ELLIPTICAL GRAPHICAL MODELLING

DANIEL VOGEL AND ROLAND FRIED

Abstract. We propose elliptical graphical models based on conditional uncorrelatedness as a generalization of Gaussian graphical models by letting the population distribution be elliptical instead of normal, allowing the fitting of data with arbitrarily heavy tails. We 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 chi square 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. The latter is likelihood-based and known to be very sensitive to model misspecification and outlying observations.

1. Introduction and notation

The statistical theory of undirected graphical models for continuous variables is usually based on the assumption of multivariate normality. In practice, data may deviate from the normal model in various ways. Outliers and heavy tails pose a problem of particular gravity: they frequently occur, and the normal likelihood methods, such as the sample covariance matrix, are very susceptible to them. 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 case where we have only continuous variables and only undirected edges. Joint multivariate normality is often assumed in this situation, and the statistical methodology is called Gaussian graphical modelling. We propose the class of elliptical distributions as a more general model and call our approach elliptical graphical modelling.

The lack of robustness of Gaussian graphical modelling has been noted by several authors. Four proposals of robust approaches to Gaussian graphical modelling are known to us: Becker (2005) and Gottard and Pacillo (2010) suggest replacing the sample covariance matrix by the reweighted minimum covariance determinant estimator. Miyamura and Kano (2006) propose an alternative M-type estimation, and Finegold & Drton (arXiv:1009.3669) consider robustified versions of the graphical lasso by Friedman et al. (2008).

This article delivers a systematic treatment of the plug-in approach used in the first two references. We show that the sample covariance matrix may be replaced by 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 necessary for the data situation at hand. In order to reduce the search space, graphical modelling is often restricted to decomposable graphical models, which allow better interpretability, cf. Whittaker (1990, Chapter 12), but are also easier to handle mathematically. For conciseness we restrict our derivations to decomposable models.

We close this section by introducing some mathematical notation. Depending on the context, the symbol ~ means distributed as or asymptotically equivalent. Finite index sets are denoted by 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 \( \alpha \) and all columns that are not in \( \beta \). Similarly, the \( p \times p \) matrix \( (S_{\alpha,\beta})^{(p)} \) is obtained from \( S \) by putting all rows not
in \(\alpha\) and all columns not in \(\beta\) to zero. We 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 \(\alpha\) and \(\beta\). Of course, the latter is not well defined without the former, but this allows us to write \((S^{-1}_{\alpha,\beta})_{(p)}\), for example. Subscripts have priority over superscripts, \(S^{-1}_{\alpha,\beta}\) stands for \((S_{\alpha,\beta})^{-1}\). Let \(\mathcal{S}_p\) and \(\mathcal{S}_p^+\) 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 \(e_1, \ldots, e_p\) be the unit vectors in \(\mathbb{R}^p\) and \(I_p\) the \(p\)-vector consisting only of ones. Define the matrices:

\[
J_p = \sum_{i=1}^{p} e_i e_i^T, \quad K_p = \sum_{i=1}^{p} \sum_{j=1}^{p} e_i e_j^T e_j e_i^T, \quad M_p = \frac{1}{2} (I_{p^2} + K_p),
\]

where \(I_{p^2}\) denotes the \(p^2 \times p^2\) identity matrix; \(K_p\) is also called the commutation matrix. Finally, 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. Elliptical graphical models

We introduce elliptical graphical models in analogy to Gaussian graphical models. For details on the latter see Whittaker (1990), Cox and Wermuth (1996), Lauritzen (1996) or Edwards (2000). Consider the class \(\mathcal{E}_p\) 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 density \(f\) of the form

\[
(1) \quad f(x) = \det(S)^{-1/2} g\{(x - \mu)^T S^{-1}(x - \mu)\}
\]

for some \(\mu \in \mathbb{R}^p\) and symmetric, positive definite \(p \times p\) matrix \(S\). We call \(S\) the shape matrix of \(F\), and denote the class of all continuous elliptical distributions on \(\mathbb{R}^p\) with the parameters \(\mu\) and \(S\) by \(\mathcal{E}_p(\mu, S)\). A continuous distribution on \(\mathbb{R}^p\) is called spherical if \(S\) is proportional to the identity matrix. The shape matrix \(S\) is unique only up to scale, that is, \(\mathcal{E}_p(\mu, S) = \mathcal{E}_p(\mu, cS)\) for any \(c > 0\). Several forms of standardization have been suggested in the literature. Paindaveine (2008) argues for \(\det(S) = 1\). For our considerations the standardization of \(S\) is irrelevant, and 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 (1) for a suitable function \(g\) a shape matrix of \(F\). We likewise view its inverse \(K = S^{-1}\), which we call a pseudo concentration matrix of \(F\). Furthermore let

\[
h : \mathcal{S}_p^+ \rightarrow \mathcal{S} : A \mapsto - (A^{-1})_D^{1/2} A^{-1} (A^{-1})_D^{1/2}
\]

and \(P = h(S)\). The function \(h\) is invariant to scale changes, i.e., \(P\) is a uniquely defined parameter of \(F \in \mathcal{E}_p(\mu, S)\). The diagonal elements of \(P\) are equal to \(-1\). If the second-order moments of \(X \sim F \in \mathcal{E}_p(\mu, S)\) exist, then \(\Sigma = \text{var}(X)\) is proportional to \(S\). Consequently, the element \(p_{i,j}\) of \(P\) at position \((i,j)\) is the partial correlation of \(X_i\) and \(X_j\) given the other components of \(X\) (Whittaker 1990 Chapter 5). We call \(P\) the generalized partial correlation matrix of \(F\) and refer to it as partial correlation matrix for brevity.

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 \(X_1, \ldots, X_p\) are the vertices, and an edge is drawn between \(X_i\) and \(X_j\) if and only if \(p_{i,j} \neq 0 (i, j = 1, \ldots, p; i \neq j)\). The graph \(G\) thus obtained is called the generalized partial correlation graph of \(F\). Formally we set \(V = \{1, \ldots, p\}\) and write the elements of \(E\) as unordered pairs \(\{i, j\}\) (\(i, j = 1, \ldots, p; i \neq j\)). The global and the local Markov property with respect to any generalized partial correlation graph \(G\) are equivalent for any \(F \in \mathcal{E}_p\) without any moment assumptions (Vogel and Fried 2010).

Let \(\mathcal{S}^+_p(G)\) be the subset of \(\mathcal{S}_p^+\) consisting of all positive definite matrices with zero entries at the positions specified by the graph \(G = (V,E)\), i.e.,

\[
K \in \mathcal{S}^+_p(G) \iff K \in \mathcal{S}^+, \quad k_{i,j} = 0 \quad (i \neq j, \{i, j\} \notin E),
\]
and define

$$\mathcal{E}_p(G) = \{ F \in \mathcal{E}_p(\mu, K^{-1}) \mid \mu \in \mathbb{R}^p, \ K \in \mathcal{S}_p^+(G) \}$$

to be the elliptical graphical model induced by $G$. We call the model $\mathcal{E}_p(G)$ decomposable if $G$ is decomposable, i.e., if it possesses no chordless cycle of length greater than three. For alternative characterizations and properties of decomposable graphs see e.g. [Lauritzen (1996) Chapter 2].

In the remainder of this section we discuss the interpretation of an absent edge in the partial correlation graph of $F \in \mathcal{E}_p$. 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 $X_3, \ldots, X_p$, i.e., $p_{12} = 0$, is to be understood as linear independence of $X_1$ and $X_2$ after the common linear effects of $X_3, \ldots, X_p$ have been removed. A relation of similar type is conditional independence: roughly, $X_1$ and $X_2$ are conditionally independent given $X_3, \ldots, X_p$, if the conditional distribution of $(X_1, X_2)$ is a product measure for almost all values of the conditioning variable $(X_3, \ldots, X_p)$. In comparison to partial correlation we understand conditional independence as complete independence of $X_1$ and $X_2$ after the removal of all common effects of $X_3, \ldots, X_p$.

