Small-Sample Analysis and Inference of Networked Dependency Structures from Complex Genomic Data

Juliane Stephanie Schäfer

Dissertation
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München
16. November 2005

Small-Sample Analysis and Inference of Networked Dependency Structures from Complex Genomic Data

Dissertation
zur Erlangung des Grades eines Doktors der Naturwissenschaften
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

vorgelegt von
Juliane Stephanie Schäfer
16. November 2005

1. Berichterstatter:	Prof. Dr. Ludwig Fahrmeir
2. Berichterstatter:	Prof. Dr. Ulrich Mansmann
Ausw. Berichterstatter:	Prof. Dr. Göran Kauermann
Betreuer:	Dr. Korbinian Strimmer
Tag des Rigorosums:	16. März 2006

Vorwort

Die statistische Modellierung und Inferenz hochdimensionaler Interaktionsstrukturen hat sich als eine conditio sine qua non in der Weiterentwicklung der Systembiologie und der funktionellen Genomik erwiesen. Die vorliegende Arbeit soll einen Beitrag dazu leisten. Mein Thema verstand ich stets als willkommene Herausforderung und immer nur als einen Markstein auf einem langen Weg.

Mein großer Dank gilt Herrn Dr. Korbinian Strimmer für die fordernde und fördernde Aufnahme in seine Arbeitsgruppe und für den Vorschlag meines Themas. Stets war es spannend, im hinterfragenden und in Frage stellenden Dialog mit ihm neue Ideen zu entwickeln. Für die zahlreichen weiterführenden Diskussionen, für vertiefende Gedanken und exzellente fachliche Betreuung und für manches Gespräch über den Tellerrand hinaus bin ich ihm zu besonderem Dank verpflichtet. Der mir ermöglichte Besuch nationaler und internationaler Workshops und Konferenzen hat meinen Horizont erweitert. Die Erfahrungen hieraus sind ein wertvolles Kapital für die Zukunft.

Herrn Prof. Dr. Ludwig Fahrmeir, Herrn Prof. Dr. Ulrich Mansmann und Herrn Prof. Dr. Göran Kauermann danke ich für die freundliche Übernahme der Begutachtung der Dissertation.

Für die konstruktive Kooperation zur praktischen Genexpressionsanalyse danke ich Herrn Dr. Reimar Abraham und Herrn Andreas Roidl. Herr Wolfgang Schmidt-Heck, Herr PD Dr. Reinhard Guthke und Herr Prof. Dr. Karl Bayer stellten mir freundlicherweise die *E. coli* Daten zur Verfügung.

Die der Arbeit zugrunde liegende Projektarbeit wurde finanziell gefördert durch die Deutsche Forschungsgemeinschaft (DFG) im Rahmen des Emmy Noether-Programms.

Moreover, I am grateful to Prof. Vincent Moulton for discussion and his hospitality during my 6 weeks visit at The Linnaeus Centre for Bioinformatics (Uppsala University,

Sweden), supported within the project Access to Research Infrastructures (ARI).

Herrn Rainer Opgen-Rhein danke ich für seine zuverlässige und geduldige Hilfe bei allen kleineren und größeren Computerproblemen und Frau Dr. Renata Gebhardt und Frau Susanne Heim für viele wertvolle Hinweise und aufmunternde Worte.

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 im November 2005
Juliane Schäfer

Diese Arbeit basiert in Teilen auf den folgenden begutachteten Veröffentlichungen in Zeitschriften und in Konferenzbänden:

Juliane Schäfer und Korbinian Strimmer. 2005. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology* **4**: Article 32. `http://www.bepress.com/sagmb/vol4/iss1/art32`

Juliane Schäfer und Korbinian Strimmer. 2005. An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics* **21**:754–764.

Juliane Schäfer und Korbinian Strimmer. 2005. Learning large-scale graphical Gaussian models from genomic data. In J. F. F. Mendes, S. N. Dorogovtsev, A. Povolotsky, F. V. Abreu, und J. G. Oliveira. (Eds.). *Science of Complex Networks: From Biology to the Internet and WWW*, Volume 776, AIP Conference Proceedings, Aveiro, PT, August 2004, pp. 263–276. American Institute of Physics.

Weitere begutachtete Veröffentlichungen in Zeitschriften:

Reimar Abraham, Juliane Schäfer, Mike Rothe, Johannes Bange, Pjotr Knyazev und Axel Ullrich. 2005. Identification of MMP-15 as an anti-apoptotic factor in cancer cells. *Journal of Biological Chemistry* **280**:34123–34132.

Zusammenfassung

Die vorliegende Arbeit beschäftigt sich mit der statistischen Modellierung und Inferenz genetischer Netzwerke. Assoziationsstrukturen und wechselseitige Einflüsse sind ein wichtiges Thema in der Systembiologie. Genexpressionsdaten weisen eine hohe Dimensionalität auf, die geringen Stichprobenumfängen gegenübersteht ("small n, large p"). Die Analyse von Interaktionsstrukturen mit Hilfe graphischer Modelle ist demnach ein schlecht gestelltes (inverses) Problem, dessen Lösung Methoden zur Regularisierung erfordert. Ich schlage neuartige Schätzfunktionen für Kovarianzstrukturen und (partielle) Korrelationen vor. Diese basieren entweder auf Resampling-Verfahren oder auf Shrinkage zur Varianzreduktion. In der letzteren Methode wird die optimale Shrinkage Intensität analytisch berechnet. Im Vergleich zur klassischen Stichprobenkovarianzmatrix besitzt speziell diese Schätzfunktion wünschenswerte Eigenschaften im Sinne von gesteigerter Effizienz und von kleinerem mittleren quadratischen Fehler. Außerdem ergeben sich stets positiv definite und gut konditionierte Parameterschätzungen. Zur Bestimmung der Netzwerktopologie wird auf das Konzept graphischer Gaußscher Modelle zurückgegriffen, mit deren Hilfe sich sowohl marginale als auch bedingte Unabhängigkeiten darstellen lassen. Es wird eine Methode zur Modellselektion vorgestellt, die auf einer multiplen Testprozedur mit Kontrolle der False Discovery Rate beruht. Dabei wird die zugrunde liegende Nullverteilung adaptiv geschätzt. Das vorgeschlagene Framework ist rechentechnisch effizient und schneidet im Vergleich mit konkurrierenden Verfahren sowohl in Simulationen als auch in der Anwendung auf molekulare Daten sehr gut ab.

Summary

The present work is concerned with modeling and inferring genetic networks. Association and dependency structures are ubiquitous in systems biology. Current gene expression data sets include a large number of variables, but only few samples ("small n, large p"). Thus, the application of graphical models is an ill-posed (inverse) problem that requires explicit regularization. In this thesis, I propose several novel estimators of covariance and (partial) correlation. These are based on variance reduction either by bootstrap aggregation or by shrinkage. The novel shrinkage estimator exploits an analytic formula for determining the optimal shrinkage intensity and reveals particularly distinct advantages over standard estimators: in addition to increased efficiency and accuracy, it is always positive definite and well-conditioned. For inferring network topology, specific focus is on graphical Gaussian models (GGMs) based on the concept of conditional independence. A model selection procedure is introduced that employs false discovery rate multiple testing with adaptive estimation of the null distribution. The proposed small-sample framework is computationally efficient and performs very favorably compared to competing approaches both in simulations as well as in application to real expression data.

Contents

1. Introduction 1

2. Network Biology 7

2.1. Biological Background 7

2.2. Introduction to Microarray Technology 9

2.3. Calibration and Data Transformation 11

2.4. Cellular Networks 13

3. Graphical Models for Describing Gene Dependency Networks 15

3.1. Conditional Independence Graphs 15

3.1.1. Graph Theory 16

3.1.2. Markov Properties 18

3.2. Covariance Graphs 19

3.3. Graphical Gaussian Models 20

3.4. Addressing the “Small n, Large p” Problem 23

3.4.1. Dimension Reduction Prior to Classic GGM Analysis 24

3.4.2. Limited Order Partial Correlations 25

3.4.3. Regularized GGMs 25

4. Regularized Estimation 27

4.1. Ledoit-Wolf-Type Shrinkage 31

4.1.1. Outline of Shrinkage Estimation and the Lemma of Ledoit-Wolf 31

4.1.2. Estimation of the Optimal Shrinkage Intensity 35

4.1.3. Shrinkage Estimation of the Covariance Matrix 35

4.2. Estimating Partial Correlation from Small Samples 41

5. Small-Sample Inference of Genetic Networks 43
5.1. Robbins-Efron-Type Inference of Empirical Null Distribution 45
5.2. Large-Scale GGM Selection Using Multiple Testing 47
5.2.1. The False Discovery Rate 49
5.2.2. The `locfdr` Algorithm for Estimating the Local fdr 51
5.3. Simulation Study 54
5.3.1. Simulation Setup 54
5.3.2. Performance for Synthetic Data 56

6. Lasso Regression for Large-Scale Covariance Selection 65
6.1. Model Selection Using L_1 Lasso Penalized Regression 65
6.2. Performance for Synthetic Data 66

7. Network Analysis of Molecular Data 71
7.1. Gene Interaction Structures in Breast Cancer 71
7.2. Stress Response of *Escherichia coli* 75

8. Summary and Outlook 81

A. Available Computer Software 89

List of Figures

1.1. Life's complexity pyramid composed of the molecular components of the cell (figure source: from Oltvai and Barabási, 2002). Net-like structures emerge on various levels of cellular organization. 3

4.1. Ordered eigenvalues of the sample covariance matrix $\boldsymbol{S}$ and of an alternative estimator $\boldsymbol{S}^{\star}$ proposed in this work, calculated from simulated data with various sample sizes. 28

5.1. Simulated sparse network. 55

5.2. Mean squared error of various small-sample estimators of partial correlation in dependence of the sample size. 57

5.3. Quantile-quantile plots of the observed null distribution for synthetic small-sized data. 58

5.4. Quantile-quantile plots of the observed mixture distribution for synthetic small-sized data. 60

5.5. Empirical Bayes posterior probabilities for synthetic small-sized data. 61

5.6. Performance of GGM network inference procedure: (a) Average number of edges detected as significant. (b) Power and positive predictive value (PPV) for reconstructing the GGM network topology. 63

6.1. Performance of GGM network inference procedure: (a) Average number of edges detected as significant. (b) Power and positive predictive value (PPV) for reconstructing the GGM network topology. 69

7.1. Sub-network consisting of 96 genes centered around the ESR2 gene reconstructed from the breast cancer data of West et al. (2001). 72

7.2. (a) Histogram of estimated shrinkage correlation coefficients. (b) Distribution of estimated shrinkage *partial* correlation coefficients. 75
7.3. Gene network inferred from the *E. coli* data by the shrinkage (Tab. 4.1) GGM approach. 77
7.4. Gene network inferred from the *E. coli* data by the lasso GGM approach by Meinshausen and Bühlmann (2005a). 78
7.5. Gene network inferred from the *E. coli* data by the relevance network approach with abs(r) > 0.8. 79

List of Tables

4.1. Small-sample shrinkage estimators of the unrestricted covariance and correlation matrix. 29
4.2. Six different shrinkage targets for the covariance matrix and associated estimators of the optimal shrinkage intensity. 37

5.1. Definition of quantities used for assessing GGM network reconstruction. . . . 62

1. Introduction

Over the past years research in the life sciences to understand complex biomolecular mechanisms in the cell has shifted from a hypothesis-driven to a discovery-driven science. This trend is inspired by achievements in genome sequencing and technical breakthroughs in spotting high-density hybridization probes that have led to an explosion in the availability of expression data on a genome-wide basis. While the publication of the draft sequence of the human genome (Sachidanandam et al., 2001; Venter et al., 2001) marks the culmination of several decades of work, *functional genomics* represents a more advanced state of genome analysis that is still in its infancy. It refers to the development and application of global experimental approaches to assess gene function and gain a deeper understanding of underlying cellular processes. A functional genomics approach is characterized by high-throughput or large-scale experimental methodologies combined with statistical and computational analysis of the results. The main focus is on expansion of the scope of biological investigation from studying *single* genes or proteins to studying *all* genes or proteins simultaneously in a systematic fashion. However, from a statistical point of view the transformation of observed genomic data into biological insights is a challenging task. The particular dimensionalities in current genomic data sets, namely a large number of investigated features as opposed to typically very few samples, pose novel demands on modeling and inference in bioinformatics and computational biology. Methods and tools, i.e. software implementing the methods, are required to analyze responses of thousands of genes in order to identify interesting genes or clusters of genes that may help biologists –subject to further extensive investigation– to solve real-world problems, e.g., to identify potential drug targets.

Genome data, such as protein and nucleotide sequences, have a relatively simple structure compared to the more and more complex data types that are currently emerging from novel high-throughput technologies (e.g., hybridization-based assays such as microarrays, protein assays, mass spectrometry). This development goes along with data analysis methods that are rapidly moving away from simple exploratory tools to model-based approaches. The most striking example are microarray data for which large-scale probabilistic models are becoming increasingly popular.

As the complex functions of a living cell are carried out through the concerted activity of genes and gene products, the *network* emerges as a paradigm in molecular biology. It is generally assumed that gene expression profiles tend to portray cellular functional structures. The application of cluster analysis methods to expression data was initiated in 1998 (Eisen et al.). Furthermore, gene expression is controlled as a response to internal as well as external stimuli, e.g. stress conditions. Cells' abilities to fine-tuning their intracellular programs in response to environmental conditions and to correcting internal errors, such as mutations and misfolded proteins, are amazing. Thus, it is evident that the detailed inventory of genes, transcripts, proteins, and metabolites is not sufficient to understand the cell's complexity. For understanding biology at the system level, structure and dynamics of cellular and organismal function need to be discussed as pointed out by Kitano (2002b,a). The focus of the present work is on modeling and inferring network-like interaction structures from complex high-dimensional genomic data.

Oltvai and Barabási (2002) design the so-called complexity pyramid of life (cf. Fig. 1.1) composed of the various molecular components of the cell – genes, transcripts, proteins, and metabolites – that offers an even more complex perspective on cellular organization. The different levels that administer information storage, information processing, and execution of cellular programs appear to be integrated rather than distinct. In other words, cellular functions are distributed among groups of heterogenous components that interact within large networks (Jeong et al., 2000). The elementary building blocks organize themselves in small recurrent patterns, called pathways in metabolism and motifs in gene regulatory networks that both together set up functional modules. Hierarchical nesting of modules defines the cell's large-scale architecture that is assumed to be shared across most species, while the degree of *universality* is gradually decreasing from functional modules over key metabolic pathways to the precise repertoire of components – genes, proteins, metabolites – that is *unique* to each organism.

The scale-free connectivity with embedded modularity, i.e. domination by a few highly connected nodes ("hubs") that provide the connections between modules responsible for distinct functions, appears to be valid beyond the scope of cellular organization.

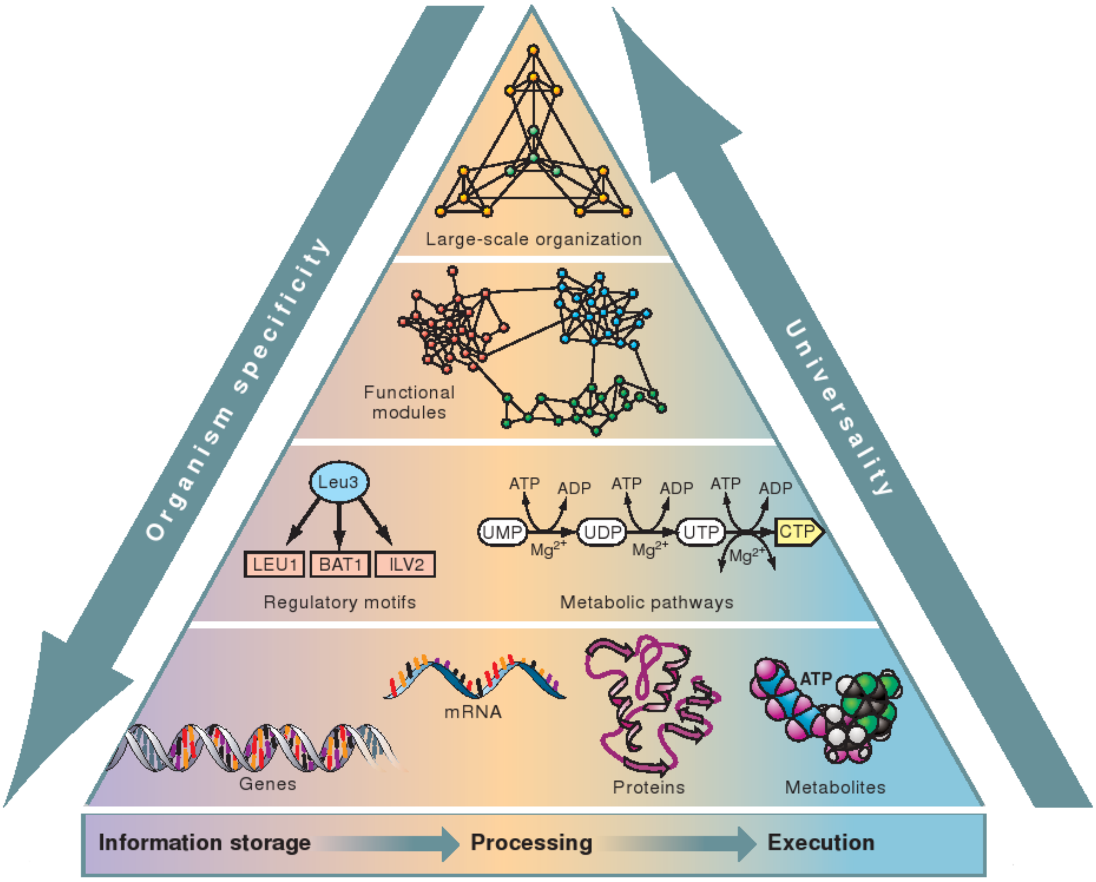


Figure 1.1.: Life's complexity pyramid composed of the molecular components of the cell (figure source: from Oltvai and Barabási, 2002). Net-like structures emerge on various levels of cellular organization.

Characteristic topologic properties are shared across networks as different as social networks and the World Wide Web.

However, creating stochastic models for inferring large-scale gene interaction on various levels is a challenging task. In principle, statistics offers a host of suitable models and inference approaches for learning genetic networks. Therein contained are various multivariate approaches, e.g., clustering and dimension reduction techniques, graphical models, time series models and many more. The big problem of eventually applying these models in systems biology is the amount of available data. Current experimental genomic data sets are huge. However, the advance in technology has increased the number of investigated features p while the number of samples n has not, and can not, similarly be increased. As a result, experimental data typically comprise measurements of between 10,000 (gene expression data) and > 1 million variables (single nucleotide polymorphisms – SNPs) for only a handful of samples (10 – 1,000).

In order to illuminate cellular and organismal function, it is crucial to choose biologically relevant models and at the same time provide methods for proper and robust inference in a "small n, large p" setting. Typically, high-dimensional experimental approaches cause ill-posed problems in statistical analysis. Referring to this, it is generally prudent to employ sparse and simple models which leads to the problem of finding a trade-off between the desired model complexity and analytical feasibility. Very generally, in a "small n, large p" data setup estimators benefit from using carefully regularized methods and exact tests only arrive at reliable conclusions. In some situations it may even be possible to exploit the high-dimensionality in genomic data sets, such that a seemingly disadvantage in the analysis, namely the large number of variables p, can effectively be turned into an advantage. Finally, it is of prime importance to make allowance for multiplicity issues.

This work concentrates on the particular demands that small-sample genomic data make with regard to modeling and inference. Regularization methods are of prime importance in this respect. Two major concepts are considered for improving upon inaccurate estimators of covariance and (partial) correlation matrices: bootstrap aggregation and shrinkage. The specific bioinformatical problem that careful attention is payed to in this work is the search for networked gene association patterns. Graphical models essentially go back to Wright's (1921) seminal work on path analysis. Subsequently, graphical modeling theory has experienced various extensions – see for example Whittaker

(1990) and Lauritzen (1996) for references – and has very generally emerged as a popular approach for elucidating stochastic associations and interdependencies in complex highly-structured multivariate data. However, given small-sample data such as from a functional genomics experimental approach, standard theories are no longer readily applicable. Instead it is necessary to introduce moderation and to provide specifically tuned methods for estimation, testing, and model selection. It is a common concern of the present work to establish methods that are straightforward to use in practice and that ensure feasible computational effort.

The outline of the thesis is as follows. Chapter 2 gives a brief introduction to molecular biology and microarray systems. Special focus is thereby on establishing an understanding of biomolecular fundamentals and of novel high-throughput technologies that provoke, e.g., issues of expression data pre-processing. In particular, the intention is to stress the distinctive dimensions of functional genomics data because it is these dimensions that motivate efforts to introduce regularization in standard modeling and inference concepts. Following some preliminary notes on graph-theoretic terminology, on (conditional) independence graphs, and on graphical models, in Chapter 3 a work review is presented with regard to the application of graphical Gaussian models (GGMs) in expression analysis and the various strategies that have emerged in this context in the literature to cope with the "small n, large p" problem.

Chapter 4 gives an extensive survey over current methods of covariance matrix estimation combined with the introduction and validation of new estimators of covariance and (partial) correlation matrices. Regularization is introduced in the form of bootstrap aggregation and of shrinkage methods combined with a recent analytic result from Ledoit and Wolf (2003) for computing the optimal shrinkage intensity. In particular the latter approach reveals distinct advantages in terms of efficiency and prediction accuracy and in terms of further favorable statistical properties such as positive definiteness.

The focus of Chapter 5 is on precise GGM network selection that is cast using a multiple testing procedure based on an *exact* correlation test for the individual edge inclusion problems. The massively parallel structure of the problem of inferring net-like interaction structures from large-scale genomic data allows for employing empirical Bayes methodology. This ensures identifiabiliy of the sampling distribution of correlation estimates under the null hypothesis. Similar to shrinkage, the idea of "borrowing strength across variables" is formalized. An indispensable prerequisite in this context is recognizing that biological knowledge can be exploited, namely the *sparsity* of molecular networks (only very few out of all possible edges in the network are expected to be truly present). In large-scale hypothesis testing situations where a selection effect is of concern rather than controlling the family-wise error rate, false discovery rate criteria are sensible. Graphical model selection using false discovery rate multiple testing constitutes an heuristic, yet fast and computationally efficient alternative to proper model search. Finally, the proposed small-sample framework for GGM network inference is investigated in an extensive simulation study with respect to estimation accuracy, model validation, and power analysis.

Chapter 6 introduces the lasso approach (Tibshirani, 1996) to covariance selection. The lasso L_1 penalty may cause some of the coefficients in the various neighborhood selection regression problems to become exactly zero. Thus, the lasso does a kind of continuous edge subset selection *per node* depending on the choice of the shrinkage factor. For synthetic data, its performance is compared to the empirical Bayes GGM network inference procedure proposed in Chapters 4 and 5.

Chapter 7 is concerned with real molecular data analysis. Firstly, the breast cancer data set published in West et al. (2001) is reanalyzed. Secondly, the shrinkage approach to GGM selection using empirical Bayes multiple testing (proposed in Chapters 4 and 5 of the present doctoral thesis), the lasso approach to GGM selection with choice of the penalty as suggested in Meinshausen and Bühlmann (2005a), and the standard and widely applied "gene relevance network" method (Butte et al., 2000) are contrasted using exemplifying gene expression data from an *E. coli* experiment (Schmidt-Heck et al., 2004). Finally, in Chapter 8 summary and discussion of the presented work lead to an exposure of possible directions for future research.

2. Network Biology: Understanding Cellular Functional Organization

Functional genomics expression profiling is promising to elucidate open problems in the life and medical sciences. For this purpose, bioanalytical efforts are undertaken, e.g., in order to detect diverse gene expression patterns in a simple pairwise comparison, comparisons under multiple conditions, or in a time course experiment. The aim is often to determine marker genes whose expression differs between different tumor types, for instance. By examining the level of gene expression, e.g., in cell populations of disease and pre-disease states, investigators aim to understand the steps of disease development and to identify the genes that are involved in it. In this context, *personalized* and *preventive medicine* put the focus of the healthcare system back on the individual patient. There is great diversity among human beings, and, in order to provide safe and efficacious treatments, it is necessary to consider their particular conditions and medical needs. For this purpose, novel biomarkers are added to the ever-increasing pharmacodiagnostical tool box. Disease markers allow for an early prediction and differential diagnosis. Efficacy and safety markers provide useful information regarding identification of patients who are responsive to a particular compound and of patients who are prone to serious adverse events. However, pharmacogenomics is a developing research field that is still in its infancy. Moreover, besides the genetic background gene-environment interactions are of prime importance, e.g. in the area of mental disorders.

2.1. Biological Background

A brief overview of the basic concepts of molecular biology that are relevant to this thesis stands at the beginning. Further details are referred to molecular biology textbooks (e.g., Li and Graur, 1991; Gibson and Muse, 2002).

Cells are the fundamental working units of every living system. All the instructions needed to direct their activities are contained within the double-stranded *DNA (deoxyribonucleic acid)* that can be found in the nucleus of all cells. Each of its strands consists of a sequence of nucleotides or bases. There are four different bases: adenine (A), thymine (T), cytosine (C), and guanine (G). The two strands of the DNA evolve from the four bases pairing in a particular manner: adenine pairs with thymine, while cytosine pairs with guanine. The human genome, for example, consists of some 3 billion base pairs. It should be noted that all of life's diversity results from the particular order of nucleotides in the genome. Put differently, the DNA is the ultimate depository of biological complexity.

The DNA is organized in physically separate molecules, the *chromosomes*. It contains instructions for the synthesis and regulation of *proteins* that determine shape, structure, and function of the cell. The instruction for making up a particular protein is coded on a segment of DNA which is called a *gene*. Each chromosome contains many genes, the basic physical and functional units of heredity. However, genes comprise only about 2% of the human genome, while the remainder consists of non-coding regions whose functions may include providing chromosomal structural integrity, for example. The human genome is estimated to contain 20,000 – 25,000 genes. However, not the number of genes, but the regulatory program is responsible for the diversity among organisms as different as worm and man.

Although genes get a lot of attention, it is the proteins that perform most life functions and make up the majority of cellular structures. Proteins are large, complex molecules made up of smaller subunits, the amino acids. Chemical properties, that distinguish 20 different amino acids, cause the protein chains to fold up into specific three-dimensional structures. The constellation of all proteins in a cell is called its *proteome*. Unlike the relatively unchanging genome, the dynamic proteome is constantly changing in response to a multitude of intra- and extracellular signals. Chemistry and behaviour of a protein are specified by the gene sequence and by the number and identities of other proteins with which it associates and reacts. Studies to explore protein structure and activities, known as *proteomics*, will be the focus of much research for decades to come and will help to elucidate the molecular basis of health and disease.

2.2. Introduction to Microarray Technology

Microarray technology is a powerful tool to monitor expressions of hundreds or thousands of genes in a single experiment. High-density array systems utilize the *central dogma* of molecular biology which is a two-step process:

1. An enzyme complex, the so-called RNA polymerase, *transcribes* the nucleotide sequence coding for a certain protein into single-stranded messenger RNA (mRNA) molecules. Transcription includes splicing in the nucleus where the large intron sequences are removed. The abundance of mRNA is widely denoted level of *gene expression*.

2. The synthesis of proteins yet requires further processing of the mRNA. Ribosomes in the cytosol *translate* the mRNA into corresponding proteins.

