Title	Bayesian variable selection for the seemingly unrelated regression model with a large number of predictors
Sub Title	
Author	安道, 知寛(Ando, Tomohiro)
Publisher	慶應義塾経営管理学会
Publication year	2010
Jtitle	慶應義塾経営管理学会リサーチペーパー・シリーズ No.103 (2010. 8)
JaLC DOI	
Abstract	Computationally efficient methods for Bayesian analysis of Seemingly unrelated regression (SCR) models are deYeloped. Cnder a Bayesian hierarchical framework where ea.ch regression function is represented as a linear combination of a large number of basis functions, the regression coefficients, the variance matrix of the errors, and a set of variables to be included in the model are estimated simultaneously. Usually the Bayesian estimation problem is solved using Markov Chain Monte Carlo (MCMC) techniques. Herein we shew how a direct Monte Carlo (DMC) technique can be employed to solve this estimation problem more efficiently."
Notes	
Genre	Technical Report
URL	https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=KO40003002-00000103-0001

Bayesian variable selection for the seemingly unrelated regression model with a large number of predictors

安道　知寛
Tomohiro Ando

慶應義塾大学大学院経営管理研究科准教授

慶應義塾経営管理学会
リサーチペーパー・シリーズ
№１０３　（2010年８月）

Bayesian variable selection for the seemingly unrelated regression model with a large number of predictors

Tomohiro Ando

Graduate School of Business Administration. Keio University 4-1-1 Hiyoshi. Kohoku-ku. Yokohama-shi. Kanagawa. 223-8526. Japan

Abstract

Computationally efficient methods for Bayesian analysis of Seemingly unrelated regression (SUR) models are developed. Under a Bayesian hierarchical framework where each regression function is represented as a linear combination of a large number of basis functions, the regression coefficients, the variance matrix of the errors, and a set of variables to be included in the model are estimated simultaneously. Usually the Bayesian estimation problem is solved using Markov Chain Monte Carlo (MCMC) techniques. Herein we show how a direct Monte Carlo (DMC) technique can be employed to solve this estimation problem more efficiently."

Key words: Bayesian estimation, Seemingly Unrelated Regression, Direct Monte Carlo, Markov Chain Monte Carlo

1 Introduction

In many areas of research and application, the seemingly unrelated regression (SUR) model, introduced by Zellner (1962), is used as a tool to study a wide range of phenomena. Many studies have contributed to the development of estimation, testing, prediction and other inference techniques for analysis of SUR models including Zellner (1962, 1963), Gallant (1975), Rocke (1989), Neudecker and Windmeijer (1991), Mandy and Martins (1993), Kurata (1999), Liu (2002), Ng (2002), Carroll, et al. (2006). Also, the SUR model and inference techniques for analyzing it are described in almost all Bayesian and

[1] *Corresponding author
Tel.:+81-45-564-2039.
E-mail address: andoh@kbs.keio.ac.jp

non-Bayesian textbooks that provide many references to the literature; see, e.g. Greene (2002), Geweke (2005), Lancaster (2004), Rossi et al. (2005) and other texts. The first analysis of the SUR model appeared in Zellner (1962, 1963) who employed a generalized least squares approach. Later, likelihood and traditional Bayesian approaches were developed followed by various other inference approaches; see ,e.g., the likelihood distributional approach (Frasera et al., 2005), Bayesian analyses, the Bayesian method of moments, van der Merwve and Viljoen (1988) and so on.

In the Bayesian analysis of the SUR model, one can apply the Gibbs algorithm of Percy (1992). However, we often want to estimate the regression coefficients, the variance matrix of the errors, and a set of variables to be included in the model simultaneously. It is obvious that the traditional method of best subset selection is computationally in feasible for high dimensional data. To solve this problem, Smith and Kohn (2000) recently introduced a Bayesian hierarchical SUR model and developed a Markov Chain Monte Carlo (MCMC) procedure to estimate it.

Although their algorithm can be applied to various types of problems, it is still not computationally efficient in some cases. This is because the use of the MCMC algorithm for drawing the regression coefficients and the variance matrix of the errors leads to very highly auto-correlated output draws in some situations (Zellner and Ando (2010c), See also Section 4). It has been shown that a direct Monte Carlo (DMC) is very efficient for drawing the posterior samples of the regression coefficients and the variance matrix of the errors (Ando and Zellner (2010), Zellner and Ando (2010a, 2010b)).

The aim of this paper is to extend their DMC approach to implement the variable selection simultaneously. We show that the developed method is more computationally efficient than Smith and Kohn (2000)'s MCMC method. The difference between our paper and Ando and Zellner (2010) and Zellner and Ando (2010a, 2010b) is that the selection of a set of variables to be included in the model was not considered in their DMC algorithm. Instead, they considered the use of some model selection criteria.

The structure of the remainder of this paper is as follows. In section 2, we briefly review the standard SUR model. Section 3 establishes an efficient Bayesian estimation procedure for the SUR model. We also provide several remarks regarding the proposed method. Numerical studies are conducted in Section 4. Section 5 concludes.

2 Overview of SUR Model

The linear SUR model involves a set of regression equations with cross-equation parameter restrictions and correlated error terms having differing variances. Algebraically, the SUR model is given by:

$$\boldsymbol{y}_j = X_j\boldsymbol{\beta}_j + \boldsymbol{u}_j,\ j = 1, ..., m, \quad \text{with} \quad E[\boldsymbol{u}_i\boldsymbol{u}_j'] = \begin{cases} \omega_{ij}I, & (i \neq j) \\ \omega_i^2 I, & (i = j) \end{cases}, \tag{1}$$

Here $\boldsymbol{y}_j$ and $\boldsymbol{u}_j$ are $n \times 1$ vectors, X_j is a $n \times p_j$ matrix of observations of rank p_j on p_j predetermined variables, and $\boldsymbol{\beta}_j$ is a p_j-dimensional coefficient vector. The domains of parameter values are given as follows: $-\infty < \beta_{jr} < \infty$, $(r = 1, ..., p_j, j = 1, ..., m)$, $-\infty < \omega_{ij} < \infty$, $(i, j = 1, ..., m, i \neq j)$ and $0 < \omega_{jj} < \infty$, $(j = 1, ..., m)$.

As shown in (1), the equations have different independent variables and variances. Also, the model permits error terms in different equations to be correlated. We can easily replace the linear combination of a set of covariates by a linear combination of basis functions. Thus, we are implicitly treating a semi-parametric model given a choice of a particular basis function.

