OULU 2018
A 726

ACTA UNIVERSITATIS OULUENSIS

Markku Kuismin

ON REGULARIZED ESTIMATION METHODS FOR PRECISION AND COVARIANCE MATRIX AND STATISTICAL NETWORK INFERENCE

UNIVERSITY OF OULU GRADUATE SCHOOL;
UNIVERSITY OF OULU,
FACULTY OF SCIENCE

ACTA UNIVERSITATIS OULUENSIS
A Scientiae Rerum Naturalium 726

MARKKU KUISMIN

ON REGULARIZED ESTIMATION METHODS FOR PRECISION AND COVARIANCE MATRIX AND STATISTICAL NETWORK INFERENCE

Academic dissertation to be presented with the assent of the Doctoral Training Committee of Technology and Natural Sciences of the University of Oulu for public defence in the OP auditorium (L10), Linnanmaa, on 24 November 2018, at 12 noon

UNIVERSITY OF OULU, OULU 2018

Acta Univ. Oul. A 726, 2018

Supervised by
Professor Mikko J. Sillanpää

Reviewed by
Associate Professor Esa Ollila
Professor Kari Auranen

Opponent
Professor Ernst Wit

ISBN 978-952-62-2079-6 (Paperback)
ISBN 978-952-62-2080-2 (PDF)

ISSN 0355-3191 (Printed)
ISSN 1796-220X (Online)

Cover Design
Raimo Ahonen

JUVENES PRINT
TAMPERE 2018

Kuismin, Markku, On regularized estimation methods for precision and covariance matrix and statistical network inference.
University of Oulu Graduate School; University of Oulu, Faculty of Science
Acta Univ. Oul. A 726, 2018
University of Oulu, P.O. Box 8000, FI-90014 University of Oulu, Finland

Abstract

Estimation of the covariance matrix is an important problem in statistics in general because the covariance matrix is an essential part of principal component analysis, statistical pattern recognition, multivariate regression and network exploration, just to mention but a few applications. Penalized likelihood methods are used when standard estimates cannot be computed. This is a common case when the number of explanatory variables is much larger compared to the sample size (high-dimensional case). An alternative ridge-type estimator for the precision matrix estimation is introduced in Article I. This estimate is derived using a penalized likelihood estimation method.

Undirected networks, which are connected to penalized covariance and precision matrix estimation and some applications related to networks are also explored in this dissertation. In Article II novel statistical methods are used to infer population networks from discrete measurements of genetic data. More precisely, Least Absolute Shrinkage and Selection Operator, LASSO for short, is applied in neighborhood selection. This inferred network is used for more detailed inference of population structures. We illustrate how community detection can be a promising tool in population structure and admixture exploration of genetic data. In addition, in Article IV it is shown how the precision matrix estimator introduced in Article I can be used in graphical model selection via a multiple hypothesis testing procedure.

Article III in this dissertation contains a review of current tools for practical graphical model selection and precision/covariance matrix estimation. The other three publications have detailed descriptions of the fundamental computational and mathematical results which create a basis for the methods presented in these articles. Each publication contains a collection of practical research questions where the novel methods can be applied. We hope that these applications will help readers to better understand the possible applications of the methods presented in this dissertation.

Keywords: covariance matrix, graphical model, high-dimensional setting, LASSO, network estimation, precision matrix, ridge

Kuismin, Markku, Rajoitetuista tarkkuus- ja kovarianssimatriisin estimointimenetelmistä sekä tilastollisesta verkkojen päättelystä.
Oulun yliopiston tutkijakoulu; Oulun yliopisto, Luonnontieteellinen tiedekunta
Acta Univ. Oul. A 726, 2018
Oulun yliopisto, PL 8000, 90014 Oulun yliopisto

Tiivistelmä

Kovarianssimatriisin estimointi on yleisesti ottaen tärkeä tilastotieteen ongelma, koska kovarianssimatriisi on oleellinen osa pääkomponenttianalyysia, tilastollista hahmontunnistusta, monimuuttujaregressiota ja verkkojen tutkimista, vain muutamia sovellutuksia mainitakseni. Sakotettuja suurimman uskottavuuden menetelmiä käytetään sellaisissa tilanteissa, joissa tavanomaisia estimaatteja ei voida laskea. Tämä on tyypillistä tilanteessa, jossa selittävien muuttujien lukumäärä on hyvin suuri verrattuna otoskokoon (englanninkielisessä kirjallisuudessa tämä tunnetaan nimellä "high dimensional case"). Ensimmäisessä artikkelissa esitellään vaihtoehtoinen harjanne (ridge)-tyyppinen estimaattori tarkkuusmatriisin estimointiin. Tämä estimaatti on johdettu käyttäen sakotettua suurimman uskottavuuden estimointimenetelmää.

Tässä väitöskirjassa käsitellään myös suuntaamattomia verkkoja, jotka liittyvät läheisesti sakotettuun kovarianssi- ja tarkkuusmatriisin estimointiin, sekä joitakin verkkoihin liittyviä sovelluksia. Toisessa artikkelissa käytetään uusia tilastotieteen menetelmiä populaatioverkon päättelyyn epäjatkuvista mittauksista. Tarkemmin sanottuna Lassoa (Least Absolute Shrinkage and Selection Operator) sovelletaan naapuruston valinnassa. Näin muodostettua verkkoa hyödynnetään tarkemmassa populaatiorakenteen tarkastelussa. Havainnollistamme, kuinka verkon kommuunien (communities) tunnistaminen saattaa olla lupaava tapa tutkia populaatiorakennetta ja populaation sekoittumista (admixture) geneettisestä datasta. Lisäksi neljännessä artikkelissa näytetään, kuinka ensimmäisessä artikkelissa esiteltyä tarkkuusmatriisin estimaattoria voidaan käyttää graafisessa mallinvalinnassa usean hypoteesin testauksen avulla.

Tämän väitöskirjan kolmas artikkeli sisältää yleiskatsauksen tämänhetkisistä työkaluista, joiden avulla voidaan valita graafinen malli ja estimoida tarkkuus- sekä kovarianssimatriiseja. Muissa kolmessa julkaisussa on kuvailtu yksityiskohtaisesti olennaisia laskennallisista ja matemaattisista tuloksista, joihin artikkeleissa esitellyt estimointimenetelmät perustuvat. Jokaisessa julkaisussa on kokoelma käytännöllisiä tutkimuskysymyksiä, joihin voidaan soveltaa uusia estimointimenetelmiä. Toivomme, että nämä sovellukset auttavat lukijaa ymmärtämään paremmin tässä väitöskirjassa esiteltyjen menetelmien käyttömahdollisuuksia.

Asiasanat: graafinen malli, high-dimensional setting, kovarianssimatriisi, Lasso, ridge, tarkkuusmatriisi, verkkojen estimointi

Tosiasioiden tunnustaminen on kaiken viisauden alku. - J.K. Paasikivi

Acknowledgements

I am deeply grateful to my supervisor, Professor Mikko J. Sillanpää, for advice, new ideas and close collaboration. I am grateful to Dr. Jukka Kemppainen and Dr. Jon Ahlinder for their collaboration. I wish to thank Dr. Tanja Pyhäjärvi for useful discussions concerning the population analysis of teosinte samples. I am very grateful to my pre-examiners, Associate Professor Esa Ollila and Professor Kari Auranen, for their careful reading of the manuscript and Professor Ernst Wit for agreeing to be my opponent. I am also grateful to Ashely Last, Adjunct Professor Markus Harju, Dr. Ilkka Launonen, Professor Phillip Watts and language consultant Joshua Ward, who have greatly helped to improve the written English of this work. This work was supported by the University of Oulu's Technology and Natural Sciences Doctoral Programme (TNS-DP, formerly Exactus).

Oulu, September 2018 Markku Kuismin

Abbreviations

AIC	Akaike Information Criterion
BH	Benjamini-Hochberg
BIC	Bayesian Information Criterion
CONE	Community Oriented Network Estimation
CRAN	Comprehensive R Archive Network
eBIC	Extended Bayesian Information Criterion
EED	Edge Exclusion Deviance
FDR	False Discovery Rate
GGM	Gaussian Graphical Model
glasso	Graphical LASSO
LASSO	Least Absolute Shrinkage and Selection Operator
LDA	Linear Discriminant Analysis
MLE	Maximum Likelihood Estimate
MSE	Mean Squared Error
PCA	Principal Component Analysis
QDA	Quadratic Discriminant Analysis
ROPE	Ridge Operated Precision Matrix Estimator
StARS	Stability Approach to Regularization Selection

List of original publications

This dissertation is based on the following articles, which are referred to in the text by their Roman numerals (I–IV):

I Kuismin M., Kemppainen J. & Sillanpää M. J. (2017) Precision matrix estimation with ROPE. Journal of Computational and Graphical Statistics 26: 682–694.

II Kuismin M., Ahlinder J. & Sillanpää, M. J. (2017) CONE: Community oriented network estimation is a versatile framework for inferring population structure in large-scale sequencing data. G3 (Bethesda) 7: 3359–3377.

III Kuismin M. & Sillanpää M. J. (2017) Estimation of covariance and precision matrix, network structure, and a view toward systems biology. Wiley Interdisciplinary Reviews: Computational Statistics: 9:e1415.

IV Kuismin M. & Sillanpää M. J. Keep it simple: Parametric network inference without FDR control improves network structure analysis. Manuscript.

Author contributions

I All authors were involved in the conception and design of the method. MK performed the example analyses, implemented the method and drafted the manuscript. All authors interpreted results and critically revised the manuscript.

II All authors were involved in the conception and design of the method. MK performed the example analyses, implemented the method and drafted the manuscript. All authors interpreted results and critically revised the manuscript.

III Both authors were involved in the conception and design of the manuscript. MK designed the examples and figures and drafted the manuscript. Both authors critically revised the manuscript.

IV Both authors were involved in the conception and design of the method. MK performed the example analyses, implemented the method and drafted the manuscript. Both authors interpreted results and critically revised the manuscript.

Contents

1 Introduction

1.1 Introduction to covariance matrix and precision matrix estimation

Whenever examining more than one variable in statistical analysis, one of the most interesting and basic statistical analyses involves determining a good quantity for the dependency and dispersion between different random variables. These quantities include the variance, covariance, Pearson's correlation coefficient (hereafter correlation) and partial correlation coefficient. Standard parameters in multivariate inference are the *covariance matrix*, *partial correlation matrix* and the inverse of the covariance matrix called the *precision matrix*. These matrices are examined in this work. We use the term *variance matrices* when referring to all of these matrices in general. In articles I, II and III we will concentrate particularly on the precision matrix. However, it is somewhat more practical and comprehensive to first introduce the basic concepts of the covariance matrix, which are closely related to all other variance matrices.

Assume that $\mathbf{Y}_k = (Y_1, \ldots, Y_p)^\top$ is a multidimensional random vector, $\mathbf{Y}_k \in \mathbb{R}^p$, where $p \geq 2$. Superscript "$\top$" denotes the vector or matrix transposition. Each entry Y_i is a scalar-valued random variable, $i = 1, \ldots, p$. Assume that $\mathbf{Y}_k$ has a multivariate probability distribution $p(\Sigma)$ with covariance matrix $\Sigma = [\sigma_{ij}]$. In general, we denote this $\mathbf{Y}_k \sim p(\Sigma)$. The covariance between random variables Y_i and Y_j is defined as

$$\sigma_{ij} = \mathbb{E}\left\{(Y_i - \mu_i)(Y_j - \mu_j)\right\}, \tag{1}$$

where μ_i is the expected value of the random variable Y_i, $\mu_i = E(Y_i)$ for $i = 1, \ldots, p$. Variances of random vectors are found on the diagonal of the covariance matrix, $\sigma_{ii} = var(Y_i)$, $i = 1, \ldots, p$, and they determine the spread of the vector entries. In general, Σ is unknown and it is estimated using a random sample from the distribution $p(\Sigma)$.