It should be noted that effectively *all* steps of expression are subject to active control by regulators cooperating in a complicated combinatorial manner. The fact that post-translational modifications are disregarded by microarray measurements is one important reason to handle these data with caution. Moreover, the modern as opposed to the above classical view of the central dogma of molecular biology suggests several mRNAs resulting from *alternative splicing* events that greatly increase the complexity of gene expression.

The process of conducting a microarray experiment is briefly outlined in the following. For high-density expression array production, gene specific probes are fixed on a solid support which is usually a glass microscope slide. More specifically, a *probe* consists of complementary DNA (cDNA) or oligonucleotides attached to the array surface.

The system of oligonucleotide expression arrays (Lockhart et al., 1996) is also known by the trademark *Affymetrix GeneChip*. Each gene is represented by 16–20 pairs of oligonucleotides with length 25 base pairs each that are referred to as *probe sets*. Each pair consists of a perfect match (PM) and a mismatch (MM) probe, where in the latter the middle base is changed with the intention to measure non-specific binding. In order to obtain an intensity value for each probe that represents the amount of corresponding mRNA in the original sample, a way has to be found to combine the 16–20 probe pair intensities. Irizarry et al. (2003) discuss the problem of defining an effective measure of gene expression using the probe level data and introduce a summary measure that is a robust multiarray average (RMA) of background-adjusted, normalized, and log-transformed PM values. The necessity of pre-processing the raw expression data is discussed in more depth later in this section.

In contrast, the underlying concept of cDNA microarrays is *competitive hybridization* between a sample that is labelled with the red-fluorescent dye cyanine 5 (Cy5) and a sample that is labelled with the green-fluorescent dye cyanine 3 (Cy3) referred to as the two channels of the two-color microarray experiment. The pairing of target samples for hybridization leads to a measure of *relative* abundance of two sets of mRNA.

For *sample* preparation, RNA is isolated from cells or tissues of interest, e.g., cells that have undergone a certain treatment *versus* control cells. RNA quality checking is sensible, e.g. by gel electrophoresis. Sample labeling follows, i.e. incorporation of fluorescent dyes or radioactivity. This process usually involves a reverse transcription step. Single color and dual color experiments are distinguished.

The so-called *target* cDNAs can *hybridize* to the complementary probe strand in compliance with the base pairing rules. Hybridization takes place during incubation of the microarrays for several hours with a hybe-mix containing the labeled cDNAs. Stringency washes to remove non-specific binding follow. After drying the array, laser scanning and quantification techniques yield an intensity value for each spot, i.e. for each gene represented on the array. The observed intensity is supposed to represent the mRNA concentration in the original sample. However, array and sample preparation as well as hybridization and subsequent steps are subject to introducing error. As a result, the noise level intrinsic to genomic data is high. *Systematic* errors can be reduced by an appropriate data "normalization" method. This aspect of expression analysis is addressed in the next section.

The problem of how to design a microarray experiment is of vital importance in order to ensure that the resulting data are amenable to statistical analysis and suitable for answering the scientific question of interest (e.g., Yang and Speed, 2002; Churchill, 2002). As a consequence, careful allocation of available ressources is necessary. Key issues include differentiation of sources of variation, namely between *biological* and *technical variation*. Biological variation is intrinsic to all organisms. It may be influenced by genetic or environmental factors. Thus, biological replicates are essential in order to draw conclusions that are valid beyond the scope of the particular samples that are assayed in the experiment. The common practice of *pooling* mRNA samples prior to hybridization profoundly affects biological variation. Contrariwise, technical variation is introduced during extraction, labeling, and hybridization of samples. Technical replicates, in turn, increase the precision of the results obtained. The ability mentioned before to directly compare two samples on the same microarray slide is a unique feature of the two-color microarray system. The repeated *dye-swap* experiment, where two arrays are used to compare two samples, is useful for reducing technical variation. On array 1, the control sample is assigned to the red dye, and the treatment sample is assigned to the green dye. On array 2, the dye assignments are reversed. Otherwise, when comparing treatment samples to a reference sample using microarrays each with the same orientation of dye labeling, dye effects are confounded with treatment effects. In general, choice of the reference RNA is a crucial issue to decide on.

Technical variation evolves from further sources: in array production, clone quality is one key issue (Halgren et al., 2001). Pre-screening of non-sequence verified clones is reasonable in order to eliminate contaminated clones from the probe set in cDNA microarray manufacturing. cDNA probe amplification (polymerase chain reaction – PCR) does not work for some clones resulting in too low concentrations of the PCR products. Furthermore, during the spotting process pin printing failures may lead to varying spot sizes and probes mixing on the slides, i.e. bad spot morphologies. Scratches, dust or other contaminations give rise to an inhomogenous array surface coating. Finally, choice of a quantification method that is not appropriate for the specific spot characteristics may result in systematically biased expression intensities.

Commercial microarray solutions are standardized with respect to array production and experimental protocols, i.e. they are subject to quality controlling. However, as these commercial solutions are expensive and nevertheless inflexible, because only standard sets of arrays are available, it is not uncommon that users opt for in-house microarray laboratories where each adheres to its own experimental protocol.

2.3. Calibration and Data Transformation

Due to the various sources of variation discussed in the previous section, the raw data are not the intended mRNA concentrations and can not directly be analyzed. Array and sample manufacturing, labeling and hybridization efficiencies, as well as further processing steps that finally yield intensity values are “not perfect”. However, the main problem is that these steps vary from array to array, i.e. experiment to experiment.

Systematic variation is characterized through similar effects regarding many measurements. Appropriate correction parameters can be estimated from the observed data. This procedure is denoted *calibration*, sometimes also “normalization”. In contrast, *stochastic* variation is too random to be explicitly accounted for.

Huber et al. (2002, 2003), e.g., provide a data pre-processing strategy that proceeds as follows to form a basis for statistical inference from microarray data.

- Firstly, each sample (array) is calibrated by an affine transformation, where it is possible to stratify the transformations within arrays. Stratification may be useful for spotted arrays, e.g., according to print-tip groups, and for oligonucleotide arrays, e.g., according to physico-chemical properties of the probe level data. The underlying assumption in the latter case is that probes of different sequence composition attract systematically different levels of (background) signal. The simplest case of only one stratum amounts to assuming that the data of one sample, i.e. all probes on an array, were subject to the same systematic effects, such that an array-wide calibration is sufficient.
- Secondly, the whole data are transformed by a variance-stabilizing transformation.

After these calibration and variance-stabilization steps, systematic array- or dye-biases should be removed, and the variance should be approximately independent of the mean expression intensity.

More specifically, if y_{ki} is the matrix of uncalibrated data, with $k, k = 1, \dots, n$, indexing the samples and $i, i = 1, \dots, p$, the genes, then the calibrated and variance-stabilized data h_{ki} are obtained through the parametric form

$$h_{ki} = \operatorname{arsinh}(a_{sk} + b_{sk} y_{ki}), \tag{2.1}$$

where $s \equiv s(i)$ is the stratum for probe i. a_{sk} and b_{sk} are the combined calibration and transformation parameters for probes from stratum s and sample k, i.e. off-set and proportionality factor, that account for non-specific and specific signal contributions, respectively. These parameters are estimated with a robust variant of maximum likelihood estimation (Huber et al., 2002, 2003). Finally, it should be noted that the approach assumes that the majority of genes is not differentially transcribed across the experiments. This needs to hold if the method is to produce meaningful results. Software is publicly available as the R package "vsn" through the Bioconductor project (`http://www.bioconductor.org`).

2.4. Cellular Networks

The behaviour of complex cellular and organismal systems emerges from the concerted activities of many interacting components such as genes and gene products. At a highly abstract level, the cooperating components can be considered as a set of vertices that are connected to each other, with links (edges) representing pairwise interactions. Vertices and edges together form a network or more formally a graph (cf. Section 3.1).

In practise, when the fairly fuzzy term of a "genetic network" is used, one of the following three types of cellular networks is meant.

Physical networks describe interactions between molecules, such as protein–protein, protein–nucleid-acid and protein–metabolite interactions that can easily be conceptualized within a graph theoretic description. Typically, in physical networks experimental approaches allow for determining the precise topology.

Metabolic networks, such as biochemical pathways, usually involve more complex functional interactions. However, they can also be looked at using the simplifying vertex-edge nomenclature. For example, substrates can be visualized as the nodes of the metabolic network, where edges represent enzyme-catalysed reactions that transform one metabolite into another. Metabolic networks are often modeled using differential equations.

In *genetic regulatory networks*, nodes represent individual genes and the respective links are derived, e.g., from (partial) correlation coefficients computed from observed microarray expression data. Special focus is on elucidating functional interaction structures and regulatory mechanisms.

In this work modeling and inferring genetic regulatory networks is considered. This type of network is also investigated by, e.g., Friedman et al. (2000), Hartemink et al.

(2002), and Dobra et al. (2004).

Networks can be directed as well as undirected. In directed networks, the interaction between any two nodes has a well-defined direction, which represents, for example, the direction of material flow from a substrate to a product in a metabolic reaction, or the direction of information flow from a transcription factor to the gene that it regulates (cited from Barabási and Oltvai, 2004). In contrast, in undirected networks, links are not assigned a direction. At first sight, this may seem less natural. However, consider as an example for the biological relevance of undirected network models protein interaction networks, where edges represent mutual binding relationships.

Despite the diversity of cellular networks, they all share a number of architectural features and are governed by a few fundamental principles that are valid even beyond the scope of network biology and equally apply to technological and social systems. Barabási and Albert (1999) introduce the concept of *scale-free* networks. The notion "scale-free" is meant to indicate the absence of a representative node in the network that can be used to characterize all nodes. The most elementary characteristic of any node in a network is its degree, i.e. connectivity k, that indicates the number of links the node has to other nodes. The degree distribution $P(k)$ gives the probability that a selected node has k links. For example, in random networks the node degree follows a Poisson distribution. Thus, most nodes have approximately the same number of links, whereas highly connected nodes, also known as hubs, are extremely rare. In contrast, power-law degree distributions with $P(k) \sim k^{-\gamma}$, where γ is the degree exponent, are characterized by a few hubs that hold together numerous small nodes. The node connectivity in cellular networks typically follows a power-law degree distribution with γ in the order of $2 < \gamma < 3$. Moreover, Hartwell et al. (1999) strongly argue for *modularity* in cellular functional organization. Scale-free network topology and modules, that are responsible for different processes in the cell, seamlessly integrate to hierarchical systems (Ravasz et al., 2002).

3. Graphical Models for Describing Gene Dependency Networks

From the previous section it becomes obvious that modeling and inferring network-like structures from genomic data is of prime importance in systems biology (cf. also Fig. 1.1). Very generally, complex stochastic associations and interdependencies can be described using graphical models (Whittaker, 1990; Lauritzen, 1996). These are parametric families of probability distributions for multivariate random vectors that obey certain (conditional) independence restrictions inherent in an independence graph. Graphical models are promising tools for the analysis of gene interaction because they allow the stochastic description of networked association and dependency structures in complex highly structured data. At the same time, graphical models offer an advanced statistical framework for inference. In theory, this makes them perfectly suited for modeling biological processes in the cell such as biochemical interactions and regulatory activities.

Consequently, many in part very complicated graphical models such as Bayesian networks (e.g., Friedman et al., 2000; Segal et al., 2003; Friedman, 2004), auto-regressive models (e.g., Yeung et al., 2002; De Hoon et al., 2003), state-space models (e.g., Murphy, 2002; Rangel et al., 2004), and graphical Gaussian models (e.g., Kishino and Waddell, 2000; Toh and Horimoto, 2002a; Wu et al., 2003; Dobra et al., 2004) have already been applied to genomic data and put to use in expression analysis.

3.1. Conditional Independence Graphs

In this section the requisite terminology and conventions concerning graphs are introduced that are relevant to this thesis. Special attention is given in Section 3.1.2 to conditional independencies reflected by graphical properties.

Very generally, for three random variables $\boldsymbol{X}, \boldsymbol{Y}$, and $\boldsymbol{Z}$ it is of interest to see wether dependence holds for one of them fixed in order to be able to distinguish direct from indirect dependencies. $\boldsymbol{X}$ and $\boldsymbol{Y}$ are *conditionally independent* given $\boldsymbol{Z}$ if and only if the density function of $\boldsymbol{X}$ conditional on $\boldsymbol{Y}$ and $\boldsymbol{Z}$, $f_{X|YZ}$, satisfies

$$f_{X|YZ}(\boldsymbol{x}; \boldsymbol{y}, z) = f_{X|Z}(\boldsymbol{x}; z)$$

for all values of $\boldsymbol{x}$ and $\boldsymbol{y}$ and for all z with $f_Z(z) > 0$. This is written as $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}$ and intuitively interpreted as follows: knowing $\boldsymbol{Z}$ renders $\boldsymbol{Y}$ irrelevant for predicting $\boldsymbol{X}$.

An equivalent characterization of $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}$ is that the joint density $f_{XYZ}(\boldsymbol{x}, \boldsymbol{y}, z)$ can be *factorized* into the product of two factors, one not involving $\boldsymbol{x}$ and the other not involving $\boldsymbol{y}$, i.e.

$$f_{XYZ}(\boldsymbol{x}, \boldsymbol{y}, z) = g(\boldsymbol{x}, z) h(\boldsymbol{y}, z), \tag{3.1}$$

where g and h are some functions.

3.1.1. Graph Theory

A *graph* $\mathcal{G} = (V, E)$ consists of a finite set of vertices $V = \{1, \ldots, p\}$ and a set of edges $E \subseteq V \times V$ corresponding to (conditional) dependencies. $\boldsymbol{X} = (X_v, v \in V)$ are the associated labeled variables. For $i, j \in V, i \neq j$, unordered pairs $\{i, j\}$ are distinguished from ordered pairs (i, j). While the former are connected through an undirected edge ("line"), the latter are connected through a directed edge ("arrow"), $i \longrightarrow j$. Although directed graphs look more intuitive, they turn out rather more subtle and complicated than undirected graphs.

A *conditional independence graph* $\mathcal{G} = (V, E)$ is *undirected* if it has only undirected edges and $\{i, j\}$ is not in the edge set if and only if $X_i \perp\!\!\!\perp X_j \mid X_{V \setminus \{i,j\}}$.

Including directed edges in graph-theoretic descriptions along with studying conditional density functions is a natural way to portray asymmetries in the roles of interacting variables. A graph $\mathcal{G} = (V, E)$ is *directed* if it has only directed edges.

For $A \subset V$, let $\boldsymbol{X}_A = (X_v, v \in A)$ and $\mathcal{X}_v$ the state space of X_v. Similarly, $\boldsymbol{x}_A = (x_v, v \in A) \in \mathcal{X}_A = \times_{v \in A} \mathcal{X}_v$. The *induced subgraph* $\mathcal{G}_A$ is defined as (A, E_A) with $E_A = (A \times A) \cap E$. A graph $\mathcal{G}$ is *complete* if all pairs of vertices are *adjacent*, i.e. joint by an undirected or directed edge. A maximally complete subgraph is denoted *clique*. Furthermore,

- the set $\mathrm{nb}(A) = \{k \in V \setminus A \mid \exists\, j \in A : \{j, k\} \in E\}$ is called the *neighbors* of A.
- the set $\mathrm{pa}(A) = \{k \in V \setminus A \mid \exists\, j \in A : (k, j) \in E\}$ is called the *parents* of A.
- the set $\mathrm{ch}(A) = \{k \in V \setminus A \mid \exists\, j \in A : (j, k) \in E\}$ is called the *children* of A.
- the set $\mathrm{bd}(A) = \mathrm{pa}(A) \cup \mathrm{nb}(A)$ is called the *boundary* of A.
- the set $\mathrm{cl}(A) = \mathrm{bd}(A) \cup A$ is called the *closure* of A.
- the set $\mathrm{an}(A) = \{k \in V \setminus A \mid \exists\, j \in A$ with a path from k to $j\}$ is called the *ancestors* of A.
- the set $\mathrm{de}(A) = \{k \in V \setminus A \mid \exists\, j \in A$ with a path from j to $k\}$ is called the *descendants* of A.
- the set $\mathrm{nd}(A) = V \setminus (\mathrm{de}(A) \cup A)$ is called the *non-descendants* of A.
- if $\mathrm{bd}(A) = \emptyset$, then A is called *ancestral.* $\mathrm{An}(A) = \mathrm{an}(A) \cup A$ is the *ancestral set* of A including A.

An ordered $(m+1)$-tuple $(j_0, \ldots, j_m)$ of distinct vertices is called a *path of length m from j_0 to j_m* if $\{j_{i-1}, j_i\} \in E$ or $(j_{i-1}, j_i) \in E$ for all $i = 1, \ldots, m$. It is called *undirected* if $\{j_{i-1}, j_i\} \in E\ \forall\, i = 1, \ldots, m$. It is called *semidirected* if $\exists\, i : \{j_{i-1}, j_i\} \notin E$ and *directed* if the latter holds for all $i = 1, \ldots, m$. A path of length m with $j_0 = j_m$ is called a *cycle*. Undirected, semidirected and directed cycles are defined analogously to the above path definitions.

However, directed cycles, that can be considered as modeling "feed-back", are not allowed because there is no well defined density function for this situation. Consequently, $\mathcal{G}$ is a *directed acyclic* graph (DAG) if it has only directed edges and no directed cycles. This is equivalent to presupposing the existence of a complete ordering of the vertices that provides each variable with a past, present and future. In the *moral graph* $\mathcal{G}^m$ of $\mathcal{G}$ parents of common children are linked and all edges are made undirected.

Finally, non-adjacent vertices i and j are *separated* by $S \subset V$ if and only if every path between i and j contains at least one element of S. It is tempting to conclude that the associated variables are independent conditional on the separating set alone. Theoretical justification for this intuitive interpretation establishes the global Markov property. For the multivariate normal distribution the proof is given by Speed and Kiiveri (1986).

3.1.2. Markov Properties

Markov properties relate graphical separations and conditional independencies inherent in the statistical model for a system V of labeled random variables $\boldsymbol{X} = (X_i,\ i \in V)$.

Markov properties for undirected graphs. Let $\mathcal{G} = (V, E)$ an undirected graph and $\boldsymbol{X} = (X_i,\ i \in V)$ a random vector with joint density function f_X. Say f_X satisfies (P) *the pairwise Markov property* if

$$\text{for any non-adjacent } i, j \in V \Longrightarrow X_i \perp\!\!\!\perp X_j \mid X_{V \setminus \{i,j\}};$$

(L) *the local Markov property* if

$$\text{for any } i \in V \Longrightarrow X_i \perp\!\!\!\perp X_{V \setminus \mathrm{cl}(i)} \mid X_{\mathrm{bd}(i)};$$

(G) *the global Markov property* if

$$\text{for any disjoint } A, B, S \subset V \text{ such that } S \text{ separates } A \text{ and } B \text{ in } \mathcal{G} \Longrightarrow X_A \perp\!\!\!\perp X_B \mid X_S\,.$$

It is a remarkable fact that the three Markov properties: pairwise Markov, local Markov and global Markov, are equivalent, when $f(\boldsymbol{x}) > 0$.

For $C \subseteq V$, $\psi_C(\boldsymbol{x})$ denotes a non-negative potential function that depends on $\boldsymbol{X}_C$ only. The density of $\boldsymbol{X}$ *factorizes* with respect to $\mathcal{G}$ or satisfies (F) if

$$f(\boldsymbol{x}) = \prod_{C \in \mathcal{C}} \psi_C(\boldsymbol{x}), \tag{3.2}$$

where $\mathcal{C}$ is the collection of cliques in $\mathcal{G}$.

In the case of positive density, $f(\boldsymbol{x}) > 0$ for all $\boldsymbol{x}$, Eq. 3.2 coincides with the three Markov properties:

$$\text{(F)} \Longleftrightarrow \text{(G)} \Longleftrightarrow \text{(L)} \Longleftrightarrow \text{(P)}.$$

It is noteworthy that a directed independence graph $\mathcal{G}$ possesses the Markov properies of its associated moral graph, $\mathcal{G}^m$ (Whittaker, 1990).

Markov properties for directed acyclic graphs. Let $\mathcal{G} = (V, E)$ a DAG and $\boldsymbol{X} = (X_i,\ i \in$

V) a random vector with joint density function f_X. Say f_X satisfies
(P) *the pairwise (directed) Markov property* if

$$\text{for any non-adjacent } i, j \in V \text{ with } j \in \mathrm{nd}(i) \Longrightarrow X_i \perp\!\!\!\perp X_j \mid \mathrm{nd}(i) \setminus \{j\}$$

(L) *the local (directed) Markov property* if

$$\text{for any } i \in V \Longrightarrow X_i \perp\!\!\!\perp \mathrm{nd}(i) \mid \mathrm{pa}(i)$$

(G) *the global (directed) Markov property* if

$$\text{for any disjoint } A, B, S \subset V \text{ such that } S \text{ separates } A \text{ and } B \text{ in } \mathcal{G}^m_{\mathrm{An}(A,B,S)} \Longrightarrow X_A \perp\!\!\!\perp X_B \mid X_S\,.$$

3.2. Covariance Graphs

In order to elucidate functional gene interaction, as well as in order to form a basis for subsequent clustering and refined network inference, an intuitive and simple idea is to look at the sample correlation between any two genes. If the computed Pearson's or Spearman's correlation coefficient, e.g., exceeds a certain a priori specified threshold (say 0.8), then an edge is drawn between the appropriate genes with the vague aim to exclude spurious edges. This approach is widely applied in the bioinformatics community, and the resulting graph is often called a *relevance network* (Butte et al., 2000). In statistical terminology it is known as *covariance graph* where missing edges denote *marginal independence*.

However, for understanding gene interaction this approach is only of limited use. For instance, a high standard correlation coefficient between two genes may be indicative of either (i) direct interaction, (ii) indirect interaction, or (iii) regulation by a common gene. In learning a genetic network from data we need to be able to distinguish among these three alternatives.

3.3. Graphical Gaussian Models

For constructing a *gene association network* where only direct interactions among genes are depicted by edges, another framework provides a better option. *Graphical Gaussian models* (GGMs), also known as *covariance selection* or *concentration graph* models, have recently become a popular tool to study gene dependency networks. The key idea behind GGMs is to use partial correlations as a measure of *conditional (in)dependence* between any two genes. This makes it straightforward to distinguish direct from indirect interactions. However, it should be noted that partial correlations are related to the inverse of the correlation matrix.

The best starting place to learn about GGMs is the paper that introduced this concept in the early 1970s (Dempster, 1972). Further details can be found in the books by Whittaker (1990) and by Edwards (1995).

GGMs are similar to the more widely known Bayesian networks in that the underlying concept is conditional independence. However, in contrast to Bayesian networks GGMs contain only undirected rather than directed edges. This makes graphical Gaussian interaction modeling on the one hand conceptually more simple, and on the other hand also potentially more widely applicable. For example, the conditional independence properties inherent in a Bayesian network model are reflected through graphical separations in the corrsponding directed acyclic graph: modeling feed-back loops is not possible.

Under the GGM approach the data $\boldsymbol{X}$ are assumed to be mutually independent and p-variate normally distributed $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with some mean vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^T$ and positive definite variance-covariance matrix $\boldsymbol{\Sigma} = (\sigma_{ij})$, where $1 \leq i, j \leq p$. Via $\sigma_{ij} = \rho_{ij}\sigma_i\sigma_j$ the covariance matrix can be decomposed into variance components σ_i^2, $i = 1, \ldots, p$, and Pearson's correlations $\boldsymbol{P} = (\rho_{ij})$.

The multivariate normal density is given as

$$f(\boldsymbol{x}) = (2\pi)^{-p/2} \, |\boldsymbol{\Sigma}|^{-1/2} \, \exp\left\{-(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})/2\right\}, \tag{3.3}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are called the moment parameters. In exponential family terminology, alternative parametrization is given through canonical parameters that are defined as $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$. Then the multivariate normal density from Eq. 3.3 can be rewritten as

$$
\begin{aligned}
f(\boldsymbol{x}) &= \exp\left\{\alpha + \boldsymbol{\beta}^T \boldsymbol{x} - \boldsymbol{x}^T \boldsymbol{\Omega} \boldsymbol{x}/2\right\} \\
&= \exp\left\{\alpha + \sum_{i=1}^{p} \beta_i x_i - \sum_{i=1}^{p} \sum_{j=1}^{p} \omega_{ij} x_i x_j / 2\right\},
\end{aligned}
\tag{3.4}
$$

where α is the normalizing constant. $\boldsymbol{\Omega} = (\omega_{ij})$ is called precision or concentration matrix. Using the factorization criterion (Eq. 3.1) it becomes obvious that the interrelation between X_i and X_j given the remaining $p-2$ variables is entirely dictated by $\omega_{ij} = 0$ or not.

In the GGM framework the strength of direct pairwise correlation is characterized by the partial correlation matrix $\tilde{\boldsymbol{P}} = (\tilde{\rho}_{ij})$. These coefficients describe the correlation between any two genes i and j conditional on all the remainder of the genes. Standard graphical modeling theory (e.g. Edwards, 1995) shows that the matrix $\tilde{\boldsymbol{P}}$ is related to the inverse of the covariance matrix $\boldsymbol{\Sigma}$. This leads to a straightforward procedure to compute $\tilde{\boldsymbol{P}}$ via the relations

$$
\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1} = (\omega_{ij}) \tag{3.5}
$$

and

$$
\tilde{\rho}_{ij} = -\omega_{ij} / \sqrt{\omega_{ii}\omega_{jj}}. \tag{3.6}
$$

It should be noted that in the inversion step (Eq. 3.5) it is equally valid to use the correlation matrix $\boldsymbol{P}$ instead of the covariance matrix $\boldsymbol{\Sigma}$. Random variables i and j are partially uncorrelated for given $V \setminus \{i, j\}$ if

$$
\omega_{ij} = 0 \text{ and } \tilde{\rho}_{ij} = 0,
$$

respectively.

Put differently, the partial correlation between random variables i and j conditional on $V \setminus \{i, j\}$ is the correlation between their residuals after linearly regressing i and j, respectively, on $V \setminus \{i, j\}$:

$$
\operatorname{Corr}\left\{X_i - \mathrm{E}(X_i \mid X_{V\setminus\{i,j\}}), X_j - \mathrm{E}(X_j \mid X_{V\setminus\{i,j\}})\right\}.
$$

Equivalently, the partial correlation between random variables i and j conditional on

$V \setminus \{i, j\}$ is defined as

$$\tilde{\rho}_{ij} = \operatorname{sign}(\beta_i^{(j)}) \sqrt{\beta_i^{(j)} \beta_j^{(i)}}, \tag{3.7}$$

where $\beta_j^{(i)}$ denotes the regression coefficient of predictor variable X_j for the response X_i, when linearly regressing each variable $i \in \{1, \dots, p\}$ on the remaining set of $p - 1$ variables. Note that while in general $\beta_i^{(j)} \neq \beta_j^{(i)}$ the signs of two non-zero coefficients are identical as $\beta_j^{(i)} = \omega_{ij}/\omega_{ii}$.

Partial correlation coefficients allow for a number of further interpretations. As the multivariate normal distribution is closed under marginalization and conditioning, the partial correlation $\tilde{\rho}_{ij}$ is the correlation coefficient of the conditional bivariate distribution for genes i and j. Furthermore, assuming normality it can be shown that two variables are conditionally independent given the remaining variables if and only if the corresponding partial correlation vanishes. Equivalently, the conditional independence graph of a jointly normal set of random variables is determined by the location of zeros in the inverse covariance matrix $\boldsymbol{\Omega}$ (Whittaker, 1990).

