Gibbs Sampling in Econometric Practice*

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

Econometric Institute and Tinbergen Institute
Erasmus University Rotterdam, The Netherlands

Econometric Institute Report 2006-13

March 21, 2006

Abstract

We present a road map for effective application of Bayesian analysis of a class of well-known dynamic econometric models by means of the Gibbs sampling algorithm. Members belonging to this class are the Cochrane-Orcutt model for serial correlation, the Koyck distributed lag model, the Unit Root 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 equation parameters and variance components and show that one should carefully scan the shape of the criterion function for irregularities before applying the Gibbs sampler. 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

Since the early business cycle analysis in macro economics, see the survey in e.g., Morgan (1990) and the references cited there, in particular Tinbergen (1939a,b), econometric practice has changed and increased substantially. These days an econometrician has an almost staggering amount of techniques at her disposal. Furthermore, modern data bank storage capacities provides an abundant store of empirical information resulting in the availability of, for example, high-frequency (stock) return data or supermarket scanner data. Due to these developments, substantial progress has been made in the empirical economic analysis in various areas such as international economics (convergence analysis between countries),

*We are very grateful to participants of the 10th International Conference on Computing in Economics and Finance, Amsterdam, 2004, and the 3rd World Conference on Computational Statistics & Data Analysis, Cyprus, 2005, 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)

finance (risk and return modelling), labour economics (duration and discrete choice models), marketing (customer choice modelling) to name just a few examples. An extensive survey on the advances in econometrics is beyond the scope of this paper, however. We refer the interested reader to textbooks such as Greene (2000) or Heij *et al.* (2004) for introductory examples and to Griliches and Intriligator (1983-1986), McFadden and Engle (1994), for more advanced surveys.

Another development is partly due to the advent of computers with ever-increasing computational power. This allows researchers to apply elaborate Bayesian simulation techniques for estimation where extensive use is made of pseudo-random number generators. One of these methods is Gibbs sampling, developed by Geman and Geman (1984), which has become a popular tool for analyzing a wide variety of problems. Judging from numerous articles in recent literature Gibbs Sampling is gaining more and more momentum. Whereas it has also found applications in classical statistics (see Tanner, 1991), since two decades the main developments using Gibbs sampling occur in Bayesian statistical and econometric analysis models. Recent textbooks such as Koop (2003) and Geweke (2005) discuss how Gibbs sampling is used in econometrics. However, an introductory analysis using some standard dynamic econometric models and simple analytical and graphical analysis is not directly available. One purpose of this paper is to fill that gap in the literature. We further note that graphics in the context of Bayesian analysis is becoming more and more important see, e.g., Murrell (2005). In our analysis we therefore also place emphasis on presenting results in a graphical way.

A second objective of this paper is to serve as a road map for econometric researchers interested in applying Gibbs sampling. One major advice for those "en route" is to investigate the shape of the criterion function of the parameters of interest (usually a posterior or predictive density). As long as one is on a path where 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. However, on the path where 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 need to be indicated. It depends on the specification of the model and the information in the data in which situation on the road map the researcher will find herself. We note that by its very nature of being introductory, this paper offers only a few guidelines on how to continue on the difficult latter path. In terms of possible guidelines one could think of a reparametrization of the model or using informative priors. Several other solutions are given in the literature such as for example the use of more flexible sampling methods, see e.g. Hoogerheide *et al.* (2006) but these are beyond the scope of this paper.

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 practice. In Section 3 we then start our analysis by applying the Gibbs sampler to a number of canonical models in econometric practice such as the Cochrane-Orcutt serial correlation model, the Koyck Distributed Lag model, the Unit Root model and Instrumental Variables models. Our focus in this section is primarily on drawing inference on equation (or level) parameters in the model and the issues involved when doing so. 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 concludes.

2 Gibbs Sampling and Typical Shapes of the Criterion Function

2.1 Gibbs Sampling

The Gibbs sampler belongs to the class of Markov Chain Monte Carlo (MCMC) simulation methods[1]. An attractive feature of MCMC techniques is that samples of random drawings can be generated from the *joint* densities of parameters of interest indirectly, without the need to specify the exact form of these densities directly[2]. The Gibbs sampler uses an iterative procedure to create Markov chains by simulating from *conditional* densities instead which are typically much easier to derive. The ultimate result is that the sets of draws that are obtained in this way can be effectively considered as samples from the joint posterior densities. Subsequently, results such as posterior means, standard deviations or density estimates can be constructed.

To illustrate the set-up of the Gibbs sampler, suppose we have a model with m parameters which are summarized in the parameter vector Θ, i.e. $\Theta = (\theta_1, \ldots, \theta_m)$. Suppose further that we have already derived the joint posterior density $p(\Theta|y)$. This density combines the data likelihood of the model given the parameters, $p(y|\Theta)$, with the prior distribution $p(\Theta)$ using Bayes' formula[3]

$$p(\Theta|y) = \frac{p(y|\Theta)p(\Theta)}{p(y)} \propto p(y|\Theta)p(\Theta) \tag{1}$$

The prior distribution reflects the researchers prior beliefs about Θ before observing the data y. These beliefs can be either informative or non-informative. In this paper we work primarily with a non-informative prior specification for Θ.

The Gibbs sampler now consists of generating samples from the joint density $p(\Theta, y)$ by iteratively drawing from the *conditional* densities $p(\theta_i|\Theta_{\setminus\{\theta_i\}}, y)$ where $\Theta_{\setminus\{\theta_i\}}$ denotes the parameter vector Θ without the i^{th} parameter. The conditional densities are to be determined from the joint posterior density. The Gibbs sampling algorithm is summarized schematically in Figure 1 and by means of the flow diagram in Figure 2.

Note that in the algorithm one always uses the most recent draws in the sense that $\Theta^{(j)}$ is constantly being updated. This is shown in more detail in the flow diagram in Figure 2 for a model with two parameters ($m = 2$). The result of the iterative procedure will be a Markov Chain consisting of a sequence of draws $\{\Theta^{(j)}\}_{j=1}^{J}$. It can be shown[4] that for large enough J this sequence of Gibbs draws is distributed according to the distribution of the marginal posterior densities $p(\theta_i|y)$, $i = 1, \ldots, m$ and can therefore be effectively considered as samples from these distributions.

Because it usually takes some time for the Markov Chain to converge it is common practice to discard the first B draws (these draws are referred to as the *burn-in draws*) are discarded. Consequently, posterior results will be based only on the draws $\{\Theta^{(j)}\}_{j=B+1}^{J}$. Furthermore, in the case of (strong) autocorrelation in the sequence of generated draws,

[1] We are necessarily brief in our explanation of the Gibbs sampler. See Casella and George (1992) for a more elaborate discussion.

[2] Of course, if one can do so then it is preferable to sample from these (marginal) densities directly.

[3] $p(y)$ is the marginal likelihood of the model. When applying the Gibbs sampler this is just a normalizing constant and can be left out. However, there are several applications, for example when computing Bayes factors, when one does need to compute the marginal likelihood. See Chib (1995) for more details.

[4] See Geweke (1999), Tierney (1994) and Smith and Roberts (1993) for details on the conditions of convergence of the Gibbs sampler.

Figure 1: Gibbs sampling: algorithm

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

2: Generate:

$\theta_1^{(j+1)}$ from $p\left(\theta_1 | \Theta_{\setminus\{\theta_1\}}^{(j)}\right)$

$\theta_2^{(j+1)}$ from $p\left(\theta_2 | \Theta_{\setminus\{\theta_2\}}^{(j)}\right)$

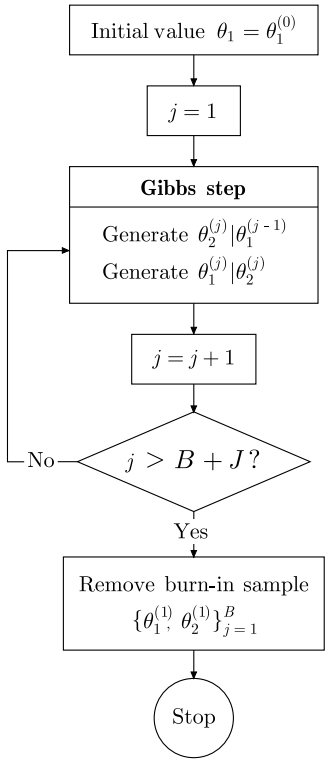
$\theta_3^{(j+1)}$ from $p\left(\theta_3 | \Theta_{\setminus\{\theta_3\}}^{(j)}\right)$

$\vdots$

$\theta_m^{(j+1)}$ from $p\left(\theta_m | \Theta_{\setminus\{\theta_m\}}^{(j)}\right)$

3: Set $j = j + 1$, go back to **2**. and repeat $B + J$ times.