Another related term is conditional uncorrelatedness: the conditional distribution of $(X_1, X_2)$ given $(X_3, \ldots, X_p)$ has correlation zero for almost all values of $(X_3, \ldots, X_p)$. There is an important qualitative difference between partial and conditional correlation: the former is a real value, the latter a function of the conditioning variable. All marginal and conditional distributions of elliptical distributions are again elliptical [Fang and Zhang (1990) Section 2.6]. Hence partial uncorrelatedness implies conditional uncorrelatedness [Baba et al. (2004)], and $p_{12} = 0$ means linear independence of $X_1$ and $X_2$ after all common effects of $X_3, \ldots, X_p$ have been removed.

However, the only spherical distributions with independent margins are Gaussian distributions, cf. [Bilodeau and Brenner (1999) p. 51]. Thus contrary to Gaussian graphical models a missing edge in the partial correlation graph 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 illusory. 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, in our case linear dependencies, are modelled parametrically, whereas other aspects remain unspecified.

Of course, non-normal data need not be elliptical. Any relevant data feature, such as non-linearities, anomalous values, etc., is of potential interest and should be analysed. If the data, say, contains strong quadratic interactions, models that incorporate them should be used, as it is described e.g. in [Cox and Wermuth (1996) Section 2.10]. We address primarily the situation where the essential structure of the data is captured by an ellipse, and the linear interactions are the prominent ones. In any case, a robust analysis of the linear effects, as proposed here, is a suitable starting point of any subsequent tests for potential non-linear effects.

3. 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 $\hat{S}_n$.

Let $X_1, \ldots, X_n$ be independent and identically distributed random vectors sampled from an elliptical distribution $F \in \mathcal{E}_p(\mu, S)$. Depending on the context, $X_k$ may denote the $k$th $p$-dimensional observation or the $k$th component of the vector $X$. Furthermore let $x_n = (X_1, \ldots, X_n)^T$ be the $n \times p$ data matrix and $\hat{S}_n = \hat{S}_n(x_n)$ be a scatter estimator. The symbol $\hat{S}_n$ may have two meanings: a function on the sample space, or as abbreviation for $\hat{S}_n(x_n)$, a random variable. We use the term
scatter estimator for any symmetric matrix-valued estimator that gives some information about
the spread of the data. We call $\hat{S}_n$ affine pseudo-equivariant, if it satisfies
\begin{equation}
(2) \quad \hat{S}_n(X_n A^T + 1_n b^T) \propto A \hat{S}_n(X_n) A^T
\end{equation}
for all $b \in \mathbb{R}^p$ and full rank $A \in \mathbb{R}^{p \times p}$. This is a generalization of the strict affine equivariance for
scatter estimators, which is obtained if $A^T \propto A$ is satisfied with equality. 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 are given
in Section 6.

[Tyler (1982)] shows that, if a strictly affine equivariant scatter estimator is evaluated at an elliptical
distribution, its first two moments, if existent, have a common structure. If the proportionality
factor in (2) is not random, the same holds true for pseudo-equivariant scatter estimators. The
following condition is therefore natural for affine pseudo-equivariant estimators at elliptical distributions $F$, and many shape estimators have been shown to satisfy it under suitable additional conditions on $F$, see also the examples in Section 6.

**Assumption 3.1.** The estimator $\hat{S}_n$ converges in probability to $\eta S$ for some $\eta \geq 0$, and there exist
$\sigma_1 \geq 0$ and $\sigma_2 \geq 0$ such that
\[n^{1/2} \text{vec}(\hat{S}_n - \eta S) \rightarrow N_{p^2}(0, \eta^2 W_S(\sigma_1, \sigma_2))\]
in distribution as $n \rightarrow \infty$, where $W_S(\sigma_1, \sigma_2) = 2 \sigma_1 M_p(S \otimes S) + \sigma_2 \text{vec}(S \text{vec}(S))^T$. The scalars $\sigma_1$ and $\sigma_2$ depend on the estimator $\hat{S}_n$, the dimension $p$ and the function $g$, but are constant with respect to the shape $S$.

We have the following implication for the derived estimators $\hat{K}_n = \hat{S}_n^{-1}$ and $\hat{P}_n = h(\hat{S}_n)$.

**Proposition 3.2.** If $\hat{S}_n$ satisfies Assumption 3.1 then with $K = S^{-1}$,
\[\begin{align*}
(i) \quad & n^{1/2} \text{vec}(\hat{K}_n - \eta^{-1} K) \rightarrow N_{p^2}(0, \eta^{-2} W_K(\sigma_1, \sigma_2)) \\
& \text{in distribution as } n \rightarrow \infty, \text{ where } W_K(\sigma_1, \sigma_2) = 2 \sigma_1 M_p(K \otimes K) + \sigma_2 \text{vec}(K \text{vec}(K))^T, \text{ and} \\
(ii) \quad & n^{1/2} \text{vec}(\hat{P}_n - P) \rightarrow N_{p^2}(0, 2 \sigma_1 \Gamma(S) M_p(K \otimes K) \Gamma(S)^T) \\
& \text{in distribution as } n \rightarrow \infty \text{ with } \Gamma(S) = (K_D^{-1/2} \otimes K_D^{-1/2}) + M_p(P \otimes K_D^{-1}) I_p.
\end{align*}\]

An important aspect of Proposition 3.2 is that under ellipticity the asymptotic covariance matrices of partial correlation estimators $\hat{P}_n$ derived from affine equivariant shape estimators $\hat{S}_n$ are proportional to each other.

### 4. Constrained estimation

In this section we treat the estimation of $P$ under a given graphical model $\mathcal{E}_p(G)$ specified
by the graph $G = (V, E)$, i.e., estimating $P$ with zero-entries. A crude approach is to put the
concerning elements in an unconstrained estimate $\hat{P}_n$ to zero, but this generally destroys the positive
definiteness of the estimate. We define the function $h_G : \mathcal{S}_p^+ \rightarrow \mathcal{S}_p^+(G) : A \mapsto A_G$ by
\begin{equation}
(3) \quad \begin{cases}
(A_G)_{i,j} = a_{i,j} & (\{i, j\} \in E \lor i = j), \\
(A_G)_{i,j} = 0 & (\{i, j\} \notin E, \ i \neq j),
\end{cases}
\end{equation}
where $a_{i,j}$ are the elements of $A$. A unique and positive definite solution $A_G$ of (3) 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 mainly deal with asymptotics, and shape estimators $\hat{S}_n$ are
usually almost surely positive definite at continuous distributions for sufficiently large $n$, we assume
positive definiteness for simplicity’s sake.

Let $G = (V, E)$ be a decomposable graph with cliques $\gamma_1, \ldots, \gamma_c$ ($c \geq 1$), and define the sequence
$\delta_1, \ldots, \delta_{c-1}$ of successive intersections by
\[
\delta_k = (\gamma_1 \cup \cdots \cup \gamma_k) \cap \gamma_{k+1} \quad (k = 1, \ldots, c - 1).
\]
We assume that the ordering \( \gamma_1, \ldots, \gamma_k \) is such that the cliques form a perfect sequence, i.e., for all \( k = 1, \ldots, c - 1 \) there is a \( j \in \{1, \ldots, k\} \) such that \( \delta_k \subset \gamma_j \). It is always possible to arrange the cliques of a decomposable graph in a perfect sequence \cite{Lauritzen1996} Prop. 2.17. For notational convenience we let

\[
\alpha_k = \begin{cases} \gamma_k & (k = 1, \ldots, c), \\ \delta_{k-c} & (k = c + 1, \ldots, 2c - 1), \end{cases} \quad \zeta_k = \begin{cases} 1 & (k = 1, \ldots, c), \\ -1 & (k = c + 1, \ldots, 2c - 1). \end{cases}
\]

Then \( h_G(A) \) allows the following explicit formulation for decomposable \( G \),

\[
h_G(A) = A_G = \left\{ \sum_{k=1}^{2c-1} \zeta_k \left( A_{\alpha_k,\alpha_k} \right) \left( p \right) \right\}^{-1} (A \in \mathcal{D}^\kappa).
\]

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

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

we have the following result about the asymptotic distribution. It is not assumed that the true shape \( S \) fits the model \( G \).

