Restricted concentration models – graphical Gaussian models with concentration parameters restricted to being equal

Søren Højsgaard
Biometry Research Unit
Danish Institute of Agricultural Sciences
Research Center Foulum, DK–8830 Tjele
Denmark

Steffen Lauritzen
Department of Statistics
University of Oxford
1 South Parks Road, Oxford OX1 3TG
United Kingdom

Abstract

In this paper we introduce restricted concentration models (RCMs) as a class of graphical models for the multivariate Gaussian distribution in which some elements of the concentration matrix are restricted to being identical is introduced. An estimation algorithm for RCMs, which is guaranteed to converge to the maximum likelihood estimate, is presented. Model selection is briefly discussed and a practical example is given.

1 Introduction

This paper introduces a class of graphical Gaussian models, Lauritzen (1996), (hereafter abbreviated GGMs) also known as covariance selection models, Dempster (1972), in which elements of the concentration matrix are restricted to being identical. Such models are denoted *restricted concentration models* and abbreviated RCMs. These models are linear in the inverse covariance matrix and can therefore be seen as instances of models discussed by Anderson (1970). Besag (1974) also studies instances of such models.

RCMs can be of relevance in a variety of different problems. An example could be gene expression data where the expression of many genes are measured. From a biological point of view it may be of interest to embody in the model that the conditional covariance between genes i and j should be the same as the conditional covariance between genes k and l. It may also be of interest (and in some cases a necessity) to impose such restrictions simply in order to reduce the dimensionality of the problem.

Models with equal conditional correlations can be constructed within RCMs but this requires restrictions on both the conditional covariances and the conditional variances. An interesting extension of RCMs would therefore be models with equal conditional correlations and no other restraints.

Finally we mention that the restrictions in RCMs can lead to some regression functions being constrained to equality as illustrated in Section 3.

2 Background and notation

The setting in GGMs is i.i.d. samples of a random vector $y = (y_1, \ldots, y_d)^\top$ following a $N_d(\mu, \Sigma)$ distribution. Let $K = \Sigma^{-1}$ denote the inverse covariance matrix, also known as the concentration matrix with elements $(k^{\alpha\beta})$. It is then well known, Lauritzen (1996), p. 130, that the partial correlation between y_1 and y_2 given all other variables is

$$\rho_{12|3\ldots d} = -k^{12}/\sqrt{k^{11}k^{22}} \tag{1}$$

Thus $k_{12} = 0$ if and only if y_1 and y_2 are independent given all other variables, and this is the traditional focus of graphical Gaussian modeling.

A GGM is often represented by an undirected graph $G = (\Gamma, E)$ where Γ is the set of nodes representing the d variables and E is the set of undirected edges representing the concentration parameters $k^{\alpha\beta}$ which are not restricted to being zero. For additional properties of GGMs we refer to Lauritzen (1996), Chapter 5. In the following we use Greek letters to refer to variables and Latin letters to refer to sets of variables.

3 The problem to be solved

The issue addressed in this paper is to estimate K when some entries $k^{\alpha\beta}$ are restricted to being equal. Such restrictions can be imposed both on the diagonal and the off–diagonal elements of K.

Example To illustrate possible implications of such restrictions, consider the model in Figure 1. The asterisks indicate the restrictions that $k^{13} = k^{14} = c_1$,

$k^{23} = k^{24} = c_2$ and $k^{33} = k^{44} = c_3$, i.e.

$$K = \begin{bmatrix} k^{11} & k^{12} & c_1 & c_1 \\ k^{12} & k^{22} & c_2 & c_2 \\ c_1 & c_2 & c_3 & 0 \\ c_1 & c_2 & 0 & c_3 \end{bmatrix}$$

If we let $a = \{1, 2\}$ and $b = \{3, 4\}$, then the regression parameters when regressing b on a are given as $-(K^{bb})^{-1}K^{ba}$. Thus the slope parameters for y_3 and y_4 become identical,

$$E(y_i|y_1, y_2) = a_i + (c_1/c_3)y_1 + (c_2/c_3)y_2 \text{ for } i = 3, 4,$$

meaning that the regression lines are parallel.