Figure 2: Gibbs sampling: flow diagram



one should consider using only every h^{th} draw with h chosen such that $\Theta^{(j)}$ does not depend on $\Theta^{(j-h)}$. An alternative for generating posterior draws is to repeat the above algorithm a large number of times and use only the final draw of each sequence. Doing so means that the researcher does not have to worry about which values to use for B and h.

However, this method is computationally intensive.

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 analysis in Gelman and Meng (1991). Suppose that we have a model with parameter vector $\Theta = (\theta_1, \theta_2)$ and that the joint posterior density is of the following form

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

where a, b, c_1 and c_2 are constants under the restriction that $a \geq 0$ and if $a = 0$ then $|b| < 1$ [5]. This class of bivariate distributions is discussed in Gelman and Meng (1991) and follows from the assumption that the random variables θ_1 and θ_2 are conditionally Normally distributed. In fact, the conditional densities $p(\theta_1|\theta_2)$ and $p(\theta_1|\theta_2)$ can be recognized from (2) as Normal densities and are given by

$$p(\theta_1|\theta_2, a, b, c_1, c_2) \sim \mathcal{N}\left(\frac{b\theta_2 + c_1}{a\theta_2^2 + 1}, \frac{1}{a\theta_2^2 + 1} \right) \tag{3}$$

$$p(\theta_2|\theta_1, a, b, c_1, c_2) \sim \mathcal{N}\left(\frac{b\theta_1 + c_2}{a\theta_1^2 + 1}, \frac{1}{a\theta_1^2 + 1} \right) \tag{4}$$

By choosing different parameter configurations for a, b, c_1 and c_2 we can construct joint posterior densities of rather different shapes. Note, however, that the conditional densities will 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 way artificial since they are not based directly on a model and data, doing so may give us some insights into the strengths but also possible weaknesses of the Gibbs sampler before we move on to examining ecnometric models in subsequent sections.

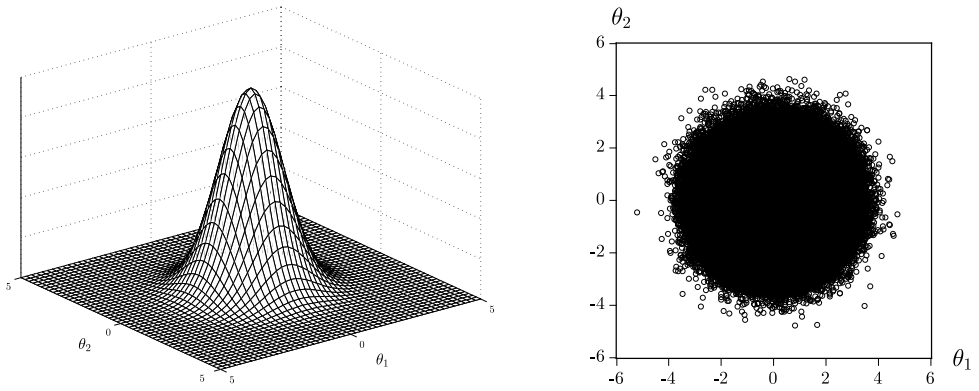
(i) Bell-shape

The first parameter configuration that we consider is the following; $(a = b = c_1 = c_2 = 0)$. In this case both conditional densities are standard Normal. The joint posterior density is bivariate Normal and is shown in Figure 3(a). Gibbs sampling simply comes down to obtaining draws by iteratively drawing from (3) and (4). A scatterplot of one million of such draws[6] 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, under 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 not necessary to use the Gibbs sampler. However, the aim of this example is simply to illustrate the straightforward approach of the Gibbs sampler and its usefulness for obtaining posterior results.

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

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

Figure 3: Gelman-Meng with $(a = b = c_1 = c_2 = 0)$



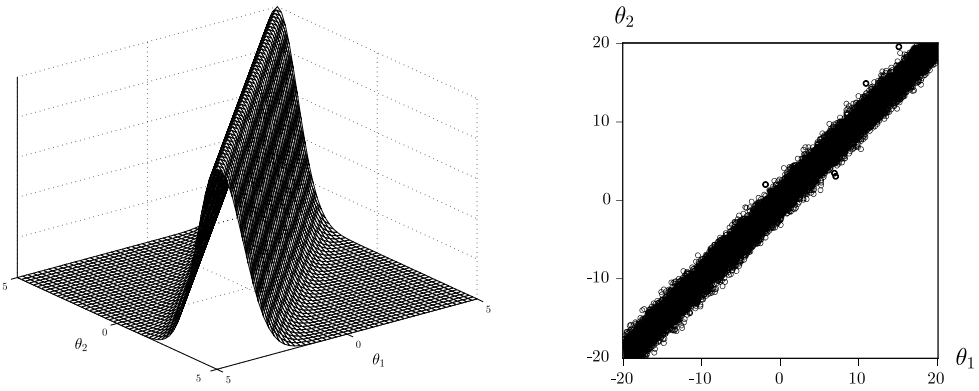
(a) Joint posterior density (b) Gibbs draws

Notes: Panel (a) shows the Gelman-Meng joint posterior density for θ_1 and θ_2 given in (2) 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 $(a = c_1 = c_2 = 0, b = 1)$. These values violate the earlier stated parameter restrictions. It is apparent from Figure 4(a) that the joint density in (2) is improper since the ridge along the line $\theta_1 = \theta_2$ causes it to be non-integrable. This is something that may not be immediately clear from the conditional

Figure 4: Gelman-Meng with $(a = c_1 = c_2 = 0, b = 1)$



(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 (2) 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.

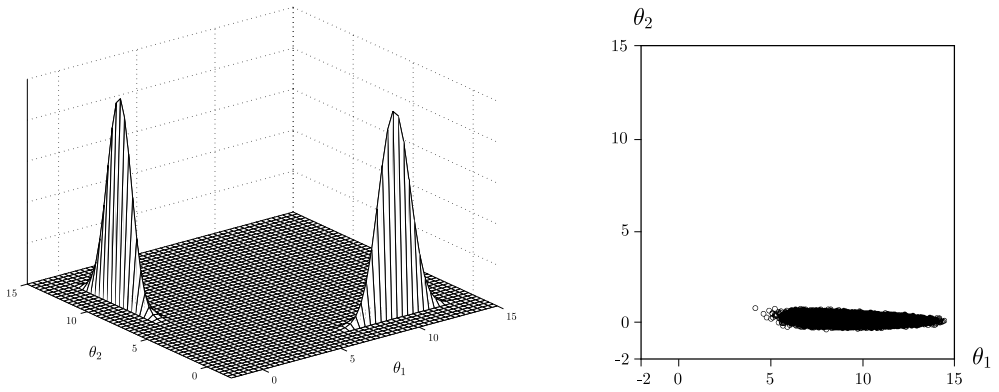
densities. The scatterplot of Gibbs draws for this example in Figure 4(b) does reveal the ridge $\theta_1 = \theta_2$ though. The posterior means and variances that we obtain from these draws will only be valid for this particular random sequence. Using different sequences

will result in completely different estimates. Although in this example it is clear that the Gibbs sampler will not converge at all this may not always be so obvious. Therefore one should carefully interpret the outcome from the Gibbs sampler as it can reveal information about a problem with the densities one is interested in.

(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 [7]. Unlike the previous example, the Gibbs draws will appear normal when in fact there is a problem. At first sight the scatterplot, shown in Figure 5(b), seems perfectly reasonable and posterior means and variances can easily be computed. However,

Figure 5: Gelman-Meng with $(a = 1, b = 0, c_1 = c_2 = 10)$



(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 (2) 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.

when inspecting the joint density as depicted in panel (a) of the same figure we see right away that the Gibbs sampler samples 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[8] which is due to the fact that the modes are too far apart in order 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 one million is already an insufficient number of draws to witness such a switch. Consequently, the Gibbs output only provides us which information of a subset of the full domain of $p(\theta_1, \theta_2)$ and posterior results are therefore incorrect. Although this example is a rather extreme case, it should be clear that multi-modality can result in very slow converge for the Gibbs sampler.

In summarizing the above examples of a bell-shaped, a ridge-shaped, and a bimodal-shaped density, we can say that although the Gibbs sampler is a straightforward technique, one should be careful not to apply it too mechanically. That is, it is essential to scrutinize a proposed model and the shape of its criterion function for any irregularities before moving

[7] See also Hoogerheide *et al.* (2006).