In order to reconstruct a GGM network from a given data set one typically employs the following procedure.

- Firstly, an estimate of the covariance matrix $\boldsymbol{\Sigma}$ is obtained, usually via the unbiased sample covariance matrix $\boldsymbol{S} = (s_{ij})$.
- Secondly, estimates of partial correlation coefficients are computed from the sample covariance matrix using Eq. 3.6.
- Thirdly, statistical tests are employed to determine which entries in the estimated partial correlation matrix $\hat{\tilde{\boldsymbol{P}}} = \tilde{\boldsymbol{R}}$ are significantly different from zero.
- Fourthly, the inferred conditional independence structure is visualized by a graph, with edges corresponding to non-zero partial correlation coefficients.

The likelihood function relates the information content in an observed sample to the unknown parameters of the statistical model under consideration. It enables us to assess which parameter values are well supported by the observed data, and which are not. Standard results and techniques of maximum likelihood estimation and likelihood ratio tests can be found in the book by Cox and Hinkley (1974). Recall that for the multivariate normal distribution, $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, inherent (conditional) independencies are expressed by the covariance matrix $\mathbf{\Sigma}$ or its inverse $\mathbf{\Omega}$. Naturally, either parameterization can be used as the correspondence between $\mathbf{\Sigma}$ and $\mathbf{\Omega}$ is one to one. Overall, there are $p(p+1)/2$ free parameters of which p are concerned with scale and $p(p-1)/2$ with interaction. It is predominantly these interaction parameters that are interesting to the graphical modeller. By contrast, the mean value, $\boldsymbol{\mu}$, is not at all important and thus, it is allowed to be entirely arbitrary. Let us consider a given conditional independence graph where each pairwise conditional independence constraint generates a constraint on the parameters, namely a zero in the corresponding entry of $\mathbf{\Omega}$. Note that the same constraint expressed in terms of the covariance parameter σ_{ij} is substantially more complicated. Speed and Kiiveri (1986) describe an iterative proportional fitting procedure for computing the maximum likelihood estimate $\boldsymbol{S}$ subject to the constraints that determine the putative graphical model. A natural way to measure the overall goodness of fit, and to compare two competing models when one is nested in the other, is the *deviance*. Its sampling distribution under the null hypothesis of the considered independence structure follows an asymptotic chi-squared distribution with degrees of freedom given by the number of constraints set on $\mathbf{\Sigma}$ and $\mathbf{\Omega}$, respectively.

3.4. Addressing the “Small n, Large p” Problem

Data dimensions peculiar to functional genomics approaches are challenging for statistical modeling and inference. Unfortunately this implies that, although graphical models are promising for the analysis of gene interaction, their practical application is currently strongly limited by the amount of available experimental data. At first, this may seem paradoxical given today’s high-throughput facilities. It should be noted however, while these technologies now allow to investigate experimentally a greatly increased number of features (genes), the number of available samples has not, and can not, similarly be expanded. As a result, in a typical microarray data set the number of genes p will exceed by far the number of sample points n. This poses a serious challenge to any statistical inference procedure, and also renders estimation of genetic networks an extremely hard problem. This is corroborated by a recent study on the popular Bayesian network method where Husmeier (2003) demonstrated that this approach tends to perform poorly on sparse microarray data.

Motivated by these challenges, great efforts are now being undertaken to further extend the theory of graphical models to allow their large-scale application on small-sample data (e.g., Wong et al., 2003; Dobra et al., 2004). In this section several recently developed approaches to small-sample inference of graphical Gaussian modeling are reviewed and strategies to cope with the high dimensionality of functional genomics data are discussed. In my understanding, all of these papers fit in one of three categories: (i) classic GGM theory, (ii) analysis using limited order partial correlations, and (iii) application of regularized GGMs.

Kishino and Waddell (2000) were the first to propose GGMs as suitable statistical models for association structures among genes. However, a number of difficulties arise when the graphical Gaussian modeling concept is applied to the analysis of high-dimensional data such as from a microarray experiment. Firstly, standard GGM theory (Whittaker, 1990) can only be applied when $n > p$, because otherwise the sample covariance and correlation matrices are not positive definite, which in turn prevents the computation of partial correlations. Moreover, there are often additional linear dependencies between the variables, which leads to the problem of multicollinearity. This, again, renders standard theory of graphical Gaussian modeling inapplicable to microarray data. Secondly, the statistical tests widely used in the literature for selecting an appropriate GGM (e.g. deviance tests) are valid only for large sample size, and hence are inappropriate for the very small sample sizes present in microarray data sets. In this case, instead of asymptotic tests an exact model selection procedure is required.

Moreover, it should be noted that the "small n, large p" problem affects both GGMs and relevance networks. Although less obvious in the latter case, it should not be overlooked that standard correlation estimates are not reliable for small sample size n. However, this fact appears to have gone largely unnoticed in the bioinformatics community.

3.4.1. Dimension Reduction Prior to Classic GGM Analysis

In order to avoid the dimensionality problems mentioned above, the most obvious and simplest approach is to restrict graphical Gaussian modeling to assess relationships among either a rather small number of genes (Kishino and Waddell, 2000; Waddell and Kishino, 2000; Bay et al., 2002; Wang et al., 2003) or among a small number of clusters of genes (Toh and Horimoto, 2002a,b; Wu et al., 2003). The number p of selected genes or gene clusters has to be chosen such that it does not exceed the sample size n.

However, this strategy is unsatisfying for a variety of reasons. Primarily, it is a matter of on-going debate how to choose reasonable (meta)-genes for inclusion in the reduced data set. The restriction to a limited number of genes risks that the estimated network topology is seriously distorted because important genes may have been excluded from the analysis. Furthermore, the resulting partial correlation coefficients for gene clusters and the corresponding conditional dependence properties are hard to understand. For instance, typically, not all the genes of one cluster will interact with all the genes of another cluster, which renders conditional dependence properties among clusters meaningless. In addition, information regarding quality and strength of the association on the gene level is lost when only clusters of genes are considered.

3.4.2. Limited Order Partial Correlations

Another possibility to tackle the “small n, large p” problem is to compute partial correlation coefficients of limited order. For instance, de la Fuente et al. (2004) propose to calculate partial correlation coefficients up to second-order only, i.e. to condition the partial correlations not on all other $p-2$ genes as in a full GGM but only on two genes at most. Similar strategies, based on first-order conditional dependence, are also employed by Wille et al. (2004),Wille and Bühlmann (2005), and Magwene and Kim (2004).

From a statistical point of view the resulting gene network constitutes something in-between a full GGM and a relevance network model based on standard correlations. It therefore remains unclear whether missing edges indicate conditional or marginal independence. A measure of distance in networks is the path length that indicates the number of links between two selected nodes. The mean path length represents the average over the shortest paths between all pairs of nodes and offers a measure of overall navigability. In genetic networks, interactions are likely to be short range. Thus, we believe that the above methods may provide a good approximation.

3.4.3. Regularized GGMs

In my opinion the statistically and also biologically most sound way to marry GGMs with small-sample modeling is to introduce regularization and moderation. In the first instance, this boils down to finding suitable estimates for the covariance matrix and its inverse when n is smaller than p. This can either be done in a full Bayesian, or in an empirical Bayesian manner. A further possibility constitutes the explicit frequentist penalization of the number of free parameters in $\tilde{\boldsymbol{P}}$. Typically, once regularized estimates of partial correlation are available, heuristic or stochastic model searches need subsequently to be employed in order to find an optimal graphical model or set of models.

Outside a genomic context using regularized GGMs was first proposed by Wong et al. (2003). For gene expression data this strategy is pursued in the paper by Dobra et al. (2004) who describe a variant of Bayesian covariance selection. However, it should be noted that full Bayesian Markov chain Monte Carlo methods such as in Dobra et al. (2004) are computationally very expensive. As efficient alternative Meinshausen and Bühlmann (2005a) employ lasso regression for covariance selection. Further details on the lasso approach to high-dimensional GGM selection are referred to Chapter 6. In Chapter 5, an empirical Bayes approach to large-scale GGM selection using false discovery rate multiple testing is proposed after introducing novel regularized estimates of covariance and (partial) correlation in the next chapter.

4. Regularized Large-Scale Covariance and (Partial) Correlation Matrix Estimation

The simple solution to obtain an accurate and reliable estimate of the population covariance matrix is to rely either on the maximum likelihood estimate $\boldsymbol{S}^{\mathrm{ML}}$ or on the related unbiased empirical covariance matrix $\boldsymbol{S} = \frac{n}{n-1}\boldsymbol{S}^{\mathrm{ML}}$, with entries defined as

$$s_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j), \tag{4.1}$$

where $\bar{x}_i = \frac{1}{n}\sum_{k=1}^{n} x_{ki}$ and x_{ki} is the k-th observation of the variable X_i. However, unfortunately both $\boldsymbol{S}$ and $\boldsymbol{S}^{\mathrm{ML}}$ exhibit serious defects in the "small n, large p" data setting commonly encountered in functional genomics problems. Specifically, in this case the empirical covariance matrix can *not* anymore be considered a good approximation of the true covariance matrix. It should be noted that this is true already for moderately sized data with $n \approx p$.

For illustration consider Fig. 4.1 where the conventional sample covariance $\boldsymbol{S}$ is compared with an alternative estimator $\boldsymbol{S}^{\star}$ developed in Subsection 4.1.3 and summarized in Tab. 4.1. Fig. 4.1 shows the sorted eigenvalues of the estimated matrices in comparison with the true eigenvalues for fixed $p = 100$ and various ratios $\frac{p}{n}$. It becomes evident that for small n the eigenvalues of $\boldsymbol{S}$ (thin black line in Fig. 4.1) are severely distorted. In addition, for $n < p$ (bottom row in Fig. 4.1) $\boldsymbol{S}$ looses its full rank as a growing number of eigenvalues become zero. This has several undesirable consequences. Firstly, $\boldsymbol{S}$ is not positive definite any more, and secondly, it can not be inverted as it becomes singular (e.g., Friedman, 1989; Hastie and Tibshirani, 2004). Now contrast the poor performance

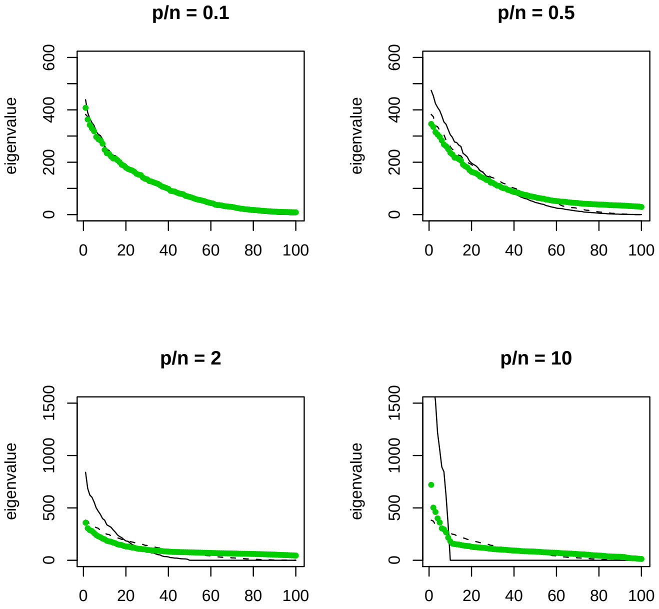


Figure 4.1.: Ordered eigenvalues of the sample covariance matrix $\boldsymbol{S}$ (thin black line) and of an alternative estimator $\boldsymbol{S}^\star$ (fat green line, for definition see Tab. 4.1), calculated from simulated data with underlying p-variate normal distribution, for $p = 100$ and various ratios p/n. The true eigenvalues are indicated by a thin black dashed line.

“Small n, Large p” Covariance and Correlation Estimators $S^\star$ and $R^\star$:

$$s^\star_{ij} = \begin{cases} s_{ii} & \text{if } i = j \\ r^\star_{ij}\sqrt{s_{ii}s_{jj}} & \text{if } i \neq j \end{cases}$$

and

$$r^\star_{ij} = \begin{cases} 1 & \text{if } i = j \\ r_{ij}\min(1, \max(0, 1 - \hat{\lambda}^\star)) & \text{if } i \neq j \end{cases}$$

with

$$\hat{\lambda}^\star = \frac{\sum_{i\neq j} \widehat{\mathrm{Var}}(r_{ij})}{\sum_{i\neq j} r_{ij}^2},$$

Table 4.1.: Small-sample shrinkage estimators of the unrestricted covariance and correlation matrix, where s_{ii} and r_{ij} denote the empirical variance (unbiased) and correlation, respectively. For details of the computation of $\widehat{\mathrm{Var}}(r_{ij})$ see the main text. Further variants of these estimators are discussed in Subsection 4.1.3.

of $\boldsymbol{S}$ with that of $\boldsymbol{S}^\star$ (fat green line in Fig. 4.1). This improved estimator exhibits none of the defects of $\boldsymbol{S}$, in particular it is more accurate, well conditioned and always positive definite – even for small sample size. Nevertheless, $\boldsymbol{S}^\star$ can be computed in only about twice the time required to calculate $\boldsymbol{S}$. These are good reasons against the blind use of the empirical covariance matrix $\boldsymbol{S}$ in data situations where it is not appropriate – noting that this affects *many* current application areas in bioinformatics.

The incontrovertible fact that the two widely-employed estimators of the covariance matrix, i.e. the maximum likelihood estimator $\boldsymbol{S}^{\mathrm{ML}}$ and the related unbiased sample covariance $\boldsymbol{S}$, are both statistically inefficient in small samples, has long been known. In a nutshell, it can be explained as the so-called “Stein phenomenon” discovered by Stein (1956) in the context of estimating the mean of a multivariate normal distribution. Stein demonstrated that in high-dimensional inference problems it is often possible to improve (sometimes dramatically!) upon the maximum likelihood estimator. This result is at first counterintuitive, as maximum likelihood can be proven to be *asymptotically* optimal, and as such it seems not unreasonable to expect that these favorable properties of maximum likelihood also extend to finite data. However, further insight into the Stein effect is provided by Efron (1982) who points out that one needs to distinguish between two different aspects of the maximum likelihood principle. Firstly, maximum likelihood as a means of summarizing the observed data and producing a *maximum likelihood summary* (MLS). Secondly, maximum likelihood as a procedure to obtain a *maximum likelihood estimate* (MLE). The conclusion is that maximum likelihood is unassailable as a data summarizer but that it has some clear limitations as an estimating procedure.

This applies directly to the estimation of covariance matrices: $\boldsymbol{S}^{\mathrm{ML}}$ constitutes the best estimator in terms of actual fit to the data. However, for medium to small sample sizes it is far from being the optimal estimator for recovering the population covariance as is well illustrated by Fig. 4.1. Fortunately, the Stein theorem also demonstrates that it is possible to construct a procedure for improved covariance matrix estimation. In addition to increased efficiency and accuracy, it is desirable for such a method to exhibit the following characteristics *not* found in $\boldsymbol{S}$ and $\boldsymbol{S}^{\mathrm{ML}}$:

1. The estimate should always be positive definite, i.e. all eigenvalues should be distinct from zero.

2. The estimated covariance matrix should be well-conditioned.

The positive definiteness requirement is an intrinsic property of the true covariance matrix that is satisfied as long as the considered random variables have non-zero variance. If a matrix is well-conditioned, i.e. if the ratio of its maximum and minimum singular value is not too large, it has full rank and can be easily inverted. Thus, by producing a well-conditioned covariance estimate one automatically also obtains an equally well-conditioned estimate of the *inverse covariance* – a quantity of crucial importance, e.g., in classification problems and in graphical models (cf. Section 3.3).

A rather naive strategy to obtain a positive definite estimator of the covariance matrix runs as follows: take the sample covariance $\boldsymbol{S}$ and apply, e.g., the algorithm by Higham (1988). This will adjust all eigenvalues to be larger than some prespecified threshold ϵ and thus guarantee positive definiteness. However, the resulting matrix will *not* be well conditioned.

4.1. Shrinkage Combined With Analytic Determination of the Intensity According to the Ledoit-Wolf Theorem

“Shrinkage” or more general “biased estimation” (e.g., Hoerl and Kennard, 1970b,a; Efron, 1975; Efron and Morris, 1975, 1977; Tikhonov and Arsenin, 1977) as a means of improvement upon unreliable estimates is investigated. From the well-known bias-variance decomposition of the mean squared error (MSE) for the sample covariance, i.e.

$$\mathrm{MSE}(\boldsymbol{S}) = \mathrm{Bias}(\boldsymbol{S})^2 + \mathrm{Var}(\boldsymbol{S}), \tag{4.2}$$

it becomes evident that in small samples a variance-reduced biased estimator for the covariance may outperform the unbiased unconstrained classical estimator $\boldsymbol{S}$. Put differently, sacrifying a little bit of bias in order to reduce the variance of the estimated high-dimensional parameter, may improve overall estimation accuracy. The most widely applied shrinkage method is ridge regression, also known as Tikhonov regularization. The ridge regression solution adds a positive constant to the diagonal of $\boldsymbol{X}^T\boldsymbol{X}$ before inversion. This makes the problem nonsingular, even if $\boldsymbol{X}^T\boldsymbol{X}$ is not of full rank, and was the main motivation for ridge regression when it was first introduced in statistics (Hoerl and Kennard, 1970b). The complexitiy parameter that controls the amount of shrinkage is typically chosen by cross-validation. A recent analytic result from Ledoit and Wolf (2003) is considered here for determining the shrinkage intensity that allows to construct an improved estimator of the covariance matrix $\boldsymbol{\Sigma}$ that is not only suitable for small sample size n and large number of variables p but at the same time is also completely inexpensive to compute.

4.1.1. Outline of Shrinkage Estimation and the Lemma of Ledoit-Wolf

In the following the *general* principles behind shrinkage estimation are reviewed and an analytic approach by Ledoit and Wolf (2003) for determining the optimal shrinkage intensity is discussed. It should be noted that the theory outlined here is not restricted to covariance estimation but applies generally to large-dimensional estimation problems.

Hence, let $\boldsymbol{\Psi} = (\psi_1, \ldots, \psi_p)$ denote the parameters of the unrestricted high-dimensional model of interest, and $\boldsymbol{\Theta} = (\theta_i)$ the matching parameters of a lower dimensional restricted submodel. For instance, $\boldsymbol{\Psi}$ could be the mean vector of a p-dimensional multivariate normal, and $\boldsymbol{\Theta}$ the vector of a corresponding constrained submodel where the means are all assumed to be equal, i.e. $\theta_1 = \theta_2 = \ldots = \theta_p$. By fitting each of the two different models to the observed data associated estimates $\boldsymbol{U} = \hat{\boldsymbol{\Psi}}$ and $\boldsymbol{T} = \hat{\boldsymbol{\Theta}}$ are obtained. Clearly, the unconstrained estimate $\boldsymbol{U}$ will exhibit a comparatively high variability due to the larger number of parameters that need to be fitted, whereas its low-dimensional counterpart $\boldsymbol{T}$ will have lower variance but potentially also considerable bias when taken as an estimator for the true $\boldsymbol{\Psi}$.

Instead of choosing between one of these two extremes, the linear shrinkage approach suggests to *combine* both estimators in a weighted average

$$\boldsymbol{U}^{\star} = \lambda \boldsymbol{T} + (1 - \lambda) \boldsymbol{U}, \tag{4.3}$$

where $\lambda \in [0, 1]$ denotes the shrinkage intensity. It should be noted that for $\lambda = 1$ the shrinkage estimate equals the shrinkage target $\boldsymbol{T}$ whereas for $\lambda = 0$ the unrestricted estimate $\boldsymbol{U}$ is recovered. The key advantage of this construction is that it offers a systematic way to obtain a regularized estimate $\boldsymbol{U}^{\star}$ that outperforms the individual estimators $\boldsymbol{U}$ and $\boldsymbol{T}$ both in terms of accuracy and of statistical efficiency.

A key question in this procedure is how to select an appropriate value for the shrinkage intensity λ. In some instances, it may suffice to fix the intensity λ at some given value, or to make it depend on the sample size according to some simple function. Often more appropriate, however, is choosing the parameter λ in a data-driven fashion by explicitly minimizing the expectation of a suitable loss function (risk function)

$$\begin{aligned} R(\lambda) &= E(L(\lambda)) \\ &= E\left(\sum_{i=1}^{p} (u_i^{\star} - \psi_i)^2 \right) \\ &= E\left(\sum_{i=1}^{p} (\lambda t_i + (1 - \lambda) u_i - \psi_i)^2 \right), \end{aligned} \tag{4.4}$$

here for example the mean squared error (MSE).

One common but computationally intensive approach is to estimate the minimizing $\lambda^\star$ by cross-validation – for an example see Friedman (1989) where shrinkage is applied in the context of regularized classification. Another widely applied route to determining λ views the shrinkage problem in an empirical Bayes context. In this case the quantity $E(\boldsymbol{T})$ is interpreted as prior mean and λ as a hyper-parameter that may be estimated from the data via optimizing the marginal likelihood (e.g., Morris, 1983; Greenland, 2000).

It is less well known that the optimal regularization parameter $\lambda^\star$ may often also be determined *analytically*. Ledoit and Wolf (2003) recently derived a simple theorem that guarantees minimal MSE without the need of having to specify any underlying distributions, and without requiring any computationally expensive procedures such as MCMC, the bootstrap, or cross-validation. Assuming that the first two moments of the distributions of $\boldsymbol{U}$ and of $\boldsymbol{T}$ exist, the expected squared error loss from Eq. 4.4 may be expanded as follows:

$$\begin{aligned} R(\lambda) &= \sum_{i=1}^{p} \operatorname{Var}(u_i^\star) + \left[E(u_i^\star) - \psi_i\right]^2 \\ &= \sum_{i=1}^{p} \operatorname{Var}(\lambda t_i + (1-\lambda)u_i) + \left[E(\lambda t_i + (1-\lambda)u_i) - \psi_i\right]^2 \\ &= \sum_{i=1}^{p} \lambda^2 \operatorname{Var}(t_i) + (1-\lambda)^2 \operatorname{Var}(u_i) + 2\lambda(1-\lambda)\operatorname{Cov}(u_i, t_i) \\ &\quad + \left[\lambda E(t_i - u_i) + \operatorname{Bias}(u_i)\right]^2 . \end{aligned} \tag{4.5}$$

Analytically minimizing this function with respect to λ gives, after some tedious algebraic calculations, a simple expression for the optimal value

$$\lambda^\star = \frac{\sum_{i=1}^{p} \operatorname{Var}(u_i) - \operatorname{Cov}(t_i, u_i) - \operatorname{Bias}(u_i)\, E(t_i - u_i)}{\sum_{i=1}^{p} E\left[(t_i - u_i)^2\right]}, \tag{4.6}$$

for which minimum MSE $R(\lambda^\star)$ is achieved. It can be shown that $\lambda^\star$ always exists and that it is unique. If $\boldsymbol{U}$ is an *unbiased* estimator of $\boldsymbol{\Psi}$, i.e. $E(\boldsymbol{U}) = \boldsymbol{\Psi}$, this equation reduces to

$$\lambda^\star = \frac{\sum_{i=1}^{p} \operatorname{Var}(u_i) - \operatorname{Cov}(t_i, u_i)}{\sum_{i=1}^{p} E\left[(t_i - u_i)^2\right]}, \tag{4.7}$$

which is – apart from some further algebraic simplification – the expression given in

Ledoit and Wolf (2003).

Closer inspection of Eq. 4.6 yields a number of insights into how the optimal shrinkage intensity is chosen:

1. The smaller the variance of the high-dimensional estimate $\boldsymbol{U}$, the smaller becomes $\lambda^\star$. Therefore, with increasing sample size the influence of the target $\boldsymbol{T}$ diminishes.

2. $\lambda^\star$ also depends on the correlation between estimation error of $\boldsymbol{U}$ and of $\boldsymbol{T}$. If both are positively correlated then the weight put on the shrinkage target decreases. Hence, the inclusion of the second term in the numerator of Eq. 4.6 adjusts for the fact that the two estimators $\boldsymbol{U}$ and $\boldsymbol{T}$ are both inferred from the same data set. It also takes into account that the "prior" information associated with $\boldsymbol{T}$ is not independent from the given data.

3. If the unconstrained estimator is biased, and the bias points already towards the target, the shrinkage intensity is correspondingly reduced.

4. With increasing mean squared difference between $\boldsymbol{U}$ and $\boldsymbol{T}$ (in the denominator of Eq. 4.6) the weight $\lambda^\star$ also decreases. Note that this automatically protects the shrinkage estimate $\boldsymbol{U}^\star$ against a misspecified target $\boldsymbol{T}$.

Furthermore, it is noteworthy that variables that by design are kept identical in the constrained and unconstrained estimators (i.e. $t_i = u_i$) play no role in determining the intensity $\lambda^\star$, as their contributions to the various sums in Eq. 4.6 cancel out.

Further generalization is possible by allowing for *multiple targets* or *different* shrinkage intensities. This is especially appropriate if there exists a natural grouping of parameters in the investigated high-dimensional model. In this case one simply computes the individual targets and applies Eq. 4.6 to each group separately. Partitioning into a small number of groups, e.g., would be conceivable according to the variables' variances $\mathrm{Var}(u_i)$ – this is typically the predominant term in determining the shrinkage level according to Eq. 4.6.

Finally, it is important to consider the transformation properties of the shrinkage procedure. From Eq. 4.6 it is clear that $\lambda^\star$ is invariant against translations. For instance, the underlying data may be centered without affecting the estimation of the optimal shrinkage intensity. However, $\lambda^\star$ is *not* generally invariant against scale transformations. This dependence on the absolute scales of the considered variables is a general property that shrinkage shares with other approaches to biased estimation, such as ridge regression and partial least squares (e.g. Hastie et al., 2001). Ultimately, it is a consequence of the selected risk function (MSE).

4.1.2. Estimation of the Optimal Shrinkage Intensity

For practical application of Eq. 4.6 one needs to obtain an estimate $\hat{\lambda}^\star$ of the optimal shrinkage intensity. In their paper Ledoit and Wolf (2003) emphasize that the parameters of Eq. 4.6 should be estimated consistently. However, this is only a very weak requirement, as consistency is an asymptotic property and a basic requirement of any sensible estimator. Furthermore, specific focus is on small sample inference. Thus, in order to compute $\hat{\lambda}^\star$, it is suggested that all expectations, variances, and covariances in Eq. 4.6 are replaced instead by their *unbiased* sample counterparts. This leads to

$$\hat{\lambda}^\star = \frac{\sum_{i=1}^{p} \widehat{\mathrm{Var}}(u_i) - \widehat{\mathrm{Cov}}(t_i, u_i) - \widehat{\mathrm{Bias}}(u_i)\,(t_i - u_i)}{\sum_{i=1}^{p}(t_i - u_i)^2}. \tag{4.8}$$

It should be noted that in finite samples $\hat{\lambda}^\star$ may exceed 1 and in some cases it may even become negative. Therefore, in order to avoid overshrinkage or negative shrinkage $\hat{\lambda}^{\star\star} = \max(0, \min(1, \hat{\lambda}^\star))$ is employed when constructing the shrinkage estimator via Eq. 4.3.

It is also noteworthy that Eq. 4.8 is valid regardless of the sample size n at hand. In particular, n may be substantially smaller than p.

4.1.3. Shrinkage Estimation of the Covariance Matrix

Estimation of the unrestricted covariance matrix requires the determination of $(p^2+p)/2$ free parameters, and thus constitutes a high-dimensional inference problem. Consequently, application of shrinkage offers a promising approach to obtain improved estimates.

