot areas (e.g., airports, train stations). If these micro cells are active all the time, there is a huge waste of energy especially in the night since the traffic load of those cells is low during that time. One solution to optimize the energy consumption is to temporary switch off base stations with low traffic load. The disadvantage of classical switching-off schemes is that they do not take care about the increase of the energy consumption on the terminal side as the competition within the remaining cells gets greater. In this report, we present the Energy-Aware Switching off Strategy (EASOS), which also takes care about the energy consumption of the terminals and only switches off a base station if the average power consumption of a device does not increase too much.

2 EASOS: Energy Aware Switching Off Strategy

Our proposed EASOS scheme is an extension to the common Traffic-Based (TR-B) switching-off strategy [2]. This strategy switches off a cell if the traffic load of this cell is below a certain value. In addition to this, the switch off should not violate the Quality of Service (QoS) requirements of the terminals in the cell.

To expand the TR-B strategy about taking the power consumption of the terminals into account, we first need to know how the switch-off of a base station influences the power consumption of the devices in the remaining cells. For this purpose, we estimate the power consumption of an LTE user terminal in relation to the number of devices inside a cell (see Eq. 1), where P_{max} stands for the maximum transmission power of a LTE terminal (for simplification, we assume that all devices send their data traffic related to the same traffic profile with an average data rate of 1 MB/min).

$$f_a(x) = \begin{cases} -0.0013 \cdot x^2 + 0.096 \cdot x + 0.078 & , 1 < x \leq 40 \\ P_{max} & , x \geq 40 \end{cases} \tag{1}$$

Now, for the derivation of EASOS, we introduce the following example scenario:

We have a geographical region with in total $(C_n + 1)$ cells. The Traffic-Based strategy would like to shut down base station z_j, which contains a number of terminals. For each of the remaining cells ($z_i, i \in \{1, 2, .., j-1, j+1, .., (C_n+1)\}$), we calculate the future average power consumption of a terminal that is located in one of the remaining cells (see Eq. 2 right), where N_{z_i} stands for the number of terminals in cell i and $M_{i,j}$ stands for the number of terminals which are added to cell i if we switch off cell j.

$$\bar{P}_{j,new} = \frac{\sum_{i=1}^{C_n} P_{i,j}}{C_n} \text{ with } P_{i,j} = f_a\left(N_{z_i} + M_{i,j}\right) \tag{2}$$

Afterwards, we have C_n power consumption values $P_{i,j}, i \in \{1.., j-1, j+1, .., (C_n+1)\}$. The average power consumption of a device in a typical cell is now the average over all $P_{i,j}$ (see Eq. 2 left), where C_n is the number of remaining cells within the region of interest.

In order that EASOS can compare the increase of the average power consumption, the strategy needs also the average power consumption of a terminal in a typical cell before disabling cell j (see Eq. 3).

$$\bar{P}_{j,old} = \frac{\sum_{i=1}^{C_n+1} f_a\left(N_{z_i}\right)}{C_n + 1} \tag{3}$$

If $\bar{P}_{j,new} < \alpha \cdot \bar{P}_{j,old}$, EASOS decides to switch off base station j, otherwise the base station stays active. The parameter α is a threshold value which the network operator can adjust to modify the system behavior.

3 Simulative Performance Evaluation of EASOS

Fig. 1 left shows the simulation scenario which we use to evaluate the performance of our proposed EASOS strategy.

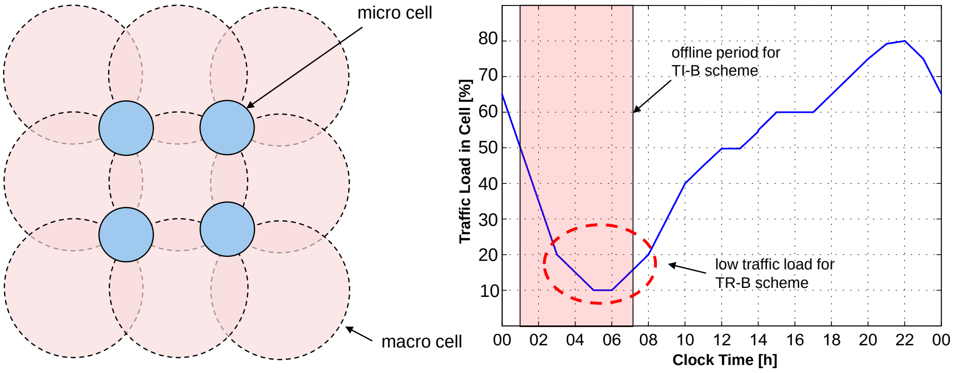


Figure 1: Simulation Scenario for Performance Analysis

The scenario consists of nine macro and four micro cells (red cicle in Fig. 1). All terminals within the cells create negative exponential distributed traffic with the mean of 1 MB/min. The active (sending) times of all terminals relate to the traffic profile displayed in Fig. 1 right with low load level during night (01:00 a.m. – 07:00 a.m.) and a peak value in the evening (10:00 p.m.). As the uplink scheduler, we use Round Robin and the bandwidth of each cell is set to 10 MHz. We compare the EASOS scheme with two traditional switching off strategies: The Time-Based Strategy (TI-B) just switches off the micro cells during night [1]. The TR-B strategy switches off the micro base stations if the cell load is below 25%. As the threshold parameter for EASOS, we use $\alpha = 1.1$.

Fig. 2 shows the average power consumption of an active terminal (left) and a base station (right) for the different switching off strategies. Without switching off any base station, the cell has a low load level during night and an efficient data transmission is possible. In the evening, there are many active terminals, which want to send data at the same time and therefore, the data transmission requires a higher power consumption. Because TI-B strategy switches off all micro cells during night, the power consumption of the terminals increase during that time, because also there is also some traffic during night in the cell and the terminals in the remaining cells suffer from the greater competition about the radio resources. The TR-B strategy switches off the micro cells to a later point of time (at 2:00 p.m.), where the traffic load in the cells is generally lower and the power consumption increases only from this time. The EASOS strategy decides to

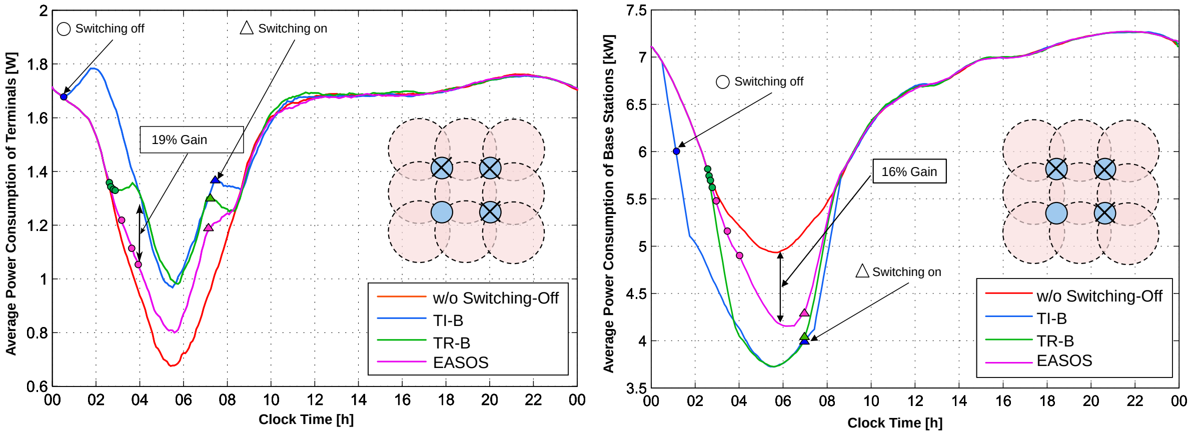


Figure 2: Average Power Consumption of Terminal (left) and all Base Stations (right)

leave the micro cell located at the bottom left of the scenario online, because then the competition between active terminals in all cells is not so high and the power consumption does not increase much. On terminal side, we gain a power consumption reduction of about 19%. On the base station side, we have an anti-proportional behavior of the curves. When the strategies switch off all base stations (TI-B and TR-B), the average power consumption of the base stations are low. When all base stations are active for the whole time (w/o Switching), we have the highest average power consumption, because all base stations permanently consume energy. Our proposed EASOS strategy is between the w/o Switching and the other curves, because EASOS only switches off three of four micro cells. This means that, compared with traditional switching-off schemes, EASOS receives a gain on the terminal as well as on the base station side and thus it optimizes the energy balance between both parties.

References

[1] A. Bousia, E. Kartsakli, L. Alonso, and C. Verikoukis. "Energy Efficient Base Station Maximization Switch Off Scheme for LTE-advanced". In *Computer Aided Modeling and Design of Communication Links and Networks (CAMAD), 2012 IEEE 17th International Workshop on*, pages 256–260, Sep. 2012.

[2] R. Sachan and N. Saxena. "Clustering Based Power Management for Green LTE Networks". In *Computer Communication and Informatics (ICCCI), 2014 International Conference on*, pages 1–3, Jan. 2014.

Poisson Graphical Models and Vehicular Traffic Prediction

Alejandro Molina
Lehrstuhl für Künstliche Intelligenz
Technische Universität Dortmund
alejandro.molina@tu-dortmund.de

This report describes a graphical model for multivariate poison distributions and their applications to the analysis of traffic data. This non-parametric model can recover dependencies that are semantically meaningful while providing means for doing inference.

1 Poisson Dependency Networks

To ease the modeling of multivariate count data, we introduced a novel family of Poisson graphical models, called Poisson Dependency Networks (PDNs) [4]. A PDN consist of a set of local conditional Poisson distributions, each one representing the probability of a single count variable given the others. Inference in this model is done by means of Gibbs Sampling. The local models of the PDNs can be as simple as log-linear Poisson models, or more complex such as Poisson regression trees or even non-parametric models trained using functional gradient ascent, i.e. boosting.

To introduce PDNs, we use X to denote a random variable and x as its instantiation. Sets of random variables are written as $\mathbf{X}$ and correspondingly their instantiations as $\mathbf{x}$. Given a set of random variables $\mathbf{X} = (X_1, \ldots, X_n)$ where each variable is defined over the natural numbers, including 0, then a Poisson Dependency Network (PDN) is a pair (G, P). Here, $G = (V, E)$ is a directed, possibly cyclic, graph with V being a set of nodes where each node corresponds to a random variable in $\mathbf{X}$. Hence, we can use nodes in G and the random variables in $\mathbf{X}$ interchangeably. $E \subseteq V \times V$ is a set of directed edges where each edge models a dependency between variables, i.e., if there is no edge between two variables X_i and X_j, the variables are conditionally independent given the other variables $\mathbf{X}_{\setminus i,j}$ in the network. Here, $\mathbf{X}_{\setminus i,j}$ is shorthand for $\mathbf{X} \setminus \{X_i, X_j\}$. We refer to the nodes that have an edge pointing to X_i as the parents of X_i, denoted as $\mathbf{pa}_i$. P is a set of conditional probability distributions for every variable in $\mathbf{X}$. For now, we assume that each variable X_i given its parents $\mathbf{pa}_i$ is Poisson distributed, i.e.,

$$p(x_i|\mathbf{pa}_i) = p(x_i|\mathbf{x}_{\setminus i}) = \frac{\lambda_i(\mathbf{x}_{\setminus i})^{x_i}}{x_i!} e^{-\lambda_i(x_{\setminus i})} \, . \tag{1}$$

Here, $\lambda_i(\mathbf{x}_{\setminus i})$ is a function which models the mean of the probability distribution of X_i.

2 Vehicular Traffic Prediction

We presented an alternative approach to modeling and predicting the traffic flow based on this probabilistic graphical model in [5]. Here, we assumed that X_i represent the traffic flow measured by detector i and x_i represents an instantiation of this random variable, i.e., a particular flow count. PDNs are well suited for this task, because they capture the appropriate nature of the variables, i.e., $x_i \in \mathbb{N}$ and Poisson distributed.

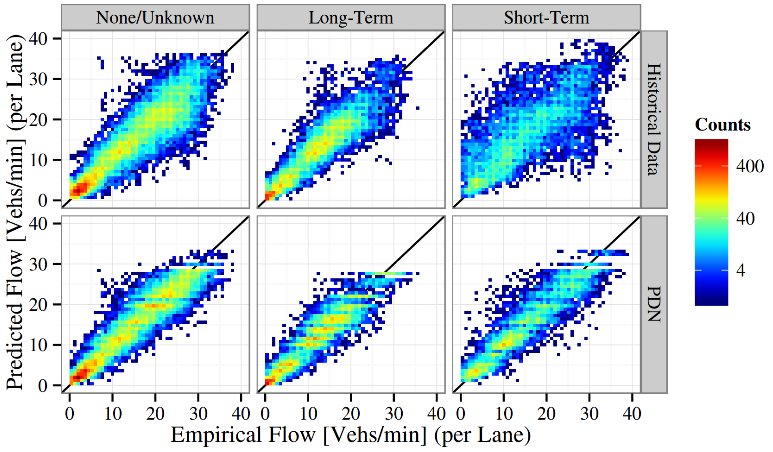


Figure 1: Comparison between Empirical and Predicted Vehicular Traffic Flows (per Lane) on Tuesday, 26 August 2014, Using the Historical Data Approach and PDNs; Characterized by Event Type. (Best viewed in color)

In this particular case, we modeled $\lambda_i(\mathbf{x}_{\setminus i})$ with Poisson regression trees (PRTs) [1], in particular implemented by [7]. We then ran a leave-one-out cross validation for the task of filling in missing data points and compared it to the work in [2]. As we can see in Fig. 1, the PDN is very competitive.

Given that the PDNs are graphical models, we can also interpret the incoming edges on the graph as important features for the prediction of a traffic flow. When we plot this

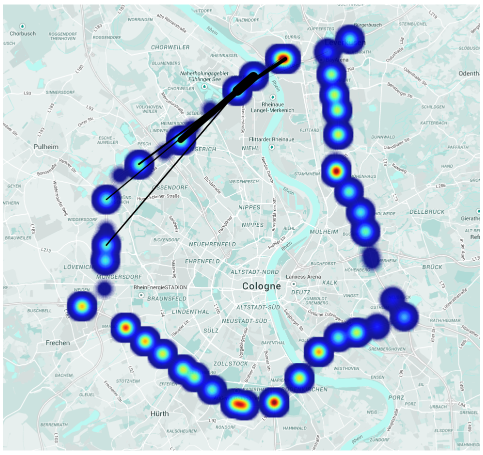


Figure 2: Map of Average Traffic (Dark Red Segments Indicate High Traffic) on the Highway at 5pm with Dependences of the Local Model Representing One Detector in the North by Black Edges. ©Google. (Best viewed in color)

on a map, we can see that the PDNs recover a semantically meaningful structure as we can see in Fig. 2.

3 Admixtures of PDNs

One further line of ongoing research is using the PDNs to model topics with directed word dependencies, we can view topic-word distributions of LDA as a special case of a Poisson graphical model.

An Admixture of PDNs [3] is a model where the local models are conditional on a topic distribution, one such model can be a parametric model similar to a PDN with GLMs, i.e.,

$$\lambda_i(\mathbf{x}^d_{\setminus i}) = \exp\left(\sum_{t \in T} z^d_t \left(\beta_{t,i} + \sum_{j \neq i} \beta_{t,i,j} x^d_j\right)\right) . \tag{2}$$

Using the Admixtures of PDNs model allows us to find topics on the traffic data where one can spot trends in time as in Fig. 3.

4 Poisson Sum Product Networks

Sum Product Networks (SPNs) [6] are graphical models that can do efficient exact inference. We are working on extending them to the Poisson distribution.

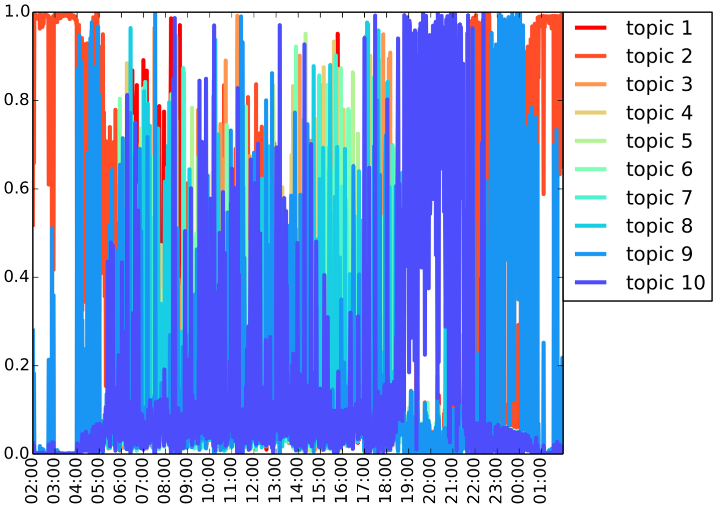


Figure 3: Plot of topic assignments found by the Admixture of Poissons on traffic data. (Best viewed in color)

References

[1] Leo Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Statistics/Probability Series. Wadsworth Publishing Company, Belmont, California, U.S.A., 1984.

[2] Johannes Brügmann, Michael Schreckenberg, and Wolfram Luther. Real-Time Traffic Information System Using Microscopic Traffic Simulation. In *8th EUROSIM Congress on Modelling and Simulation*, pages 448–453, Cardiff, Wales, September 2013.

[3] Elena Erdmann. Capturing semanticall meaningful word and topic dependencies with Poisson Dependency Networks. Master's thesis, TU Dortmund, Germany, 2015.

[4] Fabian Hadiji, Alejandro Molina, Sriraam Natarajan, and Kristian Kersting. Poisson dependency networks: Gradient boosted models for multivariate count data. *Machine Learning Journal (MLJ)*, 2015.

[5] Christoph Ide, Fabian Hadiji, Lars Habel, Alejandro Molina, Thomas Zaksek, Michael Schreckenberg, Kristian. Kersting, and Christian Wietfeld. Lte connectivity and vehicular traffic prediction based on machine learning approaches. In *IEEE 82nd Vehicular Technology Conference (VTC-Fall)*, Boston, USA, Sep. 2015.

[6] Hoifung Poon and Pedro M. Domingos. Sum-product networks: A new deep architecture. *CoRR*, abs/1202.3732, 2012.

[7] Terry M. Therneau, Beth Atkinson, and Brian Ripley. *rpart: Recursive Partitioning*, 2011.

Software-Defined Radio Solution for Lane-Specific Vehicle Positioning via Local Interference Compensation

Florian Schweikowski
Lehrstuhl für Kommunikationsnetze
Technische Universität Dortmund
florian.schweikowski@tu-dortmund.de

The steadily increasing traffic density is causing enormous negative effects such as jams, accidents or CO^2 emissions, especially in urban areas. In Europe more than 12% of the traffic network is daily congested. Hereby a reliable and dynamic congestion forecast is the key to avoid and/or compensate such locale bottleneck situations. Hence, this contribution focuses on a cloud-aided, lane-specific position determination of vehicles using a so-called *Local Interference Compensation* (LOCATe) to enable a more detailed and lane-accurate traffic prediction (e.g. detecting short-dated roadworks or car breakdowns), and by that traffic-flow manipulation in the future.

1 Local Interference Compensation (LOCATe)

The idea of LOCATe is to predict all this partially position-specific and highly variable influences for any given location. Thereby, LOCATe performs the following three steps: *Predict* all influences, *Quantify* the accruing shift and *Compensate* the overall error vector. The procedure itself is visualized in Figure 1. First of all, the defective GNSS position P_{GPS} (in this case the authors used GPS) is measured. Afterwards a reference grid around P_{GPS} (blue crosses) is generated within a 3D model for the local considerations, whereby its dimensions are correlated to the estimated accuracy of the measured position using the available key performance indicators, like the *Geometric Dilution of Precision (GDOP)*.

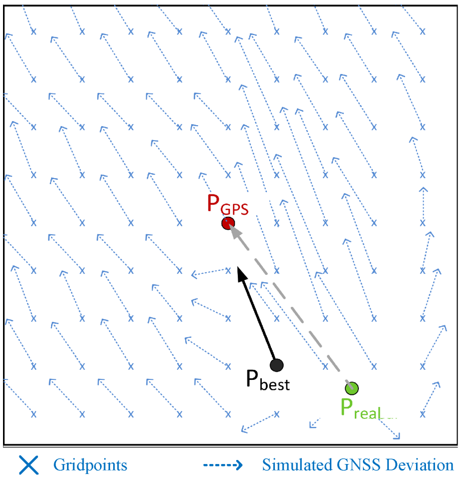


Figure 1: Evaluation of LOCATe's Enhanced GNSS Positioning

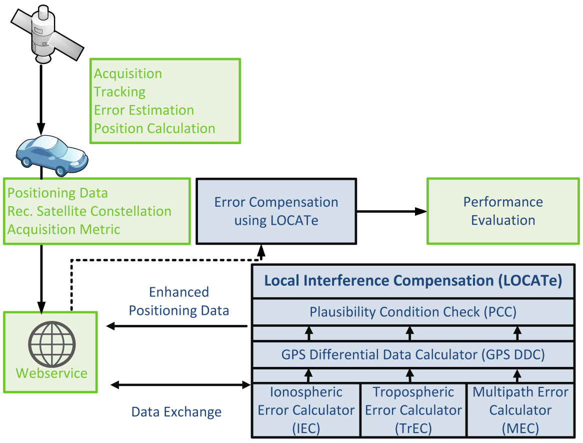

Acquisition
Tracking
Error Estimation
Position Calculation
Positioning Data
Rec. Satellite Constellation
Acquisition Metric
Error Compensation using LOCATe
Performance Evaluation
Enhanced Positioning Data
Webservice
Data Exchange
Local Interference Compensation (LOCATe)
Plausibility Condition Check (PCC)
GPS Differential Data Calculator (GPS DDC)
Ionospheric Error Calculator (IEC)
Tropospheric Error Calculator (TrEC)
Multipath Error Calculator (MEC)

Figure 2: Architecture of LOCATe

Afterwards LOCATe simulates the accruing error effects (blue arrows), atmospheric as well as multipath, for every single reference point/satellite connection using the *Multiscale Simulation Environment (MSE)* [1]. In a further step, the most probable error vector, which ends up in the measured GPS position is used as best-in-case compensation for the actual constellation. The beginning of this vector then is interpreted as the supposed position P_{best}. To evaluate the performance of LOCATe, two of measurement points specified by the land surveying office at the campus of the university in Dortmund with fixed GPS coordinates will be used. These points are used for all the measurements and therefore the real position (P_{real}) is known and the remaining positioning error may be quantified.

The architecture of the LOCATe is shown in Figure 2. Input value is the defective GPS position that is passed to the so developed *Open Source Satellite Simulator (OS^3)* [4], which is based on updated *Two-Line-Element* sets, to enable and integrate a highly accurate satellite movement. The satellite and receiver positions are passed through multiple error correction modules. The atmospheric compensation modules uses *SISNeT* to receive EGNOS messages as additional information. Subsequently, the individual error values are determined. Using the so-called *Differential Data Calculator (DDC)* the grid point with the highest probability of being the wanted real position is selected. Finally a validation (Plausibility Check Condition (PCC)) is performed, which determines whether time and/or position of sequenced measurements differ widely and are therefore might be subject to errors. A more detailed description of the error sources within a GNSS signal, the error compensation techniques as well as the validation equipment can be found in [3].

2 Accuracy Enhancement using LOCATe

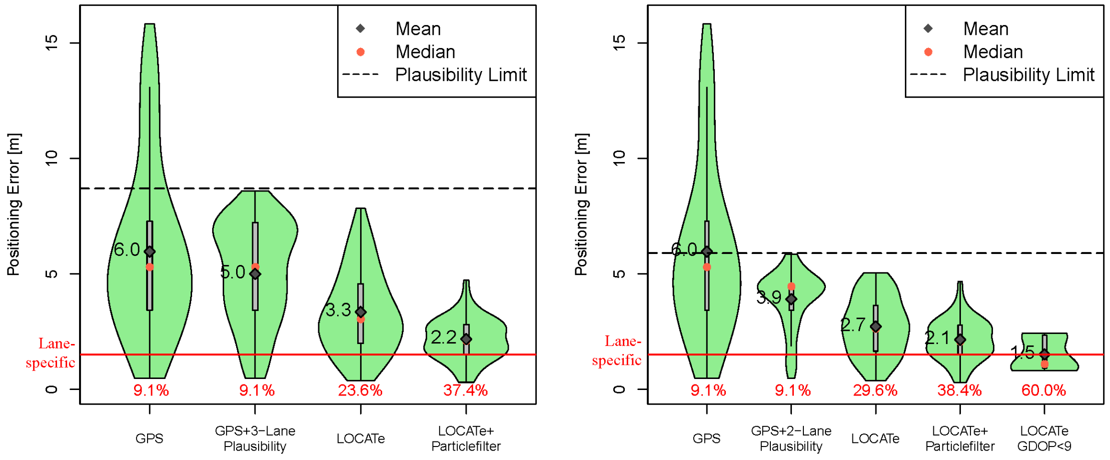
The results of analyzing different LOCATe compensation setups using more than 500 measurements on the mentioned points are shown in Figure 3a. It turns out that the best performance gain occurs when using the GPS integrated troposphere correction, whereby ionosphere and multipath effects are compensated by the LOCATe.

Pure GPS for example shows an expectable deviation for highly challenging environments, which also fits to former evaluations [2]. Based on the fact that the scenario deals with traffic situations, it is obvious to add plausibility routines, like normal GPS receiver do. E.g., on a three lane highway, deviations greater than the overall street width (e.g., a three lane highway with 2.9m each lane) may be detected easily. Hence, a corresponding filter was added to the pure GPS and set all values above the plausibility limit to the nearest lane-center value, so in this case to 7.25m $(= 3 \cdot 2.9\text{m} - (\frac{2.9\text{m}}{2}))$. Based on that, LOCATe used these values as input and it turns out that up to 23.6% of all positions can be determined lane-specific ($< 1.5m$, visualized by the red solid line) in contrast to just 9% with pure GPS and reduce the average error by more than 45 %, and even more important: LOCATe lowers the occurring peak value significantly that again clarifies the benefit.