Assume that one has n i.i.d. samples from a probability distribution. These samples can be stacked into an $n \times p$ data matrix Y, $Y = (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)^\top$. The unbiased sample covariance matrix estimator $S = [s_{ij}]$ is a $p \times p$-dimensional symmetric matrix defined as

$$S = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{Y}_i - \bar{\mathbf{Y}})(\mathbf{Y}_i - \bar{\mathbf{Y}})^\top, \tag{2}$$

where $\bar{\mathbf{Y}} = n^{-1}\sum_{i=1}^{n}\mathbf{Y}_i$ is the sample mean vector. In numerical analysis, or when the sample size n is large, it is irrelevant if one scales the right-hand side of equation (2) with $1/(n-1)$ or $1/n$.

How does one interpret the covariance? Firstly, if it looks like the variable Y_i tends to increase when the other variable Y_j also increases, then the corresponding covariance coefficient σ_{ij} is positive (without commenting on its volume). Secondly, if the other variable tends to decrease when the other increases, then the covariance coefficient is negative (without commenting on its volume).

1.1.1 Covariance matrix eigenvalues and eigenvectors

Both the covariance matrix and the sample covariance matrix are always symmetric and positive semidefinite, that is, $\Sigma = \Sigma^\top$, $S = S^\top$, $(\Sigma\mathbf{X},\mathbf{X}) \geq 0$ and $(S\mathbf{X},\mathbf{X}) \geq 0$ for an arbitrary non-zero vector $\mathbf{X} \in \mathbb{R}^p$, where $(S\mathbf{X},\mathbf{X}) = \mathbf{X}^\top S\mathbf{X}$ is the inner product. The eigenvalues of a symmetric[1] square matrix lie on the real axis. Because the (sample) covariance matrix is always positive semidefinite, its eigenvalues lie on the positive half of the real axis. The eigenvalues of a positive definite and symmetric matrix are always real numbers greater than zero. Both the covariance and the precision matrix share the same eigenvectors. The eigenvalues of the precision matrix are simply reciprocals of the eigenvalues of the covariance matrix.

Eigenvalues of the covariance matrix Σ correspond to the magnitude of the variances of variables stored into the data (matrix) Y along the respective eigenvector directions.

In this work we mainly examine random vectors $\mathbf{Y}$ that are assumed to have a p-variate multivariate normal distribution with mean vector μ and covariance matrix Σ. The probability density function $p(\mathbf{Y}|\mu,\Sigma)$ of the random vector $\mathbf{Y}$ is then

$$p(\mathbf{Y}|\mu,\Sigma) = \det(2\pi\Sigma)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{Y}-\mu)^\top\Sigma^{-1}(\mathbf{Y}-\mu)\right\}, \tag{3}$$

where $\det(A)$ is the determinant of a real-valued square matrix A. Clearly, the probability density function (3) is defined only if Σ is invertible. Thus, in this particular case one has to actually find a positive definite estimate for Σ. Without loss of generality, assume that the mean vector μ in (3) is a zero vector and denote the distribution of the vector $\mathbf{Y}$ with $\mathbf{Y} \sim N(\mathbf{0},\Sigma)$. Then the sample covariance matrix estimate $S = n^{-1}Y^\top Y$ is the

[1] In this work we will concentrate on real valued matrices. Thus we will use the term symmetric instead of Hermitian.

maximum likelihood estimate (MLE) of the covariance matrix. Figure 1 illustrates the connection between eigenvalues, eigenvectors and data dispersion of the multivariate normal distribution in a simple two-dimensional example.

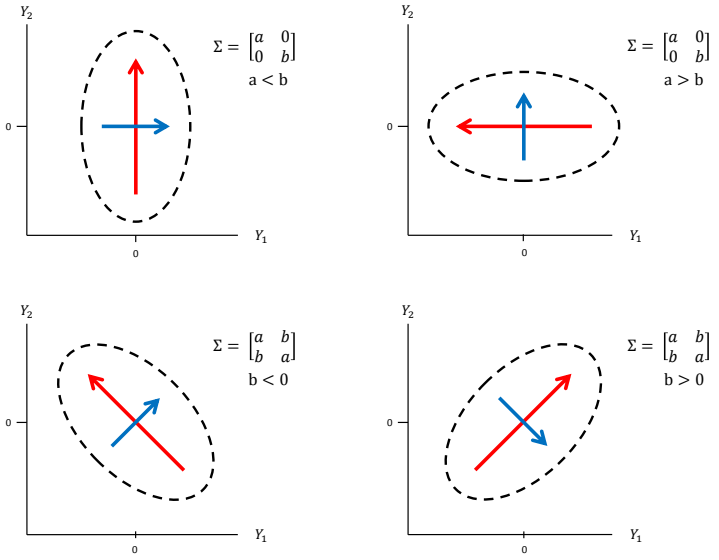


Fig. 1. A schematic illustration of the covariance matrix Σ eigenvalues, eigenvectors and data dispersion of a random sample $\mathbf{Y} = (Y_1, Y_2)^\top$. Red arrows correspond to the eigenvectors of the largest eigenvalue and blue arrows correspond to the smallest eigenvalue. Black dashed lines illustrate the general form of the data dispersion. The data are assumed to be centered around zero, $\mathbf{Y} \sim N(\mathbf{0}, \Sigma)$.

We would like to estimate Σ with a sample of size n. A natural choice is the sample covariance matrix S because it is a consistent estimate of the covariance matrix, that is

$$\mathbb{P}(|s_{ij} - \sigma_{ij}| > \varepsilon) \to 0, \quad n \to \infty, \tag{4}$$

for any small real-valued $\varepsilon > 0$, $i, j = 1, \ldots, p$. If n is sufficiently large, S can be used to estimate covariances.

According to the law of large numbers, one can compute numerically reliable estimates for the distinct covariance matrix coefficients assuming that the sample size is very large compared to the problem dimension. However, in a variety of data analysis

tasks the available resources restrict the number of samples and thus the sample size is usually very small as compared to the problem dimension p. In the worst case scenario – which is not uncommon in real life problems – the sample size is smaller than the number of variables of interest; this setting is commonly known as the *high-dimensional setting* or *"large p, small n"* scenario (with $p > n$ or even $p \gg n$) in the statistical literature. In the high-dimensional setting the sample covariance coefficients are highly biased as compared to the true covariances. Because the matrix $Y^\top Y$ has at most $min(n, p)$ non-zero eigenvalues, the inverse of the sample covariance matrix S^{-1} does not exist in the high-dimensional setting, because the sample covariance matrix is not positive definite. In addition, Ledoit & Wolf (2004a,b) have shown that when the sample size n is larger than p the eigenvalues of S are far away from the ones of Σ in a way that the small eigenvalues are underestimated and the large eigenvalues are overestimated.

In summary, if the proportion n/p is less than one or close to one, the following holds for the sample covariance matrix S:

1. The estimator is of poor quality: in other words, risk measures such as the mean squared error (MSE), $\text{MSE}(S) = \mathbb{E}\left\{||S - \Sigma||_2^2\right\}$, is large.
2. S is not positive definite and Σ^{-1} cannot be estimated by computing the inverse of S.
3. Eigenvalues are dispersed and the estimate is ill-conditioned, which means that the condition number (the largest eigenvalue divided by the smallest eigenvalue) of the matrix is large.

Above, $||A||_2^2 = \sum_{i,j=1}^{p \times p} a_{ij}^2$ is the squared Euclidean norm of a matrix A. MSE can be interpreted as a matrix risk function or measure. The above-mentioned aspects have serious effects on applications such as principal component analysis (PCA), statistical network estimation, portfolio optimization, linear and quadratic discriminant analysis (LDA, QDA) and overall in almost every field of multivariate analysis.

As a side note, there is a field of studies that investigates high-dimensional asymptotics related to high-dimensional settings, where not only the sample size n goes to infinity but also the problem dimension (the number of variables p) approaches infinity while the ratio p/n converges to a constant. Related to covariance estimators, high-dimensional asymptotics have been studied by Ledoit & Wolf (2004b, 2012, 2015) via shrinking the eigenvalues of the sample covariance matrix. In our work we have not studied high-dimensional asymptotics and we mainly concentrate on the classical asymptotic scenario where the ratio p/n converges to zero while n increases.

In the next section we demonstrate how one can gain better conditioned and positive definite estimates for both the covariance and the precision matrices. In particular, in Article I we examined how to compute positive definite estimates for the precision matrix, which we will denote by $\Sigma^{-1} = \Theta$ from now on.

2 Novel methods for covariance and precision matrix estimation

One of the basic problems in numerical matrix analysis is computing the inverse of a square matrix. This is because matrix inversion requires $\mathcal{O}(p^3)$ floating point operations ("flops"). Performing so many flops is in most cases too laborious to compute and prone to numerical errors. In this work the problem is inverting the sample covariance matrix to have an estimate for the precision matrix Θ.

In a high-dimensional setting one cannot compute the matrix inverse of the sample covariance matrix because the sample covariance matrix is singular. We cannot always perform precision matrix estimation in less than $\mathcal{O}(p^3)$ operations but we usually can estimate it without inverting the sample covariance matrix. Before we discuss the estimation of Θ we have to provide the basics which help us describe and distinguish methods from each other and talk about LASSO and *ridge* regression.

2.0.1 LASSO: Least absolute shrinkage and selection operator

Discussing LASSO (least absolute shrinkage and selection operator) regression, introduced in Tibshirani (1996), helps us understand the basic theoretical and conceptual methodology involved in the precision matrix estimation methods discussed in this dissertation. When $\mathbf{Y}$ is the $n \times 1$ outcome vector and X is the $n \times p$ covariate matrix, one would like to predict $\mathbf{Y}$ with a linear model $\mathbf{Y} = X\beta$, where $\beta = (\beta_1, \ldots, \beta_p)^\top$ is the vector of regression coefficients. LASSO estimate $\widehat{\beta}_{LASSO}(\lambda)$ is the solution of the optimization problem

$$\min ||\mathbf{Y} - X\beta||_2^2 + \lambda ||\beta||_1, \tag{5}$$

where $||\mathbf{b}||_2^2 = \sum_{i=1}^p b_i^2$ is the squared Euclidean norm of the vector and $||\beta||_1 = \sum_{i=1}^p |\beta_i|$ is the sum of the absolute values of the entries of vector β. The constraint $\lambda||\beta||_1$ is sometimes referred to as the LASSO penalty. A positive constant λ is a tuning parameter, the penalty parameter.

In general, LASSO estimates depend on λ and increasing the penalty parameter in (5) causes some elements of the LASSO estimate $\widehat{\beta}_{LASSO}(\lambda)$ to *shrink* toward zero, and some of them can be exactly zero if λ is substantially large. Because of this shrinkage

property, using LASSO to estimate linear models will produce more interpretable models and the LASSO penalty tends to reduce the prediction error depending on the value of the tuning parameter λ (see, Huang 2003, Rosset & Zhu 2004). LASSO and the properties of the LASSO penalty have been essential tools in statistical network estimation in the beginning of the *21st* century and the 2010s. We will discuss network estimation using LASSO in more detail in the *Neighborhood selection* section.

2.1 Penalized covariance and precision matrix estimation utilizing LASSO penalty

Suppose that the data matrix $Y = (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)^\top$ is a random sample from $N(\mathbf{0}, \Sigma)$ with probability distribution function (3). The likelihood function of the parameters defining the covariance matrix, based on data Y, is

$$p(Y|\Sigma) = \prod_{i=1}^{n} p(\mathbf{Y}_i|\mathbf{0}, \Sigma) = \det(2\pi\Sigma)^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} \mathbf{Y}_i^\top \Sigma^{-1} \mathbf{Y}_i \right\}. \tag{6}$$

The log-likelihood function is simply the logarithm of the likelihood function $\log p(Y|\Sigma)$. The log-likelihood function of the multivariate normal distribution can be written as a function of matrix products up to a constant,

$$\log p(Y|\Theta) \propto \log\{\det(\Theta)\} - tr(S\Theta), \tag{7}$$

where $\log(.)$ is the natural logarithm and $tr(A) = \sum_{i=1}^{p} a_{ii}$ is the trace of a symmetric matrix A. Note that equation (7) uses matrix notation Θ instead of Σ^{-1} which is commonly used in the literature related to penalized covariance and precision matrix estimation. Assume that the mean is known, and without loss of generality assume that $\mu = \mathbf{0}$ (zero vector).

