TI 2006-076/4
Tinbergen Institute Discussion Paper

On the Practice of Bayesian Inference in Basic Economic Time Series Models using Gibbs Sampling

Michiel D. de Pooter
Rene Segers
Herman K. van Dijk

Econometric Institute, Erasmus Universiteit Rotterdam, and Tinbergen Institute.

Tinbergen Institute
The Tinbergen Institute is the institute for economic research of the Erasmus Universiteit Rotterdam, Universiteit van Amsterdam, and Vrije Universiteit Amsterdam.

Tinbergen Institute Amsterdam
Roetersstraat 31
1018 WB Amsterdam
The Netherlands
Tel.: +31(0)20 551 3500
Fax: +31(0)20 551 3555

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900
Fax: +31(0)10 408 9031

Most TI discussion papers can be downloaded at http://www.tinbergen.nl.

On the Practice of Bayesian Inference in Basic Economic Time Series Models Using Gibbs Sampling*

Michiel D. de Pooter† Rene Segers Herman K. van Dijk

Econometric Institute and Tinbergen Institute
Erasmus University Rotterdam, The Netherlands

Tinbergen Institute Discussion Paper 2006-076/4

August 28, 2006

Abstract

Several lessons learned from a Bayesian analysis of basic economic time series models by means of the Gibbs sampling algorithm are presented. Models include the Cochrane-Orcutt model for serial correlation, the Koyck distributed lag model, the Unit Root model, the Instrumental Variables model and as Hierarchical Linear Mixed Models, the State-Space model and the Panel Data model. We discuss issues involved when drawing Bayesian inference on regression parameters and variance components, in particular when some parameter have substantial posterior probability near the boundary of the parameter region, and show that one should carefully scan the shape of the posterior density function. Analytical, graphical and empirical results are used along the way.

Keywords: Gibbs sampler, MCMC, serial correlation, non-stationarity, reduced rank models, state-space models, random effects panel data models.

JEL Classification Codes: C11, C15, C22, C23, C30

1 Introduction

As discussed by, for instance, Van Dijk (1999) and Hamilton (2006), the 'simulation revolution in Bayesian econometric inference' is to a large extent due to the advent of computers with ever-increasing computational power. This allows researchers to apply elaborate

*This paper is a substantial revision and extension of De Pooter *et al.* (2006). We are very grateful to participants of the 10th International Conference on Computing in Economics and Finance, Amsterdam, 2004, the 3rd World Conference on Computational Statistics & Data Analysis, Cyprus, 2005, and two anonymous referees for their helpful comments on earlier versions of the paper. All remaining errors are ours.

†Corresponding author. Tinbergen Institute, Erasmus University Rotterdam, P.O. Box 1738, NL-3000 DR Rotterdam, The Netherlands. Tel.: +31-10-4089142, fax: +31-10-4089031. *Email addresses*: depooter@few.eur.nl (M.D. de Pooter), rsegers@few.eur.nl (R. Segers), hkvandijk@few.eur.nl (H.K. van Dijk)

Bayesian simulation techniques for estimation in which extensive use is made of pseudo-random number generators. One of the most important methods is Gibbs sampling, developed by Geman and Geman (1984), Tanner and Wong (1987) and Gelfand and Smith (1990). This method has become a popular tool in econometrics for analyzing a wide variety of problems; see Chib and Greenberg (1996) and Geweke (1999). Judging from numerous articles in recent literature, Gibbs sampling is still gaining more and more momentum. Recent textbooks such as Bauwens *et al.* (1999), Koop (2003), Lancaster (2004), and Geweke (2005) discuss how Gibbs sampling is used in a wide range of econometric models, in particular in models with latent variables.

In the present paper we focus our attention on some basic models for economic time series and our aim is to investigate which lessons can be learned from applying Bayesian analysis to these models using Gibbs sampling, in particular, when some parameters have a substantial amount of posterior probability near or at the boundary of the parameter region. This feature may occur and is relevant in several economic time series. A practical example is a dynamic economic process that is possibly non-stationary, otherwise stated, substantial posterior probability of the dominant characteristic root of the process is near the boundary of unity. A second example is the case in time varying parameter models such as state space models when substantial posterior probability of the variance of the cyclical component is near zero and it is not clear how much of the variation in the economic time series is due to the cycle and how much due to noise. A third example is the presence of very weak instruments in an instrumental variable regression model.

For expository purposes two classes of canonical models are used. The first one is the linear regression model with first order serial correlation in the disturbances; the so-called Cochrane-Orcutt model; for details we refer to standard textbooks of econometrics, for instance, Heij *et al.* (2004). In this class of models the regression parameters are parameters of interest and the variances of the disturbances are nuisance parameters. We show that, for our purpose of a near-boundary analysis in the parameter region, other basic models for economic time series like the Koyck model for distributed lags and the unit root model are special cases of this first canonical model. The intrinsic econometric issues in such models may be summarized as follows. What is a plausible dynamic specification such that short and long term effects can be measured adequately (serial correlation model and distributed lag model)? Are financial markets efficient and do there exist stochastic trends in macro-economic series (unit root models)? We also treat the issue of endogeneity with weak instruments in an instrumental variable regression model for economic time series.

The second class of models deals with variance parameters as parameters of interest. We discuss how the class of linear hierarchical mixed models (HLMM) serves as a parent class for such time varying parameter models as state space models and as a parent class for panel data models. We are interested to investigate what happens when the density of one of the variance parameters is located near the zero-bound and/or when one faces the situation when there is not a sufficient number of components/groups in a panel.

Given the specification of the models one can derive their likelihoods and specify prior information. Our approach with respect to specifying prior information is to start with uniform priors on a large but bounded region. The use of such non-informative priors means that we concentrate on the information content in the likelihood function. We emphasize that, given our priors, the posterior densities may of may not exist and that it is very important to investigate their shape. As mentioned before, we are, in particular, interested in the behavior of the likelihood/posterior near and at the boundary of the parameter region.

We emphasize that for analysis we make use of an interplay of analytical techniques, by giving detailed derivations of conditional and marginal distributions, and Gibbs sampling and further note that graphics in the context of Bayesian analysis is becoming more and more important see, for example, Murrell (2005). In our analysis we therefore also place emphasis on presenting results in a graphical way. The technical level of the paper is like that of an introductory graduate econometrics course. Matrix notation is used in order to indicate the common, linear (sub)-structure of several models.

The results of our analysis may be labelled as 'lessons learned'. A summary of models used and lessons learned is presented in Section 5. The key lesson is to investigate the shape of the criterion function of the parameters of interest and classify this shape in two categories. As long as this shape is approximately elliptical and much probability mass is in the interior of the parameter region, then applying Gibbs sampling is straightforward and yields accurate results. When the criterion function has strong non-elliptical contours and substantial mass is at the boundary of the parameter region then warning signals for the researcher may need to be indicated. It depends on the specification of the model and the information in the data in which situation a researcher will find herself. In case of a regular shape one may use rather noninformative, for instance uniform, priors if one prefers to emphasize the information content of the likelihood. In the other cases uniform priors for the regression parameters are not suitable. We discuss some prior information that may regularize or smooth the shape of the criterion function, such as the class of information matrix or Jeffreys' priors, and we discuss an example of a training sample prior. For variance components models, however, uniform priors may be attractive when a researcher is interested in the probability mass near a zero variance; see also the recommendation by Gelman (2006).

This paper offers some directions on how to continue in nonstandard cases. In terms of possible further 'advice' one could think of a reparametrization of the model, the use of subjective informative priors, and for the use of predictive priors, see, for example, Geweke (2005). Several other approaches are also given in the literature such as the use of more flexible sampling methods to approximate the irregular shape of the posterior but all this is beyond the scope of this paper.

The topic of this paper should be of interest to Bayesians to see how Gibbs sampling functions in basic regression models for economic time series when the focus is on the information content of the likelihood. The topic should interest non-Bayesians who are very knowledgeable about basic econometric models and want to learn how the information in the likelihood function of such models is summarized according to Bayes' rule.

The contents of this paper is structured as follows. In Section 2 we briefly review the Gibbs sampler. Through a number of (artificial) examples we discuss several shapes of the criterion function that the researcher may encounter in econometric practice, in particular, irregular shapes near the boundary of the parameter region. In Section 3 we present our analysis by applying the Gibbs sampler to a number of canonical models such as the Cochrane-Orcutt serial correlation model, the Koyck Distributed Lag model, the Unit Root and Instrumental Variables models. Our focus in this section is primarily on drawing inference on regression parameters. Then in Section 4 we move on to studying variance components. One often used model there is the Hierarchical Linear Mixed Model (HLMM). As an application of HLMM we discuss how the Gibbs sampler performs in State-Space models and Random Effects Panel Data models. Section 5 presents a summary of models used and lessons learned.

2 Gibbs Sampling and Some Typical Shapes of the Criterion Function

In this section we briefly discuss the basic idea of the Gibbs sampling algorithm and illustrate its potential by means of several bivariate examples. We show particular shapes of posterior densities, using the Gelman-Meng model, that one may encounter in econometric practice.

2.1 Gibbs Sampling

One may characterize the Gibbs sampling algorithm as an application of the *divide-and-conquer* principle[1]. First, a z-dimensional vector $\boldsymbol{\theta}$ is divided into into m components $\theta_1, \theta_2, ..., \theta_m$, where $m \leq z$. Second, for many posterior distributions which are intractable in terms of simulation the lower-dimensional *conditional* distributions turn out to be remarkably simple and tractable. The Gibbs sampler exploits this notion, as it precisely samples from these conditional distributions. Its usefulness is, for example, demonstrated by Chib and Greenberg (1996), and Smith and Roberts (1993).

Since Gibbs sampling is based on the characterization of the joint posterior distribution by means of the complete set of conditional distributions, it follows that a requirement for application of the Gibbs sampler is that the latter distributions, described by the densities

$$p(\theta_i|\boldsymbol{\theta}_{-i}), \qquad \text{for } i = 1, ..., m, \tag{1}$$

where $\boldsymbol{\theta}_{-i}$ denotes the parameter vector $\boldsymbol{\theta}$ without the i^{th} component, can all be sampled from. The Gibbs sampling algorithm generates a sequence

$$(\theta_1^{(0)}, ..., \theta_m^{(0)}), (\theta_1^{(1)}, ..., \theta_m^{(1)}), ..., (\theta_1^{(J)}, ..., \theta_m^{(J)}) \tag{2}$$

following a process such that $\theta_i^{(j)}$ is obtained from $p(\theta_i|\boldsymbol{\theta}_{-i}^{(j-1)})$. Thus, $\theta_i^{(j)}$ is obtained conditional on the most recent values of the other components. The values $(\theta_1^{(0)}, ..., \theta_m^{(0)})$ initialize the Gibbs sequence and should be given. We may summarize the Gibbs sampling algorithm as follows

1: Specify starting values $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \ldots, \theta_m^{(0)})$ and set $j = 0$.

2: Generate:
$\theta_1^{(j+1)}$ from $p(\theta_1|\theta_2^{(j)}, \ldots, \theta_m^{(j)})$
$\theta_2^{(j+1)}$ from $p(\theta_2|\theta_1^{(j+1)}, \theta_3^{(j)}, \ldots, \theta_m^{(j)})$
$\theta_3^{(j+1)}$ from $p(\theta_3|\theta_1^{(j+1)}, \theta_2^{(j+1)}, \theta_4^{(j)}, \ldots, \theta_m^{(j)})$
$\vdots$
$\theta_m^{(j+1)}$ from $p(\theta_m|\theta_1^{(j+1)}, \ldots, \theta_{m-1}^{(j+1)})$

3: If $j < J$, set $j = j + 1$, and go back to Step **2**.

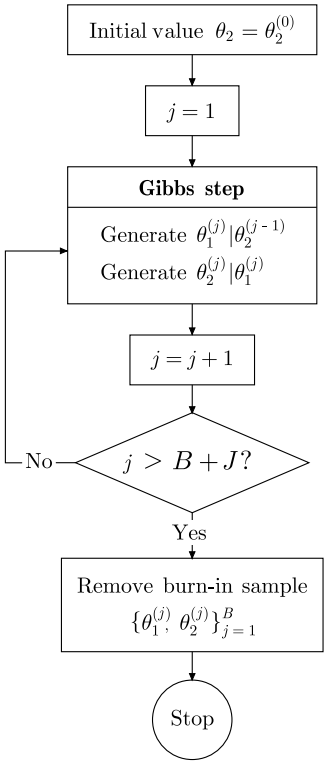
The above algorithm yields J realizations $\boldsymbol{\theta}^{(j)} = (\theta_1^{(j)}, ..., \theta_m^{(j)})$, for $j = 1, 2, ..., J$, from a Markov chain, converging to the target distribution. We will refer to Step **2**. of the algorithm as 'the Gibbs step' and for each of the models that we discuss in the subsequent

[1] We are necessarily brief in our explanation of the Gibbs sampler. See Casella and George (1992), or Hoogerheide *et al.* (2006b), among others, for a more elaborate discussion.

sections we will always indicate what the Gibbs step looks like. Note that the components of $\boldsymbol{\theta}$ do not necessarily need to be one-dimensional. Generating draws for blocks of parameters instead in which case some of the θ_i components denote a block of parameters is also possible.

The Gibbs algorithm is shown in the flow diagram in Figure 1 for a model with two parameters which are treated as separate components ($m = 2$). It is illustrated in Figure 2 where we show an example path of Gibbs sampled points, when the conditional densities of $\theta_1|\theta_2$ and $\theta_2|\theta_1$ are both standard Normal. The sample path is shown at different stages of the algorithm.

Figure 1: **Gibbs sampling: Flow diagram**

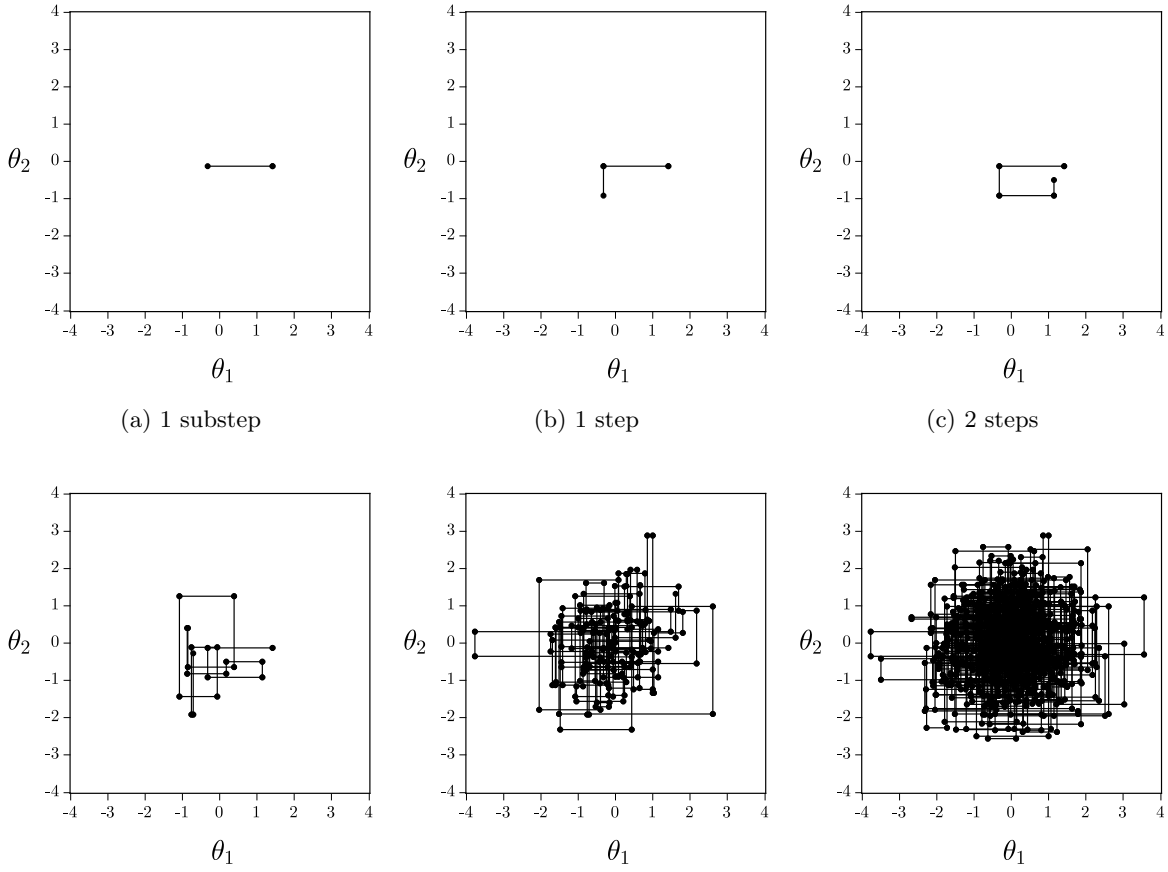


For large enough J the sequence of Gibbs draws, generated from the conditional distributions, is distributed according to the joint and marginal posterior distributions. A simple argument for a bivariate case is as follows. Suppose θ_i and $\boldsymbol{\theta}_{-i}$ have as *joint* posterior distribution with density $p(\theta_i, \boldsymbol{\theta}_{-i})$. Then $\boldsymbol{\theta}_{-i}$ has the *marginal* posterior distribution with density $p(\boldsymbol{\theta}_{-i})$. In Step **2**. of the Gibbs sampling algorithm, $\theta_i^{(j)}$ is drawn from $p(\theta_i|\boldsymbol{\theta}_{-i}^{(j-1)})$, which is the density of the conditional distribution of θ_i given $\boldsymbol{\theta}_{-i}^{(j-1)}$. The joint density of $\theta_i^{(j)}$ and $\boldsymbol{\theta}_{-i}^{(j-1)}$ is

$$p(\theta_i^{(j)}|\boldsymbol{\theta}_{-i}^{(j-1)})p(\boldsymbol{\theta}_{-i}^{(j-1)}) = p(\theta_i^{(j)}, \boldsymbol{\theta}_{-i}^{(j-1)}). \tag{3}$$

Therefore, $(\theta_i^{(j)}, \boldsymbol{\theta}_{-i}^{(j-1)})$ is distributed according to the joint posterior distribution. For a more detailed analysis on theoretical properties of the Gibbs sampler, we refer to Geweke (1999), Tierney (1994) and Smith and Roberts (1993).

Figure 2: **Gibbs sampling: Example steps**



(a) 1 substep (b) 1 step (c) 2 steps

(d) 10 steps (e) 100 steps (f) 1000 steps

Notes: Panels (a) through (f) show subsequent steps of the Gibbs sampler using two conditional posterior densities, $p(\theta_1|\theta_2)$ and $p(\theta_2|\theta_1)$ that are both standard normal.

Because in practice it may takes some time for the Markov Chain to converge it is common to discard the first B draws, where typically $B << J$. These draws are referred to as the *burn-in draws*. Consequently, posterior results will be based only on the draws $\boldsymbol{\theta}^{(B+1)}, \ldots, \boldsymbol{\theta}^{(J)}$ of the generated chain. Furthermore, the sequence of draws sometimes does display some degree of autocorrelation. When autocorrelations are significant up to the $h - 1^{\text{th}}$ lag, one should consider using only every h^{th} draw and to discard the intermediate draws (h is known as the thinning value). An altogether different approach is to generate multiple Markov Chains instead of just one and then to use only the final draw from each sequence. This means the Gibbs algorithm has to be executed a large number of times. When opting for this approach the researcher does not have to worry about which values to choose for B and h. Although the drawback is that this method can be very computationally intensive, it can help prevent posterior results from being determined by a particular set of starting values chosen for $\boldsymbol{\theta}^{(0)}$. As we will see in the next section, randomizing over $\boldsymbol{\theta}^{(0)}$ can a worthwhile endeavor when the likelihood displays signs of multimodality.

