Elliptical graphical modelling

Dissertation

zur Erlangung des Grades „Doktor der Naturwissenschaften“
an der Fakultät Statistik der Technischen Universität Dortmund
vorgelegt von

Daniel Vogel

Betreuer und Gutachter: Prof. Dr. R. Fried
Gutachter: Prof. Dr. Ch. H. Müller
Kommissionsvorsitz: PD Dr. S. Kuhnt
Abgabe der Dissertation: 25. 10. 2010
Tag der mündlichen Prüfung: 15. 12. 2010

Contents

Acknowledgement

I owe my gratitude to a lot of people. Being aware of the inherent complicacy of any personal acknowledgement, the unavoidable risk of forgetting someone, I was inclined not to mention anyone at all. On second thought, this would not be just either, and I apologize beforehand to everyone that I forget.
At the first place I thank Roland Fried for the excellent supervision over the last years, for his help and constant support, for a very interesting research topic and, not least, for his leniency in tolerating my faults.
There are many people that I have talked to, that gave me advice and shared their knowledge with me, many people I have worked with and discussed about statistics, discussions that have been very fruitful and enlightening: Christoph Croux, Herold Dehling, Michael Eichler, Anna Gottard, Sonja Kuhnt, Christine Müller, Klaus Nordhausen, Hannu Oja, Davy Paindaveine, David E. Tyler, Jürgen Vogel, Silvia Vogel and many, many more.
I would like to thank everyone at the Statistics Department at the University in Dortmund, in particular the Biosciences group, for the very pleasent and productive atmosphere. I greatly benefited from the excellent scientific infrastructure: Thorsten Bernholt and Robin Nunkesser from the Computer Science Department wrote an evolutionary algorithm to compute the Oja median, which was analyzed and made available in R by Daniel Fischer (2008) and further evolved into the R-Package OjaNP (Fischer et al., 2009). Alexander Dürre and Claudia Köllmann assisted on most of the simulations in this thesis. Their reliable help is highly appreciated. Also, the competence of Sebastian Krey, Uwe Ligges and Olaf Mersmann in all computing matters, particularly LaTeX and R, should not go accounted for.
Last and most I thank Kristine Uebelgünn for her enduring support.

Chapter 1

Introduction

1.1 Motivation and summary

In this thesis two lines of statistical research meet that, so far, have coexisted and very little interfered with each other: *graphical modelling* and *robustness*.
Graphical model is a very broad term, used whenever graphs are employed to express associations between several entities. When talking about graphical models in statistics, these entities, represented by the nodes of the graph, are random variables, and an edge between two nodes, directed or undirected, reflects some form of probabilistic dependence. In the large majority of cases the type of relation expressed by an edge is of conditional nature, for example, conditional independence of two variables given all other variables. There are several reasons that favour conditional over ordinary, marginal dependence, reasons that lie in the benefits of the graphical representation, see Section 2.2.2, especially Theorem 2.2.7, as well as reasons concerning the relevance for multivariate data analysis in general, see, e.g., the many examples in Lauritzen (1996, Chapter 4), Edwards (2000, Chapter 1) or any instance of what is known as the Yule-Simpson paradox. It can be argued that, whenever one has more than two variables of interest, conditional dependence is much more meaningful than marginal dependence.
Graphical modelling then refers to the statistical task of selecting a graph that appropriately reflects the dependence structure of a given data set, and the body of statistical tools and methods applied towards this end. While graphical models are already a wide area, generated by the many possible interpretations of an edge (neither is a graph restricted to one type of edges), graphical modelling is an even larger field. The statistical modelling substantially differs with type of distribution, e.g. if it is continuous or discrete. Moreover, it may be approached in the frequentist or the Bayesian framework. Graphical modelling is ultimately a model choice problem, and there is generally not one best answer to it.
This thesis is about the graphical modelling of continuous data. Graphical modelling of continuous data, may it be frequentist or Bayesian, is some way or another, based on Gaussianity. There is only one exception known to me, and that is the use of the skew normal distribution (Capitanio et al., 2003). Contrary to the categorical case such a restrictive distributional assumption is necessary in the continuous case in order to have some workable, useful statistical model. Nevertheless it remains a restriction, and it emerges in many statistical applications that data tends to have larger than normal tails. The driving goal of this thesis is to free graphical modelling of this dependence on Gaussianity. This is meant in two ways:

(I) Allow a larger class of distributions and devise methods that are valid and efficient on the whole

class. This can be phrased as *robustness against non-normality*.

(II) While generally sticking to the normality assumption, reduce the susceptibility towards outliers and model misspecifications of the statistical methods, which are usually likelihood based and known to be sensitive in the respect. This means robustness in the classical sense, as described e.g. in Hampel et al. (1986).

Having set the goal, the next question is where to begin its implementation. Naturally at the beginning, with the basic case[1]: we consider graphical models with *only undirected edges*, i.e. mutual dependencies, not directed influences, and with *only continuous variables*, as compared to a mixture of, say, continuous and categorical variables. In such a situation, the joint distribution of all variables is assumed to be multivariate normal, and such models go under the name *Gaussian graphical models* in the literature. Equivalently used terms are *covariance selection models* and *concentration graph models*.

The whole dependence information is then fully contained in the covariance matrix, and classical, likelihood based Gaussian graphical modelling is an analysis of the sample covariance matrix. An appealing way of robustifying Gaussian graphical modelling is thus an plug-in approach: replace the highly non-robust sample covariance matrix by an alternative, more robust scatter estimator and apply any subsequent analysis in analogous manner.

The first proposal of this kind, to my knowledge, is by Becker (2005), who suggested to use the reweighted minimum covariance determinant (RMCD) estimator, underpinned by a simulated example demonstrating that, if the contamination is severe enough, the RMCD will eventually outperform the sample covariance matrix. This thesis gives, among other things, answers to all open questions Becker (2005) poses at the end of the article.

The RMCD is not the only potential robust substitute for the sample covariance matrix and not in the focus of the thesis. We identify *proportional affine equivariance* (i.e. affine equivariance up to a multiplicative constant) as a key property that allows to formulate simple modifications of Gaussian graphical modelling tools in a unified framework. Many proposals of robust scatter estimators possessing this property have been made over the last decades, see Sections 2.4.1 and 3.4.1.

By employing the class of elliptical distributions as data model—as a convenient way of modelling large tails of several variates—we can analyze our statistical methods under non-normality and thus quantify what we lose by, say, going away from normality towards heavier tails. This motivates to consider graphical models for elliptical distributions, which we label *elliptical graphical models* in analogy to Gaussian graphical models. The thesis indeed approaches both formulated aims (I) and (II): Besides robustifying the statistical methodology used under Gaussianity we devise graphical models for the broader class of elliptical distributions and give instructions for estimating and testing within these models.

Rather than aiming at high performance at a specific distribution we are interested in a good performance over a preferably broad range of distributions. In this respect we particularly mention Tyler's scatter estimator, which is (asymptotically) distribution-free within the elliptical model.

[1] We may call it *simple case*, in the sense that more complex situations build on results from the basic case. For instance, before looking at chain graphs we should know how to analyze a chain element. This view equally encourages the terminology *fundamental case*.

1.2 Outline of the thesis

This is a cumulative thesis. It has two main parts: the actual thesis stretching over Chapters 2 to 4 and several supplements making up Chapter 5.

The **first part**: Chapters 2 and 3 are two fully self-contained expositions that may be read individually, which entails that both have a separate introduction. The notation is consistent except that vectors are not bold in Chapters 3, which is a stylistic requirement of the journal it has been submitted to. Chapter 4 uses the notation of Chapter 3. Despite the cumulative structure of the thesis all three chapters fit tightly together and build upon each other with almost no overlap. The chapters in detail:
Chapter 2 **On robust Gaussian graphical modelling** gives the current state of research on the topic. Sections 2.1 and 2.2 give an instructive introduction to Gaussian graphical models, Section 2.3 briefly recollects the important terms of robustness, particularly robust multivariate scatter estimation and surveys the (yet manageable amount of) literature on robust Gaussian graphical modelling. The chapter was written in autumn 2009 and has been published as

> Vogel, D., Fried, R.: On robust Gaussian graphical modelling. In: Devroye, L., Karasözen, B., Kohler, M., Korn, R. (eds.) Recent Developments in Applied Probability and Statistics. Dedicated to the Memory of Jürgen Lehn, pp. 155-182. Berlin, Heidelberg: Springer-Verlag (2010).

It is a review article with no genuine research results except Proposition 2.4.5 and lays the groundwork for the subsequent Chapters 3 and 4.
In Chapter 3 **Elliptical graphical modelling — the decomposable case** a new class of graphical models is proposed along with suggestions of how to estimate and test, allowing to employ basically any model selection scheme from classical Gaussian graphical modelling in an analogous manner. The main result is the validity of a generalized version of the deviance test, cf. Proposition 3.3.9. All mathematical derivations are formulated for decomposable graphs. Decomposable graphical models are of particular interest, due to better interpretability. They are at the same time better tractable mathematically and thus dominate the literature on the theoretical as well as on the applied side. Tyler's M-estimator and the RMCD are mentioned as examples of robust, affine equivariant scatter estimators. Their finite sample performance is evaluated and compared to that of the sample covariance matrix in a small simulation study. Chapter 3 was written mainly in the first half of 2010, and has been submitted for publication on September 29, 2010, under the name *Elliptical graphical modelling*.
Chapter 4 **Elliptical graphical modelling — the non-decomposable case** is the newest part, written in September 2010, and contains unsubmitted material. An explicit formula for the asymptotic covariance of a constrained shape estimator $\hat{S}_G$ for a general, not necessarily decomposable graph G is given. The formula of the asymptotic distribution of $\hat{S}_G$ for *decomposable* G in Chapter 3 (Proposition 3.3.4) is derived by means of the perfect sequence representation of the cliques of G. The corresponding formula for *general* G given in Proposition 4.1.3, which is deduced from the implicit function theorem, is in fact completely different and hardly recognizable to describe the same quantity. Both approaches are worthwhile in their own right.
I consider Corollary 4.1.5 and Proposition 4.1.9 as the most important results of this thesis.

The **second part** of the thesis, also roughly the second half in page numbers, consists of Chapter 5 **Supplements**, that assembles six manuscripts that were written from 2008 to 2010. These manuscripts do what the chapter title says: they supplement, they do not amend, complete or complement the thesis. They provide additional information and insight, e.g., further examples of scatter estimators

(Sections 5.1, 5.2) and extended simulation results (Section 5.5), they reflect earlier stages of the work (Section 5.4) and report intermediate results on work in progress that is related to, but at some point branched off of the main course of the thesis (Sections 5.3 and 5.6). The thesis may be read and judged without them. Sections 5.1, 5.2, 5.4 and 5.5 have appeared in conference proceedings, Sections 5.3 and 5.6 are not published elsewhere. A description of all manuscripts is given at the beginning of Chapter 5 on page 62.

The Appendix **On how to differentiate matrix-valued functions w.r.t. matrices** gives a brief introduction to matrix differential calculus, in particular explains how to differentiate functions of symmetric matrices, which is a vital tool of most proofs in the thesis. Section A.1 is a three-pages aggregation of the essentials of matrix differential calculus as it is explained in Magnus and Neudecker (1999). Section A.2 is a suggestion on how to deal with derivatives w.r.t. symmetric matrices, which I have not found as such in the literature.

1.3 Outlook

It is my personal impression that graphical models become increasingly popular and relevant, which is reflected in an ongoing active research in graphical modelling. The material presented here must fall short of covering more than just a tiny fraction of what we set out as our prime objective: non-Gaussian graphical modelling. We give an outlook guided by the question what still is to be done. A set of answers is generated by the limitations and alternatives of our approach. We want to name a few:

- In recent years the research on Gaussian graphical models has been particularly driven by the desire to analyze high-dimensional data sets, (e.g. Drton and Perlman, 2004; Meinshausen and Bühlmann, 2006; Castelo and Roverato, 2006; Yuan and Lin, 2007; Verzelen and Villers, 2009). Our plug-in method fails to provide a solution in the $p > n$ situation and does neither allow a simple transfer of standard techniques, like e.g. regularization, that are used in Gaussian graphical models. We face the inherent problem that any affine equivariant, robust estimator requires more than $p + 1$ data points, because the only affine equivariant scatter estimator in the $p+1 > n$ situation is the sample covariance estimator (Tyler, 2010). Droppung the affine equivariance property is inevitable for robust, high-dimensional graphical modelling, and alternative estimators should be examined, see also Section 3.4.3.

- The tests proposed in Sections 3.3.3 and 4.1.2 rely on asymptotic approximations, which give good answers if n is sufficiently large, but may be rather inaccurate for small n. This has been noted in the context of classical graphical modelling, improved small-sample approximations have been proposed (Porteous, 1985, 1989), but also the exact distribution of the deviance test statistic is known for decomposable models, cf. Lauritzen (1996, Sections 5.2.2 and 5.3.3). While the exact distribution of most robust, affine equivariant estimators is hardly accessible and eludes a unified treatment as it is possible for the asymptotics, finite sample correction techniques may be applicable in an analogous manner. It seems worthwhile to devote some further attention to the small-sample properties of the proposed elliptical graphical modelling methods.

- The class of affine equivariant scatter estimators contains the elliptical maximum likelihood estimators (MLEs). Hence, assuming knowledge of the population distribution the use of the appropriate MLE is an efficient way of unconstrained estimation. However, the constrained

estimation that we propose for elliptical distributions, see Section 3.3.2, is derived from the Gaussian likelihood equations. We may increase the efficiency of the test procedure by employing also an elliptical constrained MLE. The challenge here is less the statistical distribution theory—this is covered by the general maximum likelihood framework—, but the numerical theory: find a suitable algorithm solving the likelihood equations and prove its convergence. In a way, this situation is antipodal to what we do here: we make use of the well-developed numerical theory in the Gaussian case, where we know how to compute the estimators, but have to derive their asymptotics under different assumptions.

- We have motivated the elliptical model as a convenient way of modelling heavy tails, following the statistical modelling principle to make things as complicated as necessary but as simple as possible. However, data may deviate from normality in many ways and does not have to be elliptical. An alternative model for continuous multivariate data is the independent-components model (ICM), as it is considered in Oja et al. (2010). It is also a semiparametric model, where the dependencies are coded in the parametric part (the mixing matrix in the ICM, the shape matrix in the elliptical model). Using the ICM we may investigate the robustness against non-ellipticity of our methods. But it is much more interesting to model the conditional independence graph (CIG) of such a distribution, i.e. to study full probabilistic dependence (including linear as well as non-linear dependencies) as opposed to only linear dependencies that we consider when modelling the partial correlation graph (PCG). Note that in the elliptical model the CIG is either saturated or, in case of the normal distribution, coincides with the PCG. This leads to a completely different way of generalizing Gaussian graphical models that still holds many theoretical challenges. Some thoughts are gathered in Section 5.6.

At the beginning we hinted at the diversity of what is understood as *graphical models* and *graphical modelling*, but have also pointed out that very little research so far has been devoted to the problem of robustness in this context. Going back to the prime objectives (I) and (II) we are still left to examine all other types of graphical models with continuous variables w.r.t. their robustness properties and potential relaxation of the normality assumption, including:

- models with directed edges,
- models with both, directed and undirected edges,
- models with continuous and categorical variables and
- graphical models for dynamic data, i.e. variables recorded over time, allowing dependencies along time and across the variables. This includes static graphical models, that reflect the dependence structure for process-valued random variables (e.g. Brillinger, 1996; Dahlhaus, 2000; Fried and Didelez, 2003) as well as dynamic graphical models, allowing the dependence structure to vary over time.

On July 1, 2010, Xuming He gave a keynote lecture entitled "Robust Statistics 2020" at the 10th International Conference on Robust Statistics in Prague, Czech Republic. He was asked by the organizers to attempt a prediction of the future development of robust statistics, manifested in the session topics of a potential ICORS meeting in 2020. He particularly mentioned Gaussian graphical models and formulated a personal wish list that included a session on robust Graphical models. I consider this thesis a step in this direction.

Chapter 2

On robust Gaussian graphical modelling

2.1 Introduction

Graphical modelling is the analysis of conditional associations between random variables by means of graph theoretic methods. The graphical representation of the interrelation of several variables is an attractive data analytical tool. Besides allowing parsimonious modelling of the data it facilitates the understanding and the interpretation of the data generating process. The importance of considering *conditional* rather than marginal associations for assessing the dependence structure of several variables is vividly exemplified by Simpson's paradox, see e.g. Edwards (2000), Chap. 1.4. The statistical literature knows several different types of graphical models, differing in the type of relation coded by an edge, in the type of data and hence in the statistical methodology. In this chapter we deal with undirected graphs only, that is, the type of association we consider is mutual. Precisely, we are going to define partial correlation graphs in Sect. 2.2.2.
Undirected models are in a sense closer to the data. A directed association suggests a causal relationship. Even though it can often be justified, e.g. by chronology or knowledge about the physiological process, the direction of the effect is an additional assumption. Undirected models constitute the simplest case, the understanding of which is crucial for the study of directed models and models with both, directed and undirected edges.
Furthermore we restrict our attention to continuous data, which are assumed to stem from a multivariate Gaussian distribution. Conditional independence in the normal model is nicely expressed through its second order characteristics, cf. Sect. 2.2.3. This fact, along with its general predominant role in multivariate statistics (largely due to the Central limit theorem justification), is the reason for the almost exclusive use of the multivariate normal distribution in graphical models for continuous data.
With rapidly increasing data sizes, and on the other hand computer hardware available to process them, the need for robust methods becomes more and more important. The sample covariance matrix possesses good statistical properties in the normal model and is very fast to compute, but highly non-robust, cf. Sect. 2.4.1. We are going to survey robust alternatives to the classical Gaussian graphical modelling, which is based on the sample covariance matrix.
The paper is organized as follows. Section 2.2 introduces Gaussian graphical models (GGMs). We start by studying partial correlations, a purely moment based relation, without any distributional assumption and then examine the special case of the normal distribution where partial uncorrelatedness coincides with conditional independence. The better transferability of the former concept to more general data situations is the reason for taking this route. Section 2.3 reviews the classical, non-robust,

likelihood-based statistical theory for Gaussian graphical models. Each step is motivated, and important points are emphasized. Sections 2.2 and 2.3 thus serve as a self-contained introduction to GGMs. The basis for this first part are the books Whittaker (1990) and Lauritzen (1996). Other standard volumes on graphical models in statistics are Cox and Wermuth (1996) and Edwards (2000), both with a stronger emphasis on applications. Section 2.4 deals with robust Gaussian graphical modelling. We focus on the use of robust affine equivariant scatter estimators, since the robust estimators proposed for GGMs in the past belong to this class. As an important robustness measure we consider the influence function and give the general form of the influence functions of affine equivariant scatter estimators and derived partial correlation estimators.

We close this section by introducing some of the mathematical notation we are going to use. Bold letters $\boldsymbol{b}$, $\boldsymbol{\mu}$, etc., denote vectors, capital letters X, Y, etc., indicate (univariate) random variables and bold capital letters $\boldsymbol{X}$, $\boldsymbol{Y}$, etc., random vectors. We view vectors, by default, neither as a column nor as a row, but just as an ordered collection of elements of the same type. This makes $(\boldsymbol{X}, \boldsymbol{Y})$ again a vector and not a two-column matrix. However, if matrix notation, such as $(\cdot)^T$, is applied to vectors, they are always interpreted as $n \times 1$ matrices.

Matrices are also denoted by non-bold capital letters, and the corresponding small letter is used for an element of the matrix, e.g., the $p \times p$ matrix Σ is the collection of all $\sigma_{i,j}$, $i, j = 1, ..., p$. Alternatively, if matrices are denoted by more complicated compound symbols (e.g. if they carry subscripts already) square brackets will be used to refer to individual elements, e.g. $[\hat{\Sigma}_G^{-1}]_{i,j}$. Throughout the paper upright small Greek letters will denote index sets. Subvectors and submatrices are referenced by subscripts, e.g. for $\alpha, \beta \subseteq \{1, ..., p\}$ the $|\alpha| \times |\beta|$ matrix $\Sigma_{\alpha,\beta}$ is obtained from Σ by deleting all rows that are not in α and all columns that are not in β. Similarly, the $p \times p$ matrix $[\Sigma_{\alpha,\beta}]^p$ is obtained from Σ by putting all rows not in α and all columns not in β to zero. We want to view this matrix operation as two operations performed sequentially: first $(\cdot)_{\alpha,\beta}$ extracting the submatrix and then $[\cdot]^p$ writing it back on a "blank" matrix at the coordinates specified by α and β. Of course, the latter is not well defined without the former, but this allows us e.g. to write $[(\Sigma_{\alpha,\beta})^{-1}]^p$.

We adopt the general convention that subscripts have stronger ties than superscripts, for instance, we write $\Sigma^{-1}_{\alpha,\beta}$ for $(\Sigma_{\alpha,\beta})^{-1}$. Let $\mathscr{S}_p$ and $\mathscr{S}_p^+$ be the sets of all symmetric, respectively positive definite $p \times p$ matrices, and define for any $A \in \mathscr{S}_p^+$

$$\mathrm{Corr}(A) = A_D^{-\frac{1}{2}} A A_D^{-\frac{1}{2}}, \tag{2.1}$$

where A_D denotes the diagonal matrix having the same diagonal as A. Recall the important inversion formula for partitioned matrices. Let $r \in \{1, ..., p-1\}$, $\alpha = \{1, ..., r\}$ and $\beta = \{r+1, ..., p\}$. Then

$$\begin{pmatrix} \Sigma_{\alpha,\alpha} & \Sigma_{\alpha,\beta} \\ \Sigma_{\beta,\alpha} & \Sigma_{\beta,\beta} \end{pmatrix}^{-1} = \begin{pmatrix} \Omega^{-1} & -\Omega^{-1}\Sigma_{\alpha,\beta}\Sigma^{-1}_{\beta,\beta} \\ -\Sigma^{-1}_{\beta,\beta}\Sigma_{\beta,\alpha}\Omega^{-1} & \Sigma^{-1}_{\beta,\beta} + \Sigma^{-1}_{\beta,\beta}\Sigma_{\beta,\alpha}\Omega^{-1}\Sigma_{\alpha,\beta}\Sigma^{-1}_{\beta,\beta} \end{pmatrix}, \tag{2.2}$$

where the $r \times r$ matrix $\Omega = \Sigma_{\alpha,\alpha} - \Sigma_{\alpha,\beta}\Sigma^{-1}_{\beta,\beta}\Sigma_{\beta,\alpha}$ is called the *Schur complement* of $\Sigma_{\beta,\beta}$. The inverse exists if and only if Ω and $\Sigma_{\beta,\beta}$ are both invertible. Note that, by simultaneously re-ordering rows and columns, the formula is valid for any partition $\{\alpha, \beta\}$ of $\{1, ..., p\}$.

Finally, the Kronecker product $A \otimes B$ of two matrices $A, B \in \mathbb{R}^{p\times p}$ is defined as the $p^2 \times p^2$ matrix with entry $a_{i,j}b_{k,l}$ at position $((i-1)p + k, (j-1)p + l)$. Let $\boldsymbol{e}_1, ..., \boldsymbol{e}_p$ be the unit vectors in $\mathbb{R}^p$ and $\mathbf{1}_p$ the p vector consisting only of ones. Define further the following matrices:

$$J_p = \sum_{i=1}^{p} \boldsymbol{e}_i\boldsymbol{e}_i^T \otimes \boldsymbol{e}_i\boldsymbol{e}_i^T, \qquad K_p = \sum_{i=1}^{p}\sum_{j=1}^{p} \boldsymbol{e}_i\boldsymbol{e}_j^T \otimes \boldsymbol{e}_j\boldsymbol{e}_i^T \quad \text{and} \quad M_p = \frac{1}{2}\left(I_{p^2} + K_p\right),$$

where I_{p^2} denotes the $p^2 \times p^2$ identity matrix. K_p is also called the *commutation matrix*. Let vec(A) be the p^2 vector obtained by stacking the columns of $A \in \mathbb{R}^{p\times p}$ from left to right underneath each other. More on these concepts and their properties can be found in Magnus and Neudecker (1999).

2.2 Partial correlation graphs and properties of the Gaussian distribution

This section explains the basic concepts of Gaussian graphical models: We define the terms *partial variance* and *partial correlation* (Sect. 2.2.1), review basic graph theory terms and explain the merit of a *partial correlation graph* (Sect. 2.2.2). Gaussianity enters in Sect. 2.2.3, where we deduce the conditional independence interpretation of a partial correlation graph which is valid under normality. Statistics is deferred to Sect. 2.3.

2.2.1 Partial variance

Let $\boldsymbol{X} = (X_1, ..., X_p)$ be a random vector in $\mathbb{R}^p$ with distribution F and positive definite variance matrix $\Sigma = \Sigma_{\boldsymbol{X}} \in \mathbb{R}^{p\times p}$. The inverse of Σ is called *concentration matrix* (or *precision matrix*) of $\boldsymbol{X}$ and shall be denoted by K or $K_{\boldsymbol{X}}$.
Now let $\boldsymbol{X}$ be partitioned into $\boldsymbol{X} = (\boldsymbol{Y}, \boldsymbol{Z})$, where $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are subvectors of lengths q and r, respectively. The corresponding index sets shall be called α and β, i.e. $\alpha = \{1, ..., q\}$ and $\beta = \{q + 1, ..., q + r\}$. The variance matrix of $\boldsymbol{Y}$ is $\Sigma_{\boldsymbol{Y}} = \Sigma_{\alpha,\alpha} \in \mathbb{R}^{q\times q}$ and its concentration matrix $K_{\boldsymbol{Y}} = \Sigma^{-1}_{\alpha,\alpha} = (K^{-1}_{\boldsymbol{X}})^{-1}_{\alpha,\alpha}$. The covariance matrix of $\boldsymbol{Y}$ and $\boldsymbol{Z}$ is $\Sigma_{\alpha,\beta} \in \mathbb{R}^{q\times r}$. The orthogonal projection of $\boldsymbol{Y}$ onto the space of all affine linear functions of $\boldsymbol{Z}$ shall be denoted by $\hat{\boldsymbol{Y}}(\boldsymbol{Z})$ and is given by

$$\hat{\boldsymbol{Y}}(\boldsymbol{Z}) = \mathbb{E}\boldsymbol{Y} + \Sigma_{\alpha,\beta}\Sigma^{-1}_{\beta,\beta}(\boldsymbol{Z} - \mathbb{E}\boldsymbol{Z}). \tag{2.3}$$

This is the best linear prediction of $\boldsymbol{Y}$ from $\boldsymbol{Z}$, in the sense that the squared prediction error $\mathbb{E}\|\boldsymbol{Y} - h(\boldsymbol{Z})\|^2$ is uniquely minimized by $h = \hat{\boldsymbol{Y}}(\cdot)$ among all (affine) linear functions h. The *partial variance of $\boldsymbol{Y}$ given $\boldsymbol{Z}$* is the variance of the residual $\boldsymbol{Y} - \hat{\boldsymbol{Y}}(\boldsymbol{Z})$. It shall be denoted by $\Sigma_{\boldsymbol{Y}\bullet\boldsymbol{Z}}$, i.e.

$$\Sigma_{\boldsymbol{Y}\bullet\boldsymbol{Z}} = \mathrm{Var}(\boldsymbol{Y} - \hat{\boldsymbol{Y}}(\boldsymbol{Z})) = \Sigma_{\alpha,\alpha} - \Sigma_{\alpha,\beta}\Sigma^{-1}_{\beta\beta}\Sigma_{\beta,\alpha}. \tag{2.4}$$

The notation $\boldsymbol{Y}\bullet\boldsymbol{Z}$ is intended to resemble $\boldsymbol{Y} \mid \boldsymbol{Z}$, that is, we look at $\boldsymbol{Y}$ in dependence on $\boldsymbol{Z}$, but instead of conditioning $\boldsymbol{Y}$ on $\boldsymbol{Z}$ the type of connection we consider here is a linear regression. In particular, $\Sigma_{\boldsymbol{Y}\bullet\boldsymbol{Z}}$ is—contrary to a conditional variance—a fixed parameter and not random.
If $\boldsymbol{Y}$ is at least two-dimensional, we partition it further into $\boldsymbol{Y} = (\boldsymbol{Y}_1, \boldsymbol{Y}_2)$ with corresponding index sets $\alpha_1 \cup \alpha_2 = \alpha$ and lengths $q_1 + q_2 = q$, and define

$$\Sigma_{\boldsymbol{Y}_1,\boldsymbol{Y}_2\bullet\boldsymbol{Z}} = (\Sigma_{\boldsymbol{Y}\bullet\boldsymbol{Z}})_{\alpha_1,\alpha_2} = \Sigma_{\alpha_1,\alpha_2} - \Sigma_{\alpha_1,\beta}\Sigma^{-1}_{\beta\beta}\Sigma_{\beta,\alpha_2}$$

as the *partial covariance between $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ given $\boldsymbol{Z}$*. If $\Sigma_{\boldsymbol{Y}_1,\boldsymbol{Y}_2\bullet\boldsymbol{Z}} = \boldsymbol{0}$, we say *$\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ are partially uncorrelated given $\boldsymbol{Z}$* and write

$$\boldsymbol{Y}_1 \perp \boldsymbol{Y}_2 \bullet \boldsymbol{Z}.$$

Furthermore, if $\boldsymbol{Y}_1 = Y_1$ and $\boldsymbol{Y}_2 = Y_2$ are both one-dimensional, $\Sigma_{\boldsymbol{Y}\bullet\boldsymbol{Z}}$ is a positive definite 2×2 matrix. The correlation coefficient computed from this matrix, i.e. the $(1, 2)$ element of $\mathrm{Corr}(\Sigma_{\boldsymbol{Y}\bullet\boldsymbol{Z}})$, cf. (2.1), is called the *partial correlation (coefficient) of Y_1 and Y_2 given $\boldsymbol{Z}$* and denoted by $\varrho_{Y_1,Y_2\bullet\boldsymbol{Z}}$. This is

nothing but the correlation between the residuals $Y_1 - \hat{Y}_1(\boldsymbol{Z})$ and $Y_2 - \hat{Y}_2(\boldsymbol{Z})$ and may be interpreted as a measure of the linear association between Y_1 and Y_2 after the linear effects of $\boldsymbol{Z}$ have been removed. For $\alpha_1 = \{i\}$ and $\alpha_2 = \{j\}$, $i \neq j$, we use the simplified notation $\varrho_{i,j\bullet}$ for $\varrho_{X_i,X_j\bullet\boldsymbol{X}_{\setminus\{i,j\}}}$.

The simple identity (2.4) is fundamental and the actual starting point for all following considerations. We recognize $\Sigma_{\boldsymbol{Y}\bullet\boldsymbol{Z}}$ as the Schur complement of $\Sigma_{\boldsymbol{Z}}$ in $\Sigma_{\boldsymbol{X}}$, cf. (2.2), implying that

$$\Sigma^{-1}_{\boldsymbol{Y}\bullet\boldsymbol{Z}} = K_{\alpha,\alpha}. \tag{2.5}$$

In words: the concentration matrix of $\boldsymbol{Y} - \hat{\boldsymbol{Y}}(\boldsymbol{Z})$ is the submatrix of $K_{\boldsymbol{X}}$ corresponding to $\boldsymbol{Y}$, or—very roughly put—while marginalizing means partitioning the covariance matrix, partializing means partitioning its inverse. This has some immediate implications about the interpretation of K, which explain why K, rather than Σ, is of interest in graphical modelling.

Proposition 2.2.1 *The partial correlation $\varrho_{i,j\bullet}$ between X_i and X_j, $1 \leq i < j \leq p$, given all remaining variables $\boldsymbol{X}_{\setminus\{i,j\}}$ is*

$$\varrho_{i,j\bullet} = -\frac{k_{i,j}}{\sqrt{k_{i,i}k_{j,j}}}.$$

Another way of phrasing this assertion is to say, the matrix $P = -\mathrm{Corr}(K)$ contains the partial correlations (of each pair of variables given the respective remainder) as its off-diagonal elements. We call P the *partial correlation matrix of* $\boldsymbol{X}$. Proposition 2.2.1 is a direct consequence of (2.5) involving the inversion of a 2×2 matrix. For a detailed derivation see Whittaker (1990), Chap. 5.

2.2.2 Partial correlation graph

The partial correlation structure of the random variable $\boldsymbol{X}$ can be coded in a graph, which originates the term *graphical model*. An undirected graph $G = (V, E)$, where V is the vertex set and E the edge set, is constructed the following way: the variables $X_1, ..., X_p$ are the vertices, and an undirected edge is drawn between X_i and X_j, $i \neq j$, if and only if $\varrho_{i,j\bullet} \neq 0$. The thus obtained graph G is called the *partial correlation graph (PCG)* of $\boldsymbol{X}$. Formally we set $V = \{1, ..., p\}$ and write the elements of E as unordered pairs $\{i, j\}$, $1 \leq i < j \leq p$. Before we dwell on the benefits of this graphical representation, let us briefly recall some terms from graph theory. We only consider undirected graphs with a single type of nodes.

If $\{a, b\} \in E$, the vertices a and b are called *adjacent* or *neighbours*. The set of neighbours of the vertex $a \in V$ is denoted by ne(a). An alternative notation is bd(a), which stands for *boundary*, but keep in mind that in graphs containing directed edges the set of neighbours and the boundary of a node are generally different.

A *path of length* k, $k \geq 1$, is a sequence $(a_1, ..., a_{k+1})$ of distinct vertices such that $\{a_i, a_{i+1}\} \in E$, $i = 1, ..., k$. If $k \geq 2$ and additionally $\{a_1, a_{k+1}\} \in E$, then the sequence $(a_1, ..., a_{k+1}, a_1)$ is called a *cycle of length* $k + 1$ or a $(k + 1)$*-cycle*. Note that the length, in both cases, refers to the number of edges.

The n-cycle $(a_1, ..., a_n, a_1)$ is *chordless*, if no other than successive vertices in the cycle are adjacent, i.e. $\{a_i, a_j\} \in E \Rightarrow |i - j| \in \{1, n - 1\}$. Otherwise the cycle possesses a *chord*. All cycles of length 3 are chordless.

The graph is called *complete*, if it contains all possible edges. Every subset $\alpha \subset V$ induces a *subgraph* $G_\alpha = (\alpha, E_\alpha)$, where E_α contains those edges in E with both endpoints in α, i.e. $E_\alpha = E \cap (\alpha \times \alpha)$. A subset $\alpha \subset V$, for which G_α is complete, but adding another vertex would render it incomplete, is called a *clique*. Thus the cliques identify the maximal complete subgraphs.

The set $\gamma \subset V$ is said to *separate* the sets $\alpha, \beta \subset V$ in G, if α, β, γ are mutually disjoint and every path from a vertex in α to a vertex in β contains a node from γ. The set γ may be empty.

Definition 2.2.2 *A partition* (α, β, γ) *of* V *is a* decomposition *of the graph* G*, if*

(1) α*,* β *are both non-empty,*

(2) γ *separates* α *and* β*,*

(3) G_γ *is complete.*

If such a decomposition exists, G *is called* reducible *(otherwise* irreducible*). It can then be* decomposed into *or* reduced to *the components* $G_{\alpha\cup\gamma}$ *and* $G_{\beta\cup\gamma}$.

Our terminology is in concordance with Whittaker (1990), Chap. 12, however, there are different definitions around. For instance, Lauritzen (1996) calls a decomposition in the above sense a "proper weak decomposition". Also be aware that the expression "G is decomposable", which is defined below, denotes something different than "there exists a decomposition of G", for which the term "reducible" is used.
Definition 2.2.2 suggests a recursive application of decompositions until ultimately the graph is fully decomposed into irreducible components, which then are viewed as atomic building blocks of the graph. It is not at all obvious, if such atomic components exist or are well defined, since, at least in principle, any sequence of decompositions may lead to different irreducible components, cf. Example 12.3.1 in Whittaker (1990). With an additional constraint, the irreducible components of a given graph are indeed well defined.

Definition 2.2.3 *The system of subsets* $\{\alpha_1, ..., \alpha_k\} \subset 2^{|V|}$ *is called the (set of)* maximal irreducible components of G*, if*

(1) G_{α_i} *is irreducible,* $i = 1, ..., k$*,*

(2) α_i *and* α_j *are mutually incomparable, i.e.* α_i *is not a proper subset of* α_j *and vice versa,* $1 \leq i < j \leq k$*, and*

(3) $\bigcup_{i=1}^{k} \alpha_i = V$.

The maximal irreducible components of any graph G are unique and can be obtained by first fully decomposing the graph into irreducible components (by any sequence of decompositions) and then deleting those that are a proper subset of another one—the *maximal* irreducible components remain.

Definition 2.2.4 *The graph* G *is* decomposable*, if all of its maximal irreducible components are complete.*

Decomposability also admits the following recursive definition: G is decomposable, if it is complete or there exists a decomposition (α, β, γ) into decomposable subgraphs $G_{\alpha\cup\gamma}$ and $G_{\beta\cup\gamma}$. Another characterization is to say, a decomposable graph can be decomposed into its cliques. Figure 2.1 shows two reducible graphs and their respective maximal irreducible components. The decomposability of a graph is a very important property, with various implications for graphical models, and decomposable graphs deserve and receive special attention, cf. e.g. Whittaker (1990), Chap. 12. The most notable consequence for Gaussian graphical models is the existence of closed form maximum likelihood estimates, cf. Sect. 2.3.1. The recursive nature of Definition 2.2.4 makes it hard to determine whether a given graph is decomposable or not. Several equivalent characterizations of decomposability are given e.g. in Lauritzen (1996). We want to name one, which is helpful for spotting decomposable graphs.

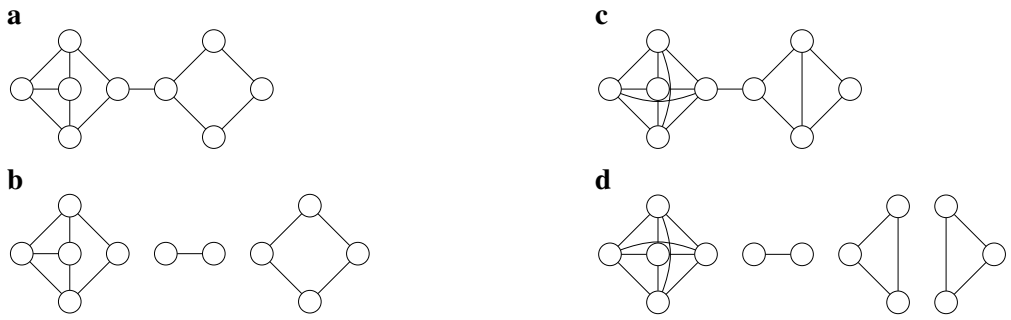

Figure 2.1: **a** a non-decomposable graph and **b** its maximal irreducible components, **c** a decomposable graph and **d** its maximal irreducible components

Definition 2.2.5 *The graph G is* triangulated*, if every cycle of length greater than* 3 *has a chord.*

Proposition 2.2.6 *A graph G is decomposable if and only if it is triangulated.*

For a proof see Lauritzen (1996), p. 9, or Whittaker (1990), p. 390.
We close this subsection by giving a motivation for partial correlation graphs. Clearly, the information in the graph is fully contained in Σ and can directly be read off its inverse K: a zero off-diagonal element at position (i, j) signifies the absence of an edge between the corresponding nodes. Of course, graphs in general are helpful visual tools. This argument is valid for representing any type of association between variables by means of a graph and is not the sole justification for partial correlation graphs. The purpose of a PCG is explained by the following theorem, which lies at the core of graphical models.

Theorem 2.2.7 (Separation theorem for PCGs) For a random vector $\boldsymbol{X}$ with positive definite covariance matrix Σ and partial correlation graph G the following is true: γ separates α and β in G if and only if $\boldsymbol{X}_\alpha \perp \boldsymbol{X}_\beta \bullet \boldsymbol{X}_\gamma$.

This result is not trivial, but its proof can be accomplished by matrix manipulation. It is also a corollary of Theorem 3.7 in Lauritzen (1996) by exploiting the equivalence of partial uncorrelatedness and conditional independence in the normal model, cf. Sect. 2.2.3. The theorem roughly tells that the association "partial uncorrelatedness" (of two random vectors given a third one) exhibits the same properties as the association "separation" (of two sets of vertices by a third one). Thus it links probability theory to graph theory and allows to employ graph theoretic tools in studying properties of multivariate probability measures. First and foremost it allows the succinct formulation of Theorem 2.2.7. The theorem lets us, starting from the pairwise partial correlations, conclude the partial uncorrelatedness $\boldsymbol{X}_\alpha \perp \boldsymbol{X}_\beta \bullet \boldsymbol{X}_\gamma$ for a variety of triples $(\boldsymbol{X}_\alpha, \boldsymbol{X}_\beta, \boldsymbol{X}_\gamma)$ (which do not have to form a partition of $\boldsymbol{X}$). It is the graph theoretic term *separation* that allows not only to concisely characterize these triples, but also to readily identify them by drawing the graph.
Finally, Theorem 2.2.7 can be re-phrased, saying that in a PCG the pairwise and the global Markov property are equivalent: We say, a random vector $\boldsymbol{X} = (X_1, ..., X_p)$ satisfies the *pairwise Markov property w.r.t. the partial correlation graph* $G = (\{1, ..., p\}, E)$, if $\{i, j\} \notin E \Rightarrow X_i \perp X_j \bullet \boldsymbol{X}_{\setminus\{i,j\}}$, that is, the edge set of the PCG of $\boldsymbol{X}$ is a subset of E. $\boldsymbol{X}$ is said to satisfy the *global Markov property w.r.t. the partial correlation graph G*, if, for $\alpha, \beta, \gamma \subset V$, "$\gamma$ separates α and β" implies $\boldsymbol{X}_\alpha \perp \boldsymbol{X}_\beta \bullet \boldsymbol{X}_\gamma$. The graph is constructed from the pairwise Markov property, but can be interpreted in terms of the global Markov property.

2.2.3 The multivariate normal distribution and conditional independence

We want to make further assumptions on the distribution F of $\boldsymbol{X}$. A random vector $\boldsymbol{X} = (X_1, ..., X_p)$ is said to have a *regular p-variate normal* (or *Gaussian*) distribution, denoted by $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$, if it possesses a Lebesgue density of the form

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = (2\pi)^{-\frac{p}{2}} (\det \Sigma)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right\}, \qquad \boldsymbol{x} \in \mathbb{R}^p, \tag{2.6}$$

for some $\boldsymbol{\mu} \in \mathbb{R}^p$ and $\Sigma \in \mathscr{S}_p^+$. Then $\mathbb{E}\boldsymbol{X} = \boldsymbol{\mu}$ and $\mathrm{Var}(\boldsymbol{X}) = \Sigma$. The term *regular* refers to the positive definiteness of the variance matrix. We will only deal with regular normal distributions—which allow the density characterization given above—without necessarily stressing the regularity.
The multivariate normal (*MVN*) distribution is a well studied object, it is treated e.g. in Bilodeau and Brenner (1999) or any other book on multivariate statistics. Of the properties of the MVN distribution the following three are of particular interest to us. Let, as before, $\boldsymbol{X}$ be partitioned into $\boldsymbol{X} = (\boldsymbol{Y}, \boldsymbol{Z})$. Then we have:

(I) The (marginal) distribution of $\boldsymbol{Y}$ is $N_q(\boldsymbol{\mu}_\alpha, \Sigma_{\alpha,\alpha})$.

(II) $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are independent (in notation $\boldsymbol{Y} \perp \boldsymbol{Z}$) if and only if $\Sigma_{\alpha,\beta} = \boldsymbol{0}$ (which is equivalent to $K_{\alpha,\beta} = \boldsymbol{0}$).

(III) The conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{Z} = z$ is

$$N_q\left(\mathbb{E}\boldsymbol{Y} + \Sigma_{\alpha,\beta}\Sigma_{\beta,\beta}^{-1}(z - \mathbb{E}\boldsymbol{Z}),\ \Sigma_{\boldsymbol{Y}\bullet\boldsymbol{Z}}\right).$$

These fundamental properties of the MVN distribution can be proved by directly manipulating the density (2.6). We want to spare a few words about assertion (III). It can be phrased as to say, the multivariate normal model is closed under conditioning—just as (I) tells that it is closed under marginalizing. Moreover, (III) gives expressions for the conditional expectation and the conditional variance:

$$\mathbb{E}(\boldsymbol{Y}|\,\boldsymbol{Z}) = \hat{\boldsymbol{Y}}(\boldsymbol{Z}) \qquad \text{and} \qquad \mathrm{Var}(\boldsymbol{Y}|\,\boldsymbol{Z}) = \Sigma_{\boldsymbol{Y}\bullet\boldsymbol{Z}}.$$

In general, $\mathbb{E}(\boldsymbol{Y}|\,\boldsymbol{Z})$ and $\mathrm{Var}(\boldsymbol{Y}|\,\boldsymbol{Z})$ are random variables that can be expressed as functions of the conditioning variable $\boldsymbol{Z}$. Thus (III) tells us that in the MVN model $\mathbb{E}(\boldsymbol{Y}|\,\cdot)$ is a *linear* function, whereas $\mathrm{Var}(\boldsymbol{Y}|\,\cdot)$ is *constant*. Further, $\mathbb{E}(\boldsymbol{Y}|\,\boldsymbol{Z})$ is the best prediction of $\boldsymbol{Y}$ from $\boldsymbol{Z}$, in the sense that $\mathbb{E}\|\boldsymbol{Y} - h(\boldsymbol{Z})\|^2$ is uniquely minimized by $h = \hat{\boldsymbol{Y}}(\cdot)$ among *all* measurable functions h. Here this best prediction coincides with the best linear prediction $\hat{\boldsymbol{Y}}(\boldsymbol{Z})$ given in (2.3). Finally, $\mathrm{Var}(\boldsymbol{Y}|\,\boldsymbol{Z})$ being constant means that the accuracy gain for predicting $\boldsymbol{Y}$ that we get from knowing $\boldsymbol{Z}$ is the same no matter what value $\boldsymbol{Z}$ takes on. It is not least this linearity of the MVN distribution that makes it very appealing for statistical modelling.
The occupation with the conditional distribution is guided by our interest in conditional independence, which is—although it has not been mentioned yet—the actual primary object of study in graphical models. Let, as in Sect. 2.2.1, $\boldsymbol{Y} = (\boldsymbol{Y}_1, \boldsymbol{Y}_2)$ be further partitioned. $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ are *conditionally independent given* $\boldsymbol{Z}$—in writing: $\boldsymbol{Y}_1 \perp \boldsymbol{Y}_2 |\, \boldsymbol{Z}$—if the conditional distribution of $(\boldsymbol{Y}_1, \boldsymbol{Y}_2)$ given $\boldsymbol{Z} = z$ is for (almost) all $z \in \mathbb{R}^r$ a product measure with independent margins corresponding to $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$. If $\boldsymbol{X}$ possesses a density $f_{\boldsymbol{X}} = f_{(\boldsymbol{Y}_1, \boldsymbol{Y}_2, \boldsymbol{Z})}$ w.r.t. some σ-finite measure, conditional independence admits

the following characterization: $\boldsymbol{Y}_1 \perp \boldsymbol{Y}_2 | \boldsymbol{Z}$ if and only if there exist functions $g : \mathbb{R}^{q_1+r} \to \mathbb{R}$ and $h : \mathbb{R}^{q_2+r} \to \mathbb{R}$ such that

$$f_{(\boldsymbol{Y}_1,\boldsymbol{Y}_2,\boldsymbol{Z})}(\boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{z}) = g(\boldsymbol{y}_1, \boldsymbol{z})h(\boldsymbol{y}_2, \boldsymbol{z}) \qquad \text{for almost all } (\boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{z}) \in \mathbb{R}^p.$$

This factorization criterion ought to be compared to its analogue for (marginal) independence. It shall serve as a definition here, saving us a proper introduction of the terms *conditional distribution* or *conditional density*.

We can construct for any random variable $\boldsymbol{X}$ in $\mathbb{R}^p$ a *conditional independence graph* (*CIG*) in an analogous way as before the partial correlation graph: We put an edge between nodes i and j unless $X_i \perp X_j | \boldsymbol{X}_{\setminus\{i,j\}}$. Then, for "nice" distributions F—for instance, if F has a continuous, strictly positive density f w.r.t. some σ-finite product measure—we have in analogy to Theorem 2.2.7 a separation property for CIGs: $\boldsymbol{X}_\alpha \perp \boldsymbol{X}_\beta | \boldsymbol{X}_\gamma$ if and only if γ separates α and β in the CIG of $\boldsymbol{X}$.

Assertions (I) to (III) are the link from conditional independence to the analysis of the second moment characteristics in Sect. 2.2.1. A direct consequence is:

Proposition 2.2.8 *If* $\boldsymbol{X} = (\boldsymbol{Y}_1, \boldsymbol{Y}_2, \boldsymbol{Z}) \sim N_p(\boldsymbol{\mu}, \Sigma)$, $\Sigma \in \mathscr{S}_p^+$, *then*

$$\boldsymbol{Y}_1 \perp \boldsymbol{Y}_2 \bullet \boldsymbol{Z} \quad \Longleftrightarrow \quad \boldsymbol{Y}_1 \perp \boldsymbol{Y}_2 | \boldsymbol{Z}.$$

In other words, the PCG and the CIG of a regular normal vector coincide. It must be emphasized that this is a particular property of the Gaussian distribution. Conditional independence and partial uncorrelatedness are generally different, cf. Baba et al. (2004), and so are the respective association graphs.

2.3 Gaussian graphical models

We have defined the partial correlation graph of a random vector and have recalled some properties of the multivariate normal distribution. We have thus gathered the ingredients we need to deal with Gaussian graphical models.

We understand a *graphical model* as a family of probability distributions on $\mathbb{R}^p$ satisfying the pairwise zero partial correlations specified by a given (undirected) graph $G = (V, E)$, i.e. for all $i, j \in V$

$$\{i, j\} \notin E \;\Rightarrow\; \varrho_{i,j\bullet} = 0. \tag{2.7}$$

If the model consists of all (regular) p-variate normal distributions satisfying (2.7) we call it a *Gaussian graphical model* (*GGM*). Another equivalent term is *covariance selection model*, originated by Dempster (1972).

We write $\mathscr{M}(G)$ to denote the GGM induced by the graph G. The model $\mathscr{M}(G)$ is called *saturated* if G is complete. It is called *decomposable* if the graph is decomposable. A Gaussian graphical model is a parametric family, which may be succinctly described as follows. Let $\mathscr{S}_p^+(G)$ be the subset of $\mathscr{S}_p^+$ consisting of all positive definite matrices with zero entries at the positions specified by G, i.e.

$$K \in \mathscr{S}_p^+(G) \;\Longleftrightarrow\; K \in \mathscr{S}_p^+ \text{ and } k_{i,j} = 0 \text{ for } i \neq j \text{ and } \{i, j\} \notin E.$$

Then

$$\mathscr{M}(G) = \left\{ N_p(\boldsymbol{\mu}, \Sigma) \,\middle|\, \boldsymbol{\mu} \in \mathbb{R}^p,\ K = \Sigma^{-1} \in \mathscr{S}_p^+(G) \right\}. \tag{2.8}$$

In the context of GGMs it is more convenient to parametrize the normal model by $(\boldsymbol{\mu}, K)$, which may be less common, but is quite intuitive considering that K directly appears in the density formula (2.6). The GGM $\mathscr{M}(G)$ is also specified by its parameter space $\mathbb{R}^p \times \mathscr{S}_p^+(G)$.
The term *graphical modelling* refers to the statistical task of deciding on a graphical model for given data and the collection of the statistical methods employed toward this end. Within the parametric family of Gaussian graphical models we have the powerful maximum likelihood theory available. We continue by stating the maximum likelihood estimates and some of their properties (Sect. 2.3.1), then review the properties of the likelihood ratio test for comparing two nested models (Sect. 2.3.2) and finally describe some model selection procedures (Sect. 2.3.3).

2.3.1 Estimation

Suppose we have i.i.d. observations $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ sampled from the normal distribution $N_p(\boldsymbol{\mu}, \Sigma)$ with $\Sigma \in \mathscr{S}_p^+$. Let furthermore $\mathbb{X}_n = (\boldsymbol{X}_1, ..., \boldsymbol{X}_n)^T$ be the $n \times p$ data matrix containing the data points as rows. We will make use of the following matrix notation. For an undirected graph $G = (V, E)$ and an arbitrary square matrix A define the matrix $A(G)$ by

$$[A(G)]_{i,j} = \begin{cases} a_{i,j} & \text{if } i = j \text{ or } \{i, j\} \in E, \\ 0 & \text{if } i \neq j \text{ and } \{i, j\} \notin E. \end{cases}$$

The saturated model We start with the saturated model, i.e. there is no further restriction on K. The main quantities of interest in Gaussian graphical models are the concentration matrix K and the partial correlation matrix P. Their computation ought to be part of any initial explorative data analysis. Both are functions of the covariance matrix Σ, thus we start with the latter.

Proposition 2.3.1 *If $n > p$, the maximum likelihood estimator (MLE) of Σ in the multivariate normal model (with unknown location $\boldsymbol{\mu}$) is*

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^T = \frac{1}{n} \mathbb{X}_n^T H_n \mathbb{X}_n,$$

where $H_n = I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$ is an idempotent matrix of rank $n - 1$. The MLEs of K and P are $\hat{K} = \hat{\Sigma}^{-1}$ and $\hat{P} = -\mathrm{Corr}(\hat{K})$, respectively.

Apparently $\mathbb{X}_n^T H_n \mathbb{X}_n$ has to be non-singular in order to be able to compute $\hat{K}$ and $\hat{P}$. It should be noted that this is also necessary for the MLE to exist in the sense that the ML equations have a unique solution. If n is strictly larger than p, this is almost surely true, but never if $n \leq p$.
We want to review some properties of these estimators. The strong law of large numbers, the continuous mapping theorem, the central limit theorem and the delta method yield the following asymptotic results, see also Propositions 3.3.3 and 3.4.2.

Proposition 2.3.2 *In the MVN model $\hat{\Sigma}$, $\hat{K}$ and $\hat{P}$ are strongly consistent estimators of Σ, K and P, respectively. Furthermore,*

(1) $\sqrt{n}\,\mathrm{vec}\left(\hat{\Sigma} - \Sigma\right) \xrightarrow{\mathscr{L}} N_{p^2}\left(\mathbf{0}, 2M_p(\Sigma \otimes \Sigma)\right)$,

(2) $\sqrt{n}\,\mathrm{vec}\left(\hat{K} - K\right) \xrightarrow{\mathscr{L}} N_{p^2}\left(\mathbf{0}, 2M_p(K \otimes K)\right)$,

(3) $\sqrt{n}\,\mathrm{vec}\left(\hat{P}-P\right) \xrightarrow{\mathscr{L}} N_{p^2}\left(\mathbf{0}, 2\Gamma M_p(K\otimes K)\Gamma^T\right)$,
where $\Gamma = (K_D^{-\frac{1}{2}} \otimes K_D^{-\frac{1}{2}}) + M_p(P \otimes K_D^{-1})J_p$.

Since the normal distribution and the empirical covariance matrix are of such utter importance, the exact distribution of the MLEs has also been the subject of study.

Proposition 2.3.3 *In the MVN model, if $n > p$, $\hat{\Sigma}$ has a Wishart distribution with parameter $\frac{1}{n}\Sigma$ and $n-1$ degrees of freedom, for which we use the notation $\hat{\Sigma} \sim W_p(n-1, \frac{1}{n}\Sigma)$.*

For a definition and properties of the Wishart distribution see e.g. Bilodeau and Brenner (1999), Chap. 7, or Srivastava and Khatri (1979), Chap. 3. It is also treated in most textbooks on multivariate statistics. The distribution of $\hat{K}$ is then called an *inverse Wishart distribution*. Of the various results on Wishart and related distributions we want to name the following three, but remark that more general results are available.

Proposition 2.3.4 *In the MVN model with $n > p$ we have*

(1) $\mathbb{E}\hat{\Sigma} = \frac{n-1}{n}\Sigma$ *and*

(2) $\mathrm{Var}(\mathrm{vec}\,\hat{\Sigma}) = \frac{2}{n}M_p(\Sigma\otimes\Sigma)$.

(3) If furthermore $\varrho_{i,j\bullet} = 0$, then

$$\sqrt{n-p}\frac{\hat{\varrho}_{i,j\bullet}}{\sqrt{1-\hat{\varrho}^2_{i,j\bullet}}} \sim t_{n-p}, \quad \text{which implies} \quad \hat{\varrho}^2_{i,j\bullet} \sim \mathrm{Beta}\left(\frac{1}{2}, \frac{n-p}{2}\right),$$

where t_{n-p} denotes Student's t-distribution with $n-p$ degrees of freedom and Beta(c,d) the beta distribution with parameters $c, d > 0$ and density

$$b(x) = \frac{\Gamma(c+d)}{\Gamma(c)\Gamma(d)}x^{c-1}(1-x)^{d-1}\mathbb{1}_{[0,1]}(x).$$

The last assertion (3) ought to be compared to the analogous results for the empirical correlation coefficient $\hat{\varrho}_{i,j} = \hat{\sigma}_{i,j}/\sqrt{\hat{\sigma}_{i,j}\hat{\sigma}_{j,j}}$: if the true correlation is zero, then

$$\sqrt{n-2}\frac{\hat{\varrho}_{i,j}}{\sqrt{1-\hat{\varrho}^2_{i,j}}} \sim t_{n-2} \qquad \text{and} \qquad \hat{\varrho}^2_{i,j} \sim \mathrm{Beta}\left(\frac{1}{2}, \frac{n-2}{2}\right).$$

Estimation under a given graphical model We have dealt so far with unrestricted estimators of Σ, K and the partial correlation matrix P. Small absolute values of the estimated partial correlations suggest that the corresponding true partial correlations may be zero. However assuming a non-saturated model, using unrestricted estimates for the remaining parameters is no longer optimal. The estimation efficiency generally decreases with the number of parameters to estimate. Also, for stepwise model selection procedures, as described in Sect. 2.3.3, which successively compare the appropriateness of different GGMs, estimates under model constraints are necessary.
Consider the graph $G = (V, E)$ with $|V| = p$ and $|E| = m$, and let $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ be an i.i.d. sample from the model $\mathscr{M}(G)$ given in (2.8). Keep in mind that K is then an element of the $(m+p)$-dimensional vector space $\mathscr{S}_p(G)$, where m may range from 0 to $p(p-1)/2$. The matrix Σ is fully determined by the $m+p$ values $k_{1,1}, ..., k_{p,p}$ and $k_{i,j}$, $\{i,j\} \in E$ (which have to meet the further restriction that K is positive definite) and in this sense has to be regarded as an $(m+p)$-dimensional object.

Theorem 2.3.5

(1) The ML estimate $\hat{\Sigma}_G$ of Σ in the model $\mathscr{M}(G)$ exists if $\hat{\Sigma} = \frac{1}{n}\mathbb{X}_n^T H_n \mathbb{X}_n$ is positive definite, which happens with probability one if $n > p$.

(2) If the ML estimate $\hat{\Sigma}_G$ exists, it is the unique solution of the following system of equations

$$[\hat{\Sigma}_G]_{i,j} = \hat{\sigma}_{i,j}, \qquad \{i,j\} \in E \ \text{ or } \ i = j,$$

$$[\hat{\Sigma}_G^{-1}]_{i,j} = 0, \qquad \{i,j\} \notin E \ \text{ and } \ i \neq j,$$

which may be succinctly formulated as

$$\hat{\Sigma}_G(G) = \hat{\Sigma}(G) \qquad \text{and} \qquad \hat{K}_G = \hat{K}_G(G), \tag{2.9}$$

where $\hat{K}_G = \hat{\Sigma}_G^{-1}$.

This result follows from general maximum likelihood theory for exponential models. The key is to observe that a GGM is a regular exponential model, cf. Dempster (1972) or Lauritzen (1996), p. 133. It is important to note that, contrary to the saturated case, the positive definiteness of $\mathbb{X}_n^T H_n \mathbb{X}_n$ is sufficient but not necessary. In a decomposable model, for instance, it suffices that n is larger than the number of nodes of the largest clique, cf. Proposition 2.3.6. Generally this condition is necessary but not sufficient. Details on stricter conditions on the existence of the ML estimate in the general case can be found in Buhl (1993) or Lauritzen (1996), p. 148.
Theorem 2.3.5 gives instructive information about the structure of $\hat{\Sigma}_G$, in particular, that it is a function of the sample covariance matrix $\hat{\Sigma}$. The relation between $\hat{\Sigma}_G$ and $\hat{\Sigma}$ is specified by (2.9), and Theorem 2.3.5 states furthermore that these equations always have a unique solution $\hat{\Sigma}_G$, if $\hat{\Sigma}$ is positive definite. What remains unclear is how to compute $\hat{\Sigma}_G$ from $\hat{\Sigma}$. This is accomplished by the *iterative proportional scaling* (*IPS*) algorithm, sometimes also referred to as *iterative proportional fitting*, which is explained in the following.

Iterative proportional scaling The IPS algorithm generally solves the problem of fitting a multivariate density that obeys a given interaction structure to specified marginal densities. Another application is the computation of the ML estimate in log-linear models, i.e. graphical models for discrete data. In the statistical literature the IPS algorithm can be traced back to at least Deming and Stephan (1940). In the case of multivariate normal densities the IPS procedure comes down to an iterative matrix manipulation. The IPS algorithm for GGMs, as it is described in the following, is mainly due to Speed and Kiiveri (1986).
Suppose we are given a graph G with cliques $\gamma_1, ..., \gamma_c$ and an unrestricted ML estimate $\hat{\Sigma} \in \mathscr{S}_p$. Then define for every clique γ the following matrix operator $T_\gamma : \mathscr{S}_p \to \mathscr{S}_p$:

$$T_\gamma(K) = K + \left[(\hat{\Sigma}_{\gamma,\gamma})^{-1}\right]^p - \left[(K^{-1})_{\gamma,\gamma}^{-1}\right]^p.$$

The operator T_γ has the following properties:

(I) If $K \in \mathscr{S}_p^+(G)$, then so is $T_\gamma K$.

(II) $(T_\gamma K)_{\gamma,\gamma}^{-1} = \hat{\Sigma}_{\gamma,\gamma}$, i.e. if the updated matrix $T_\gamma K$ is the concentration matrix of a random vector, $\boldsymbol{X}$ say, then $\boldsymbol{X}_\gamma$ has covariance matrix $\hat{\Sigma}_{\gamma,\gamma}$.

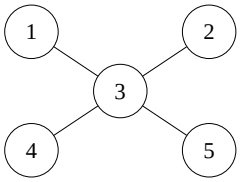

Figure 2.2: example graph

Apparently T_γ preserves the zero pattern of G. That it also preserves the positive definiteness and assertion (II) is not as straightforward, but both can be deduced by applying (2.2) to K^{-1}, cf. Lauritzen (1996), p. 135. The IPS algorithm then goes as follows: choose any $K_0 \in \mathscr{S}_p^+$, for instance $K_0 = I_p$, and repeat

$$K_{n+1} = T_{\gamma_1} T_{\gamma_2} ... T_{\gamma_c} K_n$$

until convergence is reached. If the ML estimate $\hat{\Sigma}_G$ exists (for which $\hat{\Sigma} \in \mathscr{S}_p^+$ is sufficient but not necessary), then (K_n) converges to $\hat{K}_G = \hat{\Sigma}_G^{-1}$, where $\hat{\Sigma}_G$ is the solution of (2.9), see again Lauritzen (1996), p. 135. Thus the IPS algorithm cycles through the cliques of G, in each step updating the concentration matrix K such that the clique has marginal covariance $\hat{\Sigma}_{\gamma,\gamma}$ while preserving the zero pattern specified by G.

Decomposable models In the case of decomposable models the ML estimate can be given in explicit form, and we do not have to resort to iterative approximations. As a decomposable graph can be *decomposed* into its cliques, the ML estimate of a decomposable model can be *composed* from the (unconstrained) MLEs of the cliques. Let $G = (V, E)$ be a decomposable graph with cliques $\gamma_1, ..., \gamma_c$ and $c > 1$. Define the sequence $(\delta_1, ..., \delta_{c-1})$ of successive intersections by

$$\delta_k = (\gamma_1 \cup ... \cup \gamma_k) \cap \gamma_{k+1}, \qquad k = 1, ..., c-1.$$

We assume that the numbering $\gamma_1, ..., \gamma_c$ is such that for every $k \in \{1, ..., c-1\}$ there is a $j \leq k$ with $\delta_k \subseteq \gamma_j$. It is always possible to order the cliques of a decomposable graph in such a way, cf. Lauritzen (1996), p. 18. The sequence $(\gamma_1, ...\gamma_c)$ is then said to be *perfect*, and it corresponds to a reversed sequence of successive decompositions. The δ_k do not have to be distinct. For instance, the graph in Fig. 2.2 has four cliques and, for any numbering of the cliques, $\delta_i = \{3\}$, $i = 1, 2, 3$.

Proposition 2.3.6

(1) The ML estimate $\hat{\Sigma}_G$ of Σ in the decomposable model $\mathscr{M}(G)$ exists with probability one if and only if $n > \max_{k=1,...,c} |\gamma_k|$.

(2) If the ML estimate $\hat{\Sigma}_G = \hat{K}_G$ exists, then it is given by

$$\hat{K}_G = \sum_{k=1}^{c} \left[(\hat{\Sigma}_{\gamma_k,\gamma_k})^{-1} \right]^p - \sum_{k=1}^{c-1} \left[(\hat{\Sigma}_{\delta_k,\delta_k})^{-1} \right]^p .$$

See Lauritzen (1996), p. 146, for a proof. Results on the asymptotic distribution of the restrained ML-estimator $\hat{\Sigma}_G$ in the decomposable as well as the general case can be found in Lauritzen (1996), Chap. 5. The exact, non-asymptotic distribution of $\hat{\Sigma}_G$ has also been studied. For decomposable G, it is known as the *hyper Wishart distribution* (Dawid and Lauritzen, 1993), and the distribution of $\hat{K}_G$ as *inverse hyper Wishart distribution* (Roverato, 2000).

2.3.2 Testing

We want to test a graphical model against a larger one, possibly but not necessarily the saturated model. Consider two graphs $G = (V, E)$ and $G_0 = (V, E_0)$ with $E_0 \subset E$, or equivalently $\mathscr{M}(G_0) \subset \mathscr{M}(G)$. Then the likelihood ratio for testing $\mathscr{M}(G_0)$ against the larger model $\mathscr{M}(G)$ based on the observation $\mathbb{X}_n$ reduces to

$$\mathrm{LR}(G_0, G) = \left(\frac{\det \hat{\Sigma}_G}{\det \hat{\Sigma}_{G_0}} \right)^{\frac{n}{2}},$$

small values of which suggest to dismiss $\mathscr{M}(G_0)$ in favour of $\mathscr{M}(G)$. It follows by the general theory for LR tests that the test statistic

$$D_n(G_0, G) = -2 \ln \mathrm{LR}(G_0, G) = n \left(\ln \det \hat{\Sigma}_{G_0} - \ln \det \hat{\Sigma}_G \right) \tag{2.10}$$

is asymptotically χ^2 distributed with $|E| - |E_0|$ degrees of freedom under the model $\mathscr{M}(G_0)$. The test statistic D_n may be interpreted as a measure of how much the appropriateness of model $\mathscr{M}(G_0)$ for the data deviates from that of $\mathscr{M}(G)$. It is thus also referred to as *deviance* and the LR test in GGMs is called *deviance test*.
It has been noted that the asymptotic χ^2 approximation of the distribution of D_n is generally not very accurate for small n. Several suggestions have been made on how to improve the finite sample approximation. One approach is to apply the Bartlett correction to the LR test statistic (Porteous, 1989). Another approximation, which is considerably better than the asymptotic distribution, is given by the exact distribution for decomposable models in Proposition 2.3.7 (Eriksen, 1996).

Decomposable models Again decomposable models play a special role. We are able to give the exact distribution of the deviance if both models compared are decomposable. Thus assume in the following that G and G_0 are decomposable. Then one can find a sequence of decomposable models $G_0 \subset G_1 \subset ... \subset G_k = G$ such that each successive pair (G_{i-1}, G_i) differs by exactly one edge e_i, $i = 1, ..., k$, cf. Lauritzen (1996), p. 20. Let a_i denote the number of common neighbours of both endpoints of e_i in the graph G_i.

Proposition 2.3.7 *If G_0 and G are decomposable and $G_0 \subset G$, then*

$$\frac{\det \hat{\Sigma}_G}{\det \hat{\Sigma}_{G_0}} = \exp\left(-\frac{D_n}{n} \right) \sim B_1 B_2 ... B_k,$$

where the B_i are independent random variables with $B_i \sim \mathrm{Beta}\left(\frac{n-a_i-2}{2}, \frac{1}{2} \right)$.

Since a complete graph and a graph with exactly one missing edge are both decomposable, the test of conditional independence of two components of a random vector is a special case of Proposition 2.3.7. If we let G_0 be the graph with all edges but $\{i, j\}$, some matrix calculus yields (cf. Lauritzen (1996), p. 150)

$$\frac{\det \hat{\Sigma}}{\det \hat{\Sigma}_{G_0}} = 1 - \hat{\varrho}^2_{i,j\bullet}.$$

By Proposition 2.3.7 this has a Beta$(\frac{n-p}{2}, \frac{1}{2})$ distribution, which is in concordance with Proposition 2.3.4 (3).

2.3.3 Model Selection

Contrary to estimation and statistical testing in GGMs there is no generally agreed-upon, optimal way to select a model. Statistical theory gives a relatively precise answer to the question if a certain model fits the data or not, but not which model to choose among those that fit. There are many model selection procedures (MSPs), and comparing them is rather difficult, since many aspects play a role—computing time being just one of them. Furthermore, theoretic results are usually hard to derive. For most MSPs, consistency can be shown, but distributional results are seldom available. Selecting a graphical model means to decide, based on the data, which partial correlations should be set to zero and which should be estimated freely. This decision, of course, heavily depends on the nature of the problem at hand, for example, if too few or too many edges are judged more severe. Ultimately, the choice of the MSP is a matter of personal taste, and the model selection has to be tailored to the specific situation. Expert knowledge should be incorporated to obtain sensible and interpretable models, especially when it comes to choosing from several equally adequate models.
The total number of p-dimensional GGMs is $2^{\binom{p}{2}}$, and only for very small p an evaluation of all possible models, based on some model selection criterion like AIC or BIC, is feasible. With respect to interpretability one might want to restrict the search space to decomposable models, cf. e.g. Whittaker (1990), Chap. 12, or Edwards (2000), Chap. 6. Otherwise a non-complete model search is necessary.

Model search The system of all possible models possesses itself a (directed) graph structure, corresponding to the partial ordering induced by set inclusion of the respective edge sets. A graph G_0, say, is a child of a graph G, if G has exactly one edge more than G_0. The fact that we know how to compare nested models, as described in Sect. 2.3.1, suggests a search along the edges of this lattice. A classic, simple search, known as *backward elimination*, is carried out as follows. Start with the saturated model, and in each step remove one edge. To determine which edge, compute all deviances between the current model and all models with exactly one edge less. The edge corresponding to the smallest deviance difference is deleted, unless all deviances are above the significance level, i.e. all edges are significant. Then the algorithm stops. The search in the opposite direction, starting from the empty graph and including significant edges, is also possible and known as *forward selection*. Although both schemes have been reported to produce similar results, there is a substantial conceptual difference that favours backward elimination. The latter searches among models consistent with the data, while forward selection steps through inconsistent models. The result of an LR test has no sensible interpretation if both models compared are actually invalid. On the other hand, forward selection is to be preferred for sparse graphs.
Of course, many variants exist, e.g., one may remove all non-significant edges at once, then successively include edges again, apply an alternative stopping rule (e.g. overall deviance against the saturated model) or generally alternate between elimination and selection steps. Another model search strategy in graphical models is known as the Edwards-Havránek procedure (Edwards and Havránek (1985, 1987), Smith (1992)). It is a global search, but reduces the search space, similar to the branch-and-bound principle by making use of the lattice structure.

One step model selection The simplicity of a one step MSP is, of course, very appealing. They become increasingly desirable as there has been an enormous growth in the dimensionality of data sets, and several proposals have been made in the recent past (Drton and Perlman, 2004, 2008; Meinshausen and Bühlmann, 2006; Castelo and Roverato, 2006). For instance, the SINful procedure by Drton and Perlman (2008) is a simple model selection scheme, which consists of setting all partial correlations to zero for which the absolute value of the sample partial correlation is below a certain threshold. This

threshold is determined in such a way that the overall probability of selecting incorrect edges, i.e. the probability that the estimated model is too large, is controlled.

2.4 Robustness

Most of what has been presented in the previous section, the classical GGM theory, has been developed in the seventies and the eighties of the last century. Since then graphical models have become popular tools of data analysis, and the statistical theory of Gaussian graphical models remains an active field of research. Many authors have in particular addressed the $n < p$ problem (a weak point of the ML theory) as in recent years one often encounters huge data sets, where the number of variables exceeds by far the number of observations. Another line of research considers GGMs in the Bayesian framework. It is beyond the scope of a book chapter to give an exhaustive survey of the recent approaches—even if we restrict ourselves to undirected graphical models for continuous data. We want to focus on another weak point of the normal ML theory: its lack of robustness, which has been pointed out, e.g., by Kuhnt and Becker (2003) and Gottard and Pacillo (2007).
Robustness denotes the property of a statistical method to yield good results also if the assumptions for which it is designed are violated. Small deviations from the assumed model shall have only a small effect, and robustness can be seen as a continuity property. This includes the often implied meaning of robustness as invulnerability against outliers. For example, any neighbourhood of a normal distribution (measured in the Kolmogorov metric) contains arbitrarily heavy-tailed distributions (measured in kurtosis, say). Outlier generating models with a small outlier fraction are actually very *close* to the pure data model.
There are two general conceptual approaches when it comes to robustifying a statistical analysis: identify the outliers and remove them, or use robust estimators that preferably nullify, but at least reduce the harmful impact of outliers. Graphical modelling—as an instance of the model selection problem—is a field where the advantages of the second approach become apparent. In its most general perception an outlier is a "very unlikely" observation under a given model, cf. Davies and Gather (1993). Irrespective of the particular rule applied to decide whether an observation is deemed an outlier or not, any sensible rule ought to give different answers for different models. An outlier in a specific GGM may be a quite likely observation in the saturated model.
This substantially complicates outlier detection in any type of graphical models, suggesting it must at least be applied iteratively, alternating with model selection steps. For Gaussian graphical models, however, we have the relieving fact that an outlier w.r.t. a normal distribution basically coincides with an *outlier* in its literal meaning: a point far away from the majority of the data. Hence, strongly outlying points tend to be ouliers w.r.t. any Gaussian model, no matter which—if any—conditional or marginal independences it obeys.
Our focus will therefore lie in the following on robust estimation. Note that Gaussian graphical modelling, as presented in the previous section, exclusively relies on $\hat{\Sigma}$. It is a promising approach to replace the initial estimate $\hat{\Sigma}$ by a robust substitute and hence robustify all subsequent analysis. We can make use of the well developed robust estimation theory of multivariate scatter.

2.4.1 Robust estimation of multivariate scatter

Robust estimation in multivariate data analysis has long been recognized as a challenging task. Over the last four decades much work has been devoted to the problem and many robust alternatives of the sample mean and the sample covariance matrix have been proposed, e.g. M-estimators (Maronna,

1976; Tyler, 1987a), Stahel-Donoho estimators (Stahel, 1981; Donoho, 1982; Maronna and Yohai, 1995; Gervini, 2002), S-estimators (Davies, 1987; Lopuhaä, 1989; Rocke, 1996), MVE and MCD (Rousseeuw, 1985; Davies, 1992; Butler et al., 1993; Croux and Haesbroeck, 1999; Rousseeuw and Van Driessen, 1999), τ-estimators (Lopuhaä, 1991), CM-estimators (Kent and Tyler, 1996), reweighted and and data-depth based estimators (Lopuhaä, 1999; Gervini, 2003; Zuo and Cui, 2005). Many variants exist, and the list is far from complete. For a more detailed account see e.g. the book Maronna et al. (2006) or the review article Zuo (2006).

The asymptotics and robustness properties of the estimators are to a large extent well understood. The computation often requires to solve challenging optimization problems, but improved search heuristics are nowadays available. What remains largely an open theoretical question is the exact distribution for small samples. Constants of finite sample approximations usually have to be assessed numerically.

There are several measures that quantify and thus allow to compare the robustness properties of estimators. We want to restrict our attention to the influence function, introduced by Hampel (1971). Toward this end we have to adopt the notion that estimators are functionals $S : \mathscr{F} \to \Theta$ defined on a class of distributions $\mathscr{F}$. In the case of matrix-valued scatter estimators S, the image space Θ is $\mathscr{S}_p$. The specific estimate computed from a data set $\mathbb{X}_n$ is the functional evaluated at the corresponding empirical distribution function $\mathbb{F}_n = \frac{1}{n}\sum_{i=1}^n \delta_{\boldsymbol{X}_i}$, where $\delta_{\boldsymbol{x}}$ denotes the Dirac-measure which puts unit mass at the point $\boldsymbol{x} \in \mathbb{R}^p$. For instance, the sample covariance matrix $\hat{\Sigma}$ is simply the functional $\mathrm{Var}(\cdot)$, which is defined on all distributions with finite second moments, evaluated at $\mathbb{F}_n$. The *influence function* of S at the distribution F is defined as

$$IF(\boldsymbol{x}; S, F) = \lim_{\varepsilon \searrow 0} \frac{1}{\varepsilon}\left(S(F_{\varepsilon,\boldsymbol{x}}) - S(F)\right), \qquad \boldsymbol{x} \in \mathbb{R}^p,$$

where $F_{\varepsilon,\boldsymbol{x}} = (1-\varepsilon)F + \varepsilon\delta_{\boldsymbol{x}}$. In words, the influence function is the directional derivative of the functional S at the "point" $F \in \mathscr{F}$ in the direction of $\delta_{\boldsymbol{x}} \in \mathscr{F}$. It describes the *influence* of an infinitesimal contamination at point $\boldsymbol{x} \in \mathbb{R}^p$ on the functional S, when the latter is evaluated at the distribution F. Of course, in terms of robustness, the influence of any contamination is preferably small. A robust estimator has in particular a bounded influence function, i.e. the maximal absolute influence $\sup\{\, \|IF(\boldsymbol{x}; S, F)\| \,|\, \boldsymbol{x} \in \mathbb{R}^p\}$, also known as *gross-error sensitivity*, is finite.

The influence function is said to measure the *local robustness* of an estimator. Another important robustness measure, which in contrast measures the global robustness but which we will not pursue further here, is the *breakdown point* (asymptotic breakdown point (Hampel, 1971), finite-sample breakdown point (Donoho and Huber, 1983)), see also Davies and Gather (2005). Roughly, the finite-sample replacement breakdown point is the minimal fraction of contaminated data points that can drive the estimate to the boundary of the parameter space. For details on robustness measures see e.g. Hampel et al. (1986).

It is a very desirable property of scatter estimators to transform in the same way as the (population) covariance matrix—the quantity they aim to estimate—under affine linear transformations. A scatter estimator $\hat{S}$ is said to be *affine equivariant*, if it satisfies $\hat{S}(\mathbb{X}_n A^T + \mathbf{1}_n \boldsymbol{b}^T) = A\hat{S}(\mathbb{X}_n)A^T$ for any full rank matrix $A \in \mathbb{R}^{p\times p}$ and vector $\boldsymbol{b} \in \mathbb{R}^p$. We want to make a notational distinction between S, the functional working on distributions, and $\hat{S}$, the corresponding estimator working on data (strictly speaking a series of estimators indexed by n), i.e. $S(\mathbb{F}_n) = \hat{S}(\mathbb{X}_n)$. The equivariance is indeed an important property, due to various reasons. For instance, any statistical analysis based on such estimators is independent of any change of the coordinate system, may it be re-scaling or rotations of the data. Also, affine equivariance implies that at any elliptical population distribution (such as the Gaussian distribution) indeed a multiple of the covariance matrix is unbiasedly estimated, cf. Proposition 2.4.2 below. Furthermore the estimate obtained is usually positive definite with probability one, which is

crucial for any subsequent analysis, e.g. we know that the derived partial correlation matrix estimator $-\text{Corr}(\hat{S}^{-1})$ actually reflects a "valid" dependence structure.
The classes of estimators listed above all possess this equivariance property—or at least the pseudo-equivariance described below. Historically though, affine equivariance for robust estimators is not a self-evident property. Contrary to univariate moment-based estimators (such as the sample variance), the highly robust quantile-based univariate scale estimators (such as the median absolute deviation, MAD) do not admit a straightforward affine equivariant generalization to higher dimensions.
In Gaussian graphical models we are interested in partial correlations and zero entries in the inverse covariance matrix, for which we need to know Σ only up to a constant. The knowledge of the overall scale is not relevant, and we require a slightly weaker condition than affine equivariance in the above sense, which we want to call *affine pseudo-equivariance* or *proportional affine equivariance*.

Condition C2.4.1 $\hat{S}(\mathbb{X}_n A^T + \mathbf{1}_n \boldsymbol{b}^T) = g(AA^T) A \hat{S}(\mathbb{X}_n) A^T$ *for* $\boldsymbol{b} \in \mathbb{R}^p$, $A \in \mathbb{R}^{p \times p}$ *with full rank, and* $g : \mathbb{R}^{p \times p} \to \mathbb{R}$ *satisfying* $g(I_p) = 1$.

This condition basically merges two important special cases, the proper affine equivariance described above, i.e. $g \equiv 1$, and the case of shape estimators in the sense of Paindaveine (2008), which corresponds to $g = \det(\cdot)^{-1/p}$. The following proposition can be found in a similar form in Bilodeau and Brenner (1999), p. 212.

Proposition 2.4.2 *In the MVN model, i.e.* $\mathbb{X}_n = (\boldsymbol{X}_1, ..., \boldsymbol{X}_n)^T$ *with* $\boldsymbol{X}_1, ..., \boldsymbol{X}_n \sim N_p(\boldsymbol{\mu}, \Sigma)$ *i.i.d., any affine pseudo-equivariant scatter estimator* $\hat{S} = \hat{S}(\mathbb{X}_n)$ *satisfies*

(1) $\mathbb{E}\hat{S} = a_n \Sigma$ *and*

(2) $\text{Var}(\text{vec}\, \hat{S}) = 2\, b_n\, M_p(\Sigma \otimes \Sigma) + c_n \,\text{vec}\,\Sigma(\text{vec}\,\Sigma)^T$,

where (a_n), (b_n) *and* (c_n) *are sequences of real numbers with* $a_n, b_n \geq 0$ *and* $c_n \geq -2b_n/p$ *for all* $n \in \mathbb{N}$.

Proposition 2.3.4 tells us that for $\hat{S} = \hat{\Sigma}$ we have $a_n = \frac{n}{n-1}$, $b_n = \frac{1}{n}$ and $c_n \equiv 0$. For root-n-consistent estimators the general form of the variance re-appears in the asymptotic variance, and they fulfil

Condition C2.4.3 *There exist constants* $a, b \geq 0$ *and* $c \geq -2b/p$ *such that*

$$\sqrt{n}\,\text{vec}(\hat{S} - a\,\Sigma) \xrightarrow{\mathscr{L}} N_{p^2}\left(\mathbf{0},\; 2a^2 b M_p(\Sigma \otimes \Sigma) + a^2 c\,\text{vec}\,\Sigma(\text{vec}\,\Sigma)^T\right).$$

The continuous mapping theorem and the multivariate delta method yield the general form of the asymptotic variance of any partial correlation estimator derived from a scatter estimator satisfying C2.4.3.

Proposition 2.4.4 *If* $\hat{S}$ *fulfils C2.4.3, then the partial correlation estimator* $\hat{P}^S = -\text{Corr}(\hat{S}^{-1})$ *satisfies*

$$\sqrt{n}\,\text{vec}(\hat{P}^S - P) \xrightarrow{\mathscr{L}} N_{p^2}(\mathbf{0}, 2b\Gamma M_p(K \otimes K)\Gamma^T), \tag{2.11}$$

where b *is the same as in Condition C2.4.3 and* Γ *is as in Proposition 2.3.2.*

Thus the comparison of the asymptotic efficiencies of partial correlation matrix estimators based on affine pseudo- equivariant scatter estimators reduces to the comparison of the respective values of the scalar b. For $\hat{S} = \hat{\Sigma}$ we have $b = 1$ by Proposition 2.3.2. Also, general results for the influence function of pseudo-equivariant estimators can be given, cf. Hampel et al. (1986), Chap. 5.3.

Proposition 2.4.5

(1) At the Gaussian distribution $F = N_p(\boldsymbol{\mu}, \Sigma)$ the influence function of any functional S satisfying Condition C2.4.1 has, if it exists, the form

$$IF(\boldsymbol{x}; S, F) = g(\Sigma)\left(\alpha(d(\boldsymbol{x}))(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T - \beta(d(\boldsymbol{x}))\Sigma\right), \tag{2.12}$$

where $d(\boldsymbol{x}) = \sqrt{(\boldsymbol{x}-\boldsymbol{\mu})^T K(\boldsymbol{x}-\boldsymbol{\mu})}$, g is as in Condition C2.4.1 and α and β are suitable functions $[0,\infty) \to \mathbb{R}$.

(2) Assuming that $\hat{S}$ is Fisher-consistent for $a\Sigma$, i.e. $S(F) = a\Sigma$, with $a > 0$, cf. Condition C2.4.3, the influence function of the corresponding partial correlation matrix functional $P^S = -\mathrm{Corr}(S^{-1})$ is

$$IF(\boldsymbol{x}; P^S, F) = \frac{\alpha(d(\boldsymbol{x}))g(\Sigma)}{a}\left(\frac{1}{2}\left(\Pi_D K_D^{-1} P + (\Pi_D K_D^{-1} P)^T\right) - K_D^{-\frac{1}{2}} \Pi K_D^{-\frac{1}{2}}\right),$$

where $\Pi = K(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T K$.

In the case of the sample covariance matrix $\hat{\Sigma}(\mathbb{X}_n) = \mathrm{Var}(\mathbb{F}_n)$ we have $a = 1$ and $\alpha = \beta \equiv 1$. Thus (2.12) reduces to $IF(\boldsymbol{x}; \mathrm{Var}, F) = (\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T - \Sigma$, which is not only unbounded, but even increases quadratically with $\|\boldsymbol{x}-\boldsymbol{\mu}\|$. We will now give two examples of robust affine equivariant estimators, that have been proposed in the context of GGMs.

The minimum covariance determinant (MCD) estimator The idea behind the MCD estimator is that outliers will increase the volume of the ellipsoid specified by the sample covariance matrix, which is proportional to the square root of its determinant. The MCD is defined as follows. A subset $\eta \subset \{1, ..., n\}$ of fixed size $h = \lfloor sn \rfloor$ with $\frac{1}{2} \leq s < 1$ is determined such that $\det(\hat{\Sigma}^\eta)$ with

$$\hat{\Sigma}^\eta = \frac{1}{h}\sum_{i\in\eta}(\boldsymbol{X}_i - \bar{\boldsymbol{X}}^\eta)(\boldsymbol{X}_i - \bar{\boldsymbol{X}}^\eta)^T \quad \text{and} \quad \bar{\boldsymbol{X}}^\eta = \frac{1}{h}\sum_{i\in\eta}\boldsymbol{X}_i$$

is minimal. The mean $\hat{\boldsymbol{\mu}}_{\mathrm{MCD}}$ and covariance matrix $\hat{\Sigma}_{\mathrm{MCD}}$ computed from this subsample are called the *raw MCD location*, respectively *scatter estimate*. Based on the raw estimate $(\hat{\boldsymbol{\mu}}_{\mathrm{MCD}}, \hat{\Sigma}_{\mathrm{MCD}})$ a reweighted scatter estimator $\hat{\Sigma}_{\mathrm{RMCD}}$ is computed from the whole sample:

$$\hat{\Sigma}_{\mathrm{RMCD}} = \left(\sum_{i=1}^n w_i\right)^{-1} \sum_{i=1}^n w_i(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}_{\mathrm{MCD}})(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}_{\mathrm{MCD}})^T,$$

where $w_i = 1$, if $(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}_{\mathrm{MCD}})^T \hat{\Sigma}_{\mathrm{MCD}}^{-1}(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}_{\mathrm{MCD}}) < r$ for some suitably chosen constant $r > 0$, and zero otherwise. Usually the the scatter estimate (reweighted as well as raw) is multiplied by a consistency factor (corresponding to $1/a$ in Condition C2.4.3) to achieve consistency for Σ at the MVN distribution. Since this is irrelevant for applications in GGMs we omit the details. The respective values of the constants b and c in Condition C2.4.3 as well as the function α and β in Proposition 2.4.5 are given in Croux and Haesbroeck (1999).
The reweighting step improves the efficiency and retains the high global robustness (breakdown point of roughly $1-s$ for $s \geq 1/2$) of the raw estimate. Although the minimization over $\binom{n}{h}$ subsets is of non-polynomial complexity, the availability of fast search heuristics (e.g. Rousseeuw and Van Driessen, 1999) along with the aforementioned good statistical properties have rendered the RMCD a very popular robust scatter estimator, and several authors (Becker, 2005; Gottard and Pacillo, 2010) have suggested its use for Gaussian graphical modelling.

The proposal by Miyamura and Kano Miyamura and Kano (2006) proposed another affine equivariant robust scatter estimator in the GGM framework. The idea is here a suitable adjustment of the ML equations. The Miyamura-Kano estimator $\hat{\Sigma}_{MK}$ falls into the class of M-estimators, as considered in Huber and Ronchetti (2009), and is defined as the scatter part Σ of the solution $(\boldsymbol{\mu}, \Sigma)$ of

$$\frac{1}{n}\sum_{i=1}^{n} \exp\left(-\frac{\xi\, d^2(\boldsymbol{X}_i)}{2}\right)(\boldsymbol{X}_i - \boldsymbol{\mu}) = \mathbf{0} \qquad \text{and}$$

$$\frac{1}{n}\sum_{i=1}^{n} \exp\left(-\frac{\xi\, d^2(\boldsymbol{X}_i)}{2}\right)\left(\Sigma - (\boldsymbol{X}_i - \boldsymbol{\mu})(\boldsymbol{X}_i - \boldsymbol{\mu})^T\right) = \frac{\xi}{(\xi+1)^{(p+2)/2}}\,\Sigma,$$

where $\xi \geq 0$ is a tuning parameter and $d(\boldsymbol{x})$ is, as in Proposition 2.4.5, the Mahalanobis distance of $\boldsymbol{x} \in \mathbb{R}^p$ w.r.t. $\boldsymbol{\mu}$ and Σ. Large values of ξ correspond to a more robust (but less efficient) estimate, i.e. less weight is given to outlying observations. The Gaussian likelihood equations are obtained for $\xi = 0$.

2.4.2 Robust Gaussian graphical modelling

The classical GGM theory is completely based on the sample covariance matrix $\hat{\Sigma}$: the ML estimates in Theorem 2.3.5, the deviance test statistic D_n in (2.10) and model selection procedures such as backward elimination, Edwards-Havránek or Drton-Perlman. Thus replacing the normal MLE by a robust, affine equivariant scatter estimator and applying the GGM methodology in analogous manner is an intuitive way of performing robust graphical modelling, insensitive to outliers in the data. Since the asymptotics of affine (pseudo-)equivariant estimators are well established (at the normal distribution), and, as described in Sect. 2.4.1, their general common structure is not much different from that of the sample covariance matrix, *asymptotic* statistical methods can rather easily be adjusted by means of standard asymptotic arguments.

Estimation under a given graphical model We have discussed properties of equivariant scatter estimators and indicated their usefulness for Gaussian graphical models. However they just provide alternatives for the unconstrained estimation. Whereas the ML paradigm dictates the solution of (2.9) as an optimal way of estimating a covariance matrix with a graphical model and exact normality, it is not quite clear what is the best way of robustly estimating a covariance matrix that obeys a zero pattern in its covariance. Clearly, Theorem 2.3.5 suggests to simply solve equations (2.9) with $\hat{\Sigma}$ replaced by any suitable robust estimator $\hat{S}$. This approach has the advantage that consistency of the estimator under the model is easily assessed. In case of a decomposable model the estimator can be computed by the decomposition of Proposition 2.3.6, or generally by the IPS algorithm, for which convergence has been shown and which comes at no additional computational cost. Becker (2005) proposed to apply IPS to the reweighted MCD.
However, a thorough study of scatter estimators under graphical models is still due, and it might be that other possibilities are more appropriate in certain situations. Many robust estimators are defined as the solution of a system of equations. A different approach is to alter these estimation equations in a suitable way that forces a zero pattern on the inverse. This requires a new algorithm, the convergence of which has to be assessed individually. This route has been taken by Miyamura and Kano (2006). Their algorithm performs an IPS approximation at each step and is hence relatively slow.
A problem remains with both strategies. Scatter estimators, if they have not a structure as simple as the sample covariance, generally do not possess the "consistency property" that the estimate of

a margin appears as a submatrix of the estimate of the whole vector. The ML estimate $\hat{\Sigma}_G$ in the decomposable as well as the general case is composed from the unrestricted estimates of the cliques, cf. Theorem 2.3.5 and Proposition 2.3.6, which makes it in particular possible to compute the MLE for $p \geq n$. One way to circumvent this problem is to drop the affine equivariance and resort to robust "pairwise" estimators, such as the Gnanadesikan-Kettenring estimator (Gnanadesikan and Kettenring, 1972; Maronna and Zamar, 2002) or marginal sign and rank matrices (Visuri et al., 2000), see also Section 5.2. Besides having the mentioned consistency property pairwise estimators are also very fast to compute.

Testing and model selection The deviance test can be applied analogously with minor adjustments when based on an affine equivariant scatter estimator. Similarly to the partial correlation estimator $\hat{P}^S$ in Proposition 2.4.4, the asymptotic distribution of the generalized deviance D_n^S, computed from any root-n-consistent, equivariant estimate $\hat{S}$, differs from that of the ML-deviance (2.10) only by a factor, see Tyler (1983) or Bilodeau and Brenner (1999), Chap. 13, for details. However, as noted in Sect. 2.3.2, the χ^2 approximation of the uncorrected deviance may be rather inaccurate for small n. Generalizations of finite-sample approximations or the exact test in Proposition 2.3.7 are not equally straightforward. Since the exact distribution of a robust estimator is usually unknown, one will have to resort to Monte Carlo or bootstrap methods.
Model selection procedures that only require a covariance estimate can be robustified in the same way. Besides the classical search procedures this is also true for the SINful procedure by Drton and Perlman (2008), of which Gottard and Pacillo (2010) studied a robustified version based on the RMCD.

2.4.3 Concluding remarks

The use of robust methods is strongly advisable, particularly in multivariate analysis, where the whole structure of the data is not immediately evident. Even if one refrains from relying solely on a robust analysis, it is in any case an important diagnostic tool. A single gross error or even mild deviations from the assumed model may render the results of a sample covariance based data analysis useless. The use of alternative, robust estimators provides a feasible safeguard, which comes at the price of a small loss in efficiency and a justifiable increase in computational costs.
Although there is an immense amount of literature on multivariate robust estimation and applications thereof (robust tests, regression, principal component analysis, discrimination analysis etc., see e.g. Zuo (2006) for references), the list of publications addressing robustness in graphical models is (still) rather short. We have described how GGMs can be robustified using robust, affine equivariant estimators. An in-depth study of this application of robust scatter estimation seems to be still open.
The main limitation of this approach is that it works well only for sufficiently large n, and on any account only for $n > p$, since, as pointed out above, usually an initial estimate of full dimension is required. Also note that, for instance, the computation of the MCD requires $h > p$. The finite-sample efficiency of many robust estimators is low, and with the exact distributions rarely accessible, methods based on such estimators rely even more on asymptotics than likelihood methods.
The processing of very high-dimensional data ($p >> n$) becomes increasingly relevant, and in such situations it is unavoidable and (even if n is sufficiently large) dictated by computational feasibility, to assemble the estimate of Σ, restricted to a given model, from marginal estimates. A high dimensional, robust graphical modelling, combining robustness with applicability in large dimensions, remains a challenging topic of future research.

Chapter 3

Elliptical graphical modelling — the decomposable case

Abstract. We propose *elliptical graphical models* as a generalization of Gaussian graphical models, also known as covariance selection models or concentration graph models, by letting the population distribution be elliptical instead of normal, allowing to fit data with arbitrarily heavy tails. We discuss the interpretation of an absent edge in the partial correlation graph of an elliptical distribution, which is equivalent to a zero-entry in the inverse of its shape matrix. We further study the class of proportionally affine equivariant scatter estimators and show how they can be used to perform elliptical graphical modelling, leading to a new class of partial correlation estimators and analogues of the classical deviance test. General expressions for the asymptotic variance of partial correlation estimators, unconstrained and under decomposable models, are given, and the asymptotic χ^2 approximation of the pseudo-deviance test statistic is proved. The feasibility of our approach is demonstrated by a simulation study, using, among others, Tyler's scatter estimator, which is distribution-free within the elliptical model. Our approach provides a robustification of Gaussian graphical modelling, which is likelihood-based and known to be very sensitive to model misspecifications and outlying observations.

3.1 Introduction and notation

Graphical modelling of continuous variables is almost exclusively based on the assumption of multivariate normality. This has two disadvantages: the assumption is not always met (for example, multivariate normality allows only linear dependencies among the variables), and the statistical tools are based on the normal likelihood and highly non-robust. Among the many ways that data may be non-Gaussian outliers pose a problem of particular gravity, due to two reasons: they frequently occur, may it be as contamination or as "valid" observations, and the normal likelihood methods (such as the sample covariance matrix) are particularly susceptible to outliers. Our objective is to deal with heavy-tailed data and to safeguard graphical modelling against the impact of faulty outliers. We restrict our attention to the basic yet important case where we have only continuous variables and want to model mutual dependence, rather than directed "influence", i.e we consider only undirected graphs. Traditionally, joint multivariate normality is assumed in this situation, and the statistical methodology goes under the name Gaussian graphical modelling. We propose the class of elliptical distributions as a more general data model and call our approach elliptical graphical modelling.

The lack of robustness of Gaussian graphical modelling has been noted by several authors before. Three proposals of a robust approach to Gaussian graphical modelling are known to us: Becker

(2005) and Gottard and Pacillo (2010) suggest to replace the sample covariance matrix by the reweighted MCD estimator, Miyamura and Kano (2006) propose to replace it by an M-estimator. A common feature of both estimators is affine equivariance. This article delivers a systematic and theoretically grounded treatment of the affine equivariant approach. We show that the sample covariance matrix may be substituted by basically any affine equivariant, root-n-consistent estimator. As long as ellipticity can be assumed, the classical Gaussian graphical modelling tools can be employed with simple adjustments. Thus the data analyst is free to choose the appropriate estimator, delivering the degree of robustness that seems necessary for the data situation at hand.

The paper is divided into seven sections. Section 2 defines elliptical graphical models. The subsequent Sections 3 on unconstrained estimation, 4 on constrained estimation and 5 on testing provide the basics of elliptical graphical modelling. Section 6 gives examples of affine equivariant estimators. Some deeper attention is paid to Tyler's M-estimator. Finally, in Section 7 we compare different estimators by means of simulation and bring the attention to some practical aspects. We summarize the article and discuss limitations of the approach. Proofs are deferred to the appendix.

We close this section by introducing some mathematical notation. We use $\sim$ for "distributed as", $\overset{\mathscr{L}}{=}$ for equality in distribution and $\overset{a}{\sim}$ for asymptotic equivalence, i.e. $X_n \overset{a}{\sim} Y_n \Leftrightarrow \|X_n - Y_n\| \xrightarrow{p} 0$. The symbol $\propto$ is used for "proportional to". Matrices are denoted by capital letters, the corresponding small letter is used for an element of the matrix, e.g., the $p \times p$ matrix P is the collection of all $p_{i,j}$, $i, j = 1, ..., p$. Alternatively, if matrices are denoted by more complicated compound symbols, e.g. if they carry subscripts already, square brackets will be used to refer to individual elements, e.g. $[\hat{S}_G^{-1}]_{i,j}$. Index sets are denoted by usually non-italic small Greek letters. Subvectors and submatrices are referenced by subscripts, e.g. for $\alpha, \beta \subseteq \{1, ..., p\}$ the $|\alpha| \times |\beta|$ matrix $S_{\alpha,\beta}$ is obtained from S by deleting all rows that are not in α and all columns that are not in β. Similarly, the $p \times p$ matrix $[S_{\alpha,\beta}]^p$ is obtained from S by putting all rows not in α and all columns not in β to zero. We want to view this matrix operation as two operations performed sequentially: first $(\cdot)_{\alpha,\beta}$ extracting the submatrix and then $[\cdot]^p$ writing it back on a "blank" matrix at the coordinates specified by α and β. Of course, the latter is not well defined without the former, but this allows us e.g. to write $[(S_{\alpha,\beta})^{-1}]^p$.

We adopt the general convention that subscripts have stronger ties than superscripts, for instance, we write $S^{-1}_{\alpha,\beta}$ for $(S_{\alpha,\beta})^{-1}$. Let $\mathscr{S}_p$ and $\mathscr{S}_p^+$ be the sets of all symmetric, respectively positive definite $p \times p$ matrices, and define A_D as the diagonal matrix having the same diagonal as $A \in \mathbb{R}^{p\times p}$. The Kronecker product $A \otimes B$ of two matrices $A, B \in \mathbb{R}^{p\times p}$ is defined as the $p^2 \times p^2$ matrix with entry $a_{i,j}b_{k,l}$ at position $((i-1)p+k, (j-1)p+l)$. Let $e_1, ..., e_p$ be the unit vectors in $\mathbb{R}^p$ and 1_p the p-vector consisting only of ones. Define further the following matrices:

$$J_p = \sum_{i=1}^{p} e_i e_i^T \otimes e_i e_i^T, \qquad K_p = \sum_{i=1}^{p}\sum_{j=1}^{p} e_i e_j^T \otimes e_j e_i^T \quad \text{and} \quad M_p = \frac{1}{2}\left(I_{p^2} + K_p\right),$$

where I_{p^2} denotes the $p^2 \times p^2$ identity matrix. K_p is also called the *commutation matrix*. Finally, let $\operatorname{vec} A$ be the p^2-vector obtained by stacking the columns of $A \in \mathbb{R}^{p\times p}$ from left to right underneath each other. More on these concepts and their properties can be found in Magnus and Neudecker (1999).

3.2 Elliptical graphical models

We introduce *elliptical graphical models*. Construction and terminology are in analogy to *Gaussian graphical models*. For details on the latter, also known as *covariance selection models* and *concen-*

tration graph models, see Whittaker (1990), Cox and Wermuth (1996), Lauritzen (1996) or Edwards (2000).
Consider the class $\mathscr{E}_p$ of all continuous, elliptical distributions on $\mathbb{R}^p$. A continuous distribution F on $\mathbb{R}^p$ is said to be *elliptical* if it has a Lebesgue-density f of the form

$$f(x) = \det(S)^{-\frac{1}{2}} g((x-\mu)^T S^{-1}(x-\mu)) \tag{3.1}$$

for some $\mu \in \mathbb{R}^p$ and symmetric, positive definite $p \times p$ matrix S. We call μ the *location* or *symmetry center* and S the *shape matrix* of F and denote the class of all continuous elliptical distributions on $\mathbb{R}^p$ having these parameters by $\mathscr{E}_p(\mu, S)$. A continuous distribution on $\mathbb{R}^p$ is called *spherical*, if it is elliptical with the shape matrix S proportional to the identity matrix.
In the parametrization (μ, S), the symmetry center μ is uniquely defined whereas the matrix S is unique only up to scale, that is, $\mathscr{E}_p(\mu, S) = \mathscr{E}_p(\mu, cS)$ for any $c > 0$. Some form of standardization can be imposed on S to uniquely define *the* shape matrix of an elliptical distribution. Several have been suggested in the literature, e.g., setting the trace of S to p or a specific element to 1. Paindaveine (2008) argues to choose $\det(S) = 1$.
Since the standardization of S is irrelevant for the following considerations, we will completely omit it. We understand the *shape* of an elliptical distribution as an equivalence class of positive definite random matrices being proportional to each other and call any matrix S satisfying (3.1) for a suitable function g a shape matrix of F. In the same manner we view its inverse $K = S^{-1}$, which we call *pseudo concentration matrix* of F. Let furthermore

$$h : \mathscr{S}_p^+ \to \mathscr{S}_p : A \mapsto -\left(A^{-1}\right)_D^{-\frac{1}{2}} A^{-1} \left(A^{-1}\right)_D^{-\frac{1}{2}}, \tag{3.2}$$

and $P = h(S)$. The function h is invariant to scale changes, i.e. P is a uniquely defined parameter of $F \in \mathscr{E}_p(\mu, S)$. The diagonal elements of P are equal to -1. If the second-order moments of $X \sim F \in \mathscr{E}_p(\mu, S)$ exist, then $\Sigma = \mathrm{var}(X)$ is proportional to S. Consequently, the element $p_{i,j}$ of P at position (i, j), $i \neq j$, is the partial correlation of X_i and X_j given the remaining components of X. For a definition and properties of partial correlation see Section 2.2.1. We call P the *generalized partial correlation matrix* of F and refer to it as *partial correlation matrix* for brevity, but keep in mind that partial correlations are defined for distributions with finite second-order moments only.
The qualitative information of P can be coded in an undirected graph $G = (V, E)$, where V is the vertex set and E the edge set, in the following way: the variables $X_1, ..., X_p$ are the vertices, and an undirected edge is drawn between X_i and X_j, $i \neq j$, if and only if $p_{i,j} \neq 0$. The thus obtained graph G is called the *generalized partial correlation graph* of F, where again, for brevity's sake, we will usually drop the leading *generalized* and use the abbreviation *PCG*. Formally we set $V = \{1, ..., p\}$ and write the elements of E as unordered pairs $\{i, j\}$, $1 \leq i < j \leq p$. The benefits of such a graphical representation will not be discussed here. Our focus lies on the modelling, but we should point out that the global and the local Markov property w.r.t. any PCG G are equivalent for any $F \in \mathscr{E}_p$ without any moment assumptions, cf. Theorem 2.2.7.
Let $\mathscr{S}_p^+(G)$ be the subset of $\mathscr{S}_p^+$ consisting of all positive definite matrices with zero entries at the positions specified by the graph $G = (V, E)$, i.e.

$$K \in \mathscr{S}_p^+(G) \iff K \in \mathscr{S}_p^+ \text{ and } k_{i,j} = 0 \text{ for } i \neq j \text{ and } \{i, j\} \notin E,$$

and define

$$\mathscr{E}_p(G) = \left\{ F \in \mathscr{E}_p(\mu, K^{-1}) \,\middle|\, \mu \in \mathbb{R}^p,\ K \in \mathscr{S}_p^+(G) \right\} \tag{3.3}$$

as the *elliptical graphical model* induced by G. In words, an elliptical graphical model $\mathscr{E}_p(G)$ is the collection of all p-dimensional continuous elliptical distributions that share the property that the inverse of the shape matrix has zero-entries at certain off-diagonal positions specified by G. We call the elliptical graphical model $\mathscr{E}_p(G)$ *decomposable* if the graph G is decomposable. A graph is decomposable, if it possesses no chordless cycle of length greater than 3. For alternative characterizations and properties of decomposable graphs see e.g. Lauritzen (1996), Chapter 2. Decomposable graphical models constitute an important class of models—in terms of interpretability as well as in terms of mathematical tractability, cf. Whittaker (1990), Chapter 12. Our focus will lie on decomposable models.

In the remainder of this section we discuss the interpretation of an absent edge in the PCG of $F \in \mathscr{E}_p$. Let us assume that the second-order moments of $X \sim F$ are finite. The partial uncorrelatedness of, say, X_1 and X_2 given $X_3, ..., X_p$, i.e. $p_{1,2} = 0$, is to be understood as *linear* independence of X_1 and X_2 that remains after the common *linear* effects of $X_3, ..., X_p$ have been removed. A relation of similar type is *conditional independence*: roughly, X_1 and X_2 are conditionally independent given $X_3, ..., X_p$, if the conditional distribution of (X_1, X_2) is a product measure for almost all values of the conditioning variable $(X_3, ..., X_p)$. In comparison to partial uncorrelatedness we understand conditional independence as *full* independence of X_1 and X_2 after the removal of *all* common effects of $X_3, ..., X_p$.

Another term, lying in-between, is *conditional uncorrelatedness*: the conditional distribution of (X_1, X_2) given $(X_3, ..., X_p)$ has correlation zero for almost all values of $(X_3, ..., X_p)$. We must point out an important qualitative difference between partial and conditional correlation: the former is a real value, whereas the latter is a function of the conditioning variable. For elliptical distributions it is known that all marginal and conditional distributions are again elliptical, cf. Fang and Zhang (1990), Section 2.6. It follows that partial uncorrelatedness implies conditional uncorrelatedness, cf. Baba et al. (2004). Hence $p_{1,2} = 0$ allows to conclude linear independence of X_1 and X_2 after *all* common effects of $X_3, ..., X_p$ have been removed.

On the other hand, the only spherical distributions with independent margins are Gaussian distributions, which is known as the Maxwell-Hershell-Theorem, cf. e.g. Bilodeau and Brenner (1999), p. 51. Thus contrary to Gaussian graphical models a missing edge in the PCG of an elliptical distribution can in general not be interpreted as conditional independence. It appears that, by going from the normal to the elliptical model, the gain in generality is paid by a loss in the strength of inference. But this loss is illusive. From a data modelling perspective the conditional independence interpretation of partial uncorrelatedness under normality is an assumption, not a conclusion. By modelling multivariate data by a joint Gaussian distribution one models the linear dependencies and *assumes* that there are no other than linear associations among the variables. By fitting an appropriate non-Gaussian model one may still model the linear dependencies and allow non-linear dependencies. Using semiparametric models embodies this idea: the aspects of interest (here linear dependencies) are modelled parametrically, whereas other aspects remain unspecified.

3.3 Elliptical graphical modelling: statistical theory

3.3.1 Unconstrained estimation

An important initial step towards elliptical graphical modelling is the unconstrained estimation of P. Unconstrained, since we do not assume a graphical model to hold, not forcing any constraints on P. We will consider estimators of the type $\hat{P}_n = h(\hat{S}_n)$, where $\hat{S}_n$ is a suitable estimator of a multiple of S, therefore start by considering shape estimators $\hat{S}_n$.

Now we consider i.i.d. random vectors $X_1, ..., X_n$ sampled from an elliptical distribution $F \in \mathscr{E}_p(\mu, S)$. (Depending on the context, X_k may denote the k-th p-dimensional observation or the k-th component of the vector X.) Let furthermore $\mathbb{X}_n = (X_1, ..., X_n)^T$ be the $n \times p$ data matrix containing the data points as rows and $\hat{S}_n = \hat{S}_n(\mathbb{X}_n)$ a scatter estimator. (The symbol $\hat{S}_n$ may have two meanings: a function on the sample space, or as abbreviation for $\hat{S}_n(\mathbb{X}_n)$, a random variable.) We use the term *scatter estimator* in a very informal way for any symmetric matrix-valued estimator that gives some information about the spread of the data. We restrict our attention to scatter estimators satisfying the following condition which we call *affine pseudo-equivariance*.

Assumption 3.3.1 $\hat{S}_n(\mathbb{X}_n A^T + 1_n b^T) = \xi(AA^T) A \hat{S}_n(\mathbb{X}_n) A^T$ *for* $b \in \mathbb{R}^p$, $A \in \mathbb{R}^{p \times p}$ *with full rank, and* $\xi \colon \mathbb{R}^{p \times p} \to \mathbb{R}$ *continuously differentiable, satisfying* $\xi(I_p) = 1$.

This is a generalization of the (strict) affine equivariance for scatter estimators, which corresponds to $\xi \equiv 1$. We use this weaker condition since overall scale is irrelevant for partial correlations, and we want to include estimators which only estimate shape, but not scale, and do not satisfy strict affine equivariance. Examples and further explanations are given in Section 3.4.1.
We call estimators satisfying Assumption 3.3.1 *shape estimators*. Evaluated at an elliptical distribution their first two moments (if existent) can be shown to have a common structure, the same given for strictly affine equivariant scatter estimators in Corollary 1 in Tyler (1982). The following condition is therefore natural for shape estimators at elliptical distributions F, and many shape estimators have been shown to satisfy it under suitable additional conditions on F.

Assumption 3.3.2 *There exist constants* $\eta \geq 0$, $\sigma_1 \geq 0$ *and* $\sigma_2 \geq -2\sigma_1/p$ *such that*

$$\hat{S}_n \xrightarrow{p} \eta S \qquad \text{and} \qquad \sqrt{n} \operatorname{vec}(\hat{S}_n - \eta S) \xrightarrow{\mathscr{L}} N_{p^2}\left(0, \eta^2 W_S(\sigma_1, \sigma_2)\right),$$

where $W_S = W_S(\sigma_1, \sigma_2) = 2\sigma_1 M_p(S \otimes S) + \sigma_2 \operatorname{vec} S (\operatorname{vec} S)^T$, *and the constants* σ_1 *and* σ_2 *do not depend on* S.

Tyler (1982) calls an estimator $\hat{S}_n$ satisfying this assumption to be *asymptotically of the radial type*. Under this assumption we have the following proposition, which is proved in the appendix.

Proposition 3.3.3 *If* $\hat{S}_n$ *fulfils Assumption 3.3.2, then we have with* $K = S^{-1}$ *that*

(1) the derived concentration matrix estimator $\hat{K}_n = \hat{S}_n^{-1}$ *satisfies*

$$\hat{K}_n \xrightarrow{p} \eta^{-1} K \qquad \text{and} \qquad \sqrt{n} \operatorname{vec}(\hat{K}_n - \eta^{-1} K) \xrightarrow{\mathscr{L}} N_{p^2}\left(0, \eta^{-2} W_K(\sigma_1, \sigma_2)\right),$$

where $W_K = W_K(\sigma_1, \sigma_2) = 2\sigma_1 M_p(K \otimes K) + \sigma_2 \operatorname{vec} K (\operatorname{vec} K)^T$, *and*

(2) the derived partial correlation estimator $\hat{P}_n = h(\hat{S}_n)$ *satisfies*

$$\hat{P}_n \xrightarrow{p} P \qquad \text{and} \qquad \sqrt{n} \operatorname{vec}(\hat{P}_n - P) \xrightarrow{\mathscr{L}} N_{p^2}\left(0, 2\sigma_1 \Gamma(S) M_p(K \otimes K) \Gamma(S)^T\right),$$

where

$$\Gamma(S) = (K_D^{-\frac{1}{2}} \otimes K_D^{-\frac{1}{2}}) + M_p(P \otimes K_D^{-1}) J_p. \tag{3.4}$$

An important aspect of Proposition 3.3.3 is that the asymptotic distribution of any partial correlation estimator $\hat{P}_n$ derived from an affine equivariant shape estimator $\hat{S}_n$ is a function of the shape except for the scalar σ_1.

3.3.2 Constrained estimation

In this section we deal with the task of estimating P under a given graphical model $\mathscr{E}_p(G)$ specified by the graph $G = (V, E)$, i.e. estimating P with zero-entries. A crude approach is to simply put the concerning elements in an unconstrained estimate $\hat{P}_n$ to zero. This will generally destroy the positive definiteness of the estimate. We pursue the path laid by the Gaussian MLE and define the function $h_G : \mathscr{S}_p^+ \to \mathscr{S}_p^+(G) : A \mapsto A_G$ by

$$\begin{cases} [A_G]_{i,j} = a_{i,j}, & \{i,j\} \in E \quad \text{or} \quad i = j, \\ [A_G^{-1}]_{i,j} = 0, & \{i,j\} \notin E \quad \text{and} \quad i \neq j. \end{cases} \tag{3.5}$$

It is not trivial and a deeper result of the theory of Gaussian graphical models that a unique and positive definite solution A_G of (3.5) exists for any positive definite A. The positive definiteness of A is sufficient but not necessary. For details see Lauritzen (1996), p. 133. Since we deal mainly with asymptotics, and, for sufficiently large n, shape matrix estimators $\hat{S}_n$ are usually a.s. positive definite at continuous distributions, we assume positive definiteness for simplicity's sake.
Let $G = (V, E)$ be a decomposable graph with cliques $\gamma_1, ..., \gamma_c$, $c \geq 1$, and define the sequence $\delta_1, ..., \delta_{c-1}$ of successive intersections by

$$\delta_k = (\gamma_1 \cup ... \cup \gamma_k) \cap \gamma_{k+1}, \qquad k = 1, ..., c-1.$$

We assume that the ordering $\gamma_1, ..., \gamma_k$ is such that the cliques form a *perfect sequence*, i.e. for all $k = 1, ..., c-1$ there is a $j \in \{1, ..., k\}$ such that $\delta_k \subseteq \gamma_j$. It is always possible to arrange the cliques of a decomposable graph in a perfect sequence (Lauritzen, 1996, Prop. 2.17). For notational convenience we let

$$\alpha_k = \begin{cases} \gamma_k & k = 1, ..., c, \\ \delta_{k-c} & k = c+1, ..., 2c-1, \end{cases} \quad \text{and} \quad \zeta_k = \begin{cases} 1 & k = 1, ..., c, \\ -1 & k = c+1, ..., 2c-1. \end{cases}$$

Then $h_G(A)$ allows the following explicit formulation

$$h_G(A) = A_G = \left(\sum_{k=1}^{2c-1} \zeta_k \left[A^{-1}_{\alpha_k, \alpha_k} \right]^p \right)^{-1}, \qquad A \in \mathscr{S}_p^+. \tag{3.6}$$

We will use this representation of h_G to further analyse the properties of the estimators $\hat{S}_G = h_G(\hat{S}_n)$, $\hat{K}_G = \hat{S}_G^{-1}$ and $\hat{P}_G = h(\hat{S}_G)$ for a decomposable graph G. Using the notation $S_G = h_G(S)$, $K_G = S_G^{-1}$, $P_G = h(S_G) \in \mathbb{R}^{p \times p}$ and

$$\Omega_G(S) = \sum_{k=1}^{2c-1} \zeta_k \left[S^{-1}_{\alpha_k, \alpha_k} \right]^p \otimes \left[S^{-1}_{\alpha_k, \alpha_k} \right]^p \in \mathbb{R}^{p^2 \times p^2}$$

we have the following result about the asymptotic distribution.

Proposition 3.3.4 *If $\hat{S}_n$ fulfils Assumption 3.3.2 and G is decomposable, then*

(1) $\hat{K}_G \xrightarrow{p} \eta^{-1} K_G$ *and* $\sqrt{n} \operatorname{vec}(\hat{K}_G - \eta^{-1} K_G) \xrightarrow{\mathscr{L}} N_{p^2}(0, \eta^{-2} W_{K_G}(\sigma_1, \sigma_2))$
with $W_{K_G} = W_{K_G}(\sigma_1, \sigma_2) = 2\sigma_1 M_p \Omega_G(S)(S \otimes S)\Omega_G(S) + \sigma_2 \operatorname{vec} K_G (\operatorname{vec} K_G)^T$,

(2) $\hat{S}_G \xrightarrow{p} \eta S_G$ *and* $\sqrt{n}\,\mathrm{vec}\,(\hat{S}_G - \eta S_G) \xrightarrow{\mathscr{L}} N_{p^2}(0, \eta^2 W_{S_G}(\sigma_1, \sigma_1))$ *with*

$$W_{S_G} = W_{S_G}(\sigma_1, \sigma_2) = 2\sigma_1 M_p\,(S_G \otimes S_G)\,\Omega_G(S)(S \otimes S)\Omega_G(S)\,(S_G \otimes S_G) + \sigma_2\,\mathrm{vec}\,S_G(\mathrm{vec}\,S_G)^T,$$

(3) $\hat{P}_G \xrightarrow{p} P_G$ *and* $\sqrt{n}\,\mathrm{vec}\,(\hat{P}_G - P_G) \xrightarrow{\mathscr{L}} N_{p^2}(0, W_{P_G}(\sigma_1))$ *with*
$W_{P_G} = W_{P_G}(\sigma_1) = 2\sigma_1\Gamma(S_G)M_p\Omega_G(S)(S \otimes S)\Omega_G(S)\Gamma(S_G)^T$ *and* $\Gamma(\cdot)$ *defined in (3.4).*

Since $\Omega_G(S)\,(S_G \otimes S_G)\,\Omega_G(S) = \Omega_G(S)$ for any $S \in \mathscr{S}_p^+$, which is proved in the appendix, the expressions for the asymptotic variances of the estimators simplify, if the true shape S satisfies the graph G, i.e. if $S = S_G$.

Corollary 3.3.5 *If $\hat{S}_n$ satisfies Assumption 3.3.2 with $S^{-1} \in \mathscr{S}_p^+(G)$ for a decomposable graph G, then the assertions of Proposition 3.3.4 are true with*

(1) $W_{K_G}(\sigma_1, \sigma_2) = 2\sigma_1 M_p\Omega_G(S) + \sigma_2\,\mathrm{vec}\,K(\mathrm{vec}\,K)^T$ *and*

(2) $W_{S_G}(\sigma_1, \sigma_2) = 2\sigma_1 M_p(S \otimes S)\Omega_G(S)(S \otimes S) + \sigma_2\,\mathrm{vec}\,S\,(\mathrm{vec}\,S)^T$,

(3) $W_{P_G}(\sigma_1) = 2\sigma_1\Gamma(S)M_p\Omega_G(S)\Gamma(S)^T$.

3.3.3 Testing

An essential tool of most model selection procedures is to test if a model under consideration fits the data or not. In this respect it is of particular interest to compare the fit of two nested models. Again, we restrict our attention here to the important subclass of decomposable models. For example, the `stepwise` model search routine of the MIM software, cf. Edwards (2000), by default only considers decomposable models. Models with at most one missing edge are decomposable.
We need to declare some notation first. On the set $\Pi_p = \{(i, j)|\, 1 \le i, j \le p\}$ of the positions of a $p \times p$ matrix we declare a strict ordering $\prec_p$ by

$$(i, j) \prec_p (k, l) \quad \text{if } (j-1)p + i \le (l-1)p + k \qquad \text{for } (i, j), (k, l) \in \Pi_p.$$

For any subset $Z = \{z_1, ..., z_q\} \subset \Pi_p$, where $z_k = (i_k, j_k)$, $k = 1, ..., q$, and $z_1 \prec_p ... \prec_p z_q$, define the matrix $Q_Z \in \mathbb{R}^{q \times p^2}$ as follows: each line consists of exactly one entry 1 and zeros otherwise. The 1-entry in line k is in column $(i_k - 1)p + j_k$. Thus $Q_Z\mathrm{vec}\,A$ picks the elements of A at positions specified by Z in the order they appear in $\mathrm{vec}\,A$. For a graph $G = (V, E)$ with $V = \{1, ..., p\}$ let

$$D(G) = \{(i, j)|1 \le j < i \le p, \{i, j\} \notin E\},$$

i.e. the set $D(G)$ gathers all sub-diagonal zero-positions that G enforces on a concentration matrix. Thus $F \in \mathscr{E}_p(G)$ is equivalent to $Q_{D(G)}\,\mathrm{vec}\,K = 0$.
Now let $G_0 = (V, E_0)$ and $G_1 = (V, E_1)$ be two decomposable graphs with V as above and $E_0 \subsetneq E_1$, or equivalently, $\mathscr{E}_p(G_0) \subsetneq \mathscr{E}_p(G_1)$. For notational convenience let

$$Q_0 = Q_{D(G_0)},\ Q_1 = Q_{D(G_1)},\ Q_{0,1} = Q_{D(G_0)\setminus D(G_1)},$$

furthermore

$$q_0 = |D(G_0)|,\ q_1 = |D(G_1)| \quad \text{and} \quad q_{0,1} = q_0 - q_1.$$

An intuitive approach to testing G_0 against the broader model G_1 is to reject G_0 in favour of G_1, if all entries at positions in $D(G_0) \setminus D(G_1)$ of an estimate $\hat{P}_{G_1}$ of P under G_1 are close to zero. For example, a sum of suitably weighted squared entries of $\hat{P}_{G_1}$, such as $\hat{T}_n(G_0, G_1)$ below, is a possible test statistic. Let

$$R_G(S) = \Gamma(S) M_p \Omega_G(S) \Gamma(S)^T.$$

For invertible S, $R_{G_1}(S)$ has rank $\frac{1}{2}(p-1)p - q_1$. This can be shown by applying the fact that invertible functions have full rank derivatives, which is a consequence of the chain rule, to suitably constructed functions. The proof is worked out in Section 4.1.2. Then $Q_{0,1} R_{G_1}(S) Q_{0,1}^T$ is of full rank, and the probability that the Wald-type test statistic

$$\hat{T}_n(G_0, G_1) = \frac{n}{2} \left(\operatorname{vec} \hat{P}_{G_1} \right)^T Q_{0,1}^T \left(Q_{0,1} R_{G_1}(\hat{S}_n) Q_{0,1}^T \right)^{-1} Q_{0,1} \operatorname{vec} \hat{P}_{G_1}$$

exists tends to one as $n \to \infty$. The next proposition describes the asymptotic behaviour of the test statistic $\hat{T}_n(G_0, G_1)$ under the null hypothesis that G_0 is true, part (1), and under a local alternative, part (2).

Proposition 3.3.6 *Let G_0 and G_1 be as above and $\hat{S}_n = \hat{S}_n(\mathbb{X}_n)$ satisfy Assumptions 3.3.1 and 3.3.2 for i.i.d. data $\mathbb{X}_n^T = (X_1, ..., X_n)$.*

(1) Under the model G_0, i.e. if $X_1, ..., X_n, ...$ are i.i.d. with $X_1 \sim F \in \mathscr{E}_p(\mu, S) \subset \mathscr{E}_p(G_0)$, then

$$\hat{T}_n(G_0, G_1) \xrightarrow{\mathscr{L}} \sigma_1 \chi^2_{q_{0,1}}.$$

(2) Let $X_1, ..., X_n, ...$ be as in part (1). Furthermore, for $m, k \in \mathbb{N}$, let $X_k^{(m)} \overset{\mathscr{L}}{=} S^{\frac{1}{2}} S_m^{-\frac{1}{2}} X_k$ and $X_1^{(m)}, ..., X_n^{(m)}, ...$ be independent (which implies that $X_1^{(m)}, ..., X_n^{(m)}, ...$ are i.i.d. elliptical with shape matrix S_m), where S_m is such that there exists a matrix $B \in \mathscr{S}_p$ with

$$\lim_{m \to \infty} \sqrt{m}(S_m - S) = B.$$

If, for each $n \in \mathbb{N}$, $\hat{S}_n$ is applied to $X_1^{(n)}, ..., X_n^{(n)}$, then

$$\hat{T}_n(G_0, G_1) \xrightarrow{\mathscr{L}} \sigma_1 \chi^2_{q_{0,1}} \left(\frac{\delta(B, S)}{\sigma_1} \right), \tag{3.7}$$

where

$$\delta(B, S) = \frac{1}{2} v^T Q_{0,1}^T \left(Q_{0,1} R_{G_1}(S) Q_{0,1}^T \right)^{-1} Q_{0,1} v$$

with the abbreviation $v = v(B, S) = \Gamma(S) \Omega_{G_1}(S) \operatorname{vec} B$.

Here we define the non-centrality parameter of the χ^2 distribution $\chi^2_r(\delta) \sim (N_r(\mu, I_r))^2$ as $\delta = \mu^T \mu$.

Remark 3.3.7 *In part (2) of Proposition 3.3.6 above we do not require that the sequence of alternatives "lies in" the model G_1, i.e. that $S_n^{-1} \in \mathscr{S}_p^+(G_1)$, as it is not necessary for the convergence (3.7) to hold. When choosing a model by forward selection one usually compares two wrong models, so it of interest to know the behaviour of $\hat{T}_n(G_0, G_1)$ also if G_1 is not true.*

A nuisance of the test in Proposition 3.3.6 may be the complicated formulation of the test statistic $\hat{T}_n(G_0, G_1)$. The classical test in Gaussian graphical models is the deviance test, an instance of a likelihood ratio test. The next proposition gives the analogue for elliptical graphical modelling. In order to treat parts (1) and (2) of the previous proposition simultaneously, we replace Assumptions 3.3.1 and 3.3.2 by the following assumption.

Assumption 3.3.8 *Let $\hat{S}_n$ be a sequence of almost surely positive definite random $p \times p$ matrices that satisfies*

$$\hat{S}_n \xrightarrow{p} \eta S \qquad \text{and} \qquad \sqrt{n}\,\mathrm{vec}\,(\hat{S}_n - \eta S) \xrightarrow{\mathscr{L}} N_{p^2}\left(\eta\,\mathrm{vec}\,C, \eta^2 W_S(\sigma_1, \sigma_2)\right),$$

for some $C \in \mathscr{S}_p$, $S \in \mathscr{S}_p^+$ with $S^{-1} \in \mathscr{S}_p^+(G_0)$ and suitable constants $\eta \geq 0$, $\sigma_1 \geq 0$ and $\sigma_2 \geq -2\sigma_1/p$. The matrix $W_S(\sigma_1, \sigma_2)$ is as in Assumption 3.3.2.

If $\hat{S}_n(\cdot)$ is a shape estimator satisfying Assumptions 3.3.1 and 3.3.2, then, in the situation of Proposition 3.3.6 (1), $\hat{S}_n(\mathbb{X}_n)$ fulfils Assumption 3.3.8 with $C = 0$, and, in the situation of Proposition 3.3.6 (2), $\hat{S}_n(\mathbb{X}_n)$ fulfils Assumption 3.3.8 with $C = B + cS$, cf. Lemma 3.5.1.

Proposition 3.3.9 *Let G_0 and G_1 be as above and $\hat{S}_n$ satisfy Assumption 3.3.8. Then*

$$\hat{D}_n(G_0, G_1) = n\left(\ln\det h_{G_0}(\hat{S}_n) - \ln\det h_{G_1}(\hat{S}_n)\right)$$

is asymptotically equivalent to $\hat{T}_n(G_0, G_1)$, i.e. $\hat{T}_n(G_0, G_1) - \hat{D}_n(G_0, G_1) \xrightarrow{p} 0$.

Proposition 3.3.9 implies that both assertions (1) and (2) of Proposition 3.3.6 remain true, if $\hat{T}_n(G_0, G_1)$ is replaced by $\hat{D}_n(G_0, G_1)$. In the special case that the larger model G_1 is the saturated model Proposition 3.3.9 is a corollary of Theorem 2 in Tyler (1983). We basically consider an extension of Tyler's result to the case of two nested models.

Remark 3.3.10 *Model search based on the pseudo-deviance tests is not restricted to decomposable models. The convergence of the test statistic $\hat{D}_n(G_0, G_1)$ to a χ^2 distribution holds true, also if one of G_0 and G_1 is not decomposable. The general case requires some further mathematical techniques and is treated in Chapter 4.*

3.4 Elliptical graphical modelling: practical aspects

3.4.1 Examples of affine pseudo-equivariant scatter estimators

We have been talking about affine pseudo-equivariant estimators without naming a single one, which we will make up for in this section. But first, we want to spare a few words about the relevance of Assumption 3.3.1. For practical purposes it may be replaced by the following, formally weaker condition.

Assumption 3.4.1 $\hat{S}_n(\mathbb{X}_n A^T + 1_n b^T) \propto A\hat{S}_n(\mathbb{X}_n)A^T$ *for $b \in \mathbb{R}^p$ and $A \in \mathbb{R}^{p\times p}$ with full rank.*

Assumption 3.3.1 additionally requires that the proportionality factor is a function solely of the affine linear transformation and not random, which ensures that the covariance of such an estimator has the form $W_S(\sigma_1, \sigma_2)$ as in Assumption 3.3.2. Our claim, Assumption 3.3.1 may be replaced by Assumption 3.4.1, has two justifications. *(1)* Any estimator $\hat{S}_n$ satisfying Assumption 3.4.1 can be turned into

an estimator satisfying Assumption 3.3.1, say $\tilde{S}_n$, by putting $\tilde{S}_n = \det(\hat{S}_n)^{-1/p}\hat{S}_n$, i.e. $\tilde{S}_n$ has determinant 1. *(2)* Our main results concern scale-invariant functions of $\hat{S}_n$ (partial correlation estimators in Propositions 3.3.3 and 3.3.4 and the Wald-type and deviance test statistics in Propositions 3.3.6 and 3.3.9, respectively) and directly apply to any $\hat{S}_n$ satisfying Assumption 3.4.1. While *(1)* suggests to formulate all results for thus standardized scatter estimators, *(2)* indicates why we refrain from doing so: to avoid the impression that a particular standardization was necessary.
Second, we want to point out that the class of affine pseudo-equivariant estimators is huge. Over the last decades the robustness literature has produced many proposals of affine equivariant, robust estimators that are at the same time preferably efficient and computationally feasible. Prominent classes of such estimators are M-estimators, S-estimators and Stahel-Donoho estimators, see e.g. the overview article by Zuo (2006) or the book by Maronna et al. (2006).
Let us now come to some specific examples. Of course, the classical estimator, the sample covariance matrix, is affine equivariant. The following can be found in Tyler (1982).

Proposition 3.4.2 *If $X_1, ..., X_n$ are i.i.d. with distribution $F \in \mathscr{E}_p(\mu, S)$ and $\mathbb{E}\|X_1 - \mu\|^4 < \infty$, then $\hat{\Sigma}_n = \hat{\Sigma}_n(\mathbb{X}_n)$ fulfils Assumption 3.3.2 with $\sigma_1 = 1 + \kappa/3$ and $\sigma_2 = \kappa/3$, where κ is the excess kurtosis of the first (or equivalently any other) component of X_1.*

Proposition 3.4.2 indicates the inappropriateness of the sample covariance matrix for heavy-tailed distributions: fourth moments are required to make it root-n-consistent, and its asymptotic distribution depends on the kurtosis, which may be large at heavy-tailed distributions, thus rendering this estimator rather inefficient. An alternative is Tyler's M-estimator, which is defined as the solution $\hat{T}_n = \hat{T}_n(\mathbb{X}_n)$ of

$$\frac{p}{n}\sum_{i=1}^{n}\frac{(X_i - \overline{X}_n)(X_i - \overline{X}_n)^T}{(X_i - \overline{X}_n)^T\hat{T}_n^{-1}(X_i - \overline{X}_n)} = \hat{T}_n$$

which satisfies $\det\hat{T}_n = 1$. Existence, uniqueness and asymptotic properties are treated in the original publication Tyler (1987a), where the following result is proven.

Proposition 3.4.3 *If $X_1, ..., X_n$ are i.i.d. with distribution $F \in \mathscr{E}_p(\mu, S)$, furthermore $\mathbb{E}\|X_1 - \mu\|^2 < \infty$ and $\mathbb{E}\|X_1 - \mu\|^{-\frac{3}{2}} < \infty$, then $\hat{T}_n$ fulfils Assumption 3.3.2 with $\sigma_1 = 1 + \frac{2}{p}$ and $\sigma_2 = -\frac{2}{p}\left(1 + \frac{2}{p}\right)$.*

We have the following remarks.

(I) An important aspect of Proposition 3.4.3 is that the scalars σ_1 and σ_2 are constant, irrespective of the function g, meaning that the Tyler matrix is asymptotically distribution-free within the elliptical model. This has the nice practical implication that, when carrying out any of the tests from Section 3.3.3, σ_1 does not need to be estimated.

(II) The assumption of finite second moments is only required for location estimation by the mean. The mean may be replaced by any root-n-consistent location estimator. The Hettmansperger-Randles median (Hettmansperger and Randles, 2002) is a canonical candidate, which follows a similar conceptual idea as Tyler's scatter estimator, but has turned out to be rather difficult to handle analytically. We note, however, that Tyler's matrix can cope with arbitrarily heavy tails.

(III) The inverse moment condition $\mathbb{E}\|X_1 - \mu\|^{-\frac{3}{2}} < \infty$ can generally not be dropped by choosing a different location estimator, cf. Tyler (1987a, Theorem 4.2). But this is a fairly mild condition: for $p \geq 2$ it is fulfilled if g has no singularity at 0, thus including normal and $t_{\nu,p}$-distributions.

(IV) The Tyler matrix $\hat{T}_n$ is an example of a pure shape estimator, which only gives information about the shape but none about the scale. Other such estimators are, for example, Oja sign and rank covariance matrices (Ollila et al., 2003, 2004), which are also affine pseudo-equivariant.

(V) Tyler (1987a) uses $\operatorname{tr}(\hat{T}_n) = p$ to fix the scale of $\hat{T}_n$. The estimator thus standardized, let us call it $\tilde{T}_n$, fulfils

$$\tilde{T}_n(\mathbb{X}_n A^T + 1b^T) = \frac{p}{\operatorname{tr}(A\tilde{T}_n(\mathbb{X}_n)A^T)} A\tilde{T}_n(\mathbb{X}_n)A^T$$

and is hence an example of an estimator satisfying Assumption 3.4.1, but not Assumption 3.3.1. As a consequence $\tilde{T}_n$ does not satisfy Assumption 3.3.2. Its asymptotic covariance matrix has indeed a different form than $W_S(\sigma_1, \sigma_2)$, which can be verified by the delta method.

Interestingly, Tyler's original publication, which considers $\tilde{T}_n$, contains a formula for the asymptotic covariance matrix under ellipticity of exactly the same type as in Assumption 3.3.2 (Tyler, 1987a, (3.10)). But this formula applies to the estimator $p/\operatorname{tr}(T^{-1}\tilde{T}_n)\,\tilde{T}_n$, which again fulfils Assumption 3.3.1. Here T denotes the population Tyler matrix, i.e. the shape matrix with the trace set to p.

(VI) Tyler's estimator $\hat{T}_n$ fulfils Assumption 3.3.2 with the restriction $\sigma_2 \geq -2\sigma_1/p$ being satisfied with equality. If equality holds, the matrix $W_S(\sigma_1, \sigma_2)$ has exactly one rank less than in the case of strict inequality. For details see Tyler (1982, Section 2). As a consequence the rank of the asymptotic covariance matrix of $\hat{T}_n$ at any non-degenerate elliptical distribution is by one smaller than that of the sample covariance matrix $\hat{\Sigma}_n$. This is plausible considering that $\hat{T}_n$ satisfies the additional constraint $\det(\hat{T}_n) = 1$.

Another affine equivariant estimator is the RMCD, the reweighted version of Rousseeuw's minimum covariance determinant estimator (Rousseeuw, 1985; Rousseeuw and Leroy, 1987; Croux and Haesbroeck, 1999), which has become a very popular highly robust scatter estimator and has previously been proposed in the context of graphical modelling (Becker, 2005; Gottard and Pacillo, 2010). It is defined as follows. A subset $\tau \subset \{1, ..., n\}$ of size $h = \lceil tn \rceil$, where $\frac{1}{2} \leq t < 1$ is fixed, is determined such that $\det(\hat{\Sigma}^\tau)$ with

$$\hat{\Sigma}^\tau = \frac{1}{h}\sum_{i\in\tau}(X_i - \bar{X}^\tau)(X_i - \bar{X}^\tau)^T \quad \text{and} \quad \bar{X}^\tau = \frac{1}{h}\sum_{i\in\tau} X_i$$

is minimal. The mean $\hat{\mu}_{\text{MCD}}$ and covariance matrix $\hat{\Sigma}_{\text{MCD}}$ computed from this minimizing subsample are called the *raw MCD location*, respectively *scatter estimate*. The covariance part is multiplied by a consistency factor to achieve consistency for the covariance at the Gaussian distribution. Based on the raw estimates $(\hat{\mu}_{\text{MCD}}, \hat{\Sigma}_{\text{MCD}})$ a reweighted scatter estimator $\hat{\Sigma}_{\text{RMCD}}$ is computed from the whole sample:

$$\hat{\Sigma}_{\text{RMCD}} = \left(\sum_{i=1}^n w_i\right)^{-1} \sum_{i=1}^n w_i(X_i - \hat{\mu}_{\text{MCD}})(X_i - \hat{\mu}_{\text{MCD}})^T,$$

where $w_i = 1$ if $(X_i - \hat{\mu}_{\text{MCD}})^T\hat{\Sigma}_{\text{MCD}}^{-1}(X_i - \hat{\mu}_{\text{MCD}}) < \chi^2_{p,1-\alpha}$ and zero otherwise, where α is a small "rejection probability", e.g. $\alpha = 0.05$. The reweighted covariance estimate is again multiplied by a consistency factor, but since this is not necessary for our applications we omit the details.

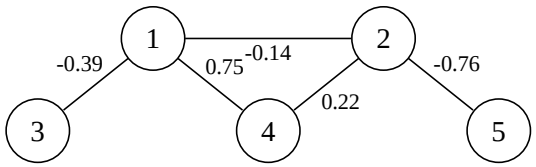

Figure 3.1: Example model, edge labels indicate non-zero partial correlations

3.4.2 Simulations

We present the results of a small simulation study, comparing several estimators, to give an impression how the approach works in practice. The set-up is as follows. For each of several elliptical distributions we sample 100 i.i.d. observations of a 5-dimensional random vector. We use the same shape matrix throughout, with equal diagonal elements and partial correlation matrix

$$P = \begin{pmatrix} -1 & & & & \\ -0.14 & -1 & & & \\ -0.39 & 0 & -1 & & \\ 0.75 & 0.22 & 0 & -1 & \\ 0 & -0.76 & 0 & 0 & -1 \end{pmatrix}.$$

Figure 3.1 shows the corresponding partial correlation graph. Of the total of ten possible edges five are present. We have chosen some variety in the magnitude of the non-zero partial correlations, their absolute values ranging from 0.14 to 0.76. So we leave the shape fixed and let the tail behaviour vary, using the normal distribution and several members of the $t_{\nu,p}$-family as a prominent example of a heavy-tailed distribution. The $t_{\nu,p}$-distribution is a p-dimensional elliptical distribution specified by, cf. (3.1),

$$g_{t_{\nu,p}}(y) = \frac{\Gamma(\frac{\nu+p}{2})}{(\nu\pi)^{\frac{p}{2}}\Gamma(\frac{\nu}{2})}\left(1 - \frac{y}{\nu}\right)^{-\frac{\nu+p}{2}},$$

where the index $\nu \in \mathbb{N}$ is referred to as the degrees of freedom. The moments of $t_{\nu,p}$ are finite only up to order $\nu - 1$. For $\nu \geq 3$ its covariance matrix is $\frac{\nu}{\nu-2}S$, and for $\nu \geq 5$ the excess kurtosis (of each component) is $\frac{6}{\nu-4}$. Hence from Propositions 3.4.2 and 3.4.3 we know that the Tyler matrix is asymptotically more efficient than the sample covariance matrix at $t_{\nu,p}$ if $\nu < p + 4$. For each distribution considered, cf. Table 3.1, we generate 2000 samples, and for each sample compute the estimates described in Section 3.4.1. Based on each estimate we select a model. By comparing the selected models to the true model we evaluate the performance of the estimators in elliptical graphical modelling.

Our model selection scheme is the simplest possible: we carry out an edge-exclusion test for each of the 10 possible edges, i.e. we test, for each pair $\{i, j\}$, the model with all edges but $\{i, j\}$ against the saturated model and exclude the edge $\{i, j\}$ if the test accepts the smaller model. The significance level $\alpha = 0.05$ is an ad hoc choice. We do neither claim that this choice is (near) optimal in any sense nor address the question of multiple testing, as e.g. in Dahlhaus (2000). More sophisticated model search procedures, such as backward elimination, showed strictly better results (in terms of mean edge difference), but of comparable magnitude and lead to the same conclusions as far as the comparison of the estimators is concerned. Our simple one-step model selection allows to concentrate on the effects of the different estimators. In our simulations the Wald-type test statistic $\hat{T}_n$ of Proposition 3.3.6 and

the deviance test statistic $\hat{D}_n$ of Proposition 3.3.9 showed a very similar behaviour. The tables below report the results of the deviance test.

We pursue two main goals: besides getting an impression of the general performance we want to examine the finite-sample behaviour of the estimators, i.e. check if the asymptotic approximations derived in Section 3.3.3 are useful in practice. A sample size of 100 seems large enough to expect some "validity" of the asymptotics. We therefore consider several criteria. The main criterion by which we measure the goodness of the model selection is the *mean edge difference*, i.e. the average number of edges that are wrongly specified in the selected model, may it be that an existing edge was rejected or an absent edge was wrongly included. Any sensible model selection must yield, in our example, a mean edge difference of less than 5, which makes it superior to random guessing. It is also of interest to know, although in our opinion less suited as an overall performance criterion, how often the edge difference is zero, i.e. how often the true model is found. Although incorrect omission of edges is usually the more severe error, because subsequent inference is then based on an incorrect model, any model selection procedure that is based on testing for zero parameters aims at controlling the probability of correctly specifying the non-edges. For instance, in our example, the probability of correctly specifying all five absent edges is at least 0.75. We may also look at how often a specific non-edge is correctly specified, which should turn out to be true in about 95% of the cases.

In a first experiment we compare the sample covariance matrix $\hat{\Sigma}_n$ to Tyler's estimator $\hat{T}_n$ with the Hettmansperger-Randles median as location estimator. The findings are summarized in Table 3.1. The benchmark is traditional graphical modelling, i.e. the performance of $\hat{\Sigma}_n$ at the normal distribution. Even in this case we find that the true model is reconstructed in only 20% of the cases (24% when using simple backward elimination). But this is not surprising considering that some alternatives are close to the null, so that the test evidently has poor power. For instance, the probability of wrongly dismissing the edge 1–2 is about 0.7. As we would expect, we see by the last two columns of Table 3.1 that the test goes wrong, if we move away from normality. We assume ellipticity but no further knowledge about the distribution and are interested in methods that are valid and preferably efficient over the whole class of elliptical distributions. As a consequence of Proposition 3.4.2 we adjust the $\hat{\Sigma}_n$-based test statistic by the sample kurtosis. This repairs the test, and does so to some extent even in the case of the t_3-distribution where the population kurtosis is not finite. But we also see that this does not necessarily imply a better model selection. The estimator $\hat{\Sigma}_n$ is inefficient under heavy tails, resulting in a test with low power. On the other hand, for Tyler's estimator we recognize the asymptotic properties: the χ^2-quantile fits, it outperforms $\hat{\Sigma}_n$ at t_ν-distributions with $\nu < 9$, requires no moment conditions and is distribution-free within the elliptical model. So we can advise to employ the Tyler matrix instead of the sample covariance matrix to perform graphical modelling of heavy-tailed data, but again, these are just two examples of affine equivariant estimators.

In a second experiment we want to know if the same robustness against heavy-tailedness may be achieved by equally simple means using other robust estimators. In particular, how do the previous proposals of robust GGM, the RMCD and the Miyamura-Kano estimator, perform in this situation? Some results are given in Table 3.2.

Outlier-robust estimators interpret the bulk of the data as approximately normal and the observations in the tails as faulty outliers, that should be downweighted or rejected. Although there are some common aspects (elliptical MLEs of heavy-tailed distributions downweight outlying observations as compared to the normal MLE), this is in principle a different situation, and it is consequently not surprising that both estimators do not meet the performance of Tyler's estimator at heavy-tailed distributions. Also, we did not estimate σ_1 from the data, but used its value for the normal distribution. For the RMCD the values can be found in Croux and Haesbroeck (1999). But also in the Gaussian case, when σ_1 was chosen "asymptotically correct", the asymptotic χ^2-distribution does not seem to provide a sensible

Table 3.1: One-step model selection based on $\hat{\Sigma}$ or $\hat{T}$

distribution	estimator	mean edge difference	% true model found	% non-edges correctly found	% ①≠⑤ correctly found
normal	$\hat{\Sigma}$	1.40	21	79	95
	$\hat{\Sigma}^*$	1.41	20	77	94
	$\hat{T}$	1.65	14	78	94
t_8	$\hat{\Sigma}$	1.65	17	64	89
	$\hat{\Sigma}^*$	1.65	15	76	93
	$\hat{T}$	1.62	13	79	94
t_5	$\hat{\Sigma}$	1.90	14	51	84
	$\hat{\Sigma}^*$	1.87	10	74	93
	$\hat{T}$	1.63	14	78	94
t_3	$\hat{\Sigma}$	2.49	8	29	72
	$\hat{\Sigma}^*$	2.28	7	71	91
	$\hat{T}$	1.65	14	78	95

approximation. This small-sample inefficiency of the RMCD is known and usually taken care of by multiplying the test statistic by a correction factor. This correction factor has to be determined numerically, some values are given in Croux and Haesbroeck (1999). Using such an appropriate finite-sample value of σ_1 (depending on n) allows to repair the test (see last column), but again, this does not improve the model selection in our example (see first column). The 50% RMCD is, even more so at small sample sizes, a relatively inefficient estimator and is only recommended when the data is heavily corrupted. For the Miyamura-Kano proposal we note that they also devise an alternative way of constrained estimation, but propose a very slow algorithm. Also, there is a tuning parameter to choose, which was set 0.3 in our experiment, following the recommendation of the authors. All calculations were done in R 2.9.1, employing routines from the packages `mvtnorm` (random sampling), `ggm` (constrained estimation, i.e. the function h_G), `ICSNP` (Tyler matrix), `rrcov` (RMCD) and `rggm` (M-K estimator).

3.4.3 Summary and discussion

We have proposed a unified framework for graphical modelling of elliptical data, generalizing Gaussian graphical modelling and allowing in particular to deal with heavy tails. As a very simple and efficient technique to safeguard graphical modelling of continuous data against the impact of heavy tails, non-normality in general and, to some degree, also faulty outliers we recommend to use Tyler's estimator in place of the empirical covariance matrix. The gain in robustness comes at a very moderate loss in efficiency, which becomes smaller with increasing dimension, and a justifiable increase in computing time. Section 5.5 reports average computing times on a 2.83 GHz Intel Core2 CPU for $n = 200$ and $p = 50$ of less than a second for the Tyler matrix, compared to less than three seconds for the RMCD. Moreover, Tyler's estimator is computable for $n > p$ (for data in general position), and its distribution generally shows an equally fast convergence to the normal limit as the law of the sample covariance matrix. Besides the convenience of this simple technique our approach allows to use any affine pseudo-equivariant, root-n-consistent estimator in an analogous way. When additional information about the data is available (concerning possible contamination, tail-behaviour,...), estimators tailored for the specific situation may be used. Alternatively, sophisticated adaptive methods, which attempt to extract such information from the data, also fall into the class under consideration.

Table 3.2: One-step model selection based on robust estimators

distribution	estimator	mean edge difference	% true model found	% non-edges correctly found	% ①≠⑤ correctly found
normal	RMCD 0.5	2.05	11	54	85
	RMCD 0.5**	2.06	5	81	94
	RMCD 0.75	1.66	15	72	92
	RMCD 0.75**	1.69	13	80	94
	M-K+	1.61	14	81	95
t_3	RMCD 0.5	2.18	9	45	82
	RMCD 0.5**	2.13	5	76	93
	RMCD 0.75	2.02	11	51	85
	RMCD 0.75**	1.96	10	61	89
	M-K+	1.82	12	67	91

** with finite-sample correction, + Miyamura & Kano (2006).

Although we have used Assumption 3.3.1 as a technical requirement at some point in the proofs, the statistical theory presented is of asymptotic nature, and Assumption 3.3.2 is the important property of the estimator $\hat{S}_n$. Our results also apply to estimators that are only asymptotically affine equivariant, like the rank-based estimation technique by Hallin et al. (2006).
Finally we should mention the main limitation of our approach. It works well only for sufficiently large n, and on any account only for $n > p$. However, the processing of very high-dimensional data, where we have more variables than observations, becomes increasingly relevant. The empirical covariance matrix possesses the nice "consistency property" that the estimate of a margin appears as a submatrix of the estimate of the whole vector. This allows the constrained estimate $\hat{\Sigma}_G$ under some model G, not necessarily decomposable, to be "assembled" from the unrestricted marginal estimates corresponding to the cliques. This makes it possible to compute the MLE for $p \geq n$. In the decomposable case it suffices to have as many observations as the size of the largest clique. For details see Lauritzen (1996), Chapter 5. Affine pseudo-equivariant shape estimators generally do not possess this consistency property, and we need an initial estimate of full size. Also note that, for instance, the computation of the 50% MCD requires more than twice as many obervations as variables. One way to tackle this problem is to drop the affine equivariance and resort to robust "pairwise" estimators, such as the Gnanadesikan-Kettenring estimator (Gnanadesikan and Kettenring, 1972; Maronna and Zamar, 2002) or marginal sign and rank matrices (Visuri et al., 2000), see also Section 5.2. Besides having the mentioned consistency property pairwise estimators are also very fast to compute.

3.5 The proofs

The following proofs repeatedly apply the delta method to functions mapping matrices to matrices. We define the derivative of such a function, say, $g : \mathbb{R}^{p\times p} \to \mathbb{R}^{p\times p}$ at point X as the derivative of $\operatorname{vec} g(X)$ w.r.t. $\operatorname{vec} X$ and denote its Jacobian at point X (which is of size $p^2 \times p^2$) by $\mathbb{D}g(X)$. The symmetry of the argument poses a technical difficulty: there are actually only $\frac{p}{2}(p+1)$ instead of p^2 variables, and the function must be viewed as a function $g : \mathbb{R}^{\frac{p}{2}(p+1)} \to \mathbb{R}^{p\times p}$ in order to sensibly define a derivative. A practical way of dealing with this issue is to compute the Jacobian of g interpreted as a function from $\mathbb{R}^{p\times p}$ to $\mathbb{R}^{p\times p}$ and post-multiply it by M_p. This is justified by the chain rule applied

to $g = g_2 \circ g_1$, where g_1 duplicates the off-diagonal elements and $g_2 : \mathbb{R}^{p\times p} \to \mathbb{R}^{p\times p}$, see Section A.2. Pre- or post-multiplying the covariance matrix of $\operatorname{vec} X$, where X is a random symmetric $p \times p$ matrix, by M_p leaves it unchanged. Hence for application of the delta method the symmetry may as well be ignored, and we will omit M_p in the derivative expressions in the proofs of Propositions 3.3.3 and 3.3.4. However, it should not be forgotten, for instance it must be included to render the formula in Theorem 1 in Tyler (1983) correct. The page numbers below refer to the textbook Magnus and Neudecker (1999). It covers most of the tools of the proofs, in particular calculation rules concerning the vec operator, the Kronecker product and derivatives of matrix functions. We repeatedly use the following without reference.

$$(A \otimes B)(C \otimes D) = AC \otimes BD, \quad (\operatorname{vec} A)^T \operatorname{vec} B = \operatorname{tr}(A^T B), \quad \operatorname{vec}(ABC) = (C^T \otimes A) \operatorname{vec} B \tag{3.8}$$

$$M_p(A \otimes A)M_p = M_p(A \otimes A) = (A \otimes A)M_p \tag{3.9}$$

for matrices $A, B, C, D \in \mathbb{R}^{p\times p}$ (MN pp. 28, 30, 31). Let $\iota : A \mapsto A^{-1}$ denote the matrix inversion. Its Jacobian is (MN p. 184)

$$\mathbb{D}\iota(A) = -(A^T)^{-1} \otimes A^{-1}. \tag{3.10}$$

Proof of Proposition 3.3.3. The weak consistency follows from the continuous mapping theorem and the asymptotic normality from the delta method. It remains to calculate the asymptotic variances. Part (1): With $K = \iota(S)$ and (3.10) application of the delta method yields

$$W_K(\sigma_1, \sigma_2) = (K \otimes K)W_S(\sigma_1, \sigma_2)(K \otimes K)$$

which is transformed to the expression given in Proposition 3.3.3 employing (3.8).
Part (2): $\hat{P}_n = \tilde{h}(\hat{K}_n)$ with $\tilde{h} : A \mapsto -A_D^{-\frac{1}{2}} A A_D^{-\frac{1}{2}}$. We want to compute the derivative of $\tilde{h}$ in order to apply the delta method. We start by considering $\tilde{h}_0 : A \mapsto A_D^{-\frac{1}{2}}$. Its Jacobian $\mathbb{D}\tilde{h}_0(A) = -\frac{1}{2}\left(A_D^{-\frac{1}{2}} \otimes A_D^{-1}\right) J_p$ is obtained by elementwise differentiation. Applying the multiplication rule to $\tilde{h}(A) = -\tilde{h}_0(A)A\tilde{h}_0(A)$ yields

$$\mathbb{D}\tilde{h}(A) = -M_p\left(\tilde{h}(A) \otimes A_D^{-1}\right) J_p \; - \; A_D^{-\frac{1}{2}} \otimes A_D^{-\frac{1}{2}}. \tag{3.11}$$

By the delta method,

$$\sqrt{n} \operatorname{vec}\left(\hat{P}_n - P\right) = \sqrt{n} \operatorname{vec}\left(\tilde{h}(\hat{K}) - \tilde{h}(\eta^{-1}K)\right)$$

converges in distribution to a p^2-dimensional normal distribution with mean zero and covariance matrix

$$\mathbb{D}\tilde{h}(\eta^{-1}K)\eta^{-2}W_K(\sigma_1, \sigma_2)\left(\mathbb{D}\tilde{h}(\eta^{-1}K)\right)^T,$$

which reduces to the expression given in Proposition 3.3.3. In particular, σ_2 vanishes. By applying (3.8) it can be seen that $\mathbb{D}\tilde{h}(K) \operatorname{vec} K = 0$. This is generally true for any scale-invariant function $\tilde{h}$, which is e.g. employed in Tyler (1983), Theorem 1. ∎

Proof of Proposition 3.3.4.
Part (1): Since $K_G = \tilde{h}_G(S)$ with

$$\tilde{h}_G : A \mapsto \sum_{k=1}^{2c-1} \zeta_k \left[A^{-1}_{\alpha_k,\alpha_k}\right]^p$$

we want to compute the derivative of $\tilde{h}_G$. Let $\tilde{h}_\alpha : A \mapsto [A^{-1}_{\alpha,\alpha}]^p$ for any subset $\alpha \subset \{1, ..., p\}$. The mapping $\tilde{h}_\alpha$ is a composition of $(\cdot)_{\alpha,\alpha}$, ι and $[\cdot]^p$. We obtain by the chain rule

$$\mathbb{D}\tilde{h}_\alpha(A) = -\left[(A^{-1}_{\alpha,\alpha})^T\right]^p \otimes \left[A^{-1}_{\alpha,\alpha}\right]^p \quad \text{and hence} \quad \mathbb{D}\tilde{h}_G(A) = -\sum_{k=1}^{2c-1} \zeta_k \left[(A^{-1}_{\alpha_k,\alpha_k})^T\right]^p \otimes \left[A^{-1}_{\alpha_k,\alpha_k}\right]^p.$$

Then $\eta^{-2} W_{K_G}(\sigma_1, \sigma_2) = \mathbb{D}\tilde{h}_G(\eta S)\eta^2 W_S(\sigma_1, \sigma_2)\left(\mathbb{D}\tilde{h}_G(\eta S)\right)^T$ is shown to have the form given in Proposition 3.3.4 (2) by noting that $\mathbb{D}\tilde{h}_G(S)\,\mathrm{vec}\, S = \mathrm{vec}\, K_G$. This holds true because

$$\left[S^{-1}_{\alpha,\alpha}\right]^p S \left[S^{-1}_{\alpha,\alpha}\right]^p = \left[S^{-1}_{\alpha,\alpha}\right]^p, \tag{3.12}$$

which is a consequence of the inversion formula for partitioned matrices.
Part (2): In analogy to the proof of Proposition 3.3.3 (1) we have to left- and right-multiply W_{K_G} by the Jacobian of ι evaluated at K_G. Note that $(S_G \otimes S_G)\,\mathrm{vec}\, K_G = \mathrm{vec}\, S_G$.
Part (3): In analogy to the proof of Proposition 3.3.3 (2) we left- and right-multiply W_{K_G} by the Jacobian of $\tilde{h}$, given in (3.11), evaluated at K_G. ∎

Proof of Corollary 3.3.5. Let $S \in \mathscr{S}_p^+$ be such that $h_G(S) = S$ and write Ω short for $\Omega_G(S)$. We want to prove

$$\Omega(S \otimes S)\Omega = \Omega. \tag{3.13}$$

As short-hand notation let $\langle \alpha, \beta \rangle_S = \left[S^{-1}_{\alpha,\alpha}\right]^p S \left[S^{-1}_{\beta,\beta}\right]^p \in \mathbb{R}^{p\times p}$ for any two subsets $\alpha, \beta \subset \{1, ..., p\}$. Equation (3.13) can thus be rewritten as

$$\sum_{j=1}^{2c-1}\sum_{k=1}^{2c-1} \zeta_j\zeta_k \left(\langle \alpha_j, \alpha_k\rangle_S \otimes \langle \alpha_j, \alpha_k\rangle_S\right) = \sum_{k=1}^{2c-1} \zeta_k \left[S^{-1}_{\alpha_k,\alpha_k}\right]^p \otimes \left[S^{-1}_{\alpha_k,\alpha_k}\right]^p. \tag{3.14}$$

We have to show that the left-hand double sum reduces to the right-hand side, and indeed, most summands cancel. By (3.12) $\langle \alpha_k, \alpha_k\rangle_S = \left[S^{-1}_{\alpha_k,\alpha_k}\right]^p$, furthermore, for $i = 1 \le i < j \le p$,

$$\langle \gamma_i, \gamma_j\rangle_S = \langle \gamma_i, \delta_{j-1}\rangle_S \quad \text{and} \quad \langle \delta_i, \gamma_j\rangle_S = \langle \delta_i, \delta_{j-1}\rangle_S.$$

This is true because δ_{j-1} separates $\gamma_j \setminus \delta_{j-1}$ and $\gamma_i \setminus \delta_{j-1}$ as well as $\gamma_j \setminus \delta_{j-1}$ and $\delta_i \setminus \delta_{j-1}$ in the graph G for $i = 1 \le i < j \le p$, cf. Lauritzen (1996), Lemma 2.11. Both sides of each pair appear with different signs in the left-hand side of (3.14). We remark furthermore that we can deduce

$$M_p\Omega(S \otimes S)\Omega = M_p\Omega,$$

which is sufficient for the proof of Corollary 3.3.5, by comparing the expression for the asymptotic covariance of the inverse of the sample covariance matrix $\hat{\Sigma}^{-1}_G$ at the normal distribution with covariance S that is given by Proposition 3.3.4 (1) in connection with Proposition 3.4.2 to the one given by formula (5.50), p. 149, in Lauritzen (1996). ∎

As an intermediate step towards the proof of Proposition 3.3.6 we note the next lemma.

Lemma 3.5.1 *Consider the situation of Proposition 3.3.6 (2). Using the notation* $\mathbb{X}_n = (X_1, ..., X_n)^T$ *and* $\mathbb{X}^{(m)}_n = (X^{(m)}_1, ..., X^{(m)}_n)^T$ *we have*

$$\sqrt{n}\,\mathrm{vec}\left(\hat{S}_n(\mathbb{X}^{(n)}_n) - \eta S\right) \xrightarrow{\mathscr{L}} N_{p^2}\left(\eta(B + cS),\, \eta^2 W_S(\sigma_1, \sigma_2)\right),$$

where $W_S(\sigma_1, \sigma_2)$ *is as in Assumption 3.3.2 and*

$$c = \lim_{n\to\infty} \sqrt{n}\left(\xi(S^{\frac{1}{2}}_n S^{-1} S^{\frac{1}{2}}_n) - 1\right) = \mathbb{D}\xi(I_p)\,\mathrm{vec}\,(S^{-\frac{1}{2}} B S^{-\frac{1}{2}}).$$

Proof of Lemma 3.5.1. The $X_k^{(m)}$, $k \in \mathbb{N}$, are independent, and $X_k^{(m)} \stackrel{\mathscr{L}}{=} S_m^{\frac{1}{2}} S^{-\frac{1}{2}} X_k$, hence $\mathbb{X}_n^{(m)} \stackrel{\mathscr{L}}{=} \mathbb{X}_n S^{-\frac{1}{2}} S_m^{\frac{1}{2}}$. We conclude from Assumption 3.3.1 that

$$\hat{S}_n(\mathbb{X}_n^{(m)}) \stackrel{\mathscr{L}}{=} \xi(S_m^{\frac{1}{2}} S^{-1} S_m^{\frac{1}{2}}) S_m^{\frac{1}{2}} S^{-\frac{1}{2}} \hat{S}_n(\mathbb{X}_n) S^{-\frac{1}{2}} S_m^{\frac{1}{2}}.$$

For brevity let $\xi_n = \xi(S_n^{\frac{1}{2}} S^{-1} S_n^{\frac{1}{2}})$. Then

$$\begin{aligned} \sqrt{n}\,\mathrm{vec}\left(\hat{S}_n(\mathbb{X}_n^{(n)}) - \eta\xi_n S_n\right) &\stackrel{\mathscr{L}}{=} \xi_n\,\mathrm{vec}\left(S_n^{\frac{1}{2}} S^{-\frac{1}{2}}\left[\sqrt{n}(\hat{S}_n(\mathbb{X}_n) - \eta S)\right] S^{-\frac{1}{2}} S_n^{\frac{1}{2}}\right) \\ &\stackrel{\mathscr{L}}{\longrightarrow} N_{p^2}\left(0,\, \eta^2 W_S(\sigma_1, \sigma_2)\right) \end{aligned}$$

follows from Assumption 3.3.2 by Slutsky's lemma. Finally

$$\sqrt{n}\,\mathrm{vec}\left(\hat{S}_n(\mathbb{X}_n^{(n)}) - \eta S\right) = \sqrt{n}\,\mathrm{vec}\left(\hat{S}_n(\mathbb{X}_n^{(n)}) - \eta\xi_n S_n\right) + \eta\sqrt{n}\,(\xi_n - 1)\,\mathrm{vec}\, S_n + \eta\sqrt{n}\,\mathrm{vec}\,(S_n - S)$$

$$\stackrel{\mathscr{L}}{\longrightarrow} N_{p^2}\left(0,\, \eta^2 W_S(\sigma_1, \sigma_2)\right) + \eta c\,\mathrm{vec}\, S + \eta B = N_{p^2}\left(\eta\,\mathrm{vec}\,(B + cS),\, \eta^2 W_S(\sigma_1, \sigma_2)\right).$$

The existence of the limit $c = \lim_{n\to\infty} \sqrt{n}(\xi_n - 1)$ follows from the continuous differentiability of the function ξ. By means of the first order Taylor expansion of $\xi(S_n^{\frac{1}{2}} S^{-1} S_n^{\frac{1}{2}})$ around I_p the limit can be identified as $c = \mathbb{D}\xi(I_p)\,\mathrm{vec}\,(S^{-\frac{1}{2}} B S^{-\frac{1}{2}})$. ∎

Proof of Proposition 3.3.6. Part (1): Since $S^{-1} \in \mathscr{S}_p^+(G_0) \subset \mathscr{S}_p^+(G_1)$, we have $S_{G_1} = h_{G_1}(S) = S$, and by Corollary 3.3.5 (3)

$$\sqrt{n}\,\mathrm{vec}\left(\hat{P}_{G_1} - P\right) \stackrel{\mathscr{L}}{\longrightarrow} N_{p^2}\left(0, 2\sigma_1 R_{G_1}(S)\right), \tag{3.15}$$

and since $Q_{0,1}\mathrm{vec}\, P = 0$,

$$\sqrt{n} Q_{0,1}\mathrm{vec}\ \hat{P}_{G_1} \stackrel{\mathscr{L}}{\longrightarrow} N_{q_{0,1}}\left(0, 2\sigma_1 Q_{0,1} R_{G_1}(S) Q_{0,1}^T\right).$$

The mapping $S \mapsto R_{G_1}(S)$ is almost surely continuous, hence by the continuous mapping theorem and Slutsky's lemma

$$\sqrt{\frac{n}{2\sigma_1}}\left(Q_{0,1} R_{G_1}(\hat{S}_n) Q_{0,1}^T\right)^{-\frac{1}{2}} Q_{0,1}\mathrm{vec}\ \hat{P}_{G_1} \stackrel{\mathscr{L}}{\longrightarrow} N_{q_{0,1}}(0, I_{q_{0,1}}),$$

and, again by the continuous mapping theorem, we conclude $\frac{1}{\sigma_1}\hat{T}_n(G_0, G_1) \stackrel{\mathscr{L}}{\longrightarrow} \chi^2_{q_{0,1}}$.
Part (2): In analogy to (3.15) we obtain from Lemma 3.5.1:

$$\sqrt{n}\,\mathrm{vec}\left(\hat{P}_{G_1} - P\right) \stackrel{\mathscr{L}}{\longrightarrow} N_{p^2}\Big(\Gamma(S)\Omega_{G_1}(S)\,\mathrm{vec}\, B,\ 2\sigma_1 R_{G_1}(S)\Big). \tag{3.16}$$

Note that $\Gamma(S)\Omega_{G_1}(S)\,\mathrm{vec}\, S = 0$. From (3.16) we proceed as from formula (3.15) in part (1) above to obtain the stated convergence result. ∎

Towards the proof of Proposition 3.3.9 we state Lemmas 3.5.2 to 3.5.4. For $A \in \mathscr{S}_p^+$ let $f_A : \mathscr{S}_p^+ \to \mathbb{R}$:

$$f_A(B) = \ln\det B + \mathrm{tr}(B^{-1}A).$$

From the theory of Gaussian graphical models we know that for any graph G and $A \in \mathscr{S}_p^+$ the matrix $A_G = h_G(A)$ is the unique solution of the constrained optimization problem

$$\begin{cases} \text{minimize} & f_A(B) \\ \text{subject to} & Q_{D(G)} \operatorname{vec} h(B) = 0,\ B \in \mathscr{S}_p^+, \end{cases} \tag{3.17}$$

because A_G is the maximum likelihood estimate of the covariance matrix under the model G at a multivariate normal distribution, if A is the observed sample covariance, cf. Lauritzen (1996, p. 133). Now consider, as in Section 3.3.3, two nested graphs $G_0 = (V, E_0)$ and $G_1 = (V, E_1)$ with $V = \{1, ..., p\}$ and $E_0 \subsetneq E_1$, and let $H_0(\cdot) = Q_{D(G_0)} \operatorname{vec} h(\cdot)$, $H_1(\cdot) = Q_{D(G_1)} \operatorname{vec} h(\cdot)$ and $H_{0,1}(\cdot) = Q_{D(G_0)\setminus D(G_1)} \operatorname{vec} h(\cdot)$.

Lemma 3.5.2 *$A_{G_0} = h_{G_0}(A)$ is a solution of the constrained optimization problem*

$$\begin{cases} \textit{minimize} & f_{A_{G_1}}(h_{G_1}(C)) \\ \textit{subject to} & H_{0,1}(h_{G_1}(C)) = 0,\ C \in \mathscr{S}_p^+. \end{cases} \tag{3.18}$$

The solution is in general not unique.

Proof. It follows from (3.17) and the defining equations (3.5) that A_{G_0} uniquely solves the constrained OP

$$\begin{cases} \text{minimize} & f_{A_{G_1}}(B) \\ \text{subject to} & H_0(B) = 0,\ B \in \mathscr{S}_p^+. \end{cases} \tag{3.19}$$

The restriction $H_0(B) = 0$ is equivalent to

$$H_1(B) = 0 \qquad \text{and} \qquad H_{0,1}(B) = 0,$$

and any matrix B with $H_1(B) = 0$ can be written as $B = h_{G_1}(C)$ for some $C \in \mathscr{S}_p^+$. Thus the sets $\mathscr{B} = \left\{ B \,\middle|\, H_0(B) = 0, B \in \mathscr{S}_p^+ \right\}$ and $\mathscr{C} = \left\{ B = h_{G_1}(C) \,\middle|\, H_{0,1}(h_{G_1}(C)) = 0, C \in \mathscr{S}_p^+ \right\}$ are equal, and so are hence the solution sets of the constrained OPs (3.19) and

$$\begin{cases} \text{minimize} & f_{A_{G_1}}(B) \\ \text{subject to} & B \in \mathscr{C}. \end{cases} \tag{3.20}$$

Thus A_{G_0} uniquely solves (3.20), and all matrices $C \in \mathscr{S}_p^+$ with $h_{G_1}(C) = A_{G_0}$, among them A_{G_0}, solve (3.18). ∎

The following asymptotic equivalence will be used in the proof of Proposition 3.3.9.

Lemma 3.5.3 *Let $H : \mathbb{R}^{p\times p} \to \mathbb{R}^q$ be continuously differentiable. Then, under Assumption 3.3.8,*

$$\sqrt{n}\left(H(\hat{S}_{G_0}) - H(\hat{S}_{G_1})\right) \overset{a}{\sim} \sqrt{n}\, \mathbb{D}H(\hat{S}_{G_0}) \operatorname{vec}\left(\hat{S}_{G_0} - \hat{S}_{G_1}\right).$$

Proof. The sequences $\sqrt{n}(\hat{S}_{G_0} - \eta S)$ and $\sqrt{n}(\hat{S}_{G_1} - \eta S)$ converge in distribution, and so do hence $\sqrt{n}(H(\hat{S}_{G_0}) - H(\eta S))$ and $\sqrt{n}(H(\hat{S}_{G_1}) - H(\eta S))$. We expand $H(\hat{S}_{G_0})$ and $H(\hat{S}_{G_1})$ both around $H(\eta S)$ to obtain

$$\sqrt{n}\left(H(\hat{S}_{G_0}) - H(\hat{S}_{G_1})\right) \overset{a}{\sim} \sqrt{n}\, \mathbb{D}H(\eta S) \operatorname{vec}\left(\hat{S}_{G_0} - \hat{S}_{G_1}\right) \overset{a}{\sim} \sqrt{n}\, \mathbb{D}H(\hat{S}_{G_0}) \operatorname{vec}\left(\hat{S}_{G_0} - \hat{S}_{G_1}\right).$$

The last equivalence holds because $\mathbb{D}H$ is continuous. ∎

The following derivatives are stated without proof. Expressions (3.22) and (3.23) can be deduced from the proofs of Propositions 3.3.3 and 3.3.4, and (3.21) can be assembled from the standard derivatives given in MN.

Lemma 3.5.4 *For $A, B \in \mathscr{S}_p^+$,*

$$\begin{aligned}
\mathbb{D}f_A(B) &= \operatorname{vec}(B-A)^T(B^{-1}\otimes B^{-1})M_p, && (3.21)\\
\mathbb{D}h_G(B) &= (h_G(B)\otimes h_G(B))\,\Omega_G(B)M_p, && (3.22)\\
\mathbb{D}H_{0,1}(B) &= Q_{0,1}\Gamma(B)\left(B^{-1}\otimes B^{-1}\right)M_p. && (3.23)
\end{aligned}$$

Proof of Proposition 3.3.9. The second order Taylor expansion of $\ln\det(\cdot)$ is

$$\ln\det(A+X) \;=\; \ln\det A \;+\; \left(\operatorname{vec}(A^T)^{-1}\right)^T \operatorname{vec} X \;-\; \frac{1}{2}\left(\operatorname{vec}(X^T)\right)^T\left((A^T)^{-1}\otimes A^{-1}\right)\operatorname{vec} X \;+\; o(\|X\|^2),$$

cf. MN pp. 108, 179, 184. Applying this to the deviance test statistic yields

$$\begin{aligned}
\hat{D}_n(G_0,G_1) &= n\left(\ln\det(\hat{S}_{G_0}) - \ln\det(\hat{S}_{G_1})\right) \;=\; -n\ln\det\left(\hat{S}_{G_1}\hat{S}_{G_0}^{-1}\right)\\
&= -n\operatorname{tr}\left(\hat{S}_{G_1}\hat{S}_{G_0}^{-1} - I_p\right) \;+\; \frac{n}{2}\operatorname{tr}\left(\left(\hat{S}_{G_1}\hat{S}_{G_0}^{-1} - I_p\right)^2\right) \;+\; o\left(n\|\hat{S}_{G_1}\hat{S}_{G_0}^{-1} - I_p\|^2\right)\\
&\overset{a}{\sim} \frac{n}{2}\left(\operatorname{vec}\left(\hat{S}_{G_1} - \hat{S}_{G_0}\right)\right)^T\left(\hat{S}_{G_0}^{-1}\otimes\hat{S}_{G_0}^{-1}\right)\operatorname{vec}\left(\hat{S}_{G_1} - \hat{S}_{G_0}\right). && (3.24)
\end{aligned}$$

The asymptotic equivalence follows because

- $\operatorname{tr}\left(\hat{S}_{G_1}\hat{S}_{G_0}^{-1} - I_p\right) \;=\; \left(\operatorname{vec}\left(\hat{S}_{G_1} - \hat{S}_{G_0}\right)\right)^T \operatorname{vec}\hat{S}_{G_0}^{-1} \;=\; 0$, which is a consequence of equations (3.5),
- $\frac{n}{2}\operatorname{tr}\left(\left(\hat{S}_{G_1}\hat{S}_{G_0}^{-1} - I_p\right)^2\right) \;=\; \frac{n}{2}\left(\operatorname{vec}\left(\hat{S}_{G_1} - \hat{S}_{G_0}\right)\right)^T\left(\hat{S}_{G_0}^{-1}\otimes\hat{S}_{G_0}^{-1}\right)\operatorname{vec}\left(\hat{S}_{G_1} - \hat{S}_{G_0}\right)$ and
- $n\|\hat{S}_{G_1}\hat{S}_{G_0}^{-1} - I_p\|^2 \;\le\; \left(\sqrt{n}\|\hat{S}_{G_1} - \eta S\| + \sqrt{n}\|\hat{S}_{G_0} - \eta S\|\right)^2\|\hat{S}_{G_0}^{-1}\|^2 = O_P(1).$

Applying Lemma 3.5.3 to $H = h_{G_1}$ and using (3.22) we find further

$$\sqrt{n}\operatorname{vec}\left(\hat{S}_{G_0} - \hat{S}_{G_1}\right) \;\overset{a}{\sim}\; \sqrt{n}\left(\hat{S}_{G_0}\otimes\hat{S}_{G_0}\right)\Omega_{G_1}(\hat{S}_{G_0})M_p\operatorname{vec}\left(\hat{S}_{G_0} - \hat{S}_{G_1}\right)$$

and from (3.24)

$$\hat{D}_n(G_0,G_1) \;\overset{a}{\sim}\; \frac{n}{2}\left(\operatorname{vec}\left(\hat{S}_{G_1} - \hat{S}_{G_0}\right)\right)^T M_p\Omega_{G_1}(\hat{S}_{G_0})\operatorname{vec}\left(\hat{S}_{G_1} - \hat{S}_{G_0}\right). \qquad (3.25)$$

Next we introduce the Lagrange multiplier, cf. MN p. 131. Since $\hat{S}_{G_0}$ solves the constrained OP (3.18) with $A = \hat{S}_n$, there is a $\lambda \in \mathbb{R}^{q_{0,1}}$ such that

$$\mathbb{D}\left(f_{\hat{S}_{G_1}}\circ h_{G_1}\right)\left(\hat{S}_{G_0}\right) \;=\; \lambda^T\mathbb{D}\left(H_{0,1}\circ h_{G_1}\right)\left(\hat{S}_{G_0}\right),$$

which transforms to, cf. Lemma 3.5.4,

$$M_p\Omega_{G_1}(\hat{S}_{G_0})\operatorname{vec}\left(\hat{S}_{G_1} - \hat{S}_{G_0}\right) \;=\; M_p\Omega_{G_1}(\hat{S}_{G_0})\Gamma(\hat{S}_{G_0})^T Q_{0,1}^T\lambda.$$

We left-multiply both sides by $(\hat{S}_{G_0}^{\frac{1}{2}}\otimes\hat{S}_{G_0}^{\frac{1}{2}})$ and solve for λ.

$$\begin{aligned}
&M_p\Omega_{G_1}(\hat{S}_{G_0})\operatorname{vec}\left(\hat{S}_{G_1} - \hat{S}_{G_0}\right)\\
&= M_p\Omega_{G_1}(\hat{S}_{G_0})\Gamma(\hat{S}_{G_0})^T Q_{0,1}^T\left[Q_{0,1}R_{G_1}(\hat{S}_{G_0})Q_{0,1}^T\right]^{-1}Q_{0,1}\Gamma(\hat{S}_{G_0})M_p\Omega_{G_1}(\hat{S}_{G_0})\operatorname{vec}\left(\hat{S}_{G_1} - \hat{S}_{G_0}\right).
\end{aligned}$$

We substitute the right-hand side for the left-hand side in (3.25), apply again Lemma 3.5.3, this time to $H = H_{0,1} \circ h_{G_1}$, which leads to

$$\sqrt{n} Q_{0,1} \operatorname{vec} \hat{P}_{G_1} \overset{a}{\sim} \sqrt{n} Q_{0,1} \Gamma(\hat{S}_{G_0}) M_p \Omega_{G_1}(\hat{S}_{G_0}) \operatorname{vec} \left(\hat{S}_{G_1} - \hat{S}_{G_0} \right),$$

and obtain

$$\hat{D}_n(G_0, G_1) \overset{a}{\sim} \frac{n}{2} (\operatorname{vec} \hat{P}_{G_1})^T Q_{0,1}^T \left[Q_{0,1} R_{G_1}(\hat{S}_{G_0}) Q_{0,1}^T \right]^{-1} Q_{0,1} \operatorname{vec} \hat{P}_{G_1}.$$

The last step is to note that $R_{G_1}(\hat{S}_{G_0}) \overset{a}{\sim} R_{G_1}(\hat{S})$, since both sides converge to $R_{G_1}(S)$. ∎

Chapter 4

Elliptical graphical modelling — the non-decomposable case

4.1 The results

In the previous chapter we have analyzed estimators $\hat{S}_G = h_G(\hat{S}_n)$ for decomposable models G, using the representation (3.6) of h_G, which is valid for decomposable G. In this chapter we treat the general case, where $G = (V, E)$ may be any graph, and derive analogues of Propositions 3.3.4 (asymptotic distribution of $\hat{S}_G$), 3.3.6 (Wald-type test) and 3.3.9 (deviance test). The corresponding results for general G are Corollary 4.1.3, Proposition 4.1.8 and Proposition 4.1.9, respectively.
We use only the definition of h_G, cf. (3.5),

$$h_G : \mathscr{S}_p^+ \to \mathscr{S}_p^+ : A \mapsto A_G,$$

where

$$\begin{cases} [A_G]_{i,j} = a_{i,j}, & \{i,j\} \in E \quad \text{or} \quad i = j, \\ [A_G^{-1}]_{i,j} = 0, & \{i,j\} \notin E \quad \text{and} \quad i \neq j. \end{cases} \tag{4.1}$$

and the knowledge that $h_G(A)$ is uniquely defined and positive definite for any $A \in \mathscr{S}_p^+$, or in other words, that such a function h_G exists, cf. Lauritzen (1996, p. 133). The main tool of the proofs is the implicit function theorem, cf. Forster (1982, pp. 66-81). The chapter has two sections: The remainder of Section 4.1 states the main results, divided into results on estimation (Subsection 4.1.1) and tests (Subsection 4.1.2). Section 4.2 contains their derivations in detail. We make use of the notation that is introduced in Sections 3.1 to 3.3.

4.1.1 Constrained estimation

Let $G = (V, E)$ be an arbitrary, potentially non-decomposable, graph with $V = \{1, ..., p\}$ and q missing edges. Related to G we define the matrices

$$Q = Q_{D(G)},$$

where $D(G)$ and $Q_{D(G)}$ are defined in Section 3.3.3, and

$$I_{p^2,G} = \text{diag}(d_{1,1}, ..., d_{1,p}, \ldots, d_{p,1}, ..., d_{p,p}) \in \mathbb{R}^{p^2}$$

with

$$d_{i,j} = \begin{cases} 1 & \text{if } \{i,j\} \in E \quad \text{or} \quad i = j, \\ 0 & \text{if } \{i,j\} \notin E \quad \text{and} \quad i \neq j. \end{cases}$$

In words, $I_{p^2,G}$ is the identity matrix with those rows that correspond to non-edges in G put to zero.

Proposition 4.1.1 *The function h_G is differentiable. Its derivative is*

$$\mathbb{D}h_G(A) = \left(I_{p^2} \; - \; M_p Q^T \left[Q M_p (A_G^{-1} \otimes A_G^{-1}) Q^T \right]^{-1} Q (A_G^{-1} \otimes A_G^{-1}) \right) M_p I_{p^2,G} \tag{4.2}$$

for $A \in \mathscr{S}_p^+$ and $A_G = h_G(A)$.

Proposition 4.1.2 *For $A \in \mathscr{S}_p$,*

(1) $\mathbb{D}h_G(A)$ *has rank* $\frac{p(p+1)}{2} - q$,

(2) $\mathbb{D}h(A) = \Gamma(A)(A^{-1} \otimes A^{-1})M_p$ *has rank* $\frac{p(p-1)}{2}$,
where h is defined in (3.2), and

(3) $\mathbb{D}\left(h \circ h_G\right)(A)$ *has rank* $\frac{p(p-1)}{2} - q$.

Corollary 4.1.3 *Let $\hat{S}_n$ be a sequence of positive definite random $p \times p$ matrices satisfying $\sqrt{n}\,\mathrm{vec}(\hat{S}_n - S) \xrightarrow{\mathscr{L}} N$ for some random variable N in $\mathbb{R}^{p^2}$. Then*

(1) $\hat{S}_G = h_G(\hat{S}_n)$ *fulfils* $\hat{S}_G \xrightarrow{p} S_G$ *and* $\sqrt{n}\,\mathrm{vec}(\hat{S}_G - S_G) \xrightarrow{\mathscr{L}} \mathbb{D}h_G(S)N$
with $S_G = h_G(S)$,

(2) $\hat{K}_G = \hat{S}_G^{-1}$ *fulfils* $\hat{K}_G \xrightarrow{p} K_G$ *and* $\sqrt{n}\,\mathrm{vec}(\hat{K}_G - K_G) \xrightarrow{\mathscr{L}} -(K_G \otimes K_G)\mathbb{D}h_G(S)N$
with $K_G = S_G^{-1}$, *and*

(3) $\hat{P}_G = h(\hat{S}_G)$ *fulfils* $\hat{P}_G \xrightarrow{p} P_G$ *and* $\sqrt{n}\,\mathrm{vec}(\hat{P}_G - P_G) \xrightarrow{\mathscr{L}} \Gamma(S_G)(K_G \otimes K_G)\mathbb{D}h_G(S)N$,
where $\Gamma(\cdot)$ is defined in (3.4).

Remark 4.1.4 In a personal communication David E. Tyler proves the following theorem.
Let $\hat{S}_n$ be as in Corollary 4.1.3, furthermore $H : \mathbb{R}^{p\times p} \to \mathbb{R}^q$ a continuously differentiable function with $H(S) = 0$ and $\mathrm{rank}(\mathbb{D}H(S)) = q$. *Then the random set*

$$\mathscr{B}_n = \arg\min \left\{ \log\det(B) + \mathrm{tr}(B^{-1}\hat{S}_n) \,\middle|\, B \in \mathscr{S}_p^+, H(B) = 0 \right\}$$

is non empty, and for any sequence of random variables B_n, where B_n lies almost surely in $\mathscr{B}_n$, $n \in \mathbb{N}$,

$$\sqrt{n}\,\mathrm{vec}(B_n - S) \xrightarrow{\mathscr{L}} \left(I_{p^2} - M_H(S)\right)N,$$

where $M_H(S) = (S \otimes S)\mathbb{D}H(S)^T \left[\mathbb{D}H(S)(S \otimes S)\mathbb{D}H(S)^T\right]^{-1} \mathbb{D}H(S)$.

Applying this in the graphical modelling context to the function $H(S) = Q\,\mathrm{vec}(S^{-1})$ yields

$$\sqrt{n}\,\mathrm{vec}(\hat{S}_G - S) \xrightarrow{\mathscr{L}} \Psi_G(S)N, \tag{4.3}$$

for any S with $H(S) = 0$, where

$$\Psi_G(A) \;=\; I_{p^2} \;-\; M_p Q^T \left[QM_p(A^{-1} \otimes A^{-1})Q^T\right]^{-1} Q(A^{-1} \otimes A^{-1})M_p. \tag{4.4}$$

Corollary 4.1.3 (1) generalizes (4.3) to any $S \in \mathscr{S}_p^+$. One apparent difference between expressions (4.4) and (4.2) is the matrix $I_{p^2,G}$ in the formula (4.2) for $\mathbb{D}h_G$. It is very reasonable that this matrix is there, since by a closer inspection of the function h_G we observe that its value $h_G(A)$ does not depend on those elements of its argument A that correspond to non-edges of G, hence the corresponding columns of the derivative $\mathbb{D}h_G(A)$ must be zero everywhere. This is exactly what right-multiplying by $I_{p^2,G}$ does: putting the rows that correspond to non-edges of G to zero.
For any S with $H(S) = 0$ ($\Leftrightarrow S^{-1} \in \mathscr{S}_p^+(G)$) we find by comparing (4.3) and 4.1.3 (1) that

$$\Psi_G(S)N \overset{\mathscr{L}}{=} \mathbb{D}h_G(S)N,$$

and since the expectation of N, apart from being symmetric, can be arbitrary, cf. Lemma 3.5.1, we have $\Psi_G(S)\,\mathrm{vec}\,C = \mathbb{D}h_G(S)\,\mathrm{vec}\,C$ for any symmetric $C \in \mathbb{R}^{p\times p}$ and hence $\Psi_G(S)M_p = \mathbb{D}h_G(S)M_p = \mathbb{D}h_G(S)$. We can deduce that for *any* $S \in \mathscr{S}_p^+$ the non-edge rows of

$$\Psi_G(S_G)M_p = \left(I_{p^2} \;-\; M_p Q^T \left[QM_p(S_G^{-1} \otimes S_G^{-1})Q^T\right]^{-1} Q(S_G^{-1} \otimes S_G^{-1})\right)M_p$$

are already zero, and $I_{p^2,G}$ may be dropped from expression (4.2).

Making use of this observation we obtain a very nice and short expression for the asymptotic variance of shape estimators under the elliptical graphical model G.

Corollary 4.1.5 *If* $\hat{S}_n$ *fulfils Assumption 3.3.2 and* $S^{-1} \in \mathscr{S}_p^+(G)$*, then*

$$\sqrt{n}\,\mathrm{vec}(\hat{S}_G - \eta S) \xrightarrow{\mathscr{L}} N_{p^2}\left(0, \eta^2 W_{S_G}(\sigma_1, \sigma_2)\right),$$

with

$$W_{S_G}(\sigma_1, \sigma_2) = 2\sigma_1 M_p\left(S \otimes S \;-\; Q^T\left[QM_p(S^{-1} \otimes S^{-1})Q^T\right]^{-1} Q\right) \;+\; \sigma_2\,\mathrm{vec}\,S\,(\mathrm{vec}\,S)^T. \tag{4.5}$$

Formula (4.5) should be compared to formula 3.3.5 (2), that gives the asymptotic covariance under the same assumptions on $\hat{S}_n$ but for decomposable G. Both formulas have been proven to be true and they describe the same quantity $W_{S_G}(\sigma_1, \sigma_2)$, but the connection between both is not at all obvious.

4.1.2 Testing

Lemma 4.1.6 *For $S \in \mathscr{S}_p^+$,*

$$R_G(S) = \mathbb{D}(h \circ h_G)(S) M_p(S \otimes S) \mathbb{D}(h \circ h_G)(S)^T. \tag{4.6}$$

has rank $\frac{p(p-1)}{2} - q$.

Considering that $R_G(S)$ is proportional to the asymptotic covariance matrix of $\operatorname{vec} \hat{P}_G$ (derived from some shape estimator $\hat{S}_n$) this is plausible: All rows (and columns) corresponding to diagonal positions of $\hat{P}_G$ and to non-edge positions are zero, since these elements of $\hat{P}_G$ do not vary. All remaining $p(p-1) - 2q$ rows appear as pairs due to the symmetry.
Let $G_0 = (V, E_0), G_1 = (V, E_1)$ be two graphs with $V = \{1, ..., p\}$, $E_0 \subsetneq E_1$ and q_0, q_1, $q_{0,1}$, Q_0, Q_1 and $Q_{0,1}$ as in Section 3.3.3.

Lemma 4.1.7 *If $\hat{S}_n$ fulfils Assumption 3.3.2, then the probability that*

$$\hat{T}_n(G_0, G_1) = \frac{n}{2} \left(\operatorname{vec} \hat{P}_{G_1}\right)^T Q_{0,1}^T \left[Q_{0,1} R_{G_1}(\hat{S}_n) Q_{0,1}^T\right]^{-1} Q_{0,1} \operatorname{vec} \hat{P}_{G_1}$$

exists converges to 1 as $n \to \infty$.

The last two propositions of this section are stated without proof. The proofs are analogous to those of Propositions 3.3.6 and 3.3.9.

Proposition 4.1.8 *Let $\hat{S}_n = \hat{S}_n(\mathbb{X}_n)$ satisfy Assumption 3.3.1 and Assumption 3.3.2 for $\mathbb{X}_n = (X_1, ..., X_n)^T$ with i.i.d. random variables $X_1, ..., X_n, ...$.*

(1) If $X_1, ..., X_n, ...$ are i.i.d. with $X_1 \sim F \in \mathscr{E}_p(\mu, S) \subset \mathscr{E}_p(G_0)$, then $\hat{T}_n(G_0, G_1) \xrightarrow{\mathscr{L}} \sigma_1 \chi^2_{q_{0,1}}$.

(2) Let $\mathbb{X}_n^{(m)} = \left(X_1^{(m)}, ..., X_n^{(m)}\right)^T$, where $X_1^{(m)}, ..., X_n^{(m)}, ...$ are i.i.d. with $X_1^{(m)} \sim F \in \mathscr{E}_p(\mu, S_m)$ and S_m is such that there exists a matrix $B \in \mathscr{S}_p$ with $\lim_{m\to\infty} \sqrt{m}(S_m - S) = B$. Then

$$\hat{T}_n(G_0, G_1) \xrightarrow{\mathscr{L}} \sigma_1 \chi^2_{q_{0,1}} \left(\frac{\delta(B, S)}{\sigma_1}\right),$$

where

$$\delta(B, S) = \frac{1}{2} v^T Q_{0,1}^T \left(Q_{0,1} R_{G_1}(S) Q_{0,1}^T\right)^{-1} Q_{0,1} v$$

with the abbreviation $v = v(B, S) = \mathbb{D}(h \circ h_G)(S) \operatorname{vec} B$.

The non-centrality parameter of the χ^2 distribution $\chi^2_r(\delta) \sim (N_r(\mu, I_r))^2$ is $\delta = \mu^T \mu$.

Proposition 4.1.9 *If $\hat{S}_n$ is a sequence of positive definite random $p \times p$ matrices such that $\sqrt{n}(\hat{S}_n - S)$ converges in distribution for some $S \in \mathscr{S}_p^+$ with $S^{-1} \in \mathscr{S}_p^+(G_0)$. Then*

$$\hat{D}_n(G_0, G_1) = n \left(\ln \det h_{G_0}(\hat{S}_n) - \ln \det h_{G_1}(\hat{S}_n)\right) \overset{a}{\sim} \hat{T}_n(G_0, G_1).$$

Proposition 4.1.9 implies that Proposition 4.1.8 remains true if $\hat{T}_n(G_0, G_1)$ is replaced by $\hat{D}_n(G_0, G_1)$.

4.2 The proofs

Initial remark. As mentioned in Section 3.5 the symmetry of the matrices generally poses some nuisance, which is not very severe once one is familiar with it. Since I am not (yet) familiar with it and neither assume the reader to be, I decided to write things down in detail in the space $\mathbb{R}^{p(p+1)/2}$.

Let $m = p(p+1)/2$ and $\mathrm{mat}_{p\times p} : \mathbb{R}^{p^2} \to \mathbb{R}^{p\times p}$ be the inverse operator to vec for $p \times p$ matrices. For a matrix $A \in \mathscr{S}_p$ let $v(A)$ be the m-vector that is obtained by deleting the super-diagonal elements of A from $\mathrm{vec}\, A$. The *duplication matrix*, cf. Magnus and Neudecker (1999, p. 49) $D_p \in \mathbb{R}^{p^2\times m}$ is the matrix that maps $v(A)$ to $\mathrm{vec}\, A$. It has exactly one 1-entry in each row and is zero otherwise. Its Moore-Penrose inverse D_p^+ then reduces $\mathrm{vec}\, A$ to $v(A)$ for any *symmetric* matrix $A \in \mathbb{R}^{p\times p}$. We will henceforth identify $A \in \mathscr{S}_p$ with $v(A) \in \mathbb{R}^m$. We denote the inverse function of $v : \mathscr{S}_p \to \mathbb{R}^m : A \mapsto D_p^+ \mathrm{vec}\, A$ by

$$\theta : \mathbb{R}^m \to \mathscr{S}_p : a \mapsto \mathrm{mat}_{p\times p}\, D_p a.$$

I try to stick to the following notational convention: For a function φ defined on $\mathscr{S}_p$, taking, say, the symmetric matrix A as argument, the corresponding function working on $\{\, a \mid \mathrm{mat}_{p\times p}(a) \in \mathscr{S}_p\} \subset \mathbb{R}^{p^2}$, which takes then $\mathrm{vec}\, A$ as argument, goes under the same name (as it is always done when we compute derivatives of matrix functions) The corresponding function defined on $\mathbb{R}^m$ applying to $v(A)$ shall be denoted by $\bar{\varphi}$.
Recall the notation introduced in Section 3.3.3, in particular the set Π_p and the ordering $\prec_p$. For a graph $G = (V, E)$ with p vertices and q absent edges let $Q_{D(G)}$ and $I_{p^2,G}$ be defined as in Section 4.1, likewise $Q_{K(G)}$, cf. Section 3.3.3, where

$$K(G) = \{\, (i,j) \mid 1 \le i < j \le p, \{i,j\} \in E\,\} \;\cup\; \{\, (i,i) \mid 1 \le i \le p\,\} \;\subset\; \Pi_p.$$

The set $K(G)$ gathers all matrix positions on the diagonal and all sub-diagonal positions that correspond to edges of G. The matrix $Q_{D(G)\cup K(G)} = Q_{K(G)\cup D(G)}$ sends $\mathrm{vec}\, A$ to $v(A)$ for any $A \in \mathbb{R}^{p\times p}$. Furthermore, let $\bar{Q}_{D(G)} = Q_{D(G)} D_p$ and $\bar{Q}_{K(G)} = Q_{K(G)} D_p$.
For vectors a and b that correspond to distinct subsets of matrix positions define the concatenation $c = (a; b)$ in such a way that its elements are ordered according to $\prec_p$, cf. Section 3.3.3, i.e. if we can write $a = Q_C \,\mathrm{vec}\, A$ and $b = Q_D \,\mathrm{vec}\, A$ for some matrix $A \in \mathbb{R}^{p\times p}$ and distinct sets $C, D \subset \Pi_p$, then $(a; b) = Q_{C\cup D} \,\mathrm{vec}\, A$.

Lemma 4.2.1 *The set* $U = \left\{ x \in \mathbb{R}^m \,\middle|\, \theta(x) \in \mathscr{S}_p^+ \right\}$ *is open in* $\mathbb{R}^m$.

Proof. Let $a \in U$, $A = \theta(a)$ and $B(p) = \{x \in \mathbb{R}^p \,|\, \|x\| = 1\}$. Then $x^T A x > 0$ for all $x \in B(p)$, and since $B(p)$ is closed, there exists an $\varepsilon > 0$ such that $x^T A x \ge \varepsilon$ for all $x \in B(p)$. Let $c \in \mathbb{R}^m$ and $C = \theta(c)$ with $\|a - c\| < \varepsilon/(2\sqrt{2})$, hence $\|A - C\| < \varepsilon/2$. Then, for $x \in B(p)$,

$$\begin{aligned} x^T C x &= x^T A x + x^T (C - A) x \;=\; x^T A x + \mathrm{tr}\left((C - A) x x^T\right) \\ &= x^T A x + \left[\mathrm{vec}(C^T - A^T)\right]^T \mathrm{vec}(x x^T) \;\ge\; x^T A x + \|\,\mathrm{vec}(C - A)\| \;\ge\; \frac{\varepsilon}{2} \end{aligned}$$

by Cauchy-Schwarz. Thus all points $c \in \mathbb{R}^m$ in an $\varepsilon/(2\sqrt{2})$-ball around a are also in U. ∎

It follows from Lemma 4.2.1 that $U_G = \{Q_{K(G)} x \,\big|\, x \in U\,\}$ is open in $\mathbb{R}^{m-q}$, since, roughly speaking, lower-dimensional projections and cuts through open sets are again open in the lower-dimensional space. We take a closer look at the function h_G and define related functions. Let

$$\bar{h}_G : U \to U : a \mapsto v\left(h_G(\theta(a))\right).$$

We observe that $\bar{h}_G(a)$ depends only on those elements of a that correspond to edges of G and define further

$$\breve{h}_G : U_G \to U : a \mapsto \bar{h}_G((a;b)),$$

where $b \in \mathbb{R}^q$ may be any vector such that $(a;b) \in U$. We furthermore observe that $\breve{h}_G(a) \in \mathbb{R}^m$ contains all components of its argument $a \in \mathbb{R}^{m-q}$. It is the other q components that we are interested. Let

$$t_G : U_G \to \mathbb{R}^q : a \mapsto \bar{Q}_{D(G)}\breve{h}_G(a).$$

The function h_G maps an unconstrained covariance estimate $\hat{\Sigma}_n$ to the corresponding constrained covariance estimate $\hat{\Sigma}_G$ under the model G. It takes $p(p+1)/2 - q$ values, p estimated variances $\hat{\sigma}_{i,i}$, $1 \le i \le p$ and $p(p-1) - q$ estimated covariances $\hat{\sigma}_{i,j}$, $\{i,j\} \in E$, and produces q new values: covariances estimates $\hat{\sigma}_{i,j}$ for $\{i,j\} \notin E$, $i \ne j$. So it is actually a function from $\mathbb{R}^{m-q}$ to $\mathbb{R}^q$ and may be reduced to the function t_G defined above.
We want to apply the implicit function theorem, precisely Satz 1, p. 68, and Satz 2, p. 71, in Forster (1982), to the function

$$\bar{H} : U \to \mathbb{R}^q : a \mapsto Q_{D(G)} \operatorname{vec}\left(\theta(a)^{-1}\right) \tag{4.7}$$

in the role of F in Satz 1 and Satz 2. Note that $\theta(a)^{-1}$ means matrix inversion, not inverse function, for which we would write $\nu(A)$ in this situation. For $a \in U \subset \mathbb{R}^m$ let

$$a_K = \bar{Q}_{K(G)}a \in \mathbb{R}^{m-q} \qquad \text{and} \qquad a_D = \bar{Q}_{D(G)}a \in \mathbb{R}^q.$$

Then, following our convention, $a = (a_K; a_D)$. Furthermore, let

$$\frac{\partial \bar{H}}{\partial x_D}(a_K; a_D) \in \mathbb{R}^{q \times q}$$

denote the matrix of all partial derivatives of $\bar{H}$ with respect to those components of its argument that are picked up by $\bar{Q}_{D(G)}$, evaluated at $a = (a_K; a_D)$, likewise $\partial \bar{H}/\partial x_K(a_K; a_D)$. The only assumption of Satz 2 that still needs to be checked is that $\partial \bar{H}/\partial x_D\,(a_K; a_D)$ is invertible.

Lemma 4.2.2

$$\frac{\partial \bar{H}}{\partial x_D}(a_K; a_D) = -Q_{D(G)}(A^{-1} \otimes A^{-1})D_p\bar{Q}_{D(G)}^T,$$

where $A = \theta(a)$, has full rank.

Proof. The derivate is a consequence of the chain rule applied to $\bar{H}$ given by (4.7). The proof of the invertibility consists of four steps.

(1) Since $a \in U$, $A = \theta(a) \in \mathscr{S}_p^+$ and $A^{-1} \otimes A^{-1}$ has full rank. Its columns are linearly independent.

(2) Each column of $(A^{-1} \otimes A^{-1})D_p \in \mathbb{R}^{p^2 \times m}$ is either a column of $(A^{-1} \otimes A^{-1})$ or the sum of two of its columns. In any case, the columns of $(A^{-1} \otimes A^{-1})D_p$ are linear combinations of mutually distinct sets of columns of $(A^{-1} \otimes A^{-1})$. Hence $(A^{-1} \otimes A^{-1})D_p$ has full column rank m.

(3) Right-multiplying $B = (A^{-1} \otimes A^{-1})D_p$ by $\bar{Q}_{D(G)}^T$ selects q out of the m linearly independent columns of B. Hence $B\bar{Q}_{D(G)}^T \in \mathbb{R}^{p^2 \times q}$ has full column rank q.

(4) $B\bar{Q}_{D(G)} \in \mathbb{R}^{p^2 \times q}$ has row rank q, hence picking any q of them, i.e. left-multiplying by $Q_{D(G)}$, yields a matrix with linearly independent rows. ∎

Proof of Proposition 4.2. The proof is divided into two parts: proof of differentiability and computation of the derivative.

Part I: differentiability.

Let $a \in U_G$ be fixed. Since h_G is well defined, there exists a unique $b \in \mathbb{R}^q$ such that $\bar{H}(a; b) = 0$ and $(a; b) \in U$. By Lemma 4.2.2,

$$\frac{\partial \bar{H}}{\partial x_D}(a; b)$$

is invertible and by Satz 2 (Forster, 1982, p. 71), there exists a continuous function

$$t_a : U_a \to \mathbb{R}^q,$$

defined on some open neighbourhood U_a of a with $U_a \subset U_G$ such that

$$t_a(a) = b \qquad \text{and} \qquad \bar{H}(x; t_a(x)) = 0 \quad \text{for all } x \in U_a.$$

By Satz 1 (Forster, 1982, p. 68), t_a is differentiable. Since t_G is the unique function defined on U_G that satisfies

$$\bar{H}(x; t_G(x)) = 0 \quad \text{for all } x \in U_G,$$

we have

$$t_a = t_G|_{U_a}.$$

This holds true for every $a \in U_G$, hence t_G is continuous and differentiable.

Part II: the derivative.

Let $a \in U_G$ and $A_G = \theta(a; t_G(a))$. By Bemerkung 1 (Forster, 1982, p. 71),

$$\mathbb{D}t_G(a) = -\left[\frac{\partial \bar{H}}{\partial x_D}(a; t_G(a))\right]^{-1} \frac{\partial \bar{H}}{\partial x_K}(a; t_G(a)).$$

We have

$$\frac{\partial \bar{H}}{\partial x_D}(a; t_G(a)) = -Q_{D(G)}(A_G^{-1} \otimes A_G^{-1})D_p\bar{Q}_{D(G)}^T,$$

$$\frac{\partial \bar{H}}{\partial x_K}(a; t_G(a)) = -Q_{D(G)}(A_G^{-1} \otimes A_G^{-1})D_p\bar{Q}_{K(G)}^T$$

and hence

$$\mathbb{D}t_G(a) = -\left[Q_{D(G)}(A_G^{-1} \otimes A_G^{-1})D_p\bar{Q}_{D(G)}^T\right]^{-1} Q_{D(G)}(A_G^{-1} \otimes A_G^{-1})D_p\bar{Q}_{K(G)}^T.$$

Next we obtain the derivative of $\check{h} : U_G \to \mathbb{R}^m : a \mapsto (a; t_G(a))$:

$$\mathbb{D}\check{h}(a) = \left(I_m - \bar{Q}_{D(G)}^T \left[Q_{D(G)}(A_G^{-1} \otimes A_G^{-1})D_p\bar{Q}_{D(G)}^T\right]^{-1} Q_{D(G)}(A_G^{-1} \otimes A_G^{-1})D_p\right)\bar{Q}_{K(G)}^T$$

by noting that

$$\bar{Q}_{D(G)}\mathbb{D}\check{h}(a) = \mathbb{D}t_G(a) \in \mathbb{R}^{q\times(m-q)},$$

$$\bar{Q}_{K(G)}\mathbb{D}\check{h}(a) = I_{m-q} \in \mathbb{R}^{(m-q)\times(m-q)},$$

which gives

$$\mathbb{D}\check{h}(a) = P^T \begin{pmatrix} I_{m-q} \\ \hdashline \mathbb{D}t_G(a) \end{pmatrix}, \qquad \text{where} \quad P = \begin{pmatrix} \bar{Q}_{K(G)} \\ \hdashline \bar{Q}_{D(G)} \end{pmatrix} \in \mathbb{R}^{m\times m}$$

is a permutation matrix and hence orthogonal. Recall that left-multiplying by a permutation matrix permutes the rows, right-multiplying the columns.
In the following $a = (a_K; a_D)$ denotes an element of U, thus a_K taking the role of a. As the next step we compute the derivative of $\bar{h}_G : U \to U : a \mapsto \bar{h}_G(a) = \check{h}_G(a_K)$. We have

$$\frac{\partial\bar{h}_G}{\partial x_K}(a_K; a_D) = \mathbb{D}\check{h}_G(a_K) \qquad \text{and} \qquad \frac{\partial\bar{h}_G}{\partial x_D}(a_K; a_D) = 0 \in \mathbb{R}^{m\times q},$$

hence

$$\mathbb{D}\bar{h}_G(a_K; a_D) = \left[\begin{array}{c:c} \mathbb{D}\check{h}_G(a_K) & 0 \end{array}\right] P = \mathbb{D}\check{h}_G(a_K)\bar{Q}_{K(G)}.$$

Finally, the derivative of $h_G : \mathscr{S}_p^+ \to \mathscr{S}_p^+ : A \mapsto \theta\left(\bar{h}_G(v(A))\right) = \mathrm{mat}_{p\times p} D_p\bar{h}_G(D_p^+ \operatorname{vec} A)$ is

$$\mathbb{D}h_G(A) = D_p\mathbb{D}\check{h}_G(a_K)\bar{Q}_{K(G)}D_p^+$$

$$= \left(I_{p^2} - D_p\bar{Q}_{D(G)}^T\left[Q_{D(G)}(A_G^{-1} \otimes A_G^{-1})D_p\bar{Q}_{D(G)}^T\right]^{-1} Q_{D(G)}(A_G^{-1} \otimes A_G^{-1})\right)D_p\bar{Q}_{K(G)}^T\bar{Q}_{K(G)}D_p^+$$

for any $A \in \mathscr{S}_p^+$, where $a_K = \bar{Q}_{K(G)}v(A) = Q_{K(G)} \operatorname{vec} A$ and $A_G = h_G(A) = \theta(a_K; t_G(a_K))$. This expression reduces to formula (4.2) by noting that

$$D_p\bar{Q}_{K(G)}^T\bar{Q}_{K(G)}D_p^+ = M_pI_{p^2,G} \quad \text{and} \quad D_p\bar{Q}_{D(G)}^T = 2M_pQ_{D(G)}$$

The matrix $M_pI_{p^2,G}$ is obtained from M_p by putting all rows, or equivalently all columns, to zero that correspond to non-edges of G. ∎

Proof of Proposition 4.1.2. Only part (3) is proven, since it is the only prerequisite for Lemma 4.1.6, and the other parts are treated likewise. The result is deduced from the fact that bijective, continuously differentiable functions have invertible Jacobi matrices. The main task of this proof is to construct a suitable invertible function, which is called ϕ below.
Let $L = \{(i,i)|1 \le i \le p\}$, $J(G) = K(G) \setminus L$, $\bar{Q}_L = Q_LD_p$ and $\bar{Q}_{J(G)} = Q_{J(G)}D_p$. The notation Q_L, $Q_{J(G)}$ is analogous to $Q_{D(G)}$ and is defined in Section 3.3.3. Let furthermore

$$V_G = \left\{a \in U \,\middle|\, \bar{Q}_La = 1, \bar{Q}_{D(G)}a = 0\right\} \subset \mathbb{R}^m$$

and

$$W_G = \left\{ \bar{Q}_{J(G)} a \,\middle|\, a \in V_G \right\} \subset \mathbb{R}^{\frac{p(p-1)}{2}-q}.$$

From any vector $b \in \mathbb{R}^{\frac{p(p-1)}{2}-q}$ we can construct a symmetric $p \times p$ matrix B as follows: The elements of b are put on the sub-diagonal positions of B that correspond to edges of G (in the right order according to $\prec_p$, cf. Section 3.3.3), the non-edge subdiagonal positions are filled up with zeros, the superdiagonal part is filled symmetrically, and the diagonal is filled up with ones. The set W_G gathers all vectors b for which the thus obtained matrix B is positive definite. Then define the function $\phi : U_G \to W_G \times (0,\infty)^p$:

$$\phi(a) = \left(Q_{J(G)} \operatorname{vec} \left[(A_G^{-1})_D^{-\frac{1}{2}} A_G^{-1} (A_G^{-1})_D^{-\frac{1}{2}} \right] , \; Q_L \operatorname{vec} \left(A_G^{-1} \right) \right),$$

where $A_G = \theta(a; t_G(a)) = (\theta \circ \breve{h})(a)$. Note that $(\cdot,\cdot)$ is the usual concatenation, that does not re-order the components. The function ϕ does the following: for given variance and covariance estimates: $\hat{\sigma}_{i,j}$, $\{i,j\} \in E \cup \{\{1\},...,\{p\}\}$ it determines the remaining covariance estimates $\hat{\sigma}_{i,j}$, $\{i,j\} \notin E$, $i \neq j$, as specified by the model G, inverts the thus generated matrix $\hat{\Sigma}_G$ and returns the non-zero partial correlation estimates as well as the diagonal values of $\hat{\Sigma}_G^{-1}$. The sets U_G and $W_G \times (0,\infty)^p$ are both open in $\mathbb{R}^m$, and ϕ is bijective and continuously differentiable, hence its derivative $\mathbb{D}\phi(a)$ has full rank $m - q$ for every $a \in U_G$. The first $p(p-1)/2 - q$ components of $\phi(a)$ are equal to

$$Q_{J(G)} \operatorname{vec} \left[(h \circ h_G)(A) \right]$$

for any $A \in \mathscr{S}_p^+$ with $Q_{K(G)} \operatorname{vec} A = \bar{Q}_{K(G)} v(A) = a$, hence deleting the last p rows of $\mathbb{D}\phi(a)$ gives

$$Q_{J(G)} \mathbb{D}(h \circ h_G)(A) D_p \bar{Q}_{K(G)}^T,$$

which consequently has $p(p-1)/2 - q$ linearly independent rows. The $p^2 \times (m-q)$ matrix

$$\mathbb{D}(h \circ h_G)(A) D_p \bar{Q}_{K(G)}^T$$

is obtained from $Q_{J(G)} \mathbb{D}(h \circ h_G)(A) Q_{K(G)}^T$ by duplicating all rows and adding some rows consisting entirely of zeros (corresponding to the diagonal- and non-edge positions), and has thus also rank $p(p-1)/2 - q$. Finally

$$\mathbb{D}(h \circ h_G)(A)$$

is formed from $\mathbb{D}(h \circ h_G)(A) D_p \bar{Q}_{K(G)}^T$ by duplicating some of its columns multiplied by $\frac{1}{2}$ and adding some zero-columns, which leaves the column rank unchanged. ∎

Remark. In the proof above it is not claimed that one may generally reconstruct a $p^2 \times p^2$ matrix, say M, from

$$Q_{J(G)} M D_p \bar{Q}_{K(G)}^T.$$

We have additional information about the structure of $\mathbb{D}(h \circ h_G)(A)$, which allows to do that. Basically, we know that the *relevant* information of $\mathbb{D}(h \circ h_G)(A)$ is contained in $Q_{J(G)} \mathbb{D}(h \circ h_G)(A) D_p \bar{Q}_{K(G)}^T$. Also keep in mind that $\mathbb{D}(h \circ h_G)$ denotes a derivative w.r.t. symmetric matrices, cf. Appendix A.2. Therefore the derivative of the first $p(p-1)/2 - q$ components of $\phi(a)$ is given by $Q_{J(G)} \mathbb{D}(h \circ h_G)(A) D_p \bar{Q}_{K(G)}^T$ and not by $Q_{J(G)} \mathbb{D}(h \circ h_G)(A) Q_{K(G)}^T$.

For the proof of Lemma 4.1.6 recall that

$$\mathbb{D}(h \circ h_G)(S)M_p = \mathbb{D}(h \circ h_G)(S),$$

which is generally true for derivatives w.r.t. symmetric matrices, and can be seen directly at formula (4.2). Hence $R_G(S)$ can be written shorter as

$$R_G(S) = \mathbb{D}(h \circ h_G)(S)(S \otimes S)\mathbb{D}(h \circ h_G)(S)^T.$$

The M_p in (4.6) is merely a reminder.

Proof of Lemma 4.1.6. We make use of the following. For any matrix A and any square, full rank matrix B,

(a) $B \otimes B$ is of full rank, cf. Magnus and Neudecker (1999, p. 28, Theorem 1) ,

(b) A and AA^T have the same rank, cf. MN. p. 8, (3), and

(c) A and AB have the same rank, cf. MN. p. 8, (5).

Then $S \otimes S$ is of full rank by (a) and hence also $(S \otimes S)^{\frac{1}{2}}$. By (c), $\mathbb{D}(h \circ h_G)(S)(S \otimes S)^{\frac{1}{2}}$ has rank $p(p-1)/2 - q$. With (b),

$$R_G(S) = \left(\mathbb{D}(h \circ h_G)(S)(S \otimes S)^{\frac{1}{2}}\right)\left(\mathbb{D}(h \circ h_G)(S)(S \otimes S)^{\frac{1}{2}}\right)^T$$

has the same rank. ∎

Proof of Lemma 4.1.7. From the proof of Proposition 4.1.2 it is clear that

$$Q_{J(G_1)}\mathbb{D}(h \circ h_{G_1})(S) \in \mathbb{R}^{(\frac{p(p-1)}{2}-q)\times p^2}$$

has full row rank. By applying the same argumentation as in the proof of Lemma 4.1.6 to this matrix instead of $\mathbb{D}(h \circ h_{G_1})(S)$ we find that

$$Q_{J(G_1)}R_{G_1}(S)Q_{J(G_1)}^T$$

has full rank $p(p-1)/2 - q$. From this matrix

$$Q_{0,1}R_{G_1}(S)Q_{0,1}^T$$

is obtained by simultaneously selecting certain rows and columns, hence it has also full rank. Since U is open, there is an ε-ball $U_\varepsilon(s)$ around $s = v(S)$ with $U_\varepsilon(s) \subset U$. By Assumption 3.3.2,

$$\mathbb{P}\left(v(\hat{S}_n) \in U_\varepsilon(s)\right) \to 1,$$

and the lemma follows with

$$\mathbb{P}\left(v(\hat{S}_n) \in U_\varepsilon(s)\right) \leq \mathbb{P}\left(v(\hat{S}_n) \in U\right) \leq \mathbb{P}\left(Q_{0,1}R_{G_1}(\hat{S}_n)Q_{0,1}^T \text{ has full rank }\right) \leq \mathbb{P}\left(\hat{T}_n(G_0, G_1) \text{ exists }\right).$$

∎

Chapter 5

Supplements

Summary

This chapter is a collection of six manuscripts, written over the years 2008 to 2010, that are connected in some way to the material presented in the previous chapters. They provide additional information directly about elliptical graphical modelling (e.g. Section 5.5) or treat related problems that emerged during my occupation with graphical models (e.g. Section 5.6). I want to emphasize that these manuscript supplement the thesis, they do not amend, complete or complement it. Four of the manuscripts (Sections 5.1, 5.2, 5.4 and 5.5) have appeared in conference proceedings.

Estimating partial correlations using the Oja sign covariance matrix (Section 5.1) was written in January 2008 and has appeared as

> Vogel, D., Fried, R.: Estimating partial correlations using the Oja sign covariance matrix. In: Brito, P. (ed.) Compstat 2008: Proceedings in Computational Statistics. Vol. II, pp. 721–729. Heidelberg: Physica-Verlag (2008).

When I started working on graphical models I was in particular considering the Oja sign covariance matrix and the Oja rank covariance matrix (Visuri et al., 2000), because of their intriguing property to estimate a multiple of the concentration matrix directly without inversion. This article examines the applicability of these estimators in elliptical graphical modelling.

Partial correlation estimates based on signs (Section 5.2) has appeared as

> Vogel, D., Köllmann, C., Fried, R.: Partial correlation estimates based on signs. In: Heikkonen, J. (ed.) Proceedings of the 1st Workshop on Information Theoretic Methods in Science and Engineering. TICSP series # 43 (2008)

and was written in summer 2008. When studying Oja signs the natural question arises how they relate to other multivariate generalizations of the sign function, the *marginal sign* and the *spatial sign*, and if correlation estimators based on such concepts may be of use in the context of graphical modelling. Using simple sign functions in data analysis means generally throwing away information. They are nevertheless relevant in nonparametric statistics, because they lead to distribution-free and robust methods. This paper gives an impression of the benefits of affine equivariance. Marginal and spatial sign methods turn out to be less suited in the Gaussian graphical modelling context due to their lack of affine equivariance.

When examining the spatial sign covariance matrix I learned that it has the same eigenvectors as the covariance matrix, but no functional relation between the eigenvalues seemed to be known. Formula (5.8) on page 75 appeared to be the first result in this direction. I was following up on that, which led to Section 5.3 and the Bachelor's thesis Dürre (2010).

The spatial sign covariance matrix in the elliptical model (Section 5.3) contains the essential part of the proof of formula (5.8) in Section 5.2 and gathers literature on the spatial median and the spatial sign covariance matrix. A first draft was written in July 2009 and has undergone some revision in February 2010. A similar result was published recently by Croux et al. (2010), and it seems that it is already contained in the Ph.D. thesis Yadine (2006).

On generalizing Gaussian graphical models (Section 5.4) has appeared as

> Vogel, D.: On generalizing Gaussian graphical models. In: Ciumara, R., Bădin, L. (eds.) Proceedings of the 16th European Young Statisticians Meeting, pp. 149–153. University of Bucharest (2009)

and was written in early summer 2009. It contains a subset of Chapter 3, showing some interim results, covering basically the unconstrained estimation.

Elliptical graphical modelling in higher dimensions (Section 5.5) was written in June 2010 and has appeared as

> Vogel, D., Dürre, A., Fried, R.: Elliptical graphical modelling in higher dimensions. In: Wessel, N. (ed.) Proceedings of International Biosignal Processing Conference, July 14-16, 2010, Berlin, Germany, 17:001–005 (2010)

Simulations results similar to those in Section 3.4.2 are presented. We consider an example graph with 50 nodes, as compared to the five nodes of the example graph in Section 3.4.2, demonstrating that the method developed in Chapter 3 is also feasible in higher dimensions. This is a decisive advantage over the proposal by Miyamura and Kano (2006).

On the hypothesis of conditional independence in the IC model (Section 5.6), written in fall 2008, tells of an entirely different route towards a generalization of Gaussian graphical modelling, on which I was led by Hannu Oja. Instead of the elliptical model the independent-components-model is considered. The multivariate normal model is the intersection of both. The work on this topic is a joint project with Hannu Oja and Klaus Nordhausen from the University of Tampere.

5.1 Estimating Partial Correlations Using the Oja Sign Covariance Matrix

Abstract. We investigate the Oja sign covariance matrix (Oja SCM) for estimating partial correlations in multivariate data. The Oja SCM estimates directly a multiple of the precision matrix and is based on the concept of Oja signs, a multivariate generalization of the univariate sign function, which obey some form of affine equivariance property. Our simulations show that the asymptotic distribution gives a good approximation of the exact finite-sample distribution already for samples of moderate size. We find it to equal the performance of the classical sample partial correlation in the normal model and outperform it in the case of heavy-tailed distributions. The high computational costs are its main disadvantage.

5.1.1 Introduction

Let $k \geq 3$ and $\boldsymbol{Z} = (\boldsymbol{X}, \boldsymbol{Y})$ with $\boldsymbol{X} = (X_1, X_2)$, $\boldsymbol{Y} = (Y_1, ..., Y_{k-2})$ be a k-dimensional random vector having a non-singular covariance matrix Σ. Let $\hat{X}_i(\boldsymbol{Y})$, $i = 1, 2$, be the projection of X_i onto the space of all affine linear functions of $\boldsymbol{Y}$. Then the *partial correlation of* X_1 *and* X_2 *given* $\boldsymbol{Y}$ is defined as

$$\varrho_{1,2\bullet Y} = \frac{\operatorname{cov}\Big(X_1 - \hat{X}_1(\boldsymbol{Y}),\, X_2 - \hat{X}_2(\boldsymbol{Y})\Big)}{\sqrt{\operatorname{var}\Big(X_1 - \hat{X}_1(\boldsymbol{Y})\Big)\operatorname{var}\Big(X_2 - \hat{X}_2(\boldsymbol{Y})\Big)}},$$

i.e. it is the correlation between the residuals $X_1 - \hat{X}_1(\boldsymbol{Y})$ and $X_2 - \hat{X}_2(\boldsymbol{Y})$. The partial correlation $\varrho_{1,2\bullet Y}$ can be computed from the covariance matrix Σ. It holds

$$\varrho_{1,2\bullet Y} = -\frac{k_{1,2}}{\sqrt{k_{1,1}k_{2,2}}}.$$

where $k_{i,j}$, $i, j = 1, ..., k$, are the elements of $K = \Sigma^{-1}$, see e.g. Whittaker (1990). The matrix K is called the *concentration matrix* (or *precision matrix*) of $\boldsymbol{Z}$.

Partial correlations play an important role for instance in graphical models, where the key notion is *conditional independence*. Roughly, a *graphical model* is a family of k-dimensional distributions of $\boldsymbol{Z} = (Z_1, ..., Z_k)$ that satisfy some given pairwise conditional independence restrictions on the components of $\boldsymbol{Z}$. One can then, based on these pairwise conditional independence conditions, draw inferences about conditional independencies between arbitrary disjoint subsets of $\{Z_1, ..., Z_k\}$ given some other subvector. The classical theory of graphical models for continuously distributed variables is built on the normality assumption. If $\boldsymbol{Z} = (X_1, X_2, \boldsymbol{Y})$ has a multivariate normal distribution, then X_1 and X_2 are conditionally independent given $\boldsymbol{Y}$ if and only if $\varrho_{1,2\bullet Y} = 0$, which is equivalent to $k_{1,2} = 0$. A Gaussian graphical model is thus specified by the concentration matrix K.

Now if we wish to estimate the partial correlation $\varrho_{1,2\bullet Y}$ from a sample of n independent realizations of the vector $\boldsymbol{Z} = (X_1, X_2, \boldsymbol{Y})$, then

$$\hat{\varrho}_{1,2\bullet Y} = -\frac{\hat{k}_{1,2}}{\sqrt{\hat{k}_{1,1}\hat{k}_{2,2}}} \tag{5.1}$$

is a natural choice of an estimator for $\varrho_{1,2\bullet Y}$, where $\hat{K} = (\hat{k}_{i,j})_{i,j}$ is a suitable estimator of the precision matrix K. Equivalent to looking at $\hat{\varrho}_{1,2\bullet Y}$ is looking at the matrix-valued estimator

$$\hat{C} = (\hat{K}_D)^{-\frac{1}{2}}\hat{K}(\hat{K}_D)^{-\frac{1}{2}}, \tag{5.2}$$

where $\hat{K}_D$ denotes the diagonal matrix having the same diagonal as K. The matrix $\hat{C}$ is 1 on the diagonal and contains the negative estimated partial correlations as its off-diagonal elements, i.e. $\hat{\varrho}_{1,2\bullet Y} = -\hat{c}_{1,2}$. The estimator $\hat{K}$ can be the inverse of basically any multivariate covariance or shape estimator. We compare the partial correlation estimator based on the Oja SCM $\hat{K}^O$ to the classical normal MLE $\hat{K}^E$. Both estimators are properly defined in Section 2. Section 3 presents asymptotic distributions and influence functions under the elliptical model. Section 4 reports the findings of some finite-sample simulations on the distribution of the estimators and their sensitivity against contaminations. Section 5 is a short summary.

5.1.2 The Oja sign covariance matrix

In this section we define the Oja sign covariance matrix (Oja SCM), as it is done in Visuri et al. (2000). For an instant suppose we have a univariate data set $\mathbb{X} = (x_1, ..., x_n)$, $n \in \mathbb{N}$. We want to call $\text{sgn}_{\mathbb{X}}(x) = \text{sgn}(x - \text{med}(\mathbb{X}))$, $x \in \mathbb{R}$, the *sign of x w.r.t. the data sample* $\mathbb{X}$. Here sgn denotes the usual univariate sign function ($\text{sgn}(x) = \frac{x}{|x|}$ if $x \neq 0$ and zero otherwise) and $\text{med}(\mathbb{X})$ the univariate median function applied to $\mathbb{X}$. There are several possibilities how to generalize this notion to the multivariate setting. One possibility is the Oja median and the Oja sign. Consider the k-variate data sample $\mathbb{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$, $n \in \mathbb{N}$, and let

$$Q_p = \{q = \{i_1, ..., i_p\} \,\Big|\, 1 \leq i_1 < ... < i_p \leq n\}, \quad 0 \leq p \leq n,$$

be the system of all subsets of $\{1, ..., n\}$ with p elements and $N_p = |\,Q_p| = \binom{n}{p}$. Then the *Oja median of the data sample* $\mathbb{X}$ is defined as

$$\mathbf{Omed}(\mathbb{X}) = \underset{x \in \mathbb{R}^k}{\arg\min} \sum_{Q_k} \Big| \det \begin{pmatrix} 1 & ... & 1 & 1 \\ \boldsymbol{x}_{i_1} & ... & \boldsymbol{x}_{i_k} & \boldsymbol{x} \end{pmatrix} \Big|,$$

and $\mathbf{omed}(\mathbb{X})$ as the gravity center of the set $\mathbf{Omed}(\mathbb{X})$. The *Oja sign of the point* $\boldsymbol{x} \in \mathbb{R}^k$ *w.r.t.* $\mathbb{X}$ is

$$\mathbf{osgn}_{\mathbb{X}}(\boldsymbol{x}) = \frac{1}{N_{k-1}} \sum_{Q_{k-1}} \nabla_{\boldsymbol{x}} \Big| \det(\boldsymbol{y}_{i_1} \; ... \; \boldsymbol{y}_{i_{k-1}} \; \boldsymbol{y}) \Big|,$$

where $\boldsymbol{y} = \boldsymbol{x} - \mathbf{omed}(\mathbb{X})$ and $\boldsymbol{y}_i = \boldsymbol{x}_i - \mathbf{omed}(\mathbb{X})$, $i = 1, ..., n$. Note that contrary to $\text{sgn}_{\mathbb{X}}$ the Oja sign $\mathbf{osgn}_{\mathbb{X}}$ does *not only* depend upon the data sample $\mathbb{X}$ through its center point $\mathbf{omed}(\mathbb{X})$. The Oja median and the Oja sign are proper multivariate generalizations of the univariate concepts, in the sense that for $k = 1$ they yield med and $\text{sgn}_{\mathbb{X}}$ as defined above. If $k = 1$,

$$\mathbf{Omed}(\mathbb{X}) = \underset{x \in \mathbb{R}}{\arg\min} \sum_{1}^{n} \Big| \det \begin{pmatrix} 1 & 1 \\ x_i & x \end{pmatrix} \Big|$$

and

$$\mathbf{osgn}_{\mathbb{X}}(x) = \frac{\partial}{\partial x} |\det(x - \mathbf{omed}(\mathbb{X}))|.$$

Finally we construct a scatter estimate based on the Oja sign. The most frequently used estimate of scatter is the *empirical covariance matrix*

$$\text{ECM}(\mathbb{X}) = \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{x}_i - \overline{\boldsymbol{x}}_n)(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_n)^T,$$

where $\overline{\boldsymbol{x}}_n$ is the mean of $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$. ECM is the (biased) MLE of the covariance matrix Σ at the normal distribution. The Oja sign covariance matrix follows the same construction principle as the ECM, with $\boldsymbol{x}_i - \overline{\boldsymbol{x}}_n$ being replaced by $\mathbf{osgn}_{\mathbb{X}}(\boldsymbol{x}_i)$. Thus we write down

$$\text{OSCM}(\mathbb{X}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{osgn}_{\mathbb{X}}(\boldsymbol{x}_i)\mathbf{osgn}_{\mathbb{X}}(\boldsymbol{x}_i)^T.$$

Now we define two estimators of the *shape* of the precision matrix, i.e. the precision matrix up to scale, $\hat{K}^E = \text{ECM}^{-1}$ and $\hat{K}^O = \text{OSCM}$. It is on purpose that OSCM($\mathbb{X}$) is not inverted. It already estimates the precision matrix up to scale, cf. section 5.1.3. We denote the corresponding estimators for C and $\varrho_{1,2\bullet Y}$ by $\hat{C}^E$ and $\hat{\varrho}^E_{1,2\bullet Y}$, respectively $\hat{C}^O$ and $\hat{\varrho}^O_{1,2\bullet Y}$. As usual $\hat{c}^E_{i,j}$ and $\hat{c}^O_{i,j}$ denote the elements of $\hat{C}^E$ and $\hat{C}^O$, respectively.

5.1.3 Some asymptotic results

A common generalization of the multivariate normal model is the family of elliptical distributions. It is often considered in multivariate data analysis since the first and second order characteristics are an intuitive description of the actual shape of the distribution.
The density f_0 of a *spherical distribution* F_0 on $\mathbb{R}^k$ is of the form $f_0(\boldsymbol{x}) = g(\boldsymbol{x}^T\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^k$, where $g : [0,\infty) \to [0,\infty)$ is such that f_0 integrates to 1. If furthermore the covariance matrix of F_0 is the identity matrix I_k, we call F_0 a *standardized spherical distribution*. In the following we assume that $\boldsymbol{X}_0 \sim F_0$, where F_0 is a standardized spherical distribution admitting the Lebesgue-density f_0. Then, for any non-singular $A \in \mathbb{R}^{k\times k}$ and $\boldsymbol{b} \in \mathbb{R}^k$ the random variable $\boldsymbol{X} = A\boldsymbol{X}_0 + \boldsymbol{b}$ has an elliptical distribution F with mean vector $\boldsymbol{b}$, non-singular covariance matrix $\Sigma = AA^T$ and density

$$f(\boldsymbol{x}) = \det(\Sigma)^{-\frac{1}{2}} g((\boldsymbol{x}-\boldsymbol{b})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{b})).$$

Following the notation of Bilodeau und Brenner (1999) we denote the class of all elliptical distributions on $\mathbb{R}^k$ having mean $\boldsymbol{b}$ and covariance Σ by $E_k(\boldsymbol{b},\Sigma)$. By choosing the function g we model the tail behaviour of the distribution F. The most prominent member of the class of elliptical distributions is the normal distribution $N_k(\boldsymbol{b},\Sigma)$, which corresponds to $g_{N_k}(y) = (2\pi)^{-\frac{k}{2}}\exp(-\frac{1}{2}y)$. Another important subclass of the elliptical model is the family of multivariate $t_{\nu,k}$-distributions with

$$g_{t_{\nu,k}}(y) = \frac{\Gamma(\frac{\nu+k}{2})}{(\nu\pi)^{\frac{k}{2}}\Gamma(\frac{\nu}{2})}(1-\frac{y}{\nu})^{-\frac{\nu+k}{2}}.$$

Here the first subscript ν denotes the degrees of freedom. The $t_{\nu,k}(\boldsymbol{b},\Sigma)$ distribution converges to $N_k(\boldsymbol{b},\Sigma)$ as $\nu \to \infty$ and is, for small ν, a popular example of a heavy-tailed distribution. Its moments are finite only up to order $(\nu-1)$.
It is considered a shortcoming of the elliptical model that it does not include independent margins, unless the margins are normal, cf. e.g. Bilodeau and Brenner (1999), page 51. Consequently, partial uncorrelatedness (i.e. an off-diagonal zero entry in the precision matrix K) does in general not mean conditional independence. It is, however, equivalent to *conditional uncorrelatedness*, cf. Baba et al. (2004). Thus in any statistical analysis incorporating only first and second order characteristics (which is very often the case) partial correlation still provides a useful measure of conditional linear independence.
The estimator $\hat{C}$ is via (5.2) a function of $\hat{K}$. Hence, if the asymptotic distribution of $\hat{K}$ is known, the asymptotic distribution of $\hat{C}$ can be assessed applying the delta method. Ollila et al. (2003) state the following lemma about the Oja SCM $\hat{K}^O$.

Lemma 5.1.1 *If $F \in E_k(\boldsymbol{b}, \Sigma)$ and F_0 is its corresponding standardized spherical distribution, furthermore $\mathbb{X}_n = (\boldsymbol{X}_1, ..., \boldsymbol{X}_n)$, $\boldsymbol{X}_i \sim F$ i.i.d., $i = 1, ..., n$, then*

(I) $\hat{K}^O(\mathbb{X}_n) \xrightarrow{p} \gamma_{F_0} \det(\Sigma)\Sigma^{-1}$,

(II) $\sqrt{n}\Big(\hat{K}^O(\mathbb{X}_n) - \gamma_{F_0} \det(\Sigma)\Sigma^{-1}\Big) \xrightarrow{\mathcal{L}} N_k(0, \Gamma)$,

where γ_{F_0} is a constant depending only on the dimension k and $\mathbb{E}\|\boldsymbol{X}_0\|$, $\boldsymbol{X}_0 \sim F_0$, and Γ can be written as a function of Σ, k and $\mathbb{E}\|\boldsymbol{X}_0\|$. Both, γ_{F_0} and Γ, are made explicit in Ollila et al. (2003).

If the true partial correlation is zero, we can also apply Slutsky's lemma to (5.1) and deduce the following from Lemma 5.1.1 by straightforward calculations.

Lemma 5.1.2 *If F, F_0 and $\boldsymbol{X}_0$ are as in Lemma 5.1.1 and $k_{1,2} = 0$ ($K = \Sigma^{-1}$), then*

$$\sqrt{n}\hat{\varrho}^O_{1,2\bullet Y} \xrightarrow{\mathcal{L}} N\Big(0, \frac{k}{k+2}\Big(\frac{4k}{(\mathbb{E}\|\boldsymbol{X}_0\|)^2} - 3\Big)\Big).$$

If $\boldsymbol{X_0} \sim N_k(0, I_k)$, then $\mathbb{E}\|\boldsymbol{X}_0\| = \sqrt{2}\frac{\Gamma(\frac{k+1}{2})}{\Gamma(\frac{k}{2})}$. The corresponding expression for the asymptotic variance can be shown to converge to 1 as $k \to \infty$. For $k = 4$ (as in Figure 5.1) it equals $\frac{4^4}{3^3\pi} - 2 \approx 1.018$. Ollila et al. (2003) also report the value of $\mathbb{E}\|\boldsymbol{X}_0\|$ at the t-distribution. In the case of $k = 4$ and $\nu = 3$ (as in Figure 5.2) it results in an asymptotic variance of $\frac{2^7}{3^3} - 2 \approx 2.741$.

Lemma 5.1.2 allows to construct an asymptotic level-α-test for conditional independence, e.g. based on $n^{-1}(\hat{\varrho}^O_{1,2\bullet Y})^2$, which – appropriately standardized – will converge to a χ^2_1-distribution. It is intuitive from the results reported here that such a test – although its properties still need to be thoroughly assessed – is at the normal model asymptotically almost as efficient as the usual normal LR-test but has better robustness properties. Furthermore this test can easily be extended to an asymptotic test for conditional uncorrelatedness in the elliptical model by additionally estimating $\mathbb{E}\|\boldsymbol{X}_0\|$.

The asymptotics of the normal MLE $\hat{K}^E$ under normality can be found in textbooks on graphical models such as Lauritzen (1996), but a rigorous treatment under elliptical distributions is not known to us. However, since $\hat{K}^E = \hat{\Sigma}^{-1}$ is a function of the covariance estimator $\hat{\Sigma}$, one can again apply the delta method. The asymptotics of $\hat{\Sigma}$ in the elliptical model can be found in textbooks on multivariate statistics such as Bilodeau and Brenner (1999). In analogy to Lemma 5.1.2 we get

Lemma 5.1.3 *If $F = N_k(\boldsymbol{b}, \Sigma)$, Σ non-singular, then*

$$\sqrt{n}(\hat{\varrho}^E_{1,2\bullet Y} - \varrho_{1,2\bullet Y}) \xrightarrow{\mathcal{L}} N(0, (1 - \varrho^2_{1,2\bullet Y})^2).$$

In the general elliptical model a similar expression for the asymptotic variance is obtained, which in addition depends on the fourth order characteristics of F. However, if the fourth moments of F are not finite, as it is the case for $t_{3,4}$ in Figure 5.2, $\hat{\varrho}^E_{1,2\bullet Y}$ will converge at a slower than the $\sqrt{n}$ rate to the true value $\varrho_{1,2\bullet Y}$, if at all.

The influence function is an important tool in robust analysis. It describes the robustness of a statistical procedure against infinitesimal contaminations. For reasons of simplicity we consider the influence functions of our estimators at the standardized spherical distribution F_0. If we write $\boldsymbol{u} = \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}$, then

$$IF(\boldsymbol{x}, \hat{C}^O, F_0) = k\Big(1 - \frac{2\|\boldsymbol{x}\|}{\mathbb{E}\|\boldsymbol{X}_0\|}\Big)\Big(\boldsymbol{u}\boldsymbol{u}^T - (\boldsymbol{u}\boldsymbol{u}^T)_D\Big)$$

whereas the influence function of $\hat{C}^E$ is

$$IF(\boldsymbol{x}, \hat{C}^E, F_0) = -\|\boldsymbol{x}\|^2\Big(\boldsymbol{u}\boldsymbol{u}^T - (\boldsymbol{u}\boldsymbol{u}^T)_D\Big),$$

cf. Croux and Haesbroeck (2000). The former is affine linear and the latter quadratic in $\|\boldsymbol{x}\|$.

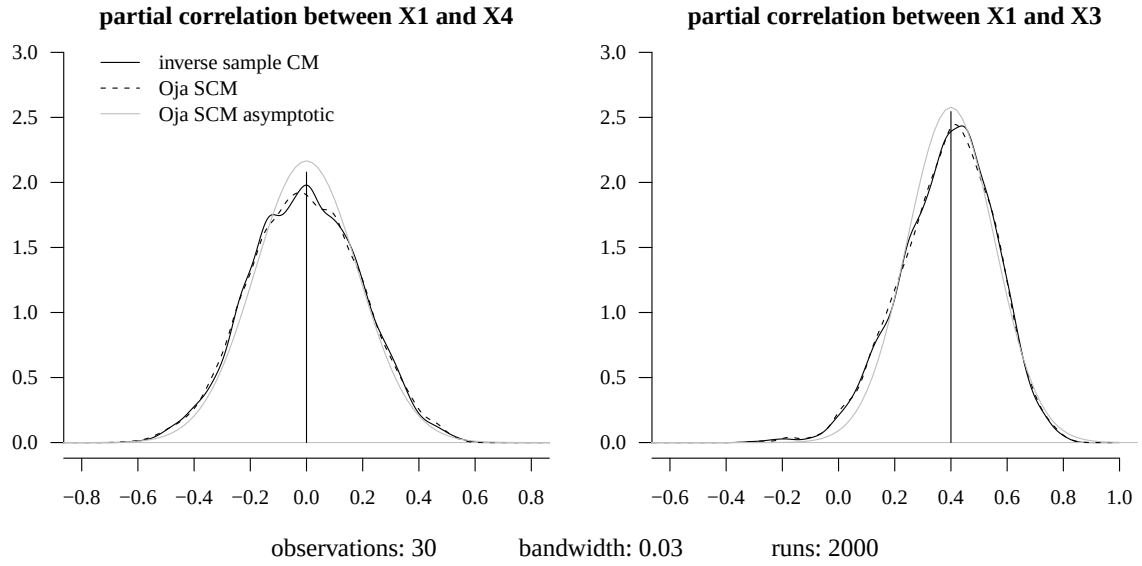

Figure 5.1: Densities of two partial correlation estimators at the multivariate normal distribution

5.1.4 Simulation results

We carried out a simulation study using several elliptical distributions to examine how the finite-sample performance relates to the asymptotics. In the examples that follow we fix the mean to zero and the covariance matrix to

$$\Sigma = \begin{pmatrix} 1 & -0.865 & 0.657 & -0.231 \\ -0.865 & 1 & -0.510 & 0.077 \\ 0.657 & -0.510 & 1 & -0.601 \\ -0.231 & 0.077 & -0.601 & 1 \end{pmatrix},$$

which corresponds to the following matrix of partial correlations

$$-C = \begin{pmatrix} -1 & -0.8 & 0.4 & 0 \\ -0.8 & -1 & 0 & -0.2 \\ 0.4 & 0 & -1 & -0.6 \\ 0 & -0.2 & -0.6 & -1 \end{pmatrix}.$$

Figure 5.1 shows the approximated densities of $-\hat{c}^O_{1,4}$ and $-\hat{c}^E_{1,4}$ (left plot) and $-\hat{c}^O_{1,3}$ and $-\hat{c}^E_{1,3}$ (right plot) calculated from 30 observations drawn from a normal distribution with covariance Σ as above. The true values to be estimated, $\varrho_{1,4 \bullet 2,3} = -c_{1,4} = 0$ and $\varrho_{1,3 \bullet 2,4} = -c_{1,3} = 0.4$, respectively, are indicated by the vertical lines. The density estimation is based on 2000 repetitions, using the R function `density()` with a Gauss kernel and bandwidth .04. The solid grey lines are the asymptotic distributions of $-\hat{c}^O_{1,4}$ (left) and $-\hat{c}^O_{1,3}$, cp. Section 5.1.3. We can not detect any relevant difference between both estimators. In fact, the asymptotic relative efficiency of $\hat{c}^O_{i,j}$ at the normal model (compared to the MLE $\hat{c}^E_{i,j}$) is more than 98%.

Figure 5.2 shows the results of an experiment with the same parameters except that the population distribution is now $t_{3,4}(\boldsymbol{b}, \Sigma)$. We find that both estimators have a higher variability (compared to the normal model), but the Oja SCM estimator $\hat{c}^O_{i,j}$ performs substantially better than the MLE $\hat{c}^E_{i,j}$. It

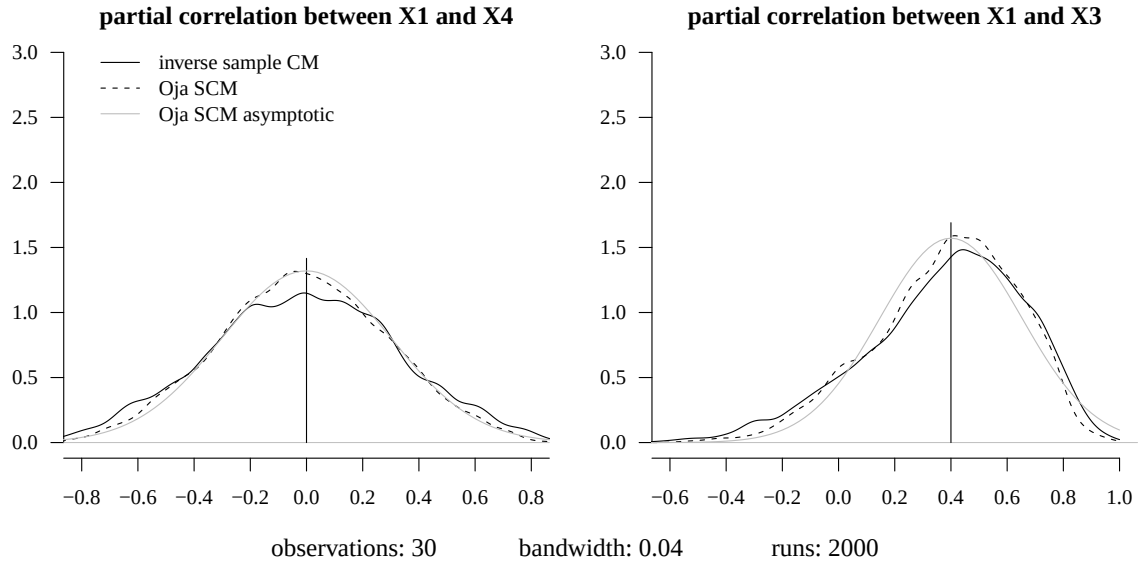

Figure 5.2: Densities of two partial correlation estimators at the t_3-distribution

should be noted, though, that in the case of light tails the picture is reversed, but still both estimators are more variable than in the normal model. Again, the solid grey lines represent the asymptotic distributions of $-\hat{c}^O_{1,4}$ and $-\hat{c}^O_{1,3}$, respectively.
In the simulation study we also examined the partial correlation estimator $\hat{C}^O$ under outlier scenarios. We found that it, though not highly robust, is less susceptible to outliers than the normal MLE $\hat{C}^E$. This is an expected behaviour considering the structure of its influence function.

5.1.5 Conclusion

The Oja SCM is well suited to the task of estimating partial correlations, in particular at heavy-tailed distributions. Note that $\sqrt{n}$-consistency of the Oja SCM only requires finite second moments. In the normal model its asymptotic and finite-sample efficiencies (almost) equal those of the MLE. The advantage is higher robustness against model misspecification. If the true distribution has heavier than Gaussian tails, the normal MLE looses strongly in efficiency whereas the Oja SCM estimator is little affected. Still, one undetected heavy outlier can make it break down.
Its major drawback remains the computational costs. Its computation necessitates the evaluation of $\binom{n}{k-1}$ $(k-1)$-dimensional hyperplanes. Using a randomized version, i.e. drawing at random a subsample of these $\binom{n}{k-1}$ hyperplanes, allows to push the limit a little bit further up to which n and k the Oja SCM is computable. In our computer experiments the approximation error turned out to be negligible compared to the estimation error up to $n = 60$ when the size of the subsample was 10%.
Finally, Visuri et al. (2000) also propose the Oja rank covariance matrix, which is based on Oja ranks instead of Oja signs. It is very similar to the Oja SCM in construction and statistical properties and exhibited the same performance in simulations.

5.2 Partial correlation estimates based on signs

Abstract. We investigate the Oja sign covariance matrix (Oja SCM) for estimating partial correlations in multivariate data. The Oja SCM estimates directly a multiple of the precision matrix and is based on the concept of Oja signs, which generalize the univariate sign function and obey some form of affine equivariance property. We compare it to the classical MLE as well as to estimates based on two alternative multivariate signs: the marginal sign and the spatial sign.

5.2.1 Introduction: partial correlation and the elliptical model

Let $k \geq 3$ and $\boldsymbol{X} = (\boldsymbol{Z}, \boldsymbol{Y})$ with $\boldsymbol{Z} = (Z_1, Z_2)$, $\boldsymbol{Y} = (Y_1, ..., Y_{k-2})$, be a k-dimensional random vector having distribution F and a non-singular covariance matrix Σ. Let furthermore $\hat{Z}_i(\boldsymbol{Y})$, $i = 1, 2$, be the projection of Z_i onto the space of all affine linear functions of $\boldsymbol{Y}$. Then the *partial correlation of* Z_1 *and* Z_2 *given* $\boldsymbol{Y}$ is defined as

$$\varrho_{1,2\bullet Y} = \frac{\operatorname{cov}\Big(Z_1 - \hat{Z}_1(\boldsymbol{Y}),\, Z_2 - \hat{Z}_2(\boldsymbol{Y})\Big)}{\sqrt{\operatorname{var}\Big(Z_1 - \hat{Z}_1(\boldsymbol{Y})\Big)\operatorname{var}\Big(Z_2 - \hat{Z}_2(\boldsymbol{Y})\Big)}},$$

i.e. it is the correlation between the residuals $Z_1 - \hat{Z}_1(\boldsymbol{Y})$ and $Z_2 - \hat{Z}_2(\boldsymbol{Y})$. The partial correlation $\varrho_{1,2\bullet Y}$ can be computed from the covariance matrix Σ of $\boldsymbol{X}$. It holds

$$\varrho_{1,2\bullet Y} = -\frac{k_{1,2}}{\sqrt{k_{1,1}k_{2,2}}},$$

where $k_{i,j}$, $i, j = 1, ..., k$, are the elements of $K = \Sigma^{-1}$, see e.g. Whittaker (1990), p. 143. K is called the *concentration matrix* (or *precision matrix*) of $\boldsymbol{X}$. The matrix

$$C = (K_D)^{-\frac{1}{2}} K (K_D)^{-\frac{1}{2}},$$

where K_D denotes the diagonal matrix having the same diagonal as K, equals 1 on the diagonal and contains the negative partial correlations as its off-diagonal elements, i.e. $\varrho_{1,2\bullet Y} = -c_{1,2}$. The correlation matrix R of $\boldsymbol{X}$ can be written as

$$R = (\Sigma_D)^{-\frac{1}{2}} \Sigma (\Sigma_D)^{-\frac{1}{2}}.$$

One easily checks that

$$C = ((M^{-1})_D)^{-\frac{1}{2}} M^{-1} ((M^{-1})_D)^{-\frac{1}{2}}$$

for any $k \times k$ matrix M that is proportional to Σ or R.
Partial correlations play an important role for instance in graphical modelling, where the key notion is *conditional independence*. Roughly, a *graphical model* is a family of k-dimensional distributions for $\boldsymbol{X} = (X_1, ..., X_k)$ that satisfy some given pairwise conditional independence restrictions on the components of $\boldsymbol{X}$. One can then, based on these pairwise conditional independence assumptions, draw inferences about conditional independencies between arbitrary disjoint subsets of $\{X_1, ..., X_k\}$. The classical theory of graphical models for continuously distributed variables is built on the normality assumption. If $\boldsymbol{X} = (Z_1, Z_2, \boldsymbol{Y})$ has a multivariate normal distribution, then Z_1 and Z_2 are conditionally independent given $\boldsymbol{Y}$ if and only if $\varrho_{1,2\bullet Y} = 0$, which is equivalent to $k_{1,2} = 0$. A Gaussian graphical model is thus specified by the concentration matrix K.

We consider the problem of estimating partial correlations, but do so in the broader situation of the elliptical model, which is a popular generalization of the multivariate normal model. Its first and second order characteristics still provide an intuitive description of the geometry of the distribution, and it is mathematically tractable. In addition it allows to model different tail behaviours.
The density f_0 of a *spherical distribution* F_0 on $\mathbb{R}^k$ is of the form $f_0(\boldsymbol{x}) = g(\boldsymbol{x}^T\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^k$, where $g : [0,\infty) \to [0,\infty)$ is such that f_0 integrates to 1. If furthermore

$$\operatorname{med}|X_1| = u_{.75}, \tag{5.3}$$

where X_1 is the first component of $\boldsymbol{X} \sim F_0$ and $u_{.75}$ the 75% quantile of the standard normal distribution, we call F_0 a *standardized spherical distribution*. In the following we assume that $\boldsymbol{X}_0 \sim F_0$, where F_0 is a standardized spherical distribution admitting the Lebesgue-density f_0. A random vector $\boldsymbol{X}$ has an elliptical distribution F if

$$\boldsymbol{X} \stackrel{\mathscr{L}}{=} S^{\frac{1}{2}}\boldsymbol{X}_0 + \boldsymbol{b}$$

for some $\boldsymbol{b} \in \mathbb{R}^k$ and symmetric, positive definite $k \times k$ matrix S. Then its density is

$$f(\boldsymbol{x}) = \det(S)^{-\frac{1}{2}} g((\boldsymbol{x}-\boldsymbol{b})^T S^{-1}(\boldsymbol{x}-\boldsymbol{b})). \tag{5.4}$$

We use the standardization assumption (5.3) in order to fix S and g in (5.4) without requiring the existence of any moments of F. It is a major advantage of sign methods that they usually work without any moment assumptions. The existence of partial correlations, of course, necessitates the existence of second moments. If expectation and variance of $\boldsymbol{X}$ exist, then $\mathbb{E}(\boldsymbol{X}) = \boldsymbol{b}$ and $\operatorname{Var}(\boldsymbol{X}) = \Sigma(F)$ is proportional to S. If F is normal, then $\Sigma(F) = S$. We call $\boldsymbol{b}$ the *symmetry center* and S the *shape matrix* of F, and – following Bilodeau and Brenner (1999) – denote the class of all elliptical distributions on $\mathbb{R}^k$ having these parameters by $E_k(\boldsymbol{b}, S)$. By choosing the function g we model the tail behaviour of the distribution F. The normal distribution $N_k(\boldsymbol{b}, \Sigma)$ corresponds to $g_{N_k}(y) = (2\pi)^{-\frac{k}{2}} \exp(-\frac{1}{2}y)$. Another important subclass of elliptical distributions is the multivariate $t_{\nu,k}$-family with

$$g_{t_{\nu,k}}(y) = c_\nu \frac{\Gamma(\frac{\nu+k}{2})}{(\nu\pi)^{\frac{k}{2}}\Gamma(\frac{\nu}{2})}(1 - \frac{c_\nu^2 y}{\nu})^{-\frac{\nu+k}{2}}.$$

Here the first subscript ν denotes the degrees of freedom. The constant $c_\nu = t_{\nu;.75}/u_{.75}$ is due to the standardization (5.3), $t_{\nu;.75}$ being the 75% quantile of the usual, univariate t_ν-distribution with ν degrees of freedom. The $t_{\nu,k}(\boldsymbol{b}, S)$ distribution converges to $N_k(\boldsymbol{b}, S)$ as $\nu \to \infty$ and is, for small ν, a popular example of a heavy-tailed distribution. Its moments are finite only up to order $(\nu - 1)$.
It is considered a shortcoming of the elliptical model that it does not include independent margins, unless the margins are normal, cf. e.g. Bilodeau and Brenner (1999), p. 51. Consequently, partial uncorrelatedness (i.e. an off-diagonal zero entry in the precision matrix K) does in general not mean conditional independence. It is, however, equivalent to *conditional uncorrelatedness*, cf. Baba et al. (2004). Thus partial correlation is a measure of conditional linear dependence.

5.2.2 Multivariate signs

A common approach in nonparametric statistics is to replace the observations by their signs or ranks. This means in general loosing efficiency under normality, but one can hope to get robust and distribution-free methods. For reasons of simplicity we only consider signs here. Since we analyse

multivariate data we are interested in multivariate signs. There are several possible generalizations of the univariate notion *sign* to the multivariate setting, three of which we want to name here: the marginal sign, the spatial sign and the Oja sign. We start by recalling the usual, univariate sign function. Suppose we have a univariate data set $\mathbb{X} = (x_1, ..., x_n)$, $n \in \mathbb{N}$. We call $\mathrm{sgn}_{\mathbb{X}}(x) = \mathrm{sgn}(x - \mathrm{med}(\mathbb{X}))$, $x \in \mathbb{R}$, the *sign of x w.r.t. the data sample* $\mathbb{X}$, where sgn is the univariate sign function ($\mathrm{sgn}(x) = \frac{x}{|x|}$ if $x \neq 0$ and zero otherwise), and $\mathrm{med}(\mathbb{X})$ is the univariate median of $\mathbb{X}$. One obvious extension of this concept to multivariate data is the component-wise application of the univariate sign, leading to the *marginal sign*. We call

$$\mathbf{msgn}_{\mathbb{X}}(\boldsymbol{x}) = \mathbf{msgn}(\boldsymbol{x} - \mathbf{mmed}(\mathbb{X}))$$

the *marginal sign of* $\boldsymbol{x} \in \mathbb{R}^k$ *w.r.t. the k-variate data sample* $\mathbb{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$, $n \in \mathbb{N}$, where $\mathbf{mmed}(\mathbb{X})$ is the component-wise, *marginal median of* $\mathbb{X}$. Another fairly straightforward multivariate generalization is obtained from the *spatial sign* function

$$\mathbf{ssgn}(\boldsymbol{x}) = \begin{cases} \frac{1}{\|\boldsymbol{x}\|}\,\boldsymbol{x} & \text{if } \boldsymbol{x} \neq \mathbf{0}, \\ \mathbf{0} & \text{if } \boldsymbol{x} = \mathbf{0}. \end{cases}$$

The spatial median $\mathbf{smed}(\mathbb{X})$ is the gravity point of $\arg\min_{\boldsymbol{x}\in\mathbb{R}^k} \left\| \sum_{i=1}^{n} \mathbf{ssgn}(\boldsymbol{x}_i - \boldsymbol{x}) \right\|$, and as before

$$\mathbf{ssgn}_{\mathbb{X}}(\boldsymbol{x}) = \mathbf{ssgn}(\boldsymbol{x} - \mathbf{smed}(\mathbb{X})), \quad \boldsymbol{x} \in \mathbb{R}^k,$$

is the *spatial sign of* $\boldsymbol{x}$ *w.r.t.* $\mathbb{X}$. A third possible multivariate extension is the Oja sign. For $0 \leq p \leq n$ let

$$Q_p = \{q = \{i_1, ..., i_p\} \,\big|\, 1 \leq i_1 < ... < i_p \leq n\}$$

be the system of all subsets of $\{1, ..., n\}$ of size p and $N_p = |\,Q_p| = \binom{n}{p}$. Then the *Oja median* $\mathbf{omed}(\mathbb{X})$ *of the data sample* $\mathbb{X}$ is defined as the gravity point of

$$\arg\min_{\boldsymbol{x}\in\mathbb{R}^k} \sum_{Q_k} \left| \det \begin{pmatrix} 1 & ... & 1 & 1 \\ \boldsymbol{x}_{i_1} & ... & \boldsymbol{x}_{i_k} & \boldsymbol{x} \end{pmatrix} \right|. \tag{5.5}$$

The *Oja sign of the point* $\boldsymbol{x} \in \mathbb{R}^k$ *w.r.t.* $\mathbb{X}$ is

$$\mathbf{osgn}_{\mathbb{X}}(\boldsymbol{x}) = \frac{1}{N_{k-1}} \sum_{Q_{k-1}} \nabla_{\boldsymbol{x}} \left| \det(\boldsymbol{y}_{i_1} \; ... \; \boldsymbol{y}_{i_{k-1}} \; \boldsymbol{y}) \right|,$$

where $\boldsymbol{y} = \boldsymbol{x} - \mathbf{omed}(\mathbb{X})$ and $\boldsymbol{y}_i = \boldsymbol{x}_i - \mathbf{omed}(\mathbb{X})$, $i = 1, ..., n$. Note that contrary to $\mathbf{msgn}_{\mathbb{X}}$ and $\mathbf{ssgn}_{\mathbb{X}}$ the Oja sign $\mathbf{osgn}_{\mathbb{X}}$ does *not only* depend upon the data sample $\mathbb{X}$ through its center point $\mathbf{omed}(\mathbb{X})$. Note that for $k = 1$ expression (5.5) comes down to

$$\arg\min_{x\in\mathbb{R}} \sum_{1}^{n} \left| \det \begin{pmatrix} 1 & 1 \\ x_i & x \end{pmatrix} \right|,$$

and

$$\mathbf{osgn}_{\mathbb{X}}(x) = \frac{\partial}{\partial x} \,|\det(x - \mathbf{omed}(\mathbb{X}))|.$$

Thus the Oja median and the Oja sign are indeed proper multivariate generalizations of med and $\text{sgn}_{\mathbb{X}}$. It should be noted that these three multivariate signs have different invariance properties. All of them are invariant w.r.t. translations. The marginal sign is also invariant w.r.t. component-wise rescaling. The spatial sign on the other hand is equivariant under orthogonal transformations, i.e. if we let $A\mathbb{X} = (A\boldsymbol{x}_1, ..., A\boldsymbol{x}_n)$ for some orthogonal matrix A, then

$$\mathbf{ssgn}_{A\mathbb{X}}(A\boldsymbol{x}) = A\mathbf{ssgn}_{\mathbb{X}}(\boldsymbol{x}).$$

The Oja sign even obeys some form of affine linear equivariance:

$$\mathbf{osgn}_{A\mathbb{X}}(A\boldsymbol{x}) = \det(A)A^{-1}\mathbf{osgn}_{\mathbb{X}}(\boldsymbol{x})$$

for any full rank $k \times k$ matrix A, cf. e.g. Ollila et al. (2003) or Hettmansperger and McKean (1998), p. 330.

5.2.3 Sign covariance matrices

Now we construct scatter estimates based on the multivariate signs introduced in the previous section: the *marginal sign covariance matrix (MSCM)*, the *spatial sign covariance matrix (SSCM)* and the *Oja sign covariance matrix (OSCM)*. All of these, along with some basic properties, can be found in Visuri et al. (2000). We start with the most frequently used estimate of scatter, the *empirical covariance matrix (ECM)*

$$\text{ECM}(\mathbb{X}) = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_n)(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_n)^T,$$

where $\overline{\boldsymbol{x}}_n$ is the mean of $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$. The ECM is the (biased) MLE of the covariance matrix Σ at the normal distribution. The sign covariance matrices follow the same construction principle as the ECM, with $\boldsymbol{x}_i - \overline{\boldsymbol{x}}_n$ being replaced by the respective signs:

$$\begin{aligned}
\text{MSCM}(\mathbb{X}) &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{msgn}_{\mathbb{X}}(\boldsymbol{x}_i)\mathbf{msgn}_{\mathbb{X}}(\boldsymbol{x}_i)^T,\\
\text{SSCM}(\mathbb{X}) &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{ssgn}_{\mathbb{X}}(\boldsymbol{x}_i)\mathbf{ssgn}_{\mathbb{X}}(\boldsymbol{x}_i)^T,\\
\text{OSCM}(\mathbb{X}) &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{osgn}_{\mathbb{X}}(\boldsymbol{x}_i)\mathbf{osgn}_{\mathbb{X}}(\boldsymbol{x}_i)^T.
\end{aligned}$$

The next lemma tells what these estimators estimate in the elliptical model. We understand the *theoretical counterpart* $\Sigma^m(F)$ of the MSCM as the functional

$$\mathbb{E}\left[\mathbf{msgn}(\boldsymbol{X}-\mathbf{mmed}(F))\,\mathbf{msgn}(\boldsymbol{X}-\mathbf{mmed}(F))^T\right]$$

with $\boldsymbol{X} \sim F$. If F is the empirical distribution function generated by the data $\mathbb{X}$, then $\Sigma^m(F) = MSCM(\mathbb{X})$. Similarly we define the theoretical counterparts of SSCM and OSCM, the latter is also explicitly stated in Ollila et al. (2003).

Lemma 5.2.1 *Let $\boldsymbol{X} \sim F$ and $\boldsymbol{X}_0 \sim F_0$ with $F \in E_k(\mathbf{0}, S)$ and F_0 the corresponding standardized spherical distribution. The theoretical counterparts of the MSCM, SSCM and OSCM at F, denoted by $\Sigma^m(F)$, $\Sigma^s(F)$ and $\Sigma^O(F)$, respectively, are given by:*

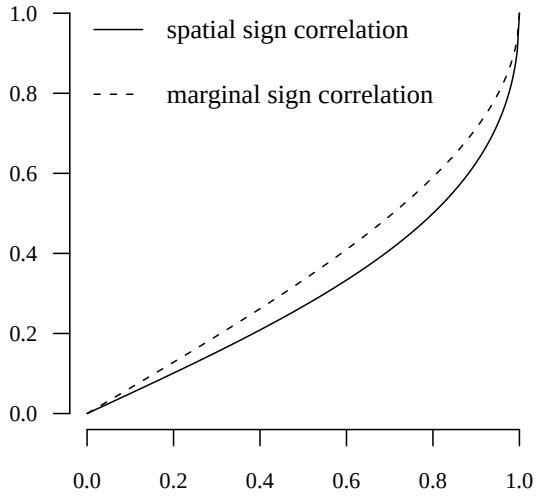

Figure 5.3: Functions σ^s (solid) and σ^m (dashed), defined in (5.6) and (5.7), respectively.

(I) $\sigma^m_{i,j}(F) = \frac{2}{\pi}\arcsin(\varrho_{i,j})$,
where $\varrho_{i,j}$, $i, j = 1, ..., k$, *are the elements of* $(S_D)^{-\frac{1}{2}} S (S_D)^{-\frac{1}{2}}$, *which equals the correlation matrix R, provided it exists.*

(II) $\Sigma^s(F) = \mathbb{E}\left(\dfrac{S^{\frac{1}{2}}\boldsymbol{X}\boldsymbol{X}^T S^{\frac{1}{2}}}{\boldsymbol{X}^T S \boldsymbol{X}}\right)$.

(III) $\Sigma^O(F) = \gamma_{F_0} \det(S) S^{-1}$,
if $\mathbb{E}\|\boldsymbol{X}_0\| < \infty$. *The constant* γ_{F_0} *depends only on* $\mathbb{E}\|\boldsymbol{X}_0\|$ *and the dimension k.*

Parts (I) and (II) are straightforward, the proof of (III) is carried out in Ollila et al. (2003), where the constant γ_{F_0} is also made explicit. Ollila et al. (2003) show furthermore that, if the second order moments of F exist, OSCM converges in probability to $\Sigma^O(F)$ and is asymptotically normal. It is intuitive that similar convergence results hold for MSCM and SSCM without any moment condition on F.
There is not such a simple relation between Σ^s and S, as there is in (I) between Σ^m and R. In particular there is in general no one-to-one correspondence between individual matrix entries. For example, in the very simple case of the 2×2 shape matrix

$$S = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \qquad -1 \leq \rho \leq 1,$$

we have

$$\Sigma^s = \frac{1}{2}\begin{pmatrix} 1 & \sigma^s(\rho) \\ \sigma^s(\rho) & 1 \end{pmatrix} \tag{5.6}$$

and

$$\Sigma^m = \begin{pmatrix} 1 & \sigma^m(\rho) \\ \sigma^m(\rho) & 1 \end{pmatrix}, \tag{5.7}$$

where $\sigma^m(\rho) = \frac{2}{\pi}\arcsin(\rho)$ and

$$\sigma^s(\rho) = \frac{2\sqrt{1+\rho}}{\sqrt{1+\rho}+\sqrt{1-\rho}} - 1. \tag{5.8}$$

Thus Figure 5.3 shows the relation of marginal-sign-correlation (also known as quadrant correlation) and spatial-sign-correlation to the usual Pearson-correlation at a two-dimensional standardized elliptical distribution. Theorem 1 in Visuri (2001) sheds some light on the structure of Σ^s in general. There is always a one-to-one connection between Σ^s and S and both matrices share the same eigenvectors.

5.2.4 Partial correlation estimators

For notational convenience we define

$$\begin{aligned}\hat{K}^e &= \mathrm{ECM}^{-1},\ \hat{K}^O = \mathrm{OSCM} \text{ and}\\ \hat{K}^m &= (h(\mathrm{MSCM}))^{-1},\end{aligned}$$

where the mapping h is the element-wise application of $x \mapsto \sin(\frac{\pi}{2}x)$. We call h(MSCM) the *modified MSCM*. From Lemma 5.2.1 we know that (if the covariance exists) $\hat{K}^e$ and $\hat{K}^O$ estimate the concentration matrix K, respectively a multiple of it, and $\hat{K}^m$ the inverse of R. From what has been said in Section 5.2.1 we can thus construct estimators of the matrix C:

$$\begin{aligned}\hat{C}^e &= (\hat{K}^e_D)^{-\frac{1}{2}}\hat{K}^e(\hat{K}^e_D)^{-\frac{1}{2}},\\ \hat{C}^O &= (\hat{K}^O_D)^{-\frac{1}{2}}\hat{K}^O(\hat{K}^O_D)^{-\frac{1}{2}},\\ \hat{C}^m &= (\hat{K}^m_D)^{-\frac{1}{2}}\hat{K}^m(\hat{K}^m_D)^{-\frac{1}{2}}.\end{aligned}$$

$\hat{C}^e$ is the normal MLE of C, see e.g. Lauritzen (1996). $\hat{C}^m$ as above is not well defined. It may happen – especially for small n – that M^m is not positive definite. The common structure of ECM and the sign covariance matrices guarantees that these matrices are always positive semi-definite, and – as long as $k < n$ and the underlying distribution F has a Lebesgue-density – ECM, OSCM and SSCM are positive definite with probability 1. This is not true for the MSCM. The additional modification step h may furthermore lead to negative eigenvalues of M^m. A remedy could be to perform an eigenvalue decomposition and set the non-positive eigenvalues to small positive values. Such a manipulation does not affect the asymptotics. We carried out a simulation study using several elliptical distributions to examine the finite-sample performance of the proposed estimators. In the examples that follow we fix the mean to $\mathbf{0}$ and the shape matrix to

$$S = \begin{pmatrix} 1 & -0.865 & 0.657 & -0.231\\ -0.865 & 1 & -0.510 & 0.077\\ 0.657 & -0.510 & 1 & -0.601\\ -0.231 & 0.077 & -0.601 & 1\end{pmatrix},$$

which corresponds to the following matrix of partial correlations

$$-C = \begin{pmatrix} -1 & -0.8 & 0.4 & 0\\ -0.8 & -1 & 0 & -0.2\\ 0.4 & 0 & -1 & -0.6\\ 0 & -0.2 & -0.6 & -1\end{pmatrix}.$$

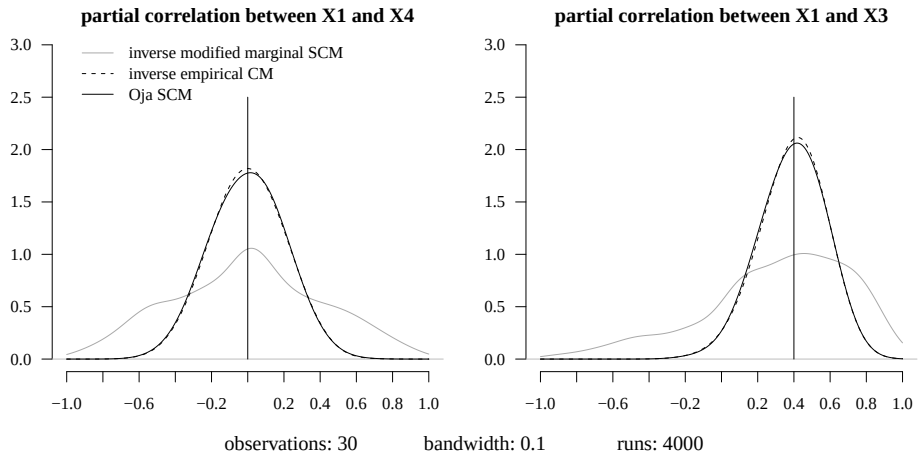

Figure 5.4: Densities of three partial correlation estimators at the multivariate normal distribution

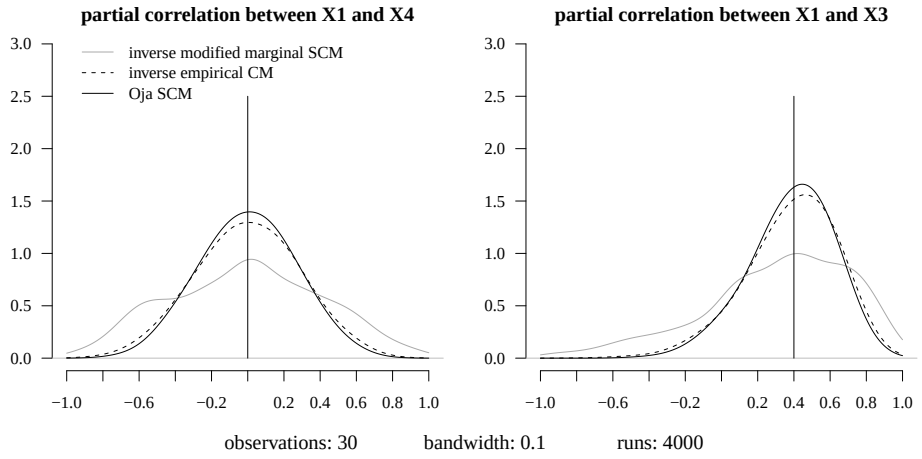

Figure 5.5: Densities of three partial correlation estimators at the t_3-distribution

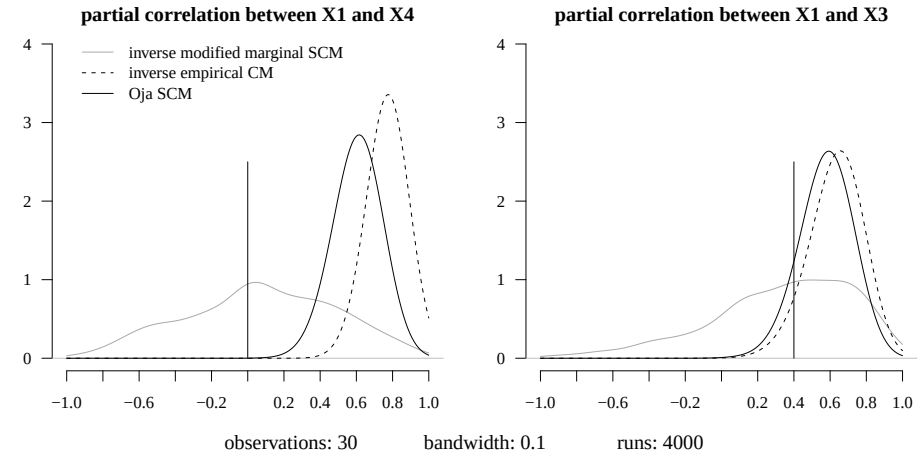

Figure 5.6: Densities of partial correlation estimators under outlier scenario

Figure 5.4 shows the estimated densities of $-\hat{c}^e_{1,4}$ $-\hat{c}^O_{1,4}$ and $-\hat{c}^m_{1,4}$ (left plot) and $-\hat{c}^e_{1,3}$, $-\hat{c}^O_{1,3}$ and $-\hat{c}^m_{1,3}$ (right plot) calculated from 30 observations drawn from a normal distribution with covariance $\Sigma = S$ as above. The true values to be estimated, $\varrho_{1,4\bullet 2,3} = -c_{1,4} = 0$ and $\varrho_{1,3\bullet 2,4} = -c_{1,3} = 0.4$, respectively, are indicated by vertical lines. The density estimation is based on 4000 repetitions, using the R function `density()` with a Gauss kernel and bandwidth .1. There does not seem to be any relevant difference between $-\hat{c}^e_{i,j}$ and $-\hat{c}^O_{i,j}$. In fact, the asymptotic relative efficiency of $\hat{c}^O_{i,j}$ at the normal model (compared to the MLE $\hat{c}^E_{i,j}$) is more than 98%, cf. Section 5.1.
In Figure 5.5 we see the results of an experiment with the same parameters except that the population distribution is now $t_{3,4}(\mathbf{0}, S)$. We find that $-\hat{c}^e_{i,j}$ and $-\hat{c}^O_{i,j}$ have a higher variability (compared to the normal model), but the Oja SCM estimator $\hat{c}^O_{i,j}$ performs substantially better than the MLE $\hat{c}^e_{i,j}$. The marginal SCM estimator $\hat{c}^m_{i,j}$ is distribution-free w.r.t. g. It should be mentioned, though, that its high variability is to a large portion due to the modification by applying h.
We also examined the behaviour of the estimators under outlier scenarios. Figure 5.6 shows the effect of a systematic outlier. We sampled again from the multivariate normal distribution (with $S = \Sigma$ as above), but added each time $(6, 0, 0, 6)$ to the first observation. The direction of this contamination was particularly aimed at destroying the partial uncorrelatedness of the variables X_1 and X_4, suggesting instead a strong positive partial dependence. $\hat{C}^m$ is little affected by the outlier. On the other hand $\hat{C}^e$ and $\hat{C}^O$ can both be made to break down by one single outlying observation, but we also find that the impact is quantitatively smaller on $\hat{C}^O$ than on $\hat{C}^e$. These findings are in agreement with the structure of the respective influence functions. At a standardized spherical distribution F_0 the influence function $IF(\boldsymbol{x}, \hat{C}^O, F_0)$ of $\hat{C}^O$ equals

$$k\Big(1 - \frac{2\|\boldsymbol{x}\|}{\mathbb{E}\|\boldsymbol{X}_0\|}\Big)(\boldsymbol{u}\boldsymbol{u}^T - (\boldsymbol{u}\boldsymbol{u}^T)_D),$$

where $\boldsymbol{u} = \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}$ and $\boldsymbol{X}_0 \sim F_0$, cf. Section 5.1. For any fixed direction $\boldsymbol{u}$ this is an affine linear function of the distance $\|\boldsymbol{x}\|$, whereas the influence functions of $\hat{C}^e$ and $\hat{C}^m$ are quadratic, respectively constant, in $\|\boldsymbol{x}\|$.

5.2.5 Conclusion

The Oja SCM is well suited to the task of estimating partial correlations in elliptical models, better than the related concepts MSCM and SSCM, since – contrary to these – it retains the whole shape information. It almost equals the efficiency of the MLE $\hat{C}^e$ in the Gaussian case, but behaves qualitatively better under model misspecifications. The loss of efficiency under heavy-tailed distributions is considerably smaller, and the same is true for the impact of outliers. We can recommend the Oja SCM as an estimator for partial correlations in graphical models, but – and this is the main drawback – only for data sets of moderate size. The reason is that its computation necessitates the evaluation of $\binom{n}{k-1}$ $(k-1)$-dimensional hyperplanes.

5.3 The spatial sign covariance matrix in the elliptical model

Abstract. This note identifies the spatial sign covariance matrix (SSCM) of a two-dimensional elliptical distribution and discusses statistical applications.

5.3.1 Definitions

For $\boldsymbol{x} \in \mathbb{R}^p$ define the *spatial sign* $\boldsymbol{s}(\boldsymbol{x})$ of $\boldsymbol{x}$ as

$$\boldsymbol{s}(\boldsymbol{x}) = \begin{cases} \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|} & \text{if } \boldsymbol{x} \neq \boldsymbol{0}, \\ \boldsymbol{0} & \text{otherwise.} \end{cases}$$

Let $\boldsymbol{X}$ be a p-dimensional random vector ($p \geq 2$) having distribution F, furthermore

$$\boldsymbol{\mu}(F) = \boldsymbol{\mu}(\boldsymbol{X}) = \underset{\boldsymbol{\mu} \in \mathbb{R}^p}{\arg\min}\, \mathbb{E}\left(\|\boldsymbol{X} - \boldsymbol{\mu}\| - \|\boldsymbol{X}\|\right)$$

the *spatial median* and

$$S(F) = S(\boldsymbol{X}) = \mathbb{E}\left(\boldsymbol{s}(\boldsymbol{X} - \boldsymbol{\mu})\boldsymbol{s}(\boldsymbol{X} - \boldsymbol{\mu})^T\right)$$

the *spatial sign covariance matrix (SSCM)* of F (or $\boldsymbol{X}$). If there is no unique minimizing point of $\mathbb{E}\left(\|\boldsymbol{X} - \boldsymbol{\mu}\| - \|\boldsymbol{X}\|\right)$, then $\boldsymbol{\mu}(F)$ is the barycenter of the minimizing set. This may only happen if F is concentrated on a line. For results on existence and uniqueness of the spatial median see Haldane (1948), Kemperman (1987), Milasevic and Ducharme (1987) or Koltchinskii and Dudley (2000).

Remarks.

(I) If the first moments of F are finite, then the spatial median allows the more descriptive characterization as $\arg\min_{\boldsymbol{\mu} \in \mathbb{R}^p} \mathbb{E}\|\boldsymbol{X} - \boldsymbol{\mu}\|$, but keep in mind that the spatial median always exists.

(II) Consider the univariate case $p = 1$ and let F be the empirical measure corresponding to the data set $x_1, ..., x_n \in \mathbb{R}$. The spatial median generalizes the idea that the (univariate) median μ has minimum average distance to all data points, mathematically speaking, that it minimizes the L_1-distance between $(\mu, ..., \mu) \in \mathbb{R}^n$ and $(x_1, ..., x_n)$. This motivates the alternative name L_1-*median*.

The term *spatial sign covariance matrix* has been introduced in Visuri et al. (2000). In the following we address the question if $S(F)$ can be given a more explicit form, say, in terms of the covariance matrix, when F belongs to a certain parametric class of distributions, e.g. the normal model.
Call $J \in \mathbb{R}^{p \times p}$ a *reflection matrix* (or *sign change matrix*), if it is a diagonal matrix with 1 or -1 on the diagonal. We say that a p-dimensional random vector $\boldsymbol{Z}$ is *reflection invariant*, if $J\boldsymbol{Z} \overset{\mathscr{L}}{=} \boldsymbol{Z}$ for every reflection matrix J. Consider the following two models.

Model M1: $\boldsymbol{X} \overset{\mathscr{L}}{=} O\boldsymbol{Z}$,
where $\boldsymbol{Z} = (Z_1, ..., Z_p)$ is a reflection invariant random vector in $\mathbb{R}^p$ and $O \in \mathbb{R}^{p \times p}$ is orthogonal.

Model M2: $\boldsymbol{X} \overset{\mathscr{L}}{=} A\boldsymbol{Y}$, where the p-dimensional random vector $\boldsymbol{Y}$ has a *spherical* distribution with center $\boldsymbol{0}$, i.e. $\boldsymbol{s}(\boldsymbol{Y}) \perp \|\boldsymbol{Y}\|$, and $A \in \mathbb{R}^{p \times p}$ is non-singular.

M2 constitutes the *elliptical* model. In this model, the matrix $V = AA^T$ is called *shape matrix* of the elliptical distribution F. Precisely we call any positive definite matrix $\tilde{V} \in \mathbb{R}^{p\times p}$ for which $\tilde{V}^{\frac{1}{2}}\boldsymbol{X}$ is spherical a shape matrix of F. The shape matrix is unique up to scale (and can thus be made unique by imposing some form of normalization, like fixing the determinant to 1, say). V is proportional to the covariance matrix, provided the latter exists. If $\boldsymbol{Y}$ has a density $f_0 : \mathbb{R}^p \to [0,\infty)$ w.r.t. the p-dimensional Lebesgue-measure, then it is of the form $f_0(\boldsymbol{y}) = g(\boldsymbol{y}^T\boldsymbol{y})$ for some suitable function $g : [0,\infty) \to [0,\infty)$, and consequently the density f of $\boldsymbol{X}$ has the form

$$f(\boldsymbol{x}) = (\det V)^{-\frac{1}{2}}\, g(\boldsymbol{x}^T V^{-1}\boldsymbol{x}).$$

Finally, let $V = U\Lambda U^T$ be an eigenvalue decomposition of V (with the usual ambiguities: permutation and the choice of eigenspace basis), where U is orthogonal and $\Lambda = \mathrm{diag}(\lambda_1, ..., \lambda_p)$. Since A is non-singular, V is positive definite, and hence $\lambda_i > 0$, $i = 1, ..., p$.

Remark. Note that M2 is a sub-model of M1: Another way of characterizing a spherical distribution is to say that it is *rotationally invariant*, i.e. $O\boldsymbol{Y} \overset{\mathscr{L}}{=} \boldsymbol{Y}$ for any orthogonal matrix $O \in \mathbb{R}^{p\times p}$. Hence it is in particular reflection invariant. Then, with the eigenvalue decomposition $V = AA^T = U\Lambda U^T$, we find that $\Lambda^{-\frac{1}{2}}U^TA$ is orthogonal. Thus $\tilde{\boldsymbol{Y}} = \Lambda^{-\frac{1}{2}}U^TA\boldsymbol{Y}$ is also spherical, and $\tilde{\boldsymbol{Y}}$ as well as $\Lambda^{\frac{1}{2}}\tilde{\boldsymbol{Y}}$ are reflection invariant. Hence $\boldsymbol{X} \overset{\mathscr{L}}{=} A\boldsymbol{Y} = U\Lambda^{\frac{1}{2}}\tilde{\boldsymbol{Y}}$ belongs to M1 with $O = U$ and $\boldsymbol{Z} = \Lambda^{\frac{1}{2}}\tilde{\boldsymbol{Y}}$.

5.3.2 Propositions

Proposition 5.3.1 *If $\boldsymbol{X}$ belongs to M1, then*

(a) $\boldsymbol{\mu}(\boldsymbol{X}) = \boldsymbol{0}$ *and*

(b) $S(\boldsymbol{X}) = \mathbb{E}\left(\frac{\boldsymbol{X}\boldsymbol{X}^T}{\|\boldsymbol{X}\|^2}\right) = O\tilde{\Lambda}O^T$ *where* $\tilde{\Lambda} = \mathrm{diag}(\tilde{\lambda}_1, ..., \tilde{\lambda}_p)$ *and*

$$\tilde{\lambda}_i = \mathbb{E}\left(\frac{Z_i^2}{\sum_{j=1}^p Z_j^2}\right), \qquad i = 1, ..., p. \tag{5.9}$$

The proof is fairly straightforward employing the definitions of $\boldsymbol{\mu}(\boldsymbol{X})$ and $S(\boldsymbol{X})$. The key is the orthogonal equivariance of the spatial median, respectively the orthogonal invariance of the spatial sign. Keep in mind that orthogonal transformations are norm preserving. Part (b) can be found in a similar form in Visuri (2001). The next result appears to be new.

Proposition 5.3.2 *If $\boldsymbol{X}$ belongs to M2 and $p = 2$, then $S(\boldsymbol{X}) = U\tilde{\Lambda}U^T$, where $\tilde{\Lambda} = \mathrm{diag}(\tilde{\lambda}_1, \tilde{\lambda}_2)$ with*

$$\tilde{\lambda}_i = \frac{\sqrt{\lambda_i}}{\sqrt{\lambda_1} + \sqrt{\lambda_2}}, \qquad i = 1, 2, \tag{5.10}$$

and $U\Lambda U^T$ is the eigenvalue decomposition of $V = AA^T$.

In words, in the elliptical model (which includes the multivariate normal model) the SSCM has the same eigenvectors as the shape matrix, and the eigenvalues transform according to (5.10). Note that $\tilde{\lambda}_1 = \tilde{\lambda}_2$ iff $\lambda_1 = \lambda_2$, thus V can be (up to scale) reconstructed from $S(\boldsymbol{X})$.

Proof of Proposition 5.3.2. We only consider the non-trivial case $\lambda_1 \neq \lambda_2$, also note that model M2 implies that both eigenvalues are strictly positive, because we require the matrix A to be of full rank. By Proposition 5.3.1 and the remark above it remains to solve the integral

$$\tilde{\lambda}_1 = \mathbb{E}\left(\frac{\lambda_1 Y_1^2}{\lambda_1 Y_1^2 + \lambda_2 Y_2^2}\right)$$

for a spherical distribution of $\boldsymbol{Y} = (Y_1, Y_2)$. The other eigenvalue $\tilde{\lambda}_2$ is obtained simultaneously, since a spherical distribution is permutation invariant (i.e. permuting the components of the vector leaves its distribution unchanged). In case $p = 2$ we also have $\tilde{\lambda}_2 = 1 - \tilde{\lambda}_1$ (and in general $\operatorname{tr}(\tilde{\Lambda}) = \operatorname{tr} S(\boldsymbol{X}) = \|\boldsymbol{s}(\boldsymbol{X})\| = 1$).
Spatial signs are distribution free within the elliptical model, i.e. $\boldsymbol{s}(\boldsymbol{X}) \overset{\mathscr{L}}{=} \boldsymbol{s}(\tilde{\boldsymbol{X}})$ for two elliptical vectors $\boldsymbol{X}$ and $\tilde{\boldsymbol{X}}$ sharing the same shape matrix V, and hence $S(\boldsymbol{X}) = S(\tilde{\boldsymbol{X}})$. The distribution of $\boldsymbol{s}(\boldsymbol{X})$ for elliptical $\boldsymbol{X}$ is also known as the *angular central Gaussian distribution*, cf. Tyler (1987b). Thus any spherical distribution can be chosen for $\boldsymbol{Y}$, for instance the uniform distribution on the unit circle with density

$$f_0(\boldsymbol{x}) = \frac{1}{\pi} \mathbb{1}_{[0,1]}(\boldsymbol{x}^T \boldsymbol{x}),$$

resulting in

$$\tilde{\lambda_1} = \frac{1}{\pi} \int_{-1}^{1} \int_{-\sqrt{1-z_1^2}}^{\sqrt{1-z_1^2}} \frac{\alpha^2 z_1^2}{\alpha^2 z_1^2 + z_2^2} \, \mathrm{d}z_2 \, \mathrm{d}z_1 \qquad \text{with} \qquad \alpha = \sqrt{\frac{\lambda_1}{\lambda_2}}.$$

Using the identity

$$\int \frac{1}{a^2 + x^2} \, \mathrm{d}x = \frac{1}{a} \arctan\left(\frac{x}{a}\right), \qquad a > 0, \tag{5.11}$$

we solve the inner integral and get with $a = \alpha z_1$:

$$\tilde{\lambda}_1 = \frac{4\alpha}{\pi} \int_0^1 z_1 \arctan\left(\frac{\sqrt{1 - z_1^2}}{\alpha z_1}\right) \mathrm{d}z_1.$$

For the remaining integral we substitute

$$x = \frac{\sqrt{1 - z_1^2}}{\alpha z_1}, \qquad 0 < z_1 \leq 1.$$

This mapping is bijective on $(0, 1]$. We get:

$$\tilde{\lambda}_1 = \frac{4\alpha}{\pi} \int_0^\infty \frac{\alpha^2 x}{(\alpha^2 x^2 + 1)^2} \arctan(x) \, \mathrm{d}x.$$

Note that $x \mapsto \frac{\alpha^2 x}{(\alpha^2 x^2 + 1)^2}$ is the derivative of $x \mapsto -\frac{1}{2(\alpha^2 x^2 + 1)}$. By means of partial integration we obtain:

$$\tilde{\lambda}_1 = \frac{2\alpha}{\pi} \int_0^\infty \frac{1}{(\alpha^2 x^2 + 1)(x^2 + 1)} \, \mathrm{d}x.$$

Employing partial fraction expansion the integrand can be written as

$$\frac{1}{(\alpha^2 x^2 + 1)(x^2 + 1)} = \frac{1}{1 - \alpha^2} \left(\frac{-\alpha^2}{\alpha^2 x^2 + 1} + \frac{1}{x^2 + 1} \right)$$

and thus integrated applying again the identity (5.11), which yields

$$\tilde{\lambda}_1 \;=\; \frac{\alpha}{1 + \alpha} \;=\; \frac{\sqrt{\lambda_1}}{\sqrt{\lambda_1} + \sqrt{\lambda_2}}.$$

This completes the proof. ∎

5.3.3 Statistical applications

Consider a p-dimensional data sample $\mathbb{X}_n = (\boldsymbol{X}_1^T, \ldots, \boldsymbol{X}_n^T)^T$ of size n, where the $\boldsymbol{X}_i$, $i = 1, ..., p$, are i.i.d., each with distribution F. Define

$$\hat{S}_n(\mathbb{X}_n; \boldsymbol{t}) = \underset{i=1,\ldots,n}{\text{ave}} \frac{(\boldsymbol{X}_i - \boldsymbol{t})(\boldsymbol{X}_i - \boldsymbol{t})^T}{\|\boldsymbol{X}_i - \boldsymbol{t}\|^2}$$

and

$$\hat{S}_n(\mathbb{X}_n; \boldsymbol{T}_n) = \underset{i=1,\ldots,n}{\text{ave}} \frac{(\boldsymbol{X}_i - \boldsymbol{T}_n)(\boldsymbol{X}_i - \boldsymbol{T}_n)^T}{\|\boldsymbol{X}_i - \boldsymbol{T}_n\|^2},$$

where $\boldsymbol{t} \in \mathbb{R}^p$, and $(\boldsymbol{T}_n)_{n\in\mathbb{N}}$ is a sequence of p-valued random vectors. For $\boldsymbol{t} = \boldsymbol{\mu}(F)$ the functional $\hat{S}_n(\mathbb{X}_n; \boldsymbol{t})$ is the empirical SSCM with known location, whereas, if $\boldsymbol{T}_n$ is a suitable location estimator, $\hat{S}_n(\mathbb{X}_n; \boldsymbol{T}_n)$ is to be interpreted as an empirical SSCM with *unknown* location. The canonical location functional in this case is the *(empirical) spatial median*

$$\hat{\boldsymbol{\mu}}_n = \hat{\boldsymbol{\mu}}_n(\mathbb{X}_n) = \min_{\boldsymbol{m}\in\mathbb{R}^p} \sum_{i=1}^{n} \|\boldsymbol{X}_i - \boldsymbol{m}\|.$$

Under regularity conditions (the data points do not lie on a line and none of them coincides with $\hat{\boldsymbol{\mu}}_n$, see Kemperman (1987), p. 228) the (empirical) spatial signs w.r.t. the (empirical) spatial median are centered, i.e.

$$\sum_{i=1}^{n} \boldsymbol{s}(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}_n) = \boldsymbol{0}.$$

Hence the (empirical) spatial sign covariance matrix $\hat{S}_n(\mathbb{X}_n; \hat{\boldsymbol{\mu}}_n)$ is indeed the covariance matrix of the spatial signs, if the latter are taken w.r.t. the spatial median.

Proposition 5.3.1 basically tells that, in the broad semiparametric model M1, the SSCM consistently estimates the eigen*vectors* (and the order of the eigenvalues) of the covariance matrix, which may be phrased as "it gives information about the *orientation* of the data" (Bensmail and Celeux, 1996, cp.)), and thus its use has been proposed for such kind of multivariate analysis that is based on this information only, most notably principal component analysis, (Marden, 1999; Locantore et al., 1999; Croux et al., 2002; Gervini, 2008). Other such applications are direction-of-arrival estimation (Visuri et al., 2001), or testing of sphericity in the elliptical model (Sirkiä et al., 2009). The latter makes use of the fact that under the null hypothesis that $\boldsymbol{X}$ is spherical, $\boldsymbol{s}(\boldsymbol{X})$ is uniformly distributed on the p-dimensional unit sphere, thus also spherical, and the covariance matrix of $\hat{S}_n$ takes on a rather simple form.
With Proposition 5.3.2 it is now possible to reconstruct the whole shape information, i.e. eigenvectors and eigenvalue ratios of the covariance matrix. Thus the SSCM can be directly employed for applications that rely on this type of information (but do not require any knowledge about the overall scale), most notably correlations and partial correlations. In higher dimensions one can, based on Proposition 5.3.2, construct a pairwise correlation estimator, which is robust, distribution-free within the elliptical model, and most of all very fast to compute.

5.4 On generalizing Gaussian graphical models

Abstract. We explore elliptical graphical models as a generalization of Gaussian graphical models, that is, we allow the population distribution to be elliptical instead of normal. Towards a statistical theory for such graphical models, consisting of estimation, testing and model selection, we consider the problem of estimating partial correlations. We derive the asymptotic distribution of a class of partial correlation matrix estimators based on affine equivariant scatter estimators.

5.4.1 Introduction: partial correlations and graphical models

Let $p \geq 3$ and $\boldsymbol{X} = (\boldsymbol{Z}, \boldsymbol{Y})$ with $\boldsymbol{Z} = (Z_1, Z_2)$, $\boldsymbol{Y} = (Y_1, ..., Y_{p-2})$, be a p-dimensional random vector having distribution F and a non-singular covariance matrix Σ. Let furthermore $\hat{Z}_i(\boldsymbol{Y})$, $i = 1, 2$, be the projection of Z_i onto the space of all affine linear functions of $\boldsymbol{Y}$. Then the *partial correlation of* Z_1 *and* Z_2 *given* $\boldsymbol{Y}$ is defined as

$$\varrho_{1,2\bullet Y} = \frac{\operatorname{cov}\Big(Z_1 - \hat{Z}_1(\boldsymbol{Y}),\, Z_2 - \hat{Z}_2(\boldsymbol{Y})\Big)}{\sqrt{\operatorname{var}\Big(Z_1 - \hat{Z}_1(\boldsymbol{Y})\Big)\operatorname{var}\Big(Z_2 - \hat{Z}_2(\boldsymbol{Y})\Big)}},$$

i.e. it is the correlation between the residuals $Z_1 - \hat{Z}_1(\boldsymbol{Y})$ and $Z_2 - \hat{Z}_2(\boldsymbol{Y})$. One can extend the definition of partial correlation (and thus partial uncorrelatedness) to vector-valued random variables in a straightforward manner. The partial correlation $\varrho_{1,2\bullet Y}$ can be computed from the covariance matrix Σ of $\boldsymbol{X}$:

$$\varrho_{1,2\bullet Y} = -\frac{k_{1,2}}{\sqrt{k_{1,1}k_{2,2}}},$$

where $k_{i,j}$, $i, j = 1, ..., p$, are the elements of $K = \Sigma^{-1}$, see e.g. Whittaker (1990), p. 143. K is called the *concentration matrix* (or *precision matrix*) of $\boldsymbol{X}$. Let

$$P = (p_{i,j})_{i,j=1,...,k} = K_D^{-\frac{1}{2}} K K_D^{-\frac{1}{2}},$$

where K_D denotes the diagonal matrix having the same diagonal as K and $K_D^{-\frac{1}{2}}$ is to be read as $(K_D)^{-\frac{1}{2}}$. The matrix P equals 1 on the diagonal and contains the negative partial correlations as its off-diagonal elements, i.e. $\varrho_{1,2\bullet Y} = -p_{1,2}$. We will also refer to P as partial correlation matrix even though it contains negative partial correlations. In this paper we consider the task of estimating P in the elliptical model, which is a popular generalization of the multivariate normal model. Its first and second order characteristics provide an intuitive description of the geometry of the distribution, and it is mathematically tractable. In addition it allows to model different tail behaviours, and is often chosen to model data with heavy tails.

Our interest in partial correlation is originated in its application in graphical models. A thorough introduction of the latter would go beyond the scope of this exposition, we refer to standard volumes, e.g. Lauritzen (1996) or Whittaker (1990). If the population distribution is jointly normal, due to the particular properties of the normal family (most notably that it is closed under conditioning, and that correlation zero implies independence) partial uncorrelatedness implies conditional independence. A spherical distribution, however, has independent margins if and only if it is a normal distribution. This is also known as the Maxwell-Hershell-theorem cf. e.g. Bilodeau and Brenner (1999), p. 51. Consequently, in the elliptical model partial uncorrelatedness (i.e. an off-diagonal zero entry in the

precision matrix K) does not imply conditional independence. It does, however, imply *conditional uncorrelatedness*, cf. Baba et al. (2004), i.e. the conditional distribution of (Z_1, Z_2) given $\boldsymbol{Y} = \boldsymbol{y}$ (which is a bivariate distribution depending on $\boldsymbol{y}$) is for almost all values $\boldsymbol{y}$ such that it has correlation zero. Thus, in the elliptical model partial correlation is a measure of conditional correlation.

5.4.2 Elliptical distributions and shape matrices

In this introduction to elliptical distributions we mainly follow the notation of chapter 13 of Bilodeau and Brenner (1999). A continuous distribution F in $\mathbb{R}^p$ is said to be *elliptical* if it has a Lebesgue-density f of the form

$$f(\boldsymbol{x}) = \det(S)^{-\frac{1}{2}} g((\boldsymbol{x}-\boldsymbol{\mu})^T S^{-1}(\boldsymbol{x}-\boldsymbol{\mu})). \tag{5.12}$$

for some $\boldsymbol{\mu} \in \mathbb{R}^p$ and symmetric, positive definite $p \times p$ matrix S. We call $\boldsymbol{\mu}$ the symmetry center and S the shape matrix of F, and denote the class of all continuous elliptical distributions on $\mathbb{R}^p$ having these parameters by $E_p(\boldsymbol{\mu}, S)$. If second-order moments of $\boldsymbol{X} \sim F$ exist, then $\mathbb{E}(\boldsymbol{X}) = \boldsymbol{\mu}$, and $\mathrm{Var}(\boldsymbol{X}) = \Sigma(F)$ is proportional to S. In the parametrization $(\boldsymbol{\mu}, S)$, the symmetry center $\boldsymbol{\mu}$ is uniquely defined whereas the matrix S is unique only up to scale, that is, $E_p(\boldsymbol{\mu}, S) = E_p(\boldsymbol{\mu}, cS)$ for any $c > 0$. One is tempted to impose some form of general standardization on S (several have been suggested in the literature, e.g., setting the trace to p or the determinant or a specific element of S to 1) and thus uniquely defining *the* shape matrix of an elliptical distribution. However, we refrain from such a standardization and call any matrix S satisfying (5.12) for a suitable function g a shape matrix of F. This allows, for example, to work always with the "simplest" function g. We want to mention two examples of elliptical distributions, the normal distribution $N_p(\boldsymbol{\mu}, S)$, which corresponds to $g_{N_p}(y) = (2\pi)^{-\frac{p}{2}} \exp(-\frac{1}{2}y)$, and the multivariate $t_{\nu,p}$-family with

$$g_{t_{\nu,p}}(y) = \frac{\Gamma(\frac{\nu+p}{2})}{(\nu\pi)^{\frac{p}{2}}\Gamma(\frac{\nu}{2})}(1-\frac{y}{\nu})^{-\frac{\nu+p}{2}}.$$

Here the first subscript ν denotes the degrees of freedom. The $t_{\nu,p}(\boldsymbol{\mu}, S)$ distribution converges to $N_p(\boldsymbol{\mu}, S)$ as $\nu \to \infty$ and is, for small ν, a popular example of a heavy-tailed distribution. Its moments are finite only up to order $(\nu - 1)$. For $\nu \geq 3$ its covariance is $\Sigma(t_{\nu,p}(\boldsymbol{\mu}, S)) = \frac{\nu}{\nu-2}S$.
We now turn to our statistical problem of interest: to estimate P in the elliptical model. Let $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ be i.i.d. random variables with elliptical distribution $F \in E_p(\boldsymbol{\mu}, S)$ and covariance matrix Σ. Let furthermore $\mathbb{X}_n = (\boldsymbol{X}_1^T, ..., \boldsymbol{X}_n^T)^T$ be the $n \times p$ data matrix containing the data points as rows and $\hat{S}_n = \hat{S}_n(\mathbb{X}_n)$ a scatter estimator. Here we use the term *scatter estimator* in a very informal way for any symmetric matrix-valued estimator that gives some form of information about the spread of the data. In a narrower sense scatter estimators aim at estimating the covariance matrix. Hence it is a desirable property of such estimators to transform in the same way as the covariance matrix under affine linear transformations, that is, they satisfy $\hat{S}_n(\mathbb{X}_n A^T + \mathbf{1}\boldsymbol{b}^T) = A\hat{S}_n(\mathbb{X}_n)A^T$ for any full rank matrix $A \in \mathbb{R}^{p\times p}$ and vector $\boldsymbol{b} \in \mathbb{R}^p$. This property of a scatter estimator is called *affine equivariance*. However, there are estimators that do not satisfy affine equivariance, but a slightly weaker condition which we want to call *affine pseudo-equivariance* or *proportional affine equivariance*.

Condition C5.4.1 $\hat{S}_n(\mathbb{X}_n A^T + \mathbf{1}\boldsymbol{b}^T) = h(A)A\hat{S}_n(\mathbb{X}_n)A^T$ *for* $\boldsymbol{b} \in \mathbb{R}^p$, $A \in \mathbb{R}^{p\times p}$ *with full rank, and* $h : \mathbb{R}^{p\times p} \to \mathbb{R}$ *satisfying* $h(H) = 1$ *for any orthogonal matrix* H.

Estimators satisfying C5.4.1 shall also be called shape estimators: they give information about the shape (orientation and relative length of the axes of the contour-ellipses of F), but not the overall scale. Since the overall scale is irrelevant for (partial) correlations, i.e.

$$P = V_D^{-\frac{1}{2}} V V_D^{-\frac{1}{2}}, \text{ where } V = S^{-1}, \tag{5.13}$$

for any shape matrix S of F, shape estimators are useful for estimating partial correlations, and we will turn our attention to this class of estimators in the following. A variety of shape estimators have been proposed and extensively studied, primarily in the robustness literature, see e.g. Zuo (2006) for a review, but also the MLE of the covariance matrix at an elliptical distribution possesses this property. Affine (pseudo-)equivariance is indeed a very handy property, and such estimators are particularly suited for the elliptical model. Their variance (which then appears as asymptotic variance if the estimator is asymptotically normal) can be shown to have a rather simple general form under elliptical population distributions, which is given below in condition C5.4.2, and is basically due to Tyler Tyler (1982). We need to introduce some matrix notation.
For matrices $A, B \in \mathbb{R}^{p\times p}$, the Kronecker product $A \otimes B$ is the $p^2 \times p^2$ matrix with entry $a_{i,j}b_{k,l}$ at position $(i(p-1)+k, j(p-1)+l)$. Let $\boldsymbol{e}_1, ..., \boldsymbol{e}_p$ be the unit vectors in $\mathbb{R}^p$ and define the following matrices: $J_p = \sum_{i=1}^p \boldsymbol{e}_i\boldsymbol{e}_i^T \otimes \boldsymbol{e}_i\boldsymbol{e}_i^T$, $K_p = \sum_{i=1}^p \sum_{j=1}^p \boldsymbol{e}_i\boldsymbol{e}_j^T \otimes \boldsymbol{e}_j\boldsymbol{e}_i^T$ (the *commutation matrix*), I_{p^2} the $p^2 \times p^2$ identity matrix and $N_p = \frac{1}{2}(I_{p^2} + K_p)$. Finally $\mathrm{vec}(A)$ is the p^2 vector obtained by stacking the columns of $A \in \mathbb{R}^{p\times p}$ from left to right underneath each other. Many shape estimators have been shown to satisfy the following condition in the elliptical model (possibly under additional assumptions on the population distribution F).

Condition C5.4.2 *There exist constants $\eta \geq 0$, $\sigma_1 \geq 0$ and $\sigma_2 \geq -2\sigma_1/p$ such that*

$$\hat{S}_n \xrightarrow{p} \eta S \qquad \text{and} \qquad \sqrt{n}\,\mathrm{vec}(\hat{S}_n - \eta S) \xrightarrow{\mathscr{L}} N_{p^2}(\mathbf{0}, W),$$

where

$$W = 2\sigma_1\eta^2 N_p(S \otimes S) + \sigma_2\eta^2\,\mathrm{vec}(S)(\mathrm{vec}(S))^T,$$

and the constants σ_1 and σ_2 do not *depend on S.*

By means of the CMT and the multivariate delta method one can derive the general form of the asymptotic variance of any partial correlation estimator derived from a scatter estimator satisfying C5.4.2.

Proposition 5.4.3 *If $\hat{S}_n$ fulfils C5.4.2, the corresponding partial correlation estimator*

$$\hat{P}_n = (\hat{S}_n^{-1})_D^{-\frac{1}{2}} \hat{S}_n^{-1} (\hat{S}_n^{-1})_D^{-\frac{1}{2}}$$

satisfies

$$\hat{P}_n \xrightarrow{p} P \quad \text{and} \quad \sqrt{n}\,\mathrm{vec}(\hat{P}_n - P) \xrightarrow{\mathscr{L}} N_{p^2}(\mathbf{0}, 2\sigma_1\Gamma N_p(V \otimes V)\Gamma^T) \tag{5.14}$$

with P and V as in (5.13) and $\Gamma = (V_D^{-\frac{1}{2}} \otimes V_D^{-\frac{1}{2}}) - N_p(P \otimes V_D^{-1})J_p$.

Remark. In the expression for the asymptotic variance of $\hat{P}_n$ the constant η obviously has to cancel out. But also the constant σ_2 does not appear. Thus the comparison of the asymptotic efficiencies of partial correlation matrix estimators based on affine (pseudo-) equivariant scatter estimators reduces to the comparison of the respective values of the scalar σ_1. This is generally true for "scale-free" functions of $\hat{S}_n$ and has already been noted by Tyler (1983).

5.4.3 Example: Tyler's M-estimator of scatter

Strictly speaking, two examples are given: Tyler's estimator mentioned in the title and, for comparison, the empirical covariance matrix $\hat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^n (\boldsymbol{X}_i - \overline{\boldsymbol{X}}_n)(\boldsymbol{X}_i - \overline{\boldsymbol{X}}_n)^T$, which is the maximum likelihood estimator for Σ at the multivariate normal distribution. $\hat{\Sigma}_n$ fulfils condition C5.4.1 with $h \equiv 1$, and we have the following asymptotic result.

Proposition 5.4.4 *If $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are i.i.d. with distribution $F \in E_p(\boldsymbol{\mu}, \alpha\Sigma)$, $\alpha > 0$, and $\mathbb{E}\|\boldsymbol{X} - \boldsymbol{\mu}\|^4 < \infty$, then $\hat{\Sigma}_n$ fulfils C5.4.2 with $\eta = \alpha^{-1}$, $\sigma_1 = 1 + \kappa/3$ and $\sigma_2 = \kappa/3$, where κ is the kurtosis excess of the first (or any other) component of $\boldsymbol{X}_1$.*

The Tyler scatter estimator $\hat{T}_n = \hat{T}_n(\mathbb{X}_n)$ is defined as the solution of

$$\frac{p}{n}\sum_{i=1}^n \frac{(\boldsymbol{X}_i - \overline{\boldsymbol{X}}_n)(\boldsymbol{X}_i - \overline{\boldsymbol{X}}_n)^T}{(\boldsymbol{X}_i - \overline{\boldsymbol{X}}_n)^T \hat{T}_n^{-1} (\boldsymbol{X}_i - \overline{\boldsymbol{X}}_n)} = \hat{T}_n \tag{5.15}$$

which satisfies $\mathrm{tr}(\hat{T}_n) = p$. It is regarded as the most robust M-estimator. Existence, uniqueness and asymptotic properties are treated in Tyler (1987a). Apparently $\hat{T}_n$ satisfies

$$\hat{T}_n(\mathbb{X}_n A^T + \mathbf{1}\boldsymbol{b}^T) = \frac{p}{\mathrm{tr}(A\hat{T}_n(\mathbb{X}_n)A^T)} A\hat{T}_n(\mathbb{X}_n)A^T$$

for $\boldsymbol{b} \in \mathbb{R}^p$ and any full rank $A \in \mathbb{R}^{p\times p}$, but not condition C5.4.1. As a consequence the asymptotic variance of $\hat{T}_n$ has a slightly different form than W in condition C5.4.2. Nonetheless the corresponding partial correlation estimator $\hat{P}_n^{(T)} = (\hat{T}_n^{-1})_D^{-\frac{1}{2}} \hat{T}_n^{-1} (\hat{T}_n^{-1})_D^{-\frac{1}{2}}$ satisfies (5.14). This is simply because, by choosing a suitable alternative normalization instead of setting the trace to p, one can obtain an estimator satisfying C5.4.1, which leads to the same partial correlation estimator $\hat{P}_n^{(T)}$. Precisely, we have the following result.

Proposition 5.4.5 *If $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are i.i.d. with distribution $F \in E_p(\boldsymbol{\mu}, \alpha\Sigma)$, $\alpha > 0$, and $\mathbb{E}\|\boldsymbol{X} - \boldsymbol{\mu}\|^2 < \infty$ and $\mathbb{E}\|\boldsymbol{X} - \boldsymbol{\mu}\|^{-\frac{3}{2}} < \infty$, then $\hat{P}_n^{(T)}$ fulfils (5.14) with $\sigma_1 = 1 + \frac{2}{p}$.*

Thus the scalar σ_1 is constant for the Tyler matrix, irrespective of the function g, i.e. the Tyler matrix (and hence the resulting partial correlation estimator) is distribution-free within the elliptical model. Moreover, it is more efficient than $\hat{\Sigma}_n$ at distributions with large (positive) kurtosis, i.e. heavy-tailed distributions. For instance, this holds true for the $t_{\nu,p}$-distribution if $\nu < p + 4$.

Final remark. Both moment conditions in Proposition 5.4.5 are only due to the location estimation in (5.15). Location estimators other than the mean are also possible and, in view of robustness, might be more appropriate, most notably the Hettmansperger-Randles median, cf. Hettmansperger and Randles (2002). However, the inverse moment condition $\mathbb{E}\|\boldsymbol{X} - \boldsymbol{\mu}\|^{-\frac{3}{2}} < \infty$ can generally not be avoided by choosing a different location estimator, cf. Tyler (1987a). But this is a fairly mild condition: for $p \geq 2$ it is fulfiled if g has no singularity at 0, thus including normal and $t_{\nu,p}$-distributions.

5.5 Elliptical graphical modelling in higher dimensions

Abstract. Simpson's famous paradox vividly exemplifies the importance of considering conditional, rather than marginal, associations for assessing the dependence structure of several variables. The study of conditional dependencies is the subject matter of graphical models. The statistical methods applied in graphical models for continuous variables rely on the assumption of normality, which leads to the term *Gaussian graphical models*. We consider *elliptical graphical models*, that is, we allow the population distribution to be elliptical instead of normal. We examine the class of affine equivariant scatter estimators and propose an adjusted version of the deviance tests, valid under ellipticity. A detailed derivation can be found in Chapters 3 and 4. In this section we report the results of a simulation study, demonstrating the feasibility of our approach also in higher dimensions. Graphical models based on classical, non-robust estimators have been used, e.g., to explore successfully the partial correlation structure within high-dimensional physiological time series (Gather et al., 2002) and within high-dimensional time series describing neural oscillators (Schelter et al., 2006).

5.5.1 Graphical models

We first introduce the basic terms and notions. Let $p \geq 3$ and $\boldsymbol{X} = (X_1, X_2, \boldsymbol{Y})$ with $\boldsymbol{Y} = (X_3, ..., X_p)$ be a p-dimensional random vector following some distribution F with non-singular covariance matrix Σ. Let $\hat{X}_i(\boldsymbol{Y})$, $i = 1, 2$, be the projection of X_i onto the space of all affine linear functions of $\boldsymbol{Y}$. Then the *partial correlation* $p_{1,2}$ *of* X_1 *and* X_2 *given* $X_3, ..., X_p$ is defined as the correlation between the residuals $X_1 - \hat{X}_1(\boldsymbol{Y})$ and $X_2 - \hat{X}_2(\boldsymbol{Y})$. The partial correlation $p_{1,2}$ can be interpreted as a measure of the *linear* association between X_1 and X_2 after the common *linear* effects of all other variables have been removed. It is a moment-based characteristic of the distribution F and can be computed from the covariance matrix Σ. It holds

$$p_{1,2} = -\frac{k_{1,2}}{\sqrt{k_{1,1}k_{2,2}}},$$

where $k_{i,j}$, $i, j = 1, ..., k$, are the elements of $K = \Sigma^{-1}$, see e.g. Whittaker (1990). The matrix K is called the *concentration matrix* (or *precision matrix*) of $\boldsymbol{X}$.

The partial correlation structure of the random variable $\boldsymbol{X}$ can be coded in a graph, which originates the term *graphical model*. An undirected graph $G = (V, E)$, where V is the vertex set and E the edge set, is constructed the following way: the variables $X_1, ..., X_p$ are the vertices, and an (undirected) edge is drawn between X_i and X_j, $i \neq j$, if and only if $p_{i,j} \neq 0$. The thus obtained graph G is called the *partial correlation graph (PCG)* of $\boldsymbol{X}$. Formally we set $V = \{1, ..., p\}$ and write the elements of E as unordered pairs $\{i, j\}$, $1 \leq i < j \leq p$. The partial correlation graph is a useful data analytical tool. It concisely displays the important aspects of the interrelations of several variables. It allows furthermore to draw conclusions about the dependence between groups of variables (note that the graph is constructed from pairwise relations between individual variables) and facilitates the understanding of the underlying physiological process. We will not dwell further on the purpose and the properties of the PCG. For more information see Section 2.2.2 or any of the classical textbooks Whittaker (1990); Cox and Wermuth (1996); Lauritzen (1996); Edwards (2000). Our concern here is the statistical modelling.

5.5.2 Gaussian graphical models

Suppose we have a data set $\mathbb{X}_n = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)^T$ of n i.i.d. realizations of the p-dimensional random vector $\boldsymbol{X}$. In order to sensibly "estimate" the PCG of $\boldsymbol{X}$ from the data, we have to make some dis-

tributional assumption about $\boldsymbol{X}$. This assumption is usually multivariate normality, i.e. $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$ for some $\boldsymbol{\mu} \in \mathbb{R}^p$ and positive definite matrix $\Sigma \in \mathbb{R}^{p\times p}$. Then the *Gaussian graphical model* $\mathscr{M}(G)$ induced by the undirected graph $G = (V, E)$ is the set of all p-dimensional Gaussian distributions satisfying the zero partial correlation restrictions specified by G. Precisely, if we denote the set of all positive definite $p \times p$ matrices by $\mathscr{S}_p^+$ and let

$$\mathscr{S}_p^+(G) = \left\{ K \in \mathscr{S}_p^+ \,\middle|\, k_{i,j} = 0 \,\forall\, i \neq j \text{ with } \{i, j\} \notin E \right\},$$

then

$$\mathscr{M}(G) = \left\{ N_p(\boldsymbol{\mu}, \Sigma) \,\middle|\, \boldsymbol{\mu} \in \mathbb{R}^p,\ K = \Sigma^{-1} \in \mathscr{S}_p^+(G) \right\}.$$

An integral part of almost any model selection scheme is the possibility to test if a model under consideration fits the data or not. In the context of Gaussian graphical models the classical tool for this purpose is the *deviance test*, which is described in the following. For any graph $G = (V, E)$ define the function $h_G : \mathscr{S}_p^+ \to \mathscr{S}_p^+ : A \mapsto A_G$ by

$$\begin{cases} [A_G]_{i,j} = a_{i,j}, & \{i, j\} \in E \quad \text{or} \quad i = j, \\ [A_G^{-1}]_{i,j} = 0, & \{i, j\} \notin E \quad \text{and} \quad i \neq j. \end{cases} \tag{5.16}$$

It is not trivial and a deeper result of the theory of Gaussian graphical models that a unique and positive definite solution A_G of (5.16) exists for any positive definite A. The solution can be found by the *iterative proportional scaling* algorithm, for which convergence has been shown, cf. Lauritzen (1996), Chap. 5. If we let further $\hat{\Sigma}_n$ denote the sample covariance matrix, then $\hat{\Sigma}_G = h_G(\hat{\Sigma}_n)$ is a sensible estimator for Σ subject to the assumption $\Sigma^{-1} \in \mathscr{S}_p^+(G)$. It is indeed the maximum likelihood estimator in the Gaussian graphical model $\mathscr{M}(G)$. Now suppose we have two nested models $\mathscr{M}(G_0) \subsetneq \mathscr{M}(G_1)$, i.e. the edge set E_0 of G_0 is a strict subset of the edge set E_1 of G_1. Let q be the number of edges that are in E_1 but not E_0. Then, under $\mathscr{M}(G_0)$,

$$\hat{D}_n(\hat{\Sigma}_n) = n \left(\ln \det h_{G_0}(\hat{\Sigma}_n) - \ln \det h_{G_1}(\hat{\Sigma}_n) \right) \tag{5.17}$$

converges to a χ^2 distribution with q degrees of freedom. This the likelihood ratio test for testing $\mathscr{M}(G_0)$ against the larger model $\mathscr{M}(G_1)$. The statistic $\hat{D}_n(\hat{\Sigma}_n)$ is also referred to as *deviance*. Many model selection procedures (backward elimination, forward selection, Edwards-Havránek,...) consist of an iterative application of this test. For details see, e.g., Edwards (2000).

5.5.3 Elliptical graphical models

A problem of the Gaussian graphical modelling described in the previous section is its lack of robustness, which is mainly due to the poor robustness of the estimator $\hat{\Sigma}_n$. Hence a promising way of robustifying the procedure is to replace $\hat{\Sigma}_n$ by a more robust scatter estimator. Over the last four decades many proposals of robust multivariate dispersion estimators have been made, for a review see, e.g., Zuo (2006). Indeed, it can be shown that the convergence of (5.17) remains true, if $\hat{\Sigma}_n$ is replaced by any scatter estimator $\hat{S}_n$ that fulfils the following regularity conditions.

(I) $\hat{S}_n$ is (at least proportionally) affine equivariant, i.e. $\hat{S}_n(\mathbb{X}_n A^T + \boldsymbol{b}) \propto A\hat{S}_n(\mathbb{X}_n)A^T$ for any $\boldsymbol{b} \in \mathbb{R}^p$ and full rank matrix $A \in \mathbb{R}^{p\times p}$, and

(II) $\hat{S}_n$ is $\sqrt{n}$*-convergent*, i.e. $\sqrt{n}(\hat{S}_n - S)$ converges in distribution, where S is some multiple of Σ.

These two conditions are natural for multivariate scatter estimators, see also Sections 2.4.1, 3.3.1 and 3.4.1. Then, under $\mathscr{M}(G_0)$,

$$\frac{1}{\sigma_1}\hat{D}_n(\hat{S}_n) = \frac{n}{\sigma_1}\left(\ln\det h_{G_0}(\hat{S}_n) - \ln\det h_{G_1}(\hat{S}_n)\right) \tag{5.18}$$

converges to a χ^2 distribution with q degrees of freedom, where $\sigma_1 > 0$ is a suitable scalar-valued constant, which is a function of the estimator $\hat{S}_n$, but does neither depend on G_0, G_1 nor the true covariance Σ. We call $\hat{D}_n(\hat{S}_n)$ *pseudo-deviance*, and the corresponding test *adjusted deviance test*, since we have to divide the test statistic by the consistency factor σ_1.
Furthermore, $\frac{1}{\sigma_1}\hat{D}_n(\hat{S}_n)$ also converges to a χ^2_q limit, if the data $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are sampled from an elliptical distribution. Then σ_1 has to be chosen accordingly, examples are given in the next section. A continuous distribution F in $\mathbb{R}^p$ is said to be *elliptical* if it has a density f of the form

$$f(\boldsymbol{x}) = \det(S)^{-\frac{1}{2}} g((\boldsymbol{x}-\boldsymbol{\mu})^T S^{-1}(\boldsymbol{x}-\boldsymbol{\mu})).$$

for some $\boldsymbol{\mu} \in \mathbb{R}^p$ and positive definite $p \times p$ matrix S. We call $\boldsymbol{\mu}$ the symmetry center and S the shape matrix of F. If the second-order moments of $\boldsymbol{X} \sim F$ exist, then $\mathbb{E}(\boldsymbol{X}) = \boldsymbol{\mu}$, and $\mathrm{Var}(\boldsymbol{X}) = \Sigma(F)$ is proportional to S. The class of all continuous, elliptical distributions constitutes a generalization of the multivariate normal model, that allows arbitrarily heavy tails and is therefore well suited to model outlying observations. The normal distribution is obtained by $g_{N_p}(y) = (2\pi)^{-\frac{p}{2}}\exp(-\frac{1}{2}y)$. A prominent example of a heavy-tailed distribution is the $t_{\nu,p}$-distribution, specified by

$$g_{t_{\nu,p}}(y) = \frac{\Gamma(\frac{\nu+p}{2})}{(\nu\pi)^{\frac{p}{2}}\Gamma(\frac{\nu}{2})}\left(1-\frac{y}{\nu}\right)^{-\frac{\nu+p}{2}},$$

where the index ν is referred to as the *degrees of freedom*. The moments of $t_{\nu,p}$ are finite only up to order $\nu - 1$. For $\nu \geq 3$ its covariance is $\Sigma = \frac{\nu}{\nu-2}S$, and for $\nu \geq 5$ the excess kurtosis (of each component) is $6/(\nu-4)$. Elliptical distributions do generally not possess finite moments, i.e. Σ does not necessarily exist. Provided $\hat{S}_n$ is $\sqrt{n}$-convergent, we may nevertheless use the adjusted deviance test to test (more generally) for a certain zero pattern in the inverse of the shape matrix S.

5.5.4 Examples of robust scatter estimators

If the fourth-order moments of $\boldsymbol{X} \sim F$ are finite, then $\hat{\Sigma}_n$ fulfils conditions (I) and (II). The corresponding value of σ_1 is $1 + \frac{\kappa}{3}$, where κ is the excess kurtosis of the first (or any other component) of $\boldsymbol{X}$. Thus, if we assume an elliptical population distribution F (with finite fourth-order moments), we may apply the adjusted deviance test, but have to divide $\hat{D}_n(\hat{\Sigma}_n)$ by a consistent estimate of σ_1. Under a heavy-tailed distribution, i.e., if κ is large, the estimator $\hat{\Sigma}_n$ is relatively inefficient, resulting in a test with poor power. An alternative, which keeps its efficiency under heavy tails, is Tyler's scatter estimator. It is defined as the solution $\hat{T}_n = \hat{T}_n(\mathbb{X}_n)$ of

$$\frac{p}{n}\sum_{i=1}^{n}\frac{(\boldsymbol{x}_i-\hat{\boldsymbol{\mu}}_n)(\boldsymbol{x}_i-\hat{\boldsymbol{\mu}}_n)^T}{(\boldsymbol{x}_i-\hat{\boldsymbol{\mu}}_n)^T\hat{T}_n^{-1}(\boldsymbol{x}_i-\hat{\boldsymbol{\mu}}_n)} = \hat{T}_n \tag{5.19}$$

which satisfies $\det\hat{T}_n = 1$. Here $\hat{\boldsymbol{\mu}}_n$ is an appropriate location estimator, which may be the mean, or, in light of robustness, the Hettmansperger-Randles median (Hettmansperger and Randles, 2002). Existence and uniqueness of a solution of (5.19) and the asymptotic properties of the estimator $\hat{T}_n$ are treated in the original publication Tyler (1987a). The estimator evidently satisfies condition (I)

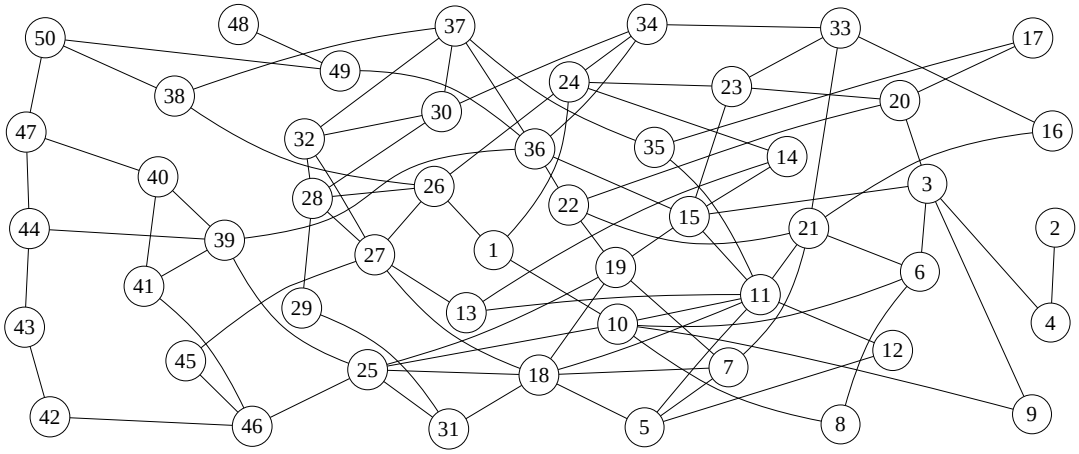

Figure 5.7: Example graph

and under some mild regularity conditions on the population distribution also condition (II). The corresponding value of σ_1 is $1 + \frac{2}{p}$, irrespective of the specific elliptical distribution. This remarkable fact may be phrased as to say the test statistic $\hat{D}_n(\hat{T}_n)$ is asymptotically distribution-free within the elliptical model, a property which it inherits from the estimator $\hat{T}_n$. This has the nice practical implication that, when carrying out the adjusted deviance test, σ_1 needs not to be estimated. Furthermore, for large p, $\hat{T}_n$ is almost as efficient as the MLE $\hat{\Sigma}_n$ at the normal distribution and outperforms $\hat{\Sigma}_n$ at distributions with slightly heavier than normal tails, e.g., at the $t_{\nu,p}$ distribution, if $\nu < p + 4$.
The third example we want to mention is the RMCD, the reweighted version of Rousseeuw's minimum covariance determinant estimator (Rousseeuw, 1985), see also Rousseeuw and Leroy (1987), Chap. 7, which has become a very popular highly robust scatter estimator. Very roughly, a subsample of size $h = \lfloor tn \rfloor$, where $\frac{1}{2} \leq t < 1$ is some fixed fraction, of the data points is chosen such that the determinant of the sample covariance matrix computed from this subsample is minimal. Afterwards a reweighting step is applied, which increases the efficiency, but maintains the high breakdown point of the initial estimator. The RMCD fulfils conditions (I) and (II). The asymptotics are treated in Butler et al. (1993) and Croux and Haesbroeck (1999). Values for σ_1 can be found in the latter. The RMCD is an appropriate estimator if the outlying observations are assumed to be contaminations, but the bulk of the data is well described by a Gaussian distribution. Similar to Tyler's estimator, the efficiency of the RMCD, relative to sample covariance matrix, increases with p.

5.5.5 Simulation study

We want to compare the estimators mentioned in the previous section in a simulation study, to give an impression of their applicability in elliptical graphical modelling. In Section 3.4.2 we report the results obtained from a small toy model consisting of five nodes and five edges. The following is aimed at complementing these numerical investigations by considering a high-dimensional example, where e.g. also run-time plays a role. Our set-up is as follows. We sample 200 i.i.d. observations of a 50-dimensional random vector that follows an elliptical distribution. For each of the elliptical distributions we consider, cf. Table 5.1, we take 1000 samples, and from each sample compute several estimates. Based on each estimate, we select a model and compare it to the true model. In all runs we use the same model (Figure 5.7) and the same shape matrix with identical diagonal elements. The partial correlation matrix is visualized in Figure 5.8: the absolute values of the matrix entries are

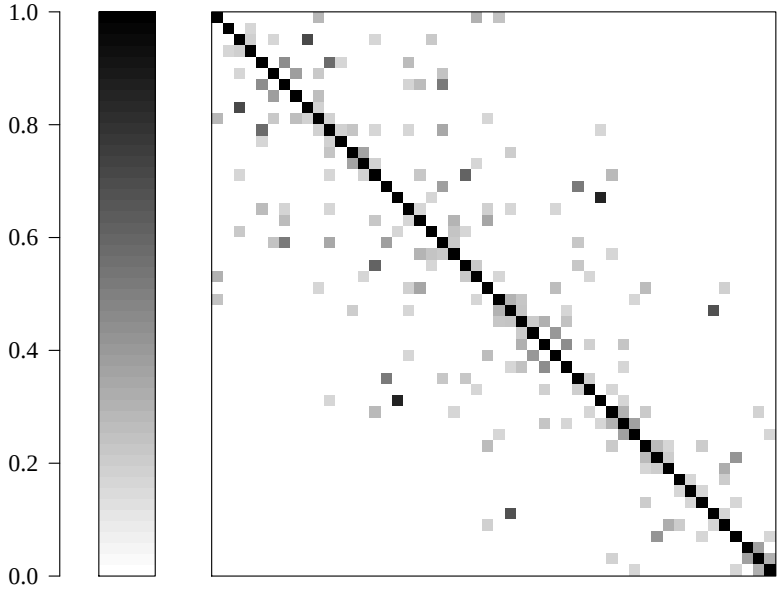

Figure 5.8: Partial correlation matrix (absolute values)

coded by different shades of gray, ranging from 0 (white) to 1 (black). Despite the many intersecting edges in Figure 5.7 this is a sparse graph. Of 1225 possible edges only 94 are present, and only two nodes (11 and 18) have more than six neighbours.

We perform a very simple model selection: we carry out an edge-exclusion test for every possible edge, i.e. we test, for each pair $\{i, j\}$, the model with all edges but $\{i, j\}$ against the saturated model and exclude the edge $\{i, j\}$, if the test accepts the smaller model. More sophisticated model search procedures generally show better results, but lead to similar conclusions as far as the comparison of the estimators is concerned. Our simple one-step model selection allows to better study the properties of the adjusted deviance tests and the effects of the choice of the scatter estimator. We perform each test at the significance level $\alpha = 0.01$, which is an ad hoc choice. It is chosen rather small due to the sparsity of the graph. Since the vast majority of possible edges is absent, identifying these non-edges correctly is of greater importance for the overall performance in this example. If we view the model selection as a multiple-testing problem, i.e. we want to restrict the probability that the fitted graph is too large, an individual significance level of $\alpha = 0.01$ is already high.

Besides getting an impression of the general performance we want to examine the finite-sample behaviour of the estimators, i.e. check if the asymptotic χ^2-approximation of the test statistics are useful in practice. A sample size of 200 seems large enough to expect some "validity" of the asymptotics. We therefore consider two criteria. The main criterion by which we measure the goodness of the model selection is the *relative mean edge difference (RMED)*, i.e. the average number of edges (averaged over all 1000 runs) that are wrongly specified in the selected model—may it be that an existing edge was rejected or an absent edge was wrongly included—divided by the total number of possible edges (1225). An RMED below 0.5 indicates that the model selection procedure is superior to random guessing. In a less complex situation it might be also of interest to know, how often the true model is found, but with 1225 test decisions in each trial we can not expect a positive number in only 1000 trials. Any model selection procedure that is based on testing for zero parameters aims at controlling the probability of correctly specifying the non-edges. Our second criterion is therefore the percentage

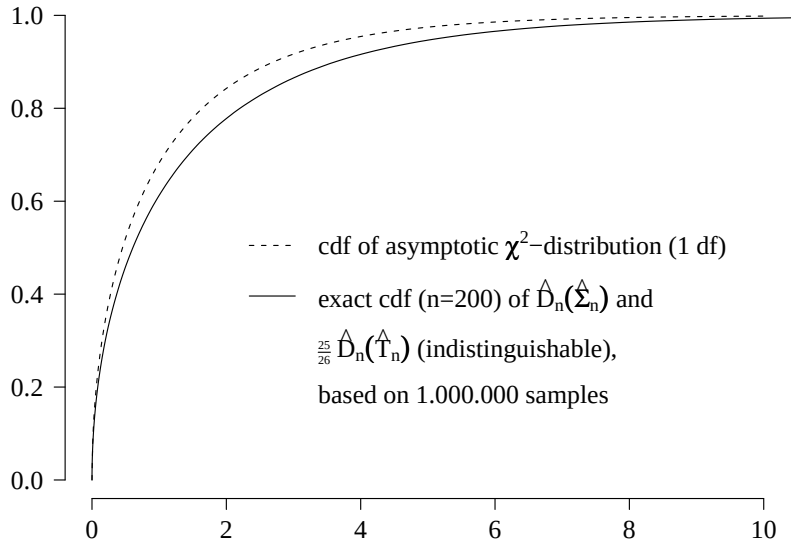

Figure 5.9: Asymptotic approximation of the test statistic for $n = 200$ at the normal distribution

of wrongly specified non-edges, which is the same as the rejection probability of the test under the null and should turn out to be about 1%.
The findings of our experiment are summarized in Table 5.1. The benchmark is traditional graphical modelling, i.e. the performance of $\hat{\Sigma}_n$ at the normal distribution, cf. first row of Table 5.1. We observe two things: First, the test is anti-conservative. The actual rejection probability under the null hypothesis is about 2.7%. The simulated cdf of the test statistic (for $n = 200$) and its limit for $n \to \infty$ are plotted in Figure 5.9. Second, the test goes wrong, if we move away from normality. We assume only ellipticity but no further knowledge about the distribution and want methods that are valid over the whole class of elliptical distributions. A possible remedy is to adjust the $\hat{\Sigma}_n$-based test statistic by an estimate of σ_1, for which we need to estimate the kurtosis κ. Since elliptical distributions have the same kurtosis in each direction, we simply take the average of the sample kurtosis of all margins. The results of this adjusted deviance test are reported in the second row of Table 5.1. This adjustment

Table 5.1: One-step model selection based on different estimators
RMED / wrongly specified non-edges (%)

distribution	normal	t_{20}	t_5	t_3
$\hat{\Sigma}$	4.9 / 2.6	5.5 / 3.2	8.1 / 6.0	11.4 / 9.5
$\hat{\Sigma}^*$	5.0 / 2.7	5.0 / 2.5	5.3 / 1.1	6.1 / 0.3
$\hat{T}$	5.1 / 2.6	5.0 / 2.6	5.0 / 2.6	5.1 / 2.6
RMCD 0.5**	6.4 / 1.0	6.1 / 0.8	6.2 / 0.9	6.3 / 1.0
RMCD 0.75**	4.2 / 1.0	4.5 / 1.3	6.0 / 2.9	8.1 / 5.2

* test statistic adjusted by estimated kurtosis
** with finite-sample correction

repairs the test, and does so surprisingly well—even in the case of the t_3-distribution, where the population kurtosis is not defined. In this case we do not have an "asymptotic justification" of the test, but we find it to be conservative. This effect, which we did not observe at the low-dimensional example, certainly deserves some further investigation.

For Tyler's estimator, there are mainly two things to note. We recognize its asymptotic efficiency properties: it almost equals the performance of $\hat{\Sigma}_n$ at the normal model, but shows no loss under larger tails. On the other hand, the test statistic shows a very similar behaviour as the $\hat{\Sigma}_n$-based deviance under normality, cf. Figure 5.9. It has in particular the same bias w.r.t. the asymptotic χ^2_1-distribution. This gives rise to the hope that finite-sample correction techniques developed for $\hat{\Sigma}_n$-based analyses, cf. e.g. Lauritzen (1996), p. 143, can be applied to $\hat{T}_n$ as well and be brought to benefit also under ellipticity.

Finally, Table 5.1 also reports results for the RMCD, with subsample fractions $t = 0.5$ and $t = 0.75$, which both exhibit generally good efficiencies, which is in contrast to the low-dimensional example. But it must be pointed out that we did not carry out an asymptotic test in this situation. It is a known problem of the RMCD that it converges very slowly to its asymptotic distribution. The "asymptotic" σ_1-value is of no use here. The problem is usually taken care of by multiplying by a correction factor which has to be determined numerically. We have chosen σ_1 such that the test delivers the desired rejection probability of 0.01 under the null at the normal model. This makes the RMCD look unjustifiably good in comparison to the other estimators.

All calculations were done in R 2.9.1, employing routines from the packages `mvtnorm` (random sampling), `ggm` (constrained estimation, i.e. the function h_G), `ICSNP` (Tyler matrix) and `rrcov` (RMCD). The simulations were run on a 2.83 GHz Intel Core2 CPU. The computation of Tyler's estimator lasted less then a second, the RMCD less than 3 seconds, all 1225 edge-exclusion tests took about 37 seconds. Figure 5.7 was created using Graphviz.

5.6 On the hypothesis of conditional independence in the IC model

Abstract. This note identifies the subset of the parameter space that is associated with hypothesis of conditional independence in the (semi-parametric) independent-components-model, and puts it into relation with other, similar hypotheses that might be of interest. In Section 5.6.4 some thoughts on a possible likelihood-ratio test procedure (assuming the marginal densities to be known) are gathered, and in Section 5.6.5 a simulated example is presented.

5.6.1 The independent-components-model (ICM)

We consider the symmetric independent components model (SICM) as described in Oja et al. (2010). Consider a p-dimensional ($p \geq 3$) random vector $\boldsymbol{X}$ that we assume to be generated by

$$\boldsymbol{X} = \Lambda \boldsymbol{Z} + \boldsymbol{\mu}, \tag{5.20}$$

where

a) $\boldsymbol{Z} = (Z_1, ..., Z_p)$ is a random vector in $\mathbb{R}^p$ with independent components,

b) each component Z_i, $i = 1, ..., p$, has a (univariate) Lebesgue-density g_i,

c) is symmetric around 0, i.e. $Z_i \sim -Z_i$, and

d) satisfies $\operatorname{med}|Z_i| = 1$.

e) The mixing matrix $\Lambda = (\lambda_{ij})_{i,j=1,...,p} \in \mathbb{R}^{p\times p}$ has full rank, and $\boldsymbol{\mu} = (\mu_1, ..., \mu_p)$ is a p-dimensional vector.

We use furthermore the following notation. Let

f) $\Gamma = (\gamma_{ij})_{i,j=1,...,p}$ be the inverse of Λ,

g) g the density of $\boldsymbol{Z}$, $g(z_1, ..., z_p) = \prod_{i=1}^{p} g_i(z_i)$, and

h) f the density of $\boldsymbol{X}$,

$$f(x_1, ..., x_p) = |\det(\Gamma)| g(\Gamma(\boldsymbol{x} - \boldsymbol{\mu})) = |\det(\Gamma)| \prod_{i=1}^{p} g_i\Big(\sum_{j=1}^{p} \gamma_{ij}(x_j - \mu_j)\Big),$$

where $\boldsymbol{x} = (x_1, ..., x_p)$.

Assumption d) is an unusual standardization condition, which does not require the existence of moments. If, however, second moments of $\boldsymbol{Z}$ exist, $\boldsymbol{X}$ has covariance matrix

$$\Sigma = \Lambda D \Lambda^T = \Big(\sum_{k=1}^{p} \operatorname{var}(Z_k)\lambda_{ik}\lambda_{jk}\Big)_{i,j=1,...,p},$$

where $D = \operatorname{diag}(d_{11}, ..., d_{pp}) = \operatorname{diag}(\operatorname{var}(Z_1), ..., \operatorname{var}(Z_p))$.

Lemma 5.6.1 *The mixing matrix Λ is unique up to permutation and sign change of the columns if and only if g fulfils condition C5.6.2.*

Condition C5.6.2 *At most one of $g_1, ..., g_p$ is Gaussian.*

Proof of Lemma 5.6.1. This is a well known result in the ICA literature, see e.g. Comon (1994), Hyvärinen et al. (2001) or also Theis (2004). ∎

The remaining paragraph of this section is spent on formalizing what is meant by "unique up to permutation and sign change of the columns". Call $P \in \mathbb{R}^{p\times p}$ a *permutation and sign change matrix (PSM)* if it has in each line and in each column exactly one non-zero element, and that is 1 or -1. Any PSM P has the following properties.

- P is orthogonal, P^{-1} is PSM, the product of two PSM's is also PSM.
- For any matrix $M \in \mathbb{R}^{p\times p}$, applying P from the right permutes and changes the signs of the *columns* of M.
- Applying P from the left permutes and changes the signs of the *rows* of M.

Now let $\boldsymbol{Z}, \Lambda, \boldsymbol{\mu}$ satisfy assumptions a)—e) and P be PSM. If $\boldsymbol{X} = \Lambda\boldsymbol{Z} + \boldsymbol{\mu}$, i.e. $\boldsymbol{Z} = \Gamma(\boldsymbol{X} - \boldsymbol{\mu})$, then, with $\tilde{\boldsymbol{Z}} = P\boldsymbol{Z}$, $\tilde{\Gamma} = P\Gamma$, and $\tilde{\Lambda} = \Lambda P^{-1}$, we also have $\boldsymbol{X} = \tilde{\Lambda}\tilde{\boldsymbol{Z}} + \boldsymbol{\mu}$, i.e. $\tilde{\boldsymbol{Z}} = \tilde{\Gamma}(\boldsymbol{X} - \boldsymbol{\mu})$, and $\tilde{\boldsymbol{Z}}, \tilde{\Lambda}, \boldsymbol{\mu}$ satisfy assumptions a)—e) as well. Lemma 5.6.1 tells that, if C5.6.2 holds, this is only true if P is PSM.

5.6.2 Conditional independence

We use $\boldsymbol{\mu}$ and Γ (rather than Λ) as parametrization, which is motivated by Lemma 5.6.4. Thus define $\vartheta = (\boldsymbol{\mu}, \operatorname{vec}\Gamma)$ and write Θ for the set of all possible parameters. The latter is a strict subset of $\mathbb{R}^{p+p^2}$, since we require Γ to be non-singular. Now let $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3)$ where $\boldsymbol{X}_1$, $\boldsymbol{X}_2$ and $\boldsymbol{X}_3$ are subvectors of sizes p_1, p_2 and p_3, respectively, with $p_i \geq 1$, $i = 1, 2, 3$, and $p_1 + p_2 + p_3 = p$. We want to test the hypothesis

$$H_0 : \ \boldsymbol{X}_1 \perp\!\!\!\perp \boldsymbol{X}_2 | \boldsymbol{X}_3,$$

that is, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are conditionally independent given $\boldsymbol{X}_3$. This can be expressed as a condition on Γ.

Condition C5.6.3

$$\min\Big\{ \sum_{j=1}^{p_1} |\gamma_{ij}|, \sum_{j=p_1+1}^{p_2} |\gamma_{ij}| \Big\} = 0 \qquad \textit{for all } i = 1, ..., p,$$

that is in words, in each row of Γ, either all elements in the first block column (of width p_1) or all elements in the second block column (of width p_2) are zero.

Call Θ_0 the set of all $\vartheta \in \Theta$ satisfying C5.6.3.

Lemma 5.6.4 *$\boldsymbol{X}_1 \perp\!\!\!\perp \boldsymbol{X}_2 | \boldsymbol{X}_3$ for all choices of g if and only if Γ fulfils C5.6.3.*

Proof. Use the density characterization of conditional independence,

$$X \perp\!\!\!\perp Y | Z \quad \Leftrightarrow \quad f_{(X,Y,Z)}(x, y, z) = h_1(x, z) h_2(y, z) \tag{5.21}$$

for some functions h_1, h_2 (cp. Lauritzen (1996), p. 29), and recall the density f of $\boldsymbol{X}$,

$$f(x_1, ..., x_p) = |\det(\Gamma)| \prod_{i=1}^{p} g_i\Big(\sum_{j=1}^{p} \gamma_{ij}(x_j - \mu_j) \Big).$$

Then, if C5.6.3 is true, f factorizes according to (5.21). On the other hand, if C5.6.3 is not true, one can find densities $g_1, ..., g_p$ such that f does not factorize according to (5.21). ∎

Remark. It has not been claimed or proven that, for any specific g that fulfils C5.6.2, $\boldsymbol{X}_1 \perp \boldsymbol{X}_2 | \boldsymbol{X}_3 \Rightarrow$ C5.6.3, but it seems plausible.

So we identify our original hypothesis $H_0 : \boldsymbol{X}_1 \perp \boldsymbol{X}_2 | \boldsymbol{X}_3$ with the equivalent hypothesis

$$H_0' : \vartheta \in \Theta_0.$$

This is not a "nice" condition, in particular Θ_0 is not a linear space, and it is not clear how to test this hypothesis. Any estimate $\hat{\Gamma}$ may deliver the rows in a totally different order than the true Γ, from which we may assume our data to be generated. In order to overcome this issue one could impose an ordering on the lines of Γ, according to how far the entries in the first block column and in the second block column are away from zero. Then one faces a variety of technical questions, such as, by which metric to measure the distance to zero and how to deal with the multiple sort criteria (first *and* second block column). Maybe the following approach is more promising. We write Θ_0 as a union of several linear hypotheses Θ_q. Let

$$Q' = \{q = (i_1, ..., i_k) \,|\, 1 \leq i_1 < ... < i_k \leq p,\ 0 \leq k \leq p\}$$

be the set of all possible ordered tuples out of $\{1, ..., p\}$, and define Θ_q for any $q = (i_1, ..., i_k) \in Q$, as follows. We say $\vartheta \in \Theta_q$ if $\vartheta \in \Theta_0$ and

$$\gamma_{ij} = 0 \ \text{ for all } (i, j) \in (\{i_1, ..., i_k\} \times \{1, ..., p_1\}) \cup ((\{1, ..., p\} \setminus \{i_1, ..., i_k\}) \times \{p_1 + 1, ..., p_2\}).$$

In words, Θ_q consists of those $\vartheta \in \Theta_0$ for which Γ has zero-entries in the first block column in all lines $i_1, ..., i_k$ in q and zero-entries in the second block column in all other lines. The tuple q then contains those rows of Γ which are zero in the first block column. Apparently

$$\bigcup_{q \in Q'} \Theta_q = \Theta_0.$$

Some of the 2^p sets Θ_q, $q \in Q'$, can be right away identified as empty. If $k < p_2$ or $k > p_2 + p_3$, then $\vartheta \in \Theta_q$ (k being the length of q) implies that Γ can not have full rank. For instance, take $k < p_2$. If $\vartheta \in \Theta_q$, then only k rows of Γ may have non-zero entries in the second block column, i.e. the p_2 column vectors in the second block column only span at most a k-dimensional subspace of $\mathbb{R}^p$, while a p_2-dimensional subspace would be needed in order to have all columns of Γ span all of $\mathbb{R}^p$. Therefore define Q as follows,

$$Q = Q(p_1, p_2, p_3) = \{q = (i_1, ..., i_k) |\ 1 \leq i_1 < ... < i_k \leq p,\ p_2 \leq k \leq p_2 + p_3\},$$

and we have

$$\bigcup_{q \in Q} \Theta_q = \Theta_0.$$

Q has $N = N(p_1, p_2, p_3) = \sum_{k=p_2}^{p_2+p_3} \binom{p}{k}$ elements.

5.6.3 Related hypotheses

In the previous section we translated our hypothesis $H_0 : \boldsymbol{X}_1 \perp\!\!\!\perp \boldsymbol{X}_2 | \boldsymbol{X}_3$ into the (somewhat unsatisfactory) condition C5.6.3 on the inverse Γ of the mixing matrix Λ. It is an open problem to translate H_0 into a workable equivalent condition on the mixing matrix Λ itself. In this section I will review two related hypotheses.
Assume in the following that the covariance matrix of $\boldsymbol{Z}$ exists. Partition Λ (and Γ as well) according to the partitioning of $\boldsymbol{X}$, i.e. let

$$\Lambda = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} & \Lambda_{13} \\ \Lambda_{21} & \Lambda_{22} & \Lambda_{23} \\ \Lambda_{31} & \Lambda_{32} & \Lambda_{33} \end{pmatrix},$$

where $\Lambda_{ij} \in \mathbb{R}^{p_i \times p_j}$, $i, j = 1, 2, 3$. Let furthermore $K = \Sigma^{-1} = \Gamma^T D^{-1} \Gamma$, the inverse of the covariance matrix of $\boldsymbol{X}$, be partitioned likewise

$$K = \begin{pmatrix} K_{11} & K_{12} & K_{13} \\ K_{21} & K_{22} & K_{23} \\ K_{31} & K_{32} & K_{33} \end{pmatrix}.$$

We call K the *concentration matrix* of $\boldsymbol{X}$. Consider the following two hypotheses.

I_0: $\Lambda_{12} = \Lambda_{21}^T = 0$, $\Lambda_{31} = 0$, $\Lambda_{32} = 0$, thus Λ looks like this:

$$\Lambda = \begin{pmatrix} \Lambda_{11} & 0 & \Lambda_{13} \\ 0 & \Lambda_{22} & \Lambda_{23} \\ 0 & 0 & \Lambda_{33} \end{pmatrix}.$$

Using some imagination we find the non-zero blocks to form an (admittedly tilted) "V", and also say "Λ has a V-shape."

J_0: $K_{12} = K_{21}^T = 0$, i.e.

$$K = \begin{pmatrix} K_{11} & 0 & K_{13} \\ 0 & K_{22} & K_{23} \\ K_{31} & K_{32} & K_{33} \end{pmatrix}.$$

This means $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are partially uncorrelated given $\boldsymbol{X}_3$. In the multivariate normal model (i.e. $g_1, ..., g_p$ are all normal) this is equivalent to $\boldsymbol{X}_1 \perp\!\!\!\perp \boldsymbol{X}_2 | \boldsymbol{X}_3$.
In the remainder of this section we prove

Proposition 5.6.5 $I_0 \overset{\Rightarrow}{\Leftarrow} H_0 \overset{\Rightarrow}{\Leftarrow} J_0$,
where we understand H_0 *as* $\boldsymbol{X}_1 \perp\!\!\!\perp \boldsymbol{X}_2 | \boldsymbol{X}_3$ *for all possible* g.

For the proof we need the following two lemmas.

Lemma 5.6.6 Λ *has V-shape if and only if its inverse* Γ *is V-shaped as well.*

Proof. By applying the following result about partitioned matrices twice

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} A^{-1} + A^{-1}BE^{-1}CA^{-1} & -A^{-1}BE^{-1} \\ -E^{-1}CA^{-1} & E^{-1} \end{pmatrix}$$

with $E = D - CA^{-1}B$, cp. e.g. Magnus and Neudecker (1999), p. 11 or Harville (1997), p. 99, we find that the inverse of

$$\begin{pmatrix} \Lambda_{11} & 0 & \Lambda_{13} \\ 0 & \Lambda_{22} & \Lambda_{23} \\ 0 & 0 & \Lambda_{33} \end{pmatrix} \quad \text{is} \quad \begin{pmatrix} \Lambda_{11}^{-1} & 0 & \Lambda_{11}^{-1}\Lambda_{13}\Lambda_{33}^{-1} \\ 0 & \Lambda_{22}^{-1} & \Lambda_{22}^{-1}\Lambda_{23}\Lambda_{33}^{-1} \\ 0 & 0 & \Lambda_{33}^{-1} \end{pmatrix}.$$

The proof is complete. ∎

Lemma 4 tells that the basic difference between I_0 and C5.6.3 is that, after appropriate ordering of the rows, we impose here on Γ the same partitioning (p_1, p_2, p_3) along the rows as along the columns. I_0 is apparently a stronger condition than C5.6.3 and hence than H_0.

Lemma 5.6.7 *J_0 holds if and only if*

$$\boldsymbol{\gamma}_{\bullet i} D^{-1} \boldsymbol{\gamma}_{\bullet j} = 0 \quad \textit{for all } i = 1, ..., p_1 \textit{ and } j = p_1 + 1, ..., p_2, \tag{5.22}$$

where $\boldsymbol{\gamma}_{\bullet i}$ denotes the i-th column of Γ, $i = 1, ..., p$. We might put (5.22) in words as: Each column in the first block column is orthogonal w.r.t. *D to every column in the second block column.*

Proof. The proof is done by writing $K = \Gamma^T D^{-1} \Gamma$ componentwise:

$$k_{ij} = \boldsymbol{\gamma}_{\bullet i} D^{-1} \boldsymbol{\gamma}_{\bullet j} = \sum_{k=1}^{p} \frac{\gamma_{ki}\gamma_{kj}}{\mathrm{var}(Z_k)}, \qquad i, j = 1, ..., p.$$

∎

Proof of Proposition 5.6.5. By Lemmas 5.6.6 and 5.6.7 we expressed I_0 and J_0, respectively, as conditions on Γ. Then it is easy to see that $H_0 \Rightarrow J_0$, $H_0 \not\Leftarrow J_0$, $I_0 \Rightarrow H_0$ and $I_0 \not\Leftarrow H_0$. As for the last result, it should be noted that any matrix Γ satisfying H_0 can not be necessarily be turned into a V-shape matrix by a PSM transformation (that is, basically, by re-ordering the rows), simply because a PSM does not change the number of zero entries. I_0 is a stricter condition than H_0 even if we allow for this PSM-ambiguity. ∎

5.6.4 Likelihood-ratio-test

In the following assume that the marginal densities $g_1, ..., g_p$ are known and satisfy C5.6.2. Then we can test

$$H_0' : \vartheta \in \Theta_0 \qquad \text{vs.} \qquad H_1' : \vartheta \notin \Theta_0$$

using the likelihood-ratio approach. Let $\mathbb{X} = (\boldsymbol{X}^{(1)}, ..., \boldsymbol{X}^{(n)})$ be an i.i.d. random sample from (5.20). Then by h) the likelihood of $\vartheta = (\boldsymbol{\mu}, \mathrm{vec}\,\Gamma)$ at $\mathbb{X}$ in the SICM is

$$L(\vartheta, \mathbb{X}) = |\det(\Gamma)|^n \prod_{\nu=1}^{n} \prod_{i=1}^{p} g_i\Big(\sum_{j=1}^{p} \gamma_{ij}(X_j^{(\nu)} - \mu_j) \Big), \tag{5.23}$$

and the negative log-likelihood is

$$l(\vartheta, \mathbb{X}) = -\log L(\vartheta, \mathbb{X}) = -n \log\Big(|\det(\Gamma)|\Big) + \sum_{\nu=1}^{n} \sum_{i=1}^{p} -\log\Big(g_i\Big(\sum_{j=1}^{p} \gamma_{ij}(X_j^{(\nu)} - \mu_j)\Big)\Big).$$

The likelihood-ratio for H_0' vs. H_1' then is

$$LR(\mathbb{X}) = \frac{\max_{\vartheta \in \Theta_0} L(\vartheta, \mathbb{X})}{\max_{\vartheta \in \Theta} L(\vartheta, \mathbb{X})},$$

where $0 \leq LR(\mathbb{X}) \leq 1$, and large values of $LR(\mathbb{X})$ suggest that H_0' is true. To solve the constrained optimization problem (OP) in the numerator, we may use

$$\bigcup_{q \in Q} \Theta_q = \Theta_0,$$

thus

$$\max_{\vartheta \in \Theta_0} L(\vartheta, \mathbb{X}) = \max_{q \in Q} \max_{\vartheta \in \Theta_q} L(\vartheta, \mathbb{X}).$$

Each of the (technically constrained) OP's $\max_{\vartheta \in \Theta_q} L(\vartheta, \mathbb{X})$ can then be solved as an unconstrained OP in a lower dimensional space by simply putting the respective entries of Γ to zero in (5.23). This can be done using standard non-linear, multivariate optimization techniques such as quasi-Newton methods. Thus $LR(\mathbb{X})$ can be determined solving (at most) $N + 1$ unconstrained (up to $(p^2 + p)$-dimensional) OP's. It is equivalent to consider the negative log-likelihood-ratios

$$\lambda_q(\mathbb{X}) = -\log\Big(\frac{\max_{\vartheta \in \Theta_q} L(\vartheta, \mathbb{X})}{\max_{\vartheta \in \Theta} L(\vartheta, \mathbb{X})}\Big) = \min_{\Theta_q} l(\vartheta, \mathbb{X}) - \min_{\Theta} l(\vartheta, \mathbb{X})$$

and

$$\lambda(\mathbb{X}) = -\log LR(\mathbb{X}) = \min_{q \in Q} \lambda_q(\mathbb{X}) = \min_{q \in Q}\Big(\min_{\Theta_q} l(\vartheta, \mathbb{X})\Big) - \min_{\Theta} l(\vartheta, \mathbb{X}). \tag{5.24}$$

It should be noted, that, by fixing the marginal densities $g_1, ..., g_p$, the rows of Γ and also the rows of any ML-estimate $\hat{\Gamma}$ are fixed, with one restriction though: only if $g_1, ..., g_p$ are all different. If two of them are identical, the corresponding rows of Γ and $\hat{\Gamma}$ can be switched without changing the distribution of $\boldsymbol{X}$, respectively the likelihood. Hence equal marginal densities reduce the number of OP's that need to be solved in (5.24). For instance, if all $g_1, ..., g_p$ are equal, then due to the symmetry of the likelihood-function, $\lambda_{q_1}(\mathbb{X}) = \lambda_{q_2}(\mathbb{X})$ for any two tuples q_1 and q_2 with $|q_1| = |q_2|$, and the corresponding ML-estimates are equal up to permutation of the rows. Hence only $p_3 + 2$, instead of $N + 1$ OP's need to be solved.

The rest of this section (and this little note) contains not-very-far-pursued ideas that have to be regarded more or less as speculation at this point. One would expect, according to general LR-theory, a result of the following type to hold. For all $\vartheta \in \Theta_q$,

$$2\lambda_q(\mathbb{X}) \xrightarrow{\mathscr{L}} \chi^2_{d(q)},$$

where $d(q)$ is the number of zero-entries in Γ associated with Θ_q, i.e. $d(q) = |q|p_1 + (p - |q|)p_2$. We can use this to device a Bonferoni-type multiple-testing procedure, i.e. we would reject H_0' at the

significance level α if we can reject $\vartheta \in \Theta_q$ for each $q \in Q$ at the significance level $\frac{\alpha}{N}$. But of course in this approach the rejection region may be unnecessarily small (having far less than probability α under the null hypothesis). We are actually interested in the asymptotic distribution of $\lambda(\mathbb{X}) = \min_{q \in Q} \lambda_q(\mathbb{X})$. Maybe one can show a result of the type

$$2\lambda(\mathbb{X}) \xrightarrow{\mathscr{L}} \chi^2_{d(\vartheta)},$$

for (at least some) $\vartheta \in \Theta_0$, but where $d(\vartheta)$ may depend on the actual choice of the true parameter ϑ. The case that ϑ lies in more than one Θ_q certainly deserves a separate treatment. From there one may derive an upper bound on the asymptotic distribution of $\lambda(\mathbb{X})$ for all $\vartheta \in \Theta_0$, something in the form of

$$2\lambda(\mathbb{X}) \xrightarrow{\mathscr{L}} Y_\vartheta \preceq Y$$

for all $\vartheta \in \Theta_0$ and some variables Y_ϑ and Y, of which it suffices to identify the distribution of Y. Here $\preceq$ shall denote *stochastically smaller*, i.e. the cdf of the left-hand side variable dominates the cdf of the right-hand side variable. This would allow to construct an asymptotic level-α-test by rejecting H_0' if $2\lambda(\mathbb{X})$ is larger than the $(1-\alpha)$-quantile of Y.

The main drawback of the LR-approach is the (unrealistic) assumption of known marginal densities $g_1, ..., g_p$ and the dependence of the results on their choice. Distribution-free methods are desirable. However, one can extend this approach by assuming $g_1, ..., g_p$ to belong to some parametric family and thus fit $g_1, ..., g_p$ along with $\boldsymbol{\mu}$ and Γ using ML. Parametric families for the marginal densities have already been considered in ICA, e.g. the Pearson system in Karvanen and Koivunen (2002) (although the estimation there is moment-based, rather than ML). Or, if we think of the data as "basically normal, but with heavier tails", we may use Student's t-family. In the univariate location-scale estimation problem the MLE of the t-distribution can be regarded as a robustified normal MLE, giving less weight to outlying observations.

5.6.5 A Monte-Carlo simulation

In order to investigate the practical feasibility of this LR-approach we carried out Monte-Carlo simulations of a few very simple situations. The results of two of them (Situation 1 and 2, see below) are reported in detail. We may take marginal densities from these popular examples of symmetric density families:

- Student's t-family (cp. e.g. Bilodeau and Brenner (1999), p. 208),

$$g_{t_\nu}(y) = c_\nu \Big(1 + \frac{y^2}{\nu}\Big)^{-\frac{\nu+1}{2}} \qquad \text{with} \qquad c_\nu = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})} \tag{5.25}$$

  The parameter ν is integer-valued. t_ν realizes heavy-tailed distributions, where t_1 is the Cauchy-distribution and $t_\nu \xrightarrow{\mathscr{L}} N(0,1)$ as $\nu \to \infty$.

- the power exponential family (cp. e.g. Bilodeau and Brenner (1999), pp. 209,239),

$$g_\alpha(y) = c_\alpha \exp\Big(-\frac{1}{2}|y|^{2\alpha}\Big) \qquad \text{with} \qquad c_\alpha = \Big(2^{(1+\frac{1}{2\alpha})}\,\Gamma(1+\frac{1}{2\alpha})\Big)^{-1} \tag{5.26}$$

  The parameter α is non-negative and g_α has lighter than normal tails if $\alpha > 1$, respectively heavier tails if $\alpha < 1$.

Of course, we can not directly take the densities above, because they do not satisfy the standardization condition d). But for any symmetric (around 0) g, the density $h = \eta g(\eta \cdot)$ does fulfil d), if $\eta = \text{med}\,|X|$, $X \sim g$, which implies $\int_{-\infty}^{\eta} h(x)dx = 0.75$.

Situation 1

The model parameters are as follows:

- $p = 3$, which is the smallest non-trivial number for the problem we investigate,
- $\boldsymbol{\mu} = \mathbf{0}$,
- $\Lambda = \Lambda_1 = \frac{1}{10}\begin{pmatrix} 5 & 1 & 2 \\ 0 & 4 & -2 \\ 0 & 2 & 4 \end{pmatrix}$, which corresponds to $\Gamma = \Gamma_1 = \begin{pmatrix} 2 & 0 & -1 \\ 0 & 2 & 1 \\ 0 & -1 & 2 \end{pmatrix}$, i.e. $\vartheta \in \Theta_q$ is true only for $q = (2, 3)$.
- The marginal densities are all equal and taken from the power exponential family (5.26) with $\alpha = 1.5$, correctly adjusted as described above. The identical densities only necessitate the computation of three optimization problems.
- We looked at two different sample sizes, $n = 30$ and $n = 60$.

We generated data from this model for each sample size 4000 times and calculated in each run the test statistic $\lambda(\mathbb{X}_n)$ as described in the previous section. This was done using the R function `optim()` from the package `stats` with the following options (as far as they differ from the default) `method="BFGS"`, `maxit=100` and the starting values $\boldsymbol{\mu}_0 = (0.5, 0.5, 0.5)$ and

$$\Gamma_0 = \begin{pmatrix} 3 & 1 & 2 \\ 2 & 1 & 3 \\ 1 & 2 & 2 \end{pmatrix}.$$

Of course, for the ML optimization under the hypotheses different initial values were used, since some entries of Γ_0 are restricted to zero, but in all cases all (non-restricted) parameters were initially set to one of 1, 2 or 3. Each of the altogether 3×4000 optimizations took on the average less than one second, resulting in a total runtime of the simulation of less than three hours (on an Intel Core2 processor with 2.66 GHz). To generate data from the power exponential distribution we used the function `rpowerexp()` from the package `rmutil` by J. K. Lindsey, which can be downloaded e.g. from `http://popgen.unimaas.nl/~jlindsey/rcode/rmutil.zip`.

Situation 2

The set-up is exactly the same as in Situation 1 (including parameters and starting values of the optimization routine), except that we take as mixing matrix

$$\Lambda = \Lambda_2 = \frac{1}{4}\begin{pmatrix} 2 & 0 & 1 \\ 0 & 2 & -1 \\ 0 & 0 & 2 \end{pmatrix}, \text{ which corresponds to } \Gamma = \Gamma_2 = \begin{pmatrix} 2 & 0 & -1 \\ 0 & 2 & 1 \\ 0 & 0 & 2 \end{pmatrix},$$

i.e. now $\vartheta \in \Theta_q$ is true for $q = (2, 3)$ and $q = (2)$.

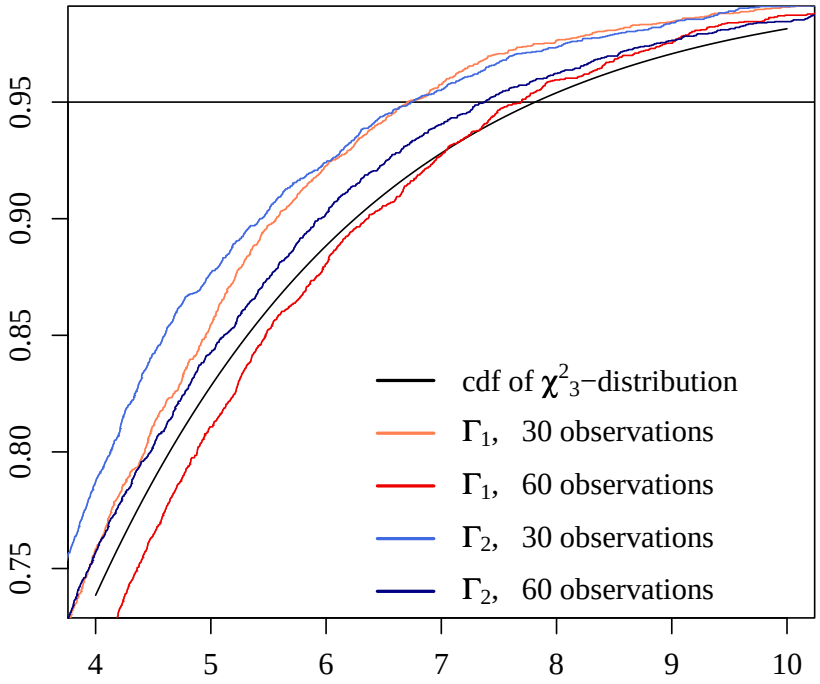

Figure 5.10: Simulated cdf of 2λ for different true Γ's and sample sizes

The results of the experiments in Situations 1 and 2 are depicted in Figure 5.10. The empirical distribution functions of $2\lambda(\mathbb{X}_n)$ (generated each time by 4000 runs) are shown along with the cdf of the conjectured asymptotic distribution χ^2_3. The reddish colors both correspond to Situation 1, the bluish colors to Situation 2. The lighter shade in each case represents the sample size $n = 30$, the darker tone the larger sample size $n = 60$.
Although these results look quite promising, it must also be mentioned that the ad-hoc optimization approach we used does not always deliver reliable results. It is sensitive to the choice of the initial value. Moreover, in similar situations where we sampled also from the power exponential distribution, but with different α-values 0.5, 1 and 1.5, the algorithm did not always converge and gave clearly wrong results (negative test statistic), although convergence was reported. Both occurred in up to 15% of the trials.

Appendix A

Matrix differential calculus

A.1 On how to differentiate matrix-valued functions w.r.t. matrices

This section is on what its title says, it deals with Jacobi-matrices, i.e. collections of partial derivatives, and how to compute them. It does not deal with theoretical background of differentiability and question like how and under what conditions the Jacobi-matrix is related to the best linear approximation of a function in the vicinity of a point. For such matters as well as for further reading we refer to the book Magnus and Neudecker (1999), abbreviated MN below. To sum it up, we are on the safe side, if all partial derivatives are continuous, the function is then called *continuously differentiable*. We will henceforth make this assumption and use the terms *derivative* and *Jacobi-matrix* synonymously. The *Jacobian* is the determinant of this matrix. Recall the important formula concerning the Kronecker product, cf. MN, p. 30,

$$\operatorname{vec}(ABC) = (C^T \otimes A) \operatorname{vec} B, \tag{A.1}$$

for matrices A, B and C of sizes such that the product ABC is defined, which implies for $A \in \mathbb{R}^{n\times p}$ and $B \in \mathbb{R}^{p\times q}$

$$\operatorname{vec}(AB) = (B^T \otimes I_n) \operatorname{vec} A = (B^T \otimes A) \operatorname{vec} I_p = (I_q \otimes A) \operatorname{vec} B. \tag{A.2}$$

The functions we wish to study of are of the type

$$F : S \subset \mathbb{R}^{p\times p} \to \mathbb{R}^{p\times p}. \tag{A.3}$$

As an example take matrix inversion, i.e. $F(X) = X^{-1}$. Then S is the set of all non-singular $p \times p$ matrices. Of course, what is said in the following also applies to functions of the more general form $F : S \subset \mathbb{R}^{n\times p} \to \mathbb{R}^{m\times q}$, thus including vectors and scalars as special cases. By abstaining from maximum generality in notation we avoid the hassle of too many indices.
Let $\mathbb{D}F(X)$ denote the derivative of F at the point X. Before we come to computing $\mathbb{D}F(X)$ we have to know what $\mathbb{D}F$ is. Clearly, $\mathbb{D}F(X)$ is the collection of all p^4 partial derivatives

$$\frac{\partial f_{i,j}(X)}{\partial x_{k,l}}, \qquad i, j, k, l = 1, ..., p.$$

The only question is how they are arranged. For example, if we agree that $\mathbb{D}F(X)$ is a $p^2 \times p^2$ matrix, what do we put in the upper left $p \times p$ block? All partials of $f_{1,1}$ with respect to X or all partials of

F with respect to $x_{1,1}$? Magnus and Neudecker (1999, pp. 95,171–174) advise to do neither, instead identify X and F with $\operatorname{vec} X$ and $\operatorname{vec} F$, respectively, and thus define $\mathbb{D}F(X)$ as the Jacobi-matrix of $\operatorname{vec} F$ with respect to $\operatorname{vec} X$. This fixes the order of the partials within $\mathbb{D}F(X)$ to

$$[\mathbb{D}F(X)]_{(i-1)p+j,(k-1)p+l} = \frac{\partial f_{i,j}(X)}{\partial x_{k,l}}.$$

There are very good reasons for this ordering, as will become apparent shortly. Behind it is the basic idea that a mapping from matrices to matrices is essentially a mapping from vector to vector. Thus, consequently using the canonical matrix-vector mapping vec matrix differential calculus in principle boils down to ordinary multivariate differential calculus. For the problem at hand matrices are most of all notational representations of vectors that allow very nice and concise formulations of complicated functional dependencies, like e.g. matrix inversion.
Now obviously we can compute $\mathbb{D}F(X)$ by determining p^4 partial derivatives. It is just as obvious that we want to avoid that. We do not even want to bother how e.g. matrix inversion is written element-wise in dimensions higher than two. In order to motivate the basic rules of matrix differential calculus we take a brief look at the univariate case. There are two fundamental rules for computing derivatives, the *chain rule*,

$$(g \circ h)'(x) = g'(h(x))h'(x), \qquad x \in \mathbb{R}, \tag{A.4}$$

and the *multiplication formula*

$$(gh)(x) = g'(x)h(x) + g(x)h'(x), \qquad x \in \mathbb{R}. \tag{A.5}$$

These two rules are indeed very powerful. Together with the basics (derivatives of linear and constants functions) and results about Taylor expansions they suffice to compute the derivative of basically any differentiable function $f : \mathbb{R} \to \mathbb{R}$. For example, differentiating both sides of

$$1 = x\frac{1}{x}$$

by making use of (A.5) yields the derivative of $f : x \mapsto \frac{1}{x}$,

$$f'(x) = -\frac{1}{x^2}.$$

Fortunately there exist multivariate analogues to (A.4) and (A.5), which we present in the following. The chain rule is straightforward, cp. MN, pp. 91, 96,

$$\mathbb{D}(G \circ H)(X) = \mathbb{D}G(H(X))\mathbb{D}H(X) \tag{A.6}$$

Note that the definition of $\mathbb{D}F(X)$ as the derivative of $\operatorname{vec} F$ w.r.t. $\operatorname{vec} X$ ensures that this formula also applies to matrix valued functions.
The core of the matrix calculus is the matrix product, for which there is indeed a differentiation rule similar to (A.5). However, its formulation can apparently not be a generalization of (A.5) as straightforward as (A.6) generalizes (A.4): Simply replacing scalar-valued functions by vector-valued functions and derivatives by Jacobi-matrices in (A.5) cannot lead to a sensible formula, since the dimensions of left- and right-hand side of the equation do not fit. In order to state a multivariate multiplication rule we have to resort to another term, the *differential*, cf. MN, p. 81. The differential

of the function $F : S \subset \mathbb{R}^{p\times p} \to \mathbb{R}^{p\times p}$ at the point $X \in \mathbb{R}^{p\times p}$ with increment $C \in \mathbb{R}^{p\times p}$ is denoted by $dF(X; C)$ and defined as

$$dF(X; C) = \underset{p\times p}{\mathrm{mat}}\,(\mathbb{D}F(X)\,\mathrm{vec}\,C). \tag{A.7}$$

Thus $F(X_0)+dF(X; X-X_0)$ is the affine linear approximation of $F(X)$ by means of F and its derivative $\mathbb{D}F$ at the point X_0. This means for "well behaved" functions F,

$$F(X) - F(X_0) - dF(X; X - X_0) = o(\|X - X_0\|) \qquad \text{as } X \to X_0.$$

Note that dF is of the same shape as F, that is a $p \times p$ matrix here.

Remark. Usually one writes dX for the increment C, which is and stays formally an arbitrary point in $\mathbb{R}^{p\times p}$ even though in light of the approximation we think of it as small in some sense. Furthermore one neglects the dependence of the differential dF on the increment dX, i.e. one writes $dF(X)$ instead of $dF(X; dX)$. Then (mis-)using X as a variable name and as a function name (i.e. depending on the context dX may denote a point in $\mathbb{R}^{p\times p}$ or the differential of the function X) is in concordance with the chain rule and leads to a (maybe mathematically slightly unsound but) very handy notation for practical purposes. Let F be a function of X, then by (A.7):

$$\mathrm{vec}\,dF(X) = \mathbb{D}F(X)\,\mathrm{vec}\,dX. \tag{A.8}$$

If now X is in turn a function of, say, Y, i.e.

$$\mathrm{vec}\,dX(Y) = \mathbb{D}X(Y)\,\mathrm{vec}\,dY, \tag{A.9}$$

then we have for the composite function $F \circ X$ via the chain rule

$$\begin{aligned}
\mathrm{vec}\,d(F \circ X)(Y) &\overset{(A.7)}{=} \mathbb{D}(F \circ X)\,\mathrm{vec}\,dY \\
&\overset{(A.6)}{=} \mathbb{D}F(X(Y))\mathbb{D}X(Y)\,\mathrm{vec}\,dY \\
&\overset{(A.9)}{=} \mathbb{D}F(X(Y))\,\mathrm{vec}\,dX(Y)
\end{aligned}$$

If we drop the Y on both sides of the equation and write $F(X)$ for $F \circ X$, then the last equation reads as

$$\mathrm{vec}\,dF(X) = \mathbb{D}F(X)\,\mathrm{vec}\,dX, \tag{A.10}$$

which is exactly the same as (A.8). Thus (A.8), which is essentially the definition of differential, is, when X is interpreted as a function, a very concise way of writing down the chain rule.

Here is the multivariate multiplication formula stated in terms of differentials

$$d(GH)(X; C) = dG(X; C)H(X) + G(X)dH(X; C) \tag{A.11}$$

or in short notation

$$d(GH) = (dG)H + GdH, \tag{A.12}$$

cp. MN, p. 148. By vectorizing (A.11) using (A.2) we obtain

$$\mathbb{D}(GH)(X) = (G(X)^T \otimes I_p)\mathbb{D}H(X) + (I_p \otimes H(X))\mathbb{D}G(X). \tag{A.13}$$

Thus we have established the chain rule (A.6) and the multiplication rule (A.13) for differentiating matrices. For practical computational matters and for memorizing the formulations in terms of differentials, (A.8) and (A.12), respectively, are much more convenient.

A.2 Differentiating w.r.t. symmetric matrices

We use the notation of Section 4.2, in particular recall $m = p(p+1)/2$, the function $\theta : \mathbb{R}^m \to \mathscr{S}_p : a \mapsto \mathrm{mat}_{p\times p}\, D_p a$ and $\mathscr{S}_p$, the set of all real, symmetric $p \times p$ matrices. Let

$$f : \mathbb{R}^{p\times p} \to \mathbb{R}^q$$

be continuously differentiable, and define

$$g : \mathscr{S}_p \to \mathbb{R}^q : A \mapsto f(A).$$

Is g differentiable? Apparently not, the set $\{\mathrm{vec}\, S \mid S \in \mathscr{S}_p\}$ contains no inner point in $\mathbb{R}^{p^2}$. However, this thesis is concerned with shape estimators, i.e. random variables with realizations in $\mathscr{S}_p$, and differentiable functions thereof. Computing derivatives of such functions is a vital part of many proofs in the thesis. We have to clarify in what sense g is *differentiable*, and what the *derivative* of g is.
Any $S \in \mathscr{S}_p$ can perceived as a *representation* of $v(S) \in \mathbb{R}^m$, and consequently g as a representation of the function

$$\bar{g} : \mathbb{R}^m \to \mathbb{R}^q : a \mapsto f(\theta(a)) = f(\mathrm{mat}_{p\times p}\, D_p a),$$

which is very well continuously differentiable. Then the derivative of g at a point $S \in \mathscr{S}_p$, denoted by $\mathbb{D}g(S)$, is a representation of the derivative $\mathbb{D}\bar{g}(v(S)) \in \mathbb{R}^{q\times m}$. This representation is a $q \times p^2$ matrix, which should satisfy

$$\mathbb{D}g(D_p s)D_p c \;=\; \mathbb{D}\bar{g}(s)c \tag{A.14}$$

for all $s, c \in \mathbb{R}^m$, meaning that the corresponding differentials, cf. (A.7), of g and $\bar{g}$ are identical. Noting that $\mathbb{D}\bar{g}(s) = \mathbb{D}f(\theta(s))D_p$ by the chain rule, (A.14) is equivalent to

$$\mathbb{D}g(S)D_p c \;=\; \mathbb{D}f(S)D_p c,$$

for all $s, c \in \mathbb{R}^m$, where $S = \theta(s)$, which is again equivalent to

$$\mathbb{D}g(S)D_p \;=\; \mathbb{D}f(S)D_p \qquad \text{for all } S \in \mathscr{S}_p \tag{A.15}$$

This relation does not uniquely specify $\mathbb{D}g(S)$ as a function of $\mathbb{D}f(S)$. If we let $[B]_k$ denote the kth column of a matrix B, then (A.15) fixes all columns

$$[\mathbb{D}g(S)]_{(i-1)p+i}, \qquad 1 \le i \le p.$$

Furthermore, for any pair $1 \le i < j \le p$,

$$[\mathbb{D}g(S)]_{(i-1)p+j} + [\mathbb{D}g(S)]_{(j-1)p+i}$$

is also fixed. Adding one more requirement, that columns $(i-1)p + j$ and $(j-1)p + i$ shall be equal, because they contain partial derivatives with respect to the same variable, fully fixes $\mathbb{D}g(S)$ to

$$\mathbb{D}g(S) = \mathbb{D}f(S)M_p \qquad \text{for all } S \in \mathscr{S}_p.$$

Final remark. This derivative for symmetric matrices is the same as defined in Srivastava and Khatri (1979) and Bilodeau and Brenner (1999).

Bibliography

Baba, K., Shibata, R., Sibuya, M.: Partial correlation and conditional correlation as measures of conditional independence. Aust. N. Z. J. Stat. **46**(4), 657–664 (2004)

Becker, C.: Iterative proportional scaling based on a robust start estimator. In: Weihs, C., Gaul, W. (eds.) Classification - The Ubiquitous Challenge, pp. 248–255. Heidelberg: Springer (2005)

Bensmail, H., Celeux, G.: Regularized Gaussian discriminant analysis through eigenvalue decomposition. J. Am. Stat. Assoc. **91**(436), 1743–1748 (1996)

Bilodeau, M., Brenner, D.: Theory of multivariate statistics. Springer Texts in Statistics. New York, NY: Springer (1999)

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

Buhl, S.L.: On the existence of maximum likelihood estimators for graphical Gaussian models. Scand. J. Stat. **20**(3), 263–270 (1993)

Butler, R.W., Davies, P.L., Jhun, M.: Asymptotics for the minimum covariance determinant estimator. Ann. Stat. **21**(3), 1385–1400 (1993)

Capitanio, A., Azzalini, A., Stanghellini, E.: Graphical models for skew-normal variates. Scand. J. Stat. **30**(1), 129–144 (2003)

Castelo, R., Roverato, A.: A robust procedure for Gaussian graphical model search from microarray data with p larger than n. J. Mach. Learn. Res. **7**, 2621–2650 (2006)

Comon, P.: Independent component analysis, a new concept? Signal Process. **36**(3), 287–314 (1994)

Cox, D.R., Wermuth, N.: Multivariate dependencies: models, analysis and interpretation. Monographs on Statistics and Applied Probability. 67. London: Chapman and Hall (1996)

Croux, C., Dehon, C., Yadine, A.: The k-step spatial sign covariance matrix. Adv. Data Anal. Classif. **4**(2-3), 137–150 (2010)

Croux, C., Haesbroeck, G.: Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. J. Multivariate Anal. **71**(2), 161–190 (1999)

Croux, C., Ollila, E., Oja, H.: Sign and rank covariance matrices: statistical properties and application to principal components analysis. In: Dodge, Y. (ed.) Statistical Data Analysis Based on the L_1-Norm and Related Methods (Papers of the 4th international conference on statistical analysis on the L_1-norm and related methods, Neuchâtel, Switzerland, August 4–9, 2002), pp. 257–269. Basel: Birkhäuser (2002)

Dahlhaus, R.: Graphical interaction models for multivariate time series. Metrika **51**(2), 157–172 (2000)

Davies, P.L.: Asymptotic behaviour of S-estimates of multivariate location parameters and dispersion matrices. Ann. Stat. **15**, 1269–1292 (1987)

Davies, P.L.: The asymptotics of Rousseeuw's minimum volume ellipsoid estimator. Ann. Stat. **20**(4), 1828–1843 (1992)

Davies, P.L., Gather, U.: The identification of multiple outliers. J. Am. Stat. Assoc. **88**(423), 782–801 (1993)

Davies, P.L., Gather, U.: Breakdown and groups. Ann. Stat. **33**(3), 977–1035 (2005)

Dawid, A.P., Lauritzen, S.L.: Hyper Markov laws in the statistical analysis of decomposable graphical models. Ann. Stat. **21**(3), 1272–1317 (1993)

Deming, W.E., Stephan, F.F.: On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. Ann. Math. Stat. **11**, 427–444 (1940)

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

Donoho, D.L.: Breakdown properties of multivariate location estimators. Ph.D. thesis, Harvard University (1982)

Donoho, D.L., Huber, P.J.: The notion of breakdown point. In: Bickel, P.J., Doksum, K.A., Hodges, J.L. (eds.) Festschrift for Erich L. Lehmann., pp. 157–183. Belmont, CA: Wadsworth (1983)

Drton, M., Perlman, M.D.: Model selection for Gaussian concentration graphs. Biometrika **91**(3), 591–602 (2004)

Drton, M., Perlman, M.D.: A SINful approach to Gaussian graphical model selection. J. Stat. Plann. Inference **138**(4), 1179–1200 (2008)

Dürre, A.: Über die ersten beiden Momente der Vorzeichen-Kovarianz-Matrix im elliptischen Modell. Bachelor's thesis, Technische Universität Dortmund, Germany (2010)

Edwards, D.: Introduction to graphical modelling. Springer Texts in Statistics. New York, NY: Springer (2000)

Edwards, D., Havránek, T.: A fast procedure for model search in multidimensional contingency tables. Biometrika **72**, 339–351 (1985)

Eriksen, P.S.: Tests in covariance selection models. Scand. J. Stat. **23**(3), 275–284 (1996)

Fang, K.T., Zhang, Y.T.: Generalized multivariate analysis. Berlin etc.: Springer-Verlag; Beijing: Science Press. (1990)

Fischer, D.: Statistische Eigenschaften des Oja-Medians mit einer algorithmischen Betrachtung. Diplomarbeit, Technische Universität Dortmund, Germany (2008)

Fischer, D., Möttönen, J., Nordhausen, K., Vogel, D.: OjaNP: Multivariate Methods Based on the Oja median and Related Concepts (2009). R package version 0.0-24

Forster, O.: Analysis 2. Differentialrechnung im $\mathbb{R}^n$, Gewöhnliche Differentialgleichungen. 5th edn. Vieweg Studium: Grundkurs Mathematik. Wiesbaden: Vieweg. (1982)

Fried, R., Didelez, V.: Decomposability and selection of graphical models for multivariate time series. Biometrika **90**(2), 251–267 (2003)

Gather, U., Imhoff, M., Fried, R.: Graphical Models for Multivariate Time Series form Intensive Care Monitoring. Statistics in Medicine **21**, 2685–2701 (2002)

Gervini, D.: The influence function of the Stahel–Donoho estimator of multivariate location and scatter. Stat. Probab. Lett. **60**(4), 425–435 (2002)

Gervini, D.: A robust and efficient adaptive reweighted estimator of multivariate location and scatter. J. Multivariate Anal. **84**(1), 116–144 (2003)

Gervini, D.: Robust functional estimation using the median and spherical principal components. Biometrika **95**, 587–600 (2008)

Gnanadesikan, R., Kettenring, J.R.: Robust estimates, residuals, and outlier detection with multiresponse data. Biometrics **28**(1), 81–124 (1972)

Gottard, A., Pacillo, S.: On the impact of contaminations in graphical Gaussian models. Stat. Methods Appl. **15**(3), 343–354 (2007)

Gottard, A., Pacillo, S.: Robust concentration graph model selection. Comput. Statist. Data Anal. **54**(12), 3070–3079 (2010)

Haldane, J.B.S.: Note on the median of a multivariate distribution. Biometrika **35**, 414–415 (1948)

Hallin, M., Oja, H., Paindaveine, D.: Semiparametrically efficient rank-based inference for shape. II: Optimal *R*-estimation of shape. Ann. Stat. **34**(6), 2757–2789 (2006)

Hampel, F.R.: A general qualitative definition of robustness. Ann. Math. Stat. **42**, 1887–1896 (1971)

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A.: Robust statistics. The approach based on influence functions. Wiley Series in Probability and Mathematical Statistics. New York etc.: Wiley (1986)

Harville, D.A.: Matrix algebra from a statistician's perspective. New York, NY: Springer. (1997)

Hettmansperger, T., Randles, R.: A practical affine equivariant multivariate median. Biometrika **89**, 851–860 (2002)

Hettmansperger, T.P., McKean, J.W.: Robust nonparametric statistical methods. Kendall's Library of Statistics. 5. London: Arnold. New York, NY: Wiley. (1998)

Huber, P.J., Ronchetti, E.M.: Robust statistics. 2nd edn. Wiley Series in Probability and Statistics. Hoboken, NJ: Wiley (2009)

Hyvärinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. Wiley (2001)

Karvanen, J., Koivunen, V.: Blind separation methods based on Pearson system and its extensions. Signal Process. **82**(4), 663–673 (2002)

Kemperman, J.H.B.: The median of a finite measure on a Banach space. In: Dodge, Y. (ed.) Statistical Data Analysis Based on the L_1-Norm and Related Methods, pp. 217–230. Amsterdam: North-Holland (1987)

Kent, J.T., Tyler, D.E.: Constrained M-estimation for multivariate location and scatter. Ann. Stat. **24**(3), 1346–1370 (1996)

Koltchinskii, V., Dudley, R.: On spatial quantiles. In: Korolyuk, V. et al. (ed.) Skorokhod's ideas in probability theory., pp. 195–210. Kiev: Institute of Mathematics of NAS of Ukraine. Proc. Inst. Math. Natl. Acad. Sci. Ukr., Math. Appl. 32 (2000)

Kuhnt, S., Becker, C.: Sensitivity of graphical modeling against contamination. In: Schader, Martin et al. (ed.) Between data science and applied data analysis (Proceedings of the 26th annual conference of the Gesellschaft für Klassifikation e. V., Mannheim, Germany, July 22–24, 2002), pp. 279–287. Berlin: Springer (2003)

Lauritzen, S.L.: Graphical models. Oxford Statistical Science Series. 17. Oxford: Oxford Univ. Press (1996)

Locantore, N., Marron, J., Simpson, D., Tripoli, N., Zhang, J., Cohen, K.: Robust principal component analysis for functional data. (With comments). Test **8**(1), 1–73 (1999)

Lopuhaä, H.P.: On the relation between S-estimators and M-estimators of multivariate location and covariance. Ann. Stat. **17**(4), 1662–1683 (1989)

Lopuhaä, H.P.: Multivariate τ-estimators for location and scatter. Can. J. Stat. **19**(3), 307–321 (1991)

Lopuhaä, H.P.: Asymptotics of reweighted estimators of multivariate location and scatter. Ann. Stat. **27**(5), 1638–1665 (1999)

Magnus, J.R., Neudecker, H.: Matrix differential calculus with applications in statistics and econometrics. 2nd edn. Wiley Series in Probability and Statistics. Chichester: Wiley (1999)

Marden, J.I.: Some robust estimates of principal components. Stat. Probab. Lett. **43**(4), 349–359 (1999)

Maronna, R.A.: Robust M-estimators of multivariate location and scatter. Ann. Stat. **4**, 51–67 (1976)

Maronna, R.A., Martin, D.R., Yohai, V.J.: Robust statistics: Theory and methods. Wiley Series in Probability and Statistics. Chichester: Wiley (2006)

Maronna, R.A., Yohai, V.J.: The behavior of the Stahel-Donoho robust multivariate estimator. J. Am. Stat. Assoc. **90**(429), 330–341 (1995)

Maronna, R.A., Zamar, R.H.: Robust estimates of location and dispersion for high-dimensional datasets. Technometrics **44**, 307–317 (2002)

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

Milasevic, P., Ducharme, G.R.: Uniqueness of the spatial median. Ann. Stat. **15**, 1332–1333 (1987)

Miyamura, M., Kano, Y.: Robust Gaussian graphical modeling. J. Multivariate. Anal. **97**(7), 1525–1550 (2006)

Oja, H., Paindaveine, D., Taskinen, S.: Parametric and nonparametric tests for multivariate independence in the independence component model. submitted. (2010)

Ollila, E., Croux, C., Oja, H.: Influence function and asymptotic efficiency of the affine equivariant rank covariance matrix. Stat. Sin. **14**(1), 297–316 (2004)

Ollila, E., Oja, H., Croux, C.: The affine equivariant sign covariance matrix: Asymptotic behavior and efficiencies. J. Multivariate Anal. **87**(2), 328–355 (2003)

Paindaveine, D.: A canonical definition of shape. Stat. Probab. Lett. **78**(14), 2240–2247 (2008)

Porteous, B.: A note on improved likelihood ratio statistics for generalized log linear models. Biometrika **72**, 473–475 (1985)

Porteous, B.T.: Stochastic inequalities relating a class of log-likelihood ratio statistics to their asymptotic χ^2 distribution. Ann. Stat. **17**(4), 1723–1734 (1989)

Rocke, D.M.: Robustness properties of S-estimators of multivariate location and shape in high dimension. Ann. Stat. **24**(3), 1327–1345 (1996)

Rousseeuw, P.J.: Multivariate estimation with high breakdown point. In: Grossmann, W., Pflug, G.C., Vincze, I., Wertz, W. (eds.) Mathematical statistics and applications, Proc. 4th Pannonian Symp. Math. Stat., Bad Tatzmannsdorf, Austria, September 4-10, 1983, Vol. B, pp. 283–297. Dordrecht etc.: D. Reidel (1985)

Rousseeuw, P.J., Leroy, A.M.: Robust regression and outlier detection. Wiley Series in Probability and Mathematical Statistics. New York etc.: Wiley. (1987)

Rousseeuw, P.J., Van Driessen, K.: A fast algoritm for the minimum covariance determinant estimator. Technometrics **41**, 212–233 (1999)

Roverato, A.: Cholesky decomposition of a hyper inverse Wishart matrix. Biometrika **87**(1), 99–112 (2000)

Schelter, B., Winterhalder, M., Hellwig, B., Guschlbauer, B., Lücking, C.H., Timmer, J.: Direct or Indirect? Graphical Models for Neural Oscillators. J. Physiol. **99**, 37–46 (2006)

Sirkiä, S., Taskinen, S., Oja, H., Tyler, D.E.: Tests and estimates of shape based on spatial signs and ranks. J. Nonparametric Stat. **21**(2), 155–176 (2009)

Smith, P.W.F.: Assessing the power of model selection procedures used when graphical modelling. In: Dodge, Y., Whittaker, J. (eds.) Coputational Statistics, Volume I, pp. 275–280. Heidelberg: Physica (1992)

Speed, T.P., Kiiveri, H.T.: Gaussian Markov distributions over finite graphs. Ann. Stat. **14**, 138–150 (1986)

Srivastava, M., Khatri, C.: An introduction to multivariate statistics. New York, Oxford: North Holland (1979)

Stahel, W.: Robust estimation: Infinitesimal optimality and covariance matrix estimation. Ph.D. thesis, ETH Zürich (1981)

Theis, F.J.: A new concept for separability problems in blind source separation. Neural Comput. **16**(9), 1827–1850 (2004)

Tyler, D.E.: Radial estimates and the test for sphericity. Biometrika **69**, 429–436 (1982)

Tyler, D.E.: Robustness and efficiency properties of scatter matrices. Biometrika **70**, 411–420 (1983)

Tyler, D.E.: A distribution-free M-estimator of multivariate scatter. Ann. Stat. **15**, 234–251 (1987a)

Tyler, D.E.: Statistical analysis for the angular central Gaussian distribution on the sphere. Biometrika **74**, 579–589 (1987b)

Tyler, D.E.: A note on multivariate location and scatter statistics for sparse data sets. Statistics & Probability Letters **80**(17-18), 1409 – 1413 (2010)

Verzelen, N., Villers, F.: Tests for gaussian graphical models. Comput. Stat. Data Anal. **53**(5), 1894–1905 (2009)

Visuri, S.: Array and multichannel signal processing using nonparametric statistics. Ph.D. thesis, Helsinki University of Technology, Helsinki, Finland (2001)

Visuri, S., Koivunen, V., Oja, H.: Sign and rank covariance matrices. J. Stat. Plann. Inference **91**(2), 557–575 (2000)

Visuri, S., Oja, H., Koivunen, V.: Subspace-Based Direction-of-Arrival Estimation Using Nonparamatric Statistics. IEEE Trans. Signal Process. **49**(9), 2060–2073 (2001)

Vogel, D.: On generalizing Gaussian graphical models. In: Ciumara, R., Bădin, L. (eds.) Proceedings of the 16th European Young Statisticians Meeting, pp. 149–153. University of Bucharest (2009)

Vogel, D., Dürre, A., Fried, R.: Elliptical graphical modeling in higher dimensions. In: Wessel, N. (ed.) Proceedings of International Biosignal Processing Conference, July 14-16, 2010, Berlin, Germany., pp. 1–5 (2010)

Vogel, D., Fried, R.: Estimating partial correlations using the oja sign covariance matrix. In: Brito, P. (ed.) Compstat 2008: Proceedings in Computational Statistics. Vol. II, pp. 721–729. Heidelberg: Physica-Verlag (2008)

Vogel, D., Fried, R.: On robust Gaussian graphical modelling. In: Devroye, L., Karasözen, B., Kohler, M., Korn, R. (eds.) Recent Developments in Applied Probability and Statistics. Dedicated to the Memory of Jürgen Lehn., pp. 155–182. Berlin, Heidelberg: Springer-Verlag (2010)

Vogel, D., Köllmann, C., Fried, R.: Partial correlation estimates based on signs. In: Heikkonen, J. (ed.) Proceedings of the 1st Workshop on Information Theoretic Methods in Science and Engineering. TICSP series # 43 (2008)

Whittaker, J.: Graphical models in applied multivariate statistics. Wiley Series in Probability and Mathematical Statistics. Chichester etc.: Wiley (1990)

Yadine, A.: Robustness and efficiency of multivariate scatter estimators. Ph.D. thesis, Université Libre de Bruxelles, Brussels, Belgium (2006)

Yuan, M., Lin, Y.: Model selection and estimation in the Gaussian graphical model. Biometrika **94**(1), 19–35 (2007)

Zuo, Y.: Robust location and scatter estimators in multivariate analysis. In: Fan, J., Koul, H. (eds.) Frontiers in statistics. Dedicated to Peter John Bickel on honor of his 65th birthday, pp. 467–490. London: Imperial College Press (2006)

Zuo, Y., Cui, H.: Depth weighted scatter estimators. Ann. Stat. **33**(1), 381–413 (2005)