

Parametric Covariance Matrix Modeling in Bayesian Panel Regression

Mickael Salabasis*

Dept. of Economic Statistics

Stockholm School of Economics

*SSE/EFI Working Paper Series in Economics and Finance
No. 565*

Abstract

The full Bayesian treatment of error component models typically relies on data augmentation to produce the required inference. Never strictly necessary a direct approach is always possible though not necessarily practical. The mechanics of direct sampling are outlined and a template for including model uncertainty is described. The needed tools, relying on various Markov chain Monte Carlo techniques, are developed and direct sampling, with and without effect selection, is illustrated.

Keywords: Bayesian; parametric covariance; model selection

JEL codes: C11, C33, C63

*Mickael Salabasis, SSE ES, P.O. Box 6501, SE-113 83 Stockholm, Sweden.

1 Introduction

Many interesting panel regression models may be defined in terms of a simple covariance structure specified by just a few parameters. Examples include, but are in no way limited to, error component models and models having serially correlated errors. Parameterizing the mean and covariance structures separately, any difficulties encountered are mostly associated with the latter. In Bayesian analysis, problems usually stem from a lack of analytical results and the modern solutions typically involve posterior simulation.

Data augmentation, introduced by Gelfand and Smith (1990), is a frequently used tool where the ability to sample, augment and condition on latent variables may result in major simplifications for convenient prior structures. Although a generally applicable method, data augmentation is not necessarily efficient. Also, while elegant and simple, for covariance modeling it is never strictly necessary as parameterization of the covariance matrix and direct sampling is always possible, albeit not always very practical. This paper shows how direct sampling can be implemented for common panel data models and further demonstrates how direct sampling is a viable alternative to data augmentation.

For direct sampling to be practical and effective, the main requirement is that the simple structure of the covariance matrix, in a broad sense, carries over to its inverse. When this is the case, convenient expressions for the full conditional posterior are readily available. Complicating the posterior simulation, these will typically not be known densities. Often involving simple, but possibly nonlinear, polynomials in the covariance parameters they may instead be used to create arbitrarily good approximations. These may in turn be used to either sample candidate values for Metropolis type updates, apply the gridy Gibbs sampler of Ritter and Tanner (1992), or form the basis of any resampling method. Whatever the preferred method, an expression for the determinant in terms of the parameters and bounded support for the parameters is helpful but not decisive; the alternative being brute force calculations.

In addition to providing a generic, if not always practical, solution to some of the problems associated with data augmentation, direct sampling brings other advantages. Many models imply restrictions on the covariance matrix, so that direct parameterizing will be efficient in terms of the number of unknowns in the model. For instance, in error component models the status of an effect as present or absent is typically governed by a single parameter, to be compared with many latent and augmented effects. This also simplifies the administration and implementation of model selection, keeping the necessary adjustments to the posterior simulation at a minimum. The changing dimensionality of the parameter vector necessitates the use of more sophisticated sampling procedures such as the reversible jump Metropolis-Hastings developed by Green (1995).

The organization of this paper is as follows. In Section 2, the mechanics of direct covariance sampling is described. The covariance matrix properties that make a direct approach practical are briefly discussed and a generic algorithm for posterior simulation is presented for the case with standard conjugated priors on all other quantities. Section 3 extends the prior to cater for model selection and outlines a generic algorithm to include it. To illustrate the methods, Section 4 walks through the necessary computations for a simple error component model. Establishing the direct approach as a viable method, its performance is compared to two standard algorithms of varying levels of sophistication in the context of one-way error component models. Finally, effect selection is then illustrated in the context of the two-way error component model. Closing, Section 5 summarizes our experiences to date and offers some ideas for future research.

2 Parametric Covariance Modelling

Consider the standard panel regression model where the response y_{it} of individual i at time t is linked to p explanatory variables x_{it} by coefficients β common across units, and is observed with error ϵ_{it}

$$y_{it}|x_{it} = x'_{it}\beta + \epsilon_{it}, \quad i = 1, \dots, N, \quad t = 1, \dots, T. \quad (1)$$

Stacking the observations associated with unit i into a $T \times 1$ vector y_i , and a $T \times p$ matrix \mathbf{X}_i Bayesian inference may be conducted within the standard linear regression framework.

Assuming $\epsilon_i \sim \mathcal{N}(0, \Sigma_\epsilon)$ and standard conjugated prior structures, analytical results are available for the two extreme models with respect to the number of parameters embodied in the covariance matrix Σ_ϵ . On one end we have the trivial model with marginally independent and homoscedastic errors, imposing a diagonal structure on the covariance matrix and contributing with a single scale parameter. On the other end we have the unrestricted model where the covariance matrix contributes with a maximal $\bar{m} = T(T+1)/2$ parameters.

Contrasting, many interesting models reduce the number of unique elements in the covariance matrix or imply exact restrictions in terms of a small number of parameters. Examples include, but are not limited to, error component models and models with serially or spatially correlated errors. While more parsimonious with respect to the number of unknowns, analytical results are typically not available. Posterior analysis is usually performed by means of approximation, for instance using Markov chain Monte Carlo (MCMC) methods. Special solutions exist for select models.

However, it is possible to treat any model of type (1) within a single unifying *direct sampling* framework which may, under certain favorable circumstances, be

efficient as well as practical. Writing the proper covariance matrix as a product of a design matrix, $\mathbf{\Lambda}_\theta$, and a scale parameter, σ^2 , the inverse may always be represented as

$$\mathbf{\Sigma}_\epsilon^{-1} = \sigma^{-2} \mathbf{\Lambda}_\theta^{-1} = \sigma^{-2} \sum_{i=1}^m g_i(\theta) \mathbf{Q}_i, \quad (2)$$

for a set of m scalar functions $g = \{g_1, \dots, g_m\}$ of a set of $k \leq m$ parameters $\theta = \{\theta_1, \dots, \theta_k\}$ and a collection of constant matrices $Q = \{\mathbf{Q}_1, \dots, \mathbf{Q}_m\}$. In the case with no restrictions, the parameter vector includes the \bar{m} unrestricted elements of the inverse and the equal number of fixed matrices are zero everywhere except either at one position on the main diagonal or two positions symmetrically to the main diagonal. In general, every restriction introduced reduces the number of parameters, functions, and matrices needed.

While the precise definition of the functions and matrices is necessary for any actual application, they are not essential to the description of direct sampling. Regardless of the form of the design matrix, the unknown parameters in (1) are $\{\beta, \sigma^2, \theta\}$ and the joint distribution of all quantities, known and unknown, is the product of the likelihood and the prior

$$p(y, \beta, \sigma^2, \theta) = \mathcal{L}(y | \beta, \sigma^2, \theta) \pi(\beta, \sigma^2, \theta), \quad (3)$$

where the conditioning on the explanatory variables is implied. The likelihood is, given a standard normality assumptions and ignoring a proportionality constant,

$$\mathcal{L}(y | \beta, \sigma^2, \theta) \propto \sigma^{-NT} |\mathbf{\Lambda}_\theta|^{-N/2} \times \exp \left\{ -0.5 \sigma^{-2} \sum_{i=1}^N e_i' \mathbf{\Lambda}_\theta^{-1} e_i \right\}, \quad (4)$$

$$e_i = y_i - \mathbf{X}_i' \beta.$$

Further, adopting the prior structure

$$\pi(\beta, \sigma^2, \theta) \sim \pi(\beta | \sigma^2, \theta) \prod_{i=1}^k \pi(\theta_i | \sigma^2) \pi(\sigma^2), \quad (5)$$

the choice of a conditional independence structure for the covariance parameters is dictated by convenience.

Completing the specification of the prior, standard conjugate choices are made for the regression coefficients and the idiosyncratic error precision

$$\pi(\beta | \sigma^2, \theta) \sim \mathcal{N}_p(b_0, \sigma^2 \mathbf{B}_0)$$

$$\pi(\sigma^{-2}) \sim \mathcal{G}(\nu_0, \nu_1),$$

The prior for the covariance parameters is left unspecified as its precise definition is problem dependent and not essential at this stage of the presentation. Typically no convenient prior is available forcing an approximation of the posterior by means of for instance posterior simulation.

The conditional conjugate prior structure avoids any unnecessary complication of the posterior simulation. In particular, the full conditional posterior of the coefficients has a normal kernel, and updating them can be performed with a Gibbs step based on a generalized least squares type quantities. Similarly, a Gibbs step can be used for updating the error precision as its full conditional posterior has a gamma kernel. This limits any remaining difficulties to the covariance parameters in θ . However, using the representation of the inverse in (2) we may begin to outline a computationally potentially rather efficient strategy.

Defining the set $\theta_{-j} = \theta \setminus \theta_j$ for any element θ_j of θ , using (2) and rearranging slightly, the full conditional posterior is proportional to

$$p(\theta_j | y, \beta, \sigma^2, \theta_{-j},) \propto |\mathbf{\Lambda}_\theta|^{-N/2} \exp \left\{ -0.5\sigma^{-2} \sum_{k=1}^m g_k(\theta) c_k \right\}, \quad (6)$$

where

$$c_k = s_k - 2\beta'v_k + tr(\mathbf{M}_k\beta\beta'),$$

$$s_k = \sum_{i=1}^N y_i' \mathbf{Q}_k y_i, \quad v_k = \sum_{i=1}^N \mathbf{X}_i' \mathbf{Q}_k y_i, \quad \mathbf{M}_k = \sum_{i=1}^N \mathbf{X}_i' \mathbf{Q}_k \mathbf{X}_i.$$

While (6) is seldom a known density being a simple function of the parameter of interest, it can be used for quick and efficient evaluation of the posterior for select values. Doing so for a reasonably large number of values, an adequate first order approximation of the full conditional posterior may be constructed. It can then in turn be used either to sample the next value immediately, applying the griddy Gibbs sampler of Ritter and Tanner (1992), to sample candidate values, for Metropolis type updates, or to serve as the basis for the implementation of any resampling method such as Acceptance-Rejection (AR) sampling. Without making any claims of optimality, Algorithm 1 describes the building blocks of one possible simple generic MCMC implementation.

Studying Algorithm 1 closer, the complexity, extent and cost of the necessary computations is increasing in the number and complexity of the functions needed to write the inverse. That is, practical and efficient direct sampling seems to require that the simple structure of the covariance matrix, in a broad sense, carries over to its inverse. In addition to an explicit structure for the inverse, computational efficiency of the direct approach described relies on the availability of an expression

Algorithm 1 A generic sampler.

1. Conditional on $\{\sigma^2, \theta\}$, sample and accept a proposal for β from its full conditional posterior. Due to the (conditional) Gaussian structure,

$$y_i | \beta, \sigma^2, \theta \sim \mathcal{N}_t(\mathbf{X}_i \beta, \sigma^2 \mathbf{\Lambda}_\theta), \quad i = 1, \dots, N$$

and combining for all units with the prior gives

$$\beta | \sigma^2, \theta \sim \mathcal{N}_p(b_1, \sigma^2 \mathbf{B}_1),$$

where

$$\mathbf{B}_1 = \left[\mathbf{B}_0^{-1} + \sum_{j=1}^m g_j(\theta) \mathbf{M}_j \right]^{-1}, \quad b_1 = \mathbf{B}_1 \left(\mathbf{B}_0^{-1} b_0 + \sum_{j=1}^m g_j(\theta) v_j \right)$$

with the \mathbf{M}_j and v_j as defined in (6).

2. Conditional on $\{\beta, \theta\}$, sample and accept a proposal for σ^{-2} from its full conditional posterior. The conjugate structure results in

$$\sigma^{-2} | \beta, \theta \sim \mathcal{G}(\nu_0 + n_0, \nu_1 + S_1),$$

$$n_0 = NT/2, \quad S_1 = \sum_{j=1}^m g_j(\theta) c_j/2,$$

with c_j defined as in (6).

3. Cycle through θ element by element and update θ_j conditional on $\{\beta, \sigma^{-2}, \theta_{-j}\}$. Selecting the desired number of nodes, approximate the full conditional posterior, $q(\theta_j)$, using the results in (6)

$$\theta_j | \beta, \sigma^2, \theta_{-j} \propto |\mathbf{\Lambda}_\theta|^{-N/2} \exp \left\{ -0.5 \sigma^{-2} \sum_{k=1}^m g_k(\theta) c_k \right\} \pi(\theta_j | \sigma^2).$$

Note that in the exponent we only need to evaluate the factors where θ_j is actually referenced in $g_k(\theta)$. Update the parameter using any preferred method such as the griddy Gibbs sampler, Metropolis-Hasting updates or resampling with AR-algorithm.

4. Repeat steps 1-3 using the most recent values of the conditioning variables until convergence or a stopping criterion is met.

for the determinant as a function of θ . Also helpful, but in no way essential, is that the support of the parameters in θ is bounded as this can simplify the construction of an approximation to the full conditional posterior.

For computational efficiency, note how the sets of scalars $s = \{s_1, \dots, s_m\}$, $p \times 1$ vectors $v = \{v_1, \dots, v_m\}$ and $p \times p$ matrices $\mathbf{M} = \{\mathbf{M}_1, \dots, \mathbf{M}_m\}$ in (6) only depend on fix quantities and need to be computed just once. Further, the set of scalars $c = \{c_1, \dots, c_m\}$ is known conditional on β and though it must be recomputed for every new instance of β , the computational burden depends on p which is typically small relative N and T . Consequently, efficiency is mainly a matter of efficient construction of and sampling from the approximation in Step 3 of Algorithm 1.

Assuming elements of θ are treated one by one, evaluating the full conditional posterior on a grid using the results in (6) is a relatively simple task. Issues like the number and placement of nodes is critical for the quality of the approximation and the classical conflict of precision and computational speed applies two ways. First, adding nodes will increase the precision of the estimated cumulative density function at the cost of more function evaluations. Second, gains in precision from more advanced integration techniques or elaborate node placement strategies are often lost to loss of speed. Further, the computational burden also depends on the method selected. For instance, adding to the overhead, Metropolis type updates require the calculation of an acceptance probability, while the implementation of the AR algorithm requires the calculation of an envelope constant. On the positive side, depending on the problem, more or less clever ways of constructing the approximation may be available.

3 Introducing Model Selection

With the elements of θ having a physical interpretation as a rule, any model under consideration may nest a number of simpler structures. At the very least, for specific values on combinations of elements it will typically collapse into the standard model with marginally independent and homoscedastic errors. Making model selection relevant, incorporating it into the direct sampling framework is simply a matter of appropriate prior extension. In the simplest case we want to distinguish two states for any covariance parameter θ_j . One special state, $\theta_j = \theta_j^*$ say, where it does not contribute to the complexity of the model, and one arbitrary active state where it does.

The desired effect is achieved with a mixed component prior for any parameter on the form

$$\pi(\theta_j | \theta_{-j}) \sim (1 - w_j) \cdot I_{\theta_j = \theta_j^*} + w_j \pi(\theta_j), \quad (7)$$

where $0 < w_j < 1$ a weight, $I_{\theta_j=\theta^*}$ the standard indicator function for when the parameter takes the special value and $\pi(\theta_j)$ an admissible prior for the variance component. Selecting weights $w = \{w_1, \dots, w_k\}$ and the respective continuous components completes the specification of the model. Lacking prior information, weights are typically all set to one half; interpreted as a reference uninformative choice. Notice how selecting a weight at the either end of the admissibility region imposes the absence ($w = 0$) or presence ($w = 1$) of an effect. Reasonable choices for the continuous components is problem dependent. Extending the prior further, to include for instance more than one special state or even parameter specific number of states, is straightforward.

Convolving the individual priors, assuming two states for each parameter, the resulting joint prior is a mixed component prior with 2^k components. Answering to a particular combination of parameters being active, each component associates with model. Implying a mixed posterior, the mixture being over the models of interest, inference conditional on or averaged over the type of model is possible. With the dimension or content of the parameter vector depending on the model, posterior simulation is performed using the reversible jump MCMC algorithm. Straightforward application of the chosen method imposes a restriction on the prior above. Completing it with the necessary continuous components, while the choice is not essential, the reversible jump requires them to be proper. Also necessary are a set of steps to update parameters conditional on the model, and a set of moves that allows the chain to go from one model to another. Algorithm 2 briefly describes a possible modification of Algorithm 1 to allow just that.

The idea is simple, always attempting to switch between the two main effect states. All the action is in Step 3 which administrates both the exploration of a given model as well as the transition between model spaces. In Step 3b, while adding to the overhead of the algorithm, updating the active effects prior to an attempted deletion improves the mixing properties of the chain. Further, while important, the role of the approximation is limited to supplying candidate effect values. Without a need to differentiate between proposals for an existing effect and proposals associated with either creating or killing an effect, sharing a single approximation of the full conditional posterior promotes computational efficacy. Modifying the algorithm for situations with more than two effect states is straightforward.

For moves associated with model space transitions it is necessary to compute an acceptance probability and, depending on it, sample a new state. If successful, the model index is then changed and the parameter is added or deleted. The acceptance probability can always be stated in the standard form of a product of a likelihood, prior, and proposal ratio; for a detailed description of the method we refer to the original presentation in Green (1995).

Algorithm 2 A generic effect selection sampler.

1. Conditional on $\{\sigma^2, \theta\}$, sample and accept a proposal for β from its full conditional posterior as in Step 1 of Algorithm 1.
 2. Conditional on $\{\beta, \theta\}$, sample and accept a proposal for σ^2 from its full conditional posterior as in Step 2 of Algorithm 1.
 3. Conditional on σ^2, β update θ . Querying the current status of the effect, attempt to change it.
 - (a) If the selected parameter is not active attempt to insert it.
Generate the appropriate approximation as in Step 3 of Algorithm 1 and sample a candidate value. Compute an acceptance probability and sample the next state. If successful, change the model index and set the parameter to the proposed value.
 - (b) If the selected parameter is active attempt to delete it.
Generate the appropriate approximation and sample a new value as in Step 3 of Algorithm 1. Attempt the inverse of Step 3a with the updated parameter value.
 4. Repeat steps 1-3 using the most recent values of the conditioning variables until convergence or a stopping criterion is met.
-

4 An Illustrative Example

4.1 The two-way error component model

To illustrate direct sampling, consider the two-way random effects model,

$$\begin{aligned}
 y_{it} &= x_{it}\beta + \varepsilon_{it}, \quad i = 1, \dots, N, \quad t = 1, \dots, T \\
 \varepsilon_{it} &= \mu_i + \lambda_t + \nu_{it} \\
 \mu_i &\sim \mathcal{N}(0, \sigma_\mu^2), \quad \lambda_t \sim \mathcal{N}(0, \sigma_\lambda^2), \quad \nu_{it} \sim \mathcal{N}(0, \sigma_\nu^2),
 \end{aligned} \tag{8}$$

where the error term ε_{it} now consists of three parts: a unit effect μ_i which is constant across time for a given unit i , a time effect λ_t which is common across units at a given date, and an idiosyncratic error ν_{it} . Each random effect is independently and identically normally distributed, and also independent of each other as well as of x_{it} . Arranging the data appropriately and stacking the individual measurements, the covariance matrix can be written as

$$\Sigma_\varepsilon = \sigma_\nu^2 (\mathbf{I}_N \otimes \mathbf{I}_T) + \sigma_\mu^2 (\mathbf{I}_N \otimes \mathbf{J}_T) + \sigma_\lambda^2 (\mathbf{J}_N \otimes \mathbf{I}_T), \tag{9}$$

where \mathbf{I}_T the identity matrix and \mathbf{J}_T a matrix of ones. Following Baltagi (1995), defining the idempotent counterparts $\bar{\mathbf{J}}_T \equiv T^{-1}\mathbf{J}_T$ and $\mathbf{E}_T \equiv \mathbf{I}_T - \bar{\mathbf{J}}_T$, substituting into (9) and collecting terms gives

$$\boldsymbol{\Sigma}_\epsilon = \sigma_\nu^2 [\mathbf{Q}_1 + \theta_\mu^{-1}\mathbf{Q}_2 + \theta_\lambda^{-1}\mathbf{Q}_3 + (\theta_\mu^{-1} + \theta_\lambda^{-1} - 1)\mathbf{Q}_4],$$

$$\begin{aligned} \theta_\mu &\equiv (T\sigma_\mu^2\sigma_\nu^{-2} + 1)^{-1}, & \theta_\lambda &\equiv (N\sigma_\lambda^2\sigma_\nu^{-2} + 1)^{-1}, & 0 < \theta_\mu, \theta_\lambda &\leq 1 \\ \mathbf{Q}_1 &= \mathbf{E}_N \otimes \mathbf{E}_T, & \mathbf{Q}_2 &= \mathbf{E}_N \otimes \bar{\mathbf{J}}_T, & \mathbf{Q}_3 &= \bar{\mathbf{J}}_N \otimes \mathbf{E}_T, & \mathbf{Q}_4 &= \bar{\mathbf{J}}_n \otimes \bar{\mathbf{J}}_t. \end{aligned}$$

The functions represent the characteristic roots, with respective multiplicity being $(N-1)(T-1)$, $N-1$, $T-1$, and 1. Also, the \mathbf{Q}_i are independent of θ , idempotent and sum to the identity matrix. Thus, the inverse and determinant are given by

$$\begin{aligned} \boldsymbol{\Sigma}_\epsilon^{-1} &= \sigma_\nu^2 \boldsymbol{\Sigma}_\theta^{-1} = \sigma_\nu^{-2} [\mathbf{Q}_1 + \theta_\mu \mathbf{Q}_2 + \theta_\lambda \mathbf{Q}_3 + (\theta_\mu^{-1} + \theta_\lambda^{-1} - 1)^{-1} \mathbf{Q}_4], \\ |\boldsymbol{\Sigma}_\epsilon| &= \sigma_\nu^{2T} |\boldsymbol{\Sigma}_\theta| = \sigma_\nu^{2T} [\theta_\mu^{-N} \theta_\lambda^{-T} (\theta_\mu + \theta_\lambda - \theta_\mu \theta_\lambda)], \end{aligned}$$

so that in terms of the quantities defined in Section 2 we have

$$\begin{aligned} m &= 4, & \theta &= \{\theta_\mu, \theta_\lambda\} \\ g &= \{1, \theta_\mu, \theta_\lambda, \theta_\mu^{-1} + \theta_\lambda^{-1} - 1\} \end{aligned}$$

and Q defined as above.

Making the reparametrization, $\{\beta, \sigma_\nu^2, \theta_\mu, \theta_\lambda\}$ are the unknown parameters of the model. Stacking the data and dropping the summation, the likelihood is as in (4). As $\theta_\mu = 1$ and $\theta_\lambda = 1$ deletes the random unit and time effect respectively, convoluting mixed component priors on the form discussed in Section 3, the implied prior adopted for the variance components is

$$\begin{aligned} \pi(\sigma_\mu^2, \sigma_\lambda^2 | \sigma_\nu^2) &= w_0 \cdot I_{\theta_\mu=1} I_{\theta_\lambda=1} + w_\mu \cdot \pi(\sigma_\mu^2 | \sigma_\nu^2) I_{\theta_\lambda=1} + w_\lambda \cdot \pi(\sigma_\lambda^2 | \sigma_\nu^2) I_{\theta_\mu=1} + \\ &+ w_{\mu\lambda} \cdot \pi(\sigma_\mu^2 | \sigma_\nu^2) \pi(\sigma_\lambda^2 | \sigma_\nu^2). \end{aligned}$$

Here $\{w_0, w_\mu, w_\lambda, w_{\mu\lambda}\}$ is a collection of prior weights corresponding to models with no random effects, only a random unit effect, only a random time effect, and both random effects present.

Completing the prior specification any continuous components must be selected together with the prior model weights. As before, the choice of continuous prior components is not essential as long as they are proper. Lacking prior information the weights are typically set equal. In the posterior simulation, the expressions in Algorithm 1 are valid after the proper redefinition of the sets s , v , and \mathbf{M} to account for the effects of the stacking performed.

4.2 Relative Performance

The merits of direct sampling are investigated in the context of the simple one-way error component model. This model has a long history in the Bayesian literature, where early examples of a full Bayesian treatment can be found in for instance Tiao and Tan (1965) and Hill (1965). Box and Tiao (1973) contains a thorough presentation. A crucial element in the modern treatment, is the method of data augmentation introduced by Tanner and Wong (1987). Inference is facilitated by the ability to sample, augment and subsequently condition on the latent random effects and Gelfand, Sahu and Carlin (1995), Vines, Gilks and Wild (1996), and Gilks and Roberts (1996) among others propose various refinements to the basic algorithm presented in Gelfand and Smith (1990). Chib and Carlin (1999) offers a brief overview and presents two of the best procedures available to date.

For direct sampling in the one-way model without model selection, the results in Section 4.1 are valid after setting $\theta_\lambda = 1$ and adjusting the prior accordingly. Its performance is compared with the standard 3-block algorithm and the improved 2-block algorithm, presented and labeled A1 and A2 in Chib and Carlin (1999). Concentrating on the variance of the random effect and the intercept, aspects of the resulting Markov chains are illustrated with the relative numerical efficiency as the main indicator of performance. An attempt to characterize the convergence properties of the used samplers is made using the Yu and Mykland (1998) CUSUM plots, the Geweke (1992) diagnostic statistic and the estimated autocorrelation function.

The illustration uses the Grunfeld panel which, having small unit and time dimensions, is a fairly typical example of the kind of panels appearing in macroeconomic applications. This choice is further motivated by the extensive presentation of results for this panel in Baltagi (1995). Grunfeld (1958) considers a simple model of corporate investment where the real gross investment of firm i at date j is assumed to be a function of the firms' real value measured by the value of outstanding shares, x_1 , and the real value of its capital stock, x_2 . The panel consists of $N = 10$ firms observed over a period of $T = 20$ years.

In all examples a zero mean g-prior is selected for the regression parameters β and gamma priors for the precision of the error components. In direct sampling this implies a prior for the structural parameter θ_μ ,

$$\pi(\theta_\mu) \propto \theta_\mu^{\nu_3-1} (1 - \theta_\mu)^{-(\nu_3+1)} \exp\{-\nu_4\theta_\mu T / (1 - \theta_\mu) / \sigma_\nu^2\},$$

where (ν_3, ν_4) are the prior parameters for the random effect precision. While perhaps not the most natural choice, it is used for comparability. As this choice results in a full conditional posterior which is very similar to a truncated gamma distribution, an attempt to use this is made in the context of Metropolis-Hastings

updates by sampling proposals from

$$q(\theta_\mu) \sim \mathcal{G}\left(\nu_3 + N/2, \left(\nu_4 T + 0.5 \sum_{i=1}^N e_i' \bar{J} e_i\right) / \sigma_\nu^2\right).$$

All runs are made using the same or equivalent prior and 50000 samples are generated in each case. The same prior parameters are used for both variance components. In particular, $\nu_1 = \nu_3 = 1$ and $\nu_2 = \nu_4 = 100$, implying a low prior precision for data that is known to be noisy. For the regression coefficients, $g = 2000$ so that, as things are defined, the contribution of the prior precision will be small.

Data augmentation is applied to establish a benchmark. While the algorithm is very fast, the results are mixed as illustrated in Figure 1 and by Table 1.

For the random effect variance the sample path of the standardized CUSUM statistic, though consistently within the standard 5% tolerance, we observe how the convergence towards the final estimate is interrupted at intervals by sudden jumps. Matters are even worse for the intercept, with the sample path making long excursions away from any smooth convergence path. Autocorrelation is a problem across the board and though small for the random effect variance it is highly persistent. Computing the Geweke (1992) diagnostic statistics is problematic due to the difficulties in estimating numerical standard errors with any confidence. Selecting a large truncation lag and using the estimated autocorrelation time as in Chib and Carlin (1999), significant diagnostics are observed even after as much as a 40% burnin.

Discarding the initial 2500 iterations as burnin, the observations above are complemented and reinforced by the results in Table 1. The performance being what it is, the cut-off point is arbitrary. Excess autocorrelation is a problem across the board and the performance, as measured by the relative numerical efficiency, is poor with the idiosyncratic error variance as the only exception.

Contrasting, the improvement offered by the 2-block algorithm is remarkable. In Figure 2, the convergence appears to be clean, smooth, and fast, in particular for the intercept. What little autocorrelation is present dies out quickly. The estimate of the numerical standard error is robust, all methods yielding similar results for the various truncation and batch sizes tried. The Geweke statistics for the random effect variance are stable and insignificant.

Discarding 2500 iterations again, the results in Table 2 show how the fortunes are completely reversed, sampling being efficient across the board. For the variance components improvement is concentrated to the random effect variance for obvious reasons. In terms of computational speed, the improvement is bought at a cost of an approximate 78% increase in execution time.

The results being so good for the 2-block algorithm leaves little, if any, room for improvement. Experimenting with various grid construction techniques and updating principles, the results are robust with respect to the choice of method and the main difference is in computational speed. Figure 3 and Table 3 illustrate the results for a simple equidistant grid on a truncated support using rejection sampling where the truncation is decided based on the output of a very short trial run using data augmentation and the 2-block algorithm. In the final run, the output of the Markov chain is monitored to ensure that the selected truncation limits do not affect the results in any obvious way.

Using 51 nodes and refreshing the approximation in every iteration, the results are at least equally good as those using the improved 2-block algorithm, depending primarily on what quantity is monitored (θ_μ or σ_μ^2). The average rejection rate is just below 0.3, meaning that on average 1.4 draws are needed to generate an accepted sample.

Convergence seems to be immediate and there is very little autocorrelation in the chain as expected. Because the regression parameters are sampled the same way, that is after marginalization of the random effects, the performance in that part of the model is the same. The numerical standard error is not sensitive to choice of method, though the spectral estimate is fragile when the truncation lag increases beyond a certain point. The Geweke diagnostics are insignificant so there are no obvious signs of trouble.

Turning to the table, the slight improvement in numerical efficiency for the random effect variance is not due to more efficient sampling of the effect parameter. Being computed as a function of the effect parameter and the idiosyncratic error variance the source of the apparent improvement is the more efficient sampling of the latter. On the downside, the execution time increased further by 8%.

The last experiment in this segment uses direct sampling truncated gamma proposals and Metropolis-Hastings updates. Cutting the execution time with more than 50% compared to the 2-block algorithm, this sampler clearly illustrated the cost of maintaining and adapting a grid in every iteration. As illustrated in Figure 4 and Table 4, the improvement is achieved with only the slightest loss of efficiency. In retrospect, this loss is of minor if any importance, especially when accounting also for the extra effort needed to fine tune the strategies that make the grid techniques work well.

Studying all the tables, it is clear how the choice of method has very little effect on the estimates of the marginal posterior distributions. Kernel density estimates based on the output of the various samplers tested indicate how the slight differences present are mainly located in the right tail.

Figure 1 Inference in the Grunfeld panel using data augmentation and the standard 3-block algorithm. Select Markov chain properties for the random effect variance σ_μ^2 and the intercept β_0 .

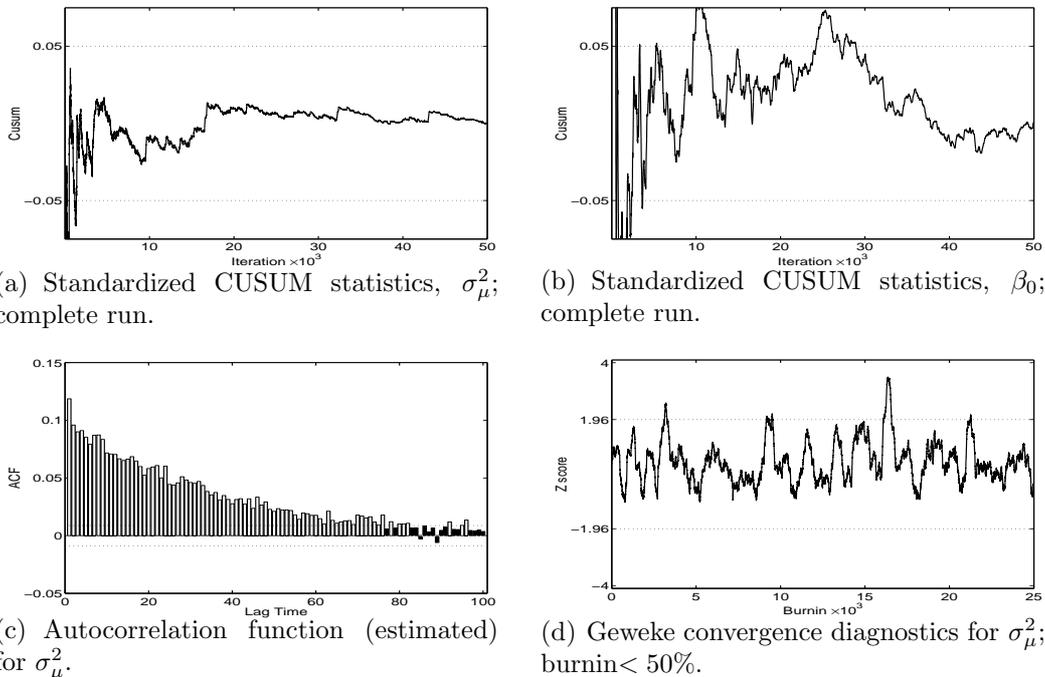


Table 1 Posterior simulation results for the Grunfeld panel; data augmentation and the simple 3-block algorithm.

	Mean	S.d	Median	Mode ^b	ρ_1	SPECTRAL ^a		BATCH	
						NSE	RNE	NSE	RNE
β_0	-60.112	28.967	-60.016	-60.249	0.97	1.155	79.46	0.977	54.08
β_1	0.109	0.010	0.109	0.110	0.90	0.217 ^c	22.53	0.211 ^c	20.28
β_2	0.308	0.017	0.308	0.308	0.31	0.125 ^c	2.66	0.127 ^c	2.61
σ_ν^2	2781.211	288.025	2761.905	2721.409	0.05	1.355	1.11	1.337	1.02
σ_μ^2	7306.572	4017.653	6338.053	5065.369	0.12	48.732	7.36	44.981	5.95

^a Calculation based on estimated autocorrelation time.

^b Estimated from the kernel density estimate.

^c All values $\times 10^3$.

Figure 2 Inference in the Grunfeld panel using data augmentation and the improved 2-block algorithm. Select Markov chain properties for the random effect variance σ_μ^2 and the intercept β_0 .

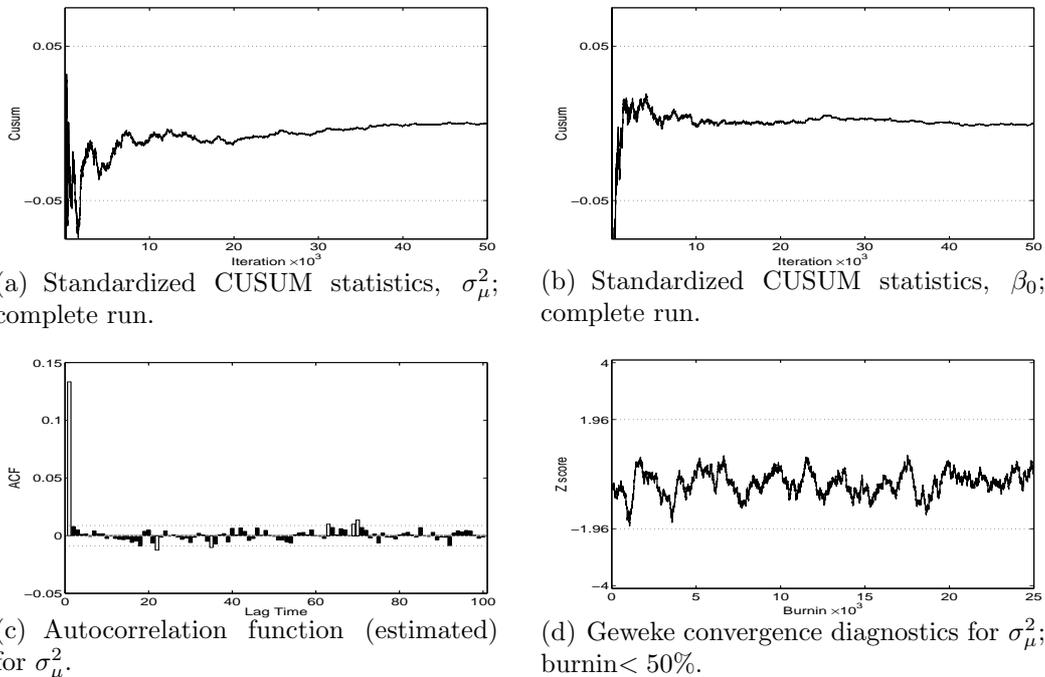


Table 2 Posterior simulation results for the Grunfeld panel; data augmentation and the improved 2-block algorithm.

	Mean	S.d	Median	Mode ^b	ρ_1	SPECTRAL ^a		BATCH	
						NSE	RNE	NSE	RNE
β_0	-60.534	28.869	-60.125	-58.267	0.01	0.133	1.06	0.128	0.93
β_1	0.109	0.010	0.109	0.109	0.01	0.046 ^c	1.00	0.048 ^c	1.02
β_2	0.308	0.017	0.308	0.309	0.01	0.074 ^c	0.94	0.078 ^c	0.97
σ_ν^2	2783.752	290.798	2764.349	2725.108	0.06	1.371	1.11	1.385	1.08
σ_μ^2	7319.079	4015.353	6355.243	5106.032	0.13	20.183	1.26	20.426	1.23

^a Calculation based on estimated autocorrelation time.

^b Estimated from the kernel density estimate.

^c All values $\times 10^3$.

Figure 3 Inference in the Grunfeld panel using direct sampling with acceptance rejection. Select Markov chain properties for the random effect variance σ_μ^2 and the intercept β_0 .

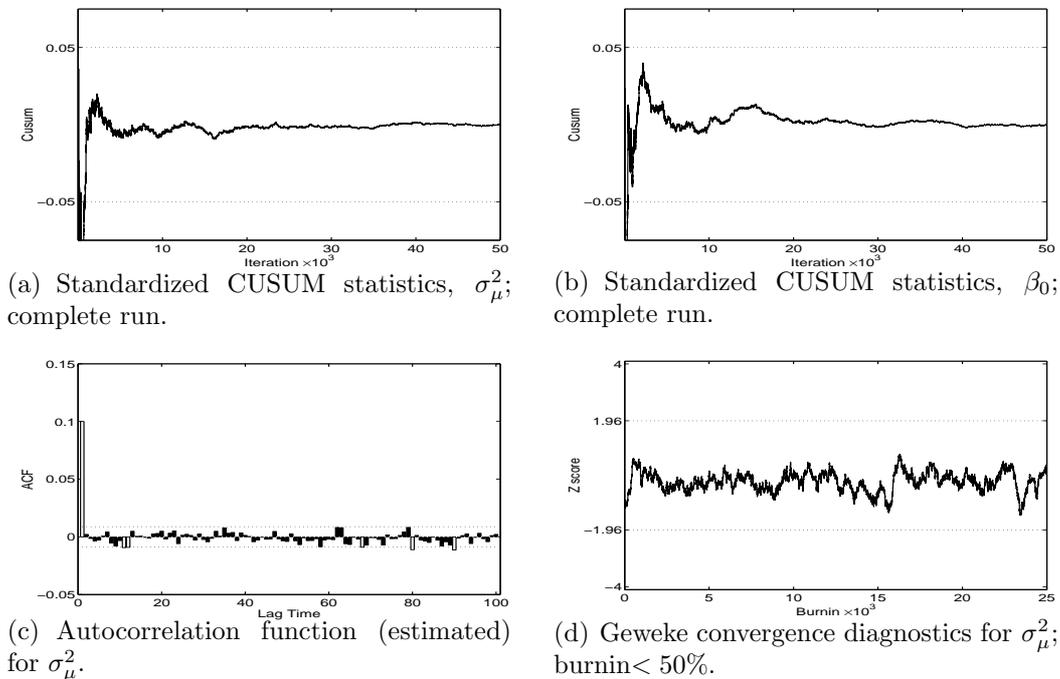


Table 3 Posterior simulation results for the Grunfeld panel; direct simulation using an equidistant grid on the truncated support and rejection resampling.

	Mean	S.d	Median	Mode ^b	ρ_1	SPECTRAL ^a		BATCH	
						NSE	RNE	NSE	RNE
β_0	-60.507	28.737	-60.149	-58.536	0.00	0.130	1.02	0.127	0.98
β_1	0.109	0.010	0.109	0.109	-0.00	0.046 ^c	0.98	0.049 ^c	1.12
β_2	0.308	0.017	0.308	0.309	-0.00	0.075 ^c	0.95	0.074 ^c	0.92
σ_ν^2	2812.870	293.840	2791.584	2746.111	0.07	1.442	1.20	1.365	1.08
σ_μ^2	7315.252	4017.892	6326.377	5123.122	0.10	18.843	1.10	19.157	1.14
θ_μ	0.023	0.010	0.022	0.018	0.13	0.052 ^c	1.26	0.051 ^c	1.21

^a Calculation based on estimated autocorrelation time.

^b Estimated from the kernel density estimate.

^c All values $\times 10^3$.

Figure 4 Inference in the Grunfeld panel using direct sampling with Metropolis-Hastings and truncated gamma proposals. Select Markov chain properties for the random effect variance σ_μ^2 and the intercept β_0 .

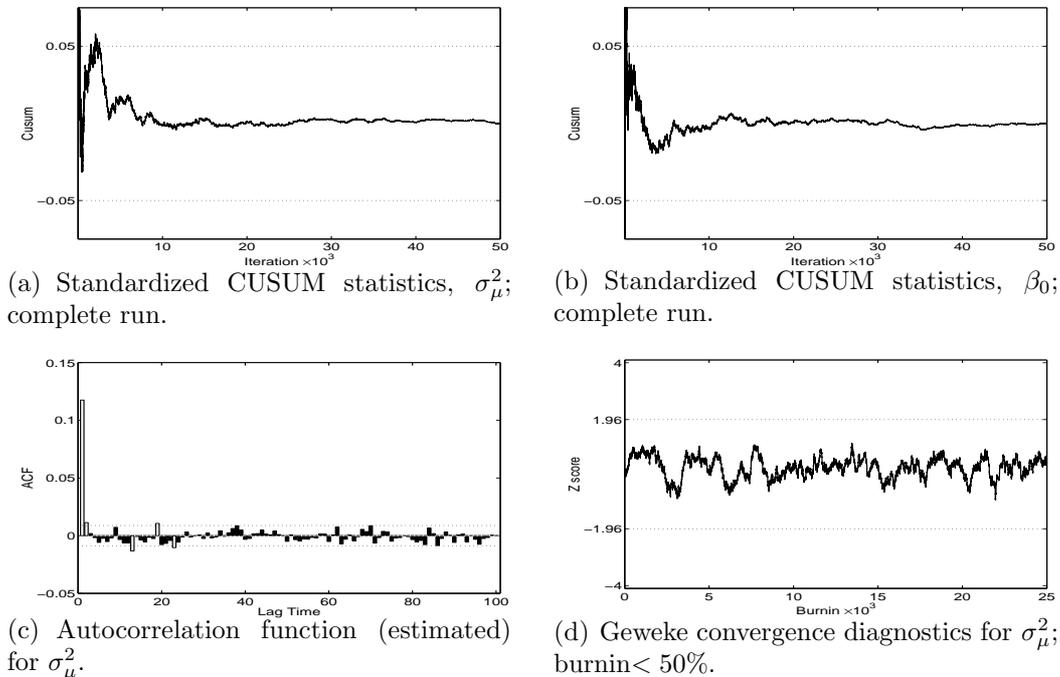


Table 4 Posterior simulation results for the Grunfeld panel; direct sampling using Metropolis-Hastings with truncated gamma proposals.

	Mean	S.d	Median	Mode ^b	ρ_1	SPECTRAL ^a		BATCH	
						NSE	RNE	NSE	RNE
β_0	-60.695	28.810	-60.265	-58.277	0.01	0.129	1.00	0.128	0.99
β_1	0.109	0.010	0.109	0.109	0.00	0.046 ^c	0.98	0.046 ^c	0.97
β_2	0.308	0.017	0.308	0.309	0.00	0.077 ^c	1.00	0.076 ^c	0.98
σ_ν^2	2813.651	292.700	2794.763	2725.891	0.07	1.395	1.14	1.373	1.10
σ_μ^2	7306.661	3979.154	6343.952	4989.580	0.12	19.958	1.26	19.556	1.21
θ_μ	0.023	0.010	0.022	0.018	0.16	0.053 ^c	1.35	0.052 ^c	1.30

^a Calculation based on estimated autocorrelation time.

^b Estimated from the kernel density estimate.

^c All values $\times 10^3$.

4.3 Model Selection

Implementing model selection with the reversible jump, for the moves associated with model space transitions computing an acceptance probability is necessary. Sampling the next state, when appropriate the model index is changed and the effect is either added or deleted.

The acceptance probability is just a product of a likelihood, prior, and proposal ratio where the ratios are evaluated based on the current and some candidate value for the effect in question. Indexing the current and proposed states with i and $i+1$, when adding an effect the product of prior and proposal ratios will always be on the form

$$\frac{w_{i+1}}{w_i} \pi(\theta_{i+1} | \sigma_\nu^2) \frac{p_{i+1,1}}{p_{i,i+1} \cdot q(\theta_{i+1})},$$

where $q(\theta_{i+1})$ the constructed approximation of the full conditional posterior and p_{st} the probability of proposing a move that attempts to move the chain from a model s to a model t . What does separate the cases is the extent to which the likelihood ratio simplifies. For example, when adding a unit random effect in the absence of any time effect the relevant ratio simplifies to

$$\theta_\mu^{N/2} \exp \left\{ -0.5\sigma^{-2} (c_1 + c_3) (\theta_\mu - 1) \right\},$$

but when the time effect is present the correct ratio is

$$\left(\frac{\theta_\mu^N}{\theta_\mu + \theta_\lambda - \theta_\mu \theta_\lambda} \right)^{1/2} \exp \left\{ -0.5\sigma^{-2} \left[c_1 + c_3 \frac{\theta_\lambda^2}{(\theta_\lambda + \theta_\mu - \theta_\mu \theta_\lambda)} \right] (\theta_\mu - 1) \right\}.$$

Similarly, when adding a time effect in the absence of a unit effect

$$\theta_\lambda^{T/2} \exp \left\{ -0.5\sigma^{-2} (c_2 + c_3) (\theta_\lambda - 1) \right\},$$

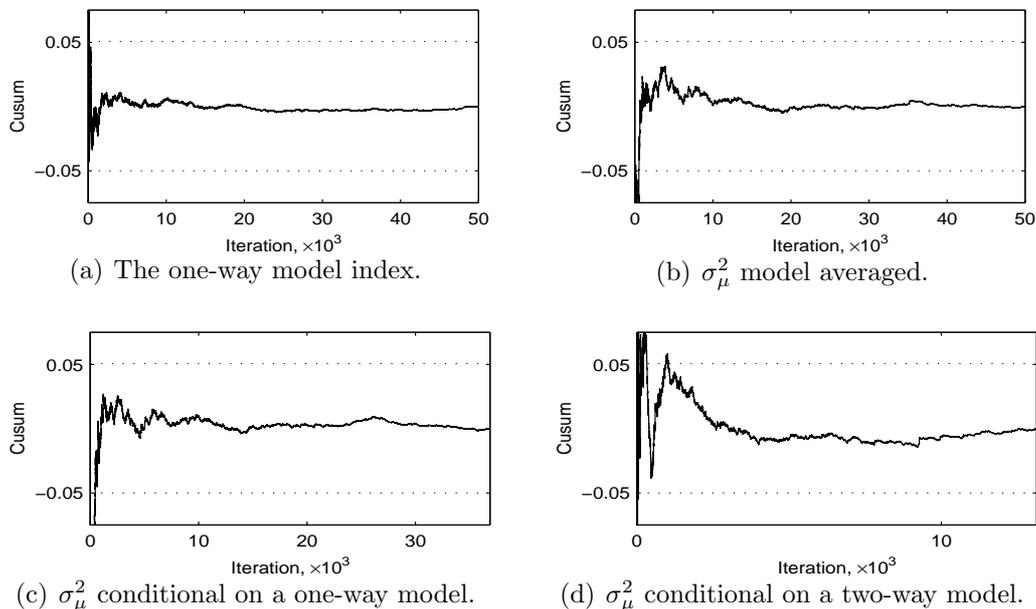
but when the unit effect is present

$$\left(\frac{\theta_\lambda^T}{\theta_\mu + \theta_\lambda - \theta_\mu \theta_\lambda} \right)^{1/2} \exp \left\{ -0.5\sigma^{-2} \left[c_2 + c_3 \frac{\theta_\mu^2}{(\theta_\lambda + \theta_\mu - \theta_\mu \theta_\lambda)} \right] (\theta_\lambda - 1) \right\}.$$

These results rely on the Q_i being symmetric, idempotent, orthogonal in pairs and, in particular, summing to the identity matrix. For the inverse moves that attempts to delete the effect, the acceptance probability is just the inverse where the current value of the relevant effect is treated as if it is sampled from the constructed proposal distribution.

The Grunfeld panel is also used to illustrate effect selection. Setting all weights equal, the sampler is run 50000 iterations and updating is performed using the

Figure 5 Standardized CUSUM plots for the posterior probability of a simple one-way model and various sections of the random unit effects variance chain.



griddy Gibbs sampler. To allow a comparison between results for the random unit effect model and those obtained in Section 4.2 the same prior parameters are used. The time effect parameters are set equal to the unit effect parameters.

One problem that occurs with model selection is that it is not longer clear what and how to monitor convergence. Conditioning on the model index, which is what we typically want to do, quantities such as estimated autocorrelations may be misleading as a gauge of performance. This as a sampler mixing over the model indexes will act as an indirect thinning process. One natural parameter to monitor is the model index. Being a quantity known to be hard to pin down with precision, apparently smoothly converging posterior model probabilities, though not definitive evidence of anything, have some value.

In Figure 5, CUSUM plots for the model index, the random effect variance conditional on model index and the random effect model averaged are illustrated. The idea is that if the chains conditional on the index as well as the model index probabilities behave well then so should also the model averaged chain. This seems to be the case, possibly with the exception of the output conditional on a two-way random effects specification. However, considering that it is based on the fewest observation in the lot, this is perhaps to be expected. That the model index converges smoothly and reasonably fast is particularly gratifying.

Table 5 Posterior simulation results for the Grunfeld panel; direct sampling using the griddy Gibbs sampler and featuring model selection.

$P(\theta_\mu < 1 \wedge \theta_\lambda = 1) = 0.74$									
	Mean	S.d	Median	Mode ^b	ρ_1	SPECTRAL ^a		BATCH	
						NSE	RNE	NSE	RNE
β_0	-60.668	28.861	-60.206	-59.421	0.01	0.131	1.03	0.151	1.02
β_1	0.109	0.010	0.109	0.109	-0.01	0.045 ^c	0.97	0.053 ^c	0.97
β_2	0.308	0.017	0.308	0.309	0.00	0.078 ^c	1.02	0.091 ^c	1.03
σ_ν^2	2806.512	293.001	2786.979	2783.182	0.04	1.360	1.08	1.553	1.04
σ_μ^2	7308.590	4072.257	6314.829	5149.746	0.06	19.272	1.12	22.534	1.13

$P(\theta_\mu < 1 \wedge \theta_\lambda < 1) = 0.26$									
	Mean	S.d	Median	Mode ^b	ρ_1	SPECTRAL ^a		BATCH	
						NSE	RNE	NSE	RNE
β_0	-63.216	29.279	-62.654	-60.189	-0.00	0.131	1.00	0.257	1.01
β_1	0.110	0.011	0.110	0.109	-0.01	0.049 ^c	1.05	0.093 ^c	0.98
β_2	0.314	0.018	0.314	0.312	-0.00	0.084 ^c	1.04	0.165 ^c	1.00
σ_ν^2	2749.408	292.844	2730.220	2703.463	0.02	1.334	1.04	2.596	1.03
σ_μ^2	7431.542	4555.323	6438.231	4986.550	0.00	20.372	1.00	39.497	0.98
σ_λ^2	88.978	51.582	77.000	48.069	0.04	0.245	1.12	0.468	1.07

^a Calculation based on estimated autocorrelation time.

^b Estimated from the kernel density estimate.

^c All values $\times 10^3$.

Studying the transition probabilities, the sampler seems to be moving between the one- and two-way specification at a healthy rate. While the survival probability of a two-way model is low this is countered by a sizeable transition probability from the one- to the two-way model. With a posterior probability of 74%, the evidence for the one-way specification is strong without being overwhelming.

In Table 5, the results conditional on the one-way model are consistent with the results obtained in Section 4.2. The transition between specifications hardly affects the parameters in the mean model and results only in slight shifts in the variance components. As calculated, the relative numerical efficiencies cannot be interpreted the usual way. Arguably, their near perfect values reflect that the chain moves between models in an unpredictable way. In general, the results are in line with some of the classical estimates reported in Baltagi (1995). Model averaged the results are similar to those obtained using iterated maximum likelihood to estimate a two-way model.

5 Final Remarks

This paper presents a direct sampling method for inference in panel regression models with parametric covariance structures. Developing the tools necessary for posterior simulation, the illustrated examples of direct sampling with and without effect selection offered many valuable insights. When practical, direct sampling is competitive, works well, is reasonably fast and can be relied on to produce the required inference. It offers the opportunity to re-examine models in search of exciting alternatives outside the mainstream. A good example of that is the Metropolis-Hastings with truncated gamma proposals tested for the one-way model which did something so rare as to combine the best of two worlds; the speed of simple data augmentation with the efficiency of the 2-block algorithm. Avoiding data augmentation, direct sampling can be quite economical, in particular when the panel size grows. It also offers a greater sense of control, making for instance model selection simpler.

There are several technical issues relating to the posterior simulation that influence the performance, the main being how to sample the variance components. In particular, how to construct reliable grids in an efficient manner are important questions. There is ample room for improvement on both counts. Experimentation showed, as expected, that adaptive grids enhanced the performance in terms of numerical efficiency at the cost of fewer iterations per time unit. However, placing nodes in a clever way is time consuming and after some point any clever procedure may end up defeating itself. Finding ways to combine quick and cheap strategies with slow, more elaborate but better methods would lead to quicker samplers. Also, finding means to generate and administrate multidimensional grids efficiently offers an interesting challenge with potentially high rewards. Other technical issues include the efficient parametrization of the models. How it is done should depend on what kind of quantities we are willing to make statements about a priori. Still, the methods outlined and the tools developed do not depend on the prior, at least not as long as the necessary structure is preserved.

While the theory for implementing model selection is standard, the application of the method revealed some practical problems. In particular, the priors on the variance components play an important role and being vague, but proper, is difficult. The mindless application of the standard more or less automatic choices, can lead to strange inference. While falling outside the scope of this paper, finding reasonable default priors should be a priority. Another interesting but more general question is when, if ever, posterior simulation should perhaps be abandoned in favor of classical numerical integration techniques. This because, in the end, doing things only because we can is usually a bad idea.

References

- Baltagi, B. H. (1995), *Econometric Analysis of Panel Data*, Wiley.
- Box, G. E. P. and Tiao, G. C. (1973), *Bayesian Inference in Statistical Analysis*, John Wiley & sons, inc, Chichester.
- Chib, S. and Carlin, B. P. (1999), ‘On MCMC sampling in hierarchical longitudinal models’, *Statistics and Computing* **9**, 17–26.
- Gelfand, A. E., Sahu, S. K. and Carlin, B. P. (1995), ‘Efficient parametrizations for generalized linear mixed models’, *Biometrika* **82**, 479–488.
- Gelfand, A. E. and Smith, A. F. M. (1990), ‘Sampling-based approaches to calculating marginal densities’, *Journal of the American Statistical Association* **85**, 398–409.
- Geweke, J. (1992), Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments, *in* J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds, ‘Bayesian Statistics 4’, Oxford University Press, Oxford, pp. 169–93.
- Gilks, W. R. and Roberts, G. O. (1996), Strategies for improving MCMC, *in* W. R. Gilks, S. Richardson and D. J. Spiegelhalter, eds, ‘Markov Chain Monte Carlo in Practice’, Chapman Hall, London, pp. 89–114.
- Green, P. J. (1995), ‘Reversible jump markov chain monte carlo computation and bayesian model determination’, *Biometrika* pp. 711–732.
- Grunfeld, Y. (1958), The Determinants of Corporate Investment, PhD thesis, University of Chicago.
- Hill, B. M. (1965), ‘Inference about variance components in the one-way model’, *Journal of the American Statistical Association* **60**, 806–825.
- Ritter, C. and Tanner, M. A. (1992), ‘Facilitating the gibbs sampler: The gibbs stopper and the griddy-gibbs sampler’, *Journal of the American Statistical Association* **87**, 861–868.
- Tanner, M. A. and Wong, W. (1987), ‘The calculation of posterior distributions by data augmentation’, *Journal of the American statistical Association* pp. 528–550.

- Tiao, G. C. and Tan, W. Y. (1965), 'Bayesian analysis of random-effects models in the analysis of variance. I: Posterior distribution of variance components', *Biometrika* **52**, 37–53.
- Vines, S. K., Gilks, W. R. and Wild, P. (1996), 'Fitting bayesian multiple random effects models', *Statistics and Computing* **6**, 337–346.
- Yu, B. and Mykland, P. (1998), 'Looking at Markov samplers through cusum path plots: A simple diagnostic idea', *Statistics and Computing* **8**(3), 275–286.