Universität Augsburg

Institut für Mathematik

Andreas Käufl

The Distribution of the Maximum Likelihood Estimator in Invariant Gaussian Graphical Models and its Application to Likelihood Ratio Tests

Preprint Nr. 16/2012 — 17. Dezember 2012

Institut für Mathematik, Universitätsstraße, D-86135 Augsburg http://www.math.uni-augsburg.de/

Impressum:

Herausgeber:

Institut für Mathematik
Universität Augsburg
86135 Augsburg
`http://www.math.uni-augsburg.de/de/forschung/preprints.html`

ViSdP:

Andreas Käufl
Institut für Mathematik
Universität Augsburg
86135 Augsburg

The Distribution of the Maximum Likelihood Estimator in Invariant Gaussian Graphical Models and its Application to Likelihood Ratio Tests

Andreas Käufl*

December 11, 2012

Abstract The distribution of the maximum likelihood estimator of the covariance matrix in a class of invariant Gaussian graphical models is determined and seen to belong to the class of generalized Riesz distributions. For testing two nested models, the distribution of the likelihood ratio statistic under the null hypothesis is shown to be of Box type, so that accurate approximation techniques are applicable.

Keywords Dimension Reduction, Wishart Distribution, Generalized Riesz Distribution, Gaussian Models, Box Approximation

1 Introduction

A graphical model only allows for distributions that respect certain conditional independence constraints. These restrictions can be represented intuitively with the help of a graph. In a graphical model with normality assumption, it can be shown that every conditional independence restriction causes an entry in the inverse of the covariance matrix to vanish. Consequently, the parameter space of a graphical model is a subset of the cone of positive definite matrices, in which certain determinants vanish. Even though in most cases a graphical model therefore has fewer parameters than the saturated model, it still may have a very high-dimensional parameter space. One way to further reduce this dimension is the introduction of symmetry restrictions in the model. This can lead to a significant reduction in the number of parameters.

In the present paper, we are concerned with statistical inference in an invariant graphical model as introduced by Madsen [1]. By definition, the distributions in such a model

*Department of Mathematics, Augsburg University, Andreas.Kaeufl@googlemail.com

respect the independence constraints given by an acyclic directed graph $\mathcal{D}$. Since we assume that $\mathcal{D}$ does not contain an immorality, the parameter space of the graphical model corresponding to $\mathcal{D}$ can also be described by an undirected graph $\mathcal{U}$ and this description will prove useful for our purposes.

Our main results are twofold: First, we compute the distribution of the maximum likelihood estimator in such a model, see Theorem 5.2. Second, this result is used to show that in an invariant graphical testing problem, the likelihood ratio statistic follows a Box-type distribution under the null hypothesis, see Theorem 6.2. This allows for very accurate approximation techniques.

In order to prove these two results, we start by establishing our notation and defining invariant graphical models in Section 2. Afterwards, we introduce the generalized Riesz distribution in Section 3. In Section 4, we present results related to maximum likelihood estimation in graphical models and invariant graphical models. This enables us to compute the distribution of the maximum likelihood estimator in an invariant graphical model in Section 5. In Section 6, we address the problem of how likelihood ratio tests between two nested invariant graphical models can be performed. In Section 7 we summarize our results and give a short outlook.

General considerations about how to include symmetries in statistical models have been made since the 1970s. An introduction to this topic is given by Andersson [2]. The commentary of Perlman [3] on the other hand offers a broader overview over the research in the area of symmetry constraints in statistical models. How invariance restrictions can be included in graphical models in particular is examined in the papers of Andersson and Madsen [4] and Madsen [1]. In this last paper, the author not only defines invariant graphical models but also develops a sufficient criterion for the existence of the maximum likelihood estimator and specifies its form. However, the question of what the distribution of this estimator is, was not addressed.

In the existing literature we found several candidates for this distribution. They all have in common that they are concentrated on certain subsets of the cone of positive definite matrices and can be seen as generalizations of the classical Wishart distribution. Examples include the papers of Dawid and Lauritzen [5], Andersson and Wojnar [6] and Letac and Massam [7]. The most recent work in this direction was developed by Andersson and Klein [8]. They define the so-called generalized Riesz distribution that turns out to be very useful for proving the two theorems mentioned above.

2 Invariant graphical models

We briefly introduce some notation and then review acyclic directed graphs. Let I be a nonempty finite index set. For the sake of readability we denote the cardinality of the set I also by I, the context will prevent misunderstandings. For a vector $x = (x_i)_{i\in I} \in \mathbb{R}^I$ and a subset $J \subseteq I$, we use the notation x_J for the subvector of x corresponding to the index set J, that is, $x_J = (x_j)_{j\in J} \in \mathbb{R}^J$. For a matrix $A = (a_{ij})_{i,j\in I} \in \mathbb{R}^{I\times I}$ we denote the submatrix corresponding to the index sets $K \subseteq I$ and $L \subseteq I$ by $A_{KL} = (a_{k\ell})_{k\in K,\ell\in L} \in \mathbb{R}^{K\times L}$.

Let $PD(I)$ be the cone of all positive definite $I \times I$ matrices. By convention, index operations are carried out before inversion or transposition, that is, $A_{KL}^{-1} = (A_{KL})^{-1}$ and $A_{KL}^T = (A_{KL})^T$ for index sets $K, L \subseteq I$.

Let V be another finite set. For every element $v \in V$ let $[v]$ be a non-empty subset $[v] \subseteq I$, such that the entire index set I is partitioned according to $I = \cup_{v \in V}[v]$. Moreover, let $E \subseteq V \times V$ with

$$\forall v, w \in V : (v, w) \in E \Rightarrow (w, v) \notin E \quad \text{and} \quad \forall v \in V : (v, v) \notin E.$$

This way, the pair (V, E) can be regarded as a directed graph $\mathcal{D}$ with vertices V and edges E. For an edge $(v, w) \in E$ we also use the more intuitive notation $v \to w \in E$.

The interpretation of this framework is as follows: We are interested in a $\mathbb{R}^I$-valued random vector X. This vector is partitioned into smaller random vectors $\{X_{[v]}\}_{v \in V}$. The marginal model for such an $X_{[v]}$ is saturated, that is, within a block $[v]$ all interactions between the components $\{X_i\}_{i \in [v]}$ are allowed. Between the blocks, on the other hand, several conditional independence restrictions are enforced. These constraints are represented by the graph $\mathcal{D}$.

Throughout this paper we assume that $\mathcal{D}$ has no immoralities, that is, no induced subgraphs of the form $\bullet \to \bullet \leftarrow \bullet$. For $v \neq w$ we write $v < w$ if w can be reached from v via a directed path, that is, there exists a $k \in \mathbb{N}$ and nodes $v_1, \ldots, v_k \in V$ such that $v = v_1 \to v_2 \to \cdots \to v_k = w$. Analogously we write $v \leq w$ if $v = w$ or $v < w$. A directed graph $\mathcal{D} = (V, E)$ is called acyclic if it does not contain a directed cycle, that is, $v \not< v$ for all nodes $v \in V$. The nodes in the set $\mathrm{pa}(v) := \{w \in V | w \to v \in E\}$ are called the parents of $v \in V$. The set $\langle v \rangle = \cup_{w \in \mathrm{pa}(v)}[w]$ is the subset of I that corresponds to the parents of v in $\mathcal{D}$. A node $v \in V$ is called maximal in $\mathcal{D}$ if no directed edge is pointing out of v.

Let $A \in \mathbb{R}^{I \times I}$ be a square matrix and let $v \in V$ be a node of $\mathcal{D}$. The submatrix of A corresponding to v is denoted by $A_{[v]} := A_{[v][v]}$. Analogously, we use the notation $A_{\langle v \rangle} := A_{\langle v \rangle \langle v \rangle}$ and $A_{[v\rangle} := A_{[v]\langle v \rangle}$. If A is positive definite, we can define the matrices $A_{[v]\bullet} := A_{[v]} - A_{[v\rangle} A_{\langle v \rangle}^{-1} A_{[v\rangle}^T \in PD([v])$ and $A_{[v\rangle\bullet} := A_{[v\rangle} A_{\langle v \rangle}^{-1} \in \mathbb{R}^{[v] \times \langle v \rangle}$ for every node $v \in V$.

