

Advances in Statistical Modeling of High Dimensional Data: Variable selection and Challenges in Image Analysis

September 17-18, 2009 in Munich

Organizing Committee: Ulrich Mansmann, Jörg Rahnenführer, Juliane Schäfer, Achim Tresch

Schedule

Thursday, Sep 17		
Time	Title	Speaker
12.55	Welcome Address	
13.00	Indirect comparison of interaction graphs	Ulrich Mansmann IBE, LMU München
13.30	Estimating high-dimensional intervention effects from observational data	Marloes Maathuis/ Peter Bühlmann ETH Zürich
14.00	Minimal Gene Set Enrichment	Julien Gagneur EMBL Heidelberg
14.30	Coffee Break	
15.00	Deep sequencing of a mixed sample	Oswaldo Zagordi/ Niko Beerenwinkel ETH Zürich/Basel
15.30	Integrated analysis of copy number alterations and gene expression	Martin Schäfer/ Katja Ickstadt TU Dortmund
16.00	Estimating Networks in a Huge Microarray Meta-Analysis with 60 Experiments and more than 7000 Microarrays	Markus Schmidberger IBE, LMU München
16.30	Coffee Break Start of the Poster Session	
17.00	Analysis of cellular genealogies	Ingo Röder IMISE, Uni Leipzig
17.30	Dynamic Nested Effects Models	Rainer Spang Uni Regensburg
18.00	Election of the new working group board and Poster Session	
19.30	Dinner	

Friday, Sep 18		
9.00	Reverse Engineering of Signaling Pathways from RNAi Data	Bettina Knapp/ Lars Kaderali Bioquant Heidelberg
9.30	Prediction of gene function by automated cellular phenotyping and genome-wide RNAi.	Grégoire Pau EMBL Heidelberg
10.00	Heterogenous population context determines cellular activity and virus infection patterns	Berend Snijder/ Lucas Pelkmans ETH Zürich
10.30	Coffee Break	
11.00	Reconstruction of signaling networks from gene intervention data. Abstract	Tim Beissbarth Uni Göttingen
11.30	Bayesian Modelling for Perfusion Imaging	Volker Schmid LMU München
12.00	Bayesian parameter estimation in signalling networks	Fabian Theis Helmholtz Zentrum München
12.30	Goodbye	

Ulrich Mansmann, LMU München

Thursday, Sep 17 13.00

Indirect comparison of interaction graphs

A new approach to test differential conditional independence structure (CIS) in two graphs is introduced. The two graphs have the same set of nodes and are estimated from data sampled under two different conditions. The statistic uses the entire pathplot in a Lasso regression as the information how a node connects with the remaining nodes in the graph. Interpreting its paths as random processes allows defining stopping times which make the statistical properties of the test statistic accessible to analytic reasoning. A resampling approach will be used to calculate p-values simultaneously for a hierarchical testing procedure. The hierarchical testing steps through a given hierarchy of clusters. First, collective effects are measured at the coarsest level possible (the global null hypothesis that no node in the graph shows a differential CIS). If the global null hypothesis can be rejected, finer resolution levels are tested for an effect until the level of individual nodes is reached. The approach can be applied to problems in molecular medicine as well as to complex phenotypes as presented for example by the International Classification of Functioning (ICF).

Marloes Maathuis / Peter Bühlmann, ETH Zürich

Thursday, Sep 17 13.30

Estimating high-dimensional intervention effects from observational data

There are established statistical methods to estimate causal effects from observational data in small systems with expert knowledge about the possible causal relationships between the variables. But in large-scale systems without expert knowledge, estimation of causal effects is much more challenging. We recently proposed a new method for this purpose. I will discuss the algorithm and some of its statistical properties, and I will illustrate it on yeast gene expression data.

Julien Gagneur, EMBL Heidelberg

Thursday, Sep 17 14.00

Minimal Gene Set enrichment

Gene group enrichment analyses often return a large number of groups making their interpretation difficult. Beyond issues of multiple testing, one reason is that these groups share genes, so that if one group turns out significant, further groups with many genes in common with it may also be significant. This is particularly relevant for the Gene Ontology which consists of nested groups and for which heuristics exploiting this structure have been previously proposed. Here we tackle the problem by turning the question differently. Instead of searching for all significantly enriched groups, we search for a minimal set of groups that can explain the data. We model the experimental observation by a set of "active" groups. Our model penalizes the number of active groups thus naturally providing parsimonious solutions.

