THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Network modeling and integrative analysis of high-dimensional genomic data

Jonatan Kallus

UNIVERSITY OF GOTHENBURG

Division of Applied Mathematics and Statistics
Department of Mathematical Sciences
University of Gothenburg and Chalmers University of Technology
Göteborg, Sweden 2020

Network modeling and integrative analysis of high-dimensional genomic data
Jonatan Kallus
Göteborg 2020
ISBN 978-91-7833-888-7 (print)
ISBN 978-91-7833-889-4 (electronic)
This thesis is available at
http://hdl.handle.net/2077/63747

Division of Applied Mathematics and Statistics
Department of Mathematical Sciences
University of Gothenburg and Chalmers University of Technology
SE-412 96 Göteborg
Sweden
Telephone +46 (0)31 772 1000

Typeset with LaTeX
Printed in Borås, Sweden 2020

Network modeling and integrative analysis of high-dimensional genomic data

Jonatan Kallus

Division of Applied Mathematics and Statistics
Department of Mathematical Sciences
University of Gothenburg and Chalmers University of Technology

Abstract

Genomic data describe biological systems on the molecular level and are, due to the immense diversity of life, high-dimensional. Network modeling and integrative analysis are powerful methods to interpret genomic data. However, network modeling is limited by the requirement to select model complexity and due to a bias towards biologically unrealistic network structures. Furthermore, there is a need to be able to integratively analyze data sets describing a wider range of different biological aspects, studies and groups of subjects. This thesis aims to address these challenges by using resampling to control the false discovery rate (FDR) of edges, by combining resampling-based network modeling with a biologically realistic assumption on the structure and by increasing the richness of data sets that can be accommodated in integrative analysis, while facilitating the interpretation of results. In paper I, a statistical model for the number of times each edge is included in network estimates across resamples is proposed, to allow for estimation of how the FDR is affected by sparsity. Accuracy is improved compared to state-of-the-art methods, and in a network estimated for cancer data all hub genes have documented cancer-related functions. In paper II, a new method for integrative analysis is proposed. The method, based on matrix factorization, introduces a versatile objective function that allows for the study of more complex data sets and easier interpretation of results. The power of the method as an explorative tool is demonstrated on a set of genomic data. In paper III, network estimation across resamples is combined with repeated community detection to compensate for the structural bias inherent in common network estimation methods. For estimation of the regulatory network in human cancer, this compensation leads to an increased overlap with a database of gene interactions. Software implementations of the presented methods have been published. The contributed methods further the understanding that can be gained from high-dimensional genomic data, and may thus help to devise new treatments and diagnostics for cancer and other diseases.

Keywords: graphical modeling, biomolecular interactions, sparsity, model selection, resampling, stability selection, community detection, matrix factorization, Euler parametrization, bi-clustering

Nätverksmodellering och integrativ analys av högdimensionell genomikdata

Jonatan Kallus

Avdelningen för tillämpad matematik och statistik
Institutionen för matematiska vetenskaper
Göteborgs universitet och Chalmers tekniska högskola

Sammanfattning

Genomikdata beskriver biologiska system på molekylär nivå och är, i och med livets enorma mångfald, högdimensionell. Nätverksmodellering och integrativ analys är kraftfulla verktyg för att tolka genomikdata. Dessa metoder begränsas dock av att modellkomplexiteten måste bestämmas och av en tendens att skatta nätverk som har en biologiskt orealistisk struktur. Dessutom finns behov av att kunna analysera en större bredd av data som beskriver olika biologiska aspekter, studier och grupper av subjekt på ett integrativt sätt. Den här avhandlingens syfte är att möta dessa utmaningar genom att använda stickprovsupprepning för kontroll av andelen felaktiga länkar i skattade nätverk, kombinera nätverksmodellering baserad på stickprovsupprepning med ett biologiskt realistiskt strukturantagande samt öka variationsrikedomen hos datamängder som är möjliga att analysera integrativt och samtidigt underlätta tolkningen av resultaten. I artikel I föreslås en statistisk modell för antalet upprepade stickprov i vilka respektive länk ingår i nätverksskattningen. Detta för att möjliggöra skattning av hur andelen felaktiga länkar i skattade nätverk påverkas av modellens gleshet. Därmed reduceras skattningsfelet jämfört med existerande metoder i forskningens framkant, och i ett nätverk skattat för cancerdata har alla hubbgener dokumenterade cancerrelaterade funktioner. I artikel II föreslås en ny metod för integrativ analys. Metoden, som baseras på matrisfaktorisering, inför en flexibel målfunktion som gör det möjligt att analysera mer komplexa datamängder och underlättar tolkningen av resultaten. Metodens användbarhet för utforskande analys demonstreras på genomikdata. I artikel III kombineras nodklustring med nätverksmodellering baserad på stickprovsupprepning för justera populära metoder så att de skattar nätverk med en mer biologiskt realistisk struktur. Vid skattning av regleringsnätverket för mänsklig cancer leder detta till ökad överensstämmelse med tidigare information om biologiska molekylära interaktioner. Programvaruimplementationer för metoderna som presenteras har publicerats. Metoderna som presenteras ökar förståelsen av högdimensionell genomikdata och har därigenom potential att bidra till utvecklingen av nya behandlingar och ny diagnostik för cancer och andra sjukdomar.

Nyckelord: grafmodellering, biomolekylära interaktioner, gleshet, modellval, stickprovsupprepning, stabilitetsselektion, nodklustring, matrisfaktorisering, Eulerparametrisering, biklustring

List of publications

This thesis is based on the work represented by the following papers:

Paper I. **Kallus, J.**, Sánchez, J., Jauhiainen, A., Nelander, S., Jörnsten, R. (2017). ROPE: high-dimensional network modeling with robust control of edge FDR. *Preprint arXiv: 1702.07685.*

Paper II. **Kallus, J.**, Johansson, P., Nelander, S., Jörnsten, R. (2019). MM-PCA: integrative analysis of multi-group and multi-view data. *Preprint arXiv: 1911.04927, Under revision for Biostatistics.*

Paper III. **Kallus, J.**, Nelander, S., Jörnsten, R. (2020). Large-scale network estimation with structure-adaptive stability selection. *Manuscript.*

Published papers not included in this thesis:

Einarsson, R., Cederberg, C., **Kallus, J.** (2018). Nitrogen flows on organic and conventional dairy farms: a comparison of three indicators. *Nutrient Cycling in Agroecosystems* 110, 25-38.

Cook, D. J., **Kallus, J.**, Jörnsten, R., Nielsen, J. (2020). Molecular natural history of breast cancer: Leveraging transcriptomics to predict breast cancer progression and aggressiveness. *Cancer Medicine* 2020;00:1-12.

Author contributions

Paper I. Shared responsibility for model development, specified and implemented the model and supporting software, designed and generated simulated data sets, evaluated the method and compared it to other methods, drafted and edited the manuscript.

Paper II. Developed and implemented the method and supporting software, designed and generated simulated data sets, evaluated the method and compared it to other methods, participated in the biological interpretation of the results, drafted and edited the manuscript.

Paper III. Developed and implemented the method, designed and generated simulated data sets, evaluated the method and compared it to other methods, drafted and edited the manuscript.

Acknowledgements

I want to thank everyone who has taken part in making this work possible. My supervisor Rebecka Jörnsten for inspiring discussions and enlightening problem-solving sessions. My co-supervisor Erik Kristiansson for friendly and reflective advice ranging from life in academia, via the philosophy of science to my own research work. My co-supervisor Sven Nelander for giving our work meaning by lucidly connecting it to the challenges of cancer research. José Sánchez, Patrik Johansson and Alexandra Jauhiainen for participating in the development of ideas in this thesis and for your friendly introduction to the research field. Sci-hub for working to remove all barriers in the way of science by providing access to scientific articles.

All of the colleagues at the department of mathematical sciences for creating such a friendly atmosphere. In particular, Tobias Ö, Fanny B, Viktor J, Olle E, Mariana P, Anna J, Anna R, Claes A, Fredrik B, Olle N, Henrik I, Henrike H, Ivar S, Sandra B, Oskar A, Felix H, Juan D, Malin P, Mikael G, Linnea H, Edvin W, Valentina F, Helga Ó, Sebastian J, Johannes B, Niek W and Andreas P for fun and interesting conversations, sometimes about research, and for cooperation in teaching and studies. Annika Lang, Marie Kühn, Loredana Colque and Liselotte Fernström for generous help with practical matters. Rasmus Einarsson for being curious and inquisitive. Your questions force me to think more clearly.

Mom and dad for support and guidance since as long ago as I can remember. Daniel for support and friendship. Josephine for loving support, patience and always believing in me.

Contents

1 Introduction

This thesis focuses on the development of statistical methods to increase the understanding of high-dimensional genomic data. Such data reflect biological systems on the molecular level and can provide insight into diseases and other aspects of living cells. From a statistical point of view, a key goal is to be able to make inference regarding the relation between covariates (i.e. transcripts or other biological compounds) and the relevance of these relations for the biological processes within the studied organism. The nature of genomic data pose challenges. First, genomic data are high-dimensional – such data can contain tens of thousands of measurements, due to the great number of biological compounds taking part in the processes of living cells. The high dimensionality poses statistical and computational challenges in itself. Secondly, due to the technical challenges in collecting such measurements, data are noisy and may suffer from unwanted variation caused by differences in laboratory procedures. Thirdly, complex interactions between covariates, such as feedback loops and non-linear dependencies caused by physical interactions between the biological compounds, calls for rich statistical models, exacerbating the challenge of high-dimensionality.

Two different, but highly related, approaches to the analysis of high-dimensional genomic data are addressed here. Network modeling aims to identify directly associated pairs of covariates among a large number of potentially confounding covariates. In this sense network modeling is an approach with a local focus. However, the focus can be shifted to the whole data set by considering the global structure of estimated networks. In contrast, integrative analysis based on low-rank matrix decomposition, instead, focuses directly on global patterns of variation in a data set. As we will see in section 2.1.3, network modeling is related to the smallest eigenvalues of the sample covariance matrix of a data set, while low-rank matrix decomposition is related to the largest eigenvalues.

New statistical methods have the potential to contribute to the understanding of the systems biology of living cells. Associations discovered in genomic data

are used to form hypotheses for biomarkers for improved disease diagnosis or the development of disease treatments. For statistical results to be useful as a guide for biological research it is important that they are accompanied by estimates of variance and accuracy. At the same time, the complicated structure of genomic data calls for new methods for explorative analysis, even if rigorous statistical theory regarding the properties of these methods does not yet exist.

This thesis is structured as follows. The next chapter gives, first, a brief background of genomic data; what it is and why it is interesting to collect and analyze. Thereafter, the challenge posed by high-dimensional data is introduced. Lastly, methods for finding associations in genomic data are reviewed, both pairwise associations through network modeling and associated groups of variables through low-rank matrix decomposition. The third chapter defines the aims of the thesis. Chapter 4 reviews resampling-based methods for network modeling to introduce the context of papers I and III. Chapter 5 introduces integrative analysis based on low-rank matrix decomposition to provide context for paper II. Chapters 4 and 5 also include methodological results not included in the papers. Chapter 6 summarizes the results of the included papers. The papers and their supplementary material are included in the thesis. Chapter 7 describes the software packages that were published along with the papers, and their capabilities and implementations are discussed. The final chapter concludes by summarizing the main contributions of the thesis.

2 Background

The diversity of living organisms, and life's ability to subsist and adapt through evolution, are astonishing. It is well known that DNA transfers information about the constitution of an organism between parent and offspring. But why are cells within a multi-cell organism so different, when they contain the same DNA? How is the information in DNA put to use? How do cells respond to changes in their environment and what has gone wrong when a cancerous cell starts to divide uncontrollably? All of these questions relate to the biochemical processes taking place within the cell, from DNA transcription to protein synthesis and function (Smith and Szathmary, 2000). Genomic data consist of measurements of the abundance of molecular components taking part in these processes. Measurements are made on samples of biological tissue, on cell colonies cultured in laboratories, or, more recently on single cells.

The central dogma of molecular biology (Crick, 1970) is the theory that genetic information is primarily transferred in the cell 1) from DNA to DNA through replication, 2) from DNA to RNA through transcription and 3) from RNA to protein through translation. Proteins are complex and diverse molecules responsible for most of the functions within cells. Figuratively, DNA is the blueprint for making proteins. Due to the role of RNA as a messenger, the abundance of a specific RNA molecule corresponds to how actively a specific piece of DNA is being transcribed, and a specific protein is being constructed. A piece of DNA that gets transcribed as a single RNA molecule is called a gene, thus gene expression is measured by RNA abundance (Smith and Szathmary, 2000).