**Proposition 4.1.** If \( \hat{S}_n \) fulfils Assumption 3.1 and \( G \) is decomposable, then

(i) \( n^{-1/2} \text{vec}(\hat{K}_G - \eta^{-1}K_G) \to N_0\{0, \eta^{-2}W_{\Omega^G}(\sigma_1, \sigma_2)\} \) in distribution as \( n \to \infty \) with \( W_{\Omega^G}(\sigma_1, \sigma_2) = 2\sigma_1 M_p G^2(\Omega G(S)) (S \otimes S) \Omega G(S) + \sigma_2 \text{vec} K^G(\text{vec} K_G)^T \),

(ii) \( n^{-1/2} \text{vec}(\hat{S}_G(\sigma_1, \sigma_2) - \eta S_G) \to N_0\{0, \eta^{-2}W_{\Omega^G}(\sigma_1, \sigma_2)\} \) in distribution as \( n \to \infty \) with \( W_{\Omega^G}(\sigma_1, \sigma_2) = 2\sigma_1 M_p (S_G \otimes S_G) \Omega G(S) (S \otimes S) \Omega G(S) (S_G \otimes S_G) + \sigma_2 \text{vec} S_G(\text{vec} S_G)^T \),

(iii) \( n^{-1/2} \text{vec}(\hat{P}_G - P_G) \to N_0\{0, W_{\Omega^G}(\sigma_1, \sigma_2)\} \) in distribution as \( n \to \infty \), where \( W_{\Omega^G}(\sigma_1, \sigma_2) = 2\sigma_1 \text{vec} (\sum_{k=1}^{2c-1} \zeta_k (S_{\alpha_k,\alpha_k})^{(p)} (S_{\alpha_k,\alpha_k})^{(p)}) \in \mathbb{R}^{p^2 \times p^2} \).

If the true shape \( S \) satisfies the graph \( G \), the expressions for the asymptotic variances simplify.

**Corollary 4.2.** If \( \hat{S}_n \) satisfies Assumption 3.1 with \( S^{-1} \in \mathcal{D}^\kappa(G) \) for a decomposable graph \( G \), then the assertions of Proposition 4.1 are true with

(i) \( W_{\Omega^G}(\sigma_1, \sigma_2) = 2\sigma_1 M_p G^2(\Omega G(S)) + \sigma_2 \text{vec} K^G(\text{vec} K_G)^T \),

(ii) \( W_{\Omega^G}(\sigma_1, \sigma_2) = 2\sigma_1 M_p (S_G \otimes S_G) \Omega G(S) (S \otimes S) + \sigma_2 \text{vec} S_G(\text{vec} S_G)^T \) and

(iii) \( W_{\Omega^G}(\sigma_1, \sigma_2) = 2\sigma_1 \text{vec} (\sum_{k=1}^{2c-1} \zeta_k (S_{\alpha_k,\alpha_k})^{(p)} (S_{\alpha_k,\alpha_k})^{(p)}) \in \mathbb{R}^{p^2 \times p^2} \).

5. Testing

An essential tool of most model selection procedures is to test if a model under consideration fits the data and to compare the fit of two nested models. On the set \( \Pi_p = \{ (i, j) \mid i, j = 1, \ldots, p \} \) of the positions of a \( p \times p \) matrix we declare a strict ordering \( \prec_p \) by

\[
(i, j) \prec_p (k, l) \iff (j - 1)p + i < (l - 1)p + k, \quad (i, j, k, l = 1, \ldots, p).
\]

For any subset \( Z = \{ z_1, \ldots, z_q \} \subset \Pi_p \), where \( z_k = (i_k, j_k) \) \( (k = 1, \ldots, q) \) and \( z_1 \prec_p \cdots \prec_p z_q \), 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 \( Q_Z \text{vec}(A) \) picks the elements of \( A \) at positions specified by \( Z \) in the order they appear in \( \text{vec}(A) \). For a graph \( G = (V, E) \) with \( V = \{1, \ldots, p\} \) let

\[
D(G) = \{(i, j) \mid i, j = 1, \ldots, p; \ (i, j) \notin E; \ j < i\},
\]

i.e., the set \( D(G) \) gathers all sub-diagonal zero-positions that \( G \) enforces on a concentration matrix. Thus \( F \in \mathcal{E}_p(G) \) is equivalent to \( Q_{D(G)} \text{vec} K = 0 \).
We require, for example, a sum of suitably weighted squared entries of $Q$ function theorem. Then if all entries at positions in $D$ of an estimate $\hat{P}_{G_i}$ of $P$ under $G_1$ are close to zero. For example, a sum of suitably weighted squared entries of $\hat{P}_{G_i}$, 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 $S$ the matrix $R_{G_i}(S)$ has rank $(p-1)p/2 - q_1$, which can be deduced from the inverse function theorem. Then $Q_{0,1}R_{G}(S)Q_{0,1}^T$ is of full rank, and the probability that the Wald-type test statistic

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

exists tends to 1 as $n \to \infty$. Proposition 5.1 describes the asymptotic behaviour of $\hat{T}_n(G_0, G_1)$ under the null hypothesis that $G_0$ is true, part [i], and under a local alternative, part [ii].

**Proposition 5.1.** Let $G_0$, $G_1$ be as above and $X_1, \ldots, X_n$ independent and identically distributed random variables with $X_1 \sim F \in \mathcal{F}_p(\mu, S) \subset \mathcal{F}_p(G_0)$. Let $\hat{S}_n$ be an affine pseudo-equivariant scatter estimator such that $\hat{S}_n(\mathbb{X}_n)$ satisfies Assumption 3.1.

(i) Then $\hat{T}_n(G_0, G_1) \to \sigma_1^2 \chi_{q_0,1}^2$ in distribution as $n \to \infty$.

(ii) For $m \in \mathbb{N}$ let $\mathbb{X}_n \sim (X_1^{(m)}, \ldots, X_n^{(m)})^T$ be distributed as $\mathbb{X}_nS^{-1/2}S_m^{1/2}$, thus $X_1^{(m)} \sim \mathcal{F}_p(\mu, S_m)$, where the sequence $S_m$ is such that $B = \lim_{m \to \infty} m^{1/2}(S_m - S)$ exists. If, for each $n \in \mathbb{N}$, $\hat{S}_n$ is applied to $\mathbb{X}_n$, then, as $n \to \infty$,

$$\hat{T}_n(G_0, G_1) \to \sigma_1^2 \chi_{q_0,1}^2 \{ \sigma_1^{-1}\delta(B, S) \}$$

in distribution, where

$$\delta(B, S) = \frac{1}{2} v^T Q_{0,1}^T \{ Q_{0,1}R_{G}(S)Q_{0,1}^T \}^{-1} Q_{0,1}v, \quad v = \Gamma(S)\Omega_{G_1}(S) \text{vec}B.$$

We have some remarks.

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

(b) We require $\hat{S}_n$ to be affine pseudo-equivariant to ensure that the convergence of $n^{1/2}\{ \hat{S}_n(\mathbb{X}_n) - \eta S_m \}$ for $n \to \infty$ is uniform in $m$.

(c) In part [ii] of Proposition 5.1, we do not require the sequence of alternatives to lie in the model $G_1$, i.e., that $S_{m}^{-1} \in \mathcal{F}_p^-(G_1)$, as it is not necessary for the convergence [ii] to hold. When choosing a model by forward selection one usually compares two wrong models, so it is of interest to know the behaviour of $\hat{T}_n(G_0, G_1)$ also if $G_1$ is not true.

A difficulty with the test in Proposition 5.1 is the complicated formulation of $\hat{T}_n(G_0, G_1)$. The classical test in Gaussian graphical models is the deviance test. The next proposition gives the analogue for elliptical graphical modelling. It treats parts [i] and [ii] of the previous proposition simultaneously.