In matrix form, the model can be expressed as $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{u}$, $\boldsymbol{u} \sim N(\boldsymbol{0}, \Omega \otimes I)$, where $N(\boldsymbol{\mu}, \Sigma)$ denotes the normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ, $\otimes$ is the tensor product, Ω is the $m \times m$ matrix with the diagonal elements $\{\omega_1^2, ..., \omega_m^2\}$, and the off-diagonal ijth elements are ω_{ij}, $\boldsymbol{y}' = (\boldsymbol{y}_1', ..., \boldsymbol{y}_m')$, $X = \text{diag}\{X_1, ..., X_m\}$, $\boldsymbol{\beta}' = (\boldsymbol{\beta}_1', ..., \boldsymbol{\beta}_m')$ and $\boldsymbol{u}' = (\boldsymbol{u}_1', ..., \boldsymbol{u}_m')$.

The likelihood function is

$$L(\boldsymbol{y}|\boldsymbol{\beta}, \Omega) = \frac{1}{(2\pi)^{nm/2}|\Omega|^{n/2}} \exp\left[-\frac{1}{2}\text{tr}\left\{R\Omega^{-1}\right\}\right],$$

where "tr" denotes the trace of a matrix, $|\Omega| = \det(\Omega)$ is the value of the determinant of Ω, the ijth element of the $m \times m$ matrix $R = (r_{ij})$ is $r_{ij} = (\boldsymbol{y}_i - X_i\boldsymbol{\beta}_i)'(\boldsymbol{y}_j - X_j\boldsymbol{\beta}_j)$.

Zellner (1971), Press (1972), Box and Tiao (1973), Percy (1992), and Srivastava and Giles (1987) studied the posterior distributions of the parameters of the normal SUR model. In the absence of prior knowledge, Bayesian analysis with noninformative priors is very common in practice. One of the most widely used noninformative priors, introduced by Jeffreys (1946, 1961), is Jeffreys's invariant prior:

$$\pi(\boldsymbol{\beta}, \Omega) = \pi(\boldsymbol{\beta})\pi(\Omega) \propto |\Omega|^{-\frac{m+1}{2}}, \tag{2}$$

which is proportional to the square root of the determinant of Fisher information matrix. One of the advantages of the use of Jeffreys's prior is that it is invariant under any one-to-one reparameterization of the model.

Because only conditional posterior probability density functions of $\boldsymbol{\beta}$ and Ω are available in analytical forms, simulation methods have to be used to produce marginal posterior densities for the parameters and future values of observations. Currently, one of the most widely used Bayesian estimation methods for the SUR model is the MCMC approach that is described and applied in many recent Bayesian econometrics and statistics texts. Because the conditional posterior densities $\pi(\boldsymbol{\beta}|\Omega, \boldsymbol{y})$ and $\pi(\Omega|\boldsymbol{\beta}, \boldsymbol{y})$ are available, the standard SUR model is also amenable to a 2-block Gibbs sampler; see, e.g. Percy (1992). It is known that the conditional posterior densities are

$$\pi(\boldsymbol{\beta}|\Omega, \boldsymbol{y}) = N\left(\hat{\boldsymbol{\beta}}, \hat{\Omega}\right) \quad \text{and} \quad \pi(\Omega|\boldsymbol{\beta}, \boldsymbol{y}) = IW\left(R, n\right), \tag{3}$$

with

$$\hat{\boldsymbol{\beta}} = \left\{X'\left(\Sigma^{-1} \otimes I\right)X\right\}^{-1} X'\left(\Sigma^{-1} \otimes I\right)\boldsymbol{y},$$
$$\hat{\Omega} = \left(X'\left(\Sigma^{-1} \otimes I\right)X\right)^{-1},$$

where $IW(\cdot, \cdot)$ denotes the inverse Wishart distribution,

Recently, Smith and Kohn (2000) introduced a Bayesian hierarchical model to explicitly parameterize the possibility that some coefficients are exactly zero. They developed a MCMC sampling scheme to estimate the SUR model. In the next section, we show how a DMC sampling procedure can be employed to obtain results more efficiently.

3 Methodology

Following the notation of Smith and Kohn (2000), we introduce a vector of binary indicator variables $\boldsymbol{\gamma}_j = (\gamma_1^j, ..., \gamma_p^j)'$ for the design matrix X_j, $j = 1, ..., m$. Here, γ_k^j corresponds to the k-th element of the coefficient vector $\boldsymbol{\beta}_j$, with $\gamma_k^j = 0$ if $\beta_j^k = 0$ and $\gamma_k^j = 1$ if $\beta_j^k \neq 0$ and by dropping the redundant terms with zero coefficients, the j-th equation can be rewritten as

$$\boldsymbol{y}_j = X_{\gamma_j}\boldsymbol{\beta}_{\gamma_j} + \boldsymbol{u}_j, \ j = 1, ..., m. \tag{1}$$

Let $q_j = \sum_{k=1}^{p_j} \gamma_k^j$. Then the design matrix X_{γ_j} is of size $n \times q_j$ and $\boldsymbol{\beta}_{\gamma_j}$ is a q_j-dimensional coefficient vector.

To complete this Bayesian hierarchical model, we use the following priors on the parameters. We use the Jeffreys's invariant prior for $\boldsymbol{\beta} = (\boldsymbol{\beta}'_{\gamma_1}, ..., \boldsymbol{\beta}'_{\gamma_m})'$ and Ω (See also Section 3.3, the use of informative prior is discussed). For the indicator variables, γ_k^j are taken a priori independently distributed, with probability that it takes 1 is $\pi(\gamma_k^j = 1|\alpha_j) = \alpha_j$. Also, the hyperparameters α_j, are taken as independent and given a non-informative uniform prior on $(0, 1)$. After we integrate the hyperparameters $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_m)'$ out, we have $\pi(\boldsymbol{\gamma}) = \int \pi(\boldsymbol{\gamma}|\boldsymbol{\alpha})\pi(\boldsymbol{\alpha})d\boldsymbol{\alpha} = \prod_{j=1}^{m} Be(q_j + 1, p_j - q_j + 1)$, where Be is the beta function. This prior is presented and used in Smith and Kohn (2000).

We generate the posterior draws by using the following MCMC sampling scheme:

(1) Using DMC algorithm, generate from $\boldsymbol{\beta}, \Omega|\boldsymbol{\gamma}, \boldsymbol{y}$.

