THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Network models with applications to genomic data: generalization, validation and uncertainty assessment

José Sánchez

Division of Mathematical Statistics
Department of Mathematical Sciences
Chalmers University of Technology and the University of Gothenburg
Göteborg, Sweden 2014

Network models with applications to genomic data: generalization, validation and uncertainty assessment
José Sánchez
ISBN 978-91-7597-112-4

Doktorsavhandlingar vid Chalmers tekniska högskola
Ny serie nr 3793
ISSN 0346-718X

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

Typeset with LaTeX.
Printed in Göteborg, Sweden 2014

Network models with applications to genomic data: generalization, validation and uncertainty assessment

José Sánchez
Division of Mathematical Statistics
Department of Mathematical Sciences
Chalmers University of Technology and the University of Gothenburg

Abstract

The aim of this thesis is to provide a framework for the estimation and analysis of transcription networks in human cancer. The methods we develop are applied to data collected by The Cancer Genome Atlas (TCGA) and supporting simulations are based on derived models in order to reflect real data structure. Nevertheless, our proposed models apply to network construction for any data type. The thesis includes four papers, all of them adressing different aspects of network estimation.

Statistical analysis of high-dimensional data requires regularization. Network model validation amounts to selection of regularization parameters which control sparsity and, possibly, some common structure across different data classes (here, types of cancer). In paper I we present a bootstrap-based method to perform sparsity selection and robust network construction. We show, by simulation studies, that our proposed methods select sparsity to control false positive rate, rather than match the size of the true underlying network.

In paper II we address the problem of uncertainty in network estimation. Since network estimation is very unstable, uncertainty is an important issue to focus on, in order to avoid overintepretation of results. Using ideas from information theory, we introduce a method that assesses uncertainty by presenting a set of network candidate estimates, rather than a single network model. The method enables us to show that different network topologies have different estimation properties, and that each network estimation method's performance depends on this topology.

It is often of interest to identify and study the commonalities and differences in network estimates across several classes (here, types of cancer) and data types. Statistical network models, like the graphical lasso, provide a framework in which several classes and data types can be integrated. Paper III makes use of such framework and presents a method that allows for large scale sparse inverse covariance estimation of several classes. Through application of priors, we account for plausible connections across different data types. The proposed method also encourages the expected modular structure of biological networks and corrects for unbalanced sample sizes

across classes. The estimated networks are part of a publicly accessible resource termed Cancer Landscapes (`cancerlandscapes.org`), which provides a setting for interactive analysis in relation of pathway and pharmacological databases, diagnoses, survival associations and drug targets.

Traditionally, the analysis of genomic data has aimed for the study of differential expression. In paper IV we propose a way to integrate differential expression analysis with network estimation. To that end we extend upon existing methods in order to jointly estimate sparse mean vectors and precision matrices across several classes, thus gaining over analyses that focus on one or the other. Additionally, by assuming a block diagonal structure in the precision matrices, the problem can be recast into an ensemble classifier where each block becomes part of either a linear or a quadratic discriminant function.

Keywords: Inverse covariance matrix, precision matrix, graphical models, high-dimension, low-sample, networks, sparsity, fused lasso, elastic net, cancer, TCGA pan cancer analysis, online resource, discriminant analysis, classification.

Acknowledgments

I have heard more than once that writing a PhD thesis can be a very solitary process, but I never felt that way. I want to thank my supervisor Rebecka Jörnsten for always being present and offering all the support I needed. Never before in my life have I exchanged so many ideas with so many bright people. For this and for the most enjoyable collaboration I thank my co-supervisor, Sven Nelander, and my co-authors Teresia Kling, Patrik Johansson, Tobias Abenius, Tatjana Pavlenko and especially Alexandra Jauhiainen – this thesis would have never seen the light of the day if it was not for you!

In these five years I have learnt many things. Especially how little I know. I want to thank Olle Nerman, Holger Rootzén, Serik Sagitov, Aila Särkkä, Sergey Zuyev, Mats Rudemo, Olle Häggström, Staffan Nilsson, Torgny Lindvall, Anastassia Baxevani, Erik Broman and the rest of the faculty at the Mathematical Sciences department for helping me to learn not only statistics, but how to see life with the eyes of a scientist.

Thanks to my fellow PhD students, past and present, Sofia Tapani, Magnus Röding, Krzysztof Bartoszek, Ottman Cronie, Daniel Ahlberg, Anton Muratov, Mariana Pereira, Henrike Häbel, Viktor Johnsson, Fredrik Boulund, and Alexey Lindo. Talking probability and statistics with you is always fun, specially when it is done over a large glass of beer!

I came to Sweden about seven years ago. It was not at all hard to adapt to my new country, particularly with the help of my bästisar Patrik and Johan. Thanks for welcoming me in Göteborg and for being there even if you have not seen much of me these last months.

And about seven years ago I left Mexico. Despite the distance and the long periods with not seeing each other, my family has always been there. I thank you all, for always supporting me to do what I want and be what I want. Fortunately, we all in Mexico have an extended family, mine is made of all the friends that I have made through life. Thanks to you all, those who are here and the one that left too soon. None of this would have been possible without you.

I must have forgotten, with probability strictly larger than one, to mention other important people. Thanks to all of you who have been part of this work.

Te quiero, con limón y sal.

José Sánchez
Göteborg, November 2014

List of Papers

This thesis includes the following papers:

I **Sánchez, J.**, Jauhiainen, A., Jörnsten, R., 2014. Sparsity selection and robust network estimation via bootstrap.

II Jauhiainen, A., **Sánchez, J.**, Jörnsten, R., 2014. Can I trust my network? Local network resolution elucidates uncertainty in estimation with different methods and across network topologies.

III Kling, T., Johansson, P., **Sánchez, J.**, Marinescu, V.D., Jörnsten, R., Nelander, S., 2014. Exploration of pan-cancer networks by generalized covariance selection and interactive web content. Submitted.

IV **Sánchez, J.**, Pavlenko, T., Jörnsten, R., 2014. Joint estimation of sparse inverse covariance and mean with applications to discriminant analysis.

Paper not included in this thesis

Abenius, T., Jörnsten, R., Kling, T., Schmidt, L., **Sánchez, J.**, Nelander, S., 2012. System-Scale Network Modeling of Cancer Using EPoC. Goryanin, I., Goryachev, A., Eds. *Advances in Systems Biology*. Advances in Experimental Medicine and Biology 736, Springer.

Contents

Chapter 1

Introduction

1.1 Background

ATTAGCACCCATATTAGCCTGATTTTTGAA. How is life encoded? Long sequences of four nucleotides; known as adenine, thymine, cytosine and guanine; form the DNA molecule and contain the inherited instructions to build and operate a living organism. For a human being, these instructions are enclosed in 23 chromosomes, each one containing between 48 to 250 million-long nucleotide sequences, adding up to 3.2 billion letters. An average book contains about half a million characters (including blank spaces).

Analyzing and understanding this code is a challenging task. Much has been done since the first organism, the *Bacteriophage MS2*, had his instruction book written or, as we more precisely say today, was sequenced (Fiers et al., 1976). Yet, a lot of work remains to be done.

In an oversimplified picture, the nucleotides in a DNA sequence are grouped into units called genes, which in turn form groups coding the instructions on how to carry out the complex tasks inside a cell. Studying how the genes interact to form these groups is one of the many ways to try to decipher and understand the instructions book.

Genes' interactions can be thought of as a network, with the genes as nodes and edges representing the nature of their interaction, if any. The problem then becomes to find (estimate) these edges using data collected from the instruction books from many individuals. Estimating networks is not a new problem; it is indeed a common one in different disciplines (engineering, computing, social sciences) and a number of methods to solve it are available.

In this thesis work we focus on the statistical estimation of genetic networks

and some of the challenges involved in the process, such as assessing uncertainty and the quality of the estimates.

1.2 Gene Regulatory Networks

A gene regulatory network is a description of how genes interact with each other to form modules and carry out cell functions. These networks can help us, by characterizing the implied dependencies for the genes, in systematically understanding complex molecular mechanisms for certain biological processes. In this thesis we have a particular interest in genes as disease drivers and model their interactions using genetic data for different types of cancers. Recent analysis has shown genes in a particular network topology, hub genes, to be possible disease drivers, identifying them as key tumorigenic genes (Kendall et al., 2005; Mani et al., 2008; Nibbe et al., 2010; Slavov and Dawson, 2009).

