Bayesian Inference in the Multivariate Probit Model

Estimation of the Correlation Matrix

by

Aline Tabet

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

in

The Faculty of Graduate Studies

(Statistics)

The University Of British Columbia

August, 2007

Abstract

Correlated binary data arise in many applications. Any analysis of this type of data should take into account the correlation structure among the variables. The multivariate Probit model (MVP), introduced by Ashford and Snowden (1970), is a popular class of models particularly suitable for the analysis of correlated binary data. In this class of models, the response is multivariate, correlated and discrete. Generally speaking, the MVP model assumes that given a set of explanatory variables the multivariate response is an indicator of the event that some unobserved latent variable falls within a certain interval. The latent variable is assumed to arise from a multivariate normal distribution. Difficulties with the multivariate Probit are mainly due to computation as the likelihood of the observed discrete data is obtained by integrating over a multidimensional constrained space of latent variables. In this work, we adopt a Bayesian approach and develop an an efficient Markov chain Monte Carlo algorithm for estimation in MVP models under the full correlation and the structured correlation assumptions. Furthermore, in addition to simulation results, we present an application of our method to the Six Cities data set. Our algorithm has many advantages over previous approaches, namely it handles identifiability and uses a marginally uniform prior on the correlation matrix directly.

Table of Contents

Appendices

List of Tables

List of Figures

Acknowledgements

I would like to thank my supervisors Dr. Arnaud Doucet and Dr. Kevin Murphy. This work would not have been possible without their valued advice and suggestions. I also thank the staff and faculty members of the Statistics Department at UBC, in particular, Dr. Paul Gustafson, Dr. Harry Joe and Dr. Matias Salibian-Barrera, for their help, advice and mentorship.

I am forever grateful to my family, Salma, Ghassan, Najat, Sal and Rhea, for their continued support and encouragement. The numerous sacrifices they made over the last few years allowed me to pursue my aspirations, and reach important milestones in my professional career.

Finally I want to thank my friends and fellow graduate students, both in the Statistics Department and in Computer Science, for providing theoretical advice, computer support and numerous help, but most importantly for making the last two years a memorable journey.

Dedication

To my mom and dad, your love and support makes everything possible.

Part I

Thesis

Chapter 1

Introduction

1.1 Motivation

Correlated discrete data, whether be it binary, nominal or ordinal, arise in many applications. Examples range from the study of group randomized clinical trials to consumer behavior, panel data, sample surveys and longitudinal studies. Modeling dependencies between binary variables can be done using Markov random fields (e.g., Ising models). However, an attractive alternative is to use a latent variable model, where the observed binary variables are assumed independent given latent Gaussian variables, which are correlated. An example of such model is the multivariate Probit model (MVP), introduced by Ashford and Snowden (1970). In this class of models, the response is multivariate, correlated and discrete. Generally speaking, the MVP model assumes that given a set of explanatory variables the multivariate response is an indicator of the event that some unobserved latent variable falls within a certain interval. The latent variable is assumed to arise from a multivariate normal distribution. The likelihood of the observed discrete data is then obtained by integrating over the multidimensional constrain space of latent variables.

$$P(Y_{ij} = 1|X_i, \beta, \Sigma) = \int_{A_{iT}} \dots \int_{A_{i1}} \phi_T(Z_i|X_i, \beta, R) dZ_1 \dots dZ_T \qquad (1.1)$$

where $i = 1, \dots, n$ indexes the independent observation, $j = 1, \dots, T$ indexes the dimension of the response, Y_{ij} is a T-dimensional vector taking values in $\{0, 1\}$, A_{ij} is the interval $(0, \infty)$ if $Y_{ij} = 1$ and the interval $(-\infty, 0]$ otherwise, β is the regression coefficient, Σ is the covariance matrix, and $\phi_T(Z_i|X_i, \beta, R)$ is the probability density function of the standard normal

distribution defined in A.1.

The MVP model has been proposed as an alternative to the multivariate logistic model, which is defined as:

$$P(Y_{ij} = 1|X_i, \beta, \Sigma) = \frac{\exp(x_i'\beta_j)}{\sum_{k=1}^{T} \exp(x_i'\beta_k)} \tag{1.2}$$

The appeal of the probit model is that it relaxes the independence of the irrelevant alternatives (IIA) property assumed by the logit model. This IIA property assumption states that if choice A is preferred to choice B out of the choice set {A,B}, then introducing a third alternative C, thus expanding the choice set to {A,B,C} must not make B preferred to A. This means that adding or deleting alternative outcome categories does not affect the odds among the remaining outcomes. More specifically in the logistic regression model, the odds of choosing m versus n does not depend on which other outcomes are possible. That is, the odds are determined only by the coefficient vectors for m and n, namely β_m and β_n:

$$\frac{P(Y_{im} = 1|X_i, \beta, \Sigma)}{P(Y_{in} = 1|X_i, \beta, \Sigma)} = \frac{\exp(x_i'\beta_m)/\sum_{k=1}^{T} \exp(x_i'\beta_k)}{\exp(x_i'\beta_n)/\sum_{k=1}^{T} \exp(x_i'\beta_k)} = \exp(X(\beta_m - \beta_n)) \tag{1.3}$$

In many cases, this is considered to be an unrealistic assumption (see for example McFadden (1974)), particularly when the alternatives are similar or redundant as is the case in many econometric applications.

Until recently, estimation of MVP models, despite its appeal, has been difficult due to computational intractability especially when the response is high dimensional. However, recent advances in computational and simulation methods made this class of models more widely used.

Both classical and Bayesian methods have been extensively developed for estimation of these models. For a low dimensional response, finding the maximum likelihood estimator numerically using quadrature methods for solving the multidimensional integral is possible, but becomes quickly intractable as the number of dimensions T increases usually past 3.

Lerman and Manski (1981) suggest the method of simulated maximum likelihood (SML). This method is based on Monte Carlo simulations to approximate the high dimensional integral to estimate the probability of each choice. McFadden (1989) introduced the method of simulated moments (MSM). This method also requires simulating the probability of each outcome based on moment conditions. Natarajan et al. (2000) introduced a Monte Carlo variant of the Expectation Maximization algorithm (MCEM) to find the maximum likelihood estimator without solving the high dimensional integral. Other frequentist methods were also developed using Generalized Estimation Equations (GEE) (eg. Chaganty and Joe (2004)).

On the Bayesian side, Albert and Chib (1993) introduced a method that involves a Gibbs Sampling algorithm using data augmentation for the univariate probit model. McCulloch and Rossi (1994) extended this model to the multivariate case. The Bayesian method entails iteratively alternating between sampling the latent data and estimating the unknown parameters by drawing from their conditional distributions. The idea is that under mild conditions, successive sampling from the conditional distributions produces a Markov chain which converges in distribution to the desired joint conditional distribution. Other work on the Bayesian side includes that of Chib and Greenberg (1998), and more recently Liu (2001), Liu and Daniels (2006), and Zhang et al. (2006). These methods will be examined in more detail in Chapter 2.

Geweke et al. (1994) compared the performance of the classical frequentist methods SML and MSM with the Bayesian Gibbs sampling method and found the Bayesian method to be superior especially when the covariates are correlated and the error variances vary across responses.

1.2 Outline

In this work we adopt a Bayesian approach for estimation in the multivariate Probit class of models. The multinomial and the ordinal models are generalizations of the binary case. The multivariate binary response is a special case of the multinomial response with only two categories. The ordinal

model is also a special case of the multinomial model, where the categories are expected to follow a certain order. All the methods developed herein are developed for the multivariate binary model, but could be easily extended to include the multinomial and ordinal cases. The aim is to find a general framework to estimate the parameters required for inference in the MVP model, especially in high dimensional problems. We particularly focus on the estimation of an identifiable correlation matrix under a full correlation assumption and a constrained partial correlation assumption.

This thesis will be structured as follows:

In Chapter 2, we introduce the notation that will be used throughout the thesis. We discuss the problem of identifiability in the MVP class of models. We briefly compare several possible choices of prior distributions for Bayesian modeling, as well as review some methods that have been proposed in the literature to deal with identifiability and prior selection.

In Chapter 3, we detail a method for estimating an identifiable correlation matrix under the saturated model. The saturated model admits a full covariance matrix where all off-diagonal elements are assumed to be non-zero. We show simulation results on a low dimensional and a higher dimensional problem. Finally, we further investigate the method, by applying it to a widely studied data set: The Six Cities Data.

In Chapter 4, we extend the method developed in Chapter 3 to the case where a structure on the partial correlation matrix is imposed. To do so, we motivate the use of Gaussian graphical models and the Hyper-inverse Wishart Distribution. We provide a general introduction to Gaussian graphical models, and we adapt the algorithm and the priors developed in Chapter 3 to the new framework. Throughout this chapter, we assume that the structure of the inverse correlation matrix is known and given. Simulation results are presented as well as an application to the Six Cities Data set from Chapter 3.

We conclude in Chapter 5, by summarizing the work and the results. We also discuss possible extensions, applications and future work.

Chapter 2

The Multivariate Probit Model

2.1 Model Specification and Notation

The multivariate Probit model assumes that each subject has T distinct binary responses, and a matrix of covariates that can be any mixture of discrete and continuous variables. Specifically, let $Y_i = (Y_{i1}, \ldots, Y_{iT})$ denote the T-dimensional vector of observed binary $0/1$ responses on the ith subject, $i = 1, \ldots, n$. Let X_i be a $T \times p$ design matrix, and let $Z_i = (Z_{i1}, \ldots, Z_{iT})'$ denote a T-variate normal vector of latent variables such that

$$Z_i = X_i\beta + \epsilon_i, \quad i = 1, \ldots, n \tag{2.1}$$

The relationship between Z_{ij} and Y_{ij} in the multivariate probit model is given by

$$Y_{ij} = \begin{cases} 1 & \text{if } Z_{ij} > 0; \\ 0 & \text{otherwise.} \end{cases} \quad j = 1, \ldots, T \tag{2.2}$$

So that

$$\begin{aligned} P(Y_i = 1|\beta, \Sigma) &= \Phi(Z_i) \\ Z_i &\sim N(X_i\beta, \Sigma) \end{aligned} \tag{2.3}$$

where Φ is the Probit link which denotes the cumulative distribution function of the normal distribution as defined in A.1. Here $\beta = (\vec{\beta_1}', \ldots, \vec{\beta_T}')$ is a $p \times T$ matrix of unknown regression coefficients, ϵ_i is a $T \times 1$ vector of residual error distributed as $N_T(0, \Sigma)$, where Σ is the $T \times T$ correlation matrix of Z_i.

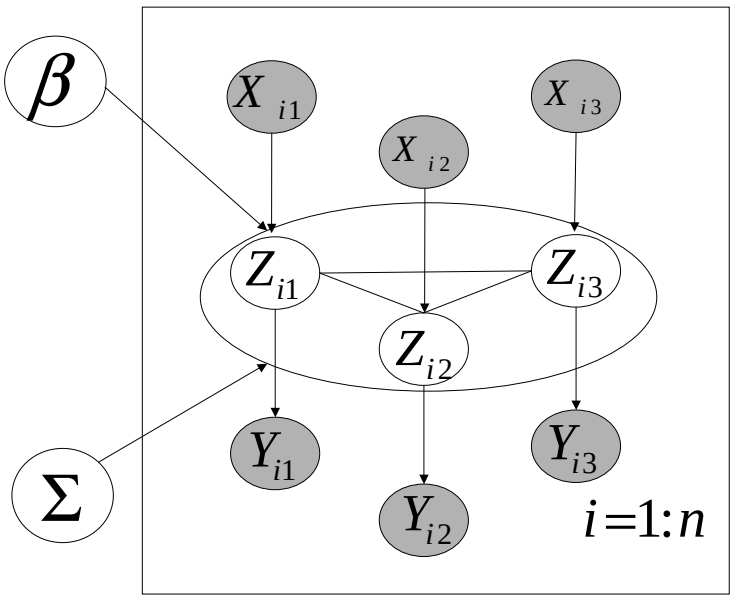


Figure 2.1: *A graphical representation of the model in 2.3 under a full correlation structure. Observed nodes are shaded.*

The posterior distribution of Z_i is given by

$$f(Z_i \,|\, Y_i, \beta, R) \propto \phi_T(Z_i|X_i, \beta, R) \prod_{j=1}^{T} \{I(z_{ij} > 0)I(y_{ij} = 1) + I(z_{ij} < 0)I(y_{ij} = 0)\} \tag{2.4}$$

This is a multivariate truncated Gaussian where $\phi_T(Z)$ is the probability density function of the normal distribution as in A.1.

The likelihood of the observed data Y is obtained by integrating over the latent variables Z:

$$P(Y_i = y_i|X_i, \beta, R) = \int_{A_{iT}} \dots \int_{A_{i1}} \Phi_T(Z_i|X_i, \beta, R) dZ_i \tag{2.5}$$

where A_{ij} is the interval $(0, \infty)$ if $Y_{ij} = 1$ and the interval $(-\infty, 0]$ otherwise.

This formulation of the model is most general, since it allows the regression parameters as well as the covariates to vary across categories T. In this work, we let the covariates vary across categories, however, we constrain the regression coefficients β to be fixed across categories by requiring $\vec{\beta}_1 = \ldots = \vec{\beta}_T = \vec{\beta}$.

2.2 Difficulty with Multivariate Probit Regression: Identifiability

In the multivariate Probit model, the unknown parameters (β, Σ) are not identifiable from the observed-data model (e.g: Chib and Greenberg (1998), Keane (1992)). This could be easily seen if we scale Z by a constant $c > 0$, we get

$$\begin{aligned} cZ &= c(X\beta + \epsilon) && (2.6) \\ &= X(c\beta) + c\epsilon && (2.7) \end{aligned}$$

from equation 2.2, clearly Y will have the same value given Z and given cZ, which means that the likelihood of $Y|X, \beta, \Sigma$ is the same as that of $Y|X, c\beta, c^2\Sigma$. Furthermore, we have no way of estimating the value of c.

In order to handle this identifiability issue in MVP, restrictions need to be imposed on the covariance matrix. In the univariate case, this restriction is handled by setting the variance to one. However, imposing such a restriction in the multivariate case is a little more complicated.

It is not uncommon to ignore the identifiability problem and perform the analysis on the unidentified model and post-process samples by scaling with the sampling variance using the separation strategy $R = D^{-1}\Sigma D^{-1}$, where D is a diagonal matrix with diagonal elements $d_{ii} = \sqrt{\Sigma_{ii}}$. This method is adopted by McCulloch and Rossi (1994), and is widely used (e.g Edwards and Allenby (2003)).

Many researchers are uncomfortable working with unidentified parameters. For instance, ignoring identifiability adds difficulty in the choice of prior

distributions, since priors are placed on unidentified parameters. Therefore, if the prior is improper, it is difficult to verify that the scaled draws are from a proper posterior distribution. Koop (2003, p. 227) gives an empirical illustration of the effect of ignoring identifiability. From simulation results, he shows that unidentifiable parameters have higher standard errors, and furthermore with non-informative priors there is nothing stopping estimates from going to infinity.

McCulloch et al. (2000) address identifiability by setting the first diagonal element of the covariance matrix $\sigma_{11} = 1$. However, this means that the standard priors for covariance could no longer be used, they propose a prior directly on the identified parameters, but their method is computationally expensive, and is slow to converge as pointed out by Nobile (2000). Nobile suggests an alternative way of normalizing the covariance by drawing from an inverse Wishart conditional on $\sigma_{11} = 1$ (Linardakis and Dellaportas, 2003). The approach of constraining one element of the covariance adds difficulty in the interpretability of the parameters and priors, and is computationally demanding and slow to converge.

Other approaches impose constraints on Σ^{-1}, the precision matrix. Webb and Forster (2006) parametrize Σ^{-1} in terms of its Cholesky decomposition: $\Sigma^{-1} = \Psi^T \Lambda \Psi^T$. In this parametrization, Ψ is an upper triangular matrix with diagonal elements equal to 1, and Λ is a diagonal matrix. The elements of Ψ could be regarded as the regression coefficients obtained by regressing the latent variable on its predecessors. Each λ_{jj} is interpreted as the conditional precision of the latent data corresponding to variable j given the latent data for all the variables preceding j in the decomposition. Identifiability is addressed in this case by setting λ_{jj} to 1. This approach only works if the data follows a specific ordering, for example time series. Dobra et al. (2004) propose an algorithm to search over possible orderings, however this becomes very computationally expensive in high dimensions.

Alternatively, identifiability could be handled by restricting the covariance matrix Σ to be a correlation matrix R (Chib and Greenberg (1998)). The correlation matrix admits additional constraints, since in addition to being positive semi-definite, it is required to have diagonal elements equal to

1 and off-diagonal elements $\in [-1, 1]$. Furthermore, just as in the covariance case, the number of parameters to be estimated increases quadratically with the dimension of the matrix.

Barnard et al. (2000) use the decomposition $\Sigma = DRD$, and place a separate prior on R and D directly. They use a Griddy Gibbs sampler (Ritter and Tanner, 1992) to sample the correlation matrix. Their approach involves drawing the correlation elements one at time and requires setting grid sizes and boundaries. This approach is inefficient, especially in high dimensions. Chib and Greenberg (1998) use a Metropolis Hastings Random Walk algorithm to sample the correlation matrix. This is more efficient than the Griddy Gibbs approach because it draws the correlation coefficient in blocks. However the resulting correlation matrix is not guaranteed to be positive definite, which requires the algorithm to have an extra rejection step. Furthermore, as with random walk algorithms in general, the mixing is slow in high dimensions.

Alternatively, some approaches use parameter expansion as described in Liu and Wu (1999) together with data augmentation, for example Liu (2001), Zhang et al. (2006), Liu and Daniels (2006), and others. The idea is to propose an alternative parametrization, to move from a constrained correlation space to sampling a less constrained covariance matrix and transform it back to a correlation matrix. These approaches differ mainly with the choice of priors and how the covariance matrix is sampled. The different possibilities for priors will be discussed in more detail in the next section, and an in-depth explanation of parameter expansion with data augmentation algorithm is in the next Chapter. Table 2.1 gives a summary of the how identifiability has been handled in the Probit model.

2.3 Bayesian Inference in Multivariate Probit Models

A Bayesian framework treats parameters as random variables and therefore requires the computation of the posterior distribution of the unknown

Table 2.1: *Summary of how identifiability has been handled in some previous work*

Identifiability	Paper
Ignored	McCulloch and Rossi (1994)
Restrict $\sigma_{11} = 1$	McCulloch et al. (2000) Nobile (2000)
Restrict $\lambda_{jj} = 1$ in $\Sigma^{-1} = \Psi^T \Lambda \Psi^T$	Webb and Forster (2006)
Restrict Σ to R	Barnard et al. (2000) Liu (2001) Liu and Daniels (2006) Zhang et al. (2006)

random parameters conditional on the data. A straightforward application of Bayes rule results in the posterior distribution of (β, R) where R is the correlation matrix, β is the matrix of regression coefficients, and D is the data.

$$\pi(\beta, R|D) \quad \propto \quad f(D|\beta, R)\pi(\beta, R) \tag{2.8}$$

In order to estimate the posterior distribution, a prior distribution on the unknown parameters β and R needs to be specified. In the absence of prior knowledge, it is often desirable to have uninformative flat priors on the parameters we are estimating

2.3.1 Prior Specification on β

It is common to assume that a priori β and R are independent. Liu (2001) propose a prior on β that depends on R to facilitate computations. There are several other choices of priors in the literature on the regression coefficients β. The most common choice is a multivariate Gaussian distribution centered at B, with known diagonal covariance matrix Ψ_β. It is typical to choose large values for the diagonal elements of Ψ_β so that the prior on β is uninformative. This is the proper conjugate prior. In addition, without loss of generality,

we could set B to 0

$$\pi(\vec{\beta}) \quad \sim \quad N_{pT}(0, \Psi_\beta \otimes I_T) \tag{2.9}$$

where $\vec{\beta}$ is the nT-dimensional vector obtained by stacking up the columns of the $p \times T$ regression coefficient matrix β. In this work, we constrain the regression parameter to be constant across T.

2.3.2 Prior Specification on the correlation matrix R

To handle identifiability, we restrict the covariance matrix Σ to be a correlation matrix, which means that the standard conjugate inverse Wishart prior for covariances cannot be used. Instead, a prior needs to be placed on R directly. However, as mentioned previously there does not exist a conjugate prior for correlation matrices.

Barnard et al. (2000) discuss possible choices of diffuse priors on R. The first is the proper jointly uniform prior:

$$\pi(R) \propto 1, \quad R \in \Re^T \tag{2.10}$$

Where the correlation matrix space $\Re^T$ is a compact subspace of the hypercube $[-1, 1]^{T(T-1)/2}$. The posterior distribution resulting from this prior is not easy to sample from. Barnard et al. use the Griddy Gibbs approach (Ritter and Tanner, 1992), which is inefficient. The approach in Chib and Greenberg (1998) uses this prior as well. Liu and Daniels (2006) use this prior for inference. However, they use a different prior to generate their sampling proposal.

It is important to note that using a jointly uniform prior would not result in uniform marginals on each r_{ij}. Barnard et al. (2000) show that a jointly uniform prior will tend to favor marginal correlations close to 0, making it highly informative, marginally. This problem becomes more apparent as T increases (see Figure 2.2).

Another commonly used uninformative prior is the Jeffrey's prior

$$\pi(R) \propto |R|^{-\frac{(p+1)}{2}} \tag{2.11}$$

This prior is used by Liu (2001). Liu and Daniels (2006) use it for generating their proposal.