Another property of this model is that some partial correlations are restricted to being equal. For example it follows directly from (1) that

$$\begin{aligned} \rho_{31|24} &= \rho_{41|23} = -c_3/\sqrt{k^{11}c_1} \text{ and} \\ \rho_{32|14} &= \rho_{42|13} = -c_3/\sqrt{k^{11}c_2}. \end{aligned}$$

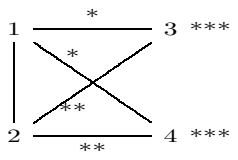

Figure 1: Graphical Gaussian model with the additional restrictions that (*) $k^{13} = k^{14} = c_1$, (**) $k^{23} = k^{24} = c_2$ and (***) $k^{33} = k^{44} = c_3$.

4 Restricted concentration models

To formalize the restrictions on the elements of K, let the edge set E be partitioned into non–empty disjoint subsets $E_1, \ldots, E_S$ each containing one or more edges. Let $E_1, \ldots, E_s$ denote those subsets containing more than one edge (those edges are said to be *marked*), and $E_{s+1}, \ldots, E_S$ be those containing only one edge (those edges are *unmarked*). Let $\Gamma_1, \ldots, \Gamma_T$ be a similar partitioning of Γ into sets of nodes, some of which are marked and some unmarked. A natural way to represent this situation graphically is to colour the edges (vertices) in the graph such that all edges (vertices) in the same edge (vertex) set have the same colour.

Let $\mathcal{R} = \{E_1, \ldots E_S, \Gamma_1, \ldots, \Gamma_T\}$ be the collection of such restrictions on K. The elements of $\mathcal{R}$ are in 1–1 correspondence with the parameters $\theta = (\theta_1, \ldots, \theta_{S+T})$ in $K = K(\theta)$. The *edge sets* $E_1, \ldots E_S$ and *vertex sets* $\Gamma_1, \ldots, \Gamma_T$ define a RCM.

5 The NIPS (Newton+IPS) algorithm

The algorithm is a combination of the classical IPS algorithm for graphical Gaussian models, Lauritzen (1996), p. 134 and the modified Newton procedure of Jensen, Johansen and Lauritzen (1991) (hereafter JJL91), see also Lauritzen (1996), p. 269.

Let $\hat{\Sigma} = \hat{K}^{-1}$ denote the current estimate of Σ at any time during the iteration, and let $f = n - 1$ where n is the number of observations, $SSD = \sum_{s=1}^{n}(y_s - \bar{y})(y_s - \bar{y})^\top$ and $S = SSD/f$.

5.1 Newton algorithm

The simplest version of the algorithm (which is just a specific version of the modified Newton algorithm of JJL91) is as follows:

Repeatedly loop through $\mathcal{R}$ until convergence doing the following: For each $s \in \mathcal{R}$ define the $d \times d$ matrix K^s as follows: 1) If s is an edge set, then K^s has entries $K^s_{\alpha\beta} = 1$ if $\{\alpha, \beta\} \in s$ and 0 otherwise. Thus K^s is the incidence matrix for the graph (Γ, s). 2) If s is a vertex set then K^s is a diagonal matrix with entries $K^s_{\alpha\alpha} = 1$ if $\alpha \in s$ and 0 otherwise. For convenience we shall identify a vertex α with a set $\{\alpha, \alpha\}$ such that vertex sets and edge sets can be treated simultaneously in the following.