Additionally, filter algorithms may be used to further improve the accuracy gain. It should be mentioned that this point is just used to clarify the additional performance possibility and shall not indicate the finest choice to work with LOCATe. Just as an example, the authors applied a particle filter (well-known method in GPS positioning techniques) to the already smoothed results from LOCATe and by that decreases the mean error by further 33%, resulting in an overall reduction of 63.3 % in average. Furthermore, a combined handling using LOCATe and additional filters, increases the lane-specific detection by four times. Using our measurements, more than 37% of all recorded points might be improved to a lane-specific positioning. In contrast to just 9% using GPS only. The same observations holds true for changing the mentioned plausibility limit to a two-lane highway, shown in Figure 3b.

Again, all values above the plausibility limit are mapped to the nearest lane-center value. In this case to 4.35m $(= 2 \cdot 2.9\text{m} - (\frac{2.9\text{m}}{2}))$. LOCATe as well as the additional particle filter again performs very well with the better input values. Furthermore the effect of less distorted input values is shown the fifth example in Figure 3b. Hereby, only those GPS values with a GDOP lower 9, were forwarded to LOCATe to clarify its performance in less challenging environments. Again, LOCATe is able to increase the positioning accuracy to 1.5m in average and locate 60% of all values lane-specific.

It turns out, that using LOCATe on its own, an enhancement of up to 45% in average is possible. By adding additional filter methods to the already smoothed results from



(a) with 3-Lane Plausibility

(b) with 2-Lane Plausibility and an Exemplary Scenario with Good Inputs (GDOP < 9)

Figure 3: Performance of LOCATe

LOCATe, an overall reduction of 63% in average is visible. In addition, LOCATe eliminated the occurring peak values in GNSS and by that, allows the use of satellite-based positioning for further applications than today, even safety-critical ones.

References

[1] Andreas Lewandowski, Brian Niehoefer, and Christian Wietfeld. Performance evaluation of satellite-based search and rescue services: Galileo vs. cospas-sarsat. In *Vehicular Technology Conference, 2008. VTC 2008-Fall. IEEE 68th*, pages 1–5. IEEE, 2008.

[2] Brian Niehoefer, Sarah Lehnhausen, and Christian Wietfeld. Combined Analysis of Local Ionospheric and Multipath Effects for Lane-Specific Positioning of Vehicles within Traffic Streams. In *6th ESA Workshop on Satellite Navigation Technologies (NaviTech)*, Noordwijk, The Netherlands, December 2012.

[3] Brian Niehoefer, Florian Schweikowski, Sarah Lehnhausen, and Christian Wietfeld. Cloud-aided sdr solution for lane-specific vehicle positioning via local interference compensation. In *Aerospace Conference, 2014 IEEE*, pages 1–7. IEEE, 2014.

[4] Brian Niehoefer, Sebastian Šubik, and Christian Wietfeld. The cni open source satellite simulator based on omnet++. In *Proceedings of the 6th International ICST Conference on Simulation Tools and Techniques*, pages 314–321. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2013.

Fractal Analysis of Traffic Time Series

Thomas Zaksek
Physik von Transport und Verkehr
Universität Duisburg-Essen
thomas.zaksek@uni-due.de

Time series can show signs of fractal and multifractal behaviour. An analysis from this perspective can unearth features of time series that remain hidden for analysis with standard statistics. We analyse the multifractal spectra of traffic time series with the help of Multifractal Detrended Fluctuation Analysis (MDFA). Empirical time series of traffic flows and velocities measured by loop detectors are compared with time series gathered from traffic simulations. As a second focus we analyse multifractal features of time series from different vehicle classes, i.e passenger and transport traffic.

Time series can not only describe systems with fractals features (e.g. systems with chaotic behaviour and strange attractors) but also the graph of a time series itself can be a fractal (e.g. time series of full developed turbulence or financial time series like the S&P500). Not only statistical self-similarity but also roughness of timeseries is quantified by signatures of fractal behaviour like box counting dimension [2].

Some studies indicate that time series of traffic data also show fractal and especially multifractal behaviour [6]. In this paper we present an analysis of empirical and simulated timeseries of German motorway traffic with focus on multifractal spectra.

The traffic data is gathered from the motorway network of the German State of North Rhine-Westphalia and traffic simulations.

We use a multifractal formalism that gives us a multifractal spectra by a hoelder grain exponent α and calculating the graph of the scaling function $\tau(\alpha)$ vs α.

A Multifractal Detrended Fluctuation Analysis [4], [5] ansatz is used here on the assumption that traffic time series have trends on many scales (e.g, intra day because of commuters, weekly because of workdays and weekend). As a first step, we "profile" the time series:

$$X_p(i) = X(i) - \langle X \rangle, \quad i = 1 \dots N$$

for a time series X with length N.
The time series is divided in dyadic intervals of degree n (i.e 2^n non overlapping intervals with length 2^{-n} each.
$\alpha_{n,k}$ is the exponent such that the coefficients

$$D(I_n(k)) = \Big(\frac{1}{\#I_n(k)} \sum_{\chi \in I_n(k)} (X_p(\chi) - P_i^{dp})^p\Big)^{1/p}$$

show a power law behaviour to the interval size:

$$\alpha_{n,k} = \frac{\log(D_x(I_n(k))}{\log(2^{-n})}$$

(with $I_n(k)$ the k-th dyadic interval at scale n, $\#I$ the cardinality of the points in I, P_i^{dp} a polynomial fit of degree dp fitting the $X \in I$).
Due to method and chosen scaling with dyadic intervals, the grain exponents α drift with scale n. It is necessary to normalize the signal (see [5]).
Now we calculate the Legendre Spectrum

$$\tau^*_{X,n}(\alpha) = \inf_{q \in \mathbb{R}} \{\alpha q - t_{x,n}(q)$$

with

$$t_{x,n}(q) = -\frac{1}{n} \log_2 \sum_{I \in \xi_n} 2^{-nq\alpha_X(I)}$$

for a range of $q \in [-50, 50]$.
Now the multifractal spectrum $\tau^*(\alpha)$ over α is analysed at different scales n. Qualitatively the width of these spectra relate to multifractal properties of the time series. The Peak (if it exists) and mean of the spectra can be both interpreted as an estimator of the likeliest fractal dimension of the time series.
As empirical data we use measurements from loop detectors of the A3 freeway north of Cologne from the German province of North-Rhine Westphalia. The time series are gathered from a detector at the second lane (three lanes altogether) and include transport traffic (including heavy trucks, light trucks and buses). The data set we analyse includes flow and velocity, separated into passenger car and transport traffic.
For comparison we use simulated traffic data from the autobahn.nrw project. The simulation uses a multilane variety of the Nagel-Schreckenberg-Model [1] [3]. The simulated network spans the whole freeway network of North-Rhine Westphalia. Every minute the simulation runs trough a tuning process where the state of the simulation is compared and balanced with realtime empirical data. In the simulated network, virtual detectors are placed to gather the time series. The placement of the virtual detectors covers approximately the same positions as the detectors on the real freeway. For an in depth description of the autobahn.nrw simulation, see [1].
We use a time series of traffic flow and velocity with 16348 data points(about 1 week of

data). Each data point corresponds to one minute. Traffic flow is cumulated over one minute, velocity is averaged over one minute.

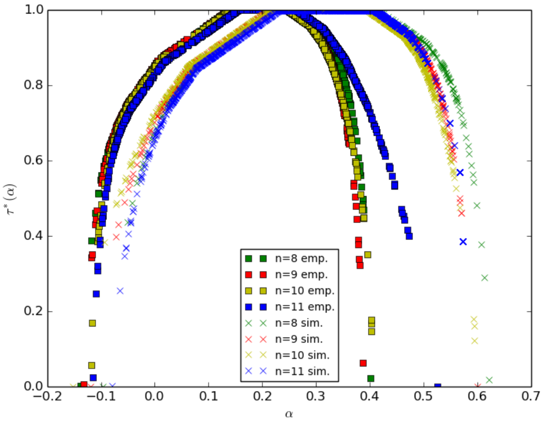


Figure 1: Multifractal spectrum of passenger traffic velocities. Empirical measurements and simulated data with 16348 data points each

Figure 1 shows a comparison of the multifractal spectra of empirical and simulated passenger car velocity time series for $n = 8, \ldots, 11$. Both spectra show signs of multifractal behaviour with a width of 0.5 respectively 0.7. The estimated likeliest fractal dimension for these time series is about 1.8 respectively 1.7.

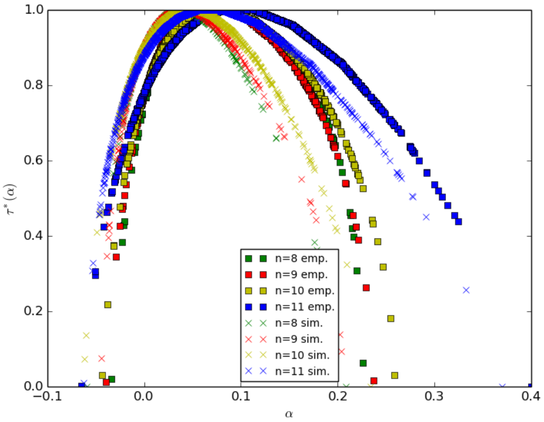


Figure 2: Multifractal spectra of traffic flows. Empirical measurements and simulated data with 16348 data points each

Figure 2 shows a comparison of the multifractal spectra of empirical and simulated traffic flow time series for $n = 8, \ldots, 11$. Both spectra show weak multifractal behaviour with a width of about 0.3 to 0.4. The estimated likeliest fractal dimension for these time series is about 1.9. For traffic flow, both spectra resemble each other much more than for the velocity time series.

We suppose that the differences between the spectra of empirical and simulated velocity time series originate in part from the fact that the cellular automata model uses a step wise car velocity and acceleration. This may lead to more unsteady behaviour of the single cars and as a result to a rougher, more volatile velocity time series. On the other hand, the traffic flow time series from simulated data resembles the time series from empirical data. This may be due to both empirical and simulated measurements are integer count data and the simulation by construction resembles the distribution of the empirical flows. The weaker signs of multifractal behaviour for the flow compared to the velocity may originate from the less volatile nature of the traffic flow and its relative insensitivity to changes of the traffic state (e.g. free flow and synchronized traffic may have the same amount of traffic flow).

References

[1] Johannes Brügmann, Michael Schreckenberg, and Wolfram Luther. A verifiable simulation model for real-world microscopic traffic simulations. *Simulation Modelling Practice and Theory*, 48(0):58–92, 2014.

[2] S. Davies and P. Hall. Fractal analysis of surface roughness by using spatial data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(1), 1999.

[3] Lars Habel and Michael Schreckenberg. Asymmetric lane change rules for a microscopic highway traffic model. In *Cellular Automata - 11th International Conference on Cellular Automata for Research and Industry, ACRI 2014*, pages 620–629, 2014.

[4] Jan W. Kantelhardt, Stephan a. Zschiegner, and H. Eugene Stanley. Multifractal detrended uctuation analysis of nonstationary time series. *Physica A*, 316:87–114, 2002.

[5] Patrick Loiseau, Claire Médigue, Paulo Gonçalves, Najmeddine Attia, Stéphane Seuret, François Cottin, Denis Chemla, Michel Sorine, and Julien Barral. Large deviations estimates for the multiscale analysis of heart rate variability. *Physica A: Statistical Mechanics and its Applications*, 391(22):5658–5671, 2012.

[6] Jing Wang, Pengjian Shang, and Xingran Cui. Multiscale multifractal analysis of traffic signals to uncover richer structures. *Physical Review E*, 89(3):032916, 2014.

SFB 876 Verfügbarkeit von
Information durch Analyse unter
Ressourcenbeschränkung

Projekt C1
Feature selection in high dimensional data for risk prognosis in oncology

Alexander Schramm Sven Rahmann
Sangkyun Lee

Unveiling Structure with Pattern Compression

Sibylle Hess
Lehrstuhl für Künstliche Intelligenz
Technische Universität Dortmund
sibylle.hess@tu-dortmund.de

We propose a new approach to a pattern-based compression of binary databases. Since Jorma Rissanen's influential work on the minimum description length principle, it has been used for model selection and data compression. Here, we aim at its use for database exploration, providing users with patterns that unveil the true underlying structure of tiles or biclusters. The level of detail in which the data is viewed shall be controllable – by specifying the amount of returned patterns. Therefore, we propose a novel boolean matrix factorization algorithm, based on optimization theoretic grounds.

Introduction In various data mining tasks such as Market Basket Analysis, Text Mining, Collaborative Filtering or DNA Expression Analysis, we are interested in the exploration of data which is represented by a binary matrix $D \in \mathbb{R}^{m\times n}$. Let's have a look at the artificially created binary database on the left in Fig. 1. Every item corresponds to a column and every transaction (a set of items) is depicted by a row. The matrix contains a visible structure of blocks which is for any data displayable by a suitable permutation of rows and columns. We model this view by a decomposition into a structural component $\theta(YX^T)$ (θ is the Heaviside step function at threshold 0.5) and noise $N \in \{-1,0,1\}^{m\times n}$, such that $D = \theta(YX^T)+N$. The matrices $X \in \{0,1\}^{n\times r}$ and $Y \in \{0,1\}^{m\times r}$, determine r block components by the itemsets $X_{\cdot s}$ and corresponding transactions $Y_{\cdot s}$. The discovery of the structure $\theta(YX^T)$ is motivated under the viewpoint of two fundamental research branches: pattern mining and clustering.

A pattern $X_{\cdot s}$ defines a block component together with the transactions $Y_{\cdot s}$ where it satisfies a given criterion of interestingness. However, individual interestingness definitions face the problem of redundant pattern generations [5]. Siebes et al. counter this problem

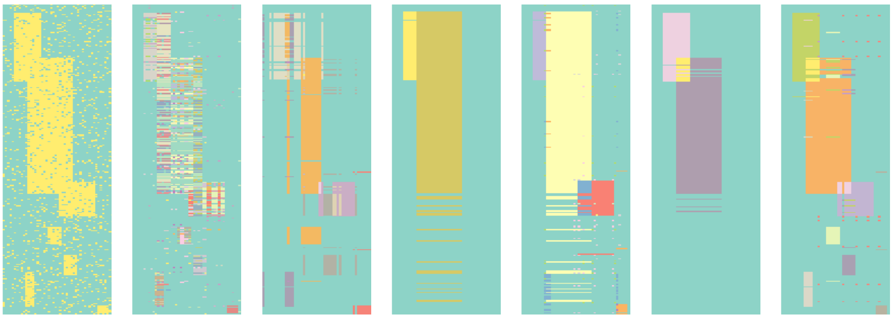


Figure 1: Example binary dataset (left), ones are depicted in yellow and zeros in blue. From left to right: Results of the algorithms Slim, BMF, Panda+ (ranks 2 and 8) and Pimp (ranks 2 and 8) are shown in different colors for each filtered structural component.

by the application of the *Minimum Description Length* (MDL) principle, using patterns for compression. The algorithm Krimp implements this method to prune a set of patterns into an informative fraction [5]. A typically more efficient compression is returned if the compressing patterns are directly mined by the algorithm Slim [4]. Its identification of the structural component in our example database, using only non-singleton patterns, is pictured by the middle matrix in Fig. 1. Derived patterns (each one has its own color) reflect the shape of blocks, but they do not generalize well.

The task of biclustering, which clusters items $X_{\cdot s}$ and transactions $Y_{\cdot s}$ simultaneously, offers another perspective. Applications include the identification of groups of highly related genes in subset gene expression samples, topic analysis of documents or the collaborative filtering of users and opinions. Zhang et al. propose in [7] a promising biclustering approach relying on numerical optimization methods for Matrix Factorization. The rightmost matrix in Fig. 1 portrays the result of their algorithm (BMF) in different colors for each bicluster. Although this method provides a more encompassing view on the structure of the data, it does not detect the rectangles correctly.

We combine the strengths of these two viewpoints in a method which is capable to unveil the underlying structure of a database in a user specified level of detail. Looking at our example matrix, we aim at discovering the three large overlapping blocks in a coarser view and some of the smaller block structures with increasing level of detail. Applying a penalty term to make numerical optimization methods well-suited for MDL encoding, we call our algorithm *Penalizing (kr)IMP* (PIMP).

Pattern Compression We formulate the objective of Krimp for the first time in terms of boolean matrix factorization and extend it to model noisy structures. As such, we use optimal codes for patterns denoted in X to encode the transactions indicated by Y. A lossless encoding is achieved by the usage of singleton codes as indicated by the noise matrix N. Therewith, the compression size of the data in bits is given as

$$L^D(X,Y,N) = -\sum_{s=1}^{r} |Y_{\cdot s}| \cdot \log\left(\frac{|Y_{\cdot s}|}{|Y|+|N|}\right) - \sum_{i=1}^{n} |N_{\cdot i}| \cdot \log\left(\frac{|N_{\cdot i}|}{|N|+|Y|}\right). \qquad (1)$$

Similarly, we compute the description size of the model $L^M(X,Y,M)$ and obtain the total description size according to the MDL principle as sum of data and model description $L(X,Y,N) = L^D(X,Y,N) + L^M(X,Y,N)$.

Existing algorithms which try to find the best describing model, follow a greedy approach and determine the usage of codes heuristically. Although this already yields useful characterizations of a dataset [5], neither overlaps nor noise in the structural components can be modeled this way. Research in the related field of tiling assesses this to return less succinct and noise susceptible models (cf. [2] and references therein).

Pimp With Pimp, we attempt to model overlapping and noisy compressing patterns without relying on a heuristic usage definition. To do so, we derive an upper bound on the compression size of the database

$$L^D(X,Y) \leq \widetilde{L^D}(X,Y) = \log(en(1+mr))\frac{1}{2}\|D - YX^T\|^2 + L^D_\epsilon(X,Y),$$

where L^D_ϵ is the approximate compression size of the database, assuming that every pattern $X_{\cdot s}$ is used at least ϵ times. This motivates the formulation of the following optimization problem, using the function ϕ to penalize non-binary matrix entries.

$$\min_{X\in\mathbb{R}^{n\times r}, Y\in\mathbb{R}^{m\times r}} \widetilde{L}(X,Y) + \phi(X) + \phi(Y). \qquad (2)$$

Minimizing $\widetilde{L}$ encompasses Nonnegative Matrix Factorization (NMF). Existing algorithms generally approximate the Gauss-Seidel scheme, exploiting the convexity of $\|D - YX^T\|^2$ in one of the matrices [6]. This can not be transferred to Problem (2) since $L_\epsilon(X,Y)$ is concave. However, the Gauss-Seidel scheme has recently been extended to nonconvex functions by the *Proximal Alternating Linearized Minimization* (PALM) [1].

To apply this minimization, we derive the proximal mapping $\text{prox}_{\alpha_k\phi}$ for a suitably defined function ϕ and the Lipschitz constant of the partial gradients $\nabla_{X/Y}\widetilde{L}(X,Y)$. This defines the method of Pimp as stated in Algorithm 1.

Experiments We analyze the quality of Pimp's compression in comparative experiments with the closest-related algorithms Krimp, Slim, and Groei [3], on six UCI datasets. On most datasets, PIMP achieves a similar compression ratio to Krimp and Slim, but Pimp requires substantially fewer patterns to do so. While Groei mirrors the low number of patterns required by Pimp, it results in a higher (worse) compression ratio.

Algorithm 1 Pimp

```
function Pimp$(D, r, K, S)$                    ▷ $K$: max iterations, $S$: set of thresholds
    Initialize $X_0 \in [0,1]^{n \times r}$ and $Y_0 \in [0,1]^{m \times r}$ randomly
    for $k \in \{0, \ldots, K-1\}$ do                         ▷ Proximal Alternating Minimization
        $\alpha_k^{-1} \leftarrow \gamma\mu\|Y_k Y_k^T\|$
        $X_{k+1} \leftarrow \mathrm{prox}_{\alpha_k \phi}\left(X_k - \alpha_k \nabla_X \widetilde{L}(X_k, Y_k)\right)$
        $\beta_k^{-1} \leftarrow \gamma\left(\mu\|X_{k+1}X_{k+1}^T\| + 2m\frac{\epsilon+1}{\epsilon^2}\right)$
        $Y_{k+1} \leftarrow \mathrm{prox}_{\beta_k \phi}\left(Y_k - \beta_k \nabla_Y \widetilde{L}(X_{k+1}, Y_k)\right)$
    end for                                            ▷ End Proximal Alternating Minimization
    $(a, b) \leftarrow \arg\min_{(a,b) \in S \times S} L\left(\theta_a(X_K), \theta_b(Y_K), D - \theta_b(Y_K)\theta_a(X_K)^T\right)$
    return $(\theta_a(X_K), \theta_b(Y_K))$
end function
```

On the artificial dataset with noisy tiles, we analyze the quality of Pimp's resulting tiles (cf. Fig. 1). Comparative experiments are run against PaNDa+ [2], which was originally evaluated similarly. Although PaNDa+ approximates the seeded structure quite agreeably, it clearly trails PIMP in terms of recovery of the four smaller tiles seeded elsewhere in the dataset.

References

[1] Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1-2):459–494, 2014.

[2] Claudio Lucchese, Salvatore Orlando, and Raffaele Perego. Mining top-k patterns from binary datasets in presence of noise. In *SDM*, volume 10, pages 165–176, 2010.

[3] Arno Siebes and René Kersten. A structure function for transaction data. In *SDM*, pages 558–569. SIAM, 2011.

[4] Koen Smets and Jilles Vreeken. Slim: Directly mining descriptive patterns. In *SDM*, pages 236–247. SIAM / Omnipress, 2012.

[5] Matthijs van Leeuwen. *Patterns that matter*. PhD thesis, Utrecht University, 2010.

[6] Yu-Xiong Wang and Yu-Jin Zhang. Nonnegative matrix factorization: A comprehensive review. *Knowledge and Data Engineering, IEEE Transactions on*, 25(6):1336–1353, 2013.

[7] Zhongyuan Zhang, Chris Ding, Tao Li, and Xiangsun Zhang. Binary matrix factorization with applications. In *ICDM*, pages 391–400, 2007.

Functional validation of mutations in neuroblastoma cell lines

Marc Schulte
Oncology Lab-Children's Hospital Essen
Universität Duisburg-Essen
marc.schulte@uk-essen.de

Neuroblastoma arises from neuroblasts of the neural crest, which remain in an immature stage. Since the probability of developing neuroblastoma decreases with increasing age, it can be assumed that for this cancer environmental influences play a minor role. Thus, neuroblastoma is an excellent model system to investigate individual oncogenes. The aim of this project is to identify genes, which are responsible for the transition from primary neuroblastoma to relapse. In 2015, we characterized 16 paired samples at diagnosis and relapse from individuals with neuroblastoma and were able to identify several potential key genes for relapse formation [7]. To investigate the biological relevance of these genes, we aim to generate target gene specific knock-out cell lines by using the CRISPR / Cas9 system [3] and compare the initial and the knock-out cell line with respect to cell migration behaviour and the ability for single cell survival. Moreover, we designed a soft agar based screening method for the identification of gene knockouts promoting a more aggressive phenotype.

Neuroblastom is the most common extra cranial solid tumour in childhood and accounts for 7 -10% of all childhood cancers. In Germany about 150 children are diagnosed with neuroblastoma every year. As a tumour of the autonomic nervous system, neuroblastoma derives from neural crest tissue and usually arises in a paraspinal location in the abdomen or chest [1,8]. Thanks to improved therapies neuroblastoma often initially responds very well to the treatment. However, at relapse there is only very little to offer for this patients and hence relapses correlate with poor prognosis and fatal outcome (Figure1a).

The median age at diagnosis is 17 months and the incidence of neuroblastoma is 10.2 cases per million children under 15 years [5].Thus, neuroblastoma develops in early life

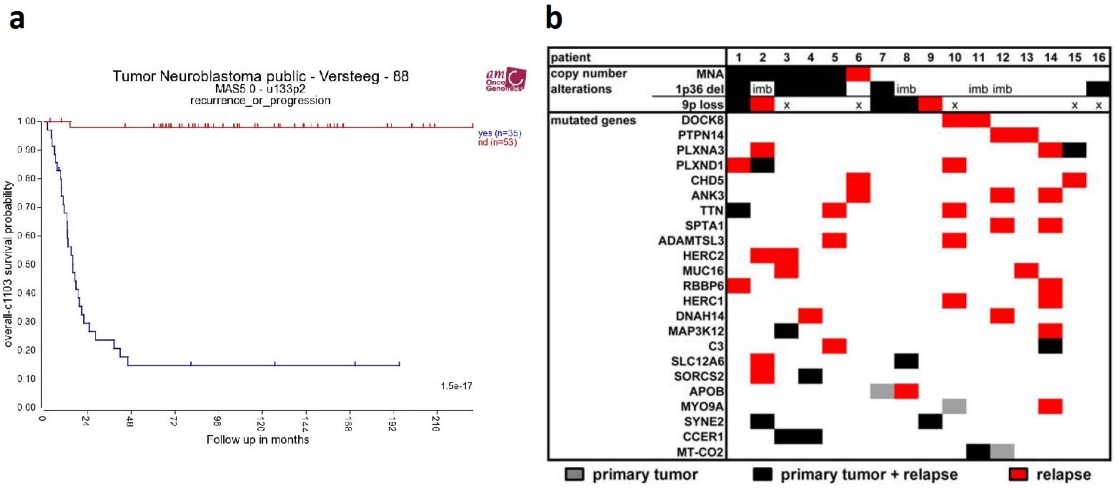


Figure 1: **Survival and gene mutations associated with neuroblastoma relapse.** (a) Kaplan Meier curve from "R2: Genomics Analysis and Visualization Platform" [4] comparing the survival of the patients with neuroblastoma reoccurrence or progression (blue) to those without (red). (b) Genomic alterations in matched primary and relapse neuroblastomas. Copy number alterations of MYCN (MNA), 1p36 and 9p were detected by array CGH (x = no samples for array CGH available) Mutated genes (SNVs) identified by exome sequencing are indicated as coloured boxes. (Adapted from Schramm et al., 2015 [7]).

and therefore one can hypothesis that environmental influences are rather unimportant for the development of this type of cancer. The fact that the genetic background seems to be the major driving force in neuroblastoma, makes it an excellent model for the investigation of individual oncogenes. In 2015, C1 subproject members used whole-exome sequencing, mRNA expression profiling, array CGH and DNA methylation analysis to characterize 16 paired samples at diagnosis and relapse from individuals with neuroblastoma. We were able to show that the mutational burden significantly increased in relapsing tumours, accompanied by altered mutational signatures and reduced subclonal heterogeneity. Furthermore, we found indications for clonal mutation selection during disease progression. In total, we identified 23 genes, which were mutated in at least two samples, as shown in Figure 1b. We hypothesize that these genes are involved in relapse formation. However, the relevance of these genes in the context of neuroblastoma needs to be validated biological. Therefore, the aim of this Project is to investigate the function of the identified genes in neuroblastoma cell lines.