(2) Generate from $\gamma_k^j|\Omega, \boldsymbol{\gamma}/\gamma_k^j, \boldsymbol{y}$ using the MCMC sampling step described in Smith and Kohn (2000).

Given value of $\boldsymbol{\gamma}_1$,...,$\boldsymbol{\gamma}_m$, our DMC approach produces independent draws and we don't have any problem with determining the "acceptance rate". On the other hand, the acceptance rates of MCMC by Smith and Kohn (2000) range between 60% and 90% (Smith and Kohn (2000)). Thus, our algorithm is much more efficient from this perspective. Also, we have checked the autocrrelation of the draws of the elements of Ω, and we found that it was much larger for the method of Smith and Kohn than for our approach.

3.1 Posterior sampling of $\boldsymbol{\beta}, \Omega|\boldsymbol{\gamma}, \boldsymbol{y}$

Recently, Zellner and Ando (2008) derived a direct Monte Carlo procedure for the Bayesian analysis of the SUR model. In their framework, the standard SUR model (4) is reformulated as follows:

$$\begin{cases} \boldsymbol{y}_1 = X_{\gamma_1}\boldsymbol{\beta}_{\gamma_1} + \boldsymbol{e}_1 \equiv Z_1\boldsymbol{b}_1 + \boldsymbol{e}_1, \\ \boldsymbol{y}_j = X_j\boldsymbol{\beta}_{\gamma_j} + \sum_{l=1}^{j-1} \rho_{jl}(\boldsymbol{y}_l - X_{\gamma_l}\boldsymbol{\beta}_{\gamma_l}) + \boldsymbol{e}_j \equiv Z_j\boldsymbol{b}_j + \boldsymbol{e}_j, \quad j = 2, ..., m, \end{cases} \tag{5}$$

where the $n \times (q_j + j - 1)$ matrices Z_j are functions of $\boldsymbol{\beta}_{\gamma_{j-1}}, ..., \boldsymbol{\beta}_{\gamma_1}$, and

$$E[\boldsymbol{e}_i\boldsymbol{e}_j'] = \begin{cases} O, & (i \neq j) \\ \sigma_i^2 I, & (i = j) \end{cases}, \quad \text{and} \quad \Sigma = \text{diag}\{\sigma_1^2, ..., \sigma_m^2\}.$$

Zellner et al (1988) and Zellner and Chen (2002) considered this transformation in the context of simultaneous equation modeling.

Zellner and Ando (2008) pointed to the capability of transforming from the parameters of the transformed model in (5) back to the parameters of the original formulation in equation (4). There is a one to one relation between the parameters of SUR model (4) and those of the transformed model (5). The likelihood function of the parameters $\boldsymbol{\theta}' = (\boldsymbol{b}_1', ..., \boldsymbol{b}_m', \sigma_1^2, ..., \sigma_m^2)'$ is

$$L(\boldsymbol{y}|\boldsymbol{b}, \Sigma, \boldsymbol{\gamma}) = \prod_{j=1}^{m} \frac{1}{(2\pi\sigma_j^2)^{n/2}} \exp\left[-\frac{(\boldsymbol{y}_j - Z_j\boldsymbol{b}_j)'(\boldsymbol{y}_j - Z_j\boldsymbol{b}_j)}{2\sigma_j^2}\right].$$

In contrast to the standard model (1), we can decompose the likelihood function thanks to $E[\boldsymbol{e}_i\boldsymbol{e}_j'] = O, (i \neq j)$. The prior density function specified in (2) expressed in terms of $\{\boldsymbol{b}, \Sigma\}$ is

$$\pi(\boldsymbol{b}, \Sigma) \propto |\Omega(\boldsymbol{b}, \Sigma)|^{-\frac{m+1}{2}}|J| = \prod_{j=1}^{m}(\sigma_j^2)^{-\frac{m+1}{2}} \times \prod_{j'=1}^{m-1}(\sigma_{j'}^2)^{m-j'} = \prod_{j=1}^{m}(\sigma_j^2)^{\frac{m-2j-1}{2}} \quad (6)$$

where $|J|$ is a Jacobian factor. The joint posterior density of parameters is then

$$\pi(\boldsymbol{b}_1, .., \boldsymbol{b}_m, \sigma_1^2, .., \sigma_m^2|\boldsymbol{\gamma}, \boldsymbol{y}) \propto \prod_{j=1}^{m}(\sigma_j)^{-(n-m+2j+1)} \exp\left[-\frac{(\boldsymbol{y}_j - Z_j\boldsymbol{b}_j)'(\boldsymbol{y}_j - Z_j\boldsymbol{b}_j)}{2\sigma_j^2}\right],$$

which is equivalent to the conditional normal inverse-gamma posterior

$$\begin{aligned} \pi(\boldsymbol{b}_j|\boldsymbol{b}_{j-1}, ..., \boldsymbol{b}_1, \sigma_j^2, \boldsymbol{\gamma}, \boldsymbol{y}) &= N\left(\hat{\boldsymbol{b}}_j, \sigma_j^2(Z_j'Z_j)^{-1}\right), \\ \pi(\sigma_j^2|\boldsymbol{b}_{j-1}, ..., \boldsymbol{b}_1, \boldsymbol{\gamma}, \boldsymbol{y}) &= IG\left(\hat{\gamma}_j/2, \hat{\nu}_j/2\right), \end{aligned}$$

where for $j = 1, ..., m$, $IG(\cdot, \cdot)$ denotes the inverse Gamma distribution, and

$$\begin{aligned} \hat{\boldsymbol{b}}_j &= \left(Z_j'Z_j\right)^{-1} Z_j'\boldsymbol{y}_j, \\ \hat{\gamma}_j &= \left(\boldsymbol{y}_j - Z_j\hat{\boldsymbol{b}}_j\right)'\left(\boldsymbol{y}_j - Z_j\hat{\boldsymbol{b}}_j\right), \\ \hat{\nu}_j &= n - m - p_j + j + 1. \end{aligned}$$

Then we can use the following direct Monte Carlo sampling procedure for the posterior sampling of $\boldsymbol{\beta}, \Omega|\boldsymbol{\gamma}, \boldsymbol{y}$.

A direct Monte Carlo sampling procedure:

Step 1 (initialization). Fix the order of a set of m equations. Set $j = 1$. Generate σ_1^2 and insert the drawn value in $\pi(\boldsymbol{b}_1|\sigma_1^2, \boldsymbol{\gamma}, \boldsymbol{y})$. Then make a draw $\boldsymbol{b}_1$ from $\pi(\boldsymbol{b}_1|\sigma_1^2, \boldsymbol{\gamma}, \boldsymbol{y})$.