**Proposition 5.2.** Let $G_0$, $G_1$ be as above and $\hat{S}_n$ a sequence of almost surely positive definite random $p \times p$ matrices, for which $n^{1/2}(\hat{S}_n - S)$ converges in distribution to a non-degenerate limit for some $S \in \mathcal{F}_p^-$ with $S^{-1} \in \mathcal{F}_p^+(G_0)$. Then, as $n \to \infty$,

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

We require $\hat{S}_n$ to be affine pseudo-equivariant to ensure that the convergence of $n^{1/2}\{ \hat{S}_n(\mathbb{X}_n) - \eta S_m \}$ for $n \to \infty$ is uniform in $m$. We have some remarks.

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

(b) We require $\hat{S}_n$ to be affine pseudo-equivariant to ensure that the convergence of $n^{1/2}\{ \hat{S}_n(\mathbb{X}_n) - \eta S_m \}$ for $n \to \infty$ is uniform in $m$.

(c) In part [ii] of Proposition 5.1, we do not require the sequence of alternatives to lie in the model $G_1$, i.e., that $S_{m}^{-1} \in \mathcal{F}_p^-(G_1)$, as it is not necessary for the convergence [ii] to hold. When choosing a model by forward selection one usually compares two wrong models, so it is of interest to know the behaviour of $\hat{T}_n(G_0, G_1)$ also if $G_1$ is not true.

A difficulty with the test in Proposition 5.1 is the complicated formulation of $\hat{T}_n(G_0, G_1)$. The classical test in Gaussian graphical models is the deviance test. The next proposition gives the analogue for elliptical graphical modelling. It treats parts [i] and [ii] of the previous proposition simultaneously.

**Proposition 5.2.** Let $G_0$, $G_1$ be as above and $\hat{S}_n$ a sequence of almost surely positive definite random $p \times p$ matrices, for which $n^{1/2}(\hat{S}_n - S)$ converges in distribution to a non-degenerate limit for some $S \in \mathcal{F}_p^-$ with $S^{-1} \in \mathcal{F}_p^+(G_0)$. Then, as $n \to \infty$,

$$\hat{D}_n(G_0, G_1) = n \left\{ \log \det h_{G_0}(\hat{S}_n) - \log \det h_{G_1}(\hat{S}_n) \right\} \sim \hat{T}_n(G_0, G_1).$$
If the larger model \( G_1 \) is the saturated model, then Proposition 5.2 is a corollary of Theorem 2 in Tyler (1983). We extend Tyler’s result to two nested models.

**Corollary 5.3.** Both assertions (i) and (ii) of Proposition 5.1 remain true, if \( \hat{T}_n(G_0, G_1) \) is replaced by \( \hat{D}_n(G_0, G_1) \).

### 6. Examples

There are many affine equivariant, robust estimators, see, for example, Zuo (2006) or Maronna et al. (2006). The comparison of asymptotic properties of such estimators in the elliptical model reduces to a comparison of the respective values of the scalars \( \sigma_1 \) and \( \sigma_2 \). Of course, the sample covariance matrix is affine equivariant. The following can be found in Tyler (1982).

**Proposition 6.1.** If \( X_1, \ldots, X_n \) are independent and identically distributed with distribution \( F \in \mathcal{E}_p(\mu, S) \) and \( E||X_1 - \mu||^4 < \infty \), then \( \hat{\Sigma}_n = \hat{\Sigma}(X_n) \) fulfills Assumption 3.1 with \( \sigma_1 = 1 + \kappa/3 \) and \( \sigma_2 = \kappa/3 \), where \( \kappa \) is the excess kurtosis of any component of \( X_1 \).

Proposition 6.1 indicates the inappropriateness of the sample covariance matrix for heavy-tailed distributions: its asymptotic distribution depends on the kurtosis, which is large at heavy-tailed distributions, rendering the estimator inefficient. An alternative is Tyler’s M-estimator, which is defined as the solution \( \hat{\Sigma} \) of

\[
\frac{p}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)(X_i - \bar{X}_n)^T = \hat{\Sigma}
\]

that satisfies \( \det \hat{\Sigma} = 1 \). Existence, uniqueness and asymptotic properties are treated in Tyler (1987), where the following result is proven.

**Proposition 6.2.** If \( X_1, \ldots, X_n \) are independent and identically distributed with distribution \( F \in \mathcal{E}_p(\mu, S) \), furthermore \( E||X_1 - \mu||^2 < \infty \) and \( E||X_1 - \mu||^{-3/2} < \infty \), then \( \hat{\Sigma} \) fulfills Assumption 3.1 with \( \sigma_1 = 1 + 2/p \) and \( \sigma_2 = -2(1 + 2/p)/p \).

We have the following remarks.

(a) In Proposition 6.1 the scalars \( \sigma_1 \) and \( \sigma_2 \) are constant, irrespective of the function \( g \), i.e., the Tyler matrix is asymptotically distribution-free within the elliptical model. Hence, when carrying out any of the tests from Section 5, \( \sigma_1 \) does not need to be estimated.

(b) Tyler’s matrix can cope with arbitrarily heavy tails. The assumption of finite second moments is only required for location estimation by the mean. It may be replaced by any root-\( n \)-consistent location estimator, for instance the Hettmansperger–Randles (2002) median. The inverse moment condition \( E||X_1 - \mu||^{-3/2} < \infty \) is fairly mild: for \( p \geq 2 \) it is fulfilled if \( g \) has no singularity at 0.

(c) The estimator \( \hat{\Sigma} \) is affine pseudo-equivariant and gives information only about the shape but none about the scale. Other such estimators are Oja sign and rank covariance matrices (Ollila et al., 2003, 2004).

The popular reweighted minimum covariance determinant estimator (Rousseeuw and Leroy, 1987; Croux and Haesbroeck, 1999) is highly robust and affine equivariant and has previously been proposed in the context of graphical modelling (Becker, 2005; Gottard and Pacillo, 2010). It is defined as follows. A subset \( \tau \subset \{1, \ldots, n\} \) of size \( h = |\tau| \), where \( 1/2 \leq t < 1 \) is fixed, is determined such that \( \det(\hat{\Sigma}_{\tau}) \) is minimal with

\[
\hat{\Sigma}_{\tau} = \frac{1}{h} \sum_{i \in \tau} (X_i - \bar{X}_\tau)(X_i - \bar{X}_\tau)^T, \quad \bar{X}_\tau = \frac{1}{h} \sum_{i \in \tau} X_i.
\]

The mean \( \hat{\mu}_{\text{MCD}} \) and covariance matrix \( \hat{\Sigma}_{\text{MCD}} \) computed from this minimizing subsample are called the raw minimum covariance determinant location and scatter estimates. The scatter part is scaled...
to achieve consistency for the covariance at the Gaussian distribution. Based on the raw estimates 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, and $\alpha$ is a small rejection probability, e.g. $\alpha = 0.05$. The reweighted covariance estimate is again scaled, but since this is not necessary for our applications we omit the details.

7. Numerical example

We present the results of a simulation study comparing several estimators. We repeatedly sample 100 independent observations of a 5-dimensional distribution. We use the same shape matrix throughout, with equal diagonal elements and the partial correlation structure represented by the graph in Figure 1. We let the tail behaviour vary, using the normal distribution and several members of the $t_{\nu,p}$ family to generate heavier tails (Bilodeau and Brenner [1999], p. 207). The index $\nu$ denotes the degrees of freedom. The moments of $t_{\nu,p}$ are finite up to order $\nu - 1$. We may talk of a fixed shape of the $t_{\nu,p}$ distribution, since $g$ is specified. For $\nu \geq 3$, its covariance matrix is $\nu(\nu - 2)^{-1} S$, and, for $\nu \geq 5$, the excess kurtosis of each component is $6/(\nu - 4)$. Propositions 6.1 and 6.2 imply 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 we generate 2000 samples, compute the estimates described in Section 6 and, based on each estimate, select a model.

