Graphical Gaussian Models with Edge and Vertex Symmetries

Søren Højsgaard
Aarhus University, Denmark
Steffen L. Lauritzen
University of Oxford, United Kingdom

Summary. In this paper we introduce new types of graphical Gaussian models by placing symmetry restrictions on the concentration or correlation matrix. The models can be represented by coloured graphs, where parameters associated with edges or vertices of same colour are restricted to being identical. We study the properties of such models and derive the necessary algorithms for calculating maximum likelihood estimates. We identify conditions for restrictions on the concentration and correlation matrices being equivalent. This is for example the case when symmetries are generated by permutation of variable labels. For such models a particularly simple maximization of the likelihood function is available.

Keywords: conditional independence; covariance selection; invariance; iterative partial maximization; patterned covariance matrices; permutation symmetry; transformation models.

1. Introduction

This paper introduces new types of graphical Gaussian models (Whittaker, 1990; Lauritzen, 1996), also known as covariance selection models (Dempster, 1972), by imposing symmetry restrictions to the concentration matrix, i.e. the inverse of the covariance matrix. Following up on Højsgaard and Lauritzen (2005) we introduce three types of restriction on graphical Gaussian models: equality among specified elements of the concentration matrix (RCON), equality among specific partial variances and correlations (RCOR), and restrictions generated by permutation symmetry (RCOP). Adding symmetry to the conditional independence restrictions of a graphical model reduces the number of parameters, useful when parsimony is needed, for example when estimating covariance matrices of large dimension with relatively few observations. The symmetry restrictions are represented by colouring of edges and vertices of the dependence graph of the model.

In the present article we introduce the model types and investigate their fundamental properties and mutual relationships whereas details of the estimation algorithms and their implementation in software are described in Højsgaard and Lauritzen (2007).

Symmetry restrictions in the multivariate Gaussian distribution have a long history (Wilks, 1946; Votaw, 1948; Olkin and Press, 1969; Andersson, 1975; Andersson et al., 1983). Such restrictions have previously been combined with conditional independence restrictions, for example in the circular stationary Markov process (Anderson, 1942; Leipnik, 1947), in

Address for correspondence: Steffen L. Lauritzen, Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, United Kingdom.
E-mail: steffen@stats.ox.ac.uk

Table 1. Empirical concentrations $\times 1000$ (on or above the diagonal) and partial correlations (below the diagonal) for the examination marks in five mathematical subjects.

	Mechanics	Vectors	Algebra	Analysis	Statistics
Mechanics	5.24	−2.44	−2.74	0.01	−0.14
Vectors	0.33	10.43	−4.71	−0.79	−0.17
Algebra	0.23	0.28	26.95	−7.05	−4.70
Analysis	−0.00	0.08	0.43	9.88	−2.02
Statistics	0.02	0.02	0.36	0.25	6.45

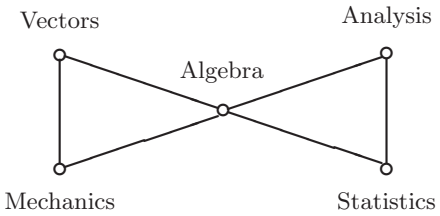

Fig. 1. Conditional independence structure of examination marks for 88 students.

spatial models based on Markov random fields (Whittle, 1954; Besag, 1974; Besag and Moran, 1975), and in combination with other graphical models and lattice independence models (Hylleberg et al., 1993; Andersson and Madsen, 1998; Madsen, 2000). Symmetry restrictions similar to those studied here are well established in the form of stationarity assumptions for time series and longitudinal data analysis.

To illustrate the different types of model considered in this article, we shall discuss three simple examples based on data which are all well known from the literature. Initially we describe the examples briefly and then revisit them later.

Mathematics marks The first example is concerned with the examination marks of 88 students in 5 different mathematical subjects. The inverse of the empirical covariance matrix is displayed in Table 1.

The data were reported in Mardia et al. (1979) and also analyzed in Whittaker (1990) as well as in Edwards (2000). Both of the latter references demonstrate an excellent fit to the model displayed in Figure 1, where absence of an edge between two variables indicates that these are conditionally independent given the remaining variables, implying that marks in Vectors and Mechanics are conditionally independent of marks in Analysis and Statistics, given the marks in Algebra. Most off-diagonal elements of the empirical concentration matrix which are not close to zero are almost equal. Indeed the data support a model with symmetry restrictions as in Figure 2, where elements of the concentration matrix corresponding to the same colours are identical, i.e. the diagonal concentrations of Vectors and Analysis are equal as indicated in green colour (or a single asterisk), as are those of Mechanics and Statistics, in blue with a double asterisk; the mixed concentration between Algebra and Analysis is different from the others, all of which have the same value, indicated with red colour and a plus sign. Here and in the following, black or white colours are considered neutral, the corresponding parameters being allowed to vary freely. A symmetry

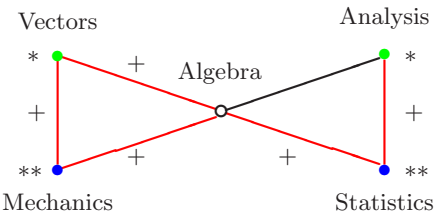

Fig. 2. Coloured graph of an RCON symmetry model for the examination marks of 88 students. Elements of the inverse concentration matrix corresponding to edges or vertices of the same colour are identical.

Table 2. Empirical concentrations ($\times 100$) (on and above diagonal), partial correlations (below diagonal), and standard deviations for personality characteristics of 684 students.

	SX	SN	TX	TN
SX (State anxiety)	0.58	−0.30	−0.23	0.02
SN (State anger)	0.45	0.79	−0.02	−0.15
TX (Trait anxiety)	0.47	0.03	0.41	−0.11
TN (Trait anger)	−0.04	0.33	0.31	0.27
Standard deviations	6.10	6.70	5.68	6.57

model of this type shall be called an *RCON model*, reflecting that the symmetry restrictions apply to the elements of the concentration matrix.

Anxiety and anger Cox and Wermuth (1993) report data on personality characteristics on 684 students. The data are concerned with anxiety and anger in a trait and state version, the trait version reflecting a disposition and the state version a more momentary concept. Each characteristic is measured on the basis of questionnaires developed by Spielberger et al. (1970, 1983), eventually leading to quantitative scores for the characteristics. Empirical concentrations and partial correlations are displayed in Table 2. As also argued by Cox and Wermuth (1993), data strongly support the conditional independence model displayed in Figure 3.

In addition, the partial correlations are strikingly similar in pairs as illustrated by the graph colouring in Figure 4.

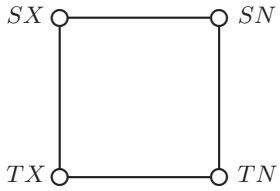

Fig. 3. Conditional independence structure for data on personality characteristics of 684 students.

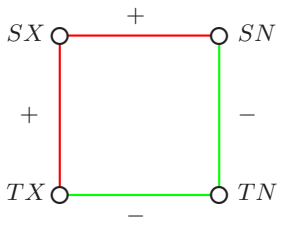

Fig. 4. Graph colouring for the RCOR model describing data on personality characteristics of 684 students, restricting *partial correlations* corresponding to edges of same colour to being identical.

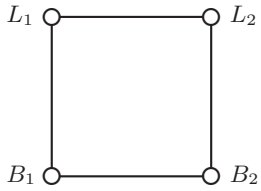

Fig. 5. Conditional independence structure for data on Frets' heads. The length and breadth of the head of the ith son are denoted L_i and B_i.

For psychological measurements such as these, the scales for the variables may not be compatible and the partial correlations may therefore be more meaningful than the concentrations. The partial correlations are invariant under changes of scale for the individual variables. Symmetry models based on identity of partial correlations are denoted RCOR models.

Frets' heads This example is concerned with data reported in Mardia et al. (1979) based on a study of heredity of head dimensions (Frets, 1921). Graphical models for these were also considered e.g. by Whittaker (1990). Length and breadth of the heads of 25 pairs of first and second sons are measured. Previous analyses of these data support a model with conditional independence relations as in Figure 5.

There is an obvious symmetry between the two sons. Hence it makes sense to investigate a model where the joint distribution is unaltered if the two sons are interchanged. A symmetry model of this type shall be called an *RCOP model*, as the symmetry restrictions are determined by permutation of the variable labels. This symmetry can be illustrated in the graph by adding colour to the graph as illustrated in Figure 6. Since the restrictions are generated by permutation symmetry, it does not matter whether we interpret the figure in RCON or RCOR terms, because all aspects of the joint distribution are unaltered when the variable labels are switched, implying that the model is both RCON and RCOR.