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

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 begin our analysis of the Gibbs sampler using typical workhorse models of econometric practice. Starting from the basic linear regression model we explore a number of extensions. First, we allow for serial correlation in the residuals of the model. Second, we examine time-series models where the regressors can be lagged values of the dependent variable. Third, we examine multivariate models where some of the dependent variables may be endogenous. The focus in this section will primarily be on drawing inference on the *equation* parameters in the models. In Section 4 we shift focus to considering variance components.

3.1 Basic Regression Model

We start our analysis with considering the basic regression model. This linear model attempts to explain the variance of a dependent variable y_t by 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):

$$y_t = x_t\beta + \varepsilon_t, \qquad t = 1, ..., T, \qquad \text{with} \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2) \tag{5}$$

To goal is to draw inference on the $(K \times 1)$ vector of equation 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

$$y = X\beta + \varepsilon \qquad \text{with} \qquad \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2 \mathrm{I}_T) \tag{6}$$

where y denotes the vector of T time-series or cross-sectional or time-series observations on the dependent variable, $y = (y_1 \ y_2 \dots y_T)'$, $X = (x_1 \ x_2 \dots x_T)'$ 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 or conditional densities. Note that in this study we use Θ to indicate the vector of model parameters. In this Θ is given by $\Theta = (\beta, \sigma_\varepsilon^2)$. Furthermore, we will always denote individuals or groups with the index i $(i = 1, \dots, N)$, time-series observations with t $(t = 1, \dots, T)$, exogenous variables and their corresponding parameter in β with k $(k = 1, \dots, K)$ and draws from the Gibbs sampler with j $(j = 1, \dots, J)$.

Gibbs Sampling

The likelihood for the model in (6) is given by

$$p(y|\Theta, X) = (2\pi\sigma_\varepsilon^2)^{-\frac{T}{2}} \exp\left[-\frac{1}{2\sigma_\varepsilon^2}(y - X\beta)'(y - X\beta)\right] \tag{7}$$

Combining the likelihood with a noninformative or uniform prior[9]

$$p(\Theta) \propto (\sigma_\varepsilon^2)^{-1} \tag{8}$$

[9] A non-informative prior for the equation 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.

yields the joint posterior density

$$p(\Theta|y,X) \propto (\sigma_\varepsilon^2)^{-\frac{(T+2)}{2}} \exp\left[-\frac{1}{2\sigma_\varepsilon^2}(y-X\beta)'(y-X\beta)\right] \tag{9}$$

A standard result is to rewrite (28) by completing the squares on β

$$p(\Theta|y,X) \propto (\sigma_\varepsilon^2)^{-\frac{(T+2)}{2}} \exp\left[-\frac{1}{2\sigma_\varepsilon^2}\left[(y-X\hat{\beta})'(y-X\hat{\beta})+(\beta-\hat{\beta})'X'X(\beta-\hat{\beta})\right]\right] \tag{10}$$

with $\hat{\beta}$ the OLS estimator $\hat{\beta} = [X'X]^{-1}X'y$. From (10) and the probability density functions given in Appendix A, it follows that the conditional density for β, given a value of σ_ε^2 is multivariate Normal with mean vector $\hat{\beta}$ and variance $\sigma_\varepsilon^2[X'X]^{-1}$, see equation (A-4). The conditional density for σ_ε^2 follows from (A-3) and is Inverted Gamma with location parameter $\frac{1}{2}(y-X\beta)'(y-X\beta)$ and $\frac{1}{2}T$ degrees of freedom. 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	$p(\sigma_\varepsilon^2\|y,X,\beta) \sim \mathcal{IG}\big(\frac{1}{2}(y-X\beta^{(j)})'(y-X\beta^{(j)}), \frac{1}{2}T\big)$

Graphical Illustration

To get a better understanding of what these densities look like graphically, we applied the model in (6) 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. To set it apart from parameters on explanatory variables, we relabel it with the symbol μ.

The 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 centered around this value. However, the value of σ_ε^2 determines the posterior variance of μ. For increasingly larger values of σ_ε^2 the posterior density clearly 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 density for σ_ε^2. Furthermore, the mean and variance of σ_ε^2 are lowest for μ close to the sample average. 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, see Appendix A, it is clear why; the value of both moments increase when μ deviates more from the sample mean.

3.2 The Cochrane-Orcutt Model

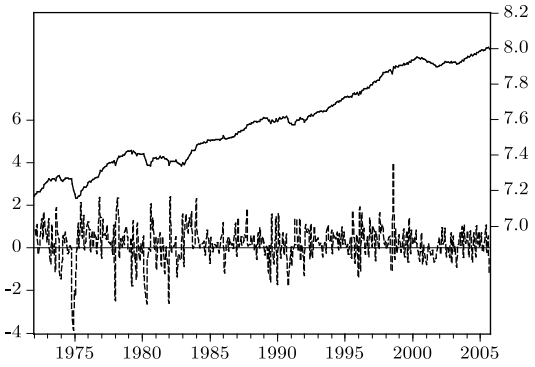
The second canonical model that we consider is the Cochrane-Orcutt model which extends the model in (5) by allowing the error term to have first order autocorrelation, that is:

$$y_t = x_t\beta + \nu_t, \qquad t = 1,...,T \tag{11}$$

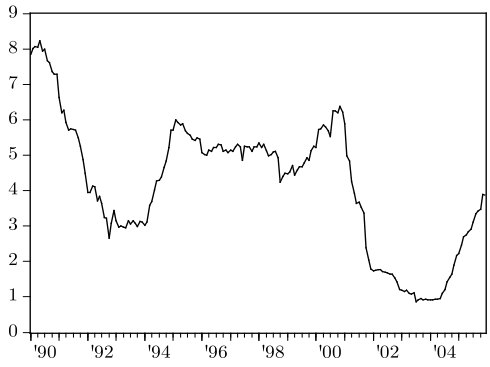
$$\nu_t = \lambda\nu_{t-1} + \varepsilon_t, \qquad \text{with} \qquad \varepsilon_t \sim \mathcal{N}(0,\sigma_\varepsilon^2) \tag{12}$$

where λ is the parameter that determines the strength of the autocorrelation. The domain of this unknown parameter is $-1 \leq \lambda \leq 1$. Θ is given by $\Theta = (\beta, \lambda, \sigma_\varepsilon^2)$. When $\lambda = 0$, the Cochrane-Orcutt model coincides with the basic regression model. As we will see later,

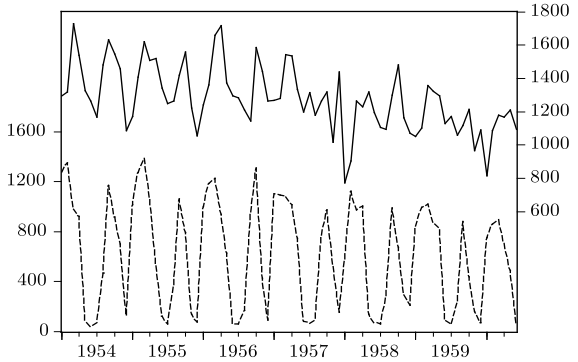
Figure 6: US Industrial Production, Lydia Pinkham Sales and Advertising series and 3-Month US Treasury Bill yield



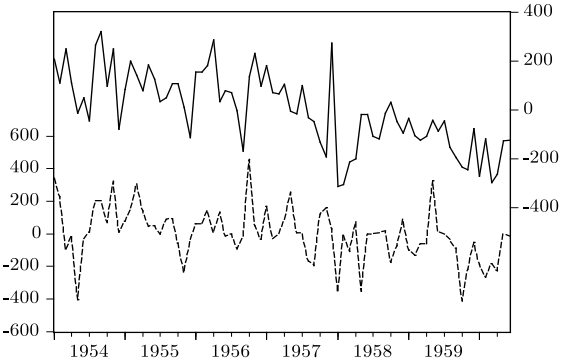
(a) US Industrial Production; level & rate (%)