Both Banerjee et al. (2008) and Friedman et al. (2008) proposed that one can produce a sparse estimate for the precision matrix by maximizing a penalized log-likelihood in terms of Θ when the penalized log-likelihood is

$$\log\{\det(\Theta)\} - tr(S\Theta) - \lambda||\Theta||_1, \tag{8}$$

where $||A||_1 = \sum_{i,j=1}^{p \times p} |a_{ij}|$ is the L_1-norm of a matrix A, sometimes referred to as the L_1-penalty. The positive parameter λ is again referred to as the penalty/tuning parameter. Penalized log-likelihood (8) resembles the LASSO minimization problem (5) and

Friedman et al. (2008) named the solution algorithm as graphical LASSO, glasso for short.

Similar to LASSO, glasso will also shrink some of the off-diagonal elements of the precision matrix Θ exactly to zero. This sparse precision matrix estimate can be used in statistical network estimation. Glasso does not require inversion of the sample covariance matrix to compute estimate $\widehat{\Theta}$ and one can actually also estimate the covariance matrix with the same algorithm without matrix inversion (Friedman et al. 2008). More surprisingly, the glasso algorithm can be computed in less than $\mathcal{O}(p^3)$ operations; the model complexity depends on the tuning parameter λ and the operations required will be greatly reduced when more zeros are induced in the glasso estimate of Θ (Witten et al. 2011). Another good property of the glasso algorithm is the fact that the estimate of the precision or the covariance matrix will always be positive definite even in the high-dimensional setting ($p \gg n$). A great deal of work related to penalized precision matrix estimation has been focused on the optimization of the solution algorithm of the L_1-regularized log-likelihood (Banerjee et al. 2008, Friedman et al. 2008, Cai et al. 2011, Witten et al. 2011, Hsieh et al. 2013, 2014, Liu & Luo 2015, Liu & Wang 2017). In addition to the above-mentioned "frequentist" methods, Khondker et al. (2013) and Wang (2012) have proposed Bayesian implementations of the glasso estimator. However, Bayesian glasso implementations lack the computational efficiency of the glasso solution algorithms when the problem dimension is very high.

We will not discuss the maximization of (8) in more detail because the theoretical properties of glasso are not examined in any article of this dissertation.

2.1.1 Ridge regression

Before we describe Article I in more detail, we have to review ridge regression (Hoerl & Kennard 1970). Assume that vectors β, $\mathbf{Y}$ and the covariate matrix X are similar as in the optimization problem (5). Ridge regression minimizes the following expression:

$$||\mathbf{Y} - X\beta||_2^2 + \lambda||\beta||_2^2, \tag{9}$$

where the constraint $\lambda||\beta||_2^2$ is sometimes called the ridge penalty. By examining the gradient function of (9), it is easy to show that the ridge solution $\widehat{\beta}_{ridge}(\lambda)$ is equal to vector $(X^\top X + \lambda I)^{-1} X^\top \mathbf{Y}$.

Ridge regression shrinks all regression coefficients in β toward zero when λ increases. This is easily seen by examining the eigenvalues of $(X^\top X + \lambda I)^{-1}$ which

will approach zero when λ approaches infinity. Naturally, this is only an abstract asymptotic result and in practical data analysis coefficients β_i will never be exactly zero. Ridge regression thus shrinks regression coefficients but does not perform model selection similar to LASSO. However, the ridge penalty reduces the mean squared error of the ridge-estimate $\widehat{\beta}_{ridge}(\lambda)$, $\text{MSE}(\widehat{\beta}_{ridge}(\lambda)) = \mathbb{E}\left\{||\widehat{\beta}_{ridge}(\lambda) - \beta||_2^2\right\}$, as compared to the least squared estimator $\widehat{\beta} = (X^\top X)^{-1}X^\top \mathbf{Y}$, depending on the value of the tuning parameter (Hoerl & Kennard 1970). One of the original motivations of Hoerl and Kennard (Hoerl & Kennard 1970) was to make $X^\top X$ invertible in the high-dimensional case.

2.1.2 Penalized covariance and precision matrix estimation utilizing ridge penalty

The simple-looking augmentation $X^\top X + \lambda I$ of the matrix product $X^\top X$ plays an important role in a class of covariance matrix estimators introduced next and gives insight into the method described in Article I. To begin with, we will discuss a more general class of estimators.

Instead of using S as the estimator of Σ one can use the following linear combination,

$$\widehat{\Sigma} = \lambda_1 I + \lambda_2 S, \tag{10}$$

where λ_1 and λ_2 are positive scalars and in general $\lambda_1 + \lambda_2 = 1$. I is the $p \times p$ identity matrix. All estimators of the form (10) are symmetric and positive definite estimates of Σ. This follows from the fact that S is a diagonalizable matrix, $S = MLM^{-1}$ which is also the *eigendecomposition* of the sample covariance matrix. In this decomposition M is actually an orthogonal matrix, $M^{-1} = M^\top$, whose columns are the eigenvectors of the sample covariance matrix. L is a diagonal matrix whose diagonal elements are the eigenvalues of S. The eigenvalue and eigenvector pair are located on the same column in both M and L. Furthermore, estimators of form (10) are rotation equivariant, meaning that they share the same eigenvectors as the sample covariance matrix.

A special class of the estimators (10) is derived by setting λ_2 equal to one and $\lambda_1 \in [0, \infty[$. The estimator (10) thus takes the form $S + \lambda I$, which resembles the augmentation proposed by Hoerl & Kennard (1970). A covariance matrix estimator of the form $\widehat{\Sigma}_{ridge} = S + \lambda I$ is called a *ridge estimator* because it resembles the augmented expression $X^\top X + \lambda I$ in the ridge solution $\widehat{\beta}_{ridge}(\lambda)$.

While glasso estimators share similar properties with LASSO, ridge estimators share similar properties with ridge regression: the estimate computed with $\widehat{\Sigma}_{ridge}$ is invertible, although S would be singular and one can compute an estimate for the precision matrix $\widehat{\Theta}_{ridge} = (S + \lambda I)^{-1}$. Similar to ridge regression, all off-diagonal elements of the penalized estimate $\widehat{\Theta}_{ridge}$ are shrunk toward zero.

Computing $\widehat{\Theta}_{ridge}$ is somewhat problematic: matrix inversion is usually avoided in practical multivariate analysis since calculations made with computers cause more or less numerical errors, not to mention how laborious and time consuming matrix inversion is. Thus inverting $\widehat{\Sigma}_{ridge}$ is not the best way to compute an estimate for the precision matrix when p is very large. Moreover, in Warton (2008) and van Wieringen & Peeters (2016) it has been shown that (10) is the maximum likelihood estimator of the penalized normal likelihood of the form

$$\log\{\det(\Theta)\} - \lambda_2 tr(S\Theta) - \lambda_1 tr(\Theta). \tag{11}$$

In particular, the ridge estimator $\widehat{\Sigma}_{ridge} = S + \lambda I$ results from the following penalized log-likelihood:

$$\log\{\det(\Theta)\} - tr(S\Theta) - \lambda tr(\Theta). \tag{12}$$

It is obvious that the penalized log-likelihood (12) does not resemble the penalized residual sum of squares equation (9) and that $\lambda tr(\Theta) = \lambda \sum_{i=1}^{p} \theta_{ii}$ is not the same as the ridge penalty $\lambda ||\beta||_2^2$.

Motivated by the inconsistency between (12) and (9), we introduced a "real" ridge estimator. In Article I we maximize a penalized log-likelihood

$$\log\{\det(\Theta)\} - tr(S\Theta) - \lambda ||\Theta||_F^2, \tag{13}$$

where $||A||_F^2 = \sum_{i,j=1}^{p} a_{ij}^2 = tr(A^2)$ is the squared Frobenius norm of a matrix A and $A^2 = A^\top A$ (symmetric matrix product).

The penalty function in (13) is more coherent with the ridge penalty when compared to (9). It is straightforward to show that the optimal solution of (13) can be found by solving a quadratic matrix equation,

$$2\lambda\Theta^2 + S\Theta - I = 0_p, \tag{14}$$

where 0_p is a $p \times p$ zero matrix. Utilizing the properties of the Riccati equations, we have shown that one can compute a penalized MLE of Θ without inverting the sample

covariance matrix. Applying the eigendecomposition of the sample covariance matrix $S = MLM^\top$ the estimator can be written as $\widehat{\Theta}_{ROPE} = MDM^\top$, where the elements of the diagonal matrix $D = \text{diag}(d_1, \ldots, d_p)$ are given by

$$d_i = \frac{2}{l_i + \sqrt{l_i^2 + 8\lambda}}, \tag{15}$$

l_i is the ith eigenvalue of the sample covariance matrix S, $i = 1, \ldots, p$ and λ is the tuning parameter in (13). We call the estimator $\widehat{\Theta}_{ROPE}$ ROPE (ridge operated precision matrix estimator) and consider it as the "true" ridge estimator of the precision matrix.

With ROPE, we can determine a positive definite estimate for both the covariance and the precision matrix without inverting the sample covariance matrix but we still need to compute its eigendecomposition. Thus more than $\mathcal{O}(p^3)$ operations are needed to compute the ROPE estimate. Some precision matrix estimates can be computed in less than $\mathcal{O}(p^3)$ flops (see Witten et al. 2011, Hsieh et al. 2013). Nevertheless, we tolerate $\mathcal{O}(p^3)$ operations because choosing the tuning parameter λ from the data (typically by cross-validation) can dispel problems with the computational speed of the algorithm. Choosing the tuning parameter usually requires determining the estimate for several different values of the tuning parameter and then finding the minimum/maximum of a goodness-of-fit measure. Thus one needs to run the estimation algorithm repeatedly for each value of λ. With ROPE, the eigendecomposition of the sample covariance matrix is not needed to be computed repeatedly. When using ROPE, matrix determinants and traces like in (7) can be computed much faster because one can utilize the eigenvalues instead of calculating the determinant of the full matrix or the product of two full matrices.

Before introducing the properties of ROPE, we note that Article I was developed independently and concurrently with the papers of van Wieringen & Peeters (2015), van Wieringen & Peeters (2016) (while writing this dissertation we also found the paper of Honorio & Jaakkola (2013) which is also closely related to our work). These authors were concerned with the same inconsistency between the ridge penalty and the penalty functions in (11) and (12). Moreover, they augmented the penalty function in (13) and maximized the following penalized log-likelihood

$$\log\{\det(\Theta)\} - tr(S\Theta) - \lambda||\Theta - T||_F^2, \tag{16}$$

where T is a symmetric and positive definite matrix referred to as the target matrix. It is a breeze to show that the sub-gradient equation when maximizing (16) can still

be expressed as a quadratic matrix equation by substituting S with $S^{'} = S - 2\lambda T$ in (14). This target matrix can be very convenient to use in some applications because van Wieringen & Peeters (2016) have shown that ROPE approaches T as λ approaches infinity. By default we assume that T is a $p \times p$ zero matrix.

ROPE shares similar characteristics with the ridge estimator of the covariance matrix. The precision matrix estimated with ROPE is always symmetric and positive definite even if S is not. This makes ROPE practicable also in high-dimensional settings. Furthermore, in van Wieringen (2017) it has been shown that under some special conditions the covariance and the precision matrix estimated with ROPE dominates the MLE of the covariance matrix in the terms of MSE. ROPE also has a Bayesian interpretation when the target matrix is chosen as a $p \times p$ zero matrix and the prior distribution is proportional to the generalized gamma distribution. However, Bayesian estimation of ROPE is not examined in this dissertation.