Daniels and Kass (2001) provide a fairly extensive review of empirical Bayes shrinkage estimators proposed in recent years. Unfortunately, most of the suggested estimators appear to suffer from at least one of the following drawbacks, which renders them unsuitable for the analysis of genomic data:

1. Typically, the application is restricted to data with $p < n$, in order to ensure that the empirical covariance $\boldsymbol{S}$ can be inverted. However, most current genomic data sets contain vastly more features than samples ($p \gg n$).

2. Many of the suggested estimators are computationally expensive due to, e.g., being based on MCMC sampling, or they require specific distributional assumptions.

These difficulties are elegantly avoided by resorting to the (almost) distribution-free Ledoit-Wolf approach to shrinkage.

In a matrix setting the equivalent to the squared error loss function is the Frobenius norm. Thus,

$$\begin{aligned} L(\lambda) &= \|\boldsymbol{S}^\star - \boldsymbol{\Sigma}\|_{\mathrm{F}}^2 \\ &= \|\lambda \boldsymbol{T} + (1-\lambda)\boldsymbol{S} - \boldsymbol{\Sigma}\|_{\mathrm{F}}^2 \\ &= \sum_{i=1}^{p}\sum_{j=1}^{p}\left(\lambda t_{ij} + (1-\lambda)s_{ij} - \sigma_{ij}\right)^2 \end{aligned} \tag{4.9}$$

is a natural quadratic measure of distance between the true and the estimated covariance matrix, $\boldsymbol{\Sigma}$ and $\boldsymbol{S}^\star$, respectively. In this formula the unconstrained unbiased empirical covariance matrix $\boldsymbol{S}$ replaces the unconstrained estimate $\boldsymbol{U}$ of Eq. 4.3.

Selecting a suitable empirical covariance target $\boldsymbol{T} = (t_{ij})$ requires some diligence. In general, the choice of the target should be guided by the presumed lower-dimensional structure in the data as this determines the increase of efficiency over the sample covariance. However, it is also a remarkable consequence of Eq. 4.6 that in fact *any* type of shrinkage will lead to a reduction in MSE, albeit only a minor one in case of a strongly misspecified target. Then $\boldsymbol{S}^\star$ will simply reduce to the unconstrained estimate $\boldsymbol{S}$.

Six commonly used covariance targets are compiled in Tab. 4.2, along with a brief description, the dimension of the target, and the resulting estimate $\hat{\lambda}^\star$. It is noteworthy that the resulting shrinkage estimators $\boldsymbol{S}^\star$ all exhibit the same order of algorithmic complexity as the standard estimate $\boldsymbol{S}$.

In order to estimate the optimal shrinkage intensity $\hat{\lambda}^\star$ (Eq. 4.8) for the various structured covariance targets listed in Tab. 4.2, it is necessary to obtain unbiased estimates

Target A: “diagonal, unit variance” 0 estimated parameters $t_{ij} = \begin{cases} 1 & \text{for } i = j \\ 0 & \text{for } i \neq j \end{cases}$ $\hat{\lambda}^\star = \frac{\sum_{i\neq j} \widehat{\mathrm{Var}}(s_{ij}) + \sum_i \widehat{\mathrm{Var}}(s_{ii})}{\sum_{i\neq j} s_{ij}^2 + \sum_i (s_{ii}-1)^2}$	**Target B:** “diagonal, common variance” 1 estimated parameter: v $t_{ij} = \begin{cases} v = \mathrm{avg}(s_{ii}) & \text{for } i = j \\ 0 & \text{for } i \neq j \end{cases}$ $\hat{\lambda}^\star = \frac{\sum_{i\neq j} \widehat{\mathrm{Var}}(s_{ij}) + \sum_i \widehat{\mathrm{Var}}(s_{ii})}{\sum_{i\neq j} s_{ij}^2 + \sum_i (s_{ii}-v)^2}$
Target C: “common (co)variance” 2 estimated parameters: v, c $t_{ij} = \begin{cases} v = \mathrm{avg}(s_{ii}) & \text{for } i = j \\ c = \mathrm{avg}(s_{ij}) & \text{for } i \neq j \end{cases}$ $\hat{\lambda}^\star = \frac{\sum_{i\neq j} \widehat{\mathrm{Var}}(s_{ij}) + \sum_i \widehat{\mathrm{Var}}(s_{ii})}{\sum_{i\neq j} (s_{ij}-c)^2 + \sum_i (s_{ii}-v)^2}$	**Target D:** “diagonal, unequal variance” p estimated parameters: s_{ii} $t_{ij} = \begin{cases} s_{ii} & \text{for } i = j \\ 0 & \text{for } i \neq j \end{cases}$ $\hat{\lambda}^\star = \frac{\sum_{i\neq j} \widehat{\mathrm{Var}}(s_{ij})}{\sum_{i\neq j} s_{ij}^2}$
Target E: “perfect positive correlation” p estimated parameters: s_{ii} $t_{ij} = \begin{cases} s_{ii} & \text{for } i = j \\ \sqrt{s_{ii} s_{jj}} & \text{for } i \neq j \end{cases}$	**Target F:** “constant correlation” $p + 1$ estimated parameters: s_{ii}, $\bar{r}$ $t_{ij} = \begin{cases} s_{ii} & \text{for } i = j \\ \bar{r}\sqrt{s_{ii} s_{jj}} & \text{for } i \neq j \end{cases}$
$f_{ij} = \frac{1}{2}\left\{ \sqrt{\frac{s_{jj}}{s_{ii}}}\widehat{\mathrm{Cov}}(s_{ii}, s_{ij}) + \sqrt{\frac{s_{ii}}{s_{jj}}}\widehat{\mathrm{Cov}}(s_{jj}, s_{ij}) \right\}$	
$\hat{\lambda}^\star = \frac{\sum_{i\neq j} \widehat{\mathrm{Var}}(s_{ij}) - f_{ij}}{\sum_{i\neq j} (s_{ij} - \sqrt{s_{ii} s_{jj}})^2}$	$\hat{\lambda}^\star = \frac{\sum_{i\neq j} \widehat{\mathrm{Var}}(s_{ij}) - \bar{r} f_{ij}}{\sum_{i\neq j} (s_{ij} - \bar{r}\sqrt{s_{ii} s_{jj}})^2}$

Table 4.2.: Six different shrinkage targets for the covariance matrix and associated estimators of the optimal shrinkage intensity. In general, target D is recommended – see the main text for discussion. *Abbreviations:* v, average of sample variances; c, average of sample covariances; $\bar{r}$, average of sample correlations.

for the variance and for the covariance of the individual entries in the matrix $\boldsymbol{S} = (s_{ij})$. Let x_{ki} be the k-th observation of the variable X_i and $\bar{x}_i = \frac{1}{n}\sum_{k=1}^{n} x_{ki}$ its sample mean. Now set $w_{kij} = (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)$ and $\bar{w}_{ij} = \frac{1}{n}\sum_{k=1}^{n} w_{kij}$. Then the unbiased empirical covariance equals

$$\widehat{\text{Cov}}(x_i, x_j) = s_{ij} = \frac{n}{n-1}\bar{w}_{ij}$$

and, correspondingly, the variance is

$$\widehat{\text{Var}}(x_i) = s_{ii} = \frac{n}{n-1}\bar{w}_{ii}.$$

The empirical unbiased variances and covariances of the *individual entries* of $\boldsymbol{S}$ are computed in a similar fashion.

$$\widehat{\text{Var}}(s_{ij}) = \frac{n^2}{(n-1)^2}\widehat{\text{Var}}(\bar{w}_{ij}) = \frac{n}{(n-1)^2}\widehat{\text{Var}}(w_{ij}) = \frac{n}{(n-1)^3}\sum_{k=1}^{n}(w_{kij} - \bar{w}_{ij})^2. \tag{4.10}$$

Similarly,

$$\widehat{\text{Cov}}(s_{ij}, s_{lm}) = \frac{n}{(n-1)^3}\sum_{k=1}^{n}(w_{kij} - \bar{w}_{ij})(w_{klm} - \bar{w}_{lm}). \tag{4.11}$$

Moments of higher order than $\widehat{\text{Var}}(s_{ij})$, in particular variances and covariances of *averages* of s_{ij}, are neglected in estimating the optimal $\hat{\lambda}^{\star}$ in Tab. 4.2.

Probably the most commonly employed shrinking targets are the identity matrix and its scalar multiple. These are denoted in Tab. 4.2 “diagonal, unit variance” (target A) and “diagonal, common variance” (target B). A further extension is provided by the two parameter covariance model that in addition to the common variance (as in target B) also maintains a common covariance (“common (co)variance”, target C). The three targets share several properties. Firstly, they are all extremely low-dimensional (0 to 2 free parameters). As a result they impose a rather strong structure which in turn requires only little data to fit. Secondly, the resulting estimators *shrink all components* of the empirical covariance matrix, i.e. both diagonal and off-diagonal entries.

In the literature it is easy to find examples where one of the above targets is employed – albeit *not* in combination with analytic estimation of the shrinkage level. For instance, the unit diagonal target A is typically used in ridge regression and the related Tikhonov regularization (e.g. Hastie et al., 2001). The target B is utilized, e.g., by Friedman

(1989) who estimates λ by means of cross-validation, by Leung and Chan (1998) who use a fixed $\lambda = \frac{2}{n+2}$, by Dobra et al. (2004) as a parameter in an inverse Wishart prior for the covariance matrix, and finally also by Ledoit and Wolf (2004b). The two-parameter target C appears not to be widely used.

Another class of covariance targets is given by the "diagonal, unequal variance" model (target D), the "perfect positive correlation" model (target E) and the "constant correlation" model (target F) of Tab. 4.2. A common feature of these three targets is that they are comparatively parameter-rich, and that they only lead to *shrinkage of the off-diagonal elements* of $\boldsymbol{S}$. The last two shrinkage targets were introduced with the purpose of modeling stock returns. These tend – on average – to be strongly positively correlated (Ledoit and Wolf, 2003, 2004a).

Special focus here is on the shrinkage target D for the estimation of covariance and of correlation matrices arising in genomics problems. This "diagonal, unequal variance" model represents a compromise between the low-dimensional targets A, B, and C and the correlation models E and F. Like the simpler targets A and B it shrinks the off-diagonal entries to zero. However, unlike shrinkage targets A and B, target D leaves diagonal entries intact, i.e. it does *not* shrink the variances. Thus, this model assumes that the parameters of the covariance matrix fall into two classes, and both are treated differently in the shrinkage process.

This clear separation also suggests that for shrinking purposes it may be useful to parameterize the covariance matrix in terms of variances and correlations (rather than variances and covariances) so that $s^\star_{ij} = r^\star_{ij}\sqrt{s_{ii}s_{jj}}$. In this formulation, shrinkage is applied to the correlations rather than covariances. This has two distinct advantages. Firstly, the off-diagonal elements determining the shrinkage intensity are all on the same scale. Secondly, the (partial) correlations derived from the resulting covariance estimator $\boldsymbol{S}^\star$ are independent of scale and location transformations of the underlying data matrix, just as is the case for those computed from $\boldsymbol{S}$.

It is this form of target D that is proposed in this work for estimating correlation and covariance matrices. For reference, the corresponding formulae are collected in Tab. 4.1. Note the remarkably simple expression for the shrinkage intensity

$$\hat{\lambda}^\star = \frac{\sum_{i\neq j} \widehat{\mathrm{Var}}(r_{ij})}{\sum_{i\neq j} r_{ij}^2} \tag{4.12}$$

– see also Tab. 4.2 (Target D). The variance Var(r_{ij}) of the empirical correlation coefficients can be estimated as follows: simply apply the above formula from Eq. 4.10 to the *standardized* data matrix. This procedure treats the estimated variances as constants and hence introduces a slight but generally negligible error. The same assumption also justifies to ignore the bias of the empirical correlation coefficients in Eq. 4.12. In this formula a concern may be the use of the empirical correlation coefficients r_{ij} – after all, these are the ones that are to be improved! Thus, it seems we face a circularity problem, namely that for an accurate estimate of the shrinkage intensity reliable estimates of correlation are needed, and vice versa. However, it is a remarkable feature of target D that it completely resolves this issue: regardless whether standard or shrinkage estimates of correlation are substituted into Eq. 4.12 the resulting $\hat{\lambda}^\star$ remains all the same.

Using the target D has another important advantage: the resulting shrinkage covariance estimate will automatically be positive definite. The target D itself is always positive definite, and the convex combination of a positive definite matrix ($\boldsymbol{T}$) with another matrix that is positive semidefinite ($\boldsymbol{S}$) always yields a positive definite matrix. Note that this is also true for targets A and B but *not* for the targets C, E, and F (consider as counterexample the target E with all variances set equal to one).

Further variants of the proposed estimator (Tab. 4.1) are easily constructed. One possible extension is to shrink the diagonal elements as well, using a different intensity for variances and for correlations. Shrinking the variances to a common mean is standard practice in genomic case-control studies (e.g. Cui et al., 2005). However, in these instances there is typically so little data that the gene-specific variances are difficult to obtain, let alone covariances. In contrast, modeling net-like relationships strongly depends on the correlations among genes. Consequently, network inference is a more demanding task than screening for differential expression, and thus requires a correspondingly larger sample size. As a result, for these data it would be expected that there is at least sufficient information to correctly estimate the variances (in which case shrinking would not be necessary).

4.2. Estimating Partial Correlation from Small Samples

In order to obtain reliable estimates of partial correlation, conceptually simple but effective variations of the standard estimate (Eq. 3.6) are considered. Firstly, when inverting the standard correlation estimator $\boldsymbol{R}$ the Moore-Penrose pseudoinverse is employed. Secondly, bootstrap aggregation (bagging) is used to stabilize the classical estimators of correlation and of partial correlation, respectively. It turns out that bagging of the sample correlation matrix $\boldsymbol{R}$ acts as an implicit regularization procedure and that the bagged estimate is always positive definite (cf. Friedman, 1989). Thirdly, shrinkage as outlined in the previous section is applied to improve upon the sample covariance and correlation estimates, respectively. The resulting estimates $\boldsymbol{S}^{\star}$ and $\boldsymbol{R}^{\star}$ (Tab. 4.1) are positive definite by construction and thus can be easily inverted.

The Moore-Penrose pseudoinverse (Penrose, 1955) is a generalization of the standard matrix inverse that can also be applied to singular matrices and that is based on the singular value decomposition (SVD). The correlation matrix $\boldsymbol{P}$ can be decomposed into $\boldsymbol{P} = \boldsymbol{U}\,\boldsymbol{D}\,\boldsymbol{V}^T$ where $\boldsymbol{D}$ is a square diagonal matrix of rank $m \leq \min(n, p)$ containing all non-vanishing singular values. The pseudoinverse $\boldsymbol{P}^{+}$ is then defined as $\boldsymbol{P}^{+} = \boldsymbol{V}\,\boldsymbol{D}^{-1}\,\boldsymbol{U}^T$ and requires only the trivial inversion of $\boldsymbol{D}$. It can be shown that the pseudoinverse $\boldsymbol{P}^{+}$ is the shortest length least squares solution of $\boldsymbol{P}\boldsymbol{P}^{+} = \boldsymbol{I}$, where $\boldsymbol{I}$ denotes the identity matrix. Hence it reduces to the standard matrix inverse where possible. Otherwise it amounts to simply ignoring all zero singular values and corresponds to 0th-order regularization.

Bootstrap aggregation offers a simple and very general nonparametric approach to variance reduction (Breiman, 1996) and thus to improve upon an unstable estimator $\hat{\theta}(y)$ for a given set of data y. The Monte Carlo algorithm proceeds as follows:

1. Generate a bootstrap sample y^{*b} with replacement from the original sample. Repeat this process $b = 1, \ldots, B$ times independently (e.g. with $B = 1000$).

2. For each bootstrap sample y^{*b} calculate the estimate $\hat{\theta}^{*b}$.

3. Compute the bootstrap mean $\frac{1}{B}\sum_{b=1}^{B} \hat{\theta}^{*b}$ to obtain the bagged estimate.

Another interpretation of the bagged estimate is as an approximate Bayesian posterior mean estimate (Hastie et al., 2001).

Shrinkage estimation combined with analytic determination of the shrinkage intensity as described above offers an appealing route to obtain reliable estimates of the covariance and correlation matrix in small samples that may prove useful beyond the scope of gene network analysis in many bioinformatical problems.

These techniques allow to construct small-sample estimators of the partial correlation matrix $\tilde{\boldsymbol{P}} = (\tilde{\rho}_{ij})$ (Eq. 3.6). In particular, in this thesis the following possibilities are considered:

$\hat{\tilde{\boldsymbol{P}}}^1$ **:** Use the pseudoinverse for inverting the sample correlation matrix $\hat{\boldsymbol{P}}$ in order to obtain an estimate of $\tilde{\boldsymbol{P}}$, without performing any form of bagging
(= "pseudoinverse partial correlation").

$\hat{\tilde{\boldsymbol{P}}}^2$ **:** Use bagging to estimate the correlation matrix $\boldsymbol{P}$, then invert the bagged correlation matrix to obtain an estimate of $\tilde{\boldsymbol{P}}$
(= "partial bagged correlation").

$\hat{\tilde{\boldsymbol{P}}}^3$ **:** Apply bagging to the estimator $\hat{\tilde{\boldsymbol{P}}}^1$, i.e. use the pseudoinverse for inverting each bootstrap replicate estimate $\hat{\boldsymbol{P}}^{*b}$, then average the results
(= "bagged partial correlation").

$\hat{\tilde{\boldsymbol{P}}}^4$ **:** Use the shrinkage covariance estimator from Eq. 4.3 comprised by target D and the estimated intensity $\hat{\lambda}^\star$ (cf. Tab. 4.1) followed by inversion to obtain an estimate of $\tilde{\boldsymbol{P}}$
(="shrinkage partial correlation").

By construction all four of these estimators can be applied to cases where the sample size is smaller than the number of variables. However, they differ drastically with respect to accuracy as can be seen below in the section on computer simulations. Moreover, especially for the very large dimensions commonly encountered in genomics problems (often with $p > 1,000$) the two bootstrap approaches are computationally very demanding.

5. Small-Sample Inference of Genetic Networks

In this chapter two simple approaches are considered for inferring networked dependency structures from complex gene expression data, both of which require as input an estimated large-scale covariance matrix. The first and conceptually simpler model is that of a "gene relevance network" that was introduced by Butte et al. (2000) and that is built in the following simple fashion. Firstly, the $p \times p$ correlation matrix $\boldsymbol{P} = (\rho_{ij})$ is estimated from the data. Secondly, a correlation test needs to be employed to the individual entries and test results should become adjusted for multiplicity. Thus, relevance networks represent the *marginal* (in)dependence structure among the p genes. In statistical terminology this type of network model is known as "covariance graph".

Despite the popularity of relevance networks which stems from the relative ease of construction there are many problems connected with their proper interpretation. For instance, the cut-off value that determines the "significant" edges is typically chosen in a rather arbitrary fashion – often simply a large value is selected (say $|r| > 0.8$) with the vague aim to exclude "spurious" edges. However, this misses the statistical interpretation of marginal correlation which takes account of both direct as well as indirect associations. As a straightforward consequence, in a reasonably well-connected genetic network most genes will by construction be correlated with each other – for an example see the analysis of the *Escherichia coli* expression data in Chapter 7. Thus, in this case even a high observed degree of correlation will provide only weak evidence for the direct dependency of any two considered genes. Instead, the *absence of correlation* will be a *strong measure of their independence*. Therefore, even ignoring the difficulties with obtaining accurate measures of correlation from small-sample data, gene relevance networks are suitable tools *not* for elucidating the dependence network among genes but rather for uncovering independence structures!

By contrast, with the class of graphical Gaussian models (GGMs), also known as "covariance selection" or "concentration graph" models, a simple statistical approach exists that allows to detect direct dependencies between genes. This "gene association network" approach is based on investigating the estimated *partial* correlations $\tilde{r}_{ij}$ for all pairs of considered genes. In a small-sample setting both the estimation of the partial correlations, i.e. connection strength in gene association networks, as well as the subsequent model selection procedure need to be suitably modified. In the following an empirical Bayes approach for identifying the precise network topology in graphical Gaussian models is discussed to allow for their large-scale application on small-sample data. The proposed network inference procedure is investigated with respect to power and other performance criteria in an extensive simulation study.

Recall that unfortunately, the naive strategy to try all potentially adequate models and to evaluate their goodness of fit is impossible given that the number of possible network topologies grows super-exponentially with the number of nodes. Thus, an exhaustive network enumeration is necessarily limited to toy cases only and by far not conceivable for the large number of investigated features in functional genomics problems. Textbook methods such as stepwise selection procedures traditionally based on asymptotic edge deletion chi-squared tests, are not reliable for the small sample sizes usually encountered in genomics problems. As an heuristic but fast and computationally efficient alternative to proper network selection, large-scale false discovery rate multiple testing of all possible edges with an *exact* correlation test is employed.

In order to address the statistical testing problem of non-zero partial correlation

$$H_0 : \tilde{\rho}_{ij} = 0 \quad \text{versus} \quad H_1 : \tilde{\rho}_{ij} \neq 0, \tag{5.1}$$

the *sampling distribution* of $\hat{\tilde{\rho}}_{ij} = \tilde{r}_{ij}$ under the null hypothesis $\tilde{\rho}_{ij} = 0$ is asked for. For convenience, subscripts ij are dropped in the following.

From Hotelling (1953) the distribution of the sample normal correlation coefficient

$\hat{\rho} = r$ is known exactly. For $\rho = 0$ we have

$$\begin{aligned} f_0(r;\kappa) &= \left(1 - r^2\right)^{(\kappa-3)/2} \frac{\Gamma\left(\frac{\kappa}{2}\right)}{\pi^{\frac{1}{2}}\Gamma\left(\frac{\kappa-1}{2}\right)} \\ &= |r| \,\mathrm{Be}\left(r^2; \frac{1}{2}, \frac{\kappa-1}{2}\right), \end{aligned} \tag{5.2}$$

where $\mathrm{Be}(x; a, b)$ is the Beta distribution and κ is the degree of freedom and the reciprocal variance of r, i.e. $\mathrm{Var}(r) = \frac{1}{\kappa}$. For the standard correlation coefficient the degree of freedom κ equals $n - 1$, i.e. is determined by the sample size n.

The sample normal *partial* correlation coefficient $\hat{\tilde{\rho}} = \tilde{r}$ is distributed precisely as the standard correlation coefficient $\hat{\rho} = r$, only that κ is reduced by the number of eliminated variables (Hotelling, 1953). Thus, if there are p variables of which $p - 2$ have to be eliminated in order to compute the pairwise partial correlation coefficients, the resulting degree of freedom is $\kappa = n - 1 - (p - 2) = n - p + 1$. Note that this relationship implies that n cannot be smaller than p if κ is to remain positive!

Furthermore, in a small-sample setting we cannot use the standard sample versions of partial correlations $\tilde{\rho}$ (Eq. 3.6) but rather have to rely on alternative estimators such as $\hat{\tilde{\boldsymbol{P}}}^1$, $\hat{\tilde{\boldsymbol{P}}}^2$, $\hat{\tilde{\boldsymbol{P}}}^3$, and $\hat{\tilde{\boldsymbol{P}}}^4$ suggested above. Unfortunately, the sampling distributions of these estimators cannot analytically be derived. However, it can be shown numerically (see section *Simulation Study* for details) that their respective simulated sampling distributions still assume the distributional form of Eq. 5.2, albeit with a smaller variance and hence with $\kappa > 0$ even for $n < p$. It should be noted that in this case the degree of freedom κ is not a simple function of n and p but rather has to be estimated itself from the data.

5.1. Robbins-Efron-Type Inference of Empirical Null Distribution

In principle, given an appropriate choice of κ, Eq. 5.2 allows to compute p-values for estimated partial correlation coefficients and thus to perform statistical testing with regard to the presence of edges in a GGM network.

As repeated estimates of the partial correlation coefficient per individual edge are not available, it is not trivial to estimate the degree of freedom κ. However, the highly parallel structure of the edge testing problem and the fact that biomolecular networks are typically *sparse* (e.g. Yeung et al., 2002) can be utilized. In a network considering p genes there is a large number $m = p(p-1)/2$ of possible edges. Only a small fraction of these will correspond to true edges, whereas for the remaining majority the corresponding true partial correlation coefficients will vanish. Therefore we may assume that the observed partial correlation coefficients $\hat{\tilde{\rho}} = \tilde{r}$ *across all edges* in the network follow a mixture density

$$f(\tilde{r}) = \eta_0 f_0(\tilde{r}; \kappa) + (1 - \eta_0) f_A(\tilde{r}), \tag{5.3}$$

where f_0 is the null distribution, η_0 is the (unknown) large proportion of "null edges" (say $\eta_0 \geq 0.9$), and f_A the distribution of observed partial correlations assigned to actually existing edges that we aim to identify. The null distribution f_0 is given by Eq. 5.2. For reasons of simplicity we may assume for the distribution of partial correlation coefficients of the true edges f_A, e.g., a simple uniform distribution from -1 to 1. However, for f_A other more complicated distributions could easily be conceived, including nonparametric estimates (Efron, 2005b).

Fitting this mixture density to the observed partial correlation coefficients (via optimizing the corresponding likelihood or an EM-type algorithm) allows to estimate the parameters η_0 and κ. In doing so one carries out the type of empirical Bayes analysis proposed by Robbins (1956) and Efron (2003). It is then straightforward to compute two-sided p-values for each possible edge in the network using the exact null distribution f_0 with $\hat{\kappa}$ as plug-in estimate. Alternatively, one may also be interested in the edge-specific "local false discovery rate" (fdr)

$$\text{Prob}(\text{null edge}|\tilde{r}) = \text{fdr}(\tilde{r}) = \frac{\eta_0 f_0(\tilde{r}; \kappa)}{f(\tilde{r})}, \tag{5.4}$$

i.e. the Bayes posterior probability of an edge being absent given $\tilde{r}$. An edge may be considered significant if its local fdr is smaller than 0.2 (Efron, 2005b). Closely related to the empirical Bayes local fdr statistic is the commonplace tail area false discovery rate (FDR) approach to multiple testing advocated by Storey (2002), also called q-value approach, and the seminal Benjamini and Hochberg (1995) FDR rule. False discovery rate methods are discussed in more detail in the next section. In practice it seems to make little difference which approach is used. However, the local fdr statistic fits more naturally with the mixture modeling setup. Moreover, it provides a measure of belief in the significance of an individual edge. An exemplifying estimation technique is discussed in Subsection 5.2.2 that takes account of the dependencies among the estimated partial correlation coefficients (Efron, 2005a).

The inference approach, though new for edge detection in graphical models, is directly inspired by similar approaches to detect differentially expressed genes (Sapir and Churchill, 2000; Efron et al., 2001; Efron, 2003). There, the mixture model represents differentially and not differentially expressed genes presupposing that the majority of investigated genes belong to the latter class.

A key element of this procedure is that it turns a seemingly disadvantage in the analysis, namely the large number of genes p in a microarray data set, into an advantage: with growing p the number of zero-edges $\eta_0 m$ becomes larger, and hence it gets easier to estimate the null distribution from the data. Note that this "Robbins-Efron-type" inference (see Efron, 2003) enables one to determine the sampling distribution f_0 from a large-dimensional point estimate. A further benefit of using an *empirical null* distribution in a large-scale testing situation is that it additionally accounts for hidden correlations and the effects of unobserved covariates (Efron, 2004, 2005a).