We use a simple one-step model selection procedure, that allows us to concentrate on the effects of the different estimators. For each pair $\{i, j\}$ we test 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. In our simulations the Wald-type test statistic $\hat{T}_n$ and the deviance test statistic $\hat{D}_n$ showed a very similar behaviour. Tables 1 and 2 report the results of the deviance test.

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, whether an existing edge was rejected or an absent edge was included. Although less suited as a performance criterion it is also of interest to know, how often the true model is found. Any model selection procedure that is based on testing for zero parameters aims at controlling the probability of correctly specifying the non-edges. We may also look at how often a single non-edge is correctly specified. This should be true in about 95% of the cases, since a sample size of 100 seems large enough to expect some validity of the asymptotics in this setting.

In Table 1 we compare the sample covariance matrix $\hat{\Sigma}_n$ to Tyler’s estimator $\hat{V}_n$ with the Hettmansperger–Randles median as location estimator. The benchmark is traditional graphical modelling, i.e., the performance of $\hat{\Sigma}_n$ at the normal distribution. The classical deviance test deteriorates, if we move away from normality. We assume only ellipticity of the distribution and hence adjust the $\hat{\Sigma}_n$-based test statistic by an estimate of $\sigma_1$, which is here the average of the sample kurtoses of all component divided by 3, cf. Proposition 6.1. This repairs the test, to some extent even in the case of the $t_3$-distribution, but does not necessarily give a better model selection. The estimator $\hat{\Sigma}_n$ is inefficient under heavy tails, resulting in a test with low power. As for

1 2
-0.39 0.75
2
3
0.14
0.75
0.22
0.76

Figure 1. Example model, edge labels indicate partial correlations
Table 1. One-step model selection based on $\hat{\Sigma}$ or $\hat{V}$

| distribution | estimator | mean edge difference | % true model found | % non-edges correctly found | % $\hat{\Sigma} \neq \hat{\xi}$ correctly found |
|--------------|-----------|----------------------|--------------------|-----------------------------|------------------------------------------------|
| normal       | $\Sigma$  | 1.40                 | 21                 | 79                          | 95                                            |
|              | $\Sigma^*$| 1.41                 | 20                 | 77                          | 94                                            |
|              | $\hat{V}$ | 1.65                 | 14                 | 78                          | 94                                            |
| $t_{25}$     | $\Sigma$  | 1.44                 | 20                 | 75                          | 93                                            |
|              | $\Sigma^*$| 1.44                 | 19                 | 78                          | 94                                            |
|              | $\hat{V}$ | 1.64                 | 14                 | 78                          | 94                                            |
| $t_{12}$     | $\Sigma$  | 1.51                 | 20                 | 71                          | 92                                            |
|              | $\Sigma^*$| 1.51                 | 18                 | 79                          | 94                                            |
|              | $\hat{V}$ | 1.66                 | 13                 | 79                          | 94                                            |
| $t_{8}$      | $\Sigma$  | 1.65                 | 17                 | 64                          | 89                                            |
|              | $\Sigma^*$| 1.65                 | 15                 | 76                          | 93                                            |
|              | $\hat{V}$ | 1.62                 | 13                 | 79                          | 94                                            |
| $t_{5}$      | $\Sigma$  | 1.90                 | 14                 | 51                          | 84                                            |
|              | $\Sigma^*$| 1.87                 | 10                 | 74                          | 93                                            |
|              | $\hat{V}$ | 1.63                 | 14                 | 78                          | 94                                            |
| $t_{3}$      | $\Sigma$  | 2.49                 | 8                  | 29                          | 72                                            |
|              | $\Sigma^*$| 2.28                 | 7                  | 71                          | 91                                            |
|              | $\hat{V}$ | 1.65                 | 14                 | 78                          | 95                                            |

* test statistic adjusted by estimated kurtosis

Tyler’s estimator, we recognize the asymptotic properties: the $\chi^2$-quantile fits, it outperforms $\hat{\Sigma}_n$ at $t_\nu$-distributions with $\nu < 9$, and it is distribution-free within the elliptical model.

In Table 2 we examine if the same robustness against heavy-tailedness may be achieved by equally simple means using other robust estimators and, in particular, how the previous proposals of robust Gaussian graphical modelling, the reweighted minimum covariance determinant estimator and the Miyamura–Kano estimator, perform in this situation. 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, this is in principle a different situation, and it is not surprising that both estimators do not meet the performance of Tyler’s estimator at heavy-tailed distributions. Also, we did not estimate $\sigma_1$ from the data, but used its value for the normal distribution. For the reweighted minimum covariance determinant the values can be found in Croux and Haesbroeck [1999]. But even in the Gaussian case, when $\sigma_1$ is chosen asymptotically correct, the asymptotic $\chi^2$-distribution does not seem to provide a sensible approximation. This small-sample inefficiency of the reweighted minimum covariance determinant estimator is usually taken care of by multiplying the test statistic by a correction factor, which has to be determined numerically (Croux and Haesbroeck 1999). Using such an appropriate finite-sample value of $\sigma_1$ repairs the test, but again, it does not improve the model selection in our example. For the Miyamura–Kano proposal we note that they devise an alternative way of constrained estimation, but propose a very slow algorithm, which makes it, at least in the R implementation we used, unfeasible in larger dimensions. There is a tuning parameter to choose, which was set to $0 \cdot 3$ in our experiment, following the recommendation of the authors. All calculations were done in R 2.9.1, employing routines from the packages mvtnorm, ggm, ICSNP, rrcov and rggm.

8. Conclusion

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
Table 2. One-step model selection based on robust estimators

| distribution | estimator          | mean difference | % true model found | % non-edges correctly found | % ∅→∅ 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 and Kano (2006)

we recommend the use of 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. Vogel et al. (2010) report 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 reweighted minimum covariance determinant estimator. Moreover, our approach allows the use of any affine pseudo-equivariant, root-$n$-consistent estimator $\hat{S}_n$ in an analogous way. Assumption 3.1 is the important prerequisite on $\hat{S}_n$, and our results also apply to estimators that are asymptotically affine equivariant, like the rank-based estimation technique of Hallin et al. (2006).

A problem that has not been addressed in this article is the accuracy of the asymptotic approximations for small to moderate sample sizes, in particular, to what extent it depends upon the ratio $p/n$. This question splits into two parts. The first is an evaluation of the finite-sample properties of the affine pseudo-equivariant scatter estimators. These may be very different and do not allow a unified treatment. Very little seems to be known theoretically, either on the exact distribution of most robust scatter estimators or the rate of convergence to the Gaussian limit. However, there is strong empirical evidence that Tyler’s estimator has excellent small-sample properties. In all our simulations the difference in the empirical distributions of any univariate function of the sample covariance matrix $\hat{\Sigma}_n$ at normality and the Tyler matrix $\hat{V}_n$ at any elliptical distribution, is fully expressed by the asymptotic scaling factor $1 + 2/p$, see also Vogel et al. (2010, Figure 2). Moreover, it is known that $\hat{V}_n$ behaves similarly to $\hat{\Sigma}_n$ when $p$ and $n$ grow large simultaneously (Dümbgen, 1998). The second task is then, given the small-sample properties of the estimators, to assess the accuracy of the asymptotic $\chi^2$ distributions of the tests. This question is of relevance also in classical graphical modelling, where it has been noted that the deviance test statistic may substantially differ from its $\chi^2$ limit for small $n$. 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). Our simulations indicate that finite-sample correction techniques used in Gaussian graphical modelling may be put to good use also under ellipticity by applying it in an analogous way to Tyler’s estimator.

The main limitation of the affine equivariant approach is that it does not provide a solution in the $p > n$ situation or allow a simple transfer of standard techniques, like regularization, that are used in Gaussian graphical modelling. 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). Dropping the affine equivariance property is inevitable for robust, high-dimensional graphical modelling.
Acknowledgement

This research was supported by the German Research Foundation. The authors gratefully acknowledge the assistance of Alexander Dürre in preparing the figures and the simulations and thank the referees, the associate editor and the editor for their helpful comments and suggestions.