For this purpose the CRISPR / Cas9 system will be used to create neuroblstoma knockout cell lines for the previously reported genes. The CRISPR / Cas9 system originated as a bacterial defence system, which cuts invading viral DNA at a defined position, and thus can effectively be used for genome editing. [3] This system consists of a protein (Cas9) and an RNA part. Cas9 binds to so-called single guide RNAs (sgRNAs), which direct the Cas9-sgRNA complex to a complementary DNA sequence. Here, Cas9 endonuclease induces double-strand breaks in the DNA. In order to create a gene knock out by using CRISPR / Cas9, guide RNAs for 12 genes were designed using a CRISPR Design website [2]. These sequences were cloned into a plasmid also coding for Cas9 protein, which is linked to a green fluorescent protein (AddGene #48140 [6]). By transfection of different neuroblastoma cell lines with this plasmid we were able to show that a Cas9 protein can be efficiently introduced in our cell lines (Figure2) of interest.

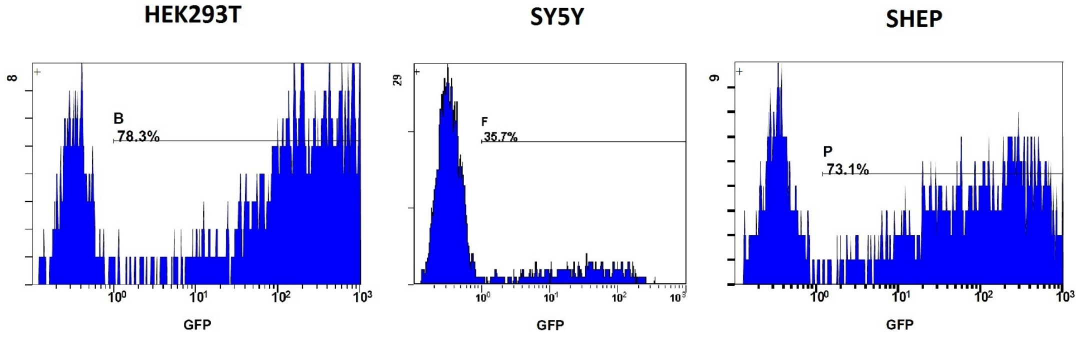


Figure 2: **Transfection efficiency of different cell lines with Cas9-GFP plasmid.** HEK 293T, SY5Y and SHEP cells are transfected with the AddGene plasmid 48140 [6] by electroporation at 1250 V ; 20 ms ; 3 Pulses. After 24h, the cells are checked for GFP fluorescence by FACS analysis, showing transfection rates of 78.3%, 35.7% and 73.1%.

However, as the exact sequence alteration caused by Cas9 induced DNA double strand breaks are not defined, in depth sequencing power is required to identify the resulting genotype of affected cells. As we are most interested in a phenotype resembling cellular state at relapse and / or metastases, we designed a new assay based on the assumption relapse formation will coincide with enhanced growth of isolated tumour cells. In this colony formation assay neuroblastoma cells, that are unable to grow in 3D spheres in “soft agar” will be transfected with a Cas9 plasmid and guide RNAs to target different genes. Cells will be then be embedded in medium containing reduced amounts of solidifying agar (“soft agar”). Only those cells, that have acquired the capability to grow as 3D spheres will be able to grow under this conditions. Isolation and propagation of these cells will enable us to identify the underlying genetic events by sequencing.

Figure 3 depicts differences in growth of neuroblastoma cell lines in soft agar, providing the prerequisites for this gain of function screen.

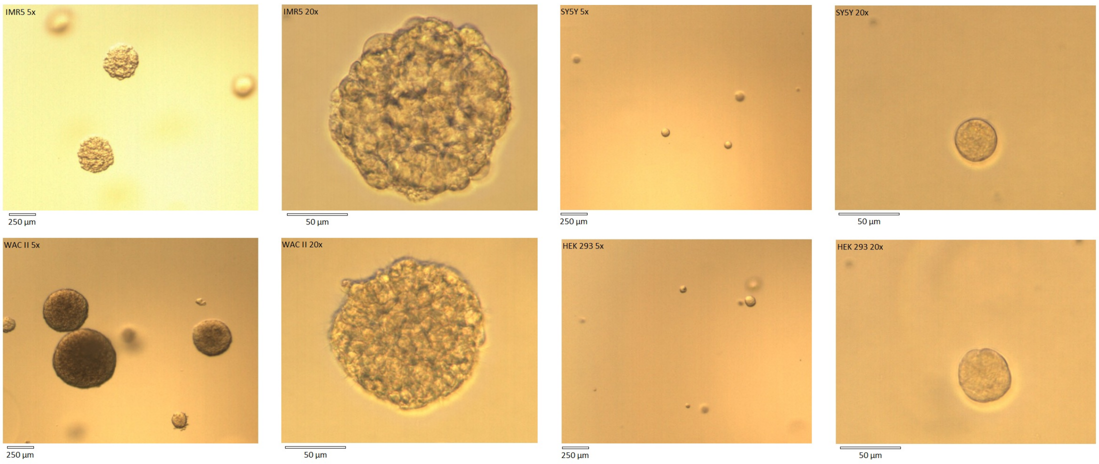


Figure 3: **Soft-Agar-Assay.** IMR5, WACII and SY5Y neuroblastoma cells as well as HEK 293T control cells were embedded in "soft agar" and incubated for 9 d. IMR5 and WACII cells develop 3D sphere-like structures, while the SY5Y and the HEK293T cells do not propagate.

References

[1] Brodeur, G. (2003). Neuroblastoma: biological insights into a clinical enigma. Nature Reviews Cancer, 3(3), pp.203-216.

[2] Crispr.mit.edu, (2015). Optimized CRISPR Design. [online] Available at: http://crispr.mit.edu/ [Accessed 12 Nov. 2015].

[3] Doudna, J. and Charpentier, E. (2014). The new frontier of genome engineering with CRISPR-Cas9. Science, 346(6213), pp.1258096-1258096.

[4] Hgserver1.amc.nl, (2015). R2: Genomics Analysis and Visualization Platform. [online] Available at: http://hgserver1.amc.nl/cgi-bin/r2/main.cgi [Accessed 12 Nov. 2015].

[5] Maris, J. (2010). Recent Advances in Neuroblastoma. New England Journal of Medicine, 362(23), pp.2202-2211.

[6] Ran, F. et al. (2013). Genome engineering using the CRISPR-Cas9 system. Nat Protoc, 8(11), pp.2281-2308.

[7] Schramm, A. et al. (2015). Mutational dynamics between primary and relapse neuroblastomas. Nature Genetics, 47(8), pp.872-877.

[8] Uni-kinderklinik3.de, (2015). Klinik für Kinderheilkunde III | Neuroblastom. [online] Available at: http://www.uni-kinderklinik3.de/haemato-onkologie/stationen-und-ambulanzen/spezialshysprechstunden/neuroblastom.html [Accessed 12 Nov. 2015].

Estimation of Genetic Diversity with ddRAD Sequencing and Error-Tolerant Hashing

Henning Timm
Genome Informatics, Institute of Human Genetics
University Hospital Essen, University of Duisburg-Essen
henning.timm@tu-dortmund.de

This report will highlight the use of error-tolerant hash functions to cluster DNA reads created by double-digest restriction site associated DNA Sequencing (ddRAD-Seq). The ddRAD-Seq protocol allows to reduce the number of DNA reads in order to analyze whole populations of organisms instead of one single organism at a time. Due to technology specific errors, and the structure of the data, the analysis of ddRAD-Seq reads can benefit greatly from an error-tolerant, hash based approach.

Through Second Generation Sequencing (SGS) the DNA sequences of organisms can by decoded and analyzed. We focus on the analysis of data generated by ddRAd-Seq, which restricts the analysis to a subset of DNA sequences. The first section describes the specific structure of ddRAD data. After that the biological, as well as the underlying computer science problem, will be pointed out. The last section focuses on our approach to the problems in context with other solutions.

The ddRAD-Seq Protocol The ddRAD-Seq protocol [2] is used in population genetics to analyze genetic diversity in and between populations of the same species. The RAD-Seq protocol confines the analysis to restriction site associated DNA, in order to manage the amount of data produced by SGS technologies. These DNA sequences are located on the genome next to a restriction site for a specific restriction enzyme. A restriction enzyme binds to the DNA strand and cuts (digests) it at an enzyme-specific sequence motif.

name	orientation	length	purpose
p5 barcode	p5	6bp	identify individual
insert	p5 and p7	0-3bp	prevent over-saturation of sequencer sensors
rc overhang	p5	3-6bp	digestion artifact
fc overhang	p7	3-6bp	digestion artifact
DBR	p7	13bp	identify PCR duplicates

Table 1: List of auxiliary and residual sequences in ddRAD reads.

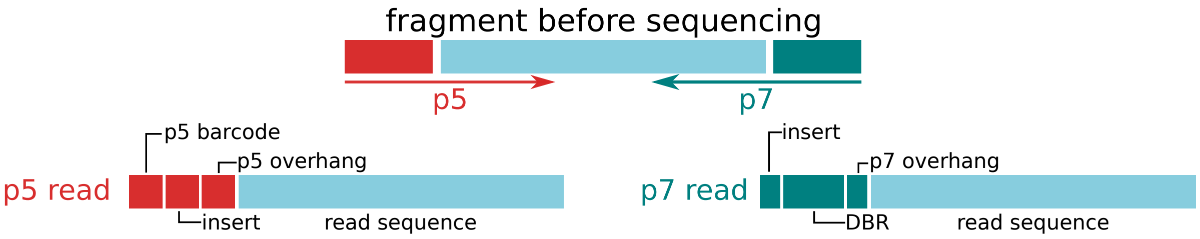


Figure 1: Structure of ddRAD fragments and the resulting reads.

In double digest RAD-Seq two different restriction enzymes are used to digest the DNA. One of these acts as a rare cutter, which binds to rare sequence motifs, and creates fragments of the DNA. The frequent cutter on the other hand binds to a common motif and trims the size of the DNA fragments.

After the treatment with both enzymes, the fragments are selected according to their size and structure. For the analysis only those fragments starting with a rare restriction site and ending with a frequent one are considered. Additionally only sequence fragments with a certain size range are considered. This is mostly due to technical limitations, as SGS technologies have a minimal and maximal length of fragments that can be analyzed. Furthermore shorter sequences do not yield any significant gain in information and might stop more useful fragments from being sequenced.

After the selection step, a number of auxiliary sequences are added to the fragments. In addition to that some artifacts from the digestion process are still part of the sequences. A list of these sequences can be found in Table 1 and the structure of the constructed reads is illustrated in Figure 1. Even more auxiliary sequences are used in the sequencing process (e.g. sequencing primers), but the ones listed here are the only ones that influence the further analysis. In this step, barcode sequences are also added to identify from which individual a read originates. Each individual in the sample can be identified by a combination of two barcodes. Additionally, a degenerate base region (DBR) [3] is added. This partially random sequence is used to uniquely identify a read, in order to remove duplicates created by the sequencer.

Finally, the fragments are sequenced using paired-end sequencing. From each fragment a p5 (or forward), beginning at the rare restriction site, and a p7 (or reverse) read, beginning at the frequent site, are generated. The reads are stored in separate FASTQ-files, split up by their p7 barcodes.

Problem Definition Given a set of ddRAD reads there are two desired levels of information. First, the origin of the reads have to be reconstructed in order to analyze the same genomic position, called a RAD locus, in several individuals. Hereafter, the level of genetic diversity can be analyzed by identifying mutations and genotypes. On top of that, populations can be inferred from the similarities of the genetic layout.

Sequence similarity is used to assign each read to its originating RAD locus. However, this can not be restricted to identical sequences, as mutations, like single nucleotide variations (SNVs), can create differences at the same locus for different individuals. An approach used to solve this is to create perfectly identical loci and merge them according to similarity [1]. This however creates quadratic runtime with respect to the number of loci, which lies in the order of 100 000 loci for smaller animal genomes e.g. caddisflies.

After the RAD loci have been assembled, the next step is to identify SNVs. Before that, several types of harmful reads have to be removed from the loci. These include reads with low sequencing quality or sequencing errors, PCR duplicates, incompletely digested fragments and reads from highly repetitive regions. All of these can skew the analysis and have to be excluded.

Error-tolerant hash functions We propose using error-tolerant hash functions to reconstruct the loci in order to speed up the analysis. By excluding possibly deviant positions from the hashing process, a fingerprint of hash values can be computed. Based on this fingerprint the reads are then assigned to RAD loci. In this context, a read is a sequence $r \in \Sigma^m, \Sigma = \{\, \mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T} \,\}$ and the read length m is assumed to be constant.

To compensate sequence differences due to SNVs only a set of probe positions of each read is considered. The subsequence induced by a set S of probe positions is extracted and hashed using a general hash function $g : \Sigma^\ell \mapsto \{\, 0..U \,\}$, where ℓ is the number of probe positions. The universe size U is chosen to be a prime number in order to minimize unwanted collisions.

We define a sub hash function $h(r; g, S) := g(r[s_0] \oplus ... \oplus r[s_\ell])$ which maps a read $r \in R$ to a hashing universe of size U. By combining several sub hash functions with different probe position sets, different patterns of SNVs can be compensated. Consider a set $\mathcal{S} = \{\, S_i \mid i \in [0, n] \,\}$ of n different probe position sets. Each position induces a sub hash function that can be used to compute the fingerprint sequence of a read r. The fingerprint is defined as:

$$f(r; g, \mathcal{S}) = \{\, h(r; g, S) \mid S \in \mathcal{S} \,\}$$

Each entry in the fingerprint supports assigning the read to one RAD locus. Two reads which differ in one position will receive different fingerprint values from all sub hash functions that cover the difference with their probe positions. The values for all other sub hash functions are identical. Based on the fingerprint, the read is then assigned to a RAD locus. Conflicts from ambiguous fingerprints are resolved using a threshold of decisive fingerprint values that has to be surpassed. Reads that fail to reach this threshold are removed from the analysis.

By using fingerprints the need to perform a pairwise comparison between RAD loci is removed. This eliminates the quadratic component, with respect to the number RAD loci, of the time required to solve this problem.

Outlook Our approach will first be evaluated on simulated data created by our tool RAGE (RAd data GEnerator), which is able to simulate the most common errors to be encountered in ddRAD data. Additionally RAGE produces a detailed annotation of the created reads and the added deviation. This simulated data will be used as a ground truth for parameter optimization and further development. An evaluation on real data from caddisflies has also been prepared. Furthermore, a theoretical analysis of precision and robustness of the approach will be performed.

References

[1] Julian Catchen, Paul A Hohenlohe, Susan Bassham, Angel Amores, and William A Cresko. Stacks: an analysis tool set for population genomics. *Molecular Ecology*, 22(11):3124–3140, 2013.

[2] Brant K Peterson, Jesse N Weber, Emily H Kay, Heidi S Fisher, and Hopi E Hoekstra. Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PloS one*, 7(5):e37135, 2012.

[3] Hannah Schweyen, Andrey Rozenberg, and Florian Leese. Detection and removal of pcr duplicates in population genomic ddrad studies by addition of a degenerate base region (dbr) in sequencing adapters. *The Biological Bulletin*, 227(2):146–160, 2014.

Projekt C3
Multi-level statistical analysis of high-frequency spatio-temporal process data

Tim Ruhe Wolfgang Rhode Katharina Morik

Mining Big Data Streams for Multiple Concepts

Christian Bockermann
Lehrstuhl für Künstliche Intelligenz, LS 8
Technische Universität Dortmund
christian.bockermann@tu-dortmund.de

The processing of continuous streams of data has become a central element in the Big Data analytics field. Complementing the traditional batch data processing, streaming platforms have found their role within the *speed layer* of the Lambda architecture, proposed by Marz and Warren [**?**]. We propose an abstraction layer on top of existing streaming platforms, allowing for a high-level specification of streaming applications by the means of their data flow. The **streams** framework aims at a platform-independent approach for designing applications from building blocks, making the power of streaming architectures accessible to end-user analysts.

1 Introduction

Looking at a generic picture of today's data applications, it is rather common to have data that is produced by some process (e.g. customers shopping) being stored in a database. Running analytical batch processes will reveal some sort of results, e.g. a daily report or a prediction model for future purchases (e.g. to improve pre-ordering). Figure 1 below outlines this scheme.

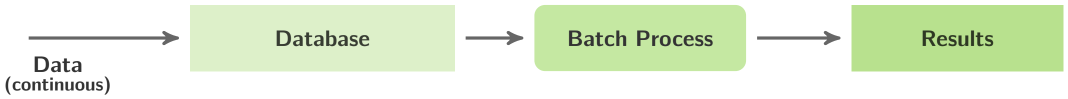


Figure 1: A generic outline of an data oriented application.

This generic architecture poses two problems, which arise when (a) data volume increases and (b) the results being required to reflect even the last minute data that most recently arrived. In the last section we briefly outlined two real world examples. The *Lambda Architecture* is a term that has first been defined by Nathan Marz in [1]. The central observation is that the answer to a query (e.g. financial report, prediction model,...) is a result computed from all the data that is available:

$$\text{result} = \text{query}(\text{all data}).$$

Computing the results by a (massively parallelized) batch job does produce a response with a significant delay, but does not include the data that has been collected in the period from starting the batch job to its completion. Therefore, the generic application scheme of Figure 1 does not meet todays data demands. The *Lambda Architecture* as proposed by Marz introduces three basic components for designing Big Data software systems: The traditional *batch layer*, the *serving layer* and the *speed layer*. Where the batch layer handles the execution of long-term jobs on history data, it typically produces intermediate results for computing the final query response. These intermediate outcomes are stored within the serving layer such that they can be queried in real-time. To bridge the gap between the continuous data that needs to be incorporated into the final query outcome, the speed layer is introduced as a streaming approach that will compute online results and feed these back into the service layer. Queries to the system are answered by aggregating intermediate results from the serving layer. Figure 2 shows the three components of the lambda architecture.

Simply building a software system using these guiding component layout does not meet the requirements of Big Data per se. The implementation of each of the layers needs to be scalable to a large amount of computing nodes and requires loosely coupled fault tolerant components to be in place. The aforementioned *Apache Hadoop* system serves as an example for the batch layer. It provides large scale distributed storage and execution of batch jobs. Software such as *Apache Cassandra* [2], *Google BigTable* [3] or the distributed full-text index *ElasticSearch* set the scene to implement a serving layer within a Big Data system. The speed layer may be provided by streaming platforms such as *Apache Storm* [4], *Apache Samza* [5], Google's *MillWheel* [6] or others.

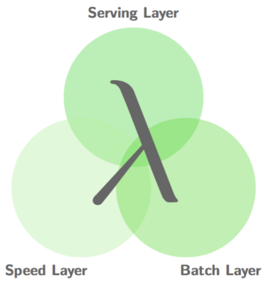


Figure 2: The components of the *Lambda Architecture*.

2 Abstracting Data Flow Definitions with streams

The underlying concept of all streaming platforms [4–6] is the definition of a data flow graph. This graph consists of source nodes and processing nodes, where each source represents an outlet for a continuous flow of data. Processing nodes are user defined functions, which are attached to the data flow, apply data tranformations or computations and in turn may emit new or enriched flows of data.

The **streams** framework provides a higher-level API to define such user-defined functions (UDFs) by means of implementations of a simple Java interface. These *processors* define a set of building blocks, which can be used to define a data flow as required for the aforementioned streaming platforms. For the definition of such *streaming applications*, **streams** offers an XML based specification scheme, that allows for directly referencing existing building blocks by their Java implementation name. Figure 3 shows the XML concept for defining streaming applications. By aiming at a high-level API and a platform neutral abstraction layer, applications defined with the **streams** framework can be mapped to data flow graphs of other platforms as well. For a single-node execution, the *streams-runtime* package offers a standalone exeuction environment that serves as the reference runtime environment for **streams** applications.

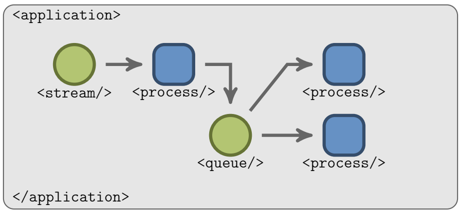


Figure 3: The outline of an application in streams – a graph of connected processes.

3 Applications to Astroparticle Physics

We successfully applied the **streams** abstraction within the analysis process of a telescope data in Cherenkov astronomy. The FACT telescope records extensive air showers as raw sampled voltages over time and requires the deployment of data calibration and feature extraction steps in a real-time pipeline as shown in Figure 4.

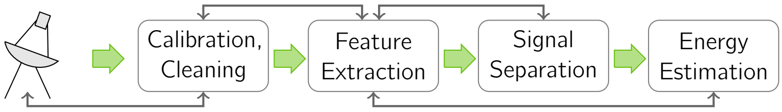


Figure 4: Data flow of the FACT telescope Data Analysis

Building on the basis of the streams framework, we developed a domain specific library of building blocks that is geared towards the FACT data and serves as a rapid prototyping environment for the complete feature extraction pipeline. This allows for processing raw data directly obtained from the telescope to a feature representation, that is suitable to apply machine learning for a advanced distinction of significant events from the unwanted background noise.
For the application of machine learning classifiers, we integrated the WEKA [7] machine learning library into **streams**. The resulting pipeline and the abstract concepts have been published at the ECML/PKDD 2015 and have been awarded the *Industrial Track Best Paper* prize [8].

References

[1] Nathan Marz and James Warren. *Big Data - Principles and best practices of scalable realtime data systems*. Manning Publications Co., 2014.

[2] Avinash Lakshman and Prashant Malik. Cassandra: A decentralized structured storage system. *SIGOPS Oper. Syst. Rev.*, 44(2):35–40, April 2010.

[3] Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach, Mike Burrows, Tushar Chandra, Andrew Fikes, and Robert E. Gruber. Bigtable: A distributed storage system for structured data. In *Proceedings of the 7th USENIX Symposium on Operating Systems Design and Implementation - Volume 7*, OSDI '06, pages 15–15, Berkeley, CA, USA, 2006. USENIX Association.

[4] Nathan Marz et.al. Twitter storm framework, 2011. `https://github.com/nathanmarz/storm/`.

[5] Samza, 2013. `http://samza.apache.org/`.

[6] Tyler Akidau, Alex Balikov, Kaya Bekiroglu, Slava Chernyak, Josh Haberman, Reuven Lax, Sam McVeety, Daniel Mills, Paul Nordstrom, and Sam Whittle. Millwheel: Fault-tolerant stream processing at internet scale. In *Very Large Data Bases*, pages 734–746, 2013.

[7] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November 2009.

[8] Christian Bockermann, Kai Brügge, Jens Buss, Alexey Egorov, Katharina Morik, Wolfgang Rhode, and Tim Ruhe. Online analysis of high-volume data streams in astroparticle physics. In *Proceedings of the European Conference on Machine Learning (ECML), Industrial Track*. Springer Berlin Heidelberg, 2015.

Measurement of the Muon Neutrino Energy Spectrum with Multiple Years of IceCube Data

Mathis Börner
Experimentelle Physik 5b
Technische Universität Dortmund
mathis.boerner@tu-dortmund.de

In my work I am investigating the muon neutrino energy spectrum from ~100 GeV up to multiple PeV. This gives deep insight into the atmospheric muon neutrino energy spectrum and the transition from a spectrum dominated by atmospheric neutrinos to a spectrum completely dominated by astrophysical neutrinos. The approach of this analysis is a signal background separation performed with methods from the field of machine learning with an subsequent unfolding. Similar analysis were performed on single years of data for the years in which the detector was build. Over the years a sophisticated course of actions were developed. This work improves it even further and applies it on a combination multiple years of data.

1 Introduction

In 2013 IceCube found evidences for high energetic astrophysical neutrinos [**?**]. This discovery opened the door to the completely new field of neutrino astronomy. Up to now the sources of those neutrinos are unsettled. In a measurement of the neutrino spectrum on earth, the spectrum is expected to be dominated by atmospheric neutrino up to energies of ~150 TeV. The exact shape of the components around the transition to a astrophysical dominated spectrum is unclear. Hence, a crucial step for the understanding of the astrophysical neutrino spectrum and the identification of sources is a detailed knowledge of the atmospheric spectrum.

The approach pursued in this work was developed by Tim Ruhe for IceCube in a configuration with ~2/3 of IceCube finals size (IC59) [**?**]. In two following analyses [**?**] the general scheme was improved adnd applied on a year with ~92 % of IceCube's final size (IC79) and the first year with a fully completed detector (IC86). In Fig. 1 the results from the previous analyses are shown. The approach was improved from year to year and gives a unique view on the muon neutrino energy spectrum. The basis of the approach is the generation of an extremely pure sample. The needed sparation between atmospheric muons and muon neutrinos is performed with techniques from the field of machine learning in the Rapidminer environment. This muon neutrinos sample is unfolded with the software TRUEE [**?**]. The muon neutrino energy spectrum is topic in several other analyses, but all of them use a parametrized fit. In contrast the unfolding does not rely on any parameterizations and therefore it gives a completely unbiased and model independent view on the spectrum. In the following the different steps of the analysis are presented with an focused on implemented and planned improvements.