In addition to gene expression data, several other types of genomic data can be collected. Variation in DNA between organisms of the same species, or between samples within the same organism, can be measured in terms of, for example, single nucleotide polymorphisms (SNP, variation at a single base-pair in the DNA), short insertions or deletions (indels), and copy number variations (CNV, longer DNA regions missing or being repeated). Epigenetic marks, such as

DNA methylation, responsible for the vast differences between different cell types despite containing identical DNA, are measured by a technique called chromatin immunoprecipitation (ChIP). ChIP measurements can be used to capture the type and genomic location of chemical interactions in connection to the DNA. Proteomics, the direct study of protein abundances, is challenging and cannot be conducted with satisfying quality at a genome-wide scale using current technology. It is, however, a fast-growing field (Richardson et al., 2016). The definition of genomic data, or genomics, used here is wide and includes types of data that are sometimes, more specifically, called e.g. transcriptomics (measurements of RNA abundance) or epigenomics (measurements of epigenetic marks).

For roughly two decades it has been possible to collect gene expression data on a massive scale. First, primarily through microarrays (Schena et al., 1995) and later through bulk RNA-Seq (Wang et al., 2009) and single-cell RNA-Seq (Hwang et al., 2018). Briefly, microarrays give a more crude estimation of the RNA abundance for a predefined set of base-pair sequences, compared to RNA-Seq which records base-pair sequences in a sample and matches them to genes afterward. RNA-Seq also has a higher dynamic range, meaning that it is able to measure both very low and very high concentrations with greater accuracy. Microarrays and bulk RNA-Seq give a measurement of the average state across all cells in a sample, whereas single-cell RNA-Seq can capture the distribution of cellular states in a sample. Human gene expression data contain measurements of the concentrations in a sample of biological tissue for about 20,000 genes. RNA is known to exhibit complex interactions with other RNA molecules and with the DNA, enabling e.g. the expression of one gene to inhibit or amplify the expression of other genes. The cancer genome atlas (TCGA) (The Cancer Genome Atlas Research Network et al., 2013) is a publicly available set of gene expression data and other genomic data from thousands of cancer patients. Covering measurements for several types of genomic data across many cancer types, TCGA allows for the comparison of different cancer types and the search for correlations between the different types of genomic data. Since it became available, the data set has often been used as an example in computational biology research. Due to the richness of the data set, methods that can be used to analyze and explore it have potential to be useful also for genomic data sets from a wide range of other scientific studies.

In statistics, a high-dimensional data set is a data set where the number of covariates (variables measured for each observation) is far greater than the number of observations. An example is RNA-Seq gene expression data for the cancer type *glioblastoma multiforme* in TCGA. It contains measurements for 20,530 genes (covariates) in 172 tumor tissue samples from human patients (observations). The statistical analysis of such data sets has become increasingly

important due to the increased ability to collect, store and transfer vast numbers of measurements. Genomics and other areas in computational biology are important examples. For the modeling of a high-dimensional data set, even the simple linear model (section 2.1.1) is too complex. Thus, the complexity of the linear model needs to be reduced further, e.g. by discarding covariates or otherwise constrain the linear model (Hastie et al., 2009).

2.1 Finding associations in genomic data

Associations in genomic data can be represented as a network, where each gene is represented by a node and nodes are connected by a link if the genes they represent are associated. Such network representations aim to raise the focus from the local associations between pairs of genes to systemic or global properties of the whole group of genes and their interactions. Low-rank matrix factorization, such as singular value decomposition or principal component analysis, is another way to model genomic data. Matrix factorization finds linear combinations of genes or of biological samples. These linear combinations can be used to find genes or samples that behave in some way that is typical in the data set or to find related groups of genes or samples. Matrix factorization can also be used to summarize high-dimensional data in fewer dimensions, to enable data exploration, through, for example, visualization.

2.1.1 Networks of pair-wise associations

Network models of human gene expressions have proven useful for the classification of cancer patients as well as for finding potential target genes for cancer therapies (Pe'er and Hacohen, 2011). Features at the network level that are of biological importance include genes that serve as network hubs and the network distance between them, as well as the betweenness-centrality of nodes (i.e. network bottlenecks). Such features can be predictive of survival time in cancer patients or be cancer-type specific (Jörnsten et al., 2011; Kling et al., 2015).

Network modeling of genetic networks concerns the estimation, from a genomic data set, of the edge set of a graph where the graph's set of nodes consists of all covariates in the data set. A graph is defined by a set of vertices V and a set of edges E, where each edge in E is a pair of vertices in V. The terms from applied fields (network, node, link) and corresponding mathematical terms (graph, vertex, edge) are used interchangeably in this thesis. The estimation

of the edge set of a graph connects to *multiple hypothesis testing* since high statistical power and asymptotically correct error control is desirable. The estimation problem also connects to *lasso* linear regression (linear regression that is constrained so that many parameters are exactly zero) since procedures based thereupon are computationally tractable. Lasso and multiple hypothesis testing are introduced in the coming subsections.

The estimation of the edge set is a high-dimensional model selection problem. Each potential edge corresponds to a parameter in a statistical model. To set a parameter to zero corresponds to not selecting the variable or edge. The lasso and related l_1-norm penalized methods are computationally and performance-wise efficient when sparsity can be assumed. Penalized methods rely on a choice of the amount of penalization, an inherently hard problem. The optimal amount of penalization depends on the number of observations and variables as well as several unknown quantities such as noise, true sparsity and variable interdependence structure. It also depends on the intended use for the network model. The choice of amount of penalization corresponds to the choice of model complexity in general model selection. The following sections provide a statistical background for network modeling and review methods for estimation of interaction networks.

Linear regression

Linear regression assumes the model $y = X\beta + \varepsilon$, where the response y and the error ε are n-dimensional vectors, the parameter β is a d-dimensional vector and $X \in \mathbb{R}^{n \times d}$ is a matrix of n observations and d covariates. The elements in ε are independent, identically distributed, independent of X and have expectation equal to zero. We can think of y as the gene expression of one gene and X as the gene expression of all other genes. Then β captures association between the gene represented in y and all other genes. With the most popular estimation method *least squares*, β is estimated by minimizing the sum of squared residuals $(y - X\beta)^T(y - X\beta)$. When $d \leq n$, X and y uniquely determines an estimate of β (assuming that X is full rank). In the high-dimensional case, however, the problem of estimating β is underdetermined. There exist infinitely many β such that $y = X\beta$ and a single solution does not say anything about the relation between X and y (Hastie et al., 2009).

To reduce model complexity, a constraint can be imposed on β. Common constraints include the l_2-constraint in *ridge regression* $\sum_{i=1}^{d} \beta_i^2 < R$ (Hoerl and Kennard, 1970) and the l_1-constraint in *lasso* $\sum_{i=1}^{d} |\beta_i| < R$ (Tibshirani, 1996). Lasso has the advantage that admissible β that minimize the sum of squared residuals are, in general, such that many elements in β are equal to

zero. This property of excluding less relevant covariates from the model is useful for the estimation of relevant covariates in genomic data sets. The lasso optimization is often formulated in the equivalent Lagrangian form

$$\min_{\beta} \left((y - X\beta)^T (y - X\beta)/2 + \lambda \sum_{i=1}^{d} |\beta_i| \right)$$

with the l_1-constraint changed into an l_1-regularization term. The regularization parameter λ corresponds to the constraining parameter R (Hastie et al., 2009). Compared to unconstrained least squares, lasso has drawbacks. First, the lasso estimate of β depends on a parameter λ. Secondly, the lasso estimation accuracy for β is less well understood (Bühlmann and van de Geer, 2011).

Graphical lasso

Assume that observations follow a multivariate Gaussian distribution, i.e. $X_i \sim N(\mu, \Sigma)\ \forall i$, where X_i is the ith row of X, μ is the mean vector and Σ is the covariance matrix. Then, if a pair of covariates are conditionally independent given all other covariates, the corresponding element in the precision matrix Σ^{-1} is zero. This allows for the modeling of gene expression data as a graph, where two genes are connected by an edge if their partial correlation is significantly non-zero. The meaningfulness of exact zeros suggests the construction of an estimator of Σ^{-1} that tends to estimate elements to be exactly zero using a lasso penalty. The log-likelihood for $\Theta = \Sigma^{-1}$, partially maximized with respect to μ, is given by $\log\det\Theta - \mathrm{tr}(S\Theta)$, where S is the empirical covariance of X and tr is the trace operator. The graphical lasso estimates a sparse graph by solving

$$\hat{\Theta} = \arg\max_{\Theta \succeq 0} (\log\det\Theta - \mathrm{tr}(S\Theta) - \lambda ||\Theta||_1)$$

where the constraint $\Theta \succeq 0$ means that Θ is constrained to be positive semidefinite and $||\Theta||_1$ is the sum of the absolute values of the elements in Θ. The maximization problem is convex and computationally tractable, although considerably slower to use than the method that is reviewed next (Friedman et al., 2008; Banerjee et al., 2008).

Neighborhood selection

Neighborhood selection was proposed before graphical lasso but is considerably faster and can be understood as an approximation of graphical lasso. It models

each covariate a with all other covariates using lasso

$$\hat{\beta}^a = \arg \min_{\{\beta:\beta_a=0\}} \left(\frac{1}{n}(X_a - X\beta)^T(X_a - X\beta) + \lambda \sum_{i=1}^{d} |\beta_i| \right)$$

where X_a is column a of X. Compared to graphical lasso, the optimization problem of neighborhood selection is computationally simpler. It is a drawback that it does not impose symmetry in gene associations directly in the optimization problem, i.e. that $\hat{\beta}^i_j = \hat{\beta}^j_i$. Symmetry is instead enforced after $\hat{\beta}$ has been computed, by letting the set $\{(i,j) : \hat{\beta}^i_j \neq 0 \vee \hat{\beta}^j_i \neq 0\}$ be the estimated edge set (Meinshausen and Bühlmann, 2006).

Multiple hypothesis testing

In mathematical statistics, decision problems are approached using hypothesis testing. When deciding if data support an association between the expression of two genes, the *alternative hypothesis* that the association is supported is posed against the *null hypothesis* that it is not. If the probability of the observed data, or a more extreme observation, under the null hypothesis is below some threshold the null hypothesis is rejected. This probability is called the p-value and the threshold is commonly 0.05 (Rice, 2006). When multiple tests are performed, such as testing the association between a gene and all other genes or even the association between all pairs of genes, the classical framework is unsatisfactory. Since the probability of failing to reject a specific hypothesis is 0.05 (if the threshold is 0.05 and the null hypothesis is true), we have to expect that 5% of all unassociated genes will be falsely deemed as associated. There is thus a risk that correctly rejected null hypotheses are lost among a large number of falsely rejected null hypotheses. Instead of focusing on the error probability in a single test it is relevant to control the total number of errors. The family-wise error rate (FWER) is the probability that at least one null hypothesis is falsely rejected. The false discovery rate (FDR) is the expected proportion of rejected null hypotheses that are falsely rejected (Hastie et al., 2009).

Parametric hypothesis testing relies on an assumption of the distribution of the test statistic under the null hypothesis. In large-scale multiple testing problems, where the proportion of alternative cases is typically less than 10%, parametric hypothesis tests can be improved by using an empirical null distribution. Empirical null distributions are generally overdispersed relative to a theoretical null distribution, for the following reasons: the existence of unobserved covariates, correlations that are not accounted for in the theoretical null distribution and

the existence of many real but uninterestingly small effects. The use of an empirical null distribution makes an important difference in multiple testing, and rich null distributions (in comparison to commonly used theoretical null distributions) are needed to capture overdispersion (Efron, 2004).

When controlling the false discovery rate, a measure of statistical significance called the *q-value* (Storey and Tibshirani, 2003) is useful. While performing multiple hypothesis significance tests, q-values are assigned to each alternative hypothesis so that if all alternative hypotheses with $q < 0.05$ were called significant, an FDR of approximately 0.05 would be achieved. Thus, q-values have the same relation to FDR as p-values have to false positive rate.