Another symmetry model In general, RCON models which restrict concentrations are different from RCOR models with the same coloured graph, but there are also other cases where a certain pattern of restriction on concentrations implies the same pattern of restrictions on the partial correlations and *vice versa*. As we shall argue later in the article, the

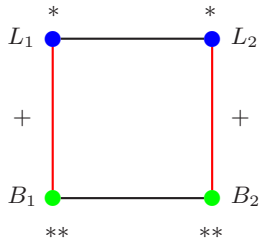

Fig. 6. Symmetry restrictions on the graphical model for Frets' heads. The first and second sons can be interchanged without altering the distribution, illustrated through graph colouring.

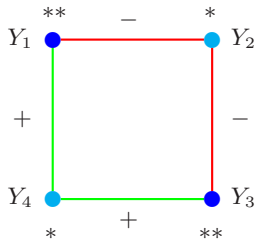

Fig. 7. Restrictions on partial correlations and variances according to this graph colouring identify the same pattern of restriction on elements of the concentration matrix and *vice versa*.

model displayed in Figure 7 is of this type although it is not generated by permutation symmetry.

2. Preliminaries and Notation

2.1. Graph colouring

For general graph terminology we refer to Lauritzen (1996) or Bollobás (1998). Consider an undirected graph $\mathcal{G} = (V, E)$. Colouring the vertices of $\mathcal{G}$ with $R \leq |V|$ different colours induces a partitioning of V into disjoint sets $V_1, \ldots, V_R$ called *vertex colour classes,* where all vertices in V_r have the same colour. Similarly, colouring the edges E with $S \leq |E|$ different colours partitions E into disjoint sets $E_1, \ldots, E_S$ called *edge colour classes,* where all edges in E_s have the same colour. We say that $\mathcal{V} = \{V_1, \ldots, V_R\}$ is a *vertex colouring,* $\mathcal{E} = \{E_1, \ldots, E_S\}$ is an *edge colouring,* and $(\mathcal{V}, \mathcal{E})$ is a *coloured graph,* referring implicitly to the induced graph $\mathcal{G} = (V, E)$ with

$$V = V_1 \cup \cdots \cup V_R, \quad E = E_1 \cup \cdots \cup E_S.$$

Note that our use of colouring differs from the standard in graph theory (Bollobás, 1998), where a colouring also satisfies that adjacent vertices as well as incident edges are of different colour, whereas we do not impose this restriction.

A colour class with a single element is called *atomic* and a colour class which is not atomic is *composite.* When visualising a coloured graph, atomic colour classes are displayed with neutral colours, i.e. black or white, whereas we apply other colours to elements of composite colour classes.

2.2. Graphical Gaussian models

Graphical Gaussian models are concerned with the distribution of a multivariate random vector $Y = (Y_\alpha)_{\alpha \in V}$ following a $\mathcal{N}_{|V|}(\mu, \Sigma)$ distribution. For simplicity of exposition we assume throughout that $\mu = 0$. In the following we use Greek letters to refer to single variables and Latin letters to refer to sets of variables. We let $K = \Sigma^{-1}$ denote the inverse covariance, also known as the *concentration* with elements $(k_{\alpha\beta})_{\alpha,\beta \in V}$. The partial correlation between Y_α and Y_β given all other variables is then

$$\rho_{\alpha\beta \,|\, V \setminus \{\alpha,\beta\}} = -k_{\alpha\beta} / \sqrt{k_{\alpha\alpha} k_{\beta\beta}}. \tag{1}$$

Thus $k_{\alpha\beta} = 0$ if and only if Y_α and Y_β are conditionally independent given all other variables.

A *graphical Gaussian model* is represented by an undirected graph $\mathcal{G} = (V, E)$ where V is a set of vertices representing the variables and E is a set of undirected edges. The graph represents the model with $K \in \mathcal{S}^+(\mathcal{G})$, the set of (symmetric) positive definite matrices which have zero elements $k_{\alpha\beta}$ whenever there is no edge between α and β in $\mathcal{G}$, i.e. partial correlations between non-neighbours in the graph are equal to zero.

For later use we recall that if we partition Y into $Y_a = (Y_\gamma, \gamma \in a)$ and $Y_b = (Y_\gamma, \gamma \in b)$, where $V = a \cup b$ and $a \cap b = \emptyset$, and partition concentration and covariance accordingly as

$$K = \begin{pmatrix} K_{aa} & K_{ab} \\ K_{ba} & K_{bb} \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix},$$

then

$$Y_b \sim \mathcal{N}_{|b|}(0, \Sigma_{bb}), \quad Y_a \,|\, Y_b = y_b \sim \mathcal{N}_{|a|}(\mu_{a|b}, \Sigma_{a|b}),$$

where

$$\mu_{a|b} = \Sigma_{ab}\Sigma_{bb}^{-1} y_b, \quad \Sigma_{a|b} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}. \tag{2}$$

Indeed it also holds that

$$K_{aa}^{-1} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba} \tag{3}$$

so that K_{aa} is equal to the concentration matrix $K_{a|b}$ for the conditional distribution of Y_a given $Y_b = y_b$, and

$$K_{aa}^{-1} K_{ab} = -\Sigma_{ab}\Sigma_{bb}^{-1}, \tag{4}$$

so then

$$\mu_{a|b} = -K_{aa}^{-1} K_{ab} y_b.$$

Consider next a sample $Y^1 = y^1, \ldots, Y^n = y^n$ of n observations of Y and let W denote the matrix of sums of squares and products

$$W = \sum_{\nu=1}^{n} Y^\nu (Y^\nu)^\top.$$

The log-likelihood function based on the sample is

$$\log L = \frac{f}{2} \log\det(K) - \frac{1}{2} \operatorname{tr}(KW), \tag{5}$$

where in this case $f = n$ is the degrees of freedom in the Wishart distribution of W. Taking into account a possible unknown mean μ and calculating W based on residuals would yield degrees of freedom $f = n - 1$.

For the facts above and additional properties of graphical Gaussian models we refer to Lauritzen (1996), Chapter 5 and Appendix C.

3. Symmetry restrictions on concentrations

3.1. Model specification

An $RCON(\mathcal{V},\mathcal{E})$ model with vertex colouring $\mathcal{V}$ and edge colouring $\mathcal{E}$ is obtained by restricting the elements of the concentration matrix K as follows:

(a) Diagonal elements of the concentration matrix K (inverse partial variances) corresponding to vertices in the same vertex colour class must be identical.

(b) Off–diagonal entries of K corresponding to edges in the same edge colour class must be identical.

Thus, the diagonal of K can be specified by an R-dimensional vector η while the off–diagonal elements are given by an S-dimensional vector δ so we can write $K = K(\eta,\delta)$. The set of positive definite matrices which satisfy these restrictions is denoted $\mathcal{S}^+(\mathcal{V},\mathcal{E})$.

Note that the restrictions so defined are linear in the concentration matrix and thus RCON models are instances of models considered by Anderson (1969, 1970, 1973).

Example 1. Consider the graph in Figure 7. The corresponding RCON model will have concentration matrix of the form

$$K = \begin{pmatrix} k_{11} & k_{12} & 0 & k_{14} \\ k_{21} & k_{22} & k_{23} & 0 \\ 0 & k_{32} & k_{33} & k_{34} \\ k_{41} & 0 & k_{43} & k_{44} \end{pmatrix} = \begin{pmatrix} \eta_1 & \delta_1 & 0 & \delta_2 \\ \delta_1 & \eta_2 & \delta_1 & 0 \\ 0 & \delta_1 & \eta_1 & \delta_2 \\ \delta_2 & 0 & \delta_2 & \eta_2 \end{pmatrix},$$

so that elements corresponding to edges or vertices in the same colour class are identical.

Note that if we calculate the regression of Y_2 on its neighbours, we get from (2) and (4) that

$$\mathbf{E}(Y_2 \,|\, Y_1 = y_1, Y_3 = y_3) = -(k_{21}/k_{22})y_1 - (k_{23}/k_{22})y_3 = -(\delta_1/\eta_2)y_1 - (\delta_1/\eta_2)y_3,$$

so that Y_3 and Y_1 contribute equally to the prediction of Y_2 and, by symmetry, also equally to the prediction of Y_4. In addition, the regressions for Y_1 and Y_3 on Y_2 and Y_4 are parallel:

$$\begin{aligned} \mathbf{E}\left\{ \begin{pmatrix} Y_1 \\ Y_3 \end{pmatrix} \middle| Y_2 = y_2, Y_4 = y_4 \right\} &= \begin{pmatrix} -(k_{12}/k_{11})y_2 - (k_{14}/k_{11})y_4 \\ -(k_{32}/k_{33})y_2 - (k_{34}/k_{33})y_4 \end{pmatrix} \\ &= \begin{pmatrix} -(\delta_1/\eta_1)y_2 - (\delta_2/\eta_1)y_4 \\ -(\delta_1/\eta_1)y_2 - (\delta_2/\eta_1)y_4 \end{pmatrix}. \end{aligned}$$

Indeed, as we are treating the case $\mu = 0$, the regressions are not just parallel, but identical. However, in the more general case they will only be parallel.