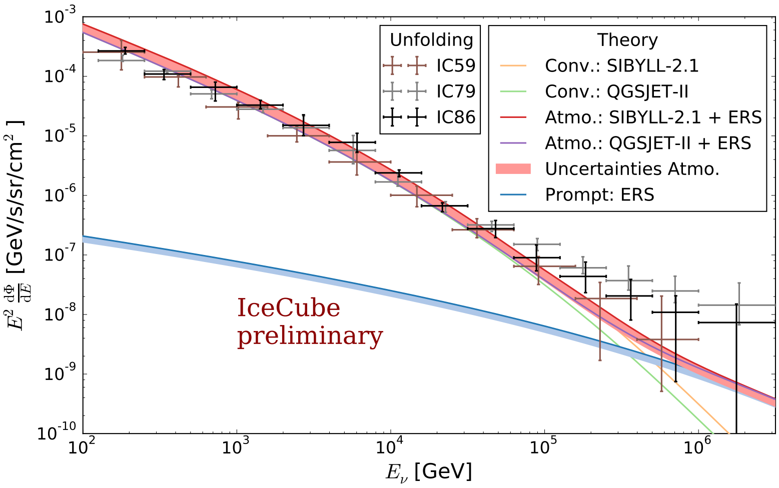


Figure 1: Unfolded muon neutrino energy spectra for IC59 [**?**], IC79 and IC86 [**?**] data compared to theoretical calculations for a purely atmospheric spectrum. For the IC79 and IC86 spectrum the transition from a atmospheric dominated spectrum to a astrophysical dominated spectrum can be observed around 100 TeV. [**?**]

2 Separation

The major challenge to generate a pure muon neutrino sample is the rejection of atmospheric muons. In IceCube the detection of neutrinos is done through charged particles produced in and after the neutrino interaction. For the muon neutrino channel the neutrino is detected through its charged counterpart the muon. The muon itself can only be characterized by its trajectory and its energy. Therefore a muon from a muon neutrino interaction looks exactly like a muon from an atmospheric air shower entering the detector. While neutrinos can propagate through large amount of matter, even high energetic muons are only able to bypass multiple tenth of kilometers through matter. Thus to get rid of the atmospheric muons the earth can be used as a shielding. This fact is utilized in this analysis by only looking for muons in the detector coming through the earth. Those muons have to be produced in a neutrino interaction close to the detector. But even after this selection, for each neutrino induced muon there are multiple tenth of thousands atmospheric muons due to misreconstructions of the particle trajectory.

The separation between well- and misrecstructed muons is performed with a sequence of data mining techniques. The data from the detector consists of ∼2000 different features for each event. After removing highly correlated and meaningless features the numbers of features in lowered to ∼300. To reduce the dimensions even further the Minimum Redundancy Maximum Relevance algorithm in the fast implementation of Schoewe et al. [**?**] is used. An improvement to the previous analyses is the implementation of new features focusing on getting better quantities for the reconstruction quality. For example several features are obtained from multiple reconstructions on a bootstrap detector signal. Those reconstructions give insight into the stability of the algorithms. In case of huge differences between those reconstructions the likelihood for a misreconstruction is expected to be high.

The selected features are used to train a Random Forest to create the final neutrino sample with a cut on the Random Forest score. A simple cut on the Random Forests score would be relative ineffective, because the the signal and background have different energy distribution and the background expectation is orders of magnitude lower for high energies. This fact can be utilized with an energy dependent cut on the Random Forest score. For this analysis a step-wise cut is implemented that is optimized for a demanded purity on small energy intervals.

In contrast to the previous analyses this work is applied on three years of data and is maybe extend on up to two more years. In total the amount of files need to be processed is above 20 TB for 3 years of data. This requires an extendable and scalable setup for the whole processing.

3 Unfolding

The unfolding will be performed with the software TRUEE [**?**]. In this analysis the energy spectrum of muon neutrinos is sought-after. It is not possible to measure the neutrino energy directly because of the indirect detection. Even the energy of the charged lepton (muon) can not be measured directly. The detector only measures the light emitted in the detector. On this detector signal sophisticated reconstruction algorithms are used to reconstructed individual events. Part of the reconstructions are observables with a high

correlation to the sough-after energy. Those observables carry transformed and smeared informations of the sought-after neutrino energy. The idea of unfolding is to revert the smearing and transformation based on simulations of the whole detection process.
The spectrum is unfolded over more than four orders of magnitude and has an expected energy dependency of $N_{\text{Events}} \propto E^{-2.7}$. The biggest limiting factor for the unfolding is the decreasing number of events in regions of high energy. With multiple years the statistical errors in those regions can be expected to be significantly lowered and maybe the spectrum can be extended to even higher energies.
The measured spectrum is the sum all processes creating muon neutrinos. The different contributions are expected to have diverse zenith dependencies. Due to the high number of events for multiple years the spectrum can be unfolded in small zenith bands. The different spectra allow to disentangle the contributing processes and provides individual information about them.

4 Outlook

The analysis of multiple years of IceCube data opens possibilities to investigate the muon neutrino energy spectrum with significantly improved precision in comparison to previous analyses. The unbiased and model independent unfolding provides unique information about the shape and normalization of the spectrum. This information can be used to test and tune theoretical predictions for the spectrum.

References

[1] M.G. Aartsen et al. Evidence for high-energy extraterrestrial neutrinos at the icecube detector. *Science*, 342(6161):1242856, 2013.

[2] M.G. Aartsen et al. Development of a general analysis and unfolding scheme and its application to measure the energy spectrum of atmospheric neutrinos with icecube. *The European Physical Journal C*, 75(3):1–14, 2015.

[3] M.G. Aartsen et al. Unfolding measurement of the atmospheric muon neutrino spectrum using icecube-79/86. *The IceCube Neutrino Observatory-Contributions to ICRC 2015 Part II*, 2015.

[4] K. Morik B. Schowe. Fast-ensembles of minimum redundancy feature selection. In *Ensembles in Machine Learning Applications*, pages 75–95. Springer, 2011.

[5] N. Mielke et al. Solving inverse problems with the unfolding program truee: Examples in astroparticle physics. *Nuclear Instruments and Methods in Physics Research*, 697:133–147, 2013.

Influence of SiPM Crosstalk on the Performance of Photon Charge Extraction in FACT-Tools

Jens Björn Buß
Experimentelle Physik 5
Technische Universität Dortmund
jens.buss@tu-dortmund.de

The First G-APD Cherenkov Telescope (FACT) is pioneering the application of silicon-based photo sensors for the imaging atmospheric Cherenkov technique. These sensors, namely silicon photo multipliers (SiPM), are more robust to bright light conditions than the established photo multiplier tubes. However, optical crosstalk is considered to be a drawback of SiPMs, as this effect adds up to the photon charge. Consequently, it is necessary to understand its impact on the charge extraction algorithm that is implemented in the data analysis software FACT-Tools. With the aim to study the systematic effects of crosstalk on the extraction, the cameras response to γ-showers was simulated for different crosstalk probabilities. With these, the performance of the extraction algorithm is determined concerning its bias and resolution, depending on the simulated number of photons.

1 Introduction

The first G-APD Cherenkov Telescope (FACT) is equipped with 1440 silicon photo multipliers (SiPM) as camera pixels. By use of these devices, the FACT collaboration successfully proved the application of semi-conductor based photon detectors (G-APDs aka. SiPMs) for the imaging atmospheric Cherenkov technique [1]. However, optical crosstalk was considered to be the major difficulty of SiPMs since it adds a background component, which cannot be distinguished from actual Cherenkov photons. Nevertheless, FACT has been able to show the minor influence of this effect on the data of an IACT, which is presented in more detail in [2].

In addition, recent improvements of the camera simulation contemplate for bringing data and Monte Carlo simulations in agreement. In turn, this allows to simulate SiPMs with different crosstalk probabilities and investigate their behavior. In the following, the effect of crosstalk on FACT's data analysis chain FACT-Tools [3] is presented, whereas this study focuses on the first step, the photon charge extraction algorithm.

2 Methods

For this study, a sample of $\approx 12{\cdot}10^6$ γ-showers is simulated to investigate the behaviour of the analsis chain for different crosstalk probabilities. Therefore, the crosstalk probability is variated in a range from 1 % to 30 %. Other effects, e.g., NSB or Dark counts, are not variated in the simulation. The cleaning levels are kept constant. More details are given in [4]. The performance of the photon charge extraction is investigated by comparing the algorithms results to the Monte Carlo truth. This is done with the equation

$$C_{div} = \frac{C_{extr} - C_{sim}}{C_{sim}} \quad , \tag{1}$$

by comparing the simulated and extracted number of Cherenkov photons in a certain pixel. Equation 1 is then applied to all events from the given data sample. The resulting distribution of C_{div} is used to compile mean and standard deviation for a certain true number of photons, which determines the bias and resolution of the respective algorithm.

3 Results

The photon charge extraction's bias and resolution are presented in Figure 1. Interestingly, no significant differences are visible for the bias and resolution in the range of 1 % to 10 % crosstalk. Only for a small photon content, a slight trend to a larger bias and a worse resolution can be accounted for larger crosstalk probabilities.

However, for a crosstalk of 30 %, a larger bias and a worse resolution for small charges of 1 p e to 10 p e is visible. With this photon quantity the bias of the extractor appears to be about 5 % worse than with a crosstalk probability below 10 %. For larger quantities this effect decreases to a one percent difference.

By contrast, the resolution is less effected for large numbers of Cherenkov photons per pixel. In the context of uncertainties, the difference to smaller crosstalk probabilities is negligible.

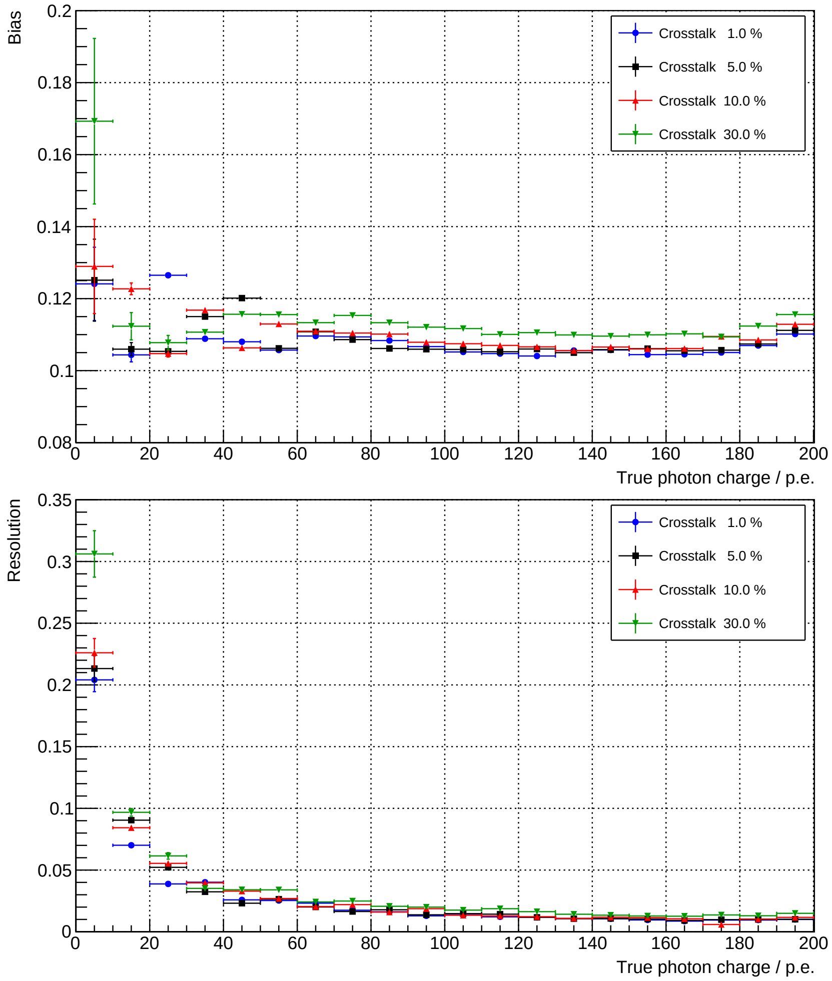


Figure 1: Dependency of bias (top) and resolution (bottom) of the charge extraction algorithm to different crosstalk probabilities, performed on cleaned shower pixels.

4 Discussion

The increased bias on the extraction of small p.e. pulses at 30 % crosstalk can be explained by the large effect of crosstalk to small signal amplitudes, which inherit from Cherenkov pulses as well as random photons. Due to crosstalk these amplitudes are raised. On the one hand, this leads to an over-estimated charge in case of Cherenkov photons. On the other hand, a random photon might be mistaken for a Cherenkov photon. Accordingly, with larger crosstalk probabilities, an over-estimated charge is more likely, and as such the bias is larger.

Concerning larger pulses, this effect is reduced, since more Cherenkov photons are involved. An extraction of NSB photons is less likely. Consequently, more photons without crosstalk contribute to the signal. Nevertheless, the residual contributions of crosstalk leave a remaining bias to the extraction. In case of smaller crosstalk probabilities these contributions are reduced and also the bias is smaller.

Regarding the declined resolution of small p.e. pulses, the main reason is that the deviation C_{div} shows larger variations. If random photons are mistaken for Cherenkov photons, a wrong charge is extracted, which leads to a fluctuation of C_{div}. With larger crosstalk, the extraction of random photons is more likely and the resolution get worse.

If more Cherenkov photons contribute to the signal, it is less likely that a random photon is extracted instead. As a consequence the deviation C_{div} shows less variations and the resolution improves in general. In this case the effect of crosstalk is negligible, since it does not cause additional fluctuations.

5 Conclusion and Outlook

Crosstalk is mainly an issue for the extraction of pulses that contain only a few Cherenkov photons, since the extraction of a random photon is more likely. Especially, large crosstalk probabilities diminish the performance of the extraction algorithm in terms of bias and resolution.

Nevertheless, it has been shown in [2] that FACT's SiPMs feature a crosstalk in a range between 11 % to 13 %. In these bounds differences of bias and resolution are negligible. In the worst case of signals with 1 p.e. to 10 p.e., they still show acceptable values of less than 14 % bias and 23 % resolution .

In order to prove the assumption that random photons in combination with crosstalk are mainly responsible for the decrease of performance, future studies will investigate the effect of random photons in more details.

References

[1] H. Anderhub et al. Design and operation of FACT – the first g-apd cherenkov telescope. *JINST*, 8(06):P06008, 2013.

[2] A. Biland et al. Calibration and performance of the photon sensor response of FACT - the first G-APD cherenkov telescope. *JINST*, 9(10):P10012, 2014.

[3] Christian Bockermann and Hendrik Blom. The streams Framework. Technical Report 5, TU Dortmund University, 2012.

[4] Jens Buss et al. Fact - influence of sipm crosstalk on the performance of an operating cherenkov telescope. *ICRC proceedings 2015*, (863), 2015.

Analysis of high energetic muons with IceCube using IC86-I

Tomasz Fuchs
Lehrstuhl für Experimentelle Physik 5
Technische Universität Dortmund
Tomasz.Fuchs@tu-dortmund.de

By analyzing high energetic muons in IceCube a measurement of the prompt muon flux is possible. To do this muon bundles are analyzed. These bundles consist mostly of just one muon which carries most of the energy of the bundle. These events are selected using the feature selection mRMR and a random forest. The selection of these events is possible using a random forest. A spectrum of the sample is then reconstructed using the unfolding software TRUEE.

1 Introduction

In the field of astro particle physics galactic or extragalactic particles reaching the earth are analysed. This can be done by satelites, balloons or ground based experiments. IceCube is a ground based particle detector at the geographical south pole and is able to detect high energetic particles which are created in the atmosphere or are extraterrestrial.

When high energetic protons or other elements enter the atmosphere multiple particles are produced. Most of these particles have a long lifetime and lose energy while propagation in the atmosphere. Due to the energy losses their energy spectrum is steeper than the spectrum of the primary particles. At higher energies particles with a short lifetime are created. These particles are called prompt because they decay before they lose energy [1]. Because the decay process occurs befor any energy losses they have the same spectral index as the primary particles.

The cross section to create these particles is not covered by accelerator experiments since the most probable angle is in forward direction. Reconstructing the flux of these prompt muons can provide knowledge about the production cross sections. Also parton distribution functions can be calculated from these cross sections.

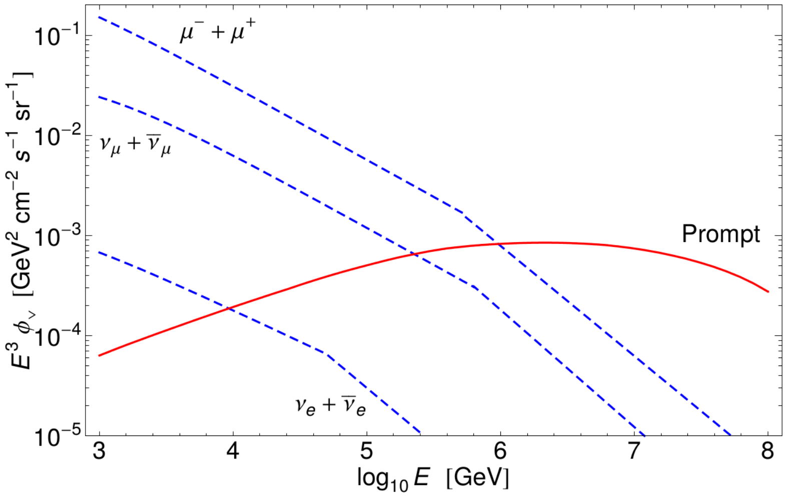


Figure 1: Flux of conventional (blue) and prompt leptons (red). [1]

2 Analysis

When muons enter the IceCube detector they arrive in a bundle. The number of muons in a single event is in a range from 1 to over 1000 muons. To detect the most energetic muon in these bundles the topology of these events is important. Two of these events are shown in figure 2. The left event of figure 2 shows a high energetic muon which carries more than 50 % of the total energy of the bundle. These events will be called leading-muon-events because they have a muon which has most of the energy of the event. The right event of this figure shows a typical muon bundle which does not contain such high energetic muon but the energy of the bundle is distrubuted between all muons in the bundle. These events will be called background-events. The leading-muon-events produce a high energetic stochastic loss in the detector. Using this high energetic stochastic loss to select such events is a good approach since only high energetic muons are able to produce these signatures.

To select the leading-muon-events an advanced data-mining approach is needed since the stochastic loss can be hidden in a mean energy loss of multiple muons. This behavior can lead to very similar topologies of these events to the background-events. To apply state of the art data mining procedures to the data and use sophisticated algorithms the software Rapidminer [2] is used.

For the selection over 400 attributes are available which describe an event reaching the detector. The first step of the analysis was to get a set of attributes which is capable to distinguish between leading-muon-events and background-events. To get this set of attributes the mRMR (minimum Redundancy Maximum Relevance) [3] feature selection is used. This algorithm selects a set of features which are highly correlated to the leading-muon-event class and are least correlated to the other features. With this it was possible to reduce the number of attributes to the most important 30 attributes.

With the selected 30 attributes one needs an algorithm to separate the leading-muon-events from the background-events. One of the best algorithms to seperate signal from background events is the random forest. In this analysis a random forest implementation by WEKA is used. This algorithm builds multiple decision trees and selects a random subset of attributes on each node which is considered for the seperation. The best of the random choosen attributes is then selected. To classify the leading-muon-events a random forest with 200 trees, 30 attributes and 5 attributes per node was trained. Also the training of the random forest was validated using a five fold cross validation.

A cut on the random forest score of 0.9 is then applied and the resulting spectrum is unfolded using the software TRUEE [4].

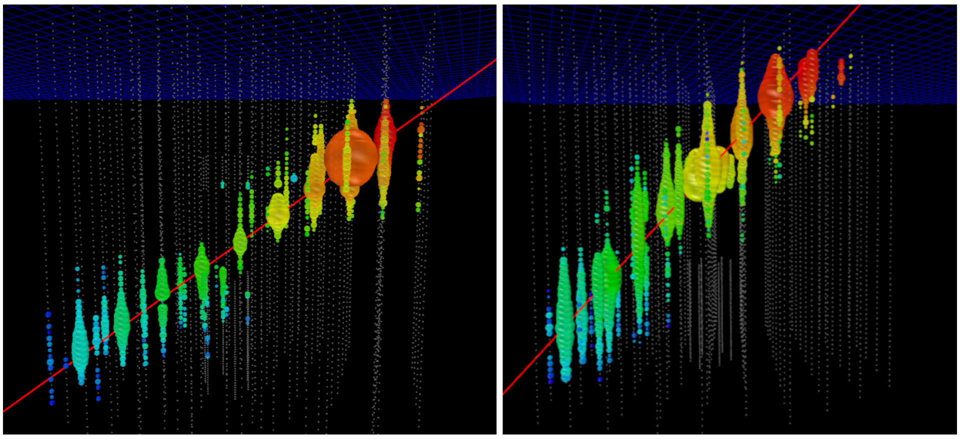


Figure 2: Event with an leading-muon-event (left) and a background-event (right).

3 Results and Outlook

The separation of leading-muon-events and background-muon-events is possible when using the random forest with the settings as mentioned in the section before. This behavior was already shown in the tech report of 2014 and compared with the distribution for data.

The resulting spectrum of high energetic muons ist shown in figure 3. The shown spectrum of Berghaus is the last measurement of high energetic muons and the blue lines are the error bands for the resulting spectrum. Sybill 2.1 is the theoretical model for high energetic muons. No break in the reconstructed spectrum of the atmospheric muons can be seen for high energies. This region is limited by the low amount of Monte-Carlo-Data which was available to reconstruct this spectrum. As last steps the data of the full year will be unfolded but no improvement is expected since the dominating factor for this analysis is the limited amount of Monte-Carlo-Data.

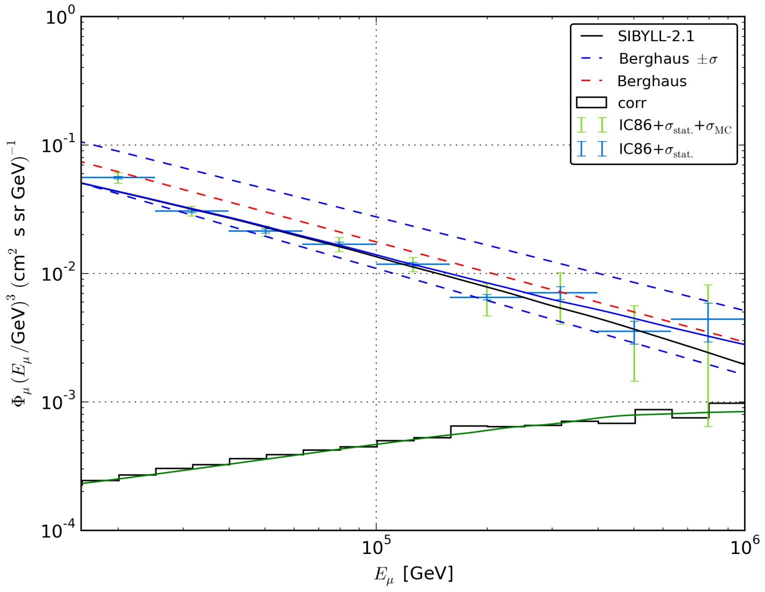


Figure 3: High energetic muon spectrum with the burn sample of the IC86-I data.

References

[1] Rikard Enberg, Mary Reno, and Ina Sarcevic. Prompt neutrino fluxes from atmospheric charm. *Physical Review D*, 78(4):043005, August 2008.

[2] Ingo Mierswa. A flexible platform for knowledge discovery experiments: Yale–yet another learning environment. 2003.

[3] Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE transactions on pattern analysis and machine intelligence*, 27(8):1226–38, August 2005.

[4] Milke, N. and et al. Solving inverse problems with the unfolding program TRUEE: Examples in astroparticle physics. 2013.

Unfolding of the Neutrino Energy Spectrum for Different Source Types

Thorben Menne
Experimentelle Physik 5b
Technische Universität Dortmund
thorben.menne@tu-dortmund.de

In this work the neutrino energy spectrum of several source types is investigated by using the stacking method. For this purpose signal regions of multiple point sources of the same type are superimposed to yield a better signal over background ratio. By using the unfolding software TRUEE the resulting neutrino flux is model independent. As no point source signal has been seen yet flux limits on the considered source types can be set only. This analysis aims to improve the flux limits originating from an earlier analysis based on data taken with the IceCube detector in the 59 string configuration. The goal is to achieve an improved unfolding procedure by combining three years of IC86 data with better statistics and smaller systematic errors. Furthermore an all-sky search for a-priori unknown source positions on the same dataset is done for comparison to other all-sky searches.

1 Introduction

The IceCube detector is a cubic kilometer sized neutrino detector located at the southpole. It consists of 5160 digital optical modules (DOMs) mounted on 86 strings in depths between 1455 m to 2450 m directly in the antarctic ice. Additionally 81 instrumented tanks of ice are installed on the surface to detect air showers. [2]

The direct detection of neutrinos is not possible because they only couple weakly to matter. Instead neutrinos are measured indirectly via secondary leptons produced in charged current interactions. Those leptons create tertiary charged particles in further interactions with the ice. All fast enough charged particles emit Cerenkov light which is measured and used to reconstruct the original neutrino properties.