Oswaldo Zagordi/Niko Beerenwinkel, ETH Zürich

Thursday, Sep 17 15.00

Deep sequencing of a mixed sample

Genetic heterogeneity within organisms of the same species plays an essential role in nature. A typical example of this heterogeneity is the HIV population present in a single patient, where the diversity can be so high to represent one of the main problems in the development of an effective vaccine. Sequencing technologies of the new generation, thanks to their higher throughput and lower cost per base, allow us to address this problem in a way that was unimaginable only a few years ago. By sequencing a genetically heterogeneous sample, in fact, we can estimate not only the amount of variation, but also the identity and frequency of the haplotypes responsible for it. This is done by means of a procedure that corrects the errors produced in the sequencing process, thus separating the technical errors from the real biological variation. After giving a brief description of the techniques used, implemented in the freely available software ShoRAH, I will present results on both simulated and real data.

Martin Schäfer/Katja Ickstadt, TU Dortmund

Thursday, Sep 17 15.30

Integrated analysis of copy number alterations and gene expression

With the rise of microarray technology, the analysis of a number of different genetic features like SNP calls, copy number variation, loss of heterozygosity, gene expression or alternative splicing has considerably increased in recent years, as well as the number of available data sets. Thus, to understand mechanisms of disease pathogenesis on a molecular basis, e.g., in cancer research, the challenge of analysing such different data types in an integrated way has become increasingly important. One focus of research has been the integration of copy number and gene expression data. Previous works on this topic generally either have

analysed copy number and gene expression consecutively, in a univariate way, or used correlation or regression procedures to assess the degree of dependence between both inputs. We will explain shortcomings of both types of approaches with respect to identifying genes or genetic regions in patients that for both inputs show abnormalities towards the same direction (e.g., underexpressed genes accompanied by a loss of DNA material). We will present a new explorative procedure, based on a modified correlation coefficient, and show that it is better able to perform this task both for real data and in simulations. Our approach is also suited for the analysis of other (in particular, continuous) data types.

Markus Schmidberger, LMU München

Thursday, Sep 17 16.00

Estimating Networks in a Huge Microarray Meta-Analysis with 60 Experiments and more than 7000 Microarrays

Public available data sets were collected from public microarray databases and preprocessed and analyzed together. Data from more than 60 experiments, more than 7000 microarrays and eight different cancer entities were used to demonstrate the power of the parallel computing with the 'affyPara' package, to discuss the difficulties of data management, and to estimate gene interaction networks with the 'pcalg' package. This is one of the first projects for analyzing this amount of public available data sets together. Therefore this talk presents more technical details and problems of performing this kind of analyses. The comparison of gene networks for different pathways and different cancer entities partly confirms established forms of gene-gene interaction.

Ingo Roeder, Uni Leipzig

Thursday, Sep 17 17.00

Analysis of cellular genealogies

The analysis of individual cell fates within a population of stem and progenitor cells is still a major experimental challenge in stem cell biology. However, new monitoring techniques, such as high-resolution time lapse video microscopy, facilitate the tracking and the quantitative analysis of single cells and their progeny. Information on cellular development, divisional history, and differentiation are naturally comprised into a pedigree-like structure, denoted as cellular genealogy. However, to extract reliable information about effecting variables and control mechanisms underlying cell fate decisions, it is necessary to analyse large numbers of cellular genealogies, which calls for the application of automatic cell

tracking algorithms. In the talk I will present methods that allow for the automatic reconstruction of cellular genealogies from time lapse video data of cultured hematopoietic stem cells. Furthermore, I will discuss a set of statistical measures that are specifically tailored for the analysis of cellular genealogies and I will show how these measures can be applied to characterize and compare cellular fates under different conditions.

Rainer Spang, Uni Regensburg

Thursday, Sep 17 17.30

Dynamic Nested Effects Models

Cellular decision making in differentiation, proliferation, or cell death is mediated by molecular signaling processes, which control the regulation and expression of genes. Vice versa, the expression of genes can trigger the activity of signaling pathways. We introduce and describe a statistical method called Dynamic Nested Effects Model (D-NEM) for analyzing the temporal interplay of cell signaling and gene expression. D-NEMs are Bayesian models of signal propagation in a network. They decompose observed time delays of multiple step signaling processes into single steps. Time delays are assumed to be exponentially distributed. Rate constants of signal propagation are model parameters, whose joint posterior distribution is assessed via Gibbs sampling. They hold information on the interplay of different forms of biological signal propagation. Molecular signaling in the cytoplasm acts at high rates, direct signal propagation via transcription and translation act at intermediate rates, while secondary effects operate at low rates. D-NEMs allow the dissection of biological processes into signaling and expression events, and analysis of cellular signal flow. An application of D-NEMs to embryonic stem cell development in mice reveals a feed-forward loop dominated network, which stabilizes the differentiated state of cells and points to Nanog as the key sensitizer of stem cells for differentiation stimuli.