We study such gene interactions through network estimation of different genomic data and several cancer classes (see Chaper 1). We focus on different problems that arise in network estimation for high-dimensional data, such as regularization, joint estimation, parameter estimation uncertainty and model validation. To this end we present novel methods and describe them in the four papers included in this thesis. We also make use of well known statistical tools, which we describe briefly in the chapters preceding the papers.

In Chapter 2 we introduce the genetic alterations and genomic data we work with. Chapter 3 contains a review of network estimation methods, in particular methods based on partial correlation. Once a network estimate is available, there are a number of tools that can be used for its analysis; we present therefore in Chapter 4 a number of graph theoretical concepts and measures that are helpful in summarizing the information contained in a network. Uncertainty in estimation and model validation are the least addressed topics in the network estimation literature, and one of our main interests. Papers I and II study this problem and make use of generalized linear models and rate-distortion theory, which we summarize in Chapter 5.

Chapter 2

Genetic alterations and genomic data

2.1 Cancer systems biology

The study of cancer systems biology involves investigating alterations that occur at a molecular level. This is a relatively new scientific area, but is well known that cancer is caused by genetic anomalies. Cancer systems biology aims thus to increase our understanding on the effect this genetic alterations have on an organism, namely, the uncontrolled growth of cancerous cells and formation of tumors (Pe'er and Hacohen, 2011; Jörnsten et al., 2011).

The central dogma of molecular biology (see Figure 2.1) describes the flow of information within a biological system and states that the transfer of information from a protein to either DNA or RNA is not possible. Following the dogma, a framework for the study of complex biological processes, such as cancer at a molecular level, can be established.

Even though most of the cells in an organism contain its entire genome, at any given time only a subset of the genes are active (expressed). Information in the genome can thus be quantitatively studied through the genes' expression levels. Expression is measured, using for example DNA microarray technology, as the relative quantities of messenger RNA (mRNA).

Complementary information can be obtained studying micro RNA (miRNA), which are small RNA molecules that can decrease (downregulate) or entirely suppress the expression of one or more genes.

As opposed to the more traditional reductionistic approach of biological and

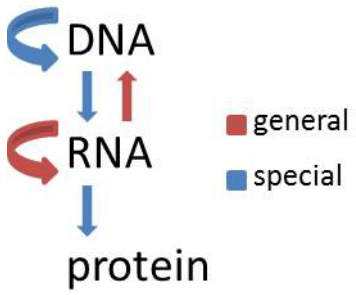

Figure 2.1: **Central dogma of molecular biology.** Blue arrows indicate general transfers of information (believed to occur in most cells). Red arrows indicate special transfers (known to occur under specific conditions such as lab experiments).

biomedical research, the systems biology one has an interdisciplinary and holistic perspective. Integration of different data types is therefore an important aspect of systems biology in general, and it is where mathematical models come to use. Gene network modelling, for example, has proved helpful in integrating several levels of genomic cancer data and addressing some important problems such as (Adler et al., 2006; Akavia et al., 2010; Garraway et al., 2005; Peng et al., 2010): (i) identification of genes with altered copy number as disease drivers; (ii) construction of features, based on molecular data, for prediction of patient survival, and (iii) discovery of possible therapeutic targets based on matching hubs in the networks to pharmacological databases.

In the following sections the genetic alterations and cancer types studied in this thesis are described briefly. We also list some of the biggest consortia working on genomic data recollection.

2.2 Genetic Alterations

The genome can undergo different alterations which are measured with raging techniques, producing in turn different data types. Below we briefly describe some of the genomic alterations that are of special interest in cancer systems biology.

- **Single Nucleotide Variants.** These are point mutations in the DNA sequence, occurring when a single nucleotide (A, T, C or G) differs between members of a pair of chromosomes.

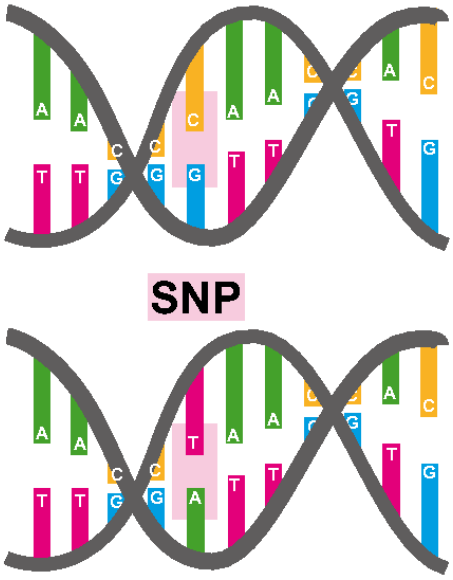

Figure 2.2: **Single Nucleotide Variants.** The two molecules of DNA differ at the highlighted base-pair location (a C/T polymorphism).

- **Copy Number Alterations.** CNA occur when the cell has an abnormal number of copies of a certain part of the DNA, sometimes of an entire gene. These can be deletions, corresponding to fewer copies than normal; or duplications, corresponding to more copies than normal.

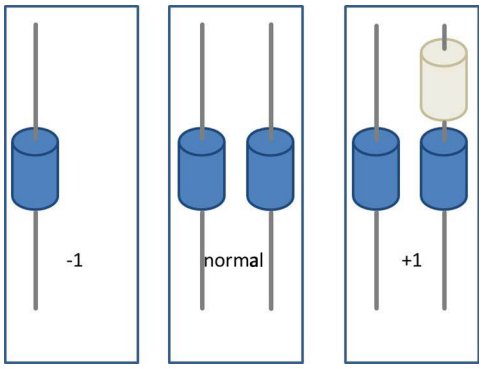

Figure 2.3: **Copy number alterations.** The cylinders represent a region of the genome. On the left side a deletion, and on the right side a duplication.

- **Loss of Heterozygosity.** Most human cells contain two copies of the genome, one from each parent. Loss of heterozygosity occurs when one parental copy of a certain region of the genome is lost.

- **Altered Methylation.** DNA methylation involves the addition of a methyl group to the cytosine (C) or guanine (G). High levels of methylation in the promotor region of a gene often results in transcriptional silencing of that gene.

2.3 Genomic data collection

The collection of genetic data is steadily growing. The consortia listed below are currently working on making available comprehensive observations of expression and genetic alterations for different types of cancers.

- **TCGA.** Since 2006, the Cancer Genome Atlas has been analysing and building up a comprehensive characterization of the genome of more than 20 cancer types. Its goal is to scientifically improve our ability to diagnose, treat and prevent cancer. The data is freely available through the TCGA Data Portal (cancergenome.nih.gov).

- **CGP.** The Cancer Genome Project (www.sanger.ac.uk) collects genomic data from 50 different types of cancers. The project focus on searching alterations that help identifying genes which are critical to the development of human cancers.

- **ICGC.** The primary goal of the International Cancer Genome Consortium (icgc.org) are to generate comprehensive catalogues of genomic abnormalities in tumors from 50 different cancer types or subtypes.

- **U-CAN.** The U-CAN (u-can.uu.se) collects and organises patient samples that are taken before, during and after cancer therapy. Patient data and radiological images are also collected. This material is in turn used to develop methods to fine-tune diagnoses and to better characterise different tumour diseases, in order to be able to choose an optimal therapy for the individual patient.

2.4 Cancer types

In this thesis we employ data from TCGA. Papers I and II focus on network modelling for one class of cancer while papers III and IV introduce joint models of up to eight cancer classes (see paper IV). The cancers in question are

- **Glioblastoma multiforme.** This is the most common fast-growing malignant brain tumor in adults. It accounts for about 15% of all brain tumors and has a poor prognosis, with a median survival time of 12 to 14 months.

- **Breast cancer.** This is the most frequently diagnosed cancer and the second most common cancer-related cause of death in women (although male breast cancer can also occur). It originates from breast tissue, most commonly inner lining of milk ducts or the lobules that supply the ducts with milk. Early detection and improvements in treatment have helped to steadily decrease the number of deaths it causes.

- **Ovarian carcinoma.** This type of ovarian cancer accounts for about 3% of all cancers in women. More than 90% of ovarian cancers are classified as epithelial and are believed to arise from the epithelium (surface) of the ovary. It has poor prognosis due to a lack of an early detection or screening test.

- **Lung squamous cell carcinoma.** This subtype of lung cancer makes up 25 to 30% of all lung cancers and its major cause is smoking. It originates from cells that replace injured cells in the lining of the bronchi.

- **Colon adenocarcinoma.** This is the most common type of gastrointestinal cancer. It originates in the glandular structures located in the inner layer of the colon. It can be treated and patients have a good prognosis if the disease is detected early.