The physics of astrophysical sources can be further studied by measuring neutrinos originating from them. Currently the neutrino signal for single source positions on the sky is to low to be seen against the large background of atmospheric neutrinos. Therefore multiple sources of the same type can be bundled into one catalog. The combined signal has a better signal over background ratio than a single source so the time needed to measure a significant signal can be reduced. This method is called stacking. [4]

However, the measured energy distribution is not the true sought after neutrino energy. It is convoluted with the charged current cross section and effects from limited detector resolution and acceptance. The used unfolding procedure provides a model independent estimation of the true neutrino energy. [1]

2 Likelihood approach and all-sky point source search

The general likelihood approach in IceCube for a point source search is described in [5]. The most basic likelihood function $\mathcal{L}$ can be expressed as

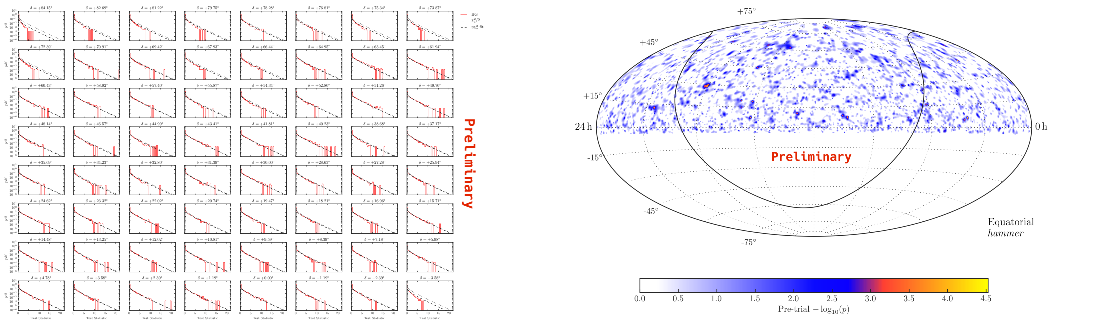
$$\mathcal{L}(n_S) = \prod_{i=1}^{N} \left[\frac{n_S}{N} \, S_i + \left(1 - \frac{n_S}{N} \right) B_i \right] \tag{1}$$

where n_S is the number of assumed signal events and S_i, B_i the expected signal and background probabilities per event i.

To search for a priori unknown point sources in the sky the likelihood is evaluated on each point in the sky on a grid finer than the spatial detector resolution. The value of the resulting test statistic is compared to the value obtained with a background only hypotheses, which yields a pre-trial p-value. The background is estimated by scrambling real data events in the azimuth which is justified by the assumption of only having a small amount of signal events. A function $\eta \, \chi^2_n$ is then fitted to describe the background only test statistic for each declination, where η accounts for underfluctuations. Figure 1a shows this dependency. From this comparison a significance sky map can be constructed as shown in figure 1b. At every point the pre-trial p-value is shown, which has still to be trial corrected before a valid conclusion on the significance can be made. The most significant spot can be seen at ($\alpha = 59.8°$ | $\delta = 6.2°$).

3 Stacking analysis

For a stacked analysis only a small modification to the likelihood function (1) is needed [3]. The signal hypotheses is replaced by a sum over M weighted hypotheses for each source



(a) Background only test statistic for different source declinations shown in red. The dotted line is the χ_1 function and the dashed line the fitted $\eta\ \chi_n$ distribution. In each declination band 10 000 random trials where used to build the test statistic.

(b) Allsky map showing the pre-trial p-values. At each point on the grid the test statistic of the unscrambled data is compared to the value from background only to obtain the p-value.

Figure 1: Background test statistic and significance sky map for 10 % of the IC86 diffuse neutrino sample.

in the catalog

$$S_i \to S_i^{\text{tot}} = \frac{\sum_{j=1}^{M} W^j R^j S_i^j}{\sum_{j=1}^{M} W^j R^j} \tag{2}$$

where W^j is the relative theoretical weight, R^j the relative detector acceptance for each source and S_i^j is the signal probability for each event i regarding source j.

Only the spatial event information is used for the signal and background hypothesis to prevent a bias for the energy unfolding. Based on the likelihood function a number of events with the highest signal to background ratio is selected and used for the unfolding step. The event selection is optimized for the unfolding step.

4 Unfolding

For the unfolding of the neutrino energy spectrum the software TRUEE [7] is used. The advantage is, that a model independent energy spectrum can be obtained. Unfolding is done by building a so called response matrix on Monte Carlo data. This matrix incorporates all measurement effects mapping the true sought after observable to the actually measured one.

Several methods are built-in to make sure the unfolding procedure yields reliable results. In test mode the Monte Carlo used to build the response matrix is unfolded with many parameter combinations to find the best settings. Pull mode is used to find any systematic bias introduced by the unfolding. It basically executes the test mode with the previously chosen parameters many times on a random subset drawn from Monte Carlo data for every subrun. This results in distributions describing the difference of the unfolded result to the MC truth and can reveal failures and systematic deviations which can then be accounted for.

5 Outlook

In this work an unfolding of the astrophysical neutrino energy spectrum on three years of IC86 data is done. To obtain a sufficient amount of events for the unfolding stacking is used to group combine multiple sources of the same type. The unfolding yields a model independent estimate of the energy spectrum. Additionally an all-sky point source search is done with a-priori unknown source positions on the same data set.

References

[1] M. G. Aartsen et al. The IceCube Neutrino Observatory Part I: Point Source Searches. In *Proceedings, 33rd International Cosmic Ray Conference (ICRC2013)*, pages 24–27, 2013.

[2] M. G. Aartsen, N. van Eijndhoven, J. C. Groh, F. Huang, H. P. Bretz, L. Classen, J. Daughhetee, G. Yodh, P. Berghaus, G. W. Sullivan, et al. Letter of intent: The precision icecube next generation upgrade (pingu). Technical report, Inst, 2014.

[3] R. Abbasi, Y. Abdou, T. Abu-Zayyad, J. Adams, J. A. Aguilar, M. Ahlers, K. Andeen, J. Auffenberg, X. Bai, M. Baker, and et al. Time-integrated Searches for Point-like Sources of Neutrinos with the 40-string IceCube Detector. *Astrophys. Journal*, 732:18, May 2011.

[4] A. Achterberg et al. On the selection of AGN neutrino source candidates for a source stacking analysis with neutrino telescopes. *Astropart. Phys.*, 26:282–300, 2006.

[5] Jim Braun, Jon Dumm, Francesco De Palma, Chad Finley, Albrecht Karle, and Teresa Montaruli. Methods for point source analysis in high energy neutrino telescopes. *Astropart. Phys.*, 29:299–305, 2008.

[6] T. K. Gaisser. *Cosmic Rays and Particle Physics*. Cambridge University Press, 1990.

[7] N. Milke, M. Doert, S. Klepser, D. Mazin, V. Blobel, and W. Rhode. Solving inverse problems with the unfolding program TRUEE: Examples in astroparticle physics. *Nuclear Instruments and Methods in Physics Research A*, 697:133–147, January 2013.

Energy Spectrum of the Crab Nebula, obtained by Analysis of FACT Data.

Fabian Temme
Experimentelle Physik 5
Technische Universität Dortmund
fabian.temme@tu-dortmund.de

The First G-APD Cherenkov Telescope (FACT) [1] is an Imaging Air Cherenkov Telescope detecting very high energy gamma rays which are emitted by astrophysical sources. FACT is located at the Observatorio del Roque de los Muchachos on the Canary Island of La Palma (Spain) and is the first telescope of its kind which uses Geiger-mode Avalanche Photo Diodes (G-APDs) as photo sensors.

The Crab Nebula, a supernova remnant with a pulsar in its center, is considered as a "standard candle" in very high energy gamma ray astronomy, due to the constant flux of very high energy gamma rays it emits. In order to evaluate the performance of an Imaging Air Cherenkov Telescope like FACT an analysis of the Crab Nebulas very high energy gamma ray flux and in particular its energy spectrum is essential.

1 Introduction

Very high energy gamma rays induce an air shower of secondary particles if they hit the atmosphere. This secondary particles emit Cherenkov light which is collected by the telescopes mirrors and registered by the photo sensors.

The first step of the analysis chain is the preprocessing of the raw data of these photo sensors. Therefore the data analysis tool FACT-Tools [3] has been developed. It performs a calibration of the raw data, an extraction of the registered cherenkov photons, a cleaning of the shower image and a parameterization of this image.

The parameters calculated by FACT-Tools are used to perform an estimation of the

energy of the primary particle by applying a random forest regression.
As the very high energy gamma rays are overwhelmed by the far more numerous charged cosmic rays, a separation between these two types of particles has to be performed. Therefore, a random forest classification is applied to the data set using the data mining framework RapidMiner [4] .
Subsequently the resulting set of gamma ray events is unfolded using the software TRUEE [5] in order to obtain the differential energy spectrum of the Crab Nebula.

2 Preprocessing and Parameterization

The preprocessing and parameterization of the raw data is performed with the data analysis tool FACT-Tools [3] . FACT-Tools is based on the *streams*-framework [2] and provides a various set of processors to analyse FACT's raw data. The first step is a calibration of the voltage curves of the photo sensors. In a second step the Cherenkov photons registered in the photo sensors are extracted. Then, pixels containing the air shower are identified by applying a cleaning algorithm to the data. In a last step, different parameters, describing the properties of the air shower image, are calculated.
Figure 1 visualizes the distributions of two exemplary parameters calculated by FACT-Tools for cosmic ray MCs and data of the Crab Nebula. The comparison between MCs and data shows an agreement in the general shape of the distributions.

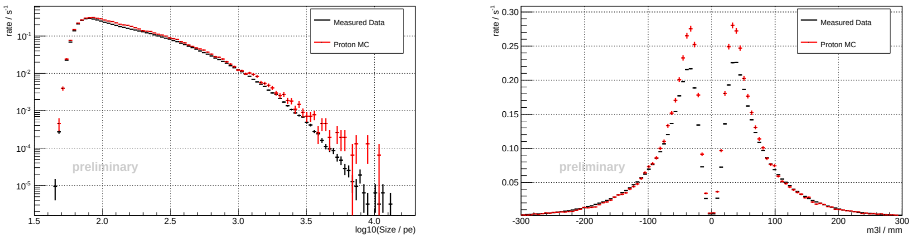


Figure 1: Comparison of parameter distributions of Crab Nebula data and proton MCs.

3 Energy Estimation

Using the parameters calculated by FACT-Tools, an estimation of the energy of the primary particle is performed. Therefore, a random forest regression is trained on gamma MCs and applied to the Crab Nebula data, as well as, to independent MCs used in the following analysis steps. Figure 2 shows the calculated estimated Energy (E_{rec}) against the true energy (E_{MC}) obtained by MC events. A clear correlation is visible although an unambiguous conclusion from the reconstructed to the true energy is not possible.

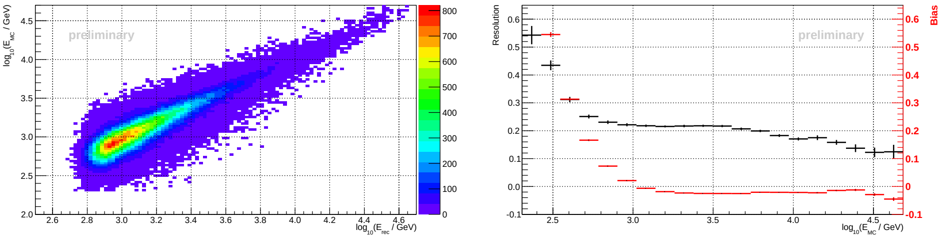


Figure 2: *Left:* Correlation between true energy (E_{MC}) and reconstructed energy (E_{rec}) by evaluating gamma MCs. *Right:* Energy resolution and bias obtained by evaluating gamma MCs.

Therefore, an unfolding of the energy distribution is necessary.
By validation of the performance of the random forest regressor, also the energy bias and energy resolution of the telescope can be estimated. Figure 2 shows the results.

4 Background Suppression

In order to separate the air showers induced by very high energy gamma rays from the far more numerous air showers induced by charged cosmic rays a random forest classification is performed. Therefore, the data mining framework RapidMiner [4] is used, which not only support the application of a large variety of multivariate methods for classification tasks, but also the validation of a given model. The random forest classificator is trained on gamma and proton MCs, and it is validated on an independent MC test sample by a 10 fold bootstrap validation.

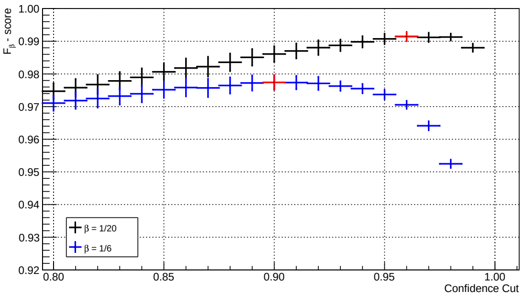


Figure 3: $F_{\frac{1}{20}}$ and $F_{\frac{1}{6}}$ score evaluated with a 10-fold validation on an independent MC set.

As a quality factor for the classification performance the F_β score [6] is used.
For different approaches the factor β can be tuned to suit the different requirements to the data set. For the source detection $\beta = \frac{1}{20}$ and for the determination of the energy spectrum $\beta = \frac{1}{6}$ are chosen.
Figure 3 shows the F_β score for different confidence cuts. The applied confidence cut is chosen according to the maximum of the F_β score.

5 Energy Spectrum

The software TRUEE [5] provides a well tested unfolding algorithm to obtain the energy distribution of the separated data set.
It is capable of using up to three image parameters during the unfolding fit. This fit is based on a regularized likelihood fit with an application of a Tikhonov regularization. TRUEE also provides different modes for the validation of the chosen settings and several quality checks for the unfolding.
The application of the unfolding algorithm to the Crab Nebula data yields to the energy spectrum shown in figure 4. It is in good agreement with results from other telescopes.

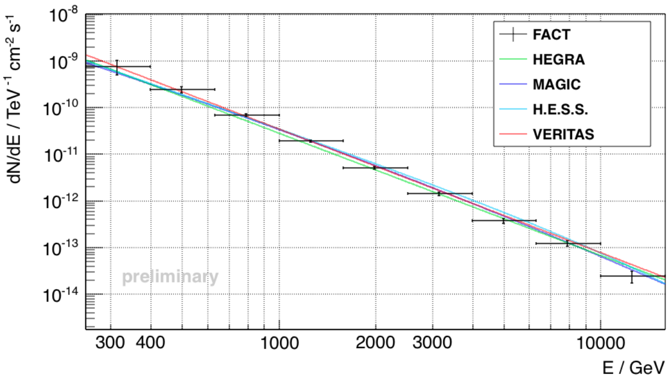


Figure 4: Unfolded energy spectrum of the Crab Nebula obtained with FACT compared to results from other telescopes.

References

[1] Anderhub et al. Design and operation of fact – the first g-apd cherenkov telescope. *JINST*, 8(06):P06008, 2013.

[2] Christian Bockermann and Hendrik Blom. The streams framework. Technical Report 5, TU Dortmund, 12 2012.

[3] Kai Brügge et al. Fact-tools: Streamed real-time data analysis. In *these proceedings*, number 865, 2015.

[4] Ingo Mierswa, Michael Wurst, Ralf Klinkenberg, Martin Scholz, and Timm Euler. Yale: Rapid prototyping for complex data mining tasks. In *Proceedings of the 12th ACM SIGKDD*, 2006.

[5] N. Milke, M. Doert, S. Klepser, D. Mazin, V. Blobel, and W. Rhode. Solving inverse problems with the unfolding program truee: Examples in astroparticle physics. *NIM A*, 697(0):133 – 147, 2013.

[6] M. Sokolova et al. Beyond accuracy, f-score and roc: A familiy of discriminant measures for performance evaluation. In *Proceedings of the 19th Australian Joint Conference on AI: Advances in Artificial Intelligence*, 2006.

Signal-Background Separation Study of FACT Data

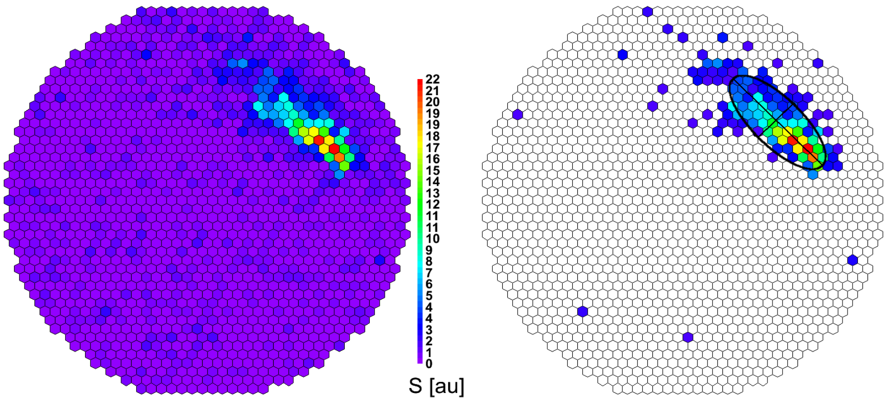
Julia Thaele
Experimentelle Physik 5
Technische Universität Dortmund
julia.thaele@tu-dortmund.de

An important aspect in astroparticle physics is the separation of signal events from background events. The First G-APD Cherenkov Telescope (FACT) detects air showers induced by gamma and hadronic particles coming from distant astrophysical sources. In order to separate the wanted gamma showers from the unwanted hadronic showers a Random Forest algorithm is trained with a set of Monte Carlo Simulations and is applied to real data recorded with FACT using the data mining environment RapidMiner. In this report the results of the training and testing of the built models are presented.

The so-called Imaging Air Cherenkov Telescopes (IACTs) are able to detect very high energy gamma-rays of galactic or extragalactic objects like supernovae or Active Galactic Nuclei (AGN). Due to the neutral electric charge gamma-rays are not influenced and deflected by intergalactic magnetic fields. Thus the direction they are coming from points directly to the astrophysical source. When very high-energetic gamma or hadronic particles are hitting the upper atmosphere layers of Earth, they induce an extensive air shower which emits a blueish light, the so-called Cherenkov light [7].

FACT is the first IACT which uses Geiger-mode Avalanche PhotoDiodes (G-APDs) instead of photomultipliers as photosensors to detect this light [1]. Due to a signal to background ratio of 1:1000 the separation of gamma showers from hadronic showers is very important to increase the sensitivity of the telescope and thus the effective observation time. The building and testing of the separation model is done with a Random Forest (RF) algorithm [3], which is implemented in the RapidMiner analytics platform [9]. In particular the RF of the implemented WEKA [8] package is used for this analysis. The models are trained and tested on gamma- and proton Monte Carlo (MC) simulations for FACT, which are further processed by the analysis software *Modular Analysis and*

Reconstruction Software (MARS) [4] and as well by the analysis software *FACTTools* developed within the SFB876. [2]. As *MARS* is already used and accepted by other working groups, it is used as a comparison to validate the results of *FACTTools*. The resulting shower images are cleaned from background noise of starlight and other non-physical data. After data processing quality cuts are applied to each data set to filter out nonphysical shower events and to cut already away a large amount of background events. In Fig.1 an exemplary shower is depicted in the uncleaned and cleaned state. The RF is trained with parameters describing distinctive features of the shower images and thus allow to distinguish between gamma showers and hadronic showers . The features for this analysis are selected within the *Minimum Redundancy Maximum Relevance (MRMR)* algorithm [5]. A grid search is applied to optimize the number of features chosen by MRMR and as well as the features selected by the RF algorithm. For the RF 100 trees are built and 18 randomly chosen feature were taken out of a total amount of 41 image parameters for *MARS*, while 12 out of 40 parameters were chosen for *FACTTools*. The training dataset consists of 40000 resp. 48000 events for each class.

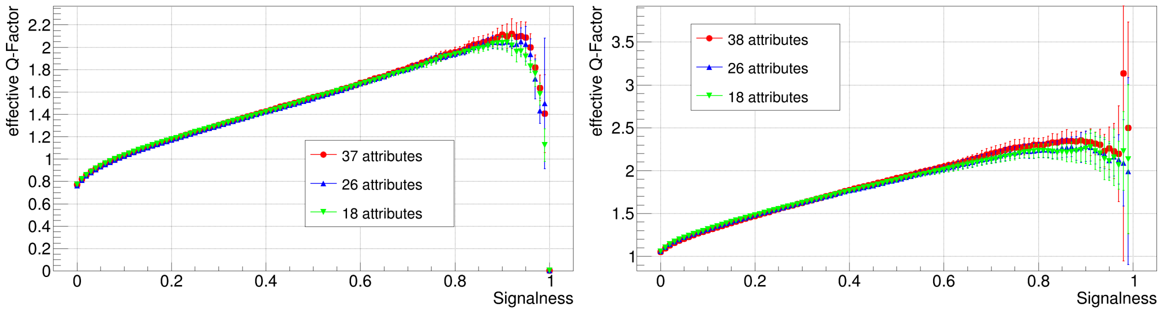


(a) Uncleaned shower image.

(b) Cleaned shower image.

Figure 1: Exemplary shower images in uncleaned and cleaned state processed by *MARS* and *FACTTools*.

To ensure and estimate the stability of the performance, the training and testing is performed within a ten-fold cross validation. In this way statistical means and error values can be determined and an overtraining can be prevented. Both software programs contain a certain intersecting-set of features for the RF training. *FACTTools* provides also a certain amount of additional features. The challenge is to find a minimum confidence cut, here described as the *Signalness*, at which not too much gamma events are cut away while the purity of the dataset is still high. One possibility to decide which *Signalness* cuts offer the best results is to determine a so-called quality factor Q, which describes basically the gain of a simple statistical significance of a gamma-ray signal coming from an astrophysical source. Thus, the *Signalness* with the maximum Q-Factor can be se-



(a) Q-Factor against Signalness for *MARS*

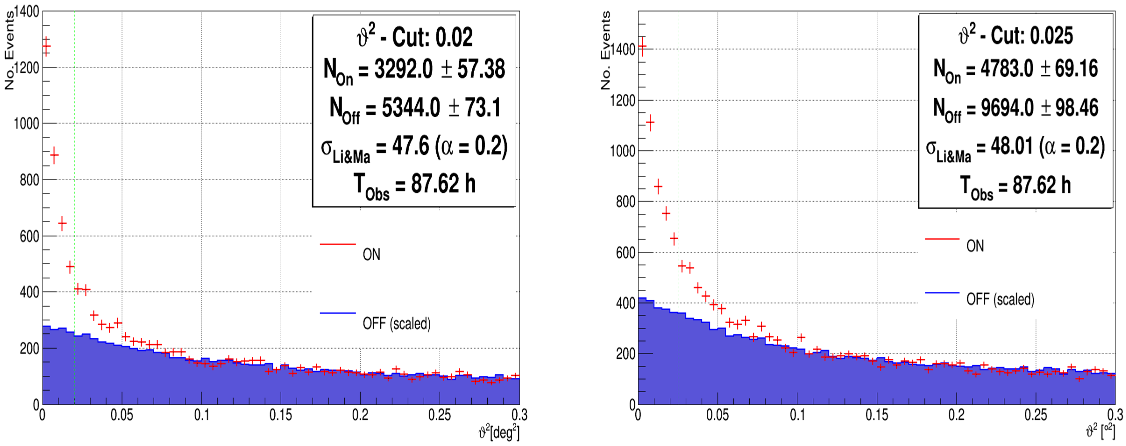
(b) Q-Factor against Signalness for *FACTTools*

Figure 2: Effective Q-Factor against Signalness of different attribute settings for *MARS* and *FACTTools*.

lected. It describes the ratio of the efficiency for gammas to the efficiency of hadronic showers and is

$$Q = \frac{E_G}{\sqrt{E_P}},$$

whereas E_G describes the gamma efficiency and E_P the proton efficiency. In Fig.2 the effective Q-Factors are depicted against the *Signalness* for different attribute settings for *MARS* and *FACTTools*. Here are all previous applied cuts taken into account. Comparing these it can be seen that the performance of *FACTTools* is slightly better. The resulting models were applied to real data of the Crab Nebula processed by *MARS* and *FACTTools*. The *Signalness* cuts were choosen to gain high Q-Factors with acceptable errorbars. In Fig.3 the distributions of the distances of the reconstructed to the real source position are shown. In both cases a significant detection can be seen.



(a) Source detection with *MARS*

(b) Source detection with *FACTTools*

Figure 3: Significant source detection of the Crab Nebula with *MARS* and *FACTTools*.

For both source detections the *Signalness* cuts with the maximum Q-Factor are applied to the data. Again, comparing both results, the significance of the source detection is not only consistent with the results from the MC simulation, but also *FACTTools* shows a slightly better significance on data. The separation leads to a detection of $5.13\,\sigma/\sqrt{h}$, leading to a significant detection of the Crab Nebula in under an hour.
A selection of *MARS* and *FACTTools* processed and separated MC simulation data are used for a new developed method for an investigation, where the RF behaves exceptionally well and bad. This was also developed within a SFB876 collaboration and the associated Techreport can be found in [6].

References

[1] Innovative camera records cosmic rays during full moon. *International Journal of High-Energy Physics*, Nov 2011.

[2] Christian Bockermann, Kai Bruegge, Jens Buss, Alexey Egorov, Katharina Morik, Wolfgang Rhode, and Tim Ruhe. Online analysis of high-volume data streams in astroparticle physics. In *Proceedings of the European Conference on Machine Learning (ECML), Industrial Track*. Springer Berlin Heidelberg, 2015.

[3] Leo Breiman. Random Forests. *Machine Learning*, 45:pp. 5–32, 2001.

[4] T. Bretz and D. Dorner. MARS - CheObs ed. - A flexible Software Framework for future Cherenkov Telescopes. *WSPC Proceedings*, Nov 2009.

[5] C. Ding and H. Peng. Minimum redundancy feature selection from microarray gene expression data. *Proceedings of the Computational Systems Bioinformatics*, pages 523–528, 2003.

[6] Wouter Duivesteijn and Julia Thaele. Understanding where your classifier does (not) work - the scape model class for exceptional model mining. Technical Report 9, TU Dortmund, 2014.

[7] Claus Grupen. *Astroteilchenphysik: Das Universum im Licht der kosmischen Strahlung*. Vieweg, 2000.

[8] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, pages 10–18, 2009.

[9] I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz, and T. Euler. Yale: Rapid prototyping for complex data mining tasks. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’06, pages 935–940, New York, NY, USA, 2006. ACM.

SFB 876 Verfügbarkeit von
Information durch Analyse unter
Ressourcenbeschränkung

Subproject C4
Regression approaches for large-scale high-dimensional data

Christian Sohler Katja Ickstadt

On embeddings for Bayesian hierarchical regression models

Leo Geppert
Lehrstuhl Mathematische Statistik und biometrische Anwendungen
Fakultät Statistik, TU Dortmund
geppert@statistik.uni-dortmund.de

In project C4, we have applied random projections to Bayesian linear regression in order to obtain an approximation of the posterior distribution on an embedded data set, while guaranteeing that the approximation error is small. This Technical Report presents an extension of the model, the application of random projections to Bayesian hierarchical linear regression models. The extension works well if the hyperparameters are used to model the location of the likelihood. For scaling parameters and generalised linear models, this does not work in general. More theoretical work and simulation studies are required for these cases.