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



(c) Unadjusted Lydia Pinkham series; sales and advertising



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

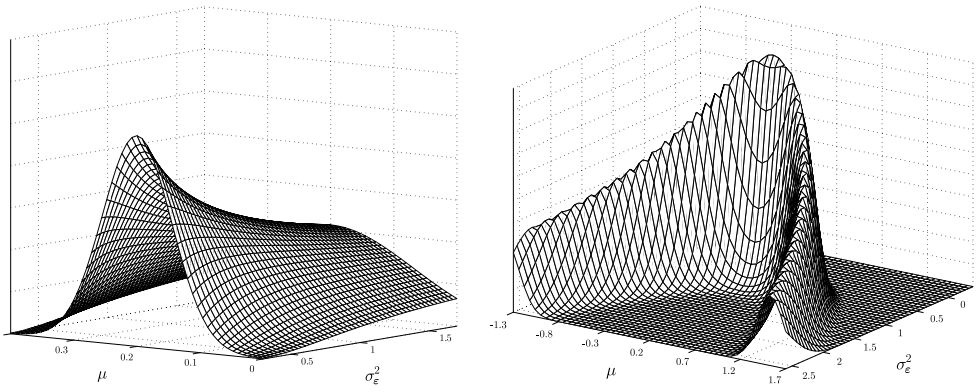
Notes: Panel (a) shows log levels (solid line) and growth rates (in % terms) for US Industrial Production (Gross Value of Products: Final products and nonindustrial supplies). The monthly series runs from January 1972 to September 2005 and was obtained from `http://www.economagic.com`. Panel (b) shows end-of month levels for the 3-Month US Treasury Bill 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 where taken from Palda (1964), Table 2, pp. 32-33.

the Gibbs sampler runs into difficulties when λ approaches the edge of its domain. By substituting (12) in (11) and rewriting the resulting expression in matrix notation, we have

$$y - \lambda y_{-1} = X\beta - X_{-1}\beta\lambda + \varepsilon, \qquad \text{with} \qquad \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2 \mathrm{I}_T) \tag{13}$$

where y_{-1} and X_{-1} contain the one-period lagged values of y and X.

Figure 7: Conditional posterior distributions



(a) Distribution of μ given σ_ε^2 (b) Distribution of σ_ε^2 given μ

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 (6), with $X=\iota_T$, to monthly US Industrial Production growth rates for the period January 1972-September 2005.

Gibbs Sampling

Combining the likelihood for this model with the same non-informative prior, as specified before in (8), the joint posterior density is as follows:

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

To facilitate the identification of the conditional densities it is convenient to rewrite (13) in two different ways. In each case we condition on one of the two types of equation parameters. First, we rewrite (14) conditional on values for λ:

$$y^* = X^*\beta + \varepsilon \qquad \text{where} \qquad \begin{cases} y^* = y^*(\lambda) \equiv y - \lambda y_{-1} \\ X^* = X^*(\lambda) \equiv X - \lambda X_{-1} \end{cases} \quad (15)$$

Second, conditional on values for β, (14) becomes:

$$\tilde{y} = \lambda\tilde{y}_{-1} + \varepsilon \qquad \text{where} \qquad \begin{cases} \tilde{y} = \tilde{y}(\beta) \equiv y - X\beta \\ \tilde{y}_{-1} = \tilde{y}_{-1}(\beta) \equiv y_{-1} - X_{-1}\beta \end{cases} \quad (16)$$

We can now use (15) to rewrite the joint posterior density. After completing the squares on β it follows immediately that the conditional density for β is Normal with mean and variance the OLS estimators in (15), i.e.

$$\hat{\beta} = (X^{*\prime}X^*)^{-1}X^{*\prime}y^* = \big[(X-\lambda X_{-1})'(X-\lambda X_{-1})\big]^{-1}(X-\lambda X_{-1})'(y-\lambda y_{-1}) \quad (17)$$

$$\sigma_\beta^2 = \sigma_\varepsilon^2(X^{*\prime}X^*)^{-1} = \sigma_\varepsilon^2\left[(X-\lambda X_{-1})'(X-\lambda X_{-1})\right]^{-1} \quad (18)$$

Using (16) it follows that the conditional density for λ is Normal with mean and variance given by

$$\hat{\lambda} = (\tilde{y}_{-1}{}'\tilde{y}_{-1})^{-1}\tilde{y}_{-1}{}'\tilde{y} = \big[(y_{-1}-X_{-1}\beta)'(y_{-1}-X_{-1}\beta)\big]^{-1}(y_{-1}-X_{-1}\beta)'(y-X\beta) \quad (19)$$

$$\sigma_\lambda^2 = \sigma_\varepsilon^2(\tilde{y}_{-1}{}'\tilde{y}_{-1})^{-1} = \sigma_\varepsilon^2\left[(y_{-1}-X_{-1}\beta)'(y_{-1}-X_{-1}\beta)\right]^{-1} \quad (20)$$

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)
λ	0.073*	(0.047)	λ	0.478***	(0.143)	λ	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 (14) with $X = \iota_T$ and $\beta = \mu$ for monthly US Industrial Production growth rates (January 1972-September 2005) in panel (a), the Koyck model (25) for the monthly Lydia Pinkham data (January 1954-June 1960) in panel (b) and the Unit Root model (43) 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. *, **, *** indicate that zero is not contained in the 90%, 95% and 99% highest posterior density (HPD) region, respectively.

Similar to the basic regression model, the conditional density for σ_ε^2 follows directly from the joint posterior density and is that of an Inverted Gamma with parameters

$$\frac{1}{2}\varepsilon'\varepsilon = \frac{1}{2}\big(y - \lambda y_{-1} - X\beta + X_{-1}\beta\lambda\big)'\big(y - \lambda y_{-1} - X\beta + X_{-1}\beta\lambda\big) \tag{21}$$

and $\frac{1}{2}T$ degrees of freedom. Summarizing, the j^{th} Gibbs step consists of

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

with the parameters of the conditional densities given in (17)-(21). Although (13) shows that the Cochrane-Orcutt model is nonlinear in the parameters β and λ which may hamper classical estimation and inference, estimating the parameters using Gibbs sampling is no problem. In fact, the joint posterior density for λ and any element of β resembles that in Figure 3(a). This is due to the fact that *conditional* on one equation parameter(s), the model for the other equation parameter(s) is again the basic linear regression model as shown in (11) and (12). Therefore, the Gibbs sampler is a very convenient approach for drawing inference on the parameters in these types of models. Furthermore, Bayesian testing for serial correlation is trivial by using Bayesian credible tests.

Empirical Illustration: US Industrial Production

To illustrate the above concepts, we 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 almost identical to the earlier reported sample average of 0.204 and the HPD region shows that there is weak evidence for serial correlation.

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. The reason is that it offers the econometrician with 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 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 and illustrate these by means of an empirical application using the well known Lydia Pinkham dataset. We also give some directions on how to prevent irregularities in the likelihood.

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

$$\begin{aligned} y_t &= \mu + \beta w_t + \nu_t, \qquad t = 1, ..., T && (22)\\ w_t &= (1-\lambda)\sum_{i=0}^{\infty} \lambda^i x_{t-i} && (23)\\ \nu_t &= \lambda \nu_{t-1} + \varepsilon_t, \qquad \text{with} \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2) && (24) \end{aligned}$$

Note that the effect of lagged values of the (single) explanatory variable x_t is determined solely by λ [11]. The parameter vector is given by $\Theta = (\beta, \lambda, \sigma_\varepsilon^2)$. Substituting (23) and (24) into (22) and rewriting the resulting expression[12] gives a first order distributed lag model for y_t in which the coefficient for x_t is a (nonlinear) function of the two equation parameters[13].

$$y_t = \lambda y_{t-1} + \beta(1-\lambda)x_t + \varepsilon_t \qquad \text{with} \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2) \tag{25}$$

In matrix notation, the model is given by:

$$y = \lambda y_{-1} + \beta(1-\lambda)x + \varepsilon, \qquad \text{with} \qquad \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2 \mathrm{I}_T) \tag{26}$$

with y_{-1} and I_T defined as before. The likelihood of the model is given by

$$p(y|\Theta, x) \propto (\sigma_\varepsilon^2)^{-\frac{T}{2}} \exp\left[-\frac{1}{2\sigma_\varepsilon^2}\big[y - \lambda y_{-1} - \beta(1-\lambda)x\big]'\big[y - \lambda y_{-1} - \beta(1-\lambda)x\big]\right] \tag{27}$$

Gibbs Sampling

Combining the likelihood with again a uniform prior yields the posterior density

$$p(\Theta|y, x) \propto (\sigma_\varepsilon^2)^{-\frac{(T+2)}{2}} \exp\left[-\frac{1}{2\sigma_\varepsilon^2}\big[y - \lambda y_{-1} - \beta(1-\lambda)x\big]'\big[y - \lambda y_{-1} - \beta(1-\lambda)x\big]\right] \tag{28}$$

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

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

[12] For ease of exposition we assume $\mu = 0$ in the remainder of the analysis.