Appendix A. Proofs

The proofs repeatedly apply the delta method to functions mapping matrices to matrices. We define the derivative of such a function, say, \( g : \mathbb{R}^{p \times p} \rightarrow \mathbb{R}^{p \times p} \) at point \( X \) as the derivative of \( \text{vec}g(X) \) with respect to \( \text{vec}(X) \) and denote its Jacobian at point \( X \), which is of size \( p^2 \times p^2 \), by \( Dg(X) \). The symmetry of the argument poses a technical difficulty: there are \( p^2 \) variables, and the function \( g \) must be viewed as a function from \( \mathbb{R}^{p(p+1)/2} \) to \( \mathbb{R}^{p \times p} \) in order to define a derivative. To deal with this issue we 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 \( g_2 : \mathbb{R}^{p \times p} \rightarrow \mathbb{R}^{p \times p} \). The derivatives below contain the right-multiplied \( M_p \) depending on whether we view the function as defined on \( \mathcal{F}_p \) or on \( \mathbb{R}^{p \times p} \). The textbook Magnus and Neudecker (1999) 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.

\[
(A \otimes B)(C \otimes D) = AC \otimes BD, \quad (\text{vec}A)^T \text{vec}B = \text{tr}(A^T B), \quad \text{vec}(ABC) = (C^T \otimes A) \text{vec}B, \\
M_p = M_p^2, \quad M_p(A \otimes A)M_p = M_p(A \otimes A) = (A \otimes A)M_p,
\]

for matrices \( A, B, C, D \in \mathbb{R}^{p \times p} \) (Magnus and Neudecker, 1999, pp. 28, 30, 31). Let \( \iota: A \mapsto A^{-1} \) denote matrix inversion. Its Jacobian matrix is (Magnus and Neudecker, 1999, p. 184)

\[
D\iota(A) = -(A^T)^{-1} \otimes A^{-1}.
\]

Proof of Proposition 3.2. Part (i) follows by straightforward calculations from the delta method.

Part (ii): We have \( \tilde{P}_n = \tilde{h}(\tilde{K}_n) \) with \( \tilde{h} : A \mapsto -A_D^{-1/2} AA_D^{-1/2} \). We need to compute the derivative of \( \tilde{h} \) in order to apply the delta method. We start by considering \( \tilde{h}_0 : A \mapsto A_D^{-1/2} \). Its Jacobian matrix \( D\tilde{h}_0(A) = -\frac{1}{2} \left\{ A_D^{-1/2} \otimes A_D^{-1} \right\} J_p \) is obtained by elementwise differentiation. Applying the multiplication rule to \( \tilde{h}(A) = -\tilde{h}_0(A)\tilde{h}_0(A) \) yields

\[
(5) \quad D\tilde{h}(A) = -M_p \left\{ \tilde{h}(A) \otimes A_D^{-1} \right\} J_p - A_D^{-1/2} \otimes A_D^{-1/2}.
\]

By the delta method,

\[
n^{1/2} \text{vec}(\tilde{P}_n - P) = n^{1/2} \text{vec}\left\{ \tilde{h}(\tilde{K}) - \tilde{h}(\eta^{-1} K) \right\}
\]

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

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

which reduces to the expression given in Proposition 3.2. In particular, \( \sigma_2 \) vanishes, since \( D\tilde{h}(K) \text{vec}K = 0 \). This is generally true for scale-invariant function \( \tilde{h} \).

Proof of Proposition 4.1. Part (ii): Since \( K_G = \tilde{h}_G(S) \) with

\[
\tilde{h}_G : A \mapsto \sum_{k=1}^{2c-1} \zeta_k \left( A_{\alpha_k \alpha_k}^{-1} \right)^{(p)}
\]
we want to compute the derivative of $\tilde{h}_G$. Let $\tilde{h}_\alpha : A \mapsto (A_{\alpha,\alpha}^{-1})^{(p)}$ for any subset $\alpha \subset \{1, \ldots, p\}$. The mapping $\tilde{h}_\alpha$ is a composition of $(\cdot)_{\alpha,\alpha}$, $\iota$ and $(\cdot)^{(p)}$. We obtain by the chain rule

$$
\mathbb{D} \tilde{h}_\alpha(A) = -\{(A_{\alpha,\alpha}^{-1})^T\}^{(p)} \otimes (A_{\alpha,\alpha}^{-1})^{(p)}, \quad \mathbb{D} \tilde{h}_G(A) = -\sum_{k=1}^{2c-1} \zeta_k \{(A_{\alpha_k,\alpha_k}^{-1})^T\}^{(p)} \otimes (A_{\alpha_k,\alpha_k}^{-1})^{(p)}.
$$

Then $\eta^{-2} W_{K_G}(\sigma_1, \sigma_2) = \mathbb{D} \tilde{h}_{G}(\eta S) \eta^2 W_S(\sigma_1, \sigma_2) \{ \mathbb{D} \tilde{h}_{G}(\eta S) \}^T$ is shown to have the form given in Proposition 4.1 by noting that $\mathbb{D} \tilde{h}_G(S) \text{vec} S = \text{vec} K_G$. This holds true because

$$(S_{\alpha,\alpha}^{-1})^{(p)} S (S_{\alpha,\alpha}^{-1})^{(p)} = (S_{\alpha,\alpha}^{-1})^{(p)},$$

which is a consequence of the inversion formula for partitioned matrices.

**Part (iii)**: Applying the delta method we have to left- and right-multiply $W_{K_G}$ by the Jacobian of $\tilde{h}$ evaluated at $K_G$. Note that $(S_G \otimes S_G) \text{vec} K_G = \text{vec} S_G$.

**Part (iv)**: We left- and right-multiply $W_{K_G}$ by the Jacobian of $\tilde{h}$, given in [5], evaluated at $K_G$. □

**Proof of Corollary 4.2**. Let $S \in \mathcal{S}_p^+$ be such that $h_G(S) = S$ and write short $\Omega$ for $\Omega_G(S)$. It suffices to show that $2M_p \Omega(S \otimes S) \Omega = 2M_p \Omega$. Proposition 4.1 in connection with Proposition 6.1 identifies the left-hand side as the asymptotic covariance of $h_G(\hat{\Sigma})$, where $\hat{\Sigma}$ is the sample covariance matrix, at the normal distribution with covariance $S$. Formula (5.50) in Lauritzen (1996) identifies the same quantity as the right-hand side. □

In the proofs of Proposition 5.1 and Corollary 5.3 we use the following lemma.

**Lemma A.1**. Let $\mathbb{X}_n$ and $\mathbb{X}^{(m)}_n$, $m, n \in \mathbb{N}$, be as in Proposition 5.1 and $\hat{S}_n$ a shape estimator such that $\hat{S}_n(\mathbb{X}_n)$ satisfies Assumption 3.1. Assume furthermore that there is a continuously differentiable function $\xi : \mathbb{R}^{p \times p} \to \mathbb{R}$ with $\xi(I_p) = 1$ such that $\hat{S}_n$ satisfies

$$(6) \quad \hat{S}_n(\mathbb{X}_n A^T + 1_n b^T) = \xi(AA^T)A\hat{S}_n(\mathbb{X}_n)A^T$$

for any data matrix $\mathbb{X}_n \in \mathbb{R}^{n \times p}$, $b \in \mathbb{R}^p$ and full rank matrix $A \in \mathbb{R}^{p \times p}$. Then

$$n^{1/2} \text{vec} \left\{ \hat{S}_n(\mathbb{X}^{(m)}_n) - \eta \Omega \right\} \to N_p^2 \left\{ \eta(B + cS), \eta^2 W_S(\sigma_1, \sigma_2) \right\}$$

in distribution as $n \to \infty$, where $B$ is as in Proposition 5.1 and $c = D\xi(I_p) \text{vec}(S^{-1/2}BS^{-1/2})$.