Bayesian hierarchical linear regression

Regression techniques are widely used for statistical analyses. Bayesian linear regression offers the opportunity to include prior information in the model. A Bayesian linear regression model is based on the same model as a classical linear regression model, $Y = X\beta + u$, where $X \in \mathbb{R}^{n\times d}$ is a matrix of observations, $Y \in \mathbb{R}^d$ is a vector containing the random dependent variables and $u \in \mathbb{R}^n$ is a random vector with independent measurement error, $u \sim N(0, \sigma_u^2 I_n)$. In a Bayesian setting, $\beta \in \mathbb{R}^d$ follows a distribution. Prior information can be incorporated in the so-called prior distribution $p(\beta)$. The aim of the analysis is to obtain the posterior distribution $p(\beta|X,Y)$, which is a compromise between the likelihood $\mathcal{L}(Y|X,\beta)$ and the prior distribution $p(\beta)$,

$$p(\beta|X,Y) \propto \mathcal{L}(Y|X,\beta) \cdot p(\beta).$$

When structural dependence between the variables is present, Bayesian hierarchical models, also called multilevel regression, are a reasonable model choice. An example of such dependencies are the success rates for heart operations which are performed on babies by different hospitals. There might be effects which affect all of the hospitals or there might be different groups of hospitals (e.g. general and specialised in care for babies). Bayesian hierarchical regression can be used to model such effects. To that end, hierarchical models introduce a population distribution, which replaces the prior distribution and captures the dependencies. The population density will typically have one or more parameters, the so-called hyperparameters. Bayesian hierarchical models thus typically have one additional layer of prior information compared to non-hierarchical models:

$$\begin{aligned} \beta &\sim (\gamma, \Sigma_b) \\ \gamma &\sim (m_\gamma, S_\gamma). \end{aligned}$$

m_γ and S_γ are assumed to be fixed hyperparameters here, but they can also be seen as random variables, thus introducing one or even multiple more hierarchical levels to the model. Gelman et al. (2014) [5] or Congdon (2010) [4] both give a comprehensive overview of hierarchical models.

Hierarchical models cannot be analysed analytically. The most common alternatives are based on Markov Chain Monte Carlo (MCMC) methods, which can be slow, but are reliable and allow for easy checking of convergence. Two possible software packages among others are OpenBUGS [7] and the R-package `rstan` [1]. In all cases, the posterior distribution is of interest and will be analysed after completing the MCMC sampling.

Random projections for Bayesian linear regression

In Bayesian analyses, the repeated evaluation of the likelihood can be a bottleneck and makes MCMC methods infeasible for very large data sets. Different approximations have been suggested as a remedy. In Geppert et al. [6], we propose obtaining a random projection of the original data set and analysing the embedded data set instead of the original one. As the size of the embedded data set is independent of the number of observations n, this reduces both the running time and the memory required for the subsequent analysis. We show that the distance between the two posterior distributions (i.e. the approximation error) is small and controlled by a parameter ε.

In our analyses, we consider three random projection methods: the Rademacher Matrix (RAD) [9], the Subsampled Randomized Hadamard Transform (SRHT) [2] and the Clarkson Woodruff Sketch (CW) [3]. For a description of these methods and their running times and dimensions of the embedded data sets, see [6].

We assume the likelihood to follow a normal distribution $N(X\beta, \Sigma_u^2)$, while the prior distribution can either follow a normal distribution $N(\mu_\beta, \Sigma_\beta)$ or a uniform distribution over $\mathbb{R}^d$. The latter choice is an improper distribution with $\int_{-\infty}^{\infty} p(\beta) = \infty$. This does not pose a problem, however, as it is guaranteed that the posterior distribution will be a proper distribution.

Application to hierarchical models

We apply all three random projection methods to hierarchical data sets. We assume the likelihood to be normally distributed, just like in the non-hierarchical case. The prior distribution necessarily changes as it now needs to incorporate the hierarchical information.

As a starting point to our research, Rathjens (2015) [8] considers a simulation study with several different settings. In all of the simulations, only the likelihood is embedded. The prior distribution is used as originally specified.

The results suggest that random projections are capable of recovering the posterior distribution. This is especially the case when the upper layers of the hierarchy model location parameters, for both homogeneous and inhomogeneous variance matrices in this layer. Scale parameters, however, are much more difficult to recover. This is not surprising as this leaves the scope of the theoretical results even for the non-hierarchical case. The differences between the three embedding methods are minimal for all of these simulations.

Another way of leaving the scope of the theoretical results is to allow for a different likelihood. This includes generalised linear models, which assume a different likelihood and often also include a link function. In general, applying random projections designed for linear regression to generalised linear models does not work well, although there are some surprisingly good results for CW sketches.

Discussion

Random projections for Bayesian linear regression can be applied to Bayesian hierarchical linear regression models. When the additional layers only model the location parameters, this seems to work very well and the resulting posterior distributions are close to the posterior distributions obtained on the original data set. However, this approach does not in general work well when scale parameters are modelled in the upper layers of the hierarchy or when there are deviations from the normally distributed likelihood, such as generalised linear models.

Current and future work aims at better defining the conditions under which random projections can be used for Bayesian hierarchical linear regression and generalised linear models (frequentist and Bayesian), both theoretically and using simulation studies. Sampling approaches may also pose an alternative, especially in the case of generalised linear regression.

Another line of work is the generalisation of Bayesian linear regression, which is based on ℓ_2 to the general case of l_p-regression. This can mean both a change from normal distribution to a distribution induced by ℓ_p, both in the likelihood and a change in the prior distribution. This is subject to current research.

References

[1] RStan: the R interface to Stan, Version 2.8.0, 2015.

[2] Nir Ailon and Edo Liberty. Fast dimension reduction using Rademacher series on dual BCH codes. *Discrete & Computational Geometry*, 42(4):615–630, 2009.

[3] Kenneth L. Clarkson and David P. Woodruff. Low rank approximation and regression in input sparsity time. In *Symposium on Theory of Computing Conference, STOC'13*, pages 81–90, 2013.

[4] Peter D. Condgon. *Applied Bayesian Hierarchical Methods*. Chapman & Hall/CRC, 2010.

[5] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. Texts in statistical science. Chapman & Hall/CRC, Boca Raton, 3rd edition, 2014.

[6] Leo N. Geppert, Katja Ickstadt, Alexander Munteanu, Jens Quedenfeld, and Christian Sohler. Random projections for Bayesian regression. *Statistics and Computing*, to appear.

[7] David Lunn, David Spiegelhalter, Andrew Thomas, and Nicky Best. The BUGS project: Evolution, critique, and future directions. *Statistics in Medicine*, 28:3049–3067, (2009).

[8] Jonathan Rathjens. Hierarchische Bayes-Regression bei Einbettung großer Datensätze. Master's thesis, TU Dortmund University, 2015.

[9] Tamás Sarlós. Improved approximation algorithms for large matrices via random projections. In *Proceedings of the 47th annual IEEE Symposium on Foundations of Computer Science, FOCS 2006*, pages 143–152, 2006.

Unimodal regression as building block for modelling multimodal data

Claudia Köllmann
Lehrstuhl Mathematische Statistik und biometrische Anwendungen
Technische Universität Dortmund
koellmann@statistik.tu-dortmund.de

Research in the field of nonparametric shape constrained regression has been extensive and there is need for such methods in various application areas. It is, for example, often natural that some intensity first increases and then decreases over time, which can be described by a unimodal shape constraint. In this technical report we describe research of project C4 that goes go beyond unimodal regression and we propose to capture multimodality by employing piecewise unimodal regression or deconvolution models based on unimodal peak shapes. The proposed methods work well on data sets from different application areas, for example, marine biology and astroparticle physics.

Introduction

Unimodal regression is a suitable choice in regression problems where the prior information about the underlying relationship between predictor and response is vague, but where it is (almost) certain that the response first increases and then decreases with higher values of the predictor. A semi-parametric spline regression approach to unimodal regression was derived in [5] and its usefulness in dose-response analysis was verified. Simulations indicated that the approach is advantageous in comparison to both parametric and nonparametric competitors. In this report we demonstrate with two real data examples from different application areas that unimodal regression is also useful in situations where the relationship between two variables is not unimodal, but multimodal.

Modelling multimodality

The unimodal regression by [5] is a semi-parametric spline regression approach and it is based on the fact that using the B-spline basis, a spline can easily be restricted to be unimodal. This penalized unimodal spline regression can be used a a building block in models for multimodal data. Multimodality can have different sources and thus, one can take different approaches when modelling multimodal data. For example, one might observe a series of well separated unimodal responses. For such data a **piecewise unimodal regression** is appropiate, that is, dividing the x-axis heuristically between each pair of modes and fitting separate unimodal splines.

Another multimodal regression approach describes the observations as a convolution of peaks, where a global response is observed from a series of overlapping and accumulating unimodal processes. Models trying to reconstruct the individual unimodal responses from the global observed signal are called deconvolution models. A representative of this model class is the **deconvolution with L_0-penalty** introduced by [3]. It describes the observed signal y as a (linear) convolution of scaled versions of a basic peak shape:

$$y_i = \sum_{j=1}^{n_g} g_j a_{i-j},$$

where $\boldsymbol{a} = (a_{-n_g+1}, \ldots, a_{n-1})' \in \mathbb{R}^{n+n_g-1}$ is the vector of the so-called input pulses, which specifies the number of peaks, their locations and heights, and $\boldsymbol{g} = (g_1, \ldots, g_{n_g})'$ describes the pointwise basic peak shape. The model can also be reformulated to a typical linear regression model, $\mathbf{y} = \boldsymbol{G}\boldsymbol{a} + \boldsymbol{\epsilon}$, where $\boldsymbol{G} \in \mathbb{R}^{n\times(n+n_g-1)}$ holds shifted copies of the peak shape $\boldsymbol{g}$ in its columns (cp. [3]).
If the peak shape $\boldsymbol{g}$ is known, the least squares estimate of $\boldsymbol{a}$ is given by $\hat{\boldsymbol{a}} = (\boldsymbol{G}'\boldsymbol{G})^{-1}\boldsymbol{G}\mathbf{y}$, but the columns of $\boldsymbol{G}$ are highly correlated. This problem was already described in [3] and the authors propose to use regularization with an L_0-penalty on $\boldsymbol{a}$, that is, using the objective function $\|\mathbf{y} - \boldsymbol{G}\boldsymbol{a}\|_2^2 + \kappa \sum_j I(a_j \neq 0)$. The regularized estimate $\hat{\boldsymbol{a}}$ is found by minimizing the objective function with an iterative procedure described in their article. Since the penalty factor is essentially the number of peaks, the regularized estimation favours sparse models: the higher the tuning parameter κ, the fewer the peaks. Altogether, the described procedure is able to estimate the number of peaks, their locations and heights simultaneously.
The authors of [3] also present an approach for cases, where the peak shape $\boldsymbol{g}$ is unknown, which is called "blind deconvolution". The idea is, starting with an initial pointwise peak shape $\boldsymbol{g}^{(0)}$, to iterate between estimation of $\boldsymbol{a}$ and $\boldsymbol{g}$. The pointwise least-squares estimate of $\boldsymbol{g}$ is given by $\hat{\boldsymbol{g}} = (\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}\mathbf{y}$, where $\boldsymbol{A}$ holds shifted copies of $\hat{\boldsymbol{a}}$ in its columns and (cp. [3]).

In this technical report we propose to estimate a continuous peak shape, in explicit, a functional and not a pointwise description of each peak, which can be done using two advanced approaches to blind deconvolution with L_0-penalty. The first one is suitable, when the peak shape is known to (approximately) follow a parametric function $g(x|\boldsymbol{\theta})$, whose parameter vector $\boldsymbol{\theta}$ has to be estimated. Starting with an initial parameter vector $\boldsymbol{\theta}^{(0)}$ and an initial peak shape $\boldsymbol{g}^{(0)} = (g(x_1|\boldsymbol{\theta}^{(0)}), \ldots, g(x_{n_g}|\boldsymbol{\theta}^{(0)}))'$ respectively, blind deconvolution is again possible by iterating between estimation of $\boldsymbol{a}$ and $\boldsymbol{\theta}$. The parameter vector $\boldsymbol{\theta}$ can be estimated using the least squares method, that is, minimizing $\sum_{i=1}^{n}(y_i - \boldsymbol{Ag})^2 = \sum_{i=1}^{n}(y_i - \boldsymbol{A}(g(x_1|\boldsymbol{\theta}), \dots, g(x_{n_g}|\boldsymbol{\theta}))')^2$ with respect to $\boldsymbol{\theta}$. If such information about the peak shape does not exist, another possibility to estimate a continuous basic shape is the application of unimodal splines. The procedure is the same as for parametric functions, choosing $g(x|\boldsymbol{\theta})$ as a spline function whose B-spline coefficients $\boldsymbol{\theta}$ are estimated with penalized unimodal spline regression as described in [5].

Applications

The aforementioned approaches to multimodal regression have been employed in two different application areas: marine biology and astroparticle physics.
For the analysis of diving behaviour time-depth-recorders (TDRs) measure the diving depth of marine animals, which repeatedly perform dives from the water surface down to various depths to find food and for other activities. Marine biologists are interested, among other things, in the detection of phases within a dive that correspond to different behaviours (see e.g. [4]). A statistical approach to determine dive phases (descent and ascent) in TDR data is implemented in the R package *diveMove* (cp. [6], [7]). In the current version (1.3.9) the procedure starts with heuristically dividing the diving depth time series into dives using a depth threshold and fitting a smoothing spline to each dive. Afterwards the derivative of the smoothing spline is used to identify the descent and ascent phase of each dive. This determination can be problematic since the uniqueness of the turning point depends on the choice of the smoothing parameter. By replacing the smoothing spline in the first analysis step by a unimodal spline (i.e., using piecewise unimodal regression) we are able to overcome this drawback. The derivative has only one sign change and in contrast to the smoothing spline approach the choice of the tuning parameter has per construction no influence on the uniqueness of the turning point.

The second application comes from astroparticle physics and data has been provided by SFB 876 project C3. The First G-APD Cherenkov Telescope (FACT, see e.g. [1]) is used by astroparticle physicists to detect cosmic rays. The camera of the telescope has several pixels and each pixel collects a signal, that is, a time series of measured voltages. Each photon that hits a camera pixel causes a change in the signal that can be described by a unimodal loading curve (see also [1]). The aim of the physicists is to detect the arrival times and numbers of photons to draw conclusions about the type of the triggering particle. Physicists have derived a parametric wave form for the change in the voltage when one or more photons arrive at a certain time (cp. [2], formula 6.11). As photons can arrive anytime, the measured voltage is an accumulation of several loading curves (each corresponding to one or more photons). This suggests using a deconvolution model with accumulated parametric waves for the analysis of a whole time series of one pixel. We applied blind L_0-deconvolution using the aforementioned parametric function and the model represented the data quite well. The approach enables simultaneous estimation of arrival times, numbers of photons and a continuous description of the loading curve.

Conclusions

We have seen that unimodal regression is not only useful when a unimodal relationship between dependent and independent variable is likely, but also as a building block in modelling situations where the relationship is multimodal. We were able to simplify analysis steps in case of the diving depth example and to enhance existing deconvolution methodology to provide functional instead of pointwise estimates. The results provide an indication for the usefulness of unimodal regression in the presented applications and beyond, but since subsequent analysis steps (e.g. classification into particle types) are common in the described situations, systematic evaluations of the impact of the modelling step on the final outcome are needed and subject to future research.

References

[1] H. Anderhub, M. Backes, A. Biland, V. Boccone, I. Braun, T. Bretz, J. Buß F. Cadoux, V. Commichau, L. Djambazov, D. Dorner, S. Einecke, D. Eisenacher, A. Gendotti, O. Grimm, H. von Gunten, C. Haller, D. Hildebrand, U. Horisberger, B. Huber, K.-S. Kim, M. L. Knoetig, J.-H. Koehne, T. Kraehenbuehl, B. Krumm, M. Lee, E.Lorenz, W. Lustermann, E. Lyard, K. Mannheim, M. Meharga, K. Meier, T. Montaruli, D. Neise, F. Nessi-Tedaldi, A.-K. Overkemping, A. Paravac, F. Pauss, D. Renker, W. Rhode, M. Ribordy, U. Roeser, J.-P. Stucki, J. Schneider, T. Steinbring, F. Temme, J. Thaele, S. Tobler, G. Viertel, P. Vogler, R. Walter, K. Warda, Q Weitzel, and M. Zaenglein. Design and operation of FACT - the first G-APD Cherenkov telescope. *J Instrum*, 8(06):P06008, 2013.

[2] J. B. Buß. Fact - signal calibration: Gain calibration and development of a single photon pulse template for the fact camera. Diploma thesis, TU Dortmund University, 2013.

[3] J. de Rooi and P. Eilers. Deconvolution of pulse trains with the l_0 penalty. *Analytica Chimica Acta*, 705:218–226, 2011.

[4] L. G. Halsey, C.-A. Bost, and Y. Handrich. A thorough and quantified method for classifying seabird diving behaviour. *Polar Biol*, 30:991–1004, 2007.

[5] C. Köllmann, B. Bornkamp, and K. Ickstadt. Unimodal regression using Bernstein-Schoenberg-splines and penalties. Biometrics, 2014. doi: 10.1111/biom.12193.

[6] S. P. Luque. Diving behaviour analysis in R. *R News*, 7(3):8–14, June 2007. Contributions from: J. P. Y. Arnould, L. Dubroca, and A. Liaw.

[7] S. P. Luque and R. Fried. Recursive filtering for zero offset correction of diving depth time series with GNU R package diveMove. *PLoS ONE*, 6(1):e15850, 01 2011.

On extending efficient Bayesian regression to distributions over general ℓ_p spaces

Alexander Munteanu
Efficient algorithms and complexity theory
Technische Universität Dortmund
alexander.munteanu@tu-dortmund.de

Our current research aims at extending previous results on efficient Bayesian regression. In [4] we showed that using random projections for approximating the structure of ℓ_2-subspaces we can achieve considerable computational speed up while preserving a very accurate approximation of the posterior distribution in any strictly Gaussian model. We generalize upon these results to distributions defined over ℓ_p spaces for $p \in [1, \infty)$.

Introduction

A Bayesian regression model can be described in the following way. We model a likelihood function $\mathcal{L}(\beta|X, Y)$ as the product of distributions from a fixed family ranging over the observed data. Additionally we assume we have some prior distribution $p_{\text{pre}}(\beta)$ over the parameter space. In this situation, the *posterior* distribution is a compromise between the prior knowledge and the observed data given by

$$p_{\text{post}}(\beta|X, Y) \propto \mathcal{L}(\beta|X, Y) \cdot p_{\text{pre}}(\beta).$$

In [4] we showed that using random projections to reduce the size of the data and thereby to form a so called ε-subspace embedding for ℓ_2-spaces, we have that the resulting posterior distribution based on the reduced dataset is very close to the original posterior up to an ε-fraction of its location and variation parameters. We can visualize the data reduction in the following diagram where $\Pi \in \mathbb{R}^{k\times n}$ is a random projection matrix, reducing from n to $k \ll n$ dimensions, drawn for example from one of the distributions described in [4]. The data $[X, Y] \in \mathbb{R}^{n\times(d+1)}$ is transformed into a *sketch*, i.e., a much smaller substitute data set $[\Pi X, \Pi Y] \in \mathbb{R}^{k\times(d+1)}$.

$$\begin{tikzcd}
{[X, Y]} \arrow[r, "\Pi"] \arrow[d] & {[\Pi X, \Pi Y]} \arrow[d] \\
p_{\text{post}}(\beta|X, Y) \arrow[r, phantom, "\approx_\varepsilon"] & p_{\text{post}}(\beta|\Pi X, \Pi Y).
\end{tikzcd}$$

The value of k is usually a small polynomial $k \in O(\text{poly}(\frac{d}{\varepsilon}))$ of degree at most 2 depending on the actual random construction and notably independent of n. This leads to a considerable speed up in the computations that need to be performed only on the reduced data.

As a first step, we showed our approximation guarantees to hold for models where the likelihood as well as the prior distributions were arbitrarily chosen from the important class of multivariate normal distributions over the parameter space, i.e., for all $\beta \in \mathbb{R}^d$. Our current research focuses on generalizing to distributions defined over ℓ_p spaces for $p \in [1, \infty)$. Important special cases include the normal distribution for $p = 2$ as well as the Laplace distribution for $p = 1$.

Extending to ℓ_p

The method can in principle be extended to Bayesian ℓ_p regression for all $p \in [1, \infty)$. However, for values of p other than $p = 2$, we have to deal with some additional problems. One thing is that the random projection matrices used for ℓ_2 must be changed to account for the differently shaped normed space. This can be handled by exploiting the concept of *p-stable distributions* [5] which settled the problem of designing subspace embeddings limited to the cases $p \in [1, 2]$. To overcome this limitation, Andoni introduced the notion of *max-stability* of reciprocal exponential random variables [1] which led to a general construction working for all $p \in [1, \infty)$ in [8]. The second issue is that the subspace approximation guarantee on the distortion of these embeddings is not bounded by $1 \pm \varepsilon$ any more. Instead, we only have weak bounds in the order of $O((d \log d)^{\frac{1}{p}})$ for dilation and contraction. Therefore, the direct embedding is used only for a coarse preprocessing step followed by weighted sampling of rows from the original input matrix. The sampling complexity depends on the weak distortion bounds. Taking the size large enough enables us to improve the approximation guarantee to form an ε-subspace embedding for the data matrix, which can be used to solve the problem efficiently and accurately. The third problem we face is that while the embedding dimension is small $k \in O(\text{poly}(\frac{d}{\varepsilon}))$ and independent of n for all $p \in [1, 2]$, the lower bounds of $\Omega(n^{1-\frac{2}{p}})$ on approximating the p^{th}-frequency moments for $p > 2$ imply that the embedding dimension must be polynomial in n and becomes linear as $p \to \infty$. We summarize the ℓ_p-subspace embedding results of Woodruff and Zhang in a simplified Theorem.

Theorem 1 (cf. [8])**.** *For every $p \in [1, \infty)$ there exists a family of random matrices $\Pi \in R^{k \times n}$ such that for any basis U of a d-dimensional p-normed subspace of* $(\mathbb{R}^n, \|\cdot\|_p)$ *we have that with constant probability*

$$\forall x \in \mathbb{R}^d : \Omega(1/(d \log d)^{\frac{1}{p}}) \, \|Ux\|_p \leq \|\Pi U x\|_q \leq O((d \log d)^{\frac{1}{p}}) \, \|Ux\|_p,$$

where

$$(q, k) = \begin{cases} (2, O(poly(d))) & \text{if } p \in [1, 2] \\ (\infty, O(n^{1-\frac{2}{p}} \log n (d \log d)^{1+\frac{2}{p}} + d^{5+4p})) & \text{if } p \in (2, \infty). \end{cases}$$

Combining this with the algorithmic scheme developed and improved in the line of research conducted in [2, 3, 7] leads us to the following algorithm. Let $M = [X, Y]$ be the input data and Π be an ℓ_p-subspace embedding matrix according to Theorem 1.

Algorithm for Bayesian ℓ_p-regression:

1. Compute $\bar{M} = \Pi M$, and its QR-decomposition $\bar{M} = QR$
2. Let $\bar{U} = MR^{-1}$, and for $i \in [n]$ let $\lambda_i = \|\bar{U}_{(i)}\|_p^p$
3. Let $D \in \mathbb{R}^{n \times n}$ be a diagonal matrix, where for sample size s
$$D_{ii} = \begin{cases} \frac{1}{p_i} & \text{with probability } p_i = \min\{1, s \cdot \frac{\lambda_i}{\sum \lambda_i}\} \\ 0 & \text{otherwise} \end{cases}$$
4. Perform Bayesian ℓ_p-regression analysis on the compressed data $C = DM$

The intuition for the algorithm is that in line 1, the data matrix is compressed such that we can efficiently perform an approximate QR-decomposition of the data matrix. Note that unlike the ℓ_2 case considered in [4], we can not conduct the regression analysis directly on $\bar{M}$, since the distortion can only be bounded in the order of $O((d \log d)^{\frac{1}{p}})$, see Theorem 1. However, in line 2 we get a coarse approximation of the *ℓ_p-leverage scores* λ_i. It is well-known that sampling rows proportional to these scores yields an accurate ℓ_p-subspace embedding formalized in the matrix D in line 3. We can argue that choosing the (expected) sample size s large enough, we can compensate for the weak approximation of the leverage scores such that we finally have the desired ε-subspace embedding property

$$\forall x \in \mathbb{R}^d : (1 - \varepsilon) \, \|Mx\|_p \leq \|DMx\|_p \leq (1 + \varepsilon) \, \|Mx\|_p,$$

with constant probability. This will imply by similar arguments as in [4] that the Bayesian regression conducted in line 4 is accurate up to an ε-fraction of the defining parameters of the distributions under study.

Conclusion

This report gives an outline of the algorithmic and theoretical ideas to extend our work on Bayesian ℓ_2-regression [4] to the general case of ℓ_p-regression for $p \in [1, \infty)$. The theoretical analysis, implementation and empirical evaluation of these ideas are subject to our current research. Other lines of research aim at extending the choices of priors to more general settings, including Bayesian analogues of LASSO-regression and important types of hierarchical priors [6]. The extension to generalized linear models can not be handled within the algorithmic framework under study. Therefore we need to develop individual methods tailored to deal with the different requirements and difficulties that arise depending on the choice of the link functions.

References

[1] Alexandr Andoni. High frequency moments via max-stability. Manuskript available at `http://web.mit.edu/andoni/www/papers/fkStable.pdf`, 2013.

[2] Kenneth L. Clarkson. Subgradient and sampling algorithms for ℓ_1-regression. In *Proceedings of the Symposium on Discrete Algorithms (SODA)*, pages 257–266, 2005.