[13] In the Koyck model without serial correlation, adding λy_{t-1} to the left and right side of (22) results in an Autoregressive Moving Average with exogenous regressors (ARMAX) specification: $y_t = \lambda y_{t-1} + \beta x_t + \nu_t - \lambda \nu_{t-1}$.

In a similar way as for the Cochrane-Orcutt model we can rewrite (26) by conditioning on one equation parameter at a time. This gives

$$y^* = \beta x^* + \varepsilon \qquad \text{where} \qquad \begin{cases} y^* = y^*(\lambda) \equiv y - \lambda y_{-1} \\ x^* = x^*(\lambda) \equiv (1-\lambda)x \end{cases} \tag{29}$$

and

$$\tilde{y} = \lambda \tilde{y}_{-1} + \varepsilon \qquad \text{where} \qquad \begin{cases} \tilde{y} = \tilde{y}(\beta) \equiv y - \beta x \\ \tilde{y}_{-1} = \tilde{y}_{-1}(\beta) \equiv y_{-1} - \beta x \end{cases} \tag{30}$$

The conditional posterior density for β again follows directly from rewriting the joint likelihood using (29) and is a Normal density with parameters

$$\hat{\beta} = (x^{*\prime}x^*)^{-1}x^{*\prime}y^* = \left[(1-\lambda)x'x\right]^{-1}x'(y-\lambda y_{-1}) \tag{31}$$

$$\sigma_\beta^2 = \sigma_\varepsilon^2(x^{*\prime}x^*)^{-1} = \sigma_\varepsilon^2\left[(1-\lambda)^2x'x\right]^{-1} \tag{32}$$

Using (30) it follows that the conditional density for λ is Normal with mean and variance given by

$$\hat{\lambda} = (\tilde{y}_{-1}'\tilde{y}_{-1})^{-1}\tilde{y}_{-1}'\tilde{y} = \left[(y_{-1}-\beta x)'(y_{-1}-\beta x)\right]^{-1}(y_{-1}-\beta x)'(y-\beta x) \tag{33}$$

$$\sigma_\lambda^2 = \sigma_\varepsilon^2(\tilde{y}_{-1}'\tilde{y}_{-1})^{-1} = \sigma_\varepsilon^2\left[(y_{-1}-\beta x)'(y_{-1}-\beta x\right]^{-1} \tag{34}$$

The conditional density for σ_ε^2 is Inverted Gamma with $\frac{1}{2}T$ degrees of freedom and parameter

$$\frac{1}{2}\varepsilon'\varepsilon = \frac{1}{2}\big(y-\lambda y_{-1}-\beta(1-\lambda)x\big)'\big(y-\lambda y_{-1}-\beta(1-\lambda)x\big) \tag{35}$$

The Gibbs step is given by

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

where the parameters are given by (31)-(35). At first sight, there may not seem to be any problems with applying the Gibbs sampler to the Koyck model. However, when examining the parameters of the conditional densities more closely it becomes evident that the Gibbs sampler runs into serious problems and can potentially even break down completely for values of λ close to 1. We will elaborate on this pathology via two directions. First, starting from (28) we work towards and subsequently analyze the marginal[14] density of λ which is a Student-t type density. Second, we consider the Information matrix for the Koyck model. Apart from providing insights on the irregularity in the likelihood for λ close to 1, the Information matrix will also provide us with a direction for a possible solution to tackle this irregularity.

In order to obtain the marginal densities for λ we first of all need to integrate out σ_ε^2 from (28). To do so we apply an Inverse Gamma integration step which can be derived

[14] We could also opt to analyze the conditional densities $p(\beta|y,x,\lambda,\sigma_\varepsilon^2)$ and $p(\lambda|y,x,\beta,\sigma_\varepsilon^2)$ directly but since such an analysis gives similar insights we do not pursue that approach here.

from (A-3), see also for example Bauwens *et al.* (2000), and which consists of the following proportionality

$$\int_0^\infty \left(\sigma^2\right)^{-\frac{(T+2)}{2}} \exp\left[-\frac{a}{2\sigma^2}\right] \mathrm{d}\sigma^2 \propto a^{-\frac{1}{2}T} \tag{36}$$

Applying this result and using the notation given in (29) we obtain the joint density of β and λ

$$p(\beta, \lambda | y, x) \propto \left[(y^* - \beta x^*)'(y^* - \beta x^*)\right]^{-\frac{T}{2}} \tag{37}$$

By completing the squares on β and some tedious but otherwise straightforward derivations we can rewrite this density as

$$p(\beta, \lambda | y, x) \propto \left[(T-1) + \frac{(\beta - \hat{\beta}) x^{*\prime} x^* (\beta - \hat{\beta})}{y^* M_{x^*} y^*}\right]^{-\frac{T}{2}} \left[\frac{x^{*\prime} x^*}{y^* M_{x^*} y^*/(T-1)}\right]^{\frac{1}{2}} \times \left[\frac{x^{*\prime} x^*}{y^* M_{x^*} y^*/(T-1)}\right]^{-\frac{1}{2}} \left[y^* M_{x^*} y^*\right]^{-\frac{T}{2}} \tag{38}$$

where we use 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 $P_A = A(A'A)^{-1}A'$.

As a final step, we need to integrate out β from (38). However, this can easily be done by recognizing that the first two terms in (38) are those of a Student-t density for β, see (A-2). Consequently,

$$\begin{aligned} p(\lambda | y, x) &\propto \left[x^{*\prime} x^*\right]^{-\frac{1}{2}} \left[y^* M_{x^*} y^*\right]^{-\frac{1}{2}(T-1)} \\ &\propto (1-\lambda)^{-1} \left[(y - \lambda y_{-1})' M_{x^*} (y - \lambda y_{-1})\right]^{-\frac{1}{2}(T-1)} \end{aligned} \tag{39}$$

After completing the squares on λ we can recognize (39), apart from the factor $(1-\lambda)^{-1}$, as another Student-t density, this time for λ. However, it is exactly this factor that can cause the Gibbs sampler to break down when λ is sufficiently close to 1 because $p(\lambda|y, x)$ is non-integrable for $\lambda = 1$. The joint density of β and λ, $p(\beta, \lambda|y, x)$, is constant at $\lambda = 1$ for $\infty < \beta < \infty$. Graphically, this means that the joint posterior density has a "wall". Figure 8(a) shows this for the joint density for a series of 3-Month Treasury Bill rates. The Fisher Information matrix, defined as minus the expectation of the matrix of second order derivatives of the log likelihood with respect to the parameter vector Θ, i.e. $\mathcal{I} = -E[\frac{\delta^2 \ln L(\Theta|y,x)}{\delta\Theta'\delta\Theta}]$, gives us a similar insight. It is fairly easy to show that the Information matrix of the Koyck model is given by[15]

$$\mathcal{I} = -E\begin{bmatrix} \frac{\delta^2 \ln L}{\delta\lambda^2} & \frac{\delta^2 \ln L}{\delta\lambda\delta\beta} & \frac{\delta^2 \ln L}{\delta\lambda\delta\sigma_\varepsilon^2} \\ \frac{\delta^2 \ln L}{\delta\beta\delta\lambda} & \frac{\delta^2 \ln L}{\delta\beta^2} & \frac{\delta^2 \ln L}{\delta\beta\delta\sigma_\varepsilon^2} \\ \frac{\delta^2 \ln L}{\sigma_\varepsilon^2\delta\lambda} & \frac{\delta^2 \ln L}{\delta\sigma_\varepsilon^2\delta\beta} & \frac{\delta^2 \ln L}{\delta\sigma_\varepsilon^4} \end{bmatrix} = \begin{bmatrix} \frac{T}{1-\lambda^2} & 0 & 0 \\ 0 & \frac{(1-\lambda)^2 x'x}{\sigma_\varepsilon^2} & 0 \\ 0 & 0 & \frac{T}{2\sigma_\varepsilon^4} \end{bmatrix} \tag{40}$$

By realizing that the Hessian is simply the inverse of the Information matrix, we can predict what happens with the Gibbs sampler when λ is close to 1. First of all, when a value for λ is drawn very close to 1 then because of the factor $(1-\lambda)^2$, the variance of

[15] We should note that we focus here on long term means only in which case $E[y] = E[y_{-q}] = X\beta$ for $q > 0$. 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.