Define a new edge set $E^\sim$ such that $(v, w) \in E^\sim$ if and only if $(v, w) \in E$ or $(w, v) \in E$. In other words, $E^\sim$ contains an undirected edge $v - w$ for every directed edge $v \to w$ in E. Therefore, the pair $\mathcal{U} = (V, E^\sim)$ can be interpreted as an undirected graph and is called the skeleton of $\mathcal{D}$. A clique of $\mathcal{U} = (V, E^\sim)$ is a maximal subset $C \subseteq V$ such that for every two nodes $v, w \in C$ the undirected edge $v - w \in E^\sim$ appears in $\mathcal{U}$. It will prove useful to consider the subsets of I that correspond to the cliques of $\mathcal{U}$,

$$\mathcal{C}(\mathcal{U}) := \{\cup_{v \in C}[v] \mid C \text{ clique of } \mathcal{U}\}. \tag{1}$$

We can now explain why we are only allowing acyclic directed graphs $\mathcal{D}$ without immoralities. Andersson et al. [9] show that an acyclic directed graph $\mathcal{D}$ is Markov equivalent to a decomposable graph $\mathcal{U}$ if and only if $\mathcal{D}$ has no immoralities. In that case the graphical model corresponding to the graph $\mathcal{D} = (V, E)$ and the graphical model given by its skeleton $\mathcal{U} = (V, E^\sim)$ are identical.

We continue by introducing automorphisms over acyclic directed graphs and examining their orbits. For an acyclic directed graph $\mathcal{D} = (V, E)$, let $Sym(V)$ be the set of all bijections $\sigma : V \to V$. A permutation $\sigma \in Sym(V)$ is called an automorphism over $\mathcal{D}$ if

$$\forall v, w \in V : v \to w \in E \Leftrightarrow \sigma(v) \to \sigma(w) \in E.$$

An automorphism σ is called cardinality-respecting, if $\sigma(v) = w$ implies that the cardinalities of the corresponding sets $[v]$ and $[w]$ are identical. In other words, a cardinality-respecting automorphism over $\mathcal{D}$ is a permutation of the nodes, that keeps the causal structure of $\mathcal{D}$ intact and only exchanges nodes v and w, for which the index sets $[v]$ and $[w]$ contain the same number of elements. The set of all cardinality-respecting automorphisms over $\mathcal{D}$ is denoted by $Aut(\mathcal{D})$. It can easily be seen that $Aut(\mathcal{D})$ is a subgroup of $Sym(V)$.

Later in this section, it will prove useful to consider the permutation matrix corresponding to a automorphism $\sigma \in Aut(\mathcal{D})$ instead of the permutation σ itself. For that reason, we define the mapping $\varrho : Aut(\mathcal{D}) \to \mathbb{R}^{I \times I}$ such that for all $\sigma \in Aut(\mathcal{D})$ and for all $v, w \in V$ we have $\varrho(\sigma)_{[v][w]} = I_{[v]}$ if $w = \sigma^{-1}(v)$ and $\varrho(\sigma)_{[v][w]} = 0$ otherwise. That is, ϱ maps a cardinality-respecting automorphism onto its corresponding permutation matrix. Here and in the following, the $k \times k$ identity matrix will be denoted by I_k for any $k \in \mathbb{N}$. The image of $Aut(\mathcal{D})$ under ϱ is called the set of allowed symmetries and is denoted by $Perm(\mathcal{D})$. Evidently, $Perm(\mathcal{D})$ inherits the group property of $Aut(\mathcal{D})$. Since $\varrho : Aut(\mathcal{D}) \to Perm(\mathcal{D})$ is a bijection, the permutation $\sigma \in Aut(\mathcal{D})$ corresponding to the matrix $\rho \in Perm(\mathcal{D})$ can be written as $\sigma = \varrho^{-1}(\rho)$.

Let $H \subseteq Aut(\mathcal{D})$ be a subgroup and $v \in V$ be a node. The orbit of v under H is the set of nodes that are mapped to v by a permutation in H, that is, $Orb(v) = \{\sigma(v) | \sigma \in H\}$. For a group of permutation matrices $G \subseteq Perm(\mathcal{D})$ we also speak of the orbit of v under G and hereby mean the orbit of v under $\varrho^{-1}(G) = \{\varrho^{-1}(\rho) | \rho \in G\}$. For a second node $w \in V$, the number of permutations in H that map w onto v is denoted by $\kappa_v(w) = |\{\sigma \in H | \sigma(w) = v\}|$. It follows that $\kappa_v(w)$ is positive if and only if w lies in the orbit of v, that is, $w \in Orb(v)$. Moreover, it is not hard to show that $\kappa_v(w) = \kappa_v(v)$ for all $w \in Orb(v)$. This particularly implies that the number of permutations in H that map w onto v is constant for all nodes w in the orbit of v. This observation together with a simple counting argument yields the relation

$$\forall v \in V : |H| = |Orb(v)| \kappa_v \text{ with } \kappa_v := \kappa_v(v). \tag{2}$$

We are now in a position to define the parameter spaces of a graphical model and an invariant graphical model respectively. Let $\mathcal{D} = (V, E)$ again be an acyclic directed graph without immoralities and let $\mathcal{U} = (V, E^{\sim})$ be the skeleton of $\mathcal{D}$. Define

$$S(\mathcal{U}) := \left\{ W \in \mathbb{R}^{I \times I}_{symm} \mid W_{[v][w]} = 0 \text{ if } v \neq w \text{ and } v - w \notin E^{\sim} \right\}$$

as the set of symmetric block matrices that contain a zero block for every missing edge in $\mathcal{U}$. It is a well-known fact that the parameter space of the graphical model given by the graph $\mathcal{U}$ can be described as

$$PD(\mathcal{U}) := \left\{ \Sigma \in PD(I) \mid \Sigma^{-1} \in S(\mathcal{U}) \right\},$$

see for example Andersson and Klein [8, p. 792]. Since in general $PD(\mathcal{U})$ is not a convex cone, it will prove helpful to work with the alternative parameter space

$$P(\mathcal{U}) := \{\Omega \in S(\mathcal{U}) \mid \Omega_{CC} \in PD(C) \text{ for all } C \in \mathcal{C}(\mathcal{U})\}$$

with $\mathcal{C}(\mathcal{U})$ as defined in equation (1). To understand the relationship between $PD(\mathcal{U})$ and $P(\mathcal{U})$, consider the mapping

$$\pi : \mathbb{R}^{I\times I}_{symm} \to S(\mathcal{U}) \quad \text{with} \quad \pi(W)_{vw} = \begin{cases} W_{vw} & v = w \text{ or } v - w \in E^{\sim} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

that deletes all entries in a symmetric matrix W that correspond to a missing edge in $\mathcal{U}$. It turns out that the restriction $\pi : PD(\mathcal{U}) \to P(\mathcal{U})$ constitutes a diffeomorphism between $PD(\mathcal{U})$ and $P(\mathcal{U})$, see Andersson and Klein [8, Proposition 3.1]. Therefore, we can use the cone $P(\mathcal{U})$ as a parameter space for the graphical model given by $\mathcal{U}$ and $\mathcal{D}$ respectively.

To include symmetries in the model, we use the following definition.