Finally, it should be noted that using the estimated degree of freedom $\hat{\kappa}$, an effective sample size $n_{\text{eff}} = \hat{\kappa} + p - 1$ can be determined. This reflects the relationship between sample size and κ for the standard normal partial correlation coefficient, but also extends to the case when other estimators such as $\hat{\tilde{\boldsymbol{P}}}^1$, $\hat{\tilde{\boldsymbol{P}}}^2$, $\hat{\tilde{\boldsymbol{P}}}^3$, and $\hat{\tilde{\boldsymbol{P}}}^4$ are employed.

5.2. Large-Scale GGM Selection Using Multiple Testing – Type I Error Rate Concepts

One simple strategy for choosing a GGM network consistent with the data is to test each of the $m = p(p-1)/2$ potential edges individually for presence in the final network, i.e. to determine whether the corresponding partial correlation coefficients differ significantly from zero (Whittaker, 1990; Drton and Perlman, 2004). GGM search by multiple testing implicitly assumes that for all cliques, i.e. fully connected subsets of nodes, of size three and more the underlying joint distribution is well approximated by the product of the bivariate marginal densities associated with the respective undirected edges (Cox and Reid, 2004).

Let $I \subseteq \{1, \ldots, m\}$ denote the set of indices of the true null hypotheses in the sense that its members describe zero edges. Note that the cardinaliy of I equals $\eta_0 m$. In order to address the various test problems of zero partial correlation, one may proceed as follows: firstly, a list of p-values $p_1, p_2, \ldots, p_m$ is calculated, where $p_i \sim \text{Un}[0, 1]$ if $i \in I$. Secondly, because of the large-scale parallel testing situation adjustment for multiplicity needs to be employed – see Shaffer (1995), Pigeot (2000), or Dudoit et al. (2003) for a review of different approaches to multiple hypothesis testing. Dudoit et al. (2003) place particular emphasis on the context of functional genomics approaches. For a given rejection region $[0, \gamma]$, let $V = V(\gamma)$ denote the number of false positives, i.e. the number of p-values p_i below γ with $i \in I$,

$$V(\gamma) = \sum_{i \in I} 1\{p_i \leq \gamma\}. \tag{5.5}$$

In terms of this random variable, the *per family error rate* (PFER) is defined as $E(V)$ and the *per comparison error rate* (PCER) as alternative criterion $E(V)/m$. Classical multiple testing procedures control the risk of committing a type I error within the tested family of hypotheses (e.g., Holm, 1979; Simes, 1986; Westfall and Young, 1993). This family-wise error rate (FWER) is defined as $\text{Prob}(V \geq 1)$ and is usually required in the strong sense, i.e. under all configurations of true and false hypotheses tested. It is well known that procedures controlling the family-wise error rate tend to have substantially less power than procedures that do not correct for multiplicity if the number of tested hypotheses is large. In the inference of large-scale GGM networks from small-sample genomic data, lack of multiplicity control would be by far too permissive. Contrariwise, full protection resulting from FWER control is too restrictive. This is valid in many instances whenever it is primarily a selection effect that is of concern. Benjamini and Hochberg (1995) introduce a less stringent criterion, the so-called *false discovery rate* (FDR), discussed in the following.

5.2.1. The False Discovery Rate

FDR multiple testing (Benjamini and Hochberg, 1995) has emerged as a strategy that is particularly useful for addressing large-scale simultaneous inference problems that are epitomized by functional genomics approaches. Very similar high-dimensional issues arise in functional neuroanatomy. Specifically, magnetic resonance imaging (MRI) and diffusion tensor imaging (DTI) produce maps of the inside of the human body. Brain anatomy is one of the most interesting areas of study. Regarding the problem of finding regions that differ between two groups of subjects, the FDR concept proves helpful (Schwartzman et al., 2005). The FDR is defined as the expected ratio of erroneous rejections to the total number of rejected hypotheses, $E(Q)$, where

$$Q = \begin{cases} V/R & \text{if } R > 0 \\ 0 & \text{if } R = 0. \end{cases} \tag{5.6}$$

$R = R(\gamma)$ denotes the number of hypotheses with p-values in a given rejection region $[0, \gamma]$, $R(\gamma) = \sum_{i \in \{1,\ldots,m\}} 1\{p_i \leq \gamma\}$. It should be noted that the FDR is equivalent to the FWER in the weak sense, i.e. if all null hypotheses are true. Otherwise it holds that FDR $\leq$ FWER (Benjamini and Hochberg, 1995). As a result, multiple comparison procedures controlling the FDR may be expected to be more powerful than the commonly used multiple comparison procedures based on the FWER. This makes it ideal for screening purposes (Storey and Tibshirani, 2003). The basic algorithm is as follows:

1. Construct the set of ordered p-values $p_{(1)}, p_{(2)}, \ldots, p_{(m)}$ with corresponding edges $e_{(1)}, e_{(2)}, \ldots, e_{(m)}$.
2. Let i_q be the largest i for which $p_{(i)} \leq iq/m$.
3. Reject the null hypothesis of zero partial correlation for edges $e_{(1)}, e_{(2)}, \ldots, e_{(i_q)}$.

It can be shown that the procedure controls the FDR at level q for independent test statistics and for any configuration of false null hypotheses (Benjamini and Hochberg, 1995).

FDR control at level q of the above sequential p-value method can be understood as follows. The p-value threshold $\gamma' = iq/m$ is estimated that controls the FDR at level q

when the number of rejected hypotheses is fixed. Thus,

$$q = \frac{m\gamma'}{i},$$

i.e. the ratio of estimated type I errors to the number of rejected hypotheses. Due to $p_{i_q} \leq \gamma'$ it holds that the estimated FDR using p_{i_q} does not exceed q. The number of tests m can be replaced by an estimator of the number of the $\eta_0 m$ true null hypotheses. Adaptive methods (Benjamini and Hochberg, 2000) will give sharper control of the FDR when $\eta_0 < 1$. Storey (2002) strongly argues for the same argument: using information in the data about the number of true null hypotheses, $\eta_0 m$, to obtain a less conservative estimator of the FDR. When rejecting all null hypotheses with p-values less than γ, he proposes an estimator of the FDR given as

$$\widehat{\mathrm{FDR}}_\lambda(\gamma) = \frac{\hat{\eta}_0(\lambda) m\gamma}{R(\gamma)}, \tag{5.7}$$

where

$$\hat{\eta}_0(\lambda) = \frac{m - R(\lambda)}{(1-\lambda)m} \tag{5.8}$$

with fine tuning parameter $\lambda \in [0, 1)$ determined using bootstrap analysis. Under an i.i.d. mixture model, the estimator from Eq. 5.7 has the property that $E[\widehat{\mathrm{FDR}}_\lambda(\gamma)] \geq \mathrm{FDR}(\gamma)$. Inclusion of $\hat{\eta}_0(\lambda)$ (Eq. 5.8), subject to the reasonable constraint that $\hat{\eta}_0 \leq 1$, is the operational difference between Eq. 5.7 and the seminal Benjamini and Hochberg (1995) approach. Note that for the most conservative choice $\hat{\eta}_0 = 1$, the two methods coincide. However, from Eq. 5.7 it becomes again obvious that otherwise a gain in power can be expected.

Moreover, in Storey (2002, 2003) it is emphasized that assuming independent tests the FDR can be written as a Bayesian posterior probability for a given significance region $[0, \gamma]$:

$$\mathrm{FDR}(\gamma) = \frac{\eta_0 \gamma}{\mathrm{Prob}(p \leq \gamma)}. \tag{5.9}$$

This comes along with the definition of the "q-value" – the FDR analogue of the p-value:

$$q(p) = \inf_{\gamma \geq p}\{\mathrm{FDR}(\gamma)\} = \inf_{\gamma \geq p}\left\{\frac{\eta_0 \gamma}{\mathrm{Prob}(p \leq \gamma)}\right\}. \tag{5.10}$$

The q-value is a multiple hypothesis testing quantity, whereas the p-value is a single hypothesis testing quantity. Moreover, it becomes evident that the FDR is justified both from a frequentist as well as from a Bayesian perspective (see also Efron et al., 2001; Efron and Tibshirani, 2002; Efron, 2003). Finally, it is noteworthy that the original independence assumption was substantially relaxed in later work (Benjamini and Yekutieli, 2001; Storey, 2003).

Different estimators of η_0 can be used in FDR controlling procedures – for recent developments see for example Meinshausen and Bühlmann (2005b) who propose an estimator under general dependence structures. In our case a suitable estimate $\hat{\eta}_0$ is available from the fit of Eq. 5.3.

5.2.2. The `locfdr` Algorithm for Estimating the Local fdr

The empirical Bayes methodology from Section 5.1 suggests a local version of the FDR (Eq. 5.4). This is an interesting extension to the above basic FDR algorithm and its variations because it accounts for localities covered by the rejection tail area approach. The close connection between the frequentist FDR rule (Benjamini and Hochberg, 1995), its "Bayesian form" (called "q-value" in Storey (2003) – Eq. 5.10), and the empirical Bayes methodology from Section 5.1, follows directly from Bayes theorem (Efron and Tibshirani, 2002). This gives theoretical justification for the intuitive interpretation that the value of the tail area false discovery rate FDR attained at a given value of the considered statistic, say $Z = z$, is the average of local false discovery rates fdr(Z) for $Z \leq z$.

Estimating the local false discovery rate fdr($\tilde{r}$) (Eq. 5.4) requires estimating the mixture density $f(\tilde{r})$ as well as $\eta_0 f_0(\tilde{r}; \kappa)$. For normalization purposes of the sampling distribution we may first apply Fisher's (1921) z-transformation

$$z = \frac{1}{2}\log\left(\frac{1+\tilde{r}}{1-\tilde{r}}\right) \tag{5.11}$$

to the estimated partial correlation coefficients $\tilde{r}$ that corresponds to the inverse hyperbolic tangent function (atanh). The histogram of resulting z-values will reveal a normal-shaped peak around zero representing the large majority of "null" edges, while the long tails chart some interesting "non-null" coefficients – the ones we intend to detect.

The `locfdr` algorithm (Efron, 2004, 2005b,a) is presented as an exemplifying approach to local fdr estimation. Software is publicly available as the R package "locfdr" from the CRAN archive (`http://cran.r-project.org`). An empirical Bayes analysis is employed of the two-class mixture model from Eq. 5.3, where $f(z)$ is estimated by fitting a smooth curve $\hat{f}(z)$ to the histogram counts of z-values using Poisson general linear model (GLM) methodology and thus transferring density estimation to the field of regression theory. By contrast, $\eta_0 f_0(z)$ is more challenging to estimate. The `locfdr` algorithm fits an *empirical null* density in order to account for inherent dependencies (Efron, 2004, 2005a). For this purpose, the algorithm exploits the sparsity of biomolecular networks in that it assumes η_0 near 1 (say $\eta_0 \geq 0.9$). Thus, using $\eta_0 = 1$ would not result in an overly conservative estimator of fdr(z). Moreover, the sparsity assumption allows to estimate the scaled null distribution $\eta_0 f_0(z)$ from the central peak in the z-values' histogram. Assuming normality for f_0 gives

$$\log f(z) = -\frac{1}{2}\left(\frac{z-\mu_0}{\sigma_0}\right)^2 + \text{constant}$$

for z near 0, so that μ_0 and σ_0 can be estimated from the observed data by fitting a quadratic polynomial to the central histogram counts $\log\hat{f}(z)$ as

$$\mu_0 = \arg\max\{f(z)\} \quad \text{and} \quad \sigma_0 = \left[-\frac{d^2}{dz^2}\log f(z)\right]_{\mu_0}^{-\frac{1}{2}}.$$

These are the crucial empirical Bayes steps that together give an estimator of fdr(z), the posterior probability of "null edge",

$$\widehat{\text{fdr}}(z) = \hat{\eta}_0 \hat{f}_0(z)/\hat{f}(z).$$

It should be noted that beyond the `locfdr` algorithm, there have recently emerged various approaches to estimating posterior probabilities in microarray experiments with special focus on identifying differentially expressed genes (e.g., Pan et al., 2003; Pounds and Morris, 2003; Pounds and Cheng, 2004; Liao et al., 2004; Scheid and Spang, 2004).

While the original frequentist FDR theorem (Benjamini and Hochberg, 1995) was proved assuming that the teststatistics are mutually independent, independence plays no essential role in the empirical Bayes approach. Moreover, it is the intention behind fitting an empirical null distribution to account for dependency structures. Thus it seems reasonable to expect accurate results under quite general conditions.

Using a multiple testing procedure for GGM selection has the advantage that it is practical and computationally efficient also for a large number of genes. Nevertheless, this is an heuristic and only an approximation to an exhaustive GGM search. However, other heuristic searches such as backward and forward selection (Whittaker, 1990) do not necessarily guarantee a better fit for large p than multiple testing (Drton and Perlman, 2004). Stochastic searches such as Bayesian Markov chain Monte Carlo sampling of GGMs may prove more effective, see Wong et al. (2003) and Dobra et al. (2004) for recent developments.

5.3. Simulation Study

In a series of extensive computer simulations the proposed small-sample GGM framework was investigated in terms of estimation accuracy, model validation, and model selection performance criteria such as power and positive predictive accuracy. Special focus is on comparing the four small-sample estimators $\hat{\hat{\boldsymbol{P}}}^1$ ("pseudoinverse"), $\hat{\hat{\boldsymbol{P}}}^2$ ("partial bagged correlation"), $\hat{\hat{\boldsymbol{P}}}^3$ ("bagged partial correlation"), and $\hat{\hat{\boldsymbol{P}}}^4$ ("shrinkage").

5.3.1. Simulation Setup

Specifically, the following algorithm is used to generate random "true" partial correlation matrices $\tilde{\boldsymbol{P}}$ that are always positive definite. It allows to control parameters of interest such as the number of features p, and the fraction of non-zero edges $\eta_A = 1 - \eta_0$.

1. Start with an empty $p \times p$ matrix.
2. Choose randomly the off-diagonal positions corresponding to the $\eta_A m$ non-zero edges, and fill in preliminary correlation values drawn from the uniform distribution between -1 and 1.
3. Compute column-wise sums of the absolute values of the matrix entries, and set the corresponding diagonal element equal to this sum plus a small constant (say 0.0001). This ensures that the resulting matrix is diagonally dominant, and thus always positive definite.
4. Standardize the matrix so that the diagonal entries all equal 1 in order to obtain the simulated "true" partial correlation matrix $\tilde{\boldsymbol{P}}$ which in turn represents the "true" GGM network.

An example of a simulated network model with $p = 100$ nodes and proportion $\eta_A = 0.02$ of non-null edges is shown in Fig. 5.1. This choice of p and η_A implies that there are 99 true edges out of 4,950 potential edges. It should be noted that even for small values of η_A the resulting "sparse" network still looks quite dense. This is because the number of available edges m grows with the square of the number of variables p. Unfortunately, further structural and distributional properties are not easily specified – see for instance

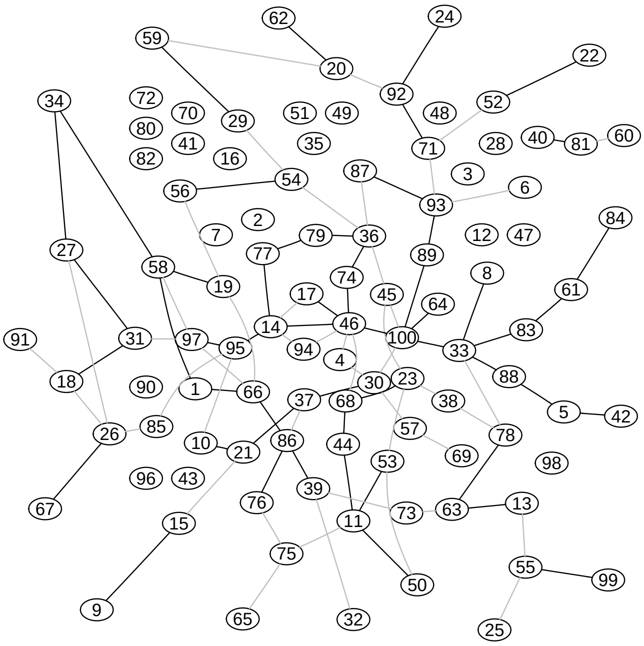


Figure 5.1.: Simulated sparse network with $p = 100$ nodes and 99 edges (corresponding to an edge fraction $\eta_A = 0.02$). Note that in this figure branch lengths are purely due to the layout of the graph and do not indicate the strength of correlation between two connected nodes. Grey lines indicate negative partial correlation, whereas edges with positive correlation are drawn in black.

Hirschberger et al. (2004). This would be desirable as the present simulation algorithm produces networks with edges that represent mostly weak links. Note that this renders their inference disproportionally hard!

Synthetic data of desired sample size n are generated as follows: from $\tilde{\boldsymbol{P}}$ the true pairwise correlation matrix $\boldsymbol{P}$ is computed via reverse application of Eq. 3.6 and Eq. 3.5. As $\tilde{\boldsymbol{P}}$ is positive definite, so is its inverse and the corresponding matrix $\boldsymbol{P}$. Subsequently, n samples are drawn from the multivariate normal distribution with zero mean vector and correlation structure $\boldsymbol{P}$.

As a measure of accuracy for the four point estimators $\hat{\tilde{\boldsymbol{P}}}^k$ $(k = 1, 2, 3, 4)$, the squared error loss $L(\hat{\tilde{\boldsymbol{P}}}^k, \tilde{\boldsymbol{P}}) = ||\hat{\tilde{\boldsymbol{P}}}^k - \tilde{\boldsymbol{P}}||_{\mathrm{F}}^2 = \sum_{i,j} (\hat{\tilde{\rho}}^k_{ij} - \tilde{\rho}_{ij})^2$ is employed. The expected loss (risk), or mean squared error (MSE), is estimated by averaging $L(\hat{\tilde{\boldsymbol{P}}}^k, \tilde{\boldsymbol{P}})$ over multiple simulation runs.

Specifically, this simulation study's setup fixes at $p = 100$, $\eta_A = 0.04$, and $n = 10, 20, \dots, 200$. A total of $R = 200$ networks, i.e. partial correlation matrices, were randomly generated per investigated sample size n and data simulated from the corresponding multivariate normal distribution. From each of the R data sets the partial correlation coefficients were estimated with the four methods "shrinkage", "pseudoinverse", $\hat{\tilde{\boldsymbol{P}}}^2$, and $\hat{\tilde{\boldsymbol{P}}}^3$. The number of bootstrap replications required for $\hat{\tilde{\boldsymbol{P}}}^2$ and $\hat{\tilde{\boldsymbol{P}}}^3$ is set to $B = 500$.

In a similar fashion, the average number of edges detected as significant, the power, and the positive predictive value (PPV), that is the number of correctly identified edges among all significant findings, were determined. The criterion for GGM selection is local fdr cut-off set to 0.2 as suggested in Efron (2005b).

5.3.2. Performance for Synthetic Data

Estimation Accuracy

In Fig. 5.2 the accuracy of the four small-sample estimators of partial correlation is contrasted. The shrinkage estimator outperforms all others regardless of sample size. The estimator $\hat{\tilde{\boldsymbol{P}}}^2$ is nearly as accurate for small sample size, however, it is much more computer expensive than the shrinkage estimator. Its good performance can be explained as follows: $\hat{\tilde{\boldsymbol{P}}}^2$ is besides the shrinkage estimator – that is guaranteed to be positive

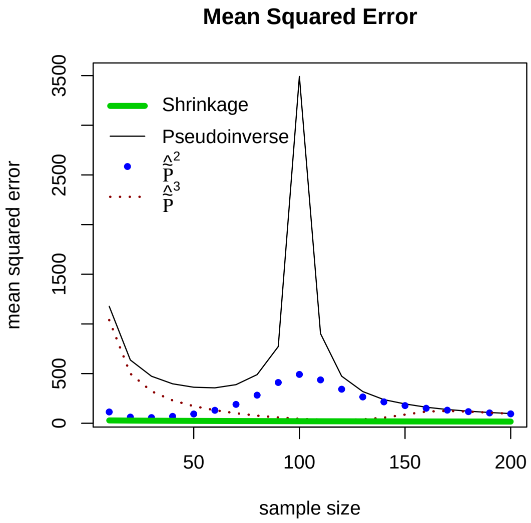


Figure 5.2.: Mean squared error of the four small-sample estimators of partial correlation: "shrinkage", "pseudoinverse", $\hat{\hat{\boldsymbol{P}}}^2$, and $\hat{\hat{\boldsymbol{P}}}^3$, in dependence of sample size n for $p = 100$ genes.

definite by construction – the only one of the investigated estimators that is based on a positive definite estimate of the correlation matrix, as averaging over bootstrap sample correlation matrices $\hat{\boldsymbol{P}}^{*b}$ acts as an implicit regularization procedure (cf. Friedman, 1989).

The peak at $n = 100$ associated with the estimator $\hat{\hat{\boldsymbol{P}}}^1$ is a dimension resonance effect (recall that $p = 100$). The mean squared error of $\hat{\hat{\boldsymbol{P}}}^1$ increases dramatically around this region, with *decreasing* error when the sample size decreases. This "peaking phenomenon" is well known in small-sample regression and classification problems and is due to the use of the pseudoinverse (Raudys and Duin, 1998; Skurichina and Duin, 2002). It can be understood as follows: for $n \approx p$ the eigenvalues of the sample correlation matrix are distorted in comparison with those of the true correlation matrix, in particular the largest and smallest eigenvalues are highly over- and underestimated, respectively (e.g. Friedman, 1989). This causes the corresponding SVD directions in the pseudoinverse to become highly overestimated. Any form of regularization of the correlation matrix (for example by bootstrap analysis) reduces this error dramatically (Skurichina and Duin, 2002). This can be seen immediately by comparing $\hat{\hat{\boldsymbol{P}}}^1$ with the

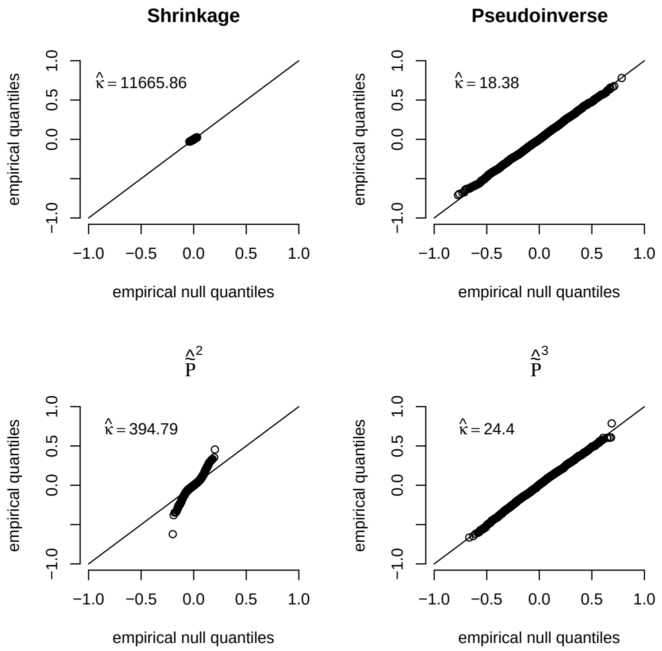


Figure 5.3.: Quantile-quantile plots of the observed null distribution of the four small-sample estimates “shrinkage”, “pseudoinverse”, $\hat{\hat{\boldsymbol{P}}}^2$, and $\hat{\hat{\boldsymbol{P}}}^3$ for $p = 100$ genes and sample size $n = 20$.

two bagged estimators, $\hat{\hat{\boldsymbol{P}}}^2$ and $\hat{\hat{\boldsymbol{P}}}^2$, and with the shrinkage estimator $\hat{\hat{\boldsymbol{P}}}^4$ that demonstrate a very good performance in the “critical n” zone with n in the order of p and exhibit a considerably lower error than $\hat{\hat{\boldsymbol{P}}}^1$.

Validation of the Empirical Null Distribution

In further studies it is verified that under the null hypothesis of zero partial correlation the proposed small-sample estimators $\hat{\hat{\boldsymbol{P}}}^1$, $\hat{\hat{\boldsymbol{P}}}^2$, $\hat{\hat{\boldsymbol{P}}}^3$, and $\hat{\hat{\boldsymbol{P}}}^4$ do indeed follow the distributional form suggested in Eq. 5.2 where $\hat{\kappa}$ is used as plug-in estimate. This model validation step is important in order to avoid systematic bias in the statistical testing of edges.

In Fig. 5.3 example quantile-quantile plots are shown comparing the observed distribution with the empirical null distribuion for small sample size ($n = 20$). The data are simulated assuming $p = 100$ genes and an empty “network” with no edges as underlying model. For $\hat{\hat{\boldsymbol{P}}}^4$ ("shrinkage"), $\hat{\hat{\boldsymbol{P}}}^1$ ("pseudoinverse"), and $\hat{\hat{\boldsymbol{P}}}^3$ clearly the observed null distributions still fit the theoretical distributional form of Eq. 5.2 well. The plot for $\hat{\hat{\boldsymbol{P}}}^2$ however indicates a stronger curtosis and broader tails of the empirical compared to the fitted empirical null distribution. Nevertheless, for the time being let us consider the fit still acceptable.

It is crucial to note that in small samples the variability of estimated partial correlation coefficients and thus the estimated degrees of freedom $\hat{\kappa}$ differ considerably among investigated estimators. Not surprisingly, for $n = 20$ and $p = 100$ the estimator $\hat{\hat{\boldsymbol{P}}}^4$ exhibits by far the smallest variance and hence largest $\hat{\kappa}$. Its successor in this context is $\hat{\hat{\boldsymbol{P}}}^2$.

Subsequently, the fit of the mixture distribution (Eq. 5.3) was also checked in the presence of true non-zero correlations. Results from a small-sample simulation with $n = 20$, $p = 100$, and $\eta_A = 0.04$ are displayed in Fig. 5.4. The quantile-quantile plots are shown of the observed distribution of partial correlation coefficients versus the fitted empirical null. We observe slightly broader tails of the empirical as compared to the empirical null distribution. This finding is as expected because in this case the empirical distribution is a mixture of the null and of the alternative distribution, from which non-zero correlations belonging to the true edges are drawn (indicated in the plots by red cross symbols). Naturally, it is assumed that non-null edges are more dispersed than nulls. However, it is worth remarking that the present simulation setting is very restrictive leading to many non-nulls that are very close to nulls. The proportion of zero edges η_0 is estimated accurately, and the estimates of the degree of freedom κ of the null distribution are similar to the corresponding estimates from Fig. 5.3.

In a similar fashion Fig. 5.5 depicts the corresponding empirical Bayes posterior probabilities of an edge being absent given $\tilde{r}$ (Eq. 5.4). The probability of an observed partial correlation to correspond to a non-existing edge is rather small for large correlation strengths and increases – more or less quickly – for smaller absolute values. Only the tails of the empirical mixture distribution contain the statistically significant edges. The width of the characteristic shape of the plotted empirical Bayes posterior probabilities is determined by the degree of freedom κ of the null distribution. It becomes evident that using an estimator with a small variance is advantageous as this allows to identify statistically significant edges even with relatively small absolute value of partial correlation.