2.2 The Gelman-Meng Example

To illustrate the workings of the Gibbs sampler we go through a number of examples which are based on the model in Gelman and Meng (1991). Suppose that we have a joint posterior density of θ_1, θ_2, which has the following form

$$p(\theta_1, \theta_2) \propto \exp\Big[-\tfrac{1}{2}[a\theta_1^2\theta_2^2 + \theta_1^2 + \theta_2^2 - 2b\theta_1\theta_2 - 2c_1\theta_1 - 2c_2\theta_2] \Big] \tag{4}$$

where a, b, c_1 and c_2 are constants under the restrictions that $a \geq 0$ and if $a = 0$ then $|b| < 1$ [2]. This class of bivariate distributions is discussed in Gelman and Meng (1991) and has as feature that the random variables θ_1 and θ_2 are conditionally Normally distributed. In fact, the conditional densities $p(\theta_1|\theta_2)$ and $p(\theta_2|\theta_1)$ can be picked off from (4) and recognized as Normal densities with the following parameters

$$p(\theta_1|\theta_2, a, b, c_1, c_2) \sim \mathcal{N}\left(\frac{b\theta_2 + c_1}{a\theta_2^2 + 1}, \frac{1}{a\theta_2^2 + 1} \right) \tag{5}$$

$$p(\theta_2|\theta_1, a, b, c_1, c_2) \sim \mathcal{N}\left(\frac{b\theta_1 + c_2}{a\theta_1^2 + 1}, \frac{1}{a\theta_1^2 + 1} \right) \tag{6}$$

Note that, typically, the joint density of (θ_1, θ_2) is not Normal. By choosing different parameter configurations for a, b, c_1 and c_2 we can construct joint posterior densities of rather different shapes, while the conditional densities remain Normal densities. In the remainder of this section we consider three types of shapes and we apply the Gibbs sampler to each of these. Although the shapes are all in a way artificial since they are not based directly on a model and data, doing so may give us some insights into the ease of but also the possible difficulties with applying the Gibbs sampler before we move on to examining econometric models in subsequent sections.

(i) Bell-shape

The first parameter configuration that we consider is the following; $(a = b = c_1 = c_2 = 0)$ in which case the joint density is given by

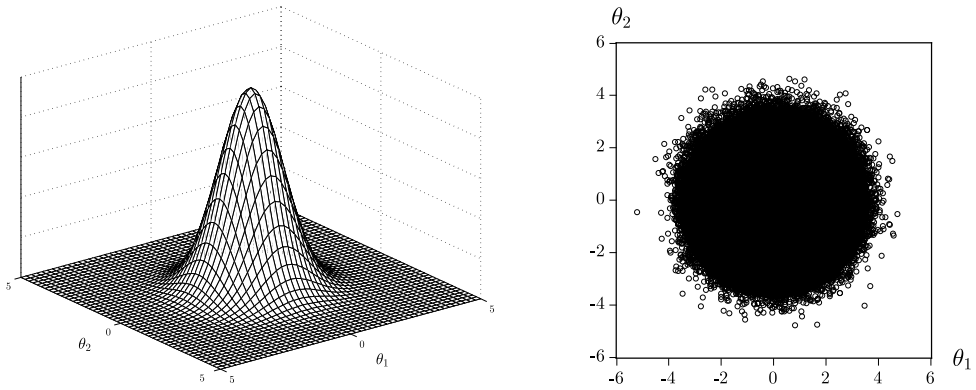
$$p(\theta_1, \theta_2) \propto \exp\big[-\tfrac{1}{2}[\theta_1^2 + \theta_2^2] \big] \tag{7}$$

Both the conditional densities and the joint density are standard Normal. The latter is depicted in Figure 3(a). Gibbs sampling simply comes down to obtaining draws by iteratively drawing from standard Normal densities[3]. A scatterplot of one million of such draws is shown in Figure 3(b). The estimated posterior means and variances are equal to 0 and 1 for both parameters. These are exactly the parameters of the marginal densities which, in this case, we know to be standard Normal. In fact, for the chosen parameter configuration, the conditional and marginal densities coincide since the conditional density for θ_1 does not depend on θ_2 and vice versa. In this particular example it is therefore obviously not necessary to use Gibbs sampling. However, the aim of this example is simply to illustrate the straightforward approach of the Gibbs sampler and its usefulness for obtaining posterior results.

[2] These restrictions are to insure that the joint density in (4) is integrable and therefore a proper probability density function.

[3] For all three examples in this section we used a burn-in period of $B = 10,000$ draws and we set the thinning value h equal to 10.

Figure 3: **Gelman-Meng: Bell-shape**



(a) Joint posterior density (b) Gibbs draws

Notes: Panel (a) shows the Gelman-Meng joint posterior density for θ_1 and θ_2 given in (4) for parameter values ($a = b = c_1 = c_2 = 0$) whereas panel (b) shows the scatterplot of one million draws from the Gibbs sampler.

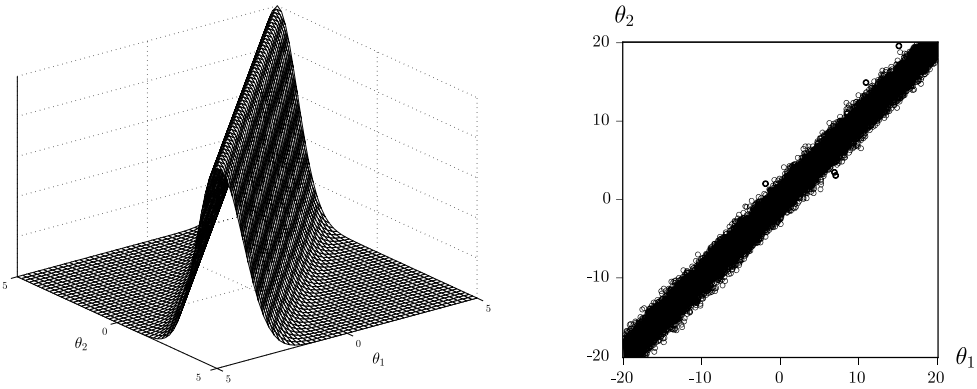
(ii) Ridges

The second parameter configuration that we examine is $(a = c_1 = c_2 = 0, b = 1)$[4]. The joint density is now given by

$$p(\theta_1, \theta_2) \propto \exp\left[-\tfrac{1}{2}[(\theta_1 - \theta_2)^2]\right] \tag{8}$$

It is apparent from Figure 4(a) that this density is improper when $-\infty < \theta_i < \infty$, for $i = 1, 2$, since the ridge along the line $\theta_1 = \theta_2$ causes it to be non-integrable. However,

Figure 4: **Gelman-Meng: Ridges**



(a) Joint posterior density (b) Gibbs draws

Notes: Panel (a) shows the Gelman-Meng joint posterior density for θ_1 and θ_2 given in (4) for parameter values ($a = c_1 = c_2 = 0$ and $b = 1$) whereas panel (b) shows the scatterplot of one million draws from the Gibbs sampler.

[4]Note that this parameter vector violates the earlier stated parameter restrictions.

on a bounded parameter region the posterior is proper. The scatterplot of Gibbs draws for this example in Figure 4(b) reveals the ridge $\theta_1 = \theta_2$. Ridges may occur in nearly nonidentified econometric models; see the next section for examples.

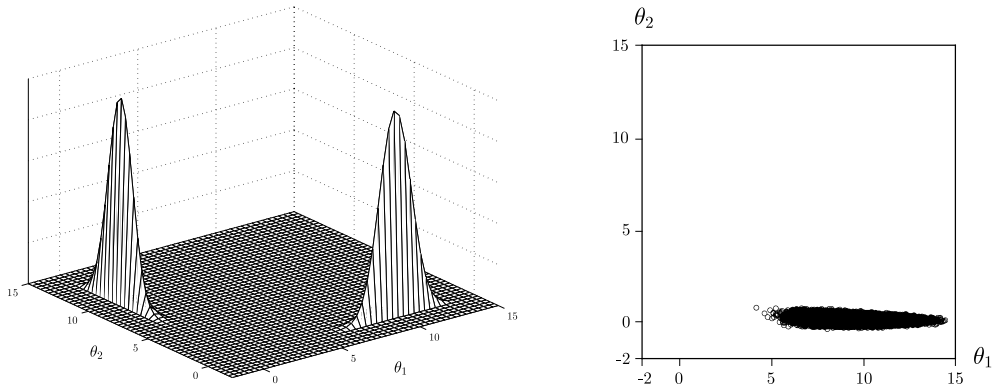
(iii) Bimodality

The third configuration we consider is $(a = 1, b = 0)$ and large, but not necessarily equal, values for c_1 and c_2 [5]. Here we select $c_1 = c_2 = 10$ which gives

$$p(\theta_1, \theta_2) \propto \exp\left[-\tfrac{1}{2}[\theta_1^2\theta_2^2 + \theta_1^2 + \theta_2^2 - 20\theta_1 - 20\theta_2] \right] \tag{9}$$

At first sight the scatterplot, shown in Figure 5(b), seems perfectly reasonable and posterior means and variances can easily be computed. However, when inspecting the joint density

Figure 5: **Gelman-Meng: Bimodality**



(a) Joint posterior density (b) Gibbs draws

Notes: Panel (a) shows the Gelman-Meng joint posterior density for θ_1 and θ_2 given in (4) with parameter values $(a = 1, b = 0$ and $c_1 = c_2 = 10)$ whereas panel (b) shows the scatterplot of one million draws from the Gibbs sampler.

as depicted in Figure 5(a) we see right away that the Gibbs sampler has only sampled from one mode of $p(\theta_1, \theta_2)$ but not from the other. Apparently it tends to get stuck in one of the two modes[6]. This is because the modes are too far apart with an insufficient amount of probability mass in between the two modes for the sampler to regularly jump from one to the other. Admittedly, increasing the number of draws substantially will eventually lead to a switch. However, one cannot be certain when this will happen. The scatterplot shows that with a single run, one million draws is already an insufficient number to witness such a switch. Therefore, the Gibbs output only provides the researcher with information on a subset of the full domain of $p(\theta_1, \theta_2)$ and posterior results are thus incorrect. One option to try and at least signal the bimodality of the likelihood is to execute the Gibbs Sampler several times with widely dispersed intial values. Although this example is a rather extreme case, it should be clear that multi-modality can result in very slow converge for

[5] See also Hoogerheide *et al.* (2006a) for a further analysis of bimodality.

[6] Which of the two modes the Gibbs sampler gets stuck in depends on the initial values $(\theta_1^{(0)}, \theta_2^{(0)})$.

the Gibbs sampler. Multimodality may occur in reduced rank models when one is close to the boundary of the parameter region.

Summarizing, the above examples of a bell-shaped, a ridge-shaped, and a bimodal-shaped density, indicate that it is essential to scrutinize a proposed model and the shape of its criterion function before moving on to drawing posterior inference on its parameters through the Gibbs sampler. In the remainder of this paper this will be our main focal point for econometric models.

3 Gibbs Sampling Within Canonical Econometric Models

We now begin our analysis of Bayesian inference by means of the Gibbs sampler using typical workhorse models of econometric practice. First we apply the Gibbs sampler to the basic linear regression model. After fixing notation and showing how straightforward it can be to apply the Gibbs sampler in order to obtain posterior results, we extend the basic model to the Cochrane-Orcutt model which allows for serial correlation in the residuals. We show that the Cochrane-Orcutt can be considered to be a template model for two famous time-series models: the Distributed Lag model and the Unit Root model. Moving from modeling a single univariate dependent variable to modeling several dependent variables at the same time we end this section with a (somewhat concise) examination of multivariate models in which some of the dependent variables may be endogenous. Throughout this section, our primary focus will be on drawing inference on the *regression* parameters in the models. In Section 4 we shift our focus to considering variance components.

3.1 Basic Linear Regression Model

We start our analysis by considering the basic regression model. This linear model attempts to explain the variance of a dependent variable y_t through a set of explanatory variables, as summarized in the $(1 \times K)$ (row-)vector x_t where K is the number of variables in x_t (including a constant):

$$y_t = x_t\beta + \varepsilon_t, \qquad t = 1, ..., T, \qquad \text{with} \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2) \tag{10}$$

To goal is to draw inference on the $(K \times 1)$ vector of regression parameters $\beta = (\beta_1 \ \beta_2 \dots \beta_k)'$ and the scalar variance parameter σ_ε^2. In matrix notation, this model is given by

$$y = X\beta + \varepsilon \qquad \text{with} \qquad \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2 \mathrm{I}_T) \tag{11}$$

where y denotes the vector of T time-series observations or cross-sectional observations on the dependent variable, $y = (y_1 \ y_2 \dots y_T)'$. $X = (x_1' \ x_2' \dots x_T')'$ denotes the matrix of observations on the explanatory variables and I_T is an $(T \times T)$ identity matrix.

For consistency we will always use matrix notation when we derive joint, conditional or marginal densities. Throughout this study we use $\boldsymbol{\theta}$ to indicate the vector of model parameters. In this instance $\boldsymbol{\theta}$ is given by $\boldsymbol{\theta} = (\beta, \sigma_\varepsilon^2)$. Furthermore, unless mentioned otherwise, we will denote individuals or groups with the index i $(i = 1, \dots, N)$, time-series observations with index t $(t = 1, \dots, T)$, exogenous variables (as well as their corresponding parameters in β) with index k $(k = 1, \dots, K)$ and draws from the Gibbs sampler with index j $(j = 1, \dots, J)$.

Gibbs Sampling

The likelihood for the basic regression model is given by

$$p(y|X,\beta,\sigma_\varepsilon^2) = (2\pi\sigma_\varepsilon^2)^{-\frac{T}{2}} \exp\left[-\frac{1}{2\sigma_\varepsilon^2}(y-X\beta)'(y-X\beta)\right] \tag{12}$$

Combining the likelihood with a noninformative or uniform prior[7]

$$p(\beta,\sigma_\varepsilon^2) \propto (\sigma_\varepsilon^2)^{-1} \tag{13}$$

yields the joint posterior density

$$p(\beta,\sigma_\varepsilon^2|y,X) \propto (\sigma_\varepsilon^2)^{-\frac{(T+2)}{2}} \exp\left[-\frac{1}{2\sigma_\varepsilon^2}(y-X\beta)'(y-X\beta)\right] \tag{14}$$

Combining the likelihood with informative (conjugate) priors is also possible of course. If one has prior information it is strictly advisable if not necessary to include this in the analysis (see the discussions in Geweke, 2005 and Lancaster, 2004). Specifying conjugate priors is, however, not an easy task especially when one is faced with a higher dimensional parameter region. Since our main question in this paper is concerned with what we can learn about the model parameters through the data likelihood we will focus primarily on non-informative priors. However, as we will see, doing so can result in the posterior density being improper. In that case one needs to go back to the drawing board and carefully specify appropriate informative prior densities to ensure that the joint density is proper.

A useful result to facilitate the derivation of the conditional and marginal posterior densities is to rewrite (14) by completing the squares on β

$$p(\beta,\sigma_\varepsilon^2|y,X) \propto (\sigma_\varepsilon^2)^{-\frac{(T+2)}{2}} \exp\left[-\frac{1}{2\sigma_\varepsilon^2}\left[(y-X\hat{\beta})'(y-X\hat{\beta}) + (\beta-\hat{\beta})'X'X(\beta-\hat{\beta})\right]\right] \tag{15}$$

with $\hat{\beta} = [X'X]^{-1}X'y$.

The only part in between brackets in (15) relevant for determining the posterior density of β conditional on a value for σ_ε^2 is that which depends on β. The first part only consists of data and does therefore not enter the parameters of the conditional density for β. From the probability density functions given in Appendix A, we can recognize a multivariate Normal density for β, given a value of σ_ε^2, which has mean vector $m = \hat{\beta}$ and variance matrix $S = \sigma_\varepsilon^2[X'X]^{-1}$, see equation (A-4). Similarly, the conditional density for σ_ε^2, given a β, follows from (A-3) and is Inverted Gamma with location parameter $m = \frac{1}{2}(y-X\beta)'(y-X\beta)$ and $\nu = \frac{1}{2}T$ degrees of freedom. Summarizing, we have

$$p(\beta|y,X,\sigma_\varepsilon^2) \sim \mathcal{N}\left(\hat{\beta},\sigma_\varepsilon^2[X'X]^{-1}\right) \tag{16}$$

$$p(\sigma_\varepsilon^2|y,X,\beta) \sim \mathcal{IG}\left(\frac{1}{2}(y-X\beta)'(y-X\beta),\frac{1}{2}T\right) \tag{17}$$

Ultimately, we are interested in learning about the properties of the marginal densities for β and σ_ε^2. In this model it is straightforward to derive these which in turn makes Gibbs

[7] A non-informative prior for the regression parameters can simply be specified as $p(\beta) \propto 1$. For a variance parameter a uniform prior comes down to $p(\sigma^2) \propto (\sigma^2)^{-1}$ which follows from specifying a uniform prior for the *logarithm* of σ^2. See Box and Tiao (1973), Chapter 1 for more details.

sampling redundant here of course. However, since we can establish beforehand what the marginal densities look like, we can easily corroborate the posterior results from the Gibbs sampler which is exactly what we will do below. But in order to do so we first need to derive the marginal densities.

To derive the marginal density for β we need to integrate out σ_ε^2 from the joint density. For this we apply the Inverse Gamma integration step which consists of the following proportionality

$$\int_0^\infty \left(\sigma^2\right)^{-\frac{(T+2)}{2}} \exp\left[-\frac{a}{2\sigma^2}\right] \mathrm{d}\sigma^2 \propto a^{-\frac{1}{2}T} \tag{18}$$

and which can be derived from (A-3), see also for example Bauwens *et al.* (1999). Applying this result to (15) gives

$$\begin{aligned} p(\beta|y,X) &= \int_0^\infty p(\beta,\sigma_\varepsilon^2|y,X)\mathrm{d}\sigma_\varepsilon^2 \\ &\propto \left[(y-X\hat{\beta})'(y-X\hat{\beta}) + (\beta-\hat{\beta})'X'X(\beta-\hat{\beta})\right]^{-\frac{T}{2}} \\ &\propto \left[y'M_{_X}y + (\beta-\hat{\beta})'X'X(\beta-\hat{\beta})\right]^{-\frac{T}{2}} \end{aligned} \tag{19}$$

In the last line of (19) we introduce the transformation matrix $M_{_X}$ which is specified in its more general form as $M_{_A} = \mathrm{I}_{_T} - P_{_A}$ with $P_{_A}$ the projection matrix defined as $P_{_A} = A(A'A)^{-1}A'$ and A a general $(T \times K)$ matrix. We can now factorize (19) as follows

$$\begin{aligned} p(\beta|y,X) \propto \left[(T-K) + \frac{(\beta-\hat{\beta})X'X(\beta-\hat{\beta})}{y'M_{_X}y/(T-K)}\right]^{-\frac{(T-K)+K}{2}} \left[\frac{X'X}{y'M_{_X}y/(T-K)}\right]^{\frac{1}{2}} \\ \times \left[\frac{X'X}{y'M_{_X}y/(T-K)}\right]^{-\frac{1}{2}} \left[y'M_{_X}y\right]^{-\frac{T}{2}} \end{aligned} \tag{20}$$

The last two factors depend only on data and will thus be part of a normalizing constant. From the first two factors, however, we recognize a multivariate Student-t density function for β with parameters $m = \hat{\beta}$, $S = \frac{X'X}{y'M_{_X}y/(T-K)}$ and $\nu = T-K$, see (A-5). The marginal density for β is therefore Student-t.

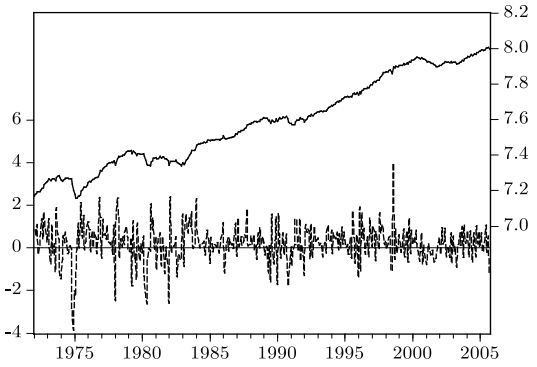
To determine the marginal density for σ_ε^2 we integrate (15) over the domain of β. Doing so gives

$$\begin{aligned} p(\sigma_\varepsilon^2|y,X) &= \int_{-\infty}^\infty p(\beta,\sigma_\varepsilon^2|y,X)\mathrm{d}\beta \\ &\propto (\sigma_\varepsilon^2)^{-\frac{(T+2)}{2}} \exp\left[-\frac{1}{2\sigma_\varepsilon^2}\left[(y-X\hat{\beta})'(y-X\hat{\beta}) + (\beta-\hat{\beta})'X'X(\beta-\hat{\beta})\right]\right] \\ &\propto (\sigma_\varepsilon^2)^{-\frac{(T-K+2)}{2}} \exp\left[-\frac{1}{2\sigma_\varepsilon^2}\left[(y-X\hat{\beta})'(y-X\hat{\beta})\right]\right] \\ &\quad \times \int_{-\infty}^\infty (\sigma_\varepsilon^2)^{-\frac{K}{2}} \exp\left[-\frac{1}{2\sigma_\varepsilon^2}(\beta-\hat{\beta})'X'X(\beta-\hat{\beta})\right] \mathrm{d}\beta \end{aligned} \tag{21}$$

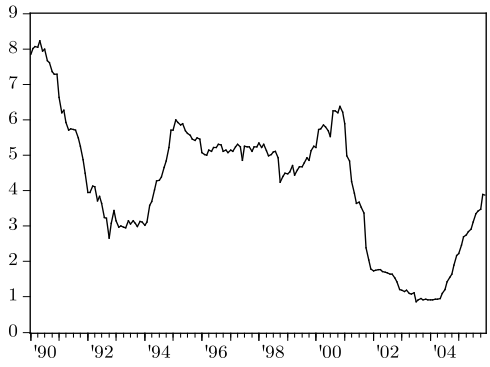
The last line of (21) is proportional to a multivariate Normal density and integrates therefore to a constant leaving

$$p(\sigma_\varepsilon^2|y,X) \propto (\sigma_\varepsilon^2)^{-\frac{(T-K+2)}{2}} \exp\left[-\frac{1}{2\sigma_\varepsilon^2}\left[(y-X\hat{\beta})'(y-X\hat{\beta})\right]\right] \tag{22}$$

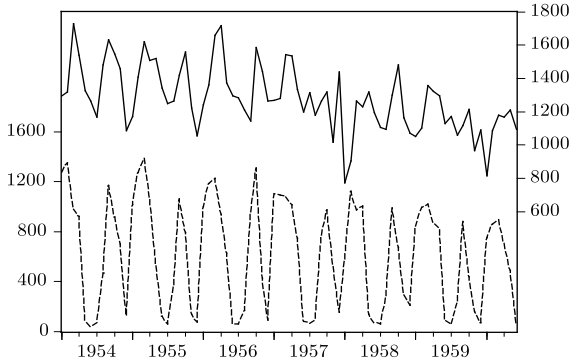
Figure 6: **US Industrial Production, Lydia Pinkham Sales and Advertising and 3-Month US Treasury Bill yield**



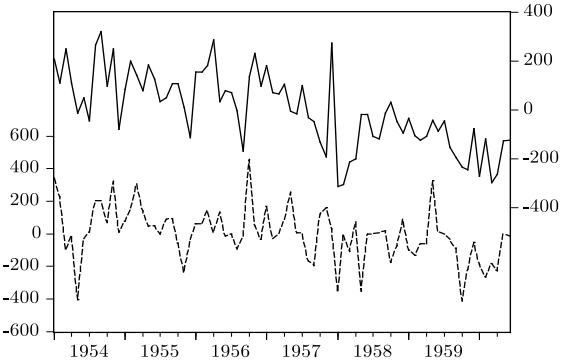
(a) US Industrial Production; level & rate (%)



(b) 3-Month US Treasury Bill yield (%)



(c) Unadjusted Lydia Pinkham series; sales and advertising



(d) Seasonally adjusted Lydia Pinkham series; sales and advertising

Notes: Panel (a) shows log levels (solid line) and growth rates (in % terms) for US Industrial Production (Gross Value of Products: Final products and nonindustrial supplies). The monthly series runs from January 1972 to September 2005 and was obtained from `http://www.economagic.com`. Panel (b) shows end-of month levels for the 3-Month US Treasury Bill yield for the period January 1990-December 2005 which were obtained from the St. Louis FED website (`http://research.stlouisfed.org/fred2`). Panel (c) shows the unadjusted Lydia Pinkham series for sales (solid lines) and advertising (dashed lines) whereas panel (d) shows the seasonally series (constructed after prefiltering the data with the results of a preliminary regression using 12 monthly dummies). The monthly series over the period January 1954-June 1960 were taken from Palda (1964), Table 2, pp. 32-33.

From (A-3) it follows that the marginal for σ_ε^2 is Inverted Gamma with parameters $m = \frac{1}{2}(y - X\hat{\beta})'(y - X\hat{\beta})$ and $\nu = \frac{1}{2}(T - K)$. The marginal densities are thus given as

$$p(\beta|y, X) \sim t\left(\hat{\beta}, \frac{X'X}{y'M_X y/(T-K)}, T-K\right) \tag{23}$$

$$p(\sigma_\varepsilon^2|y, X) \sim \mathcal{IG}\left(\frac{1}{2}(y - X\hat{\beta})'(y - X\hat{\beta}), \frac{1}{2}(T-K)\right) \tag{24}$$

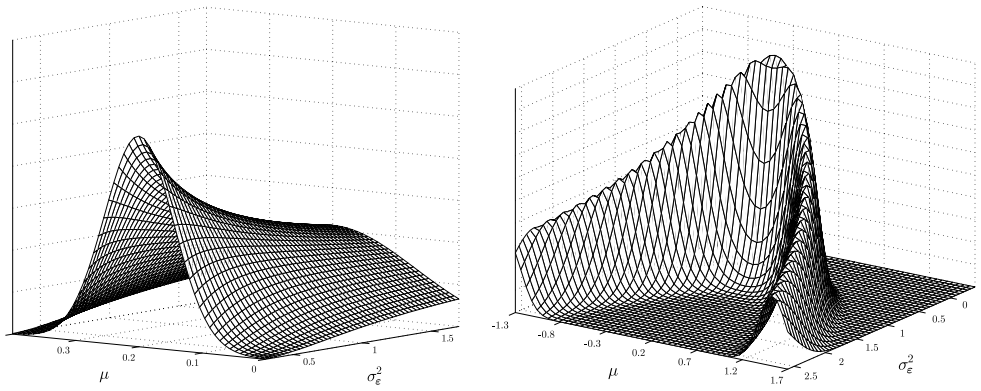
Gibbs sampling for the basic linear regression model consists of iteratively drawing from the conditional densities $p(\beta|y, X, \sigma_\varepsilon^2)$ and $p(\sigma_\varepsilon^2|y, X, \beta)$. The j^{th} Gibbs step therefore consists of

- generate	$\beta^{(j)}\|\sigma_\varepsilon^{2\,(j-1)}$	from	$p(\beta\|y, X, \sigma_\varepsilon^2) \sim \mathcal{N}\big(\hat{\beta}, \sigma_\varepsilon^{2\,(j-1)}[X'X]^{-1}\big)$
- generate	$\sigma_\varepsilon^{2\,(j)}\|\beta^{(j)}$	from	$p(\sigma_\varepsilon^2\|y, X, \beta) \sim \mathcal{IG}\big(\frac{1}{2}(y - X\beta^{(j)})'(y - X\beta^{(j)}), \frac{1}{2}T\big)$

Empirical Illustration: US Industrial Production

To get a better understanding of what the Gibbs conditional densities look like graphically and to see whether the generated Markov Chains can indeed be considered as samples from the marginal densities, we apply the linear regression model to a monthly series of US Industrial Production growth rates for the period January 1972-September 2005, shown in Figure 6 (a). Denote Industrial Production growth by the symbol y and for simplicity set $X = \iota_T$ where ι_T denotes a $(T \times 1)$ vector or ones. Therefore, the (scalar) β estimates the average growth rate of production. For convenience we relabel it with the symbol

Figure 7: **Conditional posterior densities**

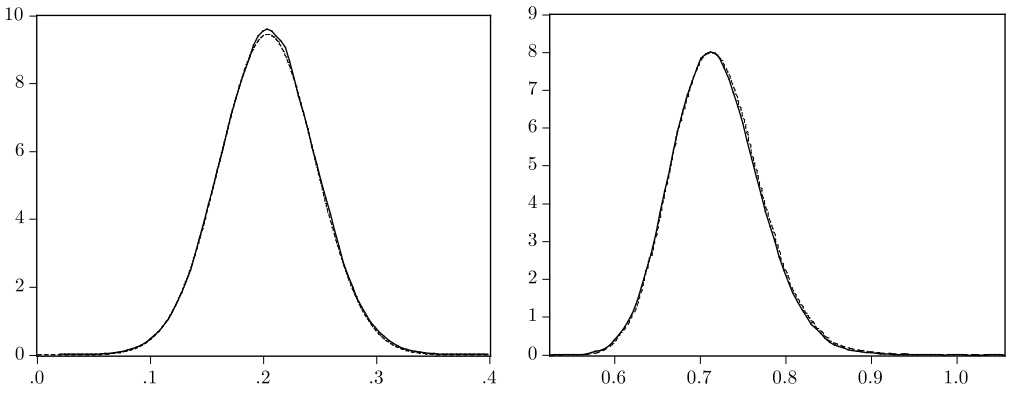


(a) $p(\mu|y, \sigma_\varepsilon^2)$ (b) $p(\sigma_\varepsilon^2|y, \mu)$

Notes: Panel (a) shows the conditional posterior density for μ for given values of σ_ε^2 and the data vector y: $p(\mu|y, \sigma_\varepsilon^2)$ whereas panel (b) shows the conditional density of σ_ε^2 for given values of μ: $p(\sigma_\varepsilon^2|y, \mu)$ when we apply the linear regression model (11), with $X = \iota_T$, to the monthly US Industrial Production growth rate for the period January 1972-September 2005.

μ. Note that the model specification we use here is not intended as a serious attempt at modelling US IP growth, it merely serves as an example. The Gibbs conditional densities are shown in Figure 7(a) and (b). The monthly average Industrial Production growth rate in our sample equals 0.204% and Figure 7(a) shows that for any given value of σ_ε^2 the conditional density for μ is nicely centered around this value. The conditional variance of μ clearly varies with the value of σ_ε^2. For increasingly larger values of σ_ε^2 the posterior density flattens out and the variance for μ will therefore increase. Figure 7(b) on the other hand shows that a given value for μ determines the location as well as the variance of the conditional density for σ_ε^2. The mean and variance of σ_ε^2 are lowest for values of μ close to the average IP growth. For all other values, both the mean and variance are higher. From the analytical expressions of the first two moments of an Inverted Gamma density,

Figure 8: **Marginal posterior densities**



(a) $p(\mu|y)$ (b) $p(\sigma_\varepsilon^2|y)$

Notes: Panel (a) shows the analytical marginal posterior density (dashed line) for μ, $p(\mu|y)$, together with a kernel density estimate (solid line) using the draws from the Gibbs sampler whereas panel (b) shows the marginal density, $p(\sigma_\varepsilon^2|y)$, and the kernel density estimate for σ_ε^2. The kernel densities are based on $100,000$ simulations after a burn-in of $B = 10,000$ draws and selecting every $h = 10^{\text{th}}$ when we apply the linear regression model (11), with $X = \iota_T$, to the monthly US Industrial Production growth rate for the period January 1972-September 2005.

see Appendix A, it is clear why; the value of both moments increase when μ deviates more from the sample mean.

Figure 8 shows kernel density estimates from 100,000 draws for μ (panel (a)) and σ_ε^2 (panel (b)) from the Gibbs sampler (solid line). Also depicted (dotted line) are the marginal densities given in (16) and (24). In either panel, both densities all but coincide. The Gibbs sampler thus provides accurate posterior results for the parameters of interest.

3.2 The Cochrane-Orcutt Model

After having derived the conditional and marginal densities for the basic linear regression model, we are now ready to consider the Cochrane-Orcutt model which extends the model in (10) by allowing the error terms to have first order autocorrelation. That is:

$$y_t = x_t\beta + \nu_t, \qquad t = 1, ..., T \tag{25}$$

$$\nu_t = \rho\nu_{t-1} + \varepsilon_t, \qquad \text{with} \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2) \tag{26}$$

where ρ is the parameter that determines the strength of the autocorrelation. The domain of this parameter is bounded to $-1 < \rho < 1$. The domain for the remaining parameter is given by $-\infty < \beta < \infty$ and $0 < \sigma_\varepsilon^2 < \infty$. $\boldsymbol{\theta}$ consists of $(\beta, \rho, \sigma_\varepsilon^2)$. When $\rho = 0$, the Cochrane-Orcutt model coincides with the basic regression model since ν reduces to a white noise series. As we will see later, difficulties occur when there is a constant term and ρ has substantial posterior density mass at the edges of its domain. By substituting (26) in (25) and rewriting the resulting expression in matrix notation, we have

$$y - \rho y_{-1} = X\beta - X_{-1}\beta\rho + \varepsilon, \qquad \text{with} \qquad \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2 \mathrm{I}_T) \tag{27}$$

where y_{-1} and X_{-1} denote the one-period lagged values of y and X. This reformulation shows that the Cochrane-Orcutt model is nonlinear in the parameters β and ρ. Although

this hampers parameter estimation and inference when using the frequentist's approach, obtaining posterior results using Gibbs sampling is straightforward as we will show below. We now turn to deriving the conditional and marginal densities and it will become apparent that the Cochrane-Orcutt model serves as a template for several other well-known econometric models.

Gibbs Sampling

Combining the likelihood for the Cochrane-Orcutt model with the same non-informative prior we specified before in (13), the joint posterior density is as follows:

$$p(\boldsymbol{\theta}|y,X) \propto (\sigma_\varepsilon^2)^{-\frac{(T+2)}{2}} \exp\left[-\frac{1}{2\sigma_\varepsilon^2}\big(y-\rho y_{-1}-X\beta+X_{-1}\beta\rho\big)'\big(y-\rho y_{-1}-X\beta+X_{-1}\beta\rho\big)\right] \quad (28)$$

To facilitate the derivation of the conditional densities it is useful to rewrite (27) in two different ways. In each case we condition on one of the two types of regression coefficients. First, we rewrite (28) conditional on values for ρ:

$$y^* = X^*\beta + \varepsilon \qquad \text{where} \qquad \begin{cases} y^* = y^*(\rho) \equiv y - \rho y_{-1} \\ X^* = X^*(\rho) \equiv X - \rho X_{-1} \end{cases} \quad (29)$$

Second, conditional on values for β, (28) becomes:

$$\tilde{y} = \rho\tilde{y}_{-1} + \varepsilon \qquad \text{where} \qquad \begin{cases} \tilde{y} = \tilde{y}(\beta) \equiv y - X\beta \\ \tilde{y}_{-1} = \tilde{y}_{-1}(\beta) \equiv y_{-1} - X_{-1}\beta \end{cases} \quad (30)$$

To derive the conditional density for β we use (29) to rewrite the joint posterior density. Doing so again gives us the joint density of the basic linear regression model so we can reuse all our earlier derivations. It therefore follows immediately that the conditional density for β is multivariate Normal with mean $m = \beta^* \equiv (X^{*\prime}X^*)^{-1}X^{*\prime}y^*$ and variance matrix $S = S_\beta \equiv \sigma_\varepsilon^2(X^{*\prime}X^*)$. Similarly, using (30) we have that the conditional density[8] for ρ is Normal with mean $m = \hat{\rho} \equiv (\tilde{y}_{-1}{}'\tilde{y}_{-1})^{-1}\tilde{y}_{-1}{}'\tilde{y}$ and variance $s^2 = \sigma_\rho^2 \equiv \sigma_\varepsilon^2(\tilde{y}_{-1}{}'\tilde{y}_{-1})^{-1}$. The conditional density for σ_ε^2 is again Inverted Gamma with parameter $m = \frac{1}{2}\varepsilon'\varepsilon \equiv \frac{1}{2}\big(y-\rho y_{-1}-X\beta+X_{-1}\beta\rho\big)'\big(y-\rho y_{-1}-X\beta+X_{-1}\beta\rho\big)$ and $\nu = \frac{1}{2}T$ degrees of freedom. The j^{th} Gibbs step thus consists of

- generate	$\beta^{(j)}\|\rho^{(j-1)},\sigma_\varepsilon^{2(j-1)}$	from	$p(\beta\|y,X,\rho,\sigma_\varepsilon^2) \sim \mathcal{N}\left(\beta^{*(j-1)}, S_\beta^{(j-1)}\right)$
- generate	$\rho^{(j)}\|\beta^{(j)},\sigma_\varepsilon^{2(j-1)}$	from	$p(\rho\|y,X,\beta,\sigma_\varepsilon^2) \sim \mathcal{N}\left(\hat{\rho}^{(j)}, \sigma_\rho^{2(j-1)}\right)$
- generate	$\sigma_\varepsilon^{2(j)}\|\beta^{(j)},\rho^{(j)}$	from	$p(\sigma_\varepsilon^2\|y,X,\beta,\rho) \sim \mathcal{IG}\left(\frac{1}{2}\varepsilon^{(j)\prime}\varepsilon^{(j)}, \frac{1}{2}T\right)$

From the conditional densities it follows that the Gibbs sampler has no difficulties with the nonlinearities in the likelihood. This is due to the fact that *conditional* on one regression parameter, the model for the other regression parameter is the basic linear regression model as shown in (25) and (26). In fact, the joint posterior density for ρ and any element of β, or the other way around, resembles the density shown in Figure 3(a). Therefore, the Gibbs sampler is a very convenient approach for drawing inference on the parameters in these types of models. We note that due to the truncation of ρ, one should ignore

[8] More precisely, we have a truncated density defined at the interval $-1 < \rho < 1$.

drawings outside the interval $(-1, 1)$. A more efficient algorithm has been developed by Geweke (1991, 1996).

Furthermore, whereas in the basic regression model Gibbs sampling was unnecessary because the marginal densities could be derived analytically, here we do need Gibbs sampling. This is because the marginal densities for β, ρ and σ_ε^2 are not a member of any known class of densities. To show why we derive the expression for $p(\beta|y, X)$ and $p(\rho|y, X)$.

After integrating out σ_ε^2 from the joint density we get

$$p(\beta, \rho|y, X) \propto \left[\left(y - \rho y_{-1} - X\beta + X_{-1}\beta\rho\right)'\left(y - \rho y_{-1} - X\beta + X_{-1}\beta\rho\right)\right]^{-\frac{T}{2}}$$

which can be rewritten in two different ways:

$$p(\beta, \rho|y, X) \propto \left[\tilde{y}' M_{\tilde{y}_{-1}} \tilde{y} + (\rho - \hat{\rho})' \tilde{y}_{-1}' \tilde{y}_{-1} (\rho - \hat{\rho})\right]^{-\frac{T}{2}} \tag{31}$$

$$p(\beta, \rho|y, X) \propto \left[y^{*\prime} M_{X^*} y^* + (\beta - \beta^*)' X^{*\prime} X^* (\beta - \beta^*)\right]^{-\frac{T}{2}} \tag{32}$$

Using the same techniques as we used before for deriving the marginal density of σ_ε^2 in the previous section we can integrate out ρ from (31) and β from (32). The resulting expressions are

$$p(\beta|y, X) \propto \left[(y - X\beta)' M_{y_{-1} - X_{-1}\beta} (y - X\beta)\right]^{-\frac{T-1}{2}} \left[(y_{-1} - X_{-1}\beta)'(y_{-1} - X_{-1}\beta)\right]^{-\frac{1}{2}} c(\beta)$$

$$p(\rho|y, X) \propto \left[(y - \rho y_{-1})' M_{X - \rho X_{-1}} (y - \rho y_{-1})\right]^{-\frac{T-K}{2}} \left[(X - \rho X_{-1})'(X - \rho X_{-1})\right]^{-\frac{1}{2}}$$

where $c(\beta)$ is given as $c(\beta) = \Phi\left(\frac{1-\hat{\rho}}{\sigma_\rho}\right) - \Phi\left(\frac{-1-\hat{\rho}}{\sigma_\rho}\right)$ and Φ stands for the standard Normal distribution function. Both these densities do not belong to any known class of density functions which means that we need Gibbs sampling to provide us with posterior results. Despite the fact that the marginal densities of β and ρ can not be determined analytically, applying the Gibbs sampler is a straightforward exercise. Furthermore, under the condition that all variables in X have some variability, there are no issues in terms of impropriety of the joint posterior density revolving ρ reaching the edges of its domain. This is also confirmed by the Fisher Information matrix which is defined as minus the expectation of the matrix of second order derivatives of the log likelihood with respect to the parameter vector $\boldsymbol{\theta}$, i.e. $\mathcal{I} = -E[\frac{\delta^2 \ln L(\boldsymbol{\theta}|y,x)}{\delta\boldsymbol{\theta}\delta\boldsymbol{\theta}'}]$. For the Cochrane-Orcutt model the Information matrix is given by[9]

$$\mathcal{I} = -E\begin{bmatrix} \frac{\delta^2 \ln L}{\delta\rho^2} & \frac{\delta^2 \ln L}{\delta\rho\delta\beta'} & \frac{\delta^2 \ln L}{\delta\rho\delta\sigma_\varepsilon^2} \\ \frac{\delta^2 \ln L}{\delta\beta\delta\rho} & \frac{\delta^2 \ln L}{\delta\beta\delta\beta'} & \frac{\delta^2 \ln L}{\delta\beta\delta\sigma_\varepsilon^2} \\ \frac{\delta^2 \ln L}{\sigma_\varepsilon^2\delta\rho} & \frac{\delta^2 \ln L}{\delta\sigma_\varepsilon^2\delta\beta'} & \frac{\delta^2 \ln L}{\delta\sigma_\varepsilon^4} \end{bmatrix} = \begin{bmatrix} \frac{T}{1-\rho^2} & 0 & 0 \\ 0 & \frac{(X-\rho X_{-1})'(X-\rho X_{-1})}{\sigma_\varepsilon^2} & 0 \\ 0 & 0 & \frac{T}{2\sigma_\varepsilon^4} \end{bmatrix} \tag{33}$$

The Hessian, which is defined as the inverse of the Information matrix, shows that even when $|\rho|$ is nearly identical to 1 none of the variances "explode". In the next two sections we will see this not always needs to be the case.

[9] We should note that we focus here on long term means only in which case $E[y] = E[y_{-l}] = X\beta$ for $l > 0$. In reality, T is finite and therefore (small) sample means should be considered. For expositional purposes, however, we focus solely on long term expectations; see Kleibergen and van Dijk (1994) for a finite sample analysis.

Table 1: **Posterior results for the Cochrane-Orcutt, Koyck and Unit Root models**

(a) Cochrane-Orcutt			(b) Koyck			(c) Unit Root		
	Posterior mean	Posterior s.d.		Posterior mean	Posterior s.d.		Posterior mean	Posterior s.d.
μ	0.202***	(0.042)	β	0.771***	(0.250)	μ	2.980*	(4.260)
ρ	0.073*	(0.047)	ρ	0.478***	(0.143)	ρ	0.986***	(0.009)
σ_ε^2	0.718***	(0.036)	σ_ε^2	1.397***	(0.187)	σ_ε^2	0.053***	(0.006)

Notes: The table shows posterior results for the Cochrane-Orcutt model (28) with $X = \iota_T$ and $\beta = \mu$ for monthly US Industrial Production growth rates (January 1972-September 2005) in panel (a), the Koyck model (37) for the monthly Lydia Pinkham data (January 1954-June 1960) in panel (b) and the Unit Root model (45) for the monthly 3-month Treasury Bill rate (January 1990-December 2005) in panel (c). All results are based on 100,000 simulations after a burn-in of $B = 10,000$ draws and selecting every $h = 10^{\text{th}}$ draw. *** indicates that zero is not contained in the 99% highest posterior density (HPD) region, ** indicates that zero is contained in the 99% but not in the 90% and 95% HPD region and * that zero is contained in the 99% and 95% but not in the 90% HPD region.

Empirical Illustration: US Industrial Production

To show the usefulness of Gibbs sampling in this context we again run the Gibbs sampler on the monthly US Industrial Production growth rates where we now allow for first order serial correlation in the error terms. Panel (a) of Table 1 reports posterior results. The posterior mean of μ is nearly identical to the earlier reported monthly sample average of 0.204%. The posterior mean of ρ is 0.073. Since zero is contained in the 95% and 99% but not in the 90% HPD region there is weak evidence for serial correlation. We note that in this particular case we treat x_t as a constant. The likelihood information is, however, such that the probability mass of ρ is not near the boundary of unity. We will treat this latter case in subsection 3.4.

In the next two subsections we discuss the Koyck Distributed Lag model and the Unit Root model. As we will show, both models are nested in the Cochrane-Orcutt model. They deviate from the latter by the way the exogenous variables are specified. Moreover, the additional structure that is imposed on the exogenous variable results in boundary issues culminating in a non-identification and non-stationary issue respectively when ρ is near 1.

3.3 The Koyck Model

A further extension of the basic linear regression model that we analyze is the univariate distributed lag model[10]. This model has proven to be one of the workhorses of econometric modelling practice since it offers the econometrician a straightforward tool to investigate the dependence of a variable on past values of the variable itself or past values of exogenous explanatory variables. Here we focus in particular on the well known Koyck model which is popular in for example marketing econometrics to investigate the dynamic link between sales and advertising. The general distributed lag model has, in principle, an infinite number of parameters. Koyck (1954) proposed a model specification in which the lag parameters are a geometric series which is governed by a single unknown parameter. The

[10]For an extensive overview of distributed lag models, see Griliches (1967).

resulting model is known as the geometric distributed lag model or simply as the Koyck model. We will discuss the difficulties that can arise when applying the Gibbs sampler to this model due to a boundary issue which results in a parameter (near) non-identification issue. We will illustrate this by means of an empirical application using the well known Lydia Pinkham dataset.

The Koyck model, in which we also allow for first order serial correlation in the error terms, is given by

$$\begin{aligned} y_t &= \beta w_t + \nu_t, \qquad t = 1, ..., T && (34) \\ w_t &= (1-\rho)\sum_{i=0}^{\infty} \rho^i x_{t-i} && (35) \\ \nu_t &= \rho\nu_{t-1} + \varepsilon_t, \qquad \text{with} \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2) && (36) \end{aligned}$$

where it is assumed that $0 < \rho < 1$, $-\infty < \beta < \infty$ and $0 < \sigma_\varepsilon^2 < \infty$.

Note that the effect of lagged values of the (single) explanatory variable x_t is determined solely by ρ and that this parameter is assumed to be equal to the first order serial correlation parameter[11]. The parameter vector is again given by $\boldsymbol{\theta} = (\beta, \rho, \sigma_\varepsilon^2)$. Substituting (36) in (34) gives the same expression as we found for the Cochrane-Orcutt model, which slightly rewritten, equals

$$y - \rho y_{-1} = \beta(w - \rho w_{-1}) + \varepsilon$$

Equation (35) puts additional structure on the term $w - \rho w_{-1}$, more specifically, $w - \rho w_{-1} = (1-\rho)x$ resulting in

$$y = \rho y_{-1} + \beta(1-\rho)x + \varepsilon, \qquad \text{with} \qquad \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2 \mathrm{I}_T) \tag{37}$$

This shows that the Koyck model is nested in the Cochrane-Orcutt model and that therefore all derivations we did for that model hold here as well. The specific structure that is placed on the exogenous variable will result in a boundary issue when ρ is close to 1. We can understand why this is so while there is no problem when this occurs in the Cochrane-Orcutt model by realizing that β will be near non-identified for ρ close to 1. This means that y effectively becomes a random walk and that exogenous variables no longer have any adjusting effect on y. We will analyze the joint, conditional and marginal densities to give insights in the consequences of the non-identification of β when applying the Gibbs sampler.

Gibbs Sampling

The Gibbs j^{th} step is given by

- generate	$\beta^{(j)} \vert \rho^{(j-1)}, \sigma_\varepsilon^{2(j-1)}$	from	$p(\beta \vert y, x, \rho, \sigma_\varepsilon^2) \sim \mathcal{N}\left(\beta^{*(j-1)}, \sigma_\beta^{2(j-1)}\right)$
- generate	$\rho^{(j)} \vert \beta^{(j)}, \sigma_\varepsilon^{2(j-1)}$	from	$p(\rho \vert y, x, \beta, \sigma_\varepsilon^2) \sim \mathcal{N}\left(\hat{\rho}^{(j)}, \sigma_\rho^{2(j-1)}\right)$
- generate	$\sigma_\varepsilon^{2(j)} \vert \beta^{(j)}, \rho^{(j)}$	from	$p(\sigma_\varepsilon^2 \vert y, x, \beta, \rho) \sim \mathcal{IG}\left(\frac{1}{2}\varepsilon^{(j)\prime}\varepsilon^{(j)}, \frac{1}{2}T\right)$

[11] The parameter ρ is usually referred to as the "retention" parameter.

where the parameters for the conditional densities of β and ρ are now specified as

$$\begin{aligned}\beta^* &= (x^{*\prime}x^*)^{-1}x^{*\prime}y^* = \left[(1-\rho)^2x'x\right]^{-1}x'(y-\rho y_{-1}) &(38)\\ \sigma_\beta^2 &= \sigma_\varepsilon^2(x^{*\prime}x^*)^{-1} = \sigma_\varepsilon^2\left[(1-\rho)^2x'x\right]^{-1} &(39)\end{aligned}$$

and

$$\begin{aligned}\hat{\rho} &= (\tilde{y}_{-1}{}'\tilde{y}_{-1})^{-1}\tilde{y}_{-1}{}'\tilde{y} = \left[(y_{-1}-\beta x)'(y_{-1}-\beta x)\right]^{-1}(y_{-1}-\beta x)'(y-\beta x) &(40)\\ \sigma_\rho^2 &= \sigma_\varepsilon^2(\tilde{y}_{-1}{}'\tilde{y}_{-1})^{-1} = \sigma_\varepsilon^2\left[(y_{-1}-\beta x)'(y_{-1}-\beta x)\right]^{-1} &(41)\end{aligned}$$

At first sight, it may seem straightforward to apply the Gibbs sampler to the Koyck model. However, upon closer inspection of the conditional density parameters it becomes clear that a problem can occur for values of ρ close to 1. Suppose that a value near 1 is drawn for ρ. The conditional variance of β given this draw goes to infinity, see (39), which means that any value along the line is likely to be drawn for β. If the next draw for β is indeed large then the conditional variance for ρ goes to zero, see (41), as a result of which the next draw for ρ is also going to be close to 1, see (40). This means that the Gibbs Markov Chain will not only converge much slower but that is convergence is not even guaranteed at all if $\rho = 1$ is acting as an absorbing state. The extent of this problem depends on how much probability mass there actually is close to $\rho = 1$.

To understand the behaviour of the Gibbs sampler here we need to examine the joint and marginal densities in detail to comprehend what is going on. The marginal densities for β and ρ are as follows

$$\begin{aligned}p(\beta|y,x) &\propto \left[(y-\beta x)'M_{y_{-1}-\beta x}(y-\beta x)\right]^{-\frac{T-1}{2}}\left[(y_{-1}-\beta x)'(y_{-1}-\beta x)\right]^{-\frac{1}{2}}c(\beta)\\ p(\rho|y,x) &\propto \left[(y-\rho y_{-1})'M_{(1-\rho)x}(y-\rho y_{-1})\right]^{-\frac{T-1}{2}}\left[x'x\right]^{-\frac{1}{2}}(1-\rho)^{-1}\end{aligned}$$

where $c(\beta)$ is given by as $c(\beta) = \Phi\left(\frac{1-\hat{\rho}}{\sigma_p}\right) - \Phi\left(\frac{-\hat{\rho}}{\sigma_p}\right)$

Focusing on the density for ρ, we can recognize it as a Student-t type density, except for the factor $(1-\rho)^{-1}$. It is exactly this factor that is causing the behaviour of the Gibbs sampler. The reason is that the joint density $p(\beta, \rho|y, x)$ is improper because it is constant at, and therefore also very close to, $\rho = 1$ for $-\infty < \beta < \infty$. Graphically, this means that the joint density has a "wall", similar to the ridge that was depicted in Figure 4. Integrating the joint density over ρ will cause the marginal density for β to potentially have infinite tails because the joint density is all but flat close to $\rho = 1$. Similarly, the marginal density for ρ will tend to infinity when ρ tends to 1.

To reiterate what we said before, the extent of the problem depends on the data at hand. If the likelihood assigns virtually no probability mass to the region close to $\rho = 1$ then the marginal for β will be indistinguishable from a Student-t density. Furthermore, the marginal density for ρ will still be infinite close to $\rho = 1$ but if ρ happens to be far out in the tail of the distribution then this should not pose a big problem. If on the other hand substantial mass is near $\rho = 1$ then action has be undertaken to prevent the Gibbs sampler from reaching that part of the domain for ρ or, alternatively, to try and regularize the likelihood. Choosing an appropriate prior density can do the trick.

Analyzing the Information Matrix gives similar insights in the irregularity in the joint density close to $\rho = 1$ and furthermore, it provides us with a direction for a possible solution to tackle this irregularity. The Information Matrix follows directly from (33) by

substituting in $X - \rho X_{-1} = (1-\rho)x$. Therefore

$$\mathcal{I} = \begin{bmatrix} \frac{T}{1-\rho^2} & 0 & 0 \\ 0 & \frac{(1-\rho)^2 x'x}{\sigma_\varepsilon^2} & 0 \\ 0 & 0 & \frac{T}{2\sigma_\varepsilon^4} \end{bmatrix} \tag{42}$$

The Information matrix again shows that for ρ close to 1, the variance of ρ is zero (inverse of the first diagonal element) whereas the variance of β is near infinity (inverse of the second diagonal element).

Potential Solutions

In order to apply the Gibbs sampler without any problems something has to be about the irregularity of the likelihood/joint density close to $\rho = 1$. A number of potential solutions have been proposed in the literature to circumvent this problem, see e.g. Schotman and van Dijk (1991) and Kleibergen and van Dijk (1994, 1998). Here we only briefly touch upon the several options to give the researcher a flavor of how to tackle the improprietry of the likelihood. One can distinguish three solution approaches: (i) truncation of the parameter space, (ii) regularization by choosing a prior that sufficiently smooths out the likelihood, (iii) use of a training sample to specify a weakly informative prior for β.

In terms of applying the first solution, one can truncate the domain of ρ near 1 and check whether there is probability mass near 1. Imposing an upper bound can be achieved by selecting for example a local uniform prior. The goal would be to only allow draws for ρ that are at least η away from 1 with $\eta > 0$ to prevent a wall in the joint posterior density. Choosing a specific value for η would necessarily be a subjective choice. But, once agreed upon a sensible value for η one can apply the Gibbs sampler. Alternatively, one can use a Metropolis-Hastings type step in which only draws that fall below $1 - \eta$ are accepted. For an example of this method, see Geman and Reynolds (1992) for an application to the (linear) image restoration problem (see also Geman and Geman, 1984) and Hurn and Jennison (1996) for a discussion on how the Truncated Gibbs Sampler fits in the Metropolis-Hastings class of sampling algorithms by choosing the proposal density such that it takes care of truncating the domain of ρ.

As for the second solution, one can try and regularize the likelihood in the neighborhood of $\rho = 1$ such that it becomes a proper density. This can be achieved by instead of an uninformative prior as in (13), using a prior that is chosen in such the way that it eliminates the factor $(1-\rho)^{-1}$. From the Information matrix in (42) we can construct the following Jeffreys' type prior for β given ρ and σ_ε^2,[12]

$$p(\beta|\rho, \sigma_\varepsilon^2) \propto \frac{(1-\rho)}{\sigma_\varepsilon^2} \qquad \text{for} \qquad 0 < \rho < 1 \tag{43}$$

Going through the derivations of the joint and marginal densities again with this prior will show that this prior eliminates the factor $(1-\rho)^{-1}$ from the marginal density of ρ. What basically happens is that the marginal density for ρ is now integrable everywhere except for $\rho = 1$ which in turn has a zero probability of occurring.

[12] In general the Jeffreys' prior is obtained from the relevant element of the square root of the determinant of the Information matrix of the considered model. For our purposes, however, we use a somewhat stronger prior because we need $(1-\rho)^1$ instead of $(1-\rho)^{\frac{1}{2}}$ to regularize the likelihood. For more details and an advanced analysis on similar Jeffreys' priors we refer to Kleibergen and van Dijk (1994, 1998).

The third solution is another way of regularizing the posterior density. One can use a training sample[13] to specify a weakly informative prior for β. Schotman and van Dijk (1991) specify the following prior

$$p(\beta|\rho,\sigma_\varepsilon^2) \propto \mathcal{N}\left(y_0, \tfrac{\sigma_\varepsilon^2}{(1-\rho)^2}\right) \qquad \text{for} \qquad 0<\rho<1 \tag{44}$$

where y_0 is the starting value for the time-series of y. The intuition behind this prior is that as ρ approaches 1 it becomes increasingly difficult to learn about β from the data since the mean of y, which depends on β, does not exist for ρ near 1. The prior is stronger for smaller values of ρ but approaches an uninformative prior for $\rho \to 1$. It is derived from the unconditional distribution of y_0 under the assumption of Normality. The effect of this Normal prior on the joint posterior density is that it eliminates the pronounced wall feature in the joint density. We will see an example of this approach when we discuss the Unit Root model

Further solutions, which we do not discuss here in the detail, are to reparameterize the model in such a way that the Gibbs sampler can be used without any problems for the reformulated model. However, one still has to translate the posterior results back to the original model. Without imposing some sort of prior, similar problems will still occur only now at a different stage in the analysis. For examples of reparametrization see for instance Gilks *et al.* (2000). Finally, modified versions of the Gibbs sampler such as the Collapsed Gibbs sampler (see Liu, 1994), where some parameters can be temporarily ignored when running the Gibbs sampler (in this case ρ) can be useful in this context as well.

Empirical Illustration: Sales and Advertising

To illustrate the behaviour of the Gibbs sampler for the Koyck model, we estimate the model (34)-(36) using the Lydia Pinkham dataset (see Palda, 1964). This dataset has been used extensively in marketing studies to investigate the dynamic relationship between advertising and sales. Figures 6(c) and 6(d) show the unadjusted and seasonally adjusted series. When we apply the Gibbs sampler on the latter series using the Gibbs conditional densities without any modifications (that is we use a uninformative prior), we run into the problems just discussed. We observe occasional extreme draws of β with an order of magnitude of $\pm 10^3$. These occur, as expected, for draws $\rho^{(j)}$ that are close to unity. However, the likelihood only assigns a small probability mass to values of ρ close to 1 so only relatively few draws of ρ are close to 1. Nevertheless, this issue should be properly addressed. We therefore truncated the domain of ρ as explained earlier by choosing $\eta = 10^{-5}$. The posterior results of the Gibbs sampler with this modification are shown in Table 1, panel (b). The posterior mean of ρ is 0.48, which implies that 90% of the advertising effect has taken place after approximately nine weeks[14]. This result is similar to the results documented in earlier studies, see Clarke (1976). For empirical applications of the Koyck model using classical (maximum likelihood) estimation techniques see for example Palda (1964), Bass and Clarke (1972) and Clarke (1976). So here truncation seems sufficient to address the impropriety of the joint density.

[13] For details on training samples we refer to O'Hagan (1994).

[14] The time period Δ_t during which $(100 \times \alpha)\%$ of the expected cumulative advertising effect has taken place can be shown to be equal to $\Delta_t = \ln(1-\alpha)/\ln(\hat{\rho}) - 1$, see Clarke (1976), pp. 348.

3.4 The Unit Root Model

A second model that turns out to be nested in the Cochrane-Orcutt model is the Unit Root model. As what we saw before with the Koyck Model, the Unit Root model puts a particular structure on the exogenous variables. In particular, x_t is imposed to be constant and therefore it holds that $X - \rho X_{-1} = (1-\rho)\iota_T$. The resulting model, after relabelling β by μ, is a first order autoregressive model for y

$$y_t - \mu = \rho(y_{t-1} - \mu) + \varepsilon_t \tag{45}$$

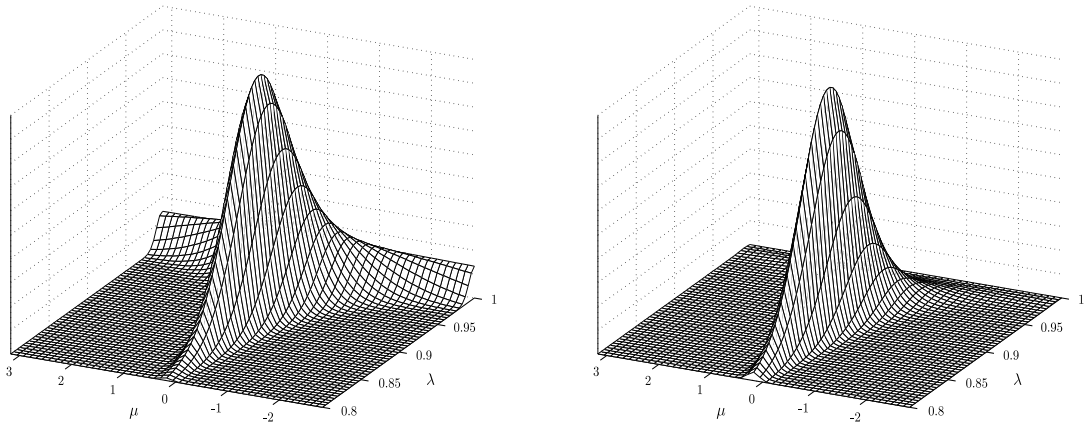
where μ is the unconditional mean of the time-series $\{y_t\}_{t=1}^T$. Similar as for the Koyck model, imposing this structure introduces a boundary issue which in this model results in non-stationarity for y which is caused by the non-identification of μ. In an AR(1) model for y_t, the interpretation of μ depends on whether the series y is stationary ($\rho < 1$) or whether it has a unit root ($\rho = 1$). In the latter case, the mean of y does not exist and μ is thus non-identified. Therefore, even when y is a weakly stationary process, with ρ close to unity, any value for μ along the real line is likely to be drawn in the Gibbs sampler when ρ is drawn close to 1. This will not only make it very difficult to pinpoint the posterior mean of μ but it also causes the sequence of draws for ρ to have difficulties moving away from $\rho = 1$ as was discussed before. Of course, ρ close to 1 can be an indication that one should model first differences of y instead of y itself which circumvents the entire issue altogether. However, for series such as interest rate levels there is no economic interpretation why they should be I(1) processes and one is left with dealing with the boundary issue nonetheless.

For series that are near unit root, substantial probability mass will lie close to $\rho = 1$ so that the impropriety of the joint posterior poses a serious issue. As an example we depicted the joint density for the unit root model for a series of monthly data on the 3-month US Treasury Bill in Figure 9(a). A time-series plot of this series is given in Figure 6(b). Figure 9(a) clearly shows the pronounced wall feature close to $\rho = 1$. In order to resolve the impropriety of the joint density a local uniform prior or truncation of the domain for ρ is unsatisfactory here because of the amount of probability mass at the edge of the domain of ρ. Using the Schotman and van Dijk (1991) prior to regularize the joint density is likely to be more promising here. In fact, the joint density which results from combining the data likelihood with this particular prior is shown in Figure 9(b). The joint density no longer has a wall close to $\rho = 1$ although it still flattens out somewhat near the edge of the domain. We note that this posterior may also be interpreted as the exact likelihood including the initial observation. For details see Schotman and van Dijk (1991).

Empirical Illustration: US Treasury Bill Rates

To illustrate, we obtain posterior results for the Unit Root model when applied to the 3-month US Treasury Bill series. This series portraits unit root type behavior as is evident from Figure 6(b). This is corroborated by posterior results from the Gibbs sampler. The posterior mean of ρ equals 0.991. However, whereas the sample mean of the T-bill series equals 4.16%, the posterior mean of μ is 1.36% and has a posterior standard deviation of a staggering 23.89%. Figure 10(a) shows that a substantial fraction of the draws for ρ are close to 1 and that the draws for μ are therefore all over the place. The scatterplot in Figure 10(b) shows that μ can be anything when ρ is drawn close to 1. As mentioned earlier, truncating the domain ρ at $1 - \eta$ might not be the best way to go here. In this case, draws for ρ close to 1 are in the region of the distribution that is of particular interest, as we expect ρ to be close to unity. Imposing the Schotman and van Dijk (1991) prior in

Figure 9: **Joint posterior density in the Unit Root model**

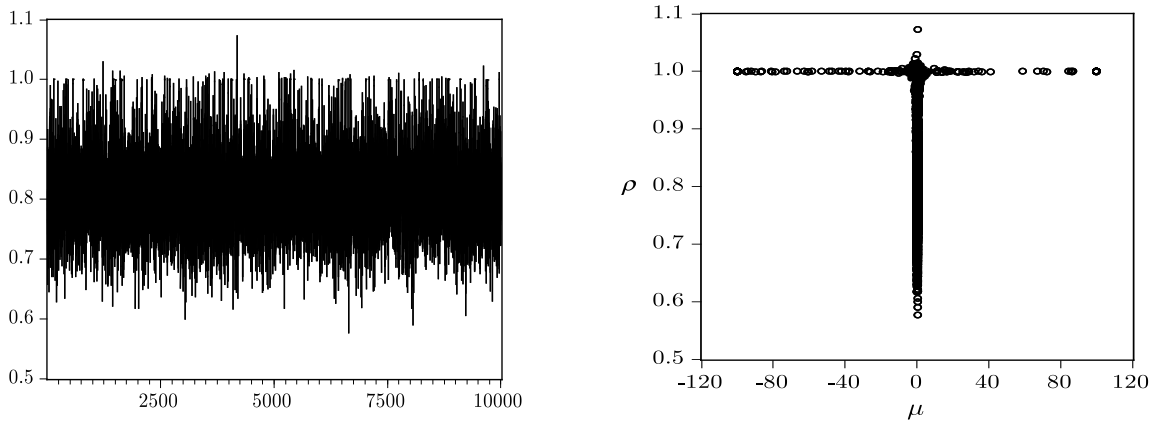


(a) Uniform prior

(b) Schotman and van Dijk (1991) prior

Notes: Panel (a) shows the joint posterior density $p(\lambda, \mu|y)$ when we use a uniform prior as in (13) whereas panel (b) shows the same posterior density however now with the prior proposed by Schotman and van Dijk (1991) as given in (44). In both panels we use the 3-Month US Treasury Bill rates for the period January 1990-December 2005 for the data vector y.

Figure 10: **Gibbs draws for the Unit Root Model with a non-informative prior**



(a) Gibbs draws for ρ

(b) Scatterplot of Gibbs draws for μ and ρ

Notes: Shown are the Gibbs draws for ρ, panel (a), and a scatterplot of the Gibbs draws for μ and ρ, panel (b). The graphs are based on the first 10,000 of a total of 100,000 draws from running the Gibbs sampler for the Unit Root model with a non-informative prior for $\boldsymbol{\theta}$. In the model we use the 3-Month US Treasury Bill rates for the period January 1990-December 2005 for the data vector y.

(44) on the other hand seems more appropriate. Doing so removes the wall in the joint density at $\rho = 1$. Table 1, panel (c) shows posterior results when the prior is imposed. The posterior mean (standard deviation) for μ and ρ are now 2.98% (4.26%) and 0.986 (0.009) respectively, which are more realistic.

3.5 The Instrumental Variables Model

The final class of models that we discuss in the current section are multivariate models. The issues involved here are similar to those surrounding univariate unit root models, i.e. non-identifiability of parameters. This will result in the Information Matrix being singular, or alternatively, in the Hessian having a reduced rank. This reduced rank problem can occur in several well-known models, such as for example Cointegration models, Vector Autoregressive (VAR) and Simultaneous Equation Models (SEM) which in turn are closely linked to Instrumental Variables (IV) models.

To show which role non-identifiability plays in these models we give an example by means of a just identified IV model and in particular we focus on the Incomplete Simultaneous Equation Model (INSEM). Our analysis, which is necessarily brief, is based on van Dijk (2003) and Hoogerheide *et al.* (2006a) and we refer to that study for a more in-depth analysis. Consider the INSEM model as it is specified in Zellner *et al.* (1988)[15]

$$y = x\beta + \varepsilon \tag{46}$$

$$x = z\pi + \nu \tag{47}$$

$$[\varepsilon \ \nu]' \sim \mathcal{N}\left([0 \ 0]', \Sigma\right) \qquad \text{with} \qquad \Sigma = \begin{bmatrix} \sigma_\varepsilon^2 & \sigma_{\varepsilon,\nu} \\ \sigma_{\varepsilon,\nu} & \sigma_\nu^2 \end{bmatrix} \tag{48}$$

with y, x and z all having dimensions $(T \times 1)$ and β and π being scalar parameters. $\boldsymbol{\theta}$ is given by $\boldsymbol{\theta} = (\beta, \pi, \Sigma)$. In this model, y is be to interpreted as the structural variable of interest, x is an endogenous variable and z is the (weakly exogenous) instrument. Similarly, β is the structural parameter of interest and π measures the quality of the instrument. Furthermore, the correlation parameter $\rho = \frac{\sigma_{\varepsilon,\nu}}{\sqrt{\sigma_\varepsilon^2 \sigma_\nu^2}}$ measures the degree of endogeneity of x in the equation for y. (46)-(48) is known as the *structural form* of the INSEM. By substituting (47) in (46) we can derive the *reduced form* which is given by

$$y = z\pi\beta + \xi \tag{49}$$

$$x = z\pi + \nu \tag{50}$$

$$[\varepsilon \ \nu]' \sim \mathcal{N}\left([0 \ 0]', \Sigma\right) \qquad \text{with} \qquad \Sigma = \begin{bmatrix} \sigma_\varepsilon^2 & \sigma_{\varepsilon,\nu} \\ \sigma_{\varepsilon,\nu} & \sigma_\nu^2 \end{bmatrix} \tag{51}$$

with $\xi = \varepsilon + \nu$. We can interpret the reduced form model as a multivariate regression model which is nonlinear in the parameters β and ρ as in (37). As was the case in the Unit Root model, this nonlinearity can lead to a non-identifiability problem. In particular, when $\pi = 0$, the joint posterior density if we assume a noninformative prior, is improper because it is flat and nonzero in the direction of β. In fact, the joint density will look very similar to that in Figure 9(a) in the sense that it has a wall at $\pi = 0$. Therefore, β is not identified when $\pi = 0$ whereas it will be for any $\beta \neq 0$. In a multivariate setting where y, x and z are all matrices and β and π are matrices as well, the identification problem of (part of the elements) of β occurs when $\pi = 0$ or when π is of reduced rank. The above problem is known as *local non-identification* and is discussed in detail in Kleibergen and van Dijk (1998).

As a result of the local-identification problem, the *marginal* density for π is non-integrable because of infinite probability mass near $\pi = 0$ (see Kleibergen and van Dijk, 1998). Whether or not the impropriety of the joint density will be revealed in the output

[15] The reason this model is called *just identified* is because there is only a single instrument, z.

from the Gibbs Sampler is unclear. Slow convergence of the Gibbs Sampler due to the fact that $\pi = 0$ is acting as an absorbing state could be an indication. Examples of bimodal posterior densities on bounded intervals are given in Hoogerheide *et al.* (2006a). A possible solution to circumvent the local non-identification problem in INSEM model would again be the specification of sensible prior densities. However, it can be an arduous task to find conjugate priors, mainly since these will have to curtail multiple parameters all at the same time.

In this section our focus has been on drawing inference on the regression parameters in univariate as well as multivariate models. In addition to the models we considered, several other model specifications, like ARMA models or error-correction models, may also be used. This is a topic for further research. As long as the researcher finds herself in a region of the parameter space where the likelihood is well behaved, for example in the case of the basic linear regression model, then the Gibbs sampler can be used in a straightforward fashion to obtain posterior results. However, when the likelihood assigns sufficient probability mass to the edges of the domain of the parameter region this may result in local identification issues, for example in the Unit Root model for ρ close to 1. Impropriety problems of the joint density can then occur and one needs to resort to measures such as imposing (weak) informative priors.

4 Gibbs Sampling Within Variance Component and Unobserved Component Models

We now switch our attention to drawing inference on variance parameters instead of regression parameters, in particular, when a variance tends towards the zero-bound and/or when a degrees of freedom restriction may be violated or an identification problem arises. We do so by analyzing again a canonical type of model, the so-called Hierarchical Linear Mixed Model (HLMM). This model is a variance components model, that is, the relative importance of several variances is the object of study. A second feature of this canonical model is the presence of unobserved components. The starting point of our analysis will be a basic specification of the HLMM. This model serves as a parent model for such extensions as a state space model and a panel data model, which we discuss subsequently.

4.1 Preliminaries

Before we specify the basic set-up of the HLMM we first discuss two preliminary models, focusing on variances of disturbances. The models serve to identify the issues involved.

Linear regression model with a small number of observations

In Section 3.1 we analyzed the basic linear regression model. Now we revisit this model which we simplify using $x_t = 1$ and $\beta = \mu$

$$y_t = \mu + \varepsilon_t, \qquad t = 1, ..., T, \qquad \text{with} \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2) \tag{52}$$

We emphasize that in this section the number of observations T may refer to the number of observation over time of a time series, to individuals or groups of individuals in a cross-section. For notational convenience, we use here the same symbol T for time series and

cross-section observations. If we use a *uniform* prior on both μ and σ_ε^2,

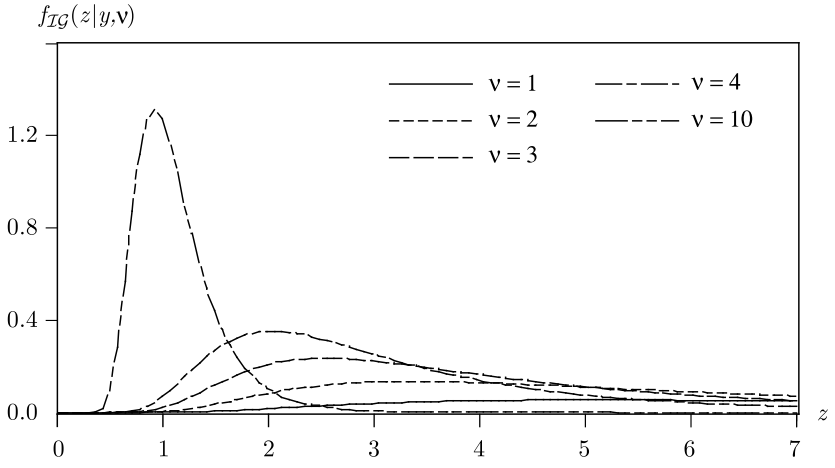
$$p(\mu, \sigma_\varepsilon^2) \propto 1 \tag{53}$$

then we can derive the marginal densities of μ and σ_ε^2 as

$$\begin{aligned} p(\mu|y) &\sim t\left(\hat{\mu}, \frac{(T-3)T}{s^2}, T-3\right) \\ p(\sigma_\varepsilon^2|y) &\sim \mathcal{IG}\left(\frac{1}{2}(y-\iota_T\hat{\mu})'(y-\iota_T\hat{\mu}), \frac{1}{2}(T-3)\right) \end{aligned}$$

with $\hat{\mu} = \frac{1}{T}\iota_T' y$ and $s^2 = y' M_{\iota_T} y$. Note that the degrees of freedom are now for the general case $T - K - 2$ since we use a uniform prior on both μ and σ_ε^2. In our case we have $K = 1$ [16]. These analytic results are necessary to analyze the convergence of the Gibbs step. From the parameters of the marginal densities and the conditions given in Appendix A it is clear that in order for these Student-t and Inverted Gamma densities to exist one needs more than 3 observations, i.e. $T > 3$. Further, in order for the first moment of each density to exist it is required that $T > 4$ for the marginal density of μ and $T > 5$ for the marginal density of σ_ε^2. Similarly, for the second moment to exist we need $T > 5$ and $T > 7$ respectively. See also the discussions in Koop (2003) and Geweke (2005). For illustration, Figure 11 shows that the right tail of an Inverted Gamma density tends to zero at a rate

Figure 11: **Inverted Gamma density**



Notes: The graph shows the Inverted Gamma density function, given by (A-3), for $y = 10$ and for a varying number of degrees of freedom, ν.

that is too small when the number of degrees of freedom is too small. For instance, for $\nu = 2$ the first moment exists but the second moment does not whereas both moments exist for any $\nu > 2$. Note that the density in (A-3) is stated in terms of ν. Therefore, from a data likelihood perspective, the relation between r and T will be $\nu = \frac{1}{2}T$. A non-flat prior will change T to for example $T + 1$ or $T + 2$.

[16] A Jeffreys' prior, $p(\sigma_\varepsilon^2) \propto 1/\sigma_\varepsilon^2$, increases the number of degrees of freedom with 1. As a result, densities now exist for $T > 1$.

The Gibbs conditional densities, using a uniform prior, are given by

$$\begin{aligned} p(\mu|y,\sigma_\varepsilon^2) &\sim \mathcal{N}\left(\hat{\mu}, \frac{1}{T}\sigma_\varepsilon^2\right) \\ p(\sigma_\varepsilon^2|y,\mu) &\sim \mathcal{IG}\left(\frac{1}{2}(y-\iota_T\mu)'(y-\iota_T\mu), \frac{1}{2}(T-2)\right) \end{aligned}$$

Only focusing on these *conditional* densities shows that $T = 3$ is a already sufficient for the Gibbs sampler to run. With a Jeffreys' prior $T = 1$ is sufficient. However, it follows from our analysis that in both situations, the *marginal* densities for μ and σ_ε^2 do not exist. Thus, we have a simple case where the Gibbs sampler can be applied as a simulation method, but the joint and marginal densities do not exist, see also the discussion in for example Koop (2003). Therefore, the generated Gibbs sample does not make sense. We emphasize that for the usual number of time series observations this degrees of freedom restriction is obviously of no significance. However, for the case of the number of groups in a panel it may become restrictive. In Section 4.4 we give an example using a panel data model.

Naive Heteroscedasticity

Consider a model in which each observation is allowed to have its own variance parameter

$$y_t = \mu + \varepsilon_t, \qquad t = 1, ..., T, \qquad \text{with} \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma_t^2), \tag{54}$$

From the analysis of the previous model we know that, depending on which (non-informative) prior specification is used, we need multiple observations to obtain sensible posterior results for $\boldsymbol{\theta} = (\mu, \sigma_1^2, \ldots, \sigma_T^2)$. One approach would be to partition the observations into groups, where it is assumed that per group the variance is constant whereas it is allowed to be different across groups. Each partition needs to be chosen in such a way that it contains a sufficient number of observations. For example, allowing for just two groups, inference in (54) is possible if we impose that

$$\sigma_t^2 = \begin{cases} \sigma_1^2 & \text{for } t = 1, ..., \tau \\ \sigma_2^2 & \text{for } t = \tau + 1, ..., T \end{cases} \tag{55}$$

for any value of τ in the open interval $(1, T)$, where we assume that τ is known and $T > 2$.

Our main point, although as trivial as it may seem, is that the degrees of freedom restriction implies that one needs multiple observations to draw inference on variance components. This may become relevant in particular in dynamic panels with groups of observations. We note that Geweke (1993) uses a weakly informative inverted Gamma density which makes the posterior more regular.

4.2 Hierarchical Linear Mixed Model (HLMM)

An example of a canonical model with at least two variances is the class of HLMM. We introduce this class through the following hierarchical model with two variance components

$$y_t = \mu_t + \varepsilon_t, \qquad \text{with} \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2) \qquad \text{for} \qquad t = 1, \ldots, T \tag{56}$$

$$\mu_t = \theta + \eta_t, \qquad \text{with} \qquad \eta_t \sim \mathcal{N}(0, \sigma_\eta^2) \qquad \text{and} \qquad \mathcal{E}[\varepsilon_t \eta_s] = 0 \tag{57}$$

with parameter vector $\boldsymbol{\theta} = (\theta, \mu, \sigma_\varepsilon^2, \sigma_\eta^2)$. μ is a vector containing the time-varying mean of y: $\mu = (\mu_1, \ldots, \mu_T)'$ and θ is the mean of the distribution of μ_t which, for any t, is Normal

with variance σ_η^2. This model serves as a parent model for more elaborate models such as state space models or panel data models. Before moving on to specifying and discussing these models, we analyze the base model by distinguishing between two cases. Each case helps to gain a better understanding of the dynamics of the HLMM class of models. Note that unless stated otherwise, we assume a flat prior for each of the variance components.

(i) $\sigma_\varepsilon^2 = 1$ and T small: a degrees of freedom bound

Because σ_ε^2 is given, the only unknown variance component is σ_η^2. The requirement on a minimum number of degrees of freedom as discussed in Section 4.1 is of importance here. Sensible posterior results can only be obtained when there is a sufficient number of observations. As before, the Gibbs sampler may work in this model even when the marginal posterior densities for θ and σ_η^2 do not exist, see Hobert and Casella (1996) for an example and discussion[17]. The conditional densities $p(\theta|\sigma_\eta^2)$ and $p(\sigma_\eta^2|\theta)$ can be derived from first substituting (57) in (56)

$$y_t = \theta + \varepsilon_t + \eta_t \qquad \text{with} \qquad \eta_t \sim \mathcal{N}(0, \sigma_\eta^2) \tag{58}$$

Since the dynamics of ε_t are known, running the Gibbs sampler consists of the following steps

- generate	$\theta^{(j)}\|\sigma_\eta^{2\,(j-1)}$	from	$p\left(\theta\|y, \sigma_\eta^2\right) \sim \mathcal{N}\left(\hat{\theta}, \frac{1}{T}\sigma_\eta^{2\,(j-1)}\right)$
- generate	$\sigma_\eta^{2\,(j)}\|\theta^{(j)}$	from	$p\left(\sigma_\eta^2\|y, \theta\right) \sim \mathcal{IG}\left(\frac{1}{2}(y - \iota_T\theta^{(j)})'(y - \iota_T\theta^{(j)}), \frac{1}{2}(T-2)\right)$

with $\hat{\theta} = \frac{1}{T}\iota_T{}'y$. Note that this is Gibbs sampling without having to concern oneself about μ. However, it is relatively easy to construct a Gibbs sampling step where μ is drawn alongside θ and σ_η^2; see Hobert and Casella (1996).

(ii) σ_ε^2 unknown and T large: an identification issue

By taking T large enough, the researcher does no longer need to worry about the marginal posterior densities possibly being non-existent. However, by making the first variance component, σ_ε^2, unknown as well introduces a new issue. More specifically, she now has to deal with an identification issue in the sense that it not possible to distinguish the two variance components from each other. Why this is the case can be made clear as follows. Note first that since T is assumed to be large enough, the marginal densities of σ_ε^2 and σ_η^2 will exist. However, respecifying the model in (58) to

$$y = \iota_T\theta + \varepsilon + \eta \qquad \text{with} \qquad \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2 \mathrm{I}_T) \qquad \text{and} \qquad \eta \sim \mathcal{N}(0, \sigma_\eta^2 \mathrm{I}_T) \tag{59}$$

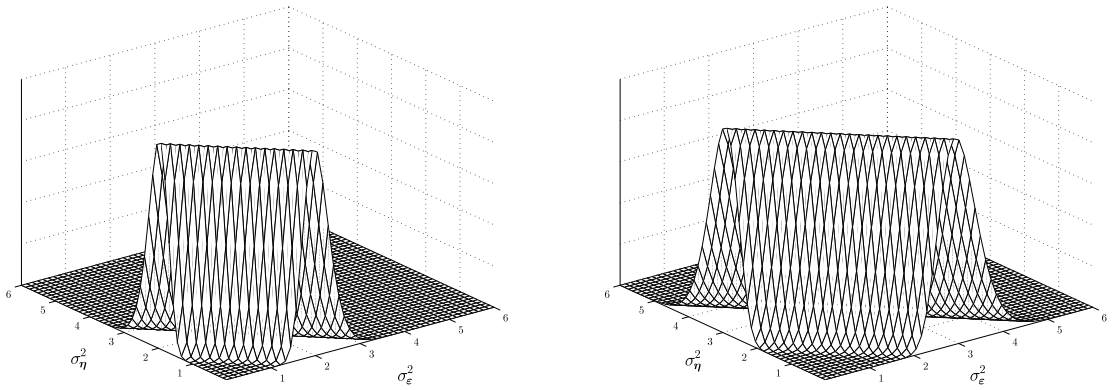
yields that the unconditional mean and variance of y are given by $\mathcal{E}[y] = \iota_T\theta$ and $\mathcal{V}[y] = (\sigma_\varepsilon^2 + \sigma_\eta^2)\mathrm{I}_T$. The same result follows from the joint posterior density which, after integrating out θ, is given by

$$p(\sigma_\eta^2, \sigma_\varepsilon^2|y) = (\sigma_\eta^2 + \sigma_\varepsilon^2)^{-\frac{1}{2}(T-1)} \exp\left(-\frac{1}{2}\frac{(y - \iota_T\hat{\theta})'(y - \iota_T\hat{\theta})}{\sigma_\eta^2 + \sigma_\varepsilon^2}\right) \tag{60}$$

Clearly, only the total variance is identified, not the individual components. Furthermore, the roles of σ_ε^2 and σ_η^2 are interchangeable. This holds true for any value of the *signal-to-noise* ratio which is defined as $\mathrm{SN} = \sigma_\eta^2/\sigma_\varepsilon^2$. Figure 12 shows the joint density for signal-to-noise ratios of 1 and 0.5. Panels (a) and (b) show that irrespective of the signal-

[17] Note that Hobert and Casella (1996) assume a Jeffreys' prior as a result of which the Inverted Gamma density for σ_η^2 has one degree of freedom since in their example $T = 2$.

Figure 12: **Joint posterior density of σ_ε^2 and σ_η^2 with a uniform prior**



(a) SN = 1 (b) SN = 0.5

Notes: Panel (a) and (b) show the joint density in (60) with a signal-to-noise ratio of 1 and 0.5 respectively. For both panels y was simulated from (56)-(57) with $\theta = 1$ and for panel (a) $\sigma_\varepsilon^2 = \sigma_\eta^2 = 1$ whereas for panel (b) $\sigma_\varepsilon^2 = 2, \sigma_\eta^2 = 1$ was used.

to-noise ratio the joint density is perfectly symmetrical. It is also clear from the figure that the joint density will always have a ridge. Note that everywhere along this ridge the sum of the variance components is the same. This becomes evident by first defining $\xi = \varepsilon + \eta$ and $\sigma_\xi^2 = \sigma_\varepsilon^2 + \sigma_\eta^2$ and then recognizing the resulting model as the basic linear regression model which only has a single variance component. The model in (56)-(57) basically splits up this single component into two components which explains the ridge. However, because this ridge is on a bounded domain the joint density is nevertheless integrable[18]. The Gibbs sampler can therefore be used to obtain posterior results. The Gibbs step is given by

- generate	$\theta^{(j)} \vert \sigma_\varepsilon^{2(j-1)}, \sigma_\eta^{2(j-1)}$	from	$p\left(\theta \vert y, \sigma_\varepsilon^2, \sigma_\eta^2\right) \sim \mathcal{N}\left(\hat{\theta}, \frac{1}{T}(\sigma_\varepsilon^{2(j-1)} + \sigma_\eta^{2(j-1)})\right)$
- generate	$\sigma_\varepsilon^{2*(j)} \vert \theta^{(j)}, \sigma_\eta^{2(j-1)}$	from	$p\left(\sigma_\varepsilon^{2*} \vert y, \theta, \sigma_\eta^2\right) \sim \mathcal{IG}\left(\frac{1}{2}(y - \iota_T\theta^{(j)})'(y - \iota_T\theta^{(j)}), \frac{1}{2}(T-2)\right)$
- generate	$\sigma_\eta^{2*(j)} \vert \theta^{(j)}, \sigma_\varepsilon^{2(j)}$	from	$p\left(\sigma_\eta^{2*} \vert y, \theta, \sigma_\varepsilon^2\right) \sim \mathcal{IG}\left(\frac{1}{2}(y - \iota_T\theta^{(j)})'(y - \iota_T\theta^{(j)}), \frac{1}{2}(T-2)\right)$

where $\sigma_\varepsilon^{2*(j)} \equiv \sigma_\varepsilon^{2(j)} + \sigma_\eta^{2(j-1)}$ and $\sigma_\eta^{2*(j)} \equiv \sigma_\eta^{2(j)} + \sigma_\varepsilon^{2(j)}$ are Inverted Gamma distributed random variables which have been shifted to the right by an amount of σ_η^2 and σ_ε^2 respectively. Note that this is again Gibbs sampling without sampling μ directly. From the latter two conditional densities it is clear that the role of the two variance components is interchangeable. The dynamic processes in (56) and (57) have an identical structure. The result is an identification issue since it is impossible to distinguish σ_ε^2 from σ_η^2.

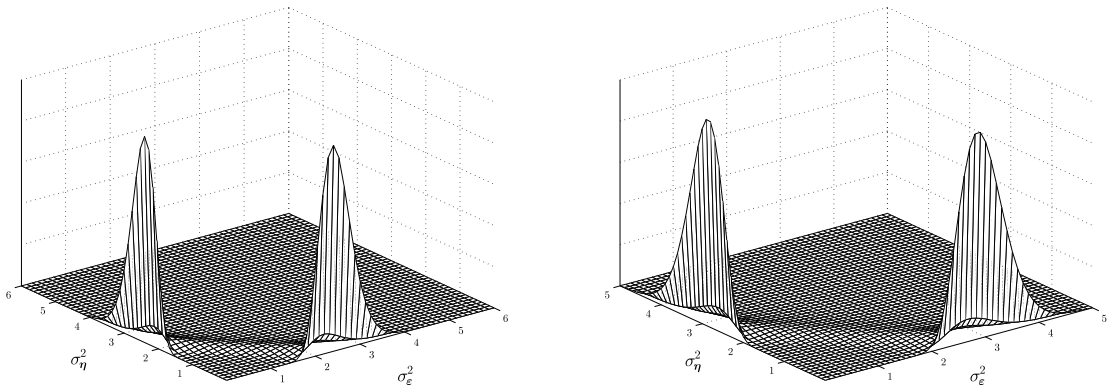
A further problem arises when instead of a uniform prior, a Jeffreys'-type prior is used, $p(\boldsymbol{\theta}) \propto \frac{1}{\sigma_\varepsilon^2}\frac{1}{\sigma_\eta^2}$, in which case the joint density becomes

$$p(\sigma_\eta^2, \sigma_\varepsilon^2|y) = \frac{1}{\sigma_\varepsilon^2}\frac{1}{\sigma_\eta^2}\left(\frac{1}{\sigma_\eta^2 + \sigma_\varepsilon^2}\right)^{\frac{1}{2}(T-1)} \exp\left(-\frac{1}{2}\frac{(y - \iota_T\hat{\theta})'(y - \iota_T\hat{\theta})}{\sigma_\eta^2 + \sigma_\varepsilon^2}\right) \tag{61}$$

[18]The density shown in Figure 4(a) on the other hand has a ridge on the domain $[0, \infty) \times [0, \infty)$ which makes it non-integrable.

Figure 13 shows that the Jeffreys' prior causes the joint density to shoot off to infinity for either $\sigma_\varepsilon^2 \to 0$ or $\sigma_\eta^2 \to 0$ [19]. Therefore, the joint posterior is now improper and the

Figure 13: **Joint posterior density of σ_ε^2 and σ_η^2 with a Jeffreys' prior**



(a) SN = 1 (b) SN = 0.5

Notes: Panels (a) and (b) show the joint density in (61) with a signal-to-noise ratio of 1 and 0.5 respectively. For both panels y was simulated from (56)-(57) with $\theta = 1$ and for panel (a) $\sigma_\varepsilon^2 = \sigma_\eta^2 = 1$ whereas for panel (b) $\sigma_\varepsilon^2 = 2$ and $\sigma_\eta^2 = 1$ was used.

Gibbs sampler will not converge. In Hobert and Casella (1996), Theorem 1, a number of conditions are stated that ensure propriety of the posterior density in HLMM models. Note that the Jeffreys' prior violates condition (a) of the theorem, while a *uniform* prior leads to a proper posterior, see also Gelman (2006).

Solutions

A number of solutions exist to prevent the problems presented in case (i) and (ii). For case (i) increasing the number of observations and assuming that the variance is identical across all observations will prevent the degrees of freedom problem. To solve the identification issue of case (ii) one can proceed in a number of ways. One possibility of dealing with this problem is to impose an identifiability constraint on the variance components, for example, $\sigma_\varepsilon^2 > \sigma_\eta^2$. Imposing this constraint in the Gibbs sampler aids in classifying the Gibbs draws to either of the variance components. However, it should be noted that 'identification' is only coming from the constraint and not in any way from the data.

Another possibility is to extend the basic HLMM in such a way that one can distinguish σ_ε^2 from σ_η^2. Two possible directions can be taken here. The first direction is to change the dynamics of μ by changing the specification of the model in (56)-(57) to that of a State-Space model. The variance components can then be identified from the additional imposed model structure. The second direction is to use a second source of information. Including additional information via more dependent variables in a Panel Data model enables one to identify σ_η^2 from the cross-sectional observations.

[19] Although Figure 13 is similar in shape as Figure 5 the two figures have a very different interpretation. Whereas Figure 5 shows a density that has two well-defined modes (albeit far apart) the density in Figure 13 is only well behaved in the domain $(\delta, \infty) \times (\delta, \infty)$ for a δ that is sufficiently far away from zero. The latter density goes to infinity when either of the variance components tends to zero.

4.3 State Space Model

Starting from the HLMM in the previous paragraph we can specify a State-Space model (SSM) by introducing time-series dynamics for the latent variable. Specifying a random walk for the state variable μ_t gives

$$\begin{aligned} y_t &= \mu_t + \varepsilon_t, \qquad \text{with} \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2) \qquad \text{and} \qquad t = 1, \ldots, T \qquad (62) \\ \mu_t &= \mu_{t-1} + \eta_t, \qquad \text{with} \qquad \eta_t \sim \mathcal{N}(0, \sigma_\eta^2) \qquad \text{and} \qquad \mathcal{E}[\varepsilon_t \eta_s] = 0 \qquad (63) \end{aligned}$$

with $\boldsymbol{\theta} = (\mu, \sigma_\varepsilon^2, \sigma_\eta^2)$. This model, which is generally known as the *local level* model, see Harvey (1989), is a basic specification of a State-Space model and has been studied extensively in the literature, e.g. Koop and van Dijk (2000).

More elaborate State-Space models are easily obtained by including explanatory variables in the measurement equation (62) and state equation (63), see Hamilton (1994) or Kim and Nelson (1999) for an overview.

The main tool for drawing inference in State-Space models is the Kalman Filter. This recursive procedure computes the optimal estimate of the unobserved state vector μ given the data y and values for the remaining parameters, see Kim and Nelson (1999) for more details. Popular algorithms for drawing Bayesian inference in State-Space models are given in Carter and Kohn (1994), De Jong and Shephard (1995) and Durbin and Koopman (2001).

The specification in (63) implies that μ_t is a random walk process which follows from recursively substituting μ_{t-1}, μ_{t-2} etc. Due to the additional structure of the State Space model one can now distinguish σ_ε^2 from σ_η^2 and therefore identify both variance components.

Gibbs Sampling

We explain the Gibbs step in a SSM by means of a model that is slightly more complicated than the local level model,

$$\begin{aligned} y_t &= x_t \beta_t + \varepsilon_t, \qquad \text{with} \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2) \qquad \text{and} \qquad t = 1, \ldots, T \qquad (64) \\ \beta_t &= \beta_{t-1} + \eta_t, \qquad \text{with} \qquad \eta_t \sim \mathcal{N}(0, \Sigma_\eta) \qquad \text{and} \qquad \mathcal{E}[\varepsilon_s \eta_{k,t}] = 0 \qquad (65) \end{aligned}$$

with x_t a $(1 \times K)$ vector of explanatory variables, β_t the $(K \times 1)$ state vector with individual elements $\beta_{k,t}$ for $k = 1, \ldots, K$ and Σ_η a $(K \times K)$ diagonal covariance matrix with diagonal elements $\sigma_{\eta,k}^2$ for $k = 1, \ldots, K$. We use this model in an empirical illustration below. It is convenient to first factorize the likelihood when deriving the Gibbs conditional densities. From the hierarchical structure of the model it follows that

$$p(y|\boldsymbol{\beta}, \sigma_\varepsilon^2, \sigma_\eta^2) = p(y|\boldsymbol{\beta}, \sigma_\varepsilon^2) p(\boldsymbol{\beta}|\sigma_\eta^2)$$

where $\boldsymbol{\beta}$ is the $T \times K$ matrix of latent states. Furthermore, we use $\boldsymbol{\beta}_k$ to denote the k^{th} column of $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_t$ to denote the t^{th} row of $\boldsymbol{\beta}$. $p(\boldsymbol{\beta}|\sigma_\eta^2)$ has to be factorized further down to individual elements $p(\beta_{k,t}|\beta_{k,t-1})$. It is now straightforward to show that the Gibbs step in this case is given by[20]

[20] If one allows for correlation between the errors in the transition equation one would have to generate draws for Σ_η from an Inverted Wishart density which is given in for example Poirier (1995).

- generate	$\boldsymbol{\beta}^{(j)}\|\sigma_{\varepsilon}^{2\,(j-1)}, \Sigma_{\eta}^{(j-1)}$	from	$p\left(\boldsymbol{\beta}\|y, \sigma_{\varepsilon}^2, \Sigma_{\eta}\right)$	$\sim$ KFS
- generate	$\sigma_{\varepsilon}^{2\,(j)}\|\boldsymbol{\beta}^{(j)}, \Sigma_{\eta}^{(j-1)}$	from	$p\left(\sigma_{\varepsilon}^2\|y, \boldsymbol{\beta}, \Sigma_{\eta}\right)$	$\sim \mathcal{IG}\left(\frac{1}{2}(y - X\boldsymbol{\beta}^{(j)})'(y - X\boldsymbol{\beta}^{(j)}), \frac{1}{2}(T-2)\right)$
- generate	$\sigma_{\eta,k}^{2\,(j)}\|\boldsymbol{\beta}^{(j)}, \sigma_{\varepsilon}^{2\,(j)}$	from	$p\left(\sigma_{\eta,k}^2\|y, \boldsymbol{\beta}, \sigma_{\varepsilon}^2\right)$	$\sim \mathcal{IG}\left(\frac{1}{2}(\boldsymbol{\beta}_k^{(j)} - \boldsymbol{\beta}_{-1,k}^{(j)})'(\boldsymbol{\beta}_k^{(j)} - \boldsymbol{\beta}_{-1,k}^{(j)}), \frac{1}{2}(T-2)\right)$

where KFS represents the Kalman Filter Sampler using one of the above mentioned algorithms.

Empirical Illustration: US Money Growth

We estimate the time-varying model parameter model used by Kim and Nelson (1989), and discussed in Kim and Nelson (1999), Application 2, pp. 44-48. Kim and Nelson (1989) use maximum likelihood estimation together with the Kalman filter to estimate the following time-varying parameter model

$$\begin{aligned} \Delta \mathrm{M}_t &= \beta_{0,t} + \beta_{1,t}\Delta i_{t-1} + \beta_{2,t}\mathrm{INF}_{t-1} + \beta_{3,t}\mathrm{SURP}_{t-1} + \beta_{4,t}\Delta \mathrm{M}_{t-1} + \varepsilon_t \quad (66) \\ \beta_{k,t} &= \beta_{k,t-1} + \eta_{k,t} \quad \text{for} \quad k = 0, \ldots, 4 \quad (67) \\ \varepsilon_t &\sim \mathcal{N}(0, \sigma_{\varepsilon}^2), \quad \eta_{k,t} \sim \mathcal{N}(0, \sigma_{\eta,k}^2) \quad \text{and} \quad \mathcal{E}[\varepsilon_t \eta_{k,s}] = 0. \quad (68) \end{aligned}$$

where ΔM_t is the M1 growth rate, Δi_{t-1} the change in the 3-Month Treasury Bill rate, INF_t the CPI inflation rate and SURP_t the detrended full employment budget surplus. The dataset used consisted of quarterly US data for the period 1964:I-1985:IV.

Here we repeat the Kim and Nelson (1989) study with two more years worth of data (1962:I-1963:IV) but, more importantly, we use Gibbs sampling to obtain posterior results. In particular, we use the Carter and Kohn (1994) algorithm to sample the time-series for the latent variables. Table 2 shows posterior moments for the variance components whereas Figure 14 shows the time-series of the posterior means for the state variables $\boldsymbol{\beta}_k$, $k = 0, \ldots, 4$.

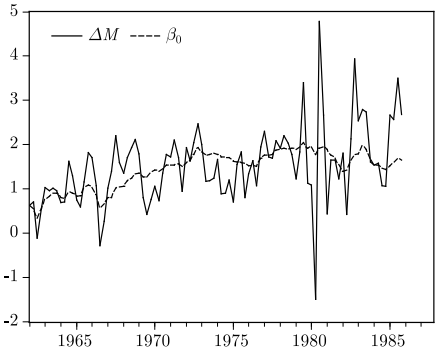
Table 2: **Posterior results for the State-Space Model**

σ_{ε}^2	$\sigma_{\eta,0}^2$	$\sigma_{\eta,1}^2$	$\sigma_{\eta,2}^2$	$\sigma_{\eta,3}^2$	$\sigma_{\eta,4}^2$
0.102	0.069	0.019	0.058	0.078	0.007
(0.060)	(0.060)	(0.026)	(0.032)	(0.087)	(0.008)

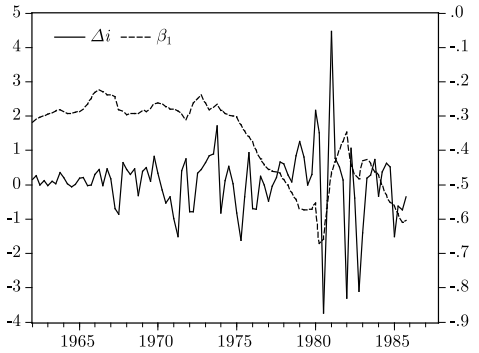
Notes: The table shows posterior means and standard deviations (in between brackets) for the variance components of model (66)-(68). Quarterly data for the period 1962:I-1985:IV for M1 growth, changes in the 3-Month Treasury Bill rate, CPI inflation rate and detrended full employment budget surplus were used. The data were obtained from the website for Kim and Nelson (1999). `http://www.econ.washington.edu/user/cnelson/SSMARKOV.htm`. Posterior results are based on 100,000 draws after a burn-in of $B = 10,000$ draws and selecting every $h = 10^{\text{th}}$ draw.

Table 2 and Figure 14 show that there is a strong indication that the coefficients in (66) are time-varying. As explained in Kim and Nelson (1999) this indicates that the way in which the US Federal Reserve reacts to changes in various macroeconomic variables, when conducting its monetary policy, varies over time. Especially the change in parameters around the Volcker period (beginning of the 1980s) is striking and very similar for all parameters.

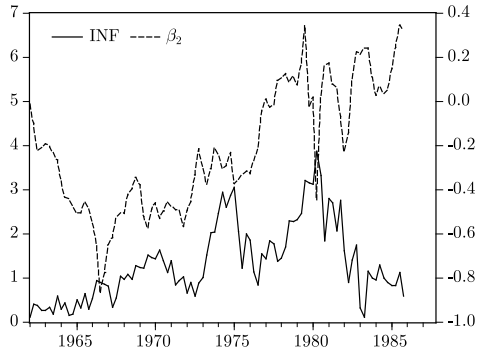
Figure 14: **Time-varying parameters in the State-Space Model**



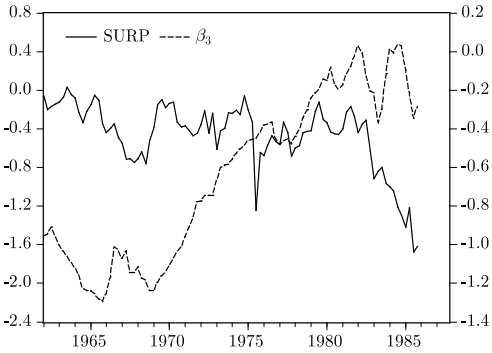
(a) ΔM and $\boldsymbol{\beta}_0$



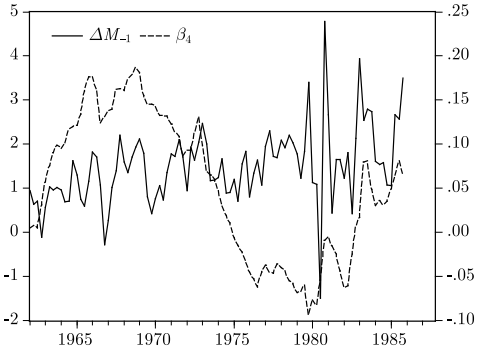
(b) Δi and $\boldsymbol{\beta}_1$



(c) INF and $\boldsymbol{\beta}_2$



(d) SURP and $\boldsymbol{\beta}_3$



(e) ΔM_{-1} and $\boldsymbol{\beta}_4$

Notes: The graphs show the posterior means for the time-varying parameters in the model (66)-(68). Panel (a) shows ΔM and $\boldsymbol{\beta}_0$ whereas panel (b)-(e) show $\boldsymbol{\beta}_k$ for $k = 1, \ldots, 4$ with the accompanying exogenous variables. In each graphs, the scale for $\boldsymbol{\beta}_k$ corresponds to the right axes. Posterior results are based on 100,000 draws after a burn-in of $B = 10,000$ draws and selecting every $h = 10^{\text{th}}$ draw.

4.4 Panel Data Model

The attractive feature of Panel Data models is that by using time-series observations as well as cross-sectional information, one can control for time-varying and cross-section specific variables as well as account for unobserved heterogeneity. The cross-sectional information results from including multiple dependent variables in the model. By grouping dependent variables that are hypothesized to have similar characteristics one can then proceed to identify the parameters for each group. Extensive discussions on panel data models can be found in recent textbooks by Baltagi (2001), Arellano (2002) and Hsiao (2003), among others. As an example of Panel Data models we discuss the following *random effects* model in which we allow for only a single group

$$\begin{aligned} y_{i,t} &= \mu_i + \varepsilon_{it}, \qquad \text{with} \qquad \varepsilon_{i,t} \sim N(0, \sigma_\varepsilon^2) \qquad \text{and} \qquad t = 1, \ldots, T,\ i = 1, \ldots, N, \quad (69) \\ \mu_i &= \theta + \eta_i, \qquad \text{with} \qquad \eta_t \sim N(0, \sigma_\eta^2) \quad (70) \end{aligned}$$

with $\boldsymbol{\theta} = (\mu, \theta, \sigma_\eta^2, \sigma_\varepsilon^2)$ where $\mu = (\mu_1\ \mu_2 \ldots \mu_N)'$. The double subscript on y reflects that one now has observations across time as well as across groups. The model allows for differences in mean, μ_i, across individuals by modelling these as random draws for a (Normal) distribution with mean θ and variance σ_η^2. As before, the vector μ, which contrary to the State-Space model is now constant over time but varies across groups, consists of latent variables and can be sampled alongside the other parameters in the Gibbs sampler. Note that inference on σ_η^2 is based on the cross-sectional observations whereas for σ_ε^2 variation across the cross-section as well as over time is utilized. Therefore, by including data on multiple individuals, the identification issues for the variance components do not exist. However, inference is only possible if a group consists of a sufficient number of individuals otherwise a degrees of freedom issue emerges. Throughout this section we assume a uniform prior on the parameters.

Gibbs Sampling

As for the State-Space model, the likelihood for the Random Effects Panel model can be factorized as

$$p(Y|\mu, \theta, \sigma_\varepsilon^2, \sigma_\eta^2) \propto p(Y|\mu, \sigma_\varepsilon^2) p(\mu|\theta, \sigma_\eta^2)$$

The matrix Y contains the observations on all individuals for all time periods. We denote the time-series observations on the i^{th} individual by y_i (column i of Y) and the observations on all individuals at time t by the vector z_t (the t^{th} row of Y). Furthermore, define the overall sum of squares as

$$E'E = \left[\text{vec}(Y) - (\mathrm{I}_N \otimes \iota_T)\mu\right]' \left[\text{vec}(Y) - (\mathrm{I}_N \otimes \iota_T)\mu\right]$$

where vec() is the operator that stacks the columns of Y into a single vector of dimensions $TN \times 1$, $\otimes$ is the Kronecker product and I_N is a $(N \times N)$ identity matrix. Given these definitions, the Gibbs step can be shown to be,

- generate	$\mu_i^{(j)} \vert \theta^{(j-1)}, \sigma_\varepsilon^{2(j-1)}, \sigma_\eta^{2(j-1)}$	from	$p\left(\mu_i \vert Y, \theta, \sigma_\varepsilon^2, \sigma_\eta^2\right) \sim \mathcal{N}\left(M_i, \frac{\sigma_\varepsilon^{2(j-1)} \sigma_\eta^{2(j-1)}}{\sigma_\varepsilon^{2(j-1)} + T\sigma_\eta^{2(j-1)}}\right)$
- generate	$\theta^{(j)} \vert \mu^{(j)}, \sigma_\varepsilon^{2(j-1)}, \sigma_\eta^{2(j-1)}$	from	$p\left(\theta \vert Y, \mu, \sigma_\varepsilon^2, \sigma_\eta^2\right) \sim \mathcal{N}\left(\frac{1}{N}\iota_N \mu^{(j)}, \frac{1}{N}\sigma_\eta^{2(j-1)}\right)$
- generate	$\sigma_\varepsilon^{2(j)} \vert \mu^{(j)}, \theta^{(j)}, \sigma_\eta^{2(j-1)}$	from	$p\left(\sigma_\varepsilon^2 \vert Y, \mu, \theta, \sigma_\eta^2\right) \sim \mathcal{IG}\left(\frac{1}{2}E^{(j)\prime}E^{(j)}, \frac{1}{2}(TN-2)\right)$
- generate	$\sigma_\eta^{2(j)} \vert \mu^{(j)}, \theta^{(j)}, \sigma_\varepsilon^{2(j)}$	from	$p\left(\sigma_\eta^2 \vert Y, \mu, \theta, \sigma_\varepsilon^2\right) \sim \mathcal{IG}\left(\frac{1}{2}(\mu^{(j)} - \iota_N \theta^{(j)})'(\mu^{(j)} - \iota_N \theta^{(j)}), \frac{1}{2}(N-2)\right)$

Table 3: Posterior results for the Random Effects Panel Data Model

Country		$N=3$	$N=4$	$N=5$	$N=6$	$N=10$	$N=17$
	$\hat{\theta}$	1.292**	1.426***	1.542***	1.667***	1.882***	1.903***
		(0.562)	(0.506)	(0.449)	(0.407)	(0.311)	(0.208)
	$\hat{\sigma}^2_\varepsilon$	50.716	45.286	39.833	37.138	47.215	38.042
		(4.251)	(3.246)	(2.568)	(2.182)	(2.146)	(1.321)
	$\hat{\sigma}^2_\eta$	4.279	2.219	1.444	1.154	0.697	0.415
		(35.062)	(7.272)	(2.215)	(1.420)	(0.532)	(0.215)
Australia	$\hat{\mu}_1$	1.525**	1.563***	1.589***	1.629***	1.731***	1.752***
		(0.633)	(0.587)	(0.543)	(0.522)	(0.533)	(0.448)
Austria	$\hat{\mu}_2$	1.765***	1.785***	1.811***	1.842***	1.907***	1.908***
Belgium	$\hat{\mu}_3$	1.610**	1.642***	1.669***	1.706***	1.795***	1.808***
Canada	$\hat{\mu}_4$		1.883***	1.906***	1.938***	1.980***	1.976***
Denmark	$\hat{\mu}_5$			1.922***	1.953***	1.989***	1.987***
Finland	$\hat{\mu}_6$				2.224***	2.210***	2.185***
France	$\hat{\mu}_7$					1.932***	1.937***
Germany	$\hat{\mu}_8$					1.831***	1.841***
Italy	$\hat{\mu}_9$					2.151***	2.133***
Japan	$\hat{\mu}_{10}$					2.464***	2.417***
Netherlands	$\hat{\mu}_{11}$						1.846***
New Zealand	$\hat{\mu}_{12}$						1.588***
Norway	$\hat{\mu}_{13}$						2.271***
Sweden	$\hat{\mu}_{14}$						1.966***
Switzerland	$\hat{\mu}_{15}$						1.873***
UK	$\hat{\mu}_{16}$						1.677***
USA	$\hat{\mu}_{17}$						1.923***

Notes: The tables shows posterior means and standard deviations (in between brackets) for the random effects panel model (69)-(70) when applied to the full panel ($N = 17$), and several subsets ($N = 3, 4, 5, 6, 10$), of annual real per capita percentage GDP growth rates for 17 OECD countries. The sample period is 1900-2000 with GDP levels for 1900-1949 obtained from Maddison (1995) whereas those for 1950-1998 were obtained from Maddison (2001). For 1999 and 2000, the data were obtained from the GGDC Total Economy Database, `http://www.ggdc.net`. All the levels are measured in 1990 US dollars converted at Geary-Khamis purchasing power parities, see Maddison (1995) for a full description. We applied a log transformation to remove the exponential trend in GDP levels across time. Posterior results are based on 100,000 draws after a burn-in of $B = 10,000$ draws and selecting every $h = 10^{\text{th}}$ draw. *** indicates that zero is not contained in the 99% highest posterior density (HPD) region, ** indicates that zero is contained in the 99% but not in the 90% and 95% HPD region and * that zero is contained in the 99% and 95% but not in the 90% HPD region. Only posterior standard deviations for Australia are given. An Inverted Gamma density with parameters $r = 10^{-5}$ and $y = 1$ was used as the prior distribution for the variance components.

where M_i, for $i = 1, \ldots, N$, is defined as

$$M_i = \frac{\sigma_\eta^{2(j-1)}}{\sigma_\eta^{2(j-1)} + (1/T)\sigma_\varepsilon^{2(j-1)}} \iota_T z_t + \frac{\sigma_\varepsilon^{2(j-1)}}{T\sigma_\eta^{2(j-1)} + \sigma_\varepsilon^{2(j-1)}} \theta_i^{(j)} \tag{71}$$

The expression in (71) shows that draws for μ_i are based on a weighted average of the information in the cross section (through $\theta_i^{(j)}$) and the information in the time-series (through z_t) and that the weights are determined by the two variance components. See also Gelfand *et al.* (1990) for more details.

Empirical Illustration: Cross-Country GDP Growth

We use the Gibbs sampler to analyze the random effects model for a panel of OECD annual real per capita Gross Domestic Product growth rates (in %). The dataset consists of 17 industrialized countries which include Australia, Canada, New Zealand, Japan, the USA and 12 Western European countries, for the period 1900-2000. It should be noted that the set-up of the panel model that we consider here is very limited. For example, we assume that growth rates are independent across countries and that there is no autocorrelation in growth rates. Nevertheless, it may serve as a good starting-point from which to consider more elaborate models.

Table 3 shows posterior results for the full panel (final column) that includes all individual countries (as a single group). In the table we only report posterior standard deviations for Australia since those for the other countries are qualitatively similar. To obtain results we used a very weakly informative Inverted Gamma prior for the variance components which parameters $r = 10^{-5}$ and $y = 1$. With these parameter values, which satisfy the conditions given in Hobert and Casella (1996), the Inverted Gamma density is similar in shape as the flat prior, but, being Inverted Gamma, it remains a proper prior.

The mean growth rate θ of the 17 countries is estimated at 1.90%. Interestingly, some part of the variation in the data is due to cross-country differences in growth, which is reflected by the estimate of σ_η^2. The Scandinavian countries seem to have experienced the highest average growth rates over the twentieth century, as well as Italy and Japan, due to their postwar growth spurt. The Australian, New Zealand, and the UK economies witnessed comparatively low growth.

Apart from including all the countries we also estimated the model with fewer countries[21]. These results, which are shown in the first five columns of Tabel 3 corroborate the analytical results from section 4.1 which for a panel model translate to a minimum required number of individuals in a group. The results for $N = 3$ show that, compared to the results for larger N, the posterior mean and standard deviation for σ_η^2 are very large. Especially the standard deviation of 35.602 seems to indicate that the second moment does not exist. In fact, we know that with $N = 3$ neither posterior mean nor posterior standard deviation exists. Including at least one additional country helps to identify the mean but still not the variance of σ_η^2. From $N = 6$ onwards the variance seems to be more reasonable, although the values are still comparatively large.

We re-emphasize that this panel model is used for illustrative purposes only. For a more detailed of cross-country growth analysis over a long period we refer to, e.g. Barro (1991), Sala-i-Martin (1994) and Quah (1997).

5 Summary of Models Used and Lessons Learned

Using a set of basic economic time series models, we presented the results of a Bayesian analysis using Gibbs sampling. We considered models ranging from the Cochrane-Orcutt model for serial correlation in a regression set up and the Hierarchical Linear Mixed Model in a variance components set-up. In particular, we treated the Koyck model for Distributed Lag analysis and the Unit Root, Instrumental Variables, State-Space and Panel Data models. The prior information is rather diffuse. A summary of models used and lessons learned is presented in Table 4.

[21] We selected countries according to their alphabetical ordering in the full panel. Although this is somewhat arbitrary we expect results using a random selection of countries to be similar.

The major lesson learned from our analysis is that we recommend that every applied researcher investigates the shape of the posterior and/or the predictive density. More specifically, we have learned the following lessons:

Lesson 1. ***Gibbs sampling for regression parameters in models with serially correlated disturbances is simple and successful.***

Gibbs sampling in the context of the famous Cochrane-Orcutt model of a regression with possible first order serial correlation of the disturbances and using flat priors is relatively easy and gives finite sample results. This is in contrast with the standard classical asymptotic results using the Durbin-Watson test.

Lesson 2. ***Approaching the boundary of the regression parameter region in nearly nonstationary and nearly nonidentified economic processes implies that irregular shapes of the posterior density may occur.***

In many macro economic processes, the information in the data is weak and the mass of the likelihood function may be close to the boundary of the parameter region using flat priors. Examples are nearly nonstationary processes or nearly non-identified processes. We indicate what problems a Bayesian Gibbs sampler faces are when she reaches the boundary of the parameter region. Empirical examples illustrate the theoretical results. We do show that Bayesian finite sample analysis is possible while classical asymptotics breaks down. The warning signal is that the Gibbs sampler is much slower in convergence.

Lesson 3. ***Use smoothness or training priors to regularize the shape of the posterior in the models of Lesson 2.***

Since the use of uniform priors with the weak information in the likelihood is the reason for these results we indicate how one can make use of smoothing or regularization priors like the information matrix prior. Alternatively, one may use a training sample or initial value prior. Gibbs sampling is feasible in such cases.

Lesson 4. ***A simple model with a time varying variance explains the structure of linear hierarchical mixed models. These latter models serve as workhorse models for state space models and panel data models.***

The basic regression model with heteroscedasticity is a good workhorse to explain the structure of linear Hierarchical Mixed Models which itself is a parent class for two applications: State Space models and Panel data models.

Lesson 5. ***Approaching the boundary of the variance parameter in case of a low number of degrees of freedom or near identification of the variance component may lead to irregular shapes of the posterior.***

The degrees of freedom problem in variance components models is analyzed in time series and cross sections models, where we concentrate on right tail behavior of the posterior density of a variance parameter. Another issue is when one approaches a variance of zero. Then the shape of the likelihood of the variances may become irregular. In such case, the priors may be selected as uniform or inverted gamma. The uniform prior is attractive

when one wants to concentrate on data information near a zero variance. Gibbs sampling becomes slow when one is near the boundary.

Lesson 6. ***Informative dynamic structure in time series and a sufficient number of units in the cross section regularize the shape of the posterior.***

This is shown theoretically and empirically. The data sets refer to growth of Gross Domestic Product of several countries, to financial data such US money growth and Treasure Bill rate and to the well-known Lydia Pinkham sales and advertising data.

We end this paper with some remarks. In terms of methods we note that several other solutions help the Gibbs sampler in terms of convergence. Reparametrization of the model and subjectively determined informative and or predictive priors may help in avoiding some irregular shapes of the criterion functions. One may also leave the shape as it is and make use of more flexible sampling methods. As far as the class of models considered, we emphasize that discrete choice and switching regression models have not been investigated. These models are relatively well-known in Bayesian statistical and econometric literature and we refer to Koop (2003), Lancaster (2004) and Geweke (2005) for a more detailed analysis.

Table 4: Models used and lessons learned

	Model	Specification	Lesson learned
1.	Cochrane-Orcutt	$y_t = x_t\beta + \varepsilon_t$ $\varepsilon_t = \rho\varepsilon_{t-1} + \eta_t$	Flat prior on $(\infty < \beta < \infty, -1 < \rho < 1)$ gives finite sample posterior analysis on serial correlation which is easy using Gibbs
2.	Koyck	Replace x_t in Model 1. by $(1-\rho)\sum_{i=0}^{\infty}\rho^i x_{t-i}$	Flat prior gives improper posterior near $\rho = 1$. Gibbs sampling is slow near this boundary. Jeffreys' prior and training sample prior regularize posterior.
3.	Unit root	Replace x_t in Model 1. by the constant 1	Same lesson as for Model 2.
4.	Weak instruments	$y_t = x_t\beta + \varepsilon_t$ $x_t = z_t\pi + \nu_t$	Same lesson as for Model 2. but now at $\pi = 0$.
5.	Naive heteroscedasticity	$y_t = x_t\beta + \varepsilon_t$ $\varepsilon_t \sim N(0, \sigma_t^2)$	Flat prior on σ_t^2 gives improper posterior. Degrees of freedom restriction implies a constant σ^2 for at least three observations.
6.	Hierarchical	$y_t = \mu_t + \varepsilon_t$ $\mu_t = \theta + \eta_t$	Model is parent model for State Space model and Random Effects Panel Data model. Parameters σ_ε^2 and σ_η^2 are not identified with uniform priors.
7.	State Space model	Replace μ_t in Model 6. $\mu_t = \mu_{t-1} + \eta_t$	Additional structure in the state equation identifies the variance parameters σ_ε^2 and σ_η^2. A flat prior may give information on probability mass near boundary.
8.	Random Effects Panel Data model	$y_{it} = \mu_i + \varepsilon_{it}$ $\mu_i = \theta + \eta_i$	Using a flat prior and sufficient number of groups or individuals $(N \geq 3)$ yields that Gibbs sampling may work well.

References

Arellano, M. (2002), *Panel Data Econometrics*, Oxford University Press, New York.

Baltagi, B. H. (2001), *Econometric Analysis of Panel Data*, second edn., John Wiley & Sons, New York.

Barro, R. J. (1991), Economic growth in a cross section of countries, *Quarterly Journal of Economics*, 106, 407–443.

Bass, F. M. and D. G. Clarke (1972), Testing Distributed Lag Models o Advertising Effects, *Journal of Marketing Research*, 9, 298–308.

Bauwens, L., M. Lubrano, and J. F. Richard (1999), *Bayesian Inference in Dynamic Econometric Models*, Oxford University Press.

Box, G. and G. Tiao (1973), *Bayesian Inference in Statistical Analysis*, Addison-Wesley.

Carter, C. K. and R. Kohn (1994), On Gibbs Sampling for State Space Models, *Biometrika*, 81, 541–553.

Casella, G. and E. George (1992), Explaining the Gibbs Sampler, *The American Statistician*, 46, 167–174.

Chib, S. and E. Greenberg (1996), Markov Chain Monte Carlo Simulation Methods in Econometrics, 12, 409–431.

Clarke, D. G. (1976), Econometric Measurement of the Duration of Advertising Effect on Sales, *Journal of Marketing Research*, 13, 345–357.

De Jong, P. and N. Shephard (1995), The Simulation Smoother for Time Series Models, *Biometrika*, 82, 339–350.

De Pooter, M., R. Segers, and H. van Dijk (2006), Gibbs Sampling in Econometric Practice, *Econometric Institute Report 2006-13*.

Durbin, J. and S. J. Koopman (2001), *Time Series Analysis by State Space Models*, Oxford Statistical Science Series, Oxford.

Gelfand, A., S. Hills, A. Racine-Poon, and A. Smith (1990), Illustration of Bayesian Inference in Normal Data Models Using Gibbs Sampling, *Journal of the American Statistical Association*, 85, 398–409.

Gelfand, A. E. and A. F. M. Smith (1990), Sampling-Based Approaches to Calculating Marginal Densities, *Journal of the American Statistical Association*, 85, 398–409.

Gelman, A. (2006), Prior distributions for variance parameters in hierarchical models, *Bayesian Analysis*, 1, 515–533.

Gelman, A. and X.-L. Meng (1991), A Note on Bivariate Distributions That Are Conditionally Normal, *The American Statistician*, 45, 125–126.

Geman, D. and G. Reynolds (1992), Constrained Restoration and the Recovery of Discontinuities, *IEEE Transactions of Pattern Analysis and Machine Intelligence*, 14, 367–383.

Geman, S. and D. Geman (1984), Stochastic Relaxations, Gibbs Distributions and the Bayesian Restoration of Images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741.

Geweke, J. (1991), Efficient Simulation from the Multivariate Normal and Student-t Distributions Subject to Linear Constraints, in *Computing Science and Statistics: Proceedings of the Twenty-Third Symposium on the Interface*, Fairfax: Iterface Foundation of North American, Inc., 571-578.

Geweke, J. (1993), Bayesian Treatment of the Independent Student-t Linear Model, *Journal of Applied Econometrics*, 8, S19–S40.

Geweke, J. (1996), Bayesian Inference for Linear Models Subject to Linear Inequality Constraints, in W. Johnson, J. Lee, and A. Zellner (eds.), *Modeling and Prediction: Honoring Seymour Geisser*, New York: Springer-Verlag, 248-263.

Geweke, J. (1999), Using Simulation Methods for Bayesian Econometric Models: Inference, Development, and Communication, *Econometric Reviews*, 18, 1–126.

Geweke, J. (2005), *Contemporary Bayesian Econometrics and Statistics*, Wiley, New Jersey.

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter (2000), *Markov Chain Monte Carlo in Practice*, Chapman & Hall/CRC.

Griliches, Z. (1967), Distributed Lags: A Survey, *Econometrica*, 35, 16–49.

Hamilton, J. D. (1994), *Time Series Analysis*, Princeton University Press, Princeton, New Jersey.

Hamilton, J. D. (2006), Computing Power and the Power of Econometrics, *Medium Econometrische Toepassingen*, 14, 32–38.

Harvey, A. (1989), *Forecasting, Structurak Time Series Models and the Kalman Filter*, Cambridge University Press.

Heij, C., P. de Boer, P. H. Franses, T. Kloek, and H. K. van Dijk (2004), *Econometric Methods with Applications in Business and Economics*, Oxford University Press, New York.

Hobert, J. P. and G. Casella (1996), The Effect of Improper Priors on Gibbs Sampling in Hierarchical Linear Mixed Models, *Journal of the American Statistical Association*, 91, 1461–1473.

Hoogerheide, L. F., J. F. Kaashoek, and H. K. van Dijk (2006a), On the Shape of Posterior Densities and Credible Sets in Instrumental Variable Regressions Models with Reduced Rank: An Application of Flexible Sampling Methods Using Neural Networks, *Journal of Econometrics*, forthcoming.

Hoogerheide, L. F., H. K. van Dijk, and R. D. van Oest (2006b), *Simulation Methods for Bayesian Econometric Inference*, chap. Handbook of Computational Economics and Statistics, Elsevier.

Hsiao, C. (2003), *Analysis of panel data*, second edn., Cambridge University Press.

Hurn, M. and C. Jennison (1996), An Extension of Geman and Reynolds' Approach to Constrained Restoration and the Recovery of Discontinuities, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18, 657–662.

Kim, C.-J. and C. R. Nelson (1989), The Time-Varying-Parameter Model for Modeling Changing Conditional Variance: The Case of the Lucas Hypothesis, *Journal of Business & Economic Statistics*, 7, 433–440.

Kim, C.-J. and C. R. Nelson (1999), *State-Space Models with Regime Switching*, MIT Press, Cambridge, Massachusetts.

Kleibergen, F. and H. K. van Dijk (1998), Bayesian Simultaneous Equations Analysis Using Reduced Rank Structures, *Econometric Theory*, 701–743.

Kleibergen, F. R. and H. K. van Dijk (1994), On the Shape of the Likelihood/Posterior in Cointegration Models, *Econometric Theory*, 10, 514–551.

Koop, G. (2003), *Bayesian Econometrics*, Wiley-Interscience.

Koop, G. and H. K. van Dijk (2000), Testing for Integration using Evolving Trend and Seasonals Models:. A Bayesian Approach, *Journal of Econometrics*, 97, 261–291.

Koyck, L. M. (1954), *Distributed Lags and Investment Analysis*, North Holland Publishing Co, Amsterdam.

Lancaster, T. (2004), *An Introduction to Modern Bayesian Econometrics*, Blackwell Publishing.

Liu, J. S. (1994), The Collapsed Gibbs Sampler in Bayesian Computations with Applications to a Gene Regulation Problem, *Journal of the American Statistical Association*, 89, 958–966.

Maddison, A. (1995), *Monitoring the World Economy 1820-1992*, OECD Development Centre, Paris.

Maddison, A. (2001), *The World Economy - A Millenial Perspective*, OECD Development Centre, Paris.

Murrell, P. (2005), *R Graphics*, CRC Computer Science & Data Analysis, Chapman & Hall.

O'Hagan, A. (1994), *Kendall's Advanced Theory of Statistics*, Volume 2B, Bayesian Inference, London: Edward Arnold.

Palda, K. S. (1964), *The Measurements of Cumulative Advertising Effects*, Englewood Cliffs, New Jersey: Prentice Hall.

Poirier, D. J. (1995), *Intermediate Statistics and Econometrics*, MIT Press, London, England.

Quah, D. T. (1997), Empirics for Growth and Distribution: Stratification, Polarization, and Convergence Clubs, *Journal of Economic Growth*, 2, 27–59.

Raiffa, H. and R. Schlaifer (1961), *Applied Statistical Decision Theory*, Harvard Business School, Boston.

Sala-i-Martin, X. (1994), Cross-sectional regression and the empirics of economic growth, *European Economic Review*, 38, 739–747.

Schotman, P. and H. K. van Dijk (1991), A Bayesian Analysis of the Unit Root in Real Exchange Rates, *Journal of Econometrics*, 49, 195–238.

Smith, A. F. M. and G. O. Roberts (1993), Bayesian Computation via the Gibbs Sampler and Related Markov Chain Monte-Carlo Methods, *Journal of the Royal Statistical Society B*, 55, 3–23.

Tanner, M. A. and W. H. Wong (1987), The Calculation of Posterior Distributions by Data Augmentation, *Journal of the American Statistical Association*, 82, 528–550.

Tierney, L. (1994), Markov Chains For Exploring Posterior Distributions, *Annals of Statistics*, 22, 1701–1762.

Van Dijk, H. K. (1999), Some Remarks on the Simulation Revolution in Bayesian Econometrics, *Econometric Reviews*, 18.

van Dijk, H. K. (2003), On Bayesian Structural Inference in a Simultaneous Equations Models, in B. Stigum (ed.), *Econometrics and the Philosophy of Economics*, Princeton, New Jersey: Princeton University Press, 642-682.

Zellner, A., L. Bauwens, and H. K. van Dijk (1988), Bayesian Specification Analysis and Estimation of Simultaneous Equations Models Using Monte-Carlo Integration, *Journal of Econometrics*, 38, 39–72.

A Probability Density Functions

In this appendix several univariate and multivariate probability density functions are given which are used throughout this paper. For univariate densities, we indicate the k^{th} moment around the mean by μ_k whereas for multivariate densities these are indicated by $\boldsymbol{\mu}_k$. Upper case symbols always indicate vectors or matrices. More properties of the below densities and concise derivations of moment(-conditions) can be found in for example Raiffa and Schlaifer (1961) or Poirier (1995).

A.1 Univariate Densities

Normal density:
If Z is univariate Normally distributed with parameters m and s^2, i.e. $Z \sim \mathcal{N}(m, s^2)$, then the density of Z and its first two moments about the mean are given by

$$f_{\mathcal{N}}(z|m,s^2) \equiv \tfrac{1}{\sqrt{2\pi s^2}} \exp\left(-\tfrac{(z-m)^2)}{2s^2}\right) \qquad \text{for} \quad \begin{array}{r} -\infty < z < \infty \\ -\infty < m < \infty \\ 0 < s^2 < \infty \end{array} \tag{A-1}$$

$$\begin{array}{rcl} \mu_1 & = & m \\ \mu_2 & = & s^2 \end{array}$$

Student-t density:
If Z is univariate Student-t distributed with parameters m, s^2 and ν, i.e. $Z \sim t(m, s^2, \nu)$, then the density of Z and its first two moments about the mean are given by

$$f_z(z|m,s^2,\nu) \equiv \tfrac{\nu^{\frac{1}{2}\nu}}{B(\frac{1}{2},\frac{1}{2}\nu)} \sqrt{s^2} [\nu + \tfrac{(z-m)^2}{s^2}]^{-\frac{1}{2}(\nu+1)} \qquad \text{for} \quad \begin{array}{r} -\infty < z < \infty \\ -\infty < m < \infty \\ 0 < s^2 < \infty \\ \nu > 0 \end{array} \tag{A-2}$$

$$\begin{array}{rcll} \mu_1 & = & m & \text{for} \quad \nu > 1 \\ \mu_2 & = & \frac{\nu s^2}{\nu - 2} & \text{for} \quad \nu > 2 \end{array}$$

with $B(\frac{1}{2}, \frac{1}{2}\nu)$ the Bessel function defined as $B(p,q) \equiv \frac{(p-1)!(q-1)!}{(p+q-1)!}$

Inverted Gamma density:
If Z is univariate inverted gamma distributed with parameters y and ν, i.e. $Z \sim \mathcal{IG}(y, \nu)$, then the density of Z and its first two moments are given by

$$f_{\mathcal{IG}}(z|m,\nu) \equiv \tfrac{m^\nu}{\Gamma(\nu)} z^{-(\nu+1)} \exp\left(-\tfrac{m}{z}\right) \qquad \text{for} \quad \begin{array}{r} t \geq 0 \\ m, \nu > 0 \end{array} \tag{A-3}$$

$$\begin{array}{rcll} \mu_1 & = & \frac{m}{\nu - 1} & \text{for} \quad \nu > 1 \\ \mu_2 & = & \frac{m^2}{(\nu-1)^2(\nu-2)} & \text{for} \quad \nu > 2 \end{array}$$

with $\Gamma\nu)$ the Gamma function defined as $\Gamma(\nu) \equiv (\nu - 1)!$

A.2 Multivariate Densities

Multivariate Normal density:
If Z is multivariate Normally distributed with parameters m and S, i.e. $Z \sim \mathcal{N}(m, S)$, where Z and m are $(N \times 1)$ and S is $(N \times N)$, then the density of Z and its first two moments about the mean are given by

$$f_{\mathcal{N}}^{(N)}(z|m,S) \equiv (2\pi)^{-\frac{1}{2}N}|S|^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2}(z-m)'S^{-1}(z-m)\right) \qquad \text{for} \quad \begin{array}{l} -\infty < z < \infty \\ -\infty < m < \infty \\ z'Sz > 0 \;\forall\; z \neq 0 \end{array} \tag{A-4}$$

$$\begin{aligned} \boldsymbol{\mu}_1 &= m \\ \boldsymbol{\mu}_2 &= S \end{aligned}$$

Multivariate Student-t density:
If Z is multivariate Student-t distributed with parameters m, S and ν, i.e. $Z \sim t(m, S, \nu)$, where Z and m are $(N \times 1)$ and S is $(N \times N)$, then the density of Z and its first two moments about the mean are given by

$$f_t^{(N)}(z|m,S,\nu) \equiv \frac{\nu^{\frac{1}{2}\nu}\Gamma(\frac{1}{2}\nu+\frac{1}{2}N)}{\pi^{\frac{1}{2}N}\Gamma(\frac{1}{2}\nu)}|S|^{-\frac{1}{2}}[\nu+(z-m)'S^{-1}(z-m)]^{-\frac{1}{2}(\nu+N)} \qquad \text{for} \quad \begin{array}{l} -\infty < z < \infty \\ -\infty < m < \infty \\ \nu > 0 \\ z'Sz > 0 \;\forall\; z \neq 0 \end{array} \tag{A-5}$$

$$\begin{aligned} \boldsymbol{\mu}_1 &= m && \text{for} && \nu > 1 \\ \boldsymbol{\mu}_2 &= S^{-1}\tfrac{\nu}{\nu-2} && \text{for} && \nu > 2 \end{aligned}$$