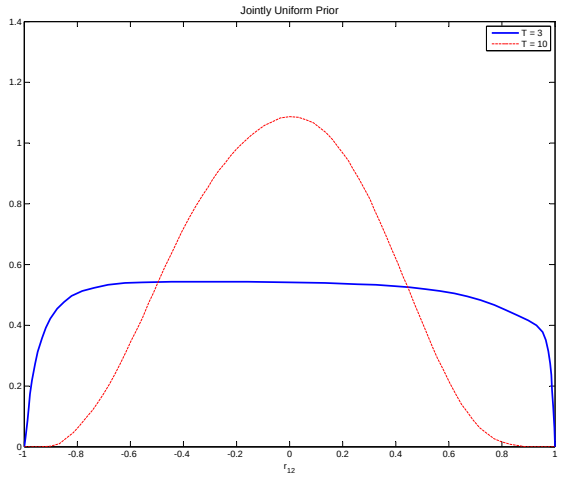


Figure 2.2: *Marginal prior density for r_{12} when $T = 3$ and $T = 10$ under the jointly uniform prior $p(R) \propto 1$, based on 2000 draws. (Figure 1 reproduced from Barnard et al. (2000))*

It has been shown that in the context of parameter expansion, this prior helps facilitate computations. However, it suffers from the disadvantage of being improper. Improper priors are not guaranteed to have a proper posterior distribution and, in addition, cannot be used for model selection due to Lindley's paradox. Furthermore, it has been shown that the use of improper priors on covariance matrices is in fact informative and tends to favor marginal correlations close to ± 1 (Rossi et al., 2005, Chapter 2).

Alternatively, Barnard et al. (2000) propose a prior on R such that marginally each r_{ij} is uniform on the interval $[-1, 1]$. This is achieved by taking the joint distribution of R to be:

$$\pi(R) \propto |R|^{\frac{T(T-1)}{2}-1}(\prod_i |R_{ii}|)^{-(T+1)/2} \tag{2.12}$$

The above distribution is difficult to sample from directly. However, they show that sampling from it can be achieved by sampling from a standard inverse Wishart with degrees of freedom equal to $\nu = T+1$ and transforming

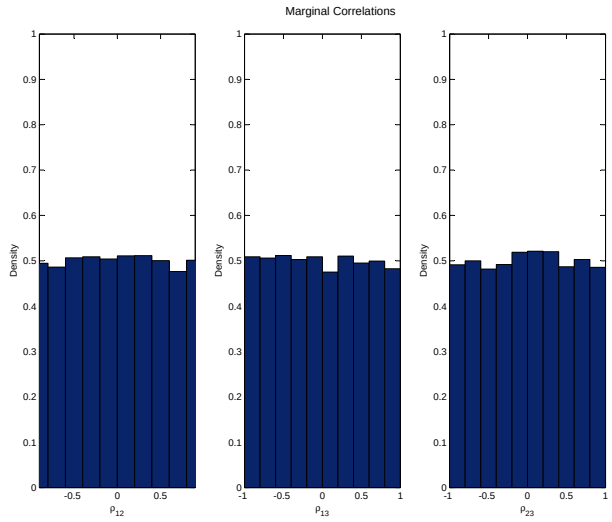


Figure 2.3: *Marginal correlations obtained using the prior in 2.12 by sampling from a standard inverse Wishart with degrees of freedom $\nu = T + 1$*

back to a correlation matrix using the separation strategy ($\Sigma = DRD$). The proof is reproduced in Appendix B and the result is illustrated in Figure 2.3.

The marginally uniform prior seems convenient, since it is proper and we are able to compute its normalizing constant. It does not push correlations toward 0 or ± 1 even in high dimensions. Most importantly, because it is proper, it opens the possibility for Bayesian model selection.

However, multiplying together the distribution of Z in equation 2.4 and the marginally uniform prior in 2.12, results in a posterior distribution that is complicated and not easily sampled from.

Nevertheless, we show in the next chapter that the marginal prior, when used in the context of parameter expansion, is actually computationally convenient for sampling from the posterior distribution.

Chapter 3

Correlation Estimation in the Saturated Model

3.1 Introduction

As we have seen from the previous chapter, inference in the MVP model is complicated due to the identifiability issue which requires constraining the covariance to be a correlation matrix. There is no conjugate prior for correlation matrices and therefore the posterior is not easily sampled from. In this Chapter, we build on previous work and adopt a Bayesian approach that uses a combination of Gibbs sampling and data augmentation. Furthermore, we use a re-parametrization leading to an expansion of the parameter space. This helps significantly with the computation of the posterior distribution. We focus on R being a full $T \times T$ correlation matrix.

3.2 Parameter Expansion and Data Augmentation

3.2.1 Data Augmentation

Data Augmentation (DA) is an algorithm introduced by Tanner and Wong (1987), very popular in statistics, used mainly to facilitate computation. These methods center on the construction of iterative algorithms by introducing artificial variables, referred to as "missing data" or latent variables. These variables may or may not have a physical interpretation but are mainly there for computational convenience.

Let Y be the observed data, and θ be the unknown parameter of interest.

If we are interested in making draws from $f(Y|\theta)$, the idea is to find a latent variable Z such that the joint distribution $f(Y, Z|\theta)$ is easily sampled from. The distribution of the observed data model is recovered by marginalizing the latent variable:

$$f(Y|\theta) = \int f(Y, Z|\theta) dZ \tag{3.1}$$

Algorithm 3.1 Data Augmentation

At iteration i

1. Draw $Z \sim f(Z|\theta, Y) \propto f(Y, Z|\theta)$
2. Draw $\theta \sim f(\theta|Z, Y) \propto f(Y, Z|\theta) f(\theta)$

The data augmentation algorithm 3.1 iterates between an imputation step where the latent variables are sampled and a posterior estimation step until convergence. The samples of the unknown parameter θ could then be used for inference.

3.2.2 Parameter Expansion for Data Augmentation

Parameter Expansion for Data Augmentation (PX-DA) , introduced by Liu and Wu (1999), is a technique usually useful for accelerating convergence. The idea is that if we can find an hidden parameter α in the complete data model $f(Y, Z|\theta)$, we can then expand this model to a larger model $p(Y, W|\theta, \alpha)$, that would preserve the distribution of the observed data model:

$$\int p(Y, W|\theta, \alpha) dW = f(Y|\theta) \tag{3.2}$$

We adopt the notation used in Liu and Wu (1999), and use W instead of Z and p instead of f to denote the latent data and the distributions under the expanded model. To implement the DA algorithm in this setting, a joint prior on the expansion parameter α and the original parameter of interest θ needs to be specified such that the prior on θ is the same under the original model and the expanded model ($\int p(\theta, \alpha) d\alpha = f(\theta)$). This can be done by maintaining the prior for θ at $f(\theta)$ and specifying a prior $p(\alpha|\theta)$.

By iterating through the steps of algorithm 3.2, we are able to achieve a faster rate of convergence than the DA algorithm in 3.1.

Algorithm 3.2 PX-DA Algorithm

At iteration i

1. Draw (α, W) jointly by drawing

$$\begin{aligned} \alpha &\sim p(\alpha|\theta) \\ W &\sim p(W|\theta, \alpha, Y) \propto p(Y, W|\theta, \alpha) \end{aligned}$$

2. Draw (α, θ) jointly by drawing

$$\alpha, \theta|Y, W \sim p(Y, W|\theta, \alpha)p(\alpha|\theta)f(\theta)$$

3.2.3 Data Transformation

Under certain conditions, an alternative view of the PX-DA treats W as the result of a transformation on the latent data Z induced by the expansion parameter α (Liu and Wu, 1999, Scheme 1). For this interpretation to hold, a transformation $Z = t_\alpha(W)$, needs to be defined such that for any fixed value of α, $t_\alpha(W)$ is a one-to-one differentiable mapping between Z and W:

$$p(Y, W|\theta, \alpha) = f(Y, t_\alpha(W)|\theta)|J_\alpha(W)| \tag{3.3}$$

where $|J_\alpha(W)|$ is the determinant of the Jacobian of the transformation T_α evaluated at W. The algorithm is detailed in 3.3. Note that in the second step of algorithm 3.3, α is sampled from its prior distribution.

This interpretation of the PX-DA algorithm is particularly useful in the case of MVP regression.

3.3 Proposed Model

In the model we are proposing, we want to use PX-DA mainly to simplify computation. We adopt the scheme described in algorithm 3.3 (correspond-

Algorithm 3.3 PX-DA Algorithm/ Data Transformation (scheme 1)

At iteration i

1. Draw $Z \sim f(Z|Y,\theta)$, compute $W = t_\alpha^{-1}(Z)$
2. Draw (α, θ) jointly conditional on the latent data

$$\alpha, \theta | Y, W \sim p(Y, t_\alpha(W)|\theta)|J_\alpha(W)|p(\alpha|\theta)f(\theta)$$

ing to scheme 1 in Liu and Wu (1999)).

3.3.1 Imputation Step

Let $\theta = (R, \beta)$, be the identifiable parameter of interest. The first step of algorithm 3.3, involves drawing Z conditional on the identifiable parameter θ. This is achieved by sampling from a multivariate truncated Gaussian as in equation (2.4).

For the generation of multivariate truncated Gaussian variables, we followed the approach outlined in Appendix D. This approach uses Gibbs steps to cycle through a series of univariate truncated Gaussians. In each step Z_{ij} is simulated from $Z_{ij}|Z_{i,-j}, \beta, R$, which is a univariate Gaussian distribution truncated to $[0, \infty)$ if $Y_{ij} = 1$ and to $(-\infty, 0]$ if $Y_{ij} = 0$. The parameters of the untruncated distribution $Z_{ij}|Z_{i,-j}, \beta, R$ are obtained from the usual formulae for moments of conditional Gaussians.

3.3.2 Posterior Sampling Step

Given the latent data sampled in step 1, we would like to draw (α, θ) from its posterior distribution. In order to implement step 2 of algorithm 3.3, we need to find an expansion parameter α, not identifiable from the observed data model, but identifiable from complete data-model. Subsequently, we need to define a transformation on the latent data.

Defining the Expansion Parameter and the Transformation

If we let

$$Z = t_\alpha(W) = D^{-1}W \tag{3.4}$$

or alternatively $W = DZ$, where D is a diagonal matrix with positive diagonal elements $d_{ii} = \sqrt{\Sigma_{ii}}$. The scale parameter D is not identifiable. For reasons which will become clear later, we could conveniently pick $\alpha = (\alpha_1, \ldots, \alpha_T)$ to be a function of D by taking

$$\alpha_i = \frac{r^{ii}}{2d_i^2} \tag{3.5}$$

where r^{ii} is the ith diagonal element of R^{-1} and d_i is the ith diagonal element of D.

In this case, for any fixed value of α, D is a one-to-one function of α and $t_\alpha(W)$ is a one-to-one differentiable mapping between Z and W.

This choice of α is not arbitrary. It is conveniently picked so that when combined with the prior of (R, β), the transformed likelihood, and the Jacobian, it results in a posterior distribution that is easily sampled from.

The Transformed Complete Likelihood: $p(Y, t_\alpha(W)|\theta)|J_\alpha(W)|$

For a given α, the determinant of the Jacobian [1] resulting by going from $(Z \to W)$ under the transformation in 3.11 is given by:

$$\begin{aligned} |J : Z \to W| &= \left|\frac{\partial(Z_1, \ldots Z_n)}{\partial(W_1 \ldots W_n)}\right| && (3.6)\\ &= \left|(I_n \otimes D^{-1})\right| && (3.7)\\ &= |D|^{-n} && (3.8) \end{aligned}$$

Combining the complete likelihood in equation 2.4 with the Jacobian, and after doing some algebra, we get:

[1] see a 3×3 example in Appendix C

$$
\begin{aligned}
p(Y, t_\alpha(W)|\beta, R) \quad & |J : Z \to W| \\
&= p(Y, Z|\beta, R) \times |J : Z \to W| \qquad (3.9) \\
&= |R|^{-\frac{n}{2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(Z_i - X_i\beta)'R^{-1}(Z_i - X_i\beta)\right) \times |J : Z \to W| \\
&= |D|^{-n}|R|^{-\frac{n}{2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(D(Z_i - X_i\beta))'(DRD)^{-1}(D(Z_i - X_i\beta))\right) \\
&= |DRD|^{-\frac{n}{2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(W_i - X_i\beta D)'(DRD)^{-1}(W_i - X_i\beta D)\right)
\end{aligned}
$$

If we define

$$
\begin{aligned}
\Sigma &= DRD \qquad (3.10) \\
\epsilon^* &= D(Z - X\beta) \qquad (3.11)
\end{aligned}
$$

We can re-write the likelihood under the expanded data model in equation 3.10 as

$$
p(Y, t_\alpha(W)|R, \beta)|J_\alpha(W)| \propto |\Sigma|^{-\frac{n}{2}} \exp tr\left(\Sigma^{-1}\epsilon^{*\prime}\epsilon^*\right) \qquad (3.12)
$$

The Prior: $p_0(\alpha|\theta)f(\theta)$

For Bayesian inference, we need to define a joint prior on $\theta = (\beta, R)$ and α. We assume that β and R are independent a priori so that $\pi(\beta, R, \alpha) = p_0(\alpha|R)f(R)f(\beta)$.

Under the transformation $\Sigma = DRD$, Barnard et al. (2000) showed that if we take Σ to be a standard inverse Wishart distribution as in A.4 we can re-write the distribution of Σ as in B.11:

$$
\pi(\Sigma) = \pi(\alpha, R) \times |J : \Sigma \to D, R| = f(R)p(\alpha|R) \qquad (3.13)
$$

Where with a particular choice of parameters, namely $\nu = T + 1$, the distribution $f(R)$ is as in 2.12 such as each r_{ij} is uniform on the interval

$[-1, 1]$. Furthermore, the distribution of $p_0(\alpha|R)$ is Gamma with shape parameter $(T+1)/2$ and rate parameter 1. Therefore, we are able to get the desired prior distribution $\pi(\alpha|R)\pi(R)$ by sampling Σ from a standard inverse Wishart with degrees of freedom $\nu = T+1$, and transforming using $\Sigma = DRD$.

Here, we like to point out that both the prior distributions of R and β are the same under the expanded model and the observed data model. This is a condition required for the PX-DA algorithm. In addition, we note that R and α are not a priori independent. The independence of these parameters is a necessary condition only to prove the optimality of the convergence of algorithm 3.3. In this case, their independence is not key since we are using PX-DA mainly for the convenience in that it results in a posterior distributions that is easily sampled from.

Posterior Distribution of (α, θ)

Now that we have specified the expanded likelihood and prior on the parameters of interest (R, β) and the expansion parameter α, the joint posterior distribution of (β, R, α) conditional on the latent data can be computed:

$$\beta, R, \alpha|Y, W \sim p(Y, t_\alpha(W)|\beta, R)|J_\alpha(W)|f(R)f(\beta)p_0(\alpha|R) \tag{3.14}$$

where $t_\alpha(W) = Z = D^{-1}W$ is the transformation of the latent data and $|J_\alpha(W)|$ is the determinant of the Jacobian of going from $Z \to W$.

We could therefore put together the likelihood in 3.12 and the marginally uniform prior on R in 2.12, the Gamma prior on α in 3.13, and the prior on β in 2.9, we get:

$$\begin{aligned} \pi(R, \alpha, \beta|Y, W) &\propto |\Sigma|^{-\frac{n}{2}} \exp tr\left(\Sigma^{-1}\epsilon^{*\prime}\epsilon^*\right) \\ &\times |R|^{\frac{T(T-1)}{2}-1}(\prod_i |R_{ii}|)^{-(T+1)/2} \times Gamma\left(\frac{T+1}{2}, 1\right) \\ &\times \exp(\beta'\psi_\beta^{-1}\beta) \end{aligned} \tag{3.15}$$

where the Gamma distribution is defined as in A.2.

In order to sample from the joint posterior distribution in 3.15, we use a Gibbs Sampling framework, where we sample $\beta|Z,R$ and then sample $R,\alpha|W$. Since given R, the parameter β is identifiable, we sample it prior to transforming the data.

Straightforward computations give the posterior distribution of $\beta|Y,Z,R$. The normal distribution is the conjugate prior, therefore the posterior distribution of β will also follow a multivariate normal distribution with mean parameters β^* and covariance Ψ^*_β where

$$\Psi^*_\beta = \Psi_\beta + \sum_{i=1}^{n} X_i' R^{-1} X_i$$

$$\beta^* = \Psi^{*-1}_\beta (\sum_{i=1}^{n} X_i' R^{-1} Z)$$

The joint posterior $\pi(R,\alpha|Y,W,\beta)$ can be obtained from 3.15:

$$\begin{aligned} \pi(R,\alpha|Y,W,\beta) &\propto |\Sigma|^{-\frac{n}{2}} \exp tr\left(\Sigma^{-1}\epsilon^{*\prime}\epsilon^*\right\} \\ &\times |R|^{\frac{T(T-1)}{2}-1} (\prod_i |R_{ii}|)^{-(T+1)/2} \times Gamma\left(\frac{T+1}{2}, 1\right) \end{aligned} \tag{3.16}$$

We perform a change of variable $\Sigma = DRD$:

$$\begin{aligned} \pi(\Sigma|Y,W,\beta) &\propto \pi(R,\alpha|Y,W,\beta) \times |J_\alpha : (D,R) \to \Sigma| \\ &= |\Sigma|^{-\frac{n}{2}} \exp tr\left(\Sigma^{-1}\epsilon^{*\prime}\epsilon^*\right\} \times |\Sigma|^{-\frac{1}{2}2(T+1)} \exp(-\frac{1}{2} tr(\Sigma^{-1})) \\ &= |\Sigma|^{-\frac{1}{2}(\nu+T+1)} \exp(-\frac{1}{2} tr(\Sigma^{-1} S)) \end{aligned} \tag{3.17}$$

This is an inverse Wishart distribution with $\nu = n+T+1$ and $S = \epsilon^{*\prime}\epsilon^*$. The second line in the equation above is obtained by reversing the steps of the proof in Appendix B.

Algorithm 3.4 Full PX-DA Sampling Scheme in Multivariate Probit

At iteration i

1. Imputation Step
   - Draw $Z \sim f(Z|Y, \beta, R)$ from a truncated Multivariate Normal distribution $TMVN(X\beta, R)$ as described in Appendix D.
2. Posterior Sampling Step Draw (β, R, α) jointly conditional on the latent data :
   - Draw $\beta|Z, Y, R$ from a Multivariate Normal distribution $\beta \sim MVN(\beta^*, \Psi^*_\beta)$
   - Draw $\alpha \sim p_0(\alpha|R)$ from a Gamma distribution $G(\frac{T+1}{2}, 1)$
   - Compute the diagonal matrix D, where each diagonal element $d_i = \sqrt{\frac{r^{ii}}{2\alpha_i}}$ and r^{ii} is the ith diagonal element of R^{-1}.
   - compute $W = t_\alpha(Z) = DZ$ or equivalently $\epsilon^* = D(Z - X\beta)$.
   - Draw $\Sigma|\beta, Y, W$ from an inverse Wishart distribution $\Sigma \sim IW(\nu, S)$ where $\nu = n + T + 1$ and $S = \epsilon^{*\prime}\epsilon^*$.
   - compute $R = D^{-1}\Sigma D^{-1}$

Repeat until convergence

3.4 Simulations

In order to test the performance of the algorithm developed in the previous section, we conduct several simulation studies first with $T = 3$ and then we increase the dimension to $T = 8$. The data is simulated as follows: we generate a design matrix with $p = 2$ covariates from a uniform distribution from $[-0.5, 0.5]$, we set the coefficients $\beta = (-1, 1)'$ and we generate random error from a multivariate Gaussian distribution centered at 0 and a full correlation matrix R. We fix R such that all ρ_{ij} off-diagonal elements are of equal value. We try for different values of ρ namely 0.2, 0.4, 0.6, and 0.8. The following two loss functions are considered to evaluate the accuracy of

the estimated correlation matrix:

$$L_1(\hat{R}, R) = tr(\hat{R}R^{-1}) - \log|\hat{R}R^{-1}| - T \quad (3.18)$$

$$L_2(\hat{R}, R) = tr(\hat{R}R^{-1} - I)^2 \quad (3.19)$$

Where $\hat{R}$ is the estimated correlation and R is the true correlation used to generate the data.

The first loss function is the entropy loss and the second is the quadratic loss. These loss functions are discussed in more detail in Yang and Berger (1994).

In each case, $N = 10000$ Gibbs samples are drawn and the first 500 are discarded as "Burn-in". We tried multiple runs, to ensure convergence of results. The correlation is always initialized at the identity matrix, and the latent variables are initialized at 0.

3.4.1 Results for $T = 3$

For $T = 3$, three parameters in the correlation matrix are estimated. Table 3.1 outlines results from the simulations for the correlation matrix. The posterior median estimate is reported, the number of parameters falling within the 95% credible interval, the average interval length, as well as the entropy loss and the quadratic loss. 95% credible intervals are calculated based on 2.5% and 97.5% quantiles of the estimates.

We can see that the likelihood carries more information with larger correlation values, estimation of the correlation becomes more accurate and confidence intervals become smaller on average. Similarly with more data, estimates become more precise and furthermore, we see a decrease in both the entropy and the quadratic loss. Except in one case ($r_{ij} = 0.2, n = 500$), the true correlation coefficient was always included in the 95% credible interval.

Figures 3.1 and 3.2, provide examples of traceplots and density plots for the correlation matrix with $\rho_{ij} = 0.4$ and $\rho_{ij} = 0.8$ respectively. Subfigures (a) and (b) in each case show how the density becomes narrower by increasing the sample size from $n = 100$ to $n = 1000$. Furthermore, we see

that the algorithm mixes very well and converges fast.