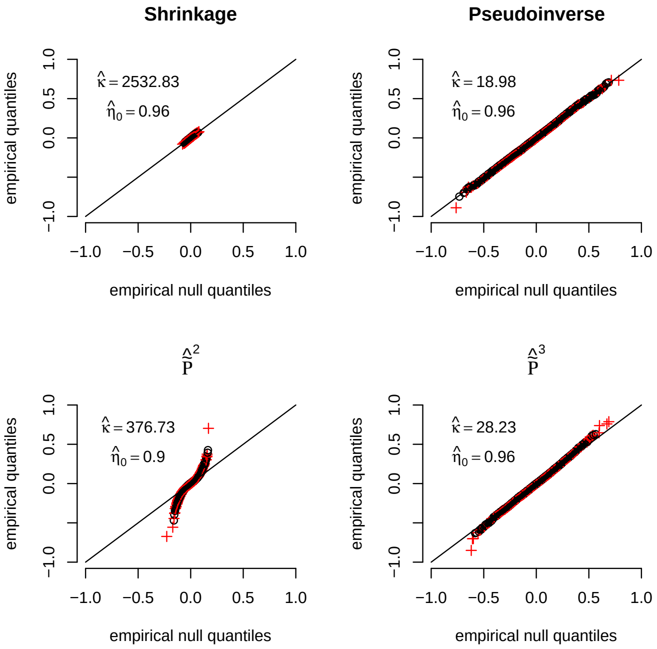


Figure 5.4.: Quantile-quantile plots of the observed mixture distribution of the four small-sample estimates "shrinkage", "pseudoinverse", $\hat{\hat{\boldsymbol{P}}}^2$, and $\hat{\hat{\boldsymbol{P}}}^3$ for $p = 100$ genes, sample size $n = 20$, and $\eta_A = 0.04$.

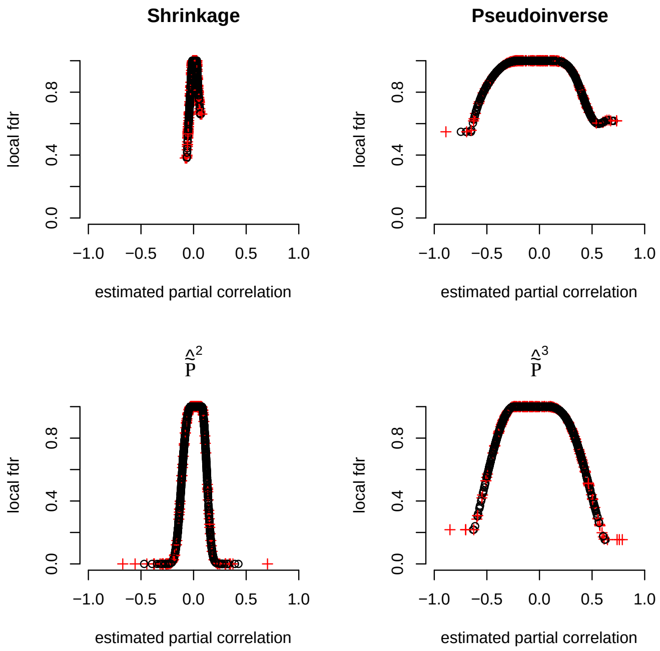


Figure 5.5.: Empirical Bayes posterior probabilities of an edge being truly absent given the corresponding entry of the estimator “shrinkage”, “pseudoinverse”, $\hat{\hat{\boldsymbol{P}}}^2$, resp. $\hat{\hat{\boldsymbol{P}}}^3$ (local false discovery rate – fdr).

Note again that it is the difficulties in random sparse correlation matrix generation that shed light on the problem of low power for small sample sizes in the present simulation studies! Put differently, if we try to report more of the non-null edges then false discovery rates grow unacceptably high, such that we would have a disproportionally high chance of pursuing artefacts, i.e. false leads.

Sensitivity and positive predictive accuracy

Finally, a large amount of computational effort was spent on simulations to investigate the statistical properties of GGM selection using local false discovery rate multiple testing. Simulations with n ranging from 10 to 200 in steps of 10, $p = 100$, and $\eta_A = 0.04$ were conducted. The GGMs were inferred by multiple testing of $m = 4,950$ edges with the desired local fdr level fixed at 0.2 (Efron, 2005b).

For each inferred network, the number of true positive features TP (correctly iden-

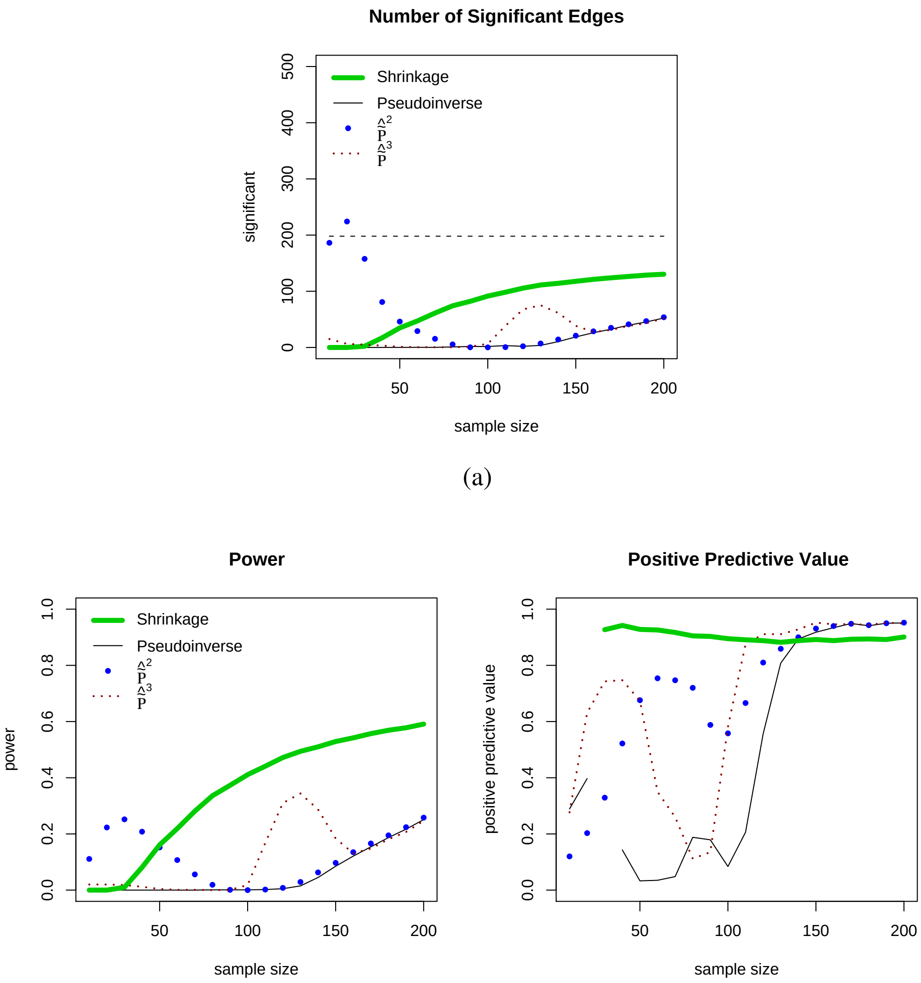
Table 5.1.: Definition of quantities used for assessing GGM network reconstruction.

Quantity	Definition
Number of true edges:	$TP + FN = \eta_A m$
Number of zero-edges:	$TN + FP = \eta_0 m$
Significant edges:	$TP + FP = S$
False positive rate:	$\mathrm{E}(\ FP/(\eta_0 m)\ ) = \alpha_I$
False negative rate:	$\mathrm{E}(\ FN/(\eta_A m)\ ) = \alpha_{II}$
True negative rate: (specificity)	$1 - \alpha_I$
True positive rate: (sensitivity, power)	$1 - \alpha_{II}$
Positive predictive value:	$\mathrm{PPV} = \mathrm{E}(TP/S)\mathrm{Prob}(S > 0)$
False discovery rate:	$\mathrm{FDR} = \mathrm{E}(FP/S)\mathrm{Prob}(S > 0)$

tified edges), false positives FP (spurious edges), true negatives TN, as well as the number of false negatives (FN) were counted. From these raw statistics, and repeated simulations of networks and data, namely $R = 200$ repetitions per investigated sample size, estimates of the false positive rate (type I error rate), power (sensitivity), and positive predictive accuracy (cf. Tab. 5.1 for the precise definitions of these quantities) were obtained for $\hat{\hat{\boldsymbol{P}}}^1$, $\hat{\hat{\boldsymbol{P}}}^2$, $\hat{\hat{\boldsymbol{P}}}^3$, and $\hat{\hat{\boldsymbol{P}}}^4$ at a given sample size n. The positive predictive value (PPV) is defined as the expected proportion of true positives among all significant findings.

Fig. 5.6a and Fig. 5.6b visualize the results with regard to GGM reconstruction. Fig. 5.6a shows the number of edges that were detected as significant using each of the four methods. For $\eta_A = 0.04$ and $p = 100$ there exist exactly 198 edges in any of the simulated networks. From $n \approx p/2$ the shrinkage estimator in comparison with $\hat{\hat{\boldsymbol{P}}}^1$, $\hat{\hat{\boldsymbol{P}}}^2$, and $\hat{\hat{\boldsymbol{P}}}^3$ typically finds the largest number of edges. The large number of significant edges for $\hat{\hat{\boldsymbol{P}}}^2$ for very small sample sizes with $n \ll p$ is a systematic bias related to the improper fit of the null model (Eq. 5.2).

Fig. 5.6b illustrates the corresponding power, i.e. the proportion of correctly identified edges, and positive predictive value (PPV). The latter quantity is of crucial practical importance as it is an estimate of the expected proportion of true edges among the list



(a)

(b)

Figure 5.6.: Performance of GGM network inference procedure: (a) Average number of edges detected as significant. Note that there are 198 true edges in the simulated network (horizontal dashed line). (b) Power and positive predictive value (PPV) for reconstructing the GGM network topology. Gaps in the curves for the PPV indicate situations in which the PPV could not be computed (no significant edges).

of edges returned as significant by the algorithm. For the shrinkage estimator the PPV – where defined – is constant across the whole range of sample sizes and close to the desired level near $1 - \mathrm{Fdr} \approx 0.9$ (Efron, 2005b). All other estimators reach the appropriate level of PPV only for $n > p$. In terms of power, the shrinkage and the $\hat{\hat{\boldsymbol{P}}}^2$-bootstrap analysis GGM approach outperform the other two investigated estimators which exhibit reasonable power only for $n > p$. However, for very small samples $\hat{\hat{\boldsymbol{P}}}^2$ liberally includes many edges in the resulting network without adequately controlling the rate of false positives among them. Thus, its PPV drops sharply: this is due to its imperfect goodness of fit with the theoretical distributional form (Eq. 5.2) under the null hypothesis. In the present simulations the shrinkage estimator has non-zero power only from $n \geq 30$ (for $p = 100$). As discussed above this should be a consequence of the simulation setup that produces partial correlation networks that are hard to infer. Put differently, all the simulations and the resulting estimates are quite conservative. This is because true GGMs are generated in such a way that they contain edges with both strong as well as many weak true correlations. The latter are notoriously difficult to detect (cf. Fig. 5.4 and Fig. 5.5) such that the test results are consequently depressed. For this reason in particular, it is crucial to appreciate the high PPV of the shrinkage estimator that indicates that if there is a significant edge then the probability is very high that it actually corresponds to a true edge.

Moreover, it should be mentioned that all four small-sample estimators exhibit the same low empirical false positive rate regardless of n (data not shown).

In summary, the obtained results particularly promote $\hat{\hat{\boldsymbol{P}}}^4$ as estimator of choice for the inference of GGM networks from small-sample gene expression data.

6. Lasso Regression for Large-Scale Covariance Selection

Alternative methodology for large-scale covariance selection is offered by penalized regression that is also a shrinkage method. Particularly promising is the lasso approach (Tibshirani, 1996) as the nature of the lasso penalty causes a kind of continuous model selection. The approach is motivated and more details are given in the next section followed by a presentation of the results of a simulation study that contrasts the empirical Bayes approach proposed in the preceding two chapters with the lasso approach to large-scale GGM selection from small-sample data.

6.1. Model Selection Using L_1 Lasso Penalized Regression

Partial correlations may not only be estimated by inversion of the covariance or correlation matrix (Eq. 3.5, Eq. 3.6). An alternative route is offered by regressing each gene's expression $X_i \in \{X_1, \ldots, X_p\}$ against the remaining set of $p-1$ variables. The estimated partial correlation coefficients are then determined as

$$\tilde{r}_{ij} = \operatorname{sign}\left(\hat{\beta}_i^{(j)}\right) \sqrt{\hat{\beta}_i^{(j)} \hat{\beta}_j^{(i)}}, \tag{6.1}$$

where $\hat{\beta}_j^{(i)}$ denotes the estimated regression coefficient of predictor variable X_j for the response X_i. Note that while in general $\hat{\beta}_i^{(j)} \neq \hat{\beta}_j^{(i)}$ the signs of these two non-zero regression coefficients are identical.

This opens the way for obtaining small-sample estimates of partial correlation and GGM inference by means of regularized regression. This avenue is pursued, e.g., by

Dobra et al. (2004) who employ Bayesian variable selection. Another possibility to determine the regression coefficients is by penalized regression, for instance ridge regression (Hoerl and Kennard, 1970a,b; Tikhonov and Arsenin, 1977) or the lasso (Tibshirani, 1996). The latter approach has the distinct advantage that it will set many of the regression coefficients (and hence also partial correlations) exactly equal to zero. Thus, for covariance selection no additional testing is required: an edge is recovered in the GGM network if both $\hat{\beta}_i^{(j)}$ and $\hat{\beta}_j^{(i)}$ differ from zero. For the standardized expression data, the lasso estimates for each gene $i \in \{1, \ldots, p\}$ are defined by

$$\begin{gathered}\hat{\boldsymbol{\beta}}^{(i)} = \arg\min_{\boldsymbol{\beta}^{(i)}} \sum_{k=1}^{n} \left(x_{ki} - \sum_{j \neq i} x_{kj} \beta_j^{(i)} \right)^2 \\ \text{subject to } \sum_{j \neq i} \left| \beta_j^{(i)} \right| \leq \lambda_i. \end{gathered} \tag{6.2}$$

GGM selection using the lasso is investigated in Meinshausen and Bühlmann (2005a) who suggest to choose the lasso penalty λ_i for regression against variable X_i according to

$$\hat{\lambda}_i = 2 \sqrt{\frac{s_{ii}^{\text{ML}}}{n}} \boldsymbol{\Phi}^{-1} \left(1 - \frac{\alpha}{2p^2} \right), \tag{6.3}$$

where $\boldsymbol{\Phi}(z)$ is the cumulative distribution function of the standard normal distribution, α is a constant (set to 0.05 in the computations below) that controls the probability of falsely connecting two distinct connectivity components (Meinshausen and Bühlmann, 2005a), and s_{ii}^{ML} is the maximum likelihood estimate of the variance of X_i. Note that this adaptive choice of penalty ensures that for small sample variance $\hat{\lambda}_i$ vanishes and hence in this case no penalization takes place. Similarly to the shrinkage approach using target D that I propose for network reconstruction (cf. Tab. 4.2), it is assumed that there is at least enough data available in order to accurately estimate the variances s_{ii}^{ML}.

6.2. Performance for Synthetic Data

In another simulation study the shrinkage and lasso approach to GGM selection were compared in terms of accuracy, power, and positive predictive accuracy.

Specifically, the simulation setup was again as follows:

1. Parameters of interest are controlled such as the number of features p, the fraction of non-zero edges $\eta_A = 1 - \eta_0$, and the sample size n of the simulated data. Specifically, parameters are fixed at $p = 100$, $\eta_A = 0.04$, and $n = 10, 20, \ldots, 200$.

2. $R = 200$ random networks were generated (i.e. partial correlation matrices) and data of size n simulated from the corresponding multivariate normal distribution.

3. From each of the R data sets the partial correlation coefficients were estimated with methods “shrinkage”, “lasso”, and $\hat{\hat{\boldsymbol{P}}}^1$. Recall that for $n > p$ the latter estimate reduces to the classical estimate of partial correlation. In this analysis the computationally inefficient bootstrap-estimators, $\hat{\hat{\boldsymbol{P}}}^2$ and $\hat{\hat{\boldsymbol{P}}}^3$, were dropped as in addition they proved to be in an inferior position compared to the shrinkage estimator in the previous simulation study.

4. Subsequently, the mean squared error was computed by comparison with the known true values.

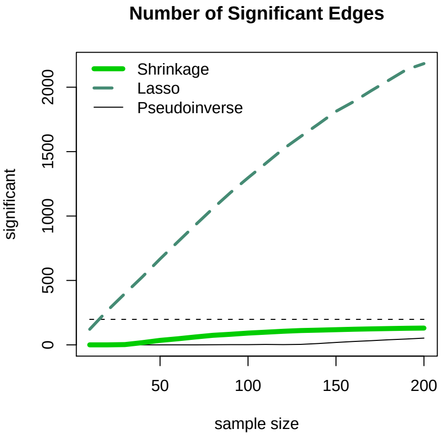
5. Similarly, the average number of edges detected as significant, the power, and the positive predictive value were calculated. Note that the latter is only reasonably defined if there is at least one significant edge. The local fdr cut-off was set to 0.2 as suggested in Efron (2005b).

In order to simulate random “true” partial correlation matrices the algorithm described in Section 5.3 producing diagonally dominant matrices was applied.

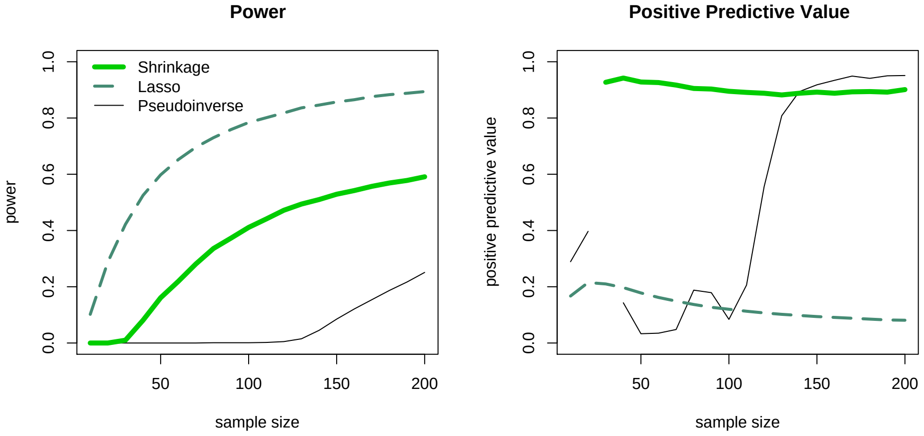
Regarding partial correlation estimation accuracy the lasso approach exhibits the same low error as the shrinkage approach (cf. Fig. 5.2). In fact the error curves depending on sample size for “shrinkage” and “lasso” completely overlap (data not shown).

Fig. 6.1a and Fig. 6.1b summarize the results with regard to GGM selection. Fig. 6.1a shows the number of edges that were detected as significant using each of the three methods. For $\eta_A = 0.04$ and $p = 100$ there exist exactly 198 edges in any of the simulated networks. The number of edges detected as significant for the shrinkage estimator remains well below this threshold, however in comparison with $\hat{\hat{\boldsymbol{P}}}^1$ it typically finds the largest number of edges. In contrast, for the simulated data the lasso GGM network approach recovers even for small sample size many more edges than are actually present. This indicates that the choice of penalization according to Eq. 6.3 may still be too permissive.

Fig. 6.1b illustrates the corresponding power (i.e. the proportion of correctly identified edges) and positive predictive value (PPV). The latter quantity is of crucial practical importance as it is an estimate of the proportion of true edges among the list of edges returned as significant by the algorithm. For the shrinkage estimator the PPV is constant across the whole range of samples sizes and close to the desired level near $1 - \text{Fdr} \approx 0.9$ (Efron, 2005b). The lasso GGM estimator exhibits a very low PPV of about 0.2 only. $\hat{\tilde{\boldsymbol{P}}}^1$ reaches the appropriate level of PPV only for $n > p$ where classical GGM theory is valid. In terms of power the shrinkage and the lasso GGM approach both outperform $\hat{\tilde{\boldsymbol{P}}}^1$ which exhibits reasonable power only for $n > p$. The power of the lasso regression approach is distinctly higher than that of the shrinkage estimator. However, this is due to the fact that the former liberally includes many edges in the resulting network without controlling the false discovery rate. The shrinkage estimator has non-zero power only from $n \geq 30$ (for $p = 100$). As discussed above this is very likely a consequence of our simulation setup which produces partial correlation networks that are hard to infer. Thus, it is crucial to note the high PPV of the shrinkage estimator indicating that if there are significant edges, then the probability is very high that these actually correspond to true edges.



(a)



(b)

Figure 6.1.: Performance of GGM network inference procedure: (a) Average number of edges detected as significant. Note that there are 198 true edges in the simulated network (horizontal dashed line). (b) Power and positive predictive value (PPV) for reconstructing the GGM network topology. Gaps in the curves for the PPV indicate situations in which the PPV could not be computed (due to zero significant edges).

7. Network Analysis of Molecular Data

For illustration the various methodologies to estimation and inference of genetic networks are now applied to molecular data. Firstly, a large-scale gene expression data set from a human breast cancer study described in West et al. (2001) is reanalyzed with respect to elucidating association structures. Secondly, the empirical Bayes methodology to large-scale sparse GGM selection, the lasso regression approach to covariance selection, and the simple relevance network approach are contrasted for a microarray experiment on the microorganism *Escherichia coli* conducted at the Institute of Applied Microbiology, University of Agricultural Sciences of Vienna (Schmidt-Heck et al., 2004).

7.1. Gene Interaction Structures in Breast Cancer

The Data – Preprocessing and Calibration

The breast cancer data set from West et al. (2001) comprises 49 tissue samples. Gene expression was measured for 7129 genes/probes using Affymetrix hu6800 chips. The corresponding CEL data were downloaded from the Duke University Center for Genome Technology (`http://data.cgt.duke.edu/West/PNASCel1.zip`). The raw data were calibrated and normalized in order to obtain robust multi-array average (RMA) expression measures (Irizarry et al., 2003). This was done using the "affy" package in Bioconductor version 1.3 (`http://www.bioconductor.org`).

Subsequently, all sequences were removed that varied only minimally or on low levels. Specifically, genes were screened out whose expression levels across all samples varied less than two-fold (corresponding to a RMA difference less than 1.0, as RMA is

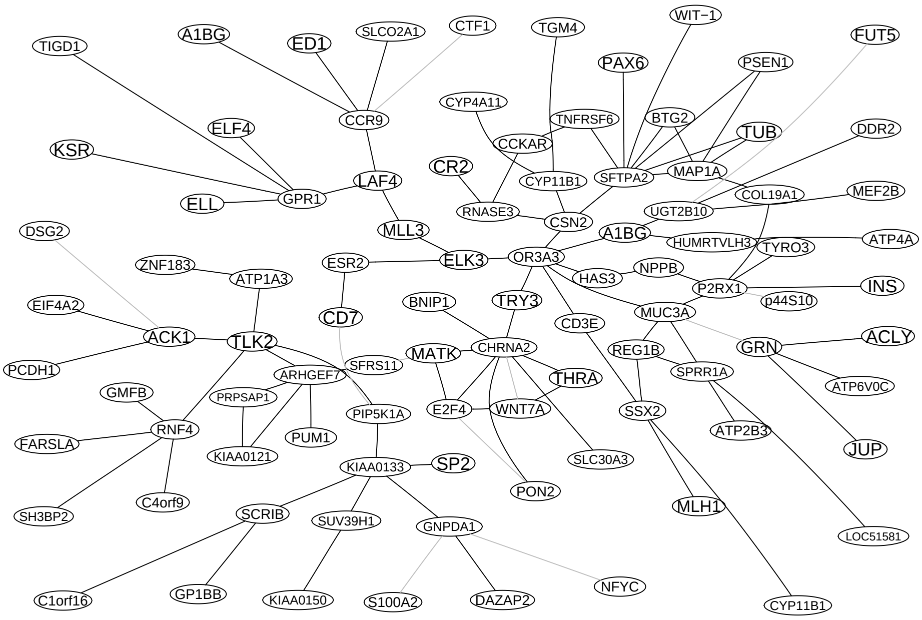


Figure 7.1.: Sub-network consisting of 96 genes centered around the ESR2 gene. This net was extracted from a global network with $p = 3,883$ genes reconstructed from the breast cancer data of West et al. (2001) using the small-sample estimator $\hat{\tilde{\boldsymbol{P}}}^2$. For a biological interpretation of selected genes neighboring ESR2 see the main text.

a measure on the log-base 2 scale) or whose maximum RMA intensity value was less than 9.0. As a result of the prescreening, gene expression data for 3,883 genes across 49 samples remained for further analysis.

Inference of Global Association Network

In order to infer the global association structure and the corresponding GGM network for all 3,883 genes, the small-sample estimator $\hat{\tilde{\boldsymbol{P}}}^{2}$ was employed with $B = 10,000$ bootstrap replications. The computation of the estimate of the partial correlation matrix –a 3,883 times 3,883 matrix with entries for 7,536,903 possible edges– required approximately 20 hours on a standard Intel Pentium 4 workstation running under the Linux operating system.

The subsequent fit of the mixture distribution (Eq. 5.3) resulted in an estimated degree of freedom $\hat{\kappa} = 4601.98$ with $\hat{\eta}_0 = 0.9924$. Using the FDR method with a desired level $q = 0.05$ 88,822 significantly non-zero coefficients were determined, corresponding to a p-value cutoff of 0.0006 and a threshold of partial correlation $\hat{\tilde{\rho}} > 0.051$.

From a statistical perspective it must be cautioned that particularly in such an extreme small-sample setting not all statistically significant edges will necessarily correspond to true edges (low PPV). To be on the conservative side, we therefore advise to take the theoretical threshold only as minimal lower bound and also to consider larger cut-off values.

CNR2 Receptor is Most-Connected Gene

Because of the large number of nodes and edges it is difficult to visualize the resulting global network structure (see below for a discussion of a sub-network). However, the degree of connectivity of each gene is more easily amenable and also highly informative.

For example, in the inferred GGM network for the investigated breast cancer data set the cannabinoid receptor 2 gene (CNR2), also known as CB2 receptor, is the best-connected gene, as it contains significant correlations with 75 (!) other genes. The "peripheral" cannabinoid receptor CNR2 is mostly expressed in the immune system, and unlike the "central" CNR1 receptor it is unrelated to cannabinoid psychoactivity.

The existence of such "super hubs" in genetic networks is well known (e.g. Barabási and Oltvai, 2004). The interesting point about CNR2 is that it seems to be directly involved in controlling tumor growth. It has been characterized as putative oncogene for acute myeloid leukemia (Jorda et al., 2003). In addition, it has been shown that targeting CNR2 can lead to induction of apoptosis in malignant lymphoblastic disease (McKallip et al., 2002). Furthermore, stimulation of CNR2 leads to a regression of skin cancer tumors (Casanova et al., 2003).

Sub-Network of the ESR2 Gene

For further illustration of the complexity of the inferred global network the genes in the immediate surroundings of the ESR2 gene (the estrogen receptor 2) are now briefly described. This gene was selected as "seed gene" for the sub-network because of its role in the pathobiology of breast cancer tumors (e.g. West et al., 2001). In Fig. 7.1 all 95 genes are shown that are correlated with ESR2 through at most five links. To reduce noise in this figure only edges with partial correlations with $\hat{\tilde{\rho}} > 0.13$ are shown. Interestingly, many close neighbors of ESR2 in this sub-network are known to be implicated in the development of malignant neuroplastic disease.