[3] Kenneth L. Clarkson, Petros Drineas, Malik Magdon-Ismail, Michael W. Mahoney, Xiangrui Meng, and David P. Woodruff. The fast cauchy transform and faster robust linear regression. In *Proceedings of the Symposium on Discrete Algorithms (SODA)*, pages 466–477, 2013.

[4] Leo N. Geppert, Katja Ickstadt, Alexander Munteanu, Jens Quedenfeld, and Christian Sohler. Random projections for Bayesian regression. *Statistics and Computing*, (to appear), 2015.

[5] Xiangrui Meng and Michael W. Mahoney. Low-distortion subspace embeddings in input-sparsity time and applications to robust linear regression. In *Proceedings of the Symposium on Theory of Computing (STOC)*, pages 91–100, 2013.

[6] Jonathan Rathjens. Hierarchische Bayes-Regression bei Einbettung großer Datensätze. Master's thesis, TU Dortmund, 2015.

[7] Christian Sohler and David P. Woodruff. Subspace embeddings for the ℓ_1-norm with applications. In *Proceedings of the Symposium on Theory of Computing (STOC)*, pages 755–764, 2011.

[8] David P. Woodruff and Qin Zhang. Subspace embeddings and ℓ_p-regression using exponential random variables. In *Proceedings of the Conference on Learning Theory (COLT)*, pages 546–567, 2013.

SFB 876 Verfügbarkeit von
Information durch Analyse unter
Ressourcenbeschränkung

Subproject C5
Real-time analysis and storage for high-volume data from particle physics

Jens Teubner Bernhard Spaan

Perspective of big data tools for the $\sin(2\beta)$ measurement with the LHCb experiment

Ulrich Eitschberger
Lehrstuhl für Experimentelle Physik 5
Technische Universität Dortmund
ulrich.eitschberger@tu-dortmund.de

The LHCb experiment is one of the four big experiments located at the Large Hadron Collider near Geneva, Switzerland. Its main focus is the search for rare decays and effects of *CP*-violation in decays of beauty and charm hadrons. In order to maximize the sensitivity with respect to these specialized targets the LHCb detector is built as a forward single arm spectrometer.

In the collisions of protons inside the vertex locator (VELO), new particles are created and decay until they finally leave traces in the various subcomponents of the detector. These traces are hits in the tracking systems (VELO, TT, T1-T3, M1-M6), clusters in the calorimeters (ECAL, HCAL) and Cherenkov radiation in the Ring Imaging Cherenkov Detectors (RICH1, RICH2). To conclude on the presence of particles the information of these subdetectors needs to be reconstructed, e.g. by fits of trajectories to ensembles of hits and pattern recognition algorithms looking for clusters of energy deposition. Finally particle candidates need to be combined to heavier particles in order to perform physics measurements on the same. The endeavour to find the particles of interest is hindered for two main reasons. Firstly, hundreds of particles are produced inside the angular acceptance which allows for a large number of combinations to be made in the reconstruction as well as the combination of particles. Secondly, the interaction rate of 50ns / 25ns together with the limitation on the bandwidth that can be written to disk enforces a fast reconstruction that leads to the selection of interesting events and the rejection of physically uninteresting ones. Both these points together set the frame for investigating these tasks in the context of resource limitation: The reconstruction and combination tasks can be parallelized and therefore performed faster.

A physics analysis has been performed to demonstrate the need for an improved processing of the data collected by the LHCb detector. This analysis studies time-dependent charge-parity (*CP*) violation which is one of the keys for understanding the matter-antimatter-asymmetry observed in our universe. In processes involving B mesons *CP* violation has been observed in the "golden mode" $B^0 \to J/\psi K^0_S$ by the BaBar and Belle experiments at the asymmetric e^+e^- colliders PEP-II and KEKB [1, 3] for the first time. As the $J/\psi K^0_S$ final state is common to both the B^0 and the $\overline{B}^0$ meson decays, the interference between the amplitudes for the direct decay and for the decay after B^0–$\overline{B}^0$ oscillation results in a decay-time dependent *CP* asymmetry,

$$\mathcal{A}(t) \equiv \frac{\Gamma(\overline{B}^0(t) \to J/\psi K^0_S) - \Gamma(B^0(t) \to J/\psi K^0_S)}{\Gamma(\overline{B}^0(t) \to J/\psi K^0_S) + \Gamma(B^0(t) \to J/\psi K^0_S)} = \frac{S \sin(\Delta m\, t) - C \cos(\Delta m\, t)}{\cosh(\frac{\Delta\Gamma\, t}{2}) + A_{\Delta\Gamma} \sinh(\frac{\Delta\Gamma\, t}{2})}. \quad (1)$$

Here, $B^0(t)$ and $\overline{B}^0(t)$ indicate the flavour of the B meson at production, while t indicates the decay time. The parameters Δm and $\Delta\Gamma$ are the mass and the decay width differences between the heavy and light mass eigenstates of the B^0–$\overline{B}^0$ system, and S, C, and $A_{\Delta\Gamma}$ are *CP* observables. As $\Delta\Gamma$ is negligible for the B^0–$\overline{B}^0$ system [2], the time-dependent asymmetry simplifies to $\mathcal{A}(t) = S \sin(\Delta m\, t) - C \cos(\Delta m\, t)$.

The analysis can be performed with $B^0 \to J/\psi K^0_S$ candidates reconstructed in the $J/\psi \to \mu^+\mu^-$ and $K^0_S \to \pi^+\pi^-$ final states. This experimentally very clean signature makes it an ideal testbed for the LHCb collaboration to prove its capability of tagged precision measurements. The complete Run I data sample is used which corresponds to an integrated luminosity of $3\,\text{fb}^{-1}$ at centre-of-mass energies of 7 and 8 TeV.

A selection is required to suppress the combinatorial background. The first step is a centrally organized preselection for all analyses performed by the LHCb collaboration. This takes several weeks. Afterwards, the individual tuple of the size of several hundred GB is produced. Another cut-based selection is applied which mainly consists of kinematical and geometrical requirements. This leaves 41 500 flavour-tagged signal candidates.

The values of the *CP* violation observables S and C are estimated by maximizing the likelihood of a probability density function (PDF) describing the unbinned distributions of the reconstructed mass m, the decay time t and its uncertainty estimate σ_t, the OS and SSπ flavour tag decisions d_{OS} and $d_{\text{SS}\pi}$, and the corresponding per- candidate mistag probability estimates η_{OS} and $\eta_{\text{SS}\pi}$. The fit is performed simultaneously in 24 independent subsamples. In each category the data distribution is modelled using a sum of two individual PDFs, one for the B^0 signal and one for the combinatorial background.

The likelihood is a function of 83 free parameters, including S and C, and 48 yield parameters for the signal and the background components. Eleven parameters are external inputs, including the production asymmetry, the flavour tagging calibration parameters, and the mass difference Δm [4]. These are constrained in the fit within their statistical uncertainties and taking their correlations into account. The complexity of the likelihood fit results in execution times of a few hours on multi-core machines.

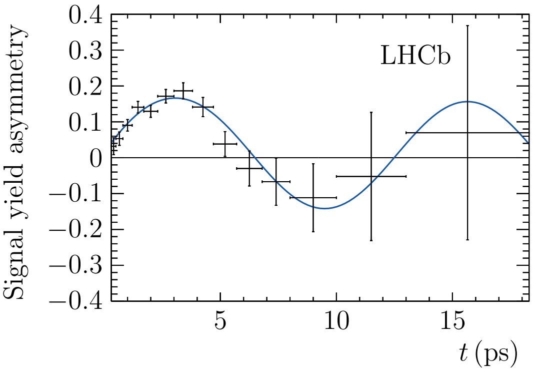


Figure 1: Time-dependent signal-yield asymmetry $(N_{\overline{B}^0} - N_{B^0})/(N_{\overline{B}^0} + N_{B^0})$. Here, N_{B^0} $(N_{\overline{B}^0})$ is the number of $B^0 \to J/\psi K^0_S$ decays with a B^0 $(\overline{B}^0)$ flavour tag. The data points are obtained with the *sPlot* technique [5], assigning signal weights to the events based on a fit to the reconstructed mass distribution. The solid curve is the projection of the signal PDF.

Various sources of systematic uncertainties on the *CP* observables are examined, in particular from mis-modelling PDFs and from systematic uncertainties on the input parameters. Adding all contributions in quadrature results in total systematic uncertainties of ± 0.020 on S and ± 0.005 on C.

In conclusion, the *CP* observables S and C are measured to be

$$\begin{aligned} S &= 0.731 \pm 0.035\,(\text{stat}) \pm 0.020\,(\text{syst}), \\ C &= -0.038 \pm 0.032\,(\text{stat}) \pm 0.005\,(\text{syst}), \end{aligned}$$

with a statistical correlation coefficient $\rho(S, C) = 0.483$. This measurement represents the most precise time-dependent *CP* violation measurement at a hadron collider to date. Fig. 1 shows the decay-time dependent signal-yield asymmetry.

The LHCb experiment will multiply the amount of collected data during the upcoming LHC Runs. To be able to fully exploit the potential of these large datasets analysis tools need to be scalable in terms of both storing and processing. In this context the usability of the open source frameworks Apache Hadoop and Apache Flink will be examined. The Apache Hadoop framework meets exactly the required criteria, as its Hadoop Distributed File System (HDFS) enables scalable distributed storing capabilities, while the processing part is based on the MapReduce programming model. Apache Flink on the other hand relies on HDFS as well, but it closes the gap between MapReduce-like systems and shared-nothing parallel data base systems, which are the fastest known solution to handle large numbers of queries on massive databases. Both frameworks are already widespread in the industry and proved their functionality in various scenarios, but they are still not being used widely in science, especially in the field of physics.

A measurement of *CP* violation as described above could profit in several ways by combining key technologies of the two frameworks in the course of the analysis. Improvements in the handling of large amounts of data would make looser preselections possible and thus retain more potential signal candidates. Necessary to achieve this is the transformation of the experiments data from existing data formats into modern versatile data formats like HDF5. The cut-based offline selection could then be replaced by a selection based on more sophisticated machine learning techniques. This makes it necessary to perform multiple computationally intensive trainings of Neural Nets and Boosted Decision Trees on large data samples. Network bottlenecks due to heavy data access can be overcome by exploiting the individual storage of cluster nodes. Making use of the MapReduce programming model at the same time should lead to reasonable processing times. In summary, the potential of big data tools usage in particle physics analyses will be investigated to pave the way for the most complex and precise future measurements at the LHC.

References

[1] K. Abe et al. Observation of large CP violation in the neutral B meson system. *Phys.Rev.Lett.*, 87:091802, 2001.

[2] Y. Amhis et al. Averages of b-hadron, c-hadron, and τ-lepton properties as of summer 2014. 2014. updated results and plots available at: `http://www.slac.stanford.edu/xorg/hfag/`.

[3] Bernard Aubert et al. Observation of CP violation in the B^0 meson system. *Phys.Rev.Lett.*, 87:091801, 2001.

[4] K. A. Olive et al. Review of particle physics. *Chin. Phys.*, C38:090001, 2014.

[5] Muriel Pivk and Francois R. Le Diberder. sPlot: A statistical tool to unfold data distributions. *Nucl.Instrum.Meth.*, A555:356–369, 2005.

Realtime Analysis and Storage for High-Volume Data in Particle Physics

Michael Kußmann
Lehrstuhl für Datenbanken und Informationssysteme
Technische Universität Dortmund
michael.kussmann@cs.tu-dortmund.de

During the last 4 months I have been looking into the offline data analysis at the LHCb-project, especially the DAVINCI-project [3]. The goal of my current research is to identify and investigate bottlenecks in the current data flow and employing modern database technologies to eliminate bottlenecks. In the long run, it is worthwhile to investigate the possibility of creating a domain specific query language to accommodate most use-cases of DAVINCI.

Introduction

The LHCb-detectors produce raw measurement data at an incredible data rate [1], but only very few events are really interesting. Figure 1 shows the different steps of reducing the amount of data. After all preliminary filtering (by the trigger) is done, around 20 PiB/year remain for offline analysis and storage. To this day, giant disk arrays and subsequently tape drives are used to permanently store the data. The Worldwide LHC Computing Grid (WLCG) [6] is then used to distribute and analyse the data. My research mainly focuses on improving the offline analysis part (marked by a rectangle in Figure 1). To accommodate the planned increase in data rate for 2020, new methods of analysing and storing measured data have to be conceived. In the current phase, bottlenecks are being identified and new data models are being investigated.

[1] 30.000.000 events per second; each event is about 150 KB big, resulting in approx. 4 TB/s total data rate.

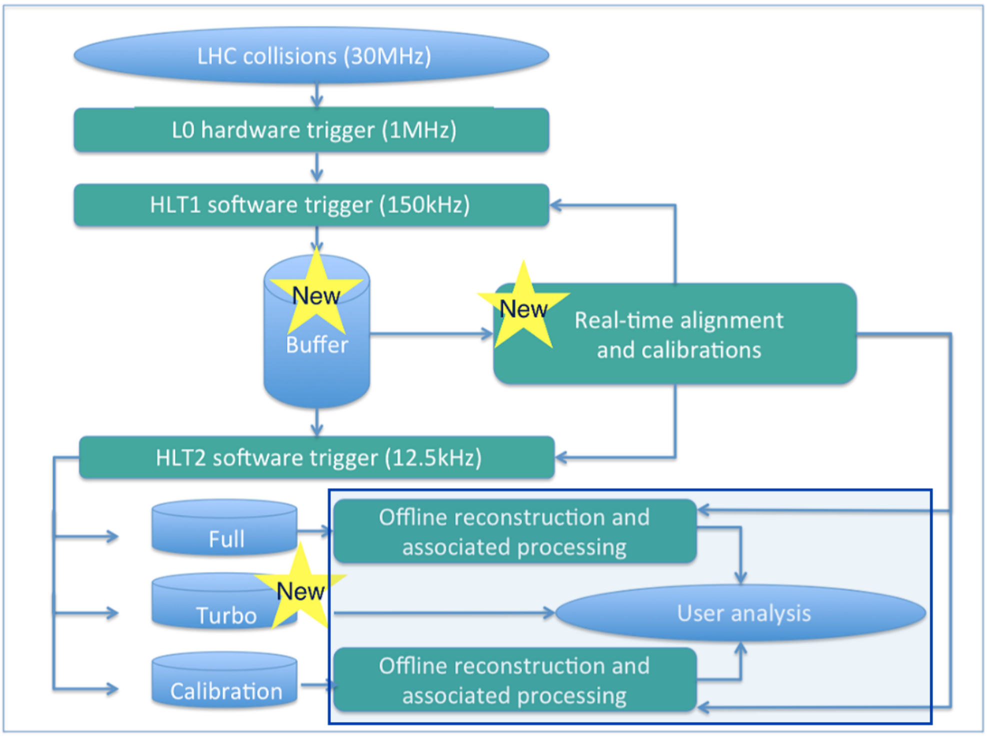

LHC collisions (30MHz)
L0 hardware trigger (1MHz)
HLT1 software trigger (150kHz)
New
Buffer
New
Real-time alignment and calibrations
HLT2 software trigger (12.5kHz)
Full
Turbo
New
Calibration
Offline reconstruction and associated processing
User analysis
Offline reconstruction and associated processing

Figure 1: LHCb Data Flow [4]

ROOT and DAVINCI

ROOT is a data analysis library being developed since 1994 at CERN. It has been designed to cope with the high amount of data the LHC would produce. Due to its old age, ROOT is not up to date with bleeding edge data processing research and Co-processor acceleration (e.g. GPUs). ROOT stores its data in flat files by serializing C++-classes and optionally employing compression. Another downside of ROOT is its lack of external data format support (like HDF5), which makes it hard to export data from ROOT-files.

DAVINCI is a framework, built on top of ROOT, that provides basic tools specialized for offline data analysis at LHC. The user interface of DAVINCI consists of Python bindings to DAVINCI's C++ code. Physicist at LHCb usually configure their analysis algorithms using Python.

Preliminary Performance Analysis

Figure 2 shows that compression is one of the major bottlenecks in offline data analysis. Nearly 30% of the program's total CPU time is spent decompressing the input data. ROOT uses a Deflate-family compression algorithm to reduce flat file size. Deflate is known to be heavy weight, for both compressing and decompressing. There are more CPU efficient general purpose algorithms (like Lempel-Ziv-Oberhumer [9]).

Another result of the profiling is that much time is spent in the new-operator for unsigned long. This looks like an artifact of the way ROOT presents data to the user: ROOT uses serialized C++-classes, which have to be rebuild upon access by the user. The profiling data suggest, that this is done in a particular inefficient way.

Function	**CPU time**
lzma_decode	254.078s
DecayTreeFitter::KalmanCalculator::updateCov	40.539s
__GI_memset	30.209s
DecayTreeFitter::KalmanCalculator::init	25.056s
operator new (unsigned long)	19.135s
[Others]	489.913s

Figure 2: Summary of hotspot analysis (Intel VTune); Total CPU time: 858.929s

Figure 3 shows that the profiled algorithm does not scale horizontally with CPU core count. This may be due to the fact, that the provided example script makes use of DAVINCI's Python API and does not detach any processes from the Python interpreter. CPython (which is used in this case) is known to not deal well with multithreading, due to its Global Interpreter Lock (GIL).

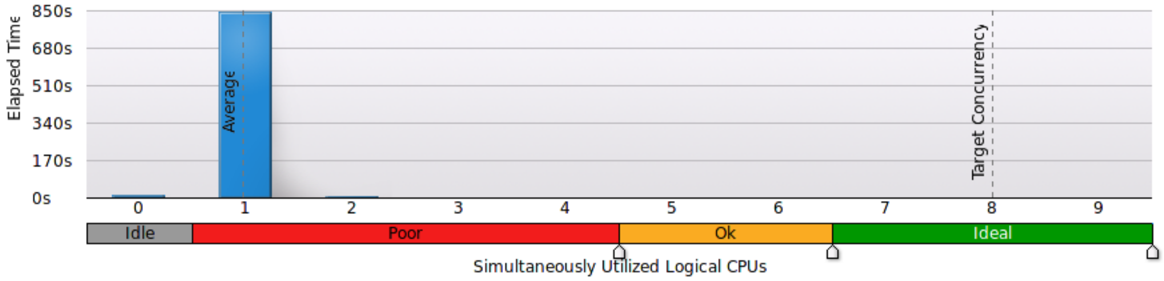


Figure 3: CPU usage histogram; Total program's thread count: 7

The profiled analysis script and input data have been kindly provided by the Particle Physics Group at TU Dortmund.

Conclusions and Future Work

So far my research indicates, that rewarding targets for optimization are the data representation and data compression of the ROOT-framework [5]. For offline analysis the measurement data is currently prefiltered and divided into many streams (called stripping lines) that are saved separately. Many stripping lines contain similar data, which implies a massive size overhead. Using modern DBMS-technologies has the potential of reducing redundancy in stripping lines, which leads to a lower annual amount of data. Additionally, the use of modern distributed data processing platforms, like Google's Dremel platform [7] or the Apache Flink platform [1] (formerly Stratosphere), seem to be promising.

The next steps include importing measurement data into a modern in-memory DBMS (like MonetDB [8] or CoGaDB [2]) and porting some analysing algorithms to make use of DBMSs. In the long run, it is worthwhile to investigate the possibility of creating a domain specific query language to accommodate most use-cases of DAVINCI.

References

[1] Alexander Alexandrov, Rico Bergmann, Stephan Ewen, Johann-Christoph Freytag, Fabian Hueske, Arvid Heise, Odej Kao, Marcus Leich, Ulf Leser, Volker Markl, Felix Naumann, Mathias Peters, Astrid Rheinländer, Matthias J. Sax, Sebastian Schelter, Mareike Höger, Kostas Tzoumas, and Daniel Warneke. The stratosphere platform for big data analytics. *The VLDB Journal*, 23(6):939–964, December 2014.

[2] Sebastian Breß. The Design and Implementation of CoGaDB: A Column-oriented GPU-accelerated DBMS. *Datenbank-Spektrum*, 14(3):199–209, 2014.

[3] CERN. The DAVINCI project. `http://lhcb-release-area.web.cern.ch/LHCb-release-area/DOC/davinci/`. Accessed: 10.11.2015.

[4] CERN. LHCb data flow. `http://lhcb-public.web.cern.ch/lhcb-public/Images2015/LHCbDatFlow.png`. Accessed: 11.11.2015.

[5] CERN. ROOT a data analysis framework. `https://root.cern.ch`. Accessed: 10.11.2015.

[6] CERN. The worldwide LHC computing grid. `http://wlcg.web.cern.ch`. Accessed: 11.11.2015.

[7] Sergey Melnik, Andrey Gubarev, Jing Jing Long, Geoffrey Romer, Shiva Shivakumar, Matt Tolton, and Theo Vassilakis. Dremel: Interactive analysis of web-scale datasets. In *Proc. of the 36th Int'l Conf on Very Large Data Bases*, pages 330–339, 2010.

[8] Stratos Idreos Fabian Groffen Niels Nes and Stefan Manegold Sjoerd Mullender Martin Kersten. MonetDB: Two decades of research in column-oriented database architectures. *Data Engineering*, page 40, 2012.

[9] Markus F.X.J. Oberhumer. LZO real-time data compression library. `http://www.oberhumer.com/opensource/lzo/`. Accessed: 12.11.2015.

Improving Efficiency of the LHCb High Level Trigger by use of Parallel Data Stream Processing on Heterogeneous Hardware

Thomas Lindemann
Lehrstuhl für Datenbanken und Informationssysteme (DBIS)
Technische Universität Dortmund
thomas.lindemann@cs.tu-dortmund.de

In the LHCb Project, a continuous stream of events is produced by the aggregates of the detector, which have to be processed in real time, since there are no capabilities to store all collision events permanently with the current storage technology. The High Level Trigger (HLT) has to select the events, which have to be stored for further analysis. In our research, we are evaluating different techniques to handle with these restrictions. A workaround is making use of heterogeneous hardware and placing operations to the hardware, which is estimated to calculate the result of the respective operation most rapidly. As a first step toward this goal, we investigated load balancing and robustness in heterogenenous co-processor systems. We also introduced new ideas for data stream processing like asynchronous copying and the processing of data in chunks of sizes that guarantee the best balance of bandwidth and processing capability.

1 Introduction

The LHCb project is a large and complex research project. Named after the b-quark, LHCb is one of the four big experiments at CERN. The general scope is to explain the matter/anti-matter asymmetry. Its main focus is the study of particle decays involving beauty and charm quarks. [1]

At the time of writing this report, we working on this project for four month. Our specific research topic is to improve the High Level Trigger (HLT). The challenge is to find new solutions for processing the big amounts of data with limited resources much faster than it has been performed in the first run of the LHCb project and allow the physicists to make experiments with more precise decisions. An idea to achieve this objective is using modern hardware components such as GPUs (Graphics Processing Units) or APUs (Accelerated Processing Units) and place the operations to the best device. This methods are already discussed in actual research in databases and information systems, thus it should be possible to apply the outcomes to a new issue.

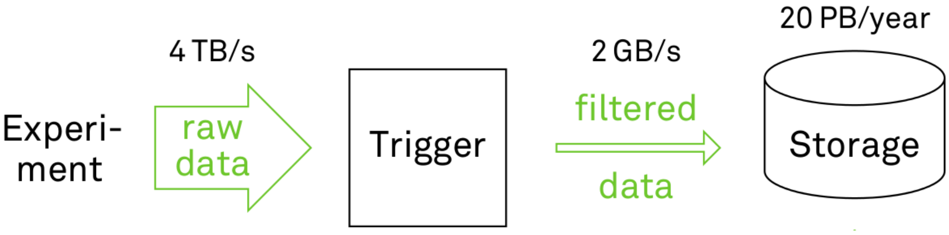


Figure 1: Trigger system setup

Another research topic is the change of the file format for the event data to make optimized declarative queries possible for fast analysis. The actual data structure is very flexible, which is obstructive for a query optimized storage.

2 Decay Event Processing on Heterogeneous Hardware

Our contributions to the project are at first a requirements analysis from a data management perspective. It comes out that the current implementation is unsuited for GPU acceleration, due to the algorithms, which are working with one event tuple at a time. Furthermore, a detailed profiling of the high level trigger and a determination of bottlenecks has been done and is still in a refinement progress. We also investigated acceleration potential on GPUs and Xeon Phi APUs. Our detailed focus at the moment lies on improving fast scan-operations on detector measurement results, especially in the data of the muon system and the reconstruction of tracks afterwards. With the reconstructed tracks, it's possible to trace back all vertices of an event, which include an interesting decay. In our research for improving the High Level Trigger, we draw on our actual research on Robust Query Processing in Co-Processor Accelerated Databases [2]. The goal is to place the operation of an algorithm either on the CPU like in the traditional way or place it to a suitable co-processor instead. Similar to database use cases, when we split large queries into smaller operations, the processing of an event in the Trigger sofware also needs different algorithms to recreate the tracks and vertices from the detector data. Thus, this approach can be followed up.

We evaluated techniques to virtualize the physical processors of a system, e.g. create multiple virtual CPUs or GPUs for a single CPU or GPU. For each device, an operator queue is maintained that executes all operators in the queue on a processor which is idle. Through this approach it is possible to schedule different jobs in a pipeline to be processed on the physical device.