$p(\beta|\lambda, \sigma_\varepsilon^2)$ goes to infinity which means that any value along the real line is equally likely to be drawn for β. Therefore, for any large draw of β, the variance of $p(\lambda|\beta, \sigma_\varepsilon^2)$ goes to zero as a result of which the next draw for λ is also going to be close to 1. This means that the Markov state is not guaranteed to converge because $\lambda = 1$ is an absorbing state which has a non-zero probability of being reached.

Potential Solutions

Our earlier analysis has shown that irregularities in the (marginal) densities are due to the factor $(1-\lambda)^{-1}$ when λ is close to 1. A number of potential solutions have been proposed in the literature to circumvent this problem. Popular approaches are either to restrict the parameter space for λ or to try to regularize the likelihood around $\lambda = 1$. Another approach is to reparameterize the model. Here we only briefly touch upon the several options to give the researcher a flavor of how to tackle pathological behaviour of the Gibbs sampler in these types of models. We do provide some references to studies that provide more in-depth analyses.

In terms of applying the first solution, one could truncate the domain of λ by imposing an upper bound which can be achieved by selecting an appropriate prior density. The goal would be to only allow draws for λ that are at least η away from 1 with $\eta > 0$. This prevents a "wall" like that in Figure 8(a). 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 with a suitable prior or, alternatively, 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 λ.

As for the second solution, the alternative could be to try and regularize the likelihood in the neighborhood of $\lambda = 1$. This can be achieved by replacing the prior in (8) with a prior that is chosen in such the way that it eliminates the factor $(1-\lambda)^{-1}$. From the Information matrix in (40) we can construct the following Jeffrey's prior[16]

$$p(\theta|\sigma_\varepsilon^2) \propto \frac{(1-\lambda)}{\sigma_\varepsilon^2} \qquad \text{for} \qquad 0 < \lambda < 1 \tag{41}$$

It is straightforward to show that with this prior the factor $(1-\lambda)^{-1}$ vanishes from (39). What basically happens is that the marginal density for λ is now integrable everywhere except for $\lambda = 1$ which in turn has a zero probability of occurring.

Another simple way of regularizing the posterior density is to use the following weakly informative prior on β as proposed in Schotman and van Dijk (1991)

$$p(\beta|\lambda, \sigma_\varepsilon^2) \propto \mathcal{N}\left(y_0, \tfrac{\sigma_\varepsilon^2}{(1-\lambda)^2}\right) \qquad \text{for} \qquad 0 < \lambda < 1 \tag{42}$$

where y_0 is the starting value for the time-series of y. The intuition behind this prior is that as λ approaches zero it becomes increasingly difficult to learn about β from the

[16] In general the Jeffrey's 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 $(1-\lambda)^1$ instead of $(1-\lambda)^{\frac{1}{2}}$ to regularize the likelihood. For more details and an advanced analysis on similar Jeffrey's priors we refer to Kleibergen and van Dijk (1994, 1998).

data since the mean of y, which depends on β, does not exist when λ equals 1. The prior is stronger for smaller values of λ but approaches an uninformative prior for $\lambda \to 1$. It it derived from the unconditional distribution of y_0 under the assumption of Normality. The effect of this Normal prior on the joint posterior density $p(\lambda, \beta|y, x, \sigma_\varepsilon^2)$ is shown in Figure 8(b). The pronounced wall feature in Figure 8(a) is no longer visible although the posterior still flattens out for $\lambda \to 1$.

Other 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 λ) 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 (22)-(24) on the famous Lydia Pinkham dataset (see Palda, 1964). This dataset has been used extensively in marketing studies to investigate the dynamic relationship between advertising and sales. Figure 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, 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 $\lambda^{(j)}$ that are close to unity. We therefore truncated the domain of λ 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 λ is 0.48, which implies that 90% of the advertising effect has taken place after approximately nine weeks[17]. This result is similar to the results documented in prior 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).

3.4 The Unit Root Model

A model that initially may seem be very different from the Koyck model is the unit root model. However, by choosing $x = \iota_T$ and relabelling β by μ in (25) we arrive at a first order autoregressive model for y

$$y_t - \mu = \lambda(y_{t-1} - \mu) + \varepsilon_t \tag{43}$$

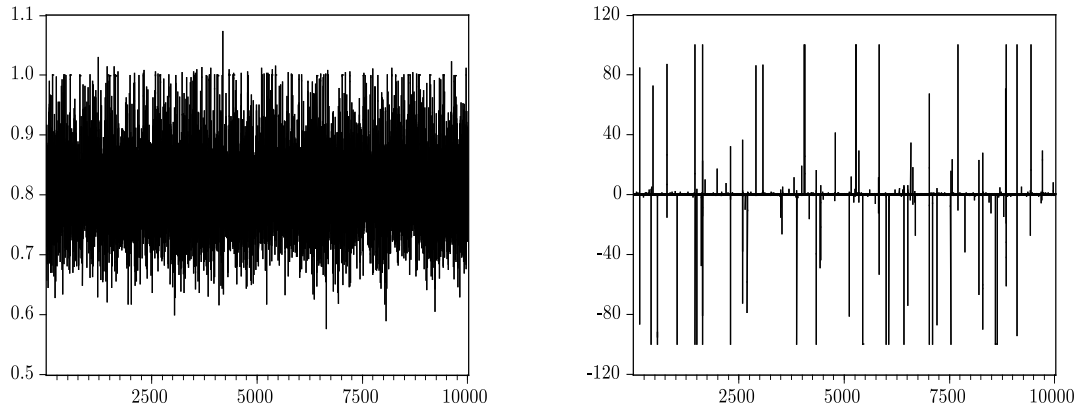
where μ is the unconditional mean of the time-series $\{y_t\}_{t=1}^T$. In this model, the interpretation of μ depends on whether the series y is stationary ($\lambda < 1$) or whether it has a unit root ($\lambda = 1$). In the latter case, the mean of y does not exist and μ is non-identified. Therefore, as discussed above, any value for μ along the real line is likely to be drawn when λ is close to unity. This will not only distort the posterior mean of λ but also causes the sequence of draws for λ to have difficulties moving away from $\lambda = 1$.

[17] 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 $\Delta_t = \ln(1-\alpha)/\ln(\hat{\lambda}) - 1$, see Clarke (1976), pp. 348.

Empirical Illustration: US Treasury Bill Rates

To illustrate, we obtain posterior results for the unit root model using monthly data on the 3-month US Treasury Bill. This series portraits unit root type behavior as is evident from Figure 6(b). This is corroborated by posteriors results from the Gibbs sampler. The posterior mean of λ 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 9(a) shows that a substantial fraction of the draws for λ are

Figure 8: Gibbs draws for the Unit Root Model with a non-informative prior



(a) Gibbs draws for λ

(b) Gibbs draws for μ

Notes: Gibbs draws for λ, panel (a), and μ, panel (b). Shown are 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 Θ. In the model we use the 3-Month US Treasury Bill rates for the period January 1990-December 2005 for the data vector y.

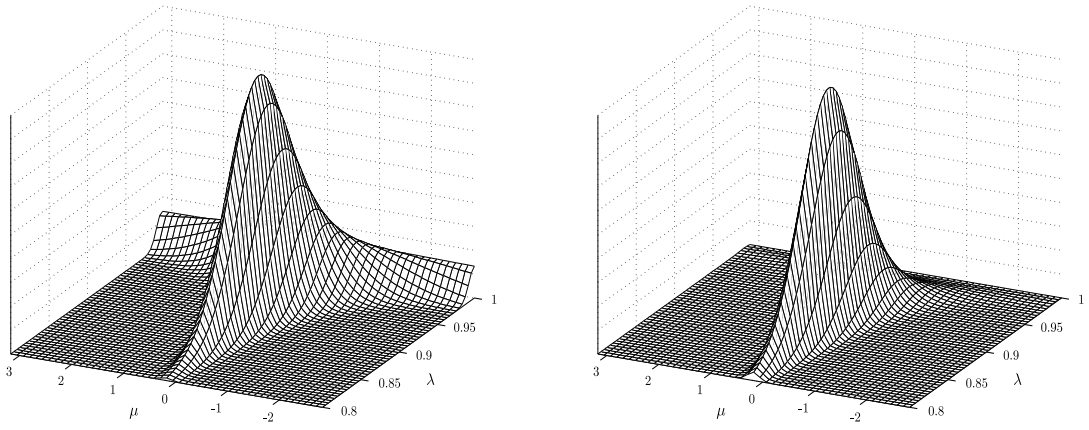
close to 1 and that the draws for μ are therefore all over the place. Note that it may not be satisfactory to truncate the domain of λ at $1-\eta$ here. In this case, draws for λ around 1 are in the region of the distribution that is of particular interest, as we expect λ to be close to unity. Imposing the Schotman and van Dijk (1991) prior in (42) on the other hand seems more appropriate. Doing so removes the wall in the joint density at $\lambda = 1$ as shown in Figure 8(b). Table 1, panel (c) shows posterior results. The posterior mean (standard deviation) for μ and λ 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 type 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 similarly, 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 Simulta-