Table 3.1: *Correlation results from simulations for $T = 3$*

Sample Size	r_{ij}	CI Contains True	Average CI Length	Entropy Loss	Quadratic Loss
	0.2	3/3	0.644	0.206	0.557
	0.4	3/3	0.580	0.122	0.296
100	0.6	3/3	0.511	0.185	0.496
	0.8	3/3	0.423	0.336	0.923
	0.2	2/3	0.290	0.064	0.135
	0.4	3/3	0.269	0.031	0.061
500	0.6	3/3	0.226	0.051	0.102
	0.8	3/3	0.164	0.127	0.329
	0.2	3/3	0.202	0.028	0.056
	0.4	3/3	0.188	0.027	0.053
1000	0.6	3/3	0.165	0.037	0.075
	0.8	3/3	0.113	0.067	0.173

Table 3.2, shows simulation results for the regression coefficients β. For each coefficient, we report the median of the posterior distribution, a 95% credible interval and the standard error

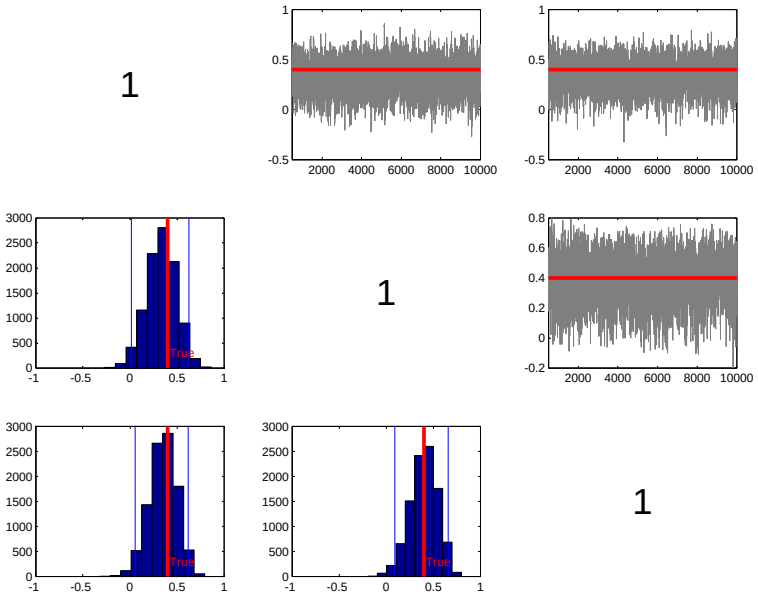
The true regression coefficients seems to always fall within the 95% credible interval. Standard errors and consequently credible intervals lengths tend to become smaller with the increase of correlation as well as the increase in sample size.

Figures 3.3, 3.4, 3.5, and 3.6 provide trace plots, density and autocorrelation plots for the regression coefficient in the case where the correlation matrix has elements $\rho_{ij} = 0.4$ and $\rho_{ij} = 0.8$ and increasing the sample size from $n = 100$ to $n = 1000$ respectively. The density becomes narrower with a larger sample size and here too, the algorithm seems to be mixing well.

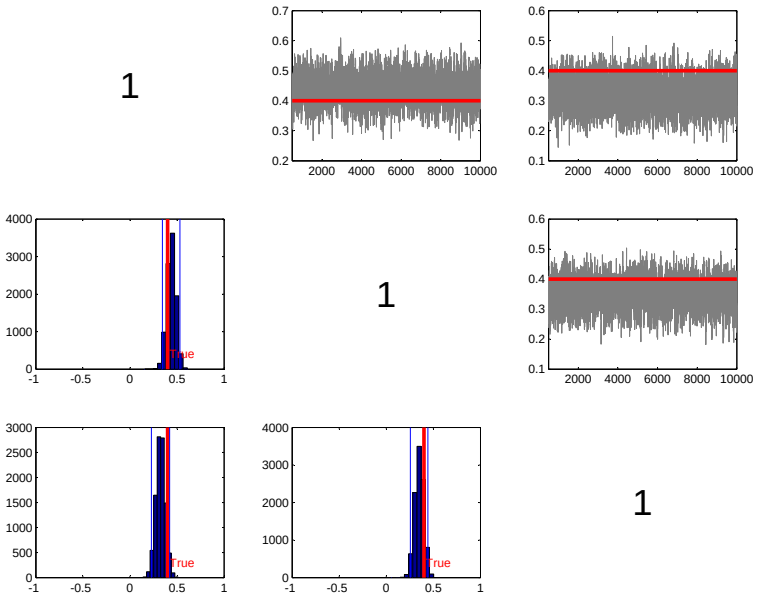
Table 3.2: *Regression coefficients results from simulations for $T = 3$*

Sample Size	r_{ij}	$\hat{\beta}_1$	Confidence Interval	Standard Error	$\hat{\beta}_2$	Confidence Interval	Standard Error
	0.2	-1.34	(-1.88,-0.83)	0.27	1.32	(0.79, 1.86)	0.27
	0.4	-1.20	(-1.72,-0.72)	0.26	0.88	(0.38, 1.37)	0.25
100	0.6	-0.99	(-1.47,-0.52)	0.24	0.88	(0.41, 1.36)	0.24
	0.8	-1.28	(-1.73,-0.82)	0.23	1.05	(0.62, 1.49)	0.22
	0.2	-1.22	(-1.45,-0.99)	0.12	1.18	(0.95, 1.40)	0.12
	0.4	-1.23	(-1.45,-1.00)	0.11	1.04	(0.82, 1.26)	0.11
500	0.6	-0.92	(-1.11,-0.71)	0.10	1.15	(0.93, 1.35)	0.11
	0.8	-1.14	(-1.33,-0.95)	0.10	0.93	(0.75, 1.11)	0.09
	0.2	-1.12	(-1.28,-0.96)	0.08	1.14	(0.98, 1.30)	0.08
	0.4	-1.09	(-1.25,-0.94)	0.08	0.96	(0.81, 1.12)	0.08
1000	0.6	-1.08	(-1.23,-0.93)	0.08	1.12	(0.97, 1.26)	0.08
	0.8	-1.09	(-1.22,-0.96)	0.07	0.98	(0.85, 1.11)	0.07

Figure 3.1: *Correlation estimates for* $\rho = 0.4$ *,* $T = 3$ *and increasing sample size from* $n = 100$ *to* n=1000

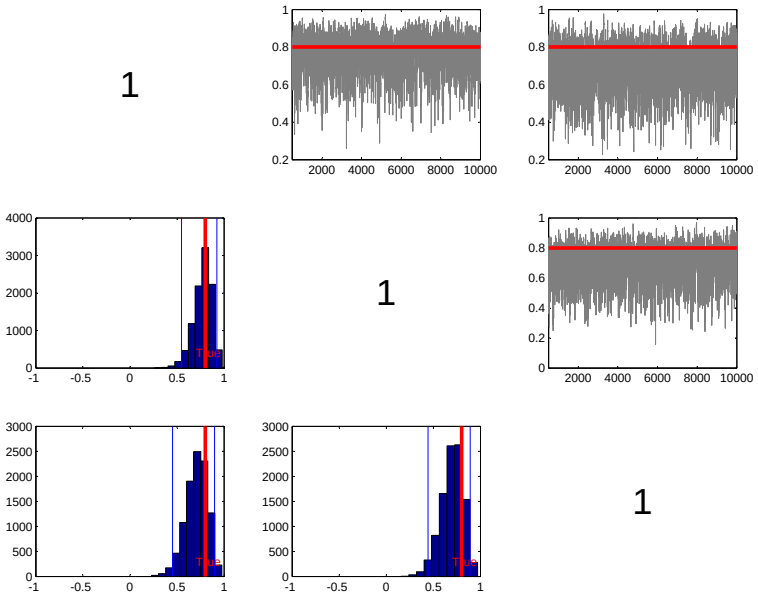


(a) $n = 100$

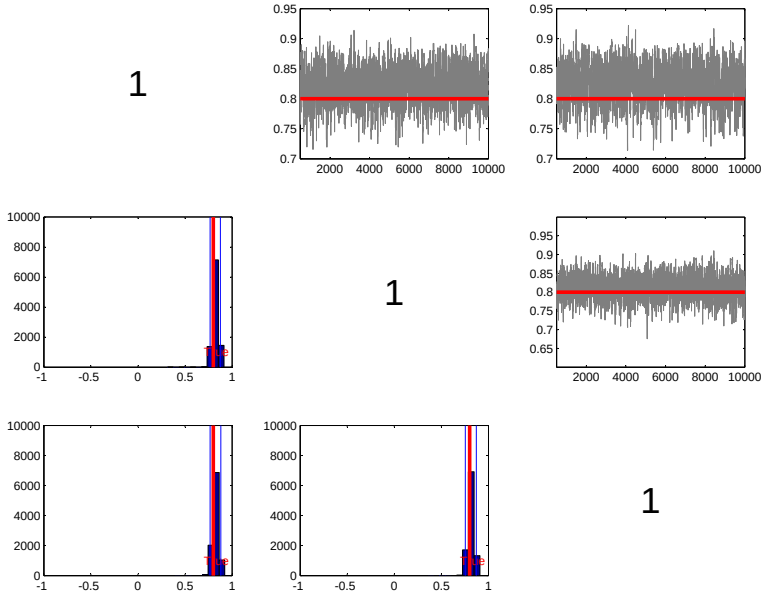


(b) $n = 1000$

Figure 3.2: *Correlation estimates for* $\rho = 0.8$ *,* $T = 3$ *and increasing sample size from* $n = 100$ *to* n=1000



(a) $n = 100$



(b) $n = 1000$

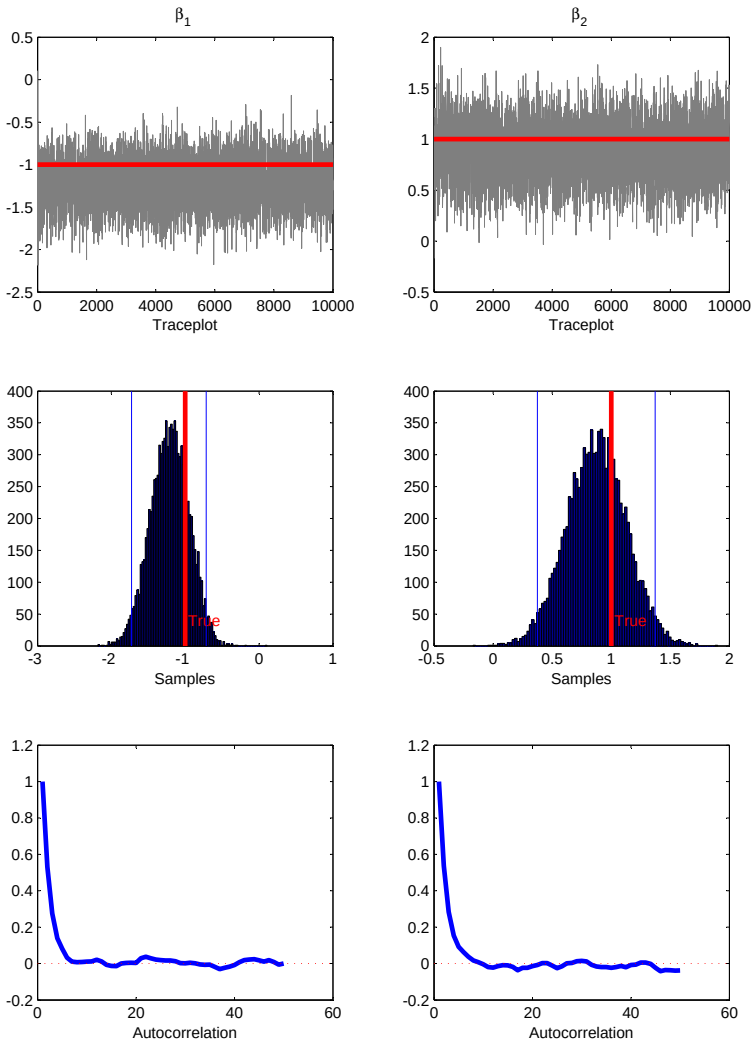


Figure 3.3: *β estimates for $\rho = 0.4$, $T = 3$ and sample size $n = 100$*

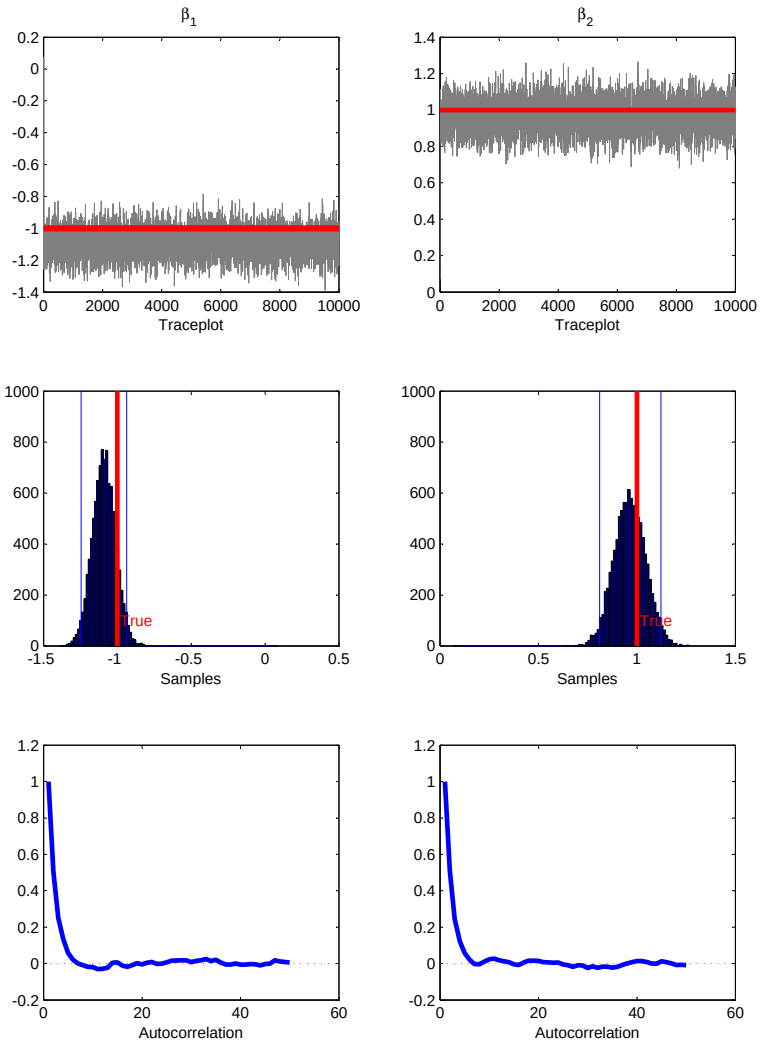


Figure 3.4: *β estimates for $\rho = 0.4$, $T = 3$ and sample size $n = 1000$*

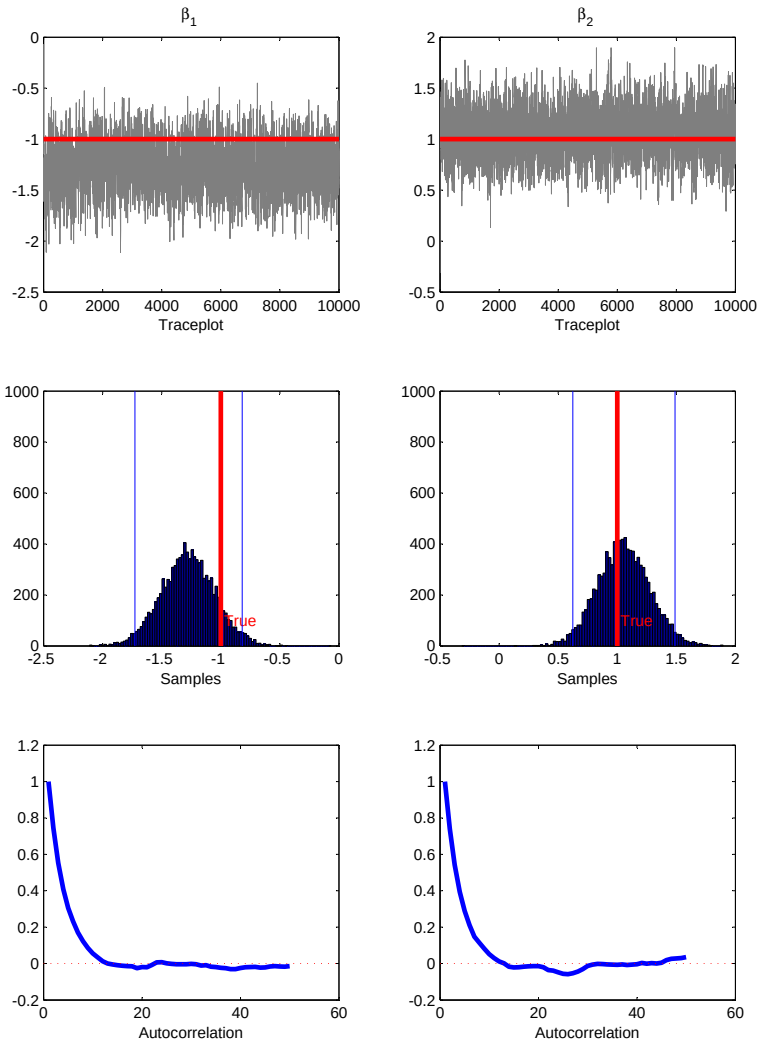


Figure 3.5: *β estimates for $\rho = 0.8$, $T = 3$ and sample size $n = 100$*

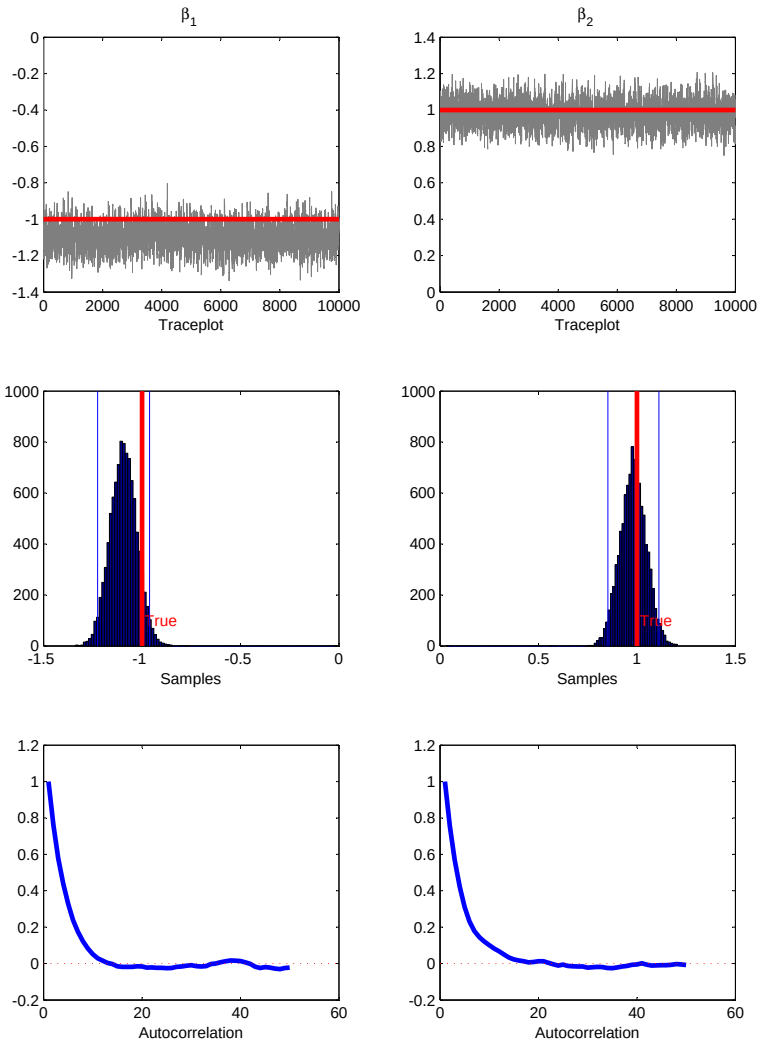


Figure 3.6: *β estimates for $\rho = 0.8$, $T = 3$ and sample size $n = 1000$*

3.4.2 Results for $T = 8$

For $T = 8$, we are estimating $T(T-1)/2 = 28$ parameters in the correlation matrix in addition to two regression coefficients. Table 3.3 shows the number of parameters falling within the 95% credible interval, the average interval length, the entropy loss and the quadratic loss.

In this case, we see that with more parameters to be estimated, we did not loose very much on the accuracy, as the average 95% credible interval length has remained within the same range as in the case of $T = 3$. We also note that five of the experiments had only one out of 28 parameters not contained in the 95% credible interval.

Furthermore, we could see that average CI length decreases with larger correlation values and a larger sample size, while loss improves only with an increased sample size.

Figure 3.7 and 3.8 show the density and the trace plot of the correlation matrices with $\rho_{ij} = 0.2$ and $\rho_{ij} = 0.6$ respectively. The density becomes more peaky and narrow with an increase in sample size as expected.

For the regression coefficients, we see from figures 3.9, 3.10, 3.11, and 3.12 that with more information in the data, the median is closer to the true parameters.

In addition, similar to the results we saw in the previous simulation, the credible intervals decrease in length with an increase in sample size. We also note from table 3.4, that the true value of β is always included in the 95% credible interval.

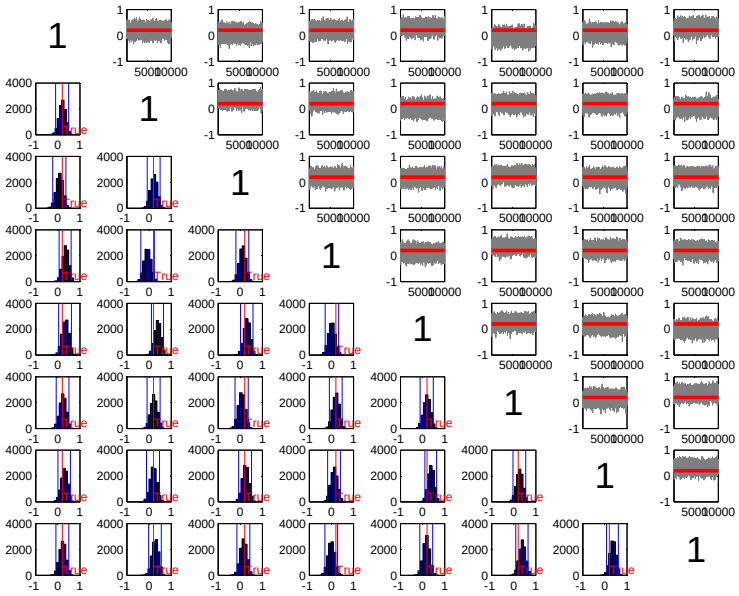
Table 3.3: *Correlation Results from simulations for $T = 8$*

Sample Size	r_{ij}	CI Contains True	Average CI Length	Entropy Loss	Quadratic Loss
	0.2	28/28	0.577	2.306	10.445
	0.4	27/28	0.511	4.327	29.895
100	0.6	28/28	0.492	4.879	75.270
	0.2	27/28	0.279	0.439	1.133
	0.4	28/28	0.254	0.489	1.312
500	0.6	27/28	0.221	0.560	1.436
	0.2	27/28	0.200	0.180	0.400
	0.4	28/28	0.181	0.186	0.421
1000	0.6	27/28	0.157	0.358	0.880

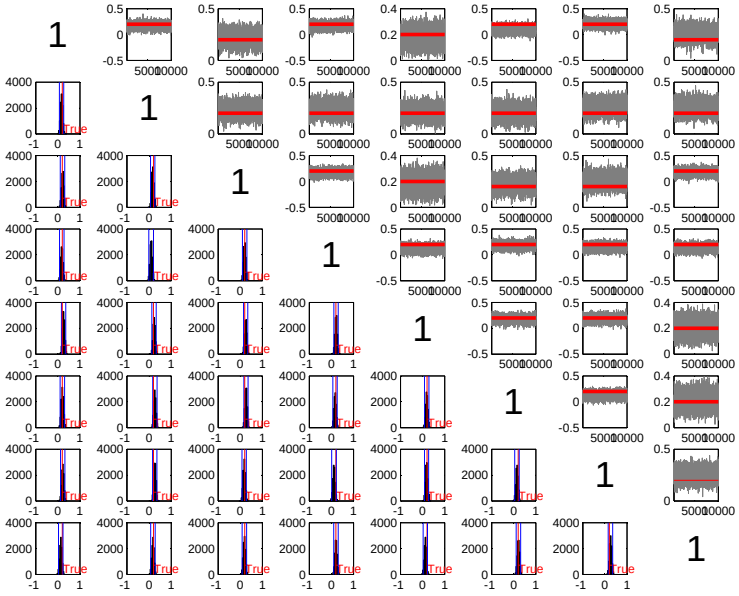
Table 3.4: *Regression coefficients results from simulations when $T = 8$*

Sample Size	r_{ij}	$\hat{\beta}_1$	Confidence Interval	Standard Error	$\hat{\beta}_2$	Confidence Interval	Standard Error
	0.2	-1.20	(-1.51,-0.89)	0.16	0.90	(0.60, 1.21)	0.15
	0.4	-1.08	(-1.37,-0.81)	0.14	0.99	(0.71, 1.27)	0.14
100	0.6	-1.06	(-1.33,-0.79)	0.14	1.04	(0.78, 1.31)	0.14
	0.2	-1.12	(-1.26,-0.98)	0.07	0.98	(0.85, 1.11)	0.07
	0.4	-1.01	(-1.14,-0.88)	0.07	0.95	(0.82, 1.08)	0.07
500	0.6	-1.04	(-1.16,-0.92)	0.06	0.97	(0.86, 1.09)	0.06
	0.2	-1.09	(-1.18,-0.99)	0.05	1.01	(0.91, 1.11)	0.05
	0.4	-1.00	(-1.09,-0.91)	0.05	0.94	(0.85, 1.03)	0.05
1000	0.6	-1.00	(-1.08,-0.91)	0.04	0.97	(0.88, 1.05)	0.04

Figure 3.7: *Correlation estimates for* $\rho = 0.2$ *,* $T = 8$ *and increasing sample size from* $n = 100$ *to* n=1000

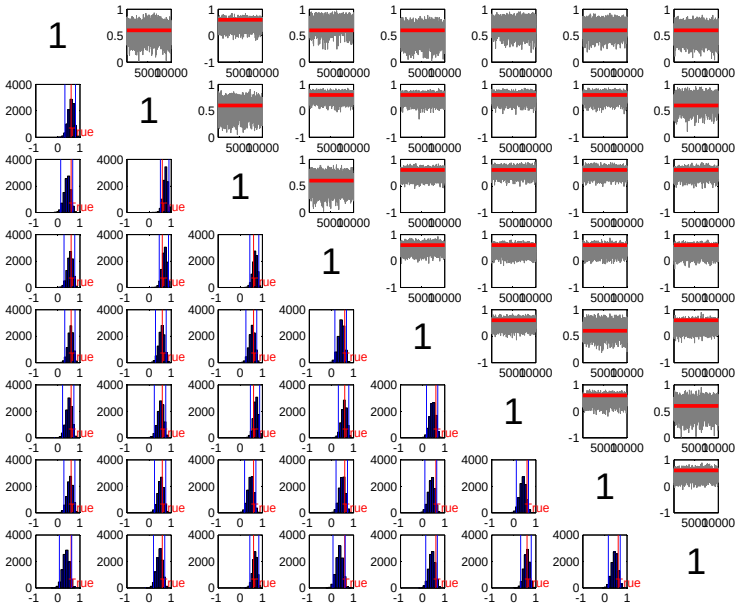


(a) $n = 100$

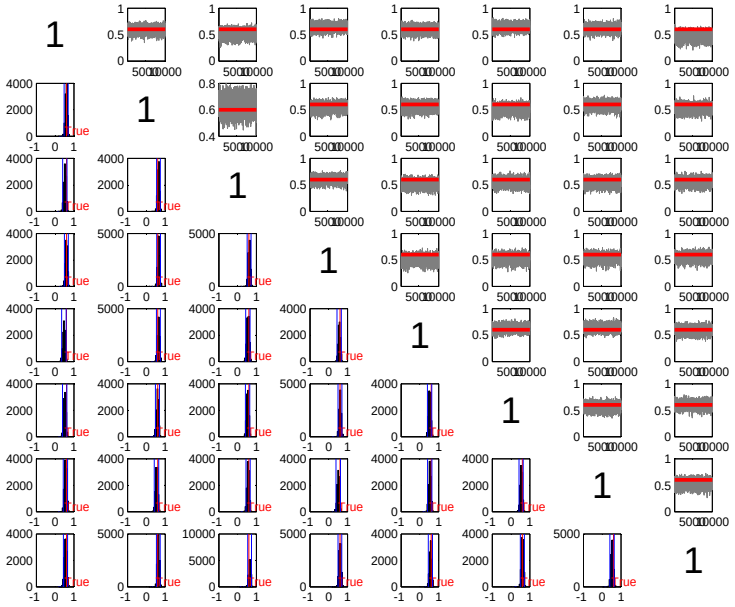


(b) $n = 1000$

Figure 3.8: *Correlation estimates for* $\rho = 0.6$ *,* $T = 8$ *and increasing sample size from* $n = 100$ *to* n=500



(a) $n = 100$



(b) $n = 500$

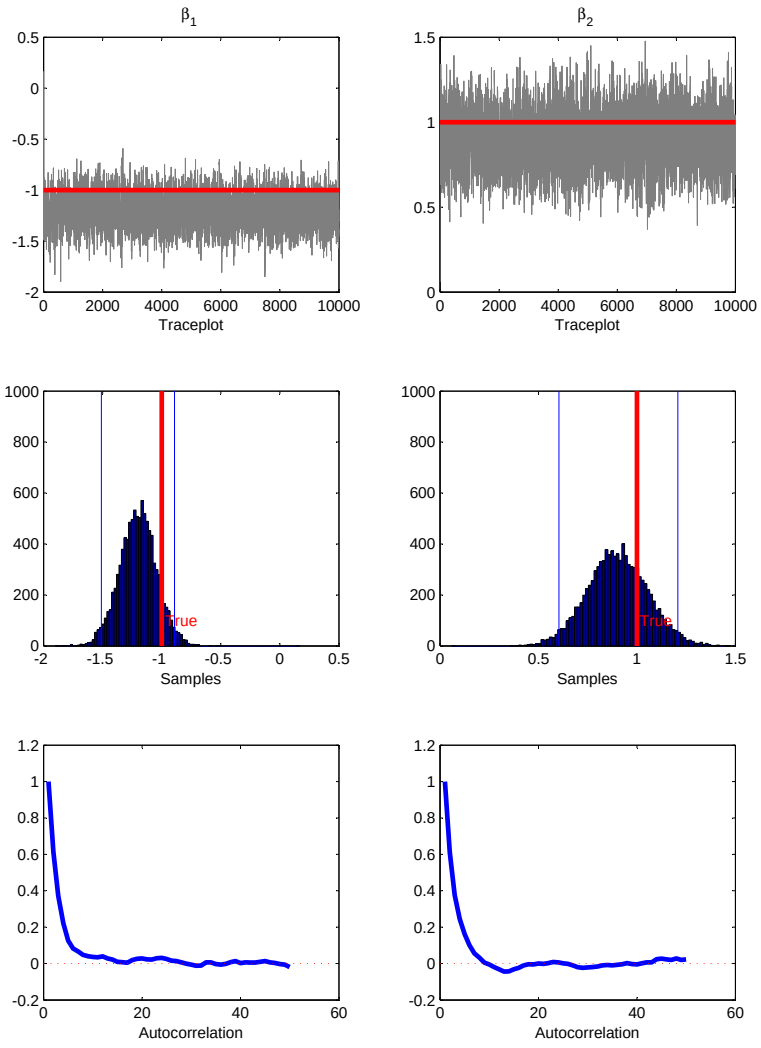


Figure 3.9: *β estimates for $\rho = 0.2$, $T = 8$ and sample size $n = 100$*

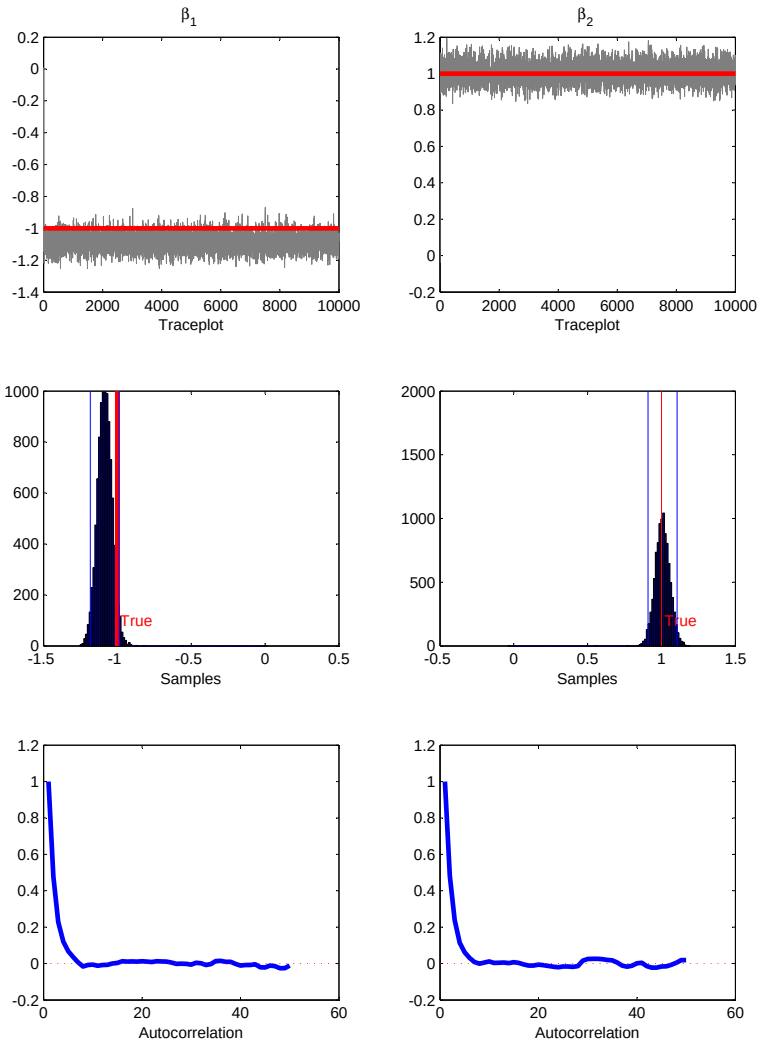


Figure 3.10: *β estimates for $\rho = 0.2$, $T = 8$ and sample size $n = 1000$*

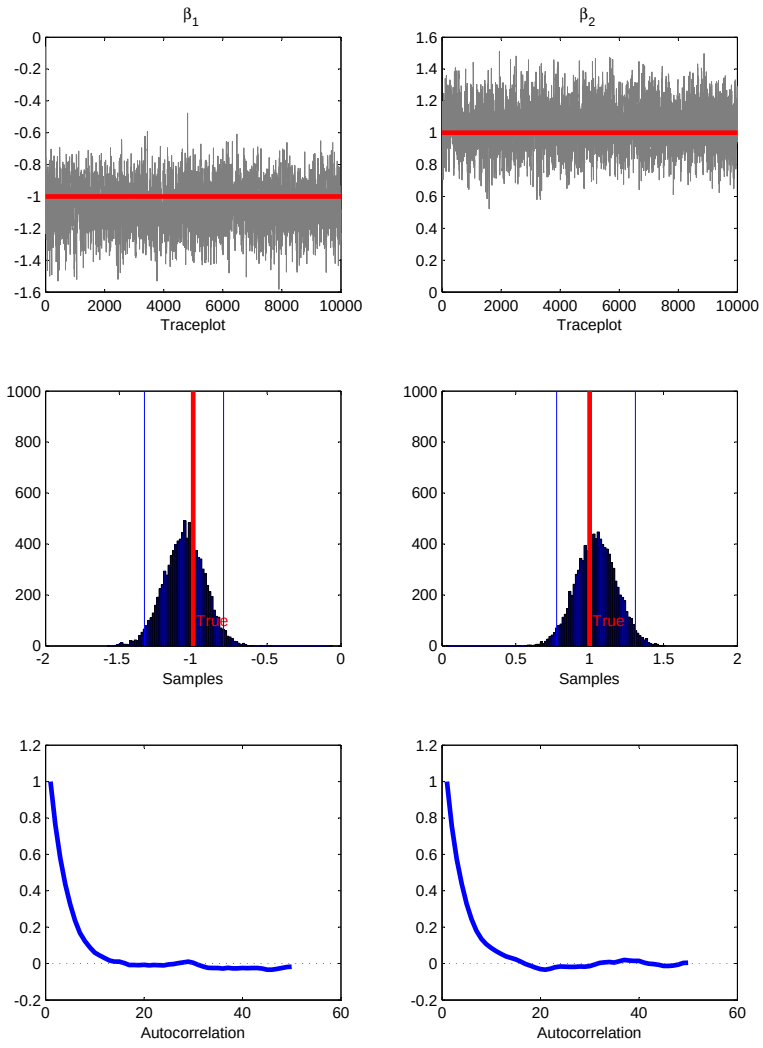


Figure 3.11: *β estimates for $\rho = 0.6$, $T = 8$ and sample size $n = 100$*

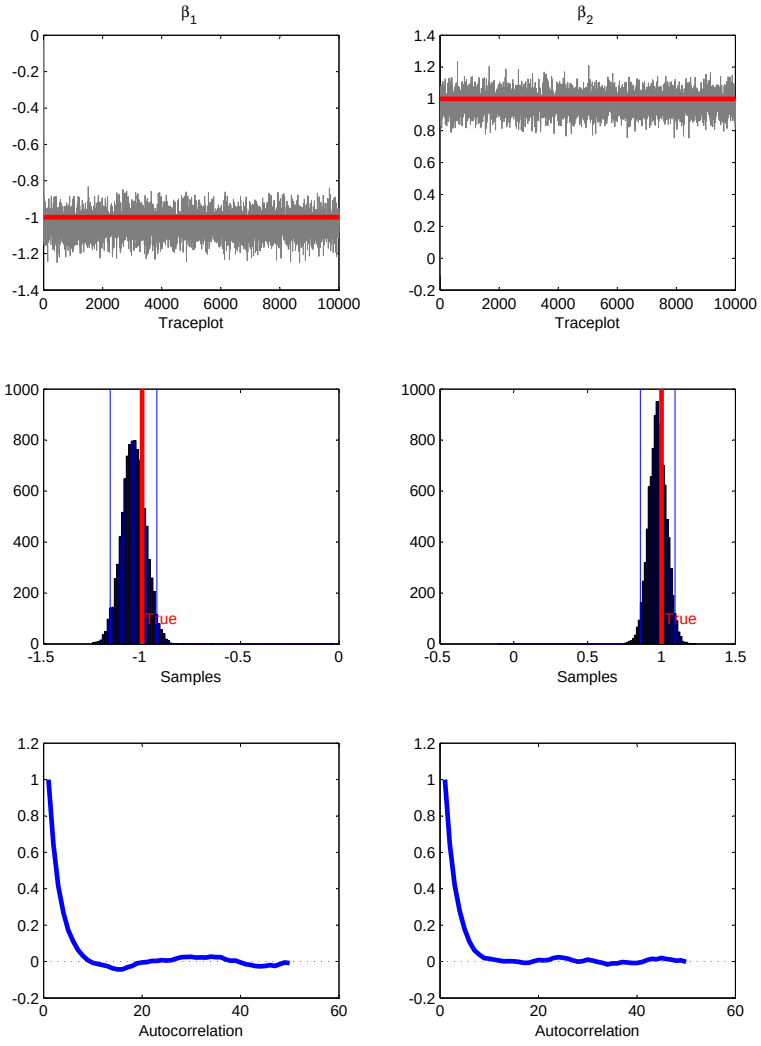


Figure 3.12: *β estimates for $\rho = 0.6$, $T = 8$ and sample size $n = 500$*

3.4.3 Convergence Assessment

As with any Markov Chain Monte Carlo Algorithm, it is important to ensure that the algorithm has converged. Unfortunately, there is no measure that can tell us definitively whether this has happened. The most common method of assessing convergence is by considering trace plots, which show the evolution of the MCMC output as a time series. These plots provide a simple way to examine the convergence behavior of the algorithm for the parameters under consideration. Trace plots are useful for immediately diagnosing lack of convergence and poor mixing, if the MCMC sampler covers the support of the posterior distribution very slowly. Poor mixing invalidates the density estimates, as it implies that the MCMC output is not a representative sample from the posterior distribution.

From previous figures, it appears that the algorithm is mixing well, since we do not see any particular trends in the time series plots, and the algorithm seems to be leveling off to a stationary state. We also consider the mixing speed by looking at the effect of drawing more samples. For examples, Figure 3.13 shows the effect of increasing the number of Gibbs draws from 500 to 5000, for the simulation where $T = 3$ and $n = 100$. We could see that with 1000 iterations post burn-in, the algorithm has started to converge to the target density.

Next we consider correlation plots of the estimated parameters. These plots depict the autocorrelation of the sequence of simulated values as a function of lag time. The autocorrelation plots can be useful in identifying slow mixing.

Figure 3.14 shows the standardized autocorrelation plots of the different values of the correlation coefficients when $T = 3$ and $n = 100$. Standardized plots mean that at lag time 0 the autocorrelation is 1. We could note that autocorrelation drops near zero at around lag 10-20. Similar results where observed for the regressions coefficients β, in previous plots.

Finally, under convergence, the estimated parameter of interest is expected to converge to a flat region near the true parameter and then fluctuate around that region. However, a statistic like the mean of the parameter

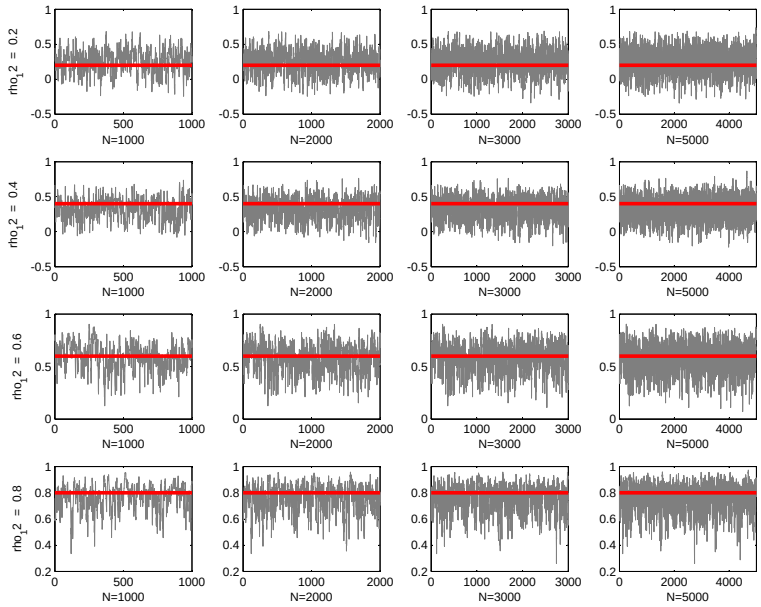


Figure 3.13: $n = 100$, $T = 3$, *Trace plots as the number of iterations increase from* $N = 500$ *to* $N = 5000$ *post "Burn-in". The algorithm has started to converge after about 1000 iteration post "Burn-in".*

is expected to converge in the limit to a constant. Diagnostic plots such as the cumulative mean and the cumulative standard deviation of the estimates provide a way to see if this has happened.

Figure 3.15 shows a plot of the cumulative means and standard deviations for the 10000 iteration from a randomly selected set of parameters from the simulation with $n = 100$ and $T = 3$. Both mean and standard deviation indeed level off to a flat line early on in the simulation. Furthermore, it appears that the value we have chosen to use for "Burn-in" is reasonable since it cuts off all the fluctuations that happen early on.

Finally, it is important to note that starting values are critical for the speed of convergence. If the starting values of a parameter are poor, it may take some time for the mean to stabilize. The fluctuations caused by the

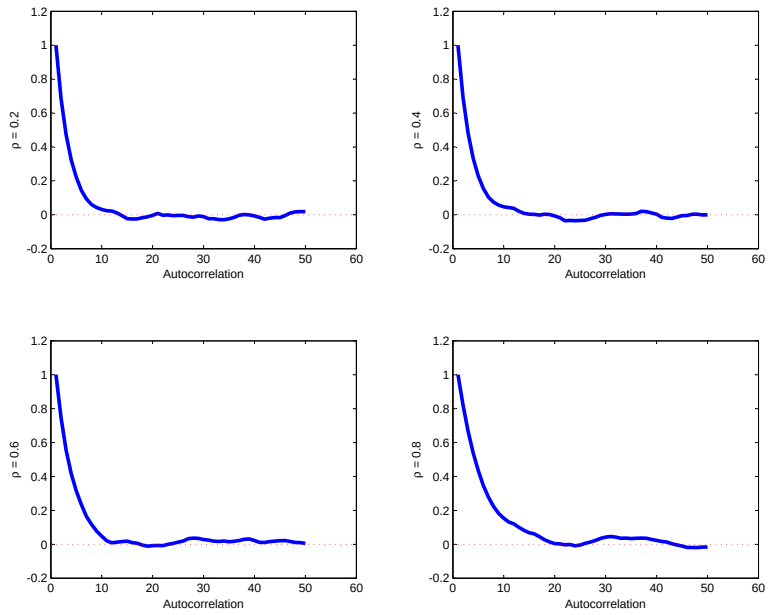


Figure 3.14: $n = 100$, $T = 3$, *Autocorrelation plots of a randomly chosen parameter from correlation matrices for the cases where the marginal correlations is* $\rho = 0.2$, $\rho = 0.4$, $\rho = 0.6$, *and* $\rho = 0.8$

poorly sampled values early on in the simulation are difficult to overcome, but in the long run, the mean will eventually stabilize. In the case of the Multivariate Probit, many have reported the sensitivity of convergence to the starting values of the parameters. In our case, many initialization values were tried, the initial values that we chose to use are equivalent to running T univariate Probit models with 0 mean, these values provided optimal results. When the algorithm was initialized randomly, it took much longer to reach convergence.

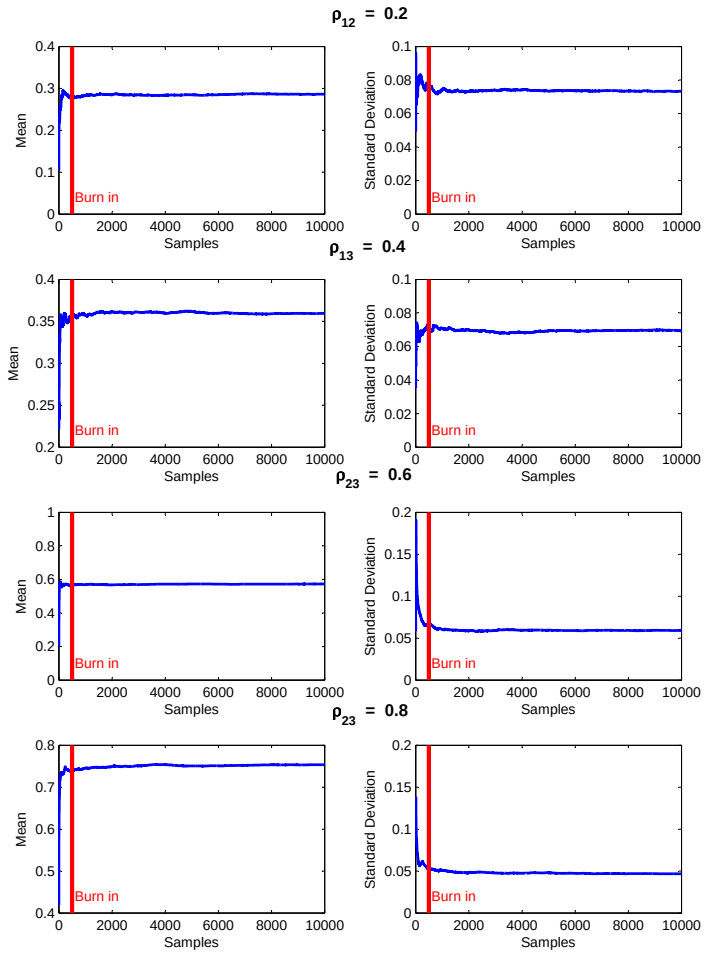


Figure 3.15: *Trace plots of the cumulative mean and cumulative standard deviation of randomly chosen parameters from correlation matrices as the correlation is varied from* $\rho = 0.2$, $\rho = 0.4$, $\rho = 0.6$, *and* $\rho = 0.8$ *and* $n = 100$, $T = 3$. *The vertical line marks the "Burn-in" value (500) used in the simulations*

3.5 Application: Six Cities Data

In order to further evaluate our method and our prior, we apply it to the Six Cities data. This data set is based on a subset of data from the Six Cities study, a longitudinal study of the health effects of air pollution, which has been analyzed by Chib and Greenberg (1998) in the context of multivariate Probit and by others (eg. Glonek and McCullagh (1995)) with a multivariate Logit model. The data contain repeated binary measures of the wheezing status (1 = yes, 0 = no) for 537 children at ages 7, 8, 9 and 10 years. The objective of the study is to model the probability of wheeze status Y_{ij} over time as a function of a binary indicator variable representing the mother's smoking habit during the first year of the study and the age of the child. We fit the same model as in Chib and Greenberg (1998):

$$P(Y_{ij} = 1|\beta, R) = \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \beta_3 X_{ij3} \tag{3.20}$$

where j in {1,2,3,4} indexes the time at which the response was observed (ages 7,8,9 and 10), and X_{ij1} is the age centered at 9 years $age_j - 9$, X_{ij2} is a binary variable representing the smoking status of the mother $X_{ij2} = X_{i2} = I_{mother-smokes}$, and X_{ij3} is the interaction between smoking status and age $X_{ij1} * X_{ij2}$. Here we would like to note that age is used both as a category in the response and as a covariate.

We use the algorithm developed in this chapter to fit a full correlation matrix among the responses. $N = 8000$ samples are obtained and the first 500 were discarded as "Burn-in". Furthermore, conforming with what was done in simulations methods, initial values for β were sampled from their prior distribution, the latent data was initialized at zero and the correlation matrix was initialized at the identity matrix.

Table 3.5 summarizes the parameters' posterior means and standard errors. It also provides, for comparative purposes, the results reported in Chib and Greenberg (1998) using both the maximum likelihood estimator and the posterior means resulting from the MCMC algorithm they develop in their paper. From comparing these results, we could see that they are very

similar for both means and standard errors. We could note however, that the estimates obtained using the joint uniform priors are smaller compared to ones we obtained using the marginally uniform prior. This is consistent with what is expected, since the jointly uniform prior will tend to favor values closer to 0.

We could also note that the intervals for β_2 and β_3 contain 0. This could be an indication that the mother's smoking habit may not have contributed to the wheezing status of the child at any age.

Table 3.5: *Six Cities Data: Posterior estimates using Marginal Prior, MLE estimate using MCEM and Posterior estimates using the Jointly Uniform Prior (Chib and Greenberg (1998))*

	Marginal Uniform Prior			MCEM		Jointly Uniform Prior	
	Mean	95% CI	s.e	MLE	s.e	Mean	s.e
$\hat{\beta}_0$	-1.13	(-1.26,-1.01)	0.06	-1.12	0.06	-1.13	0.06
$\hat{\beta}_1$	-0.08	(-0.14,-0.02)	0.03	-0.08	0.03	-0.08	0.03
$\hat{\beta}_2$	0.18	(-0.02, 0.38)	0.10	0.15	0.10	0.16	0.10
$\hat{\beta}_3$	0.04	(-0.06, 0.14)	0.05	0.04	0.05	0.04	0.05
r_{12}	0.59	(0.45, 0.73)	0.07	0.58	0.07	0.56	0.07
r_{13}	0.54	(0.38, 0.68)	0.08	0.52	0.08	0.50	0.07
r_{14}	0.55	(0.40, 0.69)	0.07	0.59	0.09	0.54	0.07
r_{23}	0.73	(0.60, 0.83)	0.06	0.69	0.05	0.66	0.06
r_{24}	0.57	(0.40, 0.69)	0.08	0.56	0.08	0.51	0.07
r_{34}	0.64	(0.40, 0.69)	0.08	0.63	0.08	0.60	0.06

In their paper, Chib and Greenberg (1998) do not show trace plots for any of their estimates and they do not discuss convergence diagnostics. It is therefore difficult to compare their algorithm to ours with that respect. Figures 3.16 and 3.17 depict the density plots and the trace plots of the correlation coefficients and the regression coefficients respectively. Trace plots do not seem to exhibit any patterns or poor mixing. The autocorrelation plot in 3.17 shows that there is a lag of 10 before autocorrelation goes down to 0.

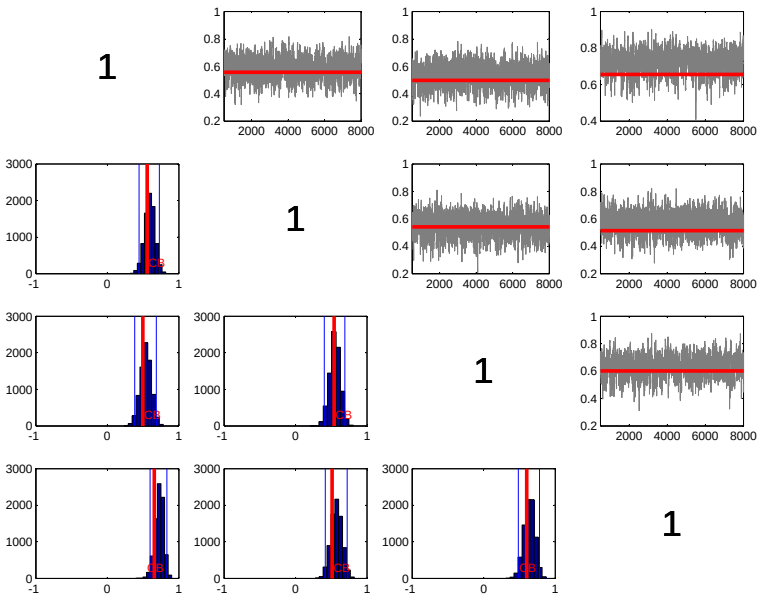


Figure 3.16: *Six Cities Data: Trace plots and density plots of the correlation coefficients. The vertical lines denote 95 % credible interval and the line in red indicates the posterior mean reported by Chib and Greenberg (1998).*

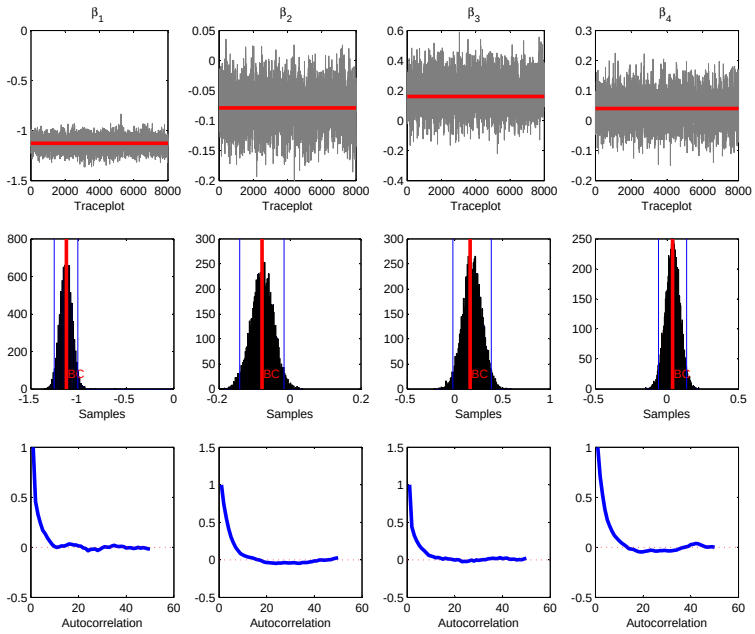


Figure 3.17: *Six Cities Data : Trace plots, density plots and autocorrelation plots of the regression coefficients. Vertical lines denote 95 % credible interval and the line in red indicates the posterior mean reported by Chib and Greenberg (1998).*

Chapter 4

Correlation Estimation in the Structured Model

4.1 Introduction

Often times, when dealing with high dimensional problems, it is useful to impose a structure of association between the outcome variables $Y_1, \ldots, Y_T$. In certain cases, the researcher may be interested in comparing different hypotheses of patterns of association. In other cases, the data might follow a natural grouping, where certain subsets of variables will express a higher degree of association within and less association between other groups of variables. For example, in longitudinal models, a variable might only be associated to the one preceding it at the previous time point. Furthermore, imposing a structure on the correlation matrix helps simplify computation. This is because the number of free parameters to be estimated is a smaller subset of the total number of parameters under the saturated model. Imposing a structure helps both statistically and computationally.

4.2 Conditional Independence

When variables exhibit a high marginal association, the pairwise estimate of their correlation is high. However, these two variables might be affected by a third variable, which acts as a mediating variable or confounder. As an example, consider the following three variables: smoking, drinking coffee and lung cancer. Drinking coffee and lung cancer might express a high degree of correlation, however once information about smoking is available, drinking coffee and lung cancer become decoupled. Therefore, smoking accounts for

the association between drinking coffee and lung cancer. In order to get around this problem, one could control for the other variables in the model, by considering the conditional dependence. Partial correlation represents the correlation between two variables conditional on all the other variables in the model. It can be obtained from the precision matrix $\Omega = \Sigma^{-1}$, with elements ω_{ij}, through the following relation:

$$\begin{cases} \tilde{\rho}_{ij} = -\frac{\omega_{ij}}{\sqrt{\omega_{ii}\omega_{jj}}} & \text{for } i \neq j; \\ \tilde{\rho}_{ii} = \rho_{ii} & \text{for } i = j. \end{cases} \tag{4.1}$$

From this, we could see that a 0 in the precision matrix Ω would result in a partial correlation of 0, meaning that the two variables are conditionally independent.

In general, the correlation coefficient is a weak criterion for measuring dependence because marginally most variables will be correlated. This implies that zero correlation is in fact a strong indicator for independence. On the other hand, partial correlation coefficients provide a strong measure of dependence and, correspondingly, offer only a weak criterion of independence.

In the multivariate Probit (MVP) class of models, the response is discrete and viewed as an indicator variable to a latent construct. Imposing a structure on the latent variables $Z_1, \ldots, Z_T$ (see figure 4.1) results in a partial correlation matrix that is sparse, however this does not imply that the correlation matrix is sparse. In this section we shift our attention to the partial correlation matrix using undirected Gaussian graphical models (GGM). We impose the structure on the partial correlation matrix and subsequently estimate marginal correlations given this structure.

4.3 Gaussian Graphical Models

Gaussian graphical models are a graphical representation of the conditional independence between multivariate Normal variables. In these graphs, variables are represented by nodes. An edge is drawn between any two nodes unless these two nodes are conditionally independent. This way, graphs pro-

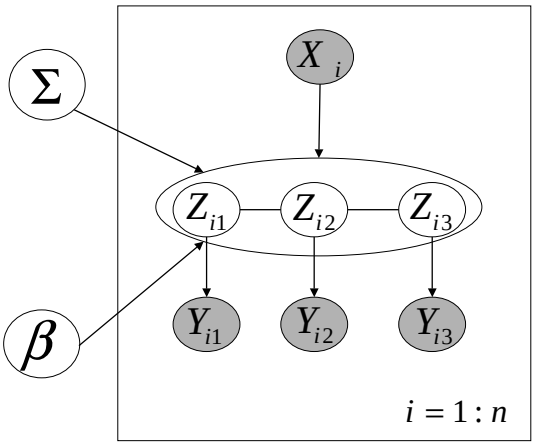


Figure 4.1: *A graphical representation of a structured MVP model for $T = 3$. The edge between Z_{i1} and Z_{i3} is missing, this is equivalent to $\tilde{r}_{13} = 0$. This structure is typical of longitudinal models where each variable is strongly associated with the one before it and after it, given the other variables in the model.*

vide a clear representation of the interrelationship between variables in the model. This approach to modeling data is known as covariance selection (Dempster, 1972). The introduction of the hyper-inverse Wishart by Dawid and Lauritzen (1993) as a conjugate prior for structured covariance matrices was central in the development of Bayesian approaches for inference in this class of models.

4.3.1 Graph Theory

In this section we review some graph theory used in this chapter, for a full account on graphical models, we refer the reader to Lauritzen (1996) or Whittaker (1990).

An undirected graph is a pair $G = (V, E)$, where V is a set of vertices representing variables and E, the edge-set, is a subset of the set of unordered distinct pair of vertices. Visually, each vertex i is a node representing the random variable i and an edge $(i, j) \in E$ is an undirected edge connect-

ing nodes i and j unless they are conditionally independent. In undirected graphical models if edge $(i, j) \in E$ then by symmetry edge $(j, i) \in E$ as well. See for example figure 4.1, Z_1 is conditionally independent of Z_3 given Z_2 and is denoted by $Z_1 \perp Z_3 | Z_2$.

Below is a list of definitions used throughout this chapter, they are illustrated in figure 4.2:

- Vertices A and B are *neighbors (nb)* or adjacent in G if there is an edge (a, b) in E.
- A *subgraph* is a graph which has as its vertices some subset of the vertices of the original graph.
- A graph or a subgraph is *complete* or fully connected if there is an edge connecting any two nodes.
- A *clique* is a complete subgraph.
- A set C is said to *separate* A from B if all paths from A to B have to go through C.
- Subgraphs (A, B, C) form a decomposition of G if $V = A \cup B, C = A \cap B$, where C is complete and separates A from B.
- A sequence of subgraphs that cannot be further decomposed are the *prime components* of a graph.
- A graph is said to be *decomposable* if every prime component is complete.
- A distribution P is Markov with respect to a graph G, if for any decomposition (A, B) of G, $A \perp B | A \cap B$, where $A \cap B$ is complete and separates A from B

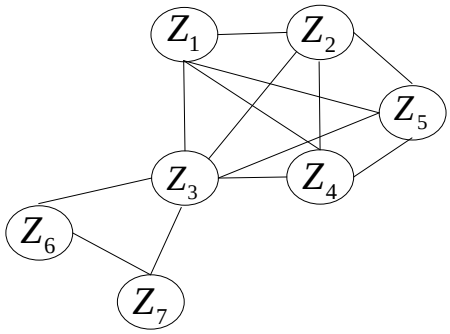


Figure 4.2: *A graphical model with $T = 7$ vertices. In this graph, Z_1 is a neighbor of Z_2. Z_3, Z_2, and Z_7 form a complete subgraph or a clique. This graph can be decomposed into two cliques $\{Z_1, Z_2, Z_3, Z_5, Z_4\}$ and $\{Z_3, Z_6, Z_7\}$. $\{Z_3\}$ separates the two cliques.*

4.3.2 The Hyper-inverse Wishart Distribution

The Hyper-inverse Wishart distribution, defined by Dawid and Lauritzen (1993), is a family of Markov probability distributions for structured covariance matrices on decomposable graphs. Given a graph structure G, the probability distribution for Σ consistent with G follows a Hyper-inverse Wishart distribution denoted by:

$$\Sigma \sim HIW_G(b, D)$$

whith degrees of freedom $b > 0$ and location parameter $D > 0$. The joint density of Σ decomposes as follows:

$$P(\Sigma|b, D) = \frac{\prod_{C \in C} P(\Sigma_C|b, D_C)}{\prod_{S \in S} P(\Sigma_S|b, D_S)} \tag{4.2}$$

For each clique C, Σ_C follows an inverse Wishart distribution $IW(b, D_C)$.

4.4 Marginally Uniform Prior for Structured Covariance

In the MVP model, the covariance matrix is restricted to a correlation matrix due to identifiability reasons discussed in Chapter 2. In order to extend the PX-DA algorithm developed in Chapter 3, it is necessary to find a marginally uniform prior on the structured correlation matrix R.

In decomposable graphical models, an edge joining any two variables i and j implies that the variables i and j belong to the same clique. This in turn means that the corresponding element in the inverse correlation matrix R^{-1} is not 0. On the other hand, if i and j do not belong to the same clique, the corresponding element in the inverse correlation matrix R^{-1} is exactly 0. The correlation matrix is obtained through matrix completion such that the resulting matrix R is positive definite. Therefore the elements of R corresponding to the marginal correlation of two variables not belonging to the same clique are not free parameters and their distributions are not of interest. On the other hand, non-zero elements in the partial correlation matrix indicate a high dependence among the corresponding variables, and since we would like an uninformative prior on the marginal correlations we would be interested in a prior on R such that:

$$\begin{cases} R^{-1}_{ij} = 0 & \text{if } i, j \notin C; \\ R_{ij} \sim U(-1, 1) & \text{if } i, j \in C. \end{cases} \tag{4.3}$$

From the properties of the Hyper inverse Wishart distribution (Roverato, 2002), we know that given a graph structure, if $\Sigma_G \sim HIW(b, D)$, then the covariance matrix of each prime component follows an inverse Wishart distribution with $\Sigma_{P_j P_j} \sim IW(b, D_{P_j P_j})$ for $j = 1, \ldots, k$ and k is the number of prime components. It is important to note here that the precision parameter b is common to all the prime components, and the location parameter $D_{P_j P_j}$ is specific to each one.

Since the covariance of each prime component follows an inverse Wishart distribution $\Sigma_P \sim IW(b, D_P)$, by taking $D_P = I_P$, we can obtain the distribution of the correlation matrix R_P for each prime component, as in Barnard

et al. (2000).

$$f_{P_j}(R|b = |P_j| + 1) \propto |R|^{\frac{|P_j|(|P_j|-1)}{2}-1} \left(\prod_i R_{ii} \right)^{-\frac{(|P_j|+1)}{2}} \tag{4.4}$$

where $|P_j|$ is the cardinality of the prime component P.

If we consider the pairwise distribution of any two variables i and j within the same clique, we could appeal to the marginalization property of the inverse Wishart to obtain the marginal distribution of each r_{ij} in that clique, based on the result used in the proof of Barnard et al. (2000) (see appendix B). Furthermore, in order to have a uniform distribution for each r_{ij} on $[-1, 1]$, we sample Σ_C from a standard inverse Wishart with degree of freedom parameter $b = |C_j| + 1$.

However, because we are restricted to have the parameter b common to all the prime components, using the parametrization in A.4 would require that all cliques have equal sizes. Alternatively, we could use the parametrization in Dawid and Lauritzen (1993) (see Appendix A.4). In that parametrization, sampling $\Sigma_C \sim IW(\delta, I_C)$ and taking $\delta = 2$ is equivalent and furthermore, it would insure that the precision parameter is independent of the size of the cliques and could therefore be common to all prime components.

Since the structure of the graph G is given, when i and j do not belong to the same clique, the corresponding element of the partial correlation matrix $\tilde{r}_{ij} = 0$ and r_{ij} is obtained through matrix completion.

To illustrate this prior, given the graph structure in 4.1, we sample from a standard hyper-inverse Wishart $\Sigma_G \sim (\delta = 2, I_C)$ and we transform back to R using the separation strategy: $\Sigma = DRD$. Figure 4.3 illustrates the marginal distribution of the elements of the correlation matrix, under a structured partial correlation assumption, based on 5000 draws.

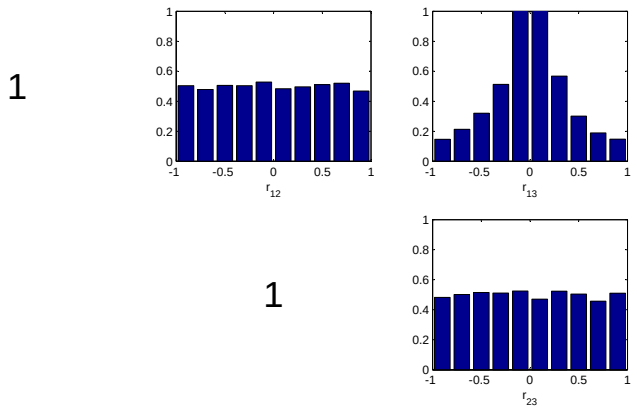
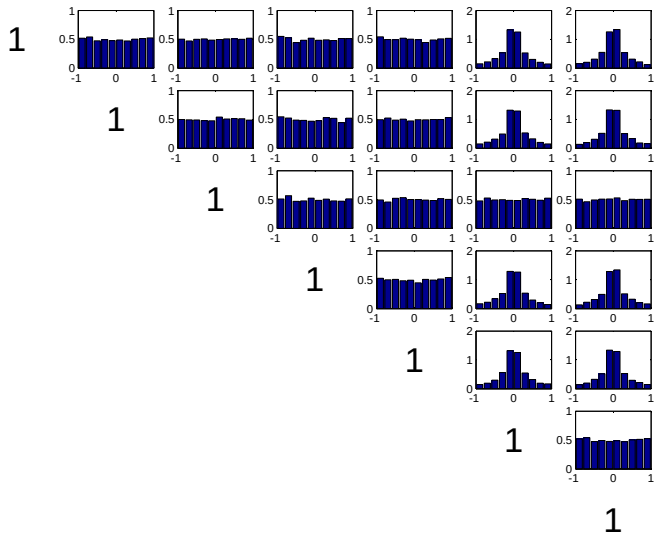


Figure 4.3: *Marginal distributions of the prior on the correlation matrix corresponding to the model in 4.1*

We could see that where there is an edge between the two variables, the corresponding element of the correlation matrix has a uniform distribution on $[-1, 1]$.

Furthermore, when two variables are conditionally independent, their marginal correlations are obtained to ensure that the correlation matrix is positive definite. Figure 4.4 illustrates the same result on a more complicated structure where the size of the cliques are not equal.

Figure 4.4: *Illustration of the marginally uniform prior on the structure of the graph in figure 4.2. In this graph we have unequal clique sizes where* $|C_1| = 5$ *and* $|C_2| = 3$



4.5 PX-DA in Gaussian Graphical Models

The PX-DA algorithm for the structured case is very similar to the one used for the saturated model in chapter 3. In the imputation step, the latent variables are sampled given a correlation matrix and regression coefficients. In the posterior sampling step, conditional on R, the regression coefficients are estimated in the same way as in chapter 3. The only difference is in sampling of the covariance matrix. Rather than sampling from an inverse Wishart distribution as before, we impose the structure on the partial correlation matrix, and subsequently expand the model by scaling it by D, a diagonal matrix, and sample the covariance from a hyper inverse Wishart distribution. Because D is diagonal, the zeros structure of the inverse correlation remains unchanged.

The multivariate Normal distribution of the latent variables W under the expanded model is also Markov with respect to G, and the joint likelihood decomposes as well:

$$P(W|\Sigma_G) = \frac{\prod_{C\in C} P(W_C|\Sigma_C)}{\prod_{S\in S} P(W_S|\Sigma_S)} \tag{4.5}$$

For Bayesian inference, the posterior distribution is given by Bayes rule:

$$P(\Sigma|W,G) \propto P(W|\Sigma,G)P(\Sigma|G)$$

The Hyper-inverse Wishart is the conjugate local prior for any graph, therefore we can compute the posterior as:

$$\begin{aligned} P(\Sigma|W,G) &= \frac{\prod_{C\in C} P(W_C|\Sigma_C)}{\prod_{S\in S} P(W_S|\Sigma_S)} \times \frac{\prod_{C\in C} P(\Sigma_C|b,D_C)}{\prod_{S\in S} P(\Sigma_S|b,D_S)} && (4.6)\\ &= \frac{\prod_{C\in C} P(W_C|\Sigma_C)P(\Sigma_C|b,D_C)}{\prod_{S\in S} P(W_S|\Sigma_S)P(\Sigma_S|b,D_S)} && (4.7) \end{aligned}$$

Since for any prime component P (cliques and separators), Σ_P is inverse Wishart, the conjugate prior for covariance for the Normal distribution, we can write for each prime component:

$$\begin{aligned} \pi(R_P,\alpha_P|Y_P,W_P) &\propto |\Sigma_P|^{-\frac{n}{2}} \exp tr\left(\Sigma^{-1}\epsilon^{*\prime}\epsilon^*\right\} \\ &\times |R_P|^{\frac{|P|(|P|-1)}{2}-1}(\prod_i |R_{Pii}|)^{-(|P|+1)/2} \times Gamma\left(\frac{|P|+1}{2},1\right) \end{aligned} \tag{4.8}$$

and doing a change of variable $D_P, R_P \rightarrow \Sigma_P$ as in (3.11), we get a posterior distribution which is an Inverse Wishart.

$$\Sigma_P \sim IW(\delta+n, D_P+S_W)$$

where n is the number of observations in W, and S_W is the cross product matrix $\epsilon^{*\prime}\epsilon^*$ under the expanded model. This way, an estimate of R is

obtained by sampling $\Sigma_P \sim HIW(\delta + n, D_P + S_Z)$ and transforming to R using the separation strategy as before (see algorithm **??**).

The algorithm to implement this method is identical to the one in **??**, except we replace the step:

- Draw $\Sigma|\beta, Y, W$ from an inverse Wishart distribution $\Sigma \sim IW(\nu, S)$ where $\nu = n + T + 1$ and $S = \epsilon^{*\prime}\epsilon^*$.

by

- Draw $\Sigma|\beta, Y, W$ from a hyper inverse Wishart distribution $\Sigma \sim HIW(\delta^*, S)$, where $\delta^* = 2 + n$ and $S = \epsilon^{*\prime}\epsilon^*$. See sampling procedure in Appendix E.

4.6 Simulations

The motivation behind imposing a structure to the model is to reduce the number of parameters to be estimated. In the saturated model, all the parameters in the correlation matrix need to be estimated. However, by imposing a structure on the inverse, the elements corresponding to correlation between conditionally independent variables are no longer free parameters. Since the number of free parameters to be estimated is reduced, the method that constrains the structure of the inverse should be more efficient at estimating the correlation matrix. This is particularly beneficial when the number of parameters is large in proportion to sample size. To illustrate this, we consider the model with $T = 8$ correlated binary responses. Under the saturated model assumption, a correlation matrix with $T(T-1)/2 = 28$ parameters needs to be estimated. Alternatively, if the longitudinal model assumption (as in figure 4.1) is made only 7 parameters would need to be estimated. This constitutes a significant reduction in the number of free parameters that could have a major impact especially for a small sample size.

All the simulations in this chapter where generated using a partial correlation matrix corresponding to the graph with a structure as in figure 4.1, with $T = 8$. Data is generated by sampling Z from a multivariate Gaussian distribution centered at 0 with sample size $n = 100$. A density model with no covariates is assumed and we set $Y = I(Z > 0)$. $N = 5000$ samples are drawn and the first 500 samples are discarded as "Burn-in".

4.6.1 Loss Under the Saturated Model and the Structured Model

In the first simulation, the PX-DA algorithm is tested under both saturated and structured covariance assumptions. To do that we generate 50 data sets from 50 different correlation structures corresponding to the graph in figure 4.1. Each time, we run our algorithm and record the entropy loss and the quadratic loss (see complete results in Appendix F). Figure 4.5 is a boxplot of the results and table 4.6.1 gives the mean and standard deviation for the

Loss	Model	Mean	s.e
Entropy	Saturated	13.875	25.164
	Constrained	1.267	1.225
Quadratic	Saturated	17098.952	119840.369
	Constrained	2.776	2.838

Table 4.1: *Simulation results: Entropy and quadratic loss averaged over 50 data sets generated by different correlation matrices with the same structure*

two estimated loss function resulting from these simulations. It is evident that the saturated model results in larger loss compared to the model where the structure is taken into account. To confirm, a paired t-test is performed. The difference in the means was significant in each case (pval $< 1 \times 10^{-5}$). We can note large variability in the loss under the saturated model.

4.6.2 Effect of Decreasing Sample Size

We would like to assess the effect of decreasing the proportion of parameters to samples size by varying the latter from $n = 100$ to $n = 200$. We generate data with the correlation matrix R, given in 4.9, such that the partial correlation matrix $\tilde{R}$, given in 4.10, has a structure corresponding to the model in figure 4.1.

$$R = \begin{bmatrix} 1.000 & 0.796 & 0.552 & 0.058 & -0.002 & -0.001 & -0.001 & -0.000 \\ 0.796 & 1.000 & 0.693 & 0.072 & -0.003 & -0.001 & -0.001 & -0.000 \\ 0.552 & 0.693 & 1.000 & 0.104 & -0.004 & -0.001 & -0.001 & -0.000 \\ 0.058 & 0.072 & 0.104 & 1.000 & -0.036 & -0.013 & -0.010 & -0.003 \\ -0.002 & -0.003 & -0.004 & -0.036 & 1.000 & 0.367 & 0.277 & 0.084 \\ -0.001 & -0.001 & -0.001 & -0.013 & 0.367 & 1.000 & 0.754 & 0.230 \\ -0.001 & -0.001 & -0.001 & -0.010 & 0.277 & 0.754 & 1.000 & 0.305 \\ -0.000 & -0.000 & -0.000 & -0.003 & 0.084 & 0.230 & 0.305 & 1.000 \end{bmatrix} \tag{4.9}$$

In table 4.6.2, we report the entropy loss and the quadratic loss defined in 3.19. We could see that under the structured model assumption, the

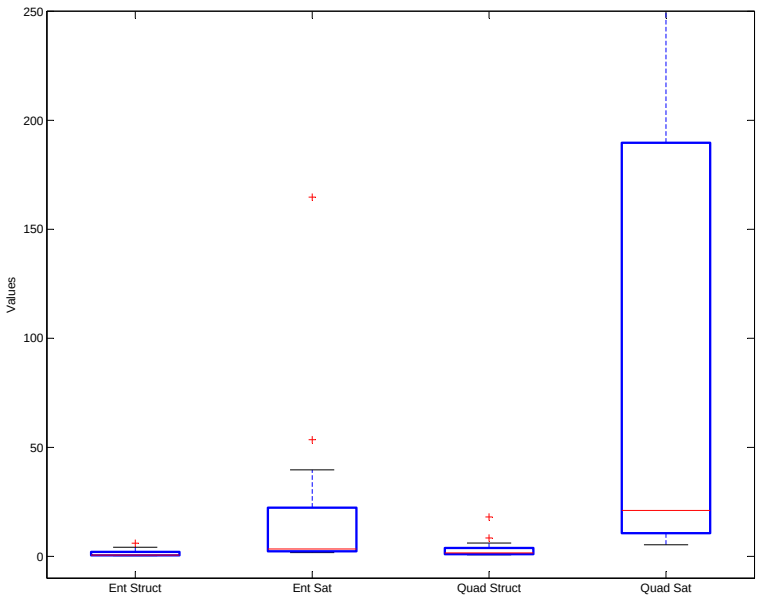


Figure 4.5: *Box plot of the entropy and quadratic loss obtained by generating data from 50 correlation structures and computing the loss function under the full correlation structure versus a structured correlation structure*

reduction in loss is significant both in estimating the marginal correlation and the partial correlation. Moreover, the loss is reduced for all cases with an increase in sample size.

We also note that under the structured model assumption, the loss in estimating the partial correlation matrix is the same as the one in estimating the correlation matrix, whereas under the saturated model assumption, the loss of estimating the partial correlation matrix is significantly larger than the loss incurred by estimating the marginal correlation matrix.

Table 4.3 and table 4.5 outline simulation results for estimating the correlation coefficients of the unconstrained parameters for $n = 100$ and $n = 200$ respectively.

$$\tilde{R} = \begin{bmatrix} 1.000 & 0.688 & -0.000 & 0.000 & 0.000 & -0.000 & 0.000 & -0.000 \\ 0.688 & 1.000 & 0.502 & -0.000 & -0.000 & 0.000 & -0.000 & 0.000 \\ 0.000 & 0.502 & 1.000 & 0.075 & 0.000 & -0.000 & 0.000 & 0.000 \\ -0.000 & 0.000 & 0.075 & 1.000 & -0.033 & -0.000 & 0.000 & -0.000 \\ 0.000 & -0.000 & 0.000 & -0.033 & 1.000 & 0.251 & 0.000 & -0.000 \\ -0.000 & 0.000 & 0.000 & -0.000 & 0.251 & 1.000 & 0.714 & -0.000 \\ 0.000 & -0.000 & -0.000 & 0.000 & 0.000 & 0.714 & 1.000 & 0.206 \\ -0.000 & 0.000 & -0.000 & -0.000 & -0.000 & -0.000 & 0.206 & 1.000 \end{bmatrix} \tag{4.10}$$

In the table, ρ^s are the correlation coefficients under a structured assumption, ρ are the correlation coefficients under the saturated model, $\tilde{\rho}$ are the partial correlation coefficients under the saturated model and $\tilde{\rho}^s$ are the partial correlation coefficients under the structured model. The results show that both the saturated and the structured model give similar results in estimating the unconstrained parameters of the marginal correlations. The standard errors and the 95% credible interval are very similar and just as we have seen in chapter 3, they are reduced by half in increasing the sample size from $n = 100$ to $n = 200$. However, in estimating the partial correlation parameters the strucured model has a smaller standard error and shorter credible intervals.

n	Correlation	Model	Entropy Loss	Quadratic Loss
100	Marginal	Saturated	2.428	11.166
	Marginal	Structured	0.415	0.830
	Partial	Saturated	11.394	679762.050
	Partial	Structured	0.415	0.839
200	Marginal	Saturated	1.179	4.139
	Marginal	Structured	0.219	0.479
	Partial	Saturated	1.145	20323.973
	Partial	Structured	0.241	0.547

Table 4.2: *Entropy and Quadratic loss obtained by estimating the true correlation and partial correlation matrix with the PX-DA algorithm under the saturated and structured model assumption*

More important differences between the structured model versus the saturated model are noted in the results of estimating the constrained parameters of the correlation and partial correlation matrix. Tables 4.4 and 4.6 show the simulation results for the constrained parameters when $n = 100$ and $n = 200$ respectively. The standard errors and 95% credible intervals are smaller under the structured model in comparison with the saturated model for marginal correlations, and the structured model gives exact results for partial correlations.

Table 4.3: *Simulation results on the unconstrained correlation coefficients corresponding to the model in 4.1, with $n = 100$, $T = 8$ based on $N = 5000$ Gibbs samples.*

	True	Mean	Median	s.e	95% CI	CIContains True	Interval Length
ρ_{12}	0.796	0.707	0.714	0.096	(0.500,0.874)	yes	0.374
ρ^s_{12}	0.796	0.701	0.707	0.092	(0.503,0.856)	yes	0.352
$\tilde{\rho}^s_{12}$	0.688	0.573	0.574	0.098	(0.378,0.758)	yes	0.380
$\tilde{\rho}_{12}$	0.688	0.707	0.714	0.096	(0.500,0.874)	yes	0.374
ρ_{23}	0.693	0.709	0.719	0.099	(0.492,0.876)	yes	0.384
ρ^s_{23}	0.693	0.695	0.704	0.095	(0.483,0.858)	yes	0.375
$\tilde{\rho}^s_{23}$	0.502	0.557	0.559	0.102	(0.351,0.743)	yes	0.392
$\tilde{\rho}_{23}$	0.502	0.709	0.719	0.099	(0.492,0.876)	yes	0.384
ρ_{34}	0.104	0.216	0.220	0.142	(-0.070,0.484)	yes	0.555
ρ^s_{24}	0.104	0.188	0.190	0.143	(-0.100,0.455)	yes	0.555
$\tilde{\rho}^s_{24}$	0.075	0.135	0.132	0.106	(-0.071,0.349)	yes	0.420
$\tilde{\rho}_{24}$	0.075	0.216	0.220	0.142	(-0.070,0.484)	yes	0.555
ρ_{45}	-0.036	-0.052	-0.051	0.149	(-0.334,0.237)	yes	0.571
ρ^s_{25}	-0.036	-0.046	-0.045	0.143	(-0.326,0.238)	yes	0.564
$\tilde{\rho}^s_{25}$	-0.033	-0.042	-0.042	0.131	(-0.298,0.215)	yes	0.513
$\tilde{\rho}_{25}$	-0.033	-0.052	-0.051	0.149	(-0.334,0.237)	yes	0.571
ρ_{56}	0.367	0.347	0.353	0.132	(0.076,0.597)	yes	0.521
ρ^s_{34}	0.367	0.326	0.332	0.136	(0.041,0.575)	yes	0.535
$\tilde{\rho}^s_{34}$	0.251	0.251	0.251	0.112	(0.031,0.473)	yes	0.442
$\tilde{\rho}_{34}$	0.251	0.347	0.353	0.132	(0.076,0.597)	yes	0.521
ρ_{67}	0.754	0.674	0.681	0.105	(0.443,0.858)	yes	0.415
ρ^s_{34}	0.754	0.643	0.650	0.106	(0.411,0.829)	yes	0.418
$\tilde{\rho}^s_{34}$	0.714	0.614	0.619	0.108	(0.385,0.806)	yes	0.421
$\tilde{\rho}_{34}$	0.714	0.674	0.681	0.105	(0.443,0.858)	yes	0.415
ρ_{78}	0.305	0.065	0.067	0.148	(-0.229,0.350)	yes	0.579
ρ^s_{34}	0.305	0.069	0.070	0.144	(-0.215,0.343)	yes	0.559
$\tilde{\rho}^s_{34}$	0.206	0.053	0.052	0.110	(-0.165,0.266)	yes	0.431
$\tilde{\rho}_{34}$	0.206	0.065	0.067	0.148	(-0.229,0.350)	yes	0.579

Table 4.4: *Simulation results on the constrained correlation coefficients corresponding to the model in 4.1, with $n = 1000$, $T = 8$ based on $N = 5000$ Gibbs samples.*

	True	Mean	Median	s.e	95% CI	CIContains True	Interval Length
ρ_{13}	0.552	0.568	0.576	0.119	(0.314,0.773)	yes	0.459
ρ^s_{13}	0.552	0.489	0.490	0.103	(0.287,0.689)	yes	0.402
$\tilde{\rho}^s_{13}$	-0.000	-0.000	-0.000	0.000	(-0.000,0.000)	yes	0.000
$\tilde{\rho}_{13}$	-0.000	0.157	0.159	0.200	(-0.260,0.524)	yes	0.785
ρ_{24}	0.072	0.032	0.031	0.148	(-0.254,0.324)	yes	0.577
ρ^s_{24}	0.072	0.130	0.129	0.102	(-0.073,0.334)	yes	0.408
$\tilde{\rho}^s_{24}$	-0.000	-0.000	-0.000	0.000	(-0.000,0.000)	yes	0.000
$\tilde{\rho}_{24}$	-0.000	-0.069	-0.075	0.196	(-0.443,0.316)	yes	0.758
ρ_{46}	-0.004	0.078	0.081	0.152	(-0.231,0.366)	yes	0.596
ρ^s_{46}	-0.004	-0.008	-0.004	0.034	(-0.089,0.057)	yes	0.146
$\tilde{\rho}^s_{46}$	0.000	-0.000	-0.000	0.000	(-0.000,0.000)	yes	0.000
$\tilde{\rho}_{46}$	0.000	0.149	0.152	0.185	(-0.225,0.507)	yes	0.732
ρ_{57}	-0.013	-0.014	-0.015	0.147	(-0.295,0.274)	yes	0.569
ρ^s_{57}	-0.013	-0.016	-0.011	0.052	(-0.129,0.086)	yes	0.215
$\tilde{\rho}^s_{57}$	-0.000	-0.000	-0.000	0.000	(-0.000,0.000)	yes	0.000
$\tilde{\rho}_{57}$	-0.000	-0.005	-0.002	0.185	(-0.368,0.347)	yes	0.715

Table 4.5: *Simulation results on the unconstrained correlation coefficients corresponding to the model in 4.1, with $n = 200$, $T = 8$ based on $N = 5000$ Gibbs samples.*

	True	Mean	Median	s.e	95% CI	CIContains True	Interval Length
ρ_{12}	0.796	0.716	0.720	0.066	(0.575,0.834)	yes	0.259
ρ^s_{12}	0.796	0.721	0.726	0.065	(0.581,0.831)	yes	0.250
$\tilde{\rho}^s_{12}$	0.688	0.577	0.579	0.073	(0.433,0.713)	yes	0.280
$\tilde{\rho}_{12}$	0.688	0.520	0.528	0.113	(0.275,0.717)	yes	0.442
ρ_{23}	0.693	0.725	0.729	0.068	(0.582,0.845)	yes	0.263
ρ^s_{23}	0.693	0.731	0.736	0.065	(0.588,0.844)	yes	0.255
$\tilde{\rho}^s_{23}$	0.502	0.591	0.593	0.074	(0.442,0.736)	yes	0.295
$\tilde{\rho}_{23}$	0.502	0.525	0.536	0.117	(0.271,0.725)	yes	0.454
ρ_{34}	0.104	0.126	0.127	0.102	(-0.076,0.322)	yes	0.398
ρ^s_{24}	0.104	0.128	0.130	0.104	(-0.082,0.325)	yes	0.407
$\tilde{\rho}^s_{24}$	0.075	0.086	0.086	0.071	(-0.053,0.225)	yes	0.278
$\tilde{\rho}_{24}$	0.075	0.127	0.128	0.135	(-0.150,0.382)	yes	0.532
ρ_{45}	-0.036	-0.163	-0.163	0.103	(-0.363,0.038)	yes	0.401
ρ^s_{25}	-0.036	-0.155	-0.157	0.103	(-0.353,0.052)	yes	0.405
$\tilde{\rho}^s_{25}$	-0.033	-0.141	-0.143	0.094	(-0.323,0.046)	yes	0.369
$\tilde{\rho}_{25}$	-0.033	-0.149	-0.151	0.117	(-0.373,0.081)	yes	0.455
ρ_{56}	0.367	0.396	0.399	0.097	(0.193,0.577)	yes	0.384
ρ^s_{34}	0.367	0.383	0.387	0.094	(0.193,0.558)	yes	0.365
$\tilde{\rho}^s_{34}$	0.251	0.263	0.263	0.073	(0.126,0.409)	yes	0.282
$\tilde{\rho}_{34}$	0.251	0.311	0.316	0.137	(0.021,0.562)	yes	0.542
ρ_{67}	0.754	0.759	0.763	0.066	(0.619,0.877)	yes	0.258
ρ^s_{34}	0.754	0.741	0.746	0.066	(0.599,0.854)	yes	0.255
$\tilde{\rho}^s_{34}$	0.714	0.695	0.699	0.071	(0.547,0.822)	yes	0.275
$\tilde{\rho}_{34}$	0.714	0.761	0.766	0.074	(0.598,0.893)	yes	0.295
ρ_{78}	0.305	0.295	0.297	0.102	(0.087,0.488)	yes	0.401
ρ^s_{34}	0.305	0.298	0.303	0.098	(0.097,0.477)	yes	0.381
$\tilde{\rho}^s_{34}$	0.206	0.205	0.205	0.072	(0.064,0.346)	yes	0.282
$\tilde{\rho}_{34}$	0.206	0.190	0.197	0.150	(-0.117,0.471)	yes	0.588

Table 4.6: *Simulation results on the constrained correlation coefficients corresponding to the model in 4.1, with $n = 200$, $T = 8$ based on $N = 5000$ Gibbs samples.*

	True	Mean	Median	s.e	95% CI	CIContains True	Interval Length
ρ_{13}	0.552	0.584	0.587	0.082	(0.411,0.732)	yes	0.322
ρ^s_{13}	0.552	0.528	0.530	0.074	(0.380,0.668)	yes	0.288
$\tilde{\rho}^s_{13}$	-0.000	-0.000	-0.000	0.000	(-0.000,0.000)	yes	0.000
$\tilde{\rho}_{13}$	-0.000	0.136	0.140	0.145	(-0.151,0.409)	yes	0.560
ρ_{24}	0.072	0.060	0.061	0.103	(-0.153,0.260)	yes	0.412
ρ^s_{24}	0.072	0.093	0.094	0.077	(-0.060,0.243)	yes	0.303
$\tilde{\rho}^s_{24}$	-0.000	-0.000	-0.000	0.000	(-0.000,0.000)	yes	0.000
$\tilde{\rho}_{24}$	-0.000	-0.067	-0.067	0.147	(-0.365,0.212)	yes	0.577
ρ_{35}	-0.004	-0.025	-0.025	0.106	(-0.229,0.174)	yes	0.403
ρ^s_{35}	-0.004	-0.020	-0.016	0.024	(-0.081,0.015)	yes	0.097
$\tilde{\rho}^s_{35}$	0.000	0.000	0.000	0.000	(-0.000,0.000)	yes	0.000
$\tilde{\rho}_{35}$	0.000	0.097	0.098	0.141	(-0.187,0.367)	yes	0.554
ρ_{46}	-0.013	-0.076	-0.076	0.108	(-0.288,0.137)	yes	0.425
ρ^s_{46}	-0.013	-0.060	-0.057	0.043	(-0.151,0.018)	yes	0.170
$\tilde{\rho}^s_{46}$	-0.000	-0.000	-0.000	0.000	(-0.000,0.000)	yes	0.000
$\tilde{\rho}_{46}$	-0.000	-0.059	-0.059	0.148	(-0.354,0.223)	yes	0.577

4.6.3 Prediction Accuracy

An important benefit of regression models, is that they allow the prediction of a new outcome Y^* given the model parameters and a new set of covariates X^*. In the multivariate Probit class of models Y is binary, therefore we are interested in $P(Y^*_{ij} = 1|X^*, Y)$. Rather than choosing a point estimator to make predictions, a Bayesian approach averages over the parameter space. This is given by the posterior predictive distribution:

$$\Pr(Y^*|Y, X, X^*) = \int \Pr(Y^*|X^*, \beta, R)p(\beta, R|Y, X)d\beta dR \quad (4.11)$$

$$= \int \Pr(Y^*|Z)p(Z|X^*, \beta, R)p(\beta, R|Y, X)d\beta dRdZ \quad (4.12)$$

Therefore the predictive probabilities can be approximated by:

$$\Pr(Y^* = 1|Y, X, X^*) \approx \frac{1}{N} \sum_{i=1}^{n} \mathbb{I}_{(Z^{(i)}>0)} \tag{4.13}$$

where

$$Z^{(i)} \sim p(Z|X^*, \beta^{(i)}, R^{(i)})$$

and N is the number of Gibbs samples collected. Therefore for each draw of β and R, we sample according to the likelihood $p(Z|X^*, \beta^{(i)}, R^{(i)})$. This involves sampling the latent variables from the multivariate normal distribution with mean $X^*\beta^{(i)}$ and correlation matrix $R^{(i)}$ and setting $Y = \mathbb{I}_{(Z>0)}$. By Averaging over Monte Carlo samples we can obtain the predictive probability $\Pr(Y^* = 1|Y, X, X^*)$ which is used to predict a new observation Y^*. The predictive accuracy is then computed by counting the number of properly predicted values and dividing by the total number of predictions.

To further investigate the advantages of estimating a structured correlation matrix, we compare the predictive accuracy under the assumption of a saturated model and a structured model. This time we use a data set with $T = 25$ and $n = 100$. In this data set, we expect the structured model to perform better than the saturated model, because the ratio of sample size to parameters to be estimated is much larger in the saturated case. In this simulation, we use a density model with no covariates and using the PX-DA algorithm, we sample $N = 5000$ draws from the joint conditional posterior distribution under both the full and the structured correlation assumption.

As expected, the results demonstrate the superiority of the structured model with 0.83 accuracy over the saturated model with 0.68 accuracy.

4.7 Application: Six Cities Data Revisited

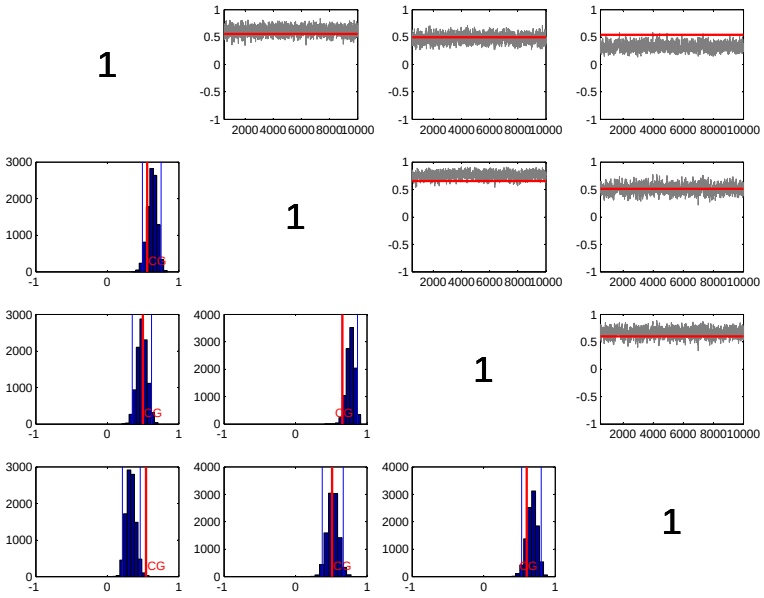
In this section, we revisit the Six Cities data, this time imposing the longitudinal structure on the covariance while fitting the exact model as 3.20. From table 4.7 and figure 4.6, we could see that the posterior mean of the marginal correlations coefficient do not correspond to the ones obtained under the

saturated model. This is particularly true for $\hat{\rho_{12}}$, $\hat{\rho_{23}}$ and $\hat{\rho_{34}}$. These correlations are now larger (posterior mean > 0.6) and the second order partial correlations are weaker and third order partial correlations are weaker still. Furthermore, the standard errors of the parameters that are constrained to zero in the inverse are smaller than the ones obtained in table 3.5. It is interesting to note however, that the estimates for β (Figure 3.17) and the standard errors have remained unchanged under the structured model. In order to assess predictive accuracy in this case, we use the method discussed in the previous section, we fit the model on a random subset of the data with $n = 100$ and we evaluate the predictive accuracy on $n = 100$ observations randomly selected from the remaining observations. The structured model has a slightly higher predictive accuracy (0.83) compared to the saturated model(0.80).

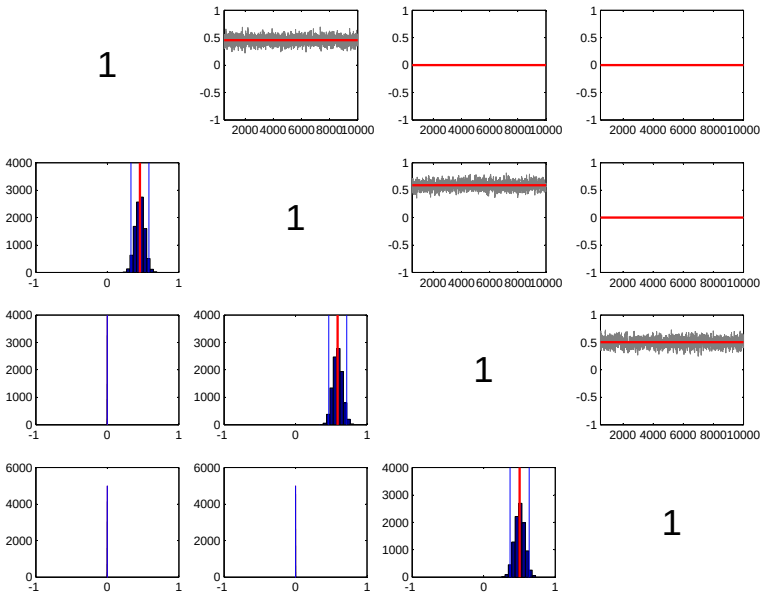
	Marginal Uniform Prior			MCEM		Jointly Uniform Prior	
	Mean	95% CI	s.e	MLE	s.e	Mean	s.e
$\hat{\beta}_0$	-1.14	(-1.26,-1.02)	0.06	-1.12	0.06	-1.13	0.06
$\hat{\beta}_1$	-0.08	(-0.15,-0.01)	0.03	-0.08	0.03	-0.08	0.03
$\hat{\beta}_2$	0.17	(-0.03, 0.37)	0.10	0.15	0.10	0.16	0.10
$\hat{\beta}_3$	0.04	(-0.07, 0.15)	0.06	0.04	0.05	0.04	0.05
r_{12}	0.63	(0.49, 0.75)	0.07	0.58	0.07	0.56	0.07
r_{13}	0.48	(0.35, 0.62)	0.07	0.52	0.08	0.50	0.07
r_{14}	0.33	(0.21, 0.46)	0.06	0.59	0.09	0.54	0.07
r_{23}	0.77	(0.66, 0.87)	0.05	0.69	0.05	0.66	0.06
r_{24}	0.52	(0.66, 0.87)	0.07	0.56	0.08	0.51	0.07
r_{34}	0.68	(0.66, 0.87)	0.07	0.63	0.08	0.60	0.06

Table 4.7: *Six Cities Data: Posterior estimates under structured model assumption, MLE estimate using MCEM and Posterior estimates using the Jointly Uniform Prior under a saturated model assumption(Chib and Greenberg (1998))*

Figure 4.6: *Six Cities Data: Correlation and partial correlation estimates*



(a) Marginal distribution and trace plots of the elements of the correlation matrix the red line denotes the estimates obtain in Chib and Greenberg (1998) by assuming the saturated model.



(b) Posterior distribution and trace plots of the elements of the partial correlation matrix.

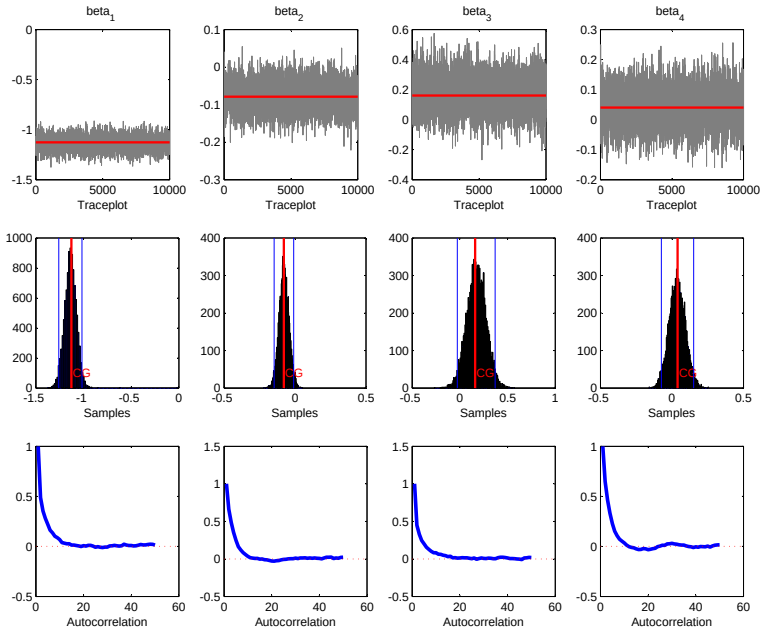


Figure 4.7: *Six Cities Data : Trace plots, density plots and autocorrelation plots of the regression coefficients under a structured model assumption. Vertical lines denote 95 % credible interval and the line in red indicates the posterior mean reported by Chib and Greenberg (1998).*

Chapter 5

Conclusion

5.1 Summary

The multivariate Probit model has several attractive features which make it particularly suitable for the analysis of correlated binary data. It relaxes the independence of the irrelevant alternatives (IIA) property assumed by the logistic model and moreover, it is a natural choice in situations where an interpretation for thresholded continuous data is possible. It allows for flexible modeling of the association structure underlying the latent data and automatically accounts for overdispersion and underdispersion.

Maximum likelihood estimation is not feasible in closed form in the multivariate Probit class of models. Likelihood based approaches for estimation in MVP are very expensive due to the intractability of the high dimensional integral that needs to be solved. The Bayesian framework is attractive because it allows the computation of a full posterior distribution on all unknown parameters. The algorithm we proposed in 3.4 and extended in chapter 4 for structured models uses parameter expansion for data augmentation, which gives full conditional posterior distributions in closed form. This allows the implementation of a Gibbs sampler. Moreover, the algorithm we developed has many desirable properties:

- It handles the identifiability problem in the MVP model by constraining the covariance to be a correlation matrix, and placing the prior directly on the identified parameters.
- The posterior distribution obtained through parameter expansion allows the use of the standard conjugate prior for the covariance. This makes the parameters easily interpretable.

- The prior is marginally uniform and does not favor marginal correlations close to 0 or ± 1 even in high dimensions. Furthermore, it is proper which makes it possible to do Bayesian model selection.
- The full Gibbs framework is convenient as it bypasses having high dimensional proposal distributions. From previous work (Zhang et al., 2006), the design of such proposal distributions is difficult and requires careful tuning of the algorithm parameters.

The extension of the algorithm provided in chapter 4, using Gaussian graphical models, greatly improves estimation and simplifies computation, especially in high dimensional space, or when the proportion of parameters to sample size is high.

Computation difficulties in our algorithm arise mainly from the sampling of univariate truncated Gaussian. We use the algorithm of Robert (1995), which is based on an accept/reject method (see appendix D). For certain simulations, accepting values was slow and this significantly slowed down the algorithm. This problem was more apparent in the estimation of a full covariance.

For $T = 8$ and $N = 5000$, the program was taking on average 5 minutes to run in `Matlab` and for $T = 25$, $N = 5000$, it was taking about 30 minutes to run. In future work, it would be important to implement a different method for sampling the univariate truncated Gaussian and try to speed up computation in order to allow for scalability of the algorithm to higher dimensions.

5.2 Extensions, Applications, and Future Work

A natural and straightforward extension of the algorithm developed here is to multinomial and ordinal Probit, where the latent variables are thresholded to multiple intervals. In addition, the extension to a response consisting of a mixture of binary and continuous data could be interesting and useful for many applications.

Furthermore, future work would include an extension of the structured algorithm of chapter 4 to the case where the structure is unknown a priori, but where we would be interested to learn it from data. The method we proposed here is particularly suitable for this task. The Gaussian graphical model framework and the choice of a proper prior make model selection feasible.

There are many applications of the multivariate Probit model, since correlated binary data arise in many settings. Biological, medical, and social studies often yield binary or dichotomous data due to the lack of adequate and direct continuous measurements. Other examples include longitudinal data, panel data, latent class models in psychology. The MVP class of models is also particularly attractive in marketing research of consumer choice which subsequently yields to market segmentation and product design. In addition, one interesting application would be Bayesian Distance Metric Learning. This application is particularly important for classification. Yang et al. (2007) developed an approach that uses logistic regression to learn a similarity metric given labeled binary examples of similar and dissimilar pairs. In their approach, they ignore identifiability, which is also a concern for logistic models, and they use a variational approximation. Our model and algorithm could be easily adapted to this problem.

Bibliography

J. Albert and S. Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422): 669–679, jun 1993.

J. Ashford and R. Snowden. Multivariate probit analysis. *Biometrics*, 26 (3):535–546, sep 1970.

J. Barnard, R. McCulloch, and X. Meng. Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, 10:1281–1311, 2000.

C. Carvalho, H. Massam, and M West. Simulation of the hyper-inverse wishart distribution in graphical models. *Biometrika*, 2007. To appear.

N. Chaganty and H. Joe. Efficiency of generalized estimating equations for binary responses. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(4):851–860, 2004.

S. Chib and E. Greenberg. Analysis of multivariate probit models. *Biometrika*, 85(2):347–361, 1998.

R. Cowell, A. Dawid, S. Lauritzen, and D. Spiegelhalter. *Probabilistic Networks and Expert Systems*. Springer, 1999.

A. Dawid and S. Lauritzen. Hyper markov laws in the statistical analysis of decomposable graphical models. *The Annals of Statistics*, 21(3):1272–1317, sep 1993.

A. Dempster. Covariance selection. *Biometrics*, 28(1):157–175, mar 1972.

D. Dobra, C. Hans, B. Jones, J. Nevins, G. Yao, and M. West. Sparse graphical models for exploring gene expression data. *J. Multivariate analysis*, 90:196–212, 2004.

Y. Edwards and G. Allenby. Multivariate analysis of multiple response data. *Journal of Marketing Research*, 40(3):321–334, 2003.

J. Geweke. Efficient simulation from the multivariate normal and student–t distributions subject to linear constraints. *Computing Science and Statistics: Proceedings of the Twenty–Third Symposium on the Interface, Alexandria, VA: American Statistical Association, pp.*, 1991.

J. Geweke, M. Keane, and D. Runkle. Alternative computational approaches to inference in the multinomial probit model. *The Review of Economics and Statistics*, 76(4):609–632, nov 1994.

G. Glonek and P. McCullagh. Multivariate logistic models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(3):533–546, 1995.

A. Gupta and D. Nagar. *Matrix Variate Distributions*. Chapman and Hall, 2000.

M. Keane. A note on identification in the multinomial probit model. *Journal of Business and Economic Statistics*, 10(2):193–200, apr 1992.

G. Koop. *Bayesian Econometrics*. John Wiley and Sons, 2003.

S. Lauritzen. *Graphical Models*. Clarendon Press, Oxford, 1996.

S. Lerman and C.F. Manski. *On the Use of Simulated Frequencies to Approximate Choice Probabilities*. MIT Press, 1981.

M. Linardakis and P. Dellaportas. Assessment of athenss metro passenger behaviour via a multiranked probit model. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 52:185–200(16), May 2003.

C. Liu. Bayesian analysis of multivariate probit models - discussion on the art of data augmentation by van dyk and meng. *Journal of Computational and Graphical Statistics*, 10:75–81, 2001.

J. Liu and Y. Wu. Parameter expansion for data augmentation. *Journal of the American Statistical Association*, 94(448):1264–1274, Dec 1999.

X. Liu and M. Daniels. A new algorithm for simulating a correlation matrix based on parameter expansion and reparameterization. *Journal of Computational and Graphical Statistics*, 15:897–914(18), December 2006.

R. McCulloch and P. Rossi. An exact likelihood analysis of the multinomial probit model. *Journal of Econometrics*, 64(1-2):207–240, 1994.

R. McCulloch, N. Polson, and P. Rossi. A bayesian analysis of the multinomial probit model with fully identified parameters. *Journal of Econometrics*, 99(1):173–193, November 2000.

D. McFadden. *Conditional Logit Analysis of Qualitative Choice Behavior*. New York: Academic Press, 1974.

D. McFadden. A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica*, 57(5):995–1026, sep 1989.

R. Natarajan, C. McCulloch, and N. Kiefer. A monte carlo em method for estimating multinomial probit models. *Computational Statistics and Data Analysis*, 34:33–50, 2000.

A. Nobile. Comment: Bayesian multinomial probit models with a normalization constraint. *Journal of Econometrics*, 99(2):335–345, December 2000.

C. Ritter and M. Tanner. Facilitating the gibbs sampler: The gibbs stopper and the griddy-gibbs sampler. *Journal of the American Statistical Association*, 87(419):861–868, sep 1992.

C. Robert. Simulation of truncated normal variables. *Statistics and Computing*, 5(2):121–125, 1995.

P. Rossi, Allenby G., and R. McCulloch. *Bayesian Statistics and Marketing*. John Wiley and Sons, 2005.

A. Roverato. Hyper inverse wishart distribution for non-decomposable graphs and its application to bayesian inference for gaussian graphical models. *Scandinavian Journal of Statistics*, 29(3):391–411, 2002.

M. Tanner and W. Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82 (398):528–540, jun 1987.

E Webb and J. Forster. Bayesian model determination for multivariate ordinal and binary data. *Work in progress*, May 2006. URL `http://www.soton.ac.uk/~jjf/Papers/Ordinal.pdf`.

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

L. Yang, R. Jin, and R. Sukthankar. Bayesian active distance metric learning. *UAI*, July 2007. URL `http://www.cse.msu.edu/~yangliu1/uai2007_bayesian.pdf`.

R. Yang and J. Berger. Estimation of a covariance matrix using the reference prior. *The Annals of Statistics*, 22(3):1195–1211, Sep 1994.

X. Zhang, W.J. Boscardin, and T. Belin. Sampling correlation matrices in bayesian models with correlated latent variables. *Journal of Computational and Graphical Statistics*, 15(4):880–896, December 2006.

Part II

Appendices

Appendix A

Distributions and Identities

A.1 The Multivariate Normal (Gaussian) Distribution

$$f_X(x_1, \ldots, x_n|\mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(\frac{1}{2}(x-\mu)^T\Sigma(x-\mu)\right) \quad \text{(A.1)}$$

where μ is the mean, and Σ is the variance-covariance matrix.

A.2 The Gamma Distribution

$$f_X(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x) \quad \text{(A.2)}$$

where α is the shape parameter and β is the rate parameter.

A.3 The Standard Inverse Wishart Distribution

There are several parametrization of the inverse Wishart distribution. We will list below the ones that we use in this work.

1. The parametrization used in Gupta and Nagar (2000) Let $\Sigma \sim IW(m, I_T)$, then

$$f_T(\Sigma|\nu) \propto |\Sigma|^{-\frac{1}{2}(m)} \exp(-\frac{1}{2} tr(\Sigma^{-1})) \quad \text{(A.3)}$$

With Expectation:

$$E(\Sigma) = \frac{I_T}{m - 2T - 2}$$

The Matlab function `invwishirnd` from the MCMC tool box (*http://www.mathworks.com/matlabcentral/fileexchange/* by David Shera),

implements sampling from this distribution.

2. The parametrization used in Barnard et al. (2000) Let $\Sigma \sim IW(\nu, I_T)$, then
$$f_T(\Sigma|\nu) \propto |\Sigma|^{-\frac{1}{2}(\nu+T+1)} \exp(-\frac{1}{2} tr(\Sigma^{-1})) \tag{A.4}$$
With Expectation:
$$E(\Sigma) = \frac{I_T}{\nu - T - 1}$$
This corresponds to A.3, with $m = \nu + T + 1$. This parametrization is implemented in the matlab function `iwishrnd` in the STAT toolbox.

3. The parametrization used in Dawid and Lauritzen (1993) Let $\Sigma \sim IW(\delta, I_T)$, then
$$f_T(\Sigma|\delta) \propto |\Sigma|^{-\frac{1}{2}(\delta+2T)} \exp(-\frac{1}{2} tr(\Sigma^{-1})) \tag{A.5}$$
With Expectation:
$$E(\Sigma) = \frac{I_T}{\delta - 2}$$
This corresponds to A.3, with $\delta = \nu - T + 1$.

In the one dimensional case all three parametrization reduce to an inverse Chi Square:
$$f(\sigma^2|v) \propto (\sigma)^{-\frac{1}{2}(v+2)} \exp(-\frac{1}{2(\sigma^2)})) \tag{A.6}$$

In this case we could see that ν and δ is parametrization 2 and 3 are equivalent and they are equal to v, and in parametrization 1, $m = v + 2$.

Appendix B

Marginal Prior on R proof from Barnard et al. (2000)

Under the transformation $\Sigma = DRD$, the Jacobian is given by

$$|J : \Sigma \to (D, R)| = \frac{\partial \sigma_{ij}}{\partial r_{ij}} \tag{B.1}$$

Where

$$\begin{cases} \sigma_{ij} = d_i d_j r_{ij} & \text{if } i \neq j; \\ \sigma_{ii} = d_i^2 & \text{if } i = j. \end{cases} \quad j = 1, \ldots, T, i = 1, \ldots, T, \tag{B.2}$$

Therefore,

$$\begin{aligned} |J| &= \begin{vmatrix} 2d_1 & & & & & & & & \\ & 2d_2 & & & & & & & \\ & & \ddots & & & & & & \\ & & & 2d_T & & & & & \\ & & & & d_1 d_2 & & & & \\ & & & & & d_1 d_3 & & & \\ & & & & & & \ddots & & \\ & & & & & & & d_{T-1} d_T \end{vmatrix} \\ &= \prod_i 2(d_i)^T = 2^T (\prod_i d_i)^T \end{aligned} \tag{B.3}$$

As an illustration consider the following example with Σ a 3×3 covariance matrix. Using the transformation $\Sigma = DRD$, we can write

$$\Sigma = \begin{bmatrix} d_1^2 & d_1 d_2 & d_1 d_3 \\ d_1 d_2 & d_2^2 & d_2 d_3 \\ d_1 d_3 & d_2 d_3 & d_3^2 \end{bmatrix} \tag{B.4}$$

The jacobian is:

$$|J| = \left| \frac{\partial(\sigma_{11}, \sigma_{22}, \sigma_{33}, \sigma_{12}, \sigma_{13}, \sigma_{23})}{\partial(d_1, d_2, d_3, r_{12}, r_{13}, r_{23})} \right| = \begin{vmatrix} 2d_1 & 0 & 0 & d_2 r_{12} & d_3 r_{13} & 0 \\ 0 & 2d_2 & 0 & d_1 r_{12} & 0 & d_3 r_{23} \\ 0 & 0 & 2d_3 & 0 & d_1 r_{13} & d_2 r_{23} \\ 0 & 0 & 0 & d_1 d_2 & 0 & 0 \\ 0 & 0 & 0 & 0 & d_1 d_3 & 0 \\ 0 & 0 & 0 & 0 & 0 & d_2 d_3 \end{vmatrix}.$$

Here we could see that the lower triangular part of the Jacobian matrix is 0 and therefore taking the determinant is equivalent to multiplying the diagonal elements, which gives us $|J| = 2^3 (d_1 d_2 d_3)^3$.

In Barnard et al. (2000), they start with $\Sigma \sim IW(\nu, I_T)$, where the inverse Wishart is defined as in A.3.

$$\begin{aligned} \pi(\Sigma|\nu) &\propto |\Sigma|^{-\frac{1}{2}(\nu+T+1)} \exp(-\frac{1}{2} tr(\Sigma^{-1})) \\ \pi(R, D|\nu) &\propto |DRD|^{-\frac{1}{2}(\nu+T+1)} \exp(-\frac{1}{2} tr(DRD)^{-1}) \times |J| \\ &\propto |R|^{-\frac{1}{2}(\nu+T+1)} (\prod_i d_i)^{-(\nu+T+1)} (\prod_i d_i)^T \exp\left(-\frac{1}{2} tr(DRD)^{-1}\right) \\ &\propto |R|^{\frac{1}{2}(\nu+T+1)} (\prod_i d_i)^{-(\nu+1)} \exp\left(-\sum_i \frac{r^{ii}}{2d_i^2}\right) \\ &\propto |R|^{\frac{1}{2}(\nu+T+1)} \prod_i \left(d_i^{-(\nu+1)} \exp\left(-\frac{r^{ii}}{2d_i^2}\right)\right) \end{aligned} \tag{B.5}$$

where r^{ii} is the i^{th} diagonal element of R^{-1}. The distribution of R is obtained by marginalizing over D:

$$f(R|\nu) = \int_0^\infty \pi(R, D|\nu) dD \tag{B.6}$$

$$\propto \int_0^\infty |R|^{\frac{1}{2}(\nu+T+1)} \prod_i \left(d_i^{-(\nu+1)} \exp\left(-\frac{r^{ii}}{2d_i^2} \right) \right) dD \quad \text{(B.7)}$$

We could perform another change of variable from $d_i \rightarrow \alpha_i$ by letting $\alpha_i = \frac{r^{ii}}{2d_i^2}$ with $\partial\alpha_i = -\frac{r^{ii}}{d_i^3}\partial d_i$ we can work through the algebra:

$$\begin{aligned}
f(R|\nu) &\propto |R|^{\frac{1}{2}(\nu+T+1)} \prod_i \int_0^\infty (d_i)^{-(\nu+1)} \exp(-\alpha_i) \frac{d_i^3}{r^{ii}} d\alpha_i \\
&\propto |R|^{\frac{1}{2}(\nu+T+1)} \prod_i \int_0^\infty \left(\frac{d_i^2}{r^{ii}} \right)^{(-\nu+2)/2} \exp(-\alpha_i) \frac{(r^{ii})}{r^{ii}}^{(-\nu+2)/2} d\alpha_i \\
&\propto |R|^{\frac{1}{2}(\nu+T+1)} (\prod_i r_{ii})^{-\frac{\nu}{2}} \prod_i \int_0^\infty (\alpha_i)^{(\nu-2)/2} \exp(-\alpha_i) d\alpha_i \\
&\propto |R|^{\frac{1}{2}(\nu+T+1)} (\prod_i r_{ii})^{-\frac{\nu}{2}} \prod_i \int_0^\infty \underbrace{\alpha_i^{(\nu-2)/2} \exp(-\alpha_i)}_{\Gamma(\frac{\nu}{2},1)} d\alpha_i \quad \text{(B.8)}
\end{aligned}$$

From the above we could see that

$$\pi(R, D) = \pi(R, \alpha) = \pi(\alpha|R)\pi(R) \quad \text{(B.9)}$$

Where

$$\begin{aligned}
\pi(\alpha_i|R) &\sim Gamma(\frac{T+1}{2}, 1) \quad \text{(B.10)} \\
\pi(R) &\propto |R|^{\frac{T(T-1)}{2}-1} (\prod_i |R_{ii}|))^{-(T+1)/2} \quad \text{(B.11)}
\end{aligned}$$

The distribution of R in B.11 is obtained by using the matrix algebra identity:

$$r^{ii} = \frac{|R_{ii}|}{|R|}$$

where R_{ii} is the principal submatrix of R.

The marginal distribution of each r_{ij} is obtained using the marginalization property of the inverse Wishart, which states that each principal submatrix of an inverse Wishart is also an inverse Wishart. This means

that this derivation could be obtained for any $T_1 \times T_1$ sub-covariance matrix. Choosing a submatrix Σ_1 of Σ, $\Sigma_1 \sim IW(\nu-(T-T_1), I)$. The density of the correlation submatrix is as in B.11, with $T = T_1$ and $\nu = \nu-(T-T_1)$. In the case where $T_1 = 2$, the marginal density is:

$$f_2(r_{ij}|\nu) = (1 - r_{ij})^{\frac{(\nu-T-1)}{2}} \tag{B.12}$$

In this case, B.12 could be viewed as $Beta\left(\frac{\nu-T+1}{2}, \frac{\nu-T+1}{2}\right)$ on $[-1, 1]$ and is uniform when $\nu = T + 1$.

Appendix C

Computation of the Jacobian $|J : Z \to W|$

This is an example of the computation of the Jacobian of the transformation $Z = D^{-1}W$ for $T = 3$, $p = 2$, and $i = 1, \ldots, n$:

$$Z_{i1} = D_{11}^{-1} W_{i1} \tag{C.1}$$

$$Z_{i2} = D_{22}^{-1} W_{i2} \tag{C.2}$$

$$Z_{i3} = D_{33}^{-1} W_{i3} \tag{C.3}$$

The Jacobian is then

$$|J| = \left| \frac{\partial(Z_{11}, Z_{12}, Z_{13}, \ldots, Z_{n1}, Z_{n2}, Z_{n3})}{\partial(W_{11}, W_{12}, W_{13}, \ldots, W_{n1}, W_{n2}, W_{n3})} \right| \tag{C.4}$$

$$\begin{aligned} |J| &= \begin{vmatrix} D_{11}^{-1} & 0 & 0 & & & & & \\ 0 & D_{22}^{-1} & 0 & & & & & \\ 0 & 0 & D_{33}^{-1} & & & & & \\ & & & \ddots & & & & \\ & & & & D_{11}^{-1} & 0 & 0 \\ & & & & 0 & D_{22}^{-1} & 0 \\ & & & & 0 & 0 & D_{33}^{-1} \end{vmatrix} \\ &= |I_n \otimes D^{-1}| = |D|^{-n} \end{aligned} \tag{C.5}$$

Appendix D

Sampling from Multivariate truncated Gaussian

In the context of Multivariate Probit, we are interested in generating random samples from a multivariate Gaussian subject to multiple linear inequality constraints. We follow the method outlined in Geweke (1991) using Gibbs sampling.

Let

$$x \sim N(\mu, \Sigma) \qquad a \leq x \leq b$$

Where Σ is an $p \times p$ covariance matrix of rank p, μ is a $p \times 1$ vector of means and a and b are vectors of lower and upper bound that can take on the values $-\infty$ and $+\infty$ respectively. This problem is equivalent to sampling from a p-variate normal distribution subject to linear constraints:

$$z \sim N(0, \Sigma) \qquad \alpha \leq z \leq \beta$$

where $\alpha = a - \mu$, and $\beta = b - \mu$.

We can then take $x = \mu + z$. To sample the z_i's, we adopt a Gibbs sampling approach that uses the property that each element of z, conditional on all of the other elements of z is a univariate truncated normal. From conditional multivariate Normal distribution theory, we have the following result:

If $z \sim N(0, T)$, the non truncated distribution

$$E(z_i | z_{-i}) = \sum_{i \neq j} c_{ij} z_j \tag{D.1}$$

where c_{ij} is defined in **??** and D.4 Then the truncated distribution has the following construction:

$$z_i = \sum_{i \neq j} c_{ij} z_j + h_i \epsilon_i, \quad \epsilon_i \sim TN\Big((\alpha_i - \sum_{i \neq j} c_{ij} z_j)/h_i, (\beta_i - \sum_{i \neq j} c_{ij} z_j)/h_i\Big) \tag{D.2}$$

Where TN is the univariate truncated normal distribution, and the vector of coefficients in the conditional mean is

$$c^i = (c_{i1}, \ldots, c_{ii-1}, c_{i+1}, \ldots c_{ip})' c1) \tag{D.3}$$

where $i = 1, \ldots, p.$ and

$$c^i = -(\Sigma^{ii})^{-1} \Sigma^{i,<i} \quad \text{and} \quad h_i^2 = (\Sigma^{ii})^{-1} \tag{D.4}$$

Where Σ^{ii} is the element in row i and column i of Σ^{-1} and $\Sigma^{i,<i}$ is row i of Σ^{-1} with Σ^{ii} deleted.

As mentioned in Geweke (1991), we only need to perform these calculation once in the beginning. We can then cycle through the Gibbs steps as follows:

- Initialize $z^{(0)} = 0$
- In the first pass, generate p successive variables from :

  $$z_i^{(1)} | (z_1^{(1)}, \ldots, z_{i-1}^{(1)}, z_{i+1}^{(0)}, \ldots, z_p^{(0)}) \sim f_i(z_1^{(1)}, \ldots, z_{i-1}^{(1)}, z_{i+1}^{(0)}, \ldots, z_p^{(0)}) \tag{D.5}$$

  where $i = 1, \ldots, p$
- Repeat the above such that at the j'th pass:

$$z_i^{(j)} | (z_1^{(j)}, \ldots, z_{i-1}^{(j)}, z_{i+1}^{(j-1)}, \ldots, z_p^{(j-1)}) \sim f_i(z_1^{(j)}, \ldots, z_{i-1}^{(j)}, z_{i+1}^{(j-1)}, \ldots, z_p^{(j-1)}) \tag{D.6}$$

- at the end of each pass we compute

$$x^{(j)} = \mu + z^{(j)} \tag{D.7}$$

There are several methods available for generating a univariate truncated Normal distribution, in our implementation, we adopt the methods used in Robert (1995).

Let

$$x \sim N(\mu, \mu^{-}, \sigma^2)$$

where μ is the mean, μ^{-} is the left truncation point and σ^2 is the variance. Robert (1995) uses a accept-reject algorithm that is more efficient than rejection sampling or the inverse *cdf* method which could be very inefficient if $\mu^{-} - \mu$ is large.

Assuming without loss of generality that $\mu = 0$ and $\sigma^2 = 1$ the algorithm proceeds as follows:

1. Generate $z \sim Exp(\alpha^*, \mu^{-})$
2. Compute $\rho(z) = \exp(-(z - \alpha^*)^2/2)$
3. Generate $u \sim U[0, 1]$ and take $x = z$ if $u \leq \rho(z)$, otherwise go back to the first step.

Where $Exp(\alpha^*, \mu^{-})$ is the translated Exponential distribution and the optimal value of $\alpha^* = \frac{\mu^{-} + \sqrt{((\mu^{-})^2 + 4)}}{2}$.

Appendix E

Sampling from the Hyper Inverse Wishart Distribution (Carvalho et al., 2007)

Let $G = (V, E)$, be a decomposable, undirected graph with $|V| = T$. If we assume that G is a Gaussian Graphical Model with a sparse structure, then

$$\Sigma \sim HIW_G(b, D)$$

Where the Hyper inverse Wishart is defined as in A.5.

A graph can be represented by a *perfect ordering* of its prime components and separators. An ordering of components $P_i \in P$ and separators $S_i \in S$, $(P_1, S_2, P_2, S_3, \ldots, P_T)$, is said to be perfect if for every $i = 2, 3, \ldots, T$ there exists a $j < i$ such that

$$S_i = P_i \; H_{i-1} \subset P_j$$

where

$$H_{i-1} = \cup_{j=1}^{i-1} P_j$$

In order to obtain the perfect ordering of prime components $P_1, \ldots, P_k$, we need to generate the junction tree of the graph (Cowell et al., 1999).

For a decomposable undirected graph G, a *junction tree* is a representation of its prime components. A *junction tree* is a subgraph of the decomposable graph that has the following characteristics: (1) it is a tree, (2) it contains all the nodes of the graph, and (3) it satisfies the *junction tree*

property: For each pair C_i, C_j of cliques with intersection S, all cliques on the path between C_i and C_j contain S.

Subsequently the joint density of Σ factorizes as:

$$p(\Sigma|b, D) = p(\Sigma_{P_1}) \prod_{i=2}^{k} p(\Sigma_{P_i}|\Sigma_{S_i}) \tag{E.1}$$

For decomposable models, all prime components are complete. Therefore standard results for the inverse Wishart (Gupta and Nagar, 2000) enables sampling from each of the distributions in the composition directly.

Where $\Sigma_{S_i,R_i} = \Sigma_{R_i,S_i}^T$. Define

$$\Sigma_{R_i.S_i} = \Sigma_{R_i} - \Sigma_{R_i,S_i}\Sigma_{S_i}^{-1}\Sigma_{S_i,R_i} \tag{E.2}$$

$$D_{R_i.S_i} = D_{R_i} - D_{R_i,S_i}D_{S_i}^{-1}D_{S_i,R_i} \tag{E.3}$$

The sampling scheme would proceed as follows:

1. sample $\Sigma_{C_1} \sim IW(b, D_{C_1})$, this gives the values of submatrix Σ_{S_2}.
2. For $i = 1, \ldots, k$, sample

$$\Sigma_{R_i.S_i} \sim IW(b + |R_i|, D_{R_i.S_i}) \tag{E.4}$$

$$U_i \sim N(D_{R_i,S_i}D_{S_i}^{-1}, \Sigma_{R_i.S_i} \otimes D_{S_i}^{-1}) \tag{E.5}$$

   We could then directly compute $\Sigma_{R_i,S_i} = U_i\Sigma_{S_i}$ and $\Sigma_{R_i} = \Sigma_{R_i.S_i} + \Sigma_{R_i,S_i}\Sigma_{S_i}^{-1}\Sigma_{S_i,R_i}$

The non-free parameters of Σ are obtained through matrix completion given the perfect ordering of its prime components (for details see Lauritzen (1996)) using:

$$\Sigma_{R_i,A_{i-1}} = \Sigma_{R_i,S_i}\Sigma_{S_i}^{-1}\Sigma_{S_i,A_{i-1}} \tag{E.6}$$

Appendix F

Simulation Results

Entropy		Quadratic	
Saturated	Structured	Saturated	Structured
22.228	2.534	154.075	3.827
28.097	3.130	237.388	4.862
36.069	3.095	424.618	4.381
23.836	3.111	189.696	4.585
27.421	3.558	226.054	4.563
21.959	3.041	149.302	4.225
39.708	4.216	1569.468	6.129
27.325	2.066	242.982	3.530
2.302	0.401	10.406	1.063
5.227	0.362	48.853	0.941
2.769	0.941	12.265	8.453
2.375	0.513	16.248	0.995
2.867	0.503	17.114	1.250
4.383	0.638	42.125	4.627
3.162	0.533	20.535	1.354
5.947	0.739	644.714	3.338
2.413	0.558	10.570	1.427
2.616	0.559	12.475	1.100
3.793	0.444	36.864	0.970
22.310	2.087	157.641	3.469
3.518	0.637	21.579	1.890
2.114	0.455	9.493	1.079
2.103	0.371	8.683	0.870

Table F.1: *Simulation results: Entropy and quadratic loss for 50 data sets generated by different correlation matrices with the same structure*

Entropy		Quadratic	
Saturated	Structured	Saturated	Structured
2.026	0.552	7.596	1.031
5.835	1.123	556.727	18.055
1.746	0.392	6.715	0.994
1.953	0.343	7.985	0.788
2.337	0.384	10.725	1.011
2.428	0.415	11.166	0.830
4.546	0.842	34.827	2.885
24.410	2.088	186.270	3.526
19.670	1.635	127.148	2.830
22.453	2.586	156.924	4.198
53.488	6.072	1066.345	5.868
29.999	2.337	278.968	3.848
26.146	2.313	211.350	3.441
1.860	0.402	6.632	0.756
2.209	0.343	10.706	0.819
2.749	0.532	14.805	0.990
1.701	0.501	5.315	0.866
1.953	0.408	8.378	0.820
2.399	0.413	10.604	0.969
5.596	1.018	67.955	3.281
1.772	0.311	7.622	0.728
2.046	0.771	6.495	1.212
2.900	0.518	16.389	1.720
164.743	0.617	847548.010	1.271
6.030	0.501	283.085	1.661
3.300	0.553	17.834	1.489
2.937	0.890	17.898	3.991

Table F.2: *Table F continued*