2.1.2 Low-rank matrix factorization

Matrix factorization is a type of exploratory analysis. It aims to reveal dominant trends in a data set rather than to answer specific questions that have been formulated beforehand. Singular matrix decomposition (SVD) factorizes a matrix $X \in \mathbb{R}^{n \times p}$ into orthonormal (pairwise orthogonal columns of unit ℓ_2-norm) matrices $U \in \mathbb{R}^{n \times n}$ and $V \in \mathbb{R}^{p \times p}$ and a non-negative diagonal matrix $D \in \mathbb{R}^{n \times p}$ such that $X = UDV^T$. The diagonal elements of D, called singular values, are commonly sorted in descending order (and the columns of U and V are sorted accordingly). The matrix product UDV^T can also be we written as a sum

$$X = \sum_{i=1}^{\min(n,p)} D_{ii} U_i V_i^T,$$

where D_{ii} are elements of D and U_i, V_i are columns of U and V. This formulation emphasizes the possibility to view SVD as a decomposition of X into a sum of rank-one matrices, called components, of varying importance (given by the magnitude of D_{ii}). By disregarding all but the first k components, a rank k approximation of X is given by $X \approx UDV^T$ with matrix sizes instead $U \in \mathbb{R}^{n \times k}$, $D \in \mathbb{R}^{k \times k}$ and $V \in \mathbb{R}^{p \times k}$ for $k \leq \min(n, p)$. In fact, this approximation of X is optimal in the ℓ_2 sense, i.e. there is no matrix $\hat{X}$ of rank k or less such that the sum of squared elements of $X - \hat{X}$ is smaller than that of $X - UDV^T$ (Golub and Van Loan, 1996). Due to this result, the SVD can be found by solving a constrained optimization problem

$$\underset{U,D,V}{\arg\min} \, ||X - UDV^T||_F$$

such that the columns of U and V have unit ℓ_2-norm and D is diagonal and non-negative, where $|| \cdot ||_F$ is the Frobenius norm, i.e. the square root of the

sum of squared elements. Orthogonality of columns of U and V is not necessary as a constraint, since the optimal solution has orthogonal columns due to the optimality of SVD.

The statistical method principal component analysis (PCA) is closely related to SVD. It is used for dimension reduction of a data matrix, commonly for visualizing the data as a scatter plot in two or three dimensions. PCA constructs new variables that are linear combinations of the original variables, such that each new variable has maximum variance while having zero covariance with the other new variables. The new variables are given by the V matrix of the SVD. The columns of V are called loadings in PCA. The SVD matrices U and D are combined with multiplication to a matrix T, called scores, such that $X \approx UDV^T = TV^T$. For PCA truncated to rank k, the k loadings hold the coefficients for constructing the new variables from the original variables and the score matrix holds the observations in terms of the k new variables. In computer science, the score matrix is often called an embedding, and observations are referred to as being embedded in a k-dimensional space.

The low-rank SVD minimizes the ℓ_2 error. Statistically, this corresponds to an assumption that each element in the data matrix has an independent normally distributed error. Rank k SVD in itself corresponds to an assumption that each variable is a linear combination of k unobserved variables. These model assumptions lead to the view of SVD as a decomposition of data into the first k components, sometimes called the signal, systematic variation or patterns, and the error, often called noise.

2.1.3 Large and small eigenvalues

The sample covariance matrix $S = X^T X/(n-1)$ is central to both network modeling and matrix factorization. It is assumed here that the columns of X have mean zero, which is natural to do since the focus is on covariance, and covariance is unaffected by mean shifts. In the least squares method, the optimal estimate of β is given by $(X^T X)^{-1} X^T y$. If X is high-dimensional, then $X^T X$ is not invertible, $(X^T X)^{-1}$ is undefined and some eigenvalues of $X^T X$ are zero. In lasso, ridge regression and graphical lasso the imposed constraints can be thought of as ways to find an inverse of an approximation of S, since the inverse of S does not exist. For an invertible matrix A, the inverse of the eigenvalues of A are equal to the eigenvalues of A^{-1}. In the constrained methods for linear regression and graphical modeling, the focus is on β or Θ which depend on $(X^T X)^{-1}$ and thus on the smallest eigenvalues of $X^T X$. By imposing constraints, an inverse is instead found to a matrix which is similar to S, where similarity is implicitly defined by the type of constraints. The

focus of these methods is local in the sense that they estimate an association between the response and each covariate, or between each pair of covariates. In matrix factorization, on the other hand, the focus is global in the sense that large groups (linear combinations) of variables are estimated. Correspondingly, in matrix factorization the focus is on the largest eigenvalues. In SVD, V holds the first k eigenvectors of X^TX, U holds the first k eigenvectors of XX^T and D holds the square roots of the k largest eigenvalues of X^TX, or equivalently of XX^T. To summarize this paragraph, matrix factorization uses the eigenvalues that can be stably calculated in high-dimensional data to capture global properties of the data but it can not well describe properties of individual pairs of covariates. Network modeling perturbs high-dimensional data to be able to estimate small eigenvalues (indirectly by approximating $(X^TX)^{-1}$) and to capture local properties of the data. In this regard, both classes of methods are necessary to gain maximum understanding from high-dimensional data.

There are several shortcomings in existing methods for network modeling and matrix factorization, that limit their usefulness in the analysis of genomic data. One example is the choice of model complexity, i.e. λ in graphical modeling and k in matrix factorization, which is difficult. Traditional methods for selecting model complexity, cross-validation and information criteria, are prone to overfit and are sensitive to outliers (Jörnsten et al., 2011). When the goal of graphical modeling is interpretation (e.g. biomarker identification or mechanistic understanding) an accurate control of the rate of falsely discovered edges (FDR) is typically more important than maximizing stability or likelihood (Storey and Tibshirani, 2003). The selection of model complexity in network modeling such that the FDR is controlled at a specific level is addressed in paper I. Existing methods for integrative analysis using matrix factorization are limited in the complexity of data relations that they can accommodate, their results are difficult to interpret and/or they do not specifically address the choice of model complexity. Matrix factorization has long been used for the analysis of single data matrices. Genomic data, however, come in the form of multiple related data matrices. These matrices contain information of different groups of studied subjects, each group described in terms of multiple, and not necessarily the same, types of genomic data. These needs are addressed in the method for integrative analysis that is developed in paper II. Integrative analysis, the use of matrix factorization for the simultaneous analysis of multiple related matrices, is an active area of research, and there is a strong need for new statistical methods (Richardson et al., 2016). The assumption of network sparsity is central in graphical modeling. However, the ℓ_1-penalty that is used to impose the assumption has the side-effect of promoting networks with a structure that is unlikely for biological networks (Tan et al., 2014). Paper III addresses this side-effect, while still allowing for control of the FDR.

3 Aims

This thesis aims to further develop statistical methodology for the understanding of high-dimensional genomic data sets by means of graphical modeling and integrative analysis. These approaches can elucidate several aspects of high-dimensional data; from local associations between pairs of covariates, to global patterns in a high-dimensional matrix and even associations between patterns in different data matrices. The high dimensionality and heterogeneity of genomic data pose statistical challenges, and a balance needs to be struck between richness and simplicity of methods. On one hand, methods need to be rich enough to be able to handle genomic data sets and to answer biological and medical questions. On the other hand, methods need to be simple enough to allow for a statistical understanding of the results produced, in order to ensure that drawn conclusions are not too strong nor too weak given the available data. More specifically the aims are:

- To develop and evaluate a new statistical model for edge selection counts in resampling-based network modeling, in order to more accurately control the false discovery rate of network edges (paper I).
- To enhance integrative analysis based on matrix factorization in order to analyze more complex data sets, and, at the same time, facilitate the choice of model complexity and the interpretation of results (paper II).
- To unify resampling-based network modeling with a biologically realistic assumption on the global structure of estimated networks, in order to increase accuracy while controlling the false discovery rate of network edges (paper III).

In all, the overarching aim of the thesis is to increase the potential to gain understanding from heterogeneous high-dimensional genomic data.

4 Resampling-based network modeling

To what extent can a network estimate be trusted? Are some or all edges strongly influenced by a few of the observations in the data set or are they representative of an entire population? To what extent can specific properties of the estimated network, or specific locations in it, be trusted? The methods reviewed in section 2.1.1 are estimators of edge sets of graphs. Given a data set $X \in \mathbb{R}^{n \times d}$ and regularization parameter λ they make an estimate $\hat{S}^\lambda(X) \in \{0,1\}^p$ of an edge set. With an indexing over all pairs of covariates in X, $\hat{S}^\lambda_i(X) = 1$ means that the ith pair of covariates is in the estimated edge set. It follows that the number of potential edges is $p = d(d-1)/2$. When using network estimates for making biological hypotheses it is beneficial to have an understanding of the distribution of such estimates. The field of statistical inference concerns the distribution of estimates such as $\hat{S}^\lambda(X)$.

A common way to estimate the distribution of $\hat{S}^\lambda$ is using bootstrap or other resampling methods. Bootstrap uses the sample X to form new samples with approximately the same distribution as X. A bootstrap sample $R(X)$ consists of n rows drawn randomly among the rows of X, with replacement. The distribution of the estimator $\hat{S}^\lambda$ can then be approximated by applying it to several resamples $R(X)$. In addition to getting an understanding of a specific estimator, this procedure can be used to compare different estimators (i.e. different levels of regularisation for a specific method or different methods). Furthermore, all of the bootstrap estimates $\hat{S}^\lambda(R_i(X))$, where R_i is the ith resample, constitutes a new data set that can be used for estimating the network. This route has the potential to improve error control and to improve robustness by decreasing sensitivity to single observations in X.

Paper I contributes a new method for resampling-based network estimation that has more exact FDR control than existing methods and is also considerably

more robust to the randomness introduced by resampling than one of the state-of-the-art methods. Paper III presents a new resampling-based method that is able to impose an assumption that networks have a community structure. In order to get a broader perspective, this chapter reviews two existing resampling based network estimators that are state-of-the-art in terms of control of the false discovery rate (FDR). This chapter also discusses how network modeling can be improved by making a biologically realistic assumption on the network structure.

Assume that bootstrap is used with B resamples. Then, estimating a graph for each bootstrap sample yields B graphs, with equal sets of nodes but different sets of edges. Thus, each potential edge i will have appeared W_i^λ times, $W_i^\lambda \in \{0, \ldots, B\}$,

$$W_i^\lambda = \sum_{j=1}^{B} \hat{S}_i^\lambda(R_j(X)).$$

W_i^λ is thus the selection count for edge i at regularization λ. Figure 4.1, where the selection count of individual edges is plotted against different levels of regularization, shows how individual edges respond to varying regularization. Figure 4.2, where a histogram shows how many edges that were selected k times for a specific λ, shows the empirical distribution of edge selection counts at one level of regularization.

4.1 Edge-wise control of the false discovery rate

Simple ways to estimate a network using selection counts W^λ would be to include all edges with $W_i^\lambda > 0$ (edges selected in at least one resample) or edges with $W_i^\lambda = B$ (edges consistently selected in all resamples) or something in between (e.g. edges selected in a majority of resamples). Stability selection (section 4.1.1), bootstrap inference for network construction (BINCO, section 4.1.2) and resampling of penalized estimates (ROPE, sections 4.1.3, 6.1 and paper I) are, however, more sophisticated. Stability selection focuses on the maximum selection count of each edge, $\max_\lambda W_i^\lambda$. BINCO fits a decreasing function to a part of the edge selection histograms. ROPE models the sequence W_i^λ for each fixed λ with a probability distribution. All three methods address the problem of selecting the level of regularization λ by using several W^λ corresponding to a range of λ values.

4.1.1 Stability selection

Stability selection uses the maximum selection count for each edge over the entire range of λ values, $\max_\lambda W_i^\lambda$. Meinshausen and Bühlmann (2010) derive an upper family-wise error rate (FWER) bound for a threshold k_t where all edges for which $\max_\lambda W_i^\lambda > k_t$ are selected (figure 4.1). It is shown in paper I that the achieved FWER is, in many cases, far below the FWER bound. This method results, thus, in too conservative choices of k_t and, in turn, too sparse network estimates.

Complementary pairs stability selection (Shah et al., 2013) introduces a less conservative choice of k_t. The method proposes complementary pairs subsampling, which means that subsampling is performed, without replacement, with two non-overlapping random subsamples at a time, each of size $\lfloor n/2 \rfloor$. Due to the subsamples being non-overlapping, the two subsamples constitute independent estimates of the population distribution. This independency is used to derive an improved FWER upper bound.

4.1.2 BINCO

BINCO (Li et al., 2013) estimates the null hypothesis distribution of edge selection counts for each value of λ (figure 4.2). The histogram estimates the distribution of edge counts, but it contains both null and alternative edges. In order to estimate a distribution that only includes potential edges that should not be included in a good network estimate, a range of selection counts is chosen that is dominated by such edges. The choice of such a range is based on the histogram having an approximate U-shape.

It is a good sign when edge selection counts are U-shaped. In an ideal case, where the network estimator estimates an identical network for each bootstrap sample, each edge will get a selection count of either 0 or B. In a slightly less ideal case, edges will be selected either a small number of times or almost B times, resulting in a U-shaped histogram. It is often the case that the mode for the distribution of false edges is larger than zero. Therefore, an assumption of U-shape in the entire range is too strong. Instead histograms are assumed to be U-shaped in a range $\{c, \ldots, B\}$, $0 \leq c < B$.

