Variable selection and Covariance selection in multivariate Regression Models.

Edward Cripps Chris Carter Robert Kohn *

January 1, 2004

Abstract

This article provides a general framework for Bayesian variable selection and covariance selection in a multivariate regression model with Gaussian errors. By variable selection we mean allowing certain regression coefficients to be zero. By covariance selection we mean allowing certain elements of the inverse covariance matrix to be zero. The prior for the regression coefficients is specified in such a way that, given the covariance matrix and a subset of predictor variables, it is noninformative relative to the likelihood. We estimate all the model parameters by model averaging using a Markov chain Monte Carlo simulation method. The methodology is illustrated by applying it to four real data sets. The effectiveness of variable selection and covariance selection in estimating the multivariate regression model is assessed by using four loss functions and four simulated data sets. Each of the simulated data sets is based on parameter estimates obtained from a corresponding real data set.

Key Words: Crossectional Regression; Longitudinal data; Model averaging; Markov chain Monte Carlo.

*Edward Cripps is PhD student, Department of Statistics, University of New South Wales, Sydney NSW 2052 Australia (ecripps@maths.unsw.edu.au). Chris Carter is Senior Research Scientist, CSIRO, Australia, (Chris.Carter@csiro.au). Robert Kohn is Professor, Faculty of Commerce and Economics, University of New South Wales, Sydney NSW 2052 Australia (R.Kohn@unsw.edu.au).

1 Introduction

This article provides a general framework for estimating a multivariate regression model. The methdology is Bayesian and allows for variable selection and covariance selection, as well as allowing some of the dependent variables to be missing. By variable selection we mean that the regression model allows for some of the regression coefficients to be identically zero. By covariance selection we mean that the model allows the off-diagonal elements of the inverse of the covariance matrix to be identically zero. We estimate all functionals of the parameters by model averaging, i.e., by taking a weighted average of the values of the functional, where the average is over the allowable configurations of the regression coefficients and the covariance matrix. The computation is carried out using a Markov chain Monte Carlo simulation method.

There is an extensive literature on Bayesian variable selection. In the univariate case, see Mitchell and Beauchamp (1988), George and McCulloch (1993, 1997), Raftery, Madigan, and Hoeting (1997), Smith and Kohn (1996), Kohn, Smith, and Chan (2001) and Hoeting, Madigan, Raftery, and Volinsky (1999) for further discussions and citations. Raftery et al. (1997) argue that if prediction is the goal of the analysis, then it may be better to use model averaging rather trying to find the "optimal" subset of variables by variable selection. Further support for model averaging is given by Brieman (1996) who argues that subset selection is unstable in the univariate linear regression case. In a series of papers, Brown, Vannucci, and Fearn (1998); Brown, Fearn, and Vannucci (1999); Brown, Vanucci, and Fearn (2002) consider variable selection and model averaging for multivariate regression models. Our approach to variable selection is multivariate models is based on Smith and Kohn (1996) and Kohn et al. (2001). We note that the model for the mean in Brown et al. (1998, 1999, 2002) is a special case of our regression model for the mean because in their model the same covariates appear in all the equations and a covariate is either in all equations or none of them. Brown et al. (1998, 1999, 2002) do not consider covariance selection.

Efficiently estimating a covariance matrix is a difficult statistical problem, especially when the dimension of the covariance matrix is large relative to the sample size (e.g. Stein, 1956; Dempster, 1969) because the number of unknown parameters in the covariance matrix increases quadratically with dimension and because it is necessary to keep the estimate of the covariance matrix positive definite. Early work on the efficient estimation of covariance matrices is by Stein (see Stein (1956) and other unpublished papers by Stein that are cited by Yang and Berger (1994)) and Efron and Morris (1976). For more recent work see Leonard and Hsu (1992) and Chiu, Leonard, and Tsui (1996) who modeled the matrix logarithm of the covariance matrix. Yang and Berger (1994) used a Bayesian approach based on a spectral decomposition of the covariance matrix. Pourahmadi (1999, 2000) estimated the covariance matrix by parameterizing the Cholesky decomposition of its inverse. Smith and Kohn (2002) used a prior that allows for zero elements in the strict lower triangle of the Cholesky decomposition of the inverse of the covariance matrix to obtain a parsimonious representation of the covariance matrix. Although the Cholesky decomposition applies to a general covariance matrix, it is most useful and interpretable for longitudinal data. Barnard, McCulloch, and Meng (2000) modeled the covariance matrix in terms of standard deviations and correlations and proposed several shrinkage estimators. Further results and simulation comparisons are given by Daniels and Kass (1999).

Dempster (1972) proposed estimating the covariance matrix parsimoniously by identifying zero elements in its inverse. He called models for the covariance matrix obtained in this way covariance selection models. His idea was that in many statistical problems the inverse pf the covariance matrix has a large number of zeros in its off-diagonal elements and these should be exploited in estimating the covariance matrix. There is a natural interpretation of such zeros: the i, jth element of the inverse is zero if and only if the partial correlation between the ith and j variables is 0, (e.g. Whittaker, 1990). This means that a covariance selection model can be interpreted as a Gaussian graphical model, (e.g. Lauritzen, 1996). Giudici and Green (1999) gave a Bayesian approach for estimating the structure of a decomposable graphical model. Their approach can be used to efficiently estimate a covariance matrix with a decomposable graphical structure, and possibly more general covariance matrices. Wong, Carter, and Kohn (2003) give a Bayesian approach for estimating a covariance selection model and it is their approach that we use in the paper. [1]

Our article makes the following contributions to the literature. First, it combines model averaging over the regression coefficients with model averaging over the inverse covariance matrix in the multivariate normal linear regression model. Second, it presents a more general approach to variable selection in the mean of the regression model than that given by Brown et al. (1998, 1999, 2002). Third, it illustrates the methdology using four real examples. Fourth, it studies whether model averaging based on variable selection or covariance selection or both improves

[1] should we reference Roverato and Delllaportas et al.

the estimation of the multivariate regression model. The assessment is based on a study of performance of four simulated examples using four loss functions. Each of the simulated examples is based on the estimates of one of the four real examples. The four loss functions consider separately the estimates of the predictive distribution, the estimates of the covariance matrix only, the estimates of the regression coefficients and the estimates of the fitted values.

The article is organized as follows. Section 2 describes the multivariate model and the priors for variable and covariance selection. Section 3 discusses the sampling scheme and computational issues. Section 4 describes the real data sets and reports on the analysis of the real data. Section 5 describes the simulated data based on the real examples and presents the results of the simulation.

2 Model Description

2.1 Introduction

For $t = 1, \ldots, n$ let Y_t be a $p \times 1$ vector of responses, X_t be a $p \times q$ design matrix and β the $q \times 1$ vector of regression coefficients. We assume the model

$$Y_t = X_t\beta + e_t, \qquad e_t \sim N(0, \Sigma)\,. \tag{2.1}$$

Let $\gamma = (\gamma_1, \ldots, \gamma_k)$ be a vector of binary variables such that the ith column of X_t is included in the regression if $\gamma_i = 1$ and is excluded if $\gamma_i = 0$. We write $X_{t,\gamma}$ for the matrix that contains all columns of X_t for which $\gamma = 1$ and β_γ for the corresponding subvector of regression coefficients. Therefore, the vector γ indexes all the mean functions for the regression model (2.1). Conditional on γ, (2.1) becomes

$$Y_t = X_{t,\gamma}\beta_\gamma + e_t, \qquad e_t \sim N(0, \Sigma)\,. \tag{2.2}$$

Models (2.1) and (2.2) contain as a special case the multivariate model

$$Y_t = Bx_t + e_t\ , \tag{2.3}$$

where B is a matrix of regression coefficients and x_t is a vector of covariates. It is clear that model (2.3) is a special case of model (2.1) by taking $X_t = x_t' \otimes I_p$ and $\beta = \mathrm{vec}(B)$. We note that $\otimes$ means Kronecker product and $\mathrm{vec}(B)$ is the vector obtained by stacking the columns of A beneath each other. The model (2.3) is used extensively in multivariateregression analysis

Mardia, Kent, and Bibby (e.g. 1979, p.) and in particular by Brown et al. (1998, 1999, 2002). We note that Brown et al. (1998, 1999, 2002) do variable selection on x_t which means that when they drop a covariate they drop a whole column of the matrix B. We show in section xx how this can be doen in general for the model (2.1).

We follow Wong et al. (2003) and parameterise $\Sigma^{-1} = \Omega$ as

$$\Omega = TCT, \tag{2.4}$$

where T is a diagonal matrix with $T_i = \Omega_{ii}^{0.5}$ such that T_i^2 is the inverse of the partial variance of $Y_{i,t}$. C is a correlation matrix. Also, the partial correlation coefficients ρ^{ij} are given by

$$\rho_{ij} = \frac{-\Omega_{ij}}{(\Omega_{ii}\Omega_{jj})^{0.5}} = -C_{ij} \ , \tag{2.5}$$

i.e. the off-diagonal elements of C are the negative of the partial correlation coefficients.

Let $Y = (Y_1', \ldots, Y_n{}')'$. From (2.2) the likelihood is

$$\begin{aligned} p(Y|\beta,\gamma,\Sigma) &= |2\pi\Omega^{-1}|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum_{t=1}^{n}(Y_t - X_{t,\gamma}\beta_\gamma)'\Omega(Y_t - X_{t,\gamma})\beta_\gamma)\right\} \\ &\propto |\Omega|^{\frac{n}{2}} \exp\left\{-\frac{1}{2}\text{trace}(\Omega S_y)\right\} \\ &\propto |T|^n |C|^{\frac{n}{2}} \exp\left\{-\frac{1}{2}\text{trace}(TCTS_y)\right\}, \end{aligned} \tag{2.6}$$

where $S_y = \sum_{t=1}^n (Y_t - X_{t,\gamma}\beta_\gamma)(Y_t - X_{t,\gamma}\beta_\gamma)'$.

2.2 Prior for the Regression Coefficients

Similarly to Smith and Kohn (1996), we define the prior for the the regression coefficients as being noninformative with respect to the likelihood and with a mode at zero. To motivate the prior, it is useful to rewrite the likelihood as follows.

$$\begin{aligned} p(Y|\beta,\gamma,\Omega) &= |2\pi\Omega|^{\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum_{t=1}^{n}(Y_t - X_{t,\gamma}\beta_\gamma)'\Omega(Y_t - X_{t,\gamma}\beta_\gamma)\right\} \\ &= |2\pi\Omega|^{\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum_{t=1}^{n} Y_t'\Omega Y_t - 2\beta_\gamma' \sum_{t=1}^{n} X_{t,\gamma}'\Omega Y_t + \beta_\gamma' \sum_{t=1}^{n} X_{t,\gamma}'\Omega X_{t,\gamma}\right\} \\ &= |2\pi\Omega|^{\frac{n}{2}} \exp\left\{-\frac{1}{2} YtOY - 2\beta_\gamma' XtOY_\gamma + \beta_\gamma' XtOX_\gamma\beta_\gamma\right\}, \end{aligned} \tag{2.7}$$

where

$$YtOY = \sum_{t=1}^{n} Y_t'\Omega Y_t, \quad XtOY_\gamma = \sum_{t=1}^{n} X_{t,\gamma}'\Omega Y_t, \quad \text{and} \quad XtOX_\gamma = \sum_{t=1}^{n} X_{t,\gamma}'\Omega X_{t,\gamma}.$$

As a function of β_γ, the likelihood is Gaussian with a mean of $(XtOX_\gamma)^{-1}XtOY_\gamma$ and covariance matrix $(XtOX_\gamma)^{-1}$.

Conditional on the binary indicator vector and the covariance matrix we take the prior for β_γ as

$$\beta_\gamma|\Sigma, \gamma \sim N(0, c(XtOX_\gamma)^{-1}) \tag{2.8}$$

and set $c = n$ such that the prior variance of β_γ stays approximately the same as n increases.