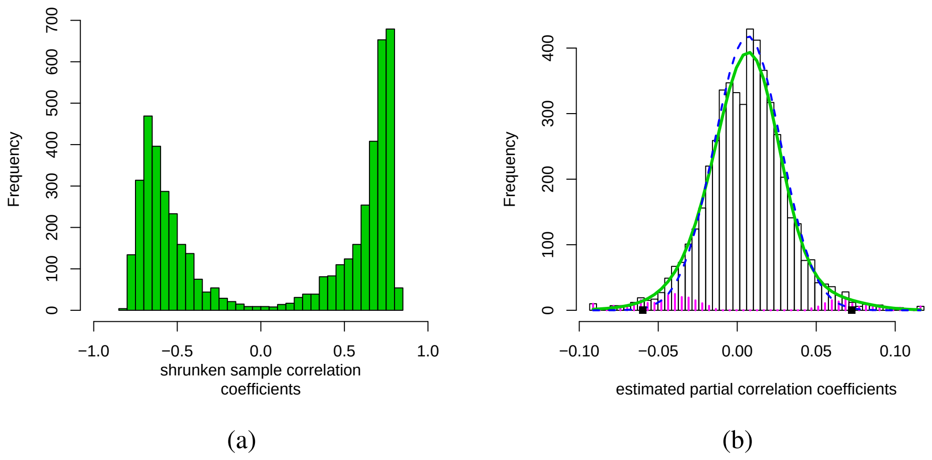
For example, ELK3 (also known as ERP, NET or SAP2) belongs to the Ets family of transcription factors. Ets proteins have been implicated in regulation of gene expression during a variety of biological processes, including growth control, transformation, and T-cell activation in many organisms. Loss of normal control is often associated with conversion to an oncoprotein (Wasylyk et al., 1993).

On the left to the ESR2 gene sits the human CD7 antigen (also known as gp40) which is a cell surface glycoprotein found on thymocytes and mature T-cells. CD7 is one of the earliest antigens to appear on cells of the T-lymphocyte lineage, and the most reliable clinical marker of T-cell acute lymphocytic leukemia (Aruffo and Seed, 1983).

The MLL3 gene, directly linked in our network with ELK3 and LADF4, is a member of the TRX/MLL gene family. It is associated with leukemia and developmental defects (Ruault et al., 2002).

Further down in the network one finds LAF4, a gene responsible for lymphocyte differentiation. Joint with MLL it is involved in lymphoblastic leukemia (von Bergh et al., 2002).

Many more genes depicted in Fig. 7.1 are related to the development of cancer (see,



(a) (b)

Figure 7.2.: (a) Histogram of the estimated shrinkage correlation coefficients computed for all $102 \times 101/2 = 5,151$ pairs of genes. (b) Distribution of the estimated shrinkage *partial* correlation coefficients (green line) after Fisher's normalizing z-transformation (atanh) was applied for normalization purposes. Also shown are the fitted null distribution (dashed blue line) and the alternative distribution (pink) as inferred by the `locfdr` algorithm (Efron, 2004, 2005b). The black squares indicate the 0.2 local fdr cut-off values for the partial correlations.

e.g., the CancerGene database at `http://caroll.vjf.cnrs.fr/cancergene/`). This justifies cautious optimism that the inferred correlation network may indeed be useful as a starting point from which to generate further medical and biochemical hypotheses.

7.2. Stress Response of *Escherichia coli*

The microarray experiment conducted at the Institute of Applied Microbiology, University of Agricultural Sciences of Vienna (Schmidt-Heck et al., 2004) was set up to measure the stress response of the microorganism *Escherichia coli* during expression of a recombinant protein. The resulting data monitor all 4,289 protein coding genes of *E. coli* 8, 15, 22, 45, 68, 90, 150, and 180 minutes after induction of the recombinant protein SOD (human superoxide dismutase). In a comparison with pooled samples before induction 102 genes were identified by Schmidt-Heck et al. (2004) as differentially expressed in one or more samples after induction. In the following we try to establish the gene network among these 102 preselected genes.

A first impression of the dependency structure can be obtained by investigating the estimated correlation coefficients. For the shrinkage approach (Tab. 4.1) $\hat{\lambda}^{\star} = 0.18$ is obtained. The resulting correlation matrix has full rank (102) with condition number equal to 386.6. In contrast, the standard correlation matrix has rank 8 only and is ill-conditioned (infinite condition number). Thus, already for calculating the correlation coefficients the benefits of using the shrinkage estimator are quite evident.

Fig. 7.2a shows the distribution of the estimated correlations across the $5,151$ pairs of genes. As can be seen most estimated correlations *differ* from zero. This is a simple consequence of that, marginally, all genes are either directly or indirectly associated with each other. Thus, constructing a traditional relevance network (Butte et al., 2000) will –at least for this data– *not* lead to uncovering of the dependency structure. This is compared with the corresponding *partial* correlation matrix. Fig. 7.2b shows the distribution of the Fisher-transformed coefficients (cf. Hotelling, 1953). The contrast with the previous figure is apparent, as the distribution of partial correlations is unimodal and centered around zero. This means that most partial correlations vanish, that the number of direct interactions is small, and hence that the resulting gene association network is sparse.

Fig. 7.3, Fig. 7.4, and Fig. 7.5 show the corresponding gene association and relevance networks. The shrinkage GGM network is depicted in Fig. 7.3 and was derived by fitting the mixture distribution defined in Eq. 5.3 to the estimated partial correlations $\hat{\hat{\boldsymbol{P}}}^{4}$ with a cut-off fdr ≤ 0.2. The network comprises 116 significant edges which amount to about 2% of the 5,151 possible edges for 102 genes. This shows that for real data – in sharp contrast to the comparable simulations – the shrinkage estimator *is* powerful for small sample size.

Several aspects of the inferred network are worth remarking. Firstly, the "hub" connectivity structure for the gene sucA is recovered. Note that this gene is involved in the citric acid cycle. The existence of these hubs is a well-known property of biomolecular networks (e.g. Barabási and Oltvai, 2004). It is a strength of the present method that these nodes can be identified without any specific additional modeling. Secondly, the edges connecting the genes lacA, lacZ, and lacY are the strongest in the network, with the largest absolute values of partial correlation, and correspondingly also with the smallest local fdr values. Interestingly, these are exactly the genes on which the experiment was based: lacA, lacY, and lacZ are induced by IPTG (isopropyl-beta-

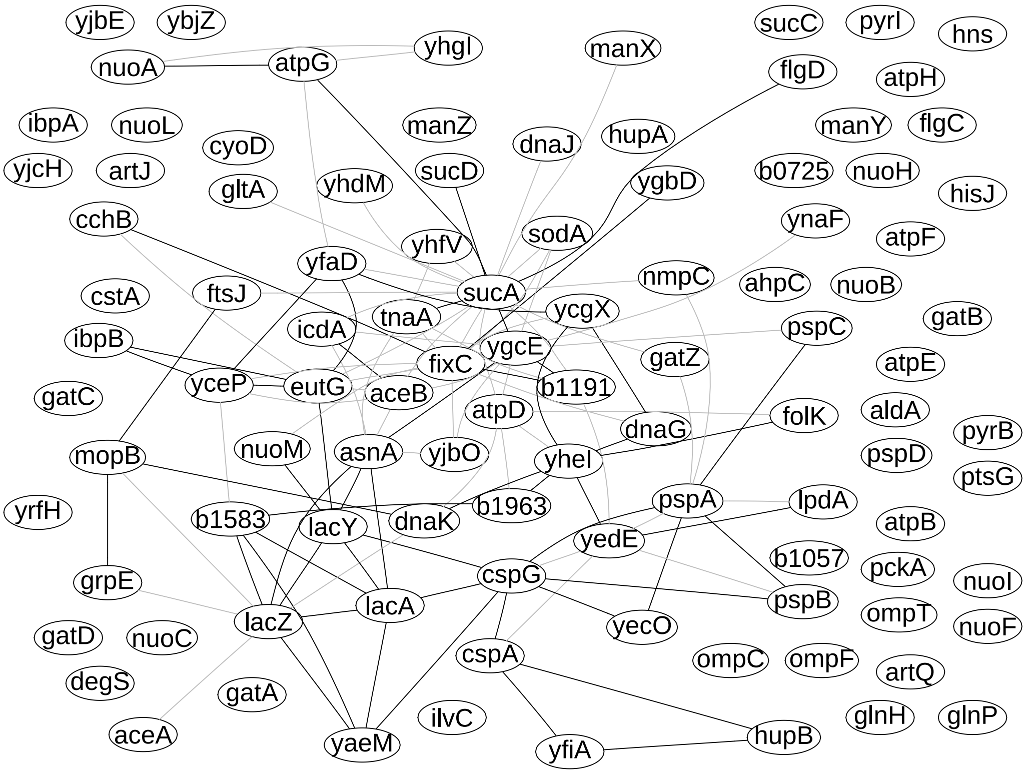


Figure 7.3.: Gene network inferred from the *E. coli* data by the shrinkage (Tab. 4.1) GGM approach. Black and grey edges indicate positive and negative partial correlation, respectively.

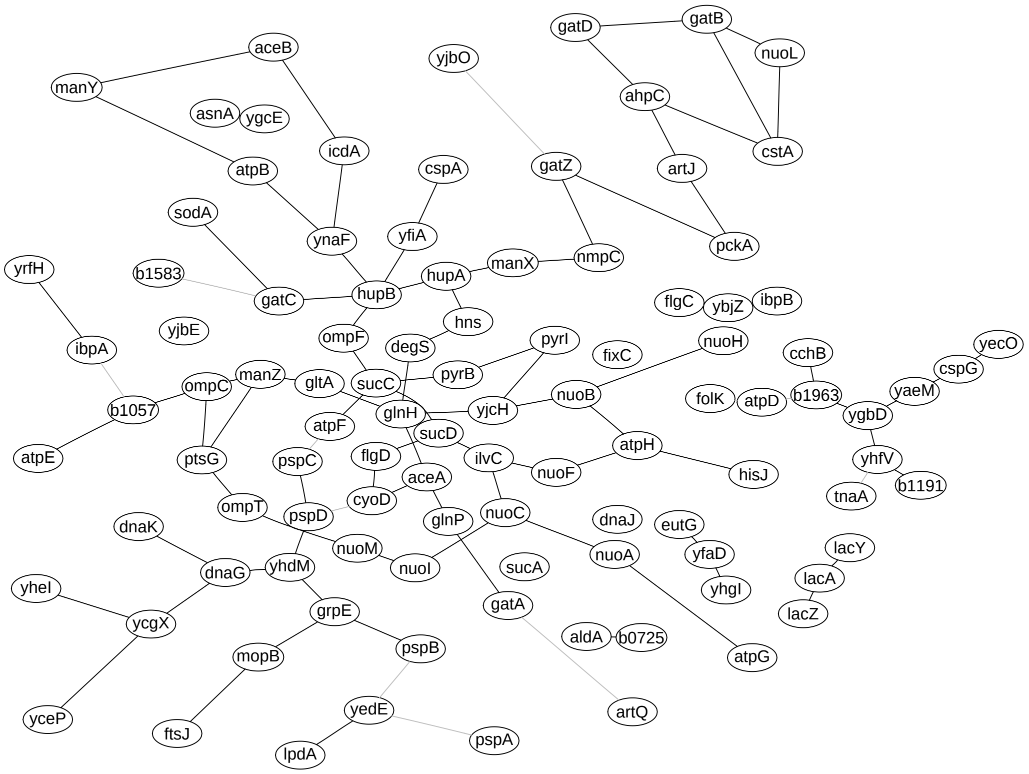


Figure 7.4.: Gene network inferred from the *E. coli* data by the lasso GGM approach by Meinshausen and Bühlmann (2005a). Black and grey edges indicate positive and negative partial correlation, respectively.

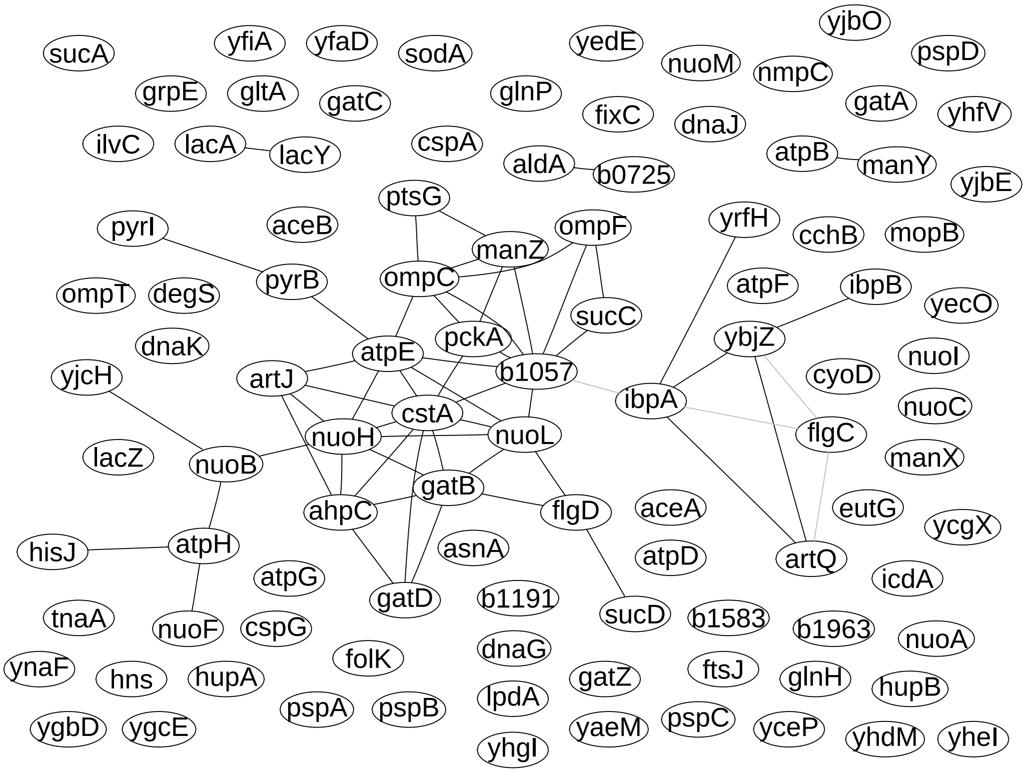


Figure 7.5.: Gene network inferred from the *E. coli* data by the relevance network approach with abs(r) > 0.8. Black and grey edges indicate positive and negative correlation, respectively.

D-thiogalactopyranoside) dosage and initiate recombinant protein synthesis (Schmidt-Heck et al., 2004).

For comparison, the lasso GGM network is shown in Fig. 7.4. It was computed from the standardized *E. coli* data by the approach by Meinshausen and Bühlmann (2005a) and contains 100 edges. Closer inspection of this network reveals an interesting structural bias introduced by the lasso regression for GGM inference. As can clearly be seen in Fig. 7.4 the lasso limits the number of edges going in and out of a node. The reason for this is that the lasso imposes sparsity on the regression coefficients *per node* so that in each regression only a few non-zero coefficients exist. As a consequence, the degree distribution of the *E. coli* lasso GGM network has an implicit upper bound. Thus, the lasso prevents the identification of hubs and also excludes power-law-type connectivity patterns. Note that in contrast in the empirical Bayes GGM approach sparsity is imposed on the network level rather than locally at node level.

Finally, Fig. 7.5 shows the relevance network obtained by applying the conventional 0.8 cut-off on the absolute values of the shrunken correlation coefficients. The resulting network contains 58 edges and bears no resemblance to the GGM networks. As is clear from inspecting Fig. 7.2a, there are many more genes that are strongly correlated, so from this network the direct dependencies among genes cannot be deduced. Instead, correlations should rather be employed for detecting *independence* among genes. The corresponding null hypothesis is that the two genes are dependent. For this purpose the mixture model of Eq. 5.3 is still applicable, except that the roles of f_0 and f_A are interchanged. Thus any edge with fdr > 0.8 (defined as in Eq. 5.4!) would be considered significant.

In the analysis it was plainly ignored that the *E. coli* data derive from a time series experiment. This appears not to be too harmful for the GGM selection process – at least part of the longitudinal correlation will be accounted for by empirically fitting the null distribution (see also Efron (2005a)).

8. Summary and Outlook

Among methods for inferring networked gene interaction structures graphical Gaussian models are becoming increasingly popular (Kishino and Waddell, 2000; Waddell and Kishino, 2000; Bay et al., 2002; Wang et al., 2003; Toh and Horimoto, 2002a,b; Wu et al., 2003; de la Fuente et al., 2004; Wille et al., 2004; Wille and Bühlmann, 2005; Magwene and Kim, 2004; Dobra et al., 2004). Their advantage over simple correlation networks, namely the ability to distinguish direct from mediated interactions, is apparent. However, their application to genome data is hampered by the "small n large p" problem which renders both estimation and inference difficult.

In this thesis a conceptually simple yet versatile and computationally fast framework was introduced for estimating and inferring large graphical Gaussian models from small-sample data. The specific bioinformatical application, that special focus is on, is the problem of inferring genetic networks from today's high-throughput genomic data. These typically contain only relatively few sample points compared to the number of investigated features. This will continue to be an important issue also in the future: sample size is primarily restricted by the availability of tissue samples, and is not necessarily increased by improved technology.

In the literature three main strategies have emerged to circumvent these dimensionality problems: dimension reduction prior to the analysis, computation of low order partial correlation coefficients, and regularized variants of graphical Gaussian modeling. In my understanding the latter approach is most promising.

The framework that is proposed in this thesis relies on three key components and novel aspects:

- Firstly, it is recognized that small-sample inference requires explicit regularization. Several novel estimators of covariance and (partial) correlation have been proposed and attention has been drawn to the problem of the widespread and largely uncritical use of standard covariance and (partial) correlation estimators in the analysis of functional genomics data. As a quick glance in any recent issue of a journal such as *Bioinformatics* or *BMC Bioinformatics* will reveal, standard correlation and covariance estimators are often rather blindly applied to large-scale problems with many variables and few sample points. For instance, consider the clustering of genes using data from a microarray experiment (e.g. Eisen et al., 1998). In order to construct a hierarchical tree describing the functional grouping of genes an estimate of the similarities between all pairs of expression profiles is needed. It is typically based on a distance measure related to the sample correlation. Thus, if p genes are analyzed (with p perhaps in the order of 1,000 to 10,000), a covariance matrix of size $p \times p$ has to be calculated. Furthermore, the covariance matrix evidently plays an important role in the classification of gene expression profiles. However, it is well known that for large-scale problems the conventional covariance and correlation estimators are not appropriate and may perform extremely poorly. For this reason, it is highly advisable to refrain from using the empirical covariance in the analysis of high-dimensional data such as from microarray or proteomics experiments.

Alternatives are readily available in the form of shrinkage estimators (e.g. Greenland, 2000). Shrinkage formalizes the idea of “borrowing strength across variables” and has proven beneficial in the problem of differential expression (e.g., Smyth, 2004; Cui et al., 2005) and of classification of transcriptome data (e.g., Tibshirani et al., 2002; Zhu and Hastie, 2004).

The shrinkage approach of Ledoit and Wolf (2003) has particularly been highlighted that allows fitting of all necessary tuning parameters in a simple *analytical* fashion. While this method appears to be little known, we anticipate that it will be helpful in many “small n, large p” inference problems.

A novel shrinkage estimator for the covariance and correlation matrix (Tab. 4.1) with guaranteed minimum MSE and positive definiteness has been introduced that is not only perfectly applicable to “small n, large p” data but can also be computed in time comparable to that of the conventional estimator. The theorem of Ledoit and Wolf (2003) to estimate the optimal shrinkage intensity demands only modest assumptions with regard to the existence of higher moments of both the unrestricted estimate and the selected shrinkage target. Consequently, computationally expensive procedures such as cross-validation are completely avoided. It should be straightforward to apply this novel shrinkage covariance estimator in different applications. For example, consider the SCRDA ("shrunken centroids regularized discriminant analysis") approach (Guo et al., 2004) that employs similar regularized covariance and correlation matrices.

- Secondly, an empirical Bayes approach has been presented to detect statistically significant edges. This allows to empirically fit from the high-dimensional point estimate of partial correlation the null distribution needed for statistical testing in its exact theoretical distributional form and to compute empirical Bayes posterior probabilities, respectively. Notice that the approach exploits the known sparse connectivity in biomolecular networks. In expression analysis similar approaches are already successfully applied in order to detect differentially expressed genes (e.g., Efron et al., 2001; Efron, 2004).

- Thirdly, an heuristic has been proposed to perform approximate model (network) selection using false discovery rate multiple testing. The frequentist FDR rule (Benjamini and Hochberg, 1995) and its variations (e.g., Benjamini and Hochberg, 2000; Storey, 2002; Storey and Tibshirani, 2003) have a Bayesian interpretation that closely connects (Efron and Tibshirani, 2002) to the above empirical Bayes framework: empirical Bayes posterior probabilities of "null" edges can be seen and interpreted as local false discovery rates.

The approach may be regarded as an extension of earlier work by Waddell and Kishino (2000), Toh and Horimoto (2002a,b), Bay et al. (2002), and Wu et al. (2003). Furthermore, extensive computer simulations have been conducted to investigate the statistical properties of the novel estimators and the performance of the proposed GGM network inference procedure. This type of power analysis should be done also for other network inference approaches where studies of this kind appear to be notably absent as pointed out before by Husmeier (2003). In the simulation studies it has been shown that using regularized estimators leads to large overall gains in prediction accuracy and in the power to recover the true network structure. This is in particular valid for the novel shrinkage estimator. Moreover, the algorithm outperforms the lasso approach to regularized GGM inference in terms of positive predictive accuracy. While the lasso approach appealingly applies shrinkage directly to the estimated partial correlations, it introduces on the other hand some interesting structural bias: sparsity is assumed on the individual gene connectivity rather than on the network level. Furthermore, GGM network inference using the Ledoit-Wolf-type shrinkage covariance estimator combined with heuristic model selection using false discovery rate multiple testing takes only a few minutes even on a slow computer – thus it is offered as fast alternative to related MCMC procedures, e.g., those by Dobra et al. (2004).

Hence, large-scale modeling and inference of graphical models, specifically GGMs, turn out possible – even for small samples. However, the assumption of linear relationships as measured by partial correlations is limiting. Non-linear interactions as well as combinatorial effects will most likely better characterize biomolecular networks. Owing to the sparsity of genomic data it is yet prudent to choose simple models that require few parameters and to act on the assumption of approximate validity. This is corroborated by several examples of successful application of graphical modeling to gene expression data (e.g., Wille et al., 2004; Magwene and Kim, 2004; Dobra et al., 2004). Although resulting GGM networks are not models of mechanistic interaction, but rather remain on a phenomenological level – similar to clustering techniques, cautious optimism is indicated that they may prove helpful in the context of gene network reconstruction and also as starting point for more complex models such as dynamic Bayesian network models.

Challenges and Outlook

All approaches have their limitations and interpretation of the results needs to be done in the light of the respective model assumptions. The proposed small-sample GGM approach to modeling and inferring networked gene associations contains a number of implicit assumptions that need to be critically assessed.

GGMs are based on the assumption of multivariate normality. Generally, this appears to be unproblematic given that calibration and normalization procedures are routinely used to preprocess gene expression measurements.

More critical is the assumption of linear relationships among the investigated variables. While this may be a good approximation in many cases, GGMs have nonetheless limited representational power if nonlinear or combinatorial effects are present. There are approaches that allow to test for deviations from linear models (Cox and Wermuth, 1994) but for small samples this may turn out to be very difficult. Note that most other statistical methods for genetic network analysis also fall into this class (e.g., D'haeseleer et al., 2000; Bay et al., 2002; De Hoon et al., 2003; Wu et al., 2003; Rangel et al., 2004; de la Fuente et al., 2004). Nevertheless, the important issue of regularization in the presence of small samples has only been discussed in a handful of papers (van Someren et al., 2001; Yeung et al., 2002; Liao et al., 2003; Dobra et al., 2004; Meinshausen and Bühlmann, 2005a).

There may be (linear) higher-order interactions among more than two variables. GGMs in general model higher-order dependencies via the notion of cliques (i.e. fully connected groups of nodes). However, the heuristic model search using multiple testing of partial correlations is based on evaluating pairwise interaction only. However, cliques can still occur in the inferred network structure, hence the approach will at least approximately detect higher-order effects.

In theory, Bayesian networks are superior to GGMs as the former allow to model non-linear relationships. If a lot of data are available, this is certainly true. In practice however, owing to the paucity of the data at hand, it is not generally possible to infer these non-linearities nor the global network structure (Husmeier, 2003; Friedman and Koller, 2003). Furthermore, the often exercised discretization causes information loss and might considerably influence the obtained results. Moreover, often Bayesian networks are in fact also linearized, which for time series data turns them into linear state-space models (Murphy, 2002). In order to analyze gene dependencies based on sparse data, it appears prudent to choose a graphical model (such as a GGM) that requires very few assumptions and only a minimal number of parameters. Note that GGMs are not endorsed as the "true model" for genetic networks.

There are many directions that can be considered for further research. Against the background of the present work three points appear particularly important.

- The small-sample approach to modeling and inferring gene networks needs to be properly adopted to time series data. Nevertheless, part of the longitudinal correlation *across* microarrays will be accounted for by the empirical fit of the null distribution, while empirical Bayes analysis does not require independence *within* a microarray (Efron, 2004, 2005a). However, explicit dynamic and temporal elements in the model will be crucial for inferring directed relationships.

The concept of graphical Gaussian modeling has been generalized to multivariate stationary processes in time (Brillinger, 1996; Dahlhaus, 2000). The corresponding models are termed *partial correlation graphs*. Consider a multivariate series $\boldsymbol{X}(t) = (X_i(t),\ i \in V)$ with components indexed by $V = \{1, \ldots, p\}$ and discrete time parameter $t = 0, \pm 1, \ldots$. In order to define partial correlation between two component series X_i and X_j, $i, j \in V$, subprocesses are considered for which the linear effects of the remaining component series have been removed. The *residual component series* $\epsilon_{i|V\setminus\{i,j\}}(t)$ is given as

$$\epsilon_{i|V\setminus\{i,j\}}(t) = X_i(t) - \mu_i^\star - \sum_{h=-\infty}^{\infty} \sum_{k \in V\setminus\{i,j\}} \phi_i^\star(t-h) X_k(h)$$

with time lags $h = 0, \pm 1, \ldots$, and where $\mu_i^\star$, $\phi_i^\star(h)$ are the values minimizing

$$E\left(X_i(t) - \mu_i - \sum_{h=-\infty}^{\infty} \sum_{k \in V\setminus\{i,j\}} \phi_i(t-h) X_k(h) \right)^2 .$$

X_i and X_j, $i, j \in V$, are *partially uncorrelated* given the remaining components $\boldsymbol{X}_{V\setminus\{i,j\}}$ if the residual component series $\epsilon_{i|V\setminus\{i,j\}}(t)$ and $\epsilon_{j|V\setminus\{i,j\}}(t+h)$ are uncorrelated at all time lags $h = 0, \pm 1, \ldots$.