Step 2 Increase the iteration index j by one $j \to (j + 1)$. Draw σ_j from the conditional inverse gamma density $\pi(\sigma_j^2|\boldsymbol{b}_{j-1}, ..., \boldsymbol{b}_1, \boldsymbol{\gamma}, \boldsymbol{y})$, and then generate $\boldsymbol{b}_j$ from $\pi(\boldsymbol{b}_j|\boldsymbol{b}_{j-1}, ..., \boldsymbol{b}_1, \sigma_j, \boldsymbol{\gamma}, \boldsymbol{y})$.

Step 3 Repeat Step 2 sequentially until $j = m$.

Step 4 Transform the generated draws $\boldsymbol{b}$ and Σ into Ω.

Step 5 Generate draws $\boldsymbol{\beta}|\Omega, \boldsymbol{\gamma}, \boldsymbol{y}$ from the conditional posterior (3).

There is the following recursive relations between $\{\boldsymbol{b}, \Sigma\}$ and Ω:

$$\omega_j^2 = \sum_{k=1}^{j-1} \rho_{jk}^2 \omega_k^2 + \sum_{k,l=1,\ k<l}^{j-1} \rho_{jk}\rho_{jl}\omega_{lk} + \sigma_j^2, \quad (j \neq 1), \tag{7}$$

$$\omega_{ji} = \sum_{k=1, k\neq i}^{j-1} \rho_{jk}\omega_{ki} + \rho_{ji}\omega_i^2, \quad (j \neq 1),$$

Thus, we can transform the generated samples $\{\boldsymbol{b}, \Sigma\}$ into Ω.

3.2 Posterior sampling of $\gamma_k^j|\Omega, \boldsymbol{\gamma}/\gamma_k^j, \boldsymbol{y}$

To speed up the generation, we use Smith and Kohn's (2000) sampling step, which is an application of the Metropolis-Hasting procedure. Let $\pi(\gamma_k^j|\Omega, \boldsymbol{\gamma}/\gamma_k^j, \boldsymbol{y})$ be the conditional posterior density of γ_k^j and $s(\gamma_k^j) = \pi(\gamma_k^j|\boldsymbol{\gamma}/\gamma_k^j)$ be its conditional prior density. Note that $\boldsymbol{\alpha}$ is integrated out in both cases.

The density $\pi(\gamma_k^j|\Omega, \boldsymbol{\gamma}/\gamma_k^j, \boldsymbol{y})$ requires calculation to enable generation of γ_k^j in the posterior sampling process. We have

$$\begin{aligned}\pi(\gamma_k^j|\Omega, \boldsymbol{\gamma}/\gamma_k^j, \boldsymbol{y}) &\propto \int L(\boldsymbol{y}|\boldsymbol{\beta}, \Omega, \boldsymbol{\gamma}) p(\boldsymbol{\beta}|\boldsymbol{\gamma}, \Omega) d\boldsymbol{\beta} \pi(\boldsymbol{\gamma}) \\ &\propto (n)^{-q_\gamma/2} \exp\left\{-\frac{1}{2} S(\boldsymbol{\gamma}, \Omega)\right\} \pi(\gamma_k^j|\boldsymbol{\gamma}/\gamma_k^j),\end{aligned}$$

where $A = \Omega^{-1} \otimes I$ and $S(\boldsymbol{\gamma}, \Omega) = \boldsymbol{y}'A\boldsymbol{y} - \boldsymbol{y}'AX_\gamma(X'AX_\gamma)^{-1}X_\gamma'A\boldsymbol{y}$ (See Smith and Kohn (2000)). Also, the conditional prior $\pi(\gamma_k^j|\boldsymbol{\gamma}/\gamma_k^j)$ can be calculated as

$$\pi(\gamma_k^j|\gamma/\gamma_k^j) \propto \int_0^1 \alpha_j^{q_\gamma^j}(1-\alpha_j)^{p_j-q_\gamma^j} d\alpha_j = Be(q_\gamma^j+1, p_j-q_\gamma^j+1),$$

and thus

$$\pi(\gamma_k^j = 1|\gamma/\gamma_k^j) = \frac{1}{1+(p_j-a_j)/(a_j+1)}$$

with $a_j = \sum_{k\neq j}\gamma_k^j$ is the number of elements of γ_j/γ_k^j that are one.

Let γ^{old} be the previous value of γ_k^j, a new value γ^{new} can then be generated. If $\gamma^{old} = 0$, then generate γ^{new} from the proposal density

$$Q(\gamma^{old} = 1 \to \gamma^{new} = 0) = s(\gamma_k^j = 0)\min\left(1, \frac{\pi(\gamma_k^j = 0)}{s(\gamma_k^j = 0)}\right)$$

If $\gamma^{new} = 1$, then accept γ^{new} with probability $\alpha = \min\{1, s(\gamma_k^j = 0)/\pi(\gamma_k^j = 0)\}$, otherwise set $\gamma^{new} = 0$.

If $\gamma^{old} = 1$, then generate γ^{new} from the proposal density

$$Q(\gamma^{old} = 0 \to \gamma^{new} = 1) = s(\gamma_k^j = 1)\min\left(1, \frac{\pi(\gamma_k^j = 1)}{s(\gamma_k^j = 1)}\right)$$

If $\gamma^{new} = 0$, then accept γ^{new} with probability $\alpha = \min\{1, s(\gamma_k^j = 1)/\pi(\gamma_k^j = 1)\}$, otherwise set $\gamma^{new} = 0$.

Smith and Kohn (2000) pointed out that the sampling method for γ_k^j is a direct application of the Metropolis-Hastings method.

3.3 Remark

An important advantage of our approach compared with the MCMC approach of Smith and Kohn (2000) is that because it decomposes the joint conditional density of the coefficient vector into a set of low-dimensional conditional densities, we can avoid large scale matrix calculations. It is well known that the computation of the inverse of a large scale matrix is a very computational intensive task. For example, with a 1G memory PC, R version 1.7 does not allow us to sample $\boldsymbol{\beta}$ from the Smith and Kohn (2000)'s algorithm to analyze the SUR model with the number of equations $m = 100$, $\boldsymbol{y}_j$ and $\boldsymbol{u}_j$ are 500×1 vectors ($n = 500$), X_j are 500×40 matrix ($p_j = 40$) and $\boldsymbol{\beta}_j$ is a 40-dimensional

vector. This is because the PC system does not accept a 40000×4000 dimensional design matrix X, although the Gibbs sampling approach of Smith and Kohn requires it. Although there are ways of simplifying the matrix inversion problem by using partitioning of the matrix and using formulas for the inverse that just involve sub-matrices of the original large matrix, such painful treatments would be time-consuming tasks.

Our approach is also applicable when one uses an informative prior for the coefficient vector $\boldsymbol{\beta}$. Zellner and Ando (2010b) have developed the DMC algorithm for the analysis of SUR model under an informative prior for the coefficient vector $\boldsymbol{\beta}$. Thus, replacing the DMC sampling procedure used in Section 3.1 by the DMC with the informative prior (Zellner and Ando (2010b)), one can easily apply our approach.

4 Simulation results

In order to assess the performance of our proposed procedures, we first present numerical results based on simulated data. We simulate data sets from the $m = 2$ dimensional SUR model. Without loss of generality in the model structure, we set the number of predictors for each of the equations to be $p_j = 100$. This model can thus be written as follows:

$$\begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{pmatrix} = \begin{pmatrix} X_1 & O \\ O & X_2 \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} + \begin{pmatrix} \boldsymbol{u}_1 \\ \boldsymbol{u}_2 \end{pmatrix}, \tag{8}$$

for $i = 1, ..., n$, where $\boldsymbol{y}_j$ and $\boldsymbol{u}_j$ are $n \times 1$ vectors, X_j is the $n \times 100$ matrix and $\boldsymbol{\beta}_j$ is the 100-dimensional vector. Each element of Ω is set to be

$$\Omega = \begin{pmatrix} \omega_1^2 & \omega_{12} \\ \omega_{21} & \omega_2^2 \end{pmatrix} = \begin{pmatrix} 0.1 & -0.05 \\ -0.05 & 0.2 \end{pmatrix}.$$

The covariate matrices X_j $j = 1, 2$ were generated from a uniform density over the interval $(-1, 1)$ The coefficient vector was set to be $\boldsymbol{\beta}_1 = (3, -2, 1, 0, 0, \cdots, 0)'$ and $\boldsymbol{\beta}_2 = (2, 1, 1, 0, 0, \cdots, 0)$. This enabled the generation of simulated response observations. In this simulation we set the number of observations to be $n = 50$. Thus, the number of covariates p_j $j = 1, 2$ are much larger than the sample size n.

To compare the accuracy of our method, we also applied the MCMC method of Smith and Kohn (2000). Following their paper, the first 1,000 iterations

are discarded as a burn-in period. The remaining 1,000 samples are used for inference. The method of Smith and Kohn allows us to compute the posterior probability that each of the predictors is included.

Many Bayesian analyses are done under the assumption that the posterior samples from the MCMC algorithms are independent samples while, as many have recognized, the generated samples exhibit autocorrelation. Figure 1 shows an autocorrelation function of successive draws of the covariance parameter ω_{12} from the MCMC output from the output of our method and from the Smith-Kohn method. As shown in Figure 1, we have to take autocorrelation into account when we compute the standard deviations and many other quantities from the MCMC posterior samples. The autocorrelations of the output of our method are generally much smaller than those of the Smith-Kohn method. We also calculated the inefficiency factor (1+sum of the squared autocorrelations from lag 1 to L. Here we set $L = 500$). It is useful as a measure of the efficiency of alternative sampling algorithms. A large value of inefficiency factor indicates that we need a large number of MCMC simulations. We found that the calculated inefficiency factor for ω_{12} from our method is 1.2791 and that from the MCMC method of Smith and Kohn (2000) is 5.8545, indicating that our procedure is more efficient. It also implies that the proposed method is much more efficient than the MCMC method of Smith and Kohn (2000).

We repeated the above Monte Carlo simulation for 100 trials. Thus, we obtain 100 posterior mean values of $\bar{\gamma}_j^k$, $j = 1, 2$, $k = 1, ..., 100$. As a result, we found that our approach resulted in the averaged posterior probability that each of the true predictors (there are 6 true predictors) is included, $\sum_{i=1}^{100} \bar{\gamma}_j^k(i)/100$, ranges from 1.00 to 0.999, exceptionally good performance. Here $\bar{\gamma}_j^k(i)$ is the posterior mean values of γ_j^k at i-th simulation. On the other hand, the method of Smith and Kohn resulted in the averaged posterior probability that each of the true predictors is included ranges from 1.00 to 0.999. It is also a good performance.

Also, we calculated the other false side by calculating the number of selected times for the unrelated predictors. With regard to unrelated predictors, the posterior probability that each of the false predictors (there are 394 such predictors) is included ranges from 0.040 to 0.085. On the other hand, those from the method of Smith and Kohn ranges from 0.040 to 0.090. Therefore, our method is alightly accurate than the Smith-Kohn (2000) method.

We also compared the mean squared errors (MSEs)

$$\text{MSE} = \sqrt{\frac{1}{n}(X_1\boldsymbol{\beta}_1 - \hat{\boldsymbol{y}}_1)'(X_1\boldsymbol{\beta}_1 - \hat{\boldsymbol{y}}_1) + \frac{1}{n}(X_2\boldsymbol{\beta}_2 - \hat{\boldsymbol{y}}_2)'(X_2\boldsymbol{\beta}_2 - \hat{\boldsymbol{y}}_2)}$$

for the true structure $X_j\boldsymbol{\beta}_j$ and the estimated structure $\hat{\boldsymbol{y}}_j$. We used the

predictive mean $\hat{\boldsymbol{y}}_j$. Because we generated 100 Monte Carlo trials, we can calculate the mean values of the MSEs and their estimated standard deviations. The mean values of the MSE are as follows: Our method: 0.084 (0.05) Smith and Kohn: 0.085 (0.05) Here the numbers in parenthetic are the estimated standard deviations. In the sense of MSE, there is no significant difference between these two methods. However, as regards computational times, we found that our method is more efficient than the method suggested by Smith and Kohn.

Using the model structure in (8), we calculated the computational times of the two methods, our approach and the method of Smith and Kohn. Without loss of generality, we set the number of predictors for each of the equations to be $p_j = 10$. For the simulated data set, we generated 10 different samples, each of size $n = 100$. As a result, 10 computational times are recorded for each of these methods. We found that the averaged time (sec.) to produce 100 posterior samples as follows: Smith and Kohn (2000): 13.91 (0.152) Our approach: 13.59 (0.066). Here the numbers in parenthetic are the standard deviations. Therefore, our method is slightly faster than the Smith-Kohn (2000) method.

In practical use of the posterior samples from MCMC outputs, researchers should take account the autocorrelations. One of the most popular approaches is to use every k-th posterior samples. Thus, the remaining samples are discarded. The number k is usually determined by considering the autocorrelation. If the autocorrelation is relatively large, the value k would become large. On the other hand, the value k would be small if the autocorrelation is small. Ideal situation is zero autocorrelation, where we can set $k = 1$ and there is no posterior samples to be discarded. Noting that the autocorrelation of the method of Smith and Kohn (2000) is much larger than that of our method, we checked the computational time to obtain 100 posterior (independent) samples. For our method, every 5-th posterior samples are stored. For the method of Smith and Kohn (2000), every 10-th posterior samples are stored. Then, 10 computational times are recorded for each of these methods. We found that the averaged time (sec.) to produce 100 posterior (independent) samples as follows: Smith and Kohn (2000): 138.05 (0.428) Our approach: 68.19 (0.158) Again, the numbers in parenthetic are the standard deviations. The required time to run MCMC method of Smith and Kohn (2000) will become larger than our method, because larger steps are needed to obtain an independent posterior samples.

5 Real data analysis

There have been a number of studies attempting to establish an excellent technique for estimating the term structure of interest rates from a cross-section of

coupon bond prices. Under the assumption that the price of a bond is equal to the present value of its future coupon payments and redemption, McCulloch (1971) regressed cash flows on a set of basis functions to estimate discount functions. Here, we shall use SUR system to capture the term structure of interest rates using a set of cross-section of coupon bond prices.

5.1 Bond equation

Let p be the price of bond, c be its coupon payment, which is paid at time $t_1,...,t_L$, let R be the redemption payment, and let L be the number of remaining payments. Following the theory of bond pricing (McCulloch, 1971), we assume that the price of a bond is equal to the present value of its future coupon payments and the redemption, i.e.,

$$p = \sum_{k=1}^{L_\alpha} c \times \delta(t_k) + R \times \delta(t_L) + \varepsilon,$$

where $\delta(\cdot)$ is the discount function. The discount function $\delta(t)$ gives the present value of a monetary unit, e.g., \$1.00 after t years. Most researchers follow McCulloch (1971) in explicitly constraining cash flows from different bonds due at the same time to be discounted at the same rate, and estimate the discount function $\delta(\cdot)$ from which the other yield curves can be derived.

We employ the most basic case where splines are placed on the discount function. In this case, $\delta(\cdot)$ is expressed as a linear combination of a set of m underlying basis functions, as follows.

$$\delta(t;\boldsymbol{\beta}) = 1 + \sum_{k=1}^{m} \beta_k b_k(t)$$

Here we shall use McCulloch (1975)'s cubic spline basis.

It then follows that the bond price model based on a linear combination of basis functions is expressed as follows.

$$p = [\boldsymbol{a}'B]\,\boldsymbol{\beta} + \varepsilon, \tag{9}$$

where $B = (\boldsymbol{b}(t_1), ..., \boldsymbol{b}(t_L))'$, $\boldsymbol{a} = (c, ..., c, c + R)'$, respectively.

Once the discount function is estimated, the zero-coupon yield and the forward rate can be obtained by transformations of the discount function. It is widely

known that the discount function $\delta(t)$ and the instantaneous forward rate $f(t)$ are related by

$$f(t) = -\delta'(t)/\delta(t),$$

where $\delta'(t)$ is the derivative of the discount function $\delta(\cdot)$ evaluated at the point t. Thus, after the discount function is obtained, we these the instantaneous forward rate $f(t)$ can be derived.

Next section describes the dataset and results.

5.2 Dataset, SUR system specification and results

As an illustration of the practical application of the proposed procedure, the method is applied to the analysis of Japanese governmental bonds trading data observed on September 2nd and 3rd, 2002. Here $n = 219$. Data is publicly available on line from the web site of Japan Securities Dealers Association.

Using the bond equation (9), we have a set of two regression equations. One is for the data traded on September 2nd and the other is for traded on September 3rd. In this case, the regression equation for the data traded on a particular date is

$$\begin{pmatrix} p_1 \\ \vdots \\ p_{219} \end{pmatrix} = \begin{pmatrix} \boldsymbol{a}_1' B_1 \\ \vdots \\ \boldsymbol{a}_{219}' B_{219} \end{pmatrix} \boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where p_α, $\boldsymbol{a}_\alpha$ and B_α are known quantities for the α-traded bond. For each date, the regression question above applies. Thus, once we allow the correlation structure between the noise terms in the regression question for the data traded on September 2nd and those in that for traded on September 3rd, this specification reduces to the SUR system.

Setting the number of basis functions to be $m = 20$ for each equation, the method is applied to this data. The first 1,000 iterations are discarded as a burn-in period. The remaining 1,000 samples are used for inference. Estimated discount functions and forward rate curves are shown in Figure 2. The solid lines are posterior mean curves. The posterior mean curves for the discount function $\delta(t)$ is given as $\frac{1}{N}\sum_{k=1}^{N} \delta(t; \boldsymbol{\beta}^{(k)})$. Here N is the number of posterior samples and $\boldsymbol{\beta}^{(k)}$ is the k-th posterior sample. Similarly, the posterior mean curves for the forward rate $f(t)$ can be calculated using the relation

$f(t) = -\delta'(t)/\delta(t)$. Two dashed lines are 95% confidence intervals. The 95% confidence intervals are estimated using the 2.5th and 97.5th percentiles of the posteriors. The results for the discount function and zero coupon yield curves are almost identical. From the forward rate curve, we can also see that the degree of uncertainty increases as the time to maturity becomes longer.

Using the posterior outputs, we can make an inference about the correlation structure. The posterior mean, the standard deviation, and 95% confidence intervals are 0.945, 0.048, and [0.807,0.984], respectively. Using the posterior draws for each of the parameters, we calculated the posterior means, the standard deviations and 95% confidence intervals. The 95% confidence intervals are estimated using the 2.5th and 97.5th percentiles of the posterior samples. Also, Figure 3 shows the estimated posterior density of the correlation. From these investigations, we can see that there is a significant correlation structure.

6 Summary and Conclusions

Computationally efficient methods for Bayesian analysis of seemingly unrelated regression (SUR) models are developed. Under a Bayesian hierarchical framework where each regression function is represented as a linear combination of a large number of basis functions, the regression coefficients, the variance matrix of the errors, and a set of variables to be included in the model are estimated simultaneously. The method is based on MCMC sampling scheme, and we employed a DMC approach for sampling efficiency.

There are several advantages of our approach compared with the MCMC approach of Smith and Kohn (2000). One is that because it decomposes the joint conditional density of the coefficient vector into a set of low-dimensional conditional densities, we can avoid large scale matrix calculations. We found that our method is more computationally efficient than the method of Smith and Kohn (2000). The autocorrelation function from our method is smaller than those from the MCMC method of Smith and Kohn (2000). We calculated an inefficiency factor and found that the calculated inefficiency factor from our method is smaller than that from the MCMC method of Smith and Kohn (2000). It implies that the proposed method is much more efficient than the MCMC method of Smith and Kohn (2000). We would recommend implementing Bayesian analysis of SUR model based on our approach.

Acknowledgement

The author is indebted to Professor Arnold Zellner, the University of Chicago, Booth School of Business for his encouragement and also for helpful discussions and suggestions.

References

[1] Ando, T. and Zellner, A. (2010) Hierarchical Bayesian analysis of the seemingly unrelated regression and simultaneous equation models using a combination of a direct Monte Carlo and an importance sampling techniques. Bayesian Analysis, in press.

[2] Box, G. E. P. and Tiao, G. C. (1973) *Bayesian Inference in Statistical Analysis*. MA: Addison-Wesley.

[3] Carroll, R. J., Doug M., Larry F. and Victor K. (2006) Seemingly unrelated measurement error models, with application to nutritional epidemiology. Biometrics, 62, 75-84.

[4] Frasera, D.A.S., Rekkasb, M. and Wong, A. (2005). Highly accurate likelihood analysis for the seemingly unrelated regression problem. Journal of Econometrics, 127, 17-33.

[5] Gallant, R. (1975). Seemingly unrelated nonlinear regressions. Journal of Econometrics, 3, 35-50.

[6] Geweke, J. (2005). *Contemporary Bayesian Econometrics and Statistics*. New York: Wiley.

[7] Greene, W.H. (2002). *Econometric Analysis* (5th ed.). New Jersey: Prentice-Hall.

[8] Jeffreys, H. (1946). An Invariant Form for the Prior Probability in Estimation Problems. Proceedings of the Royal Society of London, Series A, 196, 453–461.

[9] Jeffreys, H. (1961). *Theory of Probability* (3rd ed.). Oxford: Oxford University Press, and published in Oxford classic texts in the physical science, 1998.

[10] Kurata, H. (1999). On the efficiencies of several generalized least squares estimators in a seemingly unrelated regression model and a heteroscedastic model. Journal of Multivariate Analysis, 70, 86-94.

[11] Lancaster T. (2004). *Introduction to Modern Bayesian Econometrics*. New Jersey: Cambridge University Press.

[12] Liu, A. (2002). Efficient estimation of two seemingly unrelated regression equations. Journal of Multivariate Analysis, 82, 445-456.

[13] Mandy, D. M. and Martins-Filho, C. (1993). Seemingly unrelated regressions under additive heteroscedasticity: theory and share equation applications. Journal of Econometrics, 58, 315-346.

[14] McCulloch, J. H. (1971). Measuring the Term Structure of Interest Rates, Journal of Business, 44, 19–31.

[15] McCulloch, J. H. (1975) The Tax-Adjusted Yield Curve, Journal of Finance, 30, 811–830.

[16] Neudecker, H. and Windmeijer, F. A. G. (1991). R2 in seemingly unrelated regression equations. Statistica Neerlandica, 45, 405-411.

[17] Ng, V. M. (2002). Robust Bayesian Inference for Seemingly Unrelated Regressions with Elliptical Errors. Journal of Multivariate Analysis, 83, 409–414

[18] Percy, D. F. (1992). Predictions for Seemingly Unrelated Regressions, Journal of the Royal Statistical Society, Series, B54, 243–252.

[19] Press, S. J. (1972). *Applied Multivariate Analysis.* New York: Holt, Rinehart and Winston, Inc.

[20] Rocke, D. M. (1989). Bootstrap Bartlett adjustment in seemingly unrelated regression. Journal of the American Statistical Association, 84, 598-601.

[21] Rossi, P.E, Allenby, G. and McCulloch, R. (2005). *Bayesian Statistics and Marketing.* NJ: John Wiley and Sons.

[22] Smith, M. and Kohn, R. (2000). Nonparametric seemingly unrelated regression. Journal of Econometrics, 98, 257–282.

[23] Srivastava, V. K. and Giles, D. E. A. (1987). *Seemingly unrelated regression equations models.* New York: Dekker.

[24] van der Merwe, A. and Viljoen, C. (1988). Bayesian analysis of the seemingly unrelated regression model. Manuscript, University of the Free State, Department of Mathematical Statistics.

[25] Zellner, A. (1962). An efficient method of estimating seemingly unrelated regression equations and tests for aggregation bias. Journal of the American Statistical Association, 57, 348–368.

[26] Zellner, A. (1963). Estimators for seemingly unrelated regression equations: some exact finite sample results. Journal of the American Statistical Association, 58, 977-992.

[27] Zellner, A. (1971). *An introduction to Bayesian inference in econometrics.* New York : Wiley.

[28] Zellner, A. and Ando, T. (2008). A direct Monte Carlo approach for Bayesian analysis of the seemingly unrelated regression model. H.G.B. Alexander Research Foundation, Graduate School of Business, University of Chicago.

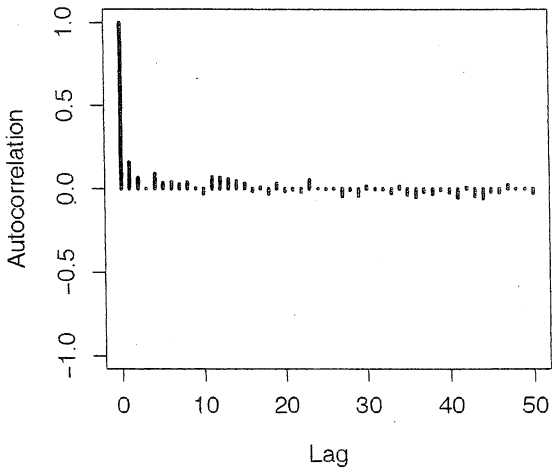
[31] Zellner, A. and Ando, T. (2010a). Bayesian and Non-Bayesian Analysis of the Seemingly Unrelated Regression Model with Student-t Errors and Its Application. International Journal of Forecasting, in press.

[30] Zellner, A. and Ando, T. (2010b). A direct Monte Carlo approach for Bayesian analysis of the seemingly unrelated regression model with informative prior (in Japanese). Journal of the Japan Statistical Society, in press.

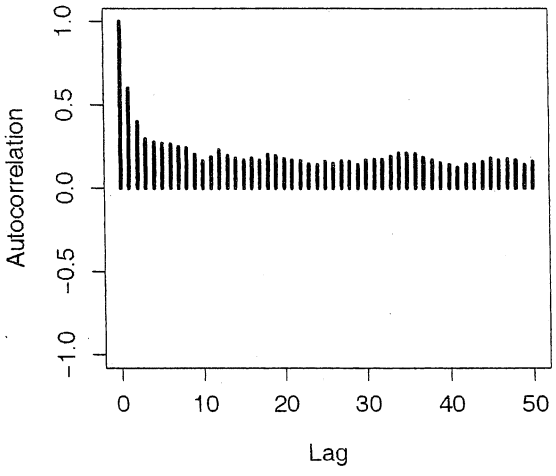
[31] Zellner, A. and Ando, T. (2010c). Rejoinder to comment on "Bayesian and Non-Bayesian Analysis of the Seemingly Unrelated Regression Model with Student-t Errors and Its Application for Forecasting". International Journal of Forecasting, in press.

[32] Zellner, A. and Chen, B, (2002). Bayesian Modeling of Economies and Data Requirements. Macroeconomic Dynamics, 5, 673–700.

[33] Zellner, A. Bauwens, L. and Van Dijk, H. K. (1988). Bayesian specification analysis and estimation of simultaneous equation models using Monte Carlo Methods. Journal of Econometrics, 38, 39-72.

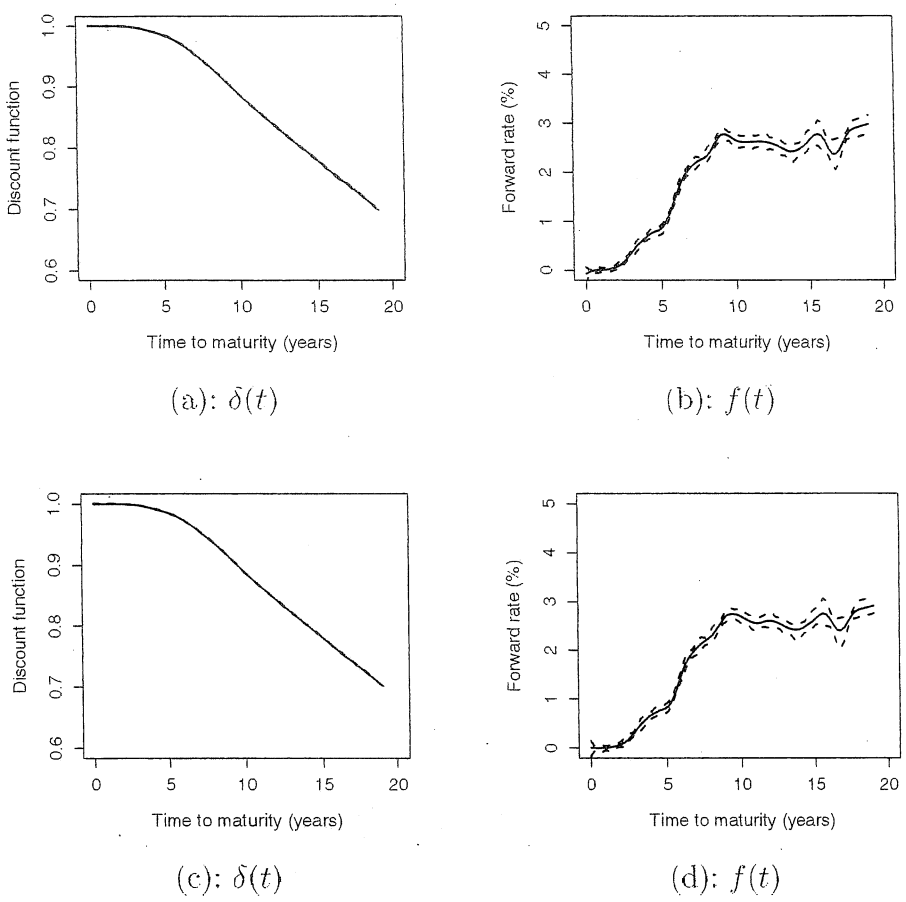


Our method



Smith and Kohn (2000)

Fig. 1. Autocorrelation function of successive draws of the covariance parameter ω_{12} from the output of our method and from that of Smith and Kohn (2000). The autocorrelation function from our method is smaller than that from the MCMC method of Smith and Kohn (2000).



(a): $\delta(t)$ (b): $f(t)$

(c): $\delta(t)$ (d): $f(t)$

Fig. 2. Discount function $\delta(t)$ and forward rate $f(t)$ for the trading date September 2nd 2002 (Figures a and b) and September 3rd 2002 (Figures c and d). The solid lines are posterior mean curves. The posterior mean curves for the discount function $\delta(t)$ is given as $\frac{1}{N}\sum_{k=1}^{N}\delta(t;\boldsymbol{\beta}^{(k)})$. Similarly, the posterior mean curves for the forward rate $f(t)$ can be calculated using the relation $f(t) = -\delta'(t)/\delta(t)$. Two dashed lines are 95% confidence intervals. The 95% confidence intervals are estimated using the 2.5th and 97.5th percentiles of the posteriors.

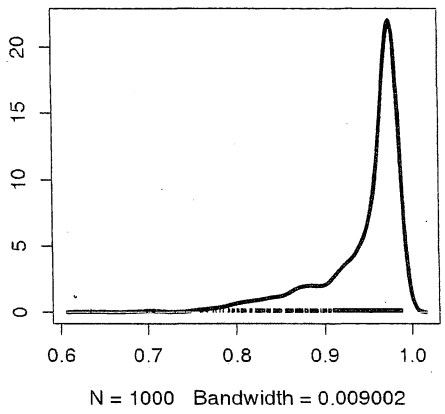


Fig. 3. Estimated posterior densities for the correlation parameter.