Bettina Knapp/Lars Kaderali, Bioquant Heidelberg

Friday, Sep 18 9.00

Reverse Engineering of Signaling Pathways from RNAi Data

The inference of signal transduction and genetic regulatory networks is a major goal in systems biology. Systematic screenings of RNA interference (RNAi) offer the possibility to identify genes related with a particular phenotype or cellular pathway of interest (Fire, 1998). The temporal and spatial placement of these genes in the respective cellular pathway remains a challenging problem (Moffat and Sabatini, 2006). While Sacher et al. (2008) cluster

phenotypes, Markowitz et al. (2007) use the nested structure of effects of different knockdowns to solve this problem. We propose a stochastic model with Boolean networks for pathway inference where the activation probabilities for each gene are described by sigmoid functions. A Markov chain Monte Carlo approach is used to infer model topology and model parameters simultaneously, by sampling from the posterior distribution over model parameters given the knockdown data in a Bayesian setting. We compute the exact transition probabilities between different network states using the effect of single- or combinatorial knockdowns. Incomplete observations are integrated out via marginalization over unobserved nodes. To address the problem of under determined model parameters we use a prior distribution on the model parameters. We then approximate the likelihood allowing to sample from the posterior distribution without explicit evaluation of the likelihood whereas the sampling from the posterior for networks with larger number of nodes is permitted. We evaluated our method on a small artificial network with five nodes and we present results from the inference of the Jak/Stat signal transduction pathway in a hepatoma cell line given the knockdown data of 11 genes of the core Jak/Stat pathway.

Grégoire Pau, EMBL Heidelberg

Friday, Sep 18 9.30

Prediction of gene function by automated cellular phenotyping and genome-wide RNAi

Phenotyping of cellular model systems through high content screening (automated microscopy and image analysis) is a powerful approach to associate genes with biological processes. It also opens the possibility to systematically assay genetic and chemical perturbations and their interactions. I will describe an approach to the complete work flow of: image segmentation and feature extraction, screen quality assessment, distance metric learning, data presentation and integrative biologic analyses.

Berend Snijder/Lucas Pelkmans, ETH Zürich

Friday, Sep 18 10.00

Heterogenous population context determines cellular activity and virus infection patterns

Single-cell heterogeneity in cell populations arises from a combination of intrinsic and extrinsic factors. This heterogeneity has been measured for gene transcription, phosphorylation, cell morphology, and drug perturbations, and used to explain various

aspects of cellular physiology. In all cases however, the causes of heterogeneity were not studied. Here we analyze for the first time the heterogeneous patterns of related cellular activities, namely virus infection, endocytosis, and membrane lipid composition in adherent human cells. We reveal correlations with specific cellular states that are defined by the population context of a cell, and we derive probabilistic models that can explain and predict the majority of cellular heterogeneity of these activities, solely on the basis of each cell's population context. We find that accounting for population-determined heterogeneity is essential for interpreting differences between the activity levels of cell populations. Finally, we reveal that synergy between two molecular components, focal adhesion kinase and the sphingolipid GM1, enhances the population-determined pattern of SV40 infection. Our findings provide an explanation for the origin of heterogeneity patterns of cellular activities in adherent cell populations.

Tim Beissbarth, Uni Göttingen

Friday, Sep 18 11.00

Reconstruction of signaling networks from gene intervention data

Signalling processes are the key to understand the functions of a living cell. These processes are often complex and little understood. In Systems Biology different levels of quantitative and qualitative modeling of these processes are proposed in order to predict the response of cellular systems or drug treatments for complex diseases. Here, we focus on methods that reconstruct the architecture of such networks. Mere correlation analysis is usually not enough to understand the complex modes of action in a living cell. Active interventions into the system, for example by up- or downregulating certain genes, are crucial to reconstruct signalling networks. RNAi has become a useful tool to quickly produce such interventions. Modern techniques for gene or protein expression analysis are efficient tools to monitor the effects of such interventions. We are working on the development of several methods to reconstruct signalling networks from interventional data: Nested Effects Models are designed to reconstruct signalling networks of few interventions based measurement of high-dimensional effects. Deterministic Effects Propagation Networks on the other hand are designed to reconstruct signalling networks from high-dimensional interventions with direct effects measurements. Here we will give a short overview of these methods and demonstrate their application on example datasets connected to the medical treatment of cancer.