Li et al. (2013) states the assumption of approximate U-shape precisely as the *proper condition*. The proper condition is satisfied when the empirical probability density function for edge selection counts is U-shaped in the limit $B \to \infty$, i.e. that when restricting the function to this interval, the function has local maxima at its endpoints, a global minimum in the interior of its domain

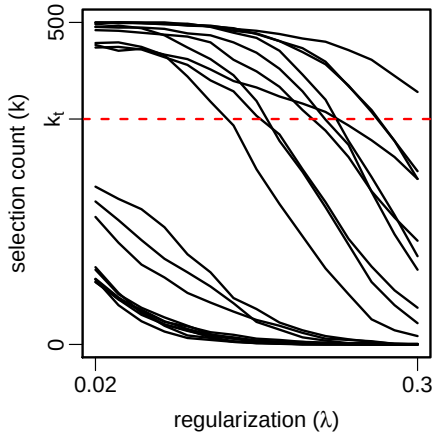

Figure 4.1: Edge selection counts k after 500 bootstraps over varying penalty parameter λ. A random subset of all edges is shown. The stability selection threshold is shown with a dashed red line. Stability selection selects all edges whose count is above a threshold k_t for at least one λ.

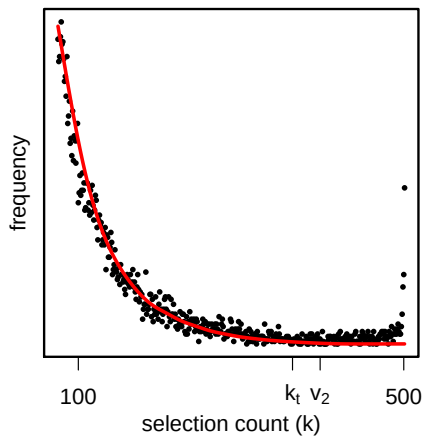

Figure 4.2: Edge selection count histogram after 500 bootstraps corresponding to one λ. The red line shows the null hypothesis distribution as estimated by BINCO. k_t shows the threshold by BINCO corresponding to an estimated FDR of 0.05. v_2 is the location of the minimum of the asymptotic distribution function estimated by BINCO.

and no other extrema. They show that the proper condition is satisfied by selection procedures for which the selection probability tends to one uniformly for alternative edges and has a limit superior strictly less than one for null edges, as $n \to \infty$. Li et al. (2013) also show that the condition is satisfied by selection procedures that are based on resampling of consistent selection procedures, such as the lasso when the irrepresentable condition (Zhao and Yu, 2006) is satisfied.

Approximate U-shape is a condition for both BINCO and ROPE. This condition excludes problems where the generating network is either extremely sparse in relation to the variance in network estimates caused by resampling, or where several different edge sets with small mutual overlap captures the data similarly well.

BINCO estimates the parameters of a modified decreasing beta-binomial density function to fit the null hypothesis distribution of edge selection counts. The modification of the density function is made to allow for overdispersion. This overdispersion may be caused by dependencies between selection counts for different edges and other reasons (see section 2.1.1). To minimize influence from edges in the alternative distribution, only the decreasing part of the histogram is used. Using the estimated null distribution and the empirical distribution, the FDR can be estimated for each threshold k_t. This procedure is repeated for a range of regularization levels. Thus, a threshold k_t is calculated for every λ, each corresponding to the same estimated FDR. To make its network estimate, BINCO uses the edge selection counts from the regularization level λ for which most edges are selected. That is, λ is selected to maximize the estimated power.

There are drawbacks in using only the decreasing range of the histogram to estimate parameters of the null distribution. Depending on the shape of histograms, thresholds corresponding to relevant false discovery rates are often located outside the range used to fit the model. When that is the case, extrapolation of the fitted model gives an unnecessarily large variance in choice of threshold. Furthermore, the presence of the alternative population in the decreasing range, especially its rightmost part, can cause an erroneous estimate of the null distribution.

4.1.3 Joint modeling across regularization levels

Instead of modeling the edge selection counts for each regularization level independently, modeling can be improved by modeling jointly across regularization levels, due to three facts and assumptions. First, it is reasonable to assume that the distribution of edge selection counts changes smoothly when the amount of regularization is changed. Secondly, an increase in regularization leads to a

sparser network. Thus the mean of the distribution of edge selection counts decreases when regularization increases. Thirdly, the proportion of potential edges that should be included in a correct network is fixed. By switching to a such global model, variance in the estimation of model parameters that is caused by the finiteness of the number of bootstraps and observations is decreased. These three relationships between selection counts and regularization are illustrated in figure 4.3. The figure also illustrates the relationship between histograms and how the curves of individual edges change as functions of regularization.

Numerical likelihood maximization for such a global model is challenging. Challenges include the large number of model parameters and a sound and efficient formulation of constraints that enforce smoothness in distribution change. The large number of potential edges and the possibility to perform many bootstraps ensures that there is much selection count data available to fit local models at each regularization level. This suggests that gains from enforcing smoothness across regularization levels are small. In ROPE we enforce the fact that the proportion of edges that should be included in the network are fixed regardless of λ. It is demonstrated in paper I that this constraint decreases bias and increases robustness.

4.2 Assumptions on network structure

The discussed methods for network estimation all rely on a sparsity assumption. The sparsity assumption is local in the sense that it affects the estimation of each edge individually and independently, only indirectly affecting the entirety of estimated networks. This section treats the imposition of an assumption on network structure, a kind of assumption that directly affects whole network estimates. The structure of a network refers to its overall shape, i.e. how nodes and edges tend to be configured in the network. Some typical network structures are represented by scale-free networks, hub networks, networks with communities and uniformly random networks. Specific networks may have a structure more or less similar to these representative networks, or have a resemblance to several of them or other structures. Network structure can most precisely be described by probabilistic generative models. In the generative model for uniformly random networks, called the Erdős-Rényi model, a random decision is made independently, and with equal probability, for each pair of nodes to be either connected or not. Scale-free networks are generated by the Barabási-Albert model, which considers nodes one at a time and connects the considered node to previously considered nodes randomly such that more highly connected nodes are preferred. Hub networks have some nodes, hub nodes, which are more highly connected than other nodes. In networks with community

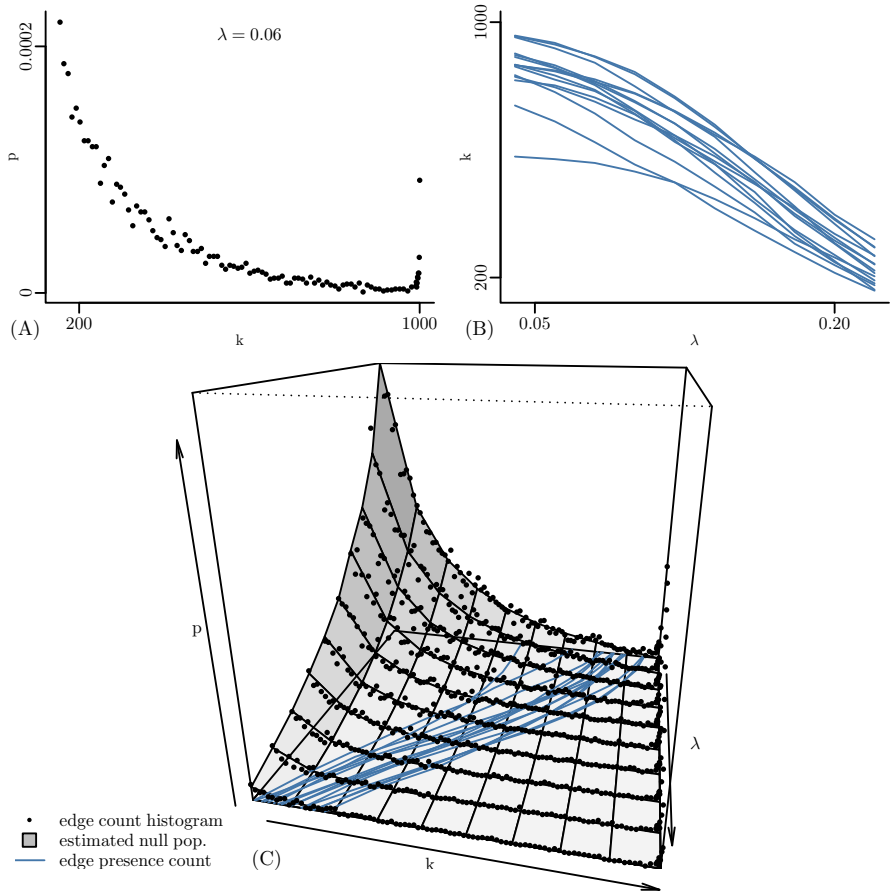

Figure 4.3: Combined two-dimensional histogram of edge selection counts and penalization parameter values, along with estimated null population density (C). Edge selection counts for some individual edges are shown as functions of the penalization parameter (B, C). If all individual edges were shown, the density of curves would have corresponded to the height of the histogram. This figure shows the relationship between selection count histograms (A) and curves (B). It also shows how histograms changes smoothly with λ. (k: edge presence count, p: frequency, λ: regularization level.)

structure, nodes are partitioned into groups, called communities, and nodes within the same group are more highly interconnected than nodes in different groups.

It has been observed that genetic networks tend to have a structure similar to scale-free, hub or community networks, or a combination thereof (Eisen et al., 1998; Albert, 2005; Hao et al., 2012). Uniformly random networks, on the other hand, do not have a structure that is likely for genetic networks. Inappropriately for estimation of genetic networks, methods for network estimation which put an individual sparsity penalty on each potential edge, such as graphical lasso and neighborhood selection (section 2.1.1), implicitly make the assumption that the estimated network has a uniformly random structure (Tan et al., 2014). Due to the small number of observations that are typically available (in relation to the large number of parameters to be estimated in network modeling) when estimating genetic networks, relevant assumptions are important. Statistically, an assumption of a particular network structure increases the likelihood for networks of that structure. This is, of course, beneficial if the unknown true network is of such structure.

4.2.1 Community detection

The attempt to find communities in a given network is called community detection. In contrast to network modeling, methods for community detection do in general treat the network as an observed truth. Community detection in a network is similar to cluster analysis of observations in a data set. Both depend on the subjective choice of a function for comparing candidate communities or clusters, based on e.g. density of network edges within the community or of observations within the cluster. Both also require a strategy for selecting the number of communities or clusters. Louvain (Blondel et al., 2008) is a method for community detection that is fast for large networks. The method uses modularity, a function of the partitioning of the nodes of a network, that increases with the edge density inside communities and decreases with the edge density between communities. Global optimization of modularity is not computationally tractable. Blondel et al. (2008) proposed a heuristic for approximately maximizing the modularity. Their algorithm iterates over all nodes repeatedly. Initially each node is in its own community. If a node has neighbors in other communities, it is moved to the neighboring community which most increases the modularity. When no move exists that increases modularity, each community is merged into a single node and the algorithm is applied again to the new smaller network. Moves and mergers are iterated until modularity no longer increases. In addition to the computational efficiency

of the algorithm, which is due to its quick reduction of the problem size, the algorithm does not require a choice of the number of communities beforehand.

4.2.2 Consensus clustering

Consensus clustering (Monti et al., 2003) is a method for the estimation of stable clusters. It is highly similar to stability selection (section 4.1.1), but is used for clustering rather than network estimation or other problems of variable selection. In consensus clustering, any method for cluster analysis is applied repeatedly to random subsamples of a set of observations. The results from all repetitions are summarized in a *consensus matrix*. The matrix is symmetric and has the same number of rows and columns as the number of observations in the initial data set. For each pair of observations, it holds the proportion of repetitions in which these observations were estimated to be in the same cluster. The repeated cluster analyses are summarized into a consensus clustering by using the consensus matrix as an observation similarity matrix. The purpose of consensus clustering is twofold. First, it aims to estimate the stable clusters, i.e. a clustering that is robust to sampling variability. Secondly, it proposes a procedure for selecting the number of clusters. In paper III, we adapt consensus clustering for use in community detection. A difference between clustering and community detection, when used in resampling-based procedures is that resampling is performed on observations, to achieve robustness to sampling variability, for both clustering and community detection, while clustering estimates a partitioning of observations and community detection estimates a partitioning of variables.

4.2.3 De-biasing of edge selection counts

The method structure-adaptive stability selection (SASS), proposed in paper III, performs consensus community detection on estimated networks and reduces the edge selection counts for edges between nodes of different communities. Thereby, the assumption of uniformly random networks inherent in methods for the estimation of sparse networks is replaced by an assumption of community-structured networks. Compared to ROPE, the histograms of edge selection counts are central to both methods. The main differences are in the different aims of the methods and in the different ways to model the histograms. Where ROPE is a novel way to control FDR in edge selection, SASS is a novel enhancement of stability selection in network modeling to enable the inclusion of a structural assumption on the estimated network. Where ROPE specifies a mixture distribution to model the histograms, SASS uses kernel density

estimation to make smooth estimates of the histograms.