Conversely it holds for an arbitrary concentration matrix that if

(i) The concentration matrix has $k_{13} = k_{24} = 0$;

(ii) Y_3 and Y_1 contribute equally to the prediction of Y_2 and to the prediction of Y_4;

(iii) the regressions for Y_1 and Y_3 on Y_2 and Y_4 are parallel,

then the concentration matrix satisfies the RCON model specified.

3.2. Likelihood equations

Since the restrictions are linear in the concentration matrix, an RCON model is a linear exponential model and the maximum likelihood estimates of the unknown parameters are uniquely determined by equating the canonical sufficient statistics to their expectation.

More specifically, for each vertex colour class $u \in \mathcal{V}$ of an RCON model $RCON(\mathcal{V}, \mathcal{E})$ let T^u be the $|V| \times |V|$ diagonal matrix with entries $T^u_{\alpha\alpha} = 1$ if $\alpha \in u$ and 0 otherwise. Similarly, for each edge colour class $u \in \mathcal{E}$ we let T^u be the $|V| \times |V|$ symmetric matrix with entries $T^u_{\alpha\beta} = 1$ if $\{\alpha, \beta\} \in u$ and 0 otherwise; hence T^u is the adjacency matrix of the edge colour class u.

We can now freely refer to a generator u for $RCON(\mathcal{V}, \mathcal{E})$ without specifying whether u refers to a vertex colour class or an edge colour class. If we rewrite (η, δ) as an $R + S$-dimensional vector θ, the concentration matrix $K = K(\theta)$ can be written $K = \sum_u \theta_u T^u$ and thus $\operatorname{tr}(KW) = \sum_u \theta_u \operatorname{tr}(T^u W)$.

If we let $t^u = \operatorname{tr}(T^u W)$, the canonical statistics are $(-t^1/2, \ldots, -t^{R+S}/2)$. The first two derivatives of the log-normalizing constant $(-f/2) \log\det(K)$ yield expectation and covariance of these statistics. It holds that

$$\frac{\partial}{\partial\theta_u} \log\det(K) = \operatorname{tr}(T^u\Sigma); \quad \frac{\partial^2}{\partial\theta_u\partial\theta_v} \log\det(K) = -\operatorname{tr}(T^u\Sigma T^v\Sigma), \tag{6}$$

which either can be derived by direct differentiation or using the moment relations in Proposition C.10 of Lauritzen (1996).

Further, the maximum likelihood estimate is the unique solution to the system of equations obtained by equating the canonical statistics to their expectation, yielding

$$\operatorname{tr}(T^u W) = f \operatorname{tr}(T^u\Sigma), \quad u \in \mathcal{V} \cup \mathcal{E}, \tag{7}$$

provided such a solution exists. The question of existence is in general non-trivial. Without symmetry restrictions, the existence is ensured with probability one if f is as least as large as the size of the largest clique in a triangulated cover of the graph (Lauritzen 1996, page 148) whereas a necessary condition is not known in the general case (Buhl, 1993).

It is an important motivation for considering these models that the conditions for maximum likelihood estimates to exist are less restrictive than for graphical Gaussian models without symmetries; typically fewer observations are needed to guarantee existence of the estimate. For example, the Gaussian graphical model in Fig. 5 demands at least $f = 3$ observations of the four-dimensional vector for the maximum likelihood estimates to exist with probability one. However, if we add symmetry restrictions by collecting all vertices into a single vertex colour class and all present edges into a single edge colour class, the maximum likelihood estimate exists with just $f = 1$ observation and is given explicitly as

$$\hat\sigma_{11} = \hat\sigma_{22} = \hat\sigma_{33} = \hat\sigma_{44} = (y_1^2 + y_2^2 + y_3^2 + y_4^2)/4,$$

$$\hat\sigma_{12} = \hat\sigma_{23} = \hat\sigma_{34} = \hat\sigma_{41} = (y_1y_2 + y_2y_3 + y_3y_4 + y_4y_1)/4,$$

$$\hat\sigma_{13} = \hat\sigma_{24} = (\sqrt{1 + 8r^2} - 1)/2,$$

where

$$r = (y_1y_2 + y_2y_3 + y_3y_4 + y_4y_1)/(y_1^2 + y_2^2 + y_3^2 + y_4^2).$$

This particular model is simultaneously RCON, RCOR, and RCOP; hence the simplicity of the estimate.

It follows from (7) that for RCON models we have

$$\operatorname{tr}(\hat{K}W) = \sum_u \hat{\theta}_u \operatorname{tr}(T^u W) = \sum_u \hat{\theta}_u f \operatorname{tr}(T^u \hat{\Sigma}) = f \operatorname{tr}(\hat{K}\hat{\Sigma}) = f|V|,$$

so that the maximized log-likelihood function can be expressed as

$$2 \log L(\hat{\theta}) = f \log \det(\hat{K}) - f|V|.$$

As a consequence, the likelihood ratio test statistic Q for comparing nested RCON models can be expressed through a ratio of determinants as

$$-2 \log \mathrm{LR} = f \log \frac{\det(\hat{K}_0)}{\det(\hat{K})} = f \log \frac{\det(\hat{\Sigma})}{\det(\hat{\Sigma}_0)} = f \log \det(\hat{\Sigma}\hat{\Sigma}_0^{-1}),$$

where $\hat{T}^0 = \hat{\Sigma}_0^{-1}$ is the maximum likelihood estimate in the smaller of the two nested RCON models.

From (6) we find the entries of the information matrix

$$I(\hat{\theta})_{uv} = f \operatorname{tr}(T^u \hat{\Sigma} T^v \hat{\Sigma})/2. \tag{8}$$

If the asymptotic covariance matrix of the maximum likelihood estimate is desired, it can be found by taking the inverse of this matrix.

3.3. Estimation algorithm

Clearly the likelihood equations can be solved by Newton iteration, provided appropriate starting values can be found. However, Jensen et al. (1991) described a globally convergent algorithm for a general linear exponential family using Newton's method on the mth root of the reciprocal likelihood function for one parameter at a time, see also Lauritzen (1996), p. 269. Here m is the number of independent observations, in our case corresponding to the degrees of freedom, $m = f$. This algorithm also avoids repeated inversion of the Fisher information matrix which is potentially of higher dimension than K.

For RCON models, this algorithm can be described as follows: For each generator u (where u can be either a vertex colour class or an edge colour class) we define the discrepancy $\Delta_u = \operatorname{tr}(T^u \hat{\Sigma}) - \operatorname{tr}(T^u W)/f$ where $\hat{\Sigma} = \hat{K}^{-1}$ denotes the current estimate of Σ at any time during the iteration. The iterative step can then be written as

$$\theta_u \leftarrow \theta_u + \frac{\Delta_u}{\operatorname{tr}(T^u \hat{\Sigma} T^u \hat{\Sigma}) + \Delta_u^2/2}. \tag{9}$$

The substitution (9) must theoretically be repeated until convergence for the set u before moving on to the next set in $\mathcal{V} \cup \mathcal{E}$. However, in most cases it is sufficient to make only a few steps, ensuring the likelihood has increased.

Thus the algorithm consists of two nested loops: an outer loop running over the elements of $\mathcal{V} \cup \mathcal{E}$, and an inner loop maximizing L with respect to θ_u while keeping all other parameters fixed. We repeatedly loop through all sets $u \in \mathcal{V} \cup \mathcal{E}$ until convergence.

In practice it can be computationally more efficient to combine this algorithm with iterative proportional scaling for graphical Gaussian models (Speed and Kiiveri, 1986; Lauritzen, 1996) and there are other ways of making the above algorithm more efficient, see Højsgaard and Lauritzen (2007) for details and implementation issues.

Table 3. Fitted concentrations $\times 1000$ (on and above diagonal) and fitted partial correlations (below diagonal) for the examination marks in five mathematical subjects assuming the RCON model with coloured graph as in Figure 2.

	Mechanics	Vectors	Algebra	Analysis	Statistics
Mechanics	6.16	−3.29	−3.29	0	0
Vectors	0.42	10.10	−3.29	0	0
Algebra	0.27	0.21	23.63	−6.51	−3.29
Analysis	0	0	0.42	10.10	−3.29
Statistics	0	0	0.27	0.42	6.16

3.4. Model properties

It follows directly from (3) that the class of RCON models is closed under conditioning in the sense that if the distribution of Y is specified through an RCON model with coloured graph $(\mathcal{V}, \mathcal{E})$, then the concentration matrix of the conditional distribution of Y_a given Y_b satisfies exactly the restrictions of $(\mathcal{V}_a, \mathcal{E}_a)$ where this is the coloured subgraph resulting by only keeping vertices in a and edges between elements of a and preserving their colours.