Figure 9: Joint posterior density of 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 (8) whereas panel (b) shows the same posterior density however now with the prior proposed by Schotman and van Dijk (1991) as given in (42). In both panels we use the 3-Month US Treasury Bill rates for the period January 1990-December 2005 for the data vector y.

neous Equation Model (INSEM). Our analysis, which is necessarily brief, is based on van Dijk (2003) and Hoogerheide *et al.* (2006) 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)[18]

$$y = w\beta + \varepsilon \tag{44}$$

$$w = x\pi + \nu \tag{45}$$

$$[\varepsilon \ \nu]' \sim \mathcal{N}\left([0 \ 0]', \Sigma\right) \qquad \text{with} \qquad \Sigma = \begin{bmatrix} \sigma_\varepsilon^2 & \sigma_{\varepsilon,\nu} \\ \sigma_{\varepsilon,\nu} & \sigma_\nu^2 \end{bmatrix} \tag{46}$$

with y, w and x all having dimensions $(T \times 1)$ and β and π being scalar parameters. Θ is given by $\Theta = (\beta, \pi, \Sigma)$. In this model, y is be to interpreted as the structural variable of interest, w is an endogenous variable and x 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 w in the equation for y. (44)-(46) is known as the *structural form* of the INSEM. By substituting (45) in (44) we can derive the *reduced form* which is given by

$$y = x\pi\beta + \xi \tag{47}$$

$$w = x\pi + \nu \tag{48}$$

$$[\varepsilon \ \nu]' \sim \mathcal{N}\left([0 \ 0]', \Sigma\right) \qquad \text{with} \qquad \Sigma = \begin{bmatrix} \sigma_\varepsilon^2 & \sigma_{\varepsilon,\nu} \\ \sigma_{\varepsilon,\nu} & \sigma_\nu^2 \end{bmatrix} \tag{49}$$

with $\xi = \varepsilon + \nu$. We can interpreted the reduced form model as a multivariate regression model which is nonlinear in the parameters β and λ as in (25). As was the case in the unit root model, this nonlinearity can lead to a non-identifiability problem. In particular,

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

when $\pi = 0$, the likelihood, and therefore also the joint posterior density if we assume a noninformative prior, is flat and nonzero in the direction of β. In fact, the joint density will look very similar to that in Figure 8(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, w and x are all matrices and therefore 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 convergence of the Gibbs sampler will be slow since $\pi = 0$ acts as a absorbing state. In fact, the *marginal* density for π is non-integrable because of infinite probability mass at $\pi = 0$ (see Kleibergen and van Dijk, 1998) which is something that may not be altogether clear from the Gibbs outcomes which should therefore be interpreted with care. A possible solution to avoid pathologies in INSEM models would again be to specify sensible prior densities. However, it is difficult to find sensible 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 equation parameters in univariate as well as multivariate models. As long as the researcher finds herself in a region of the parameter space where the likelihood is well behaved, for example in the basic linear regression model, then the Gibbs sampler can be used in a straightforward way for obtain posterior results. However, when parameters assume values for which the likelihood is irregular, as in the unit root model for $\lambda \to 1$, problems can arise. In that case any output from the Gibbs sampler has to be interpreted with care or if necessary, other measures need to be taken, such as imposing weakly informative priors.

4 Gibbs Sampling Within Variance Component and Unobserved Component Models

In this section our focus is on drawing inference on variance parameters instead of equation parameters. We do so by analyzing another workhorse model, the 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 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 small T

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

$$y_t = \mu + \varepsilon_t, \qquad t = 1, ..., T, \qquad \text{with} \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2) \tag{50}$$

If, instead of the previously used non-informative prior, we use the *uniform* prior

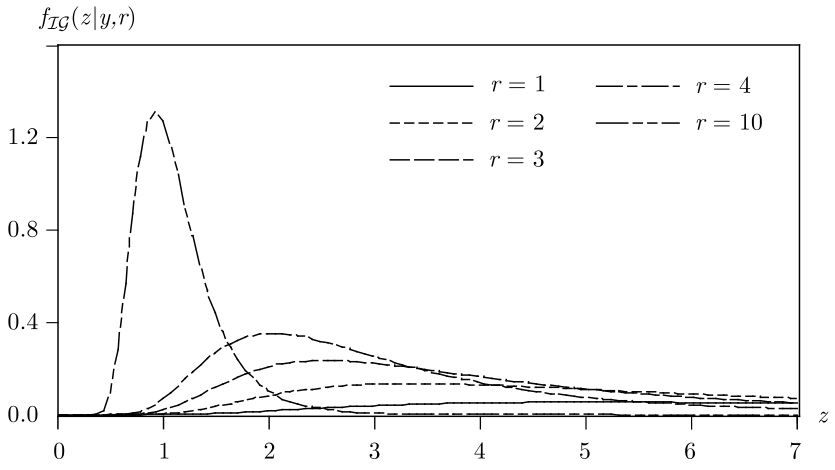
$$p(\Theta) \propto 1 \tag{51}$$

we can derive the marginal densities of μ and σ_ε^2 to be given by

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

with $\hat{\mu} = \frac{1}{T}\iota_T' y$ and $s^2 = y' M_{\iota_T} y$. These densities result from integrating the joint posterior with respect to σ_ε^2 and μ using the proportionality result of (36). These analytic results are necessary to analyze the convergence of the Gibbs step. From the conditions given in Appendix A it is clear in order for these Student-t and Inverted Gamma densities to exist that 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 case of $p(\mu|y)$ density and $T > 5$ for $p(\sigma_\varepsilon^2|y)$. Similarly, for the second moment to exist we need $T > 5$ and $T > 7$ respectively. A Jeffrey's 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 $T > 1$. For illustration, Figure 10 shows that

Figure 10: 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, r.

the right tail of an Inverted Gamma density distribution tends to zero at a too small rate, when the number of degrees of freedom of an Inverted Gamma density is too small. For instance, for $r = 2$ the first moment exists but not the second moment whereas for any $r > 2$ both moments exist. Note that the density in (A-3) is stated in terms of r. Therefore, from the likelihood of a model, the relation between r and T can be seen to be $r = \frac{1}{2}T$. A non-flat prior will change the right hand side of this relation with a scalar.

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 Jeffrey's prior $T = 1$ is sufficient. However, as follows from the analysis above, in both situations, the *marginal* densities for μ and σ_ε^2 do not exist, let alone their moments. 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. Therefore, the generated Gibbs sample does not make sense. In Section 4.4 we give an illustration using a panel data model.

Naive Heteroscedasticity

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

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

Clearly, inference in this model is impossible since there are more parameters than observations. 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 $\Theta = (\mu, \sigma_1^2, \ldots, \sigma_T^2)$. 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 such that is has a sufficient number of observations. For example, allowing for just two groups, inference in (52) is possible if it is imposed that

$$\sigma_t^2 = \begin{cases} \sigma_1^2 & \text{for } t = 1, ..., \tau \\ \sigma_2^2 & \text{for } t = \tau + 1, ..., T \end{cases} \tag{53}$$

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

Note that we specified both type of preliminary model using t as index for time-series observations ($t = 1, \ldots, T$). However, the analysis also holds for cross-sectional data ($i = 1, \ldots, N$). In the latter case, one could think of, for example, individuals or countries in a panel data model. Our main point, although as trivial as it may seem, is that one needs multiple observations to draw inference on variance components

4.2 Hierarchical Linear Mixed Model

We introduce the HLMM through the following hierarchical model with two variance components

$$y_t = \mu_t + \varepsilon_t, \qquad \text{with} \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2) \qquad \text{for} \qquad t = 1, \ldots, T \tag{54}$$

$$\mu_t = \theta + \eta_t, \qquad \text{with} \qquad \eta_t \sim \mathcal{N}(0, \sigma_\eta^2) \qquad \text{and} \qquad \mathcal{E}[\varepsilon_t \eta_s] = 0 \tag{55}$$

with $\Theta = (\theta, \mu, \sigma_\varepsilon^2, \sigma_\eta^2)$. 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 first analyze the base model in greater detail 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.e. $p(\Theta) \propto 1$.