Similar to the estimators of form (10), ROPE is a rotation equivariant estimator, depending on the choice of the target matrix. Because of this, one cannot obtain better estimates for the eigenvectors of the covariance matrix with these estimators. One could say that the ROPE and ridge estimators are actually *shrinkage methods of the eigenvalues*. Moreover, we could say that ROPE is rather a nonlinear shrinkage method whereas the ridge estimator is a linear shrinkage method. Understanding this, we need to explain the concepts of *linear* and *nonlinear* shrinkage of the eigenvalues introduced in Ledoit & Wolf (2004a,b, 2012, 2015).

2.1.3 Linear and nonlinear shrinkage of the eigenvalues

Previously we have described that LASSO and ridge regression shrink regression coefficient estimates toward zero, in other words, these coefficients become smaller in absolute value when the LASSO penalty or the ridge penalty increases. This is not the case while describing shrinkage of eigenvalues. As mentioned before, the sample covariance matrix is always positive semidefinite and the covariance matrix is positive definite. All variance matrices are also symmetric and because of this their eigenvalues lie on the positive half of the real axis.

The intuition underlying the eigenvalue shrinkage is fairly simple. As mentioned in the *Covariance matrix eigenvalues and eigenvectors* subsection, Ledoit & Wolf (2004a,b) showed that the eigenvalues of the sample covariance matrix are biased such that the small ones are biased downwards and the large ones upwards in comparison to

the true eigenvalues of Σ. The authors of Ledoit & Wolf (2004a,b) proposed a following remedy to the problem: first define an estimator as a special case of (10):

$$\widehat{\Sigma} = \lambda vI + (1-\lambda)S, \tag{17}$$

where λ is called the *shrinkage intensity* varying between 0 and 1. Scalar v is the *grand mean* of the sample eigenvalues, that is $v = tr(S)/p$. When using estimator (17) one applies a shrinkage intensity of the same level to all sample eigenvalues (relative to the distance to the grand mean) and "moves" the eigenvalues toward the grand mean of the sample eigenvalues. As described in Ledoit & Wolf (2012): "For example, if the linear shrinkage intensity is 0.5, then every sample eigenvalue is moved half-way toward the grand mean of all sample eigenvalues" (p. 2). In other words, estimator (17) pushes small eigenvalues away from zero toward the grand mean and large eigenvalues toward the grand mean (toward zero) all at the same proportional amount. The original authors of Ledoit & Wolf (2004a,b) call this *linear shrinkage* of the eigenvalues.

Nonlinear shrinkage of the eigenvalues is described in Ledoit & Wolf (2012, 2015). Nonlinear shrinkage differs from linear shrinkage as follows: rather than moving small and large eigenvalues of the sample covariance matrix away/toward zero of the same level, the amount of shrinkage of each eigenvalue is defined individually. It follows from this that small eigenvalues may not always move drastically upwards and large eigenvalues may increase. Overall, nonlinear shrinkage is far more complicated to carry out than linear shrinkage, and we will not describe it in any more detail in this dissertation.

Interpreting ROPE and ridge type estimators (10) in terms of linear and nonlinear eigenvalue shrinkage is one way to gain insight into the characteristics (and differences) of these estimators. Here we compare the eigenvalues of ROPE with the eigenvalues of the sample covariance matrix S. This is because the eigenvalues of S are somewhat easier to interpret than the eigenvalues of S^{-1}, because the latter may not exist or they can be very large, making it intuitively more difficult to understand the connection between the eigenvalues and the dispersion of the random sample.

In the following we assume that the target matrix is a $p \times p$ zero matrix, $T = 0_p$. The distance of each eigenvalue of the ridge estimator $S + \lambda I$ from the corresponding sample covariance matrix eigenvalues is λ, that is $|\sigma(\widehat{\Sigma}_{ridge} - S)_i| = \lambda$ for all $i = 1, \ldots, p$, where the set $\sigma(.)$ denotes the spectrum of a square matrix. This is not exactly the same as linear shrinkage in the sense described earlier but can be regarded as linear shrinkage:

the relationship between the sample eigenvalues and the eigenvalues of the ridge estimate is affine. ROPE, on the other hand, works differently. If the sample eigenvalue is zero, the corresponding eigenvalue of ROPE is $\sqrt{2\lambda}$. When the eigenvalues increase, the amount of upward shrinkage of large eigenvalues gets smaller. Here ROPE differs from the ridge estimator and resembles a special type of nonlinear shrinkage as the amount of shrinkage decreases for large eigenvalues, that is, the distance $|\sigma(\widehat{\Sigma}_{ROPE} - S)_i|$ is not the same for all $i = 1, \ldots, p$. In our opinion, this is a desired property compared to the ridge estimator of the covariance matrix because large eigenvalues of the sample covariance matrix are already biased upwards. Figure 2 illustrates how different methods shrink eigenvalues of the sample covariance matrix.

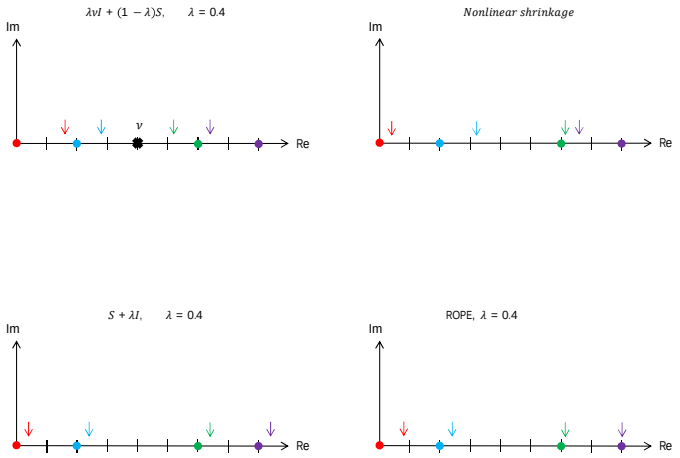


Fig. 2. A schematic illustration of the shrinkage of sample covariance matrix S eigenvalues along the real axis when the spectrum of the sample covariance matrix is $\sigma(S) = \{0, 2, 6, 8\}$. Each colored point illustrates the place of eigenvalues of S on the real axis and the arrow of the same color the place of the corresponding shrunken eigenvalues of each estimate. The linear shrinkage estimator $\lambda vI + (1-\lambda)S$ shrinks eigenvalues toward the grand mean v (the black cross mark). The nonlinear shrinkage estimator applies a different level of shrinkage to each eigenvalue and does not always shrink eigenvalues similarly to the linear shrinkage estimator. Ridge estimator $S + \lambda I$ moves each eigenvalue upwards by λ. The shrinkage induced by ROPE decreases for large eigenvalues when the tuning parameter λ is not substantially large.

2.1.4 Choosing the tuning parameter

Despite the fact that ridge estimators are easy to compute there is still the problem of how one chooses a proper value for the tuning parameter λ. Apart from the estimator of Ledoit & Wolf (2004b, 2012, 2015) which will select the tuning intensity analytically based on data, it is still somewhat an open problem to choose the value for λ. In practice, selecting the tuning parameter means scanning through tens or hundreds of different values of λ. This makes the estimation of the variance matrices computationally demanding. Because the eigenvalues of ROPE can be expressed in a closed form (15), we do not need to compute the eigendecomposition of the sample covariance matrix for each value of λ. Thus the set of solutions with different values of λ can be computed fairly cheaply. In this subsection we will briefly review the tuning parameter selection methods we have used in articles I, II and IV and mention a few alternatives.

So how does one select the tuning parameter? Typically, the tuning parameter is chosen based on the data. In articles I and IV we have mainly used 5-fold cross-validation to select the optimal value for the tuning parameter for ROPE and other ridge estimators (see, e.g., the original paper of Picard & Cook (1984) for a description of cross-validation). Performing 5-fold cross-validation can be computationally intensive even for moderately sized data sets because one has to divide the data into roughly five equal sized sets. Each of these data sets is then used as the validation set in its turn and all four remaining pieces act as the training set. After five validation-training divisions, the results are averaged to obtain one selection criterion value. The same cross-validation procedure is repeated for all candidate values of the tuning parameter. Finally, one chooses the value for λ that produces the maximal (minimal) selection criterion value. In particular, we have used the following selection criterion in articles I and IV

$$l(\widehat{\Theta}_\lambda, S_{valid}) = \log\det(\widehat{\Theta}_\lambda) - tr(S_{valid}\widehat{\Theta}_\lambda), \tag{18}$$

where $\widehat{\Theta}_\lambda$ is the precision matrix estimate computed with a penalized precision matrix estimator which depends on λ (such as ROPE) from the training set and S_{valid} is the sample covariance matrix computed from the validation set (Bien & Tibshirani 2011).

An attractive property of cross-validation is that it can be used in both covariance matrix and precision matrix estimation. In Liu & Luo (2015) it has been shown that under special conditions the sparse estimate chosen via cross-validation converges toward the true precision matrix under high-dimensional asymptotics.

In addition to k-fold cross-validation, there are the leave-one-out cross-validation, Bayesian information criterion (BIC) and Akaike information criterion (AIC) but we have not applied these methods extensively in our studies. In any case, we note that choosing the tuning parameter in a data-dependent way remains a challenging problem in high-dimensional settings.

3 Statistical network inference

3.1 Introduction to statistical network inference

To begin with, we would like to remind the reader that there is a fundamental difference between mathematical network analysis and network *inference* in statistics. In mathematical graph theory, social networks, telecommunications etc. *edges* of the network are (usually) known in advance and network analysis in these fields means more or less the analysis of the network structure. Statistical methods are used when the topological structure of the network is unknown *a priori*. Only associated variables are known beforehand but everything else has to be inferred from the data; *edges*, *communities*, *hubs* etc. The basic problem in statistical network inference is to determine whether the inferred network is a sufficient estimate of the ground truth network. In this sense, one could say that the network is estimated. In this work, the terms network and graph are used interchangeably in the same way that graphical model selection and network estimation mean the same thing. In particular, we are interested in estimation of Gaussian graphical models (GGMs).

Pinpointing the early historical work of Gaussian graphical models (GGMs) is not needed to understand the basic concepts of this dissertation. Of the pioneering articles related to GGMs, the most important is Dempster (1972) under the name of covariance selection models. More relevant articles for this dissertation are Meinshausen & Bühlmann (2006) and Friedman et al. (2008).

One does not need to delve deeply into the theoretical properties of networks and graphical models when examining undirected networks from the statistical point of view. In this work it is sufficient to determine just a few basic concepts of undirected graphical models, aka Markov random fields or Markov networks (see pp. 625–630 in Hastie et al. 2017).

A graphical model $G = (N,E)$ is determined by two sets: The capital N is the set of nodes $N = \{1,\ldots,p\}$ and E is the set of edges (i,j). The pair (i,j) – or the pair (j,i) – is included in E if and only if there is an edge between nodes i and j. In this work we have only considered undirected networks (in which the edge does not have a direction). Note that only one of the pairs (i,j) or (j,i) needs to be included in set E because in undirected networks they share the same information. Overall, the cardinality of set E can vary from 0 to $p(p-1)/2$, corresponding to an empty and a full graph, respectively.

Each element $i \in N$ corresponds to a random variable Y_i, $i = 1, \ldots, p$. One could also write $N = \{Y_1, \ldots, Y_p\}$ and we will use the i and Y_i notations interchangeably when referring to the ith node in this dissertation. In model $G = (N,E)$ set N is known whereas set E is unknown. The statistical network analysis examined in this dissertation basically culminates in the selection of the best candidate for the set of edges E.

The undirected graphical model G can be encoded into a symmetric $p \times p$ *adjacency matrix* $A = [A_{ij}]$ where

$$A_{ij} = \begin{cases} 1 & \text{if there is an edge between variables } i \text{ and } j, \\ 0 & \text{if there is no edge between variables } i \text{ and } j. \end{cases} \tag{19}$$

For an undirected graphical model, A is symmetric and there are no self-loops in the network (a node is not connected to itself with an edge) $A_{ii} = 0$, $i = 1, \ldots, p$.