From (2.7) and (2.8) we can write the density of β_γ conditional on Y, Σ and γ as

$$\beta_\gamma|\gamma, \Sigma \sim N\left(\frac{c}{1+c}(XtOX_\gamma)^{-1}XtOY_\gamma, \frac{c}{1+c}(XtOX_\gamma)^{-1}\right) \tag{2.9}$$

2.3 Prior for the vector of binary indicator variables

We first define

$$q_\gamma = \sum_{i=1}^{q} \gamma_i,$$

which is the number of columns contained in X_t specified by $\gamma_i = 1$.

We assume that γ and Σ are independent apriori and as in Kohn et al. (2001) we specify the prior for γ as

$$p(\gamma|\pi) = \pi^{q_\gamma}(1-\pi)^{q-q_\gamma}, \qquad \text{with} \qquad 0 \leq \pi \leq 1. \tag{2.10}$$

We set the prior for π as uniform, i.e. $p(\pi) = 1$ for $0 \leq \pi \leq 1$, so that

$$\begin{aligned} p(\gamma) &= \int p(\gamma|\pi)p(\pi)d\pi \\ &= \int \pi^{q_\gamma}(1-\pi)^{q-q_\gamma}d\pi \\ &= \mathrm{B}(q_\gamma + 1, q - q_\gamma + 1) \end{aligned}$$

where B is the beta function defined by

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

The likelihood for γ and Σ, with β_γ integrated out is

$$\begin{aligned} p(Y|\gamma, \Sigma) &= \int p(Y|\beta, \Sigma, \gamma)p(\beta_\gamma|\gamma, \Sigma)d\beta_\gamma \\ &\propto (1+c)^{-\frac{q_\gamma}{2}} \exp\left\{-\frac{1}{2}\left(YtOY - \frac{c}{1+c}XtOY_\gamma' XtOX_\gamma^{-1} XtOY_\gamma\right)\right\}, \end{aligned} \tag{2.11}$$

We can write the density of γ conditional on Y and Σ as

$$p(\gamma|Y,\Sigma) \propto p(Y|\Sigma,\gamma)p(\gamma),$$

and we use this density to update γ in the Markov chain Monte Carlosimulation.

2.4 Prior for Ω_{ii}

Following Wong et al. (2003), we take the prior for Ω_{ii} as a gamma distibution such that

$$\Omega_{ii} \propto \Omega_{ii}^{\tau} \exp\{-\nu\Omega_{ii}\},$$

which means that the prior for T_i is

$$\begin{aligned} p(T_i) &\propto p(\Omega_{ii})\frac{d\Omega_{ii}}{dT_i} \\ &\propto T_i^{2\tau-1}\exp\{-\nu T_i^2\} \end{aligned} \qquad (2.12)$$

To ensure the prior is noninformative we follow Wong et al. (2003) and set $a = 10^{-10}$ and $b = 10^{-8}$ in the rest of the article.

From (2.6) and (2.7) we have

$$\begin{aligned} p(T_i|Y,T_{\{-i\}},C,\beta,\gamma) &\propto P(Y|T,C,\beta,\gamma)p(T_i) \\ &\propto T_i^n \exp\left\{-\frac{1}{2}\left(T_i^2(S_y)_{ii} + 2T_i\sum_{j\neq i}^{p}(S_y)_{ij}C_{ij}T_j\right)\right\} \\ &\propto T^{n_\tau}\exp\{-aT_i^2 - 2bT_i\}, \end{aligned} \qquad (2.13)$$

where $n_\tau = n + 2\alpha - 1$, $a = \frac{(S_y)_{ii}}{2} + \nu$ and $b = 1/2\sum_{j\neq i}^{p}(S_y)_{ij}C_{ij}T_j$. This is the conditional density we use to generate T_i. Wong et al. (2003) show that this conditional density of T_i tends to normality as $n \to \infty$.

2.5 Prior for the partial correlation matrix C

We use the covariance selection prior for C in Wong et al. (2003), which allows the off-diagonal elements of C to be identically zero. This prior is similar in intention to the variable selection prior used in section 2.3, except that it is now necessary to keep the matrix C positive definite. For $j = 1,\ldots,p$ and $i < j$, we define the binary variable $J_{ij} = 0$ if C_{ij} is identically zero and

$J_{ij} = 1$ otherwise. Let $J = \{J_{ij}, i < j, j = 1, \ldots, p\}$. These binary variables are analogous to the γ_i binary variables that we use for variable selection. Let

$$S(J) = \sum_{ij} J_{ij}, \quad i < j,$$

and let $r = p(p-1/2$ making $0 \leq S(J) \leq r$. Let C_p be the set of $p \times p$ positive definite correlation matrices. Let

$$V(J^*) = \int_{C \in C_p : J(C) = J^*} \left(\prod_{i \leq j, J_i j = 1} dC_{ij} \right)$$

be the volume of the positive definite region for C, given the constraints imposed by J^*, and let

$$\overline{V}(l) = \begin{pmatrix} r \\ l \end{pmatrix}^{-1} \left(\sum_{J : S(J) = l} V(J) \right)$$

be the average volume for regions with size l.

The hierarchical prior for C is given by

$$p(dC|J) = V(J)^{-1} dC_{J=1} I(C \in C_p), \tag{2.14}$$

$$p(J|S(J) = l) = \begin{pmatrix} r \\ l \end{pmatrix}^{-1} \frac{V(J)}{\overline{V}(l)}, \tag{2.15}$$

$$p(S(J) = l|\psi) = \begin{pmatrix} r \\ l \end{pmatrix}^{-1} \psi^l (1 - \psi)^{r-l}, \tag{2.16}$$

where $0 \leq \psi \leq 1$, $I(C \in C_p) = 1$ if C is a correlation and 0 otherwise and $C_{J=1} = \{C_{ij} : C_{ij} \neq 0\}$. The parameter ψ is the probability that $J_{ij} = 1$. For the remainder of this article we take $p(\psi) = 1$. For a more extensive discussion of this prior see Wong et al. (2003).

2.6 Missing values

The Bayesian methdology coupled with the Markov chain Monte Carlosimulation method make it straightforward to handle missing values in the dependent variable as part of the estimation problem, (e.g. Gelman et al., 2000, pp. 443-447). We assume that the observations are missing at random. Suppose that Y_t^m is the subvector of Y_t that is missing and $Y_t{}^o$ is the subvector that is observed. Then, $p(Y_t^m|Y_t{}^o, \beta, \Sigma)$ is Gaussian and it is straightforward to obtain $E(Y_t^m|Y_t{}^o, \beta, \Sigma)$ and $\text{var}(Y_t^m|Y_t{}^o, \beta, \Sigma)$, and hence to generate Y_t^m.

2.7 Permanently selected variables

We frequently wish to permanently retain some variables in the regression. For example, we may wish to retain all the intercept terms in the regression. We do so by setting the indicators γ for these variables to be identically one and setting q_γ in section 2.3 to be the sum of the γ_i, excluding those γ_i that are identically 1.

If we wish to estimate the model with no variable selection, then we would set the whole γ vector to 1.

2.8 Selecting variables in groups

In some problems it is useful to add or delete a group of variables rather than a single variable. For example, suppose that in the model (2.3) we wish to add or drop elements of the vector x_t, e.g. the second element of x_t. This is equivalent to dropping columns $p+1, \dots 2p$ of the matrix X_t or, equivalently, setting the second column of the coefficient matrix B to zero. Thus, in the model (2.3) we would take the first p columns of X_t as the first group, the next p columns as group 2, etc.

The binary indicator vector γ will now refer to groups of columns of X_t rather than individual columns. Similarly to section 2.7, we can choose to retain some groups permanently, e.g. , we may wish to retain all the intercepts in the model (2.3).

2.9 Noninformative prior on Σ

In sections 4 and 5 we compare the effect of covariance selection with a prior for Σ that does not allow for covariance selection. One way to do so is to use the prior for Σ given in sections 2.4 and 2.5, but always generating the C_{ij}. The effect on the estimation of Σ was carried out by Wong et al. (2003).

In our article we use the following prior for Σ when we do not wish to perform covariance selection.

$$p(\Sigma) \propto \det(\Sigma)^{-(p+1)/2}, \tag{2.17}$$

which implies that the prior for Ω is also of this form. The prior (2.17) is uninformative.

We now show the posterior distribution of Ω is proper. For conciseness, we do so when all regressors are in the model, but the extension to the case when there is variable selection is straightforward. It is not difficult to show that

$$
\begin{aligned}
p(\Omega|Y) &\propto p(Y|\Omega)p(\Omega) \\
&\propto \int p(Y|\Omega,\beta)p(\beta|\Omega)p(\Omega)d\beta \\
&\propto \det(\Omega)^{(n-p-1)/2} \times \\
&\quad \exp\left\{-\frac{1}{2}\left(YtOY - \frac{c}{1+c}YtOX(XtOX)^{-1}XtOY\right)\right\} . \qquad (2.18)
\end{aligned}
$$

With a little algebra we can show that

$$YtOY - \frac{c}{1+c}YtOX(XtOX)^{-1}XtOY \geq \frac{1}{c+1}YtOY$$

which implies that

$$p(\Omega|Y) \leq \det(\Omega)^{(n-p-1)/2}\exp\left\{-\frac{1}{2(c+1)}YtOY\right\} ,$$

with the right side of the expression above being a proper Wishart distribution. It follows that the posterior distribution of Ω, and hence Σ, is proper.

3 Sampling Scheme

Let Y_{miss} be the vector of missing values of Y, Y_{obs} the vector of observed values of Y. We generate $\beta, \gamma_i, i = 1, \ldots, q, T_i, i = 1, \ldots, p, C_{ij}, i \leq j$ and Y_{miss} using the following Markov chain Monte Carlo scheme.

1. $\gamma|Y_{\text{obs}}, T, Y_{\text{miss}}$;
2. $\beta_\gamma|Y_{\text{obs}}, T, C, \gamma, Ymiss$
3. $C|Y_{\text{obs}}, T, \gamma, \beta_\gamma, Y_{\text{miss}}$
4. $T|Y_{\text{obs}}, C, \gamma, \beta_\gamma, Y_{\text{miss}}$;
5. $Y_{\text{miss}^{[g]}}|Y_{\text{obs}}, \gamma, T, C, \beta_\gamma$

We generate the elements of γ one at a time by calculating

$$p(\gamma_i = 1|Y, T, C, \gamma_{\{-i\}}) = \frac{p(\gamma_i = 1|Y, T, C, \gamma_{\{-i\}})}{p(\gamma_i|Y, T, C, \gamma_{\{-i\}})}; \qquad (3.1)$$

see Kohn et al. (2001) for details. β_γ and Y_{miss} are generated from their conditionals as described above.

The c_{ij} and T_i are generated one element at a time using a $Y_{\text{Metropolis Hastings}}$ step. Details are given by Wong et al. (2003).

4 Real Data

This section studies both variable selection and covariance selection on real data by analyzing four data sets using two models. The first model carries out variable selectio but not covariance selection and uses the prior xxx for Σ. We call this the $NCSVS$ model. The second model does both variable selection and covariance selection and we call it the $CSVS$ model. We report the posterior means and standard errors of the regression coefficients, the posterior probabilities of including a predictor variable in the regression and the image plots of the estimated partial correlation matrix. Also included for the $CSVS$ model are the image plots for the posterior probabilities that the elements of the partial correlation matrix are non zero. The image plots are lighter where the matrix is sparser.

4.1 Cow Milk Protein Data

This is a longitudinal data set described in Diggle et al. (2002, page 5) who analyzed it to determine the effect of diet on the protein content in cow's milk[2]. The data was collected weekly for 79 cows. Each cow was assigned to one of three diets: barley (25 cows), mixture of barley and lupins (27 cows), or lupins (27 cows). The time was measured in weeks since calving and the experiment was terminated 19 weeks after the earliest calving, resulting in 38 observations with incomplete measurements. Diggle et al. (2002, page 6) note that calving may be associated with the milk protein content and as such the incomplete data should not be ignored. There are 11 other missing data values. We treat all the missing values as described in Section 2.6.