The proof of Lemma A.1 follows by straightforward calculations and is omitted. The constant $c$ is identified by means of the first order Taylor expansion of $\xi(S_n^{-1/2}S^{-1}S_n^{1/2})$ around $I_p$.

**Proof of Proposition 5.1**. Part (i) follows by standard arguments from the asymptotic normality of the estimator $\hat{P}_G$. For any affine pseudo-equivariant estimator $\hat{S}_n$ the rescaled estimator $\hat{S}_n = \text{det}(\hat{S}_n)^{-1/p} \hat{S}_n$ satisfies (6), and the value of the test statistic $\hat{T}_n(G_0, G_1)$ is the same, if computed from $\hat{S}_n$ or $\hat{S}_G$. Applying Lemma A.1 to $\hat{S}_n$ we deduce part (iii) for analogously to part (ii). □

Towards the proof of Proposition 5.2 we state Lemmas A.2 to A.4. For $A \in \mathcal{S}_p^+$ let $f_A : \mathcal{S}_p^+ \to \mathbb{R}$: $f_A(B) = \log \text{det} B + \text{tr}(B^{-1} A)$. From the theory of Gaussian graphical models we know that for any graph $G$ and $A \in \mathcal{S}_p^+$ the matrix $A_G = h_G(A)$ is the unique solution of the constrained optimization problem

$$(7) \quad \text{minimize } f_A(B) \quad \text{subject to } Q_{D(G)} \text{vec} h(B) = 0, \ B \in \mathcal{S}_p^+;$$

because $A_G$ is the maximum likelihood estimate of the covariance matrix under the model $G$ at the normal distribution, if $A$ is the observed sample covariance, cf. Lauritzen (1996, p. 133). Now with the notation of Section 5 let $H_0(\cdot) = Q_{D(G_0)} \text{vec} h(\cdot)$, $H_1(\cdot) = Q_{D(G_1)} \text{vec} h(\cdot)$ and $H_{0,1}(\cdot) = Q_{D(G_0)\setminus D(G_1)} \text{vec} h(\cdot)$.
Lemma A.2. \( A_{G_0} = h_{G_0}(A) \) is a solution of the constrained optimization problem

\[
\begin{align*}
\text{minimize } & f_{A_{G_0}}(h_{G_0}(C)) \\
\text{subject to } & H_{0,1}(h_{G_0}(C)) = 0, \quad C \in \mathcal{A}_p^+.
\end{align*}
\]

Proof. By \((7)\) and \((3)\), \( A_{G_0} \) uniquely solves the constrained optimization problem

\[
\begin{align*}
\text{minimize } & f_{A_{G_1}}(B) \\
\text{subject to } & H_0(B) = 0, \quad B \in \mathcal{A}_p^+.
\end{align*}
\]

The restriction \( H_0(B) = 0 \) is equivalent to \( H_1(B) = 0 \) and \( H_{0,1}(B) = 0 \), and any matrix \( B \) with \( H_1(B) = 0 \) can be written as \( B = h_{G_1}(C) \) for some \( C \in \mathcal{A}_p^+ \). Thus \( \{ B \mid H_0(B) = 0, B \in \mathcal{A}_p^+ \} \) and \( \mathcal{C} = \{ B = h_{G_1}(C) \mid H_{0,1}(h_{G_1}(C)) = 0, \quad C \in \mathcal{A}_p^+ \} \) are equal, and so are the solution sets of the constrained optimization problems \((9)\) and \((10)\).

Thus \( A_{G_0} \) uniquely solves \((10)\), and all matrices \( C \in \mathcal{A}_p^+ \) with \( h_{G_1}(C) = A_{G_0} \), among them \( A_{G_0} \), solve \((8)\).

The next two lemmas are stated without proof. Expressions \((12)\) can be deduced from the proofs of Propositions \(3.2\) and \(4.1\) and \((11)\) can be assembled from the derivatives given in Magnus and Neudecker (1999) pp. 178,179).

Lemma A.3. Let \( H : \mathcal{A}_p \to \mathbb{R}^p \) be continuously differentiable and \( G_0, G_1 \) as in Section 3. Let furthermore \( \hat{S}_n \) be a sequence of almost surely positive definite random \( p \times p \) matrices, for which \( n^{1/2}(\hat{S}_n - S) \) converges in distribution for some \( S \in \mathcal{A}_p^+ \) with \( S^{-1} \in \mathcal{A}_p^+(G_0) \). Then for \( n \to \infty \)

\[
n^{1/2} \left\{ H(\hat{S}_{G_0}) - H(\hat{S}_{G_1}) \right\} \sim n^{1/2} \text{DH}(\hat{S}_{G_0}) \text{vec} \left( \hat{S}_{G_0} - \hat{S}_{G_1} \right).
\]

Lemma A.4. For \( A, B \in \mathcal{A}_p^+ \),

\[
\begin{align*}
(11) \quad & \text{Df}_A(B) = \text{vec}(B - A)^T(B^{-1} \otimes B^{-1})M_p, \\
(12) \quad & \text{Dh}_G(B) = \{ h_G(B) \otimes h_G(B) \} \Omega_G(B) M_p, \quad \text{DH}_{0,1}(B) = Q_{0,1} \Gamma(B) (B^{-1} \otimes B^{-1}) M_p.
\end{align*}
\]

Proof of Proposition 5.2. The second order Taylor expansion of \( \log \det(\cdot) \) is

\[
\log \det(A + X) = \log \det A + \{ \text{vec}(A^T)^-1 \}^T X - \frac{1}{2} \{ \text{vec}(X^T)^-1 \}^T \{ (A^T)^-1 \otimes A^{-1} \} \text{vec} X + o(\|X\|^2),
\]

cf. Magnus and Neudecker (1999) pp. 108, 179, 184). Applying this to the deviance test statistic yields

\[
\begin{align*}
\hat{D}_n(G_0, G_1) &= n \left\{ \log \det(\hat{S}_{G_0}) - \log \det(\hat{S}_{G_1}) \right\} = -n \log \det \left( \hat{S}_{G_1}, \hat{S}_{G_0}^{-1} \right) \\
&= -n \text{tr} \left( \hat{S}_{G_1}, \hat{S}_{G_0}^{-1} - I_p \right) + \frac{n}{2} \text{tr} \left( \left( \hat{S}_{G_1}, \hat{S}_{G_0}^{-1} - I_p \right)^2 \right) + o \left( n\|\hat{S}_{G_1}, \hat{S}_{G_0}^{-1} - I_p\|^2 \right) \\
&\sim \frac{n}{2} \left\{ \text{vec} \left( \hat{S}_{G_1} - \hat{S}_{G_0} \right) \right\}^T \left( \hat{S}_{G_0}^{-1} \otimes \hat{S}_{G_0}^{-1} \right) \text{vec} \left( \hat{S}_{G_1} - \hat{S}_{G_0} \right), \quad n \to \infty.
\end{align*}
\]

The asymptotic equivalence follows because

\[
\begin{align*}
(1) \quad & \text{tr} \left( \hat{S}_{G_1}, \hat{S}_{G_0}^{-1} - I_p \right) = \left\{ \text{vec} \left( \hat{S}_{G_1} - \hat{S}_{G_0} \right) \right\}^T \text{vec} \hat{S}_{G_0}^{-1} = 0, \quad \text{which is a consequence of (3)},
\end{align*}
\]

\[
\begin{align*}
(2) \quad & n\|\hat{S}_{G_1}, \hat{S}_{G_0}^{-1} - I_p\|^2 \leq \left( n^{1/2}\|\hat{S}_{G_1} - S\| + n^{1/2}\|\hat{S}_{G_0} - S\| \right)^2 \|\hat{S}_{G_0}^{-1}\|^2 = O_p(1), \quad n \to \infty.
\end{align*}
\]

Applying Lemma A.3 to \( H = h_{G_1} \) and using \((12)\) we find further

\[
\begin{align*}
n^{1/2} \text{vec} \left( \hat{S}_{G_0} - \hat{S}_{G_1} \right) \sim n^{1/2} \left( \hat{S}_{G_0} \otimes \hat{S}_{G_0} \right) \Omega_{G_1}(\hat{S}_{G_0}) M_p \text{vec} \left( \hat{S}_{G_0} - \hat{S}_{G_1} \right)
\end{align*}
\]

and from \((13)\)

\[
\begin{align*}
\hat{D}_n(G_0, G_1) \sim \frac{n}{2} \left\{ \text{vec} \left( \hat{S}_{G_1} - \hat{S}_{G_0} \right) \right\}^T M_p \Omega_{G_1}(\hat{S}_{G_0}) \text{vec} \left( \hat{S}_{G_1} - \hat{S}_{G_0} \right), \quad n \to \infty.
\end{align*}
\]
Next we introduce the Lagrange multiplier \cite{Magnus and Neudecker 1999} p. 131. Since \( \hat{S}_{G_0} \) solves the constrained optimization problem \( f_{\hat{S}_{G_0}}(S) \) with \( A = S_n \), there exists a vector \( \lambda \in \mathbb{R}^{p+1} \) such that

\[
D \left( f_{\hat{S}_{G_1}} \circ h_{G_1} \right) \left( \hat{S}_{G_0} \right) = \lambda^T D \left( h_{0,1} \circ h_{G_1} \right) \left( \hat{S}_{G_0} \right),
\]

which transforms to

\[
M_p \Omega_{G_1}(\hat{S}_{G_0}) \text{vec}(\hat{S}_{G_1} - \hat{S}_{G_0}) = M_p \Omega_{G_1}(\hat{S}_{G_0}) \Gamma(\hat{S}_{G_0})^T Q_{0,1}^T \lambda, \quad \text{cf. Lemma A.4}
\]

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

\[
M_p \Omega_{G_1}(\hat{S}_{G_0}) \text{vec}(\hat{S}_{G_1} - \hat{S}_{G_0}) = M_p \Omega_{G_1}(\hat{S}_{G_0}) \Gamma(\hat{S}_{G_0})^T Q_{0,1}^T \lambda
\]

We substitute the right-hand side for the left-hand side of this equation in (14), apply again Lemma A.3 this time to \( H = h_{0,1} \circ h_{G_1} \), which leads to

\[
n^{1/2} Q_{0,1} \text{vec} \hat{P}_{G_1} \sim n^{1/2} Q_{0,1} \Gamma(\hat{S}_{G_0}) M_p \Omega_{G_1}(\hat{S}_{G_0}) \text{vec}(\hat{S}_{G_1} - \hat{S}_{G_0}),
\]

and obtain

\[
\hat{D}_n(G_0, G_1) \sim \frac{n}{2} \left( \text{vec} \hat{P}_{G_1} \right)^T Q_{0,1} \left\{ Q_{0,1} R_{G_1}(\hat{S}_{G_0}) Q_{0,1} \right\}^{-1} Q_{0,1} \text{vec} \hat{P}_{G_1}, \quad n \to \infty.
\]

Finally \( R_{G_1}(\hat{S}_{G_0}) \sim R_{G_1}(\hat{S}_{n}) \) as \( n \to \infty \), since both sides converge to \( R_{G_1}(S) \).

\[\square\]

**Proof of Corollary 5.3** Part (1) is straightforward. For part (2) we take, as in Proposition 5.1 the detour via \( \hat{S}_n = \text{det} \left( \hat{S}_n \right)^{-1/p} \hat{S}_n \) and make use of Lemma A.1 to ensure that \( \hat{S}_n(X^{(n)}_n) \) meets the assumptions of Proposition 5.2

\[\square\]

**References**

K. Baba, R. Shibata, and M. Sibuya. Partial correlation and conditional correlation as measures of conditional independence. *Aust. N. Z. J. Stat.*, 46(4):657–664, 2004.

C. Becker. Iterative proportional scaling based on a robust start estimator. In C. Weihs and W. Gaul, editors, *Classification - The Ubiquitous Challenge*, pages 248–255. Heidelberg: Springer, 2005.

M. Bilodeau and D. Brenner. *Theory of Multivariate Statistics*. New York, NY: Springer, 1999.

D. R. Cox and N. Wermuth. *Multivariate Dependencies: Models, Analysis and Interpretation*. London: Chapman and Hall, 1996.

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

L. Dümbgen. On Tyler’s M-functional of scatter in high dimension. *Ann. Inst. Stat. Math.*, 50(3):471–491, 1998.

D. Edwards. *Introduction to Graphical Modelling*. New York, NY: Springer, 2000.

K.-T. Fang and Y.-T. Zhang. *Generalized Multivariate Analysis*. Berlin etc.: Springer-Verlag; Beijing: Science Press., 1990.

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

A. Gottard and S. Pacillo. Robust concentration graph model selection. *Comput. Statist. Data Anal.*, 54 (12):3070–3079, 2010.

M. Hallin, H. Oja, and D. Paindaveine. Semiparametrically efficient rank-based inference for shape. II: Optimal R-estimation of shape. *Ann. Stat.*, 34(6):2757–2789, 2006.

T. Hettmansperger and R. Randles. A practical affine equivariant multivariate median. *Biometrika*, 89: 851–860, 2002.

S. L. Lauritzen. *Graphical Models*. Oxford: Oxford Univ. Press, 1996.

J. R. Magnus and H. Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Chichester: Wiley, 2nd edition, 1999.

R. A. Maronna, D. R. Martin, and V. J. Yohai. *Robust Statistics: Theory and Methods*. Chichester: Wiley, 2006.

M. Miyamura and Y. Kano. Robust Gaussian graphical modeling. *J. Multivariate. Anal.*, 97(7):1525–1550, 2006. ISSN 0047-259X.
E. Ollila, H. Oja, and C. Croux. The affine equivariant sign covariance matrix: Asymptotic behavior and efficiencies. *J. Multivariate Anal.*, 87(2):328–355, 2003.

E. Ollila, C. Croux, and H. Oja. Influence function and asymptotic efficiency of the affine equivariant rank covariance matrix. *Stat. Sin.*, 14(1):297–316, 2004.

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

B. Porteous. A note on improved likelihood ratio statistics for generalized log linear models. *Biometrika*, 72:473–475, 1985.

B. T. Porteous. Stochastic inequalities relating a class of log-likelihood ratio statistics to their asymptotic $\chi^2$ distribution. *Ann. Stat.*, 17(4):1723–1734, 1989.

P. J. Rousseeuw and A. M. Leroy. *Robust Regression and Outlier Detection*. New York etc.: Wiley, 1987.

D. E. Tyler. Radial estimates and the test for sphericity. *Biometrika*, 69:429–436, 1982.

D. E. Tyler. Robustness and efficiency properties of scatter matrices. *Biometrika*, 70:411–420, 1983.

D. E. Tyler. A distribution-free M-estimator of multivariate scatter. *Ann. Stat.*, 15:234–251, 1987.

D. E. Tyler. A note on multivariate location and scatter statistics for sparse data sets. *Stat. Probab. Lett.*, 80(17-18):1409–1413, 2010.

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

D. Vogel, A. Dürre, and R. Fried. Elliptical graphical modeling in higher dimensions. In *Proceedings of International Biosignal Processing Conference, July 14-16, 2010, Berlin, Germany.*, pages 1–5, 2010.

J. Whittaker. *Graphical Models in Applied Multivariate Statistics*. Chichester etc.: Wiley, 1990.

Y. Zuo. Robust location and scatter estimators in multivariate analysis. In J. Fan and H. Koul, editors, *Frontiers in Statistics. Dedicated to Peter John Bickel on Honor of his 65th Birthday*, pages 467–490. London: Imperial College Press, 2006.

Fakultät Statistik, Technische Universität Dortmund, 44221 Dortmund, Germany
E-mail address: daniel.vogel@tu-dortmund.de
E-mail address: fried@statistik.tu-dortmund.de