- **Uterine carcinoma.** Uterine carcinoma makes up around 80% of the female reproductive system cancers. It develops from cells in the endometrium, the lining of the uterus. About 69% of the cases can be detected early giving a five-year survival rate of about 83%.

- **Kidney clear cell carcinoma.** This cancer is the most common kidney cancer. It forms in the cells lining the small tubules that filter the blood and make urine. It is more common in men and can be treated effectively if it is detected early.

- **Head and neck squamous cell carcinoma.** The membranes lining the inside of the mouth, nose and throat are made up squamous cells, where this type of cancers originate. It affects men twice as often as women and smoking and heavy drinking is associated with increased risk.

2.5 Biological pathways

Network estimation is about finding interactions between variables, given the data. A collection of already known interactions at the gene and metabolic level, the so called pathways, is publicly available at www.pathwaycommons.org. Therefore, one way to perform biological analysis of network estimates for genomic data, and further validate these estimates, is to find their overlap with the pathways. In this work, we have used the following pathway databases to perform overlap analysis:

- The Human Protein Reference Database (HPRD), contains information about interactions and disease associations for each protein in the human proteome (www.hprd.org).
- The National Cancer Institute Pathway Interaction Database (NCI-NATURE), is a collection of known biomolecular interactions and key cellular processes assembled into signaling pathways (pid.nci.nih.gov).
- REACTOME (www.reactome.org), which is a pathway database for different omics levels.
- IntAct Molecular Interaction Database (IntAct), which is a molecular interaction database (www.ebi.ac.uk).

We describe below the steps to perform network estimates overlap with the pathways. First, we map identifiers in the databases to our set of variables using the official gene symbols. We then compute the length of the shortest path (see Chapter 5), P_{ij} for gene pairs (i, j) in the database using Johnson's algorithm (Johnson, 1977). Pathway overlap is then computed as a fold enrichment, which is the expected number of times an estimated network Θ, contains an edge for genes with a path of length k. In detail, fold enrichment is defined as

$$\frac{\Pr(P_{ij} = k \mid \text{edge } (i,j) \text{ present in } \Theta)}{\Pr(P_{ij} = k \mid \text{edge } (i,j) \text{ present in } \Theta_{\text{permuted}})}.$$

The network Θ_{permuted} is obtained by randomly permuting the rows and columns in Θ, which is equivalent to randomly re-assigning gene names. We used a path-length $k = 1, 2$ in our calculations. The numerator and denominator were estimated as follows:

$$\Pr(P_{ij} = k \mid \text{edge } (i,j) \text{ present in } \Theta) = \sum_{i<j} \mathrm{I}\{P_{ij} = k \mid \theta_{ij} \neq 0\}/N,$$

where p is the number of genes in the network and $N = p(p-1)/2$ is the total number of possible edges. Similarly,

$$\Pr(P_{ij} = k \mid \text{edge } (i,j) \text{ present in } \Theta_{\text{permuted}}) = \sum_{r=1}^{R} \sum_{i<j} I\{P_{ij} = k \mid \theta_{ij,permuted} \neq 0\}/NR$$

where R is the number of random permutation graphs created.

Chapter 3

Network Estimation

Network estimation methods can be classified into four categories according to the way they measure the dependency level between variables; they are Bayesian networks, information theory-based, correlation-based and partial correlation-based methods (Allen et al., 2012). Estimation of networks is usually understood as estimation of sparse networks, since a fully connected network does not provide actionable insights into the dependency structure of the variables.

Papers I and II address a problem common to all network estimation methods and present a framework which can be applied to any of them. Papers III and IV introduce extensions to current partial correlation-based methods to better suit the analysis of biological data.

Below we present a brief summary of the first three types of estimation methods, Bayesian networks, information theory-based methods and correlation-based methods. Given the particular focus of papers III and IV in partial correlation-based methods, we present a more comprehensive review of such methods in the next section.

Bayesian networks

Construction of Bayesian networks is based on searching for a probabilistic-network structure with a high posterior probability. The solution is constrained to a graphical model that represents a set of variables and their independencies. Examples of methods to compute Bayesian networks are *BNArray* (Chen et al., 2006), *B-course* (Myllymäki et al., 2002), *BNT* (Murphy, 2001) and Werhli's implementation of *BN* (Werhli et al., 2006).

Information Theory-based Methods

This type of method uses mutual information to determine the dependencies between variables and removes indirect candidate interactions using the data processing inequality. The best known algorithm of such type is the *Algorithm for the Reconstruction of Accurate Cellular Networks*, ARACNE (Margolin et al., 2006).

Correlation-based Methods

The most straightforward way of estimating a network is by thresholding the sample covariance matrix, keeping only the strongest connections between pairs of variables. An example of a correlation-based method is the *Weighted Correlation Network Analysis*, WGCNA (Langfelder and Horvath, 2008). Another approach is to estimate the covariance matrix through penalized maximum likelihood as in Bien and Tibshirani (2011).

3.1 Review of partial correlation-based methods

Partial correlation-based methods make use of Gaussian graphical model-theory. The dependency is measured as the partial (conditional) correlation between variables, which is given by the inverse of the correlation matrix. The most common implementation of sparse partial correlation-based methods is the graphical lasso (glasso).

The genomic data we are interested in analysing are usually high-dimensional. Current technology allows for measurements of tens of thousands of genes, however, we seldom have access to more that a few hundred samples. For correlation and partial correlation-based methods it is indeed necessary to look for sparse estimates of the correlation or partial correlation matrices, since the empirical estimate of the correlation matrix is not positive definite and can be highly variable in the high-dimensional settings.

Under the assumption of normality, the problem of estimating the partial correlations is equivalent to estimating the inverse correlation matrix. For genomic data, this amounts to stating that the transcription of gene i is conditionally independent of the transcription of gene j given all the others (i.e. there is no link between i and j in the corresponding network), if and only if the (i, j)-th element in the precision matrix is zero.

For a single Gaussian graphical model, Dempster (1972) formulated this as the combinatorial problem of optimizing the location of zeros in the matrix.

Such an approach is too computationally intensive and indeed unfeasible for high dimensions. More recently, focus has shifted to models in which the the number of estimated parameters is constrained. Meinshausen and Buhlmann (2006) estimate each variable through an L_1 penalized regression on the rest of the variables. Later on, extensions and generalizations were proposed by Yuan and Lin (2007), Banerjee et al. (2008), D'Aspremont et al. (2008) and Friedman et al. (2008). All of these approaches produce estimates of the inverse covariance matrix referred to as the *graphical lasso*. Below we present some of the most recent network estimation methods based penalized maximum-likelihood.

Assume that the data X are observations from $N(0, \Sigma)$, where we assume without loss of generality that the data is centered, and Σ is the covariance matrix. Let $\Theta = \Sigma^{-1}$, penalized likelihood methods seek to optimize the function

$$l(\Theta) = \ln(\det(\Theta)) - \operatorname{tr}(S\Theta) - g(\lambda, \Theta),$$

where $S = \frac{1}{n}X^T X$ is the empirical covariance matrix, g is a suitable function of Θ which imposes the desired constrains on the model, and λ is a tuning parameter which can be a vector or a matrix.

For the glasso method $g(\lambda, \Theta) = \lambda\|\Theta\|_1 = \lambda \sum_{i \neq j} |\theta_{ij}|$ and θ_{ij} is the ij-th element of Θ. This penalty is known as the *lasso* penalty (Tibshirani, 1996), and therefore referred to as the graphical lasso in this context. The parameter λ controls the degree of sparsity in Θ, the larger λ is, the more elements in Θ will be shrunk to zero.

A similar problem, but so far only studied in the linear regression context, is the *elastic net* (Zou and Hastie, 2008), where the penalty function is given by

$$g(\lambda, \alpha, \Theta) = \lambda \sum_{i \neq j} \left[\alpha|\theta_{ij}| + (1-\alpha)\theta_{ij}^2\right].$$

The elastic net often outperforms the lasso as a variable selection method. It also has a grouping effect, in which parameter estimates for strongly correlated variables tend to be zero, or not, simultaneously.

In the presence of several classes of samples that share parameters, it is necessary to jointly estimate the precision matrices. Joint estimation models have been proposed by Guo et al. (2011), Yuan and Lin (2007), and Guo and Wang (2010). The assumptions are that the data for each group are drawn from a multivariate normal distribution with covariance matrix Σ^k, $k = 1, 2, \ldots, K$.