Exploratory analysis by Diggle et al. (2002, pages 5-9) suggests that the barley diet yields the highest mean protein content and the mixture diet yields the second highest protein content. Diggle et al. (2002, page 99) also point out that the mean response shows an initial drop in

[2] The data can be obtained from http://www.maths.lancs.ac.uk/∼diggle/lda/Datasets/.

the protein content in cows milk followed by a constant mean response over the majority of the experiment with a slight rise towards the end of the experiment. Diggle et al. (2002, pages 99 -103) create a model to determine whether diet affects the protein content in cow's milk. The model and results are presented in Diggle et al. (2002, pages 99- 103) with the conclusion that diet does affect the mean response and, of secondary interest, that there is no significant rise in the mean response towards the end of the experiment. We model the mean as a linear function of time, allowing for different regression coefficients for different diets and do not explore the presence of a nonlinear trend. For each cow t denote the vector of the milk protein content Y_t such that

$$Y_t = Z\beta_d + e_t, \qquad e_t \sim N(0, \Sigma)\,, \tag{4.1}$$

for $d = 1, 2, 3$ corresponding to the diets of barley, the mixture of barley and lupin, and lupin respectively. In (4.1) the predictor matrix Z and the vector of regression coefficients are

$$Z = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ \vdots & \vdots \\ 1 & 19 \end{bmatrix} \quad \text{and} \quad \beta_d = \begin{bmatrix} \beta_{d,0} \\ \beta_{d,1} \end{bmatrix},$$

where the second column in Z indexes the time in weeks since calving. Writing (4.1) in the more general form of (2.1) we have

$$Y_t = X_t\beta + e_t, \qquad e_t \sim N(0, \Sigma)\,, \tag{4.2}$$

where

$$\beta = (\beta_{1,0}, \beta_{1,1}, \beta_{2,0}, \beta_{2,1}, \beta_{3,0}, \beta_{31})',$$

and

$$X_t = \begin{cases} [Z \;\; 0_{19\times2} \;\; 0_{19\times2}] & \text{if } X_i \text{ recieves diet 1,} \\ [0_{19\times2} \;\; Z \;\; 0_{19\times2}] & \text{if } X_i \text{ recieves diet 2,} \\ [0_{19\times2} \;\; 0_{19\times2} \;\; Z] & \text{if } X_i \text{ recieves diet 3,} \end{cases}$$

where $0_{19\times2}$ denotes a matrix of zeros of size 19×2.

The estimated posterior means and standard errors for the regression coefficients and the posterior probabilities that the regression coefficients are non-zero are recorded in Table 1 for Model $NCSVS$ and in Table 2 for Model $CSVS$. As in Diggle et al. (2002), the magnitude of the intercept terms for both models for each of the diet groups agrees with the initial exploratory analysis. That is, the intercept for the barley diet is greater than the mixture diet which in turn is greater than the lupin diet. Tables 1 and 2 suggest the coefficients of the time trend are not significant statistically or practically for any of the three diets.

	$\beta_{1,0}$	$\beta_{1,1}$	$\beta_{2,0}$	$\beta_{2,1}$	$\beta_{3,0}$	$\beta_{3,1}$
post. mean	3.2468	0.0022	3.1972	-0.0007	3.1073	-0.0042
post. std. error	0.0663	0.0043	0.0621	0.0031	0.0711	0.0056
post. prob.	NA	0.33450	NA	0.2240	NA	0.4725

Table 1: Posterior means, standard errors and probabilities of being non-zero for the regression coefficients using Model $NCSVS$ for the cow milk protein data. NA means not available as the coefficient is always included.

	$\beta_{1,0}$	$\beta_{1,1}$	$\beta_{2,0}$	$\beta_{2,1}$	$\beta_{3,0}$	$\beta_{3,1}$
post. mean	3.3093	0.0005	3.2200	-0.0003	3.1389	-0.0032
post. std. error	0.0496	0.0023	0.0475	0.0023	0.0597	0.0049
post. prob.	NA	0.1775	NA	0.1485	NA	0.3720

Table 2: Posterior means, standard errors and probabilities of being non-zero for the regression coefficients using Model $CSVS$ for the cow milk protein data. NA means not available as the coefficient is always included.

The posterior means and standard errors of the difference between the intercepts for each diet are contained in Tables 3 and 4. The tables suggest that the difference for Model DVS between the barley and lupin diet are statistically significant, but insignificant for the remaining differences. The results for Model $NCSVS$ suggest that none of the intercepts are statistically different from each other.

	$\beta_{1,0} - \beta_{2,0}$	$\beta_{1,0} - \beta_{3,0}$	$\beta_{2,0} - \beta_{3,0}$
post. mean	0.0495	0.1394	0.0899
post. std. error	0.0639	0.0763	0.0683

Table 3: Posterior means and standard errors of the difference in the intercepts of the three diets using Model $NCSVS$ for the cow milk protein data.

	$\beta_{1,0} - \beta_{2,0}$	$\beta_{1,0} - \beta_{3,0}$	$\beta_{2,0} - \beta_{3,0}$
post. mean	0.0893	0.1704	0.0811
post. std. error	0.0535	0.0649	0.0618

Table 4: Posterior means and standard errors of the difference in the intercepts of the three diets using Model $CSVS$ for the cow milk protein data.

Figure 1 shows the image plots of the posterior means of the partial correlations and posterior probabilities that the elements in the partial correlation matrix are non-zero. Model $CSVS$ estimates a sparser partial correlation matrix than Model $NCSVS$ and the image plots of Model $CSVS$ suggest that the matrix of partial correlations is banded.

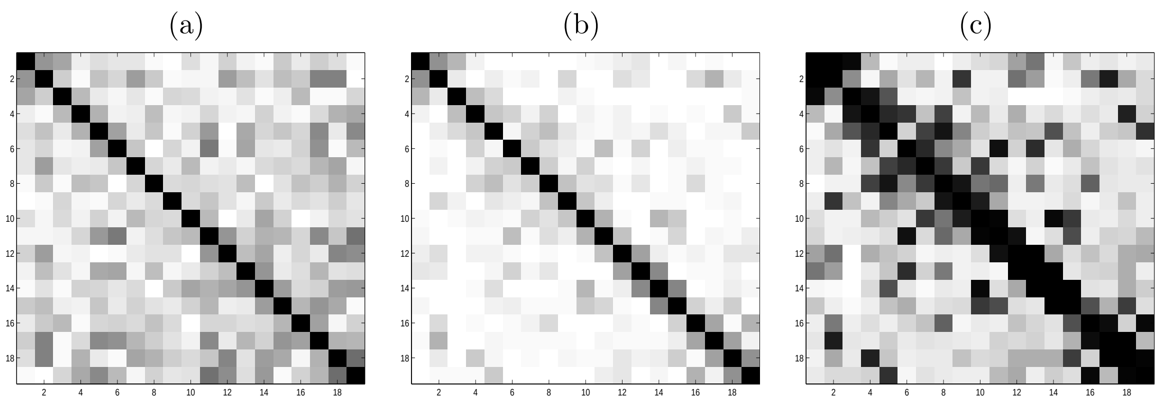

Figure 1: Image plots for the cow milk protein data. Panel (a) is the image plot of the partial correlation matrix estimated by Model $NCSVS$. Panel (b) is the image plot of the partial correlation matrix estimated by Model $CSVS$. Panel (c) is the image plot of the probabilities of the elements of the correlation matrix being non-zero as estimated by Model $CSVS$.

4.2 Hip replacement data

This data set contains observations on 30 patients who had hip replacements. The data set is in Crowder and Hand (1990, page 79). Each patient had their haematocrit levels measured four times, once before the operation and three times afterward. The main goals of this analysis are to examine differences in haematocrit levels between men and women who experienced hip replacements and to investigate whether an age effect is present. In addition to each patient's haematocrit levels their gender and age were recorded. We denote the age of individual i as a_i and take $s_i = 1$ if individual i is male and $s_i = -1$ if individual i is female.

The third time measurement had 19 out of 30 values missing which appeared to distort our results. Following Crowder and Hand (1990) we omit all the measurements at the third time point. There are two other missing values. One is for patient 8 at time 4 and the other is patient 15 at time 1. We deal with these missing values as described in Section **??**.

We assume the same mean structure as Crowder and Hand (1990) and allow for a different intercept for each time point and also include covariates for gender and age. Denote the heamatocrit levels as Y_{ti} for individual t at time i. We write the regression for observation Y_{ti} as

$$Y_{ti} = \alpha_i + \lambda_1 s_t + \lambda_2 a_t + e_{ti}, \text{ for } i = 1, 2, 3. \tag{4.3}$$

Writing (4.3) in the notation of (2.1) and denoting Y_t as the 3×1 vector of responses across time for individual t we have

$$Y_t = X_t \beta + e_t, \qquad e_t \sim N(0, \Sigma)\,, \tag{4.4}$$

where the predictor matrix for each individual X_t and the vector of regression coefficients are

$$X_t = \begin{bmatrix} 1 & 0 & 0 & s_i & a_i \\ 0 & 1 & 0 & s_i & a_i \\ 0 & 0 & 1 & s_i & a_i \end{bmatrix} \quad \text{and} \quad \beta = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \lambda_1 \\ \lambda_2 \end{bmatrix}.$$

The estimated posterior means and standard errors for the regression coefficients and the posterior probabilities that the regression coefficients are non-zero are recorded in tables 5 and 6. The estimated posterior probabilities are similar for Model $NCSVS$ and Model $CSVS$ when estimating (4.4). The estimated regression coefficients for the first three predictors for both Model

$NCSVS$ and Model $CSVS$ are comparable to those in Crowder and Hand (1990). The last two regression coefficients were estimated by Crowder and Hand (1990, page 81) as 0.807 and 0.031 respectively but do not greatly exceed their standard errors. Model $NCSVS$ and Model $CSVS$ also suggest these coefficients are not significantly different from zero and are therefore comparable with Crowder and Hand (1990, page 81). Tables 7 and 8 show the posterior means for the differences in the intercepts for haemocratic levels across each time period and show that the haemocratic levels are more significantly different when comparing the first time period with the second and third time periods than when comparing the difference between the second time period and the third time period.

	β_1	β_2	β_3	β_4	β_5
post. mean	38.8217	29.4162	31.6827	0.1674	0.0032
post. std. error	1.9205	1.8555	1.9003	0.4208	0.0257
post. prob.	NA	NA	NA	0.2165	0.1425

Table 5: Posterior means, standard errors and probabilities of being non-zero for the regression coefficients using Model $NCSVS$ for the hip replacement data. NA mean not available as the coefficient is always included.

	β_1	β_2	β_3	β_4	β_5
post. mean	38.6631	29.2353	31.5143	0.3098	0.0060
post. std. error	1.9948	1.9397	1.9946	0.5097	0.0269
post. prob.	NA	NA	NA	0.3545	0.1770

Table 6: Posterior means, standard errors and probabilities of being non-zero for the regression coefficients using Model $CSVS$ for the hip replacement data. NA mean not available as the coefficient is always included.

Figure 2 shows the image plots for the posterior means for the partial correlation matrices for both models and the posterior probabilities that the elements in the partial correlation matrix are non-zero for Model $CSVS$. The patterns in the first two image plots are similar but, as expected, the estimated partial correlation matrix for Model $CSVS$ is sparser than the estimated partial correlation matrix for Model $NCSVS$.