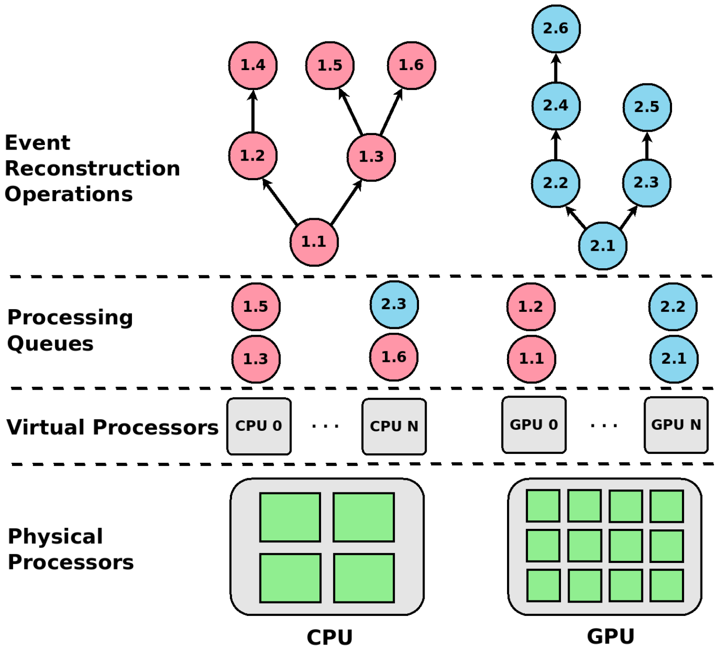


Figure 2: Operator Placement on Virtual Processors

In order to avoid unnecessary data transfers between devices, conceptions regarding data locality have to be considered by the scheduler when placing the operations on the devices. Especially the scan-operators or other random access patterns with a small portion of arithmetics are leading to bus bottlenecks. In consequence, we make some experiments according to stream processing to optimize the data transfer between main memory and the co-processor's memory. The idea to handle this kind of data streams is divided into two parts. First of all, the data for the stream processing should be transferred asynchronous while computing, which is supported by the PCI-e bus and lowers the transfer costs significantly compared to sequential processing. The theoretical PCI-e bandwidth is 0.5, 1.0, 2.0 and 4.0 GB/s per each lane in the 1., 2., 3. and 4. generation of the PCI-e standard and modern GPUs are connected through 16 lanes. In our tests with PCI-e 3rd generation GPUs, we could achieve 75% of the theoretical bandwidth using unidirectional transfer and 62.5% using bidirectional transfer. Additionally, we have shown that a minimum block size of about 1 MB is required to achieve the maximum bandwidth. Assuming that transfer and computation have the same throughput, the transfers will cost only a small delay at the beginning and the end of the stream processing. The second point depends on the data streaming, because asynchronous copying and processing at the same time requires the data to be splitted into data chunks, which will be processed at once.

One drawback on stream computations is that it depends on the main memory bandwidth, if all processor cores and co-processor devices can be used compute-bound.

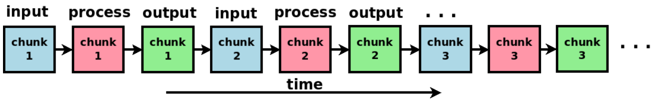


Figure 4: Sequential I/O

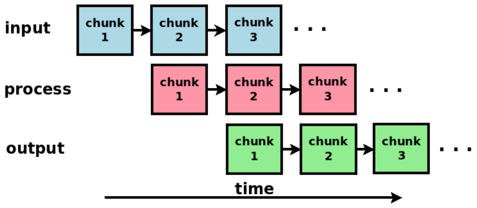


Figure 3: Overlapping asynchronous I/O

The usage of a GPU assumes that the problem is highly parallelizable, which is essential for the utilization with an SIMD-architecture. A disadvantage might be, that there is a high number of events with a relatively small amount of data. A solution might be to process many events simultaneously with the same algorithm, but there is the adverse fact that every event might be different in the number of tracks to be reconstructed and the number and kind of vertices, due to different decay channels.

Additionally, we did some experiments with Intel Xeon Phi co-processors. Our first results have show that simple generic scan-operations on a Xeon Phi are not as fast as a parallel implementation on CPUs. We also observed, that a scan-operation on a Xeon Phi accelerator uses nearly the maximum memory bandwidth, so it is unlikely to expect speedups in scan-operations with further optimizations. Thus, we will not further investigate the current generation Xeon Phis for scans. Nevertheless, there are still more capabilities to accelerate more complex operations due to high parallel execution on a large number of compute units, which a Xeon Phi provides.

3 Conclusion and Future Work

At this point of our research we are in profiling the existing trigger algorithms and deciding which operations are expected to have a good performance on the different co-processor architectures. We are in developing a cache efficient memory layout of the event data for efficient processing on modern hardware.

References

[1] The LHCb collaboration, C. Langenbruch et al. Angular analysis of the $B0 \to K^{*0}\mu^+ \mu^-$ decay. *LHCb-CONF-2015-002*, 2015.

[2] Sebastian Breß and Jens Teubner. Robust Query Processing in Memory-Constrained Co-Processor Systems. *Submitted to SIGMOD 2015*, 2015.

Determination of $\sin(2\beta)$ with the LHCb experiment

Frank Meier
Lehrstuhl für Experimentelle Physik 5
Technische Universität Dortmund
frank.meier@tu-dortmund.de

The LHCb experiment is one of the four big experiments located at the Large Hadron Collider near Geneva, Switzerland. Its main focus is the search for rare decays and effects of *CP*-violation in decays of beauty and charm hadrons. In order to maximize the sensitivity with respect to these specialized targets the LHCb detector is built as a forward single arm spectrometer.

In the collisions of protons inside the vertex locator (VELO), new particles are created and decay until they finally leave traces in the various subcomponents of the detector. These traces are hits in the tracking systems (VELO, TT, T1-T3, M1-M6), clusters in the calorimeters (ECAL, HCAL) and Cherenkov radiation in the Ring Imaging Cherenkov Detectors (RICH1, RICH2). To conclude on the presence of particles the information of these subdetectors needs to be reconstructed, e.g. by fits of trajectories to ensembles of hits and pattern recognition algorithms looking for clusters of energy deposition. Finally particle candidates need to be combined to heavier particles in order to perform physics measurements on the same. The endeavour to find the particles of interest is hindered for two main reasons. Firstly, hundreds of particles are produced inside the angular acceptance which allows for a large number of combinations to be made in the reconstruction as well as the combination of particles. Secondly, the interaction rate of 50ns / 25ns together with the limitation on the bandwidth that can be written to disk enforces a fast reconstruction that leads to the selection of interesting events and the rejection of physically uninteresting ones. Both these points together set the frame for investigating these tasks in the context of resource limitation: The reconstruction and combination tasks can be parallelized and therefore performed faster.

A physics analysis has been performed to demonstrate the need for an improved processing of the data collected by the LHCb detector. This analysis studies time-dependent charge-parity (*CP*) violation which is one of the keys for understanding the matter-antimatter-asymmetry observed in our universe. In processes involving B mesons *CP* violation has been observed in the "golden mode" $B^0 \to J/\psi K^0_S$ by the BaBar and Belle experiments at the asymmetric e^+e^- colliders PEP-II and KEKB [1, 3] for the first time. As the $J/\psi K^0_S$ final state is common to both the B^0 and the $\overline{B}^0$ meson decays, the interference between the amplitudes for the direct decay and for the decay after B^0–$\overline{B}^0$ oscillation results in a decay-time dependent *CP* asymmetry,

$$\mathcal{A}(t) \equiv \frac{\Gamma(\overline{B}^0(t) \to J/\psi K^0_S) - \Gamma(B^0(t) \to J/\psi K^0_S)}{\Gamma(\overline{B}^0(t) \to J/\psi K^0_S) + \Gamma(B^0(t) \to J/\psi K^0_S)} = \frac{S \sin(\Delta m\, t) - C \cos(\Delta m\, t)}{\cosh(\frac{\Delta\Gamma\, t}{2}) + A_{\Delta\Gamma} \sinh(\frac{\Delta\Gamma\, t}{2})}. \quad (1)$$

Here, $B^0(t)$ and $\overline{B}^0(t)$ indicate the flavour of the B meson at production, while t indicates the decay time. The parameters Δm and $\Delta\Gamma$ are the mass and the decay width differences between the heavy and light mass eigenstates of the B^0–$\overline{B}^0$ system, and S, C, and $A_{\Delta\Gamma}$ are *CP* observables. As $\Delta\Gamma$ is negligible for the B^0–$\overline{B}^0$ system [2], the time-dependent asymmetry simplifies to $\mathcal{A}(t) = S \sin(\Delta m\, t) - C \cos(\Delta m\, t)$.

The analysis can be performed with $B^0 \to J/\psi K^0_S$ candidates reconstructed in the $J/\psi \to \mu^+\mu^-$ and $K^0_S \to \pi^+\pi^-$ final states. This experimentally very clean signature makes it an ideal testbed for the LHCb collaboration to prove its capability of tagged precision measurements. The complete Run I data sample is used which corresponds to an integrated luminosity of $3\,\text{fb}^{-1}$ at centre-of-mass energies of 7 and 8 TeV.

A selection is required to suppress the combinatorial background. The first step is a centrally organized preselection for all analyses performed by the LHCb collaboration. This takes several weeks. Afterwards, the individual tuple of the size of several hundred GB is produced. Another cut-based selection is applied which mainly consists of kinematical and geometrical requirements. This leaves 41 500 flavour-tagged signal candidates.

The values of the *CP* violation observables S and C are estimated by maximizing the likelihood of a probability density function (PDF) describing the unbinned distributions of the reconstructed mass m, the decay time t and its uncertainty estimate σ_t, the OS and SSπ flavour tag decisions d_{OS} and $d_{\text{SS}\pi}$, and the corresponding per- candidate mistag probability estimates η_{OS} and $\eta_{\text{SS}\pi}$. The fit is performed simultaneously in 24 independent subsamples. In each category the data distribution is modelled using a sum of two individual PDFs, one for the B^0 signal and one for the combinatorial background.

The likelihood is a function of 83 free parameters, including S and C, and 48 yield parameters for the signal and the background components. Eleven parameters are external inputs, including the production asymmetry, the flavour tagging calibration parameters, and the mass difference Δm [4]. These are constrained in the fit within their statistical uncertainties and taking their correlations into account. The complexity of the likelihood fit results in execution times of a few hours on multi-core machines.

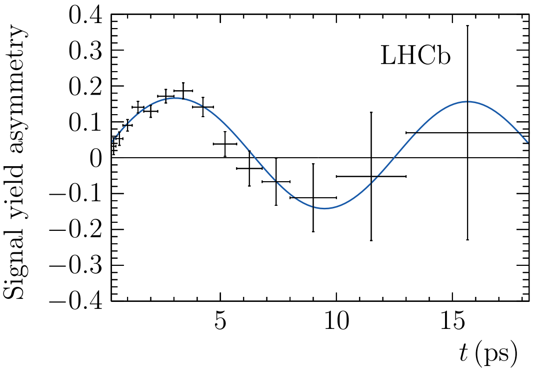


Figure 1: Time-dependent signal-yield asymmetry $(N_{\overline{B}^0} - N_{B^0})/(N_{\overline{B}^0} + N_{B^0})$. Here, N_{B^0} ($N_{\overline{B}^0}$) is the number of $B^0 \to J/\psi K^0_S$ decays with a B^0 ($\overline{B}^0$) flavour tag. The data points are obtained with the *sPlot* technique [5], assigning signal weights to the events based on a fit to the reconstructed mass distribution. The solid curve is the projection of the signal PDF.

Various sources of systematic uncertainties on the *CP* observables are examined, in particular from mis-modelling PDFs and from systematic uncertainties on the input parameters. The size of the systematic uncertainties is determined in pseudo-experiments. For this purpose data samples are generated with PDFs that differ from the nominal setup by the source under study. The nominal fit is performed on these data samples. Any deviation shows up as a bias in the pull distributions of the *CP* observables. To make sure that reliable results are produced each study is done with one thousand pseudo-experiments. As this requires lots of computing resources improvements in the minimization of the likelihood are very useful. Adding all contributions in quadrature results in total systematic uncertainties of ± 0.020 on S and ± 0.005 on C.

In conclusion, the *CP* observables S and C are measured to be

$$\begin{aligned} S &= 0.731 \pm 0.035\,(\text{stat}) \pm 0.020\,(\text{syst}), \\ C &= -0.038 \pm 0.032\,(\text{stat}) \pm 0.005\,(\text{syst}), \end{aligned}$$

with a statistical correlation coefficient $\rho(S, C) = 0.483$. This measurement represents the most precise time-dependent *CP* violation measurement at a hadron collider to date. Fig. 1 shows the decay-time dependent signal-yield asymmetry.

References

[1] K. Abe et al. Observation of large CP violation in the neutral B meson system. *Phys.Rev.Lett.*, 87:091802, 2001.

[2] Y. Amhis et al. Averages of b-hadron, c-hadron, and τ-lepton properties as of summer 2014. 2014. updated results and plots available at: `http://www.slac.stanford.edu/xorg/hfag/`.

[3] Bernard Aubert et al. Observation of CP violation in the B^0 meson system. *Phys.Rev.Lett.*, 87:091801, 2001.

[4] K. A. Olive et al. Review of particle physics. *Chin. Phys.*, C38:090001, 2014.

[5] Muriel Pivk and Francois R. Le Diberder. sPlot: A statistical tool to unfold data distributions. *Nucl.Instrum.Meth.*, A555:356–369, 2005.

Use of FPGAs at the LHCb experiment

Ramon Niet
Lehrstuhl für Experimentelle Physik 5
Technische Universität Dortmund
ulrich.eitschberger@tu-dortmund.de

The LHCb experiment is one of the four big experiments located at the Large Hadron Collider near Geneva, Switzerland. Its main focus is the search for rare decays and effects of *CP*-violation in decays of beauty and charm hadrons. In order to maximize the sensitivity with respect to these specialized targets the LHCb detector is built as a forward single arm spectrometer.

In the collisions of protons inside the vertex locator (VELO), new particles are created and decay until they finally leave traces in the various subcomponents of the detector. These traces are hits in the tracking systems (VELO, TT, T1-T3, M1-M6), clusters in the calorimeters (ECAL, HCAL) and Cherenkov radiation in the Ring Imaging Cherenkov Detectors (RICH1, RICH2). To conclude on the presence of particles the information of these subdetectors needs to be reconstructed, e.g. by fits of trajectories to ensembles of hits and pattern recognition algorithms looking for clusters of energy deposition. Finally particle candidates need to be combined to heavier particles in order to perform physics measurements on the same. The endeavour to find the particles of interest is hindered for two main reasons. Firstly, hundreds of particles are produced inside the angular acceptance which allows for a large number of combinations to be made in the reconstruction as well as the combination of particles. Secondly, the interaction rate of 50ns / 25ns together with the limitation on the bandwidth that can be written to disk enforces a fast reconstruction that leads to the selection of interesting events and the rejection of physically uninteresting ones. Both these points together set the frame for investigating these tasks in the context of resource limitation: The reconstruction and combination tasks can be parallelized and therefore performed faster.

A physics analysis has been performed to demonstrate the need for an improved processing of the data collected by the LHCb detector. This analysis studies time-dependent charge-parity (*CP*) violation which is one of the keys for understanding the matter-antimatter-asymmetry observed in our universe. In processes involving B mesons CP violation has been observed in the "golden mode" $B^0 \to J/\psi\, K^0_S$ by the BaBar and Belle experiments at the asymmetric e^+e^- colliders PEP-II and KEKB [1,3] for the first time. As the $J/\psi\, K^0_S$ final state is common to both the B^0 and the $\overline{B}{}^0$ meson decays, the interference between the amplitudes for the direct decay and for the decay after B^0–$\overline{B}{}^0$ oscillation results in a decay-time dependent CP asymmetry,

$$\mathcal{A}(t) \equiv \frac{\Gamma(\overline{B}{}^0(t) \to J/\psi\, K^0_S) - \Gamma(B^0(t) \to J/\psi\, K^0_S)}{\Gamma(\overline{B}{}^0(t) \to J/\psi\, K^0_S) + \Gamma(B^0(t) \to J/\psi\, K^0_S)} = \frac{S \sin(\Delta m\, t) - C \cos(\Delta m\, t)}{\cosh(\frac{\Delta\Gamma\, t}{2}) + A_{\Delta\Gamma} \sinh(\frac{\Delta\Gamma\, t}{2})}. \quad (1)$$

Here, $B^0(t)$ and $\overline{B}{}^0(t)$ indicate the flavour of the B meson at production, while t indicates the decay time. The parameters Δm and $\Delta\Gamma$ are the mass and the decay width differences between the heavy and light mass eigenstates of the B^0–$\overline{B}{}^0$ system, and S, C, and $A_{\Delta\Gamma}$ are CP observables. As $\Delta\Gamma$ is negligible for the B^0–$\overline{B}{}^0$ system [2], the time-dependent asymmetry simplifies to $\mathcal{A}(t) = S \sin(\Delta m\, t) - C \cos(\Delta m\, t)$.

The analysis can be performed with $B^0 \to J/\psi\, K^0_S$ candidates reconstructed in the $J/\psi \to \mu^+\mu^-$ and $K^0_S \to \pi^+\pi^-$ final states. This experimentally very clean signature makes it an ideal testbed for the LHCb collaboration to prove its capability of tagged precision measurements. The complete Run I data sample is used which corresponds to an integrated luminosity of $3\,\text{fb}^{-1}$ at centre-of-mass energies of 7 and 8 TeV.

A selection is required to suppress the combinatorial background. The first step is a centrally organized preselection for all analyses performed by the LHCb collaboration. This takes several weeks. Afterwards, the individual tuple of the size of several hundred GB is produced. Another cut-based selection is applied which mainly consists of kinematical and geometrical requirements. This leaves 41 500 flavour-tagged signal candidates.

The values of the CP violation observables S and C are estimated by maximizing the likelihood of a probability density function (PDF) describing the unbinned distributions of the reconstructed mass m, the decay time t and its uncertainty estimate σ_t, the OS and SSπ flavour tag decisions d_{OS} and $d_{\text{SS}\pi}$, and the corresponding per- candidate mistag probability estimates η_{OS} and $\eta_{\text{SS}\pi}$. The fit is performed simultaneously in 24 independent subsamples. In each category the data distribution is modelled using a sum of two individual PDFs, one for the B^0 signal and one for the combinatorial background.

The likelihood is a function of 83 free parameters, including S and C, and 48 yield parameters for the signal and the background components. Eleven parameters are external inputs, including the production asymmetry, the flavour tagging calibration parameters, and the mass difference Δm [5]. These are constrained in the fit within their statistical uncertainties and taking their correlations into account. The complexity of the likelihood fit results in execution times of a few hours on multi-core machines.

Various sources of systematic uncertainties on the CP observables are examined, in particular from mis-modelling PDFs and from systematic uncertainties on the input parameters.

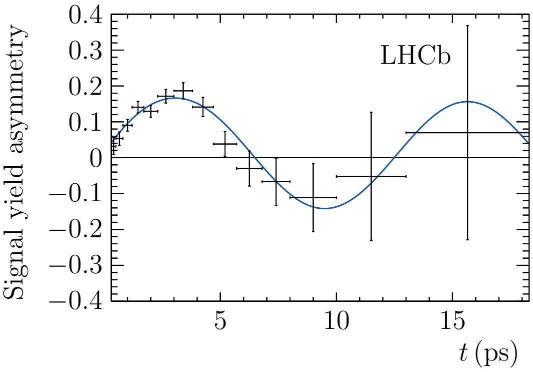


Figure 1: Time-dependent signal-yield asymmetry $(N_{\overline{B}^0} - N_{B^0})/(N_{\overline{B}^0} + N_{B^0})$. Here, N_{B^0} ($N_{\overline{B}^0}$) is the number of $B^0 \rightarrow J/\psi K^0_S$ decays with a B^0 ($\overline{B}^0$) flavour tag. The data points are obtained with the *sPlot* technique [6], assigning signal weights to the events based on a fit to the reconstructed mass distribution. The solid curve is the projection of the signal PDF.

Adding all contributions in quadrature results in total systematic uncertainties of ± 0.020 on S and ± 0.005 on C.

In conclusion, the CP observables S and C are measured to be

$$\begin{aligned} S &= 0.731 \pm 0.035\,(\text{stat}) \pm 0.020\,(\text{syst}), \\ C &= -0.038 \pm 0.032\,(\text{stat}) \pm 0.005\,(\text{syst}), \end{aligned}$$

with a statistical correlation coefficient $\rho(S, C) = 0.483$. This measurement represents the most precise time-dependent CP violation measurement at a hadron collider to date. Fig. 1 shows the decay-time dependent signal-yield asymmetry.

The uncertainties on the presented measurement will decrease with further data being recorded. In order to increase the signal yield per data taking time, an upgrade of the LHCb detector is planned to take place in 2018. The aim is to prepare the detector for a running period with increased luminosity. Part of this upgrade consists of replacing the tracking stations by a scintillating fibre detector [4]. This detector will provide a readout rate of 40 MHz to a full software trigger system. Field Programmable Gate Arrays (FPGAs) are employed in the frontend as well as the backend detector electronics, as they are highly suited to cope with huge data rates. Due to the limited space that is available on these FPGAs algorithms need to be designed in an efficient way. As an example algorithm the preclustering of hits in the different tracking modules is investigated. Reducing the number of hit candidates to form tracks at this early reconstruction stage will decrease the time that is needed to reconstruct the full event in the software trigger. Benefits of this are the opportunity to perform more sophisticated selections at the trigger stage in order to increase the signal efficiency and background rejection.

References

[1] K. Abe et al. Observation of large CP violation in the neutral B meson system. *Phys.Rev.Lett.*, 87:091802, 2001.

[2] Y. Amhis et al. Averages of b-hadron, c-hadron, and τ-lepton properties as of summer 2014. 2014. updated results and plots available at: `http://www.slac.stanford.edu/xorg/hfag/`.

[3] Bernard Aubert et al. Observation of CP violation in the B^0 meson system. *Phys.Rev.Lett.*, 87:091801, 2001.

[4] LHCb Collaboration. LHCb Tracker Upgrade Technical Design Report. Technical Report CERN-LHCC-2014-001. LHCB-TDR-015, CERN, Geneva, Feb 2014.

[5] K. A. Olive et al. Review of particle physics. *Chin. Phys.*, C38:090001, 2014.

[6] Muriel Pivk and Francois R. Le Diberder. sPlot: A statistical tool to unfold data distributions. *Nucl.Instrum.Meth.*, A555:356–369, 2005.

Improved precision measurements with the LHCb experiment

Margarete Schellenberg
Lehrstuhl für Experimentelle Physik 5
Technische Universität Dortmund
margarete.schellenberg@tu-dortmund.de

The LHCb experiment at the Large Hadron Collider (LHC) near Geneva was designed to research the asymmetry of matter and anti-matter in our universe. With precision measurements it searches for decays of beauty and charm hadrons, that show physical processes, which cannot be described by the so called Standard Model of Particle Physics. The experiment collects data in the order of Terabytes per second from collisions of high energy protons at a rate of 40 MHz. Due to the characteristics of the measured processes, the LHCb detector is constructed as a single arm forward spectrometer, cf. Figure 1.

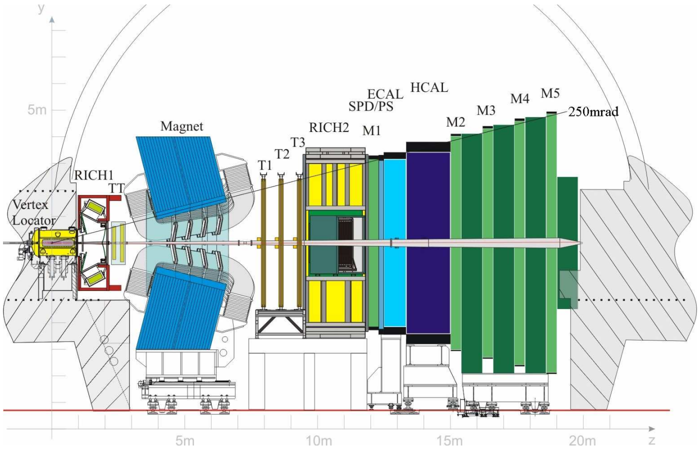


Figure 1: The LHCb detector with the various subdetectors for the identification of particles and reconstruction of their tracks.

The vertex locator (VELO) is situated at the interaction point of these collisions, where – as a cause of different physical processes – many new particles are created, which then further decay into other particles. The latter are detected in the following subsystems of the detector: the tracking system (TT, IT, OT) measures the required spatial information to later reconstruct the flight paths of particles. The particle identification system (RICH1, RICH2, ECAL and HCAL, Muon chambers) on the other hand collects information to determine the type of detected particles. The combined information of these systems can then be used to reconstruct the primary beauty and charm hadrons which are of interest for the physics analyses. The search for relevant decays is complicated because of two circumstances. On the one hand the proton collisions produce a large number of particles in the acceptance range of the detector. This allows for a large number of combinations in the reconstruction of particle tracks as well as the combination of different particles to their mother particle. On the other hand the high interaction rate together with a limited bandwidth between detector and data storages requires a fast event reconstruction to select relevant and reject physically uninteresting events.

Under the assumptions, that the dataset of one event has a size of approximately 100 Kilobytes and that the detector runs 100 of 365 days, the LHCb experiment collects around 20 Petabytes of data per year. This data has not only to be saved permanently but it also has to be available at all times for 1000 members of the LHCb collaboration with different types of queries. Existing methods of the particle physics community for data storage and usage are already limited concerning the efficient availability of data. At the moment collected data is written to tapes and only on special requests it is stored on discs, after being already preselected for a special group of decays. This fact limits the quality and possibilities of physics analyses. Therefore, the usability of distributed storage and computing architectures shall be investigated. Cloud-computing-developed techniques are able to process huge amounts of data in an efficient and fault-tolerant way. These techniques has to be tested for the use case of particle physics analyses.

In principle there are two main challenges, especially concerning the amount of data: Firstly for a given analysis only a part of the stored data is relevant. The required preselection of data is a typical data base task, only in this case it is hindered by much more complex selection criteria, which demands adequate indexing and optimization techniques. In contrast to the common cloud computing use case, the access to every aspect of the stored data is necessary and not only e.g. the metadata to perform this kind of task. Secondly, one needs to search for new methods of communication between analysis tasks and stored data. This would enable the possibility of a stepwise selection during runtime so that physical uninteresting events can be rejected at an early stage and the rest of its information has not to be processed. This means that less resources for analysis operations are needed. All in all, the usage of big data tools in particle physics analyses will be investigated to exploit the huge possibilities for the improvement of the measurements with the LHCb experiment.