Definition 2.1 *Let $\mathcal{D}$ be an acyclic directed graph without immoralities and let $\mathcal{U}$ be the skeleton of $\mathcal{D}$. Let $G \subseteq Perm(\mathcal{D})$ be a subgroup of the group of allowed symmetries. The parameter spaces of the invariant graphical model given by $\mathcal{U}$ and G are then defined as*

$$PD(\mathcal{U}, G) := \{\Sigma \in PD(\mathcal{U}) \mid \rho\Sigma\rho^T = \Sigma \ \forall \rho \in G\} \quad \textit{and}$$
$$P(\mathcal{U}, G) := \{\Omega \in P(\mathcal{U}) \mid \rho\Omega\rho^T = \Omega \ \forall \rho \in G\}.$$

Using the fact that G is a subgroup of $Perm(\mathcal{D})$, it is possible to show that the projection π commutes with the congruence transformation $W \mapsto \rho W \rho^T$ for all $\rho \in G$, that is, $\pi\left(\rho W \rho^T\right) = \rho\pi(W)\rho^T$ for all $W \in \mathbb{R}^{I\times I}_{symm}$. Using this observation, it follows that the image of $PD(\mathcal{U}, G)$ under π is the alternative parameter space $P(\mathcal{U}, G)$. Therefore, the restriction $\pi : PD(\mathcal{U}, G) \to P(\mathcal{U}, G)$ is also a diffeomorphism.

The inverse mapping $\pi^{-1} : P(\mathcal{U}) \to PD(\mathcal{U})$ can be described explicitly, see corollaries 5.1 and 5.3 in Andersson and Klein [8]. This description of $\pi^{-1}(\Omega) \in PD(\mathcal{U})$ for an $\Omega \in P(\mathcal{U})$ can be seen as a generalized Cholesky decomposition of Ω and mainly relies on the matrices $\{\Omega_{[v]\bullet}, \Omega_{[v\rangle\bullet}\}_{v\in V}$. Out of that reason, we call the matrices $\{\Omega_{[v]\bullet}, \Omega_{[v\rangle\bullet}\}_{v\in V}$ the Cholesky parameters of Ω.

Another interesting property of the Cholesky parameters becomes apparent when Ω is G-invariant, that is, $\Omega \in P(\mathcal{U}, G)$. In that case, Theorem 6.1 in Madsen [1] together with our definition of the mapping ϱ shows that

$$\Omega_{[\sigma_\rho^{-1}(v)]\bullet} = \Omega_{[v]\bullet} \quad \text{and} \quad \Omega_{[\sigma_\rho^{-1}(v)\rangle\bullet} = \Omega_{[v\rangle\bullet} \tag{4}$$

for all $\rho \in G$ and all $v \in V$. Here, σ_ρ means the permutation of nodes corresponding to the matrix $\rho \in Perm(\mathcal{D})$, that is, $\sigma_\rho = \varrho^{-1}(\rho)$. This means that symmetries in Ω are also reflected in its Cholesky parameters in the sense that all nodes in the same orbit have the same Cholesky parameters.

3 Generalized Riesz distribution

Let $\mathcal{D} = (V, E)$ be an acyclic directed graph without immoralities and let $\mathcal{U} = (V, E^{\sim})$ be its skeleton. Moreover, let $G \subseteq Perm(\mathcal{D})$ be a subgroup of the group of allowed symmetries. To describe the distribution of the maximum likelihood estimator in the graphical and the invariant graphical model respectively, we will make use of the generalized Riesz distribution. Aside from the graph $\mathcal{D}$ itself this distribution has two more parameters: An expectation parameter $\Omega \in P(\mathcal{U})$ and a shape parameter $\lambda \in \mathbb{R}^V$. We write $\mathcal{R}_p(\Omega, \lambda)$ for the generalized Riesz distribution of $p \times p$ matrices with parameters Ω and λ. Andersson and Klein define this class of distributions by specifying the density, see [8, Definition 10.1]. They also develop an alternative characterization that will prove useful for our purposes.

In order to describe this characterization we need to define some other matrix valued distributions first. For that purpose, let $n, p \in \mathbb{N}$ and $\mu \in \mathbb{R}^{n \times p}$ as well as $\Sigma \in PD(p)$ and $\Phi \in PD(n)$. By $\mathcal{N}_{n \times p}(\mu, \Phi \otimes \Sigma)$ we mean the normal distribution for $n \times p$ matrices with expectation parameter μ and dispersion parameter $\Phi \otimes \Sigma$. The parametrization of $\mathcal{N}_{n \times p}$ is the same as in Kollo and von Rosen [10, Chapter 2.2] and the symbol $\otimes$ stands for the Kronecker product.

For a centered random matrix $X \sim \mathcal{N}_{n \times p}(0, I_n \otimes \Sigma)$, the inner product $X^T X$ follows a $p \times p$ Wishart distribution with n degrees of freedom and parameter Σ, that is, $X^T X \sim \mathcal{W}_p(n, \Sigma)$. Here, we also chose the parametrization of Kollo and von Rosen [10, Chapter 2.4]. By making use of these well-known distributions, we can now state the alternative characterization of the generalized Riesz distribution.

Proposition 3.1 *Let W be a $P(\mathcal{U})$-valued random matrix and let $m \in V$ be a maximal node in $\mathcal{D}$. Let $\Omega \in P(\mathcal{U})$ and $\lambda \in \mathbb{R}^V$ with $\lambda_v > ([v] + \langle v \rangle - 1)/2$ for all $v \in V$.*

Then W follows the generalized Riesz distribution $\mathcal{R}_I(\Omega, \lambda)$ if and only if the following four conditions are met:

(i) $W_{[m]\bullet} \perp\!\!\!\perp \left\{W_{[m\rangle\bullet}, W_{I\setminus[m]}\right\}$,

(ii) $\mathcal{L}\left(W_{[m]\bullet}\right) = \mathcal{W}_{[m]}\left(2\lambda_m - \langle m\rangle, \frac{1}{2\lambda_m}\Omega_{[m]\bullet}\right)$,

(iii) $\mathcal{L}\left(W_{[m\rangle\bullet} | W_{I\setminus[m]}\right) = \mathcal{N}_{[m]\times\langle m\rangle}\left(\Omega_{[m\rangle\bullet}, \frac{1}{2\lambda_m}\Omega_{[m]\bullet} \otimes W^{-1}_{\langle m\rangle}\right)$,

(iv) $\mathcal{L}\left(W_{I\setminus[m]}\right) = \mathcal{R}_{I\setminus[m]}\left(\Omega_{I\setminus[m]}, \lambda_{V\setminus\{m\}}\right)$.

Proof
See Andersson and Klein [8, Proposition 10.2]. □

Remark 3.2 The generalized Riesz distribution $\mathcal{R}_I(\Omega, \lambda)$ not only depends on Ω and λ but also on the acyclic directed graph $\mathcal{D}$. That is, in general there are several different Riesz distributions over $P(\mathcal{U})$ corresponding to the parameter pair (Ω, λ), namely one for every acyclic directed graph $\mathcal{D}$ with skeleton $\mathcal{U}$.

For the purposes of the present paper, we do not need that much flexibility. Later on, we will work with a Riesz distributed random variable $W \sim \mathcal{R}_I(\Omega, \lambda)$ that has a

scalar shape parameter, that is, $\lambda = (\beta, \ldots, \beta)^T$ for a $\beta \in \mathbb{R}$. Computing the moment generating function $S \mapsto \mathbb{E}[\exp(\operatorname{tr}(SW))]$, we see that this function does not depend on the choice of $\mathcal{D}$ at all, see Andersson and Klein [8, Remark 8.1]. Consequently, we do not have to specify a certain choice of $\mathcal{D}$ in order to get a well-defined Riesz distribution as long as we have a scalar shape parameter λ.

4 Maximum likelihood estimation

We first describe the maximum likelihood estimator in the graphical model $P(\mathcal{U})$. To do so, let the random vectors $X_1, \ldots, X_n$ be independent and identically distributed according to a $\mathcal{N}_I(0, \Sigma)$ distribution for a $\Sigma \in PD(\mathcal{U})$. This distribution can also be parametrized by $\Omega = \pi(\Sigma) \in P(\mathcal{U})$, see equation (3) and the remark beneath it. Therefore, it suffices to specify the estimator $\hat{\Omega} \in P(\mathcal{U})$.