	$\beta_1 - \beta_2$	$\beta_1 - \beta_3$	$\beta_2 - \beta_3$
post. mean	9.4056	7.1390	-2.2665
post. std. error	1.0226	0.9727	1.0788

Table 7: Posterior means and standard errors of the difference in the intercepts of the three diets using Model $NCSVS$ for the hip replacement data.

	$\beta_1 - \beta_2$	$\beta_1 - \beta_3$	$\beta_2 - \beta_3$
post. mean	9.4278	7.1487	-2.2791
post. std. error	1.0861	1.1244	1.1102

Table 8: Posterior means and standard errors of the difference in the intercepts of the three diets using Model DVS for the hip replacement data.

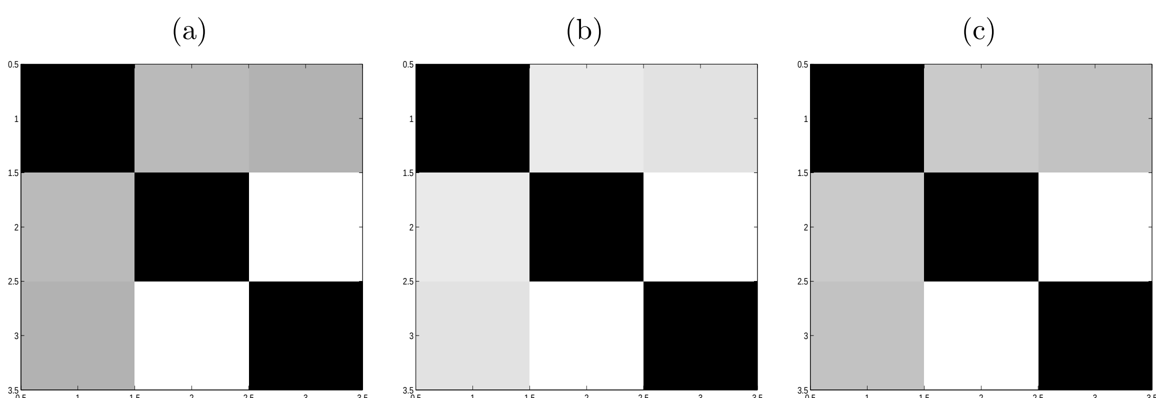

Figure 2: Image plots for the hip replacement data. Panel (a) is the image plot of the partial correlation matrix estimated by Model $NCSVS$. Panel (b) is the image plot of the partial correlation matrix estimated by Model $CSVS$. Panel (c) is the image plot of the probabilities of the elements of the correlation matrix being non-zero as estimated by Model $CSVS$.

4.3 Cross-Sectional Cow Diet Data

This data set consists of observations on 50 cows that are subjected to a diet additive. The data is cross-sectional and is described in Gelman et al. (2000, pages 213-215) [3]. The diet additive is methionine hydroxy analog and each cow is assigned to one of four different levels: 0% for the first 12 cows, 0.1% cows 13-25, 0.2% for cows 26-38 and 0.3% for the remaining 12 cows. The following variables were also recorded for each cow:

1. Lactation
2. Age (mos)
3. Initial weight (lb)
4. Mean daily dry matter consumed (kg)
5. Mean daily milk product (lb)
6. Milk fat (%)
7. Milk solids nonfat (%)
8. Final weight (lb)
9. Milk protein (%)

The first three variables were recorded before the additive was included in the diet and the last 6 variables were recorded after the additive was included in the diet. We treat the 6 post-diet variables as the multivariate response and the diet and the 3 pre-diet variables as the predictors. We model the data as in (**??**) which allows the same covariates to have different regression coefficients for each element in the vector of the responses. An interesting feature of this data is the high correlation amongst some of the predictor variables. In particular, the correlation between lactation and age is 0.9624, the correlation between lactation and initial weight is 0.7504 and the correlation between age and initial weight is 0.7808.

The response vector for the tth cow is $Y_t = (Y_{t1}, Y_{t2}, \ldots, Y_{t6})'$ where Y_t contains, in the following order, mean daily dry matter consumed, mean daily milk product, milk fat, milk solids

[3] The data is available from http://www.stat.columbia.edu/~gelman/book/data/.

nonfat, final weight and milk protein respectively. The predictor vector for the tth cow is $x_t = (x_{t0}, x_{t1}, x_{t2}, x_{t3}, x_{t4})'$, where x_t contains, in the following order, an intercept, diet, lactation, age and initial weight. The matrix of regression coefficients in (**??**) for this example is

$$B = \begin{bmatrix} \beta_{1,0} & \beta_{1,1} & \beta_{1,2} & \beta_{1,3} & \beta_{1,4} \\ \beta_{2,0} & \beta_{2,1} & \beta_{2,2} & \beta_{2,3} & \beta_{2,4} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \beta_{6,0} & \beta_{6,1} & \beta_{6,2} & \beta_{6,3} & \beta_{6,4} \end{bmatrix}$$

We specify our model such that if a variable is dropped from the regression it is dropped from all equations. To allow for this structure we write the regression equation in terms of (2.1) as

$$Y_t = X_t\beta + e_t, \qquad e_t \sim N(0, \Sigma)\,, \tag{4.5}$$

where

$$X_t = I_6 \otimes x_t' \quad \text{and} \quad \beta = vec(B'),$$

and I_6 is a 6×6 identity matrix.

To enable variables to be dropped from all regression models we group variables as outlined in Section **??** with the number of groups equal to 5. The first group refers to the intercept for each equation and is always included.

	$\beta_{1,0}$	$\beta_{1,1}$	$\beta_{1,2}$	$\beta_{1,3}$	$\beta_{1,4}$	$\beta_{2,0}$	$\beta_{2,1}$	$\beta_{2,2}$	$\beta_{2,3}$	$\beta_{2,4}$
post. mean	8.1754	0.4743	0.0000	0.0000	0.0062	24.2911	-0.1390	0.0001	0.0000	0.0271
post. std. error	2.3647	2.0404	0.0021	0.0002	0.0018	8.7645	7.2526	0.0084	0.0004	0.0069
post. prob.	NA	0.4250	0.0001	0.0001	1.0000	NA	0.4250	0.0001	0.0001	1.0000
	$\beta_{3,0}$	$\beta_{3,1}$	$\beta_{3,2}$	$\beta_{3,3}$	$\beta_{3,4}$	$\beta_{4,0}$	$\beta_{4,1}$	$\beta_{4,2}$	$\beta_{4,3}$	$\beta_{4,4}$
post. mean	2.6561	0.8378	0.0000	-0.0000	0.0006	8.5527	-0.0891	-0.0000	-0.0000	-0.0000
post. std. error	0.5166	1.0481	0.0012	0.0001	0.0004	0.3677	0.3236	0.0013	0.0001	0.0003
post. prob.	NA	0.4250	0.0001	0.0001	1.0000	NA	0.4250	0.0001	0.0001	1.0000
	$\beta_{5,0}$	$\beta_{5,1}$	$\beta_{5,2}$	$\beta_{5,2}$	$\beta_{5,4}$	$\beta_{6,0}$	$\beta_{6,1}$	$\beta_{6,2}$	$\beta_{6,3}$	$\beta_{6,4}$
post. mean	218.3280	-95.6234	0.0004	0.0001	0.8076	3.3429	-0.0735	0.0000	-0.0000	-0.0001
post. std. error	87.1405	130.8774	0.2096	0.0102	0.0668	0.2607	0.2326	0.0002	0.0000	0.0002
post. prob.	NA	0.4250	0.0001	0.0001	1.0000	NA	0.4250	0.0001	0.0001	1.0000

Table 9: Posterior means, standard errors and probabilities of being non-zeros for the regression coefficients using Model $NCSVS$ for the grouped cow diet data. NA means not available as the coefficient is always included.

	$\beta_{1,0}$	$\beta_{1,1}$	$\beta_{1,2}$	$\beta_{1,3}$	$\beta_{1,4}$	$\beta_{2,0}$	$\beta_{2,1}$	$\beta_{2,2}$	$\beta_{2,3}$	$\beta_{2,4}$
post. mean	8.1690	0.4862	0.0000	-0.0000	0.0063	24.3082	-0.1750	0.0001	-0.0000	0.0271
post. std. error	2.2344	1.9875	0.0015	0.0005	0.0017	8.3744	7.1448	0.0104	0.0023	0.0065
post. prob.	NA	0.4390	0.0001	0.0003	1.0000	NA	0.4390	0.0001	0.0003	1.0000
	$\beta_{3,0}$	$\beta_{3,1}$	$\beta_{3,2}$	$\beta_{3,3}$	$\beta_{3,4}$	$\beta_{4,0}$	$\beta_{4,1}$	$\beta_{4,2}$	$\beta_{4,3}$	$\beta_{4,4}$
post. mean	2.6559	0.8612	0.0000	0.0000	0.0006	8.5518	-0.0904	-0.0000	-0.0000	-0.0000
post. std. error	0.4951	1.0446	0.0009	0.0001	0.0004	0.3454	0.3102	0.0010	0.0001	0.0003
post. prob.	NA	0.4390	0.0001	0.0003	1.0000	NA	0.4390	0.0001	0.0003	1.0000
	$\beta_{5,0}$	$\beta_{5,1}$	$\beta_{5,2}$	$\beta_{5,2}$	$\beta_{5,4}$	$\beta_{6,0}$	$\beta_{6,1}$	$\beta_{6,2}$	$\beta_{6,3}$	$\beta_{6,4}$
post. mean	218.6719	-98.9527	-0.0008	0.0001	0.0636	0.2448	0.2242	0.0002	0.0001	0.0002
post. std. error	83.0269	130.3103	0.0893	0.0087	0.0636	0.2448	0.2242	0.0002	0.0001	0.0002
post. prob.	NA	0.4390	0.0001	0.0003	1.0000	NA	0.4390	0.0001	0.0003	1.0000

Table 10: Posterior means, standard errors and probabilities of being non-zerosfor the regression coefficients using Model $CSVS$ for the grouped cow diet data. NA means not available as the coefficient is always included.

	$\beta_{1,0}$	$\beta_{1,1}$	$\beta_{1,2}$	$\beta_{1,3}$	$\beta_{1,4}$	$\beta_{2,0}$	$\beta_{2,1}$	$\beta_{2,2}$	$\beta_{2,3}$	$\beta_{2,4}$
post. mean	10.2406	0.6445	0.0606	0.0018	0.0044	34.2230	-0.6413	0.3325	0.0094	0.0183
post. std. error	3.6447	1.7309	0.2107	0.0118	0.0031	14.6015	3.6768	1.0009	0.0534	0.0127
post. prob	NA	0.1815	0.1456	0.1155	0.7386	NA	0.0886	0.1679	0.1182	0.7436
	$\beta_{3,0}$	$\beta_{3,1}$	$\beta_{3,2}$	$\beta_{3,3}$	$\beta_{3,4}$	$\beta_{4,0}$	$\beta_{4,1}$	$\beta_{4,2}$	$\beta_{4,3}$	$\beta_{4,4}$
post. mean	2.9680	2.0439	0.0302	-0.0000	0.0001	8.5329	-0.0052	-0.0144	-0.0005	0.0000
post. std. error	0.3811	0.5958	0.0699	0.0039	0.0003	0.1885	0.1057	0.0351	0.0021	0.0002
post. prob	NA	0.9780	0.2732	0.1499	0.2199	NA	0.0691	0.2240	0.1574	0.1163
	$\beta_{5,0}$	$\beta_{5,1}$	$\beta_{5,2}$	$\beta_{5,2}$	$\beta_{5,4}$	$\beta_{6,0}$	$\beta_{6,1}$	$\beta_{6,2}$	$\beta_{6,3}$	$\beta_{6,4}$
post. mean	238.8083	-201.2394	0.4431	-0.0305	0.8041	3.3002	-0.0134	0.0011	0.0001	-0.0000
post. std. error	87.1000	117.7308	4.7685	0.3408	0.0704	0.1420	0.0877	0.0110	0.0009	0.0001
post. prob	NA	0.8181	0.0754	0.0759	1.0000	NA	0.0809	0.0867	0.0916	0.1403

Table 11: Posterior, means, standard errors and probabilities of being nonzero for the regression coefficients using Model $NCSVS$ for the cow diet data. NA means not available as the coefficient is always included.