If we change the scale of the variables Y to form Y' where $Y'_\alpha = a_\alpha Y_\alpha, \alpha \in V$ then it would typically *not* be the case that Y' satisfies the same restrictions of an RCON model as Y unless $a_\alpha = a$ for all $\alpha \in V$. Thus, when using RCON models it is generally important that all variables are on comparable scales to ensure interpretability of conclusions. In contrast, models to be investigated in the next section possess properties of invariance under rescaling.

Example 2. The marks in mathematical subjects discussed earlier have been made on comparable scales as they all are marks out of 100. The RCON model with graph displayed in Figure 2 yields an excellent fit with a likelihood ratio of $-2 \log \mathrm{LR} = 7.2$ on 6 degrees of freedom, when compared to the butterfly model without symmetry restrictions in Figure 1. The fitted concentrations and partial correlations are displayed in Table 3.

4. Symmetry restrictions on partial correlations

4.1. Model specification

An $RCOR(\mathcal{V}, \mathcal{E})$ model with vertex colouring $\mathcal{V}$ and edge colouring $\mathcal{E}$ is obtained by restricting the elements of K as follows:

(a) All diagonal elements of K (inverse partial variances) corresponding to vertices in the same vertex colour class must be identical.

(b) All partial correlations corresponding to edges in the same edge colour class must be identical.

The set of positive definite matrices which satisfy the restrictions of an $RCOR(\mathcal{V}, \mathcal{E})$ model is denoted $\mathcal{R}^+(\mathcal{V}, \mathcal{E})$.

If we let A be the diagonal matrix with entries equal to the inverse partial standard deviations, i.e.

$$a_\alpha = \sqrt{k_{\alpha\alpha}} = 1/\sqrt{\mathbf{V}(Y_\alpha \,|\, Y_{V\setminus\alpha})},$$

and use (1), we can uniquely represent any $K \in \mathcal{R}^+(\mathcal{V}, \mathcal{E})$ as

$$K = ACA,$$

where $C = \{c_{\alpha\beta}\}$ has all diagonal entries equal to one and all off-diagonal entries are the negative partial correlations

$$c_{\alpha\beta} = -\rho_{\alpha\beta \,|\, V\setminus\{\alpha,\beta\}} = k_{\alpha\beta}/\sqrt{k_{\alpha\alpha}k_{\beta\beta}} = k_{\alpha\beta}/(a_\alpha a_\beta).$$

The vertex colour classes of the RCOR-model are then restricting the elements of A, whereas the edge colour classes are restricting elements of C to have the same value for entries in the same edge colour class. More precisely, if we let $K(\eta, \delta)$ denote the matrix with entries $k_{\alpha\beta} = a_\alpha a_\beta c_{\alpha\beta}$, then $\mathcal{V}$ defines the restrictions $a_\alpha = \eta_u$ if $\alpha \in u \in \mathcal{V}$. Similarly $\mathcal{E}$ represents the restrictions $c_{\alpha\beta} = c_{\beta\alpha} = \delta_u$ if $\{\alpha, \beta\} \in u \in \mathcal{E}$ and $c_{\alpha\beta} = 0$ if $\{\alpha, \beta\} \notin E = \cup_{u\in\mathcal{E}} u$.

4.2. Likelihood equations

Although the restrictions are linear in each of A and C, they are in general not linear in K and the models are therefore only curved exponential families. To obtain simplified expressions for derivatives we let $\lambda_u = \log \eta_u = \log a_\alpha = \frac{1}{2}\log k_{\alpha\alpha}$ for $\alpha \in u \in \mathcal{V}$. Differentiation yields

$$\frac{\partial}{\partial \delta_u} \operatorname{tr}(CAWA) = \operatorname{tr}(T^u AWA); \quad \frac{\partial}{\partial \lambda_u} A = T^u A = AT^u \tag{10}$$

and further

$$\frac{\partial}{\partial \lambda_u} \operatorname{tr}(ACAW) = 2\operatorname{tr}(T^u ACAW). \tag{11}$$

Expanding (5) and using that $\operatorname{tr}(ACAW) = \operatorname{tr}(CAWA)$ yields the log likelihood function

$$\log L = \frac{f}{2}\log\det(C) + f\sum_{u\in\mathcal{V}} \lambda_u \operatorname{tr}(T^u) - \frac{1}{2}\operatorname{tr}(CAWA). \tag{12}$$

Differentiating (12) w.r.t. (δ, λ) using (6), (10), and (11) yields the likelihood equations

$$\operatorname{tr}(T^u AWA) = f\operatorname{tr}(T^u C^{-1}), u \in \mathcal{E};\ \operatorname{tr}(T^u CAWA) = f\operatorname{tr}(T^u), u \in \mathcal{V} \tag{13}$$

or alternatively, since $AT^u = T^u A$ because A and T^u are both diagonal matrices, the equations can be expressed as

$$\operatorname{tr}(T^u AWA) = f\operatorname{tr}(T^u C^{-1}), u \in \mathcal{E};\ \operatorname{tr}(T^u ACAW) = f\operatorname{tr}(T^u), u \in \mathcal{V}. \tag{14}$$

Differentiating further yields the observed information matrix

$$J(\delta, \lambda)_{uv} = \begin{cases} f\operatorname{tr}(T^u C^{-1} T^v C^{-1})/2 & \text{for } u, v \in \mathcal{E} \\ \operatorname{tr}(T^u AWAT^v)/2 & \text{for } u \in \mathcal{V}, v \in \mathcal{E} \\ \operatorname{tr}(T^u AWAT^v C) & \text{for } u, v \in \mathcal{V}, u \neq v \\ 2\operatorname{tr}(T^u AWAT^u C) & \text{for } u = v \in \mathcal{V} \end{cases} \tag{15}$$

and the Fisher information matrix is then obtained by taking expectations

$$I(\delta,\lambda)_{uv} = f\begin{cases} \operatorname{tr}(T^uC^{-1}T^vC^{-1})/2 & \text{for } u,v\in\mathcal{E} \\ \operatorname{tr}(T^uC^{-1}T^v)/2 & \text{for } u\in\mathcal{V}, v\in\mathcal{E} \\ \operatorname{tr}(T^uC^{-1}T^vC) & \text{for } u,v\in\mathcal{V}, u\neq v \\ 2\operatorname{tr}(T^uC^{-1}T^uC) & \text{for } u=v\in\mathcal{V}. \end{cases} \tag{16}$$

The matrix of second derivatives (15) is not necessarily negative definite and the log-likelihood function is therefore not in general concave in (δ,λ). It is not known to the authors whether there could be more than one possible solution to the system of likelihood equations. Below we argue that the log-likelihood function (12) is concave in λ for fixed δ and *vice versa*. This does not ensure joint concavity and thus is not in itself sufficient to establish general uniqueness. Indeed the likelihood funtion with vertex classes $\mathcal{V}=\{(1,2,3),(4,5,6)\}$ and only one edge colour class with two edges $\mathcal{E}=\{(1{:}2,5{:}6)\}$ may have multiple local maxima for the likelihood function, we omit the details of the argument. This example is due to Søren Tolver Jensen (personal communication) and conforms with remarks on pp. 230-231 of Jensen and Madsen (2004).

For fixed A we have an exponential model as before. Hence, provided the number of observations is sufficiently large so that the maximum exists, *the log-likelihood function is strictly concave in δ for fixed A* and it is maximized by the unique solution to the system of equations

$$\operatorname{tr}(T^uAWA) = f\operatorname{tr}(T^uC^{-1}), u\in\mathcal{E}. \tag{17}$$

We next consider the likelihood function for fixed C. We first observe that the submatrix of the information matrix (15) corresponding to $u,v\in\mathcal{V}$ is also positive semidefinite. To see this, we first note that $\operatorname{tr}(T^uAWAT^uC)\geq 0$ and then let $X=\sum_{u\in\mathcal{V}}x_uT^u$ so

$$\sum_u\sum_v x_ux_v\operatorname{tr}(T^uAWAT^vC) + \sum_u x_u^2\operatorname{tr}(T^uAWAT^uC) = \operatorname{tr}(XAWAXC) + \sum_u x_u^2\operatorname{tr}(T^uAWAT^uC)\geq 0$$

since C and $XAWAX$ are positive (semi)definite so also $\operatorname{tr}(XAWAXC)\geq 0$. Hence, if the number of observations is sufficiently large, *the log-likelihood function is strictly concave in λ for fixed C* and its maximum is given by the unique solution to the system of equations

$$\operatorname{tr}(T^uCAWA) = f\operatorname{tr}(T^u), u\in\mathcal{V}. \tag{18}$$

Also in the case of RCOR models, the maximized log-likelihood function can be expressed in terms of the determinant of the estimate. Using (14) and the fact that $\sum_{u\in\mathcal{V}}T^u = I$ yields

$$\operatorname{tr}(\hat{K}W) = \operatorname{tr}(\hat{A}\hat{C}\hat{A}W) = \sum_{u\in\mathcal{V}}\operatorname{tr}(T^u\hat{A}\hat{C}\hat{A}W) = f\sum_{u\in\mathcal{V}}\operatorname{tr}(T^u) = f\operatorname{tr}(I) = f|V|$$

and hence the maximized log-likelihood becomes

$$2\log L(\hat{\delta},\hat{\eta}) = f\log\det(\hat{C}) + 2f\sum_\alpha\log\hat{a}_\alpha - f|V| = f\log\det(\hat{K}) - f|V|.$$

The likelihood ratio test statistic for comparing nested RCOR models becomes

$$-2\log \mathrm{LR} = f\log\frac{\det(\hat{K}_0)}{\det(\hat{K})} = f\log\frac{\det(\hat{\Sigma})}{\det(\hat{\Sigma}_0)} = f\log\det(\hat{\Sigma}\hat{\Sigma}_0^{-1}),$$

where $\hat{K}^0 = \hat{\Sigma}_0^{-1}$ is the maximum likelihood estimate in the smaller of the two nested RCOR models.

4.3. Estimation algorithm

The likelihood equations must be solved iteratively. Clearly one can use Newton iteration based on the matrix of second derivatives (15) or Fisher's more stable method of scoring, replacing the matrix of second derivatives with the information matrix (16). The method described below avoids repeated inversions of these matrices which are potentially of higher dimension than C. The algorithm alternates between maximizing over C for fixed A and maximizing over A for fixed C.

The likelihood function is maximized in C for fixed A by solving the equations in (17) using the iteration in Section 3.3 for every edge colour class $u \in \mathcal{E}$, i.e. by iterating

$$\delta_u \leftarrow \delta_u + \frac{\Delta_u}{\mathrm{tr}(T^u\hat{C}^{-1}T^u\hat{C}^{-1}) + \Delta_u^2/2} \tag{19}$$

until convergence, with $\Delta_u = \mathrm{tr}(T^u\hat{C}^{-1}) - \mathrm{tr}(T^uAWA)/f$.

The likelihood function in η_u for a given $u \in \mathcal{V}$, keeping δ and other components of η fixed, can be maximized in closed form, observing that the equations (18) can be rewritten as

$$\eta_u^2 \sum_{\alpha\in V_u}\sum_{\beta\in V_u} Q_{\alpha\beta} + \eta_u \sum_{\alpha\in V_u}\sum_{\beta\notin V_u} Q_{\alpha\beta}a_\beta = f|V_u|, u \in \mathcal{V},$$

where Q is the Hadamard (or Schur) product of C and W with entries $Q_{\alpha\beta} = C_{\alpha\beta}W_{\alpha\beta}$. Since C and W are both positive definite this also holds for Q (Schur, 1911), see also Horn and Johnson (1985) p. 455ff. This equation has a unique positive root for η_u:

$$\eta_u \leftarrow \frac{-B + \sqrt{B^2 + 4f|V_u|D}}{2D},$$

where

$$B = \sum_{\alpha\in V_u}\sum_{\beta\notin V_u} Q_{\alpha\beta}a_\beta, \quad D = \sum_{\alpha\in V_u,\beta\in V_u} Q_{\alpha\beta},$$

where $D > 0$ for the number of observations being sufficiently large.

Cycling through all $u \in \mathcal{V}$ with this substitution and alternating between these and the iterations (19) yields an iterative partial maximization algorithm. From the appendix of Drton and Eichler (2006) it follows that if there are only a finite number of solutions to the likelihood equations (13), the algorithm will indeed converge to one of those solutions.

4.4. Model properties

As was the case for RCON models, conditioning on Y_b in an RCOR model generates an RCOR model for the conditional concentration K_{aa} of Y_a given Y_b.

Table 4. Fitted concentrations ($\times 100$) (on and above diagonal), fitted partial correlations (below diagonal), and observed concentrations using RCOR model with coloured graph as in Figure 4 for personality characteristics of 684 students.

	SX	SN	TX	TN
SX (State anxiety)	0.59	−0.31	−0.22	0
SN (State anger)	0.46	0.78	0	-0.15
TX (Trait anxiety)	0.46	0	0.40	-0.10
TN (Trait anger)	0	0.31	0.31	0.28
Observed concentrations	0.58	0.79	0.41	0.27

In contrast to the case of *RCON* models, an RCOR model is *invariant under rescaling* if variables in the same vertex colour class are changed in the same way, i.e. under transformation of the form

$$y^* \leftarrow \Psi y, \text{ for } \Psi = \textstyle\sum_{u\in\mathcal{V}} \psi_u T^u, \tag{20}$$

with $\psi_u > 0$. Such a transformation would affect only the matrix A and not C. More precisely, Y^* has parameters $K^* \sim (A^*, C^*) = (\Psi^{-1}A, C)$. The scaling obtained by choosing $\Psi = A$ shall be called the *intrinsic scale*. With this rescaling, C will be the concentration matrix for Y^*.

The set of transformations (20) form a group, A is *equivariant* and C *invariant* under the action of the group. It follows that an RCOR model is a *composite transformation model* in the sense of Barndorff-Nielsen et al. (1982) with C as index parameter and A as group parameter. Following Barndorff-Nielsen et al. (1982) an alternative method of estimation would be to use the marginal likelihood from the correlation matrix $R = S^{-1/2}WS^{-1/2}$ where S is the diagonal matrix with elements $S_{\alpha\alpha} = W_{\alpha\alpha}$. The correlation matrix R is invariant under the group action and its distribution therefore only depends on C. The intrinsic scale A could then be estimated from the conditional distribution of W given R. Although this avoids problems of non-uniqueness of estimates, it appears technically involved, due to the fact that the marginal density of R seems to have no simple structure and depends on C in a rather complex manner.

Example 3. The data described in Table 2 on anxiety and anger is an example where it is not obvious that the scalings of each of the four variables are comparable and therefore symmetry models of RCOR type are appropriate. Fitting the RCOR model displayed in Figure 4 yields a likelihood ratio of $-2\log \mathrm{LR} = 0.22$ on 2 degrees of freedom when comparing with the model in Figure 3 without symmetry restrictions. The fitted valued of the parameters are displayed in Table 4.

Note that if we transform all variables to the intrinsic scale by multiplying each with the square root of the fitted relevant diagonal element of the concentration, then the transformed data also fits the model where all diagonal elements are identical, and hence the regression interpretations in Example 1 are valid. Thus, when variables are measured on the intrinsic scale, state anger and trait anxiety contribute equally to the prediction of state anxiety and equally to the prediction of trait anger. Also, regressions of state anxiety and trait anger on trait anxiety and state anger are parallel.

5. Permutation symmetry

5.1. Model specification

Consider a permutation of the elements of V i.e. an element of the symmetric group $S(V)$, represented by a permutation matrix G with $G_{\alpha\beta} = 1$ if and only if G maps β into α. If $Y \sim \mathcal{N}_{|V|}(0, \Sigma)$ then $GY \sim \mathcal{N}_{|V|}(0, G\Sigma G^\top)$. Let now $\Gamma \subseteq S(V)$ be a subgroup of such permutations. The distribution of Y is *invariant under the action of* Γ if and only if

$$G\Sigma G^\top = \Sigma \text{ for all } G \in \Gamma. \tag{21}$$

Since a permutation matrix G satisfies $G^{-1} = G^\top$, (21) is equivalent to

$$G\Sigma = \Sigma G \text{ for all } G \in \Gamma, \tag{22}$$

i.e. that G commutes with Σ. Because Γ is a group, $G \in \Gamma \implies G^{-1} \in \Gamma$. Taking inverses on both sides of (22) and substituting $G^\top$ for G^{-1} yields that the invariance condition is equivalent to G commuting with the concentration matrix K:

$$GK = KG \text{ for all } G \in \Gamma. \tag{23}$$

To specify permutation symmetry for a graphical Gaussian model with graph $\mathcal{G} = (V, E)$ we would also insist that zero elements of K are preserved, in other words that the permutation is an *automorphism* of the graph, mapping edges to edges:

$$G(\alpha) \sim G(\beta) \iff \alpha \sim \beta \text{ for all } G \in \Gamma, \tag{24}$$

i.e. $\Gamma \subseteq \mathrm{Aut}(\mathcal{G})$, where $\mathrm{Aut}(\mathcal{G})$ is the group of automorphisms of $\mathcal{G}$. The condition (24) can also be expressed as

$$GA(\mathcal{G}) = A(\mathcal{G})G \text{ for all } G \in \Gamma,$$

where $A(\mathcal{G})$ is the adjacency matrix of $\mathcal{G}$.

An RCOP model $RCOP(\mathcal{G}, \Gamma)$ generated by $\Gamma \subseteq \mathrm{Aut}(\mathcal{G})$ is given by assuming

$$K \in \mathcal{S}^+(\mathcal{G}, \Gamma) = \mathcal{S}^+(\mathcal{G}) \cap \mathcal{S}^+(\Gamma)$$

where $\mathcal{S}^+(\Gamma)$ is the set of positive definite matrices satisfying (23).

An RCOP model can also be represented by a graph colouring. More precisely, if $\mathcal{V}$ denotes the *vertex orbits* of Γ, i.e. the equivalence classes of the relation

$$\alpha \equiv_\Gamma \beta \iff \beta = G(\alpha) \text{ for some } G \in \Gamma,$$

and similarly $\mathcal{E}$ the *edge orbits*, i.e. the equivalence classes of the relation

$$\{\alpha, \gamma\} \equiv_\Gamma \{\beta, \delta\} \iff \{\beta, \delta\} = \{G(\alpha), G(\gamma)\} \text{ for some } G \in \Gamma,$$

then we have

$$\mathcal{S}^+(\mathcal{G}, \Gamma) = \mathcal{S}^+(\mathcal{V}, \mathcal{E}) = \mathcal{R}^+(\mathcal{V}, \mathcal{E}).$$

Hence an RCOP model can also be represented as an RCON or an RCOR model with vertex orbits as vertex colour classes and edge orbits as edge colour classes. For later use we note that the permutation matrices in Γ commute with all adjacency matrices of the colour classes:

$$GT^u = T^u G \text{ for all } G \in \Gamma \text{ and all } u \in \mathcal{V} \cup \mathcal{E}. \tag{25}$$

5.2. Likelihood equations

Since an RCOP model is also an RCON and an RCOR model, the likelihood equations can be formulated in such terms and any of the algorithms previously presented can be used to calculate the maximum likelihood estimate. However, for RCOP models there is another alternative.

Representing an RCOP model as an RCON model yields the likelihood equations

$$\operatorname{tr}(T^u W) = f \operatorname{tr}(T^u \Sigma), u \in \mathcal{V} \cup \mathcal{E}; \quad \Sigma^{-1} \in \mathcal{S}^+(\mathcal{V}, \mathcal{E}). \tag{26}$$

We shall show that these equations are equivalent to the following

$$\operatorname{tr}(T^l \overline{W}) = f \operatorname{tr}(T^l \Sigma), l \in V \cup E; \quad \Sigma^{-1} \in \mathcal{S}^+(\mathcal{G}), \tag{27}$$

where the matrices $T^l, l \in V \cup E$ represent the graph $\mathcal{G}$ with neutral colours and

$$\overline{W} = \frac{1}{|\Gamma|} \sum_{G \in \Gamma} G W G^\top.$$

In other words, the equations can be solved by first taking appropriate averages of the elements in the Wishart matrix and then solving the equations for corresponding graphical Gaussian model without symmetry restrictions. Estimation relations of this type are typical in group invariance models, and (27) have been derived using general theory (Andersson, 1975; Barndorff-Nielsen et al., 1982; Andersson et al., 1983). Indeed, the relation can also be found in Hylleberg et al. (1993).

The equations (27) can be solved using iterative proportional scaling (Speed and Kiiveri, 1986; Lauritzen, 1996), or explicitly if $\mathcal{G}$ is decomposable (Lauritzen 1996, p. 146). In both cases this can lead to considerable computational savings.

To see that (26) and (27) are equivalent we first note that for $u \in \mathcal{V} \cup \mathcal{E}$

$$T^u = \sum_{l \in u} T^l; \text{ and } G T^l = T^{G(l)} G, \ l \in V \cup E.$$

Assume first that Σ is the unique solution to (27). We then get for all $l \in V \cup E$ that

$$\operatorname{tr}(T^l G \Sigma G^\top) = \operatorname{tr}(T^{G^{-1}(l)} \Sigma) = \operatorname{tr}(T^{G^{-1}(l)} \overline{W}) = \operatorname{tr}(T^l \overline{W})$$

and thus $G \Sigma G^\top$ also solves (27). Since the solution is unique, we must have $G \Sigma G^\top = \Sigma$ and hence $\Sigma \in \mathcal{S}^+(\mathcal{V}, \mathcal{E})$. That Σ is indeed a solution to (26) now follows from the calculation

$$\operatorname{tr}(T^u \Sigma) = \operatorname{tr}(T^u \overline{W}) = \frac{1}{|\Gamma|} \sum_{G \in \Gamma} \operatorname{tr}(T^u G W G^\top) = \operatorname{tr}(T^u W)$$

where we have used that $T^u G = G T^u$ and $G^\top G = I$.

Conversely, if Σ satisfies (26) we have

$$\operatorname{tr}(T^l \overline{W}) = \frac{1}{|u|} \operatorname{tr}(T^u \overline{W}) = \frac{1}{|u|} \operatorname{tr}(T^u W) = \frac{1}{|u|} \operatorname{tr}(T^u \Sigma) = \operatorname{tr}(T^l \Sigma),$$

where $u = \{G(l), G \in \Gamma\}$ is the orbit of l.

Table 5. Observed concentrations ($\times 100$) (on and above diagonal) together with fitted concentrations (below diagonal) using RCOP model with coloured graph as in Figure 6 for head dimensions of first and second sons.

	$L1$	$B1$	$L2$	$B2$
$L1$ (Length of head of first son)	3.21	−1.60	−0.78	−1.11
$B1$ (Breadth of head of first son)	−1.76	2.21	−0.50	0.48
$L2$ (Length of head of second son)	−1.41	0	2.67	−1.89
$B2$ (Breadth of head of second son)	0	−1.71	−1.76	3.37
Fitted concentrations	2.90	2.48	2.90	2.48

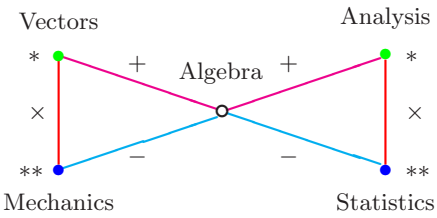

Fig. 8. Coloured graph of an RCOP symmetry model for the examination marks of 88 students. The distribution of the marks is unchanged if we simultaneously replace Vectors with Analysis and Mechanics with Statistics.

Example 4. Table 5 shows observed and fitted concentrations for head dimensions of 25 pairs of sons. The symmetry restrictions represent the natural symmetry between first and second son. The likelihood ratio for testing the validity of the model versus the saturated model without symmetries is equal to $-2\log \mathrm{LR} = 5.18$ on 5 degrees of freedom, thus representing an excellent fit.

Since the model is an RCOP model, it can be fitted with any of the three algorithms presented in the article and the symmetry restrictions will be reflected in the partial correlations, as well as in the covariance or correlation matrix.

Example 5. The symmetry model in Figure 8 is also giving an excellent fit to the data. The likelihood ratio is $-2\log \mathrm{LR} = 5.0$ on 5 degrees of freedom when compared with the butterfly model without additional symmetries. This model says that the distribution is invariant when we simultaneously replace Vectors with Analysis and Mechanics with Statistics. The fitted values of the concentration matrix are displayed in Table 6.

6. Relations between model types

As we have argued above, any RCOP model is automatically also RCON and RCOR whereas the converse is false. Also the latter two model types are generally different. However, if the coloured graph satisfies certain properties, RCON and RCOR models coincide, without necessarily being RCOP. The restrictions on the inverse partial variances $k_{\alpha\alpha}$ are clearly

Table 6. Fitted concentrations $\times 1000$ (on and above diagonal) and fitted partial correlations (below diagonal) for the examination marks in five mathematical subjects assuming the RCOP model with coloured graph as in Figure 8.

	Mechanics	Vectors	Algebra	Analysis	Statistics
Mechanics	5.75	−2.28	−3.70	0	0
Vectors	0.30	9.96	−6.44	0	0
Algebra	0.29	0.39	27.4	−6.44	−3.70
Analysis	0	0	0.39	9.79	−2.28
Statistics	0	0	0.29	0.30	5.75

identical for both model types, so the question only pertains to the edge colour classes implying the same restrictions on the concentration matrices.

To identify when this is the case, consider first two incident edges $\{\alpha, \beta\}$ and $\{\alpha, \gamma\}$ of the same colour class. Using (1) it follows directly that the restriction $k_{\alpha\beta} = k_{\alpha\gamma}$ is equivalent to the restriction $\rho_{\alpha\beta \,|\, V\setminus\{\alpha,\beta\}} = \rho_{\alpha\gamma \,|\, V\setminus\{\alpha,\gamma\}}$ if and only if $k_{\beta\beta} = k_{\gamma\gamma}$.

Similarly, for two non-incident edges $\{\alpha, \beta\}$ and $\{\gamma, \delta\}$, the restriction $k_{\alpha\beta} = k_{\gamma\delta}$ is equivalent to the restriction $\rho_{\alpha\beta \,|\, V\setminus\{\alpha,\beta\}} = \rho_{\gamma\delta \,|\, V\setminus\{\gamma,\delta\}}$ if and only if either

$$k_{\alpha\alpha} = k_{\gamma\gamma} \text{ and } k_{\beta\beta} = k_{\delta\delta} \tag{28}$$

or the similar relation with α and β interchanged

$$k_{\beta\beta} = k_{\gamma\gamma} \text{ and } k_{\alpha\alpha} = k_{\delta\delta}. \tag{29}$$

The condition for incident edges is clearly implied by those for non-incident edges. We summarize this in the following.

PROPOSITION 1. *The RCOR and RCON model determined by $(\mathcal{V}, \mathcal{E})$ yield identical restrictions*

$$\mathcal{R}^+(\mathcal{V}, \mathcal{E}) = \mathcal{S}^+(\mathcal{V}, \mathcal{E})$$

if and only if any pair of edges in the same colour class $\alpha\beta, \gamma\delta \in u \in \mathcal{E}$ connect the same vertex colour classes.

The model in Figure 7 clearly satisfies this condition. It is therefore both RCOR and RCON. However, it is not an RCOP model because the largest subgroup of $\mathrm{Aut}(\mathcal{G})$ preserving the colour symmetries has only one element other than the identity which is the permutation switching 1 and 3. However, 2 and 4 are not in the same vertex orbit of that group, so these will have different colours in the corresponding RCOP model, corresponding to Y_2 and Y_4 having different marginal *variances*. In fact, *any non-trivial RCOR model with atomic vertex classes* cannot be RCON. Indeed RCOR models with atomic vertex colour classes may seem more interpretable in situations where scale is not well defined.

Figure 9 shows another example of a model where edges in same colour class connect the same vertex colour classes and it is therefore is both RCON and RCOR. However, this model is RCOP for the group generated by simultaneously interchanging 1 with 3 and 2

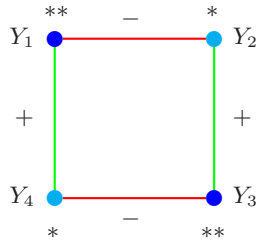

Fig. 9. Graph colouring for another model which is both RCON and RCOR. Indeed, this model is RCOP for the group generated by *simultaneously* interchanging 1 with 3 and 2 with 4.

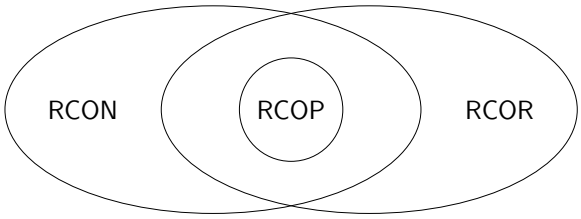

Fig. 10. Relations between symmetry models. Models given by permutation symmetry (RCOP) have similar symmetries for concentrations (RCON) and partial correlations (RCOR). RCON models are not necessarily RCOR and *vice versa* but a model can be simultaneously RCON and RCOR without being RCOP.

with 4, i.e. the nilpotent permutation matrix

$$\begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}.$$

When the RCOR and RCON interpretations of a coloured graph coincide, the RCOR model clearly inherits the property of uniqueness of the maximum likelihood estimate from its RCON manifestation. Hence in such a case both scale invariance and simple estimation prevails. The relationship between the model classes are shown graphically in Figure 10.

7. Breast cancer genes

The examples considered so far have been chosen to be as simple as possible with the purpose of illustrating the fundamental concepts and issues associated with symmetry restrictions. This example serves to show that it is indeed possible to fit high dimensional models with symmetry restrictions.

Miller et al. (2005) investigated gene expression signatures for p53 mutation status in 250 breast cancer samples. Of these, 58 samples have a mutation in the p53 sequence and data from these are considered in the following. The data have been standardized to have zero mean and unit variance. For simplicity we consider only 150 genes. Note that in this

example there are more variables than observations, so only models with a large degree of parsimony can be fitted at all. Since there is an issue of scaling of the variables, we are looking for models of RCOR type.

As it is outside the scope of the present article to develop and discuss robust and efficient model selection procedures for these models, we shall only give a very rough exploratory initial analysis.

Our first step is to identify an initial dependence graph for the gene expressions and we chose to use the lasso (Tibshirani, 1996) for this purpose, as described by Meinshausen and Bühlmann (2006): for each vertex $\alpha \in V$ we performed a lasso regression on the remaining vertices $V \setminus \{\alpha\}$. Whenever a regressor variable $\beta \in V \setminus \{\alpha\}$ is actively present in this regression, $\{\alpha, \beta\}$ was included as an edge in the graph. Using the `l1ce` function in the R-package `lasso2` (Osborne et al., 2000) with penalty 0.05 produces a sparse graph with 174 edges.

Next we initially estimated the concentration parameters $k_{\alpha\beta}$ by the marginal empirical concentrations $\{W_{c(\alpha,\beta)}/f\}^{-1}$, where $c(\alpha, \beta) = \mathrm{ne}(\alpha) \cap \mathrm{ne}(\beta) \cup \{\alpha, \beta\}$ includes the common neighbours of α and β and calculated the corresponding partial correlations when α and β were different. A rough cluster analysis of these and the diagonal elements $k_{\alpha\alpha}$ led us to choose 7 clusters for the edges and 10 clusters for the vertices, corresponding to the RCON model displayed in Figure 11.

The level of parsimony achieved by adding this symmetry to the model is a reduction of the number of parameters from 324 to 17, with a drop in the likelihood from -1396 to -1992. Although clearly significant— with a χ^2-difference of 1192 on 307 degrees of freedom— the BIC criterion corresponding to 58 observations yields a value of $1396 + 658 = 2054$ for the model without restrictions whereas the model with symmetry has a BIC value of $1992 + 35 = 2027$, suggesting that the simpler model yields a reasonable overall description of the basic features of the data. We note that the computation time required to fit the model in Figure 11 as an RCOR model was about 2 seconds on a standard laptop using the current version of the R-package described in Højsgaard and Lauritzen (2007).

8. Discussion and perspectives

We have described three classes of symmetry restrictions imposed on graphical Gaussian models. The simplest of these are given by symmetry under permutation of variable labels. Software has been developed under R which implements the methods we have described (Højsgaard and Lauritzen, 2007), excluding the specific simplifications available for the case of RCOP models.

Alternative ways of introducing symmetry would apply restrictions to marginal covariances and correlations rather than concentrations and partial correlations. This would be natural to do in connection with graphical models for marginal independence, sometimes known as covariance graph models (Kauermann, 1996; Cox and Wermuth, 1993, 1996), although these have more complex estimation properties (Drton and Richardson, 2003, 2004). Note that RCOP models would automatically also satisfy the relevant symmetry restrictions for the marginal covariances and correlations.

In general, the RCOP models are simplest and most readily interpretable, partly due to their justification through symmetries among the variables under study, partly due to the fact that there are no paradoxes of the form that identity of partial correlations and concentrations are very different assumptions in general. We only recommend use of RCON

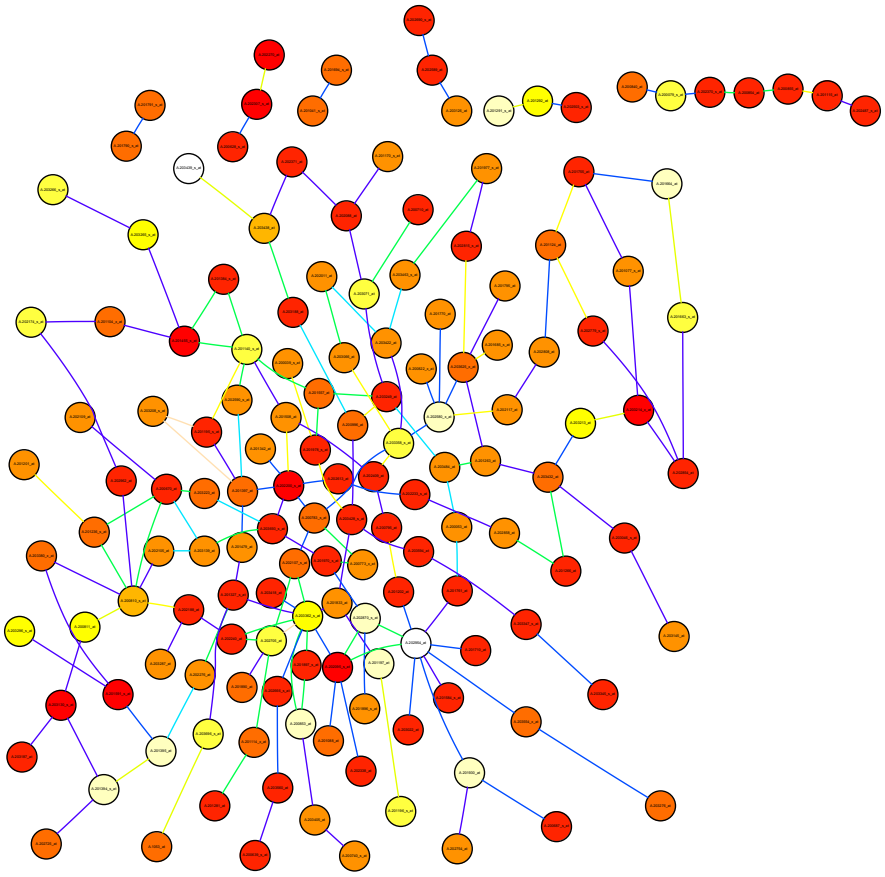

Fig. 11. RCOR model for gene expression data. There are 7 edge colour classes and 10 vertex colour classes

models in cases where the variables clearly are measured on comparable scales. For cases where measurement scales are not absolutely defined, RCOR models with only atomic vertex colour classes may be of special interest. In all cases it might be worthwhile to consider whether a minor extension/modification of an RCOR or RCON model could lead to an RCOP version of the model, and generally be very careful with interpreting results from an RCOR or RCON analysis.

For the models to become widely applicable in exploratory analysis of data with a large number of variables it is mandatory to develop algorithms for model identification which are robust, reliable, and transparent. In particular it is important that the algorithms avoid an intermediate estimation of a model without symmetries where parameters could be unreliably determined or even not estimable.

It would be of interest to study discrete symmetry models based on graphical log-linear models, extending classic models of symmetry, marginal homogeneity and quasi-symmetry

in contingency tables (Agresti, 1990).

As mentioned in the introduction, instances of the models have been considered previously in different contexts. Here we mention in addition that Wolfe (1976) considered the hypothesis of equality of two correlation coefficients in a trivariate Gaussian distributions, showing that adding additional symmetry constraints, effectively identifying an RCOP model, would lead to a test statistic with a simple distribution.

Acknowledgements

This work has been supported by a grant financed by the Directorate for Food, Fisheries and Agro Business, Denmark, Lattec I/S, the Danish Cattle Association, the Danish Institute of Agricultural Sciences, and by the Danish Natural Science Research Council. The authors have benefited from conversations with Mathias Drton, Ingram Olkin and Michael Perlman, and Chris Holmes has kindly made the breast cancer data available to us.

References

Agresti, A. (1990). *Categorical Data Analysis*. New York: John Wiley and Sons.

Anderson, R. L. (1942). Distribution of the serial correlation coefficient. *Ann. Math. Statist. 13*, 1–13.

Anderson, T. W. (1969). Statistical inference for covariance matrices with linear structure. In P. R. Krishnaiah (Ed.), *Proceedings of the Second International Symposium on Multivariate Analysis*, New York, pp. 55–66. Academic Press.

Anderson, T. W. (1970). Estimation of covariance matrices which are linear combinations or whose inverses are linear combinations of given matrices. In R. C. Bose, I. M. Chakravarti, P. C. Mahalanobis, C. R. Rao, and K. J. C. Smith (Eds.), *Essays in Probability and Statistics*, pp. 1–24. University of North Carolina Press, Chapel Hill, N.C.

Anderson, T. W. (1973). Asymptotically efficient estimation of covariance matrices with linear structure. *Ann. Statist. 1*, 135–141.

Andersson, S. A. (1975). Invariant normal models. *Ann. Statist. 3*, 132–154.

Andersson, S. A., H. Brøns, and S. T. Jensen (1983). Distribution of eigenvalues in multivariate statistical analysis. *Ann. Statist. 11*, 392–415.

Andersson, S. A. and J. Madsen (1998). Symmetry and lattice conditional independence in a multivariate normal distribution. *Ann. Statist. 26*, 525–572.

Barndorff-Nielsen, O. E., P. Blæsild, J. L. Jensen, and B. Jørgensen (1982). Exponential transformation models. *Proceedings of the Royal Society of London, Series A 379*, 41–65.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. R. Statist. Soc. B 36*, 302–309.

Besag, J. E. and P. A. P. Moran (1975). On the estimation and testing of spatial interaction in Gaussian lattice processes. *Biometrika 62*, 555–562.

Bollobás, B. (1998). *Modern Graph Theory*. New York: Springer-Verlag.

Buhl, S. L. (1993). On the existence of maximum likelihood estimators for graphical Gaussian models. *Scand. J. Statist. 20*, 263–270.

Cox, D. R. and N. Wermuth (1993). Linear dependencies represented by chain graphs (with discussion). *Statist. Sci. 8*, 204–218; 247–277.

Cox, D. R. and N. Wermuth (1996). *Multivariate Dependencies*. London, United Kingdom: Chapman and Hall.

Dempster, A. P. (1972). Covariance selection. *Biometrics 28*, 157–175.

Drton, M. and M. Eichler (2006). Maximum likelihood estimation in Gaussian chain graph models under the alternative Markov property. *Scand. J. Statist. 33*, 247–257.

Drton, M. and T. S. Richardson (2003). An algorithm for maximum likelihood estimation in Gaussian graphical models for marginal independence. In *Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence*, San Francisco, pp. 130–137. Morgan Kaufmann Publishers.

Drton, M. and T. S. Richardson (2004). Multimodality of the likelihood function in the bivariate seemingly unrelated regressions model. *Biometrika 91*, 383–392.

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

Frets, G. P. (1921). Heredity of head form in man. *Genetica 3*, 193–400.

Højsgaard, S. and S. L. Lauritzen (2005). Restricted concentration models – graphical Gaussian models with concentration parameters restricted to being equal. In R. G. Cowell and Z. Ghahramani (Eds.), *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, Hillsdale, NJ, pp. 159–187. Society for Artificial Intelligence and Statistics. Available electronically at http://www.gatsby.ucl.ac.uk/aistats/.

Højsgaard, S. and S. L. Lauritzen (2007). Inference in graphical Gaussian models with edge and vertex symmetries with the `gRc` package for R. *Journal of Statistical Software 23*; Issue 6.

Horn, R. A. and C. R. Johnson (1985). *Matrix Analysis*. Cambridge, UK: Cambridge University Press.

Hylleberg, B., M. Jensen, and E. Ørnbøl (1993). Graphical symmetry models. Master's thesis, Aalborg University, Aalborg, Denmark.

Jensen, S. T., S. Johansen, and S. L. Lauritzen (1991). Globally convergent algorithms for maximizing a likelihood function. *Biometrika 78*, 867–877.

Jensen, S. T. and J. Madsen (2004). Estimation of proportional covariances in the presence of linear restrictions. *Ann. Statist. 31*, 219–232.

Kauermann, G. (1996). On a dualization of graphical Gaussian models. *Scand. J. Statist. 23*, 105–116.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford, United Kingdom: Clarendon Press.

Leipnik, R. B. (1947). Distribution of the serial correlation coefficent in a circularly correlated universe. *Ann. Math. Statist. 18*, 80–87.

Madsen, J. (2000). Invariant normal models with recursive graphical Markov structure. *Ann. Statist. 28*(4), 1150–1178.

Mardia, K. V., J. T. Kent, and J. M. Bibby (1979). *Multivariate Analysis*. Academic Press.

Meinshausen, N. and P. Bühlmann (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist. 34*, 1436–1462.

Miller, L. D., J. Smeds, J. George, V. B. Vega, L. Vergara, Y. Pawitan, P. Hall, S. Klaar, E. T. Liu, and J. Bergh (2005). An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proceedings of the National Academy of Sciences 102*(38), 13550–13555.

Olkin, I. and S. J. Press (1969). Testing and estimation for a circular stationary model. *Ann. Math. Statist. 40*, 1358–1373.

Osborne, M. R., B. Presnell, and B. A. Turlach (2000). On the LASSO and its dual. *Proceedings of the National Academy of Sciences 9*, 319–337.

Schur, I. (1911). Bemerkungen zur Theorie der beschränkten Bilinearformen mit unendlich vielen Veränderlichen. *Journal für die reine und angewandte Mathematik 140*, 1–29.

Speed, T. P. and H. Kiiveri (1986). Gaussian Markov distributions over finite graphs. *Ann. Statist. 14*, 138–150.

Spielberger, C. D., R. L. Gorsuch, and R. E. Luschene (1970). *Manual for the State–Trait Anxiety Inventory*. Palo Alto, CA: Consulting Psychologists Press.

Spielberger, C. D., S. Russell, and R. Crane (1983). Assessment of anger. In J. N. Butcher and C. D. Spielberger (Eds.), *Advances in Personality Assessment*, Volume 2, pp. 159–187. Hillsdale, NJ: Erlbaum.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B 56*, 267–288.

Votaw, D. F. (1948). Testing compound symmetry in a normal multivariate distribution. *Ann. Math. Statist. 19*, 447–473.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Chichester, United Kingdom: John Wiley and Sons.

Whittle, P. (1954). On stationary processes in the plane. *Biometrika 41*, 439–449.

Wilks, S. S. (1946). Sample criteria for testing equality of means, equality of variances, and equality of covariances in a normal multivariate distribution. *Ann. Math. Statist. 17*, 257–281.

Wolfe, D. A. (1976). On testing equality of related correlation coefficients. *Biometrika 63*, 214–215.