GGMs are closely related to the pairwise Markov property which determines the conditional independence of two distinct variables Y_i and Y_j, that is, Y_i is conditionally independent of Y_j, given all remaining variables,

$$Y_i \perp Y_j \mid Y \setminus \{Y_i, Y_j\} \Leftrightarrow (i,j) \notin E, \tag{20}$$

where "$\Leftrightarrow$" reads as "if and only if" (see, e.g., Drton & Perlman 2007). In GGMs the pairwise Markov property is actually equivalent to the global Markov property (see pp. 434–435 in Bühlmann & Van De Geer 2011). In the terms of the adjacency matrix A, equation (20) is equivalent to

$$Y_i \perp Y_j \mid Y \setminus \{Y_i, Y_j\} \Leftrightarrow A_{ij} = 0. \tag{21}$$

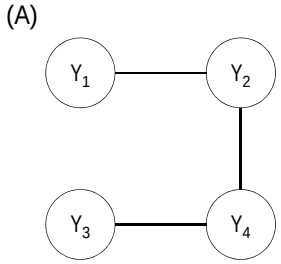
A simple GGM and corresponding adjacency matrix are illustrated in Figure 3 similar to the one in Drton & Perlman (2007).

In GGMs the conditional independence (20) can be expressed with the elements of the partial correlation matrix or with the precision matrix $\Theta = [\theta_{ij}]$. We will discuss the partial correlation matrix later in this section. For now we examine how the precision matrix is used in GGM selection.

The equations (19) and (20) are equivalent to

$$Y_i \perp Y_j \mid Y \setminus \{Y_i, Y_j\} \Leftrightarrow \theta_{ij} = 0. \tag{22}$$

Thus, one can determine the graphical model G by examining the non-zero support of the (sparse) precision matrix (see, e.g., Edwards 2000).



(B)

0	1	0	0
1	0	0	1
0	0	0	1
0	1	1	0

Fig. 3. An undirected graph (A) and corresponding adjacency matrix (B). The pairwise Markov properties are $Y_1 \perp Y_3 \mid \setminus \{Y_2, Y_4\}$**,** $Y_1 \perp Y_4 \mid \setminus \{Y_2, Y_3\}$ **and** $Y_2 \perp Y_3 \mid \setminus \{Y_1, Y_4\}$**.**

As inside information for a motivated reader, another way to infer a graphical model would be to examine linear dependencies between its variables. Then the edge set E can be inferred by examining the support of the correlation matrix $C = [c_{ij}]$, where c_{ij} is the Pearson correlation coefficient between variables Y_i and Y_j,

$$(i, j) \notin E \Leftrightarrow c_{ij} = 0. \tag{23}$$

As a curiosity, the graphical model (23) is widely used in high-dimensional gene co-expression network analysis. These correlation networks are often referred to as gene co-expression networks (see, e.g., the articles of Zhang & Horvath 2005, Langfelder & Horvath 2008 and related publications Plaisier et al. 2009, Voineagu et al. 2011, Session et al. 2016 to name a few). In this work we will concentrate on graphical models determined by the support of the sparse precision matrix and this model is utilized in more detail in Article IV.

3.2 Neighborhood selection

Although the GGM can be easily parametrized using the precision matrix Θ, the practical estimation of the support of the precision matrix becomes quickly infeasible. In the high-dimensional setting, the sample covariance matrix is singular. Even if the sample covariance matrix is non-singular, inverting it becomes quickly computationally impractical, albeit the dimension would increase just moderately. In addition, it is clear that all elements of the sample covariance matrix and its inverse are non-zero when the data are continuous. Thus one has to somehow select the graphical

model from the non-zero elements of the precision matrix. Overall, there can be $p(p-1)/2$ edges in a p-dimensional graph and each edge corresponds to a distinct upper/lower triangular element of the symmetric precision matrix. There can be $\sum_{k=0}^{p(p-1)/2} \binom{p(p-1)/2}{k} = 2^{p(p-1)/2}$ different graphical models to choose from. Clearly it would be computationally an enormous task to check all possible graphical models when the problem size p is large.

The most relevant theory considering our work is LASSO-based neighborhood selection for GGM estimation. This procedure was introduced in Meinshausen & Bühlmann (2006) and we will refer to it here as *MB-approximation* according to the initials of the original authors, Meinshausen and Bühlmann. The original work of Meinshausen & Bühlmann (2006) assumed that a p-dimensional random variable follows the multivariate normal distribution. However, the methodological background introduced in Meinshausen & Bühlmann (2006) is not limited to Gaussian models (see, e.g., Liu & Ihler 2011) and we have used it while examining discrete nominal data in Article II. The main motivation behind MB-approximation is that it is much less computationally demanding to estimate the neighborhood of each node one at a time than to compute an estimate for the full network. This approach transforms the graphical model selection problem into p lower-dimensional LASSO problems. Because this neighborhood selection method plays a big role in Article II we will introduce the theory presented in Meinshausen & Bühlmann (2006). See also (Hastie et al. 2015 pp. 254–258) for another description of the theory.

The neighborhood of node Y_i is the smallest subset $Ne(Y_i)$ of nodes $N \setminus \{Y_i\}$ so that variable Y_i is conditionally independent of all remaining variables. For example, the neighborhood of the node Y_2 in Figure 3 is the subset (subgraph) consisting of nodes Y_1 and Y_4 and the set of edges $\{(1,2),(2,4)\}$. This does not exclude the possibility that there could be an edge between nodes Y_1 and Y_4. For simplicity, we concentrate only on the set $Ne(Y_i)$ in neighborhood selection.

Building the MB-approximation begins with the next notice. Consider predicting random variable Y_i, given all the remaining variables,

$$\beta^{Y_i} = \underset{\beta:\, \beta_i = 0}{\operatorname{argmin}}\, \mathbb{E}\left\{ (Y_i - \sum_{k=1}^{p} \beta_k Y_k)^2 \right\}. \tag{24}$$

There is a connection between the elements of β^{Y_i} and the precision matrix. For $Y_j \in N \setminus \{Y_i\}$ and $\Theta = [\theta_{ij}]$ it holds that $\beta_j^{Y_i} = -\theta_{ij}/\theta_{ii}$. The set of non-zero coefficients of β^{Y_i} is identical to the set $E_i = \left\{ Y_j \in N \setminus \{Y_i\} \mid \theta_{ij} \neq 0 \right\}$. This corresponds to non-zero

entries of the ith column (or, interchangeably, the ith row) in the precision matrix Θ. Therefore, the nodes in set E_i are the neighbors of the node Y_i. Due to the connection between β^{Y_i} and θ_{ij}, the neighborhood of Y_i can be estimated by solving problem (24). Because of this connection, the neighborhood of node Y_i is determined as follows:

$$Ne(Y_i) = \left\{ Y_j \in N \setminus \{Y_i\} \mid \beta_j^{Y_i} \neq 0 \right\}. \tag{25}$$

Thus the neighborhood of Y_i and the overall support of the precision matrix can be estimated with a sparse vector of regression coefficients β^{Y_i} when $\mathbf{Y}_i$ is the $n \times 1$ outcome vector and all the remaining variables $Y \setminus \mathbf{Y}_i$ are used as predictors. The question remains of how to estimate β^{Y_i} so that some of the regression coefficients are estimated as zero. As we have discussed earlier, the LASSO estimator of a linear regression problem will shrink regression coefficients exactly to zero. Thus, LASSO can be used to estimate the non-zero coefficients of β^{Y_i}.

Let the LASSO estimator $\widehat{\beta}(\lambda)^{Y_i}$ of β^{Y_i} be the vector minimizing the optimization problem similar to (5):

$$\min_{\beta:\, \beta_i = 0} ||\mathbf{Y}_i - Y\beta||_2^2 + \lambda ||\beta||_1, \tag{26}$$

The neighborhood estimate of Y_i is defined by the non-zero coefficient estimates of $\widehat{\beta}(\lambda)^{Y_i}$,

$$\widehat{Ne}(Y_i)^{\lambda} = \left\{ Y_j \in N \setminus \{Y_i\} \mid \widehat{\beta}(\lambda)_j^{Y_i} \neq 0 \right\}, \tag{27}$$

with user-specified positive penalty parameter λ. After estimating the neighborhood of each node Y_i, $i = 1, \ldots, p$, one will have an estimate for the whole set of edges E and for the adjacency matrix A. The set E can be used to produce an estimate of the whole network $\widehat{G} = \{N, \widehat{E}\}$.

It is shown in the original paper of Meinshausen & Bühlmann (2006) that (27) is a consistent estimate of $Ne(Y_i)$ for the high-dimensional setting. Revising the consistency property is a somewhat laborious task and it is omitted here. We highlight that for the consistency of the neighborhood estimate to hold, the number of neighbors is assumed to be restricted compared to the sample size, $|Ne(Y_i)| = \mathcal{O}(n^{\kappa})$ for any $0 \leq \kappa < 1$ when $n \to \infty$ (see Meinshausen & Bühlmann 2006, Assumption 3 and other assumptions). For a more detailed description of the MB-approximation, see the original paper of Meinshausen & Bühlmann (2006) and Hastie et al. (2015). A review of a free software implementation of MB-approximation can be found in Bühlmann et al. (2014).

The MB-approximation is more like an estimate of the support of the precision matrix, not a numerical precision matrix estimate because it does not guarantee the positive definiteness of the estimate or provide reasonable numerical estimates for the entries in the precision matrix. MB-approximation can be seen as an approximation of the glasso algorithm. Furthermore, the initial neighborhood estimates might be inconsistent; if $\widehat{\beta}(\lambda)^{Y_i}$ indicates that Y_j is a neighbor of Y_i, $\widehat{\beta}(\lambda)^{Y_j}$ does not necessarily indicate that Y_i is a neighbor of Y_j. To solve this inconsistency, Meinshausen & Bühlmann (2006) proposed to make the final estimate symmetric using one of the following simple decision rules:

– AND-rule: The edge (i, j) between nodes Y_i and Y_j is in $\widehat{E}$ only when Y_i is selected as a neighbor of Y_j AND Y_j is selected as a neighbor of Y_i
– OR-rule: The edge (i, j) between nodes Y_i and Y_j is in $\widehat{E}$ when Y_i is selected as a neighbor of Y_j OR Y_j is selected as a neighbor of Y_i.

We have used the AND-rule in Article II.

3.2.1 Some extensions of the MB-approximation

MB-approximation has proven to be a convenient estimator due to its computational efficiency. A similar approach has been utilized by many GGM selection (sparse precision matrix estimation) methods, such as the graphical Dantzig selector (Yuan 2010), CLIME (Cai et al. 2011), scaled lasso (Sun & Zhang 2013), SCIO (Liu & Luo 2015), ACLIME (Cai et al. 2016), TIGER (Liu & Wang 2017), and CONE (Community Oriented Network Estimation framework) presented in Article II. In Article II we utilize the fast algorithms developed by Friedman et al. (2010) to solve multinomial regression LASSO problems.

In principle, MB-approximation is not limited to LASSO and any other sparsity-inducing regression method could be used to perform neighborhood selection. For example, MB-approximation could be performed using Bayesian LASSO (Park & Casella 2008). In Article II we have utilized the MB-approximation framework in population structure estimation for categorical data and named this framework CONE. We note that CONE is solely a network estimation method and here it differs from the other methods mentioned in the previous chapter. We note that the theoretical limit of the number of neighbors $|Ne(Y_i)|$ sets strict conditions on the consistency of our CONE framework. Nevertheless, we argue that this does not have a serious effect on CONE

because the number of individuals in each population is presumably much smaller than the total sample size.

As mentioned in the *Choosing the tuning parameter* subsection one still has to choose a suitable positive value for the tuning parameter λ when using LASSO regression. With the term "suitable" we refer to a value which produces an estimate agreeing with the assumptions of the underlying (sparse) graphical model. For example, excessively sparse networks are not biologically interesting and can be uninformative in the population estimation, which we have considered in Article II. In the next subsection, we will discuss the methods for selecting the tuning parameter as used in our work considering network inference and sparse precision matrix estimation.

3.2.2 Choosing the tuning parameter for graphical model estimators

We have mainly concentrated on data-dependent tuning parameter selection methods which would produce consistent estimates for sparse networks also in the high-dimensional setting. We have mainly used the following tuning parameter selection methods:

- Cross-validation schemes portrayed by Liu & Luo (2015)
- Extended Bayesian information criterion (eBIC) (Chen & Chen 2008, Foygel & Drton 2010)
- Stability approach to regularization selection (StARS) (Liu et al. 2010)

Of the above-mentioned methods, the StARS procedure plays an important role in Article II. Therefore, we will describe it in more detail later in its own subsection.

Both eBIC and StARS are developed utilizing the good properties of the glasso algorithm where the overall sparsity of the estimated network can be controlled with a single parameter, that is, the tuning parameter λ. In contrast, cross-validation does not take into account the possible sparsity in the structure of the precision matrix and it has been empirically shown to select extremely dense graphical models (Liu et al. 2010). On the other hand, in Article IV the cross-validation scheme of Liu & Luo (2015) seems to select extremely sparse graphical models.

The extended Bayesian information criterion (eBIC) is a modification of the more fundamental Bayesian information criterion (BIC). For comparison purposes, we have used the following formulation of the eBIC to be minimized over different values of λ,

$$eBIC_\gamma = -n \log p(Y|\widehat{\Theta}) + df(\widehat{\Theta}) \log n + 4 df(\widehat{\Theta}) \gamma \log p, \quad (28)$$

where $\log p(Y|\widehat{\Theta})$ is the log-likelihood (7), $\widehat{\Theta}$ is a sparse estimate of the precision matrix, $df(\widehat{\Theta})$ is the number of non-zero elements in $\widehat{\Theta}$, and γ is a user-specified parameter. Positive values of γ lead to more sparse estimates for the precision matrix. The original paper of Foygel & Drton (2010) proposed setting γ to a value of 0.5. Note that formulation (28) is slightly different from the one in Foygel & Drton (2010).

As an important side note we emphasize that it is possible to define more specific regularizations for each node of the network instead of controlling the general sparsity level of the precision matrix or the statistical network with just one parameter (the tuning parameter λ) (Friedman et al. 2008, Bien & Tibshirani 2011, Hsieh et al. 2014, Li & Jackson 2015, Liu & Luo 2015). However, it can be computationally very demanding to objectively determine several different regularization parameters without solid *a priori* information about the network of interest. For the present, it is more popular to use just one parameter to control for sparsity in network inference.

In addition to cross-validation, eBIC and StARS, there are several research articles dedicated to the tuning parameter selection (see, e.g., Liu & Ihler 2011, Lederer & Müller 2014, Tandon & Ravikumar 2014). In these articles, the special structure of the network of interest (mainly the so-called scale-free network) is examined. Special characteristics of the underlying network model, such as the node degree (the number of neighbors of a specific node), are taken into account while trying to choose an optimal value for the tuning parameter.

In Article II we have used the so-called "elbow method" to select the optimal value for the tuning parameter to compute more comprehensive population structure estimates. In the elbow method, one plots an error measure or error measures against the number of estimated clusters identified in the data. Usually the error measure decreases when the number of inferred clusters increases and usually there is a remarkable drop in the plot. After this drop, the error measure usually decreases more moderately. This change point is called an "elbow" and it indicates the number of inferred clusters which sufficiently describe the data (Tibshirani et al. 2001). We found the elbow method to have a nice interpretation when using the network modularity as an error measure, although it remains a somewhat nebulous concept.

3.2.3 Network community

Even if one had an exact graphical model describing every single edge between the nodes, the model itself would not include any easily interpretable information about the lower level structure of the network. Groupings or separating different nodes or plotting a visually pleasing and informative graphical representation of a high-dimensional network is not an easy task. In Article II we have examined how the nodes of a network can be divided into *communities* and utilized this community division in population structure analysis.

There is no mathematically exact definition for a network community. In Fortunato (2010) the problem of defining a network community is described as follows: "No definition is universally accepted. As a matter of fact, the definition often depends on the specific system at hand and/or application one has in mind . . . we get the notion that there must be more edges 'inside' the community than edges linking vertices of the community with the rest of the graph. This is the reference guideline at the basis of most community definitions. But many alternative recipes are compatible with it. Moreover, in most cases, communities are algorithmically defined, i.e. they are just the final product of the algorithm, without a precise *a priori* definition." (pp. 83–84). See Figure 4 for an illustration of a potential community structure of a very small network.

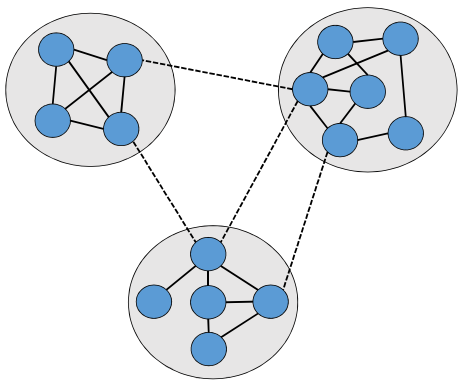


Fig. 4. A schematic illustration of a network with three communities (shaded areas). Nodes within communities share more edges (solid lines) between each other than edges from different communities (dashed lines).

In Article II we have used ready-made algorithms available for the R programming language. In particular, we have utilized tools available in the `igraph` package (Csardi & Nepusz 2006). We have mainly used the walktrap algorithm (Pons & Latapy 2006) to distinct, different communities and draw a parallel between the estimated communities and the populations from which individuals originated. In addition to the walktrap algorithm, we have also used the Fruchterman-Reingold algorithm (Fruchterman & Reingold 1991) to visualize estimated networks. Overall, there are a vast and increasing number of community detection algorithms available but examining them in more detail is beyond of the scope of this dissertation. We refer the interested reader to Fortunato (2010) and Fortunato & Hric (2016) for more information of network communities and a review of community detection algorithms.

Stability approach to regularization selection

The stability approach to regularization selection (StARS) introduced by Liu et al. (2010) is related to stability selection developed for linear model selection (Meinshausen & Bühlmann 2010). StARS uses recurrent subsampling where each sample is drawn without replacement and a network is inferred from each subsample one at a time. StARS can be used to select the optimal value of λ for a network model (sparse precision matrix) based on the matrix support instability. This is done by actually measuring the instability of the estimated graph. First, one estimates the dispersion relative to the variance of the Bernoulli indicator of a distinct edge from all inferred networks. Then one estimates the instability associated with a value of the tuning parameter. Finally, one monitors the network instability over each value of the tuning parameter λ and chooses the optimal value of λ as a trade-off between network sparsity and instability.

The main motivation behind StARS is the selection of a sparse precision matrix support which still contains the true precision matrix support with high probability. The following theorem presented by Liu et al. (2010) highlights this property:

Theorem 1 *Let $\widehat{\Lambda} = 1/\widehat{\lambda}$ be the selected regularization parameter using the StARS procedure, $b = \lfloor 10\sqrt{n} \rfloor$ is the subsample size and $\widehat{E}^b(\widehat{\Lambda})$ the set of edges selected with StARS. Then, if $p \leq \exp(n^\gamma)$ for any $\gamma < 1/2$, we have*

$$\mathbb{P}(E \subset \widehat{E}^b(\widehat{\Lambda})) \to 1 \quad as \quad n \to \infty, \tag{29}$$

under some specific assumptions.

For a more detailed description of Theorem 1, see the original paper of Liu et al. (2010). Theorem 1 guarantees that the tuning parameter chosen by StARS will select the true graph with high probability in the high-dimensional setting by adjusting just one parameter.

StARS is also more suitable for sparse precision matrix estimation than cross-validation which tends to favor dense precision matrix estimates. StARS converges to the true support of the precision matrix in terms of common asymptotics. The largest drawback of StARS is that it is more computationally intensive than cross-validation. StARS also includes a user-specified parameter β which is a cut-point value for the optimal network instability of the estimated graph. Nevertheless, β is a more easily interpretable quantity than λ and it can be set to a default value (Liu et al. 2010). Due to its good theoretical properties such as convergence under high-dimension asymptotics, adaptability to any precision matrix or statistical network estimation method, and interpetability, we have used StARS to construct population graphs in Article II.

3.3 Graph selection via multiple testing

Although the sample covariance matrix S is a consistent estimator of Σ, distinct off-diagonal elements of S^{-1} are never remotely close to zero in practice, even though the corresponding "ground truth" elements of $\Theta = [\theta_{ij}]$ would be zero. In the *Penalized covariance and precision matrix estimation utilizing LASSO penalty*i section, we described methods which utilize a regularization approach to produce sparse precision matrix estimates to be used in GGM estimation. Alternatively, one could select the GGM by simultaneously testing hypothesis $H_0 : \theta_{ij} = 0$ against hypothesis $H_A : \theta_{ij} \neq 0$ for each non-diagonal entry of the precision matrix $\Theta = [\theta_{ij}]$, $i, j = 1, \ldots, p(p-1)/2$, $i \neq j$. There is an edge between nodes Y_i and Y_j only if the null-hypothesis H_0 is rejected and we call this rejection a *significant* signal for short.

As mentioned in the *Introduction to statistical network inference* section, the conditional independence (20) between variables can be examined by the elements of the partial correlation matrix or the precision matrix. The partial correlation matrix $R = [r_{ij}]$ and the precision matrix have the following relation,

$$r_{ij} = -\frac{\theta_{ij}}{\sqrt{\theta_{ii}\theta_{jj}}}. \tag{30}$$

Equation (30) can be expressed as a product of the form

$$R = -\text{diag}(\Theta)^{-1/2}\,\Theta\,\text{diag}(\Theta)^{-1/2}, \tag{31}$$

where $\text{diag}(\Theta)$ is a diagonal matrix constructed by the diagonal elements of Θ (Ha & Sun 2014). The square root of the diagonal matrix, $\text{diag}(\Theta)^{1/2}$, is simply the square root of the diagonal elements of the precision matrix. Because of the matrix and diagonal matrix products in equation (31), the partial correlation matrix can be determined very quickly in just $\mathcal{O}(p^2)$ flops.

While trying to construct a GGM by simultaneous multiple hypothesis testing, we have examined partial correlation coefficients rather than elements of the precision matrix: $H_0 : r_{ij} = 0$ against $H_A : r_{ij} \neq 0$, $i, j = 1, \ldots, p(p-1)/2$, $i \neq j$.

One possible way to test H_0 against H_A would be to use Fisher's z-transformation as a test statistic: $0.5\log\{(1+\widehat{r}_{ij})/(1-\widehat{r}_{ij})\}$, where $\widehat{R} = [\widehat{r}_{ij}]$ is the sample partial correlation estimate. Because z is approximately normally distributed under the null hypothesis, one can test whether the sample partial correlation coefficients are in line with the null hypothesis for a pre-specified level of significance α (see, e.g., Drton & Perlman 2004 or Ha & Sun 2014). Fisher's z-transformation and the distribution of partial correlation coefficients have been utilized extensively (Schäfer & Strimmer 2005, Drton & Perlman 2007, Ha & Sun 2014, van Wieringen & Peeters 2015, van Wieringen & Peeters 2016).

In article IV, we use edge exclusion deviance (EED) as our test statistic. EED is defined as

$$EED_{ij} = -n\log(1-\widehat{r}_{ij}^2). \tag{32}$$

The asymptotic distribution of EED is the chi-squared distribution with one degree of freedom (see pp. 189–190 in Whittaker 1990).

In Drton & Perlman (2004, 2007) a multiple testing procedure was proposed but the authors did not consider a high-dimensional setting in their studies. Because the normal and chi-squared distributions are asymptotic distributions of z and EED, respectively, the asymptotics hold only for $n \gg p$. In addition, the sample partial correlation matrix $\widehat{R}$ cannot be computed in high dimensions.

Even though one cannot compute a sample estimate for the partial correlation matrix, one can always compute a penalized estimate using some estimator of the form (11). In Article IV we used ROPE to estimate the precision matrix and then derived the partial correlation matrix previously defined in Equation (31) in high dimensions. After

determining a penalized estimate (ROPE) for the partial correlation matrix R EED is used as the test statistic to carry out tests of a set of hypotheses simultaneously to determine a GGM which is in harmony with the sample.

However, for penalized estimates the asymptotics do not hold for penalized estimates. Due to the vast number of observed EED values, rather than saying that each test statistic EED_{ij} comes from the theoretical null distribution, the null distribution can be estimated empirically. Because of the empirical estimation approach, the observed EED values come from two different classes: "significant" or "nonsignificant". In Efron (2004) the author proposed that the observed test statistic values arise from a mixture distribution. Following this approach, in Article IV we have assumed that the distribution of the observed EED values across edges is a mixture density

$$f(EED) = \omega f_0(EED) + (1-\omega) f_A(EED), \tag{33}$$

where f_0 is the null distribution, ω is the (unknown) proportion of null edges in the graphical model, $\omega \in [0,1]$ and f_A is the alternative distribution of the observed EED values assigned to actually existing edges of the graphical model. The mixture density (33) can be estimated by a fit to the histogram counts. Rather than the quantile determined from the theoretical null distribution, one can use the quantile of the estimated mixture distribution to determine a desired significance level.

The good property of the multiple testing procedure is that "only" $(p^2-p)/2$ different hypotheses need to be checked. These hypotheses correspond to the upper (or lower) triangular of the sample partial correlation matrix. This is in practice a collection of simple testing procedures: "*if* $EED_{ij} < q_\alpha$, *then set* $\widehat{r}_{ij} = 0$" where q_α is the quantile of our empirical null distribution for the desired level of significance. Thus determining the support of the partial correlation matrix with multiple hypothesis testing is not a computer intensive way to select the GGM.

3.3.1 False discovery rate control

When testing multiple hypotheses simultaneously, the proportion of false positive signals (those appearing to be significant even though the null hypothesis is true) increases with the number of hypotheses tested, even when the significance level is small and the sample size is large. Thus it is reasonable to control the (proportion) of false positive signals.

A popular error measure for false positive signal examination has been the false discovery rate (FDR) proposed by Benjamini & Hochberg (1995) (see also Benjamini & Yekutieli 2001). FDR is an error measure defined as the expected proportion of falsely positive signals among all signals which are determined to be significant, that is, $FDR = \mathbb{E}(FP/M)$, where FP is the number of false positive tests and M is the number of all tests which are significant. Clearly, FDR is zero when there are no false positive signals among all significant signals. By convention, FDR is also zero when no hypotheses are rejected.

The original Benjamini-Hochberg (BH) method for FDR control assumes that the hypotheses are independent. However, the independence of hypotheses cannot always be guaranteed in high-dimensional data. For example, high correlation can have an effect on the number of false positive signals (Owen 2005, Schwartzman & Lin 2011). In addition, correlated variables can increase the variance of the FDR causing biased FDR estimators (Qiu & Yakovlev 2006, Schwartzman & Lin 2011). In Article IV we almost without exception noticed a huge difference in the sparsity level of network estimates computed with and without FDR control.

FDR control has been very popular in the "classical" statistics sense where the number of false positive signals is minimized. However, in the next section we discuss why too rigid a control of false positive signals may backfire in network estimation.

3.4 Graph under- and overselection

Before delving more deeply into the subject of this subsection, we first clarify the concept of node degree (connectivity). The node degree is the number of edges incident to the node, which is equal to the number of neighbors of the node. For example, the node degrees of nodes $\{Y_1, Y_2, Y_3, Y_4\}$ of the graph represented in Figure 3 are $\{1, 2, 1, 2\}$, respectively. The node degree can be easily determined as the column or row sum of the network adjacency matrix.

In Article IV we have focused on the graph underselection (the estimated graph is overly sparse) and graph overselection (the estimated graph is overly dense) problems. Following "classical" statistical thinking, the graph overselection is usually controlled; model (graph) overselection can be restrained in multiple hypothesis testing with the FDR control (see, e.g., Liu 2013, Xia et al. 2015). FDR control is intuitively a reasonable procedure because overly dense networks are not reasonable in many real life applications. For example in gene co-expression analysis, networks where there are

more high-degree nodes than there are low-degree nodes are not biologically meaningful networks (see, e.g., Carlson et al. 2006).

However, controlling the false discovery rate may cause nodes in the estimated graph to have very low degrees. In practice this means that network clusters or neighborhoods of (interesting) nodes are drastically underselected. This underselection is problematic because in many biological networks there are usually a few hub nodes. These hub nodes and their neighbors can possess biologically interesting information (Zhang & Horvath 2005, Carlson et al. 2006, Langfelder & Horvath 2008, Langfelder et al. 2013). If the graph is underselected, most of the neighbors of the hub nodes are undetected. Figure 5 illustrates graph under- and overselection.

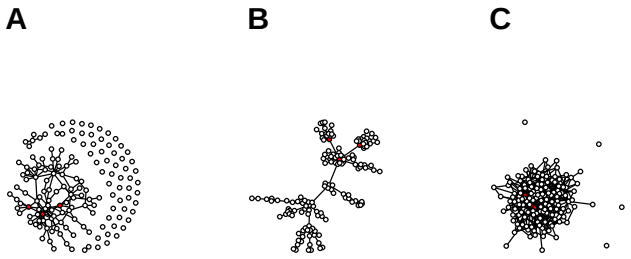


Fig. 5. (A) underselected graph (B) the true scale-free structured graph (C) overselected graph. The three nodes with the highest degrees are $\{14, 17, 21\}$, which correspond to the hub nodes of the true graph and are colored in red in each graph. The estimated node degrees of the hub nodes corresponding to the true hub nodes are $\{8, 7, 18\}$ for the underselected graph and $\{15, 15, 26\}$ for the overselected graph, respectively.

We have tried to minimize graph underselection in Article IV because we find graph underselection to be a more severe problem than moderate graph overselection; in a careful and thorough investigation it is easier to experimentally test which edges in the estimated graph represent false positive signals (this is a small or moderate size problem with a predetermined problem space) than to try to experimentally find correct positive signals (this is a huge problem with an undefined problem space). In Article IV we use simulated and real data to demonstrate that disregarding the FDR control may actually produce more meaningful network estimates while trying to find network hubs and

clusters. As a side note, in Article II graph underselection has to be controlled at some level, although this is not directly declared in Article II.

Graph underselection and graph overselection are not well studied or defined research problems in the statistical literature, considering penalized precision matrix estimation or partial correlation estimation. Either they have been mentioned indirectly besides the StARS procedure (Liu et al. 2010) or remarked upon briefly (Zhao et al. 2012). Nevertheless, we state that graph overselection and underselection with Gaussian graphical model selection methods should be investigated in more detail.

4 Conclusion

In this dissertation, we have introduced two novel methods, ROPE and CONE. We have proposed ROPE for penalized precision matrix estimation. We have applied statistical neighborhood selection methods with community detection (CONE) in population structure analysis of genetic data. We have also reviewed a collection of covariance and precision matrix estimation methods, which are publicly available on CRAN (Comprehensive R Archive Network) or MATLAB.

Estimation of the covariance and precision matrix in high dimensions is still an open and active are of research in the statistical literature. In addition to the graphical model selection problem, we have discussed graph under- and overselection, which is an unexplored field in statistical network estimation.

4.1 Future work

There are plenty of opportunities for future studies. One possibility would be to apply ROPE, introduced in Article I, to jointly estimate multiple precision matrices from different classes. Joint estimation has already been investigated by other researchers (see, e.g., Danaher et al. 2014, Bilgrau et al. 2015) and there are computer software implementations available for the R language related to these research articles.

We have also applied the network estimation framework introduced in Article II in genetic stock identification of sparrow populations which have been previously studied by Jensen et al. (2013). In the ongoing study we have used CONE to genetically identify the origin of individuals (sparrow fledgling) sampled in a mixture of individuals from genetically distinct populations.

It would be interesting to use methods such as glasso to identify specific network structures mentioned in Article III. One of these specific network structures are scale-free networks (see, e.g., Khanin & Wit 2006). We are interested in scale-free networks because they are characterized by a few nodes with large degrees. These highly connected hub nodes can possess valuable biological information about biological processes (see, e.g., Barabasi & Oltvai 2004). Although the prevailing interpretation of the importance of hub nodes has been criticized (He & Zhang 2006, Langfelder et al. 2013), it is indisputable that hub nodes in general play an essential part of network structures.

From a theoretical point of view it would be interesting to investigate how network under- and overselection should be controlled in network selection. For now the theoretical justification of the graph underselection is fairly limited in Article IV.

References

Banerjee, O., Ghaoui, L. E., & d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research*, *9*, 485–516.

Barabasi, A.-L., & Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, *5*, 101–113. http://dx.doi.org/10.1038/nrg1272

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *57*(1), 289–300. Retrieved from http://www.jstor.org/stable/2346101

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, *29*(4), 1165–1188. https://doi.org/10.1214/aos/1013699998

Bien, J., & Tibshirani, R. J. (2011). Sparse estimation of a covariance matrix. *Biometrika*, *98*(4), 807–820. http://doi.org/10.1093/biomet/asr054

Bilgrau, A. E., Peeters, C. F., Eriksen, P. S., Bøgsted, M., & van Wieringen, W. N. (2015). Targeted fused ridge estimation of inverse covariance matrices from multiple high-dimensional data classes. *arXiv preprint arXiv:1509.07982*.

Bühlmann, P., Kalisch, M., & Meinshausen, N. (2014). High-dimensional statistics with a view toward application in biology. *Annual Review of Statistics and Its Application*, *1*, 255–278. https://doi.org/10.1146/annurev-statistics-022513-115545

Bühlmann, P., & Van De Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. London, UK: Springer Science & Business Media.

Cai, T., Liu, W., & Luo, X. (2011). A constrained l_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, *106*(494), 594–607. https://doi.org/10.1198/jasa.2011.tm10155

Cai, T. T., Liu, W., & Zhou, H. H. (2016). Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation. *Annals of Statistics*, *44*(2), 455–488. https://doi.org/10.1214/13-AOS1171

Carlson, M. R., Zhang, B., Fang, Z., Mischel, P. S., Horvath, S., & Nelson, S. F. (2006). Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. *BMC Genomics*, *7*, 40. http://doi.org/10.1186/1471-2164-7-40

Chen, J., & Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, *95*(3), 759–771. https://doi.org/10.1093/biomet/asn034

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*, *Complex Systems*, 1695.

Danaher, P., Wang, P., & Witten, D. M. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*(2), 373–397. https://doi.org/10.1111/rssb.12033

Dempster, A. P. (1972). Covariance selection. *Biometrics*, *28*(1), 157–175. doi:10.2307/2528966

Drton, M., & Perlman, M. D. (2004). Model selection for Gaussian concentration graphs. *Biometrika*, *91*(3), 591–602. https://doi.org/10.1093/biomet/91.3.591

Drton, M., & Perlman, M. D. (2007). Multiple testing and error control in Gaussian graphical model selection. *Statistical Science*, *22*(3), 430–449. https://doi.org/10.1214/088342307000000113

Edwards, D. (2000). *Introduction to Graphical Modelling*. New York, USA: Springer, 2nd ed.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null distribution. *Journal of the American Statistical Association*, *99*(465), 96–104. https://doi.org/10.1198/016214504000000089

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, *486*(3), 75–174. https://doi.org/10.1016/j.physrep.2009.11.002

Fortunato, S., & Hric, D. (2016). Community detection in networks: A user guide. *Physics Reports*, *659*, 1–44. https://doi.org/10.1016/j.physrep.2016.09.002

Foygel, R., & Drton, M. (2010). Extended Bayesian information criteria for Gaussian graphical models. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, & A. Culotta (Eds.) *Advances in Neural Information Processing Systems 23*, (pp. 604–612). USA: Curran Associates, Inc.

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

https://doi.org/10.1093/biostatistics/kxm045

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22. http://dx.doi.org/10.18637/jss.v033.i01

Fruchterman, T. M., & Reingold, E. M. (1991). Graph drawing with force-directed placement. *Software: Practice and Experience*, *21*(11), 1129–1164. https://doi.org/10.1002/spe.4380211102

Ha, M. J., & Sun, W. (2014). Partial correlation matrix estimation using ridge penalty followed by thresholding and re-estimation. *Biometrics*, *70*(3), 765–773. https://doi.org/10.1111/biom.12186

Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The Elements of Statistical Learning*. New York, USA: Springer series in statistics, Corrected 12th printing - Jan 13, 2017, Second ed.