Define the discrepancy $\Delta = tr(K^s\hat{\Sigma}) - tr(K^s S)$. For each element s do a sequence of Newton steps

$$\begin{aligned} \theta_s^{n+1} &\leftarrow \theta_s^n + \frac{\Delta}{tr(K^s\hat{\Sigma}K^s\hat{\Sigma}) + f\Delta^2/2}, \\ k^{\alpha\beta} &\leftarrow \theta_s^{n+1} \text{ for all } \{\alpha, \beta\} \in s. \end{aligned} \qquad (2)$$

The substitution (2) is repeated until convergence for the set s before moving on to the next set in $\mathcal{R}$. Thus the algorithm consists of two nested loops: 1) An outer loop running over the elements of $\mathcal{R}$ and 2) an inner loop maximizing L with respect to θ_s while keeping all other parameters fixed. Below it is shown that this algorithm in some cases can be speeded up by replacing the inner loop by a direct line search.

The likelihood equations are obtained as follows: With the definition of the matrices K^s for all $s \in \mathcal{R}$ given above, the concentration matrix can be written $K = \sum_s \theta_s K^s$. Let SS denote the sums–of–squares matrix. Then $tr(KSS) = \sum_s \theta_s tr(K^s SS)$. Let $t^s = \sum_s tr(K^s SS)$. Hence $(-t^1/2, \ldots, -t^{S+T}/2, \bar{y})$ is a set of canonical statistics, and these are to be equated with their expectation.

To do so, we exploit the following: The multivariate normal distribution is a regular k–dimensional exponential family. Therefore the maximum likelihood estimate (MLE) exists and is unique, provided that the

sufficient statistic is contained in its convex support. By Theorem 2 in Jensen *et al.* (1991), the MLE can be found by iteratively maximizing over each canonical parameter, keeping the others fixed. Note that when keeping all parameters but one at fixed values we get a regular one–dimensional exponential family. By Theorem 1 in JJL91, their modified Newton algorithm applied to a one–dimensional regular exponential family converges to the MLE for any starting value.

Following Lauritzen (1996), p. 133, $\hat{\mu} = \bar{y}$ so what remains is to maximize $L(\theta, \hat{\mu})$ over Θ which is an $S+T$ dimensional space restricted only by the requirement that $K(\theta)$ must be positive definite for all $\theta \in \Theta$. For any $\theta^* \in \Theta$ and any $s \in \mathcal{R}$, define

$$\Theta_s(\theta^*) = \{\theta \in \Theta | \theta_r = \theta_r^* \text{ for } r \neq s\}.$$

Then L is maximized by cyclically maximizing L over $\Theta_s(\theta^*)$, Lauritzen (1996), p. 270. For practical reasons we have chosen to fit the model on S rather than on SS. Following Lauritzen (1996) p. 259, $\tau^s = E(t^s) = -\frac{1}{2}tr(K^s\Sigma)$ and $v^s = Var(t^s) = \frac{1}{2}tr(K^s\Sigma K^s\Sigma)$. The modified Newton algorithm of JJL91 consists in updating θ as

$$\theta^{n+1} = \theta^n + \frac{t^s - \tau^s}{v^s + (t^s - \tau^s)^2}$$

which specializes to (2) in this context.

Convergence The parameter space Θ is restricted by $K(\theta)$ having to be positive definite. We have not shown that the Newton steps are guaranteed to keep K positive definite and this should therefore strictly speaking be checked at each Newton step, decreasing the step length appropriately if the condition is no longer satisfied, see JJL91. Empirical evidence suggests however, that K indeed remains positive definite.

5.2 IPS algorithm

For a GGM (without restrictions of the kind discussed in this paper) let $a = \{\alpha, \beta\}$ be an edge in the graph and let b denote the complement to a. Then in the IPS algorithm, see e.g. Lauritzen (1996) p. 134 ff, can be used for updating the parameters $k^{\alpha\alpha}, k^{\beta\beta}$ and $k^{\alpha\beta}$ by updating the 2×2 submatrix K^{aa} of K as

$$K^{aa} \leftarrow (S^{aa})^{-1} + K^{ab}(K^{bb})^{-1}K^{ba}. \qquad (3)$$

Note that in this step both the conditional variances and conditional covariances are updated. This IPS step maximizes the likelihood over the particular section of the parameter space given by $k^{\alpha\alpha}, k^{\beta\beta}$ and $k^{\alpha\beta}$ and thus no iteration is needed. This operation can also be performed on a single vertex α, which gives an update of the 1×1 submatrix $K^{\alpha\alpha}$.

5.3 NIPS algorithm

Considerable computational savings can be achieved by combining the Newton sequence (2) with the IPS step (3) and this combination constitutes the NIPS (=Newton+IPS) algorithm. The combination is straight forward and most easily explained by an example: The graph in Figure 2 has cliques [12][23][34][45]. The asterisks indicate that the edges [12] and [23] and the vertices 2 and 3 are *marked*, i.e. the restrictions $k^{12} = k^{23}$ and $k^{22} = k^{33}$.

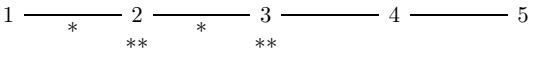

Figure 2: RCM with the additional restrictions that (*) $k^{12} = k^{23}$ and (**) $k^{22} = k^{33}$.

The marked entries can be updated using the Newton sequence while k^{11} is unrestricted and can be updated using an IPS step on a 1×1 matrix. The edge [45] (comprising the parameters k^{44}, k^{45} and k^{55} can also be updated in a single IPS step on a 2×2 matrix. Left to consider is therefore only k^{34}. Even though no restriction is put onto this parameter it can not immediately be updated using an IPS step (3) because that would also update k^{33} (and k^{44}) which is constrained. Therefore this parameter is updated using a Newton sequence. This constitutes one full cycle of the inner loop of the NIPS algorithm. Note that it is easy to keep track of such restrictions: Whenever an edge $\{\alpha, \beta\}$ contains a marked vertex, the edge must itself be marked.

Computational Savings The following considerations can lead to substantial computational savings:

1. Computational savings can be achieved when calculating $\Delta = tr(K^s\hat{\Sigma}) - tr(K^sS)$. The incidence matrix K^s serves to pick out (and sum the correct way) the relevant entries of S and $\hat{\Sigma}$. For a fixed edge set $s \in \mathcal{R}$, let a denote the set of vertices in s and let b be the complement of a. Let $\tilde{A}^s$ be the incidence matrix for the graph (a, s) and let finally $\hat{\Sigma}^{aa}$ and S^{aa} denote the corresponding submatrices of $\hat{\Sigma}$ and S. It it then straight forward to see that $tr(K^sS) = tr(\tilde{A}^sS^{aa})$ and hence $\Delta = tr(\tilde{A}^s\hat{\Sigma}^{aa}) - tr(\tilde{A}^sS^{aa})$. The modification when s is a vertex set is straight forward.

2. After updating entries of K in a NR step, one need not find $\Sigma = K^{-1}$. The relevant part Σ^{aa} can be found as $(K^{aa} - K^{ab}(K^{bb})^{-1}K^{ba})^{-1}$, and here it is noted that 1) $K^{ab}(K^{bb})^{-1}K^{ba}$ is fixed throughout the whole Newton sequence and 2) the dimension of Σ^{aa} is often much smaller than the dimension

of Σ. A similar construct can be used when calculating the value of the likelihood function.

3. Convergence is sometimes speeded up when replacing the Newton steps in (2) by an alternative line search algorithm of the form

$$\begin{aligned} \theta_s^{n+1} &\leftarrow \theta_s^n + \alpha \cdot p, \\ k^{\alpha\beta} &\leftarrow \theta_s^{n+1} \text{ for all } \{\alpha, \beta\} \in s \end{aligned} \qquad (4)$$

where $p = \frac{\Delta}{tr(A^s \hat{\Sigma} A^s \hat{\Sigma}) + f\Delta^2/2}$ and α is chosen to maximize L in the direction defined by $\theta_s^n + tp$.

4. If a clique consists exclusively of unmarked edges/vertices, then it is more computationally efficient to update the entire clique using IPS at one time rather than working the way through the edges one at the time.

6 Implementation

The algorithm has been implemented in the general statistical package `R`, R Development Core Team (2004).

7 Model selection issues

The number of different models which can be formed by colouring edges/vertices in a given graph is enormous. To illustrate the complexity, consider graphs with vertices, 1,2 and 3 (for which there are 8 different graphs). There are 5 possible vertex sets: $\{123\}, \{12, 3\}, \{1, 23\}, \{13, 2\}$ and $\{1, 2, 3\}$. A tedious calculation shows that there are in total (over all 8 graphs) 15 possible vertex sets giving $5 \times 15 = 75$ different models! Therefore, good model selection strategies become important. Here we shall just outline some ideas:

Often in model selection in graphical models one consider the operations `dropEdge` and `addEdge`. For RCMs there are four additional operations which are natural to consider: `joinEdgeSet` and `splitEdgeSet` (and similarly for vertices). In connection with a backward model search where edges are successively deleted it is tempting to supplement with the possibility of joining two edge sets. If there are p edge sets then there are $p(p-1)/2$ pairwise comparisons of the corresponding parameters and this can be done by e.g. calculating Wald statistics (which requires $Var(\hat{\theta})$ to be computed).

A more brute force approach is to search for a graphical model and then apply a clustering algorithm to the diagonal of K and to the non–zero off–diagonal elements of K.

One motivation for considering RCMs is applications where data is sparse, i.e. where $n < d$. In this case S is singular and hence $K = S^{-1}$ does not exist. One option in this case is to start from the independence model and do a forward selection possibly supplied with joining operations as discussed above.

8 Example: measurements on pig carcasses

To illustrate the developments in this paper we consider a prediction problem: In slaughter pig production, prediction of the lean meat content is important 1) to ensure fair payment to the producers and 2) to ensure an appropriate processing of the meat afterwards. The task is to predict the lean meat percentage y on the basis of a set of predictor variables denoted by x. In modern carcass grading, the predictor variables are often obtained e.g. by ultra sound measurements on the carcass and hence the number of predictor variables can be very large – and much larger than the sample size.

For simplicity, we here consider the `carcass` data set contained in the `mimR` package in `R`, see Højsgaard (2004). This data set contains measurements of the thickness of the meat and fat layer at three locations on the back of 340 carcasses. The data also contains the lean meat percentage determined by dissection. The response variable is the meat percentage, $y = MP$ while x denotes the measurements of thickness of meat and fat layers. The regression coefficients for the prediction are $\Sigma_{yx}\Sigma_{xx}^{-1} = -(K^{yy})^{-1}K^{yx}$. The problem in such prediction problems is that either Σ_{xx} is singular or it is very ill–conditioned because the predictor variables often are very correlated.

To accommodate for this, one often make a principal component regression or a partial least squares regression to obtain the regression coefficients. Other alternatives are ridge regression and the lasso, see e.g. Hastie, Tibshirani and Friedman (2001), pp. 59 for a description of these methods.

8.1 Selection of different models

The saturated model (which has Table 1 as concentration matrix) is in the following denoted $\mathcal{M}1$. Table 1 shows that the fat–concentration parameters all tend to be of the same size (conditional variances as well as covariances) and so do the meat concentration parameters. Similarly, the concentration parameters between the fat measurements and the lean meat percentage appear identical and so do (to a lesser extent) the concentration parameters between the meat measurements and the lean meat percentage. The model

with these constraints is denoted $\mathcal{M}1r$ and the estimated concentration matrix is shown in Table 2.

Table 1: Empirical concentration matrix for the carcass data (multiplied by 10).

	F1	F2	F3	M1	M2	M3	MP
F1	4.36	-1.99	-1.58	0.28	-0.73	0.41	0.99
F2	-1.99	5.35	-2.09	-0.26	0.64	-0.53	0.88
F3	-1.58	-2.09	5.57	-0.56	-0.06	0.26	0.71
M1	0.28	-0.26	-0.56	1.58	-0.60	-0.56	-0.33
M2	-0.73	0.64	-0.06	-0.60	1.35	-0.88	-0.04
M3	0.41	-0.53	0.26	-0.56	-0.88	1.57	-0.14
MP	0.99	0.88	0.71	-0.33	-0.04	-0.14	2.63

Table 2: Estimated concentration matrix for the carcass data (multiplied by 10) under the model $\mathcal{M}1r$ with parameters restricted to being equal.

	F1	F2	F3	M1	M2	M3	MP
F1	4.83	-1.77	-1.77	0.30	-0.86	0.42	0.88
F2	-1.77	4.83	-1.77	-0.27	0.58	-0.37	0.88
F3	-1.77	-1.77	4.83	-0.30	-0.04	0.04	0.88
M1	0.30	-0.27	-0.30	1.40	-0.64	-0.64	-0.16
M2	-0.86	0.58	-0.04	-0.64	1.40	-0.64	-0.16
M3	0.42	-0.37	0.04	-0.64	-0.64	1.40	-0.16
MP	0.88	0.88	0.88	-0.16	-0.16	-0.16	2.64

Starting with the independence model and doing a forward selection we get the model $\mathcal{M}2$ with concentration matrix in Table 3. Then we applied a clustering algorithm to the diagonal and to the off–diagonals to identify possible edge sets and vertex sets. Inspired by Table 1, we asked for 3 clusters on the diagonal and 5 clusters on the off–diagonal. The model with these restrictions is $\mathcal{M}2r$ and the estimated concentrations are presented in Table 4.

This scheme was repeated with a backward selection starting from the saturated model giving model $\mathcal{M}3$. Clustering the entries as described above gave $\mathcal{M}3r$. (The estimated concentration matrices have been omitted).

Table 3: Estimated concentration matrix for the carcass data (multiplied by 10) for $\mathcal{M}2$.

	F1	F2	F3	M1	M2	M3	MP
F1	4.06	-1.68	-1.53	0.00	-0.18	0.00	1.08
F2	-1.68	5.04	-2.12	0.00	0.00	0.00	0.78
F3	-1.53	-2.12	5.54	-0.39	0.00	0.00	0.75
M1	0.00	0.00	-0.39	1.52	-0.56	-0.56	-0.27
M2	-0.18	0.00	0.00	-0.56	1.22	-0.79	0.00
M3	0.00	0.00	0.00	-0.56	-0.79	1.51	-0.26
MP	1.08	0.78	0.75	-0.27	0.00	-0.26	2.68

8.2 Model comparisons – predictive performance

To evaluate the feasibility of the various models, we took a cross validation approach as follows: Out of the 340 carcasses we took a random sample of size $N = 8, 10, 15, 20, 30$ and fitted the models to these training data. Then we predicted MP for the validation data consisting of $340–N$ carcasses and calculated the mean squared prediction error (MSPE) defined as $\frac{1}{340-N}\sum_i (y_i - \hat{y}_i)^2$. This scheme was repeated $M = 5$ times and at the end average MSPE was calculated. To provide a benchmark for comparison we also made a principal component regression (PCR) and a partial least squares regression (PLS). Højsgaard, Jørgensen, Olsen and Busk (2004) have found that 3 components were optimal in PLS and PCR for predictions of these data, and therefore 3 components have been used here. To ease the comparison the MSPEs were all calculated relative to the MSPE for the PCR model.

Table 4: Estimated concentration matrix for the carcass data (multiplied by 10) for $\mathcal{M}2r$.

	F1	F2	F3	M1	M2	M3	MP
F1	4.60	-2.00	-2.00	0.00	-0.20	0.00	0.77
F2	-2.00	4.60	-1.11	0.00	0.00	0.00	0.77
F3	-2.00	-1.11	4.60	-0.20	0.00	0.00	0.77
M1	0.00	0.00	-0.20	1.06	-0.47	-0.47	-0.20
M2	-0.20	0.00	0.00	-0.47	1.06	-0.47	0.00
M3	0.00	0.00	0.00	-0.47	-0.47	1.06	-0.20
MP	0.77	0.77	0.77	-0.20	0.00	-0.20	2.39

8.3 Results

The relative MSPEs are presented in Table 5. Within each sample size, we find the following: It is always beneficial to reduce the saturated model $\mathcal{M}1$ to the restricted model $\mathcal{M}1r$, and for small samples ($N = 8, 10$) the improvement is quite dramatic. (Note that when $N = 8$ the saturated model is just identifiable as there are 7 variables in the model).

A comparison of models $\mathcal{M}i$ and $\mathcal{M}ir$ for $i = 2, 3$ yields no clear picture, but it suggests that there is a place for refinement of the brute force clustering approach used in getting from $\mathcal{M}i$ and $\mathcal{M}ir$. For each sample size, one of the RCMs always performs at least as well or better than the traditional regression methods PLS and PCR. Finally it is noted that when sample size increases the models perform more and more similarly, which was to be expected.

8.4 Computing time

Compared with the IPS algorithm used for GGMs the NIPS algorithm presented here is somewhat more time consuming. For example, fitting $\mathcal{M}3$ (which is a GGM) took 1.27 seconds while fitting the RCM $\mathcal{M}3r$ took 4.87 seconds.

Table 5: Relative mean squared prediction error (MSPE) (calculated relative to MSPE for principal component regression) for different models and different sizes of the training data sets.

	Sample size				
	8	10	15	20	30
$\mathcal{M}1$	4.53	1.08	1.10	1.06	1.01
$\mathcal{M}1r$	0.99	0.92	0.99	0.99	0.99
$\mathcal{M}2$	1.11	1.03	1.04	1.03	1.01
$\mathcal{M}2r$	1.18	0.99	1.03	0.99	1.00
$\mathcal{M}3$	1.20	1.04	1.04	1.07	1.00
$\mathcal{M}3r$	1.20	0.95	1.01	1.00	1.01
PLS	1.16	1.01	1.03	1.04	1.01
PCR	1.00	1.00	1.00	1.00	1.00

9 Discussion and directions for future work

This paper has presented an estimation algorithm for restricted concentration models (RCMs), and it has been proven empirically that important gains in terms of prediction precisions can be achieved from such models.

It is emphasized, that to use the result in JJL91 we should strictly speaking check that the concentration matrix stays positive definite in each step (2) and, if not, only move half of the distance to the associated boundary point of the parameter space. We have not seen an example where the positive definiteness has been violated, but we have not been able to prove theoretically that this cannot happen. For practical purposes we therefore suggest that this check is only performed occasionally.

To make RCMs of practical importance, it is important to investigate possible model selection strategies for RCMs, and this is a subject of future work. In this connection it will become important to make a fast implementation of the NIPS algorithm.

References

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

Besag, J. E. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society, Series B* **36**, 192–236.

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

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer.

Højsgaard, S. (2004). The mimR package for graphical modelling in R. *Journal of Statistical Software* .

Højsgaard, S., Jørgensen, E., Olsen, E. V. and Busk, H. (2004). A comparison of latent variable models and partial least squares regression – with an application to pig carcass grading. *Livestock Production Science* Manuscript submitted.

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

Lauritzen, S. L. (1996). *Graphical Models*. Oxford University Press.

R Development Core Team (2004). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-00-3.