(i) $\sigma_\varepsilon^2 = 1$ and T small

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. Only when there is a sufficient number of observations can sensible posterior results be obtained. 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[19]. The conditional densities $p(\theta|\sigma_\eta^2)$ and $p(\sigma_\eta^2|\theta)$ can be derived from first substituting (55) in (54)

$$y_t - \varepsilon_t = \theta + \eta_t \qquad \text{with} \qquad \eta_t \sim \mathcal{N}(0, \sigma_\eta^2) \tag{56}$$

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 $y_t^* \equiv y_t - \varepsilon_t$ and $\hat{\theta}$ the OLS estimator. Note that this is Gibbs sampling without having to concern oneself about μ. However, Gibbs sampling where μ is drawn alongside θ and σ_η^2 is also possible, see Hobert and Casella (1996).

(ii) σ_ε^2 unknown and T large

By taking T large enough, the researcher does no longer need to worry about possible non-existence of the posterior marginal densities. 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 a labeling issue in the sense that it not possible to distinguish the variance components from each other[20]. 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, rearranging the model (57) to

$$y = \iota_T\theta + \varepsilon + \eta \qquad \text{with} \qquad \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2 \mathrm{I}_T) \qquad \text{and} \qquad \eta \sim \mathcal{N}(0, \sigma_\eta^2 \mathrm{I}_T) \tag{57}$$

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

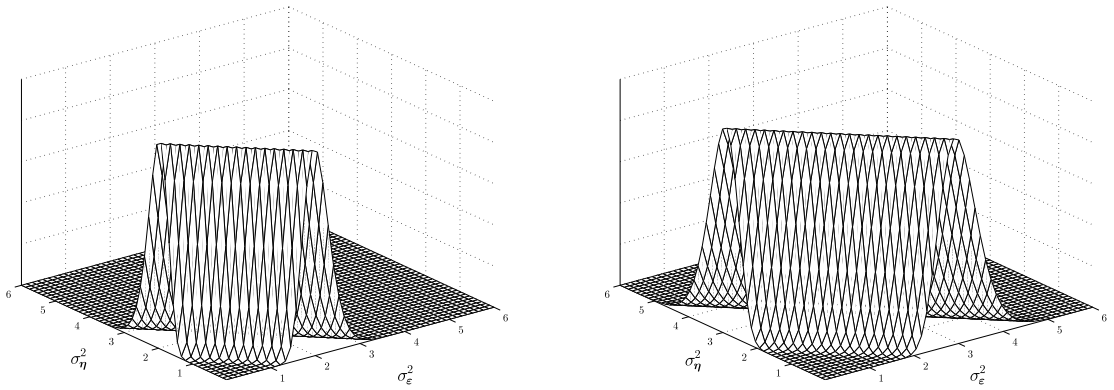
$$p(\sigma_\eta^2, \sigma_\varepsilon^2|y) = (\sigma_\eta^2 + \sigma_\varepsilon^2)^{-\frac{1}{2}(T-1)} \exp\left(-\frac{1}{2}\frac{(y - \iota_T\hat{\theta})'(y - \iota_T\hat{\theta})}{\sigma_\eta^2 + \sigma_\varepsilon^2} \right) \tag{58}$$

Clearly, 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 11 shows the joint density for signal-to-noise ratios of 1 and 0.5. Panels (a) and (b) show that irrespective of the signal-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

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

[20] A related problem is that of *label-switching*, see for example Redner and Walker (1984) or Frühwirth-Schnatter (2001).

Figure 11: 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 (58) with a signal-to-noise ratio of 1 and 0.5 respectively. For both panels y was simulated from (54)-(55) 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.

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 (54)-(55) 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[21]. 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 again obvious that the role of the two variance components is interchangeable. Therefore, it is impossible to distinguish σ_ε^2 from σ_η^2. The resulting labeling issue is caused by the fact that the variance components in (54) and (55) have exactly the same dynamics. Both are white noise processes across *time*. Because their dynamics are the same, one can never be certain which variance component is which in the Gibbs sampler.

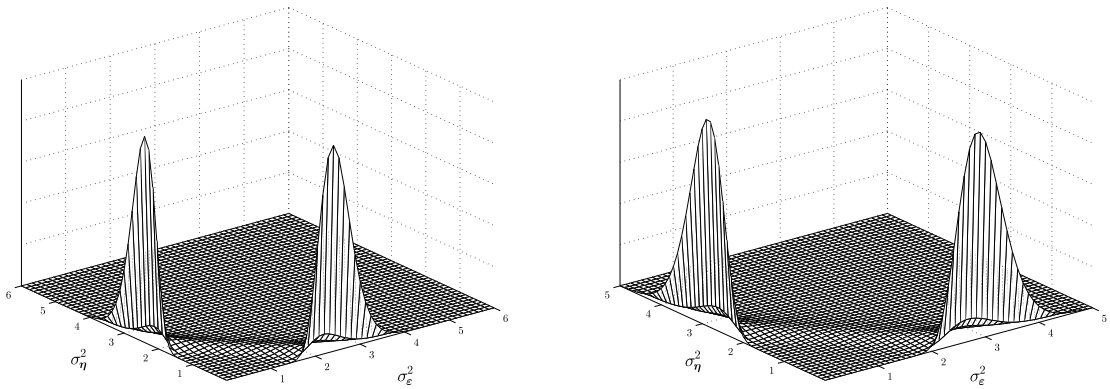
A further problem arises when instead of a uniform prior, a Jeffrey's-type prior is used, $p(\Theta) \propto \frac{1}{\sigma_\varepsilon^2}\frac{1}{\sigma_\eta^2}$, in which case the joint density becomes

$$p(\sigma_\eta^2, \sigma_\varepsilon^2|y) = \frac{1}{\sigma_\varepsilon^2}\frac{1}{\sigma_\eta^2}\left(\frac{1}{\sigma_\eta^2 + \sigma_\varepsilon^2}\right)^{\frac{1}{2}(T-1)} \exp\left(-\frac{1}{2}\frac{(y - \iota_T\hat{\theta})'(y - \iota_T\hat{\theta})}{\sigma_\eta^2 + \sigma_\varepsilon^2}\right) \tag{59}$$

[21]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 12 shows that the Jeffrey's prior causes the joint density to shoot off to infinity for either $\sigma_\varepsilon^2 \to 0$ or $\sigma_\eta^2 \to 0$ [22]. Therefore, the joint posterior is now improper and the

Figure 12: Joint posterior density of σ_ε^2 and σ_η^2 with a Jeffrey's prior



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

Notes: Panels (a) and (b) show the joint density in (59) with a signal-to-noise ratio of 1 and 0.5 respectively. For both panels y was simulated from (54)-(55) 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 Jeffrey's prior violates condition (a) of the theorem, while a uniform prior leads to a proper posterior.

Solutions

A number of solutions exist to prevent the problems presented in case (i) and (ii). For case (i) increasing the number of time-series observations and assuming that the variance is identical across all observations, will prevent the degrees of freedom problem. To solve the labeling 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.

Another possibility is to extend the basic HLMM in such a way that one can distinguish the dynamic processes for σ_ε^2 and σ_η^2. Two possible directions can be taken here. The first direction is to change the dynamics of η by changing the specification of the model in (54)-(55) to that of a State-Space model. The result will be that η is no longer a white noise process. The variance components can then be identified from the different processes processes of ε and η. 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.

[22] Although Figure 12 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 12 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 goes to zero.

To summarize, we can consider the basic HLMM to be a general specification and a parent model for more elaborate types of models such as State-Space and Panel Data models. The latter types of models are heavily used in econometric modelling practice and we therefore discuss these in somewhat more detail below.

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

$$\begin{aligned} y_t &= \mu_t + \varepsilon_t, \qquad \text{with} \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2) \qquad \text{and} \qquad t = 1, \ldots, T \qquad (60) \\ \mu_t &= \mu_{t-1} + \eta_t, \qquad \text{with} \qquad \eta_t \sim \mathcal{N}(0, \sigma_\eta^2) \qquad \text{and} \qquad \mathcal{E}[\varepsilon_t \eta_s] = 0 \qquad (61) \end{aligned}$$

with $\Theta = (\mu, \sigma_\varepsilon^2, \sigma_\eta^2)$. 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, in e.g. Koop and van Dijk (2000).

More elaborate State-Space models are easily obtained by including explanatory variables in the measurement equation (60) and state equation (61), 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), Jong and Shephard (1995) and Durbin and Koopman (2001).

The specification in (61) implies that η_t is a random walk process which follