Aspects of Statistical Learning in Complex Systems

Rainer Opgen-Rhein

München 2007

Aspects of Statistical Learning in Complex Systems

Rainer Opgen-Rhein

Inaugural-Dissertation
zur Erlangung des Grades Doctor oeconomiae publicae
(Dr. oec. publ.)
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig–Maximilians–Universität München

vorgelegt von
Rainer Opgen-Rhein

München, 2007

Referent: Prof. Dr. Ludwig Fahrmeir
Koreferent: PD Dr. Christian Heumann
Betreuer: Dr. Korbinian Strimmer
Promotionsabschlußberatung: 18. Juli 2007

Contents

Part I

Statistical Learning in Complex Systems

Chapter 1

Introduction

1.1 Overview

A great challenge in science is the analysis of complex systems. Traditionally, it was only possible to successively examine small parts of these systems and assort the results. Nevertheless, due to possible interactions across all parts of the systems, knowledge of the components alone cannot lead to a full understanding of complex systems. The recent arising of new technologies however made it possible to simultaneously observe a large amount of variables as well as analyze the attained data.

The high dimensional data structure causes traditional methods no longer to be directly applicable for statistical learning. Most notably, research in systems biology stimulated the development of new methods for learning in complex systems. However, similar problems arise in various areas of scientific research like economics, finance, astronomy, meteorology or medicine.

This thesis is concerned with developing interpretable models for prediction and inference in complex systems that primarily build on a Stein-type shrinkage approach. It is based on the following seven articles:

Article A: Accurate Ranking of Differentially Expressed Genes by a Distribution-Free Shrinkage Approach, by Rainer Opgen-Rhein and Korbinian Strimmer, published in *SAGMB* (Volume 6, Issue 1, Article 9, 2007); (Opgen-Rhein and Strimmer, 2006a)

Article B: Inferring Gene Dependency Networks from Genomic Longitudinal Data: A Functional Data Approach, by Rainer Opgen-Rhein and Korbinian Strimmer, published in *REVSTAT* (Volume 4, Number 1, pp. 53–65, March 2006); (Opgen-Rhein and Strimmer, 2006b)

Article C: Using Regularized Dynamic Correlation to infer Gene Dependency Networks from time-series Microarray Data, by Rainer Opgen-Rhein and Korbinian Strimmer, refereed conference proceedings of *WCSB 2006*, pp. 73–76; (Opgen-Rhein and Strimmer, 2006e)

Article D: Condition Number and Variance Inflation Factor to Detect Multicollinearity, by Rainer Opgen-Rhein, technical report 01/07; (Opgen-Rhein, 2007)

Article E: Learning Causal Networks from Systems Biology Time Course Data: An Effective Model Selection Procedure for the Vector Autoregressive Process, by Rainer Opgen-Rhein and Korbinian Strimmer, published in *BMC Systems Biology* (Volume 8, Supplement 2, S3): Proceedings of PMSB 2006 ("Probabilistic Modeling and Machine Learning in Structural and Systems Biology"), Tuusula, Finland, 17-18 June 2006 ; (Opgen-Rhein and Strimmer, 2007b)

Article F: From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data, by Rainer Opgen-Rhein and Korbinian Strimmer, published in *BMC Systems Biology* (Volume 1, Article 37, 2007); (Opgen-Rhein and Strimmer, 2007a)

Article G: Reverse Engineering Genetic Networks using the **GeneNet** package, by Juliane Schäfer, Rainer Opgen-Rhein and Korbinian Strimmer, published in *R News* (Volume 6, Number 5, December 2006, pp. 50-53); (Schäfer et al., 2006b)

The articles are henceforth referred to by their character. A short description of the individual articles can be found in appendix (i).

This part I of the thesis, "Statistical Learning in Complex Systems", offers a summary of the theory and the methods of the articles. It elucidates the structure of the arguments that render learning in complex systems possible. Additionally, it provides some theory and further background for the methods developed in the articles. Part II, "Articles" provides the essentially unmodified publications, merely their layout was adjusted.

1.2 Contribution of this thesis

We will now summarize the contributions of this thesis. The new theory, methods and algorithms developed in this work are the following:

- a new Stein-type shrinkage estimator for the variance and the covariance matrix for large dimensional data sets (*Article A*)
- an extension of the shrinkage estimator, called *limited translation* for the covariance matrix, which takes into account the risk of individual components (*Article A*)
- development of a *shrinkage t* statistic for high-dimensional case-control analysis (*Article A*)
- introduction of *dynamical correlation*, a consistent extension of the concept of correlation for longitudinal data (*Article B*)
- development of a Stein-type shrinkage estimator for the dynamical correlation applicable in the "small n, large p" setting (*Article C*)

- a new interpretation and decomposition of the regression coefficient in linear regression (*Article D*)
- introduction of small sample *shrinkage regression* (*Article E*)
- a new small sample model selection method for the VAR-process (*Article E*)
- a new method of estimating partially causal networks (*Article F*)

The new methods were tested in simulations and applied to analyze the following high-dimensional data sets:

- *Arabidopsis thaliana* data set by Smith et al. (2004) in *Article E* and *Article F*
- *Escherichia coli* data set by Schmidt-Heck et al. (2004) in *Article G*
- *human T-cell* data set by Rangel et al. (2004) in *Article B* and *Article C*
- *Affymetrix spike-in study* by Cope et al. (2004) in *Article A*
- *"golden spike" Affymetrix experiment* by Choe et al. (2005) in *Article A*
- *HIV-1 infection study* by van 't Wout et al. (2003) in *Article A*

All statistical procedures described are implemented in packages of the computer program R (R Development Core Team, 2006), which is available under the terms of the GNU General Public License and can be found in the CRAN archive (`http://cran.r-project.org`). Specifically, the methods can be found in the following packages:

- *GeneNet*: Modeling and inferring gene networks (Opgen-Rhein et al., 2007)
- *corpcor*: Efficient estimation of covariance and (partial) correlation (Schäfer et al., 2006a)
- *st*: "shrinkage t" statistic (Opgen-Rhein and Strimmer, 2006d)
- *longitudinal*: Analysis of multiple time course data (Opgen-Rhein and Strimmer, 2006c)

A more detailed description of the packages can be found in appendix (ii). The packages and data sets can also be found on the following website: `http://strimmerlab.org/`.

Chapter 2

Complex Systems

2.1 The challenge of learning in complex systems

The focus of this thesis lies in gaining knowledge about complex systems, in which a large number of components interact and conjointly affect the state of the system in future points of time. The term "complex systems" itself is not explicitly defined in science and leaves room for interpretation. Generally, a complex system can be understood to be one which properties cannot be fully explained by a knowledge merely of the component parts (Gallagher and Appenzeller, 1999).

This character of complex systems renders the inference of their structure challenging, as all parts of the system have to be taken into consideration simultaneously to achieve reliable information. With the appearance of new high throughput technologies, e.g., in biology or astronomy, much progress has been made in the measurement of the components, so that – for large systems – hundreds of variables can be observed at the same time. At the same time, the number of repetitions for the observation of each item is often restricted, so that the number of variables possibly by far exceeds the number of observations. The key problem for learning in complex systems is that in this "small n, large p" paradigm standard estimation procedures tend to be unreliable or to be even inapplicable (West et al., 2000).

The methods developed in this thesis are applied to complex systems in biology which mainly appear in a discipline called systems biology. Nevertheless, they are not restricted to this scientific area. In contrast, the same data structure is encountered in various fields of contemporary research. Examples are finance, where, e.g., portfolio optimization requires the estimation of covariances between hundreds of stocks (Mantegna and Stanley, 2000; Ruppert, 2004; Ledoit and Wolf, 2003a), management, e.g., for business optimization (Huang et al., 2006; Bickel and Levina, 2006), astronomy (e.g. Kabán et al., 2006; Patat, 2003), meteorology (e.g., Murphy and Wilks, 1998; Levine and Berliner, 1999) or medicine (e.g. Efron, 2005c).

However, the development of new or improvement of established statistical methods of learning in complex systems was heavily stimulated by systems biology, where microarray

technology initialized the production of large data sets, and where new technologies like time of flight spectroscopy, proteomic devices or flow cytometry are likely to generate even larger quantities of data (Efron, 2005c).

These developments are challenges for statistical inference. Nevertheless, the consequences of multidimensional data is often avoided, either by ignoring the statistical problems arising by large scale estimation (e.g. Mantegna and Stanley (2000)) or by a deterministic reduction of the dimensionality like concentrating on prominent variables and forgoing spatial details as done for weather forecast in Pappenberger et al. (2005).

An extensive overview of the methods used in systems biology can be found in Klipp et al. (2005). In this work the focus lies on regularization techniques based on James-Stein estimation (James and Stein, 1961). Although the inference techniques introduced in the following chapters will be applied to microarray data, it is – again – to stress that they are suited for all kind of multidimensional data in the "small n, large p" paradigm.

2.2 The use of models in systems biology

Generally, the interest of research lies in understanding the true nature of a system. Nevertheless, it is epistemological knowledge that this cannot be accomplished; it is only possible to observe phenomena produced by it. These observations can be used to create models of the system. Ideally, this model should be of the same complexity as the system itself to include all structures of the system and to be able to make reliable predictions in all circumstances. However, this is neither possible due to incomplete knowledge and limited computational capacities nor is it desirable: to gain an understanding of complex systems, this complexity has to be reduced, which is an important function of models. Different models concentrate on different aspects of complex systems. In this sense, a model cannot be right or wrong, it can be only more or less useful for different purposes.

We will elaborate this for genetic networks in systems biology. Important aspects are the structures (e.g. the network of gene interactions and biochemical pathways), the dynamics (behavior over time), the control methods (mechanisms controlling the state of the cell) or the design principles (the strategies of modification and construction of biological systems) (Kitano, 2002).

One possibility in modeling complex systems is to focus on different levels of scaling. An example is given in Fig. 2.1. In the left column the modeling is concentrated on describing the molecular details of the characteristics of a single gene (e.g. with differential equations). The remainder of the system (all the other genes) is ignored, as the extrapolation of this detailed model would be prohibitively complicated. The center left column focuses on the circuit of a few genes, and the center right and right column deal with increasingly larger networks of genes. Nevertheless, the details of kinetic properties of genes have to be ignored when modeling the relations among them.

The different models introduced in this thesis all try to capture the relationships among a large set of variables (here: genes), but take different points of view. Section 4.2, e.g., concentrates on correlation networks, which are extensively used in biology. They allow

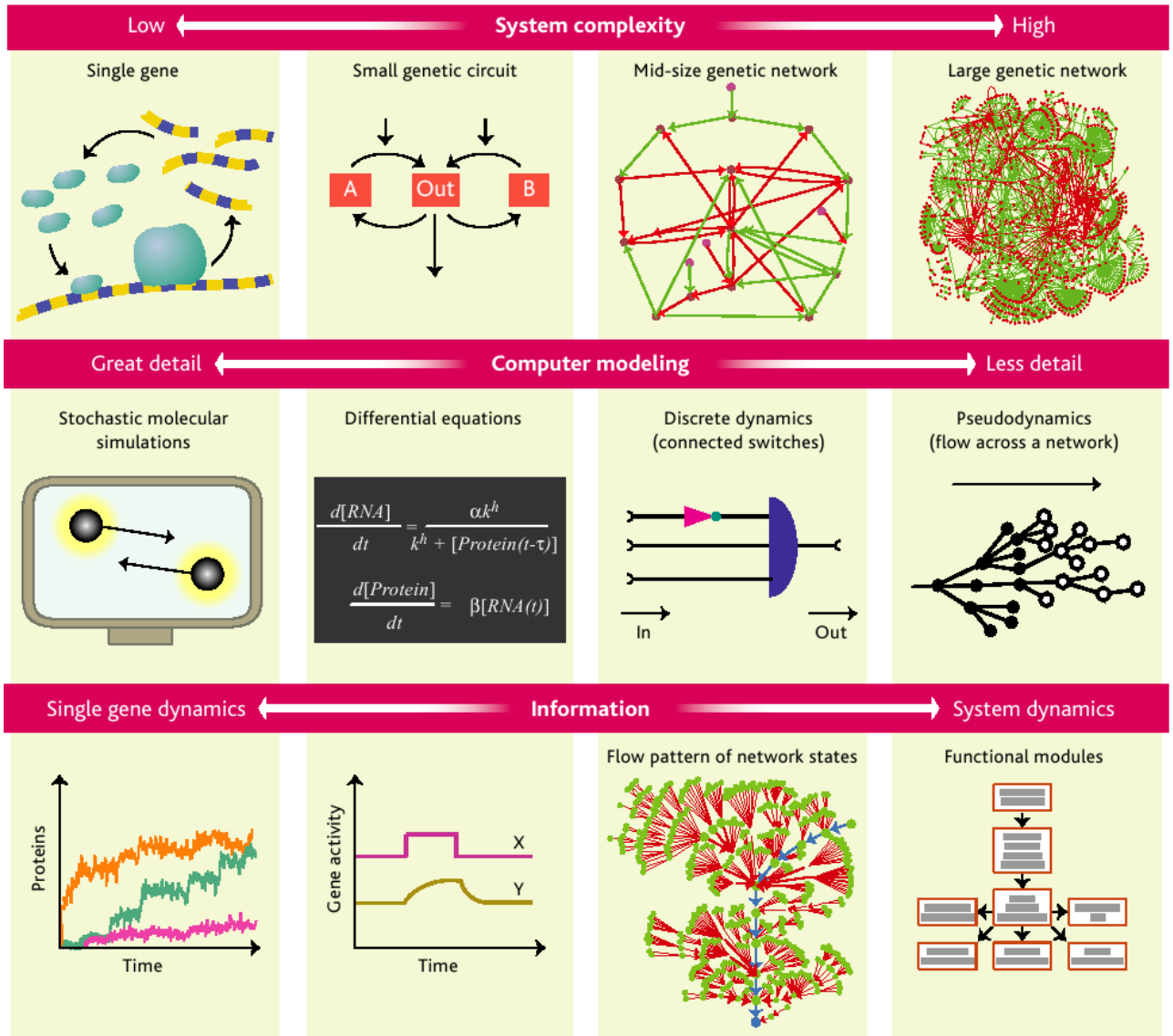

Figure 2.1: Different levels of description in models of genetic networks (figure source: Bornholdt, 2005)

identifying strongly correlated genes. This model is not wrong, but does not serve the intention of discovering functional relations between genes, which is usually the interest in systems biology. For this purpose, graphical Gaussian models will be introduced in section 4.3. Another important aspect lies in the incorporation of dynamical aspects into models as done in chapter 5 or in the identification of causal structures (chapter 6).

The next section introduces microarray data, which has a data structure typical for complex systems, as the number of variables is much larger as the number of observations. This data will be used to demonstrate the methods developed in this thesis.

2.3 Microarray data and their biological background

To understand microarray data and their connection to genetic networks, some biological background has to be provided. The particulars of the processes of molecular biology and the technical methods utilized to gain microarray data can only be indicated, for more details see, e.g., Graur and Li (2000); Gibson and Muse (2004); Klipp et al. (2005).

Cells are the basic structure and the functional units of every living system. All information needed to direct their activities and therefore to sustain life is contained in a sequence of four different bases or nucleotides: adenine (A), thymine (T), cytosine (C), and guanine (G), which compose the deoxyribonucleic acid (DNA). It consists of two long strands that have the shape of a double helix. The double-stranded structure evolves, as each type of nucleotide on one strand pairs with just one other type on the other strand: adenine forms a bond with thymine and cytosine with guanine. A part of the DNA (the coding region) consists of genes, segments that contain instructions about the synthesis and regulation of proteins. Proteins control the structure and function of a cell and are the essential parts in building complex living systems. An idea of the increasingly complex structures of life, all building up on genes, can be found in Fig. 2.2. The human genome is estimated to contain about 20,000 – 25,000 genes. However, not the number of genes is decisive: the difference between distinct species lies mainly in the regulatory program. This means that the creation, function and different amounts of gene products in the living system can only be understood by taking the activation and interaction of different genes into consideration.

The activation of genes, called gene expression, can be measured using microarray technology; the interaction between different genes has to be inferred from the gene expression of a large set of genes using statistical methods. Microarray technology is therefore the method to generate the data for statistical analysis. It builds upon the central dogma of molecular biology (Crick, 1958), which describes the process of synthesizing proteins from genes. The idea is the following:

- The first step is *transcription:* RNA polymerase, an enzyme complex, and transcription factors transfer the information of a gene to the single-stranded messenger RNA (mRNA). This often includes further processing via alternative splicing, where parts of the mRNA are removed and rearranged.

- Afterwards, *translation* takes place: ribosomes read the mRNA and build the protein by adding amino acids in the sequence given by the gene and by subsequently folding it into the correct conformation.

The term *gene expression* usually refers to the amount of mRNA. Microarray technology, which measures the gene expression, takes advantage of the fact that the single-stranded mRNA binds to oligonucleotides with complementary nucleotide sequence. A large number of gene-specific *probes* consisting of a complementary DNA (cDNA) are attached to the array surface.

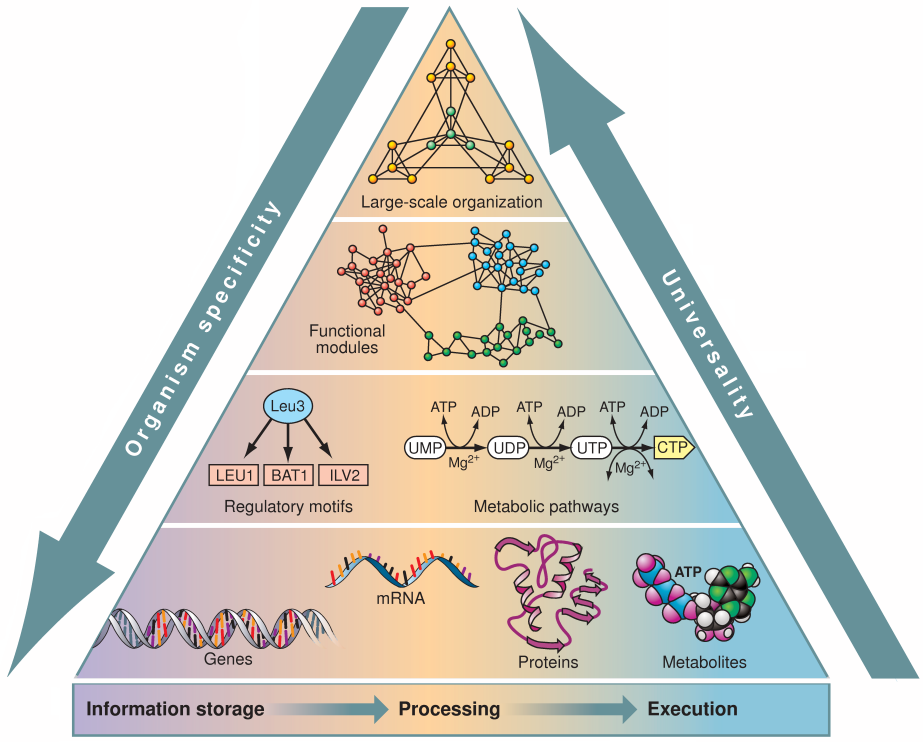

Figure 2.2: Life's complexity pyramid (figure source: Oltvai and Barabási, 2002)

The prepared sample, called *target* cDNA, consists of preprocessed RNA from the cells or tissues of interest. It is brought in contact to the array and *hybridization* can take place. The methods to achieve measurements of the gene expression differ. Oligonucleotide expression arrays (Lockhart et al., 1996), also known by the trademark *Affymetrix GeneChip*, consists of pairs of oligonucleotides. In the spots for different genes are located not only oligonucleotides which match perfectly to the specific gene (perfect match - PM), but they are paired with a probe called mismatch (MM), in which the middle base of the 25 base pairs of the oligonucleotide is changed. This allows measuring non-specific binding. In contrast, cDNA microarrays use two differently marked samples called the two channels of a two-color microarray experiment, that hybridize competitively with the target. A laser determines the intensity value for each spot, which is supposed to represent the expression for each gene.

It is to remark that the design of microarray experiments is an important factor for the quality of the resulting data (Yang and Speed, 2002; Churchill, 2002). This is especially important, as the biological as well as the technical variation tends to be high in microarray experiments. This is also a reason for the fact that the data cannot be used directly but has to be preprocessed before analyses can take place, e.g. the data has to be calibrated and transformed (Huber et al., 2002, 2003).

Nevertheless, microarray data of even high quality should be handled with care. Post-translational modifications, alternative splicing and other modifications that take place in

all steps of the expression process do not necessarily allow to draw direct conclusions about the interaction of the genes on the one hand nor about the interplay of proteins on the other hand. A full understanding of the dynamical behavior of living systems can only be achieved by taking all steps of the life's complexity pyramid of Fig. 2.2 into view and by studying the biological system at different cellular levels. Proteomics, the large-scale study of proteins in a cell, especially their structures, functions and interaction is often considered the next step in systems biology (Auerbach et al., 2002). However, the data structure of proteomics is even more demanding than microarray data.

We will now introduce the microarray data used to demonstrate the methods of high-dimensional inference developed in this thesis.

2.4 Microarray data sets analyzed in this thesis

2.4.1 Microarray time course data sets

Arabidopsis thaliana time course data

The dataset that will be mainly used to study the different models for learning in complex systems is a microarray time course data set of the gene expression of *Arabidopsis thaliana*. Specifically, we reanalyze expression time series resulting from an experiment investigating the impact of the diurnal cycle on the starch metabolism of *Arabidopsis thaliana* (Smith et al., 2004).

For this, we downloaded the calibrated signal intensities for 22,814 probes and 11 time points for each of the two biological replicates from the NASCArrays repository (`http://affymetrix.arabidopsis.info/narrays/experimentpage.pl?experimentid=60`). After log-transforming the data we filtered out all genes containing missing values and whose maximum signal intensity value was lower than 5 on a log-base 2 scale. Subsequently, we applied the periodicity test of Wichert et al. (2004) to identify the probes associated with the day-night cycle. As a result, we obtained a subset of 800 genes. An example can be found in Fig. 2.3, where the expression levels for the first nine genes are displayed. The subset of 800 genes of the *Arabidopsis thaliana* data is used throughout this work.

The *Arabidopsis thaliana* data were measured at $0, 1, 2, 4, 8, 12, 13, 14, 16, 20,$ and 24 hours after the start of the experiment. Nevertheless, in chapter 4 we assume no temporal structure in the data. To compare these methods with learning in dynamic systems, we still use the Arabidopsis dataset, but ignore the dynamic aspects of the data and assume that all measurements were taken at the same time point. This leads to 22 observations (2 replications times 11 time points) for each of the 800 genes.

Throughout this work we mainly assume the dependencies among the variables in the complex systems to be essentially linear. This can easily be checked by inspecting the pairwise scatter plots of the data. In Fig. 2.4 the correlations and the pairwise scatter plots of the calibrated expression levels are depicted. We see that for the 800 considered *Arabidopsis thaliana* genes the linearity assumption is indeed satisfied.

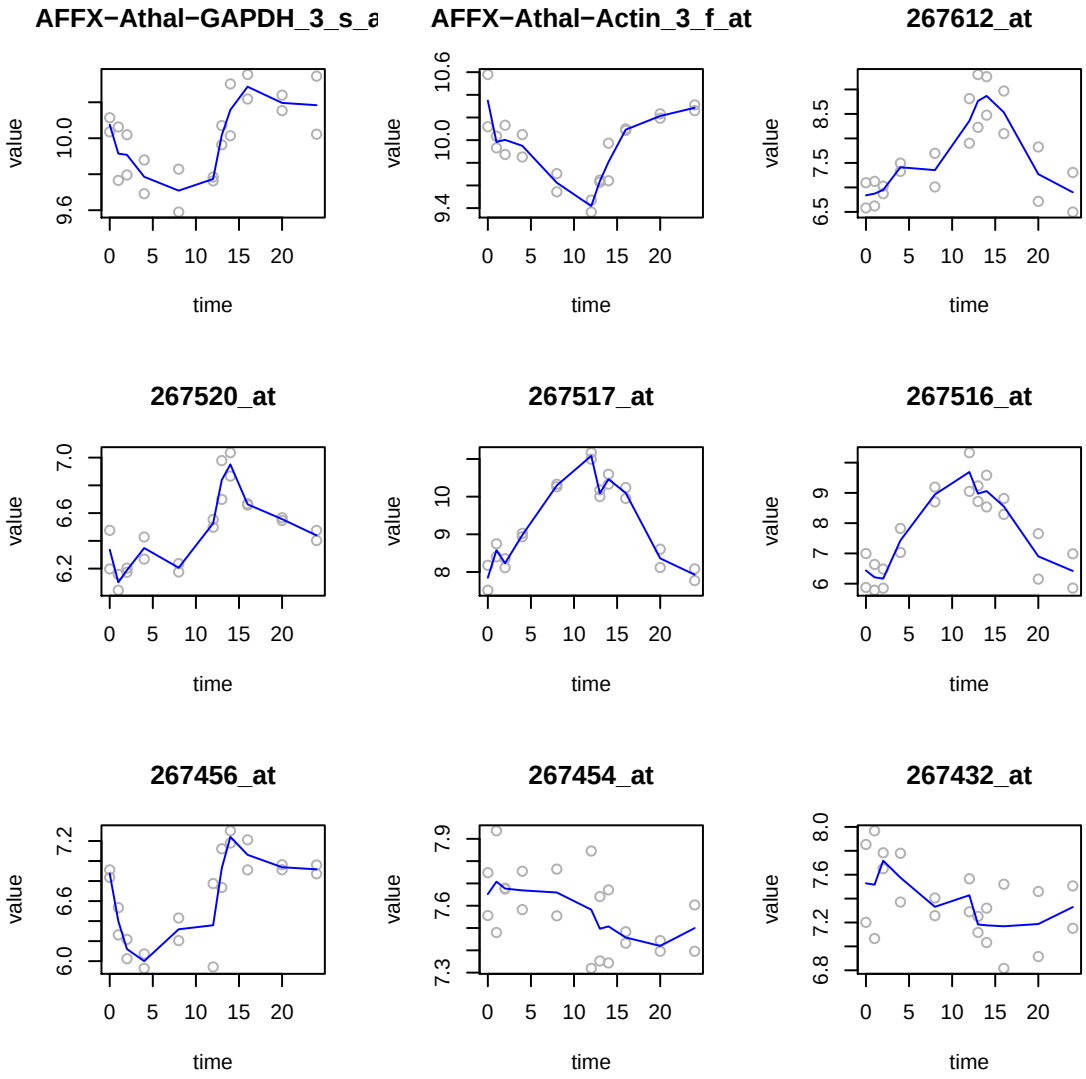

Figure 2.3: *Arabidopsis thaliana* time course data: expression levels

The annotation and a short description of most of the nodes which are shown in the inferred networks throughout this work can be found in the supplementary information of *Article E*. A complete description of all genes can be retrieved along with the *Arabidopsis thaliana* data set from the R-package "GeneNet" (Opgen-Rhein et al., 2007).

Escherichia coli data set

The *Escherichia coli* data set by Schmidt-Heck et al. (2004) describes a stress response experiment for the microorganism *Escherichia coli*. Under a steady state condition an overexpression of the recombinant protein SOD (human superoxide dismutase) was induced

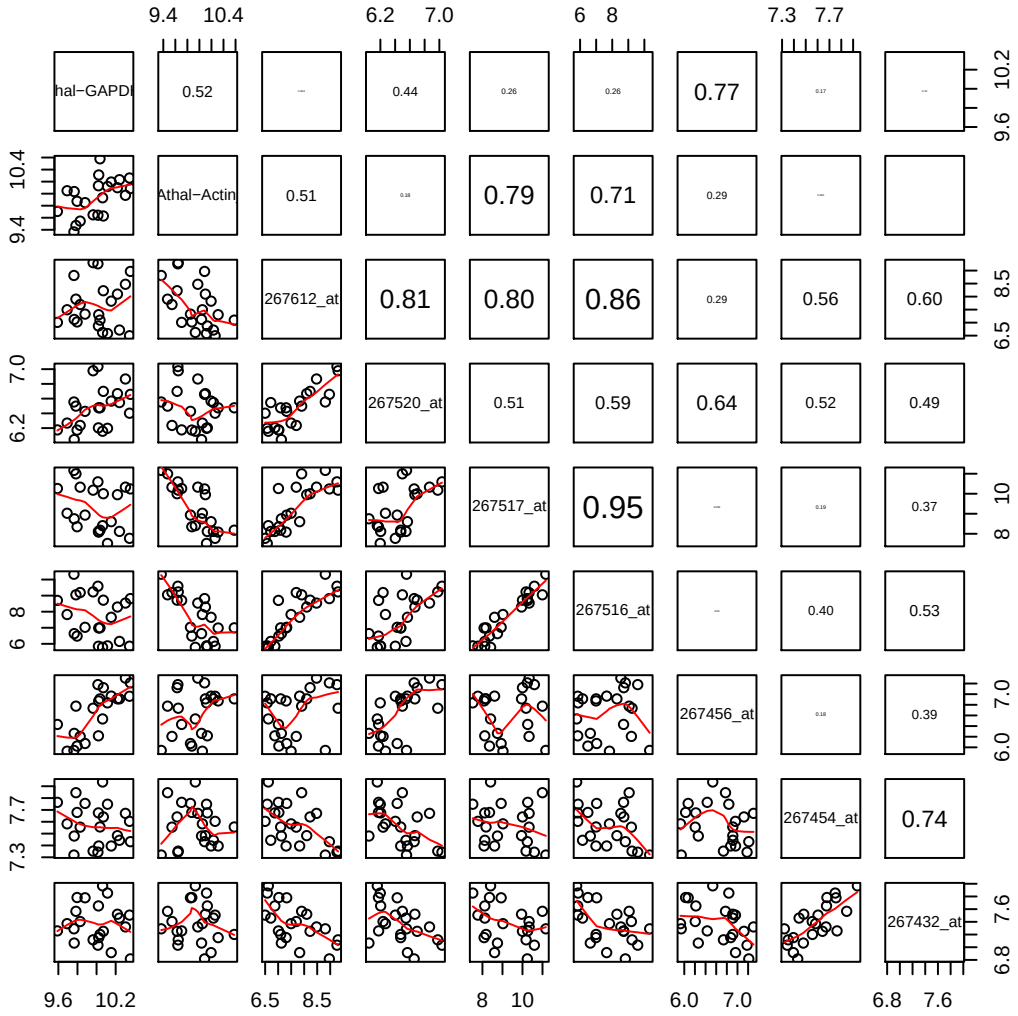

Figure 2.4: *Arabidopsis thaliana* time course data: pairwise correlations

by a dosage of IPTG (isopropyl-beta-D-thiogalactopyranoside, a nonmetabolizable analog of the normal substrate, lactose). All 4289 protein-coding genes were monitored during the process using microarrays. Samples were taken at 9 time points, at 8, 15, 22, 45, 68 90, 150 and 180 minutes after induction. Normalization by the LOWESS method (Cleveland, 1979) leads to 102 differentially expressed genes which form the *Escherichia coli* data set.

Human T-cell data set

The *human T-cell* data set was introduced by Rangel et al. (2004). These data characterize the response of a *human T-cell* line (Jirkat) to a treatment with PMA and ioconomin. After

preprocessing the time course data consist of 58 genes measured across 10 time points with 44 replications. The measurements in the experiment were taken at unequally spaced time points, i.e. after 0, 2, 4, 6, 8, 18, 24, 32, 48, and 72 hours after treatment. A peculiarity of the data set is the large number of replications compared to the number of variables. This allows introducing the concept of dynamical correlation in *Article B* without the necessity of regularization.

2.4.2 Microarray case-control data sets

The following three data sets are used in section 4.1 to compute gene ranking of differential expressed genes. In these data sets the differentially expressed genes are known.

The first data set is the well-known *Affymetrix spike-in study* that contains 12,626 genes, 12 replicates in each group, and 16 known differentially expressed genes (Cope et al., 2004).

The second data is a subset of the *"golden spike" Affymetrix experiment* of Choe et al. (2005). From the original data we removed the 2,535 probe sets for spike-ins with ratio 1:1, leaving in total 11,475 genes with 3 replicates per group, and 1,331 known differentially expressed genes. We note that excluding the 1:1 spike-ins is important as these comprise a severe experimental artifact (Irizarry et al., 2006). Both the Choe et al. (2005) data and the Affymetrix spike-in data were calibrated and normalized using the default methods of the "affy" R package (Gautier et al., 2004).

The third data set is from the *HIV-1 infection study* of van 't Wout et al. (2003). It contains 4 replicates per group, and 13 of the 4,608 genes have been experimentally confirmed to be differentially expressed.

Chapter 3

Parameter estimation in complex systems

This chapter introduces a Stein-type shrinkage method for regularized inference. Stein estimators are a compromise between Bayesian estimation and frequentistic methods. First, the principle of Stein estimation is explained. Afterwards the necessity of regularized estimation of the covariance matrix in the "small n, large p" paradigm is demonstrated. Subsequently, Stein-type estimators for variance shrinkage and for the covariance matrix are developed. The methods are mainly described in *Article A* and *Article E*.

3.1 Stein-type shrinkage

3.1.1 Stein-phenomenon

High dimensional data in the "small n, large p" paradigm often requires the estimation of a large number of variables, e.g. the covariances between gene expression profiles. For their estimation a counterintuitive phenomenon called "Stein-phenomenon" can be exploited: even if each single variable is estimated using the maximum likelihood estimator, it is possible to construct estimators that (sometimes extremely) improve the estimation of all variables in terms of total risk. The risk is measured by a risk function, for example the total mean squared error

$$\mathrm{R}(\hat{\boldsymbol{\theta}}) = \mathrm{E}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^2, \tag{3.1}$$

which is the sum of the squared deviances of the p estimated parameters $\hat{\theta}_i$ from the p true parameters θ_i. We will throughout this work concentrate on this risk measure, but other risk functions could also be applied.

The basic idea of Stein estimation is that all separate estimators (e.g. the maximum likelihood estimators) of the different parameters are shrunken towards a common target (hence the expression "shrinkage"-estimation or "regularized" estimation). This effect was first demonstrated by Charles Stein (Stein, 1956) for the estimation of parameters from independent normal observations and further developed by James and Stein (1961). For a

detailed description of the original estimators see, e.g., Efron and Morris (1975) or Lehmann and Casella (1998).

The phenomenon does not rely on a connection among the estimated parameters: Stein (1956) demonstrated the construction of the improved estimator for *independent* parameters. The effect can rather be understood by recalling that the mean squared error (MSE) of an estimator can be decomposed into the sum of its variance and its squared bias

$$\mathrm{MSE}(\hat{\boldsymbol{\theta}}) = \mathrm{Bias}(\hat{\boldsymbol{\theta}})^2 + \mathrm{Var}(\hat{\boldsymbol{\theta}}). \tag{3.2}$$

James-Stein estimation increases the bias (or introduces bias for an unbiased estimator), but this can already be outbalanced by a reduction of the variance if the length of the estimated parameter vector $\hat{\boldsymbol{\theta}}$ is greater than two, if the focus lies on the total MSE.

In the following section we will introduce a general shrinkage estimation procedure that is based on James-Stein estimation, and subsequently apply it to the estimation of covariance matrices.

3.1.2 Distribution-Free Shrinkage Estimation

A James-Stein type shrinkage estimator can by constructed from an arbitrary unregularized estimator. Unlike many other estimation methods like Bayes-estimation no distribution of the data or the model parameters have to be assumed. The first step is to suppose the availability of an unregularized estimation rule

$$\delta^0 = \hat{\boldsymbol{\theta}}, \tag{3.3}$$

e.g., the maximum-likelihood or the minimum variance unbiased estimate. It is important here that $\hat{\boldsymbol{\theta}}$ is a *vector* $(\hat{\theta}_1, \ldots, \hat{\theta}_k, \ldots, \hat{\theta}_p)^T$. Then the James-Stein ensemble shrinkage estimation rule may be written as

$$\begin{aligned} \delta^\lambda &= \delta^0 - \lambda \boldsymbol{\Delta} \\ &= \hat{\boldsymbol{\theta}} - \lambda \, (\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}^{\mathrm{Target}}). \end{aligned} \tag{3.4}$$

In other words, the shrinkage estimate δ^λ is the linear combination $\lambda\hat{\boldsymbol{\theta}}^{\mathrm{Target}} + (1-\lambda)\hat{\boldsymbol{\theta}}$ of the original estimator $\hat{\boldsymbol{\theta}}$ and a target estimate $\hat{\boldsymbol{\theta}}^{\mathrm{Target}}$. The parameter λ determines the extent to which these estimates are pooled together. If $\lambda = 1$ then the target dominates completely, whereas for $\lambda = 0$ no shrinkage occurs.

An example can be seen in Fig. 3.1 where the mean of 10 independent normally distributed variables $X_{1,\ldots,10} \sim N(0,1)$ is estimated using $n = 1$ observation for each variable and three different targets: $\hat{\boldsymbol{\theta}}^{\mathrm{Target}} = 0$, $\hat{\boldsymbol{\theta}}^{\mathrm{Target}} = 1$ and $\hat{\boldsymbol{\theta}}^{\mathrm{Target}} = 2$. It is obvious that shrinkage towards the true value ($\hat{\boldsymbol{\theta}}^{\mathrm{Target}} = 0$) improves the MSE for larger λ. The decisive point in Stein-estimation is that any target, even the grossly misspecified target $\hat{\boldsymbol{\theta}}^{\mathrm{Target}} = 2$, has the potential to reduce the total risk. We see that the shrinkage estimation requires two tasks: first, the selection of a suitable target, as a well-chosen target allows a potentially greater improvement of MSE than a bad target, and secondly determining a

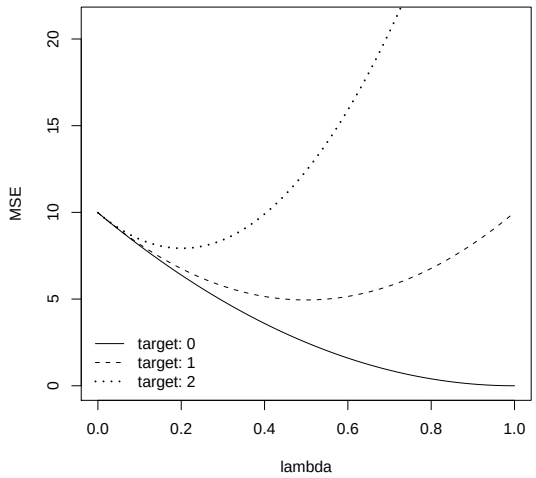

Figure 3.1: Principle of shrinkage estimation.

shrinkage intensity λ which exploits as much as possible of the potential reduction of risk for a given target.

In James-Stein estimation the choice of a target has to be done due to practical considerations. For correlations, e.g., it is often suitable to shrink towards zero; in other applications, the mean or the median of the unregulated estimators could be a useful target.

Once a target is chosen, the search for the optimal shrinkage intensity λ is considered from a decision theoretic perspective. First, a loss function is selected (e.g. the squared error). Second, λ is chosen such that the corresponding risk of δ^λ, i.e. the expectation of the loss with respect to the data (e.g. the mean squared error, MSE), is minimized. It turns out that for squared error loss this can be done *without* any reference to the unknown true value $\boldsymbol{\theta}$, as the MSE of δ^λ may be written as follows:

$$\begin{aligned} \mathrm{MSE}(\delta^\lambda) = \mathrm{MSE}(\hat{\boldsymbol{\theta}}) &+ \lambda^2 \sum_{k=1}^{p} \{E((\hat{\theta}_k - \hat{\theta}_k^{\mathrm{Target}})^2)\} \\ &- 2\lambda \sum_{k=1}^{p} \{\mathrm{Var}(\hat{\theta}_k) - \mathrm{Cov}(\hat{\theta}_k, \hat{\theta}_k^{\mathrm{Target}}) \\ &\qquad + \mathrm{Bias}(\hat{\theta}_k)\ E(\hat{\theta}_k - \hat{\theta}_k^{\mathrm{Target}})\} \\ =:\ & c + \lambda^2\, b - 2\lambda\, a. \end{aligned} \tag{3.5}$$

As already suspected by looking at Fig. 3.1, it turns out that the MSE risk curve has the shape of a parabola whose parameters a, b, and c are completely determined by only the

first two distributional moments of $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}^{\text{Target}}$.

Taking the first derivative of $\text{MSE}(\delta^\lambda)$,

$$\frac{d\ \text{MSE}(\delta^\lambda)}{d\lambda} = 2\lambda\ b - 2a, \tag{3.6}$$

gives some further insights concerning the shrinkage rule Eq. 3.4: Only a and b determine the risk improvement of δ^λ compared to the MSE of the unregularized estimate δ^0, which is given by c: $\text{MSE}(\delta^0) = c$. The same MSE as that of the unregularized estimate δ^0 can be achieved for $\lambda = 2\frac{a}{b}$. *Any* value of λ in the range between 0 (which yields the unregularized estimator) and $2\frac{a}{b}$ leads to a decrease in MSE.

Nevertheless, the aim lies in an optimal shrinkage intensity that results in overall minimum MSE. This is given by the simple formula

$$\lambda^\star = \frac{a}{b}. \tag{3.7}$$

In this case the savings relative to the unshrunken estimate amount to $\text{MSE}(\hat{\boldsymbol{\theta}}) - \text{MSE}(\delta^{\lambda^\star}) = \frac{a^2}{b}$. The factor b has a special interpretation: it measures the misspecification of target and estimate and therefore plays the role of a precision for λ.

Note that it is possible to allow for multiple shrinkage intensities. For instance, if the model parameters fall into two natural groups, each could have its own target and its own associated shrinkage intensity. In the extreme case each parameter could have its own λ.

Further discussion and interpretation of Eq. 3.7 may be found in Schäfer and Strimmer (2005b). It should only be noted here that special versions of the rule $\lambda^\star = \frac{a}{b}$ are well known, see, e.g., Ledoit and Wolf (2003b) who describe the multivariate case but require an unbiased $\hat{\boldsymbol{\theta}}$, or Thompson (1968) who considers only the univariate case and a non-stochastic target and also restricts to $\text{Bias}(\hat{\boldsymbol{\theta}}) = 0$.

3.1.3 Construction of a Shrinkage Estimator

In actual application of the shrinkage rule (Eq. 3.4), the variances and covariances in Eq. 3.5 are unknown and the pooling parameter λ therefore needs to be estimated from the data. Inevitably, this leads to an *increase* of the total risk of the resulting shrinkage estimator. However, it is a classic result by Stein that the cost of estimating the shrinkage intensity is already (and always!) offset by the savings in total risk when the dimension p is larger than or equal to three (e.g. Gruber, 1998).

One straightforward way to estimate the optimal $\lambda^\star$ is to replace the variances and covariances in Eq. 3.5 by their unbiased empirical counterparts, i.e. $\hat{\lambda}^\star = \frac{\hat{a}}{\hat{b}}$. The alternative, a search for an unbiased estimate for the whole fraction $\frac{a}{b}$, will only be possible in some special cases.

Despite its simplicity, the rule for the construction of Stein-type estimator $\delta^\lambda = \lambda\hat{\boldsymbol{\theta}}^{\text{Target}} + (1-\lambda)\hat{\boldsymbol{\theta}}$ (Eq. 3.4), together with an estimated version of $\lambda^\star = \frac{a}{b}$ (Eq. 3.7), provides instant access to several classic shrinkage estimators, and offers a simple and unified framework for their derivation.

For instance, consider the old problem of Stein (1956, 1981) of inferring the mean of a p-dimensional multivariate normal distribution with unit-diagonal covariance matrix from a single ($n = 1$) vector-valued observation – clearly an extreme example of the "small n, large p" setting. In this case the maximum-likelihood estimate equals the vector of observations, i.e. $\hat{\theta}_k^{\mathrm{ML}} = x_k$. However, shrinkage estimators with improved efficiency over the ML estimator are easily constructed. With the covariance being the identity matrix ($\mathrm{Var}(x_k) = 1$ and $\mathrm{Cov}(x_k, x_l) = 0$) and the target being set to zero ($\hat{\boldsymbol{\theta}}^{\mathrm{Target}} = 0$) one finds $\hat{a} = p$ and $\hat{b} = \sum_{k=1}^{p} x_k^2$ which results in the shrinkage estimator

$$\hat{\theta}_k^{\mathrm{JS}} = \left(1 - \frac{p}{\sum_{k=1}^{p} x_k^2}\right) x_k. \tag{3.8}$$

If we follow Lindley and Smith (1972) and shrink instead towards the mean across dimensions $\bar{x} = \frac{1}{p}\sum_{k=1}^{p} x_k$ we get $\hat{a} = p - 1$ and $\hat{b} = \sum_{k=1}^{p}(x_k - \bar{x})^2$ and obtain

$$\hat{\theta}_k^{\mathrm{EM}} = \bar{x} + \left(1 - \frac{p-1}{\sum_{k=1}^{p}(x_k - \bar{x})^2}\right)(x_k - \bar{x}) \tag{3.9}$$

It is noteworthy that these are *not* the original Stein estimators given in James and Stein (1961) and Efron and Morris (1973) but instead are exactly the shrinkage estimators of Stigler (1990) derived using a regression approach. We point out that the Stigler and our versions have the advantage that they are applicable also for $p = 1$ and $p = 2$.

Some further remarks are interesting:

- As we will see by inferring a regularized estimator for the covariance matrix, using the above equation leads to an almost automatic procedure for shrinkage estimation
- The construction of the estimator assumes at no point a normal or any other distribution
- Shrinkage estimation can also be understood in terms of hierarchical models, where λ compromises between unpooled and completely pooled estimates: for $\lambda = 0$ (no shrinkage), every variable is treated separately, and for $\lambda = 1$ (target dominates completely), all variables belong to the same group (Gelman and Pardoe, 2006; Raudenbush and Bryk, 2001)
- Many empirical Bayes estimators can be put into the form of Eq. 3.4, (e.g. Gruber, 1998). Note that using Eq. 3.5 allows deriving these estimators without first going through the full Bayesian formalism!

In the following section we describe a Stein-type estimation of the covariance matrix, an essential tool for analyzing complex systems.

3.2 Shrinkage estimation of the covariance matrix

The main application for Stein-type inference in this work is the estimation of the covariance structure of the variables. The reason for the central role of the covariance matrix is the possibility to infer all linear relationships among the variables *directly* from the covariance matrix, without further estimation.

A suitable estimator for the covariance matrix is not only required to minimize the total MSE, but should also guarantee a well-conditioned matrix, as most of the methods developed here require the inversion of the covariance matrix. The usual sample covariance estimator does not guarantee this, and will almost for sure produce ill-conditioned covariance matrices in the "small n, large p" paradigm. The Stein-type shrinkage estimator for the covariance matrix developed in this chapter not only dramatically reduces the total risk (also called ensemble risk) compared to the sample covariance, but also *guarantees* well-conditioning.

As a motivating example take a look at Fig. 3.2, where the total MSE of the estimation of a covariance matrix $\mathbf{\Sigma}$ with a number of variables $p = 100$ using a varying sample size $n = 10$ to $n = 200$ is displayed. To generate the covariance matrix, a sparse correlation structure was simulated (only 5 percent of the partial correlations – the correlations between two variables conditioned on all others – are non-zero) and the variances were simulated from a scale-inverse chi-squared distribution Scale-inv-$\chi^2(d_0, s_0^2)$ with $s_0^2 = 4$ and degrees of freedom $d_0 = 100$. To estimate the covariance matrix, n samples of the vector X were drawn from the multinormal distribution using the simulated covariance matrix: $X \sim N(0, \mathbf{\Sigma})$.

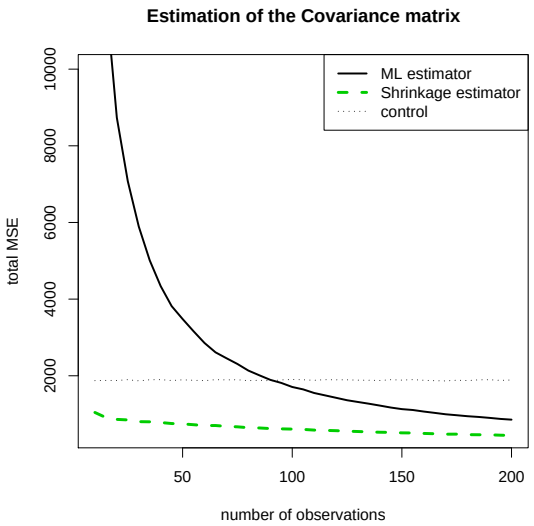

Figure 3.2: Example: Estimation of a covariance matrix.

The covariance matrix was estimated using the unbiased empirical estimator of the covariance matrix $\boldsymbol{S} = \frac{1}{n-1}(\boldsymbol{X} - \bar{\boldsymbol{X}})^T(\boldsymbol{X} - \bar{\boldsymbol{X}})$ (the ML-estimator) and the shrinkage estimator $\boldsymbol{S}^\star$, which is developed in this chapter. Additionally, a control estimator was

introduced consisting of a diagonal matrix (implying a variance of one for each variable and no correlation among the variables). An estimator having greater total MSE as this estimator can be seen as incapable of retrieving any information from the data. The simulation was repeated 100 times for each sample size, and the average total MSE for each estimator computed.

Fig. 3.2 clarifies the need for improved estimators in the "small n, large p" paradigm as the unbiased empirical estimator is highly unreliable for small sample sizes and cannot any longer extract any information from the data if the number of observations n is only slightly smaller than the number of variables p. The Stein-type shrinkage estimator on the other hand works well even if n is much smaller than p.

We will now proceed to infer the shrinkage estimator for covariance matrices. As the covariance matrix can be separated into two natural groups – the variances of the variables and the correlation between the variables – we will estimate the shrinkage intensities separately. The advantage is that this approach makes the estimation of the intensities for the correlation matrix and the partial correlation matrix independent of scale and local transformation of the data (Schäfer and Strimmer, 2005b).

3.2.1 Variances

The shrinkage rule in Eq. 3.4 allows to calculate a shrinkage estimator of the variance vector. From given data with p variables (e.g. genes) we first compute the usual unbiased empirical variances $v_1^2, \ldots, v_p^2$. These provide the components for the unregularized vector estimate $\hat{\boldsymbol{\theta}}$ of Eq. 3.3. For the shrinkage target we suggest using the median value of all v_k. In the exploration of possible other targets we considered also shrinking against zero and towards the mean of the empirical variances. However, these two alternatives turned out to be either less efficient (zero target) or less robust (mean target) than shrinking towards the median.

Following the recipe outlined above, we immediately obtain the shrinkage estimator

$$v_k^\star = \hat{\lambda}^\star v_{\text{median}} + (1 - \hat{\lambda}^\star) v_k. \tag{3.10}$$

The optimal shrinkage intensity is given by Eq. 3.7,

$$\lambda^* = \frac{\sum_{k=1}^p \text{Var}(v_k) - \text{Cov}(v_k, v_k^{\text{Target}}) + \text{Bias}(v_k) E(v_k - v_k^{\text{Target}})}{\sum_{k=1}^p E\left[(v_k - v_k^{\text{Target}})^2\right]}, \tag{3.11}$$

This formula can be simplified: note that we use the unbiased empirical variances, therefore: $\text{Bias}(v_k) = 0$. Furthermore we use the approximation $\text{Cov}(v_k, v_{\text{median}}) \approx 0$. The optimal estimated pooling parameter can now be written as:

$$\hat{\lambda}^\star = \frac{\sum_{k=1}^p \widehat{\text{Var}}(v_k)}{\sum_{k=1}^p (v_k - v_{\text{median}})^2}. \tag{3.12}$$

The computation of a sample version of $\widehat{\mathrm{Var}}(v_k)$ is straightforward. Defining $\bar{x}_k = \frac{1}{n}\sum_{i=1}^{n} x_{ik}$, $w_{ik} = (x_{ik} - \bar{x}_k)^2$, and $\bar{w}_k = \frac{1}{n}\sum_{i=1}^{n} w_{ik}$, we have $v_k = \frac{n}{n-1}\bar{w}_k$ and $\widehat{\mathrm{Var}}(v_k) = \frac{n}{(n-1)^3}\sum_{i=1}^{n}(w_{ik} - \bar{w}_k)^2$.

Eq. 3.12 has an intuitive interpretation. If the empirical variances v_k can be reliably determined from the data, and consequently exhibit only a small variance themselves, there will be little shrinkage, whereas if $\widehat{\mathrm{Var}}(v_k)$ is comparatively large pooling across the parameters will take place. Furthermore, the denominator of Eq. 3.12 is an estimate of the misspecification between the target (here $v^{\mathrm{Target}} = v_{\mathrm{median}}$) and the v_k. Hence, if the target is incorrectly chosen then no shrinkage will take place either.

3.2.2 Correlation structure

The calculation of a shrinkage estimator of the correlation matrix $\boldsymbol{R}^{\star}$ is done equivalently:

$$\boldsymbol{R}^{\star} = \lambda \boldsymbol{R}^{\mathrm{Target}} + (1-\lambda)\boldsymbol{R}, \tag{3.13}$$

with the optimal shrinkage intensity

$$\lambda^{*} = \frac{\sum_{k\neq l}\mathrm{Var}(r_{kl}) - \mathrm{Cov}(r_{kl}, r_{kl}^{\mathrm{Target}}) + \mathrm{Bias}(r_{kl})E(r_{kl} - r_{kl}^{\mathrm{Target}})}{\sum_{k\neq l} E\left[(r_{kl} - r_{kl}^{\mathrm{Target}})^2\right]}, \tag{3.14}$$

where only the off-diagonal elements of the sample correlation matrix $\boldsymbol{R}$ are used to calculate λ. Using the unbiased empirical estimator of the correlation matrix

$$\boldsymbol{R} = \frac{1}{n-1}\left(\boldsymbol{X} - \bar{\boldsymbol{X}}\right)^{T}\left(\boldsymbol{X} - \bar{\boldsymbol{X}}\right),$$

again simplifies the calculation of λ, as the bias of r_{kl} is zero and therefore $\mathrm{Bias}(r_{kl})E(r_{kl} - r_{kl}^{\mathrm{Target}}) = 0$. A straightforward target $\boldsymbol{R}^{\mathrm{Target}}$ is the identity matrix $\boldsymbol{I}$ (Schäfer and Strimmer, 2005b), which leads to $\mathrm{Cov}(r_{kl}, r_{kl}^{\mathrm{Target}}) = 0$ and $E[(r_{kl} - r_{kl}^{\mathrm{Target}})^2] = E[(r_{kl})^2]$.

As the moments of r_{kl} are unknown, they have to be estimated from the data. The variance of the entries of the empirical covariance matrix $\widehat{\mathrm{Var}}(r_{kl})$ can be found by

$$\widehat{\mathrm{Var}}(r_{kl}) = \frac{n^2}{(n-1)^2}\ \ \widehat{\mathrm{Var}}(\bar{w}_{kl}) = \frac{n}{(n-1)^2}\ \ \widehat{\mathrm{Var}}(w_{kl}) = \frac{n}{(n-1)^3}\sum_{i=1}^{n}(w_{ikl} - \bar{w}_{kl})^2 \tag{3.15}$$

with $w_{ikl} = x_{ik}^{s}x_{il}^{s}$ and $\bar{w}_{kl} = \frac{1}{n}\sum_{i=1}^{n} w_{ikl}$, whereby x^s are the standardized data vectors $x^s = \frac{x_i - \bar{x}}{\sigma}$. A similar formula can be derived for the variance of the entries of the empirical covariance matrix (cf. Schäfer and Strimmer (2005b)). Using additionally $\hat{E}[(r_{kl})^2] = r_{kl}^2$, we can compute the sample approximation of the shrinkage intensity

$$\hat{\lambda}^{*} = \frac{\sum_{k\neq l}\widehat{\mathrm{Var}}(r_{kl})}{\sum_{k\neq l} r_{kl}^2}, \tag{3.16}$$

which in turns allows calculating the matrix of shrunken correlation coefficients.

3.2.3 Covariance matrix

The shrinkage estimations of the variance vector and the correlations can now be put together to infer a Stein-type estimation of the covariance matrix $\boldsymbol{S}^\star$. We summarize:

$$s^\star_{kl} = r^\star_{kl}\sqrt{v^\star_k v^\star_l} \tag{3.17}$$

with

$$r^\star_{kl} = (1-\hat{\lambda}^\star_1) r_{kl} \tag{3.18}$$

$$v^\star_k = \hat{\lambda}^\star_2 v_{median} + (1-\hat{\lambda}^\star_2) v_k \tag{3.19}$$

and

$$\hat{\lambda}^\star_1 = \min\left(1, \frac{\sum_{k\neq l}\widehat{\mathrm{Var}}(r_{kl})}{\sum_{k\neq l} r^2_{kl}}\right) \tag{3.20}$$

$$\hat{\lambda}^\star_2 = \min\left(1, \frac{\sum_{k=1}^{p}\widehat{\mathrm{Var}}(v_k)}{\sum_{k=1}^{p}(v_k - v_{\mathrm{median}})^2}\right). \tag{3.21}$$

Note that we truncated the estimated $\hat{\lambda}$ at one, which results in the so-called positive part James-Stein estimator $\delta^{\hat{\lambda}+} = \delta^0 - \min(1,\hat{\lambda})\boldsymbol{\Delta}$. It dominates the unrestricted shrinkage estimator of Eq. 3.4 in terms of statistical efficiency (Barachnik, 1970).

3.3 Component Risk Protection by Limited Translation

The efficiency of the above shrinkage estimator can be further improved by restricting the translation allowed for individual components. The original James-Stein procedure is geared, in the terminology of Efron and Morris (1975), towards producing estimators with good *ensemble* risk properties. This means that it aims at minimizing the total risk accumulated over all parameters. However, in some instances this may occur at the expense of individual parameters whose risks may even increase (!). Therefore, in Stein estimation (and indeed also in hierarchical Bayes estimation) individual components of a parameter vector need to be protected against too much shrinkage.

"Limited translation" (Efron and Morris, 1972, 1975) is a simple way to construct estimators that exhibit both, good ensemble risk as well as favorable component risk properties. One example of a protected shrinkage rule is

$$\delta^{\hat{\lambda}+,M}_k = \delta^0_k - \min(1,\hat{\lambda})\min(1,\frac{M}{|\Delta_k|})\Delta_k, \tag{3.22}$$

which ensures that we always have $|\delta^{\hat{\lambda}+,M}_k - \delta^0_k| \leq M$, where M is a cutoff parameter chosen by the user. A convenient selection of M is, e.g., the 99 percent quantile of the

distribution of the absolute values $|\Delta_k|$ of the components of the shrinkage vector $\mathbf{\Delta}$. In the terminology of Efron and Morris (1972), the term $\min(1, \frac{M}{|\Delta_k|})$ constitutes the *relevance function* that determines the degree to which any particular component is affected by the ensemble-wide shrinkage.

Finally, we point out an interesting connection with soft thresholding, as the above limited translation shrinkage rule may also be written as

$$\delta_k^{\hat{\lambda}+,M} = \delta_k^{\hat{\lambda}+} + \min(1, \hat{\lambda})(|\Delta_k| - M)_+ \mathrm{sgn}(\Delta_k), \tag{3.23}$$

where the subscript "+" denotes truncation at zero.

We will now proceed to apply the shrinkage estimation for learning in complex systems without a temporal structure.

Chapter 4

Statistical Learning in High-dimensional Data Sets

In this chapter we concentrate on learning from data without temporal information. The first application of Stein-type inference are case-control analyses for high-dimensional problems. For this, we introduce a *"shrinkage t"* statistic that is based on the shrinkage estimator of the variance vector of the parameters. This method is published in *Article A*. Simulations and applications to real data show that our algorithm is competitive with the best currently available methods.

Afterwards we concentrate on models that analyse the connections among all variables in a complex system. The interest of research lies in determining which variables influence each other. A common tool for this are *correlation networks*. Nevertheless, we will show that using this model has considerable drawbacks and hence introduce graphical Gaussian models to infer influence networks from data. As graphical Gaussian models are the basis of most of the methods developed in this thesis, descriptions can be found in *Article B*; *Article C*; *Article E*; *Article F* and *Article G*.

4.1 High dimensional case-control analyses

High-dimensional case-control analysis is a key problem that has many applications, most prominently in computational genomics. The most well-known example is that of ranking genes according to differential expression, but there are many other instances that warrant similar statistical methodology, such as the problem of detecting peaks in mass spectrometric data or finding genomic enrichment sites.

All these problems have in common that they require a variant of a (regularized) t statistic that is suitable for high-dimensional data and large-scale multiple testing. For this purpose, in the last few years various test statistics have been suggested, which may be classified as follows:

i) Simple methods: fold change, classical t statistic.

ii) Ad hoc modifications of ordinary t statistic: Efron's 90% rule (Efron et al., 2001), SAM (Tusher et al., 2001).

iii) (Penalized) likelihood methods, e.g.: Ideker et al. (2000), Wright and Simon (2003), Wu (2005).

iv) Hierarchical Bayes methods, e.g.: Newton et al. (2001), Baldi and Long (2001), Lönnstedt and Speed (2002), Newton et al. (2004), "moderated t" (Smyth, 2004), Cui et al. (2005), Fox and Dimmic (2006).

For an introductory review of most of these approaches see, e.g., Cui and Churchill (2003) and Smyth (2004).

Current good practice in gene expression case-control analysis favors the empirical or full Bayesian approaches (item iv) over other competing methods. The reason behind this is that Bayesian methods naturally allow for information sharing across genes, which is essential when the number of samples is as small as in typical genomic experiments. Specifically, the estimation of gene-specific variances profits substantially from pooling information across genes (e.g., Wright and Simon, 2003; Smyth, 2004; Delmar et al., 2005; Cui et al., 2005). On the other hand, Bayesian methods can become computationally quite expensive, and more importantly, typically rely on a host of very detailed assumptions concerning the underlying data and parameter generating models.

We will now use the shrinkage estimation of the variance vector to introduce a "shrinkage t" approach that is as simple as the ad hoc rules (item ii) but performs as well as fully Bayesian models (item iv), even in simulation settings that are favorable to the latter. Other algorithms also based on shrinkage have been suggested (Cui et al., 2005; Tong and Wang, 2007), but rely on complicated models, whereas our shrinkage t statistic exhibits both, conceptual simplicity and high performance.

4.1.1 The "Shrinkage t" Statistic

To construct the shrinkage t statistic, we use the shrinkage estimator of Eq. 3.10,

$$v_k^\star = \hat{\lambda}^\star v_{\text{median}} + (1 - \hat{\lambda}^\star) v_k, \tag{4.1}$$

to estimate the gene-specific variances efficiently. Subsequently, the shrinkage variance estimate is plugged into the ordinary t statistic. With the sample sizes in groups 1 and 2 denoted as n_1 and n_2 the shrinkage t statistic is given by

$$t_k^\star = \frac{\bar{x}_{k1} - \bar{x}_{k2}}{\sqrt{\frac{v_{k1}^\star}{n_1} + \frac{v_{k2}^\star}{n_2}}}. \tag{4.2}$$

We consider two variants of this statistic, one where variances are estimated separately in each group (hence with two different shrinkage intensities), and the other one using a pooled estimate (i.e. with one common shrinkage factor).

Note that the shrinkage t statistic essentially provides a compromise between the standard t statistic (to which it reduces for $\lambda = 0$) and the fold change or difference of means statistic ($\lambda = 1$).

Closely related to the shrinkage t statistic are in particular two approaches, "moderated t" (Smyth, 2004) and the Stein-type procedure by Cui et al. (2005), which will here be called Cui et al. t. The essential feature characteristic for both of these methods is, again, variance shrinkage, albeit done in a different fashion compared to shrinkage t. Nevertheless, both the moderated t and the Cui et al. t rely on some form of distributional assumption, whereas the shrinkage t statistic is derived without any such consideration.

4.1.2 Assessment of Quality of Gene Ranking

To assess the quality of gene ranking provided by the shrinkage t statistic in comparison to other competing scores, we performed computer simulations and analyzed experimental gene expression data.

For the *simulations* we considered in total 2,000 genes, 100 of which were randomly assigned to be differentially expressed. The variances across genes where chosen to be (A) highly similar, (B) balanced, and (C) different.

Additionally we computed the gene ranking for three *experimental case-control data sets* with known differentially expressed genes, the well-known Affymetrix spike-in study (Cope et al., 2004), a subset of the "golden spike" Affymetrix experiment of Choe et al. (2005), and the HIV-1 infection study of van 't Wout et al. (2003). Detailed information on the simulations can be found in *Article A* and on the experimental data sets in section 2.4.2.

These data formed the basis for computing various gene ranking scores. Specifically, we compared the following statistics: fold change, ordinary t, moderated t (Smyth, 2004), Cui et al. t statistic (i.e. the unequal variance t statistic regularized by using the variances estimated by the method of Cui et al. (2005)), Efron's 90% rule (Efron et al., 2001), Wu's improved SAM statistic (Wu, 2005), and the shrinkage t statistic (with both equal and unequal variances). As reference we also included random ordering in the analysis. For these different ways of producing rankings we computed false positives (FP), true positives (TP), false negatives (FN), and true negatives (TN) for all possible cut-offs in the gene list (1-2000).

This procedure was repeated 500 times for each test statistic and variance scenario, to obtain estimates of the true discovery rates $E(\frac{TP}{TP+FP})$ and ROC curves describing the dependence of sensitivity $E(\frac{TP}{TP+FN})$ and specificity $E(\frac{TN}{TN+FP})$.

4.1.3 Performance of Gene Ranking Statistics

The results from simulations and data analysis are summarized in Fig. 4.1 and Fig. 4.2. In each figure the first row shows the fraction of correctly identified differentially expressed genes in relation to the number of included genes (i.e. the true discovery rate, or posi-

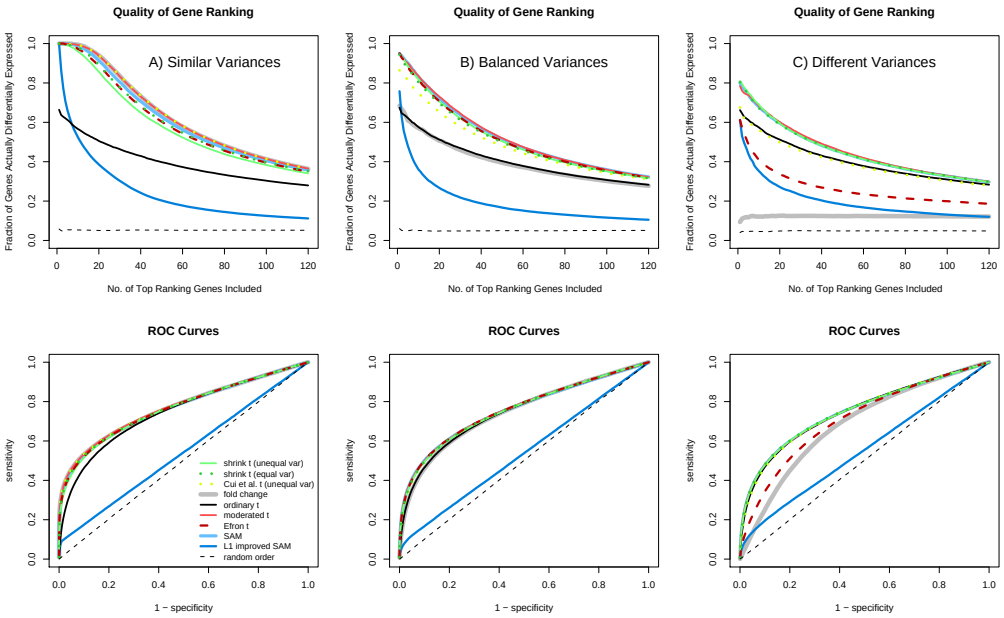

Figure 4.1: True discovery rates and ROC curves computed for simulated data under three different scenarios for the distribution of variances across genes. See main text for details of simulations and analysis procedures.

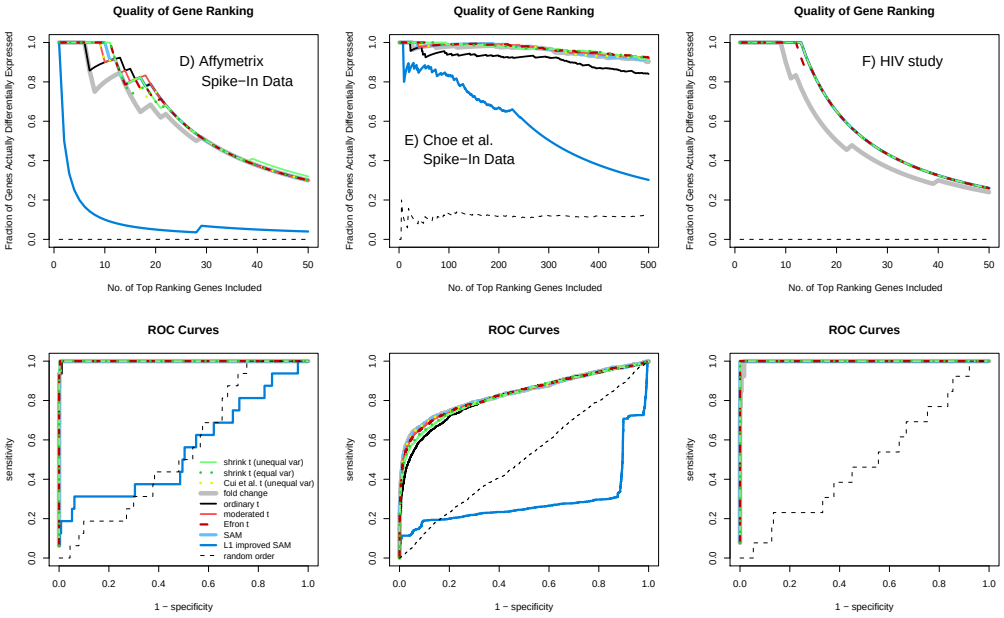

Figure 4.2: True discovery rates and ROC curves for the three investigated experimental data sets.

tive predictive value), whereas the second row depicts the corresponding receiver-operator characteristic (ROC).

We see that the shrinkage t and the moderated t statistics are the only two methods that perform optimally in all three simulation settings. Moreover, both produce accurate rankings also for the three experimental data, perhaps with a small edge for shrinkage t. If we compare the two methods to apply the shrinkage t statistic, we see that the shrinkage t statistic with unequal variances performs nearly as well as shrinkage t with equal variances, even though the former has only half the sample size available for estimation of variances.

We have to note that the moderated t statistic is based exactly on the model employed in the simulations and the Cui et al. t statistic also incorporates prior knowledge on the distribution of variances across genes. Therefore, it is easy to understand why moderated t and the Cui et al. t statistic perform well. On the other hand, recall that the shrinkage estimation of the variance vector is not specifically tailored to any particular distributional setting.

We see that the proposed method based on the shrinkage estimation of the covariance matrix provides highly accurate rankings of differential expression both for simulated and real data, on par with much more complicated models, but without relying on computational expensive procedures such as MCMC or optimization.

We now proceed to the analysis of the relations among all parameters of a high-dimensional data set.

4.2 Correlation networks

Using networks to visualize the connection between parameters is a common and popular method in many areas of research. The nodes represent the variables, e.g., genes in systems biology, and the links or edges a connection between them. The specific type of connection depends on the underlying model. An arrow from a node A to a node B constitutes directed networks and imply that node A influences node B. Directed networks hence introduce a causal meaning.

A simple type of a network are correlation graphs (also called relevance networks). The idea of is this method to look at the sample correlation matrix of the variables. If the correlation between two variables is greater than a specified threshold, an edge is drawn between them. There are different methods to choose the threshold. A simple example is order the strength of the correlation and to choose a certain number of the highest correlations. A statistical method is based on the p-values, e.g. by choosing a 95% level of significance (sometimes with a correction for multiple testing). A more elaborated method is based on the false discovery rate (see section 4.3.2).

Correlation graphs are very popular, especially for gene expression analysis and metabolic data. An example is given in Fig. 4.3 for the gene, protein and lipid expression of a transgenic mouse (Clish et al., 2004), where a relevance network is generated from the correlation matrix of the variables and a treshold for correlations that are considered in the network of 0.8.

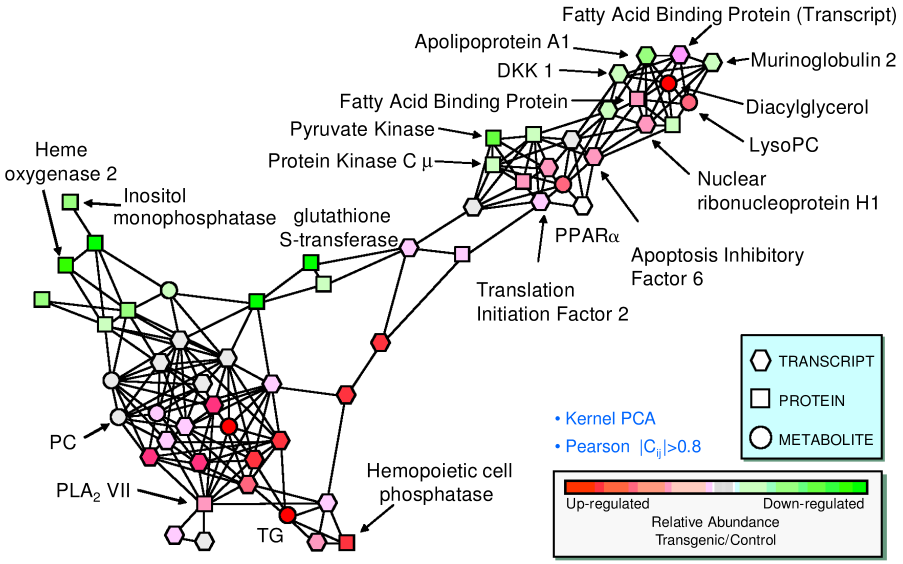

Figure 4.3: Correlation network of select expressed genes, proteins and lipids. The shading inside the box indicates the relative amount in the transgenic animals compared to wild type controls (red5 higher level; green5 lo wer level) and a line connecting two entities indicates a high level of correlation (a Pearson correlation coefficient of 0.8 was used as a cut-off). This figure and its description can be found in Clish et al. (2004).

Although correlation graphs are widespread and very easy to construct, the underlying model of using the correlation between two variables have a serious drawback. Usually the interest lies in a direct connection between two variables. Nevertheless, two variables can also be correlated due to indirect interaction or due to the regulation by a third variable (see Fig. 4.4). Hence correlation tells rather little about direct dependence between variables. In the best case, it indicates independence (two variables with a zero correlation coefficient can still be dependent, e.g., in a nonlinear way). Using relevance networks may therefore lead to wrong conclusions about the dependence of the variables.

Alternatives are graphical models, which are able to determine direct dependence between two variables. The methods that are developed in the following chapters are based on these graphical models.

4.3 Graphical Gaussian networks

A graphical model is a representation of stochastic conditional dependencies between the investigated variables. That means that only direct dependencies between variables are taken into account. Recall the middle and three right part of Fig. 4.4: given C (resp. D), the correlation between the variables A and B is conditioned away.

This property makes graphical models very attractive for the examination of complex systems. Especially in systems biology, where the identification of networked genetic inter-

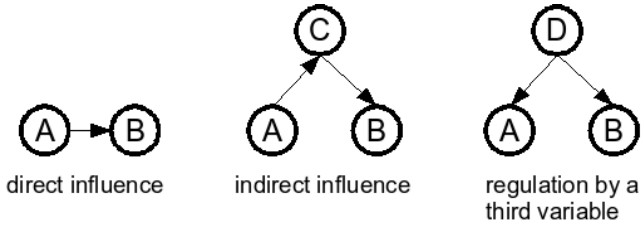

Figure 4.4: *Sources of correlation:* A correlation between two Variables A and B can originate from a direct influence of one variable onto the other, here: A influences B (left figure). The main interest usually lies in identifying these effects. Nevertheless, correlation can also exist due to an indirect influence via other variables, here via node C (middle figure), or due to a common regulation by a third variable, here node D (right figure).

dependencies that form the basis of cellular regulation is one of the key issues, many authors have followed this statistical approach to estimate genetic networks from high-throughput data (e.g., Hartemink et al., 2002; Friedman, 2004; Schäfer and Strimmer, 2005a).

Among the simplest graphical models is the class of graphical Gaussian models (GGMs) – see, e.g., Whittaker (1990). These are based on the partial correlation matrix: the strength of these coefficients indicates the presence or absence of a direct association between each pair of genes. In the following section we concentrate on the partial correlation and also on the partial variance, which is based on the same idea of conditioning on other variables.

4.3.1 The partial counterparts of correlation and variance

We demonstrate the inference of the partial correlation/variance starting from the true covariance matrix. In the usual notation, the covariance matrix and its inverse can be displayed as

$$\begin{aligned} \mathrm{Cov}(X_k, X_l) &= \sigma_{kl}, \\ \boldsymbol{\Sigma} &= (\sigma_{kl}), \\ \boldsymbol{\Omega} &= (\omega_{kl}) = \boldsymbol{\Sigma}^{-1}. \end{aligned} \tag{4.3}$$

The covariance matrix implies a correlation matrix and its inverse:

$$\mathrm{cor}(X_k, X_l) = \rho_{kl} = \sigma_{kl}(\sigma_{kk}\sigma_{ll})^{-1/2}, \tag{4.4}$$

$$\begin{aligned} \boldsymbol{P} &= (\rho_{kl}), \\ \boldsymbol{Q} &= (q_{kl}) = \boldsymbol{P}^{-1}. \end{aligned} \tag{4.5}$$

The partial correlations can be inferred by

$$\mathrm{cor}(X_k, X_l|\mathrm{rest}) = \tilde{\rho}_{kl} = -\frac{\omega_{kl}}{(\omega_{kk}\omega_{ll})^{-1/2}}, \tag{4.6}$$

$$\tilde{\boldsymbol{P}} = (\tilde{\rho}_{kl}),$$

and the partial variance by

$$\sqrt{\text{Var}(X_k|\text{rest})} = \tilde{\sigma}_{kk} = \omega_{kk}^{-1} = \text{diag}(\boldsymbol{\Omega})^{-1}, \tag{4.7}$$

whereby the tilde denotes the "partial". The partial correlation matrix can also be displayed computationally efficient in terms of $\boldsymbol{\Omega}$ or $\boldsymbol{Q}$ in a single expression:

$$\begin{aligned} \tilde{\boldsymbol{P}} &= -\boldsymbol{\Omega} \odot \sqrt{\text{diag}(\boldsymbol{\Omega})^{-1}(\text{diag}(\boldsymbol{\Omega})^{-1})^T}, \\ \tilde{\boldsymbol{P}} &= -\boldsymbol{Q} \odot \sqrt{\text{diag}(\boldsymbol{Q})^{-1}(\text{diag}(\boldsymbol{Q})^{-1})^T}. \end{aligned}$$

The symbol $\odot$ marks the elementwise product, also called Hadamard product. Note that we have to change the sign of the diagonal.

Apart from the definition via the inverse of the covariance matrix, partial correlations and partial variance can also be seen from a perspective of linear regression. For the correlation this can be done in two different ways:

- If ϵ_1 are the residuals from a linear regression of X_1 against $X_3, \ldots, X_p$, and equivalently ϵ_2 are the residuals from regressing X_2 against $X_3, \ldots, X_p$, then the correlation between the residuals are the partial correlations between X_1 and X_2,

$$\tilde{\rho}_{12} = \text{Cor}(\epsilon_1, \epsilon_2).$$

- If $\beta_2^{(1)}$ is the regression coefficient of X_2 in a first regression $X_1 = \beta_2^{(1)} X_2 + \text{rest}$, and $\beta_1^{(2)}$ the regression coefficient X_1 of a second regression $X_2 = \beta_1^{(2)} X_1 + \text{rest}$, then the partial correlation between X_1 and X_2 can be displayed using $\beta_2^{(1)}$ and $\beta_1^{(2)}$:

$$\tilde{\rho}_{12} = \sqrt{\beta_1^{(2)} \beta_2^{(1)}} \text{sign}(\beta_1^{(2)}). \tag{4.8}$$

  The absolute value of partial correlation is the square root of the product of the regression coefficients. The partial correlation is positive if the regression coefficients are positive, and negative otherwise.

The partial variance can also be defined via linear regression: if ϵ_1 are the residuals from a linear regression of X_1 against $X_2, \ldots, X_p$, then the variance of ϵ_1 is the partial variance of X_1:

$$\tilde{\sigma}_{11}^2 = \text{Var}(\epsilon_1)$$

The partial variance of a variable X_k can therefore be seen as the variance of X_k that cannot be explained by the other variables.

So far the partial counterparts of the covariance structure were derived from the true covariance structure $\boldsymbol{\Sigma}$. For practical application an estimated covariance matrix $\boldsymbol{S}$ has to be employed. As already demonstrated in section 3.2, it is crucial to apply shrinkage or other regularization techniques (Dobra et al., 2004; Schäfer and Strimmer, 2005b) to yield an accurate estimation for the "small n, large p" paradigm encountered in complex

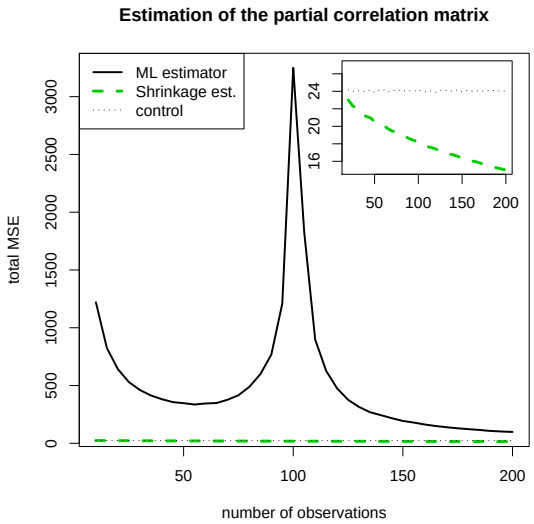

Figure 4.5: Example: Estimation of a partial correlation matrix.

systems (West et al., 2000). This is even more important if the partial counterparts of the variance and correlation have to be estimated.

As a demonstration for the relevance of regularization when calculating partial correlations and partial variances, we will again use the simulation of section 3.2. In Fig. 4.5 the MSE of the estimation of the partial correlation is displayed. Note that we have to use the generalization of the inverse, the Moore-Penrose pseudoinverse, if the empirical covariance matrix of the empirical estimator is ill-conditioned (namely if $n < p$). The simulation reveals two insights for estimating the partial correlations based on the sample covariance estimator. Firstly, the ML-estimator is not able to extract any information from the data even if the number of observations is considerable larger than the number of variables. Secondly, there exists a peak for $n \approx p$ where the error increases dramatically. This is a result of a dimension resonance effect due to the use of the pseudoinverse (see Schäfer and Strimmer (2005a) for a discussion of this phenomenon and references). The estimation of the partial correlation based on the Stein-type estimator shrinkage on the other hand is able to avoid the resonance effect and is able to give a reasonable estimation of the partial correlation matrix in the "small n, large p" paradigm.

A real world example for the need of regularization is the application of GGMs in systems biology: using GGMs for microarray data was first proposed in 2000. Nevertheless, this method seemed to work only for either very small number of genes (Waddell and Kishino, 2000; Kishino and Waddell, 2000) or for small numbers of clusters of genes (Toh and Horimoto, 2002a,b). The shrinkage estimator for the covariance/correlation matrix proposed in section 3.1.3 is able to achieve a stable and accurate estimation even if the number of variables p is much larger than the number of observations n.

An example for the inference of the correlations and partial correlations can be seen in Fig. 4.6 for the *Arabidopsis thaliana* data. The correlation matrix is estimated using

the shrinkage estimator of section 3.1.3 and the partial correlations are inferred using the method described above. The density of the distribution of the correlations is shown on left side of Fig. 4.6 and the partial correlations on the right side. We see that most of the correlations vanish and therefore resulted from indirect influences between two genes.

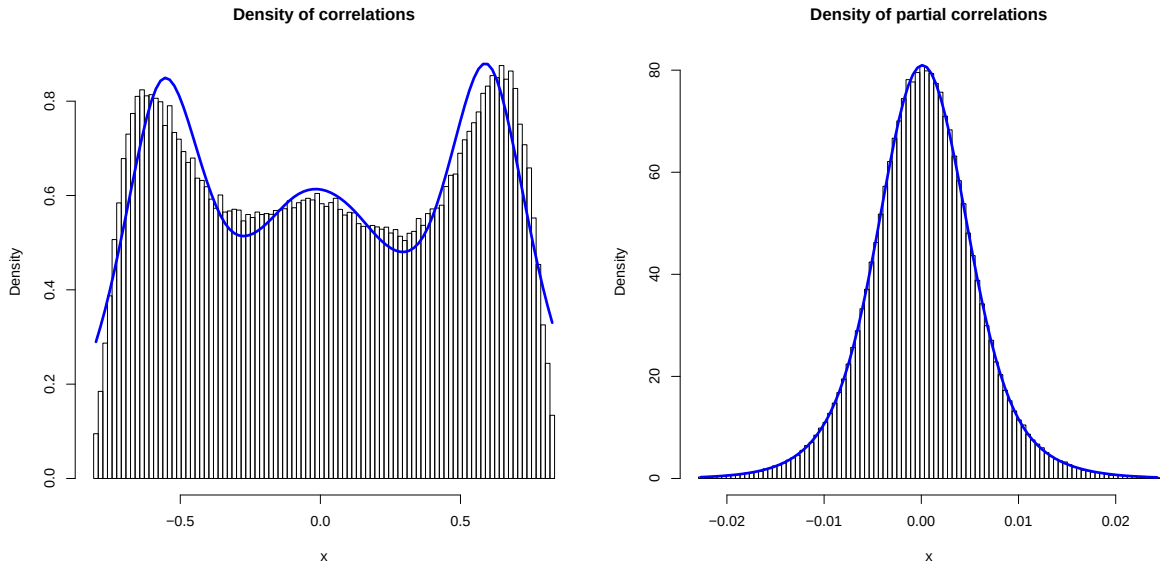

Figure 4.6: Correlations and partial correlations for the *Arabidopsis thaliana* data

The shrinkage estimator of the partial correlation coefficients is a good estimator for the strength of the connection between two variables. Nevertheless, we now face the problem which edges to include in the network, that means which of the partial correlation coefficients can be seen as significant. This will be done using *local fdr* (Efron, 2004b).

4.3.2 Model selection using local fdr

Identifying the correct network topology of the graphical Gaussian model can be seen from two perspectives:

- *Model selection:* The whole network is taken into consideration and the network with the best goodness of fit is chosen.
- *Hypothesis testing:* Each partial correlation coefficient is examined separately and only the significant coefficients are included as edges into the network.

A straightforward method for the model selection perspective is to evaluate the goodness of fit for all potentially adequate graphical models. Nevertheless, an exhaustive search is impossible even for medium-sized networks, as the number of potential networks increases dramatically with the number of variables (the number of potential networks is 2^p). This

problem can be reduced by using numerical optimization methods (see, e.g., Nocedal and Wright, 2000), but these are computationally very demanding and break down for large-scale systems.

The hypothesis testing perspective tries to identify significant genes separately. In the standard textbook method the hypothesis is formulated (the partial correlation between two variables is zero) which is subsequently tried to reject on a certain significance level (e.g. 95%). Nevertheless, the small sample size and the necessity to regard for the simultaneous calculation of possibly thousands of hypothesis tests renders standard methods unreliable.

For the identification of large scale networks we use a third option, an exact correlation test, which is based on hypothesis testing, but nevertheless takes all other possible edges of the network into account and is employed here as a model selection procedure. It is called *local* fdr (Efron, 2004b) and is an empirical Bayes version of the false discovery rate (FDR) by (Benjamini and Hochberg, 1995). The methodology of FDR focuses on the tail areas, whereas fdr is based on the densities.

Generally, the global false discovery rate (FDR) focuses on a hypothesis different to the usual tests. Instead of controlling the probability of a type 1 error of a single hypothesis, the FDR is defined as the expected ratio of erroneous rejections to the total number of rejected hypothesis, $E(Q)$. Technically,

$$Q = \begin{cases} \frac{V}{R} & \text{if } R > 0 \\ 0 & \text{if } R = 0, \end{cases} \tag{4.9}$$

where $R = R(\gamma)$ are the number of hypothesis with p-values in a specified rejection region $[0, \gamma]$. The FDR therefore provides – loosely speaking – "interesting" cases: with a threshold of, e.g., 0.2, 80% of the identified edges are known to result from non-zero partial correlations between the variables.

We will now concentrate on the local fdr algorithm Efron (2004a, 2005b,c) to identify significant edges. It is based on the idea of separating the distribution of the estimations of the partial correlations resulting from variables without any connections (the null distribution) from the distribution of the estimations from variables that have a non-zero partial correlation. As we will apply the local fdr algorithm in chapter 6 to a different test problem, a test for causality in the network, we will describe local fdr for arbitrary test problems.

Assume a large number of test statistics $z_1, \ldots, z_k$ where we have to decide whether or not z_i is generated from the null hypothesis H_0. Specifically, we have to separate the null distribution $f_0(z)$ (in which applies: $z_i \sim H_0$) from the alternative distribution $f_A(z)$, where $z_i \sim H_A$. To achieve this, it is possible to use the empirical Bayes procedure for model selection by Efron (2003) and Robbins (1956), as elaborated in Efron (2005c). First, we have to fit the empirical null distribution across all edges assuming a mixture

$$f(z) = \eta_0 f_0(z) + \eta_A f_A(z), \tag{4.10}$$

with $\eta_0 + \eta_A = 1$. Note that we assume that most of the z_i belong to the null distribution and therefore $\eta_0 >> \eta_A$, to ensure that the null distribution can be identified correctly.

It is now possible to compute the empirical posterior probability that z_i is significantly different from H_0:

$$\text{Prob}(z_i \sim H_1|\tilde{\rho}) = 1 - \widehat{\text{fdr}}(z) = 1 - \frac{\eta_0 f_0(z)}{f(z)} \tag{4.11}$$

A good cutoff value is 0.8, corresponding to a local fdr smaller than 0.2, as suggested and discussed in Efron (2005c).

The relation between local fdr and FDR is discussed in detail in Efron et al. (2001). The principle is that FDR focuses on the tail-area of the cumulative distribution functions of the densities f_0 and f: the FDR(z_0) is the conditional expectation of $\text{fdr}(z) \equiv \eta_0 f_0(z)/f(z)$ given $z \leq z_0$,

$$\text{FDR}(z_0) = \frac{\int_{-\infty}^{z_0} \text{fdr}(z) f(z) dz}{\int_{-\infty}^{z_0} f(z) dz}. \tag{4.12}$$

The null distribution, the density under null hypothesis, depends on the test problem. For identifying significant edges of the GGM, we have to focus on the distribution of the partial correlation coefficients with the null hypothesis $\tilde{\rho} = 0$. In chapter 6 we will use local fdr for assessing causal directions in a network, with the null hypothesis that no causal direction can be identified.

But we will now concentrate on testing for significant edges in the network. The null distribution of the normal (partial) correlation coefficients r (Hotelling, 1953) is known and depends only on a single parameter κ:

$$f_0(r;\kappa) = (1-r^2)^{(\kappa-3)/2} \frac{\Gamma(\frac{\kappa}{2})}{\pi^{\frac{1}{2}}\Gamma(\frac{\kappa-1}{2})} \tag{4.13}$$

For correlations, if $\rho = 0$, the degree of freedom κ is equal to the inverse of the variance, i.e. $\text{Var}(\rho) = \frac{1}{\kappa}$, and to sample size minus one ($\kappa = n - 1$). For partial correlations, κ can be calculated similarly: $\kappa = n - 1 - (p - 2) = n - p + 1$ (the relation to the variance still holds: $\text{Var}(\tilde{\rho}) = \frac{1}{\kappa}$). Nevertheless, for complex systems, where $n < p$, that implies that κ is negative! The explanation is that the *effective* degree of freedom can differ from the theoretical degree of freedom: the formula $\kappa = n-1-(p-2) = n-p+1$ demands that the test statistics are mutually independent, which is often not satisfied even for large sample sizes. To infer partial correlations for $n < p$, we have to use the shrinkage estimator for the correlation matrix which always introduces dependence between the variables. Therefore, the theoretical degree of freedom cannot be used in the "small n, large p" paradigm. However, it is still possible to fit an empirical null hypothesis, see Efron (2004a, 2005c).

An example of the Efron-Robbins procedure can be seen in Fig. 4.7 for the *Arabidopsis thaliana* data set. The continuous line smoothing the histogram shows the distribution of the estimated shrinkage partial correlation coefficients (note that in contrast to the right part of Fig. 4.6 Fisher's normalizing z-transformation was applied for normalization purposes). The fitted null distribution is depicted by the dashed line, and the alternative distribution by the dark area. The triangles indicate the 0.2 local fdr cut-off values for the partial correlations.

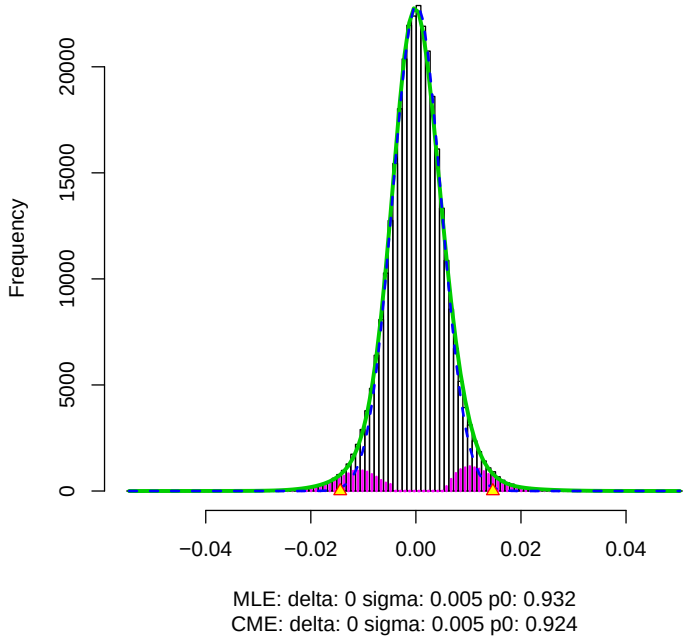

Figure 4.7: The local fdr algorithm (Efron, 2004a, 2005c) for the *Arabidopsis thaliana* data set.

4.3.3 Inferring a graphical Gaussian network

We now summarize the procedure to infer a graphical Gaussian network using shrinkage estimation of the correlation matrix and local fdr model selection. The algorithm is used to infer the genetic network of the *Arabidopsis thaliana* data set.

1. The correlation matrix of the data, here the correlations of the expression among the different genes of the Arabidopsis data set, is estimated. For complex systems a shrinkage estimator of the correlation matrix (section 3.1.3) has to be employed.

2. The estimated correlations allow calculating the partial correlation matrix (section 4.3.1).

3. The local fdr algorithm provides significant partial correlations (section 4.3.2)

4. Significant partial correlations are included as edges in the network, which connect the variables (the nodes), here the different genes of the *Arabidopsis thaliana* data set.

The inferred genetic network of the *Arabidopsis thaliana* data set using a graphical Gaussian model can be seen in Fig. 4.8. In total, 7163 of the 319600 possible edges are

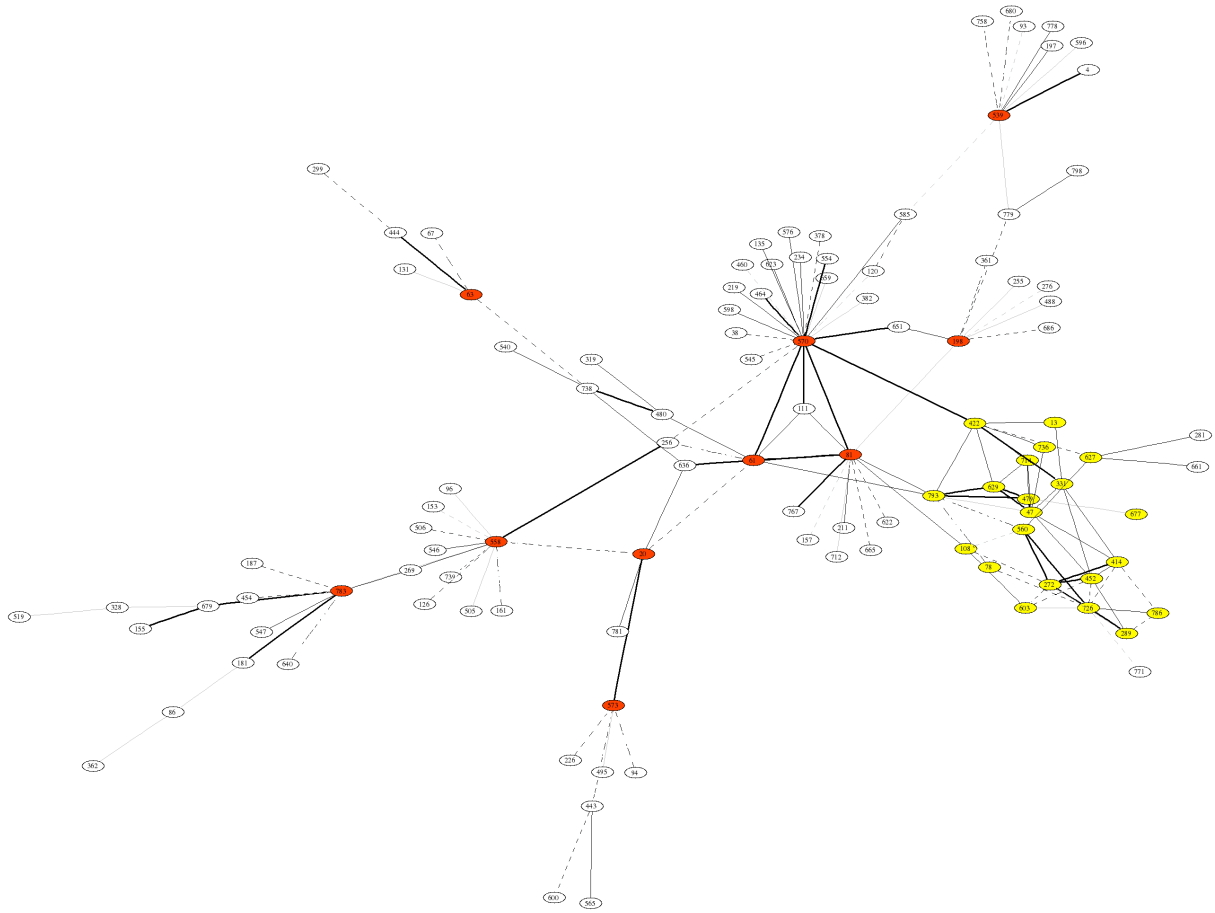

Figure 4.8: *Arabidopsis thaliana* GGM network

significant, connecting 693 of the 800 preselected genes[1]. For purposes of clarity only the 150 strongest connections are depicted in Fig. 4.8. The structure of the inferred network corresponds to the kind expected for biological networks (see Ravasz et al. (2002); Barabási and Oltvai (2004) for some discussion about network biology). We can clearly identify the hub structure of the network where a few "hub" nodes exist, which are connected to a large number of nodes. The corresponding genes will probably play a large role in the regulation of the cell. This hub connectivity structure is particularly striking for the gene 570 (an AP2 transcription factor) and 81 (a gene involved in DNA-directed RNA polymerase), but also for the genes 198, 558 or 783 and a few others. These hubs are marked in red. Moreover, it is possible to identify a hypothesized functional module in the upper right corner (encircled and marked in light-yellow), where a large number of genes are strongly connected, but have only a few connections with nodes outside the module.

[1]Note: The number of significant edges is different to the results of *Article E*. This is due to a change of the implementation of the local fdr algorithm: in *Article E*, the R package "locfdr" (Efron et al., 2006) was used, here, the R package "fdrtool" (Strimmer, 2007).

In this chapter we estimated a graphical Gaussian model to visualize connections between variables. The methods developed here are used as a basis to employ more elaborated models developed in the following chapter, where we expand the principle of GGMs to time series data and infer techniques that regard the dynamic structure of these kind of data.

Chapter 5

Analysis of longitudinal data sets in complex systems

In the last chapter we assumed the data of complex systems to have no temporal structure. The reason for this is that the GGM approach (as well as other graphical models such as Bayesian networks) relies on the assumption of independent and identically distributed (i.i.d.) data. Nevertheless, we often observe the behavior of complex system through time, resulting in time course data. Especially in systems biology, an increasing proportion of microarray expression experiments are concerned with *longitudinal* measurements of mRNA and protein concentrations. For instance, stress response and cell cycle experiments like the diurnal cycle of the *Arabidopsis thaliana* data set by design produce time course data. Moreover, a further characteristic of these data is that the time points at which the experiments are conducted are almost always not equidistant, but irregularly spaced.

In this chapter we try to incorporate time series aspects in the analysis of complex networks and introduce two possibilities to account for longitudinal data. The first is a functional data approach which is employed to expand the concept of correlation to *dynamical* correlation. It allows describing time course data consistently in a correlation matrix (*Article B*; *Article C*; *Article G*). The second approach is to use a vector autoregressive model, which characterizes the connection between different time points and therefore inevitably introduces a causal meaning (*Article D*; *Article E*).

5.1 Dynamical correlation

To account for time series data of complex systems, we now investigate GGM network inference from the perspective of functional data analysis (Ramsay and Silverman, 2005). Specifically, we describe a graphical model that treats the observed gene expression over time as realizations of random curves, rather than to describe the individual time points separately. This approach is based on the notion of *dynamical correlation*, a concept introduced in this chapter. It provides a similarity score for pairs of groups of randomly sampled curves. Once the dynamical correlation matrix is obtained, partial dynamical

correlations can be computed and the identification of the associated network structure be inferred according to the inference of GGMs from the correlation structure as described in section 4.3.

The concept of dynamical correlation was first introduced by Dubin and Müller (2005). Although based on the same idea, we will define it differently, so that our approach is conceptually an extension of the usual correlation. This means that in the limiting case (the number of time points is reduced to one) our definition – unlike the one by Dubin and Müller (2005) – reduces to the sample correlation.

We will first summarize the basic notation for functional data analysis (FDA) and also introduce the functional inner product. Next, we discuss the concept of dynamical correlation of which we describe two different variants, the one introduced here and the one by Dubin and Müller (2005). The problem of the "small n, large p" paradigm renders a regularized inference of the dynamical correlation necessary, which will be derived subsequently. Afterwards, the concept of dynamical correlation is used to infer a network of the *Arabidopsis thaliana* data, which explicitly takes the longitudinal data structure of the experiment into consideration.

5.1.1 The Concept of Dynamical Correlation

Setup and Notation

To introduce the notation required for the inference of dynamical correlation we assume that a number of p variables and n replications are measured over a time interval $[A, B]$. This can be for example data from a typical gene expression time course experiment where for p genes and n subjects mRNA concentrations are measured. This results in functional observations $f_{ik}(t)$ where $1 \leq i \leq n$ and $1 \leq k, l \leq p$. We assume all functions $f_{ik}(t)$ to be square-integrable so that the functional inner product

$$\langle g(t), h(t) \rangle = \frac{1}{B-A} \int_A^B g(t)h(t)dt \tag{5.1}$$

exists, where $g(t)$ and $h(t)$ are any of the observed functions. The time average of $f_{ik}(t)$ may then be conveniently expressed by $\langle f_{ik}(t), 1 \rangle$. The average over the n replicates gives the empirical mean function $\bar{f}_k(t) = \frac{1}{n} \sum_{i=1}^n f_{ik}(t)$.

In practice, however, the functions $f_{ik}(t)$ are not continuously measured but rather obtained by experiments at discrete time points t_j, with $1 \leq j \leq m$ and $A = t_1 < t_2 < \ldots < t_{m-1} < t_m = B$. Note that the time points need not be equidistant. If one assumes a linear approximation of $g(t)$ and $h(t)$ the inner product of Eq. 5.1 turns into the weighted sum

$$\langle g(t), h(t) \rangle \approx \sum_{j=1}^{m} g(t_j)h(t_j) \frac{\delta_j + \delta_{j+1}}{2(B-A)} \tag{5.2}$$

where the $\delta_j = t_j - t_{j-1}$ are the time differences between subsequent measurements (with $\delta_1 = \delta_{m+1} = 0$).

In the random effects representation of Dubin and Müller (2005) each observed $f_{ik}(t)$ is a realization of the random function

$$f_k(t) = \mu_k(t) + \mu_{0k} + \epsilon_{0k} + \sum_{u=1}^{\infty} \epsilon_{uk}\eta_u(t), \tag{5.3}$$

where ϵ_{0k} and ϵ_{uk} are random variables with $E(\epsilon_{0k}) = 0$ and $E(\epsilon_{uk}) = 0$, $\mu_k(t)$ is the fixed time dependent mean function with zero time average $\langle \mu_k(t), 1\rangle = 0$, $\mu_{0k} + \epsilon_{0k}$ represents the static random part and the remaining terms describe the dynamic random part. In Eq. 5.3 the $\eta_u(t)$ are orthonormal basis functions with zero time average $\langle \eta_u(t), 1\rangle = 0$.

In this notation the empirical mean function $\bar{f}_k(t)$ is an estimate of $E(f_k(t)) = \mu_k(t) + \mu_{0k}$. As $\mu_k(t)$ has time average zero we are also able to identify the two components of $E(f_k(t))$ by using $\hat{\mu}_{0k} = \langle \bar{f}_k(t), 1\rangle$ and $\hat{\mu}_k(t) = \bar{f}_k(t) - \hat{\mu}_{0k}$.

Measuring similarity between two exactly known curves

Suppose for a moment that we have sufficient data to estimate the expression levels through time of two genes k and l *exactly*, i.e. that we know the mean functions $E(f_k(t))$ and $E(f_l(t))$. In order to understand the functional connection between these two variables a measure of similarity between the two curves is required. Dubin and Müller (2005) suggest to introduce the notion of *dynamical correlation* with the informal proposition that "if both trajectories tend to be mostly on the same side of their time average (a constant) then the dynamical correlation is positive; if the opposite occurs, then dynamical correlation is negative".

This immediately leads to the following straightforward definition of dynamical correlation between two curves $g(t)$ and $h(t)$. First, calculate the time-centered functions $g^C(t) = g(t) - \langle g(t), 1\rangle$ and $h^C(t) = h(t) - \langle h(t), 1\rangle$. Then define the variances as

$$\mathrm{Var}(g(t)) = \langle g^C(t), g^C(t)\rangle \tag{5.4}$$

and

$$\mathrm{Var}(h(t)) = \langle h^C(t), h^C(t)\rangle. \tag{5.5}$$

Finally, compute the standardized functions

$$g^S(t) = g^C(t)/\sqrt{\mathrm{Var}(g(t))} \tag{5.6}$$

and

$$h^S(t) = h^C(t)/\sqrt{\mathrm{Var}(h(t))}, \tag{5.7}$$

and obtain the correlation by

$$\mathrm{Cor}(g(t), h(t)) = \langle g^S(t), h^S(t)\rangle. \tag{5.8}$$

The general case including sampling error

The above definition of dynamical correlation for a single curve extends in a straightforward fashion to the case where each observed time course f_{ik} represents a noisy realization of the mean function $E(f_k)$.

In order to estimate the correlation between two variables k and l we first define the simultaneously time- *and* space-centered functions according to $f_{ik}^C(t) = f_{ik}(t) - \langle \bar{f}_k(t), 1 \rangle$. Note that here the inner product is computed over the mean function $\bar{f}_k(t)$. Based on the $f_{ik}^C(t)$ the empirical estimate of the variance of variable k is then given by

$$\widehat{\mathrm{Var}}_k = \hat{\sigma}_{kk} = s_{kk} = \frac{1}{n} \sum_{i=1}^{n} \langle f_{ik}^C(t), f_{ik}^C(t) \rangle. \tag{5.9}$$

This allows to compute standardized residual functions $f_{ik}^S(t) = f_{ik}^C / \sqrt{s_{kk}}$ that form the basis for the estimate of dynamical correlation

$$\widehat{\mathrm{Cor}}_{kl} = \hat{\rho}_{kl} = r_{kl} = \frac{1}{n} \sum_{i=1}^{n} \langle f_{ik}^S(t), f_{il}^S(t) \rangle. \tag{5.10}$$

Correspondingly, the estimated dynamical covariance between variables k and l is simply

$$\widehat{\mathrm{Cov}}_{kl} = \hat{\sigma}_{kl} = s_{kl} = r_{kl} \sqrt{s_{kk} s_{ll}}. \tag{5.11}$$

This simple estimator of dynamical correlation exhibits several attractive properties. In particular, it is a generalization of the standard correlation for cross-sectional data. Specifically, if $m = 1$ and $n > 1$ then it reduces to the usual maximum-likelihood estimator of correlation. Furthermore, it is also applicable if there is only a single realization of each time series available ($n = 1$, $m > 1$).

The Dubin-Müller definition of dynamical correlation

The definition of dynamical correlation by Dubin and Müller (2005) is related, but nevertheless different. They propose to compute the standardized residual functions according to

$$f_{ik}^S(t) = q_{ik}(t) / \sqrt{\langle q_{ik}(t), q_{ik}(t) \rangle} \tag{5.12}$$

using

$$q_{ik}(t) = f_{ik}(t) - \bar{f}_{ik}(t) - \langle f_{ik}(t), 1 \rangle + \langle \bar{f}_{ik}(t), 1 \rangle. \tag{5.13}$$

This definition has the drawback that it is only defined if both $m > 1$ and $n > 1$. As we will exemplify below, it also produces counter-intuitive correlations.

5.1.2 Regularized Inference of the Dynamical Correlation

The above definition allows the inference of correlations between sets of curves. However, in the "small n, large p" paradigm, we need a regularized estimator of the dynamical correlation.

To construct a Stein-type shrinkage estimate $\boldsymbol{S}^\star$ of the dynamic covariance matrix we can follow the recipe provided in chapter 3.1 and combine the unregularized estimator $\boldsymbol{S} = (\hat{\rho}_{kl})$ of the above chapter 5.1.1 and a suitable target $\boldsymbol{S}^{\text{Target}}$. This results in a shrinkage estimate $\boldsymbol{S}^\star = \lambda \boldsymbol{S}^{\text{Target}} + (1-\lambda)\boldsymbol{S}$.

In section 3.2 we calculated separate shrinkage parameters λ_i for the correlation structure and for the variances. Scaling reasons (Schäfer and Strimmer, 2005b) suggest that this is also advisable for the dynamical correlation. We will therefore apply shrinkage to the sample dynamical correlation matrix directly and infer the shrinkage estimate $\boldsymbol{R}^\star = \lambda \boldsymbol{R}^{\text{Target}} + (1-\lambda)\boldsymbol{R}$. A shrinkage estimation of the variances is not needed, as only the dynamical correlations are necessary to infer the genetic network.

The selection of the shrinkage parameter λ for the dynamical correlation matrix has to take place in a data-driven fashion. We will again estimate the optimal shrinkage intensity by minimizing the MSE risk function

$$\text{MSE}(\lambda) = E\left(\sum_{k=1}^{p}\sum_{l=1}^{p}(r_{kl}^\star - r_{kl})^2\right). \tag{5.14}$$

Recall Eq. 3.14: the minimum mean squared error $R(\lambda^\star)$ is achieved *exactly* and uniquely for the choice

$$\lambda^\star = \frac{\sum_{k\neq l}\text{Var}(r_{kl}) - \text{Cov}(r_{kl}, r_{kl}^{\text{Target}}) + \text{Bias}(r_{kl})E(r_{kl} - r_{kl}^{\text{Target}})}{\sum_{k\neq l} E\left[(r_{kl} - r_{kl}^{\text{Target}})^2\right]}, \tag{5.15}$$

whereby we choose $\boldsymbol{R}^{\text{Target}}$ to be the identity matrix I. Defining

$$\overline{f_{kl}} = \sum_{i=1}^{n}\sum_{j=1}^{m} \underbrace{f_{ik}^S(t_j) f_{il}^S(t_j)}_{f_{ijkl}} \underbrace{\frac{\delta_j + \delta_{j+1}}{2(B-A)n}}_{w_{ij}} \tag{5.16}$$

and the sum of squared weights

$$\tau = \sum_{i=1}^{n}\sum_{j=1}^{m} w_{ij}^2 = \frac{1}{n}\sum_{j=1}^{m}\left(\frac{\delta_j + \delta_{j+1}}{2(B-A)}\right)^2, \tag{5.17}$$

the *unbiased* empirical correlation equals

$$\widehat{\text{Cor}}(g(t), h(t)) = r_{kl} = \frac{1}{1-\tau}\overline{f_{kl}} \tag{5.18}$$

and find after some calculation the individual entries for

$$\widehat{\mathrm{Var}}(r_{kl}) = \frac{\tau}{(1-\tau)^3} \sum_{i=1}^{n} \sum_{j=1}^{m} w_{ij} \left(f_{ijkl} - \overline{f_{kl}} \right)^2 . \tag{5.19}$$

This leads to the sample approximation of the shrinkage intensity

$$\hat{\lambda}^{\star} = \frac{\sum_{k \neq l} \widehat{\mathrm{Var}}(r_{kl})}{\sum_{k \neq l} r_{kl}^2}, \tag{5.20}$$

which in turns allows calculating the matrix of shrunken dynamical correlation coefficients. This regularized method of inferring the dynamical correlation matrix extends the application of dynamical correlation to complex systems.

5.1.3 Applications of dynamical correlation

In the following we first apply our method of computing dynamical correlation to example data to clarify our definition and to compare it with the related concept of Dubin and Müller (2005). Subsequently, we infer the gene association network for the *Arabidopsis thaliana* data set using dynamical correlation.

Illustrative Example

In order to understand the concept of dynamical correlation and to illustrate the difference between our definition (Eq. 5.10) and that of Dubin and Müller (2005) we first consider a set of artificial examples. These are shown in Fig. 5.1 where two negatively dependent variables are depicted. For instance, this may, in systems biology, represent the case where one gene is up-regulated and the other is correspondingly down-regulated. For each gene there are two measured curves, and there are three slightly different ways in which the sampled curves relate to each other (Fig. 5.1a, b, and c). The exact definition of the curves can be found in Tab. 5.1. Note that the two realizations are paired, i.e. the upper lines belong to individual 1 and the lower ones to individual 2.

Intuitively, one would expect that the dynamical correlation between the two variables is strongly negative in all three cases. For our definition of dynamical correlation according to Eq. 5.10 this is indeed the case: the correlations for the three examples cases Fig. 5.1a, b, and c are -0.946, -0,973, and -0.947, respectively. In contrast, the dynamical correlation of Dubin and Müller (2005) behaves in a completely different fashion. For Fig. 5.1a it is not defined, for case b) it is equal to +1 and for case c) it is equal to -1.

Therefore, it is easy to see that the Dubin and Müller (2005) estimator is *not* suited for extending the concept of correlation between variables to be applicable to time series data. Nevertheless, this is often needed in complex systems, e.g. for the detection of functional dependencies in genomic longitudinal data. This is because that estimator is geared towards detecting changes in the relative trends of the individual realizations, rather than between the common trends. However, note that this is generally not the

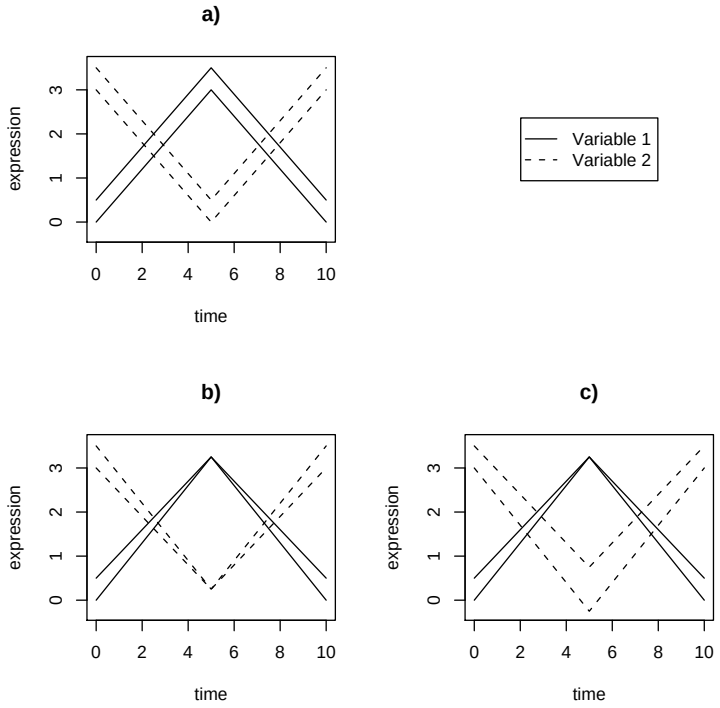

Figure 5.1: Toy example to illustrate the concept of dynamical correlation between two variables ("genes"). In all three cases a), b) and c) there are two realizations ("individuals"). See main text for details, and Tab. 5.1 for the underlying data.

effect one wants to identify, especially in systems biology when gene interaction is looked for. In addition, the Dubin and Müller (2005) definition of dynamical correlation has the additional disadvantage over that of Eq. 5.10 that it is not defined if there is only a single time course per gene available. In contrast, the above toy examples show that our definition of dynamical correlation is able to detect the main trend of positive or negative dependency between two variables, and is not susceptible to the small changes in the sampled curves.

Data		**Variable 1**			**Variable 2**		
Time points		*0*	*5*	*10*	*0*	*5*	*10*
Fig. 5.1a	*Realization 1*	0	3	0	3	0	3
	Realization 2	0.5	3.5	0.5	3.5	0.5	3.5
Fig. 5.1b	*Realization 1*	0	3.25	0	3	0.25	3
	Realization 2	0.5	3.25	0.5	3.5	0.25	3.5
Fig. 5.1c	*Realization 1*	0	3.25	0	3	-0.25	3
	Realization 2	0.5	3.25	0.5	3.5	0.75	3.5

Table 5.1: Data points of the toy examples in Fig. 5.1.

Estimating Gene Association Networks Using Dynamical Correlation

The procedure of learning gene association networks of longitudinal data using the concept of dynamical correlations corresponds to the algorithm of inferring GGM described in section 4.3.3. The only difference is to utilize the (regularized) estimator of dynamical correlation matrix for time course data sets instead of the (regularized) correlation matrix, which can only be used for data obtained from a single time point.

To demonstrate the inference of a gene association network we use the *Arabidopsis thaliana* dataset. Note that the time series is unequally spaced, which is typical for microarray time course data. Nevertheless, the present functional data approach allows the incorporation of arbitrary time distances between subsequent measurements.

As approximation of the temporal expression of the 800 genes we used a linear spline, which allows applying the method described in section 5.1.2 to infer the dynamical correlation matrix. After computing the associated partial correlation coefficients employing Eq. 4.5 and Eq. 4.6 the significant edges are inferred using local fdr (see section 4.3.2). The network inferred with dynamical correlation consists of 669 of the 800 preselected genes connecting 6102 edges. The network itself is displayed in Fig. 5.2 (for clarity only the 150 most significant edges are displayed).

The network inferred with dynamical correlation exhibits many similarities to the network inferred with the static correlation (Fig. 4.8): both share the same structure where many "hub" genes play a central role in both networks (e.g. gene 81, 570 or 783) and both possess the hypothesized functional module. This indicates that the usage of graphical Gaussian networks is a reasonable approach even for longitudinal data. Nevertheless, it can be argued, that our method of estimating the network using dynamical correlations is able to identify time-varying components of the interaction between the investigated genes.

5.1.4 Remarks

In this chapter we introduced a method to infer a network of the dependencies among variables from functional data. In this approach time course experiments are seen as a realization of random curves. The method described generalizes the widely used static GGM approach (see the corresponding references in Schäfer and Strimmer, 2005a) and is able to unravel the dependency structure of longitudinal data across the whole time series rather than at single time points. Note that in FDA unequal time points are accounted for by the weights employed in the functional inner product. Furthermore, unlike many other time series method the functional data approach does not require equally spaced measurements. In addition, our algorithm is easily implemented and computationally inexpensive (the calculation of the above gene dependency network takes only a fraction of a second). Shrinkage allows to improve the precision of the estimation and to extend the method to high dimensional data.

An important extension for estimating networks from time course data is the inclusion of autoregressive aspects (Diggle et al., 2002). While the method described in this chapter covers the dynamical correlation through time it is not able to account, e.g., for a time

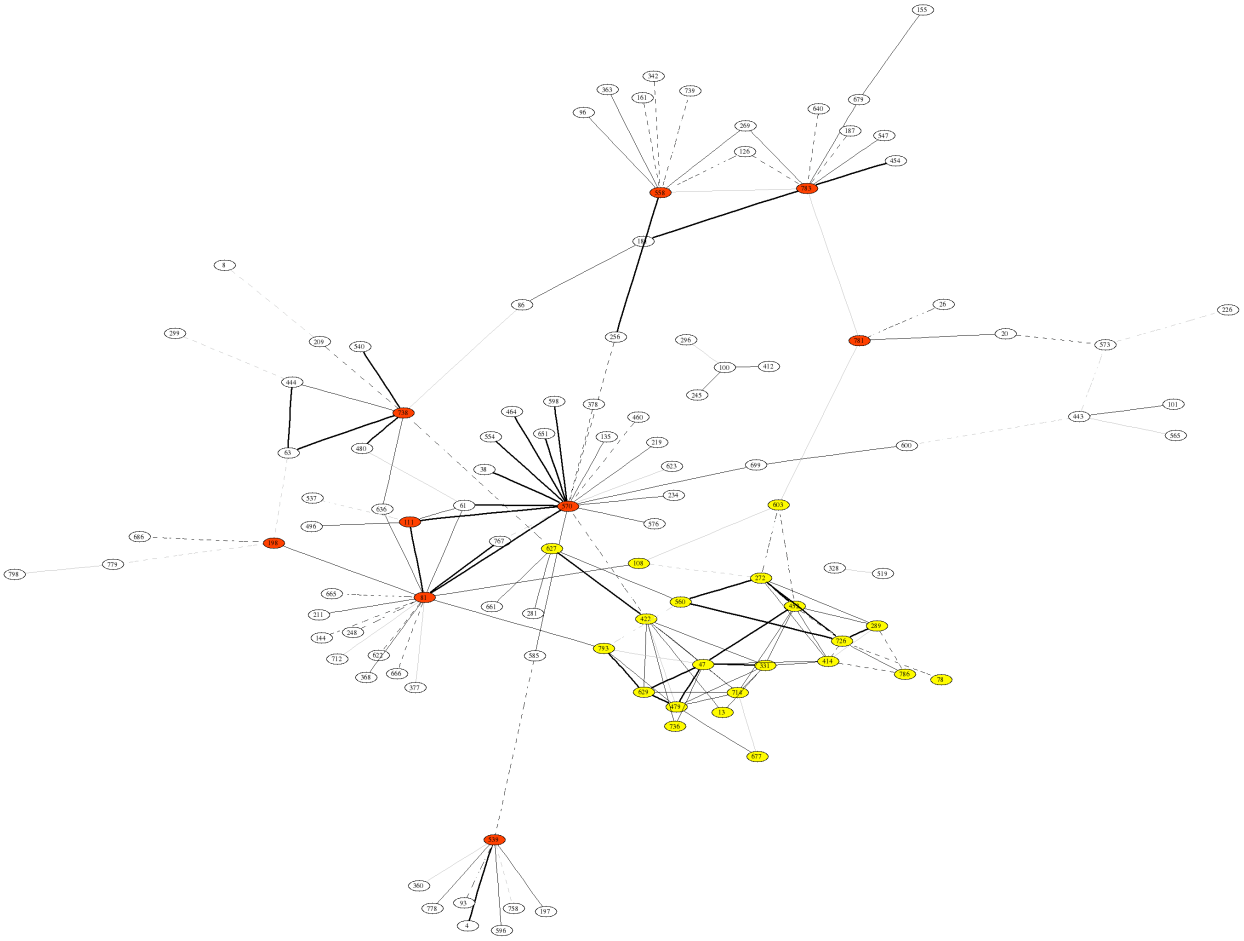

Figure 5.2: *Arabidopsis thaliana* gene association network using dynamical correlation

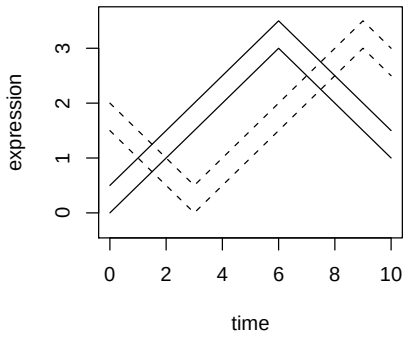

Figure 5.3: Example with a fixed time lag between the two variables.

shift between any two variables. This is illustrated in Fig. 5.3, which is a variation of the toy examples presented in section 5.1.3. For this data the Dubin and Müller (2005) estimate is (again) not defined and our suggested dynamical estimator results in very small correlation close to zero, even though it is clear by inspection that the two depicted variables are strongly connected. These dependencies and the associated time shifts could be accounted for by modeling the temporal mean via a system of differential equation or in the discrete case by some autoregressive process.

In the next chapter we will concentrate on the vector autoregressive model. For this, we first have to develop a regression method, called shrinkage regression, that is applicable to high-dimensional data.

5.2 Using a vector autoregressive model for analyzing time course data

Using a vector autoregressive (VAR) model allows describing the influences of the variables of a certain time point onto future time points. Linear regression is an essential tool in estimating these models; therefore it will be analyzed at the beginning of this section. The insights gained allow developing a Stein-type regularization for linear regression leading to the method of *shrinkage regression*, a regression method adapted to inference in the "small n, large p" paradigm. Subsequently the VAR-process is described and shrinkage regression applied to estimate the VAR regression coefficients. A VAR network model selection method provides an algorithm for inferring a directed network that describes the influences between variables through time.

5.2.1 Linear regression

A linear regression model assumes that a response variable Y can be described by a linear combination of a set of predictor variables X,

$$Y = a + Xb + \epsilon, \qquad \epsilon \sim N(0, \sigma_\epsilon^2), \tag{5.21}$$

with ϵ being the residuals and a an intercept. We will first deal with multiple regression, where one response variable depends on several predictor variables. As we will later extend the analysis to a regression with more than one response variable we write the response with a capital letter Y. The maximum likelihood (ML) estimator for the regression coefficients b can be calculated by

$$b = (X^T X)^{-1} X^T Y. \tag{5.22}$$

Note that X and Y have to be centered. The intercept can be then calculated by $a = \frac{1}{n}\sum y_i - \frac{1}{n}\sum \hat{y}_i$.

Nevertheless, X and Y represent only observations of the true structure of the variables. We now examine the regression model given the true structure of the data and show that all linear dependencies can be derived using only the covariance matrix of the variables.

Our starting point is the *true* covariance matrix $\mathbf{\Sigma}$ of some variables X. The actual estimation of this covariance matrix is a different problem to be dealt with later. We stress that this covariance matrix is the only object required for the calculation done here, nothing else is given or estimated.

For the regression model, we use a subscript if necessary: $\mathbf{\Sigma}_X$ is, e.g., the covariance matrix of the variables X, $\mathbf{\Sigma}_{YX}$ the covariance matrix of the combined variables Y and X: $[YX]$.

By applying partial variances and partial correlations, it is possible to infer the regression model *exactly* using the covariance/correlation structure of $[YX]$. The best linear predictor of the regression coefficients in terms of mean squared error is

$$b_k = \tilde{\rho}_{YX_k} \sqrt{\frac{\tilde{\sigma}_{YY}}{\tilde{\sigma}_{X_k X_k}}}, \tag{5.23}$$

$$b_k = \tilde{\rho}_{YX_k} \sqrt{\frac{q^{-1}_{YY}}{q^{-1}_{X_k X_k}}} \sqrt{\frac{\sigma_{YY}}{\sigma_{X_k X_k}}}. \tag{5.24}$$

The last representation of the predictor is especially interesting. The first term represents the partial correlation between two variables and displays the strength of the connection between them. In the middle term we see q^{-1}_{YY} and $q^{-1}_{X_k X_k}$. These are – for each variable – the partial variances divided by the variances of the respective variables:

$$\mathrm{SPV}_Y = q^{-1}_{YY} = \frac{\tilde{\sigma}_{YY}}{\sigma_{YY}},$$

$$\mathrm{SPV}_{X_k} = q^{-1}_{X_k X_k} = \frac{\tilde{\sigma}_{X_k X_k}}{\sigma_{X_k X_k}}.$$

We can see q^{-1} as the standardized partial variance (SPV) and interpret it as the fraction of the variance of each variable that cannot be explained by all others. The middle term therefore represents the ratio of this fraction for two variables and indicates which of them can be seen as dependent and which of them as independent variable. The last term is the only one that depends on the scaling of the variables and is necessary for scaling reasons (the regression coefficients are given in absolute terms). We will return to the decomposition of the regression coefficient and especially to the interpretation of the middle term in chapter 6, where it is used to infer causal dependencies between variables only by analyzing the covariance matrix.

We see that the true covariance structure allows inferring all linear dependencies in a model. Nevertheless, the true values for the covariances are – in empirical problems – unknown and have to be estimated. For linear regressions models where the number of observations by far exceeds the number of variables, the true covariance matrix $\mathbf{\Sigma}$ for the variables X can simply be replaced by its empirical counterpart $\boldsymbol{S}$

$$\boldsymbol{S} = \hat{\mathbf{\Sigma}} = \frac{1}{n-1}(X - \bar{X})^T(X - \bar{X}). \tag{5.25}$$

This applies analogue to the covariance matrix $\mathbf{\Sigma}_{YX}$ of the combined matrix of response and observation vectors. This is equivalent to the ordinary least squares (OLS) solution for linear regression.

5.2.2 Shrinkage regression

Using the analysis of the last section, it is straightforward to develop a regularized version of linear regression. As an estimator for the covariance matrix used to infer the regression coefficients, the Stein-type shrinkage estimator $\boldsymbol{S}^\star$ of the covariance matrix (see section 3.1.3) is employed.

Up to now, we assumed that there is a fixed response variable Y which is regressed on predictor variables X. Nevertheless, it is often not possible to determine a priori which variable is influenced by the others, that means that it is not possible to decide which is the response and which are the predictor variables. This implies that the strict separation between predictor and response variables is not necessarily justified.

To abstain from a strict distinction between response and predictors has a decisive influence if we take more than one response variable into account. In this multivariate case, two regression methods are possible:

- *Multivariate regression:* This is the usual method, where all response variables Y_i are separately regressed on the matrix of the predictor variables X.
- *Block regression (Cox and Wermuth, 1993):* The response variable Y_i is not only regressed on X but also on all remaining components $Y_{j\setminus i}$ of the response variables.

The difference between the two methods lies in the relation between two response variables Y_k and Y_l. In multivariate regression, the relation between Y_k and Y_l is measured essentially

by the correlation of Y_k and Y_l given all variables X. In block regression, it can be measured by the partial correlation of Y_k and Y_l given all remaining variables $Y_{i\setminus k,l}$ and X. Thus, block regression is more adequate in situations where a connection between the response variables cannot be ruled out. This is a situation typical for complex systems, where the causal structure is essentially unknown.

The method of shrinkage regression described here uses the approach of block regression. We will first elaborate block regression and subsequently infer shrinkage regression for the multivariate case. First, the response and predictor variables are combined into a single matrix $\Phi = [YX]$. It is then possible to infer all regression structures in Φ: all Φ_j are subsequently regressed on $\Phi_{i\setminus j}$. Assume that the true covariance matrix of Φ is given by $\boldsymbol{\Sigma}$, its inverse by $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ and the true partial correlation matrix by $\tilde{\boldsymbol{P}}$. A matrix of regression coefficients B for all regression structures can be exactly inferred by:

$$\begin{aligned} B &= \tilde{\boldsymbol{P}} \odot \sqrt{\operatorname{diag}(\boldsymbol{\Omega})^{-1}\operatorname{diag}(\boldsymbol{\Omega})^T} \\ B &= \tilde{\boldsymbol{P}} \odot \sqrt{\operatorname{diag}(\boldsymbol{Q})^{-1}\operatorname{diag}(\boldsymbol{Q})^T} \odot \sqrt{\operatorname{diag}(\boldsymbol{\Sigma})(\operatorname{diag}(\boldsymbol{\Sigma})^{-1})^T}, \end{aligned} \tag{5.26}$$

with $\odot$ marking the elementwise or Hadamard product. This means, e.g., that we find on row j the regression coefficients for the regression of all $\Phi_{i\setminus j}$ onto Φ_j plus 1 in the column j (the diagonal). We can also infer B using covariance/correlation directly, but we have to change the sign of the diagonal (as we have done by inferring the partial correlations in section 4.3.1):

$$\begin{aligned} B^d &= -\boldsymbol{\Omega} \odot (\operatorname{diag}(\boldsymbol{\Omega})^{-1}\mathbf{1}^T) \\ B^d &= -\boldsymbol{Q} \odot (\operatorname{diag}(\boldsymbol{Q})^{-1}\mathbf{1}^T) \odot \sqrt{\operatorname{diag}(\boldsymbol{\Sigma})(\operatorname{diag}(\boldsymbol{\Sigma})^{-1})^T} \\ B &= \begin{cases} B^d_{kl} & \forall \quad k \neq l \\ -B^d_{kl} & \forall \quad k = l \end{cases} \end{aligned} \tag{5.27}$$

The vectors $b_1, b_2, \ldots, b_m$ of the regression coefficients for the regression of the response variables $Y_1, Y_2, \ldots, Y_m$ can then simply be inferred by taking the first m rows of B (without the diagonal).

The development of a regularized version of block regression, the shrinkage regression for the multivariate case, is straightforward: Instead of using the sample covariance $\boldsymbol{S}$ to estimate $\boldsymbol{\Sigma}$, we apply the shrinkage estimator $\boldsymbol{S}^\star$ of section 3.1.3 for the covariance matrix of Φ.

The covariance matrix allows two other views on shrinkage regression. Note that the empirical estimator $\boldsymbol{S}$ of the combined response and predictor variables Φ contains two submatrices $\boldsymbol{S}_1 = \boldsymbol{X}^T\boldsymbol{X}$ and $\boldsymbol{S}_2 = \boldsymbol{X}^T\boldsymbol{Y}$. The OLS estimate of the regression coefficients can then be written as $\hat{\boldsymbol{B}}^{\text{OLS}} = (\boldsymbol{S}_1)^{-1}\boldsymbol{S}_2$. Replacing $\boldsymbol{S}$ by the shrinkage estimator $\boldsymbol{S}^\star$ leads to the first alternative expression of the shrinkage estimation of the regression coefficients $\hat{\boldsymbol{B}}^{\text{Shrink}} = (\boldsymbol{S}^\star_1)^{-1}\boldsymbol{S}^\star_2$.

The second interpretation can be found by decomposing $\boldsymbol{S}^\star$ using the SVD or Cholesky algorithm. It is then possible to reconstruct regularized "pseudodata" matrices $\boldsymbol{X}^\star$ and

$\boldsymbol{Y}^\star$. We will demonstrate this using the SVD:

$$S^\star = VDV^T = \Phi^{\star T}\Phi^\star \tag{5.28}$$

$$[X^\star Y^\star] = \Phi^\star = \sqrt{D}V^T \tag{5.29}$$

The shrinkage regression can be interpreted as OLS or normal-distribution ML based on these pseudodata:

$$\hat{B}^\star = (X^{\star T}X^\star)^{-1}X^{\star T}Y^\star \tag{5.30}$$

Before concentrating on the vector autoregressive model, it should be noted that the main area of application of shrinkage regression is in the "small n, large p" paradigm. If the number of observations is considerably larger than the number of variables, ordinary least square or Ridge regression is advantageously, as the latter try to optimize the estimation of the regression coefficients directly, whereas shrinkage regression goes an indirect way: first, a regularized covariance matrix is estimated, and subsequently the regression coefficients are inferred. The target of the optimization lies in the covariance matrix, not in the regression coefficients. Simulations showed that this leads to inferior results compared to ordinary least square or Ridge regression if $n > p$. Nevertheless, in the "small n, large p" paradigm, shrinkage regression provides a simple and superior method for estimating a linear model (see section 5.2.5).

We will now apply shrinkage regression to estimate the vector autoregressive model. The next section introduces the model. Afterwards, shrinkage estimation of the VAR coefficients is presented.

5.2.3 Shrinkage estimation of the vector autoregressive model

Definition of the VAR model

To introduce the vector autoregressive model, we consider vector-valued time series data $\boldsymbol{x}(t) = (x_1(t), \ldots, x_p(t))$. Each component of this row vector corresponds to a variable of interest, e.g., the expression level of a specific gene or the concentration of some metabolite in dependence of time. The vector autoregressive model specifies that the value of $\boldsymbol{x}(t)$ is a linear combination of those of earlier time points, plus noise,

$$\boldsymbol{x}(t) = \boldsymbol{c} + \sum_{i=1}^{m} \boldsymbol{x}(t - iL)\ \boldsymbol{B}_i + \boldsymbol{\epsilon}_i. \tag{5.31}$$

In this formula m is the order of the VAR process, L the time lag, and $\boldsymbol{c}$ a $1 \times p$ vector of means. The errors $\boldsymbol{\epsilon}_i$ are assumed to have zero mean and a $p \times p$ positive definite covariance matrix $\boldsymbol{\Sigma}$. The matrices $\boldsymbol{B}_i$ with dimension $p \times p$ represent the dynamical structure and thus contain the information relevant for reading off the causal relationships.

The autoregressive model has the form of a standard regression problem. Therefore, estimation of the matrices $\boldsymbol{B}_i$ is straightforward. For demonstration, we will set both m and L to 1. Then the above equation reduces to the VAR(1) process

$$\boldsymbol{x}(t+1) = \boldsymbol{c} + \boldsymbol{x}(t)\ \boldsymbol{B} + \boldsymbol{\epsilon}. \tag{5.32}$$

We now denote the centered *matrices of observations* corresponding to $\boldsymbol{x}(t+1)$ and $\boldsymbol{x}(t)$ by $\boldsymbol{X}_f$ ("future") and $\boldsymbol{X}_p$ ("past"), respectively, i.e. $\boldsymbol{X}_p = \begin{bmatrix} \boldsymbol{x}(1) \\ \vdots \\ \boldsymbol{x}(n-1) \end{bmatrix}$ and $\boldsymbol{X}_f = \begin{bmatrix} \boldsymbol{x}(2) \\ \vdots \\ \boldsymbol{x}(n) \end{bmatrix}$. In this notation the ordinary least squares (OLS) estimate can be written as

$$\hat{\boldsymbol{B}}^{\mathrm{OLS}} = ({\boldsymbol{X}_p}^T \boldsymbol{X}_p)^{-1} {\boldsymbol{X}_p}^T \boldsymbol{X}_f. \tag{5.33}$$

This is also the maximum likelihood (ML) estimate assuming the normal distribution. The coefficients of higher-order VAR models may be obtained in a corresponding fashion (Lütkepohl, 1993).

Shrinkage Estimation of VAR Coefficients

Small sample shrinkage estimates of VAR regression coefficients may be obtained by applying the shrinkage regression of section 5.2.2 appropriately to the VAR process. The specific algorithm is based on the second interpretation of shrinkage regression:

1. Combine the centered observations $\boldsymbol{X}_p$ and $\boldsymbol{X}_f$ into $\Phi = [\boldsymbol{X}_p \boldsymbol{X}_f]$.
2. The $(n-1)$ multiple of the *empirical* covariance matrix, $\boldsymbol{S} = \Phi^T \Phi$, contains $\boldsymbol{S}_1 = {\boldsymbol{X}_p}^T \boldsymbol{X}_p$ and $\boldsymbol{S}_2 = {\boldsymbol{X}_p}^T \boldsymbol{X}_f$.
3. Calculate the shrinkage estimate $\boldsymbol{S}^\star$ for the covariance matrix of Φ.
4. Determine the submatrices $\boldsymbol{S}_1^\star$ and $\boldsymbol{S}_2^\star$ and compute the estimates $\hat{\boldsymbol{B}}^{\mathrm{Shrink}} = (\boldsymbol{S}_1^\star)^{-1} \boldsymbol{S}_2^\star$.

5.2.4 VAR network model selection

The network representing potentially directed causal influences is given by the non-zero entries in the matrix of VAR coefficients. For an extensive discussion of the meaning and interpretation of the implied Granger (non)-causality we refer to (Granger, 1980).

As $\hat{\boldsymbol{B}}^{\mathrm{Shrink}}$ is an estimate it is unlikely that any of its components are exactly zero. Therefore, we need to statistically test whether the entries of $\hat{\boldsymbol{B}}^{\mathrm{Shrink}}$ are vanishing. However, instead of inspecting regression coefficients directly, it is preferably to test the corresponding partial correlation coefficients: as shown in section 5.2.1, the partial correlations represent the strength of the connection.

Specifically, consider in the VAR(1) model the multiple regression that connects the first variable $x_1(t+1)$ at time $t+1$ with all variables $x_1(t), \ldots, x_p(t)$ at the previous time t,

$$x_1(t+1) = c + \beta_k^1 x_k(t) + \sum_{j=1, j\neq k}^{p} \beta_j^1 x_j(t) + \text{error}. \tag{5.34}$$

If in this equation the roles of $x_k(t)$ and $x_1(t+1)$ are reversed,

$$x_k(t) = c + \beta_1^k x_1(t+1) + \sum_{j=1, j\neq k}^{p} \beta_1^j x_j(t) + \text{error}, \tag{5.35}$$

the partial correlation between the two variables is the geometric mean of the corresponding regression coefficients, times their sign, i.e. $\sqrt{\beta_1^k \beta_k^1}\text{sgn}(\beta_1^k)$, for details see section 4.3.1 and Whittaker (1990).

Once the partial correlations in the VAR model are computed, we use the local fdr approach of section 4.3.2 to identify significant partial correlations. The network can be inferred in the same way as in 4.3.3, nevertheless, unlike in a graphical Gaussian model, the edges in a VAR network are by design directed.

5.2.5 Applications

Simulation Study

In a comparative simulation study we investigated the power of diverse approaches to recovering the true VAR network. We simulated VAR(1) data of different sample size, with n varying between 5 and 200, for 100 randomly generated true networks with 200 edges and $p = 100$ nodes. The 200 nonzero regression coefficients were drawn uniformly from the intervals $[-1; -0.2]$ and $[0.2; 1]$.

In addition to the shrinkage procedure we estimated regression coefficients by ordinary least squares (OLS) and by ridge regression (RR). All these three regression strategies were applied in conjunction with the above VAR model selection based on partial correlations, with a cutoff value for the "local fdr" statistic set at 0.2 – the recommendation of Efron (2005c). As a fourth method we employed L1 regression (Tibshirani, 1996), also called the LASSO method, to estimate VAR regression coefficients. Note that in the latter instance there is no need for additional model selection, as the LASSO method combines shrinkage and model selection and automatically sets many regression coefficients identically to zero.

In the simulations we ran OLS only for $n > 100$, as for small sample size the corresponding empirical covariance matrix is singular and consequently the OLS regression is ill-posed. The penalty for the LASSO regression was chosen as in Meinshausen and Bühlmann (2006). The regularization parameter in RR was determined by generalized cross validation (Golub et al., 1979). Unfortunately, even GCV turned out to be computationally expensive, so that for RR we conducted only 10 repetitions, rather than the 100 considered for the other methods.

The results of the simulations are summarized in Figure 5.4. The left box shows the positive predictive value, or true discovery rate of the four methods. This is the proportion of correctly identified edges in relation to all significant edges. Our proposed shrinkage algorithm is the only method achieving around 80% positive predictive value regardless of the sample size. Note that this is exactly the theoretically expected value, given the specified "local fdr" cutoff of 0.2. In contrast, the RR and LASSO methods perform

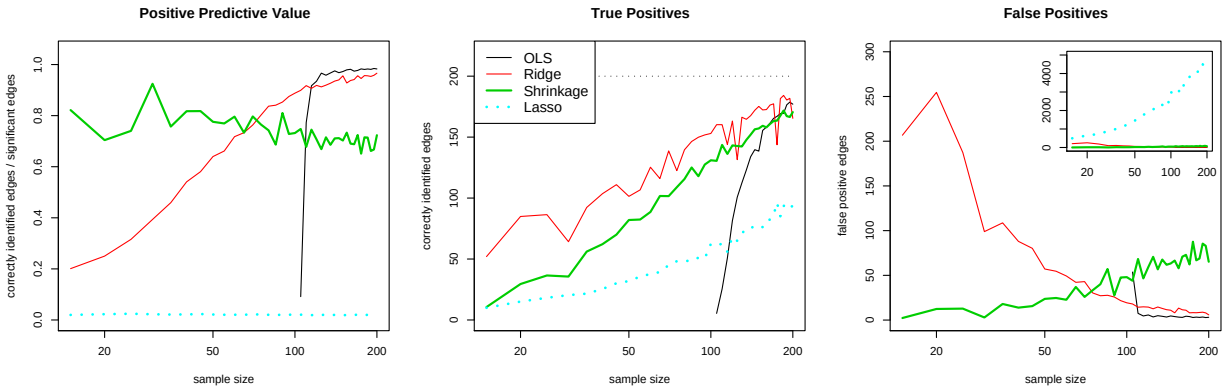

Figure 5.4: Relative performance of the four investigated methods for learning VAR networks in terms of positive predictive value (true discovery rate) and the number of true and false edges. The thin dotted line in the middle box at 200 corresponds to the true number of edges in the simulated networks.

remarkably poor at small sample size, with much lower true discovery rates. For medium to large sample size the OLS estimation dominates RR, LASSO and the shrinkage approach. This is easily explained by the fact that OLS has no parameters to optimize and that it is asymptotically optimal. However, it is bothering that for both the RR and the OLS approach the false discovery rate appears not to be properly controlled. Finally, for large sample size the Stein-type estimator appears to be prone to overshrinking, which leads to an increase of false positives.

The relative performance of the four approaches to VAR estimation can be further explained by considering the relative amount of true and false positive edges (Figure 5.4, middle and right box). The shrinkage method generally produces very few false positives. In contrast, the RR and LASSO methods lead to a large number of false edges, especially for small sample size. This is particularly pronounced for the LASSO regression, as can be seen in the differently scaled inlay plot contained in the right box of Figure 5.4, indicating that the penalty applied in the L1 regression may not be sufficient in this situation. In terms of the number of correctly identified edges the RR and shrinkage approach are the two top performing methods. However, even though RR finds a considerable number of true edges even at very small sample size, this has little impact on its true discovery rate because of the high number of false positives.

In summary, the simulation results suggest using for small sample size the James-Stein-type shrinkage procedure, and for $n > p$ the traditional OLS approach.

Analysis of a Microarray Time Course Data Set

For further illustration we apply the VAR shrinkage approach to the *Arabidopsis thaliana* data set. We note that an assumption of the VAR model is that time points are equidistant – see Eq. 5.31. Recall that this is not the case for the *Arabidopsis thaliana* data which

were measured at $0, 1, 2, 4, 8, 12, 13, 14, 16, 20,$ and 24 hours. However, as the intensity of the biological reactions is likely to be higher at the change points from light to dark periods (time points 0 and 12), one may argue that assuming equidistant measurements is justifiable at least in terms of equal relative reaction rate.

To infer the VAR network, we estimated the regularized regression coefficients and the corresponding partial correlations, and identified the significant edges of the VAR causal graph as described above. We found a total number of 14967 significant edges connecting 669 nodes[1]. In Fig. 5.5 we show for reasons of clarity only the subnetwork containing the 150 most significant edges, which connect 92 nodes. The "hub" connectivity structure (nodes filled with red) already encountered in the GGM and the dynamical correlation network can again be identified.

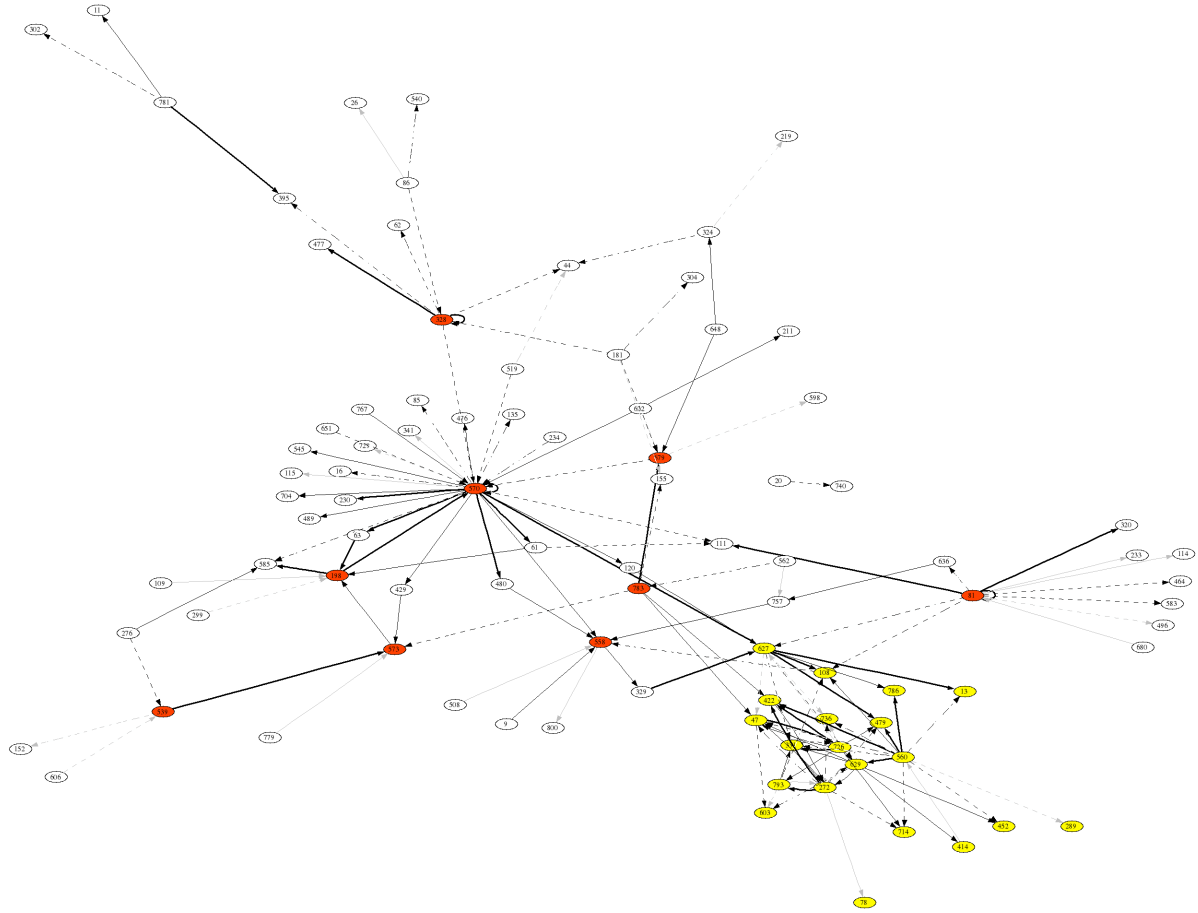

Figure 5.5: Directed VAR network inferred from the *Arabidopsis thaliana* data. The solid and dotted lines indicate positive and negative regression coefficients, respectively, and the line intensity denotes their strength.

[1]Note: The number of significant edges is different to the results of *Article E*. This is due to a change of the implementation of the local fdr algorithm: in *Article E*, the R package "locfdr" (Efron et al., 2006) was used, here, the R package "fdrtool" (Strimmer, 2007).

As the VAR network contains directed edges it is possible to distinguish genes that have mostly outgoing arcs, which could be indicative for key regulatory genes, from those with mostly ingoing arcs. In the graph of Figure 5.5 node 570, an AP2 transcription factor, and node 81, a gene involved in DNA-directed RNA polymerase, belong to the former category, whereas for instance node 558, a structural constituent of ribosome, seems to be part of the latter. Node 627 is another hub in the VAR network, which according to the annotation of Smith et al. (2004) encodes a protein of unknown function. The hypothesized functional module (marked in yellow in the lower right corner of Figure 5.5) is again visible.

We note that the VAR network visualizes influences of the genes over time; hence a VAR graph can also include directed loops and even genes that act upon themselves. Although the VAR graph shows many similarities to the GGM and the network inferred from dynamical correlation, it is not possible to infer the causal structure of the network using the latter.

Finally we remark that the current algorithm employs a fixed "one step ahead" time lag. One strategy to generalization to arbitrary time lags may be to consider functional data and combine the algorithm of chapter 5.1 with the perspective of the autoregressive model. This would have the additional benefit to suitable deal with non-equally spaced measurements, a common characteristic of many biological experiments.

In this chapter we developed methods for learning in complex model for longitudinal data. The VAR model of chapter 5.2 allowed a causal interpretation due to the time structure of the data. In the next chapter we develop a method to infer (partially) causal graphs merely from the covariance matrix.

Chapter 6

Discovering causal structure in high-dimensional data

The interest of research in complex systems ultimately lies in the discovery of causal structures. Although causality is often avoided in statistics, there exist a few models (e.g. simultaneous equations models or Bayesian networks) that introduce directionality in networks and can therefore be interpreted as causal models. However, none of them seems appropriate for inference in the "small n, large p" paradigm. In this chapter, we will introduce a heuristic for discovering causal structure in high-dimensional data, which is based on a decomposition of regression coefficients.

For this, we first give a short description of the possibility of causal inference, and discuss the application of directed networks for an understanding of underlying causal processes in complex systems. Subsequently, we describe the theoretical basis of our learning method and outline the algorithm. After a statistical interpretation, we will apply our method to infer a partially directed network from the *Arabidopsis thaliana* data set. This approach of discovering causal structure in complex models can be found in *Article F*.

6.1 Causality

In statistics, causal models are usually treated with caution and are the object of much controversy (Freedman, 1987; Pearl, 1993). This is based in the belief, that data drawn from observations allow to identify patterns (e.g. correlation between variables), to make predictions (e.g. test problems) or to assume models that generated the data, but that it is ultimately impossible to disclose the true nature beyond perception, which is supposed to be needed to discover causal relations between components of a system. However, this is not required for causality.

Causality understood in a modern sense is based on David Hume, who emphasized the importance of the relation between cause and effect. The relation between two events can be discovered by experience, but the conclusion of cause and effect are drawn by men: if we repeatedly observe two events following one after another, we relate them (Hume,

1740, 1777). Immanuel Kant also locates the idea of causality in human reasoning, but in contrast to Hume, he does not build merely on experience, but shows that the law of causality is inevitably necessary for perception; otherwise it would not be possible to understand the world. Modern physics still builds on Kant's definition of causality, which is based on the succession of time according to the law of causality and the connection of cause and effect that determines all changes. Causality is a pure a priori representation, and can be compared to a model in statistic with which we structure the world we perceive (Kant, 1783, 1787).

As causality is located in human reasoning, it is obvious that different mathematical conceptualizations of causality are possible. A formalization of causality on a statistical basis can be found in the common cause principle by Reichenbach (1956). It states that for two correlated simultaneous events A and B, there exists a prior common cause C which screens off this correlation, that means that A and B are uncorrelated conditioned on C. A causal concept in econometrics is called Granger causality and is a regression model based on the idea of which variables can be used to predict other variables (Granger, 1969). This interpretation was essentially used in chapter 5.2 for the directed network inferred from the VAR model. Another model of causality is the closest-world-concept of David Lewis (Lewis, 1973) who argues with counterfactual conditionals. It finds a statistical concretization in Pearl (2000).

In the following we will introduce a method to infer causal relations only from the correlation matrix of variables. It draws ideas from all the methods mentioned in the last paragraph, e.g., the conditional independence of the common cause principle, the regression model of Granger causality or the analysis of Graphical models by Pearl (2000). In spite of this, the algorithm is simple and computationally efficient. It is based on the idea of ordering variables due to their degree of dependence in the regression model. Nevertheless, we will first discuss the usage of networks for discovering causal structures.

6.2 Causality in directed networks

In the last chapters we used network models to gain knowledge about complex systems. For causal understanding and modeling of the underlying processes in complex systems, the network topology (the structure of nodes and edges) as well as the directionality of the edges is important (e.g. Freedman, 2005; Pearl, 2000). The latter aspect is completely ignored in correlation networks (section 4.2), GGM networks (section 4.3) and networks inferred from dynamical partial correlations (section 5.1). The VAR network (section 5.2) introduces directions in the network. However, this is possible only due to the given time structure of the data (past and future time points). Two widely used classes of directed networks do not require this assumption for the data to assign directionality to the edges in a network: simultaneous equations models (SEMs) and Bayesian networks. These two models are closely related (Wermuth, 1980), albeit the former emphasizes the functional relations whereas the latter focuses on conditional independence properties (Studený, 2005).

Chain graphs Cox and Wermuth (1993) are networks that contain both directed and

undirected edges, and thus generalize both GGMs and Bayesian nets. For this reason, they are natural candidates for a general functional causal model where a *partially* directed network is employed as representative of the system of (potentially nonlinear) causal relations (e.g. Pearl, 2003). However, until now only a handful of algorithms for statistical learning of chain graphs have appeared in the literature (Drton and Eichler, 2006; Aburatani et al., 2006). None of them appears to be particularly suited for producing large-scale graphs, or to be appropriate for small sample analysis.

We will now introduce our method for efficiently constructing partially directed causal graphs from high-dimensional data. As input, it requires only a positive definite estimate of the matrix of pairwise correlations. Computationally, the complexity of our approach is surprisingly modest: it requires no more effort than computing the (partial) correlation graph. In contrast to the correlation network, the resulting graph has a meaningful interpretation in terms of direct interactions, and in contrast to the GGM or the network inferred from dynamical partial correlations it features (partially) directed edges.

First, we describe the theoretical basis of our method, which relies on a certain decomposition of the regression coefficient. Second, we outline the proposed discovery algorithm for selecting large-scale partial causal graphs. Third, we provide a statistical interpretation of this network. Finally, after discussion of various properties of the method we illustrate the approach by learning the partial causal network for the *Arabidopsis thaliana* gene expression experiment.

6.3 Algorithm for discovering causal stucture

6.3.1 Theoretical Basis

Our method for learning chain graphs relies on the decomposition of the standard regression coefficient introduced in section 5.2.1. We will first recapitulate the linear regression model.

We consider the linear regression with Y as response and X_k as covariates. Both X_k and Y are assumed to be random variables with known variances $\mathrm{Var}(Y)$ and $\mathrm{Var}(X_k)$ and with covariance $\mathrm{Cov}(Y, X_k)$ between Y and X_k. The best linear predictor of Y in terms of the X_k that minimizes the MSE of $\sum_k \beta_k X_k - Y$ is given (e.g. ref. Cox and Wermuth, 1993, p. 206) by

$$\beta_k^y = \tilde{\rho}_{yk} \sqrt{\frac{\tilde{\sigma}_y^2}{\tilde{\sigma}_k^2}}, \tag{6.1}$$

where $\tilde{\rho}_{yk}$ is the partial correlation between Y and X_k, and $\tilde{\sigma}_y^2$ and $\tilde{\sigma}_k^2$ are the respective partial variances (see Eq. 5.23). Therefore, β_k^y is completely determined by the joint covariance matrix of Y and X_k. Note that if the usual empirical covariance matrix is substituted into Equation 6.1 the standard OLS estimator for β_k^y is recovered. If there is only a single dependent variable, the above equation reduces to the well-known relation $\beta = \rho_{yk}\sqrt{\sigma_y^2/\sigma_k^2}$, an expression containing only the unconditioned correlation and variances (without the tilde).

Equation 6.1 can be further rewritten by introducing a scale factor (see Eq. 5.24). Denoting the standardized partial variance $(q_k^2)^{-1} = \tilde{\sigma}_k^2/\sigma_k^2$ by SPV_k we can decompose the regression coefficient into the product

$$\beta_k^y = \underbrace{\tilde{\rho}_{yk}}_{\mathcal{A}} \underbrace{\sqrt{\frac{\mathrm{SPV}_y}{\mathrm{SPV}_k}}}_{\mathcal{B}} \underbrace{\sqrt{\frac{\sigma_y^2}{\sigma_k^2}}}_{\mathcal{C}} . \tag{6.2}$$

Note that SPV_y and SPV_k take on values from 0 to 1. We can recall and expand the interpretation of all three factors indicated in section 5.2.1:

$\mathcal{A}$: This factor determines whether there is a direct association between Y and the covariate X_k. If the partial correlation between X_k and Y vanishes, so will also the two corresponding regression coefficients β_k^y and β_y^k. In a partial correlation graph an edge is drawn between two nodes Y and X_k if $\mathcal{A} \neq 0$.

$\mathcal{B}$: This factor adjusts the regression coefficient for the relative reduction in variance of Y and X_k due to the respective other covariates. In the algorithm outlined below a test of $\log(\mathcal{B})$ establishes the directionality of edges of a partially causal network.

$\mathcal{C}$: This is a scale factor correcting for different units in Y and X_k.

The product $\mathcal{AB} = \beta_k^y \sqrt{\sigma_k^2/\sigma_y^2}$ is also known as the standardized regression coefficient. Note that for computing both $\mathcal{A}$ and $\mathcal{B}$ only the correlation matrix is needed, as the variance information is already accounted for by factor $\mathcal{C}$.

6.3.2 Discovery Algorithm

The above decomposition immediately suggests a simple discovery algorithm for statistical learning of a partially causal network from high-dimensional data. By multiple testing of $\mathcal{A} = 0$ we determine the network topology, and by multiple testing of $\log(\mathcal{B}) = 0$ we establish a partial ordering of the nodes, and hence the directionality of the network.

Specifically, our proposed algorithm proceeds as follows:

1. The computation of the partial variances and partial correlations requires as input a positive definite estimate $\boldsymbol{R}$ of the correlation matrix. If the sample size is large ($n >> p$, many more observations than variables) the usual empirical correlation matrix is sufficient. However, for small samples a regularized estimator, e.g., the shrinkage estimator of chapter 3, is needed to improve efficiency and to guarantee positive definiteness. In addition, if the samples are longitudinal it may be necessary to adjust the correlations for autocorrelation using the dynamical correlation of section 5.1.

2. From the estimated correlations we compute plug-in estimates of the factors $\mathcal{A}$ and $\mathcal{B}$ for all possible edges (see section 4.3.1). In this calculation each variable assumes in turn the role of the response Y. Note that $\mathcal{B}$ can be efficiently calculated by taking the square root of the diagonal of the inverse of the estimated correlation matrix, and computing the corresponding pairwise ratios.

3. Following the algorithms outlined in sections 4.3.2 and 4.3.3, we obtain the partial correlation graph by multiple testing of the corresponding coefficients $\mathcal{A}$. Note that for large dimension p the distribution of partial correlations across edges allows to determine the variance parameter of the null hypothesis from the data.

4. In a similar fashion we conduct multiple testing of all $\log(\mathcal{B})$. As $\mathcal{B}$ is the ratio of two variances with the same degrees of freedom $\log(\mathcal{B})$, it is approximately normal distributed with an unknown variance parameter (Fisher, 1924). Using a similar fitting technique as described in Efron (2004b) we obtain an estimate of this parameter, and subsequently compute on this basis (local) false discovery rates for the test $\log(\mathcal{B}) = 0$. Note that we include the values of $\mathcal{B}$ of all edges for inferring the null model and the fdr values, regardless of the corresponding value of $\mathcal{A}$ or the outcome of the test $\mathcal{A} = 0$.

5. Finally, a partially directed network is constructed as follows. All significant edges ($\mathcal{A} \neq 0$) are assembled to graph. Edges in the correlation graph with significant $\log(\mathcal{B}) \neq 0$ are directed such that the arrow points from the variable with the larger standardized partial variance ("exogenous" variable) to the variable with smaller standardized partial variance ("endogenous" variable). The other edges with $\log(\mathcal{B}) \approx 0$ remain undirected.

6.4 Results

6.4.1 Statistical Interpretation

The reasoning behind the above procedure is best understood by considering the connections between systems linear regressions and graphical Gaussian models.

The partial correlation graph is an undirected graphical model whose edges correspond to pairs of variables with non-vanishing partial correlation (Whittaker, 1990). Equation 6.2 points to another interpretation. The partial correlation is the geometric mean of β_k^y and the corresponding reciprocal coefficient β_y^k, i.e.

$$\sqrt{\beta_y^k \beta_k^y} = |\tilde{\rho}_{yk}| \tag{6.3}$$

(see Eq. 4.8). In this light, the GGM represents a system of linear regression equations, where each node is in turn taken as a response variable and regressed against the other remaining nodes. Hence, an undirected edge between node A and B in a GGM is in fact

rather a bi-directed edge, in the sense, that A influences B and vice versa in the underlying system of regression. Thus, the directionality induced by testing $\mathcal{B} = 1$ actually *removes* one of these two directions, namely the arc from the well explained variable with smaller SPV value to the less well explained variable with the larger SPV, rather than adding directionality.

The choice whether or not the bi-directed edge is reduced to a single directed arc hinges on the value of $\mathcal{B}$, a statistic that compares the standardized partial variances of the two variable connected by the investigated edge. If numerator and denominator in $\mathcal{B}$ are not significantly different (i.e. if $\log B \approx 0$) there will be no preferred direction, and hence the edge remains undirected. Otherwise, Equation 6.2 suggests that only the relative variance reduction compared between the two variables involved need to be considered for establishing the direction of an edge.

6.4.2 Further Remarks and Properties

The suggested algorithm and the inferred partially causal networks exhibit several important properties:

1. The computational complexity of the proposed heuristic is $O(p^3)$, where p is the number of nodes. This is no more expensive than computing the partial correlation graph, and allows for estimation of networks containing in the order of thousands and more nodes, even on a small computer. In contrast, this is not possible with traditional search algorithms for Bayesian networks (Lauritzen, 1996) or chain graphs.

2. The estimated partially directed network cannot contain any directed cycles. For instance, it is not possible for a graph to contain a pattern such as $A \to B \to A$. This example would imply $\mathrm{SPV}_A > \mathrm{SPV}_B > \mathrm{SPV}_A$, which is a contradiction.

3. The assignment of directionality is transitive. If there is a directed edge from A to B and from B to C then there must also be a directed edge from A to C (if that edge has non-zero partial correlation).

4. As the algorithm relies on correlations as input, causal models that produce the same correlation matrix are indistinguishable. The existence of such equivalence classes is well known for SEMs (Bollen, 1989) and also for Bayesian belief networks. Consider as simple example the case of only two variables, with the two indistinguishable models $A \to B$ and $B \to A$. In the present approach this is automatically accounted for. In the example the value of $\mathcal{B}$ is identically 1, and hence no directionality is imposed.

5. Both sub-algorithms (for inferring the network topology and for directing the edges) are scale-invariant, meaning that a (linear) change of scale in any of the measured units of the data has no effect on the overall estimated partially causal network.

6. By construction, the topology of the resulting graph is identical with that of the corresponding partial correlation graph.

7. The network discovery algorithm is agnostic with respect to the specific choice of estimator for the correlation matrix, as long as it is positive definite, so that estimates of $\mathcal{A}$ and $\mathcal{B}$ can be derived in a plug-in fashion. In our implementation of the algorithm in the "GeneNet" R package (Schäfer et al., 2006b) we employ the shrinkage estimator of section 3.1.3 for the usual correlation and the shrinkage estimator of section 5.1.2 for the dynamical correlation, which are both suitable for "small n, large p" data.

8. Finally, we remark that the proposed algorithm for learning partially directed graphs might also be understood as variable selection method in linear regression. Interestingly, our method implies that if only a certain one-dimensional regression with a specific response variable is in the center of focus, for the decision for inclusion of a variable (edge) one must still rely on the complete system of equations, through multiple testing of all coefficients $\mathcal{A}$ and $\mathcal{B}$!

6.4.3 Application

To illustrate our algorithm for discovering causal structure, we apply the approach to the *Arabidopsis thaliana* data set. To infer an estimate for the correlation matrix of the *Arabidopsis thaliana*, we use the dynamical correlation shrinkage estimator of section 5.1.2 Opgen-Rhein and Strimmer (2006e).

Subsequently, we employ our discovery algorithm to the estimated (dynamical) correlation matrix. We first take a look at factor $\mathcal{A}$, which determines the existence of an association between genes. We find a total number of 6102 significant edges connecting 669 nodes – the network topology inferred from dynamical correlation, see the network of section 5.1.3. Factor $\mathcal{B}$ determines whether edges are directed. Figure 6.1 elucidates the distribution of $\log \mathcal{B}$, which can be split in a null component and an alternative component. The null distribution (dashed line) follows a normal distribution and characterizes the edges that cannot be directed. The alternative distribution (solid line) can be arbitrarily and signifies the directed edges. In total, we found 15928 significant directions, which corresponds to 0.0498% of potential directions that can be inferred from the correlation matrix.

To construct the network, we assemble the significant edges of factor $\mathcal{A}$ with the significant directions of factor $\mathcal{B}$. In the network of significant associations, 1216 directions are significant, which corresponds to 0.1993% of the possible directions in the network. Note that the fraction of significant directions is by far greater in the subset of the significant partial correlations than in the complete set of all partial correlations. This agrees with the intuitive notion, that causal influences can only be attributed to existing connections between variables.

The network is presented in Figure 6.2. For reasons of clarity we show only the subnetwork containing the 150 most significant edges, which connect 107 nodes. The structure of

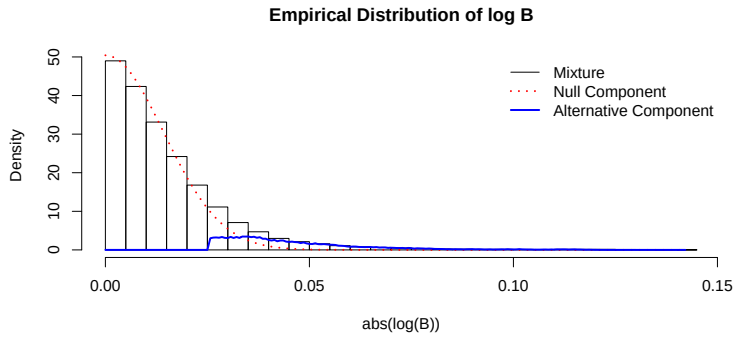

Figure 6.1: Distribution of log $\mathcal{B}$ for the *Arabidopsis thaliana* data. The null distribution is depicted by the dashed line; it follows a normal distribution with zero mean and a standard deviation of 0.014. The solid line signifies the alternative distribution. The empirical distribution (indicated by the histogram) is composed of the null distribution ($\eta_0 = 0.8995$) and of the alternative distribution ($\eta_A = 0.1005$.)

the network (nodes and edges) is equal to the network inferred from dynamical correlation. We see that our algorithm for discovering causal structure can be understood as a directing of a partial correlation network. We will therefore discuss only the *directions* of the edges (for the network topology, see section 5.1.3). We note that many of the hub nodes (colored red) have mostly outgoing arcs. Hubs can be hypothesized to be mainly regulatory genes: genes that influence a large number of other genes. We see that the directions our algorithm assigns the edges of the hub genes are exactly what can be expected from a biological perspective.

In the "functional module" (colored yellow in the lower right corner of Figure 6.2), it is not possible to determine any directions, which could be due to complex interactions among the nodes of the module, so that none of these genes play a key role in the regulation of the other genes in the module.

We see that the partially directed network *Arabidopsis thaliana* contains both, directed and undirected nodes. It is a great advantage of this model, that it provides a causal meaning for the edges of a network, but unlike, e.g., a vector autoregressive model, it does not *force* directions for the edges. A causal meaning is given to edges only on a basis of statistical testing.

For comparison, we display the correlation graph for the *Arabidopsis thaliana* data in Figure 6.3. Specifically, we display the 150 edges that have the largest correlations. In contrast to the partially directed network of Figure 6.2 and to the general expectation for biological networks (Ravasz et al., 2002; Barabási and Oltvai, 2004), we find no hub structure in the network. We see that our method not only allows causal interpretations of the interaction between genes, but is also makes much more sense biologically.

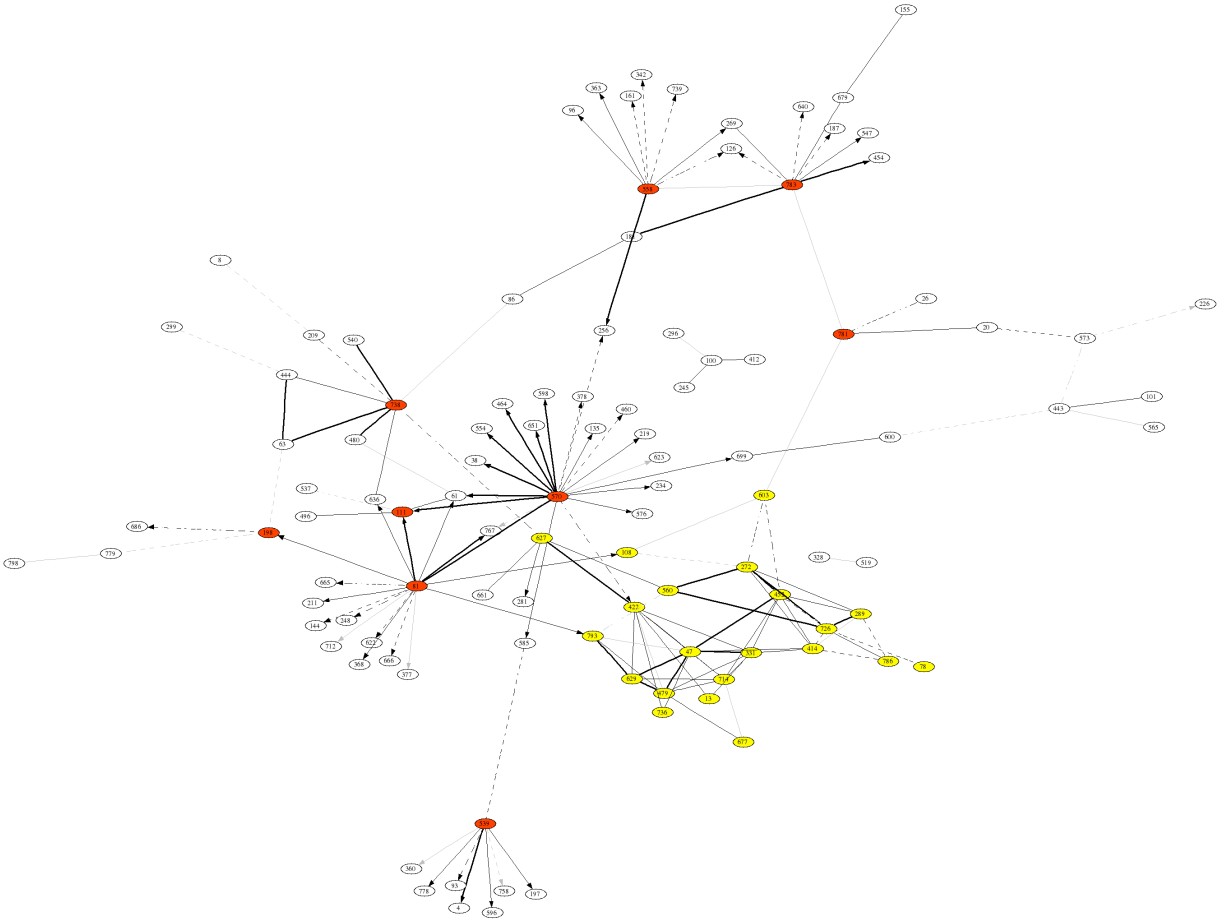

Figure 6.2: Partially causal network inferred from the *Arabidopsis thaliana* data. The solid and dotted lines indicate positive and negative partial correlation coefficients, respectively, and the line intensity denotes their strength. Significant directions are depicted by an arrow.

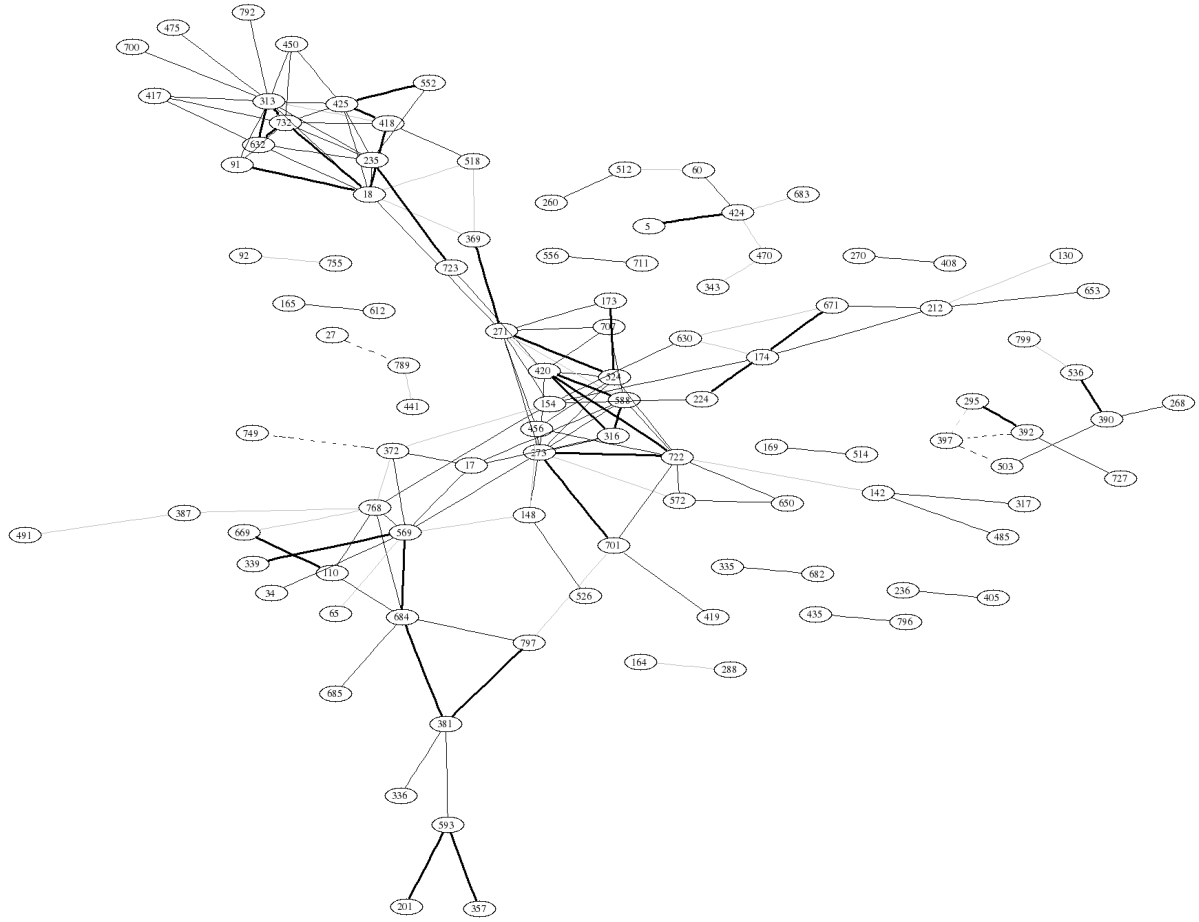

Figure 6.3: Correlation network inferred from the *Arabidopsis thaliana* data. The solid and dotted lines indicate positive and negative correlation coefficients, respectively, and the line intensity denotes their strength. The network displays the 150 edges that have the largest correlations.

6.5 Discussion

Methods for exploring causal structures in high-dimensional data are growing in importance, particularly in the study of complex biological, medical and financial systems. Existing approaches for interpreting these data using correlation networks are inherently problematic, as these tend to obscure rather than uncover direct associations and interactions. Current algorithms for learning partially causal networks are also not suited for high-dimensional data.

We suggest here a simple yet versatile heuristic for screening large-scale data set for causal structure. Our procedure is based on projecting an estimated (partial) ordering of nodes onto a partial correlation graph. It can be furnished both in a population version (large n) and in a small sample version and it is computationally efficient. Moreover, it does not require any interventional data, and in applications has produced highly informative causal networks.

We are aware that more sophisticated algorithms may be devised, e.g., by accounting for non-linear effects, by employing entropy criteria, or by refining causal models (e.g. Pearl, 2000; Spirtes et al., 2001). However, our approach has the crucial advantage of being both computationally as well as algorithmically nearly as simple as a correlation network, thus it is readily applicable to many kinds of data. At the same time, our discovery algorithm produces networks with highly relevant association structure and with causal interpretation.

Chapter 7

Outlook

Learning in complex systems is a challenge for statistical analysis. In this thesis we used a Stein-type shrinkage approach to extend the area of application of traditional methods to high-dimensional data sets. The combination of frequentist and Bayesian ideas Stein-type shrinkage builds upon seems very promising in solving the problems that learning in complex systems implicates (Efron, 2005a).

For high-dimensional data sets, we introduced a new method for case-control analysis and discussed graphical Gaussian models to infer a network, which elucidates the relationships between variables. The concept of dynamical correlation and the application of the vector autoregressive model allowed extending graphical Gaussian modeling to time course data sets. The introduction of an algorithm for inferring (partially) directed graphs provided a method for discovering causal structures in high-dimensional data sets.

The methods and ideas introduced in this thesis allow further research in many directions. It is for example conceivable to combine the functional data approach of dynamical correlation with the vector autoregressive model to extend the analysis of longitudinal data to data sets with unequally spaced time points. Another approach could be to further improve the efficiency of the Stein estimation by gearing it towards achieving the best results for the particular application (like estimation of regression coefficients) instead of founding the shrinkage estimation only on improving the efficiency of the estimation of covariance matrices. Another interesting area could be to further investigate the concept of causal inference used to infer partially directed networks.

In any case, the development of learning methods for complex systems will continue to be an important area of research.

Appendix i

Description of the articles

- *Article A*: *Accurate Ranking of Differentially Expressed Genes by a Distribution-Free Shrinkage Approach*
  This article discusses the method of Stein-type regularization and introduces variance shrinkage. This provides in combination with shrinkage of the correlation matrix a regularized estimator for the covariance matrix. We also describe limited translation, a version of shrinkage that not only controls the ensemble risk, but also takes the risk of individual components into account. Furthermore we introduce the "shrinkage t" statistic, a high-dimensional case-control analysis. This computationally inexpensive method is compared with alternative methods for synthetic and real expression data.

- *Article B*: *Inferring Gene Dependency Networks from Genomic Longitudinal Data: A Functional Data Approach*
  Here we introduce dynamical correlation, a generalization of the usual correlation applicable to time series data. It is compared with an alternative definition of dynamical correlation, demonstrated on a toy data set and applied to real data.

- *Article C*: *Using Regularized Dynamic Correlation to infer Gene Dependency Networks from time-series Microarray Data*
  This article elaborates the method of *Article B* and infers a Stein-type shrinkage estimator for dynamical correlation. This allows applying the concept of dynamical correlation to high-dimensional data sets.

- *Article D*: *Condition Number and Variance Inflation Factor to Detect Multicollinearity*
  This technical report examines measures for multicollinearity in linear regression. The contribution to this thesis lies in a new interpretation and decomposition of the regression coefficient. This can be used for a test statistic for correlation coefficients and allows causal interpretation in regression models.

- *Article E*: *Learning Causal Networks from Systems Biology Time Course Data: An Effective Model Selection Procedure for the Vector Autoregressive Process*

This article introduces a regularized regression method, called shrinkage regression, which allows to estimate the coefficients of a vector autoregressive model for high-dimensional data. Combined with a local fdr we infer a model selection algorithm for the vector autoregressive process.

- *Article F*: *From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data*
  In this article, we introduce causality in graphical Gaussian models. We develop a method to order variables according to the degree of dependence in linear regression. This allows assigning directions to the GGM. The method is demonstrated by applying it to a real data set.

- *Article G*: *Reverse Engineering Genetic Networks using the* **GeneNet** *package*
  This article describes the implementation of the shrinkage estimation of the graphical Gaussian model for both, the empirical and the dynamical correlation in the package **GeneNet** for the statistical computer language **R**.

Appendix ii

Description of the R packages

- *"corpcor": Efficient estimation of covariance and (partial) correlation*
  This package implements the Stein-type shrinkage estimation for the variance, the correlation and the covariance. Moreover it provides functions for calculating their partial counterparts, for both shrinkage and empirical estimators. Additionally, it contains some further tools for matrix calculation

- *"longitudinal": Analysis of multiple time course data.*
  This package allows working with time series data using dynamical correlation. It provides a data structure for longitudinal data including some utility functions. Furthermore, it contains the *human t-cell* data set by Rangel et al. (2004).

- *"st": R package implementing the "shrinkage t" statistic*
  This package provides the shrinkage t statistic, which allows high-dimensional case-control analysis. It also implements functions that allow accessing all the alternative methods for analyzing case-control data mentioned in *Article A*. Additionally, it provides the *"golden spike" Affymetrix* data set by Choe et al. (2005).

- *"GeneNet": Modeling and inferring gene networks*
  This package is concerned with the inference of graphical Gaussian models. It allows calculating the strength of the connection between two edges of the network, that is, their (regularized) partial correlation. It implements functions to test for significant edges using the local fdr algorithm and subsequently to plot the network. Furthermore, it contains some utility functions and the *Escherichia coli* data set by Schmidt-Heck et al. (2004) as well as the *Arabidopsis thaliana* data set by Smith et al. (2004).

Appendix iii

Definitions, Notations, and Abbreviations

Definitions

	Definition	True value	Estimate
Covariance matrix:	$\mathrm{Cov}(X_k, X_l) = \sigma_{kl}$	$\boldsymbol{\Sigma} = (\sigma_{kl})$	$\boldsymbol{S} = (s_{kl})$
Concentration matrix:	$\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$	$\boldsymbol{\Omega} = (\omega_{kl})$	
Variances:	$\mathrm{Var}(X_k) = \sigma_{kk} = \sigma_k^2$	σ_{kk}	s_{kk}
Correlation matrix:	$\mathrm{Corr}(X_k, X_l) = \rho_{kl}$ $= \sigma_{kl}(\sigma_{kk}\sigma_{ll})^{-1/2}$ $\boldsymbol{Q} = (q_{kl}) = \boldsymbol{P}^{-1}$	$\boldsymbol{P} = (\rho_{kl})$	$\boldsymbol{R} = (r_{kl})$
Partial variances	$\mathrm{Var}(X_k \mid X_{\neq k}) = \tilde{\sigma}_{kk} = \tilde{\sigma}_k^2 = \omega_{kk}^{-1}$	$\tilde{\sigma}_{kk}$	$\tilde{s}_{kk}$
Partial correlations:	$\mathrm{Corr}(X_k, X_l \mid X_{\neq k,l}) = \tilde{\rho}_{kl}$ $= -\omega_{kl}(\omega_{kk}\omega_{ll})^{-1/2}$	$\tilde{\boldsymbol{P}} = (\tilde{\rho}_{kl})$	$\tilde{\boldsymbol{R}} = (\tilde{r}_{kl})$

Notations and Abbreviations

n	sample size (number of observations)
p	dimension (number of variables)
X	data matrix ($n \times p$)
θ	vector of parameters
$\hat{}$	estimated quantities, e.g.: $\hat{\theta}$ is the estimated vector of parameters
$\tilde{}$	"partial" quantities, e.g.: $\tilde{\boldsymbol{R}}$ is the partial correlation matrix
*	shrinkage quantities, e.g.: $\lambda^\star$ is the optimal shrinkage parameter
λ	shrinkage parameter
$< \cdot >$	functional inner product
δ^0	unregularized estimation rule

δ^{λ}	James-Stein shrinkage estimation rule
θ^{Target}	target estimate
$\odot$	Hadamard (elementwise) product
i.i.d.	independent and identically distributed
MSE	mean squared error
FDR	false discovery rate
fdr	local false discovery rate
SPV_k	standardized partial variance: $\text{SPV}_k = q_k^{-1}$
GGM	graphical Gaussian model
VAR	vector autoregressive model
SEM	structural equations model

List of Figures

Bibliography

Aburatani, S., Saito, S., Toh, H., and Horimoto, K. (2006). A graphical chain model for inferring regulatory system networks from gene expression profiles. *Statist. Meth.*, 3:17–28.

Auerbach, D., Thaminy, S., Hottinger, M. O., and Stagljar, I. (2002). The post-genomic era of interactive proteomics: Facts and perspectives. *Proteomics*, 2(6):611–623.

Baldi, P. and Long, A. D. (2001). A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics*, 17:509–519.

Barabási, A.-L. and Oltvai, Z. N. (2004). Network biology: Understanding the cell's functional organization. *Genetics*, 5:101–113.

Barachnik, A. J. (1970). A family of minimax estimators of the mean of a multivariate normal distribution. *Ann. Math. Statist.*, 41:642–645.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B*, 57:289–300.

Bickel, P. J. and Levina, E. (2006). Regularized estimation of large covariance matrices. Technical Report 716, Berkeley University.

Bollen, K. A. (1989). *Structural Equations With Latent Variables*. John Wiley & Sons.

Bornholdt, S. (2005). Less is more in modeling large genetic networks. *Science*, 310(5747):449–451.

Choe, S. E., Boutros, M., Michelson, A. M., Church, G. M., and Halfon, M. S. (2005). Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control data set. *Genome Biology*, 6:R16.

Churchill, G. A. (2002). Fundamentals of experimental design for cDNA microarrays. *Nature Genetics*, 32:490–495.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.*, 74:829–836.

Clish, C., Davidov, E., Oresic, M., Plasterer, T., Lavine, G., Londo, T., Meys, M., Snell, P., Stochaj, W., and Adourian, A. (2004). Integrative biological analysis of the APOE*3-Leiden transgenic mouse. *Omics: A Journal of Integrative Biology*, 8(1):3–13.

Cope, L. M., Irizarray, R. A., Jaffee, H. A., Wu, Z., and Speed, T. P. (2004). A benchmark for Affymetrix GeneChip expression measures. *Bioinformatics*, 20:323–331.

Cox, D. R. and Wermuth, N. (1993). Linear dependencies represented by chain graphs. *Statistical Science*, 8(3):204–218.

Crick, F. H. (1958). On protein synthesis. *Symp. Soc. Exp. Biol.*, 12:138–163.

Cui, X. and Churchill, G. A. (2003). Statistical test for differential expression in cDNA microarray experiments. *Genome Biology*, 4:R210.

Cui, X., Hwang, J. T. G., Qiu, J., Blades, N. J., and Churchill, G. A. (2005). Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics*, 6:59–75.

Delmar, P., Robin, S., Tronik-Le Roux, D., and Daudin, J. J. (2005). Mixture model on the variance for the differential analysis of gene expression data. *Appl. Statist.*, 54:31–50.

Diggle, P., Heagerty, P., Liang, K.-Y., and Zeger, S. (2002). *Analysis of Longitudinal Data*. Oxford University Press.

Dobra, A., Hans, C., Jones, B., Nevins, J. R., Yao, G., and West, M. (2004). Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*, 90:196–212.

Drton, M. and Eichler, M. (2006). Maximum likelihood estimation in gaussian chain graph models under the alternative markov property. *Scand. J. Statist*, 33:247–257.

Dubin, J. A. and Müller, H.-G. (2005). Dynamical correlation for multivariate longitudinal data. *Journal of the American Statistical Association*, 100(471):872–881.

Efron, B. (2003). Robbins, empirical Bayes and microarrays. *Annals of Statistics*, 31(2):366–378.

Efron, B. (2004a). The estimation of prediction error: Covariance penalties and cross-validation. *Journal of the American Statistical Association*, 99(467):619–642.

Efron, B. (2004b). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *Journal of the American Statistical Association*, 99(465):96–104(9).

Efron, B. (2005a). Bayesians, frequentists, and scientists. *Journal of the American Statistical Association*, 100(469):1–5(5).

Efron, B. (2005b). Correlation and large-scale simultaneous significance testing. Technical report, Dept. of Statistics, Stanford University.

Efron, B. (2005c). Local false discovery rates. Technical Report 2005-20B/234, Dept. of Statistics, Stanford University.

Efron, B. and Morris, C. (1972). Limiting the risk of Bayes and empirical Bayes estimators – part II: The empirical Bayes case. *Journal of the American Statistical Association*, 67:130–139.

Efron, B. and Morris, C. N. (1973). Stein's estimation rule and its competitors – an empirical Bayes approach. *J. Amer. Statist. Assoc.*, 68:117–130.

Efron, B. and Morris, C. N. (1975). Data analysis using Stein's estimator and its generalizations. *J. Amer. Statist. Assoc.*, 70:311–319.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.*, 96:1151–1160.

Efron, B., Turnbull, B. B., and Narasimhan, B. (2006). *locfdr: Computes local false discovery rates*. R package version 1.1-3.

Fisher, R. A. (1924). On a distribution yielding the error functions of several well known statistics. *Proc. Intl. Congr. Math.*, 2:805–813.

Fox, R. J. and Dimmic, M. W. (2006). A two-sample Bayesian t-test for microarray data. *BMC Bioinformatics*, 7:126.

Freedman, D. A. (1987). As others see us: A case study in path analysis. *Journal of Educational Statistics*, 12(2):101–128.

Freedman, D. A. (2005). *Statistical Models: Theory and Practice*. Cambridge University Press, Cambridge, UK.

Friedman, N. (2004). Inferring cellular networks using probabilistic graphical models. *Science*, 303(5659):799–805.

Gallagher, R. and Appenzeller, T. (1999). Beyond reductionism. *Science*, 284(5411):79.

Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. (2004). Affy - analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, 20:307–315.

Gelman, A. and Pardoe, I. (2006). Bayesian measures of explained variance and pooling in multilevel (hierarchical) models. *Technometrics*, 48:241–251.

Gibson, G. and Muse, S. V. (2004). *A Primer of Genome Science*. Sinauer Associates, 2nd edition.

Golub, G. H., Heath, M., , and Wahba, G. (1979). Generalized crossvalidation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223.

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438.

Granger, C. W. J. (1980). Testing for causality, a personal viewpoint. *J. Econom. Dyn. Control*, 2:329–352.

Graur, D. and Li, W.-H. (2000). *Fundamentals of Molecular Evolution*. Sinauer Associates, 2nd edition.

Gruber, M. H. J. (1998). *Improving Efficiency By Shrinkage*. Marcel Dekker, Inc., New York.

Hartemink, A. J., Gifford, D. K., Jaakkola, T. S., and Young, R. A. (2002). Bayesian methods for elucidating genetic regulatory networks. *IEEE Intelligent Systems*, 17(2):37–43.

Hotelling, H. (1953). New light on the correlation coefficient and its transforms. *Journal of the Royal Statistical Society. Series B (Methodological)*, 15(2):193–232.

Huang, J. Z., Liu, N., Pourahmadi, M., and Liu, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, 93:85–98.

Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A., and Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18 Suppl. 1:96–104.

Huber, W., von Heydebreck, A., Sültmann, H., Poustka, A., and Vingron, M. (2003). Parameter estimation for the calibration and variance stabilization of microarray data. *Statist. Appl. Genet. Mol. Biol.*, 2:3.

Hume, D. (1740). *A Treatise of Human Nature*. Reprinted by Hard Press 2006. Written 1739–1740.

Hume, D. (1777). *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*. Reprinted by Oxford University Press 1975.

Ideker, T., Thorsson, V., Seigel, A. F., and Hood, L. E. (2000). Testing for differentially expressed genes by maximum likelihood analysis of microarray data. *J. Comp. Biol.*, 7:805–817.

Irizarry, R. A., Cope, L., and Wu, Z. (2006). Feature-level exploration of a published control data set. *Genome Biology*, 7:404.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In Neyman, J., editor, *Proc. Fourth Berkeley Symp. Math. Statist. Probab.*, volume 1, pages 361–379, Berkeley. Univ. California Press.

Kabán, A., Sun, J., Raychaudhury, S., and Nolan, L. (2006). On class visualisation for high dimensional data: Exploring scientific data sets. *Lecture Notes in Computer Science*, 4265:125–136.

Kant, I. (1783). *Prolegomena zu einer jeden künftigen Metaphysik, die als Wissenschaft wird auftreten können (Prolegomena to any Future Metaphysics).* Reprinted by Felix Meiner Verlag 2001.

Kant, I. (1787). *Kritik der reinen Vernunft (Critique of Pure Reason).* Reprinted by Felix Meiner Verlag 1998.

Kishino, H. and Waddell, P. J. (2000). Correspondence analysis of genes and tissue types and finding genetics links from microarray data. *Genome Informatics*, 11:83–95.

Kitano, H. (2002). Systems biology: A brief overview. *Science*, 295(5747):1662–1664.

Klipp, E., Herwig, R., and Kowald, A. (2005). *Systems Biology in Practice. Concepts, Implementation and Application.* Wiley-VCH.

Lauritzen, S. (1996). *Graphical Models.* Oxford University Press, Oxford.

Ledoit, O. and Wolf, M. (2003a). Honey, I shrunk the sample covariance matrix. Economics Working Papers 691, Department of Economics and Business, Universitat Pompeu Fabra.

Ledoit, O. and Wolf, M. (2003b). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *J. Empir. Finance*, 10:603–621.

Lehmann, E. and Casella, G. (1998). *Theory of Point Estimation.* Springer, 2nd edition.

Levine, R. A. and Berliner, L. M. (1999). Statistical principles for climate change studies. *Journal of Climate*, 12:564.

Lewis, D. K. (1973). *Counterfactuals.* Blackwell & Harvard U.P.

Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for the linear model. *J. R. Statist. Soc. B*, 34:1–72.

Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Norton, H., and Brown, E. L. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, 14:1675–1680.

Lönnstedt, I. and Speed, T. (2002). Replicated microarray data. *Statistica Sinica*, 12:31–46.

Lütkepohl, H. (1993). *Introduction to Multiple Time Series Analysis*. Springer, Berlin.

Mantegna, R. N. and Stanley, H. E. (2000). *In Introduction to Econophysics: Correlations and Complexity in Finance*. Cambridge University Press.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Ann. Statist.*, 34:1436–1462.

Murphy, A. H. and Wilks, D. S. (1998). A case study of the use of statistical models in forecast verification: Precipitation probability forecasts. *Weather Forecast*, 13(2:3):795–810.

Newton, M. A., Kendziorski, C. M., Richmond, C. S., Blattner, F. R., and Tsui, K. W. (2001). On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *J. Comp. Biol.*, 8:37–52.

Newton, M. A., Noueiry, A., Sarkar, D., and Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, 5:155–176.

Nocedal, J. and Wright, S. J. (2000). *Numerical Optimization*. Springer.

Oltvai, Z. N. and Barabási, A.-L. (2002). Life's complexity pyramid. *Science*, 298(5594):763–764.

Opgen-Rhein, R. (2007). Condition number and variance inflation factor to detect multicollinearity. Technical report 01/07.

Opgen-Rhein, R., Schäfer, J., and Strimmer, K. (2007). *GeneNet: Modeling and Inferring Gene Networks*. R package version 1.1.0.

Opgen-Rhein, R. and Strimmer, K. (2006a). Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach. *Statist. Appl. Genet. Mol. Biol.*, 6(1):9.

Opgen-Rhein, R. and Strimmer, K. (2006b). Inferring gene dependency networks from genomic longitudinal data: a functional data approach. *REVSTAT*, 4:53–65.

Opgen-Rhein, R. and Strimmer, K. (2006c). *longitudinal: Analysis of Multiple Time Course Data*. R package version 1.1.3.

Opgen-Rhein, R. and Strimmer, K. (2006d). *st: Shrinkage t Statistic*. R package version 1.0.0.

Opgen-Rhein, R. and Strimmer, K. (2006e). Using regularized dynamic correlation to infer gene dependency networks from time-series microarray data. In *Proceedings of the 4th International Workshop on Computational Systems Biology (WCSB 2006)*, volume 4, pages 73–76, Tampere.

Opgen-Rhein, R. and Strimmer, K. (2007a). From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. *BMC Systems Biology*, 1:37.

Opgen-Rhein, R. and Strimmer, K. (2007b). Learning causal networks from systems biology time course data: an effective model selection procedure for the vector autoregressive process. *BMC Bioinformatics*, 8 (Suppl. 2):S3.

Pappenberger, F., amd N.M. Hunter, K. B., Bates, P., Gouweleeuw, B., Thielen, J., and de Roo, A. (2005). Cascading model uncertainty from medium range weather forecasts (10 days) through a rainfall-runoff model to flood inundation predictions within the European flood forecasting system (effs). *Hydrology and Earth System Sciences*, 9(4):381–393.

Patat, F. (2003). Ubvri night sky brightness during sunspot maximum at eso-paranal. *Astronomy & Astrophysics*, 400:1183–1198.

Pearl, J. (1993). Aspects of graphical models connected with causality. In *Proceedings of the 49th Session of the International Statistical Institute*, pages 391–401.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, UK.

Pearl, J. (2003). Statistics and causal inference: a review. *TEST*, 2:281–345.

R Development Core Team (2006). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer, 2nd edition.

Rangel, C., Angus, J., Ghahramani, Z., Lioumi, M., Sotheran, E., Gaiba, A., Wild, D. L., and Falciani, F. (2004). Modeling t-cell activation using gene expression profiling and state-space models. *Bioinformatics*, 20(9):1361–1372.

Raudenbush, S. W. and Bryk, A. S. (2001). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage Publications.

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabási, A.-L. (2002). Hierarchical organization of modularity in metabolic networks. *Science*, 297:1551–1555.

Reichenbach, H. (1956). *The Direction of Time*. Berkeley, University of Los Angeles Press.

Robbins, H. (1956). An empirical Bayes approach to statistics. In *Proc. 3rd Berkeley Sympos. Math. Statist. Probability 1*, pages 157–163. Berkeley, University of California Press.

Ruppert, D. (2004). *Statistics and Finance: an introduction*. Springer.

Schäfer, J. (2006). *Small-Sample Analysis and Inference of Networked Dependency Structures from Complex Genomic Data*. PhD thesis, LMU München.

Schäfer, J., Opgen-Rhein, R., and Strimmer, K. (2006a). *corpcor: Efficient Estimation of Covariance and (Partial) Correlation*. R package version 1.4.4.

Schäfer, J., Opgen-Rhein, R., and Strimmer, K. (2006b). Reverse engineering genetic networks using the **GeneNet** package. *R News*, 6(5):50–53.

Schäfer, J. and Strimmer, K. (2005a). An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, 21(6):754–764.

Schäfer, J. and Strimmer, K. (2005b). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4(1):32.

Schmidt-Heck, W., Guthke, R., Toepfer, S., Reischer, H., Dürrschmid, K., and Bayer, K. (2004). Reverse engineering of the stress response during expression of a recombinant protein. In *Proceedings of the EUNITE symposium*, pages 407–412, Aachen, Germany. Verlag Mainz.

Smith, S. M., Fulton, D. C., Chia, T., Thorneycroft, D., Chapple, A., Dunstan, H., Hylton, C., and Smith, S. C. Z. A. M. (2004). Diurnal changes in the transcriptom encoding enzymes of starch metabolism provide evidence for both transcriptional and post-transcriptional regulation of starch metabolism in Arabidopsis leaves. *Plant Physiol.*, 136:2687–2699.

Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statist. Appl. Genet. Mol. Biol.*, 3:3.

Spirtes, P., Glymour, C., and Scheines, R. (2001). *Causation, Prediction, and Search*. MIT Press, 2nd edition. edition.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In Neyman, J., editor, *Proc. Third Berkeley Symp. Math. Statist. Probab.*, volume 1, pages 197–206, Berkeley. Univ. California Press.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.*, 6:1135–1151.

Stigler, S. M. (1990). A Galtonian perspective on shrinkage estimators. *Statistical Science*, 5:147–155.

Strimmer, K. (2007). *fdrtool: Estimation and Control of (Local) False Discovery Rates*. R package version 1.1.0.

Studený, M. (2005). *Probabilistic Conditional Independence Structures*. Springer.

Thompson, J. R. (1968). Some shrinkage techniques for estimating the mean. *J. Amer. Statist. Assoc.*, 63(321):113–122.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B*, 58:267–288.

Toh, H. and Horimoto, K. (2002a). Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modeling. *Bioinformatics*, 18:287–297.

Toh, H. and Horimoto, K. (2002b). System for automatically inferring a genetic network from expression profiles. *J. Biol. Physics*, 28:449–464.

Tong, T. and Wang, Y. (2007). Optimal shrinkage estimation of variances with applications to microarray data analysis. *J. Amer. Statist. Assoc.*, 12(477):113–122.

Tusher, V., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA*, 98:5116–5121.

van 't Wout, A. B., Lehrman, G. K., Mikheeva, S. A., O'Keeffe, G. C., Katze, M. G., Bumgarner, R. E., Geiss, G. K., and Mullins, J. I. (2003). Cellular gene expression upon human immunodeficiency virus type 1 infection of CD4(+)-T-cell lines. *J Virol*, 77(2):1392–1402.

Waddell, P. J. and Kishino, H. (2000). Cluster inference methods and graphical models evaluated on NCI60 microarray gene expression data. *Genome Informatics*, 11:129–140.

Wermuth, N. (1980). Linear recursive equations, covariance selection, and path analysis. *J. Amer. Statist. Assoc.*, 75:963–972.

West, M., Nevins, J. R., Marks, J. R., Spang, R., and Zuzan, H. (2000). Bayesian regression analysis in the "large p, small n" paradigm with application in DNA microarray studies. *Technical Report*.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley, New York.

Wichert, S., Fokianos, K., and Strimmer, K. (2004). Identifying periodically expressed transcripts in microarray time series data. *Bioinformatics*, 20(1):5–20.

Wright, G. W. and Simon, R. M. (2003). A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics*, 19:2448–2455.

Wu, B. (2005). Differential gene expression detection using penalized linear regression models: the improved SAM statistic. *Bioinformatics*, 21:1565–1571.

Yang, Y. H. and Speed, T. (2002). Design issues for cDNA microarray experiments. *Nature reviews (Genetics)*, 3(8):579–588.

Part II

Articles

Article A

Accurate Ranking of Differentially Expressed Genes by a Distribution-Free Shrinkage Approach

Published in Statistical Applications in Genetics and Molecular Biology (SAGMB) *(Volume 6, Issue 1, Article 9, 2007)*

Authors: Rainer Opgen-Rhein *and* Korbinian Strimmer

Abstract:

- High-dimensional case-control analysis is encountered in many different settings in genomics. In order to rank genes accordingly, many different scores have been proposed, ranging from ad hoc modifications of the ordinary t statistic to complicated hierarchical Bayesian models.

  Here, we introduce the "shrinkage t" statistic that is based on a novel and model-free shrinkage estimate of the variance vector across genes. This is derived in a quasi-empirical Bayes setting. The new rank score is fully automatic and requires no specification of parameters or distributions. It is computationally inexpensive and can be written analytically in closed form.

  Using a series of synthetic and three real expression data we studied the quality of gene rankings produced by the "shrinkage t" statistic. The new score consistently leads to highly accurate rankings for the complete range of investigated data sets and all considered scenarios for across-gene variance structures.

Key-Words:

- *High-dimensional case-control data, James-Stein shrinkage, limited-translation, quasi-empirical Bayes, regularized t statistic, variance shrinkage.*

A.1 Introduction

High-dimensional case-control analysis is a key problem that has many applications in computational genomics. The most well-known example is that of ranking genes according to differential expression, but there are many other instances that warrant similar statistical methodology, such as the problem of detecting peaks in mass spectrometric data or finding genomic enrichment sites.

All these problems have in common that they require a variant of a (regularized) t statistic that is suitable for high-dimensional data and large-scale multiple testing. For this purpose, in the last few years various test statistics have been suggested, which may be classified as follows:

i) Simple methods: fold change, classical t statistic.

ii) Ad hoc modifications of ordinary t statistic: Efron's 90% rule (Efron et al., 2001), SAM (Tusher et al., 2001).

iii) (Penalized) likelihood methods, e.g.: Ideker et al. (2000), Wright and Simon (2003), Wu (2005).

iv) Hierarchical Bayes methods, e.g.: Newton et al. (2001), Baldi and Long (2001), Lönnstedt and Speed (2002), Newton et al. (2004), "moderated t" (Smyth, 2004), Cui et al. (2005), Fox and Dimmic (2006).

For an introductory review of most of these approaches see, e.g., Cui and Churchill (2003) and Smyth (2004).

Current good practice in gene expression case-control analysis favors the empirical or full Bayesian approaches (item iv) over other competing methods. The reason behind this is that Bayesian methods naturally allow for information sharing across genes, which is essential when the number of sample is as small in typical genomic experiments. Specifically, the estimation of gene-specific variances profits substantially from pooling information across genes (e.g., Wright and Simon, 2003; Smyth, 2004; Delmar et al., 2005; Cui et al., 2005). On the other hand, Bayesian methods can become computationally quite expensive, and more importantly, typically rely on a host of very detailed assumptions concerning the underlying data and parameter generating models.

In this paper we introduce a novel "shrinkage t" approach that is as simple as the ad hoc rules (item ii) but performs as well as fully Bayesian models (item iv), even in simulation settings that are favorable to the latter. Moreover, the new gene ranking statistic is fully analytic, requires no computer-intensive procedures, and is derived without any specific distributional assumptions. In this sense, it is a further development of the quasi-likelihood approach of Strimmer (2003) but with additional regularization.

The shrinkage t statistic is developed in the framework of James-Stein-type analytic shrinkage (e.g., Gruber, 1998; Schäfer and Strimmer, 2005). This approach offers a highly efficient means for regularized inference, both in the statistical and computational sense.

It is complementary to more well-known alternatives such as Bayesian and penalized likelihood inference. Nevertheless, the resulting estimators are typically very hard to improve (Yi-Shi Shao and Strawderman, 1994). James-Stein shrinkage estimation may also be understood as a "quasi-empirical Bayes" method as only information concerning second moments rather than fully specified distributions are used. In short, analytic shrinkage estimators combine properties that render them very attractive for analyzing large-dimensional genomic assays.

In the context of differential expression a similar approach was suggested before only by Cui et al. (2005) who also employ James-Stein estimation to obtain shrinkage estimates of the gene-specific variances. Our approach shares many aspects with that of Cui et al. (2005). However, our estimator for variance shrinking is different in that absolutely no distributional assumptions are involved (not even for hyperparameters). Moreover, it is applied on the original data scale and thus requires no transformations. Finally, it is derived via a rather general route for constructing Stein-type estimators, and results in a very compact, fully analytic, and yet still highly efficient estimator for variance shrinkage (Eq. A.10 and Eq. A.11).

The remainder of this paper is structured as follows. In the next section we briefly review analytic shrinkage estimation. Subsequently, we develop an estimator for inference of gene-specific variances and construct the shrinkage t score for ranking differentially expressed genes. Subsequently, we investigate the performance of this statistic in simulations and in extensive data analysis in comparison relative to a number of competing statistics. The final section contains a discussion of the results.

A.2 Distribution-Free Shrinkage Estimation

In this section we describe how analytic James-Stein-type shrinkage estimators may be constructed from an arbitrary unregularized estimator, without assuming any distributions for data or the model parameters.

A.2.1 James-Stein Shrinkage Rules

Initially, we assume that an unregularized estimation rule

$$\delta^0 = \hat{\boldsymbol{\theta}}, \tag{A.1}$$

is available, e.g., the maximum-likelihood or the minimum variance unbiased estimate. It is important here that $\hat{\boldsymbol{\theta}}$ is a *vector* $(\theta_1, \ldots, \theta_k, \ldots, \theta_p)^T$. (In the specific example of the present article this vector contains all gene-specific empirical variances.) Then the James-Stein ensemble shrinkage estimation rule may be written as

$$\begin{aligned} \delta^\lambda &= \delta^0 - \lambda \boldsymbol{\Delta} \\ &= \hat{\boldsymbol{\theta}} - \lambda \, (\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}^{\text{Target}}). \end{aligned} \tag{A.2}$$

In other words, the shrinkage estimate δ^λ is the linear combination $\lambda\hat{\boldsymbol{\theta}}^{\text{Target}} + (1-\lambda)\hat{\boldsymbol{\theta}}$ of the original estimator $\hat{\boldsymbol{\theta}}$ and a target estimate $\hat{\boldsymbol{\theta}}^{\text{Target}}$. The parameter λ determines the extent to which these estimates are pooled together. If $\lambda = 1$ then the target dominates completely, whereas for $\lambda = 0$ no shrinkage occurs.

In James-Stein estimation the search for the optimal shrinkage intensity λ is considered from a decision theoretic perspective. First, a loss function is selected (e.g. the squared error). Second, λ is chosen such that the corresponding risk of δ^λ, i.e. the expectation of the loss with respect to the data (e.g. the mean squared error, MSE), is minimized.

Interestingly, it turns out that for squared error loss this can be done *without* any reference to the unknown true value $\boldsymbol{\theta}$, as the MSE of δ^λ may be written as follows:

$$\begin{aligned}\text{MSE}(\delta^\lambda) = \text{MSE}(\hat{\boldsymbol{\theta}}) &+ \lambda^2 \sum_{k=1}^{p}\{E((\hat{\theta}_k - \hat{\theta}_k^{\text{Target}})^2)\} \\ &- 2\lambda \sum_{k=1}^{p}\{\text{Var}(\hat{\theta}_k) - \text{Cov}(\hat{\theta}_k, \hat{\theta}_k^{\text{Target}}) \\ &\qquad + \text{Bias}(\hat{\theta}_k)\; E(\hat{\theta}_k - \hat{\theta}_k^{\text{Target}})\} \\ =:\; c &+ \lambda^2\, b - 2\lambda\, a.\end{aligned} \tag{A.3}$$

Hence, the MSE risk curve has the shape of a parabola whose parameters a, b, and c are completely determined by only the first two distributional moments of $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}^{\text{Target}}$. This allows a number of further insights concerning the shrinkage rule Eq. A.2:

- The risk improvement of δ^λ compared to MSE of the unregularized estimate δ^0 is determined only by a and b (note that $\text{MSE}(\delta^0) = c$).
- *Any* value of λ in the range between 0 and $2\frac{a}{b}$ leads to a decrease in MSE.
- The optimal shrinkage intensity that results in overall minimum MSE is given by the simple formula
$$\lambda^\star = \frac{a}{b}. \tag{A.4}$$
- In this case the savings relative to the unshrunken estimate amount to $\text{MSE}(\hat{\boldsymbol{\theta}}) - \text{MSE}(\delta^{\lambda^\star}) = \frac{a^2}{b}$.
- The factor b, which measures the misspecification of target and estimate, plays the role of a precision for λ.

Further discussion and interpretation of Eq. A.4 may be found in Schäfer and Strimmer (2005). We only note here that special versions of the rule $\lambda^\star = \frac{a}{b}$ are well known, see, e.g., Ledoit and Wolf (2003) who describe the multivariate case but require an unbiased $\hat{\boldsymbol{\theta}}$, or Thompson (1968) who considers only the univariate case and a non-stochastic target and also restricts to $\text{Bias}(\hat{\theta}) = 0$.

A.2.2 Construction of Shrinkage Estimator

In actual application of the shrinkage rule (Eq. A.2) the pooling parameter λ needs to be estimated from the data. Inevitably, this leads to an *increase* of the total risk of the resulting shrinkage estimator. However, it is a classic result by Stein that the cost of estimating the shrinkage intensity is already (and always!) offset by the savings in total risk when the dimension p is larger than three (e.g. Gruber, 1998).

One straightforward way to estimate the optimal $\lambda^\star$ is to replace the variances and covariances in Eq. A.3 by their unbiased empirical counterparts, i.e. $\hat{\lambda}^\star = \frac{\hat{a}}{\hat{b}}$. Alternatively, an unbiased estimate for the whole fraction $\frac{a}{b}$ may be sought – but this will only be possible in some special cases. We would also like to point out that we do *not* recommend the suggestion of Thompson (1968) who employ a biased estimate for the denominator (b) to Eq. A.4 – this will lead to a potentially very inaccurate shrinkage estimator.

Despite its simplicity, rule Eq. A.2 together with an estimated version of Eq. A.4 provides instant access to several classic shrinkage estimators, and offers a simple and unified framework for their derivation.

For instance, consider the old problem of Stein (1956, 1981) of inferring the mean of a p-dimensional multivariate normal distribution with unit-diagonal covariance matrix from a single ($n = 1$) vector-valued observation – clearly an extreme example of the "small n, large p" setting. In this case the maximum-likelihood estimate equals the vector of observations, i.e. $\hat{\theta}_k^{\mathrm{ML}} = x_k$. However, shrinkage estimators with improved efficiency over the ML estimator are easily constructed. With the covariance being the identity matrix ($\mathrm{Var}(x_k) = 1$ and $\mathrm{Cov}(x_k, x_l) = 0$) and the target being set to zero ($\hat{\boldsymbol{\theta}}^{\mathrm{Target}} = 0$) one finds $a = p$ and $\hat{b} = \sum_{k=1}^{p} x_k^2$ which results in the shrinkage estimator

$$\hat{\theta}_k^{\mathrm{JS}} = \left(1 - \frac{p}{\sum_{k=1}^{p} x_k^2}\right) x_k. \tag{A.5}$$

If we follow Lindley and Smith (1972) and shrink instead towards the mean across dimensions $\bar{x} = \frac{1}{p}\sum_{k=1}^{p} x_k$ we get $a = p - 1$ and $\hat{b} = \sum_{k=1}^{p}(x_k - \bar{x})^2$ and obtain

$$\hat{\theta}_k^{\mathrm{EM}} = \bar{x} + \left(1 - \frac{p-1}{\sum_{k=1}^{p}(x_k - \bar{x})^2}\right)(x_k - \bar{x}) \tag{A.6}$$

It is noteworthy that these are *not* the original Stein estimators given in James and Stein (1961) and Efron and Morris (1973) but instead are exactly the shrinkage estimators of Stigler (1990) derived using a regression approach. We point out that the Stigler and our versions have the advantage that they are applicable also for $p = 1$ and $p = 2$.

A.2.3 Positive Part Estimator and Component Risk Protection by Limited Translation

The efficiency of the above shrinkage estimator can be further improved by two simple measures.

Firstly, by truncating the estimated $\hat{\lambda}$ at one,

$$\delta^{\hat{\lambda}+} = \delta^0 - \min(1, \hat{\lambda})\boldsymbol{\Delta}, \tag{A.7}$$

which results in the so-called positive part James-Stein estimator that dominates the unrestricted shrinkage estimator of Eq. A.2 in terms of statistical efficiency (Barachnik, 1970).

Secondly, by restricting the translation allowed for individual components. The original James-Stein procedure is geared, in the terminology of Efron and Morris (1975), towards producing estimators with good *ensemble* risk properties. This means that it aims at minimizing the total risk accumulated over all parameters. However, in some instances this may occur at the expense of individual parameters whose risks may even increase (!). Therefore, in Stein estimation (and indeed also in hierarchical Bayes estimation) individual components of a parameter vector need to be protected against too much shrinkage.

"Limited translation" (Efron and Morris, 1972, 1975) is a simple way to construct estimators that exhibit both good ensemble risk as well as favorable component risk properties. One example of a protected shrinkage rule is

$$\delta_k^{\hat{\lambda}+,M} = \delta_k^0 - \min(1, \hat{\lambda}) \min(1, \frac{M}{|\Delta_k|})\Delta_k, \tag{A.8}$$

which ensures that we always have $|\delta_k^{\hat{\lambda}+,M} - \delta_k^0| \leq M$, where M is a cutoff parameter chosen by the user. A convenient selection of M is, e.g., the 99 percent quantile of the distribution of the absolute values $|\Delta_k|$ of the components of the shrinkage vector $\boldsymbol{\Delta}$. In the terminology of Efron and Morris (1972), the term $\min(1, \frac{M}{|\Delta_k|})$ constitutes the *relevance function* that determines the degree to which any particular component is affected by the ensemble-wide shrinkage.

Finally, we point out an interesting connection with soft thresholding, as the above limited translation shrinkage rule may also be written as

$$\delta_k^{\hat{\lambda}+,M} = \delta_k^{\hat{\lambda}+} + \min(1, \hat{\lambda})(|\Delta_k| - M)_+ \mathrm{sgn}(\Delta_k), \tag{A.9}$$

where the subscript "+" denotes truncation at zero.

A.2.4 Further Remarks

In order to complete the discussion of analytic James-Stein shrinkage estimators we would like to remark on the following additional points:

- It is interesting to note that many empirical Bayes estimators can be put into the form of Eq. A.2, (e.g. Gruber, 1998). Note that using Eq. A.3 allows to derive these estimators without first going through the full Bayesian formalism!
- Using the above equation leads to an almost automatic procedure for shrinkage estimation.

- The construction of the estimator assumes at no point a normal or any other distribution.

- Note that it is possible to allow for multiple shrinkage intensities. For instance, if the model parameters fall into two natural groups, each could have its own target and its own associated shrinkage intensity. In the extreme case each parameter could have its own λ.

A.3 The "Shrinkage t" Statistic

A.3.1 Shrinkage Estimation of Variance Vector

Within the above framework for distribution-free shrinkage it is straightforward to construct an efficient estimator of gene-specific variances.

From given data with p variables (genes) we first compute the usual unbiased empirical variances $v_1^2, \ldots, v_p^2$. These provide the components for the unregularized vector estimate $\hat{\boldsymbol{\theta}}$ of Eq. A.1. Subsequently, we choose a suitable shrinkage target. For this we suggest using the median value of all v_k. In the exploration of possible other targets we considered also shrinking against zero and towards the mean of the empirical variances. However, these two alternatives turned out to be either less efficient (zero target) or less robust (mean target) than shrinking towards the median.

Following the recipe outlined above, we immediately obtain the shrinkage estimator

$$v_k^\star = \hat{\lambda}^\star v_{\text{median}} + (1 - \hat{\lambda}^\star) v_k \tag{A.10}$$

with optimal estimated pooling parameter

$$\hat{\lambda}^\star = \min\left(1, \frac{\sum_{k=1}^p \widehat{\text{Var}}(v_k)}{\sum_{k=1}^p (v_k - v_{\text{median}})^2}\right), \tag{A.11}$$

Note that in this formula we have used the approximation $\text{Cov}(v_k, v_{\text{median}}) \approx 0$.

Eq. A.11 has an intuitive interpretation. If the empirical variances v_k can be reliably determined from the data, and consequently exhibit only a small variance themselves, there will be little shrinkage, whereas if $\widehat{\text{Var}}(v_k)$ is comparatively large pooling across genes will take place. Furthermore, the denominator of Eq. A.11 is an estimate of the misspecification between the target and the v_k. Hence, if the target is incorrectly chosen then no shrinkage will take place either.

The computation of a sample version of $\widehat{\text{Var}}(v_k)$ is straightforward. Defining $\bar{x}_k = \frac{1}{n}\sum_{i=1}^n x_{ik}$, $w_{ik} = (x_{ik} - \bar{x}_k)^2$, and $\bar{w}_k = \frac{1}{n}\sum_{i=1}^n w_{ik}$, we have $v_k = \frac{n}{n-1}\bar{w}_k$ and $\widehat{\text{Var}}(v_k) = \frac{n}{(n-1)^3}\sum_{i=1}^n (w_{ik} - \bar{w}_k)^2$. A similar formula can be derived for the variance of the entries of the empirical covariance matrix (cf. Schäfer and Strimmer (2005)).

A.3.2 Construction of "Shrinkage t" Statistic

The "shrinkage t" statistic considered in the following is obtained by plugging the above shrinkage variance estimate (Eq. A.10 and Eq. A.11) into the ordinary t statistic. With the sample sizes in groups 1 and 2 denoted as n_1 and n_2 the shrinkage t statistic is given by

$$t_k^\star = \frac{\bar{x}_{k1} - \bar{x}_{k2}}{\sqrt{\frac{v_{k1}^\star}{n_1} + \frac{v_{k2}^\star}{n_2}}}. \tag{A.12}$$

We consider two variants of this statistic, one where variances are estimated separately in each group (hence with two different shrinkage intensities), and the other one using a pooled estimate (i.e. with one common shrinkage factor).

Note that the shrinkage t statistic essentially provides a compromise between the standard t statistic (to which it reduces for $\lambda = 0$) and the fold change or difference of means statistic ($\lambda = 1$).

A.3.3 Other Regularized t Statistics

Closely related to the shrinkage t statistic are in particular two approaches, "moderated t" (Smyth, 2004) and the Stein-type procedure by Cui et al. (2005). The essential feature characteristic for both of these methods is, again, variance shrinkage, albeit done in a different fashion compared to shrinkage t:

1. The moderated t method assumes a scale-inverse-chi-square distribution as distribution for the variances across genes. The corresponding parameters are estimated by empirical Bayes, and the resulting variance estimates are plugged into the t-statistic.

2. The variance shrinking procedure of Cui et al. (2005) is essentially the classic Stein estimator on the log-scale, complemented by a bias correction and by an estimation of the variance factor (numerator in the Stein formula) by simulation from a chi-square distribution whose degrees of freedom depends on the sample size. The t statistic resulting from using this kind of variance shrinkage will be called "Cui et al. t" in this paper.

Therefore, both the moderated t and the Cui et al. t rely on some form of distributional assumption, whereas the shrinkage t statistic is derived without any such consideration.

A.4 Results

A.4.1 Assessment of Quality of Gene Ranking

In this section we describe the results from computer simulations and analysis of experimental gene expression data that we conducted to assess the quality of gene ranking provided by the shrinkage t statistic in comparison to other competing scores.

Setup of Simulations

In the simulations we followed closely the setup specified in Smyth (2004):

- Variances across genes were assumed to follow a scale-inverse-chi-square distribution Scale-inv-$\chi^2(d_0, s_0^2)$ with $s_0^2 = 4$ and three different settings for the degrees of freedom d_0: highly similar variances across genes ($d_0 = 1000$), balanced variances ($d_0 = 4$), and different variances across genes ($d_0 = 1$).

- In total 2,000 genes were considered, 100 of which were randomly assigned to be differentially expressed.

- The differences in group means for the 100 differentially expressed genes were determined by drawing from a Normal distribution with mean zero and the gene-specific variance, whereas for the non-differentially expressed genes it was set to zero.

- Finally, synthetic data matrices were obtained by sampling for each gene and separately for the control and case groups three independent observations from a Normal distribution with the respective gene-specific variances and means.

These data formed the basis for computing various gene ranking scores. Specifically, we compared the following statistics: fold change, ordinary t, moderated t (Smyth, 2004), Cui et al. t statistic (i.e. the unequal variance t statistic regularized by using the variances estimated by the method of Cui et al. (2005)), Efron's 90% rule (Efron et al., 2001), Wu's improved SAM statistic (Wu, 2005), and the shrinkage t statistic (with both equal and unequal variances). As reference we also included random ordering in the analysis. For these different ways of producing rankings we computed false positives (FP), true positives (TP), false negatives (FN), and true negatives (TN) for all possible cut-offs in the gene list (1-2000).

This procedure was repeated 500 times for each test statistic and variance scenario, to obtain estimates of the true discovery rates $E(\frac{TP}{TP+FP})$ and ROC curves describing the dependence of sensitivity $E(\frac{TP}{TP+FN})$ and specificity $E(\frac{TN}{TN+FP})$.

Experimental Data Sets

In addition to the simulations we computed the gene ranking for three experimental case-control data sets with known differentially expressed genes.

The first data set studied is the well-known Affymetrix spike-in study that contains 12,626 genes, 12 replicates in each group, and 16 known differentially expressed genes (Cope et al., 2004).

The second investigated data is a subset of the "golden spike" Affymetrix experiment of Choe et al. (2005). From the original data we removed the 2,535 probe sets for spike-ins with ratio 1:1, leaving in total 11,475 genes with 3 replicates per group, and 1,331 known differentially expressed genes. We note that excluding the 1:1 spike-ins is important as these comprise a severe experimental artifact (Irizarry et al., 2006). Both the Choe et al.

(2005) data and the Affymetrix spike-in data were calibrated and normalized using the default methods of the "affy" R package (Gautier et al., 2004).

The third data set is from the HIV-1 infection study of van 't Wout et al. (2003). It contains 4 replicates per group, and 13 of the 4,608 genes have been experimentally confirmed to be differentially expressed.

For reproducibility, these three experimental test data sets are available for download from `http://strimmerlab.org/data.html` exactly in the form used in this article, including all preprocessing.

Performance of Gene Ranking Statistics

The results from simulations and data analysis are summarized in Fig. A.1 and Fig. A.2. In each figure the first row shows the fraction of correctly identified differentially expressed genes in relation to the number of included genes (i.e. the true discovery rate, or positive predictive value), whereas the second row depicts the corresponding receiver-operator characteristic (ROC).

If the variances are highly similar across genes (Fig. A.1, first column) the best methods are fold change, moderated t, Efron t, Cui et al. t, and shrinkage t, all of which provide in this case similarly accurate rankings. Only the ordinary t statistics and Wu's improved SAM statistic are not efficient in this setting. If variances are balanced (Fig. A.1, second column), the best rankings are given by moderated t, Efron t and shrinkage t. The gene ranking accuracy of fold-change is dropping to that of the standard t statistic. If the variances are highly different (Fig. A.1, column 3), fold change ranking is close to random, and Efron's 90% percent rule also becomes inefficient. Only moderated t and shrinkage t are offering optimal gene rankings in this setting. The Cui et al. t and the ordinary t test produce similar and the second best rankings in this case.

In the analysis of the Affymetrix spike-in data the methods with largest true discovery rates independent of the chosen cutoff are the shrinkage t statistic (unequal variance), Efron t, shrinkage t statistic (equal variance), moderated t, and Cui et al. t. For the Choe et al. (2005) data the shrinkage t statistic shows the best performance along with moderated t, Efron t, Cui et al. t, and fold change, whereas the ordinary t statistic and the improved SAM statistic don't perform well. Thus, this data set resembles the situation in the simulations where all variances were highly similar. Finally, for the van 't Wout et al. (2003) study all methods except for fold-change provide optimal rankings, with the Efron t statistic being slightly less accurate than the remainder of the methods.

This generally confirms earlier findings presented in Smyth (2004). We emphasize in addition the following points:

- The ordinary t statistic shows average though never optimal performance regardless of the variance structure across genes.

- Using fold change is only a good idea if variances are all fairly similar; the same is true for Efron t statistic.

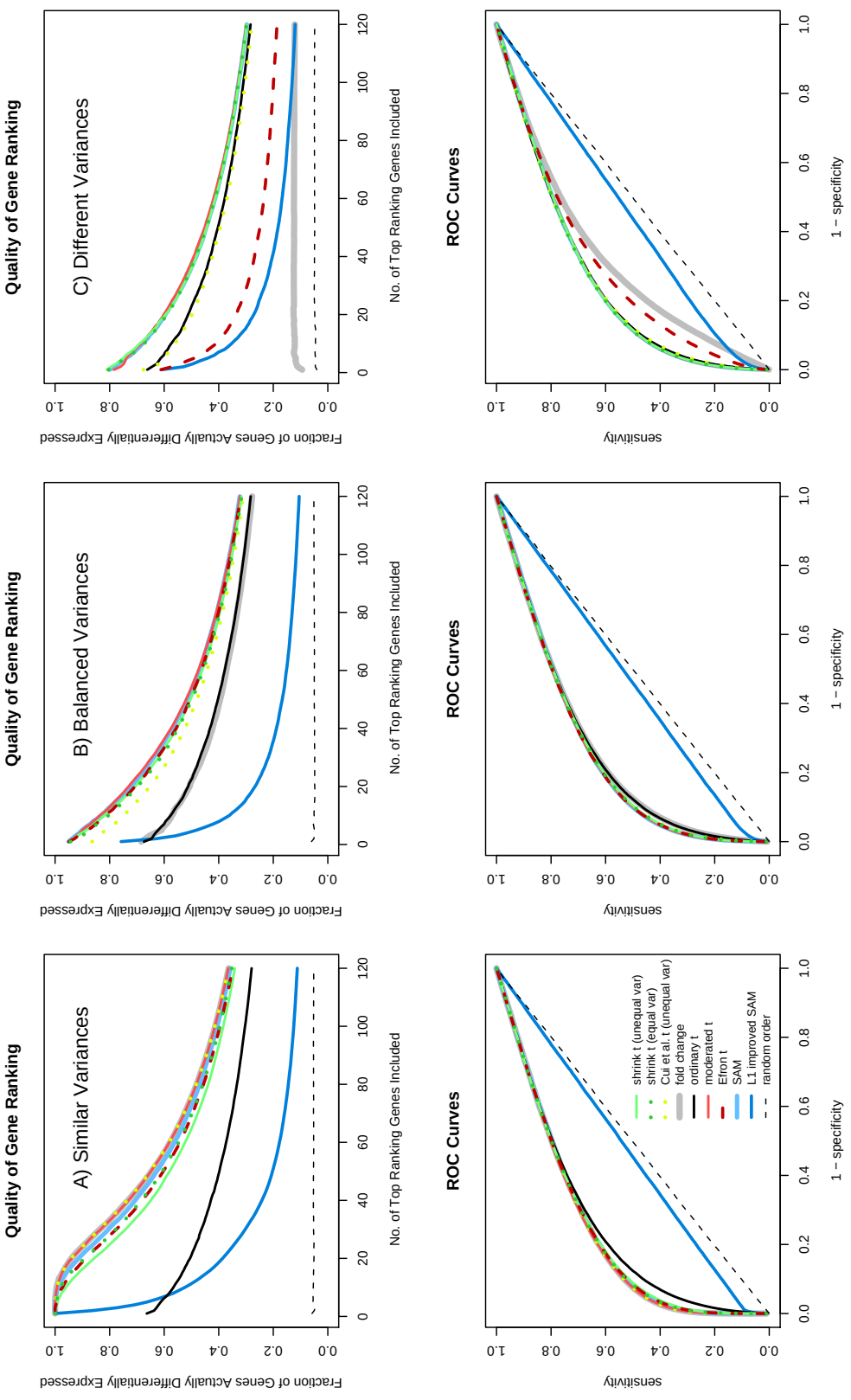

Figure A.1: True discovery rates and ROC curves computed for simulated data under three different scenarios for the distribution of variances across genes. See main text for details of simulations and analysis procedures.

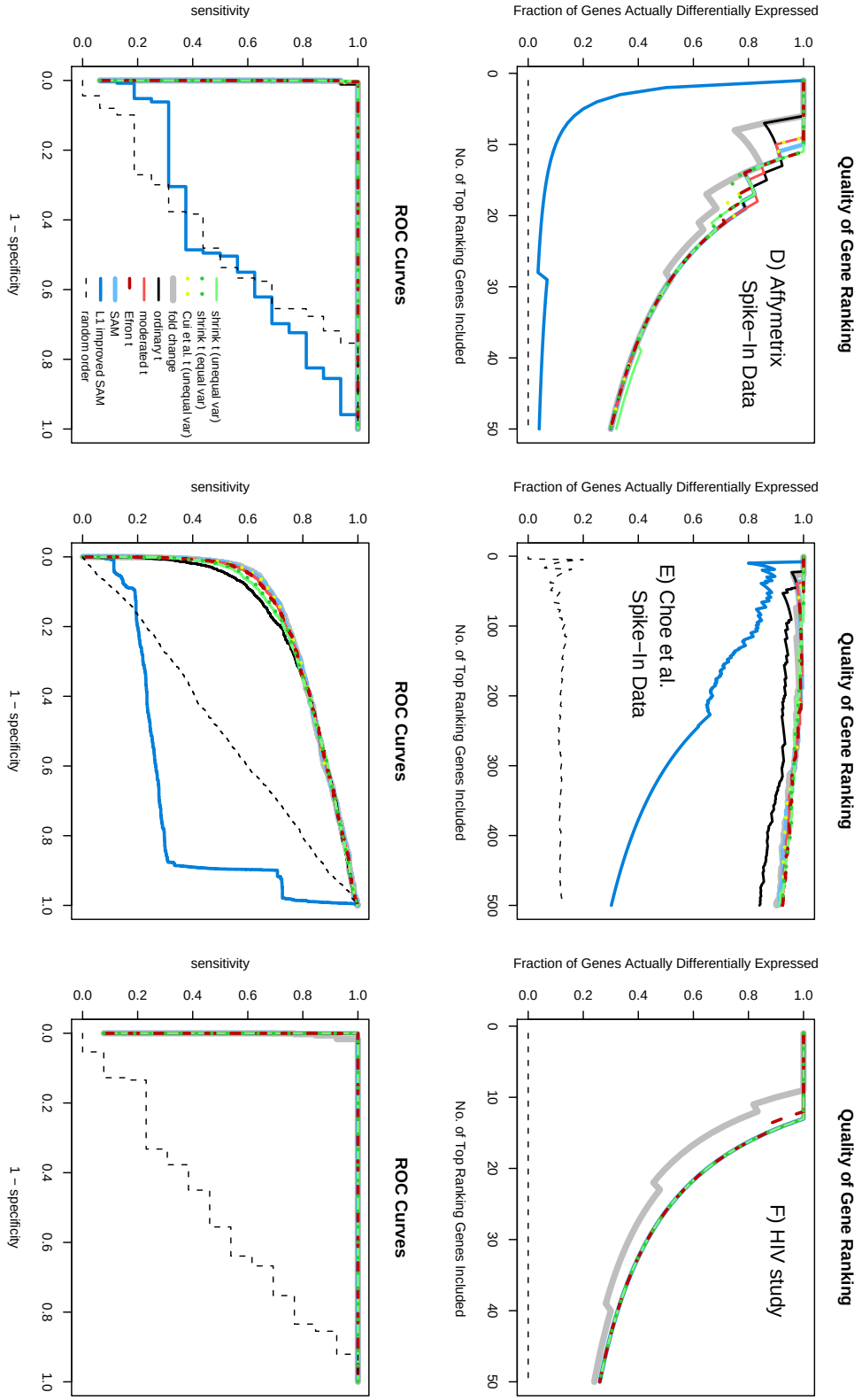

Figure A.2: True discovery rates and ROC curves for the three investigated experimental data sets.

- The improved SAM statistic by Wu (2005) generally provides very poor rankings of genes. This is due to the fact that most genes with differential expression are assigned a zero score, and hence are ordered randomly (i.e. in input order).
- ROC curves appear to be only of limited help for assessing performance. Instead, we suggest relying on quality of ranking plots based on estimated true discovery rates.
- The shrinkage t and the moderated t statistics are the only two methods that perform optimally in all three simulation settings. However, the Cui et al. t statistic is nearly as good, showing only a slightly decreased accuracy when variances are strongly heterogeneous.
- Both moderated t and shrinkage t produce accurate rankings also for the three experimental data, perhaps with a small edge for shrinkage t in the Affymetrix spike-in study.
- The shrinkage t statistic with unequal variances performs nearly as well as shrinkage t with equal variances, even though the former has only half the sample size available for estimation of variances.

We also note that the moderated t statistic is based exactly on the model employed in the simulations (i.e. scale-inverse-chi-square distribution for the variance). The Cui et al. t statistic also incorporates prior knowledge on the distribution of variances across genes in its simulation step employing a chi-square distribution.

Therefore, it is easy to understand why moderated t and the Cui et al. t statistic perform well. On the other hand, we point out that it is quite remarkable that shrinkage t, a simple analytic statistic not specifically tailored to any particular distributional setting, can fully match the performance of the moderated t approach.

A.5 Discussion

In this paper we have introduced a novel gene ranking statistic for genomic case control studies. This method is based on a James-Stein-type shrinkage approach that, unlike its Bayesian or empirical Bayes cousins, is fully analytic and does not rely on explicit priors or other distributional assumptions. Hence, this approach is potentially more flexible, for instance when variance scenarios differ (e.g. Gelman, 2006), the underlying models are misspecified, or when variances are unequal. Most importantly, the proposed method provides highly accurate gene rankings both for simulated and real data, on par with much more complicated models, but without relying on computational expensive procedures such as MCMC or optimization.

From our simulations it is also interesting to learn that there seems to exist an optimality limit with regard to producing accurate rankings. In our comparative evaluation the moderated t statistic and the shrinkage t statistic were the only two methods that achieved that limit for all considered scenarios.

In this paper, we haven't raised at all the issue of (multiple) testing needed to determine an appropriate cut-off value. This is typically done by controlling the false discovery rate. Our preferred tools for this task are mixture models (e.g., Sapir and Churchill, 2000; Dean and Raftery, 2005) and the "local fdr" approach (see, e.g. Efron, 2005). The latter procedure has the advantage of being adaptive with regard to the null hypothesis. This means that it will automatically take account of correlation among the genes, and also accommodate for the decreased variance of the null-distribution of the shrinkage t statistic, which by construction is smaller than that of the standard t statistic. However, with Fig. A.1 in mind we caution that it often is not possible to guarantee a prescribed false discovery rate even in optimal circumstances - as this crucially depends on the capability of the underlying statistic to produce an accurate ranking!

In summary, with few exceptions (e.g., Cui et al., 2005; Schäfer and Strimmer, 2005) James-Stein-type estimation appears to have been somewhat overlooked in the recent efforts for analyzing high-dimensional systems (it is not mentioned in the reference text by Hastie et al. (2001), for instance). In this respect, the shrinkage t approach demonstrates that statistics derived in this fashion may indeed compare very favorable to penalized ML or Bayesian methods. Indeed, our proposed variance shrinkage procedure may be useful not only in simple t test situations but also in more general ANOVA-type analyses (Smyth, 2004; Cui et al., 2005).

Computer Implementation and Availability

All statistical procedures described have been implemented in computer programs that are available under the terms of the GNU General Public License.

The "shrinkage t" statistic is implemented in the R package "st" which is available from the CRAN archive (`http://cran.r-project.org`) and from web page `http://strimmerlab.org/software/st/`. This package also contains wrapper functions for a number of other regularized t statistics.

The shrinkage variance estimator of Eq. A.10 and Eq. A.11 is contained in the R package "corpcor" that is available from `http://strimmerlab.org/software/corpcor/` and also from CRAN.

Acknowledgments

This research was supported by an Emmy Noether excellence grant from the Deutsche Forschungsgemeinschaft. We thank Gary Churchill and an anonymous referee for valuable comments.

Bibliography

Baldi, P. and Long, A. D. (2001). A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics*, 17:509–519.

Barachnik, A. J. (1970). A family of minimax estimators of the mean of a multivariate normal distribution. *Ann. Math. Statist.*, 41:642–645.

Choe, S. E., Boutros, M., Michelson, A. M., Church, G. M., and Halfon, M. S. (2005). Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control data set. *Genome Biology*, 6:R16.

Cope, L. M., Irizarray, R. A., Jaffee, H. A., Wu, Z., and Speed, T. P. (2004). A benchmark for Affymetrix GeneChip expression measures. *Bioinformatics*, 20:323–331.

Cui, X. and Churchill, G. A. (2003). Statistical test for differential expression in cDNA microarray experiments. *Genome Biology*, 4:R210.

Cui, X., Hwang, J. T. G., Qiu, J., Blades, N. J., and Churchill, G. A. (2005). Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics*, 6:59–75.

Dean, N. and Raftery, A. E. (2005). Normal uniform mixture differential gene expression detection for cDNA microarrays. *Genome Biology*, 6:173.

Delmar, P., Robin, S., Tronik-Le Roux, D., and Daudin, J. J. (2005). Mixture model on the variance for the differential analysis of gene expression data. *Appl. Statist.*, 54:31–50.

Efron, B. (2005). Local false discovery rates. Technical Report 2005-20B/234, Dept. of Statistics, Stanford University.

Efron, B. and Morris, C. (1972). Limiting the risk of Bayes and empirical Bayes estimators – part II: The empirical Bayes case. *Journal of the American Statistical Association*, 67:130–139.

Efron, B. and Morris, C. N. (1973). Stein's estimation rule and its competitors – an empirical Bayes approach. *J. Amer. Statist. Assoc.*, 68:117–130.

Efron, B. and Morris, C. N. (1975). Data analysis using Stein's estimator and its generalizations. *J. Amer. Statist. Assoc.*, 70:311–319.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.*, 96:1151–1160.

Fox, R. J. and Dimmic, M. W. (2006). A two-sample Bayesian t-test for microarray data. *BMC Bioinformatics*, 7:126.

Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. (2004). Affy - analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, 20:307–315.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3):515–533.

Gruber, M. H. J. (1998). *Improving Efficiency By Shrinkage.* Marcel Dekker, Inc., New York.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning.* Springer Verlag, New York.

Ideker, T., Thorsson, V., Seigel, A. F., and Hood, L. E. (2000). Testing for differentially expressed genes by maximum likelihood analysis of microarray data. *J. Comp. Biol.*, 7:805–817.

Irizarry, R. A., Cope, L., and Wu, Z. (2006). Feature-level exploration of a published control data set. *Genome Biology*, 7:404.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In Neyman, J., editor, *Proc. Fourth Berkeley Symp. Math. Statist. Probab.*, volume 1, pages 361–379, Berkeley. Univ. California Press.

Ledoit, O. and Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *J. Empir. Finance*, 10:603–621.

Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for the linear model. *J. R. Statist. Soc. B*, 34:1–72.

Lönnstedt, I. and Speed, T. (2002). Replicated microarray data. *Statistica Sinica*, 12:31–46.

Newton, M. A., Kendziorski, C. M., Richmond, C. S., Blattner, F. R., and Tsui, K. W. (2001). On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *J. Comp. Biol.*, 8:37–52.

Newton, M. A., Noueiry, A., Sarkar, D., and Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, 5:155–176.

Sapir, M. and Churchill, G. A. (2000). Estimating the posterior probability of differential gene expression from microarray data. Poster, Jackson Laboratory, Bar Harbor.

Schäfer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4(1):32.

Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statist. Appl. Genet. Mol. Biol.*, 3:3.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In Neyman, J., editor, *Proc. Third Berkeley Symp. Math. Statist. Probab.*, volume 1, pages 197–206, Berkeley. Univ. California Press.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.*, 6:1135–1151.

Stigler, S. M. (1990). A Galtonian perspective on shrinkage estimators. *Statistical Science*, 5:147–155.

Strimmer, K. (2003). Modeling gene expression measurement error: a quasi-likelihood approach. *BMC Bioinformatics*, 4:10.

Thompson, J. R. (1968). Some shrinkage techniques for estimating the mean. *J. Amer. Statist. Assoc.*, 63(321):113–122.

Tusher, V., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA*, 98:5116–5121.

van 't Wout, A. B., Lehrman, G. K., Mikheeva, S. A., O'Keeffe, G. C., Katze, M. G., Bumgarner, R. E., Geiss, G. K., and Mullins, J. I. (2003). Cellular gene expression upon human immunodeficiency virus type 1 infection of CD4(+)-T-cell lines. *J Virol*, 77(2):1392–1402.

Wright, G. W. and Simon, R. M. (2003). A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics*, 19:2448–2455.

Wu, B. (2005). Differential gene expression detection using penalized linear regression models: the improved SAM statistic. *Bioinformatics*, 21:1565–1571.

Yi-Shi Shao, P. and Strawderman, W. E. (1994). Improving on the James-Stein positive-part estimator. *Ann. Statist.*, 22:1517–1538.

Article B

Inferring Gene Dependency Networks from Genomic Longitudinal Data: A Functional Data Approach

Published in REVSTAT *(Volume 4, Number 1, March 2006)*

Authors: Rainer Opgen-Rhein *and* Korbinian Strimmer

Abstract:

- A key aim of systems biology is to unravel the regulatory interactions among genes and gene products in a cell. Here we investigate a graphical model that treats the observed gene expression over time as realizations of random curves. This approach is centered around an estimator of dynamical pairwise correlation that takes account of the functional nature of the observed data. This allows to extend the graphical Gaussian modeling framework from i.i.d. data to analyze longitudinal genomic data. The new method is illustrated by analyzing highly replicated data from a genome experiment concerning the expression response of human T-cells to PMA and ionomicin treatment.

Key-Words:

- *Graphical model, longitudinal data, dynamical correlation, gene dependency networks*

AMS Subject Classification:

- 37N25, 62M10, 92B15, 92D10.

B.1 Introduction

The identification of networked genetic interdependencies that form the basis of cellular regulation is one of the key issues in systems biology. Consequently, many authors have investigated statistical approaches such as graphical models to estimate genetic networks

from high-throughput data (e.g., Hartemink et al., 2002; Friedman, 2004; Schäfer and Strimmer, 2005a).

A graphical model is a representation of stochastic conditional dependencies between the investigated variables. Among the simplest graphical models is the class of graphical Gaussian models (GGMs) – see, e.g., Whittaker (1990). In this framework gene network may be constructed as follows. First, a positive definite and well-conditioned estimate $\boldsymbol{R} = (r_{kl})$ of the linear correlation matrix $\boldsymbol{P} = (\rho_{kl})$ is inferred from the data. Second, the standardized inverse of this matrix gives an estimate $\tilde{\boldsymbol{R}} = (\tilde{r}_{kl})$ of the *partial* correlations $\tilde{\boldsymbol{P}} = (\tilde{\rho}_{kl})$. The strength of these coefficients indicate the presence or absence of a direct association between each pair of genes. For large sample size computation of covariances and GGM selection can be conducted using classical estimation and testing theory as outlined in Whittaker (1990). However, the small sample size relative to the large number of genes typically considered in genome experiments requires the additional application of shrinkage and other regularization techniques (Dobra et al., 2004; Schäfer and Strimmer, 2005b).

A drawback shared by the GGM approach and other graphical models such as Bayesian networks is that these methods rely on the assumption of identically and independently distributed (i.i.d.) data. However, an increasing proportion of microarray expression experiments are concerned with *longitudinal* measurements of mRNA and protein concentrations. For instance, stress response and cell cycle experiments by design produce time course data. A further characteristic of these data is that the time points at which the experiments are conducted are almost always not equidistant but irregularly spaced.

In order to avoid these issues, in this paper we investigate GGM network inference from the perspective of functional data analysis (Ramsay and Silverman, 2005). Specifically, we describe a graphical model that treats the observed gene expression over time as realizations of random curves, rather than to describe the individual time points separately. This approach is based on the notion of *dynamical correlation* which provides a similarity score for pairs of groups of randomly sampled curves. Subsequently, it allows computation of partial dynamical correlations and the identification of the associated network structure.

The remainder of the paper is organized as follows. In the next section we summarize the basic notation for functional data analysis and also introduce the functional inner product. Next, we discuss the concept of dynamical correlation of which we describe two different variants, one introduced in this paper and one by Dubin and Müller (2005). Subsequently, the dynamical correlation is employed for GGM network selection. Finally, in order to compare the traditional GGM method with the present approach we reanalyze data from a human T-cell experiment with 58 genes, 10 time points, and 44 replications (Rangel et al., 2004), and compare the networks resulting from dynamical correlation with those from static correlation.

B.2 Methods

B.2.1 Setup and Notation

We consider data from a typical gene expression time course experiment. For p genes (variables) and n subjects (replications) mRNA concentrations are measured over a time interval $[A, B]$. This results in functional observations $f_{ik}(t)$ where $1 \leq i \leq n$ and $1 \leq k, l \leq p$. We assume all functions $f_{ik}(t)$ to be square-integrable so that the functional inner product

$$\langle g(t), h(t) \rangle = \frac{1}{B-A} \int_A^B g(t)h(t)dt \tag{B.1}$$

exists, where $g(t)$ and $h(t)$ are any of the observed functions. The time average of $f_{ik}(t)$ may then be conveniently expressed by $\langle f_{ik}(t), 1 \rangle$. The average over the n replicates gives the empirical mean function $\bar{f}_k(t) = \frac{1}{n} \sum_{i=1}^n f_{ik}(t)$.

In practice, however, the functions $f_{ik}(t)$ are not continuously measured but rather obtained by experiments at discrete time points t_j, with $1 \leq j \leq m$ and $A = t_1 < t_2 < \ldots < t_{m-1} < t_m = B$. Note that the time points need not be equidistant. If one assumes a linear approximation of $g(t)$ and $h(t)$ the inner product of Eq. B.1 turns into the weighted sum

$$\langle g(t), h(t) \rangle \approx \sum_{j=1}^m g(t_j)h(t_j) \frac{\delta_j + \delta_{j+1}}{2(B-A)} \tag{B.2}$$

where the $\delta_j = t_j - t_{j-1}$ are the time differences between subsequent measurements (with $\delta_1 = \delta_{m+1} = 0$).

In the random effects representation of Dubin and Müller (2005) each observed $f_{ik}(t)$ is a realization of the random function

$$f_k(t) = \mu_k(t) + \mu_{0k} + \epsilon_{0k} + \sum_{u=1}^{\infty} \epsilon_{uk}\eta_u(t), \tag{B.3}$$

where ϵ_{0k} and ϵ_{uk} are random variables with $E(\epsilon_{0k}) = 0$ and $E(\epsilon_{uk}) = 0$, $\mu_k(t)$ is the fixed time dependent mean function with zero time average $\langle \mu_k(t), 1 \rangle = 0$, $\mu_{0k} + \epsilon_{0k}$ represents the static random part and the remaining terms describe the dynamic random part. In Eq. B.3 the $\eta_u(t)$ are orthonormal basis functions with zero time average $\langle \eta_u(t), 1 \rangle = 0$.

In this notation the empirical mean function $\bar{f}_k(t)$ is an estimate of $E(f_k(t)) = \mu_k(t) + \mu_{0k}$. As $\mu_k(t)$ has time average zero we are also able to identify the two components of $E(f_k(t))$ by using $\hat{\mu}_{0k} = \langle \bar{f}_k(t), 1 \rangle$ and $\hat{\mu}_k(t) = \bar{f}_k(t) - \hat{\mu}_{0k}$.

B.2.2 Dynamical Correlation

Measuring similarity between two exactly known curves

Suppose for a moment that we have sufficient data to estimate the expression levels through time of two genes k and l *exactly*, i.e. that we know the mean functions $E(f_k(t))$ and

$E(f_l(t))$. In order to understand the functional connection between these two variables a measure of similarity between the two curves is required. Dubin and Müller (2005) suggest to introduce the notion of *dynamical correlation* with the informal proposition that "if both trajectories tend to be mostly on the same side of their time average (a constant) then the dynamical correlation is positive; if the opposite occurs, then dynamical correlation is negative".

This immediately leads to the following straightforward definition of dynamical correlation between two curves $g(t)$ and $h(t)$. First, compute the time-centered functions $g^C(t) = g(t) - \langle g(t), 1\rangle$ and $h^C(t) = h(t) - \langle h(t), 1\rangle$. Then define the variances as

$$\mathrm{Var}(g(t)) = \langle g^C(t), g^C(t)\rangle$$

and

$$\mathrm{Var}(h(t)) = \langle h^C(t), h^C(t)\rangle.$$

Finally, compute the the standardized functions $g^S(t) = g^C(t)/\sqrt{\mathrm{Var}(g(t))}$ and $h^S(t) = h^C(t)/\sqrt{\mathrm{Var}(h(t))}$, and obtain the correlation by

$$\mathrm{Cor}(g(t), h(t)) = \langle g^S(t), h^S(t)\rangle.$$

The general case including sampling error

The above definition of dynamical correlation for a single curve extends in a straightforward fashion to the case where each observed time course f_{ik} represents a noisy realization of the mean function $E(f_k)$.

In order to estimate the correlation between two variables k and l we first define the simultaneously time- *and* space-centered functions according to $f_{ik}^C(t) = f_{ik}(t) - \langle \bar{f}_k(t), 1\rangle$. Note that here the inner product is computed over the mean function $\bar{f}_k(t)$. Based on the $f_{ik}^C(t)$ an estimate of the variance of variable k is then given by

$$\widehat{\mathrm{Var}}_k = \hat{\sigma}_{kk} = s_{kk} = \frac{1}{n-1}\sum_{i=1}^{n}\langle f_{ik}^C(t), f_{ik}^C(t)\rangle. \tag{B.4}$$

This allows to compute standardized residual functions $f_{ik}^S(t) = f_{ik}^C/\sqrt{s_{kk}}$ that form the basis for the estimate of dynamical correlation

$$\widehat{\mathrm{Cor}}_{kl} = \hat{\rho}_{kl} = r_{kl} = \frac{1}{n-1}\sum_{i=1}^{n}\langle f_{ik}^S(t), f_{il}^S(t)\rangle. \tag{B.5}$$

Correspondingly, the estimated dynamical covariance between variables k and l is simply

$$\widehat{\mathrm{Cov}}_{kl} = \hat{\sigma}_{kl} = s_{kl} = r_{kl}\sqrt{s_{kk}s_{ll}}. \tag{B.6}$$

This simple estimator of dynamical correlation exhibits several attractive properties. In particular, it is a generalization of the standard correlation for cross-sectional data. Specifically, if $m = 1$ and $n > 1$ then it reduces to the usual maximum-likelihood estimator of correlation. Furthermore, it is also applicable if there is only a single realization of each time series available ($n = 1$, $m > 1$).

The Dubin-Müller definition of dynamical correlation

Another related but different definition of dynamical correlation is given by Dubin and Müller (2005). They propose to compute the standardized residual functions according to

$$f_{ik}^S(t) = q_{ik}(t)/\sqrt{\langle q_{ik}(t), q_{ik}(t)\rangle} \tag{B.7}$$

using

$$q_{ik}(t) = f_{ik}(t) - \bar{f}_{ik}(t) - \langle f_{ik}(t), 1\rangle + \langle \bar{f}_{ik}(t), 1\rangle. \tag{B.8}$$

This definition has the drawback that it is only defined if both $m > 1$ and $n > 1$. As we will exemplify below, it also produces counter-intuitive correlations.

B.2.3 Estimating Gene Association Networks Using Dynamical Correlation

The basic idea to infer a network from the pairwise dynamical correlation is to refer to the genes as the nodes and to the correlations as the connectivity strengths assigned to the edges of the network. However, we cannot use the correlations directly, because they represent only marginal dependencies and also include indirect interactions between two variables. Instead, we need to rely on the concept of *partial* correlation which describe the correlation between any two variables i and j conditioned on all the other variables. It is straightforward to compute the matrix of partial dynamical correlations $\tilde{\boldsymbol{P}} = (\tilde{\rho}_{kl})$ from the correlation coefficients $\boldsymbol{P} = (\rho_{kl})$ via the inverse relationship

$$\boldsymbol{\Omega} = \boldsymbol{P}^{-1} = (\omega_{ij}) \tag{B.9}$$

$$\tilde{\rho}_{kl} = -\frac{\omega_{kl}}{\sqrt{\omega_{kk}\omega_{ll}}} \tag{B.10}$$

(Edwards, 1995). Applying these equations to estimates $\boldsymbol{R} = (r_{kl})$ of (dynamical) correlations allows to obtain estimates $\tilde{\boldsymbol{R}} = (\tilde{r}_{kl})$ of the associated partial (dynamical) correlations.

In order to test the significance of the correlations and to decide which of the possible edges to include in the resulting gene association network statistical tests are needed. In this paper we employ the "local fdr" network search as proposed by Schäfer and Strimmer (2005a,b). The false discovery rate (fdr) is the expected proportion of false positives among the proposed edges. The local fdr is an empirical Bayes estimator of the false discovery rate proposed by Efron (2004, 2005). This method computes the posterior probability for an edge to be present or absent, and takes account of the multiplicity in the simultaneous testing of edges. The final network is obtained by visualizing all significant edges in an undirected graph.

B.3 Results

In the following section we first apply our method of computing dynamical correlation to example data to clarify our definition and to compare it with the related concept of Dubin

and Müller (2005). Subsequently, we infer the gene association network for a longitudinal gene expression data set described in Rangel et al. (2004).

B.3.1 Illustrative Example

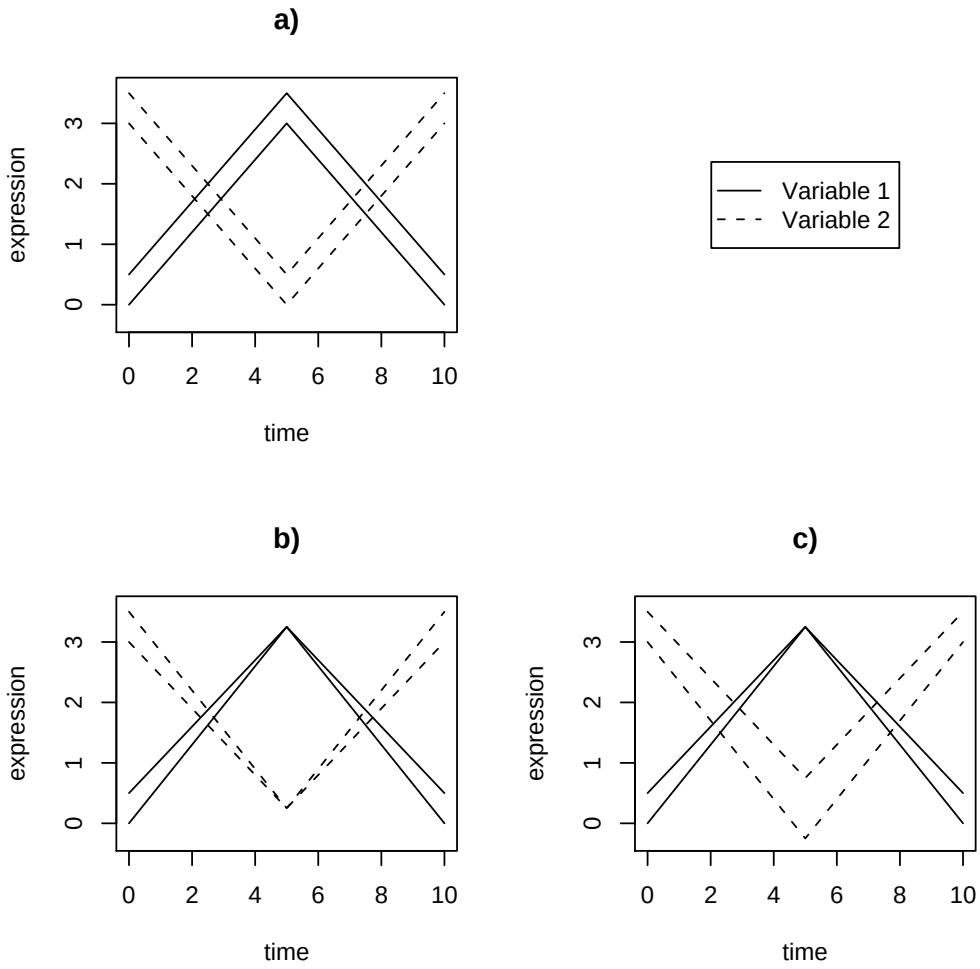

Figure B.1: Toy example to illustrate the concept of dynamical correlation between two variables ("genes"). In all three cases a), b) and c) there are two realizations ("individuals"). See main text for details, and Tab. B.1 for the underlying data.

In order to understand the concept of dynamical correlation and to illustrate the difference between our definition (Eq. B.5) and that of Dubin and Müller (2005) we first consider a set of artificial examples. These are shown in Fig. B.1 where two negatively

Data		**Variable 1**			**Variable 2**		
Time points		*0*	*5*	*10*	*0*	*5*	*10*
Fig. B.1a	*Realization 1*	0	3	0	3	0	3
	Realization 2	0.5	3.5	0.5	3.5	0.5	3.5
Fig. B.1b	*Realization 1*	0	3.25	0	3	0.25	3
	Realization 2	0.5	3.25	0.5	3.5	0.25	3.5
Fig. B.1c	*Realization 1*	0	3.25	0	3	-0.25	3
	Realization 2	0.5	3.25	0.5	3.5	0.75	3.5

Table B.1: Data points of the toy examples in Fig. B.1.

dependent variables are depicted. For instance, this may represent the case where one gene is up-regulated and the other is correspondingly down-regulated. For each gene there are two measured curves, and there are three slightly different ways in which the sampled curves relate to each other (Fig. B.1a, b, and c). The exact definition of the curves can be found in Tab. B.1. Note that the two realizations are paired, i.e. the upper lines belong to individual 1 and the lower ones to individual 2.

Intuitively, one would expect that the dynamical correlation between the two variables is strongly negative in all three cases. For our definition of dynamical correlation according to Eq. B.5 this is indeed the case: the correlations for the three examples cases Fig. B.1a, b, and c are -0.946, -0,973, and -0.947, respectively. In contrast, the dynamical correlation of Dubin and Müller (2005) behaves in a completely different fashion. For Fig. B.1a it is not defined, for case b) it is equal to +1 and for case c) it is equal to -1.

Therefore, it is easy to see that the Dubin and Müller (2005) estimator is *not* suited for detecting functional dependencies in genomic longitudinal data. This is because that estimator is geared towards detecting changes in the relative trends of the individual realizations, rather than between the common trend. However, note that this is generally not the effect one wants to identify when looking for gene interaction. In addition, the Dubin and Müller (2005) definition of dynamical correlation has the additional disadvantage over that of Eq. B.5 that it is not defined if there is only a single time course per gene available. In contrast, the above toy examples show that our definition of dynamical correlation is able to detect the main trend of positive or negative dependency between two variable, and is not susceptible to the small changes in the sampled curves.

B.3.2 Gene Expression Time Course Data

We now employ our method of estimation of the (partial) dynamical correlation to a real world example and compare it with the results of the traditional GGM method. Specifically, we reanalyzed a microarray time series data set described in detail in Rangel et al. (2004). These data characterize the response of a human T-cell line (Jirkat) to a treatment with PMA and ioconomin. After preprocessing the time course data consist of 58 genes measured

across 10 time points with 44 replications. Rangel et al. (2004) used a state space model to estimate the influence between genes and measured a genetic network by combining direct effects and indirect effects via hidden states. This approach is generally very time-consuming due to the necessity of using of the EM algorithm for optimization. A peculiarity of the Rangel et al. (2004) data is also that the measurements in the experiment were taken at unequally spaced time points, i.e. after 0, 2, 4, 6, 8, 18, 24, 32, 48, and 72 hours after treatment. This was neglected in the original state-space analysis which assumed equally spaced data. In contrast, note that the present functional data approach allows the incorporation of arbitrary time distances between subsequent measurements.

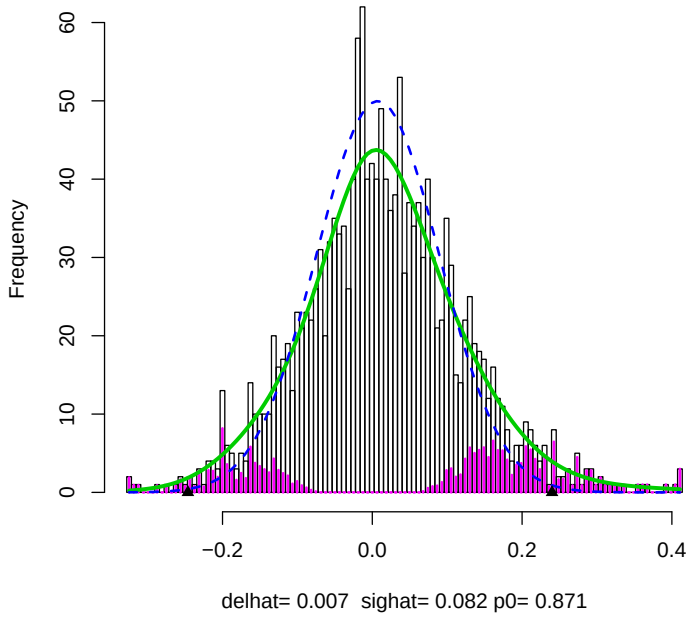

Figure B.2: Histogram of the Fisher z-transformed estimated partial dynamical correlations. Values left and right the two black triangles are considered significantly different from zero, and thus correspond to edges in a gene dependency network.

As approximation of the temporal expression of the 58 genes we used a linear spline and employed Eq. B.2 for the functional inner product. After estimating the dynamical correlations with Eq. B.5 we computed the associated partial correlation coefficients employing Eq. B.9 and Eq. B.10. Fig. B.2 shows the histogram of the estimated partial correlation coefficients after Fisher's normalizing z-transformation. Also depicted in this plot are the fitted overall distribution (fat line) and the null (dashed line) and alternative distribution (filled histogram)as estimated by the locfdr algorithm (Efron, 2004, 2005). The 0.2 local

fdr cut-off values for the partial correlations are indicated by the black triangles. As expected, the distribution of the partial correlations is centered around zero and most of the coefficients are not significant. Consequently, the resulting network is sparse and there are only 54 significant edges. The network itself is displayed in Fig. B.3b.

It is instructive to compare the genetic network inferred with dynamical correlation to the gene association network obtained by the classic GGM approach. For this analysis we ignored the dynamic aspects of the data and assumed that all measurements were taken at the same time point, which leads to 440 observations (44 replications times 10 time points) for each of the 58 genes. As this number of observations is not small in comparison to the number of the genes no regularization is needed (cf. Schäfer and Strimmer (2005b) for the opposite case). From the empirical correlation matrix we proceeded as above, obtaining estimates of partial correlation and a static GGM network. This is displayed on the left side of figure B.3. For comparison, the network estimated with dynamical correlation in shown on its right side. For clarity only the nodes which have at least one connection are displayed.

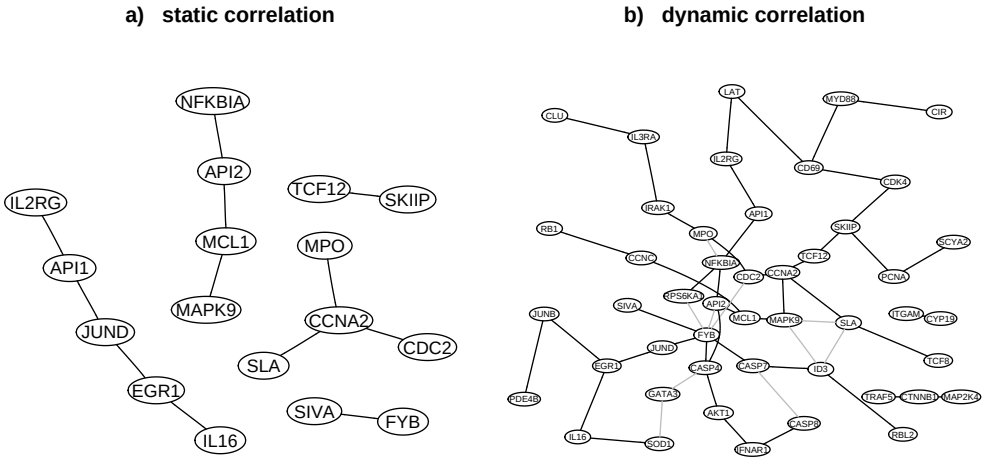

Figure B.3: Gene dependency networks inferred from human T-cell data (Rangel et al., 2004) using (a) static correlation and (b) dynamical correlation.

The network inferred with static correlation consists of 17 nodes with 12 edges, a much sparser network than the one computed with dynamical correlation. This indicates that our dynamical estimator is able to identify additional time-varying components of the interaction between the investigated genes.

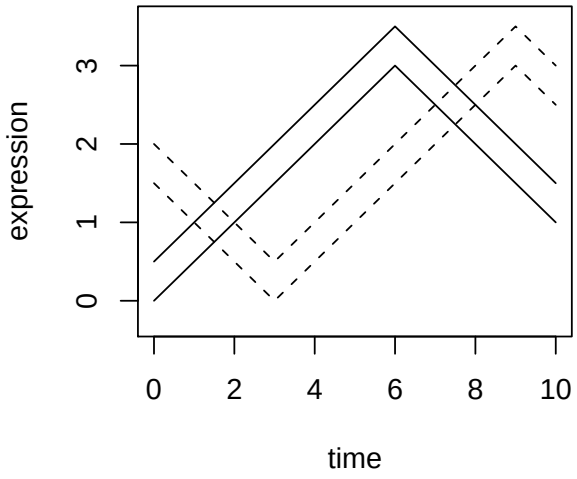

Figure B.4: Example with a fixed time lag between the two variables.

B.4 Discussion

A growing interest in genetics lies in observing and inferring the gene interactions over time. Here, we introduced a method to infer a gene dependency network from functional data. In this approach time course experiments are seen as a realization of random curves. The method described generalizes the widely used static GGM approach (see the corresponding references in (Schäfer and Strimmer, 2005a)) and is able to unravel the dependency structure of longitudinal data across the whole time series rather than at single time points. Furthermore, unlike many other time series method the functional approach does not require equally spaced measurements. In addition, our algorithm is easily implemented and computationally inexpensive (the calculation of the above gene dependency network takes only a fraction of a second).

In order to further develop our approach many extensions are conceivable. For instance, in the above analysis of human T-cells the data was highly replicated. In genomics, however, it is more typical that the sample size is very small compared to the number of genes (this is the so-called "small n, large p" paradigm). In this case, the empirical covariance is a highly inefficient estimator, and needs to be regularized (Schäfer and Strimmer, 2005b). For small n this will also be the case with our estimate of dynamical correlation (Eq. B.5). Thus, shrinkage techniques similar to those of Schäfer and Strimmer (2005b) are needed.

A further important extension is the inclusion of autoregressive aspects (Diggle et al., 2002). While our method covers the dynamical correlation through time it is not able to

account, e.g., for a time shift between any two variables. This is illustrated in Fig. B.4 which is a variation of the toy examples presented in section B.3. For this data the Dubin and Müller (2005) estimate is (again) not defined and our suggested dynamical estimator results in very small correlation close to zero, even though it is clear by inspection that the two depicted variables are strongly connected. These dependencies and the associated time shifts could be accounted for by modeling the temporal mean via a system of differential equation (or in the discrete case by some autoregressive process). We also note that for this reason we have also refrained here from a comparison of the gene association network inferred from dynamical correlation (Fig. B.3b) with the state space network presented by Rangel et al. (2004). Future work should regard for these aspects.

Acknowledgments

K.S. thanks the organizers of the "Workshop on Statistics in Genomics and Proteomics (WSGP 2005)" at Monte Estoril, Portugal (5-8 October 2005) for a stimulating meeting. This work was supported by Deutsche Forschungsgemeinschaft (DFG) Emmy-Noether research award to K.S.

Bibliography

Diggle, P., Heagerty, P., Liang, K.-Y., and Zeger, S. (2002). *Analysis of Longitudinal Data*. Oxford University Press.

Dobra, A., Hans, C., Jones, B., Nevins, J. R., Yao, G., and West, M. (2004). Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*, 90:196–212.

Dubin, J. A. and Müller, H.-G. (2005). Dynamical correlation for multivariate longitudinal data. *Journal of the American Statistical Association*, 100(471):872–881.

Edwards, D. (1995). *Introduction to Graphical Modelling*. Springer, New York.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *Journal of the American Statistical Association*, 99(465):96–104(9).

Efron, B. (2005). Local false discovery rates. Technical Report 2005-20B/234, Dept. of Statistics, Stanford University.

Friedman, N. (2004). Inferring cellular networks using probabilistic graphical models. *Science*, 303(5659):799–805.

Hartemink, A. J., Gifford, D. K., Jaakkola, T. S., and Young, R. A. (2002). Bayesian methods for elucidating genetic regulatory networks. *IEEE Intelligent Systems*, 17(2):37–43.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer, 2nd edition.

Rangel, C., Angus, J., Ghahramani, Z., Lioumi, M., Sotheran, E., Gaiba, A., Wild, D. L., and Falciani, F. (2004). Modeling t-cell activation using gene expression profiling and state-space models. *Bioinformatics*, 20(9):1361–1372.

Schäfer, J. and Strimmer, K. (2005a). An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, 21(6):754–764.

Schäfer, J. and Strimmer, K. (2005b). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4(1):32.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley, New York.

Article C

Using Regularized Dynamic Correlation to infer Gene Dependency Networks from time-series Microarray Data

Refereed Conference Proceedings of the 4th International Workshop on Computational Systems Biology, WCSB 2006 *(June 12-13, 2006, Tampere, Finland), pp. 73-76*

Authors: Rainer Opgen-Rhein *and* Korbinian Strimmer

Abstract:

- Graphical models allow to understand regulatory interactions among genes and gene products in a cell, and hence contribute to an enhanced understanding of systems biology. Here we investigate a graphical model that treats the observed gene expression over time as realizations of random curves. This approach is centered around a regularized estimator of dynamical pairwise correlation that takes account of the functional nature of the observed data. The new method is illustrated by analyzing highly replicated gene expression time series data.

C.1 Introduction

The identification of networked genetic interdependencies that form the basis of cellular regulation is one of the key issues in systems biology. Consequently, many authors have investigated statistical approaches such as graphical models to estimate genetic networks from high-throughput data (e.g., Schäfer and Strimmer, 2005a). Among the simplest graphical models is the class of graphical Gaussian models (GGMs) – see, e.g., Whittaker (1990).

A drawback shared by the GGM approach and other graphical models such as Bayesian networks is that these methods rely on the assumption of i.i.d. data. However, an in-

creasing proportion of microarray expression experiments are concerned with *longitudinal* measurements of mRNA and protein concentrations.

In order to account for this we investigate GGM network inference from the perspective of functional data analysis (FDA) (Ramsay and Silverman, 2005; Opgen-Rhein and Strimmer, 2006). Specifically, we describe a graphical model that treats the observed gene expression over time as realizations of random curves, rather than to describe the individual time points separately.

The remainder of the paper is organized as follows. In the next section following (Opgen-Rhein and Strimmer, 2006) we summarize the basic notation for functional data analysis and also introduce the functional inner product. Next, we discuss the concept of dynamical correlation and introduce a regularization technique for the "small n, large p" domain (Dobra et al., 2004). Subsequently, the dynamical correlation is employed for GGM network selection. Finally, we analyze data from a human T-cell experiment (Rangel et al., 2004).

C.2 Methods

C.2.1 Setup and Notation

We consider data from a typical gene expression time course experiment. For p genes (variables) and n subjects (replications) mRNA concentrations are measured over a time interval $[A, B]$. This results in functional observations $f_{ik}(t)$ where $1 \leq i \leq n$ and $1 \leq k, l \leq p$. We assume all functions $f_{ik}(t)$ to be square-integrable so that the functional inner product

$$\langle g(t), h(t) \rangle = \frac{1}{B-A} \int_A^B g(t)h(t)dt \tag{C.1}$$

exists, where $g(t)$ and $h(t)$ are any of the observed functions. The time average of $f_{ik}(t)$ may then be conveniently expressed by $\langle f_{ik}(t), 1 \rangle$. The average over the n replicates gives the empirical mean function $\bar{f}_k(t) = \frac{1}{n} \sum_{i=1}^{n} f_{ik}(t)$.

In practice, however, the functions $f_{ik}(t)$ are not continuously measured but rather obtained by experiments at discrete time points t_j, with $1 \leq j \leq m$ and $A = t_1 < t_2 < \ldots < t_{m-1} < t_m = B$. Note that the time points need not be equidistant. If one assumes a linear approximation of $g(t)$ and $h(t)$ the inner product of Eq. C.1 turns into the weighted sum

$$\langle g(t), h(t) \rangle \approx \sum_{j=1}^{m} g(t_j)h(t_j) \frac{\delta_j + \delta_{j+1}}{2(B-A)} \tag{C.2}$$

where the $\delta_j = t_j - t_{j-1}$ are the time differences between subsequent measurements (with $\delta_1 = \delta_{m+1} = 0$).

In the random effects representation of Dubin and Müller, 2005 Dubin and Müller

(2005) each observed $f_{ik}(t)$ is a realization of the random function

$$f_k(t) = \mu_k(t) + \mu_{0k} + \epsilon_{0k} + \sum_{u=1}^{\infty} \epsilon_{uk}\eta_u(t), \tag{C.3}$$

where ϵ_{0k} and ϵ_{uk} are random variables with $E(\epsilon_{0k}) = 0$ and $E(\epsilon_{uk}) = 0$, $\mu_k(t)$ is the fixed time dependent mean function with zero time average $\langle \mu_k(t), 1\rangle = 0$, $\mu_{0k} + \epsilon_{0k}$ represents the static random part and the remaining terms describe the dynamic random part. In Eq. C.3 the $\eta_u(t)$ are orthonormal basis functions with zero time average $\langle \eta_u(t), 1\rangle = 0$.

In this notation the empirical mean function $\bar{f}_k(t)$ is an estimate of $E(f_k(t)) = \mu_k(t) + \mu_{0k}$. As $\mu_k(t)$ has time average zero we are also able to identify the two components of $E(f_k(t))$ by using $\hat{\mu}_{0k} = \langle \bar{f}_k(t), 1\rangle$ and $\hat{\mu}_k(t) = \bar{f}_k(t) - \hat{\mu}_{0k}$.

C.2.2 Dynamical Correlation

Measuring similarity between two exactly known curves

Suppose for a moment that we have sufficient data to estimate the expression levels through time of two genes k and l *exactly*, i.e. that we know the mean functions $E(f_k(t))$ and $E(f_l(t))$. In order to understand the functional connection between these two variables a measure of similarity between the two curves is required. Dubin and Müller, 2005 Dubin and Müller (2005) suggest to introduce the notion of *dynamical correlation* with the informal proposition that "if both trajectories tend to be mostly on the same side of their time average (a constant) then the dynamical correlation is positive; if the opposite occurs, then dynamical correlation is negative".

This immediately leads to the following straightforward definition of dynamical correlation between two curves $g(t)$ and $h(t)$. First, calculate the time-centered functions $g^C(t) = g(t) - \langle g(t), 1\rangle$ and $h^C(t) = h(t) - \langle h(t), 1\rangle$. Then define the variances as

$$\mathrm{Var}(g(t)) = \langle g^C(t), g^C(t)\rangle$$

and

$$\mathrm{Var}(h(t)) = \langle h^C(t), h^C(t)\rangle.$$

Finally, compute the standardized functions

$$g^S(t) = g^C(t)/\sqrt{\mathrm{Var}(g(t))}$$

and

$$h^S(t) = h^C(t)/\sqrt{\mathrm{Var}(h(t))},$$

and obtain the correlation by

$$\mathrm{Cor}(g(t), h(t)) = \langle g^S(t), h^S(t)\rangle.$$

The general case including sampling error

The above definition of dynamical correlation for a single curve extends in a straightforward fashion to the case where each observed time course f_{ik} represents a noisy realization of the mean function $E(f_k)$.

In order to estimate the correlation between two variables k and l we first define the simultaneously time- *and* space-centered functions according to $f_{ik}^C(t) = f_{ik}(t) - \langle \bar{f}_k(t), 1 \rangle$. Note that here the inner product is computed over the mean function $\bar{f}_k(t)$. Based on the $f_{ik}^C(t)$ the empirical estimate of the variance of variable k is then given by

$$\widehat{\mathrm{Var}}_k = \hat{\sigma}_{kk} = s_{kk} = \frac{1}{n} \sum_{i=1}^{n} \langle f_{ik}^C(t), f_{ik}^C(t) \rangle. \tag{C.4}$$

This allows to compute standardized residual functions $f_{ik}^S(t) = f_{ik}^C / \sqrt{s_{kk}}$ that form the basis for the estimate of dynamical correlation

$$\widehat{\mathrm{Cor}}_{kl} = \hat{\rho}_{kl} = r_{kl} = \frac{1}{n} \sum_{i=1}^{n} \langle f_{ik}^S(t), f_{il}^S(t) \rangle. \tag{C.5}$$

Correspondingly, the estimated dynamical covariance between variables k and l is simply

$$\widehat{\mathrm{Cov}}_{kl} = \hat{\sigma}_{kl} = s_{kl} = r_{kl} \sqrt{s_{kk} s_{ll}}. \tag{C.6}$$

This simple estimator of dynamical correlation exhibits several attractive properties. In particular, it is a generalization of the standard correlation for cross-sectional data. Specifically, if $m = 1$ and $n > 1$ then it reduces to the usual maximum-likelihood estimator of correlation. Furthermore, it is also applicable if there is only a single realization of each time series available ($n = 1$, $m > 1$).

Regularization

The above definition allows the inference of correlations between sets of curves. However, if there are only a few observations for a large number of variables ("small n, large p"–problem), the unbiased empirical estimator is suboptimal in the sense that other, biased estimators may be constructed that are more efficient and exhibit higher accuracy in terms of MSE (Stein, 1956). The pivotal element in successful learning of complex models from sparse data is regularization. It is possible to achieve a better estimation of dynamic correlation by means of *shrinkage*.

In the present case, we can construct a shrinkage estimate S^* of the dynamic covariance matrix by the convex combination $S^* = \lambda S^{\mathrm{Target}} + (1 - \lambda) S$ of the unregularized estimator S and a suitable target S^{Target}. The selection of the shrinkage parameter λ will have to take place in a data-driven fashion and has to meet some requirements. For instance, given a large sample size the shrinkage intensity λ must vanish. A simple rule to estimate the

optimal shrinkage intensity can be found by minimizing the MSE risk function

$$R(\lambda) = E\left(\sum_{k=1}^{p}\sum_{l=1}^{p}(s_{kl}^{*} - s_{kl})^2\right). \tag{C.7}$$

It can be shown (Schäfer and Strimmer, 2005b) that the minimum mean squared error $R(\lambda^*)$ is achieved *exactly* and uniquely for the choice

$$\lambda^* = \frac{\sum_{k=1}^{p}\sum_{l=1}^{p}\mathrm{Var}(s_{kl}) - \mathrm{Cov}(s_{kl}, s_{kl}^{\mathrm{Target}}) + \mathrm{Bias}(s_{kl})E(s_{kl} - s_{kl}^{\mathrm{Target}})}{\sum_{k=1}^{p}\sum_{l=1}^{p} E\left[(s_{kl} - s_{kl}^{\mathrm{Target}})^2\right]}. \tag{C.8}$$

Here we choose S^{Target} to be the diagonal matrix with the variances s_{kk} on the diagonal. Defining

$$\overline{f_{kl}} = \sum_{i=1}^{n}\sum_{j=1}^{m} \underbrace{f_{ik}^{C}(t_j) f_{il}^{C}(t_j)}_{f_{ijkl}} \underbrace{\frac{\delta_j + \delta_{j+1}}{2(B-A)n}}_{w_{ij}} \tag{C.9}$$

and the sum of squared weights

$$\tau = \sum_{i=1}^{n}\sum_{j=1}^{m} w_{ij}^2 = \frac{1}{n}\sum_{j=1}^{m}\left(\frac{\delta_j + \delta_{j+1}}{2(B-A)}\right)^2, \tag{C.10}$$

the *unbiased* empirical covariance equals

$$\widehat{\mathrm{Cov}}(g(t), h(t)) = s_{kl} = \frac{1}{1-\tau}\overline{f_{kl}} \tag{C.11}$$

and find after some calculation the individual entries for

$$\widehat{\mathrm{Var}}(s_{kl}) = \frac{\tau}{(1-\tau)^3}\sum_{i=1}^{n}\sum_{j=1}^{m} w_{ij}\left(f_{ijkl} - \overline{f_{kl}}\right)^2. \tag{C.12}$$

For scaling reasons (Schäfer and Strimmer, 2005b) we apply shrinkage to the correlation matrix. The variances $\mathrm{Var}(r_{kl})$ of the empirical correlation coefficients can be estimated by applying the above formulae to the *standardized* data (f_{ik}^{S}). This leads to the sample approximation of the shrinkage intensity

$$\hat{\lambda}^* = \frac{\sum_{k\neq l}\widehat{\mathrm{Var}}(r_{kl})}{\sum_{k\neq l} r_{kl}^2}, \tag{C.13}$$

which in turns allows to calculate the matrix of shrunken dynamical correlation coefficients.

C.2.3 Estimating Gene Association Networks Using Dynamical Correlation

The basic concept behind inferring a gene dependency network from the pairwise dynamical correlation is to investigate the correlation structure. However, we cannot simply use the correlations directly, because these represent only marginal dependencies and also include indirect interactions between two variables. Instead, we need to rely on the concept of *partial* correlation which describe the correlation between any two variables i and j conditioned on all the other variables. It is straightforward to compute the matrix of partial dynamical correlations $\tilde{\boldsymbol{P}} = (\tilde{\rho}_{kl})$ from the correlation coefficients $\boldsymbol{P} = (\rho_{kl})$ via the inverse relationship

$$\boldsymbol{\Omega} = \boldsymbol{P}^{-1} = (\omega_{kl}) \tag{C.14}$$

$$\tilde{\rho}_{kl} = -\frac{\omega_{kl}}{\sqrt{\omega_{kk}\omega_{ll}}} \tag{C.15}$$

(Edwards, 1995). Applying these equations to estimates $\boldsymbol{R} = (r_{kl})$ of (dynamical) correlations allows to obtain estimates $\tilde{\boldsymbol{R}} = (\tilde{r}_{kl})$ of the associated partial (dynamical) correlations.

In order to test the significance of the correlations and to decide which of the possible edges to include in the resulting gene association network statistical tests are needed. In this paper we employ the "local fdr" network search Schäfer and Strimmer (2005a,b). The false discovery rate (fdr) is the expected proportion of false positives among the proposed edges. The local fdr is an empirical Bayes estimator of the false discovery rate (Efron, 2004). In the network search the local fdr is utilized to compute the posterior probability for an edge to be present or absent, and takes account of the multiplicity in the simultaneous testing of edges. The final network is obtained by visualizing all significant edges in an undirected graph.

C.3 Results

We now employ shrinkage estimation of the (partial) dynamical correlation to a real world example and compare it with the results of the traditional GGM method. Specifically, we reanalyzed a microarray time series data set Rangel et al. (2004). These data characterize the response of a human T-cell line (Jirkat) to a treatment with PMA and ioconomin, and consist of 10 time points with 44 replications each.

As approximation of the temporal expression of the 58 genes we used a linear spline and employed Eq. C.2 for the functional inner product. After estimation of the dynamical correlations with Eq. C.5 and regularization (section C.2.2) we computed the associated partial correlation coefficients employing Eq. C.14 and Eq. C.15. Using the locfdr algorithm (Efron, 2004) we then identified significant edges. The resulting network is displayed in Fig. C.1d.

For comparison we also compute the network as obtained by the classic GGM approach. For this analysis we ignored the dynamic aspects of the data and assumed that all mea-

surements were taken at the same time point. Furthermore, we examine the influence of shrinking. This leads to the four networks displayed in Fig. C.1.

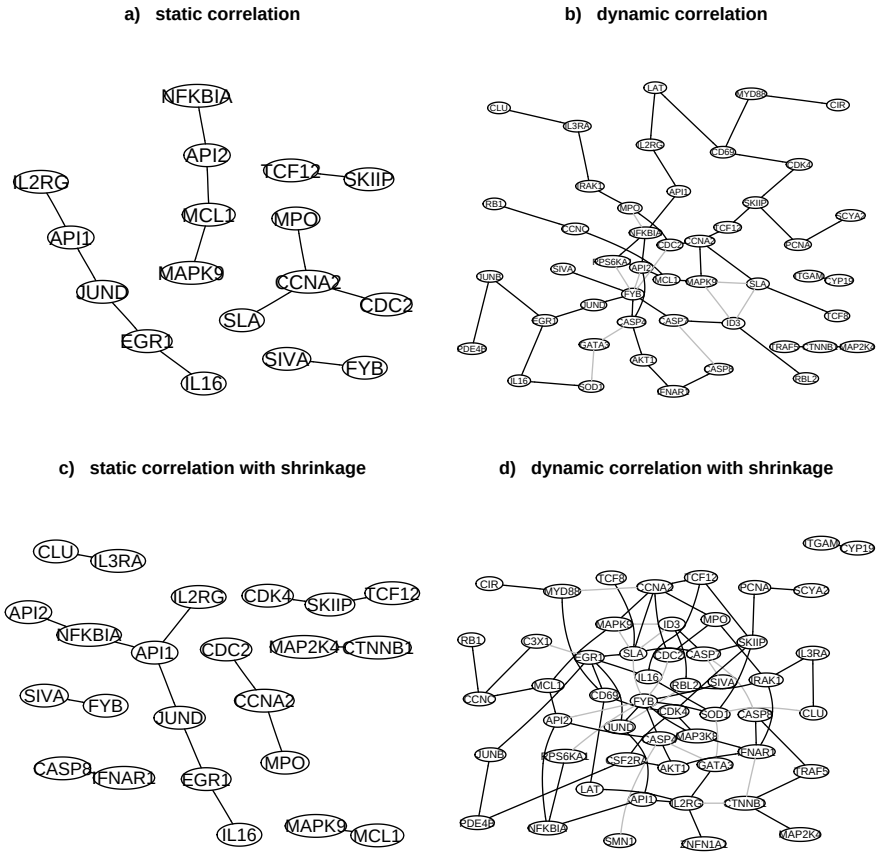

Figure C.1: Gene dependency networks inferred from human T-cell data (Rangel et al., 2004).

Ignoring the time series aspects and using static correlation leads to less well-connected networks compared with the ones calculated by dynamical correlation. This indicates that our dynamical FDA-based estimator is able to extract additional information about the interaction among the investigated genes. Furthermore, shrinkage also improves the power of the network reconstruction. Hence, we conclude that the best of the four investigated methods to infer gene association networks is the one relying on regularized dynamical correlation.

C.4 Conclusion

A growing interest in genetics lies in observing and inferring the gene interactions over time. Here, we introduced a method to infer a *regularized* gene dependency network

from functional data. It generalizes the static regularized GGM approach (Schäfer and Strimmer, 2005a) and is able to unravel the dependency structure of longitudinal data across the whole time series. Furthermore, unlike many other time series methods FDA does not require equally spaced measurements. Note that in FDA unequal time points are accounted for by the weights employed in the functional inner product. Furthermore, our algorithm is easily implemented and computationally inexpensive. Shrinkage allows to improve the precision of the estimation and to extend the method to high dimensional data. In order to further develop our approach many extensions are conceivable. An important topic is the inclusion of auto-regressive aspects. While our method covers the dynamical correlation through time it is not able to account, e.g., for a time shift between any two variables. These dependencies and the associated time shifts could be accounted for by modeling the temporal mean via a system of differential equation.

Acknowledgments

This work was supported by Deutsche Forschungsgemeinschaft (DFG) Emmy-Noether research award to K.S.

Bibliography

Dobra, A., Hans, C., Jones, B., Nevins, J. R., Yao, G., and West, M. (2004). Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*, 90:196–212.

Dubin, J. A. and Müller, H.-G. (2005). Dynamical correlation for multivariate longitudinal data. *Journal of the American Statistical Association*, 100(471):872–881.

Edwards, D. (1995). *Introduction to Graphical Modelling*. Springer, New York.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *Journal of the American Statistical Association*, 99(465):96–104(9).

Opgen-Rhein, R. and Strimmer, K. (2006). Inferring gene dependency networks from genomic longitudinal data: a functional data approach. *REVSTAT*, 4:53–65.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer, 2nd edition.

Rangel, C., Angus, J., Ghahramani, Z., Lioumi, M., Sotheran, E., Gaiba, A., Wild, D. L., and Falciani, F. (2004). Modeling t-cell activation using gene expression profiling and state-space models. *Bioinformatics*, 20(9):1361–1372.

Schäfer, J. and Strimmer, K. (2005a). An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, 21(6):754–764.

Schäfer, J. and Strimmer, K. (2005b). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4(1):32.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In Neyman, J., editor, *Proc. Third Berkeley Symp. Math. Statist. Probab.*, volume 1, pages 197–206, Berkeley. Univ. California Press.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley, New York.

Article D

Condition Number and Variance Inflation Factor to Detect Multicollinearity

Technical Report 01/07

Authors: Rainer Opgen-Rhein

Abstract:

- Multicollinearity can be a serious problem in regression models. We analyze linear regression in a manner that allows separating multicollinearity in a part originating in the estimation and a structural part. We clarify the difference of a regression model with and without an intercept. Two indices for multicollinearity, the variance inflation factor and the condition number are examined. We see that the problem of collinearity can be split in a part caused by the true structure of the data and in a part originated in the observation of the data. For the indices of multicollinearity we derive that centered and not centered data should be employed according to the model used for the regression. A simulation study is conducted to illustrate the insights.

Key-Words:

- *Multicollinearity, linear regression, condition number, variance inflation factor*

D.1 Introduction

Regression is used to describe a response variable with the help of some predictor variables. A problem often encountered in practical applications is multicollinearity. This means that the data used to predict the response is itself in some way connected, which can seriously affect the estimation of the regression model.

Here we will evaluate two measures of multicollinearity, the variance inflation factor and the condition number. For this we will first analyze the regression model extensively. This

allows clarifying the problem of multicollinearity and the meaning of these two indexes. We will furthermore go into the detail of the difference between a regression with and without a constant term. A simulation study is used to illustrate these insights. But at first we will concentrate on the regression model.

D.2 Linear Regression

A linear regression model assumes, that a response variable Y can be described by a linear combination of a set of predictor variables X

$$Y = a + Xb + \epsilon \qquad \epsilon \sim N(0, \sigma_\epsilon^2) \tag{D.1}$$

with ϵ being the residuals and a an intercept which is – depending on the model – included or not. We will deal with multiple regression, where one response variable depends on several predictor variables. Nevertheless, all analyses can also be applied to a multivariate regression with more than one response variables. The maximum likelihood (ML) estimator for the regression coefficients b can be calculated by:

$$b = (X'X)^{-1}X'Y \tag{D.2}$$

This represents the model without an intercept. The model with an intercept can be calculated in two ways. The first possibility is to add a column of 1 to the predictors, whereby the intercept is the regression coefficient of this column. The second possibility is to center the data matrix X, to apply the ML estimator and to infer the intercept by $a = \frac{1}{n}\sum y_i - \frac{1}{n}\sum \tilde{y}_i$.

Nevertheless, the data matrix X represents only observations of the true structure of the variables. In the next section we examine the regression model given the true structure of the data. We show that all linear dependencies in a model with an intercept can be derived using only the covariance matrix of the variables. This applies analogous to a model without an intercept using the true "crossproduct structure". These considerations allow splitting the problem of collinearity in a part caused by the true structure of the data and in a part originated in the observation of the data.

D.2.1 Linear regression based on the true model

Our starting point is the *true* covariance matrix of some variables X. The actual estimation of this covariance matrix is a different problem to be dealt with later. We stress that this covariance matrix is the only object required for the calculation done here, nothing else is given or estimated. In the usual notation, the covariance matrix and its inverse can be displayed as

$$\begin{aligned} \mathrm{Cov}(X_k, X_l) &= \sigma_{kl}, \\ \boldsymbol{\Sigma} &= (\sigma_{kl}), \\ \boldsymbol{\Omega} &= (\omega_{kl}) = \boldsymbol{\Sigma}^{-1}. \end{aligned} \tag{D.3}$$

The covariance matrix implies a correlation matrix and its inverse:

$$\begin{aligned} \operatorname{cor}(X_k, X_l) &= \rho_{kl} = \sigma_{kl}(\sigma_{kk}\sigma_{ll})^{-1/2}, && \text{(D.4)} \\ \boldsymbol{P} &= (\rho_{kl}), \\ \boldsymbol{Q} &= (q_{kl}) = \boldsymbol{P}^{-1}. && \text{(D.5)} \end{aligned}$$

For the regression model, we use a subscript if necessary: $\boldsymbol{\Sigma}_X$ is, e.g., the covariance matrix of the variables X, $\boldsymbol{\Sigma}_{YX}$ the covariance matrix of the combined variables Y and X: $[YX]$.

An important concept is that of partial variances and partial correlations, which are the variances and correlations between two variables conditional on all other variables. The partial variance can be inferred by

$$\operatorname{Var}(X_k|\text{rest}) = \tilde{\sigma}_{kk} = \omega_{kk}^{-1} = \operatorname{diag}(\boldsymbol{\Omega})^{-1} \tag{D.6}$$

and the partial correlations by

$$\begin{aligned} \operatorname{cor}(X_k, X_l|\text{rest}) &= \tilde{\rho}_{kl} = -\frac{\omega_{kl}}{(\omega_{kk}\omega_{ll})^{-1/2}}, && \text{(D.7)} \\ \tilde{\boldsymbol{P}} &= (\tilde{\rho}_{kl}) \end{aligned}$$

the tilde denotes "partial. The partial correlation matrix can also be displayed in terms of $\boldsymbol{\Omega}$ or $\boldsymbol{Q}$:

$$\begin{aligned} \tilde{\boldsymbol{P}} &= -\boldsymbol{\Omega} \odot \sqrt{\operatorname{diag}(\boldsymbol{\Omega})^{-1}(\operatorname{diag}(\boldsymbol{\Omega})^{-1})'}, \\ \tilde{\boldsymbol{P}} &= -\boldsymbol{Q} \odot \sqrt{\operatorname{diag}(\boldsymbol{Q})^{-1}(\operatorname{diag}(\boldsymbol{Q})^{-1})'}. \end{aligned}$$

The symbol $\odot$ marks the elementwise product, also called Hadamard product. Note that we have to change the sign of the diagonal.

Having defined partial variances and partial correlations, we can infer the regression model *exactly* using the covariance/correlation structure of $[YX]$. The best linear predictor of the regression coefficients in terms of mean squared error is

$$b_k = \tilde{\rho}_{YX_k}\sqrt{\frac{\tilde{\sigma}_{YY}}{\tilde{\sigma}_{X_kX_k}}}, \tag{D.8}$$

$$b_k = \tilde{\rho}_{YX_k}\sqrt{\frac{q_{YY}^{-1}}{q_{X_kX_k}^{-1}}}\sqrt{\frac{\sigma_{YY}}{\sigma_{X_kX_k}}}. \tag{D.9}$$

The ordinary least squares solution of the regression model can be derived by substituting the true correlations/covariances with their corresponding estimators. The last representation of the predictor is especially interesting. The first term represents the partial correlation between two variables and displays the strength of the connection between them. In the middle term we see q_{YY}^{-1} and $q_{X_kX_k}^{-1}$. These are – for each variable – the partial variances

divided by the variances of the respective variables:

$$\begin{aligned} q^{-1}_{YY} &= \frac{\tilde{\sigma}_{YY}}{\sigma_{YY}}, \\ q^{-1}_{X_k X_k} &= \frac{\tilde{\sigma}_{X_k X_k}}{\sigma_{X_k X_k}}. \end{aligned}$$

We can see q^{-1} as the standardized partial variance and interpret it as the fraction of the variance of each variable that cannot be explained by all others. The middle term therefore represents the ratio of this fraction for two variables and indicates which of them can be seen as dependent and which of them as independent variable. The last term is the only one that depends on the scaling of the variables and is necessary there for scaling reasons (the regression coefficients are given in absolute terms).

The variance of the residuals is usually estimated by

$$\hat{\sigma}^2 = \frac{1}{n-p-1} \underbrace{e'e}_{\text{RSS}}.$$

The residual sum of square $e'e$ depends on the true covariance structure of $[YX]$ so that

$$\hat{\sigma}^2 = \frac{n-1}{n-p-1} \text{diag}(\mathbf{\Omega}_{YX})^{-1}_{(1,1)} = \frac{n-1}{n-p-1} \tilde{\sigma}_{YX_{(1,1)}}.$$

It becomes clear that the estimated variance of the residuals corresponds to the partial variance of the response in the exact model $\sigma^2 = \tilde{\sigma}_{YX_{(1,1)}}$.

If the covariance structure is unknown and the regression coefficients have to be estimated, the variance of a single regression coefficient is influenced by the true covariance structure of the variables and the number of observations:

$$\begin{aligned} \hat{\sigma}^2_k &= \frac{1}{n-p-1} \text{diag}(\mathbf{\Omega}_{YX})^{-1}_{(1,1)} \text{diag}(\mathbf{\Sigma}^{-1}_X)_{(k,k)} \\ &= \frac{1}{n-p-1} \frac{\tilde{\sigma}_{YX_{(1,1)}}}{\tilde{\sigma}_{X_{(k,k)}}} \\ &= \frac{1}{n-p-1} \frac{\tilde{\sigma}_{YX_{(1,1)}}}{\sigma_{YX_{(1,1)}}} \left(\frac{\tilde{\sigma}_{X_{(k,k)}}}{\sigma_{X(k,k)}} \right)^{-1} \frac{\sigma_{YX(1,1)}}{\sigma_{X(k,k)}}. \end{aligned} \tag{D.10}$$

The last form allows some interpretation. The first term contains the degrees of freedom. The second term is the fraction of the variance of the response not explained by the regression. This is divided by the third term, which is the fraction of the variance of the predictor variable k that cannot be explained by the other predictor variables. This implies that the more the predictors are connected, the greater the variance of the regression coefficients. The last term is necessary for scaling reasons.

We see that the true covariance structure allows inferring all linear dependencies in a model. Nevertheless, this true structure is – for real life problems – unknown, and has to be estimated. This is described in the next section.

D.2.2 Empirical estimation - with and without intercept

In empirical problems, the true values for the covariances are unknown and have to be estimated. For regressions models conducted with an intercept, the true covariance matrix $\mathbf{\Sigma}$ for the variables X can simply be replaced by its empirical counterpart $\boldsymbol{S}$

$$\boldsymbol{S} = \hat{\mathbf{\Sigma}} = \frac{1}{n-1}(X - \overline{X})'(X - \overline{X}). \tag{D.11}$$

This applies analogue to the covariance matrix $\mathbf{\Sigma}_{YX}$ of the combined matrix of response and observation vectors.

It is clear that accurate estimation of the covariance structure is crucial for efficient estimation of the regression model. Increasing the number n of the observations serves this purpose, but often this is not possible. In these cases, regularization is necessary. The well-known Ridge regression (Hoerl and Kennard, 1970) is implicitly a regularized estimation of the covariance matrix. An explicit regularization of the covariance matrix utilized for regression is conducted by shrinkage regression (Opgen-Rhein and Strimmer, 2007), which relies on a Stein-type regularization method.

Estimation of a regression model without an intercept can be done in the same way as described above. The only difference that instead of using centered data for the estimation, only the crossproduct is employed

$$\boldsymbol{S}^* = \frac{1}{n}X'X. \tag{D.12}$$

We are now able to discuss the problem of multicollinearity in the data and describe some measures of diagnosis.

D.3 Detecting Multicollinearity

D.3.1 Multicollinearity

Exact multicollinearity between data vectors X_i is present, if one or more of the data vectors are a linear combination of the others. Multicollinearity in a broader sense means that the data vector X_k can to some extent be expressed by other vectors (Belsley, 1991, p. 19). The main problem of multicollinearity is that regression coefficients have large variances and that test statistics are unreliable (Mansfield and Helms, 1982).

Multicollinearity is asserted to be a "data problem, not a statistical problem"(Belsley, 1991, p. 20). Nevertheless the statistical analysis conducted in chapter D.2 allows us to segment multicollinearity in a true collinear structure (leading to large variances of the regression coefficients) and instability generated by the observation (due to inefficient estimation of the covariance structure). This clarifies the effects on the estimation of the linear model and helps to deal with multicollinearity.

We will now focus on two different measures of multicollinearity, the variance inflation factor and the condition number.

D.3.2 Variance inflation factor

The "variance inflation factor" (VIF) for a variable X_j is defined as

$$\mathrm{VIF}_j = \frac{1}{1 - R_j^2} \tag{D.13}$$

(Smith and Campbell, 1980). R_j^2 is the multiple correlation coefficient for the regression of the variable X_j on the other variables. However, it is difficult to derive the meaning of the variance inflation factor using this description. Another formulation is that the variance inflation factors are the diagonal elements of the inverse of the correlation matrix $\mathrm{VIF}_j = C_{jj}^{-1}$ (Mansfield and Helms, 1982). We will derive an alternative expression for the VIF, which already implicitly appeared in our exact reconstruction of linear regression.

Define Σ as the true correlation structure of the prediction variables and its inverse $\Omega = \Sigma^{-1}$. This allows inferring the matrix of regression coefficients for all regressions of $X_{i\setminus j}$ onto X_j simultaneously and *exactly*

$$B = -\Omega \odot (\mathrm{diag}(\Omega)^{-1}\mathbf{1}'). \tag{D.14}$$

Multiplying the diagonal of B by -1, this means, e.g., that we find on row j the regression coefficients for the regression of all $X_{i\setminus j}$ onto X_j plus 1 in the column j (the diagonal). The vector of the residual sum of squares (RSS) can be displayed as

$$\mathrm{RSS}_X = \sum_{\text{columnwise}} (X - \hat{X})^2 = \sum_{\text{columnwise}} (2X - XB')^2 = \mathrm{diag}(\Omega)^{-1}. \tag{D.15}$$

The RSS in the equation above applies to a specific set of observations. Nevertheless, it is possible to infer the true RSS for a correlation structure by noting that it is constructed from the residuals which in turn only depends on the correlation structure Σ

$$\mathrm{RSS} = n \cdot \mathrm{diag}(\Omega)^{-1}. \tag{D.16}$$

As the total sum of square can be displayed as $\mathrm{SS}_{\text{total}} = n \cdot \mathrm{diag}(\Sigma)^{-1}$ the vector of multiple correlation coefficients for a certain correlation structure results in

$$\mathrm{R}^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{SS}_{\text{total}}} = 1 - \mathrm{diag}(\Omega)^{-1} \odot \mathrm{diag}(\Sigma)^{-1}, \tag{D.17}$$

and finally vector of true variance inflation factors

$$\mathrm{VIF} = 1 - \frac{\mathrm{RSS}}{\mathrm{SS}_{\text{total}}} = \mathrm{diag}(\Omega) \odot \mathrm{diag}(\Sigma). \tag{D.18}$$

This expression allows inferring the true meaning of the variance inflation factor: $\mathrm{diag}(\Omega)^{-1}$ are the *partial* variances $\tilde{r}_{ii}$ of X, which describe the correlation between any two variables X_i and X_j conditioned on all the other variables. Therefore VIF_i is simply

$$\mathrm{VIF}_i = \frac{\sigma_{ii}}{\tilde{\sigma}_{ii}}, \tag{D.19}$$

the variances divided by the partial variances. By the definition of the partial variances it follows that the VIF increases, the more of the variance of X_j can be explained by the Variables $X_{i\setminus j}$. It is possible to demonstrate why high variance inflation factors affect the regression. Recall the estimation of the variance of the regression coefficients

$$\begin{aligned}\hat{\sigma}_k^2 &= \frac{1}{n-p-1}\,\frac{\tilde{\sigma}_{YX_{(1,1)}}}{\sigma_{YX_{(1,1)}}}\,\left(\frac{\tilde{\sigma}_{X_{(k,k)}}}{\sigma_{X(k,k)}}\right)^{-1}\,\frac{\sigma_{YX(1,1)}}{\sigma_{X(k,k)}} \\ &= \frac{1}{n-p-1}\,\frac{\tilde{\sigma}_{YX_{(1,1)}}}{\sigma_{YX_{(1,1)}}}\,\mathrm{VIF}_k\frac{\sigma_{YX(1,1)}}{\sigma_{X(k,k)}}. \end{aligned} \tag{D.20}$$

It becomes clear that large variance inflation factors lead to a high variance in the estimation of the regression coefficients and therefore to unstable results.

We see that the true variance inflation only depends on the covariance structure, not on the sample size and is likely to increase by adding an intercept, as more of the variance can be explained by a better fit of the regression of $X_{i\setminus j}$ onto X_j usually obtained by adding an intercept.

As the true covariance structure of the observation is usually unknown, it has to be estimated from the data. Increasing the sample size of the observation therefore leads to a reduction of the variance of the estimations of the VIFs, but we can only improve the estimation of the true VIFs, not reduce the variance inflation itself.

We can again distinguish between regression with an intercept and without an intercept: to estimate the VIFs, we use in the first case the sample covariance and in the latter the sample matrix without centering.

D.3.3 Condition number

The condition number $K(X)$ of the matrix X of the observations is another measure for collinearity in the data. It is defined as the ratio of the biggest and the smallest singular value of X (Belsley, 1991, p.40 et sqq.):

$$K(X) = \frac{\mu_{\max}}{\mu_{\min}}. \tag{D.21}$$

This can also be displayed in terms of the eigenvalues of $X'X$:

$$\begin{aligned} X &= UDV', \\ X'X &= VDU'UDV' = VD^2V', \\ K(X) &= \sqrt{\frac{d^2_{\max}}{d^2_{\min}}}. \end{aligned} \tag{D.22}$$

There are two interpretations of the condition number given in Belsley (1991): it can be seen as a "measure of potential sensitivity of a system of linear equations to perturbations in its data" and as a "relative measure of X from exact collinearity" (p. 55).

We will give an additional interpretation, which will also clarify the notion that the "sensitivity is further affected by the strength of the linear relation between Y and X" and the remark that even well-conditioned data can show high sensitivity (p. 54 et sq.).

The inverse of the covariance matrix Σ_{YX}defined in chapter D.2 can also be displayed in terms of eigenvalues and eigenvectors. Note that we use the combined matrix of response and predictors.

$$\begin{aligned} \Omega_{YX} &= V\frac{1}{D^2}V', \\ \omega_{kl} &= \sum_i \frac{V_{ki}V_{li}}{D_i^2}. \end{aligned}$$

Some calculations lead to the regression coefficients b_k in terms of eigenvalues and eigenvectors:

$$b_k = -\frac{\sum_i \frac{V_{ki}V_{li}}{D_i^2}}{\sum_i \frac{V_{ki}^2}{D_i^2}}. \tag{D.23}$$

We see that near zero singular values of X (or eigenvalues of $X'X$) lead to an explosion of the nominator as well as the denominator. The condition number of $[YX]$ measures the ill-conditioning of the matrix YX. We are now also able to explain the remarks about condition number of X. A well-conditioned matrix X does not imply a well-conditioned matrix $[YX]$, but an ill-conditioned matrix X necessarily leads to an ill-conditioned matrix $[YX]$. Nevertheless, as the condition number as a measure of multicollinearity for a regression is defined only for the matrix X, we will use the condition number of X for further examinations.

Moreover, the idea of an alternative measure which uses the matrix $[YX]$ to calculate a condition number $K(YX)$ also has its severe drawback: the better the response can be explained by the predictors, the more one singular value of $[YX]$ approximates zero which in turn increases the condition number $K(YX)$.

We can again separate multicollinearity in ill-conditioned true structure and the problems of estimating these true structures. As before, we can make the distinction between a regression model with an intercept (where we take the empirical covariance matrix) and without an intercept (where we estimate the crossproduct structure).

D.4 Simulation

D.4.1 Simulation Models

For further examination, we conducted simulations with two different setups inspired by the simulations in Belsley (1991, p.79 et sqq.). In the first study, only a few variables are multicollinear, in the second, there is complex multicollinearity between all data vectors. We will first describe the setup in detail and subsequently present the results.

Simulation 1 We first generated a basic data structure $p = 10$ variables and $n = [20; 100; 500]$ observations normally distributed with $X \sim N(0, 1)$. To allow substantially differences between centered and not centered data in the later examination a constant $d \sim \text{Unif}(-5, 5)$ was added to each column. Multicollinearity is introduced by replacing the first column X_1 of the data with X_1^*, a mixture by X_1 and a combination of X_5, X_6 and X_7

$$X_1^* = (1 - \lambda)X_1 + \lambda \frac{1}{\sum g_i}(g_1 X_5 + g_2 X_6 + g_3 X_7) \tag{D.24}$$

The variable g is uniformly distributed with $a \sim \text{Unif}(0, 1)$ and λ, the strength of multicollinearity, varies between $[0; 0.9]$.

For the regression, $p + 1$ regression coefficients were generated by $a, b \sim Unif(-5, 5)$ with a being the intercept. The response Y was generated by

$$Y = a + Xb + \epsilon \qquad \epsilon \; N(0, \sigma_\epsilon)$$

with a signal to noise ratio of 3 to 1. Each simulation was repeated 250 times and the VIFs, the condition numbers, and the Mean Square Errors (MSEs) calculated, each for an estimation with and without an intercept.

Simulation 2 The second simulations study was conducted analogue to the first one, only the multicollinearity involves all variables. Given the basic data structure X, the multicollinear data is generated according to

$$X_1^* = (1 - \lambda)X_1 + \lambda \frac{1}{\sum g_i^1}(g_1^1 X_5 + g_2^1 X_6 + g_3^1 X_7), \tag{D.25}$$

$$X_2^* = (1 - \lambda)X_2 + \lambda \frac{1}{\sum g_i^2}(g_1^2 X_8 + g_2^2 X_9 + g_3^2 X_{10}), \tag{D.26}$$

$$X_3^* = (1 - \lambda)X_3 + \lambda \frac{1}{\sum g_i^3}(g_1^3 X_6 + g_2^3 X_8 + g_3^3 X_{10}), \tag{D.27}$$

$$X_4^* = (1 - \lambda)X_4 + \lambda \frac{1}{\sum g_i^4}(g_1^4 X_1^* + g_2^4 X_2^* 8 + g_3^4 X_3^*). \tag{D.28}$$

Again, the variable g_i^j is uniformly distributed with $a \sim \text{Unif}(0, 1)$ and λ varies between $[0; 0.9]$.

D.4.2 Results

The results of the first simulation are shown in figure D.1 for a sample size of 20 observations, in figure D.2 for a sample size of 100 and in figure D.3 for a sample size of 500. If we look in detail at figure D.1 we see in the upper left box the variance inflation factors of the regression coefficients for a regression model including an intercept. We can clearly

distinguish the variance inflation factor of Variable X_1 (the dotted lines reflects the standard deviation of the estimation: $VIF_1 \pm \text{sd}(VIF_1)$), which explodes for increasing λ, the strength of multicollinearity in the model (note that we have a log scale on the y-axis). The variable X_1 is composed by a linear combination of the variables X_5, X_6, X_7, which are also clearly distinguishable for higher multicollinearity. The upper right box shows the same for the uncentered data, hence for the estimation without an intercept. We see that the variance inflation factors are higher even for $\lambda = 0$. This was obviously introduced by the simulation setup where we added a random constant to each variable. Nevertheless, the behavior of the VIFs for increasing λ is analogue to the centered data.

In the lower left box we see the condition numbers for the centered and uncentered data (the standard deviations are depicted by the dotted lines). The condition number shows a similar behavior to the variance inflation factors: both, the centered number for the estimation with an intercept and the uncentered number for the estimation without an intercept increase heavily for $\lambda \to 1$, but the condition number from the uncentered data has a higher value to begin with. This shrinkage of the collinearity indices is described in Stewart (1987) in detail, but our analysis showed that this can be no claim only to use only the uncentered data as suggested in (Belsley, 1984).

The lower right box shows the MSE of the regression coefficients. The dotted line indicates the MSE for the zero model (all regression coefficients equal zero). As expected, the regression with an intercept shows a better fit of the model and the MSE increases for larger multicollinearity until no information can be gathered from the data (the MSE of the regression is worse than the MSE of the zero model).

Figure D.2 and D.3 show the same simulation model with an increased sample size of 100 and 500. We see the same pattern of behavior of the collinearity indexes and the MSE of the regression coefficients. Nevertheless, we can clearly see the distinction made in chapter D.2 between the observational and structural problems: the absolute values of the collinearity indexes as well as the influence on the fit of the model decrease with increasing sample size. This can be attributed to the better estimation of the true covariance and autocorrelation matrices.

Figure D.4, D.5 and figure D.6 show the result of the second simulation, where multicollinearity involves all data vectors. This is reflected in the variance inflation factors where all VIFs increase with higher λ, the strength of multicollinearity. The condition number is not able to distinguish between the two model setups and exhibits similar behavior even in absolute values.

Figure D.7 explicitly concentrates on the difference between estimation and structural problems. The estimation of the real matrices improves with increasing sample size. In D.7 the second simulation setup is used with a fixed λ of 0.7 and a varying sample size. For small sample sizes the collinearity indexes have high values and standard deviations, which reflects the uncertainty in the estimation of the covariance matrix and the autocorrelation matrix. For an increasing sample size the VIF and condition numbers approach a fixed value. This reflects the multicollinearity originated by the true structure of the system and cannot be eliminated by additional observations.

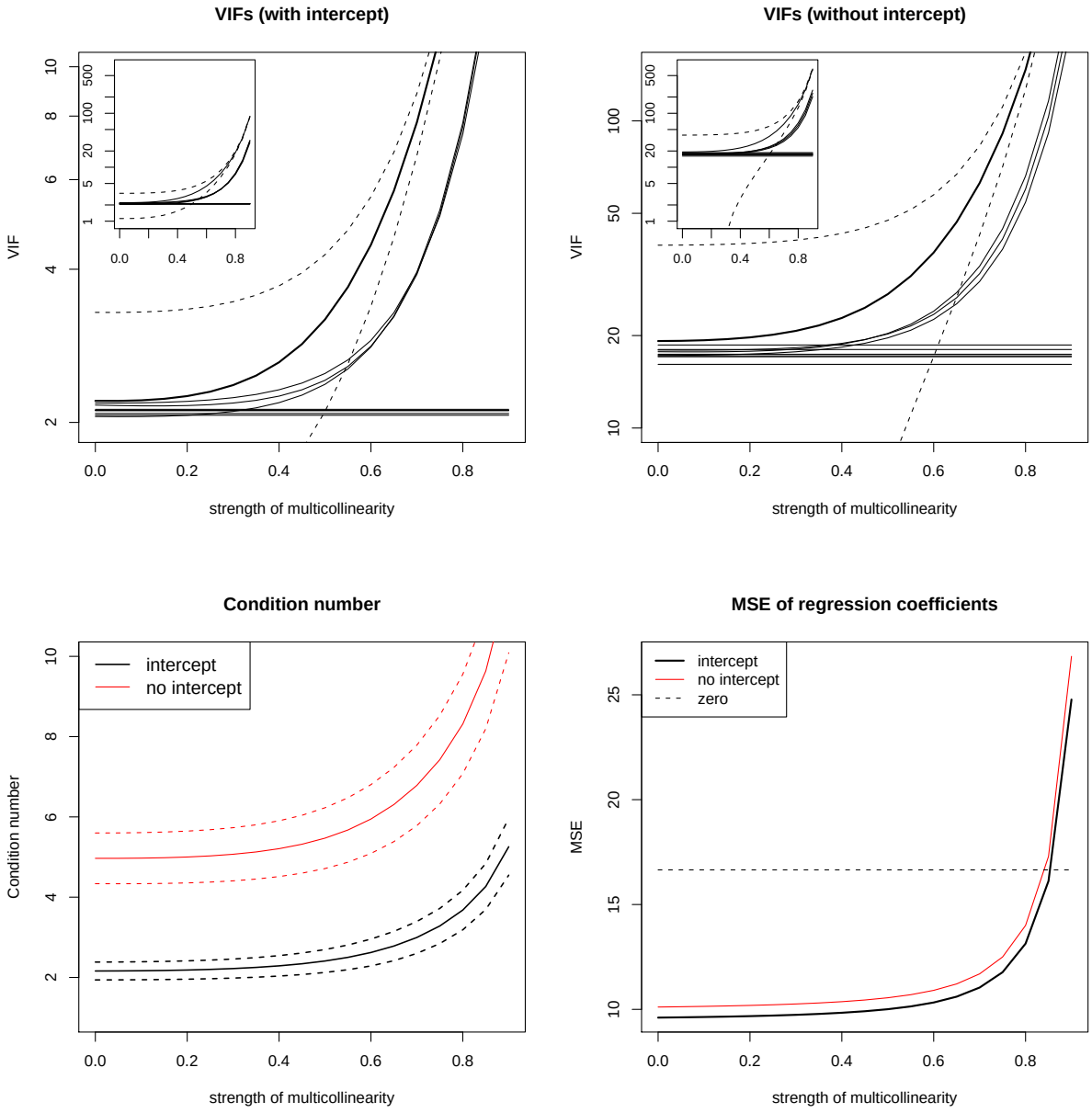

Figure D.1: First simulation (weak structural multicollinearity), 20 observations

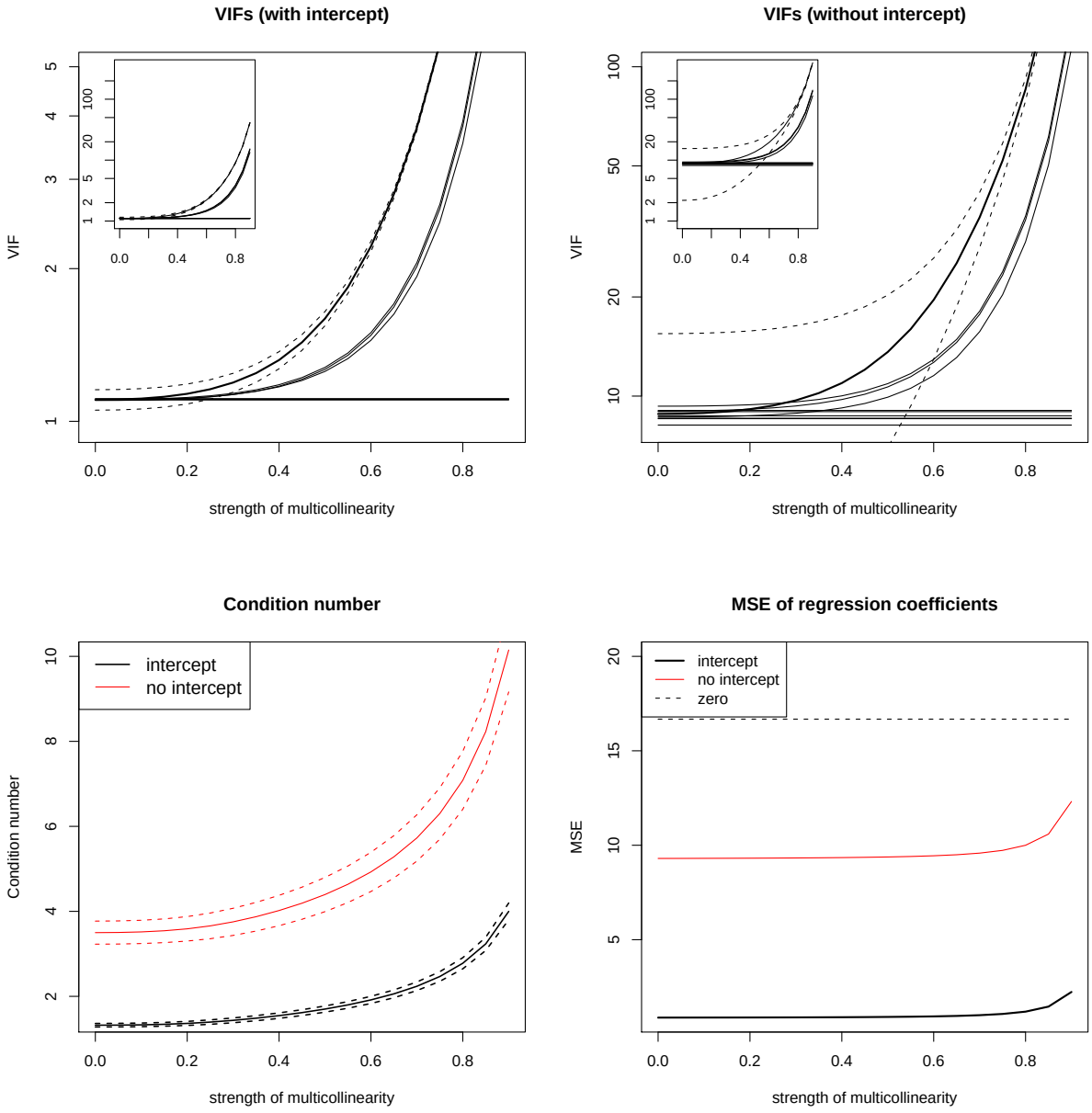

Figure D.2: First simulation (weak structural multicollinearity), 100 observations

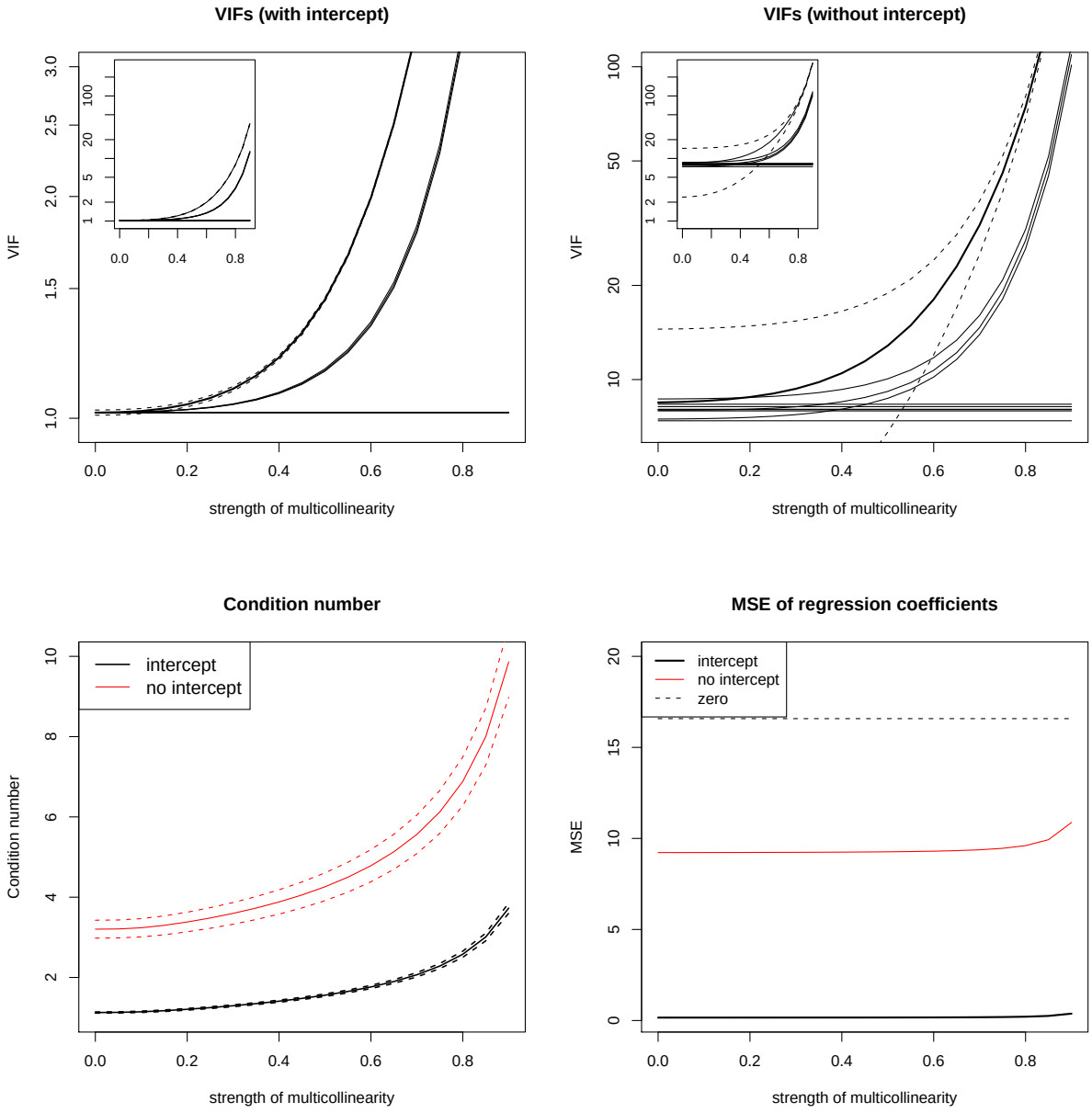

Figure D.3: First simulation (weak structural multicollinearity), 500 observations

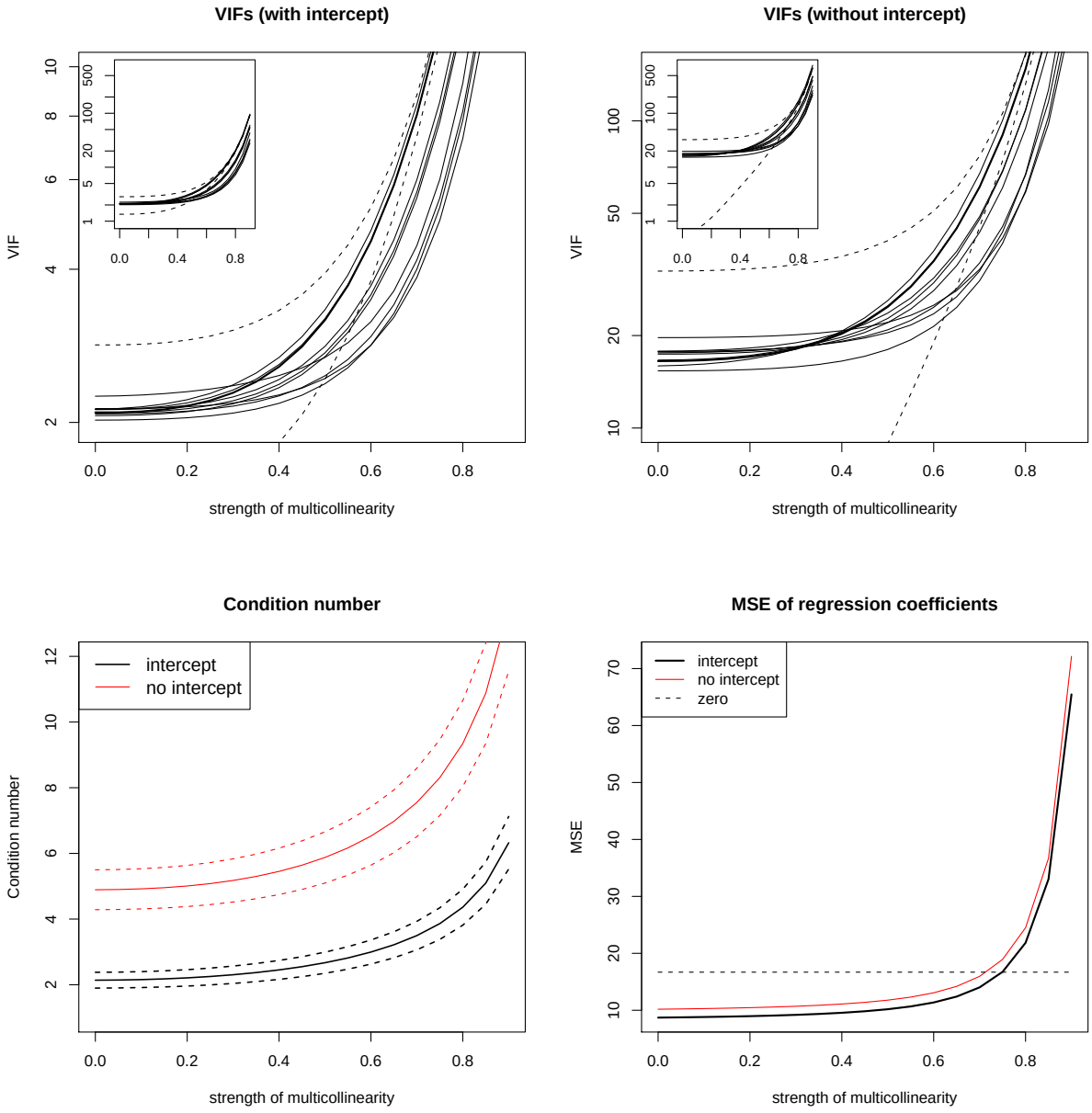

Figure D.4: Second simulation (strong structural multicollinearity), 20 observations

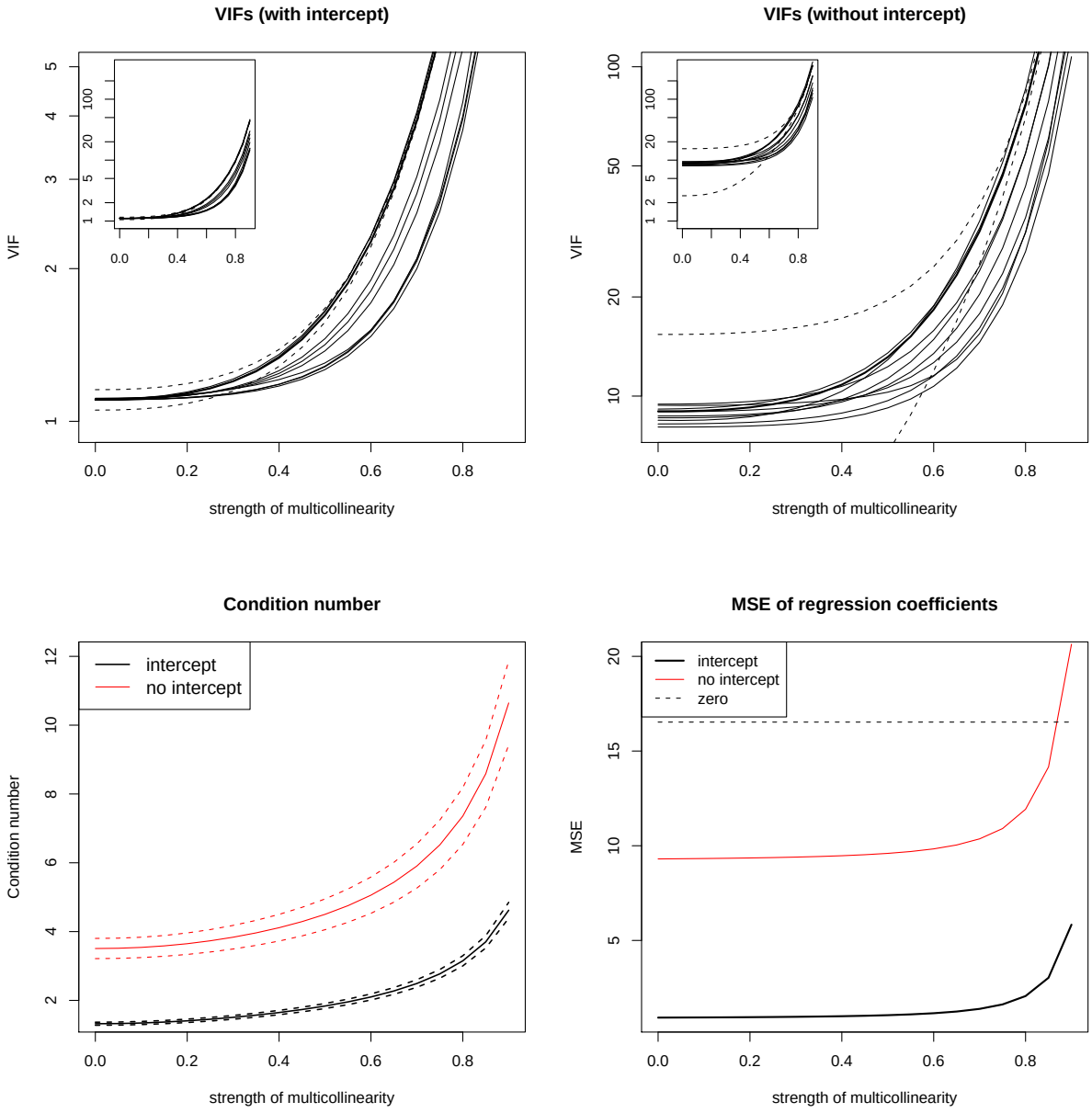

Figure D.5: Second simulation (strong structural multicollinearity), 100 observations

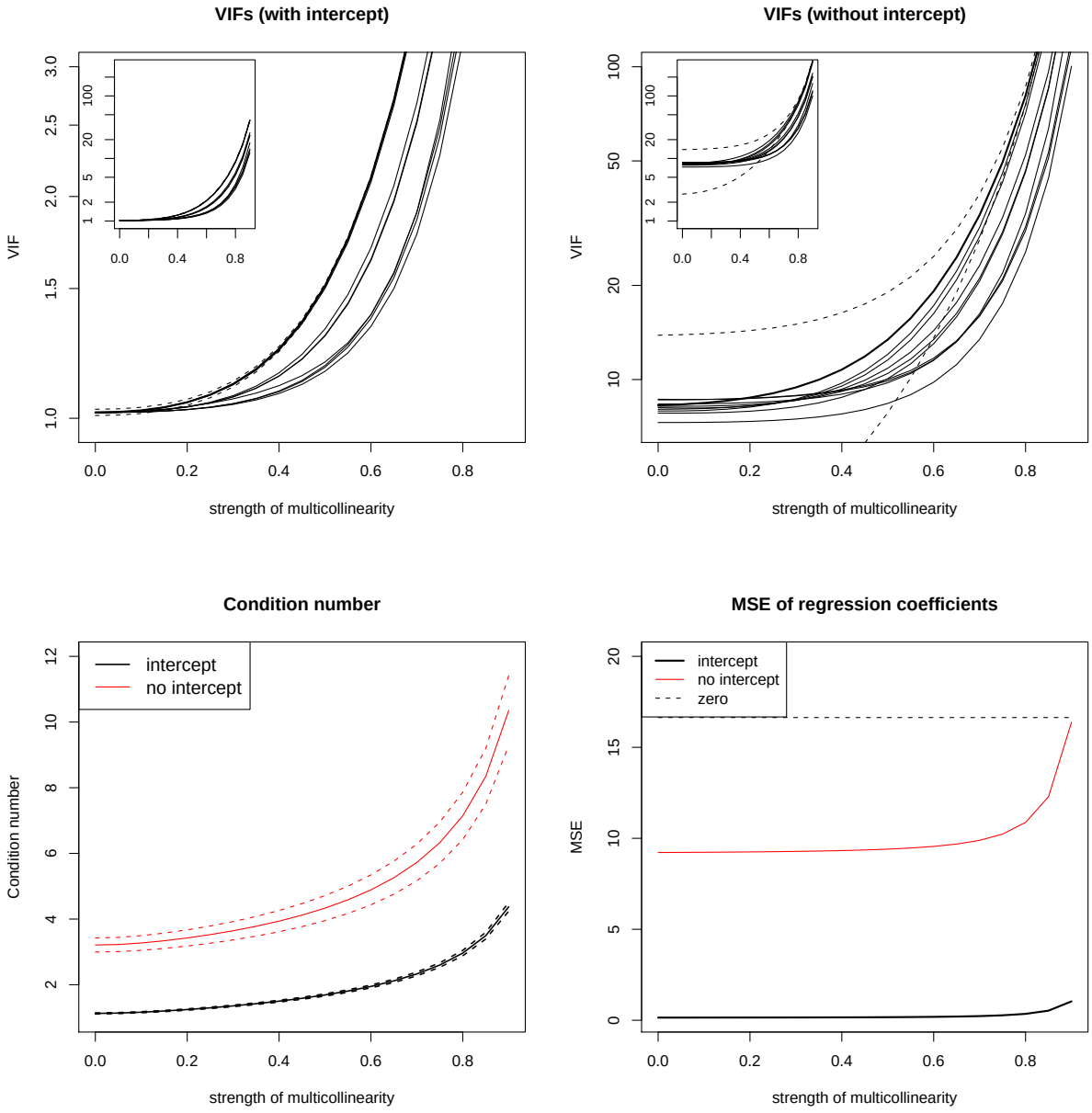

Figure D.6: Second simulation (strong structural multicollinearity), 500 observations

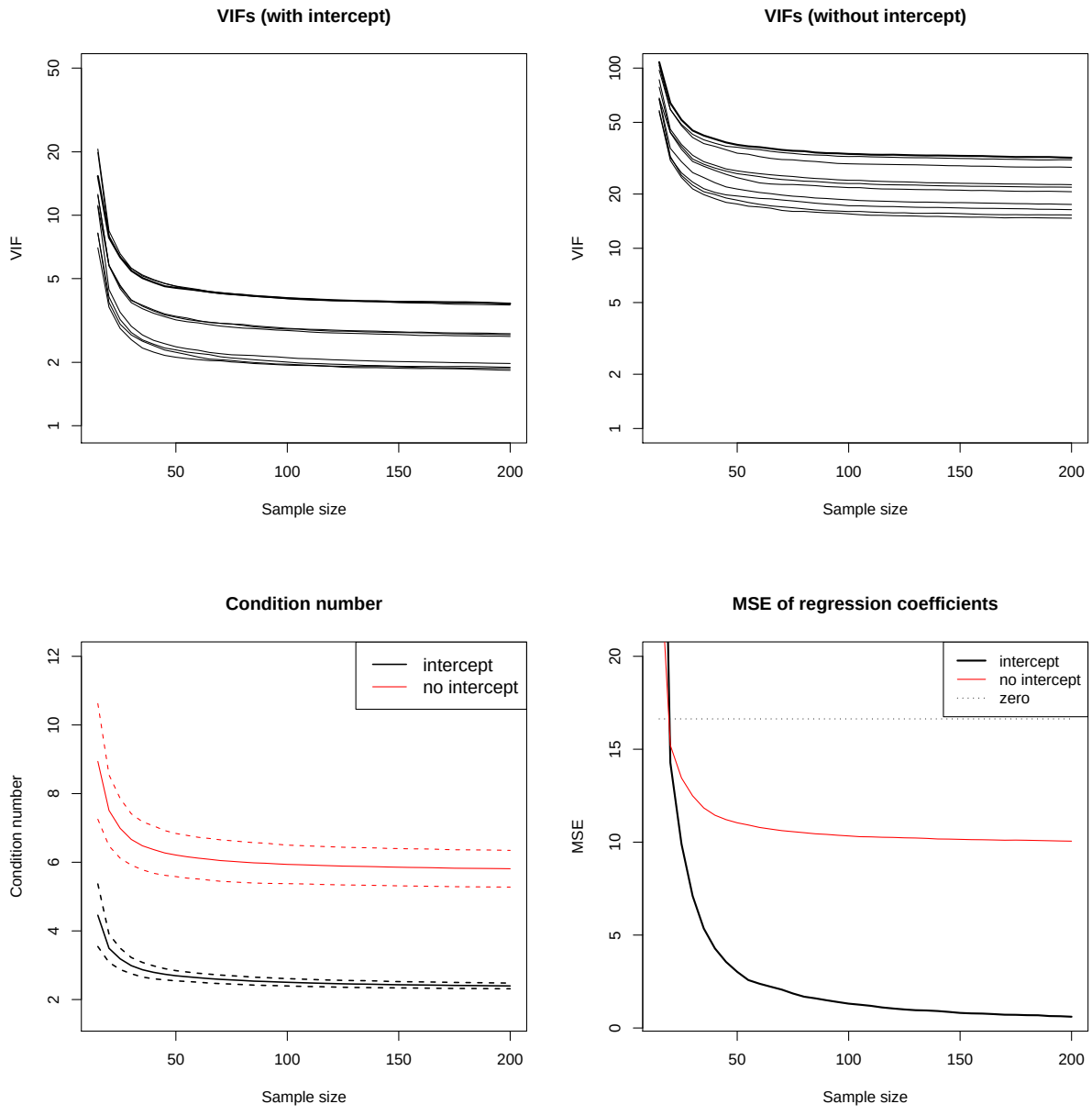

Figure D.7: Second simulation (strong structural multicollinearity), dependence on sample size, lambda=0.7

D.5 Discussion

The insights gathered in the analysis of multicollinearity can be summarized as follows.

The first result is the importance of the distinction between the true and the estimated structure of the model. The observational problems can be reduced by a better estimation of the true structure, either by increasing sample size or by the usage of more efficient estimation methods than ordinary least square. The latter is especially possible for small sample sizes, if the number of observations cannot be increased. The structural multicollinearity cannot be eliminated, it decreases the efficiency of the regression, but – on the other hand – it gives further knowledge of the connection between the variables.

The second important fact is the distinction between centered and not centered data. Centered data is used for the estimation of a regression model with an intercept, not centered data help to identify a regression model without a constant. The important point is that the collinearity indexes should reflect this distinction. Centered or uncentered data should be used for the VIF or the condition number according to the model used for the regression, especially as the indexes estimated from uncentered data depend on the choice of the origin.

The last observation concerns the difference between the VIF and the condition number. We saw that these measures are both able to identify multicollinearity and – furthermore – behave similar in the extent of the identification. Nevertheless, the VIFs allow identifying the data vectors involved in the multicollinearity. It is to mention, that the VIF is typically used with centered data, the condition number with not centered data. This might – in some cases – lead to a multicollinearity identified by the condition index, but not by the VIF. However, this is only based on the difference between centered and not centered data explained in the second result.

In the end, there are no fixed values of the collinearity indexes that indicate unacceptable multicollinearity. Nevertheless our analysis allowed measuring the influence of multicollinearity on the regression.

Bibliography

Belsley, D. A. (1984). Demeaning conditioning diagnostics through centering. *American Statistician*, 38(2):73–77.

Belsley, D. A. (1991). *Conditioning Diagnostics: collinearity and weak data in regression*. Wiley Interscience.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Applications to nonorthogonal problems. *Technometrics*, 12(1):69–82.

Mansfield, E. R. and Helms, B. P. (1982). Detecting multicollinearity. *American Statistician*, 36(3, Part 1):158–160.

Opgen-Rhein, R. and Strimmer, K. (2007). Learning causal networks from systems biology time course data: an effective model selection procedure for the vector autoregressive process. *BMC Bioinformatics*, 8 (Suppl. 2):S3.

Smith, G. and Campbell, F. (1980). A critique of some ridge regression methods. *Journal of the American Statistical Association*, 75(369):74–81.

Stewart, G. W. (1987). Collinearity and least squares regression. *Statistical Science*, 2(2):68–84.

Article E

Learning Causal Networks from Systems Biology Time Course Data: An Effective Model Selection Procedure for the Vector Autoregressive Process

Published in BMC Systems Biology *(Volume 8, Supplement 2, S3) Proceedings of PMSB 2006 ("Probabilistic Modeling and Machine Learning in Structural and Systems Biology"), Tuusula, Finland, 17-18 June 2006*

Authors: Rainer Opgen-Rhein *and* Korbinian Strimmer

Abstract:

- *Background:* Causal networks based on the vector autoregressive (VAR) process are a promising statistical tool for modeling regulatory interactions in a cell. However, learning these networks is challenging due to the low sample size and high dimensionality of genomic data.

  Results: We present a novel and highly efficient approach to estimate a VAR network. This proceeds in two steps: (i) improved estimation of VAR regression coefficients using an analytic shrinkage approach, and (ii) subsequent model selection by testing the associated partial correlations. In simulations this approach outperformed for small sample size all other considered approaches in terms of true discovery rate (number of correctly identified edges relative to the significant edges). Moreover, the analysis of expression time series data from *Arabidopsis thaliana* resulted in a biologically sensible network.

  Conclusions: Statistical learning of large-scale VAR causal models can be done efficiently by the proposed procedure, even in the difficult data situations prevalent in genomics and proteomics.

  Availability: The method is implemented in R code that is available from the authors on request.

E.1 Background

The vector autoregressive regression (VAR) model is an approach to describe the interaction of variables through time in a complex multivariate system. It is very popular in economics (Sims, 1980) but with few exceptions (Bay et al., 2004) it has not been widely used in systems biology, where it could be employed to model genetic networks or metabolic interactions. One possible reason for this is that while the statistical properties of the VAR model are well explored (Lütkepohl, 1993), its estimation from sparse data and subsequent model selection is very challenging due to the large number of parameters involved (Ni and Sun, 2005).

In this paper we develop a procedure for effectively learning the VAR model from small sample genomic data. In particular, we describe a novel model selection procedure for learning causal VAR networks from time course data with only a few time points, and no or little replication. This procedure is based on regularized estimation of VAR coefficients, followed by subsequent simultaneous significance testing of the corresponding partial correlation coefficients.

Once the VAR model has been learned from the data, it allows to elucidate possible underlying causal mechanisms by inspecting the Granger causality graph implied by the non-zero VAR coefficients.

The remainder of the paper is organized as follows. In the next section we first give the definition of a vector autoregressive process and recall the standard estimation. Subsequently, we describe our approach to regularized inference and to network model selection. This is followed by computer simulations comparing a variety of alternative approaches. Finally, we analyze data from an *Arabidopsis thaliana* expression time course experiment.

E.2 Methods

E.2.1 Vector Autoregressive Model

We consider vector-valued time series data $\boldsymbol{x}(t) = (x_1(t), \ldots, x_p(t))$. Each component of this row vector corresponds to a variable of interest, e.g., the expression level of a specific gene or the concentration of some metabolite in dependence of time. The vector autoregressive model specifies that the value of $\boldsymbol{x}(t)$ is a linear combination of those of earlier time points, plus noise,

$$\boldsymbol{x}(t) = \boldsymbol{c} + \sum_{i=1}^{m} \boldsymbol{x}(t - iL)\, \boldsymbol{B}_i + \boldsymbol{\epsilon}_i. \tag{E.1}$$

In this formula m is the order of the VAR process, L the time lag, and $\boldsymbol{c}$ a $1 \times p$ vector of means. The errors $\boldsymbol{\epsilon}_i$ are assumed to have zero mean and a $p \times p$ positive definite covariance matrix $\boldsymbol{\Sigma}$. The matrices $\boldsymbol{B}_i$ with dimension $p \times p$ represent the dynamical structure and thus contain the information relevant for reading off the causal relationships.

The autoregressive model has the form of a standard regression problem. Therefore, estimation of the matrices $\boldsymbol{B}_i$ is straightforward. A special case considered in this paper is when both m and L are set to 1. Then the above equation reduces to the VAR(1) process

$$\boldsymbol{x}(t+1) = \boldsymbol{c} + \boldsymbol{x}(t)\ \boldsymbol{B} + \boldsymbol{\epsilon}. \tag{E.2}$$

We now denote the centered *matrices of observations* corresponding to $\boldsymbol{x}(t+1)$ and $\boldsymbol{x}(t)$ by $\boldsymbol{X}_f$ ("future") and $\boldsymbol{X}_p$ ("past"), respectively, i.e. $\boldsymbol{X}_p = \begin{bmatrix} \boldsymbol{x}(1) \\ \dots \\ \boldsymbol{x}(n-1) \end{bmatrix}$ and $\boldsymbol{X}_f = \begin{bmatrix} \boldsymbol{x}(2) \\ \dots \\ \boldsymbol{x}(n) \end{bmatrix}$. In this notation the ordinary least squares (OLS) estimate can be written as

$$\hat{\boldsymbol{B}}^{\mathrm{OLS}} = ({\boldsymbol{X}_p}^T \boldsymbol{X}_p)^{-1} {\boldsymbol{X}_p}^T \boldsymbol{X}_f. \tag{E.3}$$

This is also the maximum likelihood (ML) estimate assuming the normal distribution. The coefficients of higher-order VAR models may be obtained in a corresponding fashion (Lütkepohl, 1993).

E.2.2 Small Sample Estimation Using James-Stein-Type Shrinkage

Genomic time course data contain only few time points (typically around $n = 10$) and often little or no replication – hence the above restriction on VAR(1) models with unit lag. Furthermore, it is known that for small sample size the least squares and maximum likelihood methods lead to statistically inefficient estimators. Therefore, application of the VAR model to genomics data requires some form of regularization. For instance, a full Bayesian approach could be used. However, for the VAR model the choice of a suitable prior is difficult (Ni and Sun, 2005).

Here, as a both computationally and statistically efficient alternative, we propose to apply James-Stein-type shrinkage, a method related to empirical Bayes (Efron and Morris, 1973; Opgen-Rhein and Strimmer, 2006a). This procedure has the advantage that it is computationally as simple as OLS, yet still produces efficient estimates for small samples.

Following (Opgen-Rhein and Strimmer, 2006a; Schäfer and Strimmer, 2005b) we now review how an unconstrained covariance matrix may be estimated using shrinkage. In the next section we then show how this result may be used to obtain shrinkage estimates of VAR coefficients.

Assuming centered data $\boldsymbol{X}$ for p variables (columns) the unbiased empirical estimator of the covariance matrix is $\boldsymbol{S} = \frac{1}{n-1}\boldsymbol{X}^T\boldsymbol{X}$ For small number of observations $\boldsymbol{S}$ is known to be inefficient and also ill-conditioned (singular!) for $n < p$. A more efficient estimator may be furnished by shrinking the empirical correlations r_{ij} towards zero and the empirical variances v_i against their median. This leads to the following expressions for the components of a shrinkage estimate $\boldsymbol{S}^\star$:

$$s^\star_{kl} = r^\star_{kl}\sqrt{v^\star_k v^\star_l} \tag{E.4}$$

with

$$r^{\star}_{kl} = (1 - \hat{\lambda}^{\star}_1) r_{kl} \tag{E.5}$$

$$v^{\star}_{k} = \hat{\lambda}^{\star}_2 v_{median} + (1 - \hat{\lambda}^{\star}_2) v_k \tag{E.6}$$

and

$$\hat{\lambda}^{\star}_1 = \min(1, \frac{\sum_{k \neq l} \widehat{\mathrm{Var}}(r_{kl})}{\sum_{k \neq l} r^2_{kl}}) \tag{E.7}$$

$$\hat{\lambda}^{\star}_2 = \min(1, \frac{\sum_{k=1}^{p} \widehat{\mathrm{Var}}(v_k)}{\sum_{k=1}^{p} (v_k - v_{\mathrm{median}})^2}). \tag{E.8}$$

The particular choice of the shrinkage intensities $\hat{\lambda}^{\star}_1$ and $\hat{\lambda}^{\star}_2$ is aimed at minimizing the overall mean squared error.

E.2.3 Shrinkage Estimation of VAR Coefficients

Small sample shrinkage estimates of VAR regression coefficients may be obtained by appropriately substituting the empirical by the shrinkage covariance. More specifically, we need to proceed as follows:

1. We combine the centered observations $\boldsymbol{X}_p$ and $\boldsymbol{X}_f$ into a joint matrix $\Phi = [\boldsymbol{X}_p \boldsymbol{X}_f]$. Note that Φ contains twice as many columns as either $\boldsymbol{X}_p$ or $\boldsymbol{X}_f$.

2. Next, we consider the $(n-1)$ multiple of the *empirical* covariance matrix, $\boldsymbol{S} = \Phi^T \Phi$, noting that $\boldsymbol{S}$ contains the two submatrices $\boldsymbol{S}_1 = {\boldsymbol{X}_p}^T \boldsymbol{X}_p$ and $\boldsymbol{S}_2 = {\boldsymbol{X}_p}^T \boldsymbol{X}_f$. This allows to write the OLS estimate of VAR coefficients as $\hat{\boldsymbol{B}}^{\mathrm{OLS}} = (\boldsymbol{S}_1)^{-1} \boldsymbol{S}_2$.

3. We replace the empirical covariance matrix $\boldsymbol{S}$ by a shrinkage estimate.

4. From $\boldsymbol{S}^{\star}$ we determine the submatrices $\boldsymbol{S}^{\star}_1$ and $\boldsymbol{S}^{\star}_2$ which in turn allow to compute the estimates $\hat{\boldsymbol{B}}^{\mathrm{Shrink}} = (\boldsymbol{S}^{\star}_1)^{-1} \boldsymbol{S}^{\star}_2$.

By decomposing $\boldsymbol{S}^{*}$ using the SVD or Cholesky algorithm it is possible to reconstruct pseudodata matrices $\boldsymbol{X}^{\star}_f$ and $\boldsymbol{X}^{\star}_p$. The above algorithm may be interpreted as OLS or normal-distribution ML based on these pseudodata.

E.2.4 VAR Network Model Selection

The network representing potential directed causal influences is given by the non-zero entries in the matrix of VAR coefficient. For an extensive discussion of the meaning and interpretation of the implied Granger (non)-causality we refer to (Granger, 1980).

As $\hat{\boldsymbol{B}}^{\text{Shrink}}$ is an estimate it is unlikely that any of its components are exactly zero. Therefore, we need to statistically test whether the entries of $\hat{\boldsymbol{B}}^{\text{Shrink}}$ are vanishing. However, instead of inspecting regression coefficients directly, it is preferably to test the corresponding partial correlation coefficients: this facilitates small-sample testing and additionally allows to accommodate for dependencies among the estimated coefficients (Schäfer and Strimmer, 2005a).

Specifically, consider in the VAR(1) model the multiple regression that connects the first variable $x_1(t+1)$ at time $t+1$ with all variables $x_1(t), \ldots, x_p(t)$ at the previous time t,

$$x_1(t+1) = c + \beta_k^1 x_k(t) + \sum_{j=1, j\neq k}^{p} \beta_j^1 x_j(t) + \text{error}. \tag{E.9}$$

If in this equation the roles of $x_k(t)$ and $x_1(t+1)$ are reversed,

$$x_k(t) = c + \beta_1^k x_1(t+1) + \sum_{j=1, j\neq k}^{p} \beta_1^j x_j(t) + \text{error}, \tag{E.10}$$

the partial correlation between the two variables is the geometric mean of the corresponding regression coefficients, times their sign, i.e. $\sqrt{\beta_1^k \beta_k^1}\text{sgn}(\beta_1^k)$ Whittaker (1990).

Once the partial correlations in the VAR model are computed, we use the *"local fdr"* approach of (Efron, 2005) to identify significant partial correlations, similar to the network model selection for graphical Gaussian models (GGMs) of Schäfer and Strimmer (2005a). Note that unlike in a GGM the edges in a VAR network are directed.

We point out that recently two papers have appeared describing related strategies for VAR model selection. As in our algorithm the strategies pursued in both (Demiralp and Hoover, 2003) and (Moneta, 2004) also consist in choosing the VAR network by selecting the appropriate underlying partial correlations. However, the key advantage of our variant of VAR network search is that it is specifically designed to meet small sample requirements, by using shrinkage estimators of regression coefficients and partial correlation, and due to the adaptive nature (i.e. the automatic estimation of the empirical null) of the "local fdr" algorithm (Efron, 2005).

E.3 Results and Discussion

E.3.1 Simulation Study

In a comparative simulation study we investigated the power of diverse approaches to recovering the true VAR network. We simulated VAR(1) data of different sample size, with n varying between 5 and 200, for 100 randomly generated true networks with 200 edges and $p = 100$ nodes. The 200 nonzero regression coefficients were drawn uniformly from the intervals $[-1; -0.2]$ and $[0.2; 1]$.

In addition to the shrinkage procedure we estimated regression coefficients by ordinary least squares (OLS) and by ridge regression (RR). All these three regression strategies were

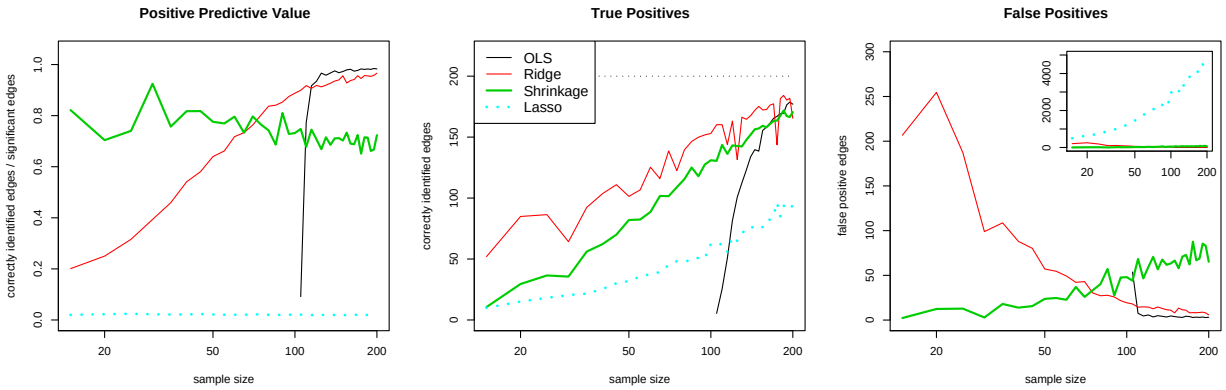

Figure E.1: Relative performance of the four investigated methods for learning VAR networks in terms of positive predictive value (true discovery rate) and the number of true and false edges. The thin dotted line in the middle box at 200 corresponds to the true number of edges in the simulated networks.

applied in conjunction with the above VAR model selection based on partial correlations, with a cutoff value for the "local fdr" statistic set at 0.2 – the recommendation of (Efron, 2005). As a fourth method we employed L1 regression (Tibshirani, 1996) (LASSO) to estimate VAR regression coefficients. Note that in the latter instance there is no need for additional model selection, as the LASSO method combines shrinkage and model selection and automatically sets many regression coefficients identically to zero.

In the simulations we ran OLS only for $n > 100$, as for small sample size the corresponding empirical covariance matrix is singular and consequently the OLS regression is ill-posed. The penalty for the LASSO regression was chosen as in (Meinshausen and Bühlmann, 2006). The regularization parameter in RR was determined by generalized cross validation (Golub et al., 1979). Unfortunately, even GCV turned out to be computationally expensive, so that for RR we conducted only 10 repetitions, rather than the 100 considered for the other methods.

The results of the simulations are summarized in Figure E.1. The left box shows the positive predictive value, or true discovery rate of the four methods. This is the proportion of correctly identified edges in relation to all significant edges. Our proposed shrinkage algorithm is the only method achieving around 80% positive predictive value regardless of the sample size. Note that this is exactly the theoretically expected value, given the specified "local fdr" cutoff of 0.2. In contrast, the RR and LASSO methods perform remarkably poor at small sample size, with much lower true discovery rates. For medium to large sample size the OLS estimation dominates RR, LASSO and the shrinkage approach. This is easily explained by the fact that OLS has no parameters to optimize and that it is asymptotically optimal. However, it is bothering that for both the RR and the OLS approach the false discovery rate appears not to be properly controlled. Finally, for large sample size the Stein-type estimator appears to be prone to overshrinking, which leads to

an increase of false positives.

The relative performance of the four approaches to VAR estimation can be further explained by considering the relative amount of true and false positive edges (Figure E.1, middle and right box). The shrinkage method generally produces very few false positives. In contrast, the RR and LASSO methods lead to a large number of false edges, especially for small sample size. This is particularly pronounced for the LASSO regression, as can be seen in the differently scaled inlay plot contained in the right box of Figure E.1, indicating that the penalty applied in the L1 regression may not be sufficient in this situation. In terms of the number of correctly identified edges the RR and shrinkage approach are the two top performing methods. However, even though RR finds a considerable number of true edges even at very small sample size, this has little impact on its true discovery rate because of the high number of false positives.

In summary, the simulation results suggest to apply for small sample size the James-Stein-type shrinkage procedure, and for $n > p$ the traditional OLS approach.

E.3.2 Analysis of a Microarray Time Course Data Set

For further illustration we applied the VAR shrinkage approach to a real world data example. Specifically, we reanalyzed expression time series resulting from an experiment investigating the impact of the diurnal cycle on the starch metabolism of *Arabidopsis thaliana* (Smith et al., 2004).

We downloaded the calibrated signal intensities for 22,814 probes and 11 time points for each of the two biological replicates from the NASCArrays repository (`http://affymetrix.arabidopsis.info/narrays/experimentpage.pl?experimentid=60`). After log-transforming the data we filtered out all genes containing missing values and whose maximum signal intensity value was lower than 5 on a log-base 2 scale. Subsequently, we applied the periodicity test of (Wichert et al., 2004) to identify the probes associated with the day-night cycle. As a result, we obtained a subset of 800 genes that we further analyzed with the VAR approach.

We note that a tacit assumption of the VAR model is that time points are equidistant – see Eq. E.1. This is not the case for the *Arabidopsis thaliana* data which were measured at $0, 1, 2, 4, 8, 12, 13, 14, 16, 20,$ and 24 hours after the start of the experiment. However, as the intensity of the biological reactions is likely to be higher at the change points from light to dark periods (time points 0 and 12), one may argue that assuming equidistant measurements is justifiable at least in terms of equal relative reaction rate.

A further implication of the VAR model (and indeed of many other graphical models) is that dependencies among genes are essentially linear. This can easily be checked by inspecting the pairwise scatter plots of the calibrated expression levels. For the 800 considered *Arabidopsis thaliana* genes we verified that the linearity assumption of the VAR model is indeed satisfied.

Subsequently, we estimated from the replicate time series of the 800 preselected genes the regularized regression coefficients and the corresponding partial correlations, and identified the significant edges of the VAR causal graph as described above. We found a total

number of 7,381 significant edges connecting 707 nodes. In Figure E.2 we show for reasons of clarity only the subnetwork containing the 150 most significant edges, which connect 92 nodes. Note that this graph exhibits a clear "hub" connectivity structure (nodes filled with red color) , which is particularly striking for the node 570 but also for nodes 81, 558, 783 and a few other genes (for annotation of the nodes see the *Supplementary Information*, Table 1).

As the VAR network contains directed edges it is possible to distinguish genes that have mostly outgoing arcs, which could be indicative for key regulatory genes, from those with mostly ingoing arcs. In the graph of Figure E.2 node 570, an AP2 transcription factor, and node 81, a gene involved in DNA-directed RNA polymerase, belong to the former category, whereas for instance node 558, a structural constituent of ribosome, seems to be part of the latter. Node 627 is another hub in the VAR network, which according to the annotation of (Smith et al., 2004) encodes a protein of unknown function. Another interesting aspect of the VAR network is the web of highly connected genes (encircled and colored yellow in the lower right corner of Figure E.2) which we hypothesize to constitute some form of a functional module.

Finally, we note that the VAR network visualizes influences of the genes over time, hence a VAR graph can also include directed loops and even genes that act upon themselves. In contrast, the GGM graphs discussed in (Schäfer and Strimmer, 2005b,a) visualize the partial correlation with no time lag involved. For comparison, we display the GGM graph for the *Arabidopsis thaliana* data in Figure E.3. As expected, both graphs share the same structure (main hubs and the module of highly connected genes): if one node influences another in the next timepoint with a constant regression coefficient (VAR-model), they also tend to be significantly partially correlated in the same time point (GGM-model). However, using a GGM it is not possible to infer the causal structure of the network.

E.4 Conclusions

We have presented a novel algorithm for learning VAR causal networks. This is based on James-Stein-type shrinkage estimation of covariances between different time points of the conducted experiment, that in turn leads to improved estimates of the VAR regression coefficients. Subsequent VAR model selection is conducted by using "local fdr" multiple testing on the corresponding partial correlations.

We have shown that this approach is well suited for the small sample sizes encountered in genomics. In addition, the approach is computationally very efficient, as no computer intensive sampling or optimization is needed: the inference of the directed network for the *Arabidopsis thaliana* data with 640,000 potentially directed edges takes about one minute on a standard desktop computer. While we have illustrated the approach by analyzing a microarray expression data set, it is by no means restricted to this kind of data - we expect that our VAR network approach performs equally well for similar high dimensional time series data from metabolomic or proteomic experiments.

The current algorithm employs a fixed "one step ahead" time lag. One strategy to gen-

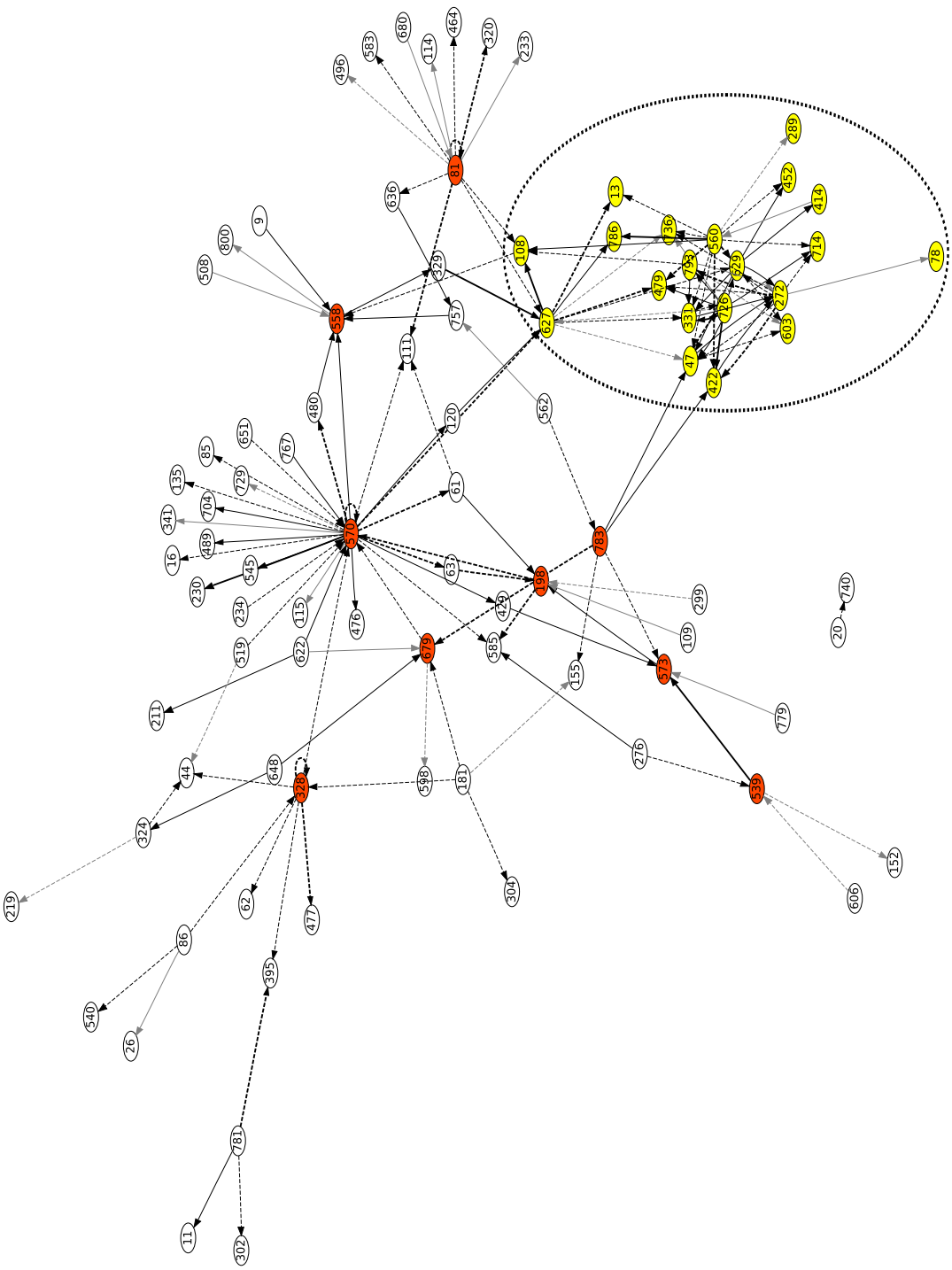

Figure E.2: Directed VAR network inferred from the *Arabidopsis thaliana* data. The solid and dotted lines indicate positive and negative regression coefficients, respectively, and the line intensity denotes their strength. For annotation of the nodes see the *Supplementary Information*, Table 1. The color code of the nodes is explained in the main text.

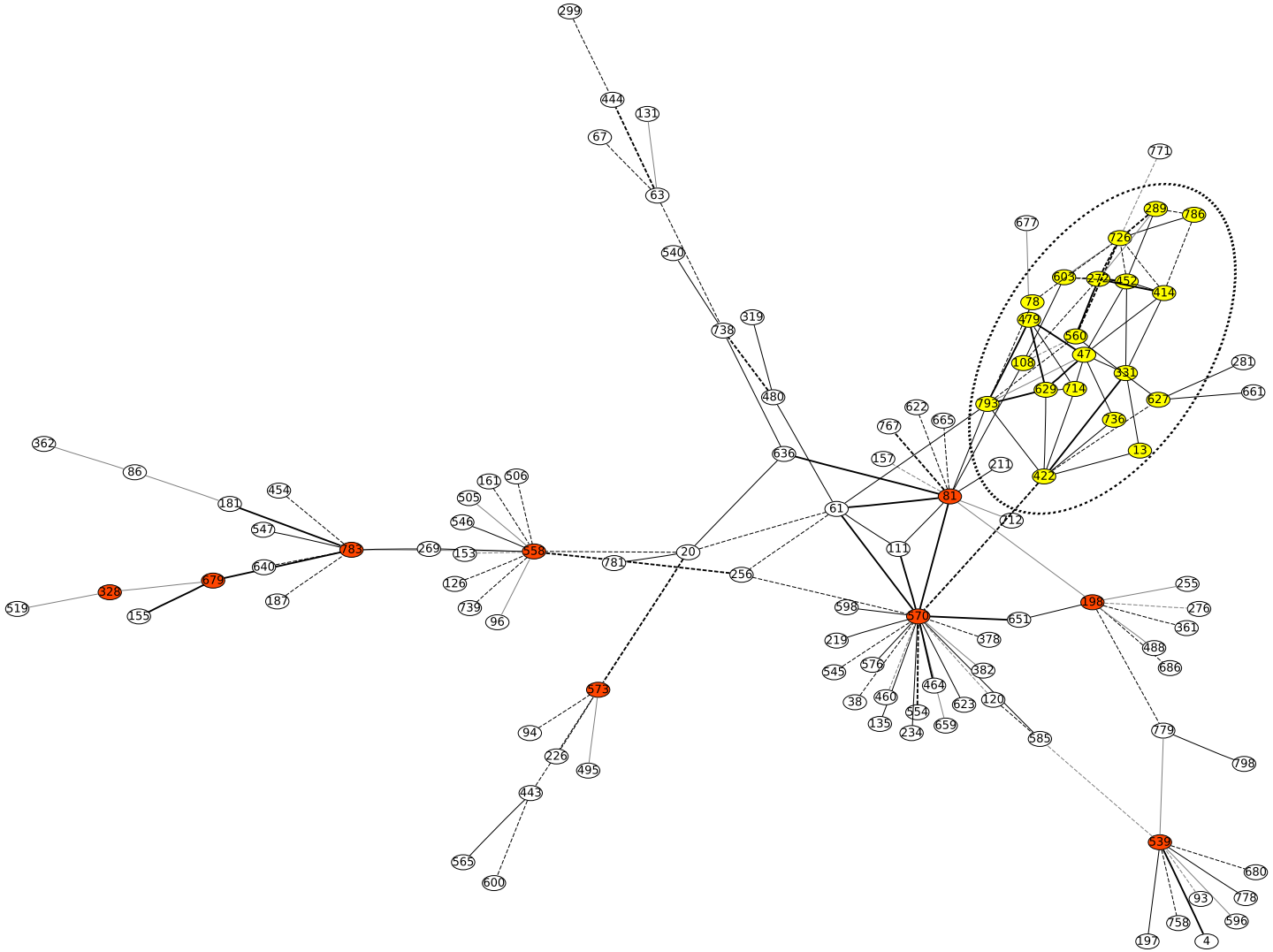

Figure E.3: Undirected GGM network inferred from the *Arabidopsis thaliana* data using the algorithm of Schäfer and Strimmer (2005a,b). The solid and dotted lines indicate positive and negative partial correlation coefficients, respectively, and the line intensity denotes their strength.

eralization to arbitrary time lags may be to consider functional data – see, e.g., (Opgen-Rhein and Strimmer, 2006b,c). This would have the additional benefit to suitable deal with non-equally spaced measurements, a common characteristic of many biological experiments.

Authors contributions

Both authors participated in the development of the methodology and wrote the manuscript. R.O. carried out all analyzes and simulations. All authors approved of the final version of the manuscript.

Acknowledgements

We thank Papapit Ingkasuwan for pointing us to the *Arabidopsis thaliana* data set, and the referees for their valuable comments. This work was supported by the Deutsche Forschungsgemeinschaft (DFG) via an Emmy-Noether research grant to K.S.

Supplementary Information

Annotation data (probe IDs, gene names, description) for the 92 genes displayed in the VAR network of Figure 2.

Node	Probe ID	Gene Name	Description
9	267432_at	At2g35020-MinT-G	putative UDP-N-acetylglucosamine pyrophosphorylase
11	267383_at	AT2G44360-TAIR-G	unknown protein; expressed protein
13	267341_at	AT2G44200-TAIR-G	unknown protein; expressed protein
16	267123_at	AT2G23560-TAIR-G	catalytic/ hydrolase; hydrolase, alpha/beta fold family protein, similar to ethylene-induced esterase (Citrus sinensis) GI:14279437, polyneuridine aldehyde esterase (Rauvolfia serpentina) GI:6651393; contains Pfam profile PF00561: hydrolase, alpha/beta fold family
20	266995_at	At2g34500-MinT-G	putative cytochrome P450
26	266897_at	AT2G45820-TAIR-G	DNA binding; DNA-binding protein, putative, identical to DNA-binding protein gi—601843—gb—AAA57124 (Arabidopsis thaliana); contains Pfam domain, PF03766: Remorin, N-terminal region; contains Pfam domain, PF03763: Remorin, C-terminal region

44	266314_at	AT2G27040-TAIR-G	AGO4 (ARGONAUTE 4); PAZ domain-containing protein / piwi domain-containing protein, similar to SP—Q9QZ81 Eukaryotic translation initiation factor 2C 2 Rattus norvegicus; contains Pfam profiles PF02171: Piwi domain, PF02170: PAZ domain
47	266247_at	AT2G27660-TAIR-G	unknown protein; DC1 domain-containing protein, contains Pfam profile PF03107: DC1 domain
61	265721_at	AT2G40090-TAIR-G	ATATH9; member of ATH subfamily
62	265695_at	AT2G24490-TAIR-G	nucleic acid binding; replication protein, putative, similar to replication protein A 30kDa (Oryza sativa (japonica cultivar-group)) GI:13516746; contains InterPro entry IPR004365: OB-fold nucleic acid binding domain
63	265674_at	AT2G32190-TAIR-G	unknown protein; expressed protein
78	264986_at	At1g27130-MinT-G	glutathione transferase, putative
81	264924_at	AT1G60620-TAIR-G	ATRPAC43; DNA binding / DNA-directed RNA polymerase; DNA-directed RNA polymerase, putative, identical to RNA polymerase subunit (Arabidopsis thaliana) GI:514324; contains Pfam profile PF01000: RNA polymerase Rpb3/RpoA insert domain
85	264837_at	AT1G03600-TAIR-G	unknown protein; photosystem II family protein, similar to SP:P74367 Synechocystis sp.; similar to ESTs emb—Z27038, gb—AA451546, emb—Z29876, gb—T45359 and gb—R90316
86	264832_at	AT1G03660-TAIR-G	unknown protein; expressed protein, contains 2 predicted transmembrane domains; expression supported by MPSS
108	264179_at	AT1G02180-TAIR-G	unknown protein; ferredoxin-related, similar to Ferredoxin. (SP:O78510) (Cryptomonas phi) Guillardia theta
109	264131_at	At1g79150-MinT-G	hypothetical protein
111	264063_at	At2g27910-MinT-G	unknown protein
114	264038_at	At2g03690-MinT-G	putative ubiquinone biosynthesis protein
115	264004_at	At2g22425-MinT-G	unknown protein
120	263852_at	At2g04450-MinT-G	putative mutT domain protein
135	263489_at	At2g31830-MinT-G	putative inositol polyphosphate 5'-phosphatase
152	263193_at	AT1G36050-STG	unknown protein
155	263047_at	AT2G17630-TAIR-G	phosphoserine transaminase/ transaminase; phosphoserine aminotransferase, putative, similar to Phosphoserine aminotransferase, chloroplast precursor (PSAT) (SP:Q96255) (Arabidopsis thaliana); contains TIGRFAM TIGR01364: phosphoserine aminotransferase; contains Pfam PF00266: aminotransferase, class V

181	262501_at	At1g21690-MinT-G	putative replication factor C subunit
198	262134_at	AT1G77990-TAIR-G	AST56; sulfate transporter; cDNA encoding a sulfate transporter.
211	261810_at	At1g08130-MinT-G	DNA ligase
219	261696_at	AT1G08470-TAIR-G	strictosidine synthase; strictosidine synthase family protein, similar to strictosidine synthase (Rauvolfia serpentina)(SP—P15324)
230	261484_at	AT1G14400-TAIR-G	UBC1 (UBIQUITIN CARRIER PROTEIN 1); ubiquitin conjugating enzyme/ ubiquitin-like activating enzyme; ubiquitin carrier protein (UBC1) mRNA, complete cds
233	261440_at	AT1G28510-TAIR-G	unknown protein; expressed protein
234	261425_at	At1g18880-MinT-G	unknown protein
272	260143_at	AT1G71880-TAIR-G	SUC1; carbohydrate transporter/ sucrose:hydrogen symporter/ sugar porter; member of Sucrose-proton symporter family
276	260099_at	AT1G73180-TAIR-G	unknown protein; eukaryotic translation initiation factor-related, similar to eukaryotic translation initiation factor 2A (GI:21956484) (Homo sapiens); similar to Eukaryotic translation initiation factor 3 subunit 9 (eIF-3 eta) (eIF3 p116) (eIF3 p110) (eIF3b) (Swiss-Prot:P55884) (Homo sapiens)
289	259875_s_at	No gene	
299	259645_at	AT1G69010-TAIR-G	DNA binding / transcription factor; basic helix-loop-helix (bHLH) family protein, contains Pfam profile: PF00010 helix-loop-helix DNA-binding domain
302	259488_at	AT1G15780-TAIR-G	unknown protein; expressed protein
304	259406_at	At1g17690-MinT-G	unknown protein
320	259111_at	At3g05520-MinT-G	alpha subunit of F-actin capping protein
324	259069_at	At3g11710-MinT-G	lysyl-tRNA synthetase
328	258871_at	AT3G03060-TAIR-G	ATP binding / ATPase/ nucleoside-triphosphatase/ nucleotide binding; AAA-type ATPase family protein, contains a ATP/GTP-binding site motif A (P-loop), PROSITE:PS00017
329	258781_at	AT3G11740-TAIR-G	unknown protein; expressed protein, similar to expressed protein [Arabidopsis thaliana] (TAIR:At5g01750.2); similar to hypothetical protein [Oryza sativa (japonica cultivar-group)] (GB:XP_470102.1); contains InterPro domain Protein of unknown function DUF567 (InterPro:IPR007612)
331	258764_at	AT3G10720-TAIR-G	enzyme inhibitor/ pectinesterase; pectinesterase, putative, contains similarity to pectinesterase from Vitis vinifera GI:15081598, Prunus persica SP—Q43062; contains Pfam profile PF01095 pectinesterase
341	258432_at	At3g16570-MinT-G	unknown protein

395	256626_at	AT3G20015-TAIR-G	pepsin A; similar to aspartyl protease family protein [Arabidopsis thaliana] (TAIR:At3g18490.1); similar to putative aspartic proteinase nepenthesin I [Oryza sativa (japonica cultivar-group)] (GB:NP_909181.1); contains InterPro domain Aspartic protease A1, pepsin (InterPro:IPR001461)
414	256169_at	At1g51800-MinT-G	unknown protein
422	255844_at	AT2G33580-TAIR-G	kinase; protein kinase family protein / peptidoglycan-binding LysM domain-containing protein, protein kinase (Arabidopsis thaliana) GI:2852449; contains Pfam profiles PF01476: LysM domain, PF00069: Protein kinase domain
429	255645_at	AT4G00880-TAIR-G	unknown protein; auxin-responsive family protein, similar to small auxin up RNA (GI:546362) Arabidopsis thaliana
452	254785_at	At4g12730-MinT-G	fasciclin-like arabinogalactan protein FLA2
464	254515_at	At4g20270-MinT-G	CLV1 receptor kinase like protein
476	254239_at	AT4G23400-TAIR-G	PIP1;5/PIP1D; water channel; major intrinsic family protein / MIP family protein, contains Pfam profile: MIP PF00230
477	254233_at	AT4G23800-TAIR-G	transcription factor; high mobility group (HMG1/2) family protein, similar to HMG2B (Homo sapiens) GI:32335; contains Pfam profile PF00505: HMG (high mobility group) box
479	254211_at	AT4G23570-TAIR-G	SGT1A; Closely related to SGT1B, may function in SCF(TIR1) mediated protein degradation
480	254162_at	At4g24440-MinT-G	transcription factor IIA small subunit
489	253927_at	AT4G26710-TAIR-G	hydrogen-transporting ATP synthase, rotational mechanism / hydrogen-transporting ATPase, rotational mechanism; ATP synthase subunit H family protein, contains similarity to Swiss-Prot:O15342 Vacuolar ATP synthase subunit H (V-ATPase H subunit) (Vacuolar proton pump H subunit) (V-ATPase M9.2 subunit) (V-ATPase 9.2 kDa membrane accessory protein) (Homo sapiens)
496	253708_at	At4g29210-MinT-G	gamma-glutamyltransferase-like protein
508	253523_at	AT4G31340-TAIR-G	unknown protein; similar to DNA repair ATPase-related [Arabidopsis thaliana] (TAIR:At2g24420.2); similar to DNA repair ATPase-related [Arabidopsis thaliana] (TAIR:At2g24420.1); similar to unknown [Oryza sativa (japonica cultivar-group)] (GB:AAO72581.1)
519	253202_at	AT4G34555-TAIR-G	structural constituent of ribosome; 40S ribosomal protein S25, putative

539	252427_at	AT3G47640-TAIR-G	DNA binding / transcription factor; basic helix-loop-helix (bHLH) family protein, contains Pfam profile: PF00010 helix-loop-helix DNA-binding domain
540	252420_at	At3g47530-MinT-G	putative protein
545	252098_at	At3g51330-MinT-G	predicted GPI-anchored protein
558	251834_at	AT3G55170-TAIR-G	structural constituent of ribosome; 60S ribosomal protein L35 (RPL35C), various ribosomal L35 proteins
560	251786_at	At3g55270-MinT-G	MAP kinase phosphatase (MKP1)
562	251768_at	AT3G55940-TAIR-G	phosphoinositide phospholipase C/ phospholipase C; phosphoinositide-specific phospholipase C, putative, similar to phosphoinositide specific phospholipase C GI:857374 from (Arabidopsis thaliana)
570	251598_at	At3g57600-MinT-G	AP2 transcription factor - like protein
573	251483_at	At3g59650-MinT-G	unknown protein
583	251227_at	At3g62700-MinT-G	ABC transporter-like protein
585	251123_at	AT5G01030-TAIR-G	unknown protein; expressed protein
598	250661_at	At5g07030-MinT-G	nucleoid DNA-binding-like protein
603	250520_at	At5g08470-MinT-G	putative protein
606	250433_at	AT5G10400-TAIR-G	DNA binding; histone H3, identical to several histone H3 proteins, including Zea mays SP—P05203, Medicago sativa GI:166384, Encephalartos altensteinii SP—P08903, Pisum sativum SP—P02300; contains Pfam profile PF00125 Core histone H2A/H2B/H3/H4
622	249997_at	AT5G18620-TAIR-G	ATP binding / ATP-dependent helicase/ DNA binding / DNA-dependent ATPase/ helicase/ nucleic acid binding; DNA-dependent ATPase, putative, similar to DNA-dependent ATPase SNF2H (Mus musculus) GI:14028669; contains Pfam profiles PF00271: Helicase conserved C-terminal domain, PF00176: SNF2 family N-terminal domain, PF00249: Myb-like DNA-binding domain
627	249817_at	AT5G23820-TAIR-G	unknown protein; MD-2-related lipid recognition domain-containing protein / ML domain-containing protein, contains Pfam profile PF02221: ML domain
629	249777_at	AT5G24210-TAIR-G	triacylglycerol lipase; lipase class 3 family protein, contains Pfam profile PF01764: Lipase
636	249569_at	At5g38070-MinT-G	putative protein
648	249315_at	At5g41190-MinT-G	unknown protein
651	249211_at	At5g42680-MinT-G	putative protein
679	248295_at	At5g53070-MinT-G	unknown protein (At5g53070)

680	248248_at	AT5G53120-TAIR-G	SPDS3 (SPERMIDINE SYNTHASE 3); encodes a novel spermine synthase and is a paralog of previously characterized spermidine synthases, SPDS1 and SPDS2. SPDS3 forms heterodimers with SDPS2, which in turn forms heterodimers with SDPS1 in vivo. The gene does not complement speDelta3 deficiency of spermidine synthase in yeast but DOES complement speDelta4 deficiency.
704	247791_at	AT5G58710-TAIR-G	ROC7; peptidyl-prolyl cis-trans isomerase; cyclophilin (ROC7) mRNA, complete cds
714	247554_at	AT5G61010-TAIR-G	protein binding; similar to exocyst subunit EXO70 family protein [Arabidopsis thaliana] (TAIR:At3g29400.1); similar to putative leucine zipper-containing protein [Oryza sativa (japonica cultivar-group)] (GB:XP_465879.1); contains InterPro domain Exo70 exocyst complex subunit (InterPro:IPR004140)
726	247097_at	AT5G66460-TAIR-G	hydrolase, hydrolyzing O-glycosyl compounds; (1-4)-beta-mannan endohydrolase, putative, similar to (1-4)-beta-mannan endohydrolase (Coffea arabica) GI:10178872; contains Pfam profile PF00150: Cellulase (glycosyl hydrolase family 5)
729	247055_at	At5g66740-MinT-G	putative protein
736	246976_s_at	At5g24810-MinT-G	unknown protein
740	246837_at	At5g26670-MinT-G	pectin acetylesterase precursor - like protein
757	246421_at	At5g16880-MinT-G	TOM (target of myb1) -like protein
767	246043_at	AT5G19380-TAIR-G	unknown protein; expressed protein
779	245619_at	AT4G13990-TAIR-G	catalytic; exostosin family protein, contains Pfam profile: PF03016 exostosin family
781	245435_at	At4g17130-MinT-G	hypothetical protein
783	245404_at	AT4G17610-TAIR-G	RNA binding / RNA methyltransferase; tRNA/rRNA methyltransferase (SpoU) family protein, similar to TAR RNA loop binding protein (Homo sapiens) GI:1184692; contains Pfam profile PF00588: SpoU rRNA Methylase (RNA methyltransferase, TrmH) family
786	245347_at	AT4G14890-TAIR-G	electron carrier/ electron transporter/ iron ion binding; ferredoxin family protein, similar to SP—P00252 Ferredoxin I from Nostoc muscorum, SP—P00248 Ferredoxin from Mastigocladus laminosus, SP—P00244 Ferredoxin I from Aphanizomenon flos-aquae; contains Pfam profile PF00111 2Fe-2S iron-sulfur cluster binding domain
793	245218_s_at	No gene	
800	244996_at	ATCG00160-TAIR-G	RPS2 (RIBOSOMAL PROTEIN S2); Chloroplast ribosomal protein S2

Bibliography

Bay, S., Chrisman, L., Pohorille, A., and Shrager, J. C. (2004). Temporal aggregation bias and inference of causal regulatory networks. *J. Comp. Biol.*, 11(5):971–985.

Demiralp, S. and Hoover, K. D. (2003). Searching for the causal structure of a vector autoregression. *Oxford Bull. Econom. Statist.*, 65:745–767.

Efron, B. (2005). Local false discovery rates. Technical Report 2005-20B/234, Dept. of Statistics, Stanford University.

Efron, B. and Morris, C. N. (1973). Stein's estimation rule and its competitors – an empirical Bayes approach. *J. Amer. Statist. Assoc.*, 68:117–130.

Golub, G. H., Heath, M., , and Wahba, G. (1979). Generalized crossvalidation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223.

Granger, C. W. J. (1980). Testing for causality, a personal viewpoint. *J. Econom. Dyn. Control*, 2:329–352.

Lütkepohl, H. (1993). *Introduction to Multiple Time Series Analysis*. Springer, Berlin.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Ann. Statist.*, 34:1436–1462.

Moneta, A. (2004). Graphical models for structural vector autoregressions.

Ni, S. and Sun, D. (2005). Bayesian estimates for vector autoregressive models. *Journal of Business & Economic Statistics*, 23(1):105–117.

Opgen-Rhein, R. and Strimmer, K. (2006a). Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach. *Statist. Appl. Genet. Mol. Biol.*, 6(1):9.

Opgen-Rhein, R. and Strimmer, K. (2006b). Inferring gene dependency networks from genomic longitudinal data: a functional data approach. *REVSTAT*, 4:53–65.

Opgen-Rhein, R. and Strimmer, K. (2006c). Using regularized dynamic correlation to infer gene dependency networks from time-series microarray data. In *Proceedings of the 4th International Workshop on Computational Systems Biology (WCSB 2006)*, volume 4, pages 73–76, Tampere.

Schäfer, J. and Strimmer, K. (2005a). An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, 21(6):754–764.

Schäfer, J. and Strimmer, K. (2005b). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4(1):32.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica*, 48:1–48.

Smith, S. M., Fulton, D. C., Chia, T., Thorneycroft, D., Chapple, A., Dunstan, H., Hylton, C., and Smith, S. C. Z. A. M. (2004). Diurnal changes in the transcriptom encoding enzymes of starch metabolism provide evidence for both transcriptional and post-transcriptional regulation of starch metabolism in Arabidopsis leaves. *Plant Physiol.*, 136:2687–2699.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B*, 58:267–288.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley, New York.

Wichert, S., Fokianos, K., and Strimmer, K. (2004). Identifying periodically expressed transcripts in microarray time series data. *Bioinformatics*, 20(1):5–20.

Article F

From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data

Published in BMC Systems Biology *(Volume 1, Article 37, 2007)*

Authors: Rainer Opgen-Rhein *and* Korbinian Strimmer

Abstract:

- *Background:* The use of correlation networks is widespread in the analysis of gene expression and proteomics data, even though it is known that correlations not only confound direct and indirect associations but also provide no means to distinguish between cause and effect. For "causal" analysis typically the inference of a directed graphical model is required. However, this is rather difficult due to the curse of dimensionality.

  Results: We propose a simple heuristic for the statistical learning of a high-dimensional "causal" network. The method first converts a correlation network into a partial correlation graph. Subsequently, a partial ordering of the nodes is established by multiple testing of the log-ratio of standardized partial variances. This allows identifying a directed acyclic causal network as a subgraph of the partial correlation network. We illustrate the approach by analyzing a large *Arabidopsis thaliana* expression data set.

  Conclusions: The proposed approach is a heuristic algorithm that is based on a number of approximations, such as substituting lower order partial correlations by full order partial correlations. Nevertheless, for small samples and for sparse networks the algorithm not only yield sensible first order approximations of the causal structure in high-dimensional genomic data but is also computationally highly efficient.

  Availability and requirements: The method is implemented in the "GeneNet" R package (version 1.2.0), available from CRAN and from `http://strimmerlab.org/software/genets/`. The software includes an R script for reproducing the network analysis of the *Arabidopsis thaliana* data.

F.1 Background

Correlation networks are widely used to explore and visualize high-dimensional data, for instance in finance (Mantegna and Stanley, 2000; Onnela et al., 2004; Boginski et al., 2005), ecology (Shipley, 2000), gene expression analysis (Butte et al., 2000; Oldham et al., 2006), or metabolomics Steuer (2006). Their popularity is owed to a large extent to the ease with which a correlation network can be constructed, as this requires only two simple steps: i) the computation of all pairwise correlations for the investigated variables, and ii) a thresholding or filtering procedure (Tumminello et al., 2005) to identify significant correlations, and hence edges, of the network.

However, for shedding light on the causal processes underlying the observed data, correlation networks are only of limited use. This is due to the fact that correlations not only confound direct and indirect associations but also provide no means to distinguish between response variables and covariates (and thus between cause and effect).

Therefore, causal analysis requires tools different from correlation networks: much of the work in this area has focused on Bayesian networks (Pearl, 2000) or related regression models such as systems of recursive equations (Freedman, 2005; Wermuth, 1980) or influence diagrams (Schachter and Kenley, 1989). All of these models have in common that they describe causal relations by an underlying directed acyclic graph (DAG).

There already exist numerous methods for learning DAGs from observational data – see for instance the summarizing review in (Tsamardinos et al., 2006) and the references therein. However, with few exceptions (e.g., the PC algorithm, Spirtes et al., 2001; Kalisch and Bühlmann, 2007) virtually all of these methods have been devised for comparatively small numbers of variables and with large sample size in mind. For instance, the numerical example of the recently proposed algorithm described in (Shimizu et al., 2006) uses $n = 10,000$ observations for $p = 7$ variables. Unfortunately, the data that would be most interesting to explore with causal methods, namely those commonly visualized by correlation networks (see above), have completely different characteristics, in particular they are likely of high dimension.

In this paper we follow (Kalisch and Bühlmann, 2007) and focus on modeling large-scale linear recursive systems. Specifically, we present a simple discovery algorithm that enables the inference of causal relations from small sampled data and for large numbers of variables. It proceeds in two steps as follows:

- First, the correlation network is transformed into a partial correlation network, which is essentially an undirected graph that displays the direct linear associations only. This type of network model is also known under the names of graphical Gaussian model (GGM), concentration graph, covariance selection graph, conditional independence graph (CIG), or Markov random field. Note that there is a simple relationship between correlation and partial correlation. Moreover, in recent years there has been much progress with regard to statistical methodology for learning large-scale partial correlation graphs from small samples (e.g., de la Fuente et al., 2004; Dobra et al., 2004; Schäfer and Strimmer, 2005a,b; Wille and Bühlmann, 2006; Li and Gui, 2006).

Here we employ the approach described in (Schäfer and Strimmer, 2005b).

- Second, the undirected GGM is converted into a *partially* directed graph. This is done by estimating a pairwise ordering of the nodes from the data using multiple testing of the log-ratios of standardized partial variances, and by subsequent projection of this partial ordering onto the GGM. The inferred causal network is the subgraph containing all the directed edges.

Note that this algorithm is similar to the PC algorithm in that edges are being removed from the independence graph to obtain the underlying DAG. However, our criterion for eliminating an edge is distinctly different from that of the PC algorithm.

The remainder of the paper is organized as follows. First, we describe the methodology. Second we consider its statistical interpretation and further properties. Subsequently, we illustrate the approach by analyzing an 800 gene data set from a large-scale *Arabidopsis thaliana* gene expression experiment. Finally, we conclude with some discussion of the method, commenting also on the limitations of the approach.

F.2 Methods

F.2.1 Theoretical basis

Consider a linear regression with Y as response and $X_1, \ldots, X_k, \ldots, X_K$ as covariates. We assume that X_k and Y are random variables with known variances $\mathrm{Var}(Y)$ and $\mathrm{Var}(X_k)$ and with covariance $\mathrm{Cov}(Y, X_k)$. The best linear predictor of Y in terms of the X_k that minimizes the MSE of $\sum_k \beta_k X_k - Y$ is given by (e.g. ref. Cox and Wermuth, 1993, p. 206)

$$\beta_k^y = \tilde{\rho}_{yk} \sqrt{\frac{\tilde{\sigma}_y^2}{\tilde{\sigma}_k^2}}, \tag{F.1}$$

where $\tilde{\rho}_{yk}$ is the *partial* correlation between Y and X_k, and $\tilde{\sigma}_y^2$ and $\tilde{\sigma}_k^2$ are the respective *partial* variances.

The partial correlation is the correlation that remains between two variables if the effect of the other variables has been regressed away. Likewise, the partial variance is the variance that remains if the influences of all other variables are taken into account. Table F.1 lists the definitions and formulas for the computation of these quantities (note that in our notation a tilde on top of a symbol indicates "partial").

From Equation F.1 it is immediately clear that the complete linear system and thus all β_k^y are determined by the joint covariance matrix of Y and X_k (see also, e.g., Whittaker, 1990; Studený, 2005). For only a single dependent variable Equation F.1 reduces to the well-known relation $\beta_x^y = \rho_{yx}\sqrt{\sigma_y^2/\sigma_x^2}$, which contains only the unconditioned correlation and variances (without the tilde).

We emphasize that Equation F.1 has a direct relation with the usual ordinary least squares (OLS) estimator for the regression coefficient. This is recovered if the empirical

Table F.1: Formulas for computing partial variances and partial correlations.

	Definition	True value	Estimate
Covariance matrix:	$\mathrm{Cov}(X_k, X_l) = \sigma_{kl}$	$\boldsymbol{\Sigma} = (\sigma_{kl})$	$\boldsymbol{S} = (s_{kl})$
Concentration matrix:	$\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$	$\boldsymbol{\Omega} = (\omega_{kl})$	
Variances:	$\mathrm{Var}(X_k) = \sigma_{kk} = \sigma_k^2$	σ_{kk}	s_{kk}
Correlation matrix:	$\mathrm{Corr}(X_k, X_l) = \rho_{kl}$ $= \sigma_{kl}(\sigma_{kk}\sigma_{ll})^{-1/2}$	$\boldsymbol{P} = (\rho_{kl})$	$\boldsymbol{R} = (r_{kl})$
Partial variances	$\mathrm{Var}(X_k \vert X_{\neq k}) = \tilde{\sigma}_{kk} = \tilde{\sigma}_k^2 = \omega_{kk}^{-1}$	$\tilde{\sigma}_{kk}$	$\tilde{s}_{kk}$
Partial correlations:	$\mathrm{Corr}(X_k, X_l \vert X_{\neq k,l}) = \tilde{\rho}_{kl}$ $= -\omega_{kl}(\omega_{kk}\omega_{ll})^{-1/2}$	$\tilde{\boldsymbol{P}} = (\tilde{\rho}_{kl})$	$\tilde{\boldsymbol{R}} = (\tilde{r}_{kl})$

Index i runs from 1 to n (sample size), and indices k and l run from 1 to p (dimension). A tilde denotes a "partial" quantity.

covariance matrix is plugged into Equation F.1. However, note that Equation F.1 also remains valid if other estimates of the covariance are used, such as penalized or shrinkage estimators (note that there is no hat on β_k^y).

For the following it is important that Equation F.1 can be further rewritten by introducing a scale factor. Specifically, by abbreviating the standardized partial variance $\tilde{\sigma}_k^2/\sigma_k^2$ by SPV_k, we can decompose the regression coefficient into the simple product

$$\beta_k^y = \underbrace{\tilde{\rho}_{yk}}_{\mathcal{A}} \underbrace{\sqrt{\frac{\mathrm{SPV}_y}{\mathrm{SPV}_k}}}_{\mathcal{B}} \underbrace{\sqrt{\frac{\sigma_y^2}{\sigma_k^2}}}_{\mathcal{C}} . \tag{F.2}$$

Note that SPV_y and SPV_k take on values from 0 to 1. All three factors have an immediate and intuitive interpretation:

$\mathcal{A}$: This factor determines whether there is a direct association between Y and the covariate X_k. If the partial correlation between X_k and Y vanishes, so will also the two corresponding regression coefficients β_k^y and β_y^k. In a partial correlation graph an edge is drawn between two nodes Y and X_k if $\mathcal{A} \neq 0$.

$\mathcal{B}$: This factor adjusts the regression coefficient for the relative reduction in variance of Y and X_k due to the respective other covariates. In the algorithm outlined below a test of $\log(\mathcal{B})$ establishes the directionality of edges of a partially causal network.

$\mathcal{C}$: This is a scale factor correcting for different units in Y and X_k.

The product $\mathcal{AB} = \beta_k^y \sqrt{\sigma_k^2/\sigma_y^2}$ is also known as the standardized regression coefficient. Note that for computing both $\mathcal{A}$ and $\mathcal{B}$ only the correlation matrix is needed, as the variance information is already accounted for by the third factor $\mathcal{C}$.

In this context it is also helpful to recall the diverse statistical interpretations of SPV:

- SPV is the *proportion* of variance that remains (unexplained) after regressing against all other variables.
- For the OLS estimator SPV is equal to $1 - R^2$, where R is the usual coefficient of determination.
- SPV is the inverse of the diagonal of the inverse of the *correlation* matrix. Thus, if there is no correlation (unit diagonal correlation matrix) the partial variance equals the variance, and hence SPV = 1.
- SPV may also be estimated by 1/VIF, where VIF is the usual variance inflation factor (cf. Stewart, 1987).

F.2.2 Heuristic algorithm for discovering approximate causal networks

The above decomposition (Equation F.2) suggests the following simple strategy for statistical learning of causal networks. First, by multiple testing of $\mathcal{A} = 0$ we determine the network topology, i.e. we identify those edges for which the corresponding partial correlation is not vanishing. Second, by subsequent multiple testing of $\log(\mathcal{B}) = 0$ we establish a partial ordering of the nodes, which in turn imposes a partial directionality upon the edges.

In more detail, we propose the following five-step algorithm:

1. First, it is essential to determine an accurate and positive definite estimate $\boldsymbol{R}$ of the correlation matrix. Only if the sample size is large with many more observations than variables ($n >> p$) the usual empirical correlation estimate will be suitable. In all other instances, the use of a regularized estimator is absolutely vital (e.g., the Stein-type shrinkage estimator of (Schäfer and Strimmer, 2005b)) in order to improve efficiency and to guarantee positive definiteness. In addition, if the samples are longitudinal it may be necessary to adjust for autocorrelation (Opgen-Rhein and Strimmer, 2006a).

2. From the estimated correlations we compute the partial variances and correlations (see Table F.1), and from those in turn plug-in estimates of the factors $\mathcal{A}$ and $\mathcal{B}$ of Equation F.2 for all possible edges. Note that in this calculation each variable assumes in turn the role of the response Y. An efficient way to calculate the various $\mathcal{B}$ is given by taking the square root of the diagonal of the inverse of the estimated correlation matrix, and computing the corresponding pairwise ratios.

3. Subsequently, we infer the partial correlation graph following the algorithm described in (Schäfer and Strimmer, 2005a). Essentially, we perform multiple testing of all partial correlation coefficients $\mathcal{A}$. Note that for high dimensions (large p) the null

distribution of partial correlations across edges can be determined from the data, which in turn allows the adaptive computation of corresponding false discovery rates (Efron, 2004).

4. In a similar fashion we then conduct multiple testing of all $\log(\mathcal{B})$. As $\mathcal{B}$ is the ratio of two variances with the same degrees of freedom, it is implicit that $\log(\mathcal{B})$ is approximately normally distributed (Fisher, 1924), with an unknown variance parameter θ. Thus, the observed $z = \log(\mathcal{B})$ across all edges follow a mixture distribution

$$f(z) = \eta_0 N(0, \theta) + (1 - \eta_0) f_A(z). \tag{F.3}$$

Assuming that most z belong to the null model, i.e. that most edges are undirected, it is possible to infer non-parametrically the alternative distribution $f_A(z)$, the proportion η_0, as well as the variance parameter θ – for an algorithm see (Efron, 2004). From the resulting densities and distribution functions local and tail-area-based false discovery rates for the test $\log(\mathcal{B}) = 0$ are computed. Note that in this procedure we include all edges, regardless of the corresponding value of $\mathcal{A}$ or the outcome of the test $\mathcal{A} = 0$.

5. Finally, a partially directed network is constructed as follows. All edges in the correlation graph with significant $\log(\mathcal{B}) \neq 0$ are directed in such a fashion that the direction of the arrow points from the node with the larger standardized partial variance (the more "exogenous" variable) to the node with the smaller standardized partial variance (the more "endogenous" variable). The other edges with $\log(\mathcal{B}) \approx 0$ remain undirected. The subgraph consisting of all directed edges constitutes the inferred causal network. Note that this does not necessarily include all nodes that are contained in the GGM network.

F.3 Results and discussion

F.3.1 Interpretation of the resulting graph

The above algorithm returns a partially directed partial correlation graph, whose directed edges form a causal network.

This procedure can be motivated by the following connection between partial correlation graph and a system of linear equations, where each node is in turn taken as a response variable and regressed against all other remaining nodes. In this setting the partial correlation coefficient is the geometric mean of β_k^y and the corresponding reciprocal coefficient β_y^k, i.e.

$$\sqrt{\beta_y^k \beta_k^y} = |\tilde{\rho}_{yk}| \tag{F.4}$$

(see also equation 16 of ref. Schäfer and Strimmer, 2005b). In this light, an undirected edge between two nodes A and B in a partial correlation graph may also be interpreted as

bidirected edge, in the sense that A influences B and vice versa in the underlying system of regression. Therefore, the test $\mathcal{B} = 1$ can be understood as *removing* one of these two directions, where Equation F.2 suggests that only the relative variance reduction between the two involved nodes needs to be considered for establishing the final direction.

F.3.2 Reconstruction efficiency and approximations underlying the algorithm

Topology of the network

The proposed algorithm is an extension of the GGM inference approach of (Schäfer and Strimmer, 2005a,b). Its accuracy of correctly recovering the *topology* of the partial correlation graph has been established, e.g., in (Werhli et al., 2006).

However, it is well known that a directed Bayesian network and the corresponding undirected graph are not necessarily topologically identical: in the undirected graph for computing the partial correlations one conditions on all other nodes whereas in the directed graph one conditions only on a subset of nodes, in order to avoid conditioning "on the future" (i.e. on the dependent nodes). Therefore, it is critical to evaluate to what extent full order partial correlations are reasonable approximations for lower order partial correlations. This has already been investigated intensively by (Castelo and Roverato, 2006) who showed that in certain situations (sparse graphs, faithfulness assumption etc.) lower order partial correlations may be used as approximate substitute of full conditional correlations. Therefore, in the proposed algorithm we adopt the very same argument but apply it in the different direction, i.e. we approximate lower order partial correlation by full order partial correlation.

Node ordering

A second approximation implicit in our algorithm concerns the determination of the ordering of the nodes, which is done by multiple testing of pairwise ratios of standardized partial variances. We have conducted a number of numerical simulations (data not shown) that indicate that for randomly simulated DAGs the ordering of the nodes is indeed well reflected in the partial variances, as expected.

However, from variable selection in linear models it is also known that the partial variance (or the related R^2) may not always be a reliable indicator for variable importance. Nevertheless, the partial ordering of nodes according to SPV and the implicit model selection in the underlying regressions is a very different procedure in comparison to the standard variable selection approaches, in which the increase or decrease of the R^2 is taken as indicator of whether or not a variable is to be included, or a decomposition of R^2 is sought (for a review see, e.g., Grömping, 2006). The distinctive feature of our procedure is that by performing all tests $\log(\mathcal{B}) \neq 0$ simultaneously we consider all p regression equations at once, even if the final feature selection occurs only locally on the level of an individual regression.

It is also noteworthy that, as we impose directionality from the less well explained variable (large SPV, "exogenous", "independent") to the one with relatively lower SPV (well explained, "endogenous", "dependent" variable), we effectively choose the direction with the relatively *smaller* regression coefficient (conditional that the corresponding partial correlation is also significant).

F.3.3 Further properties of the heuristic algorithm and of the resulting graphs

The simple heuristic network discovery algorithm exhibits a number of further properties worth noting:

1. The estimated partially directed network cannot contain any (partially) directed cycles. For instance, it is not possible for a graph to contain a pattern such as $A \rightarrow B \rightarrow A$. This example would imply $\mathrm{SPV}_A > \mathrm{SPV}_B > \mathrm{SPV}_A$, which is a contradiction. As a consequence, the subgraph containing the directed edges only is also acyclic (and hence a DAG).

2. The assignment of directionality is transitive. If there is a directed edge from A to B and from B to C then there must also be a directed edge from A to C. Note however, that actual inclusion of a directed edge into the causal network is conditional on a non-zero partial correlation coefficient.

3. As the algorithm relies on correlations as input, causal processes that produce the same correlation matrix lead to the same inferred graph, and hence are indistinguishable. The existence of such equivalence classes is well known for SEMs (Bollen, 1989) and also for Bayesian belief networks (Chickering, 2002).

4. The proposed algorithm is scale-invariant by construction. Hence, a (linear) change in any of units of the data has no effect on the overall estimated partially directed network, and the implied causal relations.

5. We emphasize that the partially directed network is *not* the chain graph representing the equivalence class of the causal network that is obtained by considering only its directed edges – see (Chickering, 2002).

6. The computational complexity of the algorithm is $O(p^3)$. Hence, it is no more expensive than computing the partial correlation graph, and thus allows for estimation of networks containing in the order of thousands and more nodes.

F.3.4 Analysis of a plant expression data set

To illustrate our algorithm for discovering causal structure, we applied the approach to a real world data example. Specifically, we reanalyzed expression time series resulting from

an experiment investigating the impact of the diurnal cycle on the starch metabolism of *Arabidopsis thaliana* (Smith et al., 2004). This is the same data set we used in a sister paper concerning the estimation of a vector autoregressive model (Opgen-Rhein and Strimmer, 2007).

The data are gene expression time series measurements collected at 11 different time points ($0, 1, 2, 4, 8, 12, 13, 14, 16, 20$, and 24 hours after the start of the experiment). The corresponding calibrated signal intensities for 22,814 genes / probe sets and for two biological replicates are available from the NASCArrays repository, experiment no. 60 (Craigon et al., 2004). After log-transforming the data we filtered out all genes containing missing values and whose maximum signal intensity value was lower than 5 on a log-base 2 scale. Subsequently, we applied the periodicity test of (Wichert et al., 2004) to identify the probes associated with the day-night cycle. As a result, a subset of 800 genes remained for further analysis.

In order to estimate the correlation matrix for the 800 genes described by the data set we employed the dynamical correlation shrinkage estimator of (Opgen-Rhein and Strimmer, 2006b) as this takes account of the autocorrelation. The corresponding correlation graph is displayed in Figure F.1. It shows the 150 edges with the largest absolute values of correlation. This graph is very hard to interpret, the branches do not have any immediate or intuitive meaning (a complete annotation of the nodes can be found along with the dataset itself in the R package "GeneNet" (Schäfer et al., 2006)). For instance, there are no hubs as typically observed in biological networks (Ravasz et al., 2002; Barabási and Oltvai, 2004).

This is in great contrast to the partially directed partial correlation graph. For this specific data set, by multiple testing of the factor $\mathcal{A}$ we identified $6,102$ significant edges connecting 669 nodes. For the second factor $\mathcal{B}$, determined whether edges are directed, the distribution of $\log(\mathcal{B})$ is displayed in Figure F.2. The null distribution (dashed line) follows a normal distribution and characterizes the edges that cannot be directed. The alternative distribution (solid line) coincides with the directed edges. In total, we found $15,928$ significant directions.

To construct the network, we projected upon the significant edges (factor $\mathcal{A}$) the significant directions (factor $\mathcal{B}$). In the network of significant associations, $1,216$ directions were significant. Note that the fraction of significant directions is by far greater in the subset of the significant partial correlations than in the complete set of all partial correlations. This agrees with the intuitive notion, that causal influences can only be attributed to existing connections between variables.

The resulting partially causal network is shown in Figure F.3. For reasons of clarity we show only the subnetwork containing the 150 most significant edges, which connect 107 nodes. This graph exhibits a clear "hub" connectivity structure (nodes filled with red color). A prominent example for this is node 570, others are 81, 558, 783 and a few more genes. We see that many of the hub nodes have mostly outgoing arcs, which is indicative for key regulatory genes. This applies, e.g., to node 570, an AP2 transcription factor, or to node 81, a gene involved in DNA-directed RNA polymerase. An interesting aspect of the partially causal network is the web of highly connected genes (colored yellow in the lower

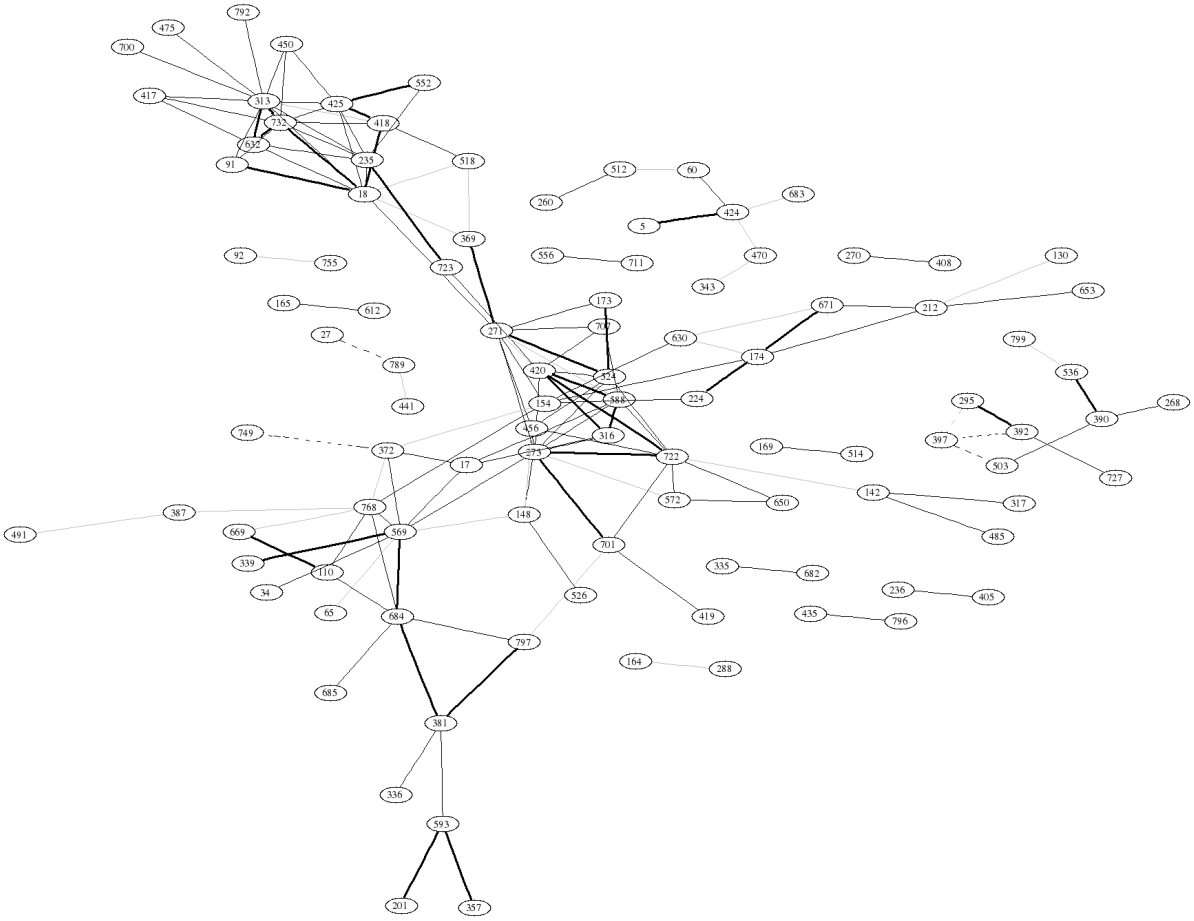

Figure F.1: Correlation network inferred from the *Arabidopsis thaliana* data. The solid and dotted lines indicate positive and negative correlation coefficients, respectively, and the line intensity denotes their strength. The network displays the 150 edges that have the largest correlations. For annotation of the nodes in this graph see the electronic information contained in the R package "GeneNet" (Schäfer et al., 2006) and the original data paper (Smith et al., 2004).

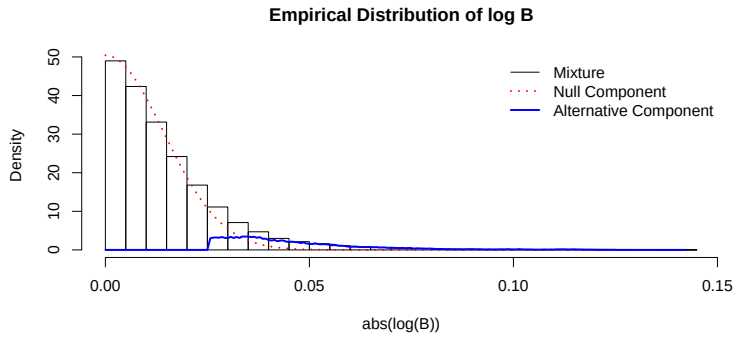

Figure F.2: Distribution of log $\mathcal{B}$ for the *Arabidopsis thaliana* data. The null distribution is depicted by the dashed line; it follows a normal distribution with zero mean and a standard deviation of 0.014. The solid line signifies the alternative distribution. The empirical distribution (indicated by the histogram) is composed of the null distribution ($\eta_0 = 0.8995$) and of the alternative distribution ($\eta_A = 0.1005$.)

right corner of Figure F.3), which we hypothesize to constitute some form of a functional module. In this module, it is not possible to determine any directions, which could be due to complex interactions among the nodes of the module. Node 627 is another hub in the network that connects the functional module with the rest of the network and which according to the annotation of (Smith et al., 2004) encodes a protein of unknown function.

We also see that the partially directed network contains both directed and undirected nodes. This is a distinct advantage of the present approach. Unlike, e.g., a vector autoregressive model (Opgen-Rhein and Strimmer, 2007), it does not *force* directions onto the edges.

Finally, in order to investigate the stability of the inferred partial causal network, we randomly removed data points from the sample, and repeatedly reconstructed the network from the reduced data set. In all cases the general topological structure of the network remained intact, which indicates that this is a signal inherent in the data. This is also confirmed by the analysis using vector autoregressions (Opgen-Rhein and Strimmer, 2007).

F.4 Conclusions

Methods for exploring causal structures in high-dimensional data are growing in importance, particularly in the study of complex biological, medical and financial systems. As a first (and often only) analysis step these data are explored using correlation networks.

Here we have suggested a simple heuristic algorithm that, starting from a (positive definite) correlation matrix, infers a partially directed network that in turn allows generating causal hypotheses of how the data were generated. Our approach is approximate, but it

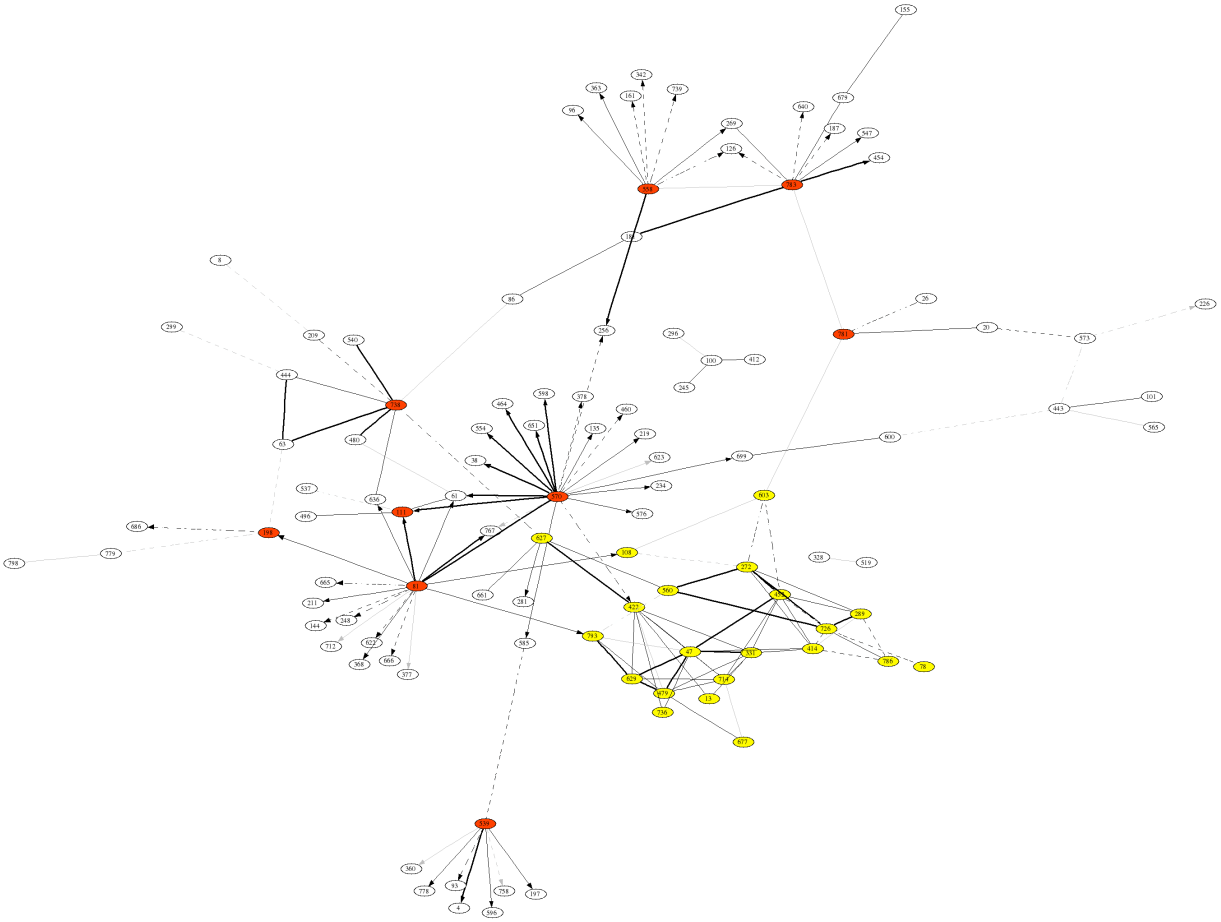

Figure F.3: Partially causal network inferred from the *Arabidopsis thaliana* data by the method introduced in this paper – note the difference to the correlation network of Figure F.1. The topology of the partially causal network is identical to that of a partial correlation graph (GGM, CIG). However, edges with significant directionality (as indicated by a factor $\mathcal{B}$ that is significantly smaller or larger than one) are oriented.

allows analysis of high-dimensional small sampled data, and its computational complexity is very modest. Thus, our heuristic is likely to be applicable whenever a correlation network is computed, and therefore is suitable for screening large-scale data set for causal structure.

Nevertheless, there a several lines along which this method could be extended. For instance, non-linear effects could be accounted for by employing entropy criteria, or by using higher order moments (Shimizu et al., 2006). Furthermore, more sophisticated algorithms may be used to enhance the approximation of lower order partial correlations or the inference of the ordering of the nodes. However, ultimately this would lead to a method similar to the PC algorithm (Spirtes et al., 2001; Kalisch and Bühlmann, 2007).

Note that the PC algorithm is more refined than our algorithm, primarily due to additional steps that aim at removing spurious edges (i.e. those edges that are induced between otherwise uncorrelated parent nodes by conditioning on a common child node). However, these iterative refinements may be very time consuming, in particular for high-dimensional graphs.

In contrast, our procedure is non-iterative and therefore both computationally and algorithmically (nearly) as simple as a correlation network. Nevertheless, it still enables the discovery of partially directed processes underlying the data.

In summary, we recommend our approach as a procedure for exploratory screening for causal mechanisms. Subsequently, the resulting hypotheses may then form the basis for more refined analyzes, such as full Bayesian network modeling.

Authors' contributions

Both authors participated in the development of the methodology and wrote the manuscript. R.O. carried out all analyzes. All authors approved of the final version of the manuscript.

Acknowledgements

This work was in part supported by an "Emmy Noether" excellence grant of the Deutsche Forschungsgemeinschaft (to K.S.).

Bibliography

Barabási, A.-L. and Oltvai, Z. N. (2004). Network biology: Understanding the cell's functional organization. *Genetics*, 5:101–113.

Boginski, V., Butenko, S., and Pardalos, P. M. (2005). Statistical analysis of financial networks. *Comp. Stat. Data Anal.*, 48:431–443.

Bollen, K. A. (1989). *Structural Equations With Latent Variables*. John Wiley & Sons.

Butte, A. J., Tamayo, P., Slonim, D., Golub, T. R., and Kohane, I. S. (2000). Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc. Natl. Acad. Sci. USA*, 97:12182–12186.

Castelo, R. and Roverato, A. (2006). A robust procedure for Gaussian graphical model search from microarray data with p larger than n. *J. Machine Learn. Res.*, 7.

Chickering, D. M. (2002). Learning equivalence classes of Bayesian-network structures. *J. Machine Learn. Res.*, 2:445–498.

Cox, D. R. and Wermuth, N. (1993). Linear dependencies represented by chain graphs. *Statistical Science*, 8(3):204–218.

Craigon, D. J., James, N., Okyere, J., Higgins, J., Jotham, J., and May, S. (2004). NASCArrays: a repository for microarray data generated by NASC's transcriptomics service. *Nucleic Acids Research*, 32:D575–D577.

de la Fuente, A., Bing, N., Hoeschele, I., and Mendes, P. (2004). Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics*, 20:3565–3574.

Dobra, A., Hans, C., Jones, B., Nevins, J. R., Yao, G., and West, M. (2004). Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*, 90:196–212.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *Journal of the American Statistical Association*, 99(465):96–104(9).

Fisher, R. A. (1924). On a distribution yielding the error functions of several well known statistics. *Proc. Intl. Congr. Math.*, 2:805–813.

Freedman, D. A. (2005). *Statistical Models: Theory and Practice*. Cambridge University Press, Cambridge, UK.

Grömping, U. (2006). Relative importance in linear regression in R: the package relaimpo. *J. Statist. Soft.*, 17:1.

Kalisch, M. and Bühlmann, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *J. Machine Learn. Res.*, 8:613–636.

Li, H. and Gui, J. (2006). Gradient directed regularization for sparse Gaussian concentration graphs, with applications to inference of genetic networks. *Biostatistics*, 7:302–317.

Mantegna, R. N. and Stanley, H. E. (2000). *In Introduction to Econophysics: Correlations and Complexity in Finance*. Cambridge University Press.

Oldham, M., Horvath, S., and Geschwind, D. (2006). Conservation and evolution of gene coexpression networks in human and chimpanzee brains. *Proc. Natl. Acad. Sci. USA*, 103(47):17973–17978.

Onnela, J. P., Kaski, K., and Kertész, J. (2004). Clustering and information in correlation based financial networks. *Eur. Phys. J. B*, 38:353–362.

Opgen-Rhein, R. and Strimmer, K. (2006a). Inferring gene dependency networks from genomic longitudinal data: a functional data approach. *REVSTAT*, 4:53–65.

Opgen-Rhein, R. and Strimmer, K. (2006b). Using regularized dynamic correlation to infer gene dependency networks from time-series microarray data. In *Proceedings of the 4th International Workshop on Computational Systems Biology (WCSB 2006)*, volume 4, pages 73–76, Tampere.

Opgen-Rhein, R. and Strimmer, K. (2007). Learning causal networks from systems biology time course data: an effective model selection procedure for the vector autoregressive process. *BMC Bioinformatics*, 8 (Suppl. 2):S3.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, UK.

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabási, A.-L. (2002). Hierarchical organization of modularity in metabolic networks. *Science*, 297:1551–1555.

Schachter, R. D. and Kenley, C. R. (1989). Gaussian influence diagrams. *Management Sci.*, 35:527–550.

Schäfer, J., Opgen-Rhein, R., and Strimmer, K. (2006). Reverse engineering genetic networks using the **GeneNet** package. *R News*, 6(5):50–53.

Schäfer, J. and Strimmer, K. (2005a). An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, 21(6):754–764.

Schäfer, J. and Strimmer, K. (2005b). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4(1):32.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., and Kerminen, A. (2006). A linear non-Gaussian acyclic model for causal discovery. *J. Machine Learn. Res.*, 7:2003–2030.

Shipley, B. (2000). *Cause and Correlation in Biology*. Cambridge University Press.

Smith, S. M., Fulton, D. C., Chia, T., Thorneycroft, D., Chapple, A., Dunstan, H., Hylton, C., and Smith, S. C. Z. A. M. (2004). Diurnal changes in the transcriptom encoding enzymes of starch metabolism provide evidence for both transcriptional and post-transcriptional regulation of starch metabolism in Arabidopsis leaves. *Plant Physiol.*, 136:2687–2699.

Spirtes, P., Glymour, C., and Scheines, R. (2001). *Causation, Prediction, and Search*. MIT Press, 2nd edition. edition.

Steuer, R. (2006). On the analysis and interpretation of correlations in metabolomic data. *Brief. Bioinform.*, 151:151–158.

Stewart, G. W. (1987). Collinearity and least squares regression. *Statistical Science*, 2(2):68–84.

Studený, M. (2005). *Probabilistic Conditional Independence Structures*. Springer.

Tsamardinos, I., Brown, L. E., and Aliferis, C. F. (2006). The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 65:31–78.

Tumminello, M., Aste, T., Di Matteo, T., and Mantegna, R. N. (2005). A tool for filtering information in complex systems. *Proc. Natl. Acad. Sci. USA*, 102:10421–10426.

Werhli, A. V., Grzegorczyk, M., and Husmeier, D. (2006). Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical Gaussian models and Bayesian networks. *Bioinformatics*, 22(20):2523–2531.

Wermuth, N. (1980). Linear recursive equations, covariance selection, and path analysis. *J. Amer. Statist. Assoc.*, 75:963–972.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley, New York.

Wichert, S., Fokianos, K., and Strimmer, K. (2004). Identifying periodically expressed transcripts in microarray time series data. *Bioinformatics*, 20(1):5–20.

Wille, A. and Bühlmann, P. (2006). Low-order conditional independence graphs for inferring genetic networks. *Statist. Appl. Genet. Mol. Biol.*, 5:1.

Article G

Reverse Engineering Genetic Networks using the GeneNet package

Published in R News *(Volume 6, Number 5, December 2006, pp. 50-53)*

Authors: Juliane Schäfer, Rainer Opgen-Rhein *and* Korbinian Strimmer

GeneNet is a package for analyzing high-dimensional (time series) data obtained from high-throughput functional genomics assays, such as expression microarrays or metabolic profiling. Specifically, **GeneNet** allows to infer large-scale gene association networks. These are graphical Gaussian models (GGMs) that represent multivariate dependencies in biomolecular networks by means of partial correlation. Therefore, the output of an analysis conducted by **GeneNet** is a graph where each gene corresponds to a node and the edges included in the graph portray direct dependencies between them.

GeneNet implements a specific learning algorithm that allows to estimate GGMs from small sample high-dimensional data that is both computationally as well as statistically efficient. This approach relies on analytic shrinkage estimation of covariance and (partial) correlation matrices and on model selection using (local) false discovery rate multiple testing. Hence, **GeneNet** includes a computational algorithm that decides which edges are to be included in the final network, in dependence of the *relative* values of the pairwise partial correlations.

In a recent comparative survey (Werhli et al., 2006) the **GeneNet** procedure was found to recover the topology of gene regulatory networks with similar accuracy as computationally much more demanding methods such as dynamical Bayesian networks (Friedman, 2004).

We note that the approach implemented in **GeneNet** should be regarded as an exploratory approach that may help to identify interesting genes (such as "hubs") or clusters of genes that are functionally related or co-regulated, rather than that it yields the precise network of mechanistic interactions. Therefore, the resulting network topologies need be interpreted and validated in the light of biological background information, ideally accompanied by further integrative analysis employing data from different levels of the cellular system.

G.1 Prerequisites

GeneNet is available from the CRAN repository and from the webpage `http://strimmerlab.org/software/genenet/`. It requires prior installation of four further R packages also found on CRAN: **corpcor**, **longitudinal**, **fdrtool**, and **locfdr** (Efron, 2004).

For installation of the required packages simply enter at the R prompt:

```
> install.packages( c("corpcor",
     "longitudinal", "fdrtool",
     "locfdr", "GeneNet") )
```

G.2 Preparation of Input Data

The input data must be arranged in a matrix where columns correspond to genes and where rows correspond to the individual measurements. Note that the data must already be properly preprocessed, i.e. in the case of expression data calibrated and normalized.

In the following we describe an example for inferring the gene association network among 102 genes from a microarray data set on the microorganism *Escherichia coli* with observations at 9 time points (Schmidt-Heck et al., 2004). These example data are part of **GeneNet**:

```
> library("GeneNet")
> data(ecoli)
> dim(ecoli)
[1]   9 102
```

G.3 Shrinkage Estimators of Covariance and (Partial) Correlation

The first step in the inference of a graphical Gaussian model is the reliable estimation of the partial correlation matrix:

```
> inferred.pcor <- ggm.estimate.pcor(ecoli)
> dim(inferred.pcor)
[1] 102 102
```

For this purpose, the function `ggm.estimate.pcor` offers an interface to a shrinkage estimator of partial correlation implemented in the **corpcor** package that is statistically efficient and can be used for analyzing small sample data. By default, the option `method="static"` is selected, which employs the function `pcor.shrink`. Standard graphical modeling theory (e.g. Whittaker, 1990) shows that the matrix of partial correlations $\tilde{\mathbf{P}} = (\tilde{\rho}_{ij})$ is related

to the inverse of the covariance matrix $\mathbf{\Sigma}$. This relationship leads to the straightforward estimator

$$\tilde{r}_{ij} = -\hat{\omega}_{ij}/\sqrt{\hat{\omega}_{ii}\hat{\omega}_{jj}}, \tag{G.1}$$

where

$$\hat{\mathbf{\Omega}} = (\hat{\omega}_{ij}) = \hat{\mathbf{\Sigma}}^{-1}. \tag{G.2}$$

In Equation G.2, it is absolutely crucial that the covariance is estimated accurately, and that $\hat{\mathbf{\Sigma}}$ is well conditioned – otherwise the above formulae will result in a rather poor estimate of partial correlation (cf. Schäfer and Strimmer, 2005a). For this purpose, the `pcor.shrink` function uses an analytic shrinkage estimator of the correlation matrix developed in Schäfer and Strimmer (2005b). This linearly combines the unrestricted sample correlation with a suitable correlation target in a weighted average. Selecting this target requires some diligence: specifically, we choose to shrink the empirical correlations $\mathbf{R} = (r_{ij})$ towards the identity matrix, while empirical variances are left intact. In this case the analytically determined shrinkage intensity is

$$\lambda^{\star} = \frac{\sum_{i \neq j} \mathrm{VAR}(r_{ij})}{\sum_{i \neq j} r_{ij}^2}. \tag{G.3}$$

The resulting shrinkage estimate exhibits a number of favorable properties. For instance, it is much more efficient, always positive definite, and well conditioned. It is inexpensive to compute and does not require any tuning parameters, as the analytically derived optimal shrinkage intensity is estimated from the data. Moreover, there are no assumptions about the underlying distributions of the individual estimates, except for the existence of the first two moments. These properties carry over to derived quantities, such as partial correlations. Furthermore, the resulting estimates are in a form that allows for fast computation of their inverse using the Woodbury matrix identity.

Note that the function `ggm.estimate.pcor` also allows the specification of a `protect` argument, with default value `protect=0`. This imposes limited translation (Efron and Morris, 1972) onto the specified fraction of entries of the estimated shrinkage correlation matrix, thereby protecting those components against overshrinkage (see also (Opgen-Rhein and Strimmer, 2006a)).

G.4 Taking Time Series Aspects Into Account

Standard Gaussian graphical models assume i.i.d. data whereas in practice many expression data sets result from time course experiments. One possibility to generalize the above procedure correspondingly is to employ dynamic (partial) correlation (Opgen-Rhein and Strimmer, 2006b). This is available in the function `ggm.estimate.pcor` by specifying the option `method="dynamic"`, which in turn relies on the **longitudinal** package for computation.

The key difference between dynamical and i.i.d. correlation is that the former takes into account the time that has elapsed between two subsequent measurements. In particular,

dynamical correlation allows for unequally spaced time points as often encountered in genomic studies. All small sample learning procedures (shrinkage) developed for i.i.d. correlation also carry over to dynamical correlation (Opgen-Rhein and Strimmer, 2006c).

G.5 Network Search and Model Selection

The second crucial part of gene association network inference is model selection, i.e. assigning statistical significance to the edges in the GGM network:

```
> test.results <-
             ggm.test.edges(inferred.pcor)
> dim(test.results)
[1] 5151    6
```

For this purpose a mixture model,

$$f(\tilde{r}) = \eta_0 f_0(\tilde{r}; \kappa) + (1 - \eta_0) f_{\mathrm{A}}(\tilde{r}) \,, \tag{G.4}$$

is fitted to the observed partial correlation coefficients $\tilde{r}$ using the subroutine `cor.fit.mixture`. f_0 is the distribution under the null hypothesis of vanishing partial correlation, η_0 is the (unknown) proportion of "null edges", and f_{A} the distribution of observed partial correlations assigned to actually existing edges. The latter is assumed to be an arbitrary nonparametric distribution that vanishes for values near zero. This allows for κ, η_0, and even f_{A} to be determined from the data – see Efron (2004) for an algorithm.

Subsequently, two-sided p-values corresponding to the null hypothesis of zero partial correlation are computed for each potential edge using the function `cor0.test`. Large-scale simultaneous testing is then conducted by obtaining q-values via the function `fdr.control` with the specified value of η_0 taken into account. `fdr.control` uses the algorithms described in Benjamini and Hochberg (1995) and Storey (2002). An alternative to the q-value approach is to use the empirical Bayes local false discovery rate (fdr) statistic (Efron, 2004). This fits naturally with the above mixture model setup, and in addition takes account of the dependencies among the estimated partial correlation coefficients. The posterior probability that a specific edge exists given $\tilde{r}$ equals

$$\mathbb{P}(\text{non-null edge}|\tilde{r}) = 1 - \mathrm{fdr}(\tilde{r}) = 1 - \frac{\eta_0 f_0(\tilde{r}; \kappa)}{f(\tilde{r})}. \tag{G.5}$$

Following (Efron, 2005), we typically consider an edge to be "significant" if its local fdr is smaller than 0.2, or equivalently, if the probability of an edge to be "present" is larger that 0.8:

```
> signif <- test.results$prob > 0.80
> sum(signif)
[1] 66
> test.results[signif,]
```

G.6 Network Visualization

The network plotting functions in **GeneNet** rely extensively on the infrastructure offered by the **graph** and **Rgraphviz** packages (cf. contribution of Seth Falcon in this R News issue).

First, a `graph` object must be generated containing all significant edges:

```
> node.labels <- colnames(ecoli)
> gr <- ggm.make.graph(
     test.results[signif,],
     node.labels)
> gr
A graphNEL graph with undirected edges
Number of Nodes = 102
Number of Edges = 66
```

Subsequently, the resulting object can be inspected by running the command

```
> show.edge.weights(gr)
```

Finally, the gene network topology of the graphical Gaussian model can be visualized using the function `ggm.plot.graph`:

```
> ggm.plot.graph(gr,
      show.edge.labels=FALSE,
      layoutType="fdp")
```

The plot resulting from the analysis of the `ecoli` data is shown in Figure G.1. For `show.edge.labels=TRUE` the partial correlation coefficients will be printed as edge labels. Note that on some platforms (e.g. Windows) the default `layoutType="fdp"` may not yet be available. In this case an alternative variant such as `"layoutType=neato"` needs to be specified.

G.7 Release History of GeneNet and Example Scripts

The package **GeneNet** emerged from a reorganization of the (now obsolete) package **Gene TS**. This was split into the **GeneNet** part dealing with gene network reconstruction, and the package **GeneCycle** for cell cycle and periodicity analysis (Wichert et al., 2004; Ahdesmäki et al., 2005).

On the home page of **GeneNet** we collect example scripts in order to guide users of **GeneNet** when conducting their own analyses. Currently, this includes the above *E. coli* data but for instance also a network analysis of *A. thaliana* diurnal cycle genes. We welcome further contributions from the biological community.

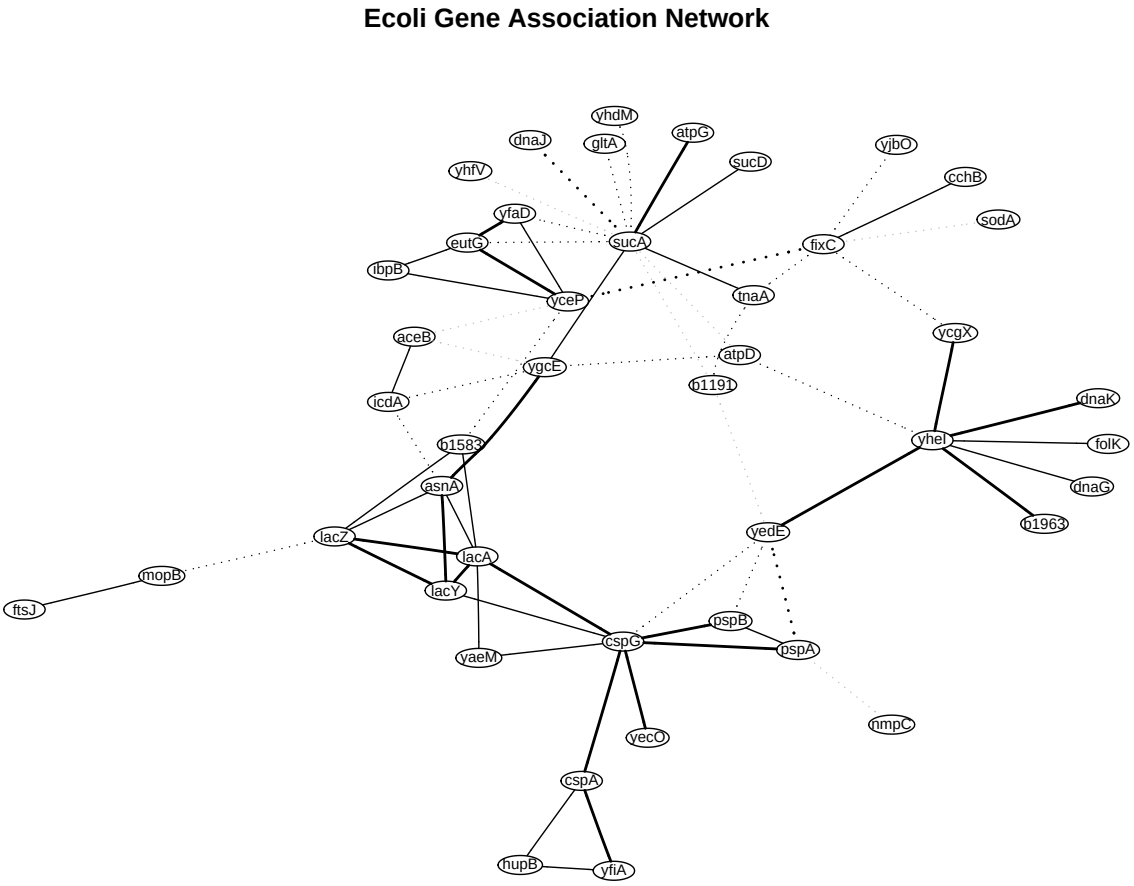

Figure G.1: Sparse graphical Gaussian model for 102 genes inferred from an *E. coli* microarray data set with 9 data points. Full and dotted lines indicate positive and negative partial correlation, respectively.

Bibliography

Ahdesmäki, M., Lähdesmki, H., Pearson, R., Huttunen, H., and Yli-Harja, O. (2005). Robust detection of periodic time series measured from biological systems. *BMC Bioinformatics*, 6:117.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B*, 57:289–300.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *Journal of the American Statistical Association*, 99(465):96–104(9).

Efron, B. (2005). Local false discovery rates. Technical Report 2005-20B/234, Dept. of Statistics, Stanford University.

Efron, B. and Morris, C. (1972). Limiting the risk of Bayes and empirical Bayes estimators – part II: The empirical Bayes case. *Journal of the American Statistical Association*, 67:130–139.

Friedman, N. (2004). Inferring cellular networks using probabilistic graphical models. *Science*, 303(5659):799–805.

Opgen-Rhein, R. and Strimmer, K. (2006a). Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach. *Statist. Appl. Genet. Mol. Biol.*, 6(1):9.

Opgen-Rhein, R. and Strimmer, K. (2006b). Inferring gene dependency networks from genomic longitudinal data: a functional data approach. *REVSTAT*, 4:53–65.

Opgen-Rhein, R. and Strimmer, K. (2006c). Using regularized dynamic correlation to infer gene dependency networks from time-series microarray data. In *Proceedings of the 4th International Workshop on Computational Systems Biology (WCSB 2006)*, volume 4, pages 73–76, Tampere.

Schäfer, J. and Strimmer, K. (2005a). An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, 21(6):754–764.

Schäfer, J. and Strimmer, K. (2005b). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4(1):32.

Schmidt-Heck, W., Guthke, R., Toepfer, S., Reischer, H., Dürrschmid, K., and Bayer, K. (2004). Reverse engineering of the stress response during expression of a recombinant protein. In *Proceedings of the EUNITE symposium*, pages 407–412, Aachen, Germany. Verlag Mainz.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal Of The Royal Statistical Society Series B*, 64(3):479–498.

Werhli, A. V., Grzegorczyk, M., and Husmeier, D. (2006). Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical Gaussian models and Bayesian networks. *Bioinformatics*, 22(20):2523–2531.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley, New York.

Wichert, S., Fokianos, K., and Strimmer, K. (2004). Identifying periodically expressed transcripts in microarray time series data. *Bioinformatics*, 20(1):5–20.

Danksagung

Diese Arbeit entstand während meiner Tätigkeit als Mitarbeiter der Arbeitsgruppe „Statistics and Computational Biology“ am Institut für Statistik der Ludwig-Maximilians-Universität München und ich möchte mich an dieser Stelle bei den Personen bedanken, die auf unterschiedliche Art und Weise einen wesentlichen Beitrag zum Entstehen dieser Arbeit geleistet haben.

Mein besonderer Dank gilt vor allem Herrn Prof. Dr. Korbinian Strimmer für die Gelegenheit, in seiner Arbeitsgruppe mitzuwirken. Durch seine exzellente fachliche Betreuung ermöglichte er es mir, mich in die interessanten Themen meiner Dissertation einzuarbeiten und durch regen Dialog und konstantes Hinterfragen neue Ideen zu entwickeln. Er verhalf mir nicht nur zu einem vertieften Einblick in das spannende Feld der angewandten Statistik, sondern auch zu vielen neuen Kenntnissen in Computersystemen und Anwendungsprogrammen. Große Freude bereiteten mir auch die die zahlreichen weiterführenden Diskussionen und manches Gespräch auch über den Tellerrand hinaus. Der mir ermöglichte Besuch nationaler und internationaler Workshops und Konferenzen hat mir zu vielen neuen Erfahrungen verholfen, von denen ich auch in Zukunft profitieren werde.

Mein Dank gilt Frau Dr. Juliane Schäfer, mit der ich bis zur Fertigstellung ihrer Promotion das Zimmer teilte und die mir bei meiner Arbeit und meinen Fragen immer eine große Hilfe war. Genauso bedanke ich mich bei allen anderen Mitgliedern unserer Arbeitsgruppe, mit denen ich zusammenarbeiten durfte.

Herrn Prof. Dr. Ludwig Fahrmeir und Herrn PD Dr. Christian Heumann danke ich für die freundliche Übernahme der Begutachtung der Dissertation.

Nicht zuletzt möchte ich meinen Eltern danken, die immer volles Vertrauen in meine universitäre Laufbahn setzten und diese überhaupt ermöglichten.

Die der Arbeit zugrunde liegende Projektarbeit wurde finanziell gefördert durch die Deutsche Forschungsgemeinschaft (DFG) im Rahmen des Emmy-Noether-Programms.

Allen Kolleginnen und Kollegen im Institut für Statistik der Ludwig-Maximilians-Universität München, die hier nicht namentlich erwähnt sind, danke ich für ihre Diskussionsbereitschaft und Hilfsbereitschaft, für die gute Arbeitsatmosphäre und für eine insgesamt unvergessliche Zeit.

München, August 2007

Rainer Opgen-Rhein

Lebenslauf

Persönliche Daten

Name:	Rainer Opgen-Rhein
Geburtsdatum:	29.08.1978
Geburtsort:	München

Schulausbildung

1985 – 1989	Grundschule Baldham
1989 – 1998	Gymnasium Vaterstetten

Hochschulausbildung

November 1999 – April 2004	Studium der Betriebswirtschaftslehre an der Ludwig-Maximilians-Universität München
April 2002 – Juni 2005	Studium der Philosophie an der Ludwig-Maximilians-Universität München
April 2004 – Juli 2007	Promotionsstudium zum Dr. oec. publ. am Institut für Statistik der Ludwig-Maximilians-Universität München

Berufstätigkeit

April 2001 – Juli 2002	studentische Hilfskraft am Institut für Statistik der Ludwig-Maximilians-Universität München
März 2004 – Juni 2007	Wissenschaftlicher Mitarbeiter am Institut für Statistik der Ludwig-Maximilians-Universität München