In addition to using the repeated resampling to perturb network estimates, the repeated resampling is also used, in SASS, to perform consensus community detection. Using the consensus matrix, the method decides if communities seem to be present in the estimated network and, if so, divides all node pairs into pairs estimated to belong to the same community and pairs estimated to belong to different communities. This stratification of potential edges is used to construct two histograms of edge selection counts, one for within community edges and one for between community edges. The bias of methods for sparse network modeling toward a homogeneous edge density in all parts of networks is manifested by different shapes of the two histograms. Edge selection counts for between-community edges are reduced to make up for this shape difference. Previously existing methods for imposing structure assumptions in network modeling rely on modifying objective functions or estimation procedures of existing methods for sparse network modeling. SASS instead combines complementary pairs stability selection and consensus community detection with existing methods for sparse network modeling.

5 Integrative analysis

Integrative analysis, as the term is used in this thesis and generally in the field of high-dimensional statistics, is the simultaneous analysis of multiple data matrices with the intent to be more informative than separate analyses of each matrix. A prerequisite for this to be possible is that the different matrices contain information that is somehow related. For example, two matrices may contain measurements for the same set of features (e.g. expression levels for a set of genes) for two different groups of observations (e.g. two patient cohorts stratified by disease type). The aim of integrative analysis is to identify structure that is consistent across multiple data matrices, and also to use similarities between matrices to increase statistical power to identify structure that is present only in individual matrices. The identified structure may subsequently be used to e.g. find clusters of variables or observations or to identify differences or similarities between the given groups of variables or observations. Ideally, integrative analysis of heterogeneous matrices should not obscure structure that is not present in all analyzed matrices. Compared to data analysis in general, which uses a model to separate the structure from the random noise, integrative analysis further separates the structure according to which matrices it is present in.

In this thesis, the focus is on integrative analysis by means of matrix factorization. Such analysis is based on the simultaneous factorization of multiple data matrices so that each matrix is approximated by a sum of low-rank matrices. Each low-rank matrix is, in turn, a product of two matrices (called factors). The decomposition is made such that identical factors take part in the approximation of different data matrices. The pattern in which factors influence the approximation holds information of which data matrices that have a similar structure and how the structure is similar. In addition to this pattern of similarities between data matrices, integrative analysis by means of matrix factorization also yields the understanding that can be gained by ordinary methods for matrix factorization of a single data matrix, such as PCA

or SVD.

The term integrative analysis is sometimes used with different meanings than the one that is used here. In particular, within systems biology the term is often used to refer to any methods or procedures developed for the analysis of data regarding different aspects of the same set of subjects. In contrast to the approaches discussed in this thesis, such methods are often specialized for a specific problem or type of data.

5.1 Applications

Many scientific fields have use for the analysis of multiple related high-dimensional data matrices with complex relations between groups of observations and groups of variables. In the last decades, however, methods for integrative analysis have primarily been used and developed within chemometrics, systems biology, computer science and statistics (see references in table 2 in the supplementary material to paper II).

Within systems biology, much focus on integrative analysis has been driven by the aim to use the data set released by The Cancer Genome Atlas (TCGA) to improve our understanding of cancer biology. TCGA is a project aimed at compiling and publishing large sets of data that describe cancer genetics and the cell biology of cancer on a molecular level. The released data consist of thousands of biological samples of cancerous tissue, sometimes matched with samples of normal tissue. The biological samples are divided into 33 different types of cancer. Multiple aspects of the cellular biology are measured, e.g. gene expressions, mutations and methylation, as well as clinical data regarding the patients, their disease type, their treatment and the disease progression. Integrative analysis of this data set has the potential to reveal connections between the different types of measurements that may be mechanistic, and thus have relevance for the development of new therapies and improved methods for diagnosis. It also has the potential to find subsets of patients that may benefit from particular types of treatments, due to e.g. biological mechanisms that are active in that specific group. A hypothetical example is shown in figure 5.1, which illustrates six matrices, where the three matrices to the left measure copy number variations (CNV), a type of mutation, at a number of chromosomal locations and the three matrices to the right measure expression levels for a number of genes (RNA). The two top matrices regard biological samples from patients diagnosed with the cancer type glioblastoma multiforme (GBM), the two middle matrices regard samples from healthy patients and the two bottom matrices regard patients diagnosed with breast cancer (BRCA). The colored

bars within matrices illustrate columns of factor matrices that the matrices have been decomposed into by an integrative approximation. White bars illustrate columns of factor matrices that are individual to only one matrix and colored bars illustrate columns of factor matrices that capture structure that is common to at least two matrices. This example will be made more concrete in section 5.2. The figure shows structure that is shared between CNV and RNA data for two of the patient groups (purple and green bars), suggesting a connection between the two data types that are present for GBM and healthy patients, but not for breast cancer patients. RNA structure is similar for the two cancer types (blue bars), and could be cancer-related. CNV structure is similar for all groups of observations (red bars), but there is also structure that is individual to the healthy patients. This is an illustrative example to describe the potential for integrative analysis and is not informative for the biology of cancer.

The division of variables (or observations) into distinct groups facilitates interpretation and enables analysis of data sets where entire sets of variables are missing for some groups of observations. In systems biology, it is natural to treat data regarding different biological components or laboratory platforms (gene expression, mutations, methylation etc.) as distinct groups of variables. This is both because there is reason to believe that data quality (in terms of signal-to-noise ratio) differs and because it is of interest to focus analysis on covariation between specific groups of variables rather than between individual variables. Covariation between groups of variables can in some cases be used to form hypotheses of mechanistic influence based on the central dogma of molecular biology (section 2). For similar reasons, it is useful to divide observations (e.g. biological samples) into groups by disease type or other important properties.

Integrative analysis is also used in other fields. Within computer science, integrative analysis has mainly been used in so-called recommender systems. Typically, recommender systems are used to predict to which extent each individual in a group of users would enjoy each item in a set of items, in order to automatically recommend items to users. A simple example involves three matrices: one containing features of each user, one containing features of each item and the third containing grades that users have given to items that they have already experienced. The third matrix would have missing data for each element that corresponds to an item that the user has not yet experienced, and the recommender system would be used to impute the missing data. This focus on prediction by imputation is in contrast to the focus, in other fields, on the interpretation of shared structure and the pattern in which it is shared.

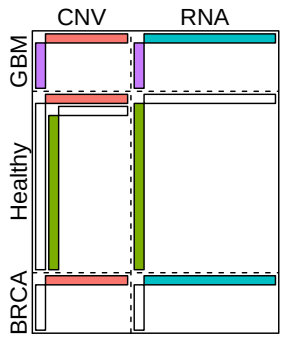

Figure 5.1: Example for illustrative purposes. Six data matrices that are vertically or horizontally related. Integrative analysis can identify structure that is shared between data matrices (colored bars) and structure that is specific to an individual data matrix (white bars)

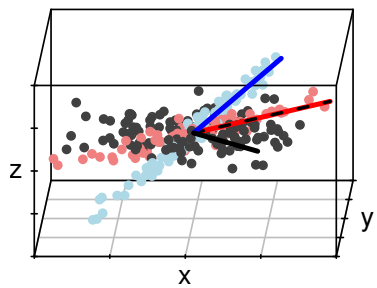

Figure 5.2: Example of vertical integrative analysis of a low-dimensional data set. Data consist of three groups of observations (black, red and blue). Colored lines show dominating directions of variation for each associated group. One of the directions describes variation in both the red and the black group. An individual PCA of each data set would not have aligned one of the black lines with the red line.

5.2 Integrative low-rank decomposition

SVD (section 2.1.2) can be utilized for integrative analysis by simultaneously decomposing multiple related matrices while enforcing similarity between the decompositions. A simple setting is when all matrices describe the same set of observations. The convention in statistics is to have data matrices with one row for each observation and one column for each variable. Thus, in this setting, all matrices have the same number of rows, and similarity between U-matrices of each decomposition can be enforced. This is called horizontal integrative analysis. Analogously, it is called vertical integrative analysis when multiple matrices describing the same set of variables for multiple groups of observations are analyzed. In vertical integrative analysis, similarity between V-matrices is instead enforced. Figure 5.2 shows a simple example of vertical integrative analysis. It shows three groups of observations (colored black, red and blue) in three variables (x, y and z). The loadings (columns of V-matrices) are depicted by colored lines. The observations of the blue and red groups lie approximately along two lines, one for each group, and the directions of their first loadings are not much affected by noise. The observations of the black group are spread in the vicinity of a disc. These observations are described approximately as well by any two diagonal loadings as long as they are in the same plane as the observations. The direction of the first loading for the black group is highly dependent on random noise. In the example, an integrative analysis has found that the red group can be well described by a loading that also captures much variability in the black group. The blue group, on the other hand, is spread along a line that is not shared with any other group. This example shows why an integrative analysis can not, in general, be performed by first performing SVD on each group separately and then searching for similarities of scores and loadings. An individual decomposition of the black group could very well have found loadings quite far from the loading of the red group. Similar difficulties are exacerbated in data sets of higher dimension.

Bi-directional integrative analysis combines horizontal and vertical integrative analysis. Figure 5.1, discussed previously, shows an example. With the concepts from this section, the meaning of the colored bars in the figure can be made more concrete. The matrix holding CNV data for healthy patients has been approximated by an SVD of rank two. The other matrices are approximated by rank one SVD, since they have only one horizontal and one vertical bar each. Out of the two U-columns for healthy CNV data one is equal to the single U-column for healthy RNA data, since they share the same color. Augmented multi-view data (AMD), introduced by Klami et al. (2014), is more general than bi-directional data. In bi-directional data, each set of entities (e.g. genes or patients) must be associated with either rows or columns of matrices. In

AMD, the same set of entities can instead be associated with rows of some matrices and columns of other matrices. For example, a matrix that holds information on how closely related variables of different groups are, such as the chromosomal distance between gene locations and the location of probes for measurement of methylation, would have rows associated with one group of variables and columns associated with another group of variables. AMD is the most general setting for CMF (section 5.5.2) and MM-PCA (section 5.5.3). AMD is treated in section 6.2 and in paper II (see in particular figure 1D in paper II). A related setting is data in tensor form. Where a data matrix holds relationships between two sets of entities, a data tensor holds relationships between an arbitrary number of sets of entities. Tensors can be analyzed with horizontal or vertical integrative analysis, but doing so disregards that *both* the columns of each tensor slice describe the same set of entities and the rows of each tensor slice describe the same set of entities. Methods for low-rank decomposition of single tensors exist, but no general method exists that can analyze multiple related tensors integratively without disregarding any of the given relations.

5.3 Interpretation

Methods for integrative analysis all result in estimated matrices U, D and V, or similar decompositions, and an estimation of the structure in which these matrices are equal for different data matrices. These results need to be interpreted. This section describes two strategies for interpretation. First, results can be interpreted similarly to the results of a PCA. Such interpretation, which focuses on scores and loadings, is enriched by the structure in which these scores and loadings are shared among data matrices. Secondly, results can be interpreted as a general dimension reduction of data and focus on e.g. clustering.

As in PCA, loadings describe linear combinations of variables that show the most variation over the observations, and scores describe the direction and magnitude of variation for each observation in terms of each such combination. For a matrix of gene expressions, for example, the loadings (each column of V) can be thought of as a virtual gene that is made up of a linear combination of the genes in the data set. These virtual genes are sometimes called eigengenes, due to the loadings being the eigenvectors of the data matrix' sample covariance. In addition to PCA-like interpretation, interpretation of integrative analysis also focuses on the structure in which scores and loadings are shared between data matrices. Different methods make different kinds of assumptions on how structure is shared. Trivially, the two extremes, no shared structure or all

structure globally shared, correspond to making an individual analysis of each matrix or to concatenate all matrices and analyze the concatenated matrix. The first is a too weak assumption and would lead to methods that are unable to use or find patterns that are shared between matrices. The latter is a too strong assumption and would lead to methods that miss patterns that are specific to some matrices. Several methods (section 5.5.1, supplementary material to paper II) make the assumption that patterns are either individual or globally shared between all matrices, so that each matrix is approximated by a sum of individual components, globally shared components and noise. This is still a strong assumption, especially for the analysis of many matrices, as it is enough for patterns to be missing in only one matrix to force the patterns to be approximated individually for each matrix. CMF (section 5.5.2) and MM-PCA (section 5.5.3) can find patterns that are shared between subsets of matrices. Thus, these methods address the more difficult problem of finding the structure in which patterns are shared, as opposed to merely separating and estimating individual and globally shared structure. The structure in which scores and loadings are shared between matrices is not a binary relation where structure is either shared or not. The amount of structure that is shared is described both by the number of components that are shared and the weights of these components in proportion to weights of other components and noise.

The dimension reduction performed with integrative analysis is useful for making sense of high-dimensional data. It can be used for visual exploration, for example by making scatter plots in two or three dimensions. A decomposition of data into signal and noise allows for focusing on what is estimated to be noise-free signal. Methods that allow for missing data in data matrices can be used to impute the missing values. Clustering is closely related to dimension reduction. Joint and individual clustering (JIC) (Hellton and Thoresen, 2016) uses the results of JIVE by performing k-means clustering on the JIVE scores. This allows for integrative clustering, where clusters of observations can be found and associated with values of specific sets of variables or with all sets of variables, globally. Lee et al. (2010) showed how scores and loadings of sparse SVD, where solutions with exact zeros in scores and loadings are encouraged, can be used to perform bi-clustering of an individual data matrix. This is utilized in MM-PCA to perform integrative bi-clustering. Bi-clustering simultaneously clusters both variables and observations, such that elements that belong to the same combination of variable cluster and observation cluster are similar. Integrative bi-clustering is bi-clustering where clusters may be equal across multiple matrices or specific to one matrix.

5.4 The Euler parametrization

In MM-PCA, we use a parametrization of orthonormal matrices that has not previously been used to solve optimization problems stemming from integrative analysis. The parametrization, called generalized Euler parametrization (Hoffman et al., 1972), is defined in paper II. Here, an efficient algorithm for computing an orthonormal matrix given the generalized Euler parameters is presented, after a brief description of the parametrization. The objective function of MM-PCA is based on the objective function from the formulation of SVD as an optimization problem (section 2.1.2). When the SVD objective function is modified, as is done in MM-PCA, the optimal matrices U and V are no longer guaranteed to be orthogonal. Thus, orthogonality needs to be enforced. Orthogonality constraints are, however, difficult to handle efficiently in numerical optimization. Therefore, we use the generalized Euler parametrization, which allows us to formulate an unconstrained optimization problem that still optimizes over the space of orthonormal matrices. At the same time, the parametrization reduces the number of parameters of the optimization problem.

The generalized Euler parametrization expresses any orthonormal matrix $V \in \mathbb{R}^{p\times k}$, $p \geq k$, in $m = pk - k(k+1)/2$ parameters $\xi_1, \ldots, \xi_m$ as a product of m matrices $R_1(\xi_1), \ldots, R_m(\xi_m)$ and a matrix I_{pk}, i.e., $V(\xi) = R_1(\xi_1)R_2(\xi_2)\cdots R_m(\xi_m)I_{pk}$, where I_{pk} is the first k columns of the p-dimensional identity matrix. The matrices $R_i(\xi_i)$ are called Givens rotations and are matrices of the form

$$R_i(\xi_i) = \begin{bmatrix} I & 0 & 0 & 0 & 0 \\ 0 & \cos\xi_i & 0 & -\sin\xi_i & 0 \\ 0 & 0 & I & 0 & 0 \\ 0 & \sin\xi_i & 0 & \cos\xi_i & 0 \\ 0 & 0 & 0 & 0 & I \end{bmatrix},$$

where I and 0 are identity matrices and zero matrices of varying sizes, such that each Givens rotation affects a unique pair of dimensions.

Naively calculating V by matrix multiplication from left to right would require $(m-1)p^3 + p^2k$ scalar multiplication operations: $(m-1)$ matrix multiplications of $p \times p$-size matrices and one matrix multiplication of a $p \times p$-size matrix with a $p \times k$-size matrix. Using the definition of m, this is a time complexity in p of $\mathcal{O}(p^4)$. A simple improvement is to calculate V by matrix multiplication starting instead with the right-most matrix product. This requires mp^2k scalar multiplication operations, time complexity $\mathcal{O}(p^3)$, a substantial improvement since $p >> k$ in practice. Further major improvements can be achieved by utilizing the sparsity and structure of Givens rotation matrices. In the implementation of MM-PCA

the following algorithm is used (here in C-like pseudo-code):

```
for (j = k-1; j >= 0; j--) {
  for (i = p-1; i >= j+1; i--) {
    cx = cos(xi[i, j]);
    sx = sin(xi[i, j]);
    for (a = j; a < k; a++) {
      tmp = v[i, a];
      v[i, a] = -sx * v[j, a] + cx * v[i, a];
      v[j, a] = cx * v[j, a] + sx * tmp;
    }
  }
}
```

Variables `p` and `k` hold the values of p and k, `xi` is a $p \times k$-size matrix of parameters of which the diagonal and the upper triangular part are not used and `v` is initialized before the algorithm to be equal to I_{pk}. After the algorithm has been run, `v` holds the orthonormal matrix V. To my knowledge, this algorithm has not been presented before, nor have more efficient algorithms. The algorithm requires less than $4k^2p$ scalar multiplication operations, a time complexity of $\mathcal{O}(p)$.

Similar algorithms for the inverse of $V(\xi)$ and for the derivatives of the MM-PCA objective function, involving the derivatives of V, are used in the published implementation of MM-PCA.

5.5 Methods

A review of methods for integrative analysis is given in the supplementary material to paper II. Here, MM-PCA is briefly introduced along with two other methods which influenced the development of MM-PCA and several other methods for integrative analysis.

5.5.1 JIVE

Joint and individual variation explained (JIVE) (Lock et al., 2013) has gained some popularity in the field of systems biology. It addresses horizontal integrative analysis (one group of observations and an arbitrary number of groups of variables) and decomposes each data matrix into individual components,

components with globally shared scores and noise. The model is estimated iteratively by performing rank k SVD individually for each matrix without the global components, followed by estimating the k global components given the individual components. This is repeated until convergence. The rank k is selected using an approach based on permutation testing. The simplicity of the model eases interpretation, but the assumption that patterns are either individual or global is unreasonably inflexible when analyzing more than a few matrices.

5.5.2 Group-wise sparse CMF

Group-wise sparse collective matrix factorization (CMF) (Klami et al., 2014) addresses AMD, and can identify structure that is shared between subsets of data matrices. CMF specifies a Bayesian model with prior distributions for loadings, scores and noise. Each set of observations or variables has one matrix that encodes its scores or loadings in all data matrices concerning that set. A data matrix of relations between set i and set j is modeled as $X_{ij} = U_i U_j^T + \varepsilon_{ij}$, where U_i, U_j and ε_{ij} are matrices with normally distributed elements. A prior distribution called the automatic relevance determination (ARD) prior is used for scores and loadings. It causes some columns of each U to be active (non-zero) in only a subset of its associated data matrices. Thereby, a structure of shared patterns is estimated for the analyzed matrices.

The model of CMF can identify components that are individual to one data matrix, shared globally by all data matrices or shared by other subsets of data matrices. The model can, however, not identify *any* subset of data matrices with a shared component, as shown by the following counterexample. Consider four data matrices, two groups of variables and two groups of observations. Let the observation groups have indices 1 and 2, and the groups of variables have indices 3 and 4. Then, the four data matrices are X_{13}, X_{14}, X_{23} and X_{24}. We need only consider the norm of columns of the U-matrices and a model of rank one. Let u_i be the norm of the only column of U_i. If all u_i are non-zero, then the only component in this example is shared by all data matrices. The component can be specific to one data matrix, for example if u_1 and u_3 are non-zero while u_2 and u_4 are zero. It can be specific to the matrices associated with one group of variables or observations, for example if u_1 is zero while u_2, u_3 and u_4 are non-zero. It can, however, not be shared by three of the four data matrices, since no u_1, u_2, u_3, u_4 exist such that exactly one of the following products is zero: $u_1u_3, u_1u_4, u_2u_3, u_2u_4$.

5.5.3 MM-PCA

Multi-group and multi-view principal component analysis (MM-PCA) is presented in paper II, and summarized in section 6.2. In the paper, an objective function based on the singular value decomposition for each group is defined. Its optimum corresponds to low-rank orthonormal bases for the row or column space for each group of observations or variables. The generalized Euler parametrization (section 5.4) is used to reduce the number of parameters of the optimization problem and to eliminate the need for orthogonality constraints. Like CMF, MM-PCA addresses AMD and can identify structure that is shared between subsets of data matrices. In contrast to CMF, the MM-PCA model is specified as an objective function rather than with a Bayesian model. The objective function combines the loss (size of the approximation error) of the model with penalty terms that facilitate the use of MM-PCA and the interpretation of its results. One penalty term fills the same function as the ARD prior of CMF. Additional penalty terms encourage sparse loadings and scores, and perform selection of the rank of the model.

6 Summary of papers

This chapter summarizes the three papers included in this thesis. The problem that each paper addresses is described, and the proposed solutions are outlined.

6.1 Paper I

In paper I, we introduce the method resampling of penalized estimates (ROPE) for robust network modeling with false discovery rate (FDR) of edges controlled at a desired level. The use of network modeling to estimate genetic networks is hampered by estimation instability, due to a relatively small sample size, and a strong dependency on the level of regularization, which is difficult to select. With ROPE, these problems are addressed by the use of a statistical model for the number of times each edge is estimated to be present across bootstraps. Like stability selection and BINCO, our method uses bootstrap samples of data to produce multiple network estimates for several values of the regularization parameter. These estimates are aggregated to selection frequencies for all edges and simultaneously analyzed across all levels of sparsity. Unlike previous methods, this global modeling approach is based on a joint beta-binomial mixture of edge selection frequencies. The edge false discovery rate estimates are based on the regularization parameter value that best separates the mixture components ("true" and "false" edges) as well as information about the true level of sparsity obtained from a range of regularization levels. We show that ROPE outperforms state-of-the-art methods in terms of FDR control and robust performance across data sets. The evaluation is performed on simulated data sets and on glioblastoma tumor gene expression data from TCGA.

We propose a statistical model for selection counts, and enable a simultaneous interpretation of selection counts for different levels of regularization. The sequence $\{W_i^\lambda : i = 1, \ldots, p\}$ is modeled as coming from a mixture of beta-

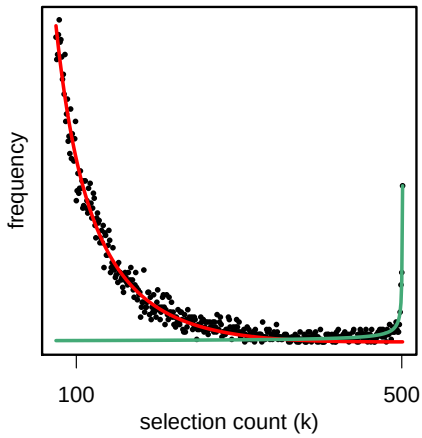

Figure 6.1: Edge selection counts histogram after 500 bootstraps corresponding to one regularization level λ. A mixture distribution with two components is estimated by ROPE. The red line shows the component that estimates the null hypothesis distribution. The green line shows the component that estimates the alternative hypothesis distribution. While only the null distribution is needed to estimate the FDR of a selection threshold, having a model that captures both populations decreases bias and avoids several model estimation difficulties.

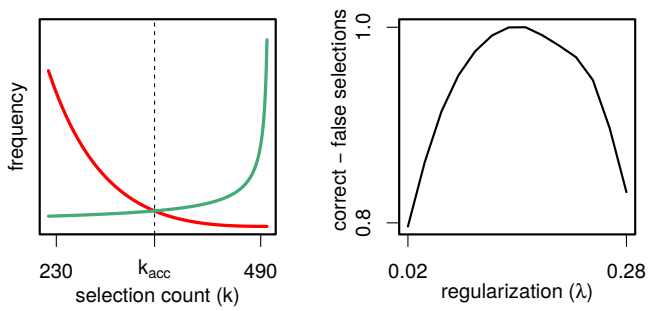

Figure 6.2: Illustration of procedure to choose which level of regularization to use for edge selection. The left panel shows the model fitted to a histogram for one level of regularization. It also shows k_{acc}, the selection count threshold that maximizes accuracy. Assuming the fitted model as truth, the right panel shows the difference between numbers of correctly and incorrectly selected edges. The difference has been normalized to have maximum 1. The procedure estimates how separated (non-overlapping) the two distributions are.

binomial distributions, with components capturing either the population of null edges or the population of alternative edges. Fitting this distribution makes it straight forward to choose a threshold $k_t \in [0, B]$ corresponding to a given false discovery rate such that edges i with $W_i^\lambda > k_t$ are declared significant. A range of regularization is chosen to minimize the overlap of mixture components. In this range, the ratio of alternative edges is constrained to be constant (for any regularization λ).

The mixture model

$$\begin{aligned} z|\pi &\sim \text{Bernoulli}(\pi) \\ y_j|\mu_j, \sigma_j &\sim \text{Beta}(\mu_j, \sigma_j),\ j = 0, 1 \\ W_i^\lambda|y, z &\sim \text{Bin}(B, y_z) \end{aligned}$$

is fitted to edge counts for each level of regularization λ. The model has five parameters: π the proportion of true edges, μ_1, σ_1 the mean and standard deviation of the probability of true edges to be selected and μ_2, σ_2 corresponding mean and standard deviation for false edges. An example of the two components of the mixture model can be seen in figure 6.1. This model is extended to allow for overdispersion, for details see paper I.