Volker Schmid, LMU München

Friday, Sep 18 11.30

Bayesian Modelling for Perfusion Imaging

Perfusion imaging aims to investigate the kinetics in human tissue in vivo. This is of interest in particular in oncology, cardiology and neurology. Using a (magnetic) contrast agent, a series of (magnetic resonance) images is obtained, which show the distribution of contrast agent in the tissue over time. Such scans are typically analysed using kinetic models composed of a (maybe unknown) input function and an (usually exponential) response function. In order to assess the tissue perfusion, one has to perform deconvolution or optimize the highly non-linear convolved model. The latter approach is typically prone to convergence problems, whereas deconvolution often is affected by numerical instability. We present a Bayesian approach to model perfusion images. Prior knowledge about kinetic parameters allows for a more robust estimation of these parameters and the computation of interval estimators. Contextual information can be used to make the models even more robust and/or to find connections between tissue voxels. Spatial priors (Gaussian Markov random fields) are used to account for spatial context and to reduce estimation errors. Temporal smoothing priors (penalty splines) are used to reduce observation noise. At last, we present a comprehensive model for the analysis of complete drug studies with perfusion imaging.

Fabian Theis, Helmholtzzentrum München

Friday, Sep 18 12.00

Independent subspace analysis and extraction

Matrix factorization algorithms, in particular nonnegative matrix factorization, principal and independent component analysis (ICA), have recently found successful, important applications in the analysis of biological recordings such as microarray data and fluorescence image stacks. Here we focus on separation based on statistical independence. Separation using independence may only be applied to data following the generative ICA model in order to guarantee algorithm-independent and theoretically valid results. Subspace ICA models generalize the assumption of component independence to independence between groups of components. They are attractive candidates for dimensionality reduction methods, however are currently limited by the assumption of equal group sizes or less general semi-parametric models. By introducing the concept of irreducible independent subspaces or components, we present a generalization to a parameter-free mixture model, and prove separability. More generally, we ask how to identify and extract subspaces in data based on statistical properties such as non-Gaussianity or signal color (autocorrelations). In the first part of my talk, I will review some matrix factorization techniques and results with a focus towards ICA. Then I will focus on subspace extraction for dimension reduction and finally for independent subspace analysis itself.