It is a well-known fact that this estimator exists with probability one if and only if the number of observations is not smaller than the number of nodes in the largest clique of $\mathcal{U}$, that ist, $n \geq \max\{|C| : C \in \mathcal{C}(\mathcal{U})\}$. In this case, the maximum likelihood estimator of Ω is given by

$$\hat{\Omega} = \pi(S) \in P(\mathcal{U}),$$

where $S = \sum_{i=1}^n X_i X_i^T / n$ is the sample covariance matrix. For a proof of this result, see for example Lauritzen [11, Proposition 5.9].

At this point, we benefit from introducing the generalized Riesz distribution in Section 3. It turns out that under the assumptions above, the estimator $\pi(S)$ follows a generalized Riesz distribution with expectation parameter $\Omega \in P(\mathcal{U})$ and shape parameter $\lambda = (n/2, \ldots, n/2) \in \mathbb{R}^V$, that is, $\pi(S) \sim \mathcal{R}_I(\Omega, \lambda)$.

Next, we turn to the maximum likelihood estimator in an invariant graphical model $P(\mathcal{U}, G)$. For that purpose, we need the balancing function ψ that averages over the orbit when the matrices $\rho \in G$ act by congruence,

$$\psi : PD(I) \to PD(I),\ W \mapsto \frac{1}{|G|} \sum_{\rho \in G} \rho W \rho^T \tag{5}$$

Note that ψ in particular depends on the group G. The restriction of ψ on $P(\mathcal{U})$ is of special interest since it is easy to show that this is a mapping $\psi : P(\mathcal{U}) \to P(\mathcal{U}, G)$, see Madsen [1, Proposition 6.1].

To describe the maximum likelihood estimator, let $X_1, \ldots, X_n$ be independent and identically distributed according to a $\mathcal{N}_I(0, \pi^{-1}(\Omega))$ distribution for an $\Omega \in P(\mathcal{U}, G)$. Corollary 7.2 in Madsen [1] shows that $\hat{\Omega}$ exists and is unique if $\psi(\pi(S))_{[v] \cup \langle v \rangle}$ is positive definite for all nodes $v \in V$. In that case, we have

$$\hat{\Omega} = \psi(\pi(S)) \in P(\mathcal{U}, G).$$

We want to stress that this result only gives a sufficient criterion for the existence of the maximum likelihood estimator. Throughout the rest of this paper we assume that this criterion is met, that is, the estimator $\hat{\Omega} \in P(\mathcal{U}, G)$ exists. The following section addresses the problem of computing the distribution of $\hat{\Omega}$.

5 Distribution of the maximum likelihood estimator

In this section we determine the distribution of the maximum likelihood estimator $\hat{\Omega} = \psi(\pi(S))$ in the invariant graphical model $P(\mathcal{U}, G)$. To make the notation a little easier, we address a more general problem. Let W follow a generalized Riesz distribution over $P(\mathcal{U})$ with expectation parameter $\Omega \in P(\mathcal{U})$ and a scalar shape parameter $\lambda = (\beta, \ldots, \beta)^T \in \mathbb{R}^V$ with $2\beta \in \mathbb{N}$ and $\beta > ([v]+\langle v\rangle - 1)/2$ for all $v \in V$. Our goal is to use this distribution to compute the distribution of $\psi(W)$. This directly gives us the distribution of $\psi(\pi(S))$ since $\pi(S)$ follows a generalized Riesz distribution with a scalar shape parameter, see Section 4.

Before we state the main result of this section, we prove the following technical proposition.

Proposition 5.1 *Let $\mathcal{D} = (V, E)$ be an acyclic directed graph without immoralities and let $\mathcal{U} = (V, E^{\sim})$ be its skeleton. Let $G \subseteq Perm(\mathcal{D})$ be a group of allowed symmetries. For any matrix $W \in P(\mathcal{U})$ and any node $m \in V$ we have*

$$\psi(W)_{[m]} = \frac{1}{|Orb(m)|} \sum_{v \in Orb(m)} \left(W_{[v]\bullet} + W_{[v\rangle\bullet} W_{\langle v\rangle} W_{[v\rangle\bullet}^T \right),$$

$$\psi(W)_{[m\rangle} = \frac{1}{|Orb(m)|} \sum_{v \in Orb(m)} W_{[v\rangle\bullet} W_{\langle v\rangle},$$

$$\psi(W)_{\langle m\rangle} = \frac{1}{|Orb(m)|} \sum_{v \in Orb(m)} W_{\langle v\rangle}.$$

Proof
The definition of the mapping ψ in equation (5) yields

$$\psi(W)_{[m]} = \frac{1}{|G|} \sum_{\rho \in G} (\rho W \rho^T)_{[m]} = \frac{1}{|G|} \sum_{\rho \in G} W_{[\sigma_\rho^{-1}(m)]},$$

where $\sigma_\rho = \varrho^{-1}(\rho)$ is the automorphism of $\mathcal{D}$ corresponding to the permutation matrix $\rho \in G$, see Section 2. In the second equality, we make use of the fact that $\rho_{[m][\sigma_\rho^{-1}(m)]}$ is the identity matrix for every $\rho \in Perm(\mathcal{D})$, see the definition of ϱ. Equation (2) yields

$$\psi(W)_{[m]} = \frac{1}{|G|} \sum_{v \in Orb(m)} \kappa_m W_{[v]} = \frac{1}{|Orb(m)|} \sum_{v \in Orb(m)} W_{[v]}.$$

Now, the first assertion follows from the definition of $W_{[m]\bullet}$ and $W_{[m\rangle\bullet}$. The remaining claims follow analogously. □

We are now in a position to prove the following theorem.

Theorem 5.2 *Let $\mathcal{D} = (V, E)$ be an acyclic directed graph without immoralities and let $\mathcal{U} = (V, E^{\sim})$ be its skeleton. Let $G \subseteq Perm(\mathcal{D})$ be a group of allowed symmetries. Let*

W follow a generalized Riesz distribution with expectation parameter $\Omega \in P(\mathcal{U}, G)$ and scalar shape parameter $\lambda = (\beta, \ldots, \beta)^T \in \mathbb{R}^V$ with $2\beta \in \mathbb{N}$ and $\beta > ([v] + \langle v \rangle - 1)/2$ for all $v \in V$.

Then for a maximal node $m \in V$ in $\mathcal{D}$ and the corresponding index set $I_{Orb} = I \backslash \cup_{v \in Orb(m)}[v]$ the following properties hold:

(i) $\psi(W)_{[m]\bullet} \perp\!\!\!\perp \left\{\psi(W)_{[m\rangle\bullet}, \psi(W)_{I_{Orb}}\right\}$,

(ii) $\mathcal{L}\left(\psi(W)_{[m]\bullet}\right) = \mathcal{W}_{[m]}\left(2\beta|Orb(m)| - \langle m\rangle, \frac{1}{2\beta|Orb(m)|}\Omega_{[m]\bullet}\right)$,

(iii) $\mathcal{L}\left(\psi(W)_{[m\rangle\bullet}|\psi(W)_{I_{Orb}}\right) = \mathcal{N}_{[m]\times\langle m\rangle}\left(\Omega_{[m\rangle\bullet}, \frac{1}{2\beta|Orb(m)|}\Omega_{[m]\bullet} \otimes \psi(W)^{-1}_{\langle m\rangle}\right)$.

Remark 5.3 Before proving the theorem, we comment on its content. Property (i) shows that the independence structure of $\psi(W)$ is inherited from W itself, see Proposition 3.1. Properties (ii) and (iii) state that the distributions of $\psi(W)_{[m]\bullet}$ and $\psi(W)_{[m\rangle\bullet}$ are Wishart and normal respectively, as is the case for $W_{[m]\bullet}$ and $W_{[m\rangle\bullet}$. The only difference is that the parameters of the distributions differ. While the shape parameter β of W does not depend on the node m, the transformed random matrix $\psi(W)$ has shape parameters depending on the concrete node m through the size of its orbit $|Orb(m)|$.