	$\beta_{1,0}$	$\beta_{1,1}$	$\beta_{1,2}$	$\beta_{1,3}$	$\beta_{1,4}$	$\beta_{2,0}$	$\beta_{2,1}$	$\beta_{2,2}$	$\beta_{2,3}$	$\beta_{2,4}$
hline post. mean	9.5310	0.5236	0.0623	0.0010	0.0050	30.7891	-0.4925	0.2332	0.0084	0.0213
post. std.	3.3086	1.5375	0.2179	0.0119	0.0029	13.0697	3.1737	0.8675	0.0527	0.0116
post. prob	NA	0.1598	0.1453	0.1027	0.8208	NA	0.0816	0.1340	0.1116	0.8339
	$\beta_{3,0}$	$\beta_{3,1}$	$\beta_{3,2}$	$\beta_{3,3}$	$\beta_{3,4}$	$\beta_{4,0}$	$\beta_{4,1}$	$\beta_{4,2}$	$\beta_{4,3}$	$\beta_{4,4}$
post. mean	2.9592	2.0622	0.0316	0.0001	0.0001	8.5256	-0.0074	-0.0109	-0.0004	0.0000
post. std.	0.3669	0.5725	0.0688	0.0040	0.0003	0.1730	0.1014	0.0310	0.0019	0.0002
post. prob	NA	0.9833	0.2918	0.1614	0.2203	NA	0.0663	0.1848	0.1436	0.1046
	$\beta_{5,0}$	$\beta_{5,1}$	$\beta_{5,2}$	$\beta_{5,2}$	$\beta_{5,4}$	$\beta_{6,0}$	$\beta_{6,1}$	$\beta_{6,2}$	$\beta_{6,3}$	$\beta_{6,4}$
post. mean	232.9853	-191.9362	0.3351	-0.0346	0.8080	3.2934	-0.0119	0.0009	0.0001	-0.0000
post. std.	82.8676	120.2146	4.5379	0.3559	0.0671	0.1259	0.0848	0.0103	0.0008	0.0001
post. prob	NA	0.7952	0.0717	0.0746	1.0000	NA	0.0780	0.0809	0.0819	0.1212

Table 12: Posterior, means, standard errors and probabilities of being nonzero for the regression coefficients using Model $CSVS$ for the cow diet data. NA means not available as the coefficient is always included.

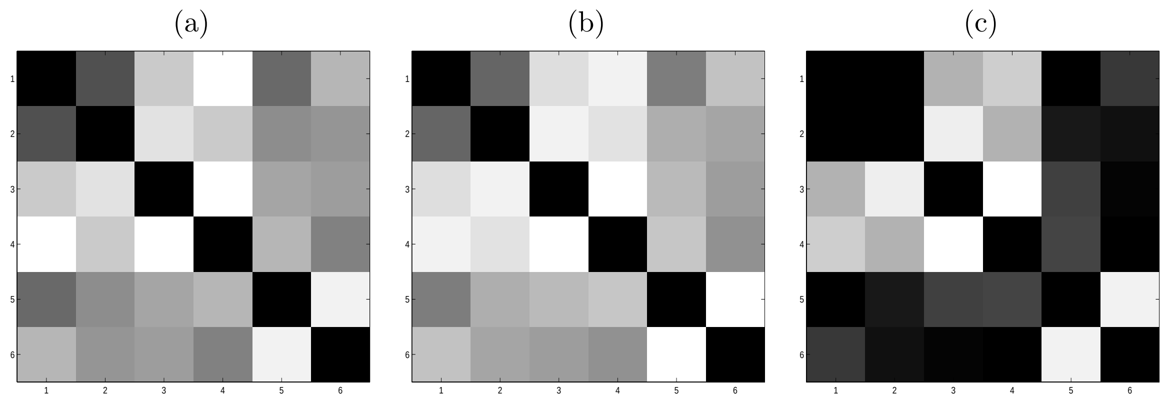

Figure 3: Image plots for the grouped cow milk protein data. Panel (a) is the image plot of the partial correlation matrix estimated by Model $NCSVS$. Panel (b) is the image plot of the partial correlation matrix estimated by Model $CSVS$. Panel (c) is the image plot of the probabilities of the elements of the correlation matrix being non-zero as estimated by Model $CSVS$.

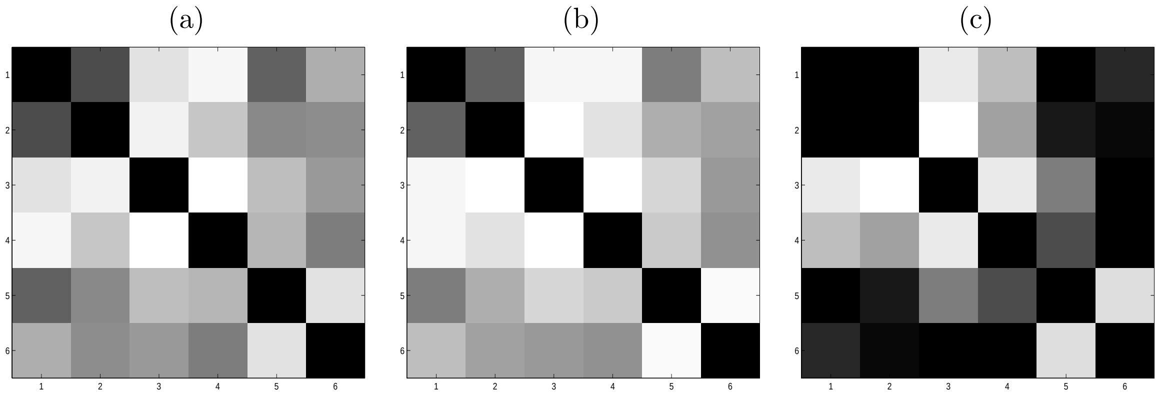

Figure 4: Image plots for the cow milk protein data. Panel (a) is the image plot of the partial correlation matrix estimated by Model $NCSVS$. Panel (b) is the image plot of the partial correlation matrix estimated by Model $CSVS$. Panel (c) is the image plot of the probabilities of the elements of the correlation matrix being non-zero as estimated by Model $CSVS$.

4.4 Pig Bodyweight Data

This longitudinal data set contains observations on 48 pigs measured over 9 successive weeks. It is described in Diggle et al. (2002, pages 34-35) who analyzed it to examine the growth rates of pigs[4]. Diggle et al. (2002, page 34) note that the trend in pig growth rates is approximately linear but each individual pig varies in both initial weight and in its growth rate. As a result Diggle et al. (2002, pages 76-77) structure the pig growth rate data as a random effects model. In this article we model the mean function of the pigs growth rate as a piecewise linear trend such that at each time point we allow the slope to change. Diggle et al. (2002, page 35) contains a plot of the pig bodyweights across time. The plot reveals that while time periods prior to the fourth week appear have a constant slope, after the fourth period the individual trajectories exhibit more variation and hence perhaps have different slopes for different time periods.

Let Y_{ti} be the response for pig t at time i and write the piecewise linear time trend as

$$\beta_0 + \beta_1 i + \beta_2 (i-2)_+ + \beta_3 (i-3)_+ + \ldots + \beta_8 (i-8)_+, \text{ for } i = 1, \ldots, 9 \tag{4.6}$$

where

$$x_+ = \begin{cases} x & \text{if } x \geq 0, \\ 0 & \text{if } x < 0. \end{cases}$$

[4]The data can be obtained from http://www.maths.lancs.ac.uk/∼diggle/lda/Datasets/.

Writing (4.6) in the notation of (2.1)

$$Y_t = X_t\beta + e_t, \qquad e_t \sim N(0, \Sigma)\,, \tag{4.7}$$

where the predictor matrix X_t is

$$X_t = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 3 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 4 & 2 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 5 & 3 & 2 & 1 & 0 & 0 & 0 & 0 \\ 1 & 6 & 4 & 3 & 2 & 1 & 0 & 0 & 0 \\ 1 & 7 & 5 & 4 & 3 & 2 & 1 & 0 & 0 \\ 1 & 8 & 6 & 5 & 4 & 3 & 2 & 1 & 0 \\ 1 & 9 & 7 & 6 & 5 & 4 & 3 & 2 & 1 \end{bmatrix}$$

and β is the corresponding 9×1 vector of regression coefficients.

The estimated posterior means and standard errors for the regression coefficients and the posterior probabilities that the regression coefficients are non-zero are recorded in Table 13 for Model $NCSVS$ and in Table 14 for Model $CSVS$. For both models the coefficient β_2 is significant but the remaining regression coefficients contained in Table 13 suggest that the change in slope is only significant at the fourth time interval, a conclusion that is consistent with the time series plot of the pigs growth rate presented by Diggle et al. (2002, page 34).

	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9
post. mean	17.7743	6.7040	0.0643	-1.2002	0.1451	0.2608	-0.0396	0.4606	-0.5418
post. std. error	0.4323	0.1807	0.1902	0.3035	0.2357	0.3617	0.1983	0.4190	0.4328
post. prob.	NA	1.0000	0.2950	0.9980	0.4455	0.5065	0.3040	0.6720	0.7295

Table 13: Posterior means, standard errors and probabilities of being non-zero for the regression coefficients using Model $NCSVS$ for the pig growth rate data. NA means not available as the coefficient is always included.

Figure 6 shows the image plots of the posterior means of the partial correlations and the posterior probabilities that the elements in the partial correlation matrix are non-zero. The plots suggest that the partial correlations have an autoregressive type structure.

	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9
post. mean	17.8010	6.6829	0.1251	-1.2903	0.1648	0.3164	-0.0604	0.6258	-0.7225
post. std	0.3903	0.1630	0.2238	0.2920	0.2357	0.3759	0.2233	0.4031	0.4614
post. prob.	1.0000	1.0000	0.3950	1.0000	0.5185	0.5905	0.3465	0.8180	0.8145

Table 14: Posterior means, standard errors and probabilities of being non-zero for the regression coefficients using Model $CSVS$ the pig growth rate data. NA means not available as the coefficient is always included.

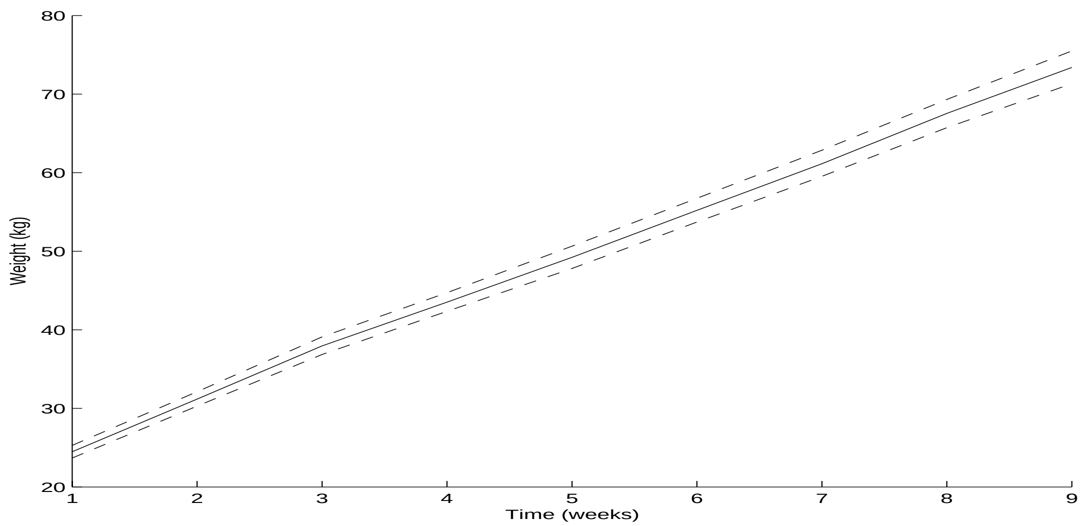

Figure 5: Posterior mean (solid line) with a 95 percent credible region (dotted lines) of the time trend for pig growth rate data estimated by Model $NCSVS$.