Guo et al. (2011) proposed a model in which common sparsity structure (location of zeros in the precision matrices) is achieved. They find the solution

to this problem by iteratively optimizing the K likelihood functions

$$l(\Theta^k) = \ln\left(\det\left(\Theta^k\right)\right) - \operatorname{tr}\left(S^k\Theta^k\right) - \lambda\sum_{i\neq j}\omega_{ij} \mid \theta_{ij}^k \mid,$$

where $\Theta^k = \left(\Sigma^k\right)^{-1}$ and $\omega_{ij} = \left(\sum_{k=1}^K \mid \theta_{ij}^k \mid\right)^{-1/2}$. The problem can thus be solved by repeatedly applying *glasso* to the precision matrix and updating the penalty so it decreases in each iteration for links that must be present across all classes.

A similar approach that also guarantees a common sparsity pattern, but not equal values, is the sparse *group lasso* (Yuan and Lin, 2006). It optimizes the likelihood function

$$\begin{aligned} l\left(\{\Theta\}\right) = &\sum_{k=1}^K n_k\left[\ln\left(\det\left(\Theta^k\right)\right) - \operatorname{tr}\left(S^k\Theta^k\right)\right] \\ &- \lambda_1\sum_{k=1}^K\sum_{i\neq j} \mid \theta_{ij}^k \mid - \lambda_2\sum_{i\neq j}\sqrt{\sum_{k=1}^K\left(\theta_{ij}^k\right)^2}, \end{aligned}$$

where $\{\Theta\} = \{\Theta^1, \Theta^2, \ldots, \Theta^K\}$. Here, the first term in the second row is the lasso penalty, controlled by λ_1; and the second term corresponds to the so called *group* penalty (an L_2 type norm), controlled by λ_2. The group penalty considers the vectors $(\theta_{ij}^1, \theta_{ij}^2, \ldots, \theta_{ij}^K)$ for all $i \neq j$ and penalizes their Euclidean norm, thus achieving common zeros across classes for large enough values of λ_2.

Going one step further, some methods look for exactly equal values of the estimated parameters across classes, instead of a common pattern of zeros only. The octagonal shrinkage and clustering algorithm for regression, *OSCAR*, described in Bondel and Reich (2008), has been proposed to solve this problem in the context of linear regression. The OSCAR penalty can be extended to network estimation just as the lasso was extended to graphical models. The resulting log-likelihood function is

$$\begin{aligned} l\left(\{\Theta\}\right) = &\sum_{k=1}^K n_k\left[\ln\left(\det\left(\Theta^k\right)\right) - \operatorname{tr}\left(S^k\Theta^k\right)\right] \\ &- \lambda_1\sum_{k=1}^K\sum_{i\neq j} \mid \theta_{ij}^k \mid - \lambda_2\sum_{k<k'}\sum_{i\neq j}\max\{\theta_{ij}^k, \theta_{ij}^{k'}\}. \end{aligned}$$

Like the group lasso, it contains a lasso penalty controlled by λ_1 to tune the network sparsity. The difference is that is adds a second penalty (an L_∞ type norm) on the maximum for all pairwise combinations across classes of

the precision matrices' elements. Large enough values of λ_2 have the effect of shrinking the values of θ_{ij}^k and $\theta_{ij}^{k'}$ towards each other.

OSCAR has the drawback of being difficult to solve. A more tractable way to approach the problem of joint estimation with equal values across classes is suggested by Danaher et al. (2014). In this case, the log-likelihood takes the form

$$l\left(\{\Theta\}\right) = \sum_{k=1}^{K} n_k \left[\ln\left(\det\left(\Theta^k\right)\right) - \operatorname{tr}\left(S^k\Theta^k\right)\right] - \lambda_1 \sum_{k=1}^{k}\sum_{i\neq j} \mid \theta_{ij}^k \mid -\lambda_2 \sum_{k<k'}\sum_{i,j} |\theta_{ij}^k - \theta_{ij}^{k'}|.$$

Like group lasso and OSCAR, sparsity is controlled by a lasso penalty (λ_1) but this time the equality in parameter values is encouraged by a fused penalty (an L_1 type norm, Hoefling (2010)). The fused penalty is controlled by λ_2 which, like OSCAR, shrinks θ_{ij}^k and $\theta_{ij}^{k'}$ towards each other for large enough values.

All of these methods have in common the need to select the value for the tuning parameters λ_1 and λ_2, which induce different sparsity and commonality patterns across classes, thus producing different models which require validation.

In the next chapter we review some of the model validation techniques and describe some common statistical tools we employ in papers I and II, in order to make the material in them easily accessible to the reader.

Chapter 4

Network model validation and uncertainty in estimation

A number of different methods are available to tune the parameters λ_1 and λ_2 for the network estimation methods described in the previous chapter. This problem, often referred to as the model validation problem, is complex and subject of ongoing research. Below we present common model validation approaches: cross-validation (CV) and Bayesian Information Criterion (BIC). We also describe some common statistical tools used in papers I and II.

4.1 Cross-validation

Probably the simplest validation method is K-fold CV. K-fold CV involves the following steps:

- Split the data X in K equal (or almost equal) sizes by randomly assigning them to groups (folds) $X^1, X^2, \ldots, X^K$. Let X^{-k} be the data without the k-th fold.
- Select a sequence of size M of tuning parameters of interest.
- For $k = 1, 2, \ldots K$ do
    1. Estimate the network models $m_1, m_2, \ldots, m_M$ using X^{-k}.
    2. Compute $L_k(m_j)$, the likelihood function for model m_j evaluated on data set X^{-k}.
- Compute the total likelihood as $\sum_{k=1}^{K} L_k(m_j)$.

- Select the model m_* which maximizes the total likelihood from previous step.

4.2 Bayesian Information Criterion

In the case of the graphical lasso, the BIC for estimate $\hat{\Theta}_\lambda$ is defined as:

$$\text{BIC}(\lambda) = -n \ln(\det(\hat{\Theta}_\lambda)) + n\text{tr}(S\hat{\Theta}_\lambda) + \ln(n) \sum_{i<j} \text{I}\left\{\hat{\theta}_{ij,\lambda} \neq 0\right\}$$

where n is the sample size, S is the empirical covariance estimate and the number of non-zero elements in $\hat{\Theta}_\lambda$ are the degrees of freedom for the model (Schwarz, 1978). BIC can easily be generalized for joint models such as the fused or the group lasso, by defining the degrees of freedom as the number of unique (differential) non-zero estimated parameters in the model.

Another measure is the Akaike Information Criterion (AIC), defined as (Danaher et al., 2014):

$$\text{AIC}(\lambda_1, \lambda_2) = \sum_{k=1}^{K} \left[-n \ln(\det(\hat{\Theta}^k_{\lambda_1,\lambda_2})) + n\text{tr}(S^k\hat{\Theta}^k_{\lambda_1,\lambda_2}) + 2\sum_{i<j} \text{I}\left\{\theta^k_{ij,\lambda_1,\lambda_2} \neq 0\right\} \right]$$

4.3 The bootstrap

Approaches to network model validation at edge level have also been considered. In de Matos Simoes and Emmert-Streib (2012), subsampling is used to collect frequency statistics on edge presence/absence. The frequency statistics are then for testing the edges to be part of the final estimate.

The bootstrap is a method estimate the distribution function of the data and use it for subsampling, which in turn can be used to estimate statistics of interest, e.g. the variance or the bias of an estimate. Below is a brief description on how the bootstrap can be used in general.

Let $X = (x_1, x_2, \ldots, x_n)$ be a random sample of size n drawn from the distribution F. The empirical distribution function $\hat{F}$ is defined as the function that assigns probability $1/n$ to each x_i, $i = 1, 2, \ldots, n$. The natural extension to an event A is thus $\Pr_{\hat{F}}(A) = \#\{x_i \in A\}/n$.

The parameters of the distribution F (a function t of F), can be estimated through the plug-in principle. The plug-in estimate of a parameter $\theta = t(F)$

is defined to be $\hat{\theta} = t(\hat{F})$. A bootstrap sample is defined to be a random sample of size n from $\hat{F}$. In practical terms, this means that a bootstrap sample $X^b = (x_1^b, x_2^b, \ldots, x_n^b)$ is sample of size n, drawn with replacement from the original data $X = (x_1, x_2, \ldots, x_n)$. For an estimator $\hat{\theta} = s(X)$ we define a bootstrap replication as $\hat{\theta}^b = s(X^b)$. The (approximate) bootstrap estimate $\hat{\theta}_B$ is an aggregated estimate of bootstrap replications $\hat{\theta}^b$ for $b = 1, 2, \ldots, B$.

