THESIS FOR THE DEGREE OF LICENTIATE OF PHILOSOPHY

Comparative network analysis of human cancer: sparse graphical models with modular constraints and sample size correction

José Sánchez

Division of Mathematical Statistics
Department of Mathematical Sciences
Chalmers University of Technology and the University of Gothenburg
Göteborg, Sweden 2013

Comparative network analysis of human cancer: sparse graphical models with modular constraints and sample size correction
José Sánchez
NO 2013:5
ISSN 1652-9715

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 2013

Comparative network analysis of human cancer: sparse graphical models with modular constraints and sample size correction

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

Abstract

In the study of transcriptional data for different groups (e.g. cancer types) it's reasonable to assume that some dependencies between genes on a transcriptional or genetic variants level are common across groups. Also, that this property is preserved locally, thus defining a modular structure in the model networks. For ease of interpretation, sparsity in the resulting model is also desirable. In this thesis we assume genomic data to have a multivariate normal distribution and estimate the networks by optimization of a penalized log-likelihood function for the corresponding inverse covariance matrices. We apply the fused elastic net penalty for sparsity and commonality. To achieve modular topology we propose a novel adaptive penalty. This adaptive penalty is computed from an initial zero-consistent solution. We also propose a generalization of the method which allows for fusion penalties defined by a graph. This method can be used to correct estimates when the groups have different sample sizes. It can also be use to correctly penalize in the presence of ordered variables such as survival. We optimize the penalized log-likelihood using the alternating directions method of multiplier (ADMM). Simulation studies show that our method more accurately identifies differential connectivity (network edges that differ between cancer classes) compared with standard methods. We also apply our method to the investigation of tumor data in glioblastoma, breast and ovarian cancer, integrating two types of data, mRNA (messenger RNA expression) and CNA (copy number aberration), by defining a prior distribution of the plausible links in the corresponding networks.

Keywords: Inverse covariance matrix, precision matrix, graphical models, high-dimension, low-sample, networks, sparsity, fused lasso, elastic net, cancer.

Acknowledgments

I would like to thank my supervisor Rebecka Jörnsten for her support and the fruitful discussions. She is always full of new ideas to explore, even when the deadline is a few hours ahead! I would also like to thank my co-supervisor Sven Nelander for his lively-biological contributions. This work is focus on solving a mathematical problem that, turns out, has important applications in systems biology as I have learnt from Sven. But I'm not alone with the two captains of this ship, there are also my fellow team members Teresia Kling, Patrik Johansson and Tobias Abenius. This work is, in the end, the result of a great collaboration with them.

I have always thought that collaboration with people working in different areas enriches research. For this reason, and many others, I would like to thank my fellow Ph.D. students and colleagues at the department of Mathematical Sciences at Chalmers.

Apparently, the first written reference to the sentence "last but not least" is found in John Lyly's Euphues and His England, 1580.

I have heard oftentimes that in love there are three things for to be used: if time serve, violence, if wealth be great, gold, if necessity compel, sorcery. But of these three but one can stand me in stead - the last, but not the least'; which is able to work the minds of all women like wax.

I don't quite agree with Lyly about the use of sorcery, but everybody knows I have to finish my acknowledgments with a "last but not least" sentence. So, last but not least, thanks to my family and friends for support, patience and the magic they work in my life. Hopefully my own mind hasn't turned to wax after this!

Contents

Chapter 1

Introduction

1.1 Background

According to the central dogma of molecular biology (Figure 1.1) which describes the flow of information within a biological system, the transfer of information from a protein to either DNA or RNA is not possible. This fact establishes a framework for the study of complex biological processes, such as cancer, at a molecular level.

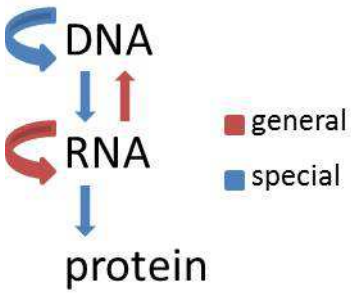


Figure 1.1: Blue arrows indicate general transfers of information (believed to occur in most cells). Red arrows indicate special transfers (known to occur under specific conditions such as lab experiments).

The study of cancer at molecular level is a relatively new area. It is known, however, that cancer is caused by anomalies in the genome (or genetic alterations) that result in an uncontrolled growth of cells. One of the main challenges of cancer systems biology is to understand the complex molecular changes that cancerous cells and tissues undergo during the formation of a

tumor and how these events are connected. This information can be used, in turn, in the development of new targeted therapies.

1.1.1 Genetic Alterations

The genetic alterations that take place during the formation of a tumor can be of different types, such as single nucleotide variants (SNV), copy number alterations (CNA), loss of heterozygosity (LOH) or altered methylation levels, among others.

- **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.

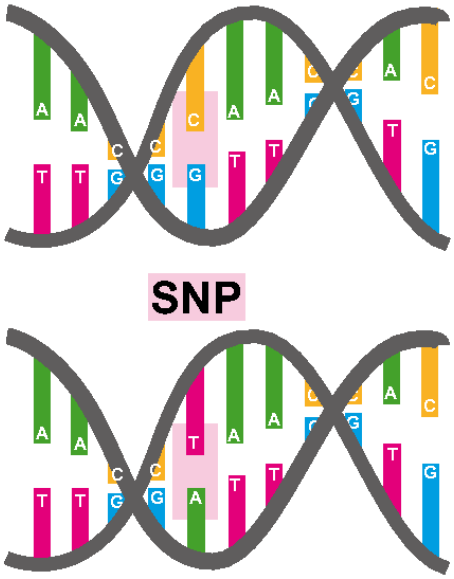


Figure 1.2: SNV: The two molecules of DNA differ at the highlighted base-pair location (a C/T polymorphism).

- **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). The most common ones are deletions (thus fewer copies than normal) and duplications (more than the normal number).

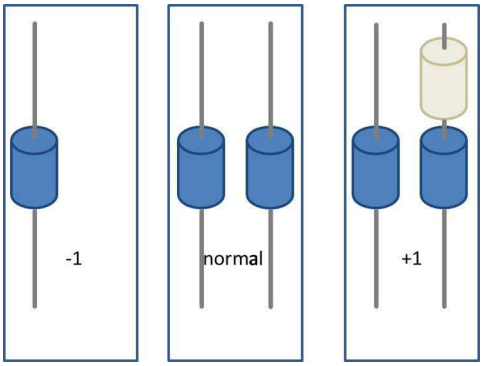


Figure 1.3: The cylinders represent a region of the genome. On the left side a deletion, and on the right side a duplication.

- **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.

1.1.2 Genetical Data

Collection of genetic data has grown in recent years and will continue. In the next few years we will have access to comprehensive observations of molecular changes for different types of cancers thanks to the work of consortia such as the Cancer Genome Atlas (TCGA), the Cancer Genome Project (CGP), the International Cancer Genome Consortium (ICGC) and the Uppsala-Umeå Comprehensive Cancer Consortium (U-CAN).

- **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.
- **CGP.** The Cancer Genome Project is using the human genome sequence and high throughput mutation detection techniques to identify somatically acquired sequence variants/mutations and hence identify genes which are critical to the development of human cancers. This initiative will ultimately provide the paradigm for the detection of

germline mutations in non-neoplastic human genetic diseases through genome-wide mutation detection approaches.

- **ICGC.** The primary goals of the International Cancer Genome Consortium are to generate comprehensive catalogues of genomic abnormalities (somatic mutations, abnormal expression of genes, epigenetic modifications) in tumors from 50 different cancer types or subtypes.

- **U-CAN.** The U-CAN 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.

The systems biology approach has an interdisciplinary perspective, as opposed to a more traditional reductionistic one, to biological and biomedical research. In this sense, integration of different data types is an important aspect of systems biology and it is where mathematical models come to use. Gene network modeling, for example, has proved helpful to integrate several levels of genomic cancer data and helped in some important problems such as (Abenius et al., 2012):

- identification of genes with altered copy number as disease drivers,

- construction of features, based on molecular data, for prediction of patient survival, and

- discovery of possible therapeutic targets based on matching hubs in the networks to pharmacological databases.

1.1.3 Cancer types

Here we focus on transcriptional data, mRNA, and one genetic alteration, namely copy number alteration, CNA. In this thesis we will use data from TCGA for glioblastoma multiforme, breast cancer and ovarian cancer.

- **Glioblastoma multiforme** is the most common and aggressive malignant brain tumor in adults. It affects 2/100000 to 3/100000 people per year in Europe and North America. The prognosis is poor, with a median survival time of 12 to 14 months.

- **Breast cancer.** The great majority of breast cancer cases occur in women, but 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. Survival rates in the western world are high compared to those of other cancer types.

- **Ovarian cancer.** 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 because it lacks any clear early detection or screening test.

1.2 Gene Regulatory Networks

A gene regulatory network describes how genes interact with each other to form modules and carry out cell functions. They can help us, by describing the implied dependencies for the genes, in systematically understanding complex molecular mechanisms for certain biological processes.

Of special interest are genes that interact with many others, called hub genes. Recent analysis of hub genes has shown them to be possible disease drivers, particularly identifying them as key tumorigenic genes (Kendall et al., 2005; Mani et al., 2008; Nibbe et al., 2010; Slavov and Dawson, 2009).

The methods used for estimation of gene regulatory networks can be classified in four categories (Allen et al., 2012): Bayesian networks, information theory-based, correlation-based and partial correlation-based methods. The method proposed in this thesis falls into the latter category. Furthermore, we are interested in the joint estimation of multiple gene regulatory networks. More precisely, we will use partial correlation-based methods to jointly estimate multiple Gaussian graphical models. Due to the nature of the data we work with, we find biologically relevant to introduce some constraints on equality of the links across classes and modularity.

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* (Myllymaki et al., 2002), *BNT* (Murphy, 2001) and Werhli's implementation of *BN* (Werhli et al., 2006).

Information Theory-based Methods

These type of methods use mutual information to determine the dependencies between genes and remove 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 gene regulatory networks is by thresholding the sample covariance matrix to keep only the strongest connections between pairs of genes. An example of a correlation-based method is the *Weighted Correlation Network Analysis*, WGCNA (Langfelder and Horvath, 2008).

Partial Correlation-based Methods

Partial correlation-based methods make use of Gaussian graphical model-theory. The goal is to estimate the partial (conditional) correlation between genes given by the non-zero elements of the inverse covariance matrix.

In the next chapter we will give a review of the latest proposals in partial correlation-based methods for estimation of both single and multiple Gaussian graphical models. We will also explain in more detail our constraints of equality and modularity.

Chapter 2

Gene Network Estimation

Under the assumption of normality, the problem of estimating the partial correlations is equivalent to estimating the inverse covariance matrix (also called the precision matrix). For a single Gaussian graphical model this can been done in many different ways. Dempster (1972) formulated it as the combinatorial problem of optimizing the location of zeros in the matrix. Since such methods don't scale up to high dimensions, more recently the focus has been shifted to optimization of penalized likelihood functions. In Meinshausen and Buhlmann (2006), each variable is estimated through an L_1 penalized regression on the rest of the variables. Later on, extensions and generalizations were proposed by Yuan and Lin (2007a), Banerjee et al. (2008), D'Aspremont et al. (2008) and Friedman et al. (2008). All of these produce estimates of the inverse covariance matrix referred to nowadays as the *graphical lasso*.

Regarding the estimation of multiple graphical models, a relevant problem in the presence of data from several classes that share variables, but not necessarily structure, Guo et al. (2011), Yuan and Lin (2007a), and Guo and Wang (2010) have proposed methods to achieve a common structure (without equal values), but not necessarily common modules.

2.1 Review of Existing Methods

If we assume that transcriptional data from different groups (for example different cancer types) can be modeled as a realization of a multivariate normal distribution with mean μ and covariance matrix Σ^k, for all groups $k = 1, 2, \ldots, K$, then the problem of estimating the transcription networks is equivalent to estimating the precision matrices $\Theta^k = \left(\Sigma^k\right)^{-1}$,

$k = 1, 2, \ldots, K$. Specifically, transcription of gene i is conditionally independent of 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.

Assume for the moment we have only one class and a data set X of observations from $N(0, \Sigma)$ (we assume without loss of generality that the data is centered). Most of the recent methods to estimate the precision matrix $\Theta = \Sigma^{-1}$ are based on the optimization of a penalized version of the likelihood function (see for example Friedman et al. (2008))

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

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. When the number of variables p is larger than the number of observations n (precisely the case we are interested in here), a penalty that imposes sparsity is needed since the usual maximum likelihood estimate $\hat{\Sigma} = S$ is not positive definite and suffers from very high variance.

Many authors who have studied this problem optimize the penalized function

$$l(\Theta) = \ln(\det(\Theta)) - \operatorname{tr}(S\Theta) - \lambda\|\Theta\|_1,$$

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

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

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

According to the authors, the elastic net often outperforms the lasso while enjoying similar sparsity structure. It also has a grouping effect, in which strongly correlated variables tend to be zero, or not, simultaneously.

Here we are interested in common networks across K classes. Guo et al. (2011) proposed a method to achieve a common structure (without equal values), but not necessarily common modules. They find an approximate

solution to this problem by iteratively optimizing the K likelihood functions

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

where $\omega_{ij}^k = \left(\sum_{k=1}^K \mid \theta_{ij}^k \mid\right)^{-1/2}$. Thus the problem can 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. However, this approach doesn't fulfil all the constraints we are interested in, since it doesn't encourage modularity, nor does it guarantee equal values of the common entries of the precision matrices, only a certain number of common non-zeros.

A similar approach that also guarantees a similar pattern of sparsity, but not equal values for some of the non-zeros, is the sparse *group lasso* (Yuan and Lin, 2007b). 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 $\{\Theta\} = \{\Theta^1, \Theta^2, \ldots, \Theta^K\}$.

Guo and Wang (2010) suggested one way to introduce modularity, understood as a partition of the nodes in disjoint sets. First they estimated the precision matrix using *glasso* and used this estimate to define a dissimilarity matrix which was, in turn, used to find clusters of nodes in the corresponding network. In a second step, they estimated again the precision matrix by penalizing the log-likelihood function with either a double regularization penalty or a group penalty. By means of independent tuning parameters they control the sparsity within and between clusters. This procedure doesn't guarantee equality of common values.

One way to guarantee equal values for the common links, if not the common modules, is the *OSCAR*, described in Bondel and Reich (2008) in the context of linear regression. The *OSCAR* penalty can, in principle, be applied to our problem, which will then require optimizing the following function:

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

Here, the L_∞ norm encourages equality of coefficients. The drawback here is of practical nature, since this is a complicated function to optimize.

A more tractable way to approach the problem is suggested in Danaher et al. (2011). There, the log-likelihood takes the form

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

This time the equality in link values is encouraged by the fused penalty. Although the authors provide a closed solution to for the case with two classes, they don't provide an efficient algorithm for the general case.