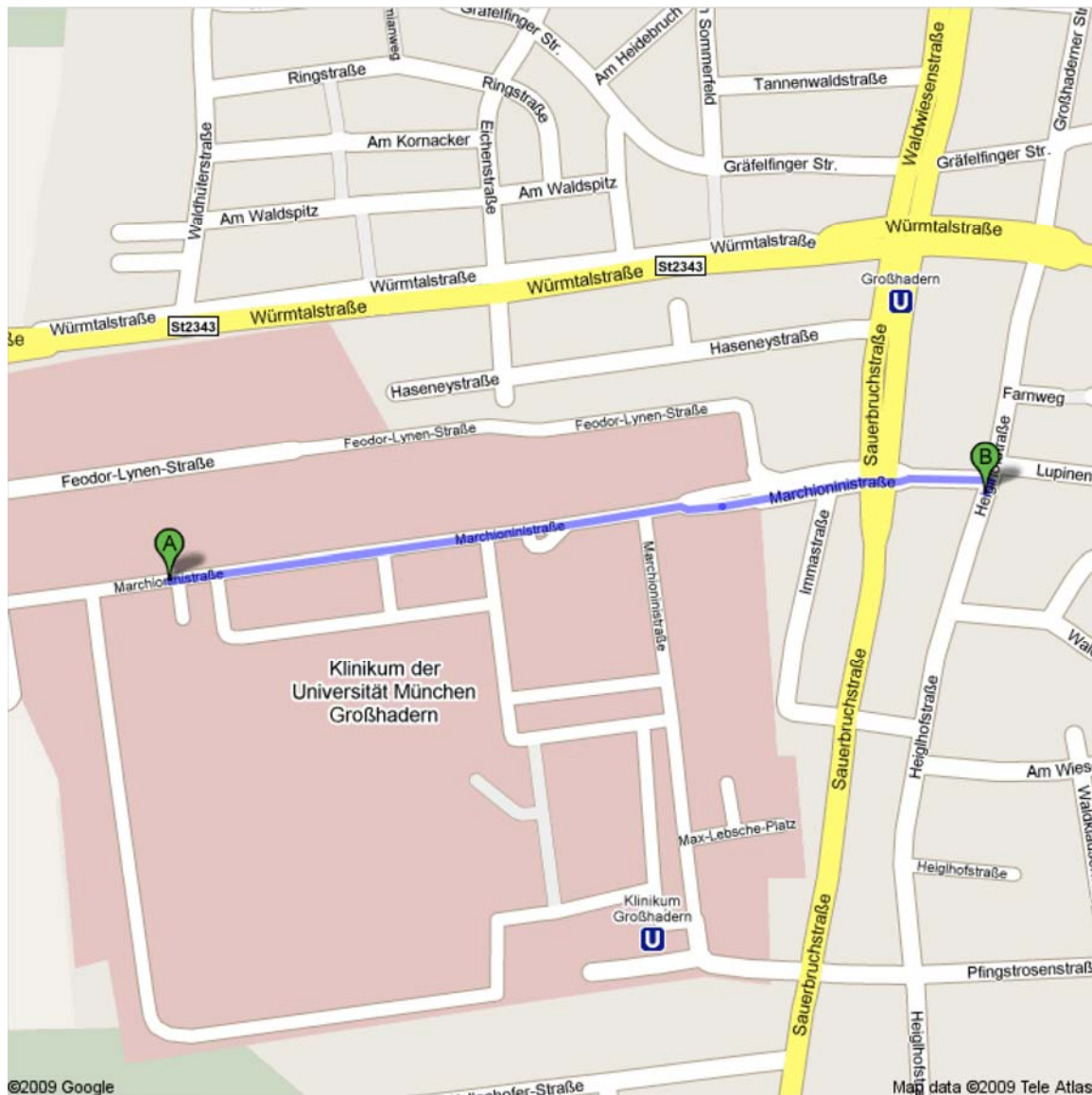
Dinner: Thursday, Sep 17 19.30h

Gasthaus Erdinger Weissbräu (NOT Weißes Bräuhaus)

Heighofstr. 13

81377 München

Phone: 089 7194300



1. Richtung **Ost** auf **Marchioninstraße** Ca. 4 Minuten

2. Bei **Heighofstraße** **rechts** abbiegen

List of participants:

Name	Affiliation
Achim Tresch	LMU München
André König	Uni Dortmund
Anja von Heydebreck	Merck Serono, Darmstadt
Anne-Laure Boulesteix	LMU München
Benedict Anchang	Uni Regensburg
Berend Snijder	ETH Zürich
Bernd Fellinghauer	ETH Zürich
Bettina Knapp	University of Heidelberg
Björn Schwalb	LMU München
Christoph Bernau	LMU München
Daniel Schmidl	TU München
Dominik Wittmann	Helmholtz München
Edgar Delgado-Eckert	ETH Zürich
Esther Herberich	LMU München
Fabian Theis	Helmholtz München
Gregoire Pau	EMBL Heidelberg
Ingo Röder	Uni Leipzig
Ingo Ruczinski	Johns Hopkins University Baltimore
Johannes Soeding	LMU München
Jörg Rahnenführer	Uni Dortmund
Juliane Schaefer	Uniklinik Basel
Julien Gagneur	EMBL Heidelberg
Kai Kammers	Uni Dortmund
Katja Ickstadt	Uni Dortmund
Katrin Knies	Uni Dortmund
Klaus Schliep	Massey University, Palmerston North, New Zealand
Lucas Pelkmans	ETH Zürich
Ludwig Fahrmeir	LMU München
Maren Vens	Universität zu Lübeck
Markus Schmidberger	LMU München
Marloes Maathuis	ETH Zürich
Martin Schäfer	Uni Dortmund
Matthias Schmid	Friedrich-Alexander-Universität

Matthias Siebert	Erlangen-Nürnberg
Mohammed Sadeh	LMU München
Nicole Radde	Uni Regensburg
Niko Beerenwinkel	Uni Stuttgart
Olga Ermakova	ETH Basel/Zürich
	Russian Academy of Science, Ekaterinburg (Ural Department)
Pauli Rämö	ETH Zürich
Rainer Spang	Uni Regensburg
Ramin Norousi	HTW Aalen
Roman Pahl	Uni Marburg
Theresa Niederberger	LMU München
Tim Beissbarth	Uni Göttingen
Ulrich Mansmann	LMU München
Volker Schmid	LMU München
Wanseon Lee	Helmholtz München
Willi Sauerbrei	Uni Freiburg