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Boca Raton, USA: CRC press.

He, X., & Zhang, J. (2006). Why do hubs tend to be essential in protein networks? *PLOS Genetics*, *2*(6), e88. https://doi.org/10.1371/journal.pgen.0020088

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67. doi:10.2307/1267351

Honorio, J., & Jaakkola, T. S. (2013). Inverse covariance estimation for high-dimensional data in linear time and space: Spectral methods for Riccati and sparse models. In A. Nicholson, & P. Smyth (Eds.) *Proceedings of the Twenty-Ninth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-13)*, (pp. 291–300). USA: AUAI Press.

Hsieh, C.-J., Sustik, M. A., Dhillon, I. S., & Ravikumar, P. (2014). QUIC: Quadratic approximation for sparse inverse covariance estimation. *Journal of Machine Learning Research*, *15*, 2911–2947.

Hsieh, C.-J., Sustik, M. A., Dhillon, I. S., Ravikumar, P. K., & Poldrack, R. (2013). BIG & QUIC: Sparse inverse covariance estimation for a million variables. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.) *Advances in Neural Information Processing Systems 26*, (pp. 3165–3173). USA: Curran Associates, Inc.

Huang, F. (2003). Prediction error property of the lasso estimator and its generalization. *Australian & New Zealand Journal of Statistics*, *45*(2), 217–228. https://doi.org/10.1111/1467-842X.00277

Jensen, H., Moe, R., Hagen, I. J., Holand, A. M., Kekkonen, J., Tufto, J., & Sæther, B.-E. (2013). Genetic variation and structure of house sparrow populations: is there an island effect? *Molecular Ecology*, *22*(7), 1792–1805. https://doi.org/10.1111/mec.12226

Khanin, R., & Wit, E. (2006). How scale-free are biological networks. *Journal of Computational Biology*, *13*(3), 810–818. https://doi.org/10.1089/cmb.2006.13.810

Khondker, Z. S., Zhu, H., Chu, H., Lin, W., & Ibrahim, J. G. (2013). The Bayesian covariance lasso. *Statistics and its Interface*, *6*(2), 243–259.

Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, *9*, 559. https://doi.org/10.1186/1471-2105-9-559

Langfelder, P., Mischel, P. S., & Horvath, S. (2013). When is hub gene selection better than standard meta-analysis? *PLOS ONE*, *8*(4), e61505. https://doi.org/10.1371/journal.pone.0061505

Lederer, J., & Müller, C. (2014). Topology adaptive graph estimation in high dimensions. *arXiv preprint arXiv:1410.7279*.

Ledoit, O., & Wolf, M. (2004a). Honey, I shrunk the sample covariance matrix. *Journal of Portfolio Management*, *30*(4), 110–119. https://doi.org/10.3905/jpm.2004.110

Ledoit, O., & Wolf, M. (2004b). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, *88*(2), 365–411. https://doi.org/10.1016/S0047-259X(03)00096-4

Ledoit, O., & Wolf, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *Annals of Statistics*, *40*(2), 1024–1060. https://doi.org/10.1214/12-AOS989

Ledoit, O., & Wolf, M. (2015). Spectrum estimation: A unified framework for covariance matrix estimation and PCA in large dimensions. *Journal of Multivariate Analysis*, *139*, 360–384. https://doi.org/10.1016/j.jmva.2015.04.006

Li, Y., & Jackson, S. A. (2015). Gene network reconstruction by integration of prior biological knowledge. *G3 (Bethesda)*, *5*(6), 1075–1079. https://doi.org/10.1534/g3.115.018127

Liu, H., Roeder, K., & Wasserman, L. (2010). Stability approach to regularization selection (StARS) for high dimensional graphical models. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, & A. Culotta (Eds.) *Advances in Neural Information Processing Systems 23*, (pp. 1432–1440). USA: Curran Associates, Inc.

Liu, H., & Wang, L. (2017). TIGER: A tuning-insensitive approach for optimally estimating Gaussian graphical models. *Electronic Journal of Statistics*, *11*(1), 241–294. https://doi.org/10.1214/16-EJS1195

Liu, Q., & Ihler, A. (2011). Learning scale free networks by reweighted l_1 regularization. In G. Gordon, D. Dunson, & M. Dudík (Eds.) *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, vol. 15 of *Proceedings of Machine Learning Research*, (pp. 40–48). USA: PMLR.

Liu, W. (2013). Gaussian graphical model estimation with false discovery rate control. *Annals of Statistics*, *41*(6), 2948–2978. https://doi.org/10.1214/13-AOS1169

Liu, W., & Luo, X. (2015). Fast and adaptive sparse precision matrix estimation in high dimensions. *Journal of Multivariate Analysis*, *135*, 153–162. https://doi.org/10.1016/j.jmva.2014.11.005

Meinshausen, N., & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the LASSO. *Annals of Statistics*, *34*(3), 1436–1462. https://doi.org/10.1214/009053606000000281

Meinshausen, N., & Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *72*(4), 417–473. https://doi.org/10.1111/j.1467-9868.2010.00740.x

Owen, A. B. (2005). Variance of the number of false discoveries. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(3), 411–426. https://doi.org/10.1111/j.1467-9868.2005.00509.x

Park, T., & Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, *103*(482), 681–686. https://doi.org/10.1198/016214508000000337

Picard, R. R., & Cook, R. D. (1984). Cross-validation of regression models. *Journal of the American Statistical Association*, *79*(387), 575–583. doi:10.2307/2288403

Plaisier, C. L., Horvath, S., Huertas-Vazquez, A., Cruz-Bautista, I., Herrera, M. F., Tusie-Luna, T., Aguilar-Salinas, C., & Pajukanta, P. (2009). A systems genetics approach implicates USF1, FADS3, and other causal candidate genes for familial combined hyperlipidemia. *PLOS Genetics*, *5*(9), e1000642. https://doi.org/10.1371/journal.pgen.1000642

Pons, P., & Latapy, M. (2006). Computing communities in large networks using random walks. *Journal of Graph Algorithms and Applications*, *10*(2), 191–218. http://dx.doi.org/10.7155/jgaa.00124

Qiu, X., & Yakovlev, A. (2006). Some comments on instability of false discovery rate estimation. *Journal of Bioinformatics and Computational Biology*, *4*(5), 1057–1068.

https://doi.org/10.1142/S0219720006002338

Rosset, S., & Zhu, J. (2004). Corrected proof of the result of 'A prediction error property of the Lasso estimator and its generalization' by Huang (2003). *Australian & New Zealand Journal of Statistics*, *46*(3), 505–510. https://doi.org/10.1111/j.1467-842X.2004.00347.x

Schäfer, J., & Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, *4*(1), –. https://doi.org/10.2202/1544-6115.1175

Schwartzman, A., & Lin, X. (2011). The effect of correlation in false discovery rate estimation. *Biometrika*, *98*(1), 199–214. http://doi.org/10.1093/biomet/asq075

Session, A. M., Uno, Y., Kwon, T., Chapman, J. A., Toyoda, A., Takahashi, S., Fukui, A., Hikosaka, A., Suzuki, A., Kondo, M., et al. (2016). Genome evolution in the allotetraploid frog *Xenopus laevis*. *Nature*, *538*(7625), 336–343. http://doi.org/10.1038/nature19840

Sun, T., & Zhang, C.-H. (2013). Sparse matrix inversion with scaled lasso. *Journal of Machine Learning Research*, *14*, 3385–3418.

Tandon, R., & Ravikumar, P. (2014). Learning graphs with a few hubs. In E. P. Xing, & T. Jebara (Eds.) *Proceedings of the 31st International Conference on Machine Learning*, vol. 32 of *Proceedings of Machine Learning Research*, (pp. 602–610). USA: PMLR.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *58*(1), 267–288. Retrieved from http://www.jstor.org/stable/2346178

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(2), 411–423. https://doi.org/10.1111/1467-9868.00293

van Wieringen, W., & Peeters, C. (2016). Ridge estimation of inverse covariance matrices from high-dimensional data. *Computational Statistics and Data Analysis*, *103*, 284 – 303. https://doi.org/10.1016/j.csda.2016.05.012

van Wieringen, W. N. (2017). On the mean squared error of the ridge estimator of the covariance and precision matrix. *Statistics & Probability Letters*, *123*, 88–92. https://doi.org/10.1016/j.spl.2016.12.002

van Wieringen, W. N., & Peeters, C. F. W. (2015). Application of a new ridge estimator of the inverse covariance matrix to the reconstruction of gene-gene interaction networks. In C. DI Serio, P. Liò, A. Nonis, & R. Tagliaferri (Eds.) *Computational*

Intelligence Methods for Bioinformatics and Biostatistics, vol. 8623 of *Lecture Notes in Computer Science*, (pp. 170–179). Germany: Springer International Publishing.

Voineagu, I., Wang, X., Johnston, P., Lowe, J. K., Tian, Y., Horvath, S., Mill, J., Cantor, R. M., Blencowe, B. J., & Geschwind, D. H. (2011). Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature*, *474*(7351), 380–384. http://doi.org/10.1038/nature10110

Wang, H. (2012). Bayesian graphical lasso models and efficient posterior computation. *Bayesian Analysis*, *7*(4), 867–886. doi:10.1214/12-BA729

Warton, D. I. (2008). Penalized normal likelihood and ridge regularization of correlation and covariance matrices. *Journal of the American Statistical Association*, *103*(481), 340–349. https://doi.org/10.1198/016214508000000021

Whittaker, J. (1990). *Graphical Models*. West Sussex, England: John Wiley & Sons.

Witten, D. M., Friedman, J. H., & Simon, N. (2011). New insights and faster computations for the graphical lasso. *Journal of Computational and Graphical Statistics*, *20*(4), 892–900. https://doi.org/10.1198/jcgs.2011.11051a

Xia, Y., Cai, T., & Cai, T. T. (2015). Testing differential networks with applications to the detection of gene-gene interactions. *Biometrika*, *102*(2), 247–266. http://doi.org/10.1093/biomet/asu074

Yuan, M. (2010). High-dimensional inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research*, *11*, 2261–2286.

Zhang, B., & Horvath, S. (2005). A general framework for weighted gene coexpression network analysis. *Statistical Applications in Genetics and Molecular Biology*, *4*(1), –. https://doi.org/10.2202/1544-6115.1128

Zhao, T., Liu, H., Roeder, K., Lafferty, J., & Wasserman, L. (2012). The huge package for high-dimensional undirected graph estimation in R. *Journal of Machine Learning Research*, *13*, 1059–1062.

Original publications

I Kuismin M., Kemppainen J. & Sillanpää M. J. (2017) Precision matrix estimation with ROPE. Journal of Computational and Graphical Statistics 26: 682–694.

II Kuismin M., Ahlinder J. & Sillanpää, M. J. (2017) CONE: Community oriented network estimation is a versatile framework for inferring population structure in large-scale sequencing data. G3 (Bethesda) 7: 3359–3377.

III Kuismin M. & Sillanpää M. J. (2017) Estimation of covariance and precision matrix, network structure, and a view toward systems biology. Wiley Interdisciplinary Reviews: Computational Statistics: 9:e1415.

IV Kuismin M. & Sillanpää M. J. Keep it simple: Parametric network inference without FDR control improves network structure analysis. Manuscript.

Original publications are not included in the electronic version of the dissertation.

Book orders:
Granum: Virtual book store
http://granum.uta.fi/granum/

UNIVERSITY OF OULU P.O. Box 8000 FI-90014 UNIVERSITY OF OULU FINLAND

ACTA UNIVERSITATIS OULUENSIS

UNIVERSITY of OULU
OULUN YLIOPISTO

ISBN 978-952-62-2079-6 (Paperback)
ISBN 978-952-62-2080-2 (PDF)
ISSN 0355-3191 (Print)
ISSN 1796-220X (Online)