In our case, for the model to have a relevant and useful interpretation from the biological point of view, some constraints must be imposed:

- A full (non-sparse) precision matrix is informative for output prediction but is hard to interpret. It also contains correlations of all strength levels, therefore sparsity restrictions must be imposed to pick out the strong correlations of interest.
- The diversity observed in cancer biology makes it plausible to assume that different cancers have different regulators, but it's possible that some of them are shared and their transcription will be, therefore, common across cancer types. From the network point of view, this means that some links will be unique, whereas some will be common to some or all cancer types.
- It is biologically sensible that the type of a given link (common across cancer types or unique) is a property that should be preserved locally, thus defining a *module* or a *modular* network structure.

In the next chapter we describe a method to solve the problem, which into account all these constraints.

Chapter 3

The adaptive penalty method

In the previous chapter we described the problem we are interested in and its equivalence to a constrained optimization problem. Sparsity and commonality of links are taken care of by the lasso and the fused penalty, respectively. Here we present details on an adaptive penalty that will encourage modularity.

3.1 Adaptive penalty schemes

Consider K data sets X^1, $X^2, \ldots, X^K$ with $K \geq 2$ corresponding to K classes. Data set X^k consists of n_k observations and p variables, which are common to all K data sets. We assume the observations within each data set to be i.i.d. $N(0, \Sigma^k)$. Let $\Theta^k = \left(\Sigma^k\right)^{-1}$ and S^k be the empirical covariance matrix for the k class, $k = 1, 2, \ldots, K$. We propose to optimize the penalized likelihood function

$$
\begin{aligned}
l\left(\{\Theta\}\right) = & \sum_{k=1}^{K} n_k \left[\ln\left(\det\left(\Theta^k\right)\right) - \operatorname{tr}\left(S^k\Theta^k\right)\right] \\
& - \lambda_1 \sum_{k=1}^{K} \sum_{i \neq j} \left[\alpha\left|\theta_{ij}^k\right| + (1-\alpha)\left(\theta_{ij}^k\right)^2\right] \\
& - \lambda_2 \sum_{k<k'} \sum_{i,j} \omega_{ij}^{kk'} \left|\theta_{ij}^k - \theta_{ij}^{k'}\right|,
\end{aligned}
\tag{3.1}
$$

where θ_{ij}^k is the ij element of Θ^k.

The first term of the penalty is the elastic net, which controls the overall sparsity level of all covariance matrices: the larger the tuning parameter λ_1 is,

the sparser the Θ^k are. The second term is the fused lasso penalty, there the parameter λ_2 controls the degree of commonality (links with the same value) across the K classes: the larger it is, the more the link value is preserved across classes. These two penalties take care of the first two constraints mentioned in the previous section however, one of the main contributions of this paper is given by the adaptivity factor $\omega_{ij}^{kk'}$, which will take care of the third constraint described in the previous section, and that we describe below.

To compute the adaptivity parameter we proceed in a similar way to Zou (2006) and his *adaptive lasso* scheme. The idea is to adapt the tuning parameter for the fused penalty, λ_2, based on an initial estimate of the network. Now, for the adaptive lasso to possess the oracle property as described in Fan and Li (2001), it is required that the initial estimate of the network is zero-consistent (estimators of zero link converge to zero in probability and estimators of non-zero links do not converge to zero). The fused lasso has this property (Sharma et al. (2012)) and, for that reason, we optimize (3.1) in two steps. First we compute the initial estimate of the networks using the usual fused lasso (that is, with $\omega_{ij}^{kk'} = 1\ \forall i, j$ and $\forall k, k'$) which we then use to compute a new $\omega_{ij}^{kk'}$. Then we optimize again (3.1) using these new values for the adaptivity parameter.

We propose four different adaptivity schemes, two global and two local ones.

Adaptivity I

We consider ω_{ij} to be the same for all pairwise comparisons in the fused penalty. For this reason we can drop the superindices kk' and define it as follows

$$\omega_{ij} = \left[\sum_{k<k'} |\tilde{\theta}_{ij}^k - \tilde{\theta}_{ij}^{k'}| \sum_{k<k'} \sum_{l \in N_{ij}} \left(|\tilde{\theta}_{il}^k - \tilde{\theta}_{il}^{k'}| + |\tilde{\theta}_{jl}^k - \tilde{\theta}_{jl}^{k'}| \right) \right]^{-\gamma}, \qquad (3.2)$$

where the $\tilde{\theta}_{ij}^k$ are the initial estimates of the network, N_{ij} denotes the set of neighbors of link (i, j), that is, the set of links connected to nodes i and j; γ is a positive tuning parameter. This adaptivity factor encourages fusing of link (i, j) for classes k and k' when they are already close or when its neighbors are (that is, when they have similar values) across classes, thus encouraging even more similarity.

As an example, consider Figure 3.1, where an initial estimate of a network is shown. The neighbors of link (i, j) (that between genes i and

j) are links marked as 1, 2, 3, and 4. Whenever link (i, j) itself is similar across some classes, the inverse $\sum_{k<k'} |\tilde{\theta}_{ij}^{k} - \tilde{\theta}_{ij}^{k'}|$ will be large. On the other hand, whenever links 1, 2, 3, or 4 are similar, the inverse of $\sum_{k<k'} \sum_{l \in N_{ij}} \left(|\tilde{\theta}_{il}^{k} - \tilde{\theta}_{il}^{k'}| + |\tilde{\theta}_{jl}^{k} - \tilde{\theta}_{jl}^{k'}| \right)$ will be large. The result is that ω_{ij} will be large whenever link (i, j) or its neighbors are similar, thus encouraging a local fused structure (modularity). If either link (i, j) or its neighbors are equal across all classes, ω_{ij} becomes infinite, in which case we define it as some large value.

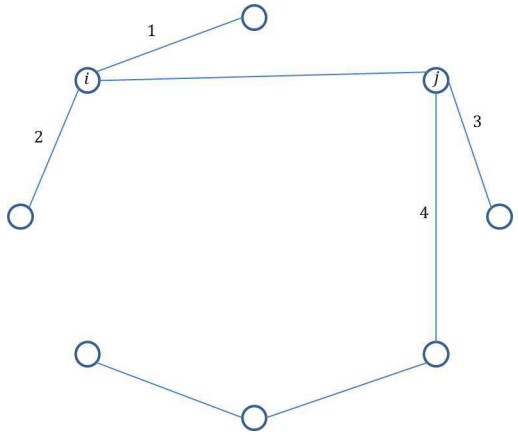


Figure 3.1: In this network the neighborhood of link (i, j) is comprised of links 1, 2, 3, and 4.

Adaptivity II

We want to further refine the adaptivity penalty. We notice that when using the above criterion to define ω_{ij}, we increase (or decrease) the fused penalty equally for all pairwise differences. This can result in neighbor links being equal across different subsets of all classes. To avoid that, we define pairwise specific $\omega_{ij}^{kk'}$ as

$$\omega_{ij}^{kk'} = \left[|\tilde{\theta}_{ij}^{k} - \tilde{\theta}_{ij}^{k'}| \sum_{l \in N_{ij}} \left(|\tilde{\theta}_{il}^{k} - \tilde{\theta}_{il}^{k'}| + |\tilde{\theta}_{jl}^{k} - \tilde{\theta}_{jl}^{k'}| \right) \right]^{-\gamma}. \tag{3.3}$$

Figure 3.2 shows an example for such situation. In this case link (i, j) is equal for classes 2, 4, and 6 so we would like to encourage fusing across these classes. However, its neighbors are equal across a different subset of classes and we would like to take that into account to obtain a "cleaner" module, in the sense that its links are equal across the same subset (or as similar as

possible). In this example ω_{ij}^{12} will be larger than any other $\omega_{ij}^{kk'}$, since 3 links in the neighborhood of link (i, j) are equal for classes 1 and 2.

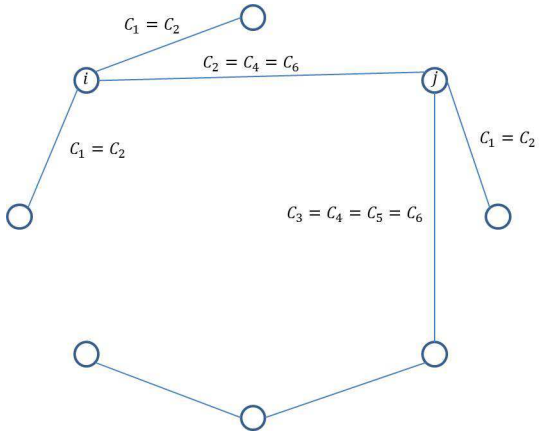


Figure 3.2: A network where the links in the neighborhood of link (i, j) are not equal across the same subset of classes.

Adaptivity III and IV

The last two adaptivity schemes are just local versions of the previous adaptivity factors. The idea is to encourage modularity even more by creating local versions the ω_{ij} and $\omega_{ij}^{kk'}$ defined above. To do that we first cluster the variables for all classes, using the union network of the initial estimate, thus finding, say, M clusters. We then define

$$\omega_{ij}^{(m)} = \left[\sum_{k<k'} |\tilde{\theta}_{ij}^{k} - \tilde{\theta}_{ij}^{k'}| \sum_{k<k'} \sum_{l \in N_{ij}^{(m)}} \left(|\tilde{\theta}_{il}^{k} - \tilde{\theta}_{il}^{k'}| + |\tilde{\theta}_{jl}^{k} - \tilde{\theta}_{jl}^{k'}| \right) \right]^{-\gamma}, \qquad (3.4)$$

where i, j and l belong to the m-th cluster and $m = 1, 2, \ldots, M$. Similarly,

$$\left(\omega_{ij}^{kk'}\right)^{(m)} = \left[|\tilde{\theta}_{ij}^{k} - \tilde{\theta}_{ij}^{k'}| \sum_{l \in N_{ij}^{(m)}} \left(|\tilde{\theta}_{il}^{k} - \tilde{\theta}_{il}^{k'}| + |\tilde{\theta}_{jl}^{k} - \tilde{\theta}_{jl}^{k'}| \right) \right]^{-\gamma}. \qquad (3.5)$$

Here too i, j and l belong to the m-th cluster and $m = 1, 2, \ldots, M$.

We will show in Chapter 5 simulation studies and results on real data using schemes I and II. However, we leave further analysis of schemes III and IV for future work. In the next chapter we present an iterative method to optimize the adaptive penalized log-likelihood using the alternating directions method of multipliers.

Chapter 4

Estimation of parameters

We are interested in applying the fused elastic net penalty to the log-likelihood of the inverse covariance matrices. To further encourage modularity we want to adapt the fused penalty using the adaptivity schemes introduced in Chapter 3. We optimize this penalized log-likelihood function in two steps. First we obtain a zero-consistent estimate from the regular fused elastic net (that is, without applying any adaptivity scheme) which is used to computed either (3.2) or (3.3), then we estimate again the networks using this adaptive penalty.

Each optimization step is done using the *alternating directions method of multipliers*, ADMM. For a complete description of the method see Boyd et al. (2011). We will present first an algorithm to solve the problem when the elastic net penalty is not class specific (that is, λ_1 is equal $\forall k = 1, 2, \ldots, K$) and the fused penalty is not pair-specific (that is, $\omega_{ij}^{kk'}$ is equal $\forall k, k' = 1, 2, \ldots, K$) since it's simpler. Nevertheless, adaptivity scheme II (and IV) require pair-specific penalties. Moreover, as we will see later, relative sample sizes have an effect in the sparsity structure that can be alleviated by using class-specific elastic net penalties. For this reason, we will present a slightly different algorithm to solve this problem with more general penalties.

4.1 The ADMM algorithm

To optimize the penalized likelihood problem using ADMM we note first that it can be written as

$$\begin{aligned} \underset{\{\Theta\},\{Z\}}{\text{minimize}} \quad & f(\{\Theta\}) + g(\lambda, \{Z\}) \\ \text{subject to} \quad & \Theta^k = Z^k,\ k = 1, \ldots, K. \end{aligned}$$

(Danaher et al., 2011) where

$$f(\{\Theta\}) = \sum_{k=1}^{K} n_k \left[\mathrm{tr}(S^k \Theta^k) - \ln \left(\det(\Theta^k) \right) \right]$$
$$g(\lambda, \{Z\}) = \lambda_1 \sum_{k=1}^{K} \sum_{i \neq j} \left[\alpha \left| Z_{ij}^k \right| + (1-\alpha) Z_{ij}^2 \right] + \lambda_2 \sum_{k<k'} \sum_{i,j} \left| Z_{ij}^k - Z_{ij}^{k'} \right|.$$

ADMM solves this problem by defining the scaled augmented lagrangian as follows:

$$L(\{\Theta\}, \{Z\}, \{U\}) = f(\{\Theta\}) + g(\lambda, \{Z\}) + \frac{\rho}{2} \sum_{k=1}^{K} \|\Theta^k - Z^k + U^k\|_F^2,$$

where U^k are the dual variables.

Then minimization is done iteratively in three steps. At iteration m, the variables $\{\Theta\}$, $\{Z\}$ and $\{U\}$ are updated according to

1. $\Theta_m^k \leftarrow \arg\min_{\{\Theta\}} \{L(\{\Theta\}, \{Z_{m-1}\}, \{U_{m-1}\})\}$
2. $Z_m^k \leftarrow \arg\min_{\{Z\}} \{L(\{\Theta_m\}, \{Z\}, \{U_{m-1}\})\}$
3. $U_m^k \leftarrow U_{m-1}^k + \Theta_m^k - Z_m^k$

for $k = 1, \ldots, K$. We now present details of the first two steps. We omit the iteration subindex to simplify the notation.

For the first step, function g is a constant, so the problem is to minimize the function

$$\sum_{k=1}^{K} n_k \left[\mathrm{tr}(S^k \Theta^k) - \ln \left(\det(\Theta^k) \right) \right] + \frac{\rho}{2} \sum_{k=1}^{K} \|\Theta^k - Z^k + U^k\|_F^2 \qquad (4.1)$$

with respect to Θ. Let VDV^T be the SVD decomposition of $\rho/n_k(Z^k - U^k) - S^k$. The minimizer of (4.1) is given (Witten and Tibshirani, 2009) by $V\tilde{D}V^T$ where $\tilde{D}$ is a diagonal matrix with elements $n_k/2\rho(D_{jj} + \sqrt{D_{jj}^2 + 4\rho/nk})$.

For the second step, function f is a constant, so the problem is to minimize

the function

$$
\begin{aligned}
& g(\lambda, \{Z\}) + \frac{\rho}{2} \sum_{k=1}^{K} \|\Theta^k - Z^k + U^k\|_F^2 \\
=& \frac{\rho}{2} \sum_{k=1}^{K} \|Z^k - A^k\|_F^2 + \lambda_1 \sum_{k=1}^{K} \sum_{i \neq j} \left[\alpha |Z_{ij}^k| + (1-\alpha) \left(Z_{ij}^k \right)^2 \right] \\
& + \lambda_2 \sum_{k<k'} \sum_{i,j} |Z_{ij}^k - Z_{ij}^{k'}|
\end{aligned}
$$

with respect to Z, where $A^k = \Theta^k + U^k$.

This problem is separable for each element (i, j), so we can solve separately as

$$
\begin{aligned}
& \underset{\{Z_{ij}\}}{\text{minimize}} \Bigg\{ \frac{1}{2} \sum_{k=1}^{K} \left(Z_{ij}^k - A_{ij}^k \right)^2 \\
& + \frac{\lambda_1}{\rho} \mathrm{I}_{i \neq j} \sum_{k=1}^{K} \left[\alpha |Z_{ij}^k| + (1-\alpha) \left(Z_{ij}^k \right)^2 \right] + \frac{\lambda_2}{\rho} \sum_{k<k'} |Z_{ij}^k - Z_{ij}^{k'}| \Bigg\}
\end{aligned}
\tag{4.2}
$$

This is a case of the *fused lasso signal approximator* (Hoefling, 2010) where all pairwise differences are penalized and regularized with the elastic net. When $K = 2$, and $\lambda_1 = 0$, the problem has a closed form solution (Danaher et al., 2011),

$$
(Z_{ij}^1, Z_{ij}^2) = \begin{cases} \left(A_{ij}^1 - \lambda_2/\rho, A_{ij}^2 + \lambda_2/\rho\right) & \text{if} \quad A_{ij}^1 > A_{ij}^2 + 2\lambda_2/\rho \\ \left(A_{ij}^1 + \lambda_2/\rho, A_{ij}^2 - \lambda_2/\rho\right) & \text{if} \quad A_{ij}^2 > A_{ij}^1 + 2\lambda_2/\rho \\ \left(\frac{1}{2}(A_{ij}^1 + A_{ij}^2), \frac{1}{2}(A_{ij}^1 + A_{ij}^2)\right) & \text{if} \quad |A_{ij}^1 - A_{ij}^2| \leq \lambda_2/\rho, \end{cases} \tag{4.3}
$$

and the solution for $\lambda_1 > 0$, Z_{ij}^*, can be found by soft-thresholding (Friedman et al., 2007; Zou and Hastie, 2008) (4.3) according to $Z_{ij}^* = \alpha ST_{\lambda_1}(Z_{ij})/(1 + (1-\alpha)\lambda_1)$. The soft-threshold function is defined as: $ST_\lambda(x) = \text{sign}(x) \max(|x| - \lambda)$.

To solve the case when $K > 2$ we follow Hoefling (2012). We focus first on the fused penalty and drop the subindexes ij and ρ to simplify notation. Proposition 1 in Hoefling (2012) ensures that the order of the coefficients in the solution Z^k is the same as the order of the A^k. Assume now, without loss of generality, that $A^1 \leq A^2 \leq \ldots \leq A^K$. Then, taking $\lambda_1 = 0$ (as before, the solution for $\lambda_1 > 0$ can be found afterwards by soft-thresholding), the objective function becomes

$$
\begin{aligned}
L(\{Z\}) = \frac{1}{2} \sum_{k=1}^{K} \left(Z^k - A^k \right)^2 + \lambda_2 \sum_{k>k'} (Z^k - Z^{k'}) = \\
\sum_{k=1}^{K} \left[\frac{1}{2} \left(Z^k - A^k \right)^2 + \lambda_2 \left(2k - (K+1) \right) Z^k \right]
\end{aligned}
\tag{4.4}
$$

subject to $Z^1 \leq Z^2 \leq \ldots \leq Z^K$, and we see that the variables are separable.

If the constraint on the order of Z is fulfilled, then the global solution, found by setting the derivative of (4.4) with respect to Z^k equal to zero, is given by $Z_*^k = A^k - \lambda_2 (2k - (K+1))$. If, for a certain index k_0, $Z_*^{k_0} > Z_*^{k_0+1}$, then Proposition 2 in Hoefling (2012) proves that $Z^{k_0} = Z^{k_0+1} = (Z_*^{k_0} + Z_*^{k_0+1})/2$.

Algorithm for the adaptive penalty method

We sketch below the algorithm to solve the initial fused elastic net problem.

Algorithm 1

for $k = 1 \rightarrow K$ **do**
$Z^k \leftarrow (S^k + \epsilon \mathrm{I})^{-1}$
$U^k \leftarrow \mathbf{0}$
end for
while convergence$\neq$TRUE **do**
for $k = 1 \rightarrow K$ **do**

$$\Theta_i^K \leftarrow \arg\min_{\{\Theta\}} \left\{ \sum_{k=1}^{K} n_k \left[\mathrm{tr}(S^k \Theta^k) - \ln\left(\det(\Theta^k)\right) \right] + \frac{\rho}{2} \sum_{k=1}^{K} \|\Theta^k - Z_{i-1}^k + U_{i-1}^k\|_F^2 \right\}$$

$$Z_i^K \leftarrow \arg\min_{\{Z\}} \left\{ \frac{\rho}{2} \sum_{k=1}^{K} \|\Theta_i^k - Z^k + U_{i-1}^k\|_F^2 + \lambda_1 \sum_{k=1}^{K} \sum_{i \neq j} \left[\alpha |Z_{ij}^k| + (1-\alpha) \left(Z_{ij}^k\right)^2 \right] + \lambda_2 \sum_{k<k'} \sum_{i,j} |Z_{ij}^k - Z_{ij}^{k'}| \right\}$$

$$U_i^k \leftarrow U_{i-1}^k + \Theta_i^k - Z_i^k$$

end for
end while

To solve the problem with adaptive penalty we just need to run Algorithm 1 first, update the penalties according to (3.2) and then run again Algorithm 1 using these updated penalties. Furthermore, we can decrease the execution time by using the solution from Algorithm 1 as a warm start. The complete algorithm is thus:

Algorithm 2

$\{\Theta\} \leftarrow$ Algorithm 1$(\{S\}, \lambda_1, \lambda_2)$
$\lambda_2 \leftarrow$ update(λ_2)
$\{\Theta\} \leftarrow$ Algorithm 1$(\{\Theta\}, \lambda_1, \lambda_2)$

where the update of λ_2 is computed, according to (3.2).

4.2 Faster computations through vectorization

The second step of the ADMM algorithm, where the matrix Z is updated, assumes elementwise updating of each element ij, since the K classes need to be sorted so that $A_{ij}^1 \leq A_{ij}^2 \leq \ldots \leq A_{ij}^K$. This results in very slow computations when the number of variables p is large. Here we present a vectorization method for faster computations of the update of Z.

1. Let B be a reshaped version of A, where B is a $p^2 \times K$ matrix containing all elements from A as rows and the K classes as columns:

$$B_{(i-1)p+j,k} = A_{i,j}^k.$$

2. Sort each row of B and save the order of the sorting for future use.

3. Let I be the $p^2 \times K$ matrix

$$I = \begin{bmatrix} 1 & 2 & 3 & \cdots & K \\ 1 & 2 & 3 & \cdots & K \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 2 & 3 & \cdots & K \end{bmatrix}.$$

4. Calculate the derivative solutions (Z_*) in a matrix form: $b = B - \lambda_2(2I - K - 1)$.

5. Initialize the following variables:

   - bI a $p^2 \times K$ boolean matrix with FALSE.
   - mv a $p^2 \times 1$ vector with zeros.
   - lf a $p^2 \times 1$ vector with ones.
   - l a $p^2 \times 1$ vector with ones.

6. Apply the following algorithm:

for $iter = 1 \rightarrow K$ **do**
  for $k = 1 \rightarrow K - 1$ **do**
    $fuse \leftarrow b_{.k} > b_{.k+1}$ ▷ $fuse$ = boolean $p^2 \times 1$ vector
    $bI_{.k} \leftarrow fuse$
    $bI_{.k+1} \leftarrow fuse$
    $lf(fuse) \leftarrow lf(fuse) + 1$ ▷ Counter of how many classes that are being *fused*
    $lf(notfuse) \leftarrow 1$

$mv(fuse) \leftarrow \frac{b(fuse,k).*(lf(fuse)-1)+b(fuse,k+1)}{lf(fuse)}$ ▷ Calculates the new values for the *fused* elements
$update \leftarrow repmat(mv, 1, K). * bI$
$b(bI) \leftarrow update(bI)$
end for
end for

7. Reorder b with the help of the sorting kept from step 2.

8. To get the final solution Z for the current iteration, reshape b into K $p \times p$ matrices.

4.3 The effect of different sample sizes

The sparsity and fusing levels of the solution depend on the sample sizes for each class. To see this, consider the penalized log-likelihood function of our problem (3.1). It can be rewritten as

$$\begin{aligned} l\left(\{\Theta\}\right) = & \sum_{k=1}^{K} n_k \left[\ln\left(\det\left(\Theta^k\right)\right) - \operatorname{tr}\left(S^k\Theta^k\right)\right] \\ & - \lambda_1 \sum_{k=1}^{K}\sum_{i\neq j}\left[\alpha\left|\theta_{ij}^k\right| + (1-\alpha)\left(\theta_{ij}^k\right)^2\right] \\ & - \frac{\lambda_2}{2}\sum_{k=1}^{K}\sum_{k'=1}^{K}\sum_{i,j}\omega_{ij}^{kk'}\left|\theta_{ij}^k - \theta_{ij}^{k'}\right| \\ & = \sum_{k=1}^{K} n_k \Bigg\{ \ln\left(\det\left(\Theta^k\right)\right) - \operatorname{tr}\left(S^k\Theta^k\right) \\ & - \frac{\lambda_1}{n_k}\sum_{i\neq j}\left[\alpha\left|\theta_{ij}^k\right| + (1-\alpha)\left(\theta_{ij}^k\right)^2\right] \\ & - \frac{\lambda_2}{2n_k}\sum_{k'=1}^{K}\sum_{i,j}\omega_{ij}^{kk'}\left|\theta_{ij}^k - \theta_{ij}^{k'}\right| \Bigg\}. \end{aligned}$$

The effective elastic net penalty for class k is thus given by λ_1/n_k. This penalty makes the estimates of classes with smaller sample sizes sparser, in comparison to those with larger sample sizes.

Consider now a specific pair of classes k and k' and a specific link (i, j), the

elements in the fused penalty which include these are

$$\begin{aligned}
&\frac{\lambda_2}{n_k}\sum_{l\neq k}^{K}\left|\theta_{ij}^{k}-\theta_{ij}^{l}\right|+\frac{\lambda_2}{n_{k'}}\sum_{l\neq k'}^{K}\left|\theta_{ij}^{k'}-\theta_{ij}^{l}\right| \\
&=\cdots+\frac{\lambda_2}{n_k}\left|\theta_{ij}^{k}-\theta_{ij}^{k'}\right|+\frac{\lambda_2}{n_{k'}}\left|\theta_{ij}^{k'}-\theta_{ij}^{k}\right|+\cdots \\
&=\lambda_2\left(\frac{1}{2n_k}+\frac{1}{2n_{k'}}\right)\left|\theta_{ij}^{k}-\theta_{ij}^{k'}\right|=\lambda_2\frac{n_k+n_{k'}}{2n_kn_{k'}}\left|\theta_{ij}^{k}-\theta_{ij}^{k'}\right|,
\end{aligned}$$

so the effective fused penalty for the pair k, k' is $\lambda_2\frac{n_k+n_{k'}}{2n_kn_{k'}}$. This penalty makes classes with smaller sample sizes fuse faster towards each other than to classes with larger sample sizes.

We think is reasonable to assume that the networks for different cancer classes have similar sparsity levels. Our approach is then to *correct* the sparsity and the fusing tuning parameters λ_1 and λ_2, respectively, so that classes with larger sample sizes don't dominate the estimation of all networks. This way we obtain similar sparsity levels. We think, however, that this correction shouldn't be used when the sample sizes are too different, since it can, potentially, create many false positives.

A natural way to correct for the sample size effects is to multiply the tuning parameters λ_1 and λ_2 by the inverses of n_k and $\frac{n_k+n_{k'}}{2n_kn_{k'}}$ respectively, however, in our experience this can overcorrect. We proceed instead as follows. Define an effective sample size n_k^e for class k as $n_k^e=\bar{n}^{\delta}n_k^{(1-\delta)}$, where $\bar{n}=\frac{1}{K}\sum_{k=1}^{k}n_k$ and $0\leq\delta\leq 1$ is the correction parameter. We penalize the log-likelihood with the effective penalty parameters $\lambda_1^k=\lambda_1 n_k^e$ and $\lambda_2^{kk'}=\lambda_2\frac{2n_k^en_{k'}^e}{n_k^e+n_{k'}^e}$. In the process, the tuning parameter for the elastic net penalty becomes class specific and the tuning parameter for the fused penalty becomes pair specific. This way, when $\delta=1$, λ_1^k are all equal $\forall k=1,2,\dots K$ and no sample size correction is done. When $\delta=0$ $n_k^e=n_k$, using the real sample size as correction factor. Selecting a value $0\leq\delta\leq 1$ we can achieve similar sparsity levels, but of course it will be data dependent.

An example, we apply the correction method for a real data set with 6 classes: glioblastoma, ovarian, breast, head and neck, uterine and kidney cancer. The sample sizes of these classes are 254, 307, 337, 394, 498 and 646 respectively. In Table 4.1 we show the number of links in the estimated networks for different values of δ.

It is difficult to obtain networks with exactly the same sparsity levels, since the sample sizes are so different (the largest sample is about 2.5 times larger than the smallest). Taking $\delta=0$, or too small, can overcorrects and give opposite results, making the estimates of the classes with larger sample sizes

δ	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6
0	494	63	368	69	97	219
0.02	352	57	318	67	115	247
0.04	245	49	271	65	128	277
0.06	154	39	223	65	153	371
0.08	104	30	193	64	70	467
0.1	61	23	154	62	189	599
0.12	37	18	125	59	81	673
0.14	28	15	100	55	230	922
0.16	20	12	78	54	254	939
0.18	8	10	61	50	101	1058
0.2	1	8	47	50	304	1080
0.22	1	6	41	48	334	1313
0.24	0	5	31	47	362	1256
0.26	0	4	22	45	399	1194
0.28	0	1	14	43	423	1309
0.3	0	1	12	41	463	4375

Table 4.1: Number of links present in estimated networks for different values of δ. $\delta = 0.04$ minimizes the variance in estimated number of links.

to be the most sparse. Not doing correction results in empty networks for the classes with smaller sample sizes. For this particular data set, $\delta = 0.04$ is the best value since it reduces the variability in the sparsity of the estimated networks.

4.4 ADMM algorithm for class and pair-specific penalties

We now present an algorithm to optimize the log-likelihood function with class and pair specific penalties. With this algorithm we have the possibility of using adaptive scheme II (3.3) and correct for unequal sample sizes.

The log-likelihood function with class-specific elastic net penalty and pair-

specific fuse penalty is

$$
\begin{aligned}
l\left(\{\Theta\}\right) = & \sum_{k=1}^{K} n_k \left[\ln\left(\det\left(\Theta^k\right)\right) - \operatorname{tr}\left(S^k\Theta^k\right)\right] \\
& - \sum_{k=1}^{K}\sum_{i\neq j} \lambda_{1,ij}^{k}\left[\alpha\left|\theta_{ij}^{k}\right| + (1-\alpha)\left(\theta_{ij}^{k}\right)^2\right] \\
& - \sum_{k<k'}\sum_{i,j} \lambda_{2,ij}^{kk'}\left|\theta_{ij}^{k} - \theta_{ij}^{k'}\right|.
\end{aligned}
\tag{4.5}
$$

We proceed similarly to Algorithm 1 to minimize (4.5), but this time the update for the $\{Z\}$ matrices is also done by ADMM. We present now the details for this update following Ye and Xie (2011).

Here we drop the subindex ij to simplify notation, but we keep the superindex that denotes class, thus, in the following, Z and A should be interpreted as vectors in $\mathbb{R}^K$). Consider thus a link (i, j), we need to solve the problem

$$
\begin{aligned}
\underset{Z}{\text{minimize}} \Bigg\{ & \frac{1}{2}\sum_{k=1}^{K}\left(Z^k - A^k\right)^2 \\
& + \sum_{k=1}^{K}\lambda_1^k\left[\alpha|Z^k| + (1-\alpha)\left(Z^k\right)^2\right] + \sum_{k<k'}\lambda_2^{kk'}|Z^k - Z^{k'}|\Bigg\}.
\end{aligned}
\tag{4.6}
$$

Let

$$
\begin{aligned}
f(Z) &= \frac{1}{2}\sum_{k=1}^{K}\left(Z^k - A^k\right)^2 \\
g(Z) &= \sum_{k=1}^{K}\lambda_1^k\left[\alpha|Z^k| + (1-\alpha)\left(Z^k\right)^2\right] \\
h(Z) &= \sum_{k<k'}\lambda_2^{kk'}|Z^k - Z^{k'}| = \|\Lambda_2 L Z\|_1,
\end{aligned}
$$

where $\Lambda_2 = (\lambda_2^{kk'})$ is a vector of dimension $\frac{1}{2}K(K+1)$ and L is a $\frac{1}{2}K(K+1)$-by-K matrix with values in $\{-1, 0, 1\}$ corresponding to the pairwise differences to be penalized. Problem (4.6) can be written as

$$
\begin{aligned}
\underset{Z}{\text{minimize}} \quad & f(Z) + g(V) + h(W) \\
\text{subject to} \quad & V = Z \\
& W = LZ.
\end{aligned}
$$

In this occasion we need to introduce two dual variables P and Q for the augmented Lagrangian. At iteration m we update the values of the variables according to

$$\begin{aligned}
Z_m &\leftarrow \text{argmin}_Z \Big\{ f(Z) + \langle P_{m-1}, Z - V_{m-1} \rangle + \langle Q, LZ - W_{m-1} \rangle \\
&\qquad + \frac{\rho_1}{2} \|Z - V_{m-1}\|_2^2 + \frac{\rho_2}{2} \|LZ - W_{m-1}\|_2^2 \Big\} \\
V_m &\leftarrow \text{argmin}_V \left\{ g(V) + \langle P_{m-1}, Z_m - V \rangle + \frac{\rho_1}{2} \|Z_m - V\|_2^2 \right\} \\
W_m &\leftarrow \text{argmin}_W \left\{ h(W) + \langle Q_{m-1}, LZ_m - W \rangle + \frac{\rho_2}{2} \|LZ_m - W\|_2^2 \right\} \\
P_m &\leftarrow P_{m-1} + \rho_1 (Z_m - V_m) \\
Q_m &\leftarrow Q_{m-1} + \rho_2 (LZ_m - W_m),
\end{aligned}$$

where $\langle x, y \rangle$ denotes the inner product of vectors x and y.

The solution to the updating problem for Z_m, which corresponds to a system of linear equations; and the solutions to the updating problems for V_m and W_m, which are given by soft-thresholding, can be computed as follows:

$$\begin{aligned}
Z &= \left[(\rho_1 + 1) I + \rho_2 L^T L \right]^{-1} \left[A + \rho_1 \left(V - \frac{1}{\rho_1} P \right) + \rho_2 L^T \left(W - \frac{1}{\rho_2} Q \right) \right] \qquad (4.7) \\
V &= ST_{\lambda_1/\rho_1} \left(Z + \frac{1}{\rho_1} P \right) \\
W &= ST_{\lambda_2/\rho_2} \left(LZ + \frac{1}{\rho_2} Q \right).
\end{aligned}$$

Although problem (4.6) and its ADMM updates (4.7) were specified for one gene only, we can solve for all the genes in our data set at the same time. We reshape the data matrices $\{A\}$ to a rectangular matrix with K rows, corresponding to each one of the classes, and $\frac{1}{2}p(p+1)$ columns, the links in the lower triangular part across all $\{A\}$. The update for Z in (4.7) requires us to solve $\frac{1}{2}p(p+1)$ linear systems. All of these have the same left-hand-side matrix, $\left((\rho_1 + 1)I + \rho_2 L^T L\right)$, and $\frac{1}{2}p(p+1)$ right-hand-sides, which are the columns of the reshaped matrices $\{A\}$. This requires of course changing the dimensions of the other variables so that $V, P \in \mathbb{R}^{K \times 1/2p(p+1)}$ and $W, Q \in \mathbb{R}^{1/2K(K-1) \times 1/2p(p+1)}$.

4.4.1 Generalization to specific pairwise penalties

For some problems it could be required penalize only specific classes, in order to avoid certain fusings. Consider for example K cancer classes for which there's available survival data. The samples can be grouped in T survival levels, and then be treated as KT cancer classes. Survival is, however, an

ordered variable which means that, for a given cancer class k, network links can be fused only for consecutive survival levels t and $t+1$ (or $t-1$ and t). The pairwise differences that we want to penalize here can be better visualized by the following graph (vertices represent cancer classes and links represent allowed penalties):

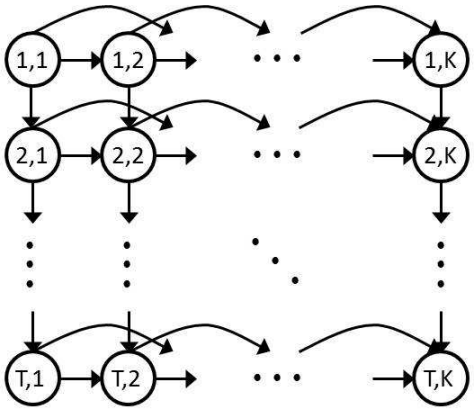


Figure 4.1: Allowed penalties for K cancer classes grouped in T survival levels. Penalties for all cancer types are allowed, while for survival levels only consecutive penalties are.

When using such penalties, a link in the estimated network is allowed, for a given cancer class, to be equal across all survival levels, and will actually be so for a sufficiently large value of the fuse penalty parameter. Alternatively it can be equal for a certain number of consecutive survival levels.

We can extend this approach to a more general situation in which we have K cancer classes and a graph structure which describes the allowed pairwise fusings. To solve this problem we apply the ADMM algorithm for class and pair-specific penalties taking care in specifying the allowed fusings in the matrix L in (4.6).

This concludes the presentation of our methods. In the next chapter we show, by means of simulations, how our adaptive penalty has better performance than the regular fused elastic net in the presence of modularity. We also present an analysis of tumor data, mRNA and CNA, for three types of cancer: breast cancer, glioblastoma and ovarian cancer.

Chapter 5

Results

5.1 Simulation Study

Construction of ROC curves

To compare the performance of our method with the regular fused elastic net, we consider a problem with three classes each with different sparsity and modularity settings.

We present the *receiver operating characteristic* (ROC) curves for the differential networks, that is, for the networks composed of links that are unique to a certain class. For the construction of the ROC curves we need to define the false positive rate (FPR) and the true positive rate (TPR). To do that we need to define, in turn, the number true positives (TP) and the number false positives (FP). Since we are interested in the discovery of differential networks, those will be the ones we will consider as positives, while fused links will be negatives.

To define the false positives and true positives we proceed as follows. Let $F_r \subseteq \{1, 2, \ldots, K\}$ for $r = 1, 2, \ldots, R$ where $1 \leq R \leq K$. For a given link (i, j) in the true networks, we can define $\mathfrak{F}_{\text{ij}} = \{F_1, F_2, \ldots, F_R\}$ as the set of (mutually disjoint) groups of classes for which the link is fused. That is, link (i, j) has exactly the same value for classes in set F_r and exactly the same values in set F_s but the value for the first group is different from the value for the second group. This implies that $F_r \bigcap F_s = \emptyset$ for all $r, s = 1, 2, \ldots, R$. The idea is to make pairwise comparisons between classes and label an estimated link as a true positive if its values are not fused for a pair of classes where the corresponding real link is not fused. Specifically, let θ_{ij}^k be the true value of link (i, j) in class k and $\tilde{\theta}_{ij}^k$ its estimate, the number of true positives and

True positives	False positives
1, 4	1, 2
1, 5	1, 3
1, 6	2, 3
2, 4	4, 5
2, 5	
2, 6	
3, 4	
3, 5	
3, 6	
4, 6	
5, 6	

Table 5.1: True positives and false positives for a link present in six classes and fusing structure {{1,2,3},{4,5},{6}}.

false negatives are defined as

$$TP = \sum_i \sum_j \sum_{k<k'} \mathrm{I}(\tilde{\theta}_{ij}^k \neq \tilde{\theta}_{ij}^{k'}, \theta_{ij}^k \neq \theta_{ij}^{k'})$$
$$FP = \sum_i \sum_j \sum_{k<k'} \mathrm{I}(\tilde{\theta}_{ij}^k \neq \tilde{\theta}_{ij}^{k'}, \theta_{ij}^k = \theta_{ij}^{k'})$$

respectively. Similarly, the number of true negatives and false negatives are defined as

$$TN = \sum_i \sum_j \sum_{k<k'} \mathrm{I}(\tilde{\theta}_{ij}^k = \tilde{\theta}_{ij}^{k'}, \theta_{ij}^k = \theta_{ij}^{k'})$$
$$FN = \sum_i \sum_j \sum_{k<k'} \mathrm{I}(\tilde{\theta}_{ij}^k = \tilde{\theta}_{ij}^{k'}, \theta_{ij}^k \neq \theta_{ij}^{k'}).$$

The FPR and TPR are defined as $TPR = TP/(TP + FN)$ and $FPR = FP/(FP + TN)$.

Consider the following example. We have 6 classes, and we know that, for a certain link (i, j), we have $\mathfrak{F}_{\mathrm{ij}} = \{\{1, 2, 3\}, \{4, 5\}, \{6\}\}$. That is, the link has the same value for classes 1, 2 and 3; the same value (but different to that in classes 1, 2 and 3) for classes 4 and 5 and a unique value (different than those for all the other classes) for class 6. The true positives and false positives are shown in Table 5.1. Specifically, each link that is not fused for any of the pairs of classes in the true positives column will be classified as a true positive. Similarly, each link that is fused for any of the pairs of classes in the true negative column will be classified as such.

Data preprocessing

We will investigate tumor data from glioblastoma, breast cancer and ovarian data sets in the Cancer Genome Atlas (TCGA). The number of genes and samples available from each cancer type are shown in Table 5.2.

	Glioblastoma	Breast cancer	Ovarian cancer
Genes	12042	20502	12042
Samples	529	779	568

Table 5.2: Variable and sample sizes for TCGA data.

The data sets contain 10321 genes common to all cancer types. For our simulation study, we randomly select 150 genes and apply our method to generate three precision matrices Θ^k, $k = 1, 2, 3$. We use different values for the tuning parameters to obtain different sparsity, fusing and modularity patterns. We then compute the covariance matrices $\Sigma^k = \left(\Theta^k\right)^{-1}$, $k = 1, 2, 3$, in order to simulate samples from $\mathrm{N}(0, \Sigma^k)$, $k = 1, 2, 3$.

Simulation results

Figure 5.1 shows the precision matrices for non-modular data. Blue dots represent the differential networks (they have different values across classes) and red dots represent the common (fused) network links. The number of non-zeros for the differential networks are 1548, 1732 and 1156 respectively. The common network comprises 1046 non-zeros.

In Figure 5.2 we show the ROC curves, averaged over 50 simulation runs, for the regular fused elastic net (no adaptivity) and Adaptivity I and II.

Figure 5.2(a) compares the regular fused elastic net to Adaptivity I for two different values of the adaptivity parameter: $\gamma = 1$ and $\gamma = 0.5$. Figure 5.2(b) compares the regular fused elastic net with Adaptivity I, for $\gamma = 1$, and Adaptivity II for $\gamma = 0.1$. We chose a smaller value for γ for Adaptivity II since, in the absence of modularity, this adaptivity scheme can otherwise be too aggressive and perform worse than the regular fused elastic net.

The ROC curves summarize the performance over a wide range of FPR whereas, in practice, we are more interested in the methods' performance for small values of FPR. We therefore also present results where we control FPR to approximately 0.1 and record the corresponding TPR for each method.

Figure 5.3 shows box plots of TPR, over 100 replications, when the FPR

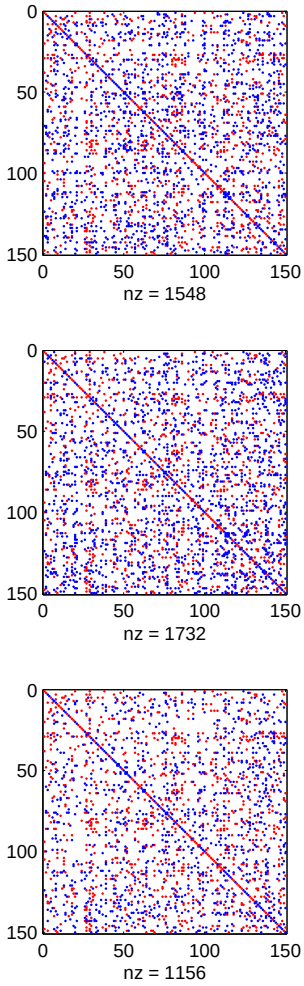
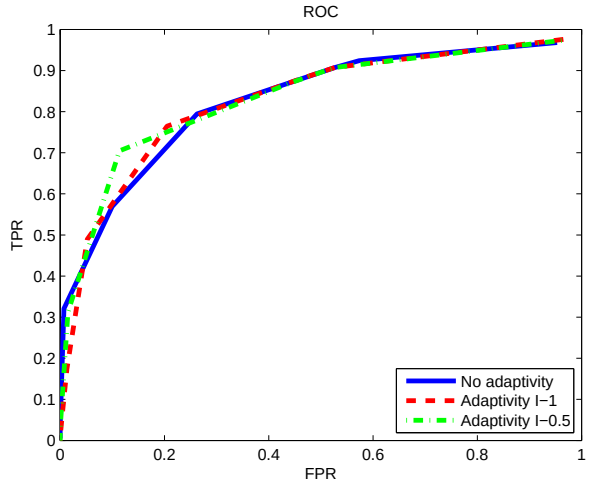


Figure 5.1: Precision matrices for a non-modular simulation. Differential networks in blue and common network in red.

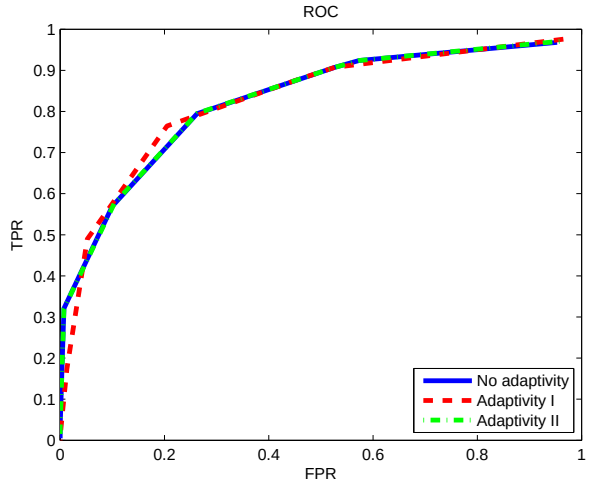
is fixed at approximately 0.1. In the absence of a modular structure, the performance of the three methods is very similar, with the two adaptivity schemes exhibiting a slight advantage over regular fused elastic net.

Figure 5.4 shows the precision matrices for networks with common and unique modules. The number of non-zeros for the differential networks are 1549, 1741 and 1111 respectively. The common network has 857 non-zeros.

The ROC curves are shown in Figure 5.5. Here, Adaptivity I performs better,



(a) Adaptivity I



(b) Adaptivity I and II

Figure 5.2: ROC curves for non-modular simulated data. (a) Comparison of fused elastic net and Adaptivity I for two different rates of adaptivity. (b) Comparison of regular fused elastic net and Adaptivity I and II. The performance of both Adaptivity I and II is as good as that of the regular fused elastic net.

even for a modest rate of adaptivity $\gamma = 0.5$, than the regular fused elastic net. This is what we expected for a modular network with common links across all classes. Adaptivity II performs similarly to the regular fused elastic net.

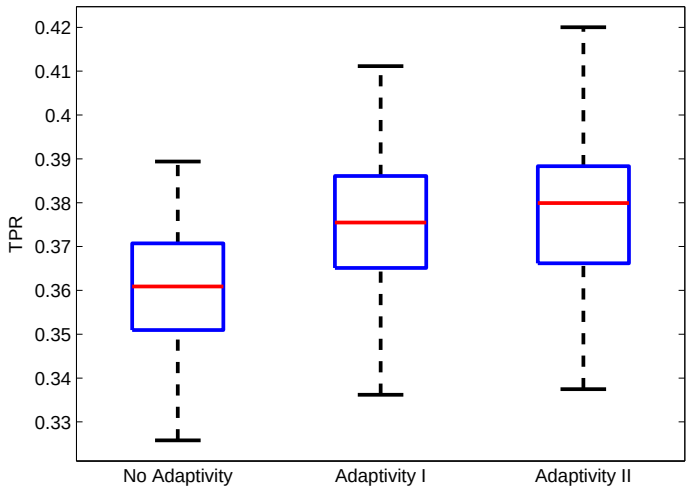


Figure 5.3: Box plots for TPR with FPR≈0.1. All methods perform similarly.

The box plots, Figure 5.6, show results consistent with the ROC curves, clearly indicating the advantage of Adaptivity I compared with regular fused elastic net, and no loss of performance using Adaptivity II.

The last simulation is presented in Figure 5.7. Here the modularity structure is more complicated, in the sense that there are modules that are common across all three classes, or common for only two classes. Blue represents unique modules, while the other colors represent common modules for the corresponding classes. The differential networks have 1395, 687 and 574 non-zeros respectively. The red common network has 744 non-zeros, the green one 241, and the yellow common network has 828 non-zeros.

In this case, Adaptivity I performs as well as the regular fused elastic net, for $\gamma = 0.5$. Adaptivity II, which is designed to handle this kind of data structure, performs better than regular fused elastic net (see Figure 5.8).

The results are even more apparent in Figure 5.9. The TPR box plots, with FPR constraint at 0.1, clearly show the advantage of Adaptivity II in the presence of a complex fuse pattern in data sets.

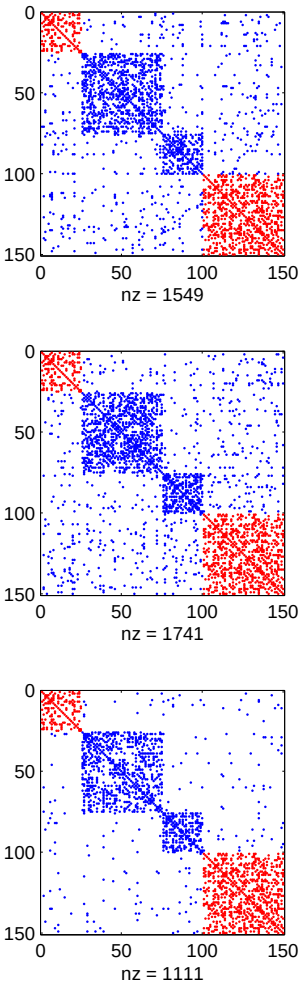
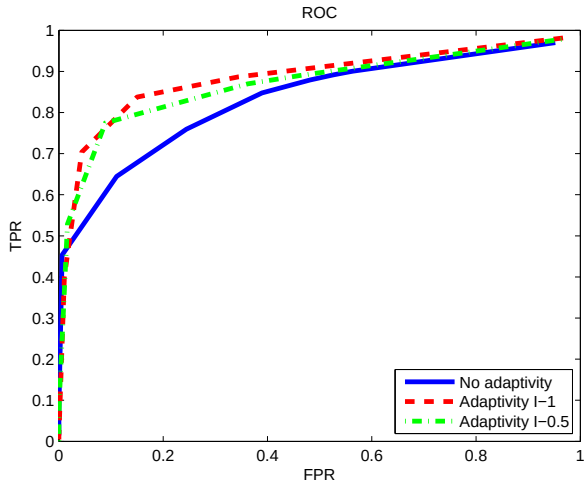
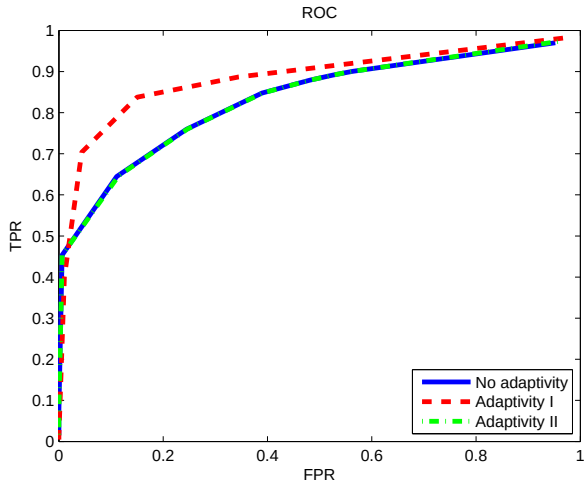


Figure 5.4: Precision matrices for a modular simulation. Differential networks in blue and common network in red.



(a) Adaptivity I



(b) Adaptivity I and II

Figure 5.5: ROC curves for modular simulated data. (a) Comparison of fused elastic net and Adaptivity I for two different rates of adaptivity. (b) Comparison of regular fused elastic net and Adaptivity I and II. The performance of Adaptivity I is better than that of the regular fused elastic net.

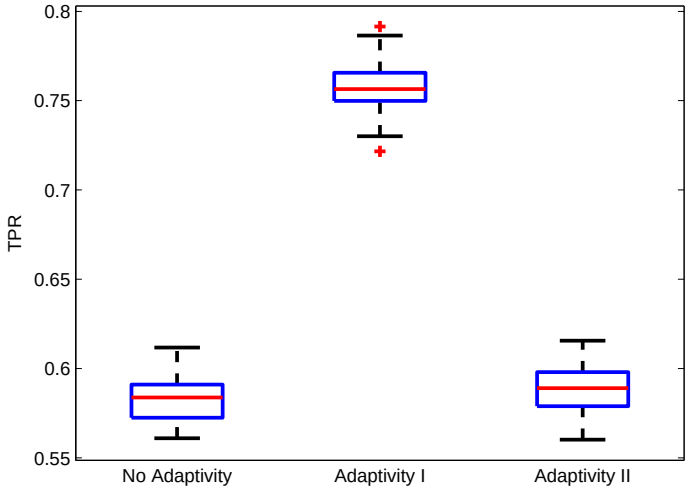


Figure 5.6: Box plots for TPR with FPR≈0.1. Performance of Adaptivity I is superior.

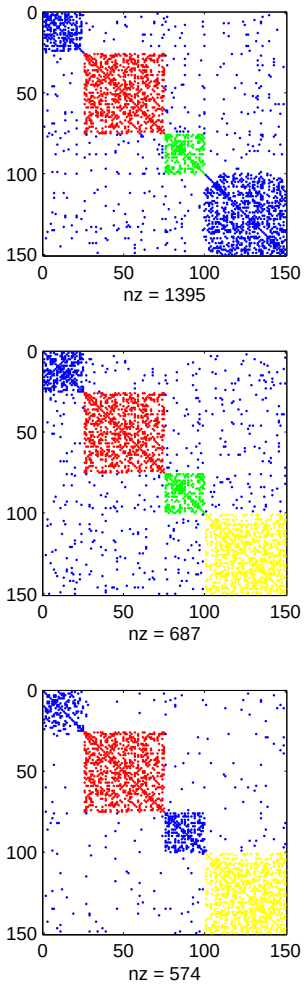
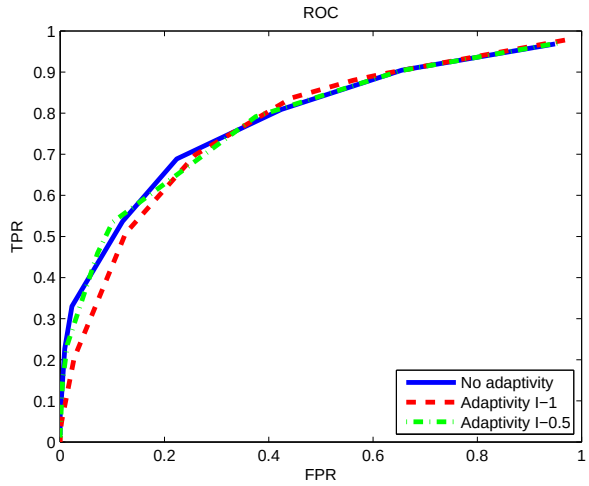
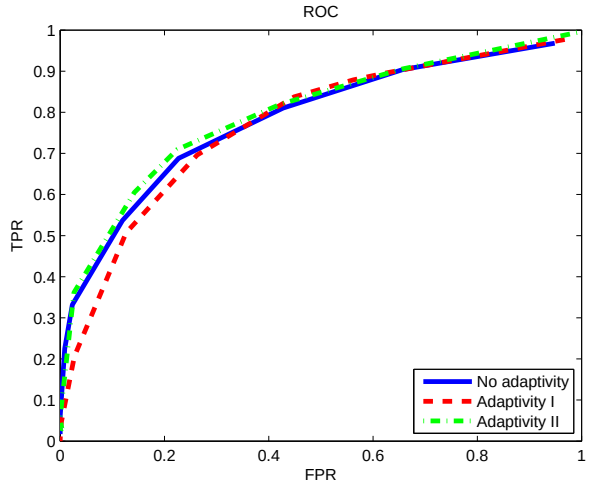


Figure 5.7: Precision matrices for a modular simulation. Differential networks in blue and common network in other colors.



(a) Adaptivity I



(b) Adaptivity I and II

Figure 5.8: ROC curves for non-modular simulated data. (a) Comparison of fused elastic net and Adaptivity I for two different rates of adaptivity. (b) Comparison of regular fused elastic net and Adaptivity I and II. The performance of Adaptivity II is better than that of the regular fused elastic net.

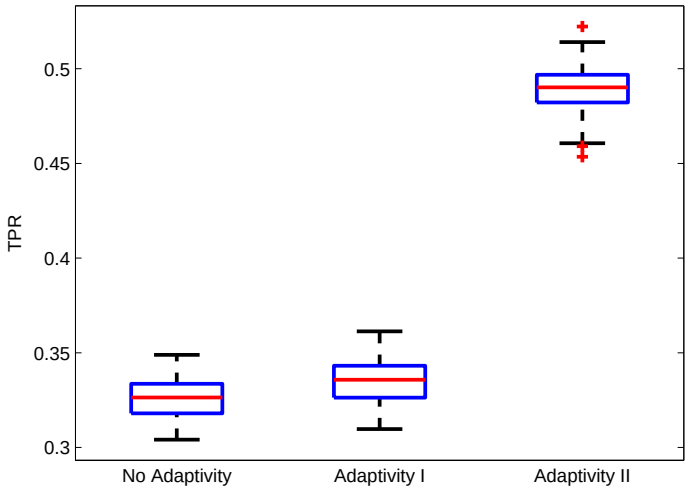


Figure 5.9: Box plots for TPR with FPR≈0.1. Performance of Adaptivity II is superior.

5.2 Real data analysis

5.2.1 Data preparation

We downloaded data sets for breast cancer, glioblastoma (gbm) and ovarian cancer from the TCGA (`http://cancergenome.nih.gov`) database. The TCGA data are organized into technical platforms, and we chose the platform for each data type and cancer to maximize the number of patients in that data set. The chosen platforms for each cancer are shown in Table 5.3 and the number of patients available for each combination of data types is shown in Table 5.4. All measurements were downloaded as TCGA level 3 data and were post-processed as described below. The processed data was assembled in a mySQL database to enable fast creation of data matrices during simulations.

mRNA. The level 3 mRNA data provided by TCGA is a list, for each patient, of known protein-coding genes with their corresponding estimated mRNA expression values. All values where logged for all Illumina RNA Sequencing platforms. Additionally, all data was quantile normalized within each cancer and platform.

CNA. The level 3 CNA (genetic copy number aberration) information provided in TCGA is, for each patient, the amplitude and genetic positions of the beginning and end of DNA segments that have gained or lost copies. Each gene available in NCBI human Build 36.1 was mapped to the segments and assigned the amplitude of the corresponding segment. Where multiple segments cover the gene, the average amplitude was used, weighted proportionally to the length of the parts of the segments. Genes with a CNA value but lacking a mRNA measurement were discarded from the analysis.

Data type	**Center**	**Platform**
mRNA	Broad Institute	Affymetrix HT Human Genome U133 Array Plate Set
CNA	Broad Institute	Affymetrix Genome-Wide Human SNP Array 6.0

Table 5.3: TCGA platforms chosen for each data type.

	Breast	Ggb	Ovarian
Sample size	766	509	560

Table 5.4: Number of patients for each cancer type.

5.2.2 Estimation

To reduce the execution time in the estimation of our networks, we adopted the method of Danaher et al. (2011). They proceed by dividing the matrix S into separable subproblems. The criteria for separability for the different described models follows below:

Fused Elastic Net for 2 classes. Let C_1 and C_2 be two disjoint sets of the variables such that $C_1 \bigcup C_2 = \{1, 2, \ldots, p\}$. For the set of variables in C_1 to be fully disconnected from the variables in C_2, the following conditions must be fulfilled:

$|n_2 S_{ij}^{(1)}| \leq \lambda_1 \alpha + \lambda_2,$
$|n_1 S_{ij}^{(2)}| \leq \lambda_1 \alpha + \lambda_2,$
$|n_1 S_{ij}^{(1)} + n_2 S_{ij}^{(2)}| \leq 2\lambda_1 \alpha,$
for all $i \in C_1$ and $j \in C_2$.

Fused Elastic Net for $k > 2$ classes. For the set of variables in C_1 to be fully disconnected from the variables in C_2, the following conditions must be fulfilled, assuming that the elements $n_k S_{ij}^k$ are sorted so that $n_1 S_{ij}^1 \leq n_2 S_{ij}^2 \leq ... \leq n_K S_{ij}^K$:

$|\sum_{k=1}^{t} n_k S_{ij}^{(k)}| \leq t\lambda_1 \alpha + \lambda_2, 1 \leq t \leq K-1,$
$|\sum_{k=t_1}^{t_2} n_k S_{ij}^{(k)}| \leq (t_2 - t_1 + 1)\lambda_1 \alpha + 2\lambda_2, 1 \leq t_1 \leq t_2 \leq K-1,$
$|\sum_{k=t}^{K} n_k S_{ij}^{(k)}| \leq (K - t + 1)\lambda_1 \alpha + \lambda_2, 2 \leq t \leq K,$
$|\sum_{k=1}^{K} n_k S_{ij}^{(k)}| \leq K\lambda_1 \alpha.$ for all $i \in C_1$ and $j \in C_2$.

Selection of penalty parameters. The selection and validation of penalty parameters λ_1 and λ_2 is a difficult problem. In our experience (Jornsten et al., 2011; Abenius et al., 2012) BIC tends to underfit, by selecting very sparse networks, while cross validation overfits, by selecting very dense networks. We believe that the selection should take into account how stable the estimates are. As a function of the penalty parameters, both the likelihood function and the number of unique estimated parameters (sparsity level) tend to stabilize for moderate values of λ_1 and small values of λ_2. An interval where the likelihood and sparsity levels are stable corresponds to values of the penalty parameters such that random links, or false positives, have been eliminated whereas real links, or true positives, have not.

In the present study we select $\lambda_1 = 0.5$ and $\lambda_2 = 0.01$, which belong to the stable region mentioned above. This way we obtain networks where enough connections exist that can be cross checked with known pathways. A complete analysis would require us to estimate and validate, experimentally

or via large-scale simulations, the networks for different combinations the penalty parameters. We leave this analysis for future work.

Prior distribution. For the integration of mRNA and CNA we require a prior distribution on the allowed connections between nodes. Consider the inverse covariance matrix for a specific cancer type. We can divide it in four blocks, or just three if we take into account its symmetry, corresponding to the mRNA-mRNA, the mRNA-CNA and the CNA-CNA connections. All mRNA-mRNA connections are potentially possible, so there's no need of specific penalization for that block other than λ_1. CNA should be allowed links only with its corresponding mRNA, so we enforce a diagonal structure for the mRNA-CNA. This is achieved by increasing the sparsity penalty λ_1 for the off-diagonal elements of that block. Finally, connections between CNAs far away from each other (in chromosomal location) shouldn't exist. As the genes are arranged by chromosomal location, we thus define a tridiagonal prior for the CNA-CNA block.

Robust network estimation via bootstrap. We generate 200 bootstrap samples choosing randomly 90% of the patients in each cancer from the original sample. For each bootstrap sample we estimate the corresponding network with the selected penalty parameters. Some links appear frequently (true positives or real links), and some show up seldom (false positives). This behaviour motivates us to use frequency statistics for final network construction (e.g. (de Matos Simoes and Emmert-Streib, 2012)). We define two thresholds T_1 and T_2 on the frequency statistics to control the sparsity and the fusing of the links in the final estimate.

Consider first the sparsity problem. The idea is that, for a given threshold T_1, a link will be present in the final estimate if it is present in $100T_1\%$ of the bootstrap estimates. Specifically, for a cancer class $k = 1, 2, 3$ let

$$n_{ij}^k = \frac{\sum_{b=1}^{200} \mathrm{I}(\theta_{ij,b}^k \neq 0)}{D},$$

where $\theta_{ij,b}^k$ is the b-th bootstrap estimate for link (i, j) in class k. This n_{ij}^k is an estimate of the probability of presence of link (i, j) in cancer class k. This link will be present in class k if and only if $n_{ij}^k \geq T_1$.

Let us now consider the fusing of edges, where further complexity arises. We proceed by estimating the edge difference probability for all cancer pairs. We distinguish 4 different cases.

Case 1. For the set of links not present in any cancer class, that is, those for which $\sum_k \mathrm{I}(n_{ij}^k \geq T_1) = 0$ no further work is needed.

Case 2. Consider the set of links that are present only in one class, that is,

those for which $\sum_k \mathrm{I}(n_{ij}^k \geq T_1) = 1$. In this case no fusing is required and the links stay present with the already estimated values.

Case 3. Here we consider the set of links for which $\sum_k \mathrm{I}(n_{ij}^k \geq T_1) = 2$, that is, those that are present in two classes, say k and k', and absent in the third one. We need now to decide whether the link should stay fused in classes k and k'(null hypothesis) or whether they should be differential (alternative hypothesis). Let θ_{ijb}^k to be the b-th bootstrap estimate for link (i, j) in class k. We compute

$$n_{ij}^{kk'} = \frac{\sum_{b=1}^{200} \mathrm{I}(\theta_{ij,b}^k \neq \theta_{ij,b}^{k'}, \theta_{ij,b}^k \neq 0, \theta_{ij,b}^{k'} \neq 0)}{\sum_{b=1}^{200} \mathrm{I}(\theta_{ij,b}^k \neq 0, \theta_{ij,b}^{k'} \neq 0)}.$$

This is an estimate of the probability that link (i, j) is differential in classes k and k' given it is present in both classes. If $n_{ij}^{kk'} \geq T_2$, then link (i, j) is differential in classes k and k' in the final estimates, otherwise it is fused.

Case 4. For the set of links where $\sum_k \mathrm{I}(n_{ij}^k \geq T_1) = 2$, that is, the set of links present across all three classes, we compute

$$n_{ij}^{kk'k''} = \frac{\sum_{b=1}^{200} \mathrm{I}(\theta_{ij,b}^k \neq \theta_{ij,b}^{k'} = \theta_{ij,b}^{k''}, \theta_{ij,b}^k \neq 0, \theta_{ij,b}^{k'} \neq 0, \theta_{ij,b}^{k''} \neq 0)}{\sum_{b=1}^{200} \mathrm{I}(\theta_{ij,b}^k \neq 0, \theta_{ij,b}^{k'} \neq 0, \theta_{ij,b}^{k''} \neq 0)},$$

where, for instance, $k = 1$, $k' = 2$ and $k'' = 3$. This makes $n_i j^{kk'k''}$ an estimate of the probability that link (i, j) is fused in classes 1 and 2 (breast and gbm respectively) and not fused, or differential, in class 3 (ovarian cancer). We also need an estimate of the probability that the link is differential for all classes, given by

$$n_{ij}^{kk'k''} = \frac{\sum_{b=1}^{200} \mathrm{I}(\theta_{ij,b}^k \neq \theta_{ij,b}^{k'} \neq \theta_{ij,b}^{k''}, \theta_{ij,b}^k \neq 0, \theta_{ij,b}^{k'} \neq 0, \theta_{ij,b}^{k''} \neq 0)}{\sum_{b=1}^{200} \mathrm{I}(\theta_{ij,b}^k \neq 0, \theta_{ij,b}^{k'} \neq 0, \theta_{ij,b}^{k''} \neq 0)}.$$

The final decision on the fusing structure is given by the maximum of $n_{ij}^{kk'k''} \geq T_2$ for all triads k, k', k''. If, conversely, $n_i j^{kk'k''} < T_2$ for all triads, we conclude that the link is fused for all cancer classes.

Selection of thresholds T_1 and T_2. Rigorous selection of T_1 and T_2 is another validation problem, but it is simpler compared to that of validation for the penalty parameters λ_1 and λ_2.

Figure 5.10 shows the histogram for frequency of link presence for all cancer types. The estimated probability of presence is 0 or 1 for many links, we omit them to clearly see the shape of the histogram.

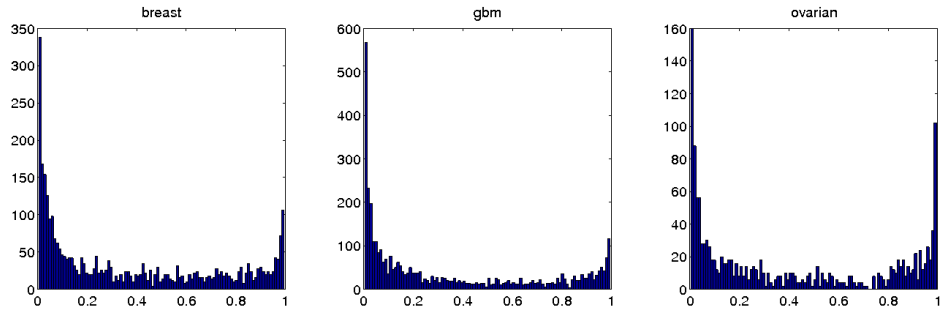


Figure 5.10: Frequency of link presence across classes. The U shape suggests links below 0.2 to be false positives and not reproducible while links above 0.8 are true positives.

All three histograms show a U shape with a decrease from probability 0 to 0.2. This interval comprises links that appear across some bootstrap estimates by chance and therefore they are not reproducible. Links above 0.8 appear consistently in our bootstrap estimates, thus these belong to the set of true positives. We conclude then that reasonable values for T_1 are above at least 0.2.

Figure 5.11 shows the histograms for frequency of pairwise link fusing. To see the shape of the histogram clearly we omit again links with estimated probability of fusing equal to 0 or 1.

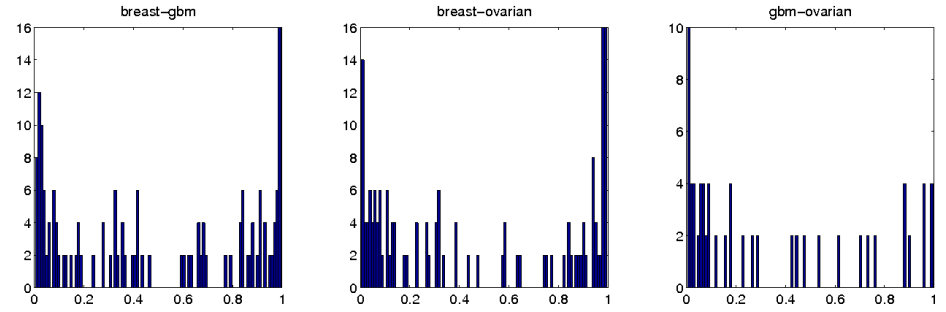


Figure 5.11: Frequency of pairwise fusing across classes. The U shape suggests links below 0.2 to be false positives and not reproducible while links above 0.8 are true positives.

Similarly as for link presence, the U shape of these histograms suggests that differential edge values that appear in less that 20% of our bootstrap estimates are random (false positives). Differential edge values above the 0.8 threshold appear consistently across bootstrap estimates, thus comprising true positives. Based on these results, we decide to take $T_2 \geq 0.2$.

5.2.3 Analysis

In collaboration with the Nelander laboratory, IGP and SciLife, Uppsala University, we have developed an on-line visualization tool (administrator Patrik Johansson) called Cancer Landscapes (`http://cancerlandscapes.org/demo/`). A screenshot of the tool is shown in Figure 5.12.

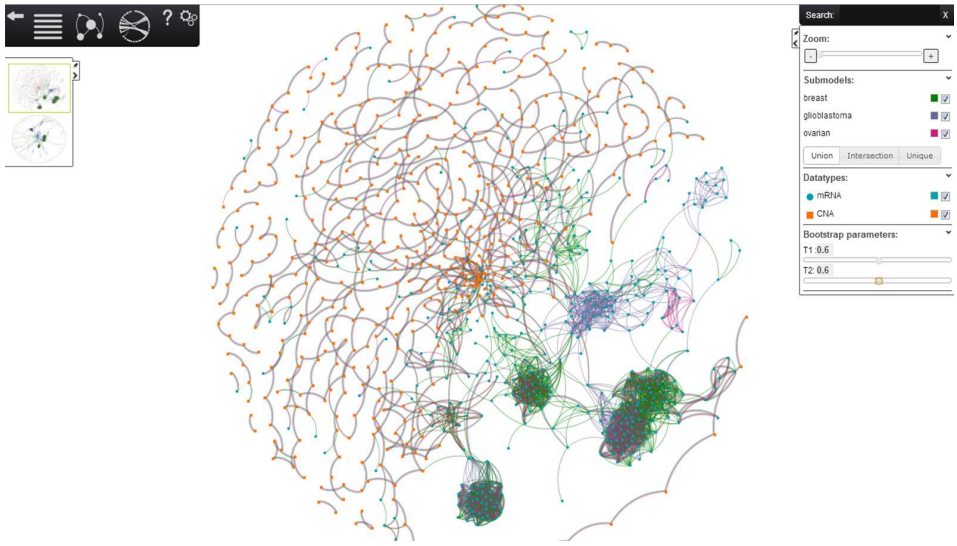


Figure 5.12: Cancer Landscapes. Initial configuration of web-tool for model Breast-Ggb-Ovarian 500 genes.

Cancer Landscapes has a variety of features. On the upper right corner we can select which cancer classes to display and data types to display. We can also choose among the union, intersection or unique networks. At the bottom there is a slide bar to select the frequency statistic threshold parameters T_1 and T_2. As described above, T_1 controls the sparsity level (the larger it is, the sparser the network), and T_2 controls the fusing level (the larger it is, the more the present links will be fused across classes).

It is also possible to choose among different topologies for the networks. Since the genes we are working with have been ordered by chromosomal location and we have a banded prior for the CNA network, it is possible to organize the CNA nodes in a ring surrounding the mRNA nodes. Figure 5.13 shows the final estimated network in this fashion.

Whenever two genes are connected in a cancer class, a link of the corresponding color is present. This way, connected genes can have 1, 2 or 3 links. Using display option "fused edges", fused links are plotted next to

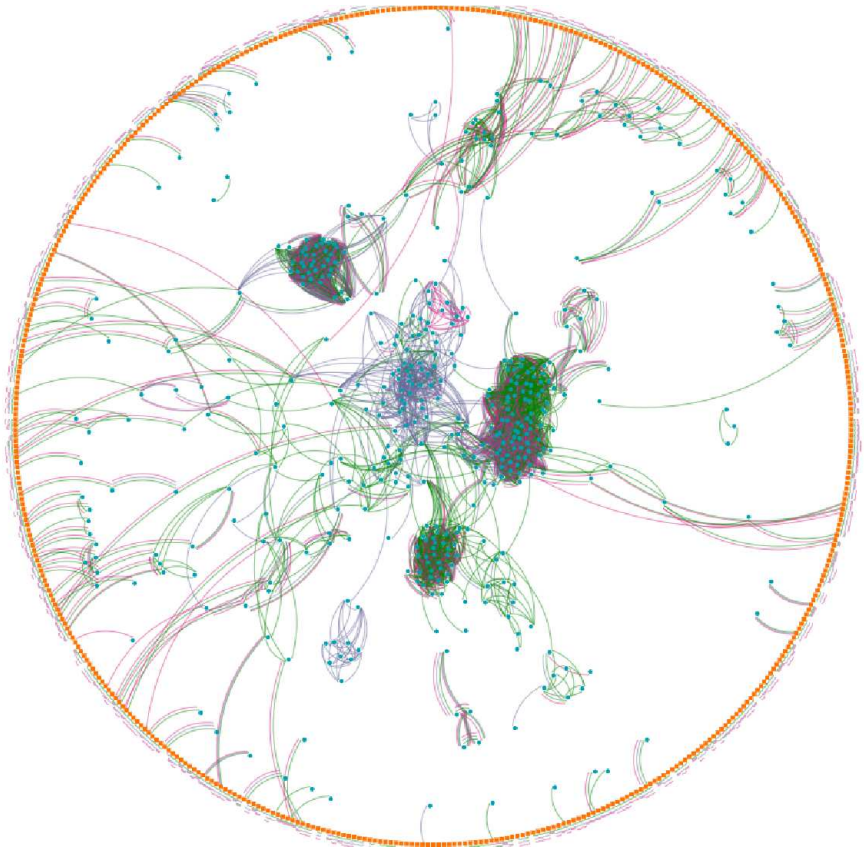


Figure 5.13: mRNA and CNA network for gbm, breast and ovarian cancer. mRNA genes in blue, CNA genes in orange. Breast cancer links in green, glioblastoma links in lilac and ovarian cancer genes in magenta.

each other, whereas differential edges are depicted with a space between them.

A complete analysis will require careful exploration of the network for all values of the threshold parameters T_1 and T_2. We are currently preparing a manuscript on a more complete analysis of cancer networks (Kling et al., 2013). In this thesis, we focus on the network for $T_1 = 0.6$ and $T_2 = 0.6$ and make a short analysis to demonstrate the kind of findings that are possible with Cancer Landscapes.

We begin by showing the fusing structure in Table 5.5.

The first row, *Total*, is the total number of links present in the corresponding cancer class. The *Unique* row lists the number of all links that are differential in each cancer. They comprise various subtypes of links:

	Breast	Gbm	Ovarian
Total	4828	3674	2896
Unique	3308	1804	1202
Fused	1166	1166	1166
Present in two classes			
Breast	-		
Gbm	976	-	
Ovarian	1122	1098	-
Present in two classes, not present in third class			
Breast	-		
Gbm	326	-	
Ovarian	532	102	-
Present in 3 classes			
Breast	-		
Gbm	650	-	
Ovarian	590	996	-
Fused in 2 classes			
Breast	-		
Gbm	220	-	
Ovarian	134	484	-
Fused in two classes, not present in third class			
Breast	-		
Gbm	144	-	
Ovarian	118	62	-
Fused in two classes, present in third class			
Breast	-		
Gbm	76	-	
Ovarian	16	422	-

Table 5.5: Fusing patterns for $T_1 = 0.6$ and $T_2 = 0.6$.

- Links that are present in one cancer class only.
- Links that are present in two cancer classes and are not fused.
- Links that are present in the all cancer classes and have different values for the three of them.

The second part of the table, *Present in two classes* shows the number of links that are present in each pair of cancers, independently of them being fused or differential and independently of a link being present or not in the third class. The situations in which the third class is present or not are shown below.

Section *Fused in two classes*, presents the number of links that are fused for the corresponding pair of cancers, independently of the link being present or not in the third class. These two subcases are listed below. For example,

breast and ovarian cancer have a total of 134 fused links. The majority of them, 118, correspond to glioblastoma links being also present, while only 16 of them are present when there are no glioblastoma links.

Differential links can be deduced from this table. For instance, there's a total of 976 links present in both breast and glioblastoma. Of those, 220 are fused, therefore 756 of them are differential, independently of the presence or absence of a link in ovarian cancer.

We proceed now with a descriptive analysis of our network using some of the capabilities of Cancer Landscapes.

Genes belonging to important know pathways can be highlighted. In Figure 5.14 we show the genes present in our network that are part of the Cell Cycle.

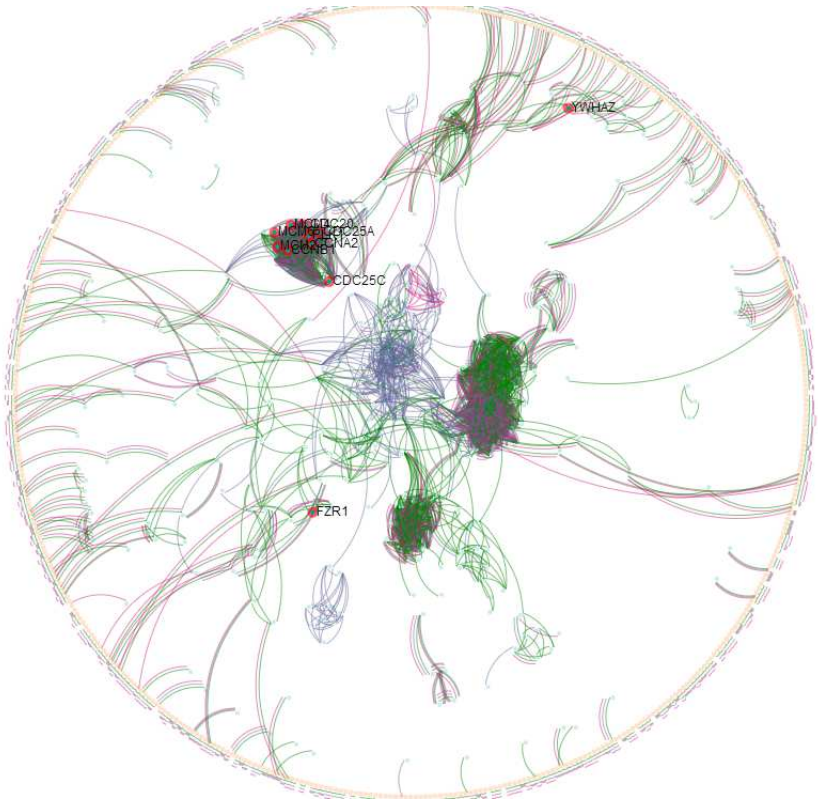


Figure 5.14: Genes belonging to Cell Cycle pathway concentrated in a module.

Most of the genes are concentrated in a module. Figure 5.15 shows a zoom to the module.

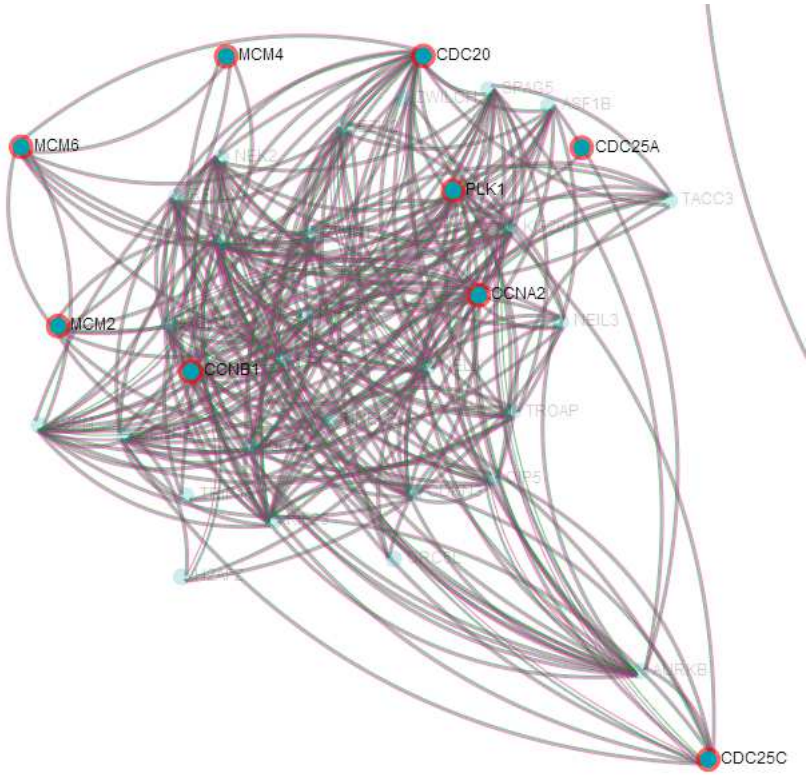


Figure 5.15: A module of fused links across all cancer classes. Genes belonging to the Cell Cycle pathway are highlighted.

Almost all of the links in the figure are fused across all classes. This is an expected result since the cell cycle should be a common process to all cancers. Results of this kind also validate our selection of penalty parameters.

Another interesting pathway, the ECM-receptor interaction, is also well located in a module of our network (see Figure 5.16). This pathway is related to cellular activities such as proliferation and apoptosis. As seen in the figure, this module is predominantly associated with breast cancer and ovarian cancer (fused).

So far we have used the term "module" to refer to sets of genes tightly connected. However, it is also interesting to look a clusters of genes, as defined in the traditional way. With Cancer Landscapes it is possible to cluster genes, using the network structure and the jaccard index as a distance metric. We cluster one cancer type at the time, and identify pathways located in those modules. Figure 5.17 shows a module (green background shade) of breast cancer that comprises genes known to be part of pathways of colorectal cancer, basal cell carcinoma and endometrial cancer.

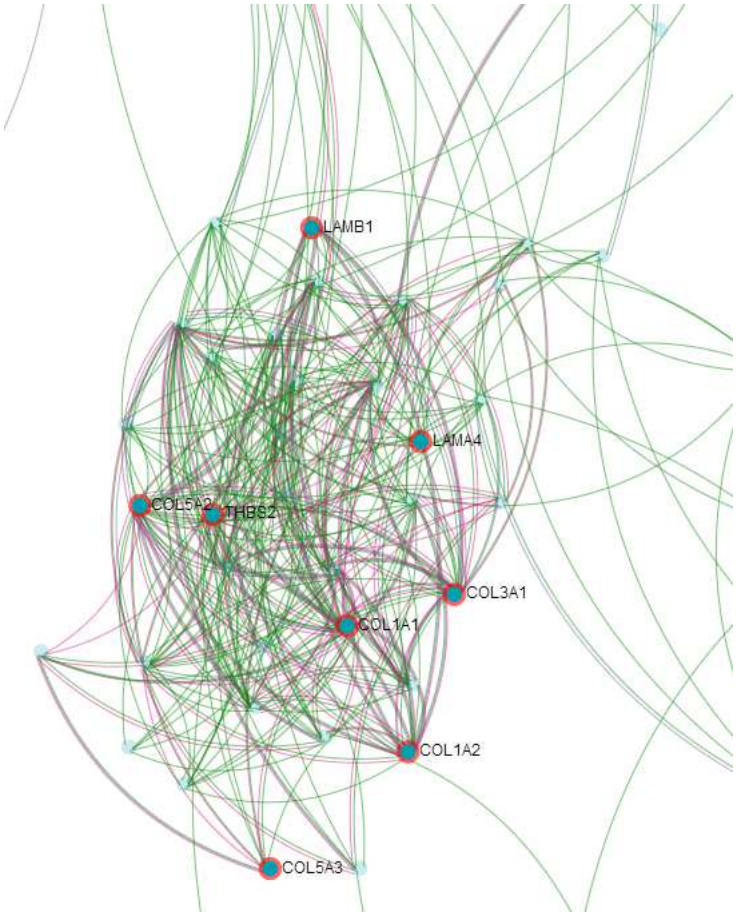


Figure 5.16: A module with some fused and some differential links. Genes belonging to the ECM-receptor interaction pathway are highlighted.

As a summary of the structural differences in the clusters, we can request the alluvial diagram (Rosvall and Bergstrom, 2010). Figure 5.18 shows the alluvial diagram for all clusters found in the three cancer networks. We have highlighted a particular cluster that is common to all three cancers. Cancer Landscapes computes the overlap of these clusters with known pathways and gene functional categories.

Next to the alluvial diagram we show the pathways identified for a particular cluster that appears in all cancers. The genes in this cluster are associated with e.g. the FOXM1 transcription factor network, a known human proto-oncogene.

Clusters reveal sets of genes well connected to each other. They can't, however, detect directly genes that have a large number of connections. These "hub" genes are important, since they are known to be possible disease

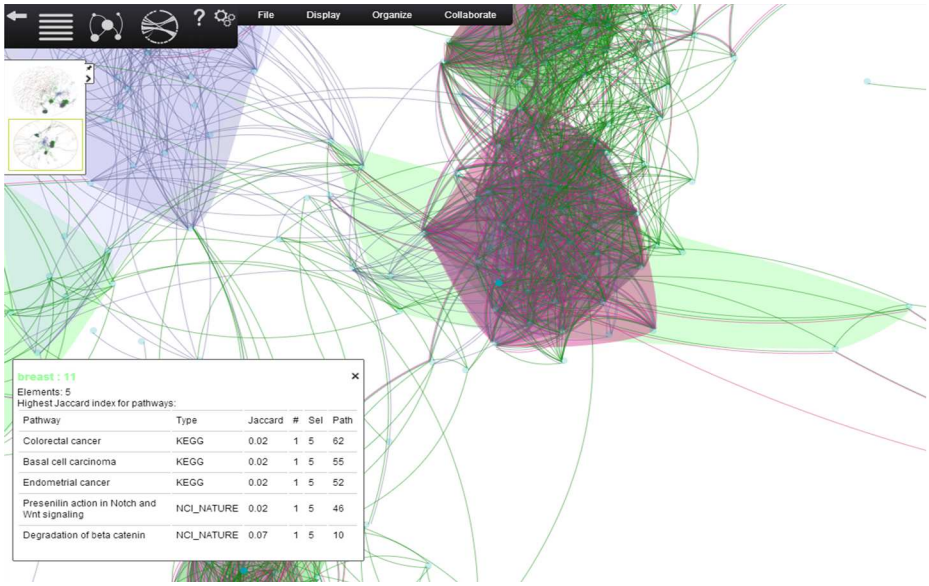


Figure 5.17: A cluster of breast cancer genes (green background shade) that contains cancer specific genes. The table shows the pathways associated to those genes.

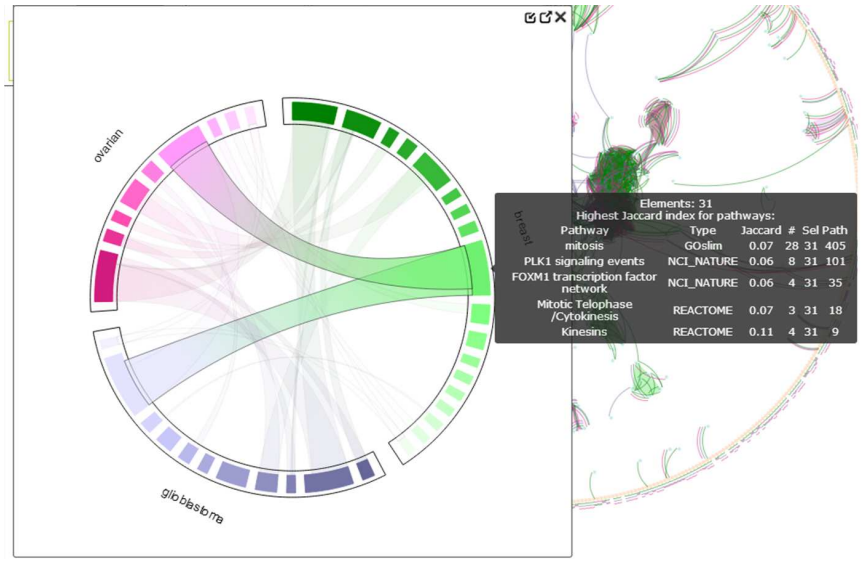


Figure 5.18: Alluvial diagram for all clusters found in the three cancer networks (a subset of which is shown in Figure 5.17). The table shows the pathways associated to the genes in the highlighted cluster, which in this case is common to all three cancers.

drivers. This motivates using node degree and some other network-theoretic

measures of node centrality in our networks.

Cancer Landscapes has three network-theoretic measures implemented, namely, node degree, PageRank and Centrality-Betweenness. We show first the node degree for our network. For a given node, this measure is defined as the number of nodes directly connected with it. Figure 5.19 shows our network next to a table with the top 10 genes with highest degree and highlights the top ranking node, CD53. This gene is associated with cell development, activation, growth and motility.

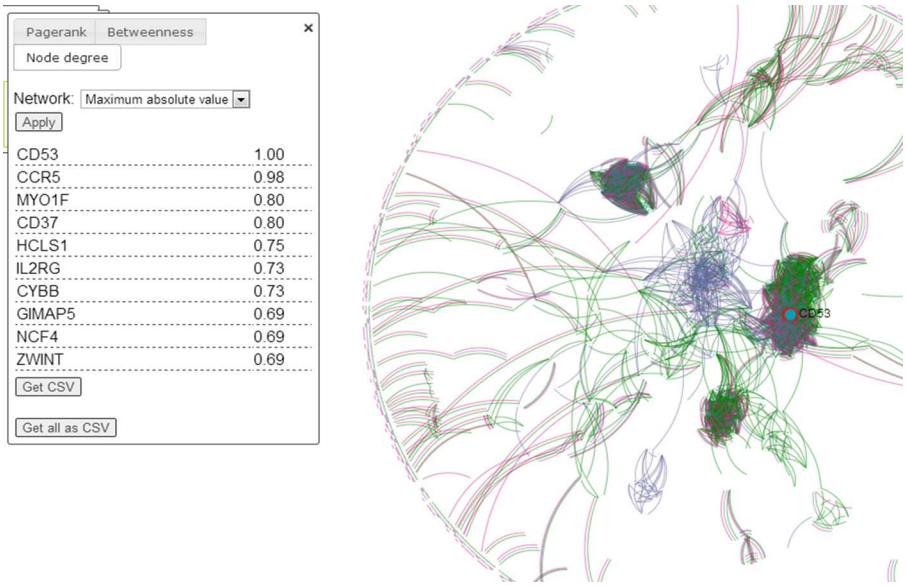


Figure 5.19: Top 10 genes with highest degree. The top ranking node, CD53, is highlighted in the network.

PageRank is another way of measuring the relative importance of a network node and it can be considered as a generalization of node degree. Instead of counting only the number of nodes directly connected (neighbors of degree 1) to the node in question, we also count the number of nodes connected to those (higher order neighbors). This way, a node with neighbors that have a high PageRank will also have high PageRank. Figure 5.20 shows the network and a table with the top 10 genes with highest PageRank. The top ranking node, CDC20 (highlighted in the network), is an important regulator in cell division.

The last measure is Centrality-Betweenness. For a given node, this property is computed by counting the number of times a node is part of the shortest path connecting any other two nodes. The more shortest paths that travel through a node, the higher its betweenness. Figure 5.21 shows our network

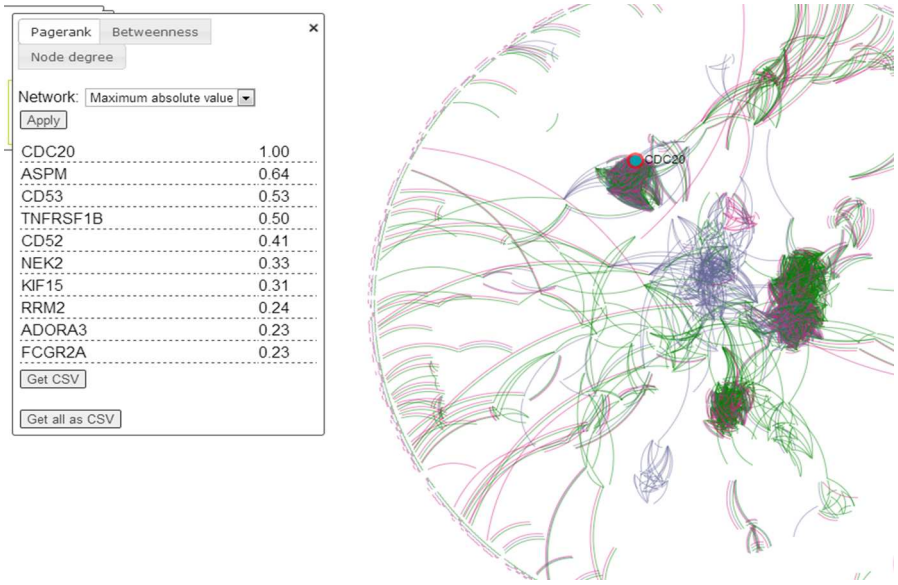


Figure 5.20: Top 10 genes with highest PageRank. The top ranking node, CDC20, is highlited in the network.

and a table with the top 10 genes with highest betweenness. It is expected that such genes will be part of the path between two modules, as highlighted in the figure. The Nelander lab has experimentally validated a subset of genes with high PageRank and betweenness. Using siRNA (silencing RNA), such genes were knocked down in glioma cell culture studies. It was found (preliminary data not shown), that genes with high betweenness were associated with a substantially reduced viability of the cultures. This suggests that network modeling, and network summary features like node centrality, can identify important transcripts that regulate key biological processes in diseases such as cancer.

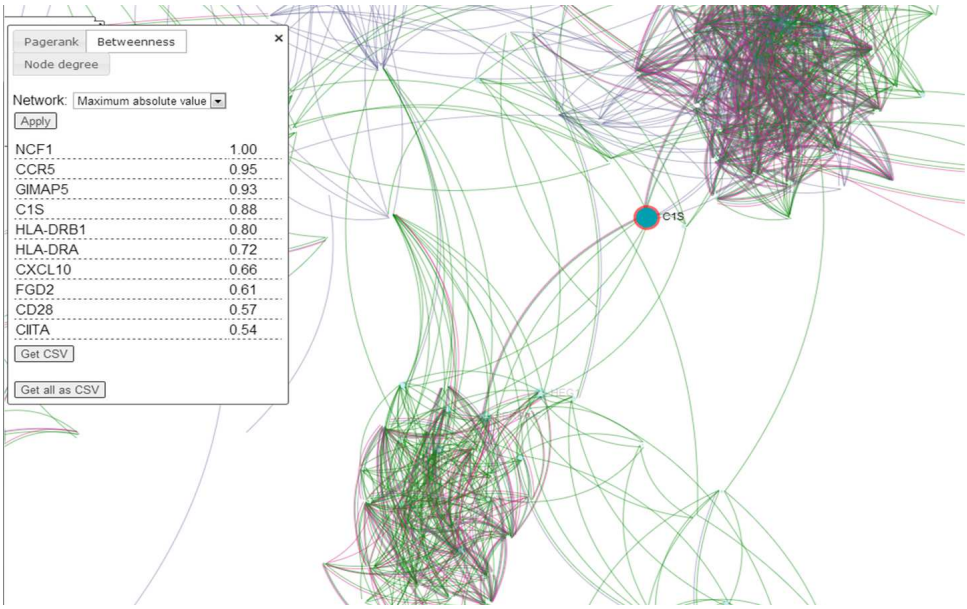


Figure 5.21: Top 10 genes with highest betweenness. A high-ranking node, CS1, is part of the path between two modules.

Chapter 6

Conclusions and future work

A network model of a biological system aims to capture functional links between some of its components. In the particular case of a cancer tumor, it is for instance important to understand how the mutations in the genome (like Copy number aberrations) are linked to downstream functional events in cells (e.g. regulation of tumor invasion). Several ideas have been put forward for network construction, such as correlations, sparse regression, and sparse graphical models (glasso). In this thesis, we have focused on extensions of sparse graphical models (sparse inverse correlation).

Recent work has extended glasso to comparative analysis (multiple cancers) using the so-called fused penalty function (Danaher et al., 2011). We have extended this framework to better account for the modularity frequently seen in biological networks. That is, the network matrix is "blocky", consisting of tightly connected groups of genes. These groups may then be connected to other groups, but usually in a much more sparse fashion (few links). We work under the hypothesis that genes within such gene modules should exhibit similar differential patterns across cancers. We have developed two adaptive penalty criteria that encourage gene modules to be (a) common across all cancers, (b) differential across a subset of cancers, or (c) unique in all cancers. The adaptive penalty criteria build on principles from adaptive lasso (Zou, 2006). We have been able to show through extensive simulation studies that our two adaptive penalty methods identify true differential connectivity between data sets (cancer classes) compared with regular fused lasso methods.

The problem of relative sample size of cancer classes has hitherto largely been ignored in the comparative modeling literature. When sample sizes differ, the effective penalty parameters will also differ (scales as λ_1/n_k), leading to overly sparse networks for cancer classes with small sample sizes. A common

way to fix this issue is to assume equal sample size for all cancer classes (Danaher et al., 2011). We have found that this method overcorrects, leading to networks that are non-sparse for cancer classes with small sample sizes. The sensitivity to sample size is a concern as it limits the interpretability of estimated models. We have examined sample size correction schemes using effective sample sizes $n_k^e = \bar{n}^\delta n_k^{(1-\delta)}$, where $\bar{n} = \frac{1}{K}\sum_{k=1}^{k} n_k$ and $0 \leq \delta \leq 1$ controls the amount of sample size correction. Our studies indicate that small values of δ close to zero produce much more balanced networks across cancers compared to the uncorrected or naive correction schemes.

Network estimation is quite unstable. To present a robust network estimate we have used bootstrap to determine (i) the presence of a link in a cancer class and (ii) the differential connectivity between cancers for a particular link. Bootstrap has been used to determine link presence in several previous publications (e.g. de Matos Simoes and Emmert-Streib (2012)). Here, we have extended this paradigm to determine the fused/differential connectivity across cancer classes.

We have visualized the final network using the web tool Cancer Landscapes. The development of this tool is ongoing and we are working closely with the Nelander lab (SciLife, IGP, Uppsala University), to improve the statistical analysis methods, features and accessibility of the tool.

Comparative network modeling is a relatively new area of research and many important problems remain to be solved. We are currently extending the framework presented here to a large-scale comparative analysis of 8 different cancer classes and 6 different data types: mRNA, CNA, microRNA, methylation, mutation and survival (Kling et al., 2013). This extension requires careful consideration of appropriate prior distribution assumptions as well as robust estimation of within and between data type correlations.

Our future methodology work will center on the extension of our methods to graph penalty models where we compare networks across both cancer classes and survival levels (Section 4.4.1). We are also planning to investigate alternative forms of adaptive penalties and sample size corrections schemes. All methods require careful validation and selection of penalty parameters. This is a very difficult problem where commonly used methods, like cross-validation and BIC model selection, perform poorly. As part of the 8-cancer network modeling project we will explore several different validation criteria, e.g. network overlap with known pathways, network overlap across different replicates or bootstrap data sets. The computational burden and the extreme dimensionality of the network problem (19,000 nodes in the 8-cancer project) makes this a challenging task indeed.

Bibliography

T. Abenius, R. Jornsten, T. Kling, L. Schmidt, J. Sánchez, and S. Nelander. System-scale network modeling of cancer using epoc. *Goryanin, I., Goryachev, A. Advances in Systems Biology. Advances in Experimental Medicine and Biology 736, Springer.*, pages 617–643, 2012.

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.

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.

S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning.*, 3(1):1–122, 2011.

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

P. Danaher, P. Wang, and D. Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *arXiv:1111.0324v1*, 2011.

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.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the America Statistical Association.*, 96: 1348–1360, 2001.

J. Friedman, T. Hastie, H. Hoefling, and R. Tibshirani. Pathwise coordinate optimization. *Annals of Applied Statistics.*, 1(2):302–322, 2007.

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

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.

H. Hoefling. personal communication. 2012.

R. Jornsten, 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.

T. Kling, P. Johansson, J. Sánchez, R. Jornsten, and S. Nelander. Cancer landscapes: global network modeling of the cancer genome atlas linked to online pathway analysis and design of anticancer perturbations. *In preparation*, 2013.

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. Myllymaki, 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.

M. Rosvall and C. T. Bergstrom. Mapping change in large networks. *PLoS ONE.*, 5(1):e8694, 2010.

D. B. Sharma, H. D. Bondell, and H. H. Zhang. Consistent group identification and variable selection in regression with correlated predictors. *Journal of Computational and Graphical Statistics. In press.*, 2012.

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.

D. Witten and R. Tibshirani. Covariance-regularized regression and classification for high-dimensional problems. *Journal of the Royal Statistical Society: Series B.*, 71(3):615–636, 2009.

G. Ye and X. Xie. Split bregman method for large scale fused lasso. *Computational Statistics and Data Analysis.*, 55(4):1552–1569, 2011.

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

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, 2007b.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association.*, 101(476):1418–1429, 2006.

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.