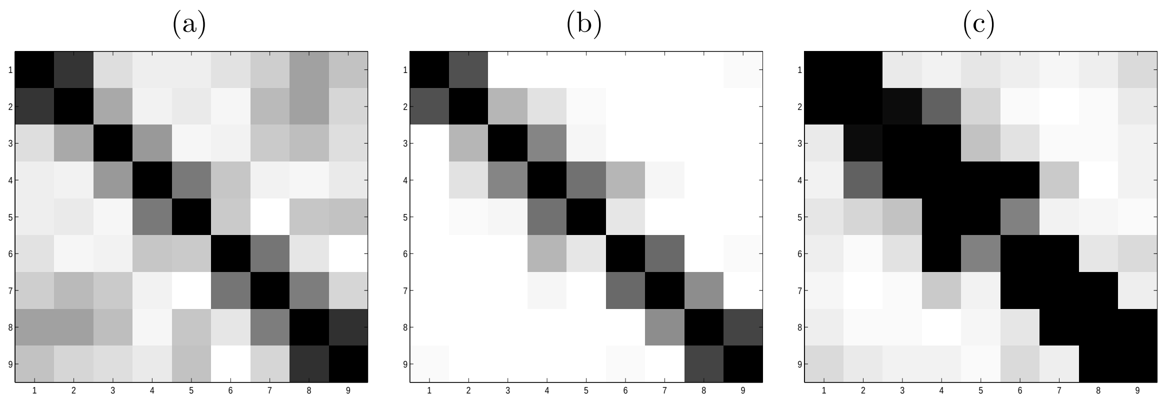

Figure 6: Image plots for pig growth rate data. Panel (a) is the image plot of the partial correlation matrix estimated by Model $NCSVS$. Panel (b) is the image plot of the partial correlation matrix estimated by Model $CSVS$. Panel (c) is the image plot of the probabilities of the elements of the correlation matrix being non-zero as estimated by Model $CSVS$.

5 Simulation study

This section studies the performance of variable selection and covariance selection either by singly or jointly. The performance is based on the effect on four different loss functions. The Kullback-Liebler loss function looks at the effect on the whole predictive density, the L_1 loss function looks at the effect on the covariance matrix, the beta loss function looks at the effect on the regression coefficients, and the fit loss function looks at the effect on the fitted values.

Our study is based four simulated data sets. Each of these data sets is based on the corresponding model that was estimated for one of the four real data sets considered in **??**. Unless stated otherwise, each of the real data sets was estimated using the $NCSVS$ model and the parameter estimates obtained were treated as the 'true model' parameters for the simulated data. Fifty replicates of data from this 'true model' were then constructed using the same values of the covariates as in the real data.

We now describe the four loss functions in detail.

1. *The KL loss function.* This is our empirical version of the Kullback-Liebler distance between the true predictive density and the estimated predictive density. We use this loss function to assess the effect of variable and covariance selection on the estimation the whole of the predictive distribution.

   Let $p(Y|Y_{\text{data}}, X)$ be the predictive density of Y given X and the observed data Y_{data} and let $p_T(Y|X)$ be the true density of Y given X. For a given value of X, the Kullback-Liebler distance between $p(Y|Y_{\text{data}}, X)$ and $p_T(Y|X)$ is [5]

   $$\int p_T(Y|X) \log \left(\frac{p(Y|Y_{\text{data}}, X)}{p_T(Y|X)} \right) dy \tag{5.1}$$

   and it can be shown KLD ≤ 0 with strict inequality unless $p(Y|Y_{\text{data}}, X) = p_T(Y|X)$ for all X.

   We cannot compute the integral in (5.1) analytically because Y is multivariate and because $p(Y|Y_{\text{data}}, X)$ is estimated by simulation. To approximate (5.1) for a given X, we generate K values of Y from $p_T(Y|X)$, which we call $Y_{k,x}, k = 1, \ldots, K$ and define the empirical Kullback-Liebler distance at X as

   $$\frac{1}{K} \sum_{k=1}^{K} \log \left(\frac{p(Y_{k,x}|Y_{\text{data}}, x)}{A_x p_T(Y_{k,x}|X)} \right) , \tag{5.2}$$

[5] Give a reference here for KL distance

where

$$A_x = \frac{1}{K}\sum_{k=1}^{K}\frac{p(Y_{k,x}|Y_{\text{data}}, X)}{p_T(Y_{k,x}|X)}.$$

It is straightforward to show using Jensen's inequality [6] that the sum in (5.2) is always less than or equal to 0 and it is strictly less than 0 unless $p(Y_{k,x}|Y_{\text{data}}, x) = p_T(Y_{k,x}|X)$ for all $k = 1, \ldots, K$. To get a representative set of values of X, we choose L values x_l of X at random from the observed covariates and define the empirical Kullback-Liebler distance as

$$KL\{p_{\text{pred}}, p_T\} = \frac{1}{KL}\sum_{l=1}^{L}\sum_{k=1}^{K}\log\left(\frac{p(Y_{k,l}|Y_{\text{data}}, x_l)}{A_l p_T(Y_{k,l}|x_l)}\right), \tag{5.3}$$

where $Y_{k,l}$ and A_l are $Y_{k,x}$ and A_x evaluated at $x = x_l$.

In the simulations we used $K = 400$, with values of K greater than 400 giving the same numerical results (to 3 decimal places) of the Kullback-Liebler distance. We also used a value of $L = 100$.

The predictive density $p(Y|X, Y)$ is estimated using the output of the Markov chain Monte Carlo simulation. Let $\beta^{[j]}, \Sigma^{[j]}, j = 1, \ldots, J$ be iterates of β and Σ generated from the posterior distribution. Then, we take

$$p(Y|Y_{\text{data}}, X) = \frac{1}{J}\sum_{j=1}^{J} p(Y|X, \beta^{[j]}, \Sigma^{[j]}) \ .$$

in the MCMC simulation scheme.

2. *The L_1 loss function.* Let $\widehat{\Sigma}$ be an estimate of the true covariance matrix Σ. The L_1 loss function for the covariance matrix is given by Yang and Berger (1994) as

$$L_1(\widehat{\Sigma}, \Sigma) = \text{tr}(\widehat{\Sigma}\Sigma^{-1}) - \log|\widehat{\Sigma}\Sigma^{-1}| - p.$$

This loss function assesses the effect of variable and covariance selection on the estimation of Σ.

3. *The beta loss function.* We write the beta loss function as

$$\text{beta}(\widehat{\beta}, \beta) = \left\{\sum_{i=1}^{q}(\widehat{\beta}_i - \beta_i)^2/q\right\}$$

where q is the length of β. This loss function assesses the effect of variable and covariance selection on the estimation of the regression coefficients.

[6] Put a reference here

4. *The fit loss function.* We write the fit loss function as

$$\mathrm{fit}(\widehat{f}(X), f(X)) = \left\{ \sum_{i=1}^{n_f} \sum_{j=1}^{p} (\widehat{f}(X)_{ij} - f(X)_{ij})^2 / (n_f \times p) \right\}^{\frac{1}{2}},$$

where $f(X)_{ij}$ is the mean response for observation i at the jth measurement, $\widehat{f}(X)_{ij}$ is the corresponding estimated mean response, n_f is the number of observations for which we calculate the fit loss function and p is the number of measurements for each observation. This loss function assesses the effect of variable and covariance selection on the mean of the regression. [7]

We estimated four different models for each data set. The first two models are Model $NCSVS$ and Model $CSVS$. These two models were explained in section xx. Teh third model we estimated carried out covariance selection but no variable selection and we call this model $NCSVS$. The fourth model did not carry out either variable selection or covariance selection and we call this model $NCSNVS$. We compare $NCSVS$to model Model $CSVS$ for a given loss function LOSS by the percentage increase (or decrease) in LOSSin going from Model $CSVS$ to say Model $NCSVS$, i.e.

$$D(NCSVS, CSVS) = \frac{\mathrm{LOSS}(NCSVS) - \mathrm{LOSS}(CSVS)}{\mathrm{LOSS}(CSVS)} \times 100. \tag{5.4}$$

If Model $CSVS$ outperforms Model $NCSVS$then $D(NCSVS, CSVS) > 0$, and conversely if model $NCSVS$ outperforms $CSVS$then $D(NCSVS, CSVS) < 0$. We carried out similar comparisons for Models $NCSNVS$ and $NCSVS$.

5.1 Longitudinal Cow Milk Protein Data

Figure 7 reports the percentage change in going from Model $CSVS$ to the other three models for the four loss functions. The figure shows that Model $CSVS$ outperforms Model $NCSNVS$ and Model $NCSVS$ under all four loss functions. The improvement is particularly pronounced for the L_1 loss function, probably due to the sparsity in Ω. There is no improvement between Model $CSVS$ over Model $NCSVS$ across all loss functions. The results imply covariance selection was beneficial compared to no covariance selection but there was no benefit in performing variable selection.

[7] I do not understand why we need the double subscript. Why don't we have a subscript l as in X_l ? What do we do in the code itself. Don't we just use $n_f = L$? In addition, should it be $f(X_{ij}$ rather than $f(X)_{ij}$.

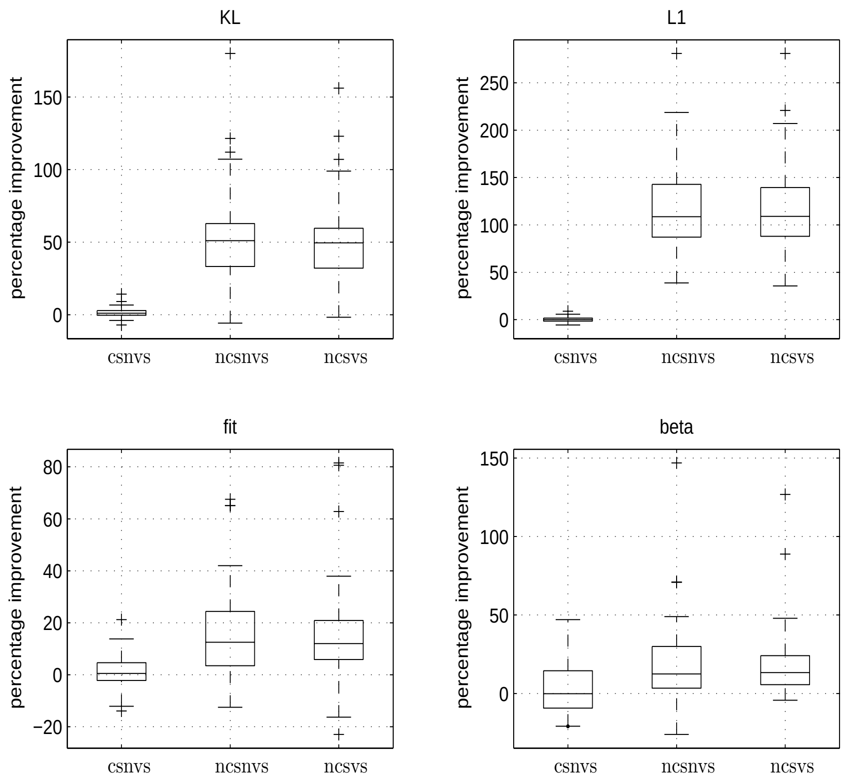

Figure 7: Longitudinal cow data. The boxplots represent the percentage change in going from Model $CSVS$ to the Model $NCSVS$, the Model $NCSNVS$ and Model $NCSVS$. From top left and reading clockwise the boxplots show the KL distance, the L_1 loss function, the fit loss function and the beta loss function.

5.2 Hip replacement data

Figure 8 is similar to Figure 7 and shows that Model $CSVS$ outperforms Model $NCSVS$ and Model $NCSNVS$ for the KL distance, the fit loss and the beta loss. The improvement is noticable but perhaps not significant for Model $CSVS$ over Model $NCSVS$ for the KL distance and the results are very similar for Model $CSVS$ and Model $NCSVS$ for the fit loss and beta loss. Model $CSVS$ offered significant improvement over Model $NCSNVS$ and Model $NCSVS$ for the L_1 loss. These results suggest that both variable selection and covariance selection made a difference in this simulated example.

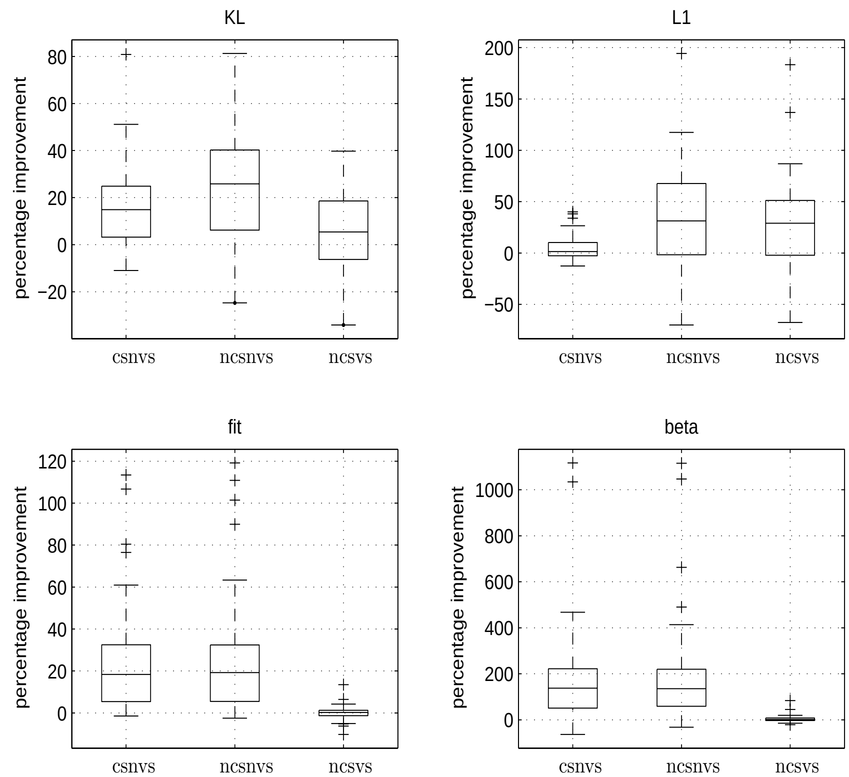

Figure 8: Longitudinal hip replacement data. The boxplots represent the percentage change in going from the $CSVS$ model to the $NCSVS$ model, the $NCSNVS$ model and the $NCSVS$ model. From top left and reading clockwise the boxplots show the KL distance, the L_1 loss function, the fit loss function and the beta loss function.

5.3 Cow cross-sectional data

Figure 9 is similar to Figures 7 and 8 and shows that the improvement in the L_1 loss function of Model DVS over the Model WVS and Model WNS is, while not excessive, significant suggesting covariance selection outperforms no covariance selection. The beta loss function shows substantial improvement of Model DVS over Model DNS and Model WNS. There is little difference between Model DVS and Model WVS. The fit loss function shows significant but small improvement of Model DVS over Model DNS and Model WNS and no difference between Model DVS and Model WVS. The KL loss function shows significant but small improvement of Model DVS over all other models. We conclude that both variable selection and covariance selection made a difference in this simulated example and that variable selection offered the most improvement.

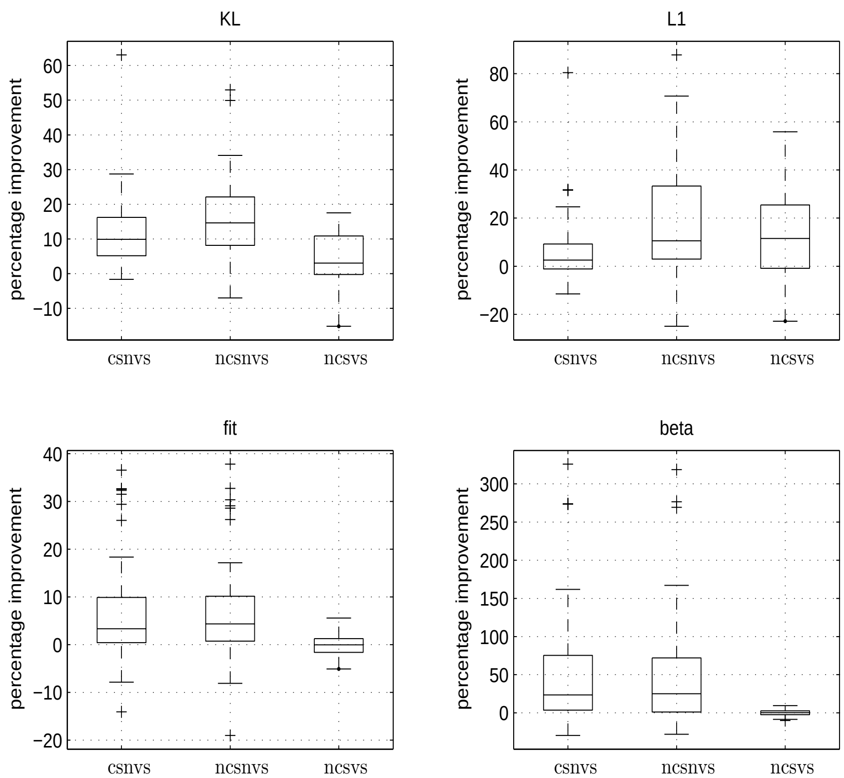

Figure 9: Cross-sectional cow data. The boxplots represent the percentage change in going from the DVS model to the DNS model, the WNS model and the WVS model. From top left and reading clockwise the boxplots show the KL distance, the L_1 loss function, the fit loss function and the beta loss function.

5.4 Pig longitudinal data

When simulating the pig growth rate data we first simulated 50 replications of the data using the original predictor matrix and the estimated parameters contained in Section 4 from Model WVS. These results are presented in Figure 10. We then simulated 50 replications of the data twice more, first with a tridiagonal inverse covariance matrix and then a diagonal inverse covariance matrix in order to show the added improvement of covariance selection as the inverse covariance matrix becomes sparser. Figures 10 to 12 report the results and are similar to the previous figures in this section. [8]

Figure 10 indicates that for the KL and beta loss functions Model DVS showed no significant improvement over Model DNS, Model WNS or Model WVS. In fact, the fit loss function indicates Model DNS and Model WNS offer improvement over Model DVS. The L_1 loss function implies Model DVS improves on Model WNS and Model WVS.

Figure 11 shows the resuls when the inverse covariance matrix is tridiagonal. The KL and L_1 loss functions show that Model DVS outperforms Model WNS and Model WVS but does not outperform the Model DNS. The beta loss function reveals Model DVS now outperforms both Model DNS and Model WNS.

Figure 12 reveals similar results for the beta and fit loss functions. However, when the inverse covriance is diagonal Model DVS shows greater improvement over Model WNS and Model WVS in the L_1 and KL loss function.

It is apparent from Figures 10 to 12 that as the inverse covariance matrix becomes sparser the improvement in the decomposition prior over the Wishart prior in the L_1 loss function increases suggesting the decomposition prior outperforms the Wishart prior when the inverse covariance matrix is sparse.

[8] We should generate data using the estimates based on the model DVS

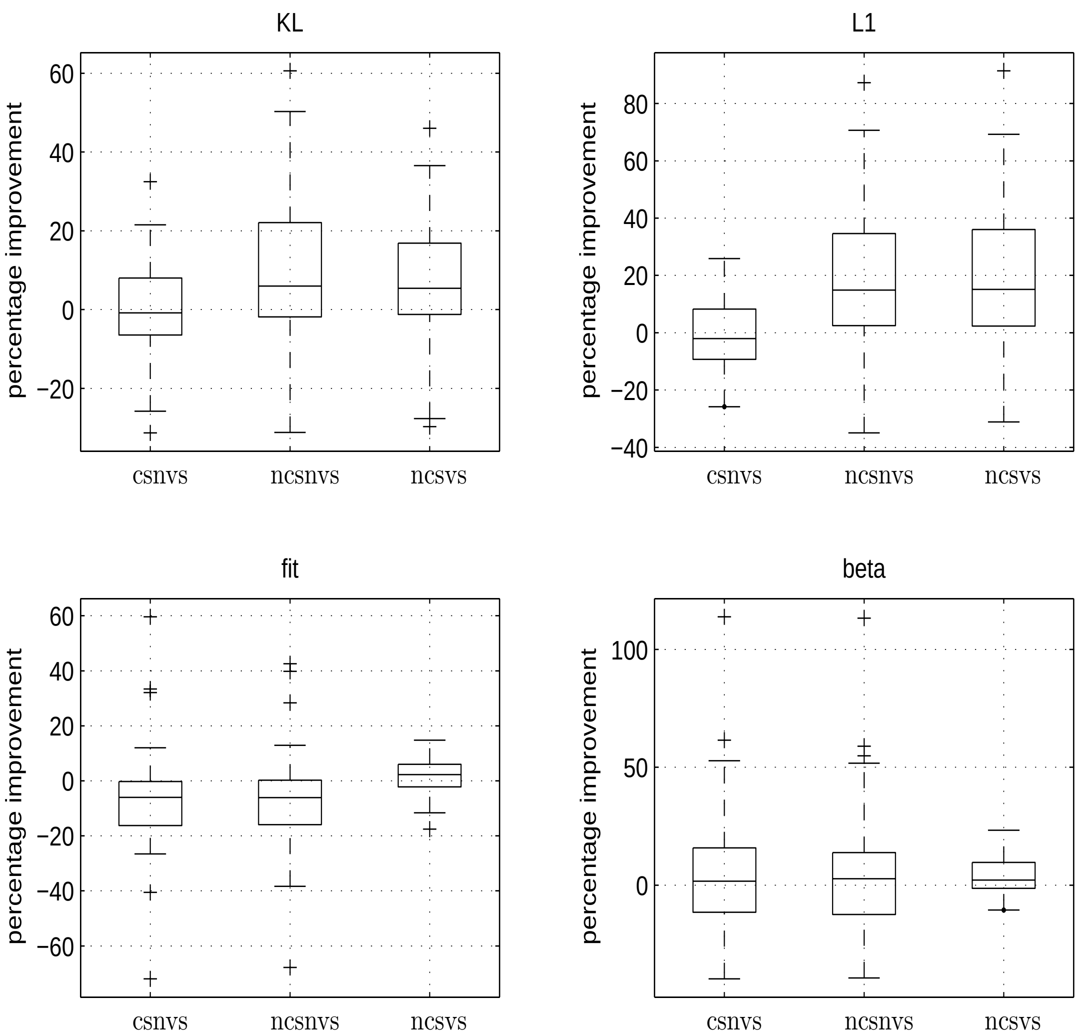

Figure 10: Longitudinal pig data with full inverse covariance matrix. The boxplots represent the percentage change in going from the DVS model to the DNS model, the WNS model and the WVS model. From top left and reading clockwise the boxplots show the KL distance, the L_1 loss function, the fit loss function and the beta loss function.

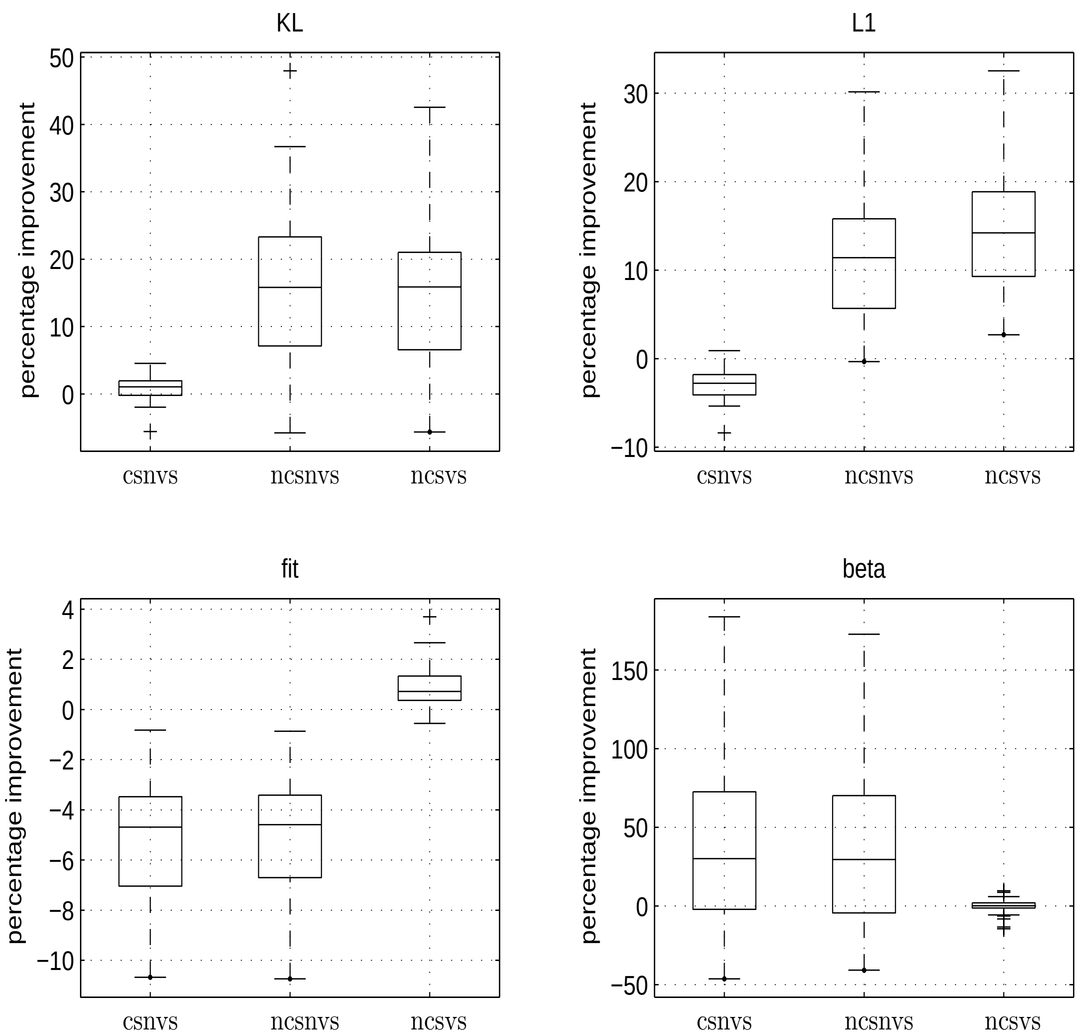

Figure 11: Longitudinal pig data with a tridiagonal inverse covariance matrix. The boxplots represent the percentage change in going from the DVS model to the DNS model, the WNS model and the WVS model. From top left and reading clockwise the boxplots show the KL distance, the L_1 loss function, the fit loss function and the beta loss function.

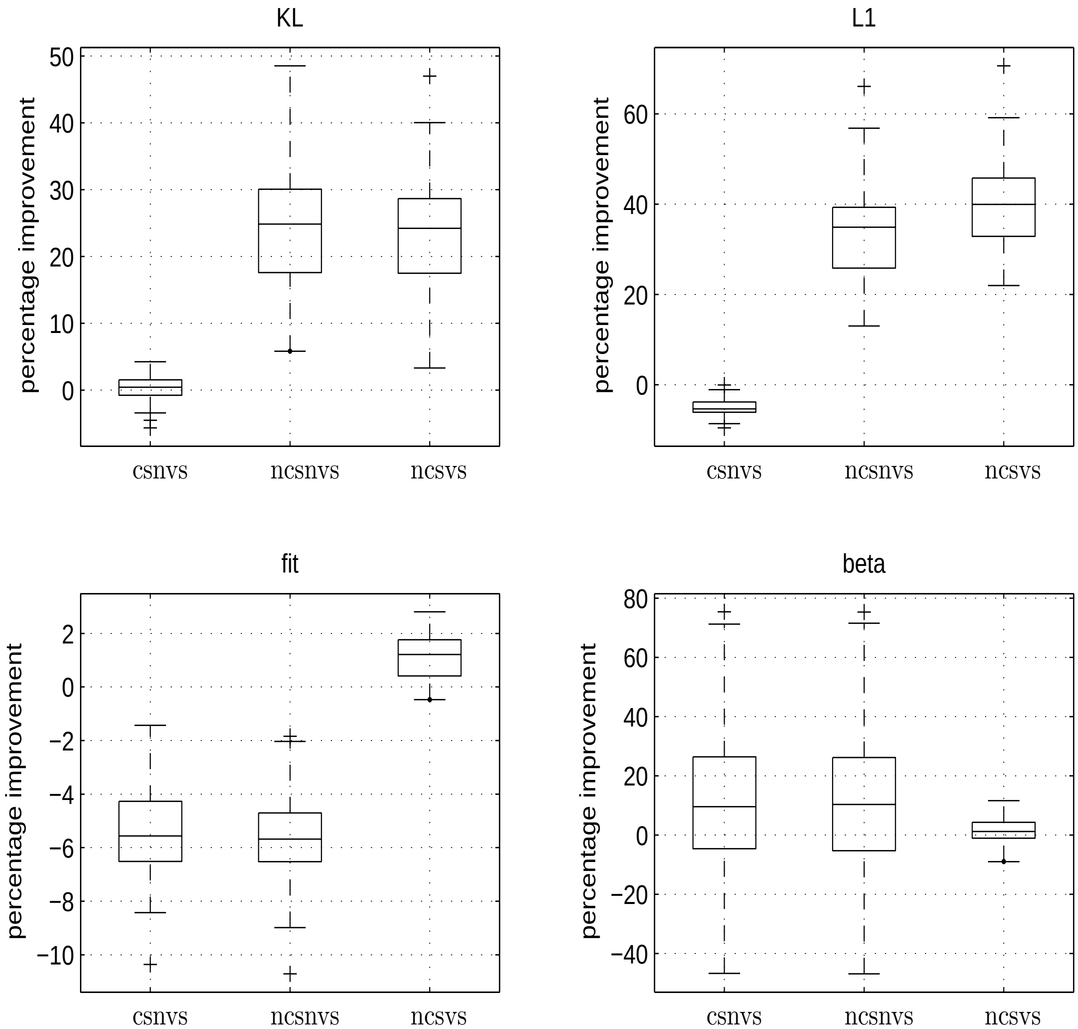

Figure 12: Longitudinal pig data with a diagonal inverse covariance matrix. The boxplots represent the percentage change in going from the DVS model to the DNS model, the WNS model and the WVS model. From top left and reading clockwise the boxplots show the KL distance, the L_1 loss function, the fit loss function and the beta loss function.

References

Barnard, J., McCulloch, R., and Meng, X. (2000), "Modeling Covariance Matrices in terms of Standard Deviations and Correlations, with Application to Shrinkage," *Statistica Sinica*, 10, 1281–1311.

Brieman, L. (1996), "Heuristics of Instability and Stabilisation in Model Melection," *Annals of Statistics*, 24, 2350–2383.

Brown, P., Fearn, T., and Vannucci, M. (1999), "The Choice of Variables in Multivariate Regression: a Non-conjugate Bayeisan Decision Theory Approach." *Biometrika*, 86, 635–648.

Brown, P., Vannucci, M., and Fearn, T. (1998), "Multivariate Bayesian Variable Selection and Prediction," *Journal of the Royal Statistical Society, Series B*, 60, 627–641.

Brown, P., Vanucci, M., and Fearn, T. (2002), "Bayes Model Averaging with Selection of Regressors," *Journal of the Royal Statistical Society, Series B*, 64, 519–536.

Chiu, T., Leonard, T., and Tsui, K. (1996), "The matrix-logarithm covariance model," *Journal of the American Statistical Association*, 81, 310–20.

Crowder, M. and Hand, D. (1990), *Analysis of Repeated Measures*, Chapman and Hall, London.

Daniels, M. and Kass, R. (1999), "Nonconjugate Bayesian estimation of covariance matrices," *Journal of the American Statistical Association*, 94, 1254–63.

Dempster, A. (1972), "Covariance Selection," *Biometrics*, 28, 157–175.

Dempster, A. P. (1969), *Elements of Continuous Multivariate Analysis*, Reading, MA: Addison-Wesley.

Diggle, P., Heagerty, P., Liang, K., and Zeger, S. (2002), *Analysis of Longitudinal Data*, Oxford University Press.

Efron, B. and Morris, C. (1976), "Multivariate Empirical Bayes and Estimation of Covariance Matrices," *The Annals of Statistics*, 4, 22–32.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2000), *Bayesian Data Analysis*, Chapman and Hall/CRC.

George, E. and McCulloch, R. (1993), "Variable Selection via Gibbs Sampling," *jasa*, 88, 881–889.

— (1997), "Approaches for Bayesian Variable Selection," *Statistica Sinica*, 7, 339–373.

Giudici, P. and Green, P. (1999), "Decomposable graphical Gaussian model determination," *Biometrika*, 86, 785–801.

Hoeting, J., Madigan, D., Raftery, A., and Volinsky, C. (1999), "Bayesian Model Averaging: A Tutorial (with discussion)," *Statistical Science*, 14, 382–417, corrected version at http://www.stat.washington.edu/www/research/online/hoeting1999.pdf.

Kohn, R., Smith, M., and Chan, D. (2001), "Nonparametric regression using linear combinations of basis functions," *Statistics and Computing*, 11, 313–322.

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

Leonard, T. and Hsu, J. S. J. (1992), "Bayesian inference for a covariance matrix," *The Annals of Statistics*, 20, 1669–96.

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

Mitchell, T. and Beauchamp, J. (1988), "Bayesian Variable Selection in Linear Regression," *Journal of the American Statistical Association*, 83, 1023–1036.

Pourahmadi, M. (1999), "Joint mean-covariance models with application to logitudinal data: Unconstrained parameterisation," *Biometrika*, 86, 677–90.

— (2000), "Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix," *Biometrika*, 87, 425–35.

Raftery, A., Madigan, D., and Hoeting, J. (1997), "Bayesian Model Averaging for Linear Regression Models," *Journal of the American Statistical Association*, 94, 179–191.

Smith, M. and Kohn, R. (1996), "Nonparametric regression using Bayesian variable selection," *Journal of Econometrics*, 75, 317–342.

— (2002), "Bayesian parsimonious covariance matrix estimation for longitudinal data," *J. Am. Statist. Assoc.*, 87, 1141–53.

Stein, C. (1956), "ome problems in multivariate analysis. Part I," Technical Report 6, Dept. Statistics, Stanford University.

Whittaker, J. (1990), *Graphical Models in Applied Mathematical Analysis*, Wiley, New York.

Wong, F., Carter, C., and Kohn, R. (2003), "Efficient Estimation of Covariance Selection Models," *Biometrika*, 90, 809–830.

Yang, R. and Berger, J. (1994), "Estimation of a Covariance Matrix using the Reference Prior," *Annals of Statistics*, 22, 1195–1211.