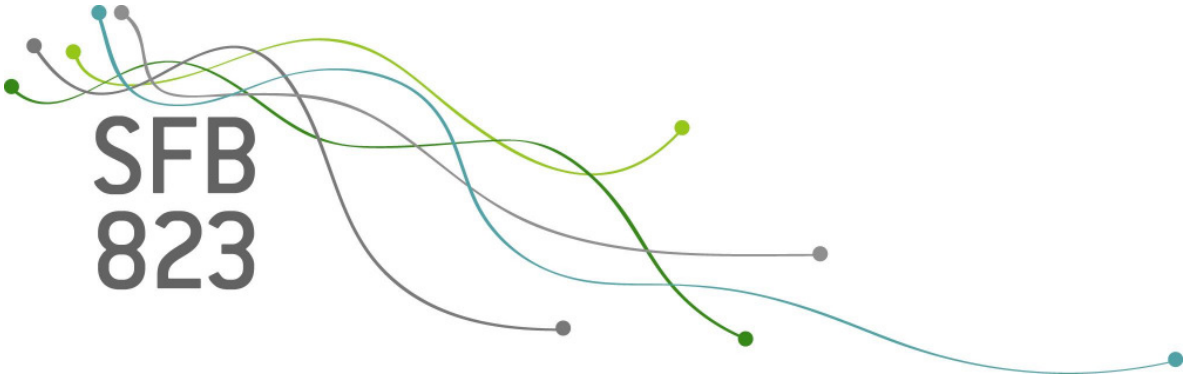
SFB 823

On robust Gaussian Graphical Modelling

Daniel Vogel, Roland Fried

Nr. 36/2009

Discussion Paper



On Robust Gaussian Graphical Modelling

Daniel Vogel and Roland Fried

Abstract The objective of this exposition is to give an overview of the existing approaches to robust Gaussian graphical modelling. We start by thoroughly introducing Gaussian graphical models (also known as *covariance selection models* or *concentration graph models*) and then review the established, likelihood-based statistical theory (estimation, testing and model selection). Afterwards we describe robust methods and compare them to the classical approaches.

1 Introduction

Graphical modelling is the analysis of conditional associations between random variables by means of graph theoretic methods. The graphical display 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.

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 chronol-

Daniel Vogel
Fakultät Statistik, Technische Universität Dortmund, e-mail: daniel.vogel@tu-dortmund.de

Roland Fried
Fakultät Statistik, Technische Universität Dortmund, e-mail: fried@statistik.tu-dortmund.de

ogy 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. 2.3. 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. 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 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 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 and 3 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 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 $\mathbf{b}$, $\boldsymbol{\mu}$, etc., denote vectors, capital letters X, Y, etc., indicate (univariate) random variables and bold capital letters $\mathbf{X}$, $\mathbf{Y}$, 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 $(\mathbf{X}, \mathbf{Y})$ 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 brack-

ets 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 $A \in \mathscr{S}_p^+$

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

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

$$\begin{pmatrix} \Sigma_{\alpha,\alpha} & \Sigma_{\alpha,\beta} \\ \Sigma_{\beta,\alpha} & \Sigma_{\beta,\beta} \end{pmatrix}^{-1} = \begin{pmatrix} \Omega^{-1} & -\Omega^{-1}\Sigma_{\alpha,\beta}\Sigma^{-1}_{\beta,\beta} \\ -\Sigma^{-1}_{\beta,\beta}\Sigma_{\beta,\alpha}\Omega^{-1} & \Sigma^{-1}_{\beta,\beta} + \Sigma^{-1}_{\beta,\beta}\Sigma_{\beta,\alpha}\Omega^{-1}\Sigma_{\alpha,\beta}\Sigma^{-1}_{\beta,\beta} \end{pmatrix}, \tag{2}$$

where the $r \times r$ matrix $\Omega = \Sigma_{\alpha,\alpha} - \Sigma_{\beta,\alpha}\Sigma^{-1}_{\beta,\beta}\Sigma_{\beta,\alpha}$ 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 $(i(p-1)+k, j(p-1)+l)$. Let $\mathbf{e}_1, ..., \mathbf{e}_p$ be the unit vectors in $\mathbb{R}^p$ and $\mathbf{1}_p$ the p vector consisting only of ones. Define further the following matrices:

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

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 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. 2.1), review basic graph theory terms and explain the merit of a *partial correlation graph* (Sect. 2.2). Gaussianity enters in Sect. 2.3, where we deduce the conditional independence interpretation of a partial correlation graph which is valid under normality. Statistics is deferred to Sect. 3.

2.1 Partial variance

Let $\mathbf{X} = (X_1, ..., X_p)$ be a random vector in $\mathbb{R}^p$ with distribution F and positive definite variance matrix $\Sigma = \Sigma_{\mathbf{X}} \in \mathbb{R}^{p\times p}$. The inverse of Σ is called *concentration matrix* (or *precision matrix*) of $\mathbf{X}$ and shall be denoted by K or $K_{\mathbf{X}}$.

Now let $\mathbf{X}$ be partitioned into $\mathbf{X} = (\mathbf{Y}, \mathbf{Z})$, where $\mathbf{Y}$ and $\mathbf{Z}$ 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 $\mathbf{Y}$ is $\Sigma_{\mathbf{Y}} = \Sigma_{\alpha,\alpha} \in \mathbb{R}^{q\times q}$ and its concentration matrix $K_{\mathbf{Y}} = \Sigma^{-1}_{\alpha,\alpha} = (K^{-1}_{\mathbf{X}})^{-1}_{\alpha,\alpha}$. The covariance matrix of $\mathbf{Y}$ and $\mathbf{Z}$ is $\Sigma_{\alpha,\beta} \in \mathbb{R}^{q\times r}$. The orthogonal projection of $\mathbf{Y}$ onto the space of all affine linear functions of $\mathbf{Z}$ shall be denoted by $\hat{\mathbf{Y}}(\mathbf{Z})$ and is given by

$$\hat{\mathbf{Y}}(\mathbf{Z}) = \mathbb{E}\mathbf{Y} + \Sigma_{\alpha,\beta}\Sigma^{-1}_{\beta,\beta}(\mathbf{Z} - \mathbb{E}\mathbf{Z}). \tag{3}$$

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

$$\Sigma_{\mathbf{Y}\bullet\mathbf{Z}} = \mathrm{Var}(\mathbf{Y} - \hat{\mathbf{Y}}(\mathbf{Z})) = \Sigma_{\alpha,\alpha} - \Sigma_{\alpha,\beta}\Sigma^{-1}_{\beta\beta}\Sigma_{\beta,\alpha}. \tag{4}$$

The notation $\mathbf{Y}\bullet\mathbf{Z}$ is intended to resemble $\mathbf{Y}|\mathbf{Z}$, that is, we look at $\mathbf{Y}$ in dependence on $\mathbf{Z}$, but instead of conditioning $\mathbf{Y}$ on $\mathbf{Z}$ the type of connection we consider here is a linear regression. In particular, $\Sigma_{\mathbf{Y}\bullet\mathbf{Z}}$ is—contrary to a conditional variance—a fixed parameter and not random.

If $\mathbf{Y}$ is at least two-dimensional, we partition it further into $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2)$ with corresponding index sets $\alpha_1 \cup \alpha_2 = \alpha$ and lengths $q_1 + q_2 = q$, and define

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

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

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

Furthermore, if $\mathbf{Y}_1 = Y_1$ and $\mathbf{Y}_2 = Y_2$ are both one-dimensional, $\Sigma_{\mathbf{Y}\bullet\mathbf{Z}}$ is a positive definite 2×2 matrix. The correlation coefficient computed from this matrix, i.e. the $(1,2)$ element of $\mathrm{Corr}(\Sigma_{\mathbf{Y}\bullet\mathbf{Z}})$, cf. (1), is called the *partial correlation (coefficient) of* Y_1 *and* Y_2 *given* $\mathbf{Z}$ and denoted by $\varrho_{Y_1,Y_2\bullet\mathbf{Z}}$. This is nothing but the correlation between the residuals $Y_1 - \hat{Y}_1(\mathbf{Z})$ and $Y_2 - \hat{Y}_2(\mathbf{Z})$ and may be interpreted as a measure of the linear association between Y_1 and Y_2 after the linear effects of $\mathbf{Z}$ have been removed. For $\alpha_1 = \{i\}$ and $\alpha_2 = \{j\}$, $i \neq j$, we use the simplified notation $\varrho_{i,j\bullet}$ for $\varrho_{X_i,X_j\bullet\mathbf{X}_{\setminus\{i,j\}}}$.

The simple identity (4) is fundamental and the actual starting point for all following considerations. We recognize $\Sigma_{\mathbf{Y}\bullet\mathbf{Z}}$ as the Schur complement of $\Sigma_{\mathbf{Z}}$ in $\Sigma_{\mathbf{X}}$, cf. (2), implying that

$$\Sigma^{-1}_{\mathbf{Y}\bullet\mathbf{Z}} = K_{\alpha,\alpha}. \tag{5}$$

In words: the concentration matrix of $\mathbf{Y} - \hat{\mathbf{Y}}(\mathbf{Z})$ is the submatrix of $K_{\mathbf{X}}$ corresponding to $\mathbf{Y}$, 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 1. *The partial correlation* $\varrho_{i,j\bullet}$ *between* X_i *and* X_j*,* $1 \leq i < j \leq p$*, given all remaining variables* $\mathbf{X}_{\setminus\{i,j\}}$ *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* $\mathbf{X}$. Proposition 1 is a direct consequence of (5) involving the inversion of a 2×2 matrix. For a detailed derivation see Whittaker (1990), Chap. 5.

2.2 Partial correlation graph

The partial correlation structure of the random variable $\mathbf{X}$ 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 $\mathbf{X}$. 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 1. 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 1 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. 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, $i = 1,...,k$,

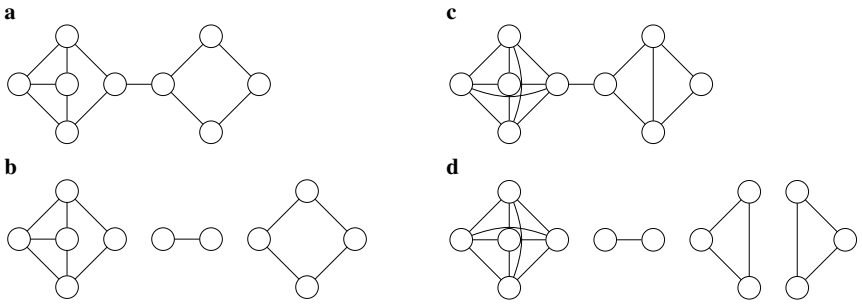


Fig. 1 **a** a non-decomposable graph and **b** its maximal irreducible components, **c** a decomposable graph and **d** its maximal irreducible components

(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 3. 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 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. 3.1. The recursive nature of Definition 3 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.

Definition 4. The graph G is *triangulated*, if every cycle of length greater than 3 has a chord.

Proposition 2. *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 1. (Separation theorem for PCGs) For a random vector $\mathbf{X}$ 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 $\mathbf{X}_\alpha \perp \mathbf{X}_\beta \bullet \mathbf{X}_\gamma$.

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. 2.3. 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 1. The theorem lets us, starting from the pairwise partial correlations, conclude the partial uncorrelatedness $\mathbf{X}_\alpha \perp \mathbf{X}_\beta \bullet \mathbf{X}_\gamma$ for a variety of triples $(\mathbf{X}_\alpha, \mathbf{X}_\beta, \mathbf{X}_\gamma)$ (which do not have to form a partition of $\mathbf{X}$). 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 1 can be re-phrased, saying that in a PCG the pairwise and the global Markov property are equivalent: We say, a random vector $\mathbf{X} = (X_1, ..., X_p)$ satisfies the *pairwise Markov property w.r.t. the partial correlation graph* $G = (\{1, ..., p\}, E)$, if $\{i, j\} \notin E \Rightarrow X_i \perp X_j \bullet \mathbf{X}_{\setminus\{i,j\}}$, that is, the edge set of the PCG of $\mathbf{X}$ is a subset of E. $\mathbf{X}$ 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 $\mathbf{X}_\alpha \perp \mathbf{X}_\beta \bullet \mathbf{X}_\gamma$. The graph is constructed from the pairwise Markov property, but can be interpreted in terms of the global Markov property.

2.3 The Multivariate Normal Distribution and Conditional Independence

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

$$f_{\mathbf{X}}(\mathbf{x}) = (2\pi)^{-\frac{p}{2}} (\det \Sigma)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}, \qquad \mathbf{x} \in \mathbb{R}^p, \tag{6}$$

for some $\boldsymbol{\mu} \in \mathbb{R}^p$ and $\Sigma \in \mathscr{S}_p^+$. Then $\mathbb{E}\mathbf{X} = \boldsymbol{\mu}$ and $\mathrm{Var}(\mathbf{X}) = \Sigma$. 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, $\mathbf{X}$ be partitioned into $\mathbf{X} = (\mathbf{Y}, \mathbf{Z})$. Then we have:

(I) The (marginal) distribution of $\mathbf{Y}$ is $N_q(\boldsymbol{\mu}_\alpha, \Sigma_{\beta,\beta})$.
(II) $\mathbf{Y}$ and $\mathbf{Z}$ are independent (in notation $\mathbf{Y} \perp\!\!\!\perp \mathbf{Z}$) if and only if $\Sigma_{\alpha,\beta} = \mathbf{0}$ (which is equivalent to $K_{\alpha,\beta} = \mathbf{0}$).
(III) The conditional distribution of $\mathbf{Y}$ given $\mathbf{Z} = \mathbf{z}$ is

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

These fundamental properties of the MVN distribution can be proved by directly manipulating the density (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:

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

In general, $\mathbb{E}(\mathbf{Y}|\mathbf{Z})$ and $\mathrm{Var}(\mathbf{Y}|\mathbf{Z})$ are random variables that can be expressed as functions of the conditioning variable $\mathbf{Z}$. Thus (III) tells us that in the MVN model $\mathbb{E}(\mathbf{Y}|\cdot)$ is a *linear* function, whereas $\mathrm{Var}(\mathbf{Y}|\cdot)$ is *constant*. Further, $\mathbb{E}(\mathbf{Y}|\mathbf{Z})$ is the best prediction of $\mathbf{Y}$ from $\mathbf{Z}$, in the sense that $\mathbb{E}\|\mathbf{Y} - h(\mathbf{Z})\|^2$ is uniquely minimized by $h = \hat{\mathbf{Y}}(\cdot)$ among *all* measurable functions h. Here this best prediction coincides with the best linear prediction $\hat{\mathbf{Y}}(\mathbf{Z})$ given in (3). Finally, $\mathrm{Var}(\mathbf{Y}|\mathbf{Z})$ being constant means that the accuracy gain for predicting $\mathbf{Y}$ that we get from knowing $\mathbf{Z}$ is the same no matter what value $\mathbf{Z}$ 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. 2.1, $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2)$ be further partitioned. $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are *conditionally independent given* $\mathbf{Z}$—in writing: $\mathbf{Y}_1 \perp\!\!\!\perp \mathbf{Y}_2|\mathbf{Z}$—if the conditional distribution of $(\mathbf{Y}_1, \mathbf{Y}_2)$ given $\mathbf{Z} = \mathbf{z}$ is for (almost) all $\mathbf{z} \in \mathbb{R}^r$ a product measure with independent margins corresponding to $\mathbf{Y}_1$ and $\mathbf{Y}_2$. If $\mathbf{X}$ possesses a density $f_{\mathbf{X}} = f_{(\mathbf{Y}_1,\mathbf{Y}_2,\mathbf{Z})}$ w.r.t. some σ-finite measure, conditional independence admits the following characterization: $\mathbf{Y}_1 \perp\!\!\!\perp \mathbf{Y}_2|\mathbf{Z}$ 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

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

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

We can construct for any random variable $\mathbf{X}$ in $\mathbb{R}^p$ a *conditional independence graph* (*CIG*) in analogous way as before the partial correlation graph: We put an edge between nodes i and j unless $X_i \perp\!\!\!\perp X_j | \mathbf{X}_{\setminus\{i,j\}}$. Then, for "nice" distributions F—for instance, if F has a continuous, strictly positive density f (w.r.t. some σ-finite measure)—we have in analogy to Theorem 1 a separation property for CIGs: $\mathbf{X}_\alpha \perp\!\!\!\perp \mathbf{X}_\beta | \mathbf{X}_\gamma$ if and only if γ separates α and β in the CIG of $\mathbf{X}$.

Assertions (I) to (III) are the link from conditional independence to the analysis of the second moment characteristics in Sect. 2.1. A direct consequence is:

Proposition 3. *If* $\mathbf{X} = (\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Z}) \sim N_p(\boldsymbol{\mu}, \Sigma)$, $\Sigma \in \mathscr{S}_p^+$, *then*

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

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.

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 $i, j \in V$

$$\{i, j\} \notin E \;\Rightarrow\; \varrho_{i,j\bullet} = 0. \tag{7}$$

If the model consists of all (regular) p-variate normal distributions satisfying (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

$$\mathscr{M}(G) = \left\{ N_p(\boldsymbol{\mu}, \Sigma) \,\middle|\, \boldsymbol{\mu} \in \mathbb{R}^p,\ K = \Sigma^{-1} \in \mathscr{S}_p^+(G) \right\}. \tag{8}$$

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. 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. 3.1), then review the properties of the likelihood ratio test for comparing two nested models (Sect. 3.2) and finally describe some model selection procedures (Sect. 3.3).

3.1 Estimation

Suppose we have i.i.d. observations $\mathbf{X}_1, ..., \mathbf{X}_n$ sampled from the normal distribution $N_p(\boldsymbol{\mu}, \Sigma)$ with $\Sigma \in \mathscr{S}_p^+$. Let furthermore $\mathbb{X}_n = (\mathbf{X}_1^T, ..., \mathbf{X}_n^T)^T$ 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 4. *If $n > p$, the maximum likelihood estimator (MLE) of Σ in the multivariate normal model (with unknown location $\boldsymbol{\mu}$) is*

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

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, cf. Vogel (2009).

Proposition 5. *In the MVN model $\hat{\Sigma}$, $\hat{K}$ and $\hat{P}$ are strongly consistent estimators of Σ, K and P, respectively. Furthermore,*

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

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

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

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 6. *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 textbook 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 7. *In the MVN model with $n > p$ we have*

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

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

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

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

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

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

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. 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 $\mathbf{X}_1, ..., \mathbf{X}_n$ an i.i.d. sample from the model $\mathscr{M}(G)$ given in (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 $p(p-1)/2$. 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.

(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

$$\hat{\Sigma}_G(G) = \hat{\Sigma}(G) \qquad \text{and} \qquad \hat{K}_G = \hat{K}_G(G), \tag{9}$$

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. 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 8. 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 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 $\hat{\Sigma}_G$ is specified by (9), and Theorem 2 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 $T_\gamma : \mathscr{S}_p \to \mathscr{S}_p$:

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

The operator T_γ has the following properties:

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

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

Apparently T_γ preserves the zero pattern of G. That it also preserves positive definiteness and assertion (II) are not as straightforward, but both can be deduced by applying (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

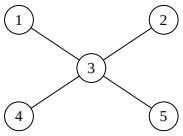
$$K_{n+1} = T_{\gamma_1} T_{\gamma_2} ... T_{\gamma_c} K_n$$

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 (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

As mentioned before, 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

Fig. 2 example graph



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 subsets of V by

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

The δ_k do not have to be distinct. For instance, the graph in Fig. 2 has four cliques and, for any numbering of the cliques, $\delta_i = \{3\}$, $i = 1, 2, 3$.

Proposition 8.

(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)).

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

$$D_n(G_0, G) = -2 \ln \mathrm{LR}(G_0, G) = n \left(\ln \det \hat{\Sigma}_{G_0} - \ln \det \hat{\Sigma}_G \right) \tag{10}$$

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 9 (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 9. *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 $B_i \sim \text{Beta}\left(\frac{n-a_i-2}{2}, \frac{1}{2}\right)$.

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 9. 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 9 this has a Beta$(\frac{n-p}{2}, \frac{1}{2})$ distribution, which is in concordance with Proposition 7 (3).

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. 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 overall probability of selecting an incorrect edge, i.e. the probability that the estimated model is too large, is controlled.

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 was 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.

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 (1987)), 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, and constants of finite sample approximations 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 $\mathbb{F}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{\mathbf{X}_i}$, where $\delta_{\mathbf{x}}$ denotes the Dirac-measure which puts unit mass at the point $\mathbf{x} \in \mathbb{R}^p$. For instance, the sample covariance ma-

trix $\hat{\Sigma}$ is simply the functional 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

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

where $F_{\varepsilon,\mathbf{x}} = (1-\varepsilon)F + \varepsilon\delta_{\mathbf{x}}$. In words, the influence function is the directional derivative of the functional S at the "point" $F \in \mathscr{F}$ in the direction of $\delta_{\mathbf{x}} \in \mathscr{F}$. It describes the *influence* of an infinitesimal contamination at point $\mathbf{x} \in \mathbb{R}^p$ 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 influence $\sup\{\|IF(\mathbf{x}; S, F)\| \,|\, \mathbf{x} \in \mathbb{R}^p\}$, 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 $\hat{S}(\mathbb{X}_n A^T + \mathbf{1}_n \mathbf{b}^T) = A\hat{S}(\mathbb{X}_n)A^T$ for any full rank matrix $A \in \mathbb{R}^{p\times p}$ and vector $\mathbf{b} \in \mathbb{R}^p$. 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 10 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 C1 $\hat{S}(\mathbb{X}_n A^T + \mathbf{1}_n \mathbf{b}^T) = g(AA^T) A \hat{S}(\mathbb{X}_n) A^T$ *for* $\mathbf{b} \in \mathbb{R}^p$, $A \in \mathbb{R}^{p \times p}$ *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 $g = 1/\det(\cdot)$. The following proposition can be found in similar form in Bilodeau and Brenner (1999), p. 212.

Proposition 10. *In the MVN model, i.e.* $\mathbb{X}_n = (\mathbf{X}_1^T, ..., \mathbf{X}_n^T)^T$ *with* $\mathbf{X}_1, ..., \mathbf{X}_n \sim N_p(\boldsymbol{\mu}, \Sigma)$ *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) $\mathrm{Var}(\mathrm{vec}\,\hat{S}) = 2\, b_n\, M_p(\Sigma \otimes \Sigma) + c_n\, \mathrm{vec}\,\Sigma (\mathrm{vec}\,\Sigma)^T$,

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 7 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 variance re-appears in the asymptotic variance, and they fulfill

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

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

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.

Proposition 11. *If* $\hat{S}$ *fulfils C2, the corresponding partial correlation estimator* $\hat{P}^S = -\mathrm{Corr}(\hat{S}^{-1})$ *satisfies*

$$\sqrt{n}\,\mathrm{vec}(\hat{P}^S - P) \xrightarrow{\mathscr{L}} N_{p^2}(\mathbf{0}, 2b\Gamma M_p(K \otimes K)\Gamma^T) \tag{11}$$

where b is the same as in Condition C2 and Γ *is as in Proposition 5.*

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 5. Also, general results for the influence function of pseudo-equivariant estimators can be given, cf. Hampel et al. (1986), Chap. 5.3.

Proposition 12.

(1) At the Gaussian distribution $F = N_p(\boldsymbol{\mu}, \Sigma)$ the influence function of any functional S satisfying Condition C1 has, if it exists, the form

$$IF(\mathbf{x}; S, F) = g(\Sigma)\Big(\alpha(d(\mathbf{x}))(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T - \beta(d(\mathbf{x}))\Sigma\Big), \tag{12}$$

where $d(\mathbf{x}) = \sqrt{(\mathbf{x}-\boldsymbol{\mu})^T K(\mathbf{x}-\boldsymbol{\mu})}$, g is as in Condition C1 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, the influence function of the corresponding partial correlation matrix functional $P^S = -\mathrm{Corr}(S^{-1})$ is

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

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

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 (12) reduces to $IF(\mathbf{x}; \mathrm{Var}, F) = (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T - \Sigma$, which is not only unbounded, but even increases quadratically with $\|\mathbf{x}\|$. 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 $\frac{1}{2} \le s < 1$ is determined such $\det(\hat{\Sigma}^\eta)$ with

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

is minimal. The mean $\boldsymbol{\mu}_{\mathrm{MCD}}$ 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 $(\boldsymbol{\mu}_{\mathrm{MCD}}, \hat{\Sigma}_{\mathrm{MCD}})$ a reweighted scatter estimator $\hat{\Sigma}_{\mathrm{RMCD}}$ is computed from the whole sample:

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

where $w_i = 1$ if $(\mathbf{X}_i - \boldsymbol{\mu}_{\mathrm{MCD}})^T \hat{\Sigma}_{\mathrm{MCD}}^{-1}(\mathbf{X}_i - \boldsymbol{\mu}_{\mathrm{MCD}}) < \chi^2_{p,0.975}$ and zero otherwise. Usually the estimate is multiplied by a consistency factor (corresponding to $1/a$ in Condition C2) 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 as well as the function α and β in Proposition 12 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 (2008)) 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

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

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

where $\xi \geq 0$ is a tuning parameter and $d(\mathbf{x})$ is, as in Proposition 12, the Mahalanobis distance of $\mathbf{x} \in \mathbb{R}^p$ 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 $\xi = 0$.

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, the deviance test statistic D_n in (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. 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 (9) as an optimal way of estimating a covariance matrix with a graphical model and exact normality, it is not quite clear what is a best way of robustly estimating a covariance matrix that obeys a zero pattern in its covariance. Clearly, Theorem 2 suggests to simply solve equations (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 8, or generally by the IPS algorithm, for which convergence has been shown and which comes at no additional computational cost. Becker (2005) has 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 and Proposition 8, 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), Vogel et al. (2008)). 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 11, 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 (10) only by a factor, see Tyler (1983) or Bilodeau and Brenner (1999), Chap. 13, for details. However, as noted in Sect. 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 9 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 (2008) studied a robustified version based on the RMCD.

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.

Acknowledgement. This work has been supported in part by the Collaborative Research Center "Statistical Modelling of Nonlinear Dynamic Processes" (SFB 823) of the German Research Foundation (DFG). The authors are grateful to Alexander Dürre for his assistance in preparing the figures.

References

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)

Bilodeau, M., Brenner, D.: Theory of multivariate statistics. Springer Texts in Statistics. New York, NY: Springer (1999)

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)

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)

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., Haesbroeck, G.: Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. J. Multivariate Anal. **71**(2), 161–190 (1999)

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)

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)
Edwards, D., Havránek, T.: A fast model selection procedure for large families of models. J. Am. Stat. Assoc. **82**, 205–213 (1987)
Eriksen, P.S.: Tests in covariance selection models. Scand. J. Stat. **23**(3), 275–284 (1996)
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)
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. **in press** (2008)
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)
Huber, P.J., Ronchetti, E.M.: Robust statistics. 2nd edn. Wiley Series in Probability and Statistics. Hoboken, NJ: Wiley (2009)
Kent, J.T., Tyler, D.E.: Constrained M-estimation for multivariate location and scatter. Ann. Stat. **24**(3), 1346–1370 (1996)
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)
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)
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)

Miyamura, M., Kano, Y.: Robust Gaussian graphical modeling. J. Multivariate. Anal. **97**(7), 1525–1550 (2006)

Paindaveine, D.: A canonical definition of shape. Stat. Probab. Lett. **78**(14), 2240–2247 (2008)

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., 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)

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)

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 (1987)

Visuri, S., Koivunen, V., Oja, H.: Sign and rank covariance matrices. J. Stat. Plann. Inference **91**(2), 557–575 (2000)

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., 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)

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)