For nonsingular spectral matrix $\boldsymbol{f}(\lambda)$ of the multivariate process $\boldsymbol{X}(t)$, the minimizing solutions $\mu_i^\star$ and $\phi_i^\star(h)$ are unique (Brillinger, 1981, Theorem 8.3.1). The entries of $\boldsymbol{f}(\lambda)$ are the (complex-valued) *cross-spectra* of the component series X_i and X_j, defined as the Fourier transform of their covariance function,

$$f_{ij}(\lambda) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} \exp\{-i\lambda h\} \operatorname{cov}\left\{X_i(t+h), X_j(t)\right\}, \ -\infty < \lambda < \infty.$$

In Dahlhaus (2000) it is shown that *partial spectral coherencies* $R_{ij|V\setminus\{i,j\}}(\lambda)$, that define an equivalent measure of partial correlation as a function of frequency λ, can be obtained as the negative values of the rescaled inverse of the spectral matrix

$\boldsymbol{f}(\lambda)$, i.e.

$$R_{ij|V\setminus\{i,j\}}(\lambda) = -d_{ij}(\lambda)$$

with

$$\boldsymbol{d}(\lambda) = \begin{pmatrix} g_{11}(\lambda)^{-1/2} & & 0 \\ & \ddots & \\ 0 & & g_{pp}(\lambda)^{-1/2} \end{pmatrix} \boldsymbol{g}(\lambda) \begin{pmatrix} g_{11}(\lambda)^{-1/2} & & 0 \\ & \ddots & \\ 0 & & g_{pp}(\lambda)^{-1/2} \end{pmatrix}$$

and

$$\boldsymbol{g}(\lambda) = \boldsymbol{f}(\lambda)^{-1}.$$

Equivalently to the above definition, X_i and X_j, $i, j \in V$, are *partially uncorrelated* given the remaining components $\boldsymbol{X}_{V\setminus\{i,j\}}$ if their partial spectral coherency $R_{ij|V\setminus\{i,j\}}(\lambda)$ vanishes for all frequencies λ ($-\infty < \lambda < \infty$).

Assuming a Gaussian process, zero partial correlation is equivalent to conditional independence. Thus, in a *partial correlation graph* an edge between two vertices i and j is defined to be missing whenever $X_i \perp\!\!\!\perp X_j \mid \boldsymbol{X}_{V\setminus\{i,j\}}$ (*pairwise Markov property*).

Note that in order to accomplish the above inversion step, $\boldsymbol{f}(\lambda)$ is required to have full rank! Thus, in a "small n, large p" setting *very* similar issues regarding modeling and inference arise as in graphical Gaussian models for i.i.d. samples and correspondingly, similar mechanisms of regularization may prove successful to overcome them.

- More research needs to be done in the field of model selection for gene regulatory networks. In particular, the quality of search heuristics such as the one presented in this work should be compared thoroughly with solutions obtained with exact approaches (this is only possible for small examples) and with those from the proposed stochastic searches (e.g. Wong et al., 2003).

- For evaluating statistical properties and performance of modeling and inference approaches, procedures and algorithms are by all means desirable that allow for a biologically more realistic simulation of random correlation structures.

The bottom line is that modern functional genomics approaches require answers to both large-scale modeling and inference based on small samples. This “small n, large p” issue will continue to be important also in the future: novel molecular biology devices, such as protein assays, even outnumber microarrays regarding the high-dimensionality of collected data. Careful statistical reasoning is the only way to see through the haze of randomness to the structure underneath (cited from Efron, 2005c). In this context areas like shrinkage and empirical Bayes, that formalize the concept of “borrowing strength across variables”, constitute promising strategies to play a role in scientific progress in understanding cellular function at the system level and in elucidating the molecular basis of health and disease.

A. Available Computer Software

The approaches proposed in this thesis to regularized estimation of covariance and of (partial) correlation matrices and to inferring gene association networks using large-scale graphical models are implemented in the R packages “corpcor” and “GeneTS”, respectively. Specifically, “GeneTS” allows fast model selection of graphical Gaussian models (GGMs) via local false discovery rate multiple testing.

Both packages require a recent version of the R software (at least version 2.0.0) and are distributed under the terms of the GNU General Public License, freely available for download from the CRAN archive (`http://cran.r-project.org`) and from `http://www.statistik.lmu.de/~strimmer/software/genets/`. “GeneTS” is also available from Bioconductor (`http://www.bioconductor.org`). The current version 2.8.0 of “GeneTS” requires installation of the R packages “corpcor” and “locfdr” (also available from CRAN). These two packages must be installed, otherwise “GeneTS” does not work.

For network visualization “GeneTS” uses the “graph” and “Rgraphviz” packages, available from Bioconductor version 1.4 and above. However, note that installation of these packages is optional and not necessary for any of the computational procedures available in “GeneTS”.

All methods available in “corpcor” and “GeneTS” are described with examples in their respective help pages.

An example session for inferring gene association networks is described in the following.

```
# load GeneTS library
> library("GeneTS")
```

Note that the pre-processed normalized data need to be arranged in a matrix where each column corresponds to a gene, and where the rows correspond to the individual measurements (e.g. time points). The exemplifying data set describes the temporal expression of 102 genes of the microorganism *E. coli* measured at 9 time points (Schmidt-Heck et al., 2004, cf. Chapter 7).

```
# load data set
> data(ecoli)
# how many samples and how many genes?
> dim(ecoli)
```

```
# define number of nodes in the network
> num.nodes <- dim(ecoli)[2]
# node labels are gene names
> node.labels <- colnames(ecoli)
```

GGM network inference essentially comprises three steps. Firstly, the partial correlation matrix is estimated. Various novel options for estimating partial correlations from small-sample data sets have been presented in Chapter 4. These methods are implemented in the function `ggm.estimate.pcor`. The basic principle behind the small-sample estimators is variance reduction, either non-parametrically (via the bootstrap) or in a shrinkage approach. The advantages of using especially the latter approach in comparison with the standard empirical estimates are that the shrinkage estimates are always positive definite, well conditioned, and exhibit (sometimes dramatically) better mean squared error. Furthermore, they are efficient to compute and independent of any tuning parameters as the shrinkage intensity is analytically estimated from the data.

```
> pcor.shrinkage <- ggm.estimate.pcor(ecoli, method=“shrinkage”)
```

Other possibilites include

```
> pcor.1 <- ggm.estimate.pcor(ecoli, method = “observed.pcor”)
> pcor.2 <- ggm.estimate.pcor(ecoli, method =
“partial.bagged.cor”, R=1000)
> pcor.3 <- ggm.estimate.pcor(ecoli, method = “bagged.pcor”,
R=1000)
```

```
# choose estimator
> inferred.pcor <- pcor.shrinkage
```

Secondly, statistical significance is assigned to the edges in the GGM network by computing p-values, q-values, and posterior probabilites for each potential edge.

```
> test.results <- ggm.test.edges(inferred.pcor,
fA.type=“nonparametric”)
# show first 20 edges with corresponding statistics
> test.results[1:20,]
```

Therefore, the subroutine `cor.fit.mixture` fits a mixture model (Eq. 5.3) to the vector of empirical partial correlation coefficients using likelihood maximization (note that `sm2vec` puts the entries in the lower triangle of a symmetric matrix into a vector). This allows to estimate both the degree of freedom κ in the null distribution and the proportion η_0 of null edges. The alternative distribution is either assumed to be the uniform distribution from -1 to 1, or an arbitrary nonparametric distribution that vanishes for values near zero.

```
> c <- cor.fit.mixture(sm2vec(inferred.pcor),
fA.type=“nonparametric”)
> c$eta0
> c$kappa
```

Thirdly, against the background of false discovery rate control it is decided which edges are included in the network.

```
# how many edges are significant based on FDR cutoff q = 0.05 ?
> significant1.idx <- test.results$qval <= 0.05
> num.significant.1 <- sum(significant1.idx)
# list significant edges with corresponding statistics
> test.results[significant1.idx,]

# how many edges are significant based on local fdr cutoff 0.2 ?
> significant2.idx <- test.results$prob > 0.80
> num.significant.2 <- sum(significant2.idx)
# list significant edges with corresponding statistics
> test.results[significant2.idx,]
```

The network plotting functions require the installation of the "graph" and "Rgraphviz" R packages. These are available from the Bioconductor website. Note that it is not necessary to install the complete set of Bioconductor packages, only "graph" and "Rgraphviz" are needed by the "GeneTS" package – together with their respective dependencies.

```
# generate graph object with all significant edges
> gr <- ggm.make.graph( test.results[significant2.idx,],
num.nodes)
> gr

# print vector of edge weights
> show.edge.weights(gr)

# plot network (cf. Fig. 7.3)
> ggm.plot.graph(gr, node.labels, show.edge.labels=FALSE)
# with partial correlations as edge labels
> ggm.plot.graph(gr, node.labels, show.edge.labels=TRUE)
```

Furthermore, `ggm.simulate.pcor` allows to randomly generate a matrix of partial correlation that corresponds to a GGM network of a given size (`num.nodes`) with a specified fraction of non-zero edges. The output is always positive definite. This is ensured by using a diagonally dominant matrix when generating the random GGM model.

```
# generate random network with 20 nodes and 10 percent edges (=19 edges)
> true.pcor <- ggm.simulate.pcor(num.nodes=20, 0.1)
# convert to edge list
> test.results2 <- ggm.test.edges(true.pcor, eta0=0.9,
kappa=1000)[1:19,]
> test.results2
# plot network
> gr2 <- ggm.make.graph(test.results2, 20)
> gr2
> ggm.plot.graph(gr2)
```

`ggm.simulate.data` takes a (randomly generated) positive definite matrix of partial correlations and produces an i.i.d. sample from the corresponding standard multivariate normal distribution. This allows to re-estimate partial correlations with the various methods described above and to investiagte the respective accuracy, e.g., in terms of squared error loss.

```
# generate random network with 40 nodes and 5 percent edges
> sim.pcor <- ggm.simulate.pcor(num.nodes=40, 0.05)
# simulate data set with 40 observations
> m.sim <- ggm.simulate.data(40, sim.pcor)
# simple estimate of partial correlations using the pseudoinverse
> estimated.pcor <- ggm.estimate.pcor(m.sim, method =
c("observed.pcor"))
# comparison of estimated and true model
> sum((sim.pcor-estimated.pcor)**2)
# bootstrap variance reduction
> estimated.pcor.2 <- ggm.estimate.pcor(m.sim, method =
c("bagged.pcor"))
> sum((sim.pcor-estimated.pcor.2)**2)
# shrinkage approach
> estimated.pcor.3 <- ggm.estimate.pcor(m.sim, method =
c("shrinkage"))
> sum((sim.pcor-estimated.pcor.3)**2)
```

Bibliography

Aruffo, A. and B. Seed (1983). Molecular cloning of two CD7 (T-cell leukemia antigen) cDNAs by a COS cell expression system. *EMBO J. 6*, 3313–3316.

Barabási, A.-L. and R. Albert (1999). Emergence of scaling in random networks. *Science 286*, 509–512.

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

Bay, S. D., J. Shrager, A. Pohorille, and P. Langley (2002). Revising regulatory networks: from expression data to linear causal models. *J. Biomed. Informatics 35*, 298–297.

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

Benjamini, Y. and Y. Hochberg (2000). The adaptive control of the false discovery rate in multiple hypotheses testing. *J. Behav. Educ. Statist. 25*, 60–83.

Benjamini, Y. and Y. Yekutieli (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist. 29*, 1165–1188.

Breiman, L. (1996). Bagging predictors. *Machine Learning 24*, 123–140.

Brillinger, D. R. (1981). *Time Series: Data Analysis and Theory*. New York: McGraw Hill.

Brillinger, D. R. (1996). Remarks concerning graphical models for time series and point processes. *Revista de Econometria 16*, 1–23.

Butte, A. J., P. Tamayo, D. Slonim, T. R. Golub, and I. S. Kohane (2000). Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc. Natl. Acad. Sci. USA 97*, 12182–12186.

Casanova, M. L., C. Blázquez, J. Martínez-Palacio, C. Villanueva, M. J. Fernández-Acenero, J. W. Huffman, J. L. Jorcano, and M. Guzmán (2003). Inhibition of skin tumor growth and angiogenesis in vivo by activation of cannabinoid receptors. *J. Clin. Invest. 111*(1), 43–50.

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

Cox, D. R. and D. V. Hinkley (1974). *Theoretical Statistics*. London: Chapman and Hall.

Cox, D. R. and N. Reid (2004). A note on pseudolikelihood from marginal densities. *Biometrika 91*, 729–737.

Cox, D. R. and N. Wermuth (1994). Tests of linearity, multivariate normality and the adequacy of linear scores. *Applied Statistics 43*, 347–355.

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

Dahlhaus, R. (2000). Graphical interaction models for multivariate time series. *Metrika 51*, 157–172.

Daniels, M. J. and R. E. Kass (2001). Shrinkage estimators for covariance matrices. *Biometrics 57*, 1173–1184.

De Hoon, M. J. L., S. Imoto, K. Kobayashi, N. Ogasawara, and S. Miyano (2003). Inferring gene regulatory networks from time-ordered gene expression data of bacillus subtilis using differential equations. *Pac. Symp. Biocomput. 8*, 17–28.

de la Fuente, A., N. Bing, I. Hoeschele, and P. Mendes (2004). Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics 20*, 3565–3574.

Dempster, A. P. (1972). Covariance selection. *Biometrics 28*, 157–175.

D'haeseleer, P., S. Liang, and R. Somogyi (2000). Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics 16*, 707–726.

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

Drton, M. and M. D. Perlman (2004). Model selection for Gaussian concentration graphs. *Biometrika 91*, 591–602.

Dudoit, S., J. Shaffer, and J. Boldrick (2003). Multiple hypothesis testing in microarray experiments. *Statist. Science 18*, 71–103.

Edwards, D. (1995). *Introduction to Graphical Modelling*. New York: Springer.

Efron, B. (1975). Biased versus unbiased estimation. *Adv. Math. 16*, 259–277.

Efron, B. (1982). Maximum likelihood and decision theory. *Ann. Statist. 10*, 340–356.

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

Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *J. Amer. Statist. Assoc. 99*, 96–104.

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

Efron, B. (2005b). Local false discovery rates. Preprint, Dept. of Statistics, Stanford University.

Efron, B. (2005c). Modern science and the Bayesian-frequentist controversy. Preprint, Dept. of Statistics, Stanford University.

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

Efron, B. and C. N. Morris (1977). Stein's paradox in statistics. *Sci. Am. 236*, 119–127.

Efron, B. and R. Tibshirani (2002). Empirical Bayes methods and false discovery rates for microarrays. *Genet. Epidemiol. 23*, 70–86.

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

Eisen, M. B., P. T. Spellman, P. O. Brown, and D. Botstein (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA 95*, 14863–14868.

Fisher, R. A. (1921). On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron 1*, 1–32.

Friedman, J. H. (1989). Regularized discriminant analysis. *J. Amer. Statist. Assoc. 84*, 165–175.

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

Friedman, N. and D. Koller (2003). Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks. *Machine Learning 50*, 95–125.

Friedman, N., M. Linial, I. Nachman, and D. Pe'er (2000). Using Bayesian networks to analyze gene expression data. *J. Comp. Biol. 7*, 601–620.

Gibson, G. and S. V. Muse (2002). *A Primer of Genome Science*. Sunderland MA: Sinauer Associates.

Greenland, S. (2000). Principles of multilevel modelling. *Intl. J. Epidemiol. 29*, 158–167.

Guo, Y., T. Hastie, and T. Tibshirani (2004). Regularized discriminant analysis and its application in microarray. Preprint, Dept. of Statistics, Stanford University.

Halgren, R. G., M. R. Fielden, C. J. Fong, and T. R. Zacharewski (2001). Assessment of clone identity and sequence fidelity for 1189 IMAGE cDNA clones. *Nucleic Acids Res. 29*, 528–588.

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

Hartwell, L. H., J. J. Hopfield, S. Leibler, and A. W. Murray (1999). From molecular to modular cell biology. *Nature 402*, C47–C52.

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

Hastie, T. and T. Tibshirani (2004). Efficient quadratic regularization for expression arrays. *Biostatistics 5*, 329–340.

Higham, N. J. (1988). Computing a nearest symmetric positive semidefinite matrix. *Linear Algebra Appl. 103*, 103–118.

Hirschberger, M., Y. Qi, and R. E. Steuer (2004). Randomly generating portfolio-selection covariance matrices with specified distributional assumption. Preprint, Terry College of Business, University of Georgia.

Hoerl, A. E. and R. W. Kennard (1970a). Ridge regression: applications to nonorthogonal problems. *Technometrics 12*, 69–82.

Hoerl, A. E. and R. W. Kennard (1970b). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics 12*, 55–67.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Statist. 6*, 65–70.

Hotelling, H. (1953). New light on the correlation coefficient and its transforms. *J. R. Statist. Soc. B 15*, 193–232.

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

Huber, W., A. von Heydebreck, H. Sültmann, A. Poustka, and M. Vingron (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics 18*, S96–S104.

Husmeier, D. (2003). Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics 19*, 2271–2282.

Irizarry, R. A., B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs, and T. P. Speed (2003). Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res. 31*, e15.

Irizarry, R. A., B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics 4*, 249–264.

Jeong, H., B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási (2000). The large-scale organization of metabolic networks. *Nature 407*, 651–654.

Jorda, M. A., N. Rayman, P. Valk, E. De Wee, and R. Delwel (2003). Identification, characterization, and function of a novel oncogene: the peripheral cannabinoid receptor CB2. *Ann. N. Y. Acad. Sci. 996*, 10–16.

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

Kitano, H. (2002a). Computational systems biology. *Nature 420*, 206–210.

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

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

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

Ledoit, O. and M. Wolf (2004a). Honey, I shrunk the sample covariance matrix. *J. Portfolio Management 30*, 110–119.

Ledoit, O. and M. Wolf (2004b). A well conditioned estimator for large-dimensional covariance matrices. *J. Multiv. Anal. 88*, 365–411.

Leung, P. L. and W. Y. Chan (1998). Estimation of the scale matrix and its eigenvalues in the Wishart and the multivariate F distributions. *Ann. Inst. Statist. Math. 50*, 523–530.

Li, W.-H. and D. Graur (1991). *Fundamentals of Molecular Evolution.* Sunderland MA: Sinauer Associates.

Liao, J. C., R. Boscolo, Y.-L. Yang, L. M. Tran, C. Sabatti, and V. P. Roychowdhury (2003). Network component analysis: Reconstruction of regulatory signals in biological systems. *Proc. Natl. Acad. Sci. USA 100*, 15522–15527.

Liao, J. G., Y. Lin, Z. E. Selvanayagam, and W. Shih (2004). A mixture model for estimating the local false discovery rate in DNA microarray analysis. *Bioinformatics 20*, 2694–2701.

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

Magwene, P. M. and J. Kim (2004). Estimating genomic coexpression networks using first-order conditional independence. *Genome Biology 5*, R100.

McKallip, R., C. Lombard, M. Fisher, B. R. Martin, S. Ryu, S. Grant, P. S. Nagarkatti, and M. Nagarkatti (2002). Targeting CB2 cannabinoid receptors as a novel therapy to treat malignant lymphoblastic disease. *Blood 100*(2), 627–634.

Meinshausen, N. and P. Bühlmann (2005a). High dimensional graphs and variable selection with the lasso. *Ann. Statist.* in press.

Meinshausen, N. and P. Bühlmann (2005b). Lower bounds for the number of false null hypotheses for multiple testing of associations under general dependence structures. *Biometrika 92*, 893–907.

Morris, C. N. (1983). Parametric empirical Bayes inference: theory and applications. *J. Amer. Statist. Assoc. 78*, 47–55.

Murphy, K. P. (2002). *Dynamic Bayesian networks: Representation, Inference and Learning (PhD Thesis).* Berkeley: Unversity of California, Computer Science Division.

Neyman, J. (Ed.) (1956). *Proc. Third Berkeley Symp. Math. Statist. Probab.*, Volume 1, Berkeley. Univ. California Press.

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

Pan, W., J. Lin, and C. T. Le (2003). A mixture model approach to detecting differentially expressed genes with microarray data. *Functional & Integrative Genomics 3*, 117–124.

Penrose, R. (1955). A generalized inverse for matrices. *Proc. Cambridge Phil. Soc. 51*, 406–413.

Pigeot, I. (2000). Basic concepts of multiple tests – a survey. *Statistical Papers 41*, 3–36.

Pounds, S. and C. Cheng (2004). Improving false discovery rate estimation. *Bioinformatics 20*, 1737–1745.

Pounds, S. and S. W. Morris (2003). Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics 19*, 1236–1242.

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

Raudys, S. and R. P. W. Duin (1998). Expected classification error of the Fisher linear classifier with pseudoinverse covariance matrix. *Patt. Recogn. Lett. 19*, 385–392.

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

Robbins, H. (1956). An empirical Bayes approach to statistics. See Neyman (1956), pp. 157–163.

Ruault, M., M. E. Brun, M. Ventura, G. Roizes, and A. De Sario (2002). MLL3, a new human member of the TRX/MLL gene family, maps to 7q36, a chromosome region frequently deleted in myeloid leukemia. *Gene 284*, 73–81.

Sachidanandam, R., D. Weissman, S. C. Schmidt, J. M. Kakol, L. D. Stein, G. Marth, S. Sherry, J. C. Mullikin, B. J. Mortimore, D. L. Willey, S. E. Hunt, C. G. Cole, P. C. Coggill, C. M. Rice, Z. Ning, J. Rogers, D. R. Bentley, P.-Y. Kwok, E. R. Mardis, R. T. Yeh, B. Schultz, L. Cook, R. Davenport, M. Dante, L. Fulton, L. Hillier, R. H. Waterston, J. D. McPherson, B. Gilman, S. Schaffner, W. J. V. Etten, D. Reich, J. Higgins, M. J. Daly, B. Blumenstiel, J. Baldwin, N. Stange-Thomann, M. C. Zody, L. Linton, E. S. Lander, and D. Altshuler (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature 409*, 928–933.

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

Scheid, S. and R. Spang (2004). A stochastic downhill search algorithm for estimating the local false discovery rate. *IEEE Transactions on Computational Biology and Bioinformatics 1*, 98–108.

Schmidt-Heck, W., R. Guthke, S. Toepfer, H. Reischer, K. Duerrschmid, and K. Bayer (2004). Reverse engineering of the stress response during expression of a recombinant protein. In *Proceedings of the EUNITE symposium, 10-12 June 2004, Aachen, Germany*, pp. 407–412. Verlag Mainz.

Schwartzman, A., R. F. Dougherty, and J. E. Taylor (2005). False discovery rate analysis of brain diffusion direction maps. Preprint, Dept. of Statistics, Stanford University.

Segal, E., M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman (2003). Module networks: identifing regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics 34*, 166–176.

Shaffer, J. (1995). Multiple hypothesis testing: a review. *Ann. Rev. Psychol. 46*, 561–584.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika 73*, 751–754.

Skurichina, M. and R. P. W. Duin (2002). Bagging, boosting and the random subspace method for linear classifiers. *Patt. Analysis and Appl. 5*, 121–135.

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

Speed, T. P. and H. T. Kiiveri (1986). Gaussian Markov distributions over finte graphs. *Ann. Statist. 14*, 138–150.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate distribution. See Neyman (1956), pp. 197–206.

Storey, J. D. (2002). A direct approach to false discovery rates. *J. R. Statist. Soc. B 64*, 479–498.

Storey, J. D. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. *Ann. Statist. 31*, 2013–2035.

Storey, J. D. and R. Tibshirani (2003). Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA 100*, 9440–9445.

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

Tibshirani, R., T. Hastie, B. Narasimhan, and G. Chi (2002). Diagnosis of multiple cancer type by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci. USA 99*, 6567–6572.

Tikhonov, A. N. and V. A. Arsenin (1977). *Solutions of ill-posed problems*. Washington: Winston and Sons.

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

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

van Someren, E. P., L. F. A. Wessels, M. J. T. Reinders, and E. Backer (2001, June). Robust genetic network modeling by adding noisy data. In *Proceeding of the Workshop on Nonlinear Signal and Image Processing (NSIP01)*. IEEE-EURASIP.

Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. G. Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. D. Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R.-R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Y. Wang, A. Wang, X. Wang, J. Wang, M.-H. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. C. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y.-H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigó, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y.-H. Chiang, M. Coyne, C. Dahlke, A. D. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham,

B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, and X. Zhu (2001). The sequence of the human genome. *Science 291*, 1304–1351.

von Bergh, A. R., H. B. Beverlooand, P. Rombout, E. R. van Wering, M. H. van Weel, G. C. Beverstock, P. M. Kluin, R. M. Slater, and E. Schuuring (2002). LAF4, an AF4-related gene, is fused to MLL in infant acute lymphoblastic leukemia. *Genes Chromosomes Cancer 35*, 92–96.

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

Wang, J., O. Myklebost, and E. Hovig (2003). MGraph: graphical model for microarray data analysis. *Bioinformatics 19*, 2210–2211.

Wasylyk, B., S. L. Hahn, and A. Giovane (1993). The Ets family of transcription factors. *Eur. J. Biochem. 211*, 7–18.

West, M., C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J. A. Olson, J. R. Marks, and J. R. Nevins (2001). Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc. Natl. Acad. Sci. USA 98*, 11462–11467.

Westfall, P. H. and S. S. Young (1993). *Resampling-based multiple testing*. New York: John Wiley and Sons.

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

Wille, A. and P. Bühlmann (2005). Low-order conditional independence graphs for inferring genetic networks. Preprint, Seminar für Statistik, ETH Zürich.

Wille, A., P. Zimmermann, E. Vranová, A. Fürholz, O. Laule, S. Bleuler, L. Hennig, A. Prelić, P. von Rohr, L. Thiele, E. Zitzler, W. Gruissem, and P. Bühlmann (2004). Sparse graphical Gaussian modeling of the isoprenoid gene network in Arabidopsis thaliana. *Genome Biology 5*, R92.

Wong, F., C. K. Carter, and R. Kohn (2003). Efficient estimation of covariance selection models. *Biometrika 90*, 809–830.

Wright, S. (1921). Correlation and causation. *J. Agricultural Research 20*, 557–585.

Wu, X., Y. Ye, and K. R. Subramanian (2003). Interactive analysis of gene interactions using graphical Gaussian model. *ACM SIGKDD Workshop on Data Mining in Bioinformatics 3*, 63–69.

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

Yeung, M. K. S., J. Tegnér, and J. J. Collins (2002). Reverse engineering gene networks using singular value decomposition and robust regression. *Proc. Natl. Acad. Sci. USA 99*, 6163–6168.

Zhu, J. and T. Hastie (2004). Classification of gene microarrays by penalized logistic regression. *Biostatistics 5*, 427–443.

Lebenslauf

Persönliche Daten

Name:	Juliane Stephanie Schäfer
Geburtsdatum:	04.04.1977
Geburtsort:	Konstanz

Schulausbildung

1983 – 1987	Franz-Schubert-Schule, Stuttgart
1987 – 1996	Eberhard-Ludwigs-Gymnasium, Stuttgart
Juli 1996	Abitur

Hochschulausbildung

November 1996 – Mai 2002	Studium der Statistik an der Ludwig-Maximilians-Universität München
Mai 2002	Diplom-Statistikerin der Ludwig-Maximilians-Universität München

Berufstätigkeit

Januar 1998 – Dezember 2001	Werkstudentin/studentische Hilfskraft am Institut für Medizinische Statistik und Epidemiologie der Technischen Universität München
seit Juni 2002	Wissenschaftliche Mitarbeiterin am Institut für Statistik der Ludwig-Maximilians-Universität München

München im November 2005
Juliane Schäfer