Using the fitted models for each λ, we estimate how separated the two components are. The estimate $g(\lambda)$ is based on the difference between the number of correctly and falsely selected edges, under the fitted model.

$$g(\lambda) = \sum_{k=k_{\text{acc}}}^{B} (f_a(k) - f_n(k))$$

where f_a and f_n are the estimated distributions for alternative and null edges respectively and k_{acc} is the threshold that maximizes accuracy, given these distributions (figure 6.2).

In contrast to BINCO, ROPE uses a statistical model for the whole range of selection counts while BINCO fits a curve only to the range where the frequency of selection counts is decreasing. The approach used in ROPE has the benefit that the subsequently estimated selection count threshold is always in the modeled range, while it is typically outside of the range that is used in BINCO. Thus, the threshold estimated with ROPE will not be based on an extrapolation, in contrast to that estimated with BINCO. Furthermore, BINCO requires an intermediate estimate of the range where the frequency of selection counts is decreasing, an extra step that increases the variability of estimated networks.

ROPE, BINCO and stability selection are evaluated with extensive simulation

studies under a range of different variable interdependence structures. Results show consistently far more correct FDR control for simulated problems, compared to BINCO and stability selection. In a scale-free network with 500 nodes and 200 observations (figure 2 in paper I), ROPE has an actual FDR that differs with at most 0.025 compared to the targeted FDR, across three targeted FDR levels (0.05, 0.1 and 0.15), the number of resamples ranging from 50 to 1000 and 20 random repetitions for each simulation setting. The network estimation accuracy of ROPE is in the range 0.6 to 0.8 depending on the FDR target. In the same settings, the achieved FDR of stability selection is around 0.025 regardless of the FDR target, resulting in an accuracy in the range 0.55 to 0.65 depending on the FDR target. Thus, stability selection is less accurate due to an edge selection that is more conservative than what is targeted. In the same settings, results of BINCO are highly unstable across random repetitions and are strongly affected by the number of bootstrap resamples that are performed.

The methods are also compared on public gene expression data from TCGA (The Cancer Genome Atlas Research Network et al., 2013). Selected network sizes and the difference between estimates for different subsets of the gene expression data suggest that BINCO fails to control FDR while stability selection is too conservative. Across 20 resamples, the agreement between networks is consistently about 0.9 for ROPE over a range of targeted FDR levels (figure 6 in paper I). For BINCO the agreement is between 0.75 and 0.6. For stability selection, an FDR of at least 0.25 needs to be targeted for the method to estimate a non-empty network, while the estimated number of edges increases linearly with the FDR target for ROPE. In the network selected by ROPE at an estimated FDR of 0.15, we found all hub genes to have documented cancer-related functions.

Lastly, in the supplementary material to paper I, we apply ROPE for classification of gene expression profiles according to their primary cancer type, illustrating that ROPE can also be applied to some variable selection problems other than graphical models. There, a multinomial logistic regression model with group lasso penalty is utilized.

6.2 Paper II

In paper II we propose the method multi-view and multi-group principal component analysis (MM-PCA) for integrative analysis of several related data matrices. The method is based on ordinary singular value decomposition (SVD), just like principal component analysis (PCA), but a novel penalty term promotes equality between singular vectors of different matrices. Existing similar methods

can be divided into two groups. First, O2-PLS (Trygg, 2002), DISCO (Schouteden et al., 2013), JIVE (Lock et al., 2013) and other methods (see paper II) are based on iteratively performing SVD on all matrices together and on the individual residuals of each matrix. The methods model data as a combination of globally joint information and information that is specific to each individual matrix. They cannot capture information that is joint between an arbitrary subset of matrices, and they are less flexible in the types of matrix relations that they allow for. Second, Bayesian methods, such as CMF (Klami et al., 2014) and MOFA (Argelaguet et al., 2018), have a flexibility similar to MM-PCA, but their results are difficult to interpret since they, in contrast to MM-PCA, lack penalties that induce sparsity, achieve variable selection and/or achieve model rank selection.

The most general set of relations that can be handled with MM-PCA was termed augmented multi-view data (AMD) by Klami et al. (2014). Data matrices in general relate the set of entities that is associated with the rows of the matrix to the set of entities that is associated with its columns. The generality of AMD allows MM-PCA to integratively analyze sets of entities and data matrices where each matrix captures a relation between any two of the sets of entities. Figure 1D in paper II shows an example and compares it to less general settings.

The MM-PCA solution is given by the lower-triangular matrices ξ_i, one for each set of entities, and the diagonal matrices D_i, also one for each set of entities, that minimizes the loss function

$$\sum_{(i,j)\in\mathcal{S}} ||X_{ij} - V(\xi_i)D_iD_jV(\xi_j)^T||_F^2 + \sum_{c=1}^{4} \lambda_c \mathrm{P}_c(\xi, D), \tag{6.1}$$

where $\mathcal{S}$ is the set of pairs of entity sets that are related, X_{ij} is the data matrix that relates entity set i with entity set j, $V(\cdot)$ is the function from angle-matrix to orthonormal matrix, ξ_i is the lower-triangular matrix of angles for entity set i, D_i is the diagonal matrix of norms for entity set i, $||\cdot||_F$ is the Frobenius norm, λ_c is the penalty parameter for penalty c and $P_c(\cdot,\cdot)$ is penalty function c. The first sum constitutes the error, i.e. the difference between data and the MM-PCA low-rank approximation of the data. The second sum consists of penalty functions, described below, that promote solutions that are more easily interpreted. The terms of the first sum are inspired by the formulation of truncated SVD as an optimization problem.

The four penalty functions in (6.1) promote 1) data integration, 2) low-rank approximation, 3) sparsity of loadings and scores and 4) variable selection. The promotion of these properties is aimed at achieving an interpretable model of the data. The function $V(\cdot)$ is the Euler parametrization discussed in section 5.4. It

is a parametrization of orthonormal matrices (matrices with pairwise orthogonal columns of unit Euclidean norm). The Euler parametrization decreases the number of parameters in the optimization problem (6.1) and eliminates the need to explicitly enforce orthonormality in it.

The optimal matrices ξ_i and D_i, $i = 1, \ldots, n_v$, (where n_v is the number of entity sets) hold a condensed encoding of the data. First, interpretation can be focused on each matrix individually. Like in PCA, the decomposition of X_{ij} into a low-rank approximation $V(\xi_i)D_iD_jV(\xi_j)^T$ expresses the data in terms of rank-one components which are composed of scores and loadings. Secondly, the exact zeros on the diagonals of matrices D_i, $i = 1, \ldots, n_v$, holds information of which scores and loadings that are shared between data matrices. A zero element on the diagonal of D_i means that the associated scores or loadings do not participate in the approximation of a matrix X_{ij}. That set of scores or loadings is thus shared among the other matrices $X_{\cdot j}$.

SVD can be used for clustering, by dividing observations (or variables) according to the signs of the associated scores (or loadings). Hierarchical clustering is achieved by dividing, first, based on the most important component, and then based on the following components in order of importance. Since this clustering can be done for both the rows and the columns of a matrix simultaneously, it can be used for so-called bi-clustering. By combining this method of clustering with information of which scores and loadings that are shared between data matrices, MM-PCA can be used to perform integrative bi-clustering. That is, bi-clustering where some clusters are equal across several matrices.

The method is evaluated in three simulation studies and its use is demonstrated in an analysis of gene expression and methylation data from cancer patients. The first simulation study focuses on the ability to find loadings that are shared among a subset of four matrices. It shows a better performance for MM-PCA than for CMF in terms of finding which loadings that are shared among which matrices. The second simulation study is simpler, in order to enable comparison with the popular method JIVE. The goal is to find scores that are common to all matrices, in the presence of noise and scores that are individual to each matrix. A range of settings, where the size of data matrices and the strength of the shared scores are varied, are studied. The experiment shows a consistently good performance for MM-PCA, while CMF and JIVE perform better in some settings and worse in other settings. The third simulation study demonstrates the ability of MM-PCA to estimate the correct rank and the correct exact zeros in scores and loadings, in the presence of noise. In the majority of simulations, MM-PCA finds the correct rank while CMF consistently overestimates it. Accuracy in correctly finding the non-zero positions of scores and loadings is also higher for MM-PCA compared to CMF. The integrative analysis of genomic data

demonstrates the use of MM-PCA in a realistic setting. It distinguishes several components that correlate significantly with clinical parameters, which were not used as input to the analysis. For example, it finds components that separate normal tissue from tumorous tissue and that separate female patients from male patients. A non-trivial structure of components shared between different subsets of matrices is revealed.

The supplementary material to paper II contains an extensive literature review and a deeper study of the Euler parametrization. Integrative analysis of several data matrices has been studied in several scientific fields. The review relates and contextualizes methods proposed in chemistry, in genomics and in data science, where the concept is often called recommender systems. A proof is given that the Euler parametrization, $V(\cdot)$, parametrizes all orthonormal matrices, and only orthonormal matrices. The inverse of the Euler parametrization is also supplied.

6.3 Paper III

In paper III we propose structure-adaptive stability selection (SASS), a method that enables the incorporation of structural assumptions in stability selection-based network estimation. In high-dimensional genomics, assumptions are needed to make network estimation feasible. Network sparsity is a very common assumption. However, the sparsity assumption in popular methods is applied independently to each potential edge in the network, leading to an implicit structural assumption that the probability for each edge to exist is independent of the existence of other edges. This implicit assumption is not feasible for biological networks, and thus negatively impacts the estimation of such networks. With SASS, we aimed to enhance stability selection, a method for stable network estimation, with the biologically feasible structural assumption that networks have a community structure with higher edge density within communities than between communities.

SASS is based on, first, repeatedly estimating networks based on random subsets of the available data (similarly to ROPE, paper I). Next, a method for community detection is repeatedly applied to the estimated networks, in order to find a stable estimate of network communities. As the third step, two separate edge selection count histograms are used: one for pairs of nodes that are estimated to be members of the same community and one for pairs of nodes that are not. These histograms are used to estimate the influence of the sparsity assumption on the network structure and to compensate for the structural bias it causes. In the paper, we also propose extensions to SASS to 1) estimate

the strengths of network connections (not only the binary decision of which edges that are present) and to 2) make a stable estimate of the hierarchical community structure of the estimated network.

The method is evaluated in simulation experiments and in the estimation of a genetic regulatory network from gene expression data. The simulation experiments show that SASS better estimates networks for which the assumption of community structure is correct (accuracy 0.825), compared to another method that makes a similar community assumption (accuracy 0.481) and compared to stability selection (accuracy 0.817). The improvement is both in terms of accuracy and in terms of structural bias. When the assumption of community structure is incorrect, the cost of making the assumption is reduced by SASS often being able to detect the lack of community structure. The lack of community structure is detected 73 or 82 times out of 100 depending on the underlying method used for network estimation. For estimation of the regulatory network in human cancer patients we show an increased overlap with a manually curated and peer-reviewed database of gene-gene interactions in 81 out of 100 repetitions on random subsamples of the data, compared to stability selection.

The proposed method adds the ability to make biologically relevant structural assumptions to stability selection, a widely used method for network estimation. The improved performance on gene expression data suggests both usefulness of SASS and biological relevance of the assumption of community structure. This enhancement of stability selection may improve the understanding of high-dimensional genomic data.

7 Software packages

Implementations of the methods presented in this thesis are publicly available, two of them as R software packages. The additional work to prepare software to be conveniently useful for other scientists is important both to enable the use of the methods in applied research and to enable evaluation of the methods by other statisticians. The following sections briefly describe the software packages, how they were implemented and how they can be used.

7.1 Model selection with FDR control of selected variables

An implementation of the method ROPE is made available as an R package. The package gives support in choosing a regularization range, using visualizations and a heuristic for automatically deciding if histograms are U-shaped. The statistical model is fitted at each regularization step using numerical optimization of the log-likelihood function. In a second round of fitting the model, information from the optimal regularization range is used to make an estimate of mixture component sizes, based on counts from several regularization levels. The package contains several visualizations to examine the goodness of model fit. The package is available at The comprehensive R archive network `https://cran.r-project.org/package=rope`.

The main function in the package is called `rope`, which performs all steps of the method. Given a matrix of variable selection counts (one column for each variable and one row for each penalization level) it computes variable selections at the requested FDR levels and q-values (section 2.1.1) for each variable. The function `rope` is accompanied by a few auxiliary functions. The function `explore` fits the mixture model for each penalization level separately in order to facilitate the choice of penalization range, without needing to run the entire

ROPE procedure. The functions `ropegraph` and `exploregraph`, corresponding to `rope` and `explore`, allow for input in the form of adjacency matrices, which may be more convenient for some users. The function `plotrope` facilitates visualization of the output from `rope` (or `ropegraph`).

The ROPE procedure includes fitting a mixture model to the selection counts for each penalization level twice: first separately and then with a constraint on the size of the mixture components across all penalization levels. The same internal function is used both times. It includes the implementation of the log-probability mass function of the mixture model and uses the standard R optimizer and its implementation of the L-BFGS-B optimization method. L-BFGS-B is a modification of the quasi-Newton method Broyden-Fletcher-Goldfarb-Shanno that allows for box-constraints. Constraints are needed due to the log-likelihood not being defined for arbitrary parameters and to avoid numerical instability. The function `isoreg`, included in R, is used to perform monotonously increasing, non-parametric regression, in order to make a conservative estimate of the proportion of edges in the alternative component (corresponding to truly existing edges) of the mixture model. Finally, the package includes code to interpret the mixture model parameters in terms of q-values and FDR-controlled variable selections.

The following is an example in the R programming language of how to use the software package. The example assumes that `x` is a data matrix with one column per variable and one row per observation, and that `net_est` is a function that outputs a network estimate in the form of an adjacency matrix. First, bootstrap is used to estimate 500 networks.

```
lambda <- seq(0.05, 0.5, 0.025)
B <- 500
n <- nrow(x)
p <- ncol(x)
W <- lapply(lambda, function(l) matrix(0, p, p))
for (i in 1:B) {
  bootstrap <- sample(n, n, replace=TRUE)
  for (j in 1:length(lambda)) {
    selection <- net_est(cov(x[bootstrap, ]), lambda[j])
    W[[j]] <- W[[j]] + selection
  }
}
```

Then, selection counts are input to the function `ropegraph` and the estimated q-values are used to make a variable selection at the FDR level 0.1.

```
result <- rope::ropegraph(W, B)
selected_edges <- result$q < 0.1
```

The matrix `selected_edges` now is an adjacency matrix for the estimated network. The documentation included in the package contains further usage details.

7.2 Integrative analysis of several related data matrices

An implementation of the method MM-PCA is made available as an R package. The package finds a set of components that approximate the given data matrices, using numerical optimization. The package can also select values for the penalty parameters by performing cross-validation. The package is available at The comprehensive R archive network `https://cran.r-project.org/package=mmpca`.

The package has one function, `mmpca`, which takes three mandatory arguments. First, it takes a list of matrices that hold the data to analyze. Second, an integer matrix of width two describes how the data matrices are related to each other. For each data matrix in the list, the corresponding row in the integer matrix gives the index of entity set that is associated with the rows (first column) and columns (second column) of the data matrix. Third, an integer gives the maximum allowed number of components in the estimated approximation of the given data matrices. There are also several optional arguments, that are described in detail in the documentation that is included in the software package. The optional arguments can be used to limit the hyperparameter search space, to enable parallelized computation or to enable caching of partially finished computations.

The analysis of several high-dimensional matrices is a highly demanding task computationally. Therefore, the central part of the computation, the numerical optimization, was implemented in the programming language C++, which is more efficient for iterative computation than R. The C++ code is compiled and made available for use within the MM-PCA R software package. Within C++, the implementation of the Broyden-Fletcher-Goldfarb-Shanno algorithm for quasi-Newton optimization in the GNU Scientific Library (Galassi and Gough, 2009) was used. The MM-PCA objective function and its gradient (including the Euler parametrization, penalty functions, the loss function and their gradients) were implemented in C++ using Eigen (Guennebaud et al.,

2010), a library for linear algebra. Functionality not used within optimization, and therefore less critical for computational efficiency, was implemented in R. This includes code for heuristic choice of initial values for numerical optimization, handling of missing values in the input data, calculating the inverse of the Euler parametrization and performing cross-validation. The mathematical details of the implementation are given in the supplementary material to paper II.

The following example of the use of `mmpca` shows an analysis of six data matrices. Their configuration is the same as in figure 5A in paper II. There are two kinds of measurements: gene expression and methylation. There are three groups of patients defined by the kinds of data that are available for each patient. Matrix `e1` contains gene expressions for patient group 1. Group 1 consists of patients for which expression data are available but not methylation data. Matrices `e2` and `m2` contain gene expressions and methylation measurements, respectively, for group 2. Group 2 consists of patients for which both kinds of data are available. Matrix `m3` contains methylation data for group 3. Group 3 consists of patients for which only methylation data are available. Finally, matrices `c12` and `c23` contain similarities between groups 1 and 2 and groups 2 and 3, respectively. The similarities are in the form of a priori covariance matrices. All matrices are assumed to have been preprocessed to have zero mean and variances of comparable magnitudes. First, all data matrices are added to a list `x` and the integer matrix `inds` is constructed to encode the relationships between the data matrices.

```
x <- list(e1, e2, m2, m3, c12, c23)
inds <- rbind(c(1, 4),
              c(2, 4),
              c(2, 5),
              c(3, 5),
              c(1, 2),
              c(2, 3))
```

Next, `mmpca` is called. The maximum rank is set to 40. If cross-validation does not estimate the rank to be less than 40 it is advisable to call the function again with a higher maximum rank.

```
result <- mmpca::mmpca(x, inds, 40)
```

A list of estimated loading matrices is now available in `result$solution$V` and the associated norms are available in `result$solution$D`.

8 Conclusion

This work has contributed three practically useful new statistical methods for the analysis of high-dimensional genomic data, along with software implementations. In paper I we show that the method ROPE outperforms state-of-the-art methods in terms of FDR control and in terms of robust performance across a range of simulation settings. It does so by using a novel statistical model for edge selection counts. The method accurately estimates the trustworthiness of each individual estimated network edge. In a set of gene expressions from cancer tumors, the method finds several connections that are known to have relevance for cancer progression. It is, thus, illustrated how ROPE can be used for principled model selection in order to find genomic associations to study further, in search of regulatory interactions.

In paper II we propose MM-PCA, a method for integrative analysis that allows for structure in data to be shared across subsets, unknown beforehand, of analyzed data matrices. A review of previously existing methods for integrative analysis based on low-rank matrix factorization is contributed in the supplementary material to paper II. Compared to existing methods, MM-PCA improves interpretability by imposing sparsity and facilitating the choice of model complexity. The imposed sparsity also enables interpretation in terms of an integrative bi-clustering of the analyzed data. In terms of method development, paper II makes two key contributions. First, it introduces a new use for the Euler parametrization, which has not previously been used to solve optimization problems stemming from integrative analysis. Secondly, it introduces a framework based on penalized optimization for the integrative analysis of high-dimensional data with complex interactions. Such integration has previously only been addressed with Bayesian methods. In a set of gene expression and methylation data from cancer tumors, MM-PCA finds scores and loadings that correlate with several relevant clinical features without having access to these features. In integrative genomics, such components shared between different data types and groups of observations can have relevance for the development

of new therapeutic solutions.

In paper III we propose a way to incorporate the biologically relevant assumption that regulatory genetic networks have a community structure, with stability selection. In a set of gene expression data from cancer tumors, the proposed method, SASS, estimates networks with a higher overlap with a human-curated database of gene interactions, compared to stability selection. Thus, the method increases the potential to gain understanding from high-dimensional data sets, such as from large-scale genomics.

For the methods presented in this thesis, practical usefulness for gaining biological understanding and the ability to facilitate the complexity of real data have been prioritized over other properties that would also be desirable. In particular, the methods are computationally heavy. Although fast solutions from e.g. closed-form expressions are helpful in research, there is a great availability of computational power both in personal computers and in computer clusters commonly available at research facilities. Thus, computer-intensiveness is a cost often worth paying when it enables a deeper understanding of more complex data. Likewise, in order to present a deeper theoretical understanding of the methods, in terms of e.g. the statistical distributions of all estimators, it would have been required to greatly simplify the methods and to limit the richness of data that can be analyzed with them. The modeling assumptions that have been made – most prominently the assumption that data have a Gaussian distribution after transformations and the assumption of independence of edge selection counts – have been evaluated to ensure that they do not substantially worsen estimates in real or realistic settings.

In summary, the work in this thesis has contributed three novel and useful methods for the exploration and understanding of high-dimensional genomic data. The focus of the methods span from the local (pairwise interactions) to whole data matrices (network structure and principal components) and further on to structure that is common between several data matrices. Such study of genomic data may help to devise new tools for treatment and diagnosis of cancer and other diseases, both by hypothesis generation in general and, more specifically, by the identification of unknown molecular interactions that are causing diseases or symptoms. After experimental validation, such interactions could, for example, be targeted in disease treatment or lead to new biological indicators of disease progression.

Bibliography

Albert, R. (2005). Scale-free networks in cell biology. *Journal of Cell Science*, 118(21):4947–4957.

Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., Buettner, F., Huber, W., and Stegle, O. (2018). Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*, 14(6):e8124.

Banerjee, O., Ghaoui, L. E., and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine learning research*, 9(Mar):485–516.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Publishing Company, Incorporated, 1st edition.

Crick, F. (1970). Central dogma of molecular biology. *Nature*, 227(5258):561–563.

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

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863–14868.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Galassi, M. and Gough, B. (2009). *GNU Scientific Library: Reference Manual*. GNU manual. Network Theory.

Golub, G. and Van Loan, C. (1996). *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press.

Guennebaud, G., Jacob, B., et al. (2010). Eigen v3. http://eigen.tuxfamily.org.

Hao, D., Ren, C., and Li, C. (2012). Revisiting the variation of clustering coefficient of biological networks suggests new modular structure. *BMC Systems Biology*, 6(1):34.

Hastie, T. J., Tibshirani, R. J., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer series in statistics. Springer, New York.

Hellton, K. H. and Thoresen, M. (2016). Integrative clustering of high-dimensional data with joint and individual clusters. *Biostatistics*, 17(3):537–548.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Hoffman, D. K., Raffenetti, R. C., and Ruedenberg, K. (1972). Generalization of Euler Angles to N-Dimensional Orthogonal Matrices. *Journal of Mathematical Physics*, 13(4):528–533.

Hwang, B., Lee, J. H., and Bang, D. (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental & molecular medicine*, 50(8):1–14.

Jörnsten, R., Abenius, T., Kling, T., Schmidt, L., Johansson, E., Nordling, T. E. M., Nordlander, B., Sander, C., Gennemark, P., Funa, K., Nilsson, B., Lindahl, L., and Nelander, S. (2011). Network modeling of the transcriptional effects of copy number aberrations in glioblastoma. *Molecular Systems Biology*, 7:486.

Klami, A., Bouchard, G., and Tripathi, A. (2014). Group-sparse Embeddings in Collective Matrix Factorization. In *Proceedings of International Conference on Learning Representations (ICLR) 2014*.

Kling, T., Johansson, P., Sánchez, J., Marinescu, V. D., Jörnsten, R., and Nelander, S. (2015). Efficient exploration of pan-cancer networks by generalized covariance selection and interactive web content. *Nucleic Acids Research*.

Lee, M., Shen, H., Huang, J. Z., and Marron, J. S. (2010). Biclustering via Sparse Singular Value Decomposition. *Biometrics*, 66(4):1087–1095.

Li, S., Hsu, L., Peng, J., and Wang, P. (2013). Bootstrap inference for network construction with an application to a breast cancer microarray study. *Ann. Appl. Stat.*, 7(1):391–417.

Lock, E. F., Hoadley, K. A., Marron, J. S., and Nobel, A. B. (2013). Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *The Annals of Applied Statistics*, 7(1):523–542.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34(3):1436–1462.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473.

Monti, S., Tamayo, P., Mesirov, J., and Golub, T. (2003). Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. *Machine Learning*, 52(1):91–118.

Pe'er, D. and Hacohen, N. (2011). Principles and strategies for developing network models in cancer. *Cell*, 144(6):864–873.

Rice, J. A. (2006). *Mathematical Statistics and Data Analysis.* Belmont, CA: Duxbury Press., third edition.

Richardson, S., Tseng, G. C., and Sun, W. (2016). Statistical Methods in Integrative Genomics. *Annual Review of Statistics and Its Application*, 3(1):181–209.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–470.

Schouteden, M., Van Deun, K., Pattyn, S., and Van Mechelen, I. (2013). SCA with rotation to distinguish common and distinctive information in linked data. *Behavior Research Methods*, 45(3):822–833.

Shah, R. D., Samworth, R. J., and Samsworth, R. J. (2013). Variable selection with error control: another look at stability selection. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 75(1):55–80.

Smith, J. M. and Szathmary, E. (2000). *The origins of life.* Oxford university press.

Storey, J. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445.

Tan, K. M., London, P., Mohan, K., Lee, S.-I., Fazel, M., and Witten, D. (2014). Learning Graphical Models with Hubs. *J. Mach. Learn. Res.*, 15(1):3297–3331.

The Cancer Genome Atlas Research Network, Weinstein, J., Collisson, E., Mills, G., Shaw, K. M., Ozenberger, B., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J. (2013). The cancer genome atlas pan-cancer analysis project. *Nat Genet*, 45(10):1113–1120.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Trygg, J. (2002). O2-PLS for qualitative and quantitative analysis in multivariate calibration. *Journal of Chemometrics*, 16(6):283–293.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1):57–63.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *J. Mach. Learn. Res.*, 7:2541–2563.