This gives a first impression on what happens to the distribution when W is transformed to $\psi(W)$: The family of distributions does not change, nor does the expectation parameter, since $\Omega \in P(\mathcal{U}, G)$ is invariant with respect to G. The shape parameter λ, on the other hand, does change since it reflects the size of the orbits under G.

Proof of Theorem 5.2
We organize the proof into five parts.

I. Let G' be the image of G under ϱ^{-1}. Since by definition ϱ maps an automorphism of $\mathcal{D}$ onto its corresponding permutation matrix, the image G' is a subgroup of $Aut(\mathcal{D})$. Let $v \in V$ be a node in the orbit of m under G'. Since every $\sigma \in G'$ is a automorphism of $\mathcal{D}$, we have $|[v]| = |[m]|$ as well as $|\langle v\rangle| = |\langle m\rangle|$ and v is maximal in $\mathcal{D}$, too. This observation together with Proposition 3.1 yields the following three properties.

$$W_{[v]\bullet} \perp\!\!\!\perp \left\{W_{[v\rangle\bullet}, W_{I\backslash[v]}\right\}, \tag{6}$$

$$\mathcal{L}\left(W_{[v]\bullet}\right) = \mathcal{W}_{[m]}\left(2\beta - \langle m\rangle, \frac{1}{2\beta}\Omega_{[m]\bullet}\right), \tag{7}$$

$$\mathcal{L}\left(W_{[v\rangle\bullet}|W_{I\backslash[v]}\right) = \mathcal{N}_{[m]\times\langle m\rangle}\left(\Omega_{[m\rangle\bullet}, \frac{1}{2\beta}\Omega_{[m]\bullet} \otimes W^{-1}_{\langle v\rangle}\right). \tag{8}$$

Here, we make use of the fact that $\Omega_{[v]\bullet} = \Omega_{[m]\bullet}$ and $\Omega_{[v\rangle\bullet} = \Omega_{[m\rangle\bullet}$ for all $v \in Orb(m)$, see equation (4). Because of equation (7), there exists a random matrix Y_v for every $v \in Orb(m)$ such that

$$W_{[v]\bullet} = Y_v^T Y_v \quad \text{and} \quad \mathcal{L}(Y_v) = \mathcal{N}_{2\beta-\langle m\rangle\times[m]}\left(0, I_{2\beta-\langle m\rangle} \otimes \frac{1}{2\beta}\Omega_{[m]\bullet}\right). \tag{9}$$

The assumption $\beta > ([m] + \langle m \rangle - 1)/2$ entails $2\beta - \langle m \rangle > 0$.

Next, observe that the set of parents $\mathrm{pa}(v)$ has to be complete in $\mathcal{D}$. Otherwise there would be an immorality in $\mathcal{D}$, contradicting the assumption. Consequently, the submatrix $W_{\langle v \rangle}$ is almost surely positive definite because the random matrix W is in $P(\mathcal{U})$ with probability one. We can therefore use a Cholesky decomposition to write $W_{\langle v \rangle} = \sqrt{W_{\langle v \rangle}}^T \sqrt{W_{\langle v \rangle}}$ for a matrix $\sqrt{W_{\langle v \rangle}} \in PD(\langle m \rangle)$. Define $Z_v := \sqrt{W_{\langle v \rangle}}(W_{[v]\bullet} - \Omega_{[m]\bullet})^T$. Using equation (8) and Theorem 2.2.2 from Kollo and von Rosen [10] we get

$$\mathcal{L}(Z_v | W_{I_{Orb}}) = \mathcal{N}_{\langle m \rangle \times [m]} \left(0, I_{\langle m \rangle} \otimes \frac{1}{2\beta} \Omega_{[m]\bullet} \right). \tag{10}$$

Let the nodes in the orbit of m be numbered according to $Orb(m) = \{v_1, \ldots, v_k\}$. Define the random matrices

$$X := \sqrt{\frac{2\beta}{|Orb(m)|}} \begin{pmatrix} Y_{v_1} \\ \vdots \\ Y_{v_k} \\ Z_{v_1} \\ \vdots \\ Z_{v_k} \end{pmatrix} \quad \text{and} \quad \mu := \sqrt{\frac{1}{|Orb(m)|}} \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \sqrt{W_{\langle v_1 \rangle}} \\ \vdots \\ \sqrt{W_{\langle v_k \rangle}} \end{pmatrix},$$

where 0 is the zero matrix of dimension $(2\beta - \langle m \rangle) \times \langle m \rangle$. We have $X \in \mathbb{R}^{2\beta |Orb(m)| \times [m]}$ and $\mu \in \mathbb{R}^{2\beta |Orb(m)| \times \langle m \rangle}$. Equation (6) yields $Y_{v_i} \perp\!\!\!\perp Y_{v_j}$ for $i \neq j$ and $Y_{v_i} \perp\!\!\!\perp Z_{v_j}$ for all $i, j = 1, \ldots, k$. Using equation (8) it is not hard to show that Z_{v_i} and Z_{v_j} are independent given $W_{I_{Orb}}$ for all $i \neq j$. Together, this makes all the submatrices appearing in the definition of X independent given $W_{I_{Orb}}$. Using elementary properties of the multivariate normal distribution, we can specify the conditional distribution of X,

$$\mathcal{L}(X | W_{I_{Orb}}) = \mathcal{N}_{2\beta |Orb(m)| \times [m]} \left(0, I_{2\beta |Orb(m)|} \otimes \frac{1}{|Orb(m)|} \Omega_{[m]\bullet} \right). \tag{11}$$

II. Using the definition of $\sqrt{W_{\langle v_i \rangle}}$, we get

$$\mu^T \mu = \frac{1}{|Orb(m)|} \sum_{i=1}^{k} \sqrt{W_{\langle v_i \rangle}}^T \sqrt{W_{\langle v_i \rangle}} = \psi(W)_{\langle m \rangle}, \tag{12}$$

see Proposition 5.1. In particular, $\mu^T \mu$ is regular with probability one. Proposition 5.1 and equation (10) imply

$$X^T \mu = \frac{\sqrt{2\beta}}{|Orb(m)|} \sum_{i=1}^{k} Z_{v_i}^T \sqrt{W_{\langle v_i \rangle}} = \sqrt{2\beta} \left(\psi(W)_{[m]} - \Omega_{[m]\bullet} \psi(W)_{\langle m \rangle} \right). \tag{13}$$

It follows that

$$\begin{aligned} X^T X &= \frac{2\beta}{|Orb(m)|} \sum_{i=1}^{k} \left\{ W_{[v_i]\bullet} + \left(W_{[v_i]\bullet} - \Omega_{[m]\bullet} \right) W_{\langle v_i \rangle} \left(W_{[v_i]\bullet} - \Omega_{[m]\bullet} \right)^T \right\} \\ &= 2\beta \left(\psi(W)_{[m]} - \psi(W)_{[m]\langle m \rangle} \Omega_{[m]\bullet}^T - \Omega_{[m]\bullet} \psi(W)_{[m]\langle m \rangle}^T + \Omega_{[m]\bullet} \psi(W)_{\langle m \rangle} \Omega_{[m]\bullet}^T \right), \end{aligned} \tag{14}$$

see Proposition 5.1 as well as the definition of X.

III. Since $\mu^T X$ is a linear transformation of the normally distributed random matrix X, we can specify its conditional distribution using equation (11),

$$\mathcal{L}(\mu^T X | W_{I_{Orb}}) = \mathcal{N}_{\langle m \rangle \times [m]} \left(0, \mu^T \mu \otimes \frac{1}{|Orb(m)|} \Omega_{[m]\bullet} \right).$$

From equations (12) and (13) it follows that the random matrix $\psi(W)_{[m\rangle\bullet}$ can also be written as $X^T \mu \psi(W)^{-1}_{\langle m \rangle} / \sqrt{2\beta} + \Omega_{[m\rangle\bullet}$. This yields the conditional distribution

$$\mathcal{L}\left(\psi(W)_{[m\rangle\bullet} | W_{I_{Orb}}\right) = \mathcal{N}_{[m]\times\langle m\rangle} \left(\Omega_{[m\rangle\bullet}, \frac{1}{2\beta|Orb(m)|} \Omega_{[m]\bullet} \otimes \psi(W)^{-1}_{\langle m \rangle} \right). \tag{15}$$

Since this distribution depends on $W_{I_{Orb}}$ only through a submatrix of $\psi(W)_{I_{Orb}}$, the distribution of $\psi(W)_{[m\rangle\bullet}$ given $\psi(W)_{I_{Orb}}$ is also given by the right-hand side of equation (15). This shows assertion (iii).

IV. Define the matrix $Q := I_{2\beta|Orb(m)|} - \mu(\mu^T\mu)^{-1}\mu^T$, that is symmetric and idempotent with rank $2\beta|Orb(m)| - \langle m \rangle$. Corollary 2.4.3.1 in Kollo and von Rosen [10] and equation (11) yield the distribution

$$\mathcal{L}\left(X^T Q X | W_{I_{Orb}}\right) = \mathcal{W}_{[m]} \left(2\beta|Orb(m)| - \langle m \rangle, \frac{1}{|Orb(m)|} \Omega_{[m]\bullet} \right).$$

Using equations (12), (13) and (14), it is not hard to show that $X^T Q X$ equals $2\beta\psi(W)_{[m]\bullet}$. Therefore, we get the distribution

$$\mathcal{L}\left(\psi(W)_{[m]\bullet} | W_{I_{Orb}}\right) = \mathcal{W}_{[m]} \left(2\beta|Orb(m)| - \langle m \rangle, \frac{1}{2\beta|Orb(m)|} \Omega_{[m]\bullet} \right). \tag{16}$$

This Wishart distribution does not depend on $W_{I_{Orb}}$ in any way. Consequently, the unconditional distribution of $\psi(W)_{[m]\bullet}$ is also given by the right-hand side of equation (16) and the assertion (ii) is proven.

V. It remains to be shown that $\{\psi(W)_{[m\rangle\bullet}, \psi(W)_{I_{Orb}}\}$ and $\psi(W)_{[m]\bullet}$ are independent. Note that equation (16) shows the independence of $\psi(W)_{[m]\bullet}$ and $W_{I_{Orb}}$. Since $\psi(W)_{I_{Orb}}$ is a mere transformation of $W_{I_{Orb}}$ it follows that $\psi(W)_{[m]\bullet} \perp\!\!\!\perp \psi(W)_{I_{Orb}}$.

By construction it holds that $\mu^T Q^T = 0 = \mu^T Q$. Theorem 2.2.4(iv) in Kollo and von Rosen [10] yields the conditional independence of $X^T Q X$ and $X^T \mu$ given $W_{I_{Orb}}$. The fact that given $W_{I_{Orb}}$, the matrices $\psi(W)_{[m]\bullet}$ and $\psi(W)_{[m\rangle\bullet}$ are linear transformations of $X^T Q X$ and $X^T \mu$ respectively, yields the conditional independence

$$\psi(W)_{[m]\bullet} \perp\!\!\!\perp \psi(W)_{[m\rangle\bullet} \mid W_{I_{Orb}}.$$

Since we already established that $\psi(W)_{[m]\bullet}$ and $W_{I_{Orb}}$ are independent, the unconditional independence of $\psi(W)_{[m]\bullet}$ and $\psi(w)_{[m\rangle\bullet}$ directly follows from the definition of conditional independence, see for example Lauritzen [11, pp. 28-29]. This proves assertion (i) and completes the proof. □

We may use Theorem 5.2 together with Proposition 3.1 to show that the estimator $\psi(W)$ follows a generalized Riesz distribution. However, since this approach needs an extensive formalism, we will not carry it out in the present paper. Instead, we illustrate the arising problems with the help of an example and refer to Käufl [12] for a general proof.

The main problem is that the parameter space $P(\mathcal{U}, G)$ in general is not the parameter space of a pure graphical model $P(\mathcal{U}')$ but a lower-dimensional subset of such a model. Therefore, there may be no well-defined generalized Riesz distribution with support $P(\mathcal{U}, G)$.

To see this, let $\mathcal{D} = 2 \leftarrow 1 \rightarrow 3$ and $\mathcal{U} = 2 - 1 - 3$. This allows the symmetry group $G = \{I_3, \rho\}$, where ρ is the permutation matrix corresponding to the permutation (132). The parameter space of the resulting invariant graphical model is

$$P(\mathcal{U}, G) = \left\{ \begin{pmatrix} \omega_{11} & \omega_{12} & \omega_{12} \\ \omega_{12} & \omega_{22} & 0 \\ \omega_{12} & 0 & \omega_{22} \end{pmatrix} \in \mathbb{R}^{3\times 3} \;\middle|\; \omega_{11}\omega_{22} - \omega_{12}^2 > 0 \right\}.$$

This is not a pure graphical model but only a lower-dimensional subset of the graphical model $P(\mathcal{U})$. The solution to this problem is to reparametrize $P(\mathcal{U}, G)$ into the set

$$\left\{ \begin{pmatrix} \omega_{11} & \omega_{12} \\ \omega_{12} & \omega_{22} \end{pmatrix} \in \mathbb{R}^{2\times 2} \;\middle|\; \omega_{11}\omega_{22} - \omega_{12}^2 > 0 \right\} = PD(2).$$

This set can be interpreted as the parameter space of the pure graphical model $P(\mathcal{U}')$ corresponding to the graph $\mathcal{U}' = 1 - 2$. It is now possible to reparametrize the maximum likelihood estimator $\psi(W)$ accordingly and to show that the reparametrized estimator follows a generalized Riesz distribution over $P(\mathcal{U}')$.

Even though this reparametrization idea also works in the general case, the required formalism is tedious and does not give any deeper insight. However, we want to stress that the change in the shape parameter mentioned in Remark 5.3 does have an important effect when trying to combine the distributions from Theorem 5.2 into a joint generalized Riesz distribution. The reason for this is that Remark 3.2 is not applicable anymore, since the new shape parameter is not scalar in general. As a consequence, one has to specify a certain acyclic directed graph $\mathcal{D}$ in order to characterize the distribution of $\psi(W)$ via a joint generalized Riesz distribution.

Instead of pursuing this approach any further, we demonstrate how to perform a likelihood ratio test between two nested invariant graphical models in the next section.

6 Likelihood ratio tests

The first task in this section is to specify to two acyclic directed graphs $\mathcal{D}_0$ and $\mathcal{D}$ with skeletons $\mathcal{U}_0$ and $\mathcal{U}$ as well as two groups G_0 and G such that the testing problem $PD(\mathcal{U}_0, G_0)$ against $PD(\mathcal{U}, G)$ is well-defined, that is, $PD(\mathcal{U}_0, G_0) \subseteq PD(\mathcal{U}, G)$.

To do so, as always let I be an index set. Let $\mathcal{D}_0 = (V_0, E_0)$ and $\mathcal{D} = (V, E)$ be two acyclic directed graphs such that I is partioned according to $\dot\cup_{v\in V}[v] = I = \dot\cup_{w\in V_0}[w]$.

Let $\varphi : V_0 \to V$ be a surjective mapping such that $[v] = \cup_{w:\varphi(w)=v}[w]$ and

$$\forall w, w' \in V_0 : w \to w' \in E_0 \Rightarrow \varphi(w) \to \varphi(w') \in E.$$

In Madsen [1], such a mapping is called a homomorphism between the graphs $\mathcal{D}_0$ and $\mathcal{D}$. Finally, let $\mathcal{U}_0$ and $\mathcal{U}$ be the skeletons of $\mathcal{D}_0$ and $\mathcal{D}$ respectively.

This construction implies that the two graphical models $PD(\mathcal{U}_0)$ and $PD(\mathcal{U})$ are nested, $PD(\mathcal{U}_0) \subseteq PD(\mathcal{U})$. Indeed, within a index set $[v]$ all interactions are allowed and that a missing directed edge corresponds to an additional independence constraint. By construction $\mathcal{D}$ has at least as many directed edges as has $\mathcal{D}_0$. Moreover, $\mathcal{D}_0$ has at least as many nodes as has $\mathcal{D}$, so that for $\mathcal{D}_0$ the variables in I are divided up into smaller subsets $[w]$. This means that the blocks in $\mathcal{D}_0$ allowing for all interactions contain fewer elements than the saturated blocks in $\mathcal{D}$.

To include symmetries in the models, let $G_0 \subseteq Perm(\mathcal{D}_0)$ and $G \subseteq Perm(\mathcal{D})$ be subgroups of allowed symmetries, such that $G \subseteq G_0$. It follows directly from definition 2.1 that the two invariant graphical models are nested, $PD(\mathcal{U}_0, G_0) \subseteq PD(\mathcal{U}, G)$, since G induces less symmetry restrictions on the covariance matrices than G_0 does.

The second task in this section is to state the likelihood ratio statistic explicitely and approximate its distribution. From now on, let $\mathcal{U}_0$, $\mathcal{U}$, $\mathcal{D}_0$, $\mathcal{D}$, G_0 and G be as described above. Let ψ_0 be the balancing function with respect to G_0 as defined in equation (5) and ψ be the balancing function with respect to G. Analogously, let π_0 and π be the projections on $S(\mathcal{U}_0)$ and $S(\mathcal{U})$ respectively, as defined in equation (3). The projections of the sample covariance matrix S are denoted by $W_0 = \pi_0(S)$ and $W = \pi(S)$. Moreover, assume that the maximum likelihood estimator $\psi_0(W_0) \in P(\mathcal{U}_0, G_0)$ in the smaller model exists. This implies the existence of the likelihood ratio test statistic

$$Q = \left(\frac{\prod_{v \in V} \det[\psi(W)_{[v]\bullet}]}{\prod_{w \in V_0} \det[\psi_0(W_0)_{[w]\bullet}]} \right)^{\frac{n}{2}},$$

see Madsen [1, p. 1177]. Caution is advised when interpreting the Cholesky parameters present in this equation. The definition of these parameters implicitly depends on the causal structure of an acyclic directed graph, see chapter 2. For $\psi(W)_{[v]\bullet}$ this corresponding graph is $\mathcal{D}$, while $\psi_0(W_0)_{[w]\bullet}$ is defined with respect to the graph $\mathcal{D}_0$. In order to keep this difference in mind, we continue to use the index "$[v]\bullet$" for nodes in $\mathcal{D}$ and the index "$[w]\bullet$" for nodes in $\mathcal{D}_0$.

To simplify the statistic Q, we can make use of the fact that the Cholesky parameters of a matrix $\Omega \in P(\mathcal{U}, G)$ are identical for all nodes in the same orbit, see equation (4). For that reason, let $V_G \subseteq V$ be a subset that includes exactly one node of every orbit of V under G. Analogously, let $V_{G_0} \subseteq V_0$ include exactly one node of every orbit of V_0 under G_0. The statistic Q can then be reformulated as

$$Q = \left(\frac{\prod_{v \in V_G} \det[\psi(W)_{[v]\bullet}]^{|Orb(v)|}}{\prod_{w \in V_{G_0}} \det[\psi_0(W_0)_{[w]\bullet}]^{|Orb(w)|}} \right)^{\frac{n}{2}}. \tag{17}$$

Now, we compute the higher moments of Q. This enables us to prove that the deviance $-2\log(Q)$ follows a Box-type distribution under the null hypothesis. The following proposition will be useful in this process.

Proposition 6.1 *Let $\mathcal{D} = (V, E)$ be an acyclic directed graph without immoralities and let $\mathcal{U} = (V, E^{\sim})$ be its skeleton. Let $G \subseteq Perm(\mathcal{D})$ be a group of allowed symmetries. Let $X_1, \ldots, X_n \sim \mathcal{N}_I(0, \Sigma)$ be independent for a $\Sigma \in PD(\mathcal{U}, G)$ with $n \geq [v] + \langle v \rangle$ for all $v \in V$. Again, we use the notation $\Omega = \pi(\Sigma) \in P(\mathcal{U}, G)$ and $W = \pi(S)$ where S is the sample covariance matrix and π the projection from equation (3).*

Then for all $\delta > \max_{v \in V_G} \left\{ \frac{[v]+\langle v \rangle - 1}{2|Orb(v)|} - \frac{n}{2} \right\}$, the following equality holds true:

$$\mathbb{E}\left[\prod_{v \in V_G} \det(\psi(W)_{[v]\bullet})^{\delta |Orb(v)|} \right]$$
$$= \frac{2^{\delta |I|} \det(\Sigma)^{\delta}}{n^{\delta |I|}} \prod_{v \in V_G} \left\{ |Orb(v)|^{-\delta [v] |Orb(v)|} \prod_{i=1}^{[v]} \frac{\Gamma\left(\left(\delta + \frac{n}{2}\right) |Orb(v)| - \frac{\langle v \rangle + i - 1}{2} \right)}{\Gamma\left(\frac{n}{2} |Orb(v)| - \frac{\langle v \rangle + i - 1}{2} \right)} \right\}$$

Proof
By assumption, W follows a generalized Riesz distribution with expectation parameter Ω and scalar shape parameter $\lambda = (n/2, \ldots, n/2)^T \in \mathbb{R}^V$, see Andersson and Klein [8, Example 17.1]. Theorem 5.2 then implies that the random matrices $\{\psi(W)_{[v]\bullet}\}_{v \in V_G}$ are all independent and follow Wishart distributions. This yields

$$\mathbb{E}\left[\prod_{v \in V_G} \det(\psi(W)_{[v]\bullet})^{\delta |Orb(v)|} \right] = \prod_{v \in V_G} \mathbb{E}\left[\det(W_v)^{\delta |Orb(v)|} \right] =: \prod_{v \in V_G} E_v, \qquad (18)$$

where W_v is Wishart distributed with $n|Orb(v)| - \langle v \rangle$ degrees of freedom and parameter matrix $\Omega_{[v]\bullet}/(n|Orb(v)|)$. Therefore, the problem of computing the expected value above simplifies to specifying the higher moments of the determinant of a Wishart matrix.

Using Corollary 2.4.4.1 in Kollo and von Rosen [10], it is not hard to show that

$$E_v = \left(\frac{2}{n|Orb(v)|} \right)^{\delta [v] |Orb(v)|} \det(\Omega_{[v]\bullet})^{\delta |Orb(v)|} \prod_{i=1}^{[v]} \frac{\Gamma\left(\delta |Orb(v)| - \frac{n|Orb(v)| - \langle v \rangle - i + 1}{2} \right)}{\Gamma\left(\frac{n|Orb(v)| - \langle v \rangle - i + 1}{2} \right)}$$

holds true for every $v \in V_G$. Also note that from equation (4) and Corollary 5.3 in Andersson and Klein [8] it follows that

$$\prod_{v \in V_G} \det(\Omega_{[v]\bullet})^{|Orb(v)|} = \prod_{v \in V} \det(\Omega_{[v]\bullet}) = \det\left[\pi^{-1}(\Omega) \right] = \det(\Sigma).$$

Together with the fact that $\sum_{v \in V_G} [v] |Orb(v)| = |I|$, this shows the assertion. Note that the restriction on δ is needed to prevent the argument of the Gamma function to become negative. □

The main step in the proof of Proposition 6.1 is to use the independence structure given by Theorem 5.2 in equation (18). The application of this result to the likelihood ratio statistic Q yields the following theorem.

Theorem 6.2 *Let $\mathcal{D}_0 = (V_0, E_0)$, $\mathcal{D} = (V, E)$, $\mathcal{U}_0$, $\mathcal{U}$, G_0 and G be as described above, so that $PD(\mathcal{U}_0, G_0) \subseteq PD(\mathcal{U}, G)$. Moreover, let $X_1, \dots, X_n \sim \mathcal{N}_I(0, \Sigma)$ be independent for a $\Sigma \in PD(\mathcal{U}, G)$. Also, let $n \geq [v] + \langle v \rangle$ for all $v \in V_0 \cup V$, so that the likelihood ratio statistic Q from equation (17) exists almost surely. Then under the null hypothesis, the identity*

$$\mathbb{E}\left[Q^t\right] = \frac{\prod_{v \in V_G} \left\{ |Orb(v)|^{-\frac{n}{2}t[v]|Orb(v)|} \prod_{i=1}^{[v]} \frac{\Gamma\left(\frac{n}{2}|Orb(v)|(t+1) - \frac{\langle v \rangle + i - 1}{2}\right)}{\Gamma\left(\frac{n}{2}|Orb(v)| - \frac{\langle v \rangle + i - 1}{2}\right)} \right\}}{\prod_{w \in V_{G_0}} \left\{ |Orb(w)|^{-\frac{n}{2}t[w]|Orb(w)|} \prod_{j=1}^{[w]} \frac{\Gamma\left(\frac{n}{2}|Orb(w)|(t+1) - \frac{\langle w \rangle + j - 1}{2}\right)}{\Gamma\left(\frac{n}{2}|Orb(w)| - \frac{\langle w \rangle + j - 1}{2}\right)} \right\}}$$

holds true for every $t > \max\left\{ \frac{[v] + \langle v \rangle - 1}{n|Orb(v)|} - 1 \;\middle|\; v \in V_0 \cup V \right\}$.

Proof
Let $\delta \in \mathbb{R}$. Under the null hypothesis, the likelihood ratio statistic Q and the Cholesky parameters $\{\psi_0(W_0)_{[w]\bullet}\}_{w \in V_0}$ of the maximum likelihood estimator $\hat{\Sigma}_0 \in PD(\mathcal{U}_0, G_0)$ are independent, see Madsen [1, Section 8]. Together with the description of Q in equation (17) this yields the identity

$$\mathbb{E}\left[Q^{\frac{2\delta}{n}}\right] = \frac{\mathbb{E}\left[\prod_{v \in V_G} \det(\psi(W)_{[v]\bullet})^{\delta|Orb(v)|}\right]}{\mathbb{E}\left[\prod_{w \in V_{G_0}} \det(\psi_0(W_0)_{[w]\bullet})^{\delta|Orb(w)|}\right]}.$$

Both the numerator and the denominator can be simplified by using Proposition 6.1. The substitution $t := 2\delta/n$ then shows the assertion. □

This theorem proves that the distribution of the deviance $-2\log(Q)$ under the null hypothesis lies in the family of Box-type distributions as introduced by Box [13]. For these distributions, very accurate approximation methods are available as described for example in Jensen [14]. We can therefore perform a likelihood ratio test by evaluating $-2\log(Q)$ with respect to its approximated distribution.

7 Conclusions

In Section 5 we showed that the maximum likelihood estimator in an invariant graphical model follows a generalized Riesz distribution. In particular, we made clear that in general this distribution has a non-scalar shape parameter λ. This retrospectively motivates the introduction of a multivariate shape parameter by Andersson and Klein [8].

Also, this result proves that a certain independence structure is present in the Cholesky parameters of the maximum likelihood estimator. With the help of this observation, we

showed in Section 6 that the likelihood ratio statistic follows a Box-type distribution under the null hypothesis. This enables us to perform a likelihood ratio test by using the accurate approximation methods that have been developed for Box-type distributions.

Finally, we want to stress that throughout this paper we are concerned with graphical models that implement conditional independence restrictions. Since these constraints are reflected as zeros in the concentration matrix, we could term these models "'graphical concentration models"'.

In recent years, research has increasingly been concerned with "'graphical covariance models"'. These models use graphs to encode marginal independencies, that is, zeros in the covariance matrix. An example of the ongoing work in this field is the paper of Drton et al. [15].

In general, graphical covariance models are not regular but curved exponential families. This makes them harder to analyze than the models examined in the present paper. It may still be worthwile to include symmetries in these models, too, in order to reduce the number of parameters. Shah and Chandrasekaran [16] take a first step in this direction and illustrate their results by numerical studies. It may then be possible to describe the exact distribution of the maximum likelihood estimator in these models as well. It seems likely that for this purpose, a new generalization of the classical Wishart distribution concentrated on a graphical covariance model is needed. In this context, we want to mention the article of Khare and Rajaratnam [17], where such a candidate is developed.

Acknowledgements

I am very grateful to Friedrich Pukelsheim and Mathias Drton for carefully reading the manuscript and giving helpful remarks.

References

[1] J. Madsen, Invariant Normal Models With Recursive Graphical Markov Structure, Annals of Statistics 28 (2000) 1150–1178.

[2] S. A. Andersson, Invariant Normal Models, Annals of Statistics 3 (1975) 132–154.

[3] M. D. Perlman, Comment: Group Symmetry Covariance Models, Statistical Science 2 (1987) 421–425.

[4] S. A. Andersson, J. Madsen, Symmetry and Lattice Conditional Independence Models in a Multivariate Normal Distribution, Annals of Statistics 26 (1998) 525–572.

[5] P. Dawid, S. Lauritzen, Hyper Markov Laws in the Statistical Analysis of Decomposable Graphical Models, Annals of Statistics 21 (1993) 1272–1317.

[6] S. A. Andersson, G. G. Wojnar, Wishart Distributions on Homogeneous Cones, Journal of Theoretical Probability 17 (2004) 781–818.

[7] G. Letac, H. Massam, Wishart Distributions for Decomposable Graphs, Annals of Statistics 35 (2007) 1278–1323.

[8] S. A. Andersson, T. Klein, On Riesz and Wishart Distributions Associated with Decomposable Undirected Graphs, Journal of Multivariate Analysis 101 (2010) 789–810.

[9] S. A. Andersson, D. Madigan, M. D. Perlman, On the Markov Equivalence of Chain Graphs, Undirected Graphs, and Acyclic Digraphs, Scandinavian Journal of Statistics 24 (1997) 81–102.

[10] T. Kollo, D. von Rosen, Advanced Multivariate Statistics with Matrices, Springer, Dordrecht, 2005.

[11] S. Lauritzen, Graphical Models, Oxford University Press, New York, 1996.

[12] A. Käufl, Statistische Inferenz in invarianten graphischen Modellen mit Normalverteilungsannahme, Ph.D. thesis, Augsburg University, 2012.

[13] G. E. P. Box, A General Distribution Theory for a Class of Likelihood Criteria, Biometrika 36 (1949) 317–346.

[14] J. L. Jensen, A Large Deviation-Type Approximation for the Box Class of Likelihood Ratio Criteria, Journal of the American Statistical Association 86 (1991) 437–440.

[15] M. Drton, R. Foygel, S. Sullivant, Global identifiability of linear structural equation models, Annals of Statistics 39 (2011) 865–886.

[16] P. Shah, V. Chandrasekaran, Group Symmetry and Covariance Regularization, Electronic Journal of Statistics 6 (2012) 1600–1640.

[17] K. Khare, B. Rajaratnam, Wishart Distributions for Decomposable Covariance Graph Models, Annals of Statistics 39 (2011) 514–555.