In some applied situations, the resampling scheme explained above may not be optimal. An example is that of genetical data (such as expression levels) where sampling with replacement is equivalent to having two or more patients with exactly the same values for all variables in the study, which is unrealistic. There are resampling schemes that address this issue, one of them being the jackknife, which predates the bootstrap and uses samples that leave one observation out of the original sample. The jackknife procedure limits the number of new samples that can be obtained much more than the bootstrap, but it still shares theoretical similarities with the bootstrap with the advantage of being simpler to compute.

Plenty of theoretical work has been published dealing with the properties of the bootstrap and the jackknife estimates for common statistics, such as the mean, the standard error and the bias. In these three cases, the jackknife provides an approximation to the corresponding bootstrap estimators, but it can be inconsistent for non-smooth statistics such as the median. One way to solve this issue is to resample by leaving out $d > 1$ observations at the time: if more than $d = \sqrt{n}$ but fewer than n observations are left out, then the jackknife estimate is consistent for the median (Efron and Tibshirani, 1994).

A suitable resampling scheme for genetic data is, for example, leave-10%-out. To perform network model validation with help of the bootstrap we first collect frequency statistics from bootstrap estimates. These can be frequency statistics of any binary decision, such presence/absence or fusing of edges. An aggregated bootstrap estimate is then constructed where the weights of the edges are equal to the proportion of bootstrap estimates in which they are present or fused. The final model is constructed by selecting cutoff thresholds $0 < t_1 \leq 1$ and $0 < t_2 \leq 1$ and keeping edges with weights at least equal to t_1, as well as fusing edges with corresponding weights larger or equal to t_2 in the aggregated bootstrap estimate.

The approach described above does not completely solve the model validation problem, since selection of the cutoff thresholds is needed. Also, its performance is dependent on the original sparsity selected for the construction of the aggregated bootstrap estimate. In paper I we describe a method in which the frequency statistics collected from bootstrap are modelled as

a mixture of Beta-Binomial distributions. This approach circumvents the selection of cutoff thresholds and provides a way to construct network estimates from bootstrap aggregation. Below we describe succinctly describe the Beta-Binomial model.

4.4 Generalized linear models for count data

Linear statistical models are a common tool for the analysis of many different types of data. Sparse versions of linear regression have been applied to the construction of networks (Meinshausen and Buhlmann, 2006) or to model the effects of certain variables into transcription networks (Jörnsten et al., 2011).

Linear models assume a response variable y to be a linear combination of certain predictors X, that is $y = X\beta + \epsilon$, where ϵ is a random variable such that $\mathrm{E}(\epsilon) = 0$ and $\mathrm{Var}(\epsilon) = \sigma^2$. In general, such linear models should be considered a first order approximation of potentially more realistic models that are nonlinear in the parameters, called generalized linear models.

A generalized linear model assumes that the response variable y has a known distribution f_Y, that is $y \sim f_Y(x; \beta)$. In this thesis work, we make use of generalized linear models to address the network validation problem (selection of sparsity level). To this end, we model frequency statistics of edge presence/absence across bootstrap estimates. A Beta-Binomial model and its zero-inflated version are the models of choice. Here we describe them briefly.

4.4.1 Beta-Binomial model

Beta-Binomial density

The Beta-Binomial model is commonly used to model overdispersed binomial data. Overdispersion is taken into account by letting the binomial probability of success p be beta distributed.

Let $Y \sim \mathrm{Bin}(n, p)$. The density for X is given by

$$f_Y(k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

The prior distribution for p is Beta(α, β) with density

$$f_p(p) = \frac{1}{\mathrm{B}(\alpha, \beta)} p^{\alpha-1}(1-p)^{\beta-1},$$

where B is the Beta function defined as $\mathrm{B}(x, y) = \int_0^1 t^{x-1}(1-t)^{y-1}$.

Note that the Beta distribution is conjugate to the Binomial, since

$$p^k(1-p)^{n-k} \sim p^{\alpha-1}(1-p)^{\beta-1}.$$

This fact allows us to compute the posterior distribution in closed form using Bayes' Theorem.

Indeed, the posterior Beta-Bin(n, μ, α, β) distribution has the density

$$\begin{aligned} f(k) &= \int_0^1 f_y(k) f_p(p) dp \\ &= \binom{n}{k} \frac{1}{\mathrm{B}(\alpha, \beta)} \int_0^1 p^{k+\alpha-1}(1-p)^{n-k+\beta-1} dp \\ &= \binom{n}{k} \frac{\mathrm{B}(k+\alpha, n-k+\beta)}{\mathrm{B}(\alpha, \beta)}. \end{aligned}$$

Zero-Inflated Beta-Binomial density

The modelling assumptions of the Beta-Binomial distribution may be insufficient when the count data has an excess of zero values. This is a relevant situation in our applications since we are interested in sparse networks, where most of the edges are absent. A zero-inflated model becomes necessary in high-dimensional settings, when the lack of edges can mask the distribution of the remaining edge presence counts.

The Zero-Inflated Beta-Binomial distribution, ZIBB$(n, \mu, \nu, \alpha, \beta)$, models overdispersed data as a mixture with two components: one for the zero-counts and another for the rest of the counts. The density is given by

$$f(y) = \nu \mathrm{I}\{y = 0\} + (1-\nu) f_{BB}(y),$$

where ν is the probability of a zero-count, and f_{BB} is the density for the Beta-Binomial distribution defined above.

4.4.2 Estimation of a mixture model's parameters: the EM algorithm

In Paper I we model the edge presence/absence counts as random observations from two populations: the positives (edges present in the true network)

and the negatives (edges absent in the true network). Each population is in turn modelled accordingly to either a Beta-Binomial or a Zero-Inflated Beta-Binomial, thus defining a mixture model with two components or three components. Estimation of the parameters is carried out by the expectation-maximization (EM) algorithm, which we describe below.

Let $Y = (y_1, y_2, \ldots, y_N)$ be the data (the observed counts in our case) and f_1 and f_2 the distributions of the two populations from which the elements in Y are drawn. Further, let Δ be some unobserved binary variable with $\Pr(\Delta = 1) = \pi$. Thus the data can be written as $Y = (1 - \Delta)Y_1 + \Delta Y_2$, where $Y_1 \sim f_1$ and $Y_2 \sim f_2$. The density of Y becomes

$$f_Y(y) = (1 - \pi)f_1(y) + \pi f_2(y). \tag{4.1}$$

Suppose we know the values of Δ_i, $i = 1, 2, \ldots, N$; then the log-likelihood of 4.1 would be

$$\begin{aligned} L(\theta; Y, \Delta) = &\sum_{i=1}^{N} \{(1 - \Delta_i) \ln [f_1(y_i|\theta_1)] + \Delta_i \ln [f_2(y_i|\theta_2)]\} \\ &+ \sum_{i=1}^{N} \{(1 - \Delta_i) \ln(1 - \pi) + \Delta_i \ln(\pi)\}, \end{aligned}$$

where $\theta = (\theta_1, \theta_2)$ and θ_1 and θ_2 are the vectors of parameters for f_1 and f_2, respectively. It can be easily seen that the MLE's of θ_1 correspond to the MLEs of f_1 computed on the data for which $\Delta_i = 0$. Similarly, the MLEs of θ_2 correspond to the MLE's of f_2 computed on the data for which $\Delta_i = 1$. The MLE for π is the proportion of the data for which $\Delta_i = 1$.

Suppose now that we know the parameters θ and π. We can substitute the values of Δ_i by their expectations, called responsibilities

$$\gamma_i(\theta) = \mathrm{E}(\Delta_i|\theta, \pi, Y) = \Pr(\Delta_i = 1|\theta, \pi, Y),$$

and estimate these as the relative density of observations in each class.

The full set of estimates is obtained by iterating these two steps until convergence. A summary of the steps is given below.

EM algorithm for two-component mixture

1. Compute initial values for the parameters $\hat{\theta_1}$, $\hat{\theta_2}$ and $\hat{\pi}$.
2. *E-Step*. Compute the responsibilities

$$\hat{\gamma_i} = \frac{\hat{\pi} f_1(y_i|\hat{\theta_1})}{(1 - \hat{\pi}) f_1(y_i|\hat{\theta_1}) + \hat{\pi} f_2(y_i|\hat{\theta_2})}, \quad i = 1, 2, \ldots, N.$$

3. *M-Step.* Compute the MLE of f_1 and f_2.

$$\hat{\theta_1} = \max_{\theta_1} \sum_{i=1}^{N} (1 - \hat{\gamma}_i) \ln [f_1(\theta_1 | y_i)]$$

$$\hat{\theta_2} = \max_{\theta_2} \sum_{i=1}^{N} \hat{\gamma}_i \ln [f_2(\theta_2 | y_i)]$$

$$\hat{\pi} = \frac{1}{N} \sum_{i=1}^{N} \hat{\gamma}_i$$

4. Iterate steps 2 and 3 until convergence.

4.5 Rate-Distortion theory

In paper II we study the problem of network estimation uncertainty. In this context, uncertainty is to be understood as the variability of the network estimate (which edges are present and which are absent). Going one step further, we claim that a network estimate has regions (modules) which have potentially different uncertainty levels. The problem of model validation becomes a question of assessing different levels of uncertainty. Similar concepts have been applied in the information theory field, where the problem of image compression is related to the variability of the image's different regions. We borrow strength from information theory and apply rate-distortion theory, used for selecting image compression rates, to do network model selection.

Rate-distortion theory has been previously used outside the information theory field for model selection in cluster analysis (Jörnsten, 2009) and sparse regression (Jauhiainen et al., 2012) on high-dimensional data. To perform model selection using rate-distortion theory we proceed as follows.

Suppose we have M models, for which we want to do simultaneous selection. In our case, the models correspond to the different network modules. Each model, is indexed by some rate value, such as the number of estimated parameters or a tuning parameter. Also, for each model, is possible to assess the distortion as the inverse of some goodness-of-fit measure, such as the likelihood or the residual sum of squares.

Our goal is to minimize the overall distortion (i.e. the inverse goodness-of-fit) constrained to a fix rate (number of parameters). In other words, we fix the total number of parameters we are willing to pay to achieve some overall explanatory power. Figure 4.1 shows, in solid lines, the rate-distortion curves for four models; the dashed lines have all a fixed slope Δ. It can be shown

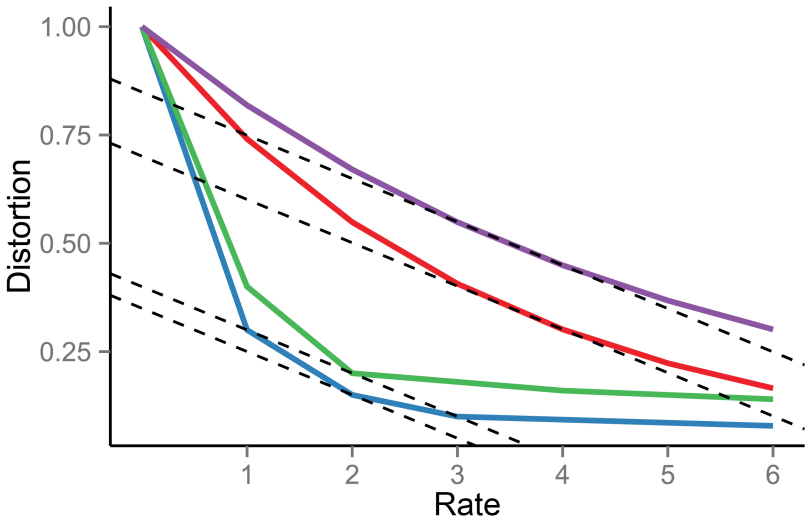

Figure 4.1: **Rate-Distortion curves.** The solid curves represent the rate-distortion curves to four different regions. The dashed curves represent a fixed slope. To minimize distortion given a fixed rate select points with equal tangent slope.

that the overall distortion is minimized by selecting the points where the dashed lines are tangents to the rate-distortion curves. At this points, the fixed rate constraint is fulfilled with a value equal to the sum of the rates for the corresponding selected points.

If any other point is selected, rather than the ones with equal tangent slope, there is always one trade-off move possible between a pair of modules that increase the overall distortion. Suppose for model 1 the rate is increased from R_1 to $R_1 + \delta_1$, leading to a distortion decrease from D_1 to $D_1 - \epsilon_1$. At the same time, decreasing the rate for model 2 from R_2 to $R_2 - \delta_2$ leads to a distortion increase from D_2 to $D_2 + \epsilon_2$ where $\epsilon_1 > \epsilon_2$.

Chapter 5

Network analysis

Once a network has been estimated, a number of tools are available for its characterization. From the biological point of view it is interesting to study the overlap of the estimated network with known pathways (possibly measured as fold enrichment) or to find the *hub* genes, which are genes with a large number of connections.

Mathematically, the study of networks or *graphs* belongs to *graph theory*. In this chapter we define networks in the context of graph theory and introduce some important graph properties which we can use to, among others, formally define the concept of hub.

5.1 Graph theory

A network or graph G is defined as a pair $G = (V, E)$, where V is the set of vertices (or nodes) and E is the set of edges (or links). Since an edge connects two vertices, the elements of set E are subsets of V with two elements each. If the set E is unordered the graph G is called undirected; if E is ordered, then G is directed. If there are no edges from a node to itself the graph is called simple. If the elements of E have real values associated to then the graph is called weighted; otherwise these values can all considered to be 1, in which case the graph is called unweighed. Here we will focus on simple, undirected, weighted and unweighted graphs.

Networks can compactly be described by their adjacency matrix A. For a simple, undirected and unweighted graph, the adjacency matrix is a square binary matrix with elements $a_{ij} = 1 = a_{ji}$, $i \neq j$, if an edge is present between nodes v_i and v_j. If no edge is present then $a_{ij} = 0 = a_{ji}$. For a

weighted network a_{ij} gives the weight for the edge between nodes v_i and v_j.

A number of properties can be defined on the nodes or edges of a graph. Below we define some of these. Let $G = (V, E)$, $V = \{v_1, v_2, \ldots, v_n\}$ and $E = \{e_1, e_2, \ldots, e_m\}$,

- Degree. The degree of a node v_i is defined a the number of edges coming into or going out from it.
- Path. A path p_{ij} between two nodes v_i and v_j is a sequence of connected nodes (there exists an edge between them) such that it connects v_i and v_j, that is, $p_{ij} = \{v_i = v_{i_0}, v_{i_1}, v_{i_2}, \ldots, v_{i_k} = v_j\}$. The length of a path is the number of elements in it.
- Shortest path. The shortest path is any path between nodes v_i and v_j with the minimum number of edges required.
- Distance. The distance between two nodes v_i and v_j is the length of the shortest path between them.
- Neighbours. The neighbours of degree d of a node v_i are the nodes reachable from v_i within distance d.
- Cycle. A cycle is a path that starts and ends at the same node.
- Farness. The farness of a node v_i is the sum of the distances from itself to all other nodes v_j.
- Centrality. The inverse of farness, also called closeness.
- Betweeness centrality. Let p_{ij} be the shortest path between nodes v_i and v_j. The betweeness of a node v_k is the number of all p_{ij} that contain v_k, that is, how often a node is part of the shortest path between any two other nodes (Freeman, 1977).
- PageRank. Originally conceived as a way to characterize the importance to web pages connected by hyperlinks, PageRank (Page et al., 1999) is an algorithm that assigns weights to nodes in a graph proportional to number of edges coming into it. Equivalently, the weight of a node is proportional to the total weight of nodes connected to it. The distribution of the weights represents the probability of arriving to a particular node by randomly moving from one node to another.

Networks can exhibit different topologies according to the number and location of their edges. A certain topology enables a network to be more suitable for a certain task. Common topologies encountered in biological networks are hubby, banded (forming chains) and scale-free networks. We describe these topologies below.

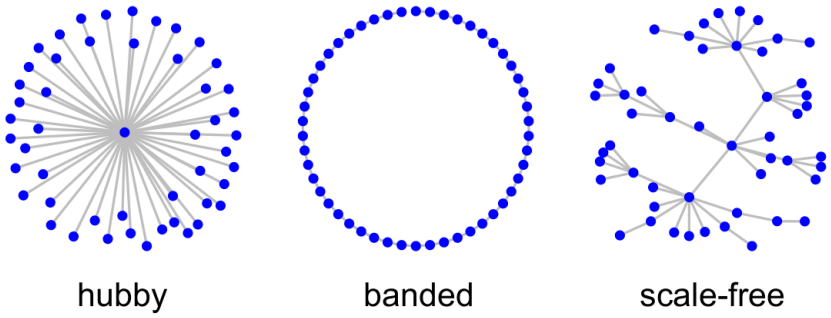

Figure 5.1: **Network topologies.**

- Hubby. Networks with a relative high number of hubs are called hubby. Hubs are the nodes with highest degree in the. The precise value of the degree depends on the number of nodes and edges in the graph, however, a node is typically considered as a hub if has a degree above the average degree for all nodes.
- Banded. A banded network, also called a chain network, where each node is connected to two and only two other nodes.
- Scale-free. In a scale-free network the degree distribution follows a power law, that is, if a node v_i is selected at random then the probability that its degree is k is proportional to $k^{-\gamma}$ (typically $2 \leq \gamma \leq 3$).

The three types of topology are illustrated in Figure 5.1.

5.2 Finding groups/modules in networks

Biological networks are not completely random. Genes form groups to perform complex tasks, therefore, the networks show a certain modular structure (also referred to as community). Communities are usually found by clustering the nodes of a network. As in any clustering method, the end result depends on how the distance between nodes is defined.

Some useful distance measures are defined below. Throughout the text we assume that we have a network with adjacency matrix $A = [a_{ij}]$.

- Correlation/partial correlation. When the adjacency matrices are given by the (sparse) correlation matrices, the dissimilarity can be defined in terms of the correlation values themselves. In this case, the communities found are groups of nodes with high correlation values. The

dissimilarity between nodes v_i and v_j is given by

$$d(i,j) = 1 - |a_{ij}|.$$

A similar definition applies for partial correlation matrices.

- Jaccard index. This measures the similarity between two nodes v_i and v_j as the quotient between the number of edges that v_i and v_j have in common, and the number of edges coming into v_i or v_j. Let $a_{i\cdot}$ be the i-th row of the adjacency matrix, the dissimilarity between nodes v_i and v_j is given by

$$d(i,j) = 1 - \frac{a_{i\cdot} \cdot a_{j\cdot}}{\|a_{i\cdot} + a_{j\cdot}\|_1}.$$

- Cosine similarity. The cosine similarity considers distance as the cosine of the angle between two vectors. The dissimilarity is given by

$$d(i,j) = 1 - \frac{a_{i\cdot} \cdot a_{j\cdot}}{\|a_{i\cdot}\|\|a_{j\cdot}\|},$$

where $a_{i\cdot}$ is defined as above.

- Topological overlap. This measure attempts to establish similarity between nodes as a function of the number of neighbours they have in common (Yip and Horvath, 2007). Let $N_d(i)$ be the set for neighbors of degree d for node v_i. The topological overlap dissimilarity is given by

$$\begin{aligned} d(i,j) &= 1 - t_d(i,j) \\ t_d(i,j) &= \begin{cases} \frac{|N_d(i) \cap N_d(j)| + a_{ij}}{\min\{N_d(i), N_d(j)\} + 1 - a_{ij}} & \text{if} \quad i \neq j \\ 1 & \text{if} \quad i = j. \end{cases} \end{aligned}$$

A different way of finding communities is proposed by Clauset et al. (2004). There, the idea is to find subsets of strongly connected nodes, but with weak connections between them. In terms of the adjacency matrix, the procedure is equivalent to finding an ordering of the rows and columns that creates a nearly block diagonal matrix.

Chapter 6

Summary of Papers

Paper I:
Sparsity selection and robust network estimation via bootstrap

Current high-throughput techniques allow for the observation of many variables. However, it is not always possible to collect a comparable number of samples. Therefore, omics data is often high-dimensional and regularization is required in the analysis. Network estimation methods are popular tools to analyze different types of omics data and regularization is imposed by the selection of a tuning parameter that controls the overall sparsity of the network estimate.

Common model validation techniques such as cross-validation and the Bayesian Information Criterion have been applied to this problem, but their performance is unsatisfactory. Also, sparsity selection has focused on selecting a threshold parameter that matches the size of the estimated network with the size of the true network. However, the size of the true network is generally unknown, and even if it could be estimated, it is not obvious that selecting a threshold in this way results in a quality estimate. Depending on sample size and signal strength, the number of false positives in such way selected estimate can be large. Instead, the goal should be to assemble a reproducible network with few false positives. Controlling the number of false positives is crucial in order to avoid overinterpretation and erroneous conclusions in the scientific disciplines where the network models are applied.

Bootstrap methods are already prevalent in the network inference literature, mainly as a tool to produce a stable final network. The network is

constructed by retaining the edges that appear frequently in the bootstrap networks, thus resulting in a more robust estimate. In Paper I we extend the bootstrap approach and develop a method that (i) selects an appropriate sparsity level for controlling the presence of false positive edges, and (ii) constructs a final network estimate that improves over naive bootstrap thresholding methods.

Through comprehensive simulation studies we show how the performance of network estimation methods depends on the true network size, signal strength and sample size. We also show that matching the sparsity of the estimate to that of the true network fails to control the false positive rate. Finally, we illustrate how our method controls the false positive rate and how it is possible to post-process an estimate and correct for an excess of false positive edges if the sparsity was not correctly selected.

Our proposed procedure is illustrated on a large-scale ovarian tumor sample data from the Cancer Genome Atlas (TCGA). By assessing the fold enrichment of functional groups (biological pathways), we show how we improve both upon commonly used methods of choosing network sparsity, as well as on the standard way of constructing a final estimate.

Paper II:
Can I trust my network? Local network resolution elucidates uncertainty in estimation with different methods and across network topologies

Once the estimate of a parameter has been obtained, it is natural to ask about its uncertainty. For point estimates, this question has been traditionally answered by providing confidence intervals. For networks, results are usually presented as a single (point) estimate, thus ignoring the instability and uncertainty of network estimation methods, leading to a substantial risk of over-interpretation of the resulting network.

In this paper we introduce a novel framework for network analysis that utilizes the estimation uncertainty in order to (i) produce a set of candidate graphs for the network, akin to a confidence set, (ii) enable a fair assessment of competing network estimation methods, and (iii) show how different structural topologies exhibit different estimation properties. The proposed framework can be used with any estimation method that produces undirected networks, e.g. likelihood based methods like glasso, or correlation based procedures like WGCNA.

The method extends concepts from information theory and image compression and assumes that the true network is modular (a common property in e.g. biological networks). Just as with regions in an image, it is reasonable to expect that modules in a network have different estimation properties. Also, just as the number of bits required to encode each region of an image is a measure of its complexity, we measure the complexity of a module as a number of candidate graphs. The candidate graphs are extracted from a set of observed bootstrap networks, which guarantees that the estimated topology is supported by the data. The number of candidates thus conveys the uncertainty in the estimation procedure, where a large number of candidate graphs implies low resolution and high uncertainty, and vice versa. We refer to this concept as network component resolution (NetCoR).

Through simulation studies we demonstrate how common topologies for biological networks (random, hubby, banded and scale-free) have different estimation properties and how estimation network performance depends on topology.

We apply our method to a data set of breast cancer tumors from The Cancer Genome Atlas. We show that high resolution modules overlap better with known pathways, whereas low resolution modules do not. This demonstrates how NetCoR can be used in real applications, enabling discrimination between network components comprising credible findings and regions that require additional experimentation before conclusions can be confidently drawn. We also classify the modules according to the four common topologies which, given the network estimation method, provides further insight into the quality of the estimates.

NetCoR is under development to be released as an R package available from the CRAN repository.

Paper III: Exploration of pan-cancer networks by generalized covariance selection and interactive web content

Current algorithms for network construction are not designed to work across multiple diagnoses and technical platforms, thus limiting their applicability to comprehensive pan-cancer datasets such as the Cancer Genome Atlas (TCGA).

Here, we introduce a strategy for pan-cancer network modeling, based on two novel contributions. First, we describe a generalization of sparse in-

verse covariance selection (SICS) designed to integrate genetic, epigenetic and transcriptional data from multiple cancers into a comparative network. The method uses a new strategy involving non-informative priors to account for the modular structure commonly observed in biological data and to integrate several data types. The method also corrects for potentially unbalanced sample sizes for different cancers. Via simulations, the algorithm is shown to be statistically robust and effective at detecting direct pathway links in TCGA data.

Second, we propose to rationalize the interpretation of the derived networks by a new and publicly accessible tool Cancer Landscapes, (cancerlandscapes.org). Cancer Landscapes is an interactive web-based tool in which derived models can be graphically explored and linked to several pathway and pharmacological databases. To evaluate the performance of the method, we constructed a model of genetic, epigenetic and transcriptional data for eight TCGA cancers, using data from 3900 patients. The derived model rediscovered known mechanisms and contained interesting predictions. Possible applications include the prediction of regulatory relationships between genes in particular cancers, comparison of network modules in across multiple forms of cancer, and identification of drug targets. The proposed algorithm may also have interesting applications in other areas of investigation, such as integration of multiple GWAS or biomarker studies.

Paper IV:
Joint estimation of sparse inverse covariance and mean with applications to discriminant analysis

Networks provide second order information, in the form of pairwise connections, allowing for the study of gene differential connectivity. Traditionally, the analysis of genomic data has been done either studying the differential connectivity or the differential expression, which provides first order information.

In Paper IV we propose to integrate and extend existing methods into a joint framework to investigate a multi-class differential connectivity and differential expression. Our method builds up the Gaussian graphical model framework, thus assuming differential connectivity to be given by partial correlations (precision matrix). With a minor additional constraint on the precision matrices to be block diagonal, we show how to construct an ensemble classifier. The classifier's quadratic component includes only differential blocks (QDA blocks) and its linear component includes only the non-differential blocks (LDA blocks).

Estimation is done via maximization of a fused lasso penalized likelihood using via alternating directions method of multipliers (ADMM). We estimate the mean vectors and the precision matrices separately by iterating ADMM maximization of the corresponding profile likelihoods.

We investigate the performance of our method through a simulation study. We show that the tuning parameters are in complex dependency, resulting in a complex dependency of network size, number of unique parameters estimated and proportion of differential parameters. Model selection becomes then a complicated problem, we show how it can be guided with the help of network estimation accuracy measures and misclassification rate.

As an application to real data, we use our method on a training data set of breast and ovarian tumours from The Cancer Genome Atlas (TCGA). We obtain estimates for differential expression levels and differential connectivity for a range of the sparsity and fuse penalties. We analyze the estimated models characteristics such as number of unique estimates parameters, network sizes and proportion LDA and QDA blocks. We measure the overlap of the estimated networks with known biological pathways as a fold enrichment and propose a strategy to perform model selection with the help of fold enrichment and misclassification rate.

Bibliography

A.S. Adler, M. Lin, H. Horlings, D.SA. Nuyten, M.J. van de Vijver, and H.Y. Chang. Genetic regulators of large-scale transcriptional signatures in cancer. *Nature genetics*, 38(4):421–430, 2006.

U.D. Akavia, O. Litvin, J. Kim, F. Sanchez-Garcia, D. Kotliar, H.C. Causton, P. Pochanard, E. Mozes, L.A. Garraway, and D. Pe'er. An integrated approach to uncover drivers of cancer. *Cell*, 143(6):1005–1017, 2010.

J. Allen, Y. Xie, M. Chen, L. Girard, and G. Xiao. Comparing statistical methods for constructing large scale gene networks. *PLoS One*, 7(1), 2012.

O. Banerjee, L. E. Ghaoui, and A. D'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine Learning Reseach*, 9:485–516, 2008.

J. Bien and R.J. Tibshirani. Sparse estimation of a covariance matrix. *Biometrika*, 98(4):807–820, 2011.

H. D. Bondel and B. J. Reich. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar. *Biometrics*, 64(1):115–123, 2008.

X. Chen, M. Chen, and K. Ning. Bnarray: an R package for constructing gene regulatory networks from microarray data by using Bayesian network. *Bioinformatics*, 22(23):2952–2954, 2006.

A. Clauset, M.EJ. Newman, and C. Moore. Finding community structure in very large networks. *Physical review E*, 70(6):066111, 2004.

P. Danaher, P. Wang, and D.M. Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):373–397, 2014.

A. D'Aspremont, O. Banerjee, and L. E. Ghaoui. First-order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and Applications*, 30:56–66, 2008.

R. de Matos Simoes and F. Emmert-Streib. Bagging statistical network inference from large-scale gene expression data. *PLoS ONE*, 7(3):6e33624, 2012.

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

Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*, volume 57. CRC press, 1994.

W. Fiers, R. Contreras, F. Duerinck, G. Haegeman, D. Iserentant, J. Merregaert, W.M. Jou, F. Molemans, A. Raeymaekers, A. Van den Berghe, et al. Complete nucleotide sequence of bacteriophage ms2 rna: primary and secondary structure of the replicase gene. *Nature*, 1976.

L.C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41, 1977.

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

L. A Garraway, H.R. Widlund, M.A. Rubin, G. Getz, A.J. Berger, S. Ramaswamy, R. Beroukhim, D.A. Milner, S.R. Granter, J. Du, et al. Integrative genomic analyses identify mitf as a lineage survival oncogene amplified in malignant melanoma. *Nature*, 436(7047):117–122, 2005.

J. Guo and S. Wang. Modularized gaussian graphical model. *Preprint submitted to Computational Statistics and Data Analysis*, 2010.

J. Guo, E. Levina, G. Michailidis, and J. Zhu. Joint estimation of multiple graphical models. *Biometrika*, 98(1):1–15, 2011.

H. Hoefling. A path algorithm for the fused lasso signal approximator. *Journal of Computational and Graphical Statistics*, 19(4):984–1006, 2010.

A. Jauhiainen, O. Nerman, G. Michailidis, and R. Jörnsten. Transcriptional and metabolic data integration and modeling for identification of active pathways. *Biostatistics*, pages 1–14, 2012.

D.B. Johnson. Efficient algorithms for shortest paths in sparse networks. *Journal of the ACM (JACM)*, 24(1):1–13, 1977.

R. Jörnsten. Simultaneous model selection via rate-distortion theory, with applications to cluster and significance analysis of gene expression data. *Journal of Computational and Graphical Statistics*, 18(3):613–639, 2009.

R. Jörnsten, T. Abenius, L. Kling, T. ans Schmidt, E. Johansson, B. Nordling, T. Nordlander, Chris. Sander, P. Gennemark, K. Funa,

B. Nilsson, L. Lindahl, and S. Nelander. Network modeling of the transcriptional effects of copy number aberrations in glioblastoma. *Molecular Systems Biology*, 7(486):485–516, 2011.

SD. Kendall, CM. Linardic, SJ. Adam, and CM. Counter. A network of genetic events sufficient to convert normal human cells to a tumorigenic state. *Cancer Research*, 65:9824–9828, 2005.

P. Langfelder and S. Horvath. Wgcna: an r package for weighted correlation network analysis. *BMC Bioinformatics*, 9(559), 2008.

KM. Mani, C. Lefebvre, K. Wang, WK. Lim, K. Baso, and et al. A systems biology approach to prediction of oncogenes and molecular perturbation targets in b-cell lymphomas. *Molecular Systems Biology*, 4(169), 2008.

AA. Margolin, I. Nemenman, K. Basso, C. Wiggins, and et al. Stolovitzky, G. Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7(Supl. 1), 2006.

N. Meinshausen and P. Buhlmann. High dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34:1436–1462, 2006.

K. Murphy. The bayes net toolbox for matlab. *Computing Science and Statistics*, 33:1024–1034, 2001.

P. Myllymäki, T. Silander, H. Tirri, and P. Uronen. B-course: A web-based tool for bayesian and causal data analysis. *International Journal on Artificial Intelligence Tools*, 11(3):369–388, 2002.

RK. Nibbe, M. Koyuturk, and MR. Chance. An integrative -omics approach to identify functional sub-networks in human colorectal cancer. *PLoS Computational Biology*, 6(1):1–15, 2010.

L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. *Stanford InfoLab Technical Report*, 1999.

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

J. Peng, J. Zhu, A. Bergamaschi, W. Han, D. Noh, J.R. Pollack, and P. Wang. Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *The annals of applied statistics*, 4(1):53, 2010.

G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6 (2):461–464, 1978.

N. Slavov and KA. Dawson. Correlation signature of the macroscopic states of the gene regulatory network in cancer. *Proceedings of the National Academy of Sciences of the United States of America*, 106(11):4079–4084, 2009.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, 58(1):267–288, 1996.

AV. Werhli, M. Grzegorczyk, and D. Husmeier. Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks. *Bioinformatics*, 22(20): 2523–2531, 2006.

A. M. Yip and S. Horvath. Gene network interconnectedness and the generalized topological overlap measure. *BMC Bioinformatics*, 8:22, 2007.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68: 49–67, 2006.

M. Yuan and Y. Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94:19–35, 2007.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B*, 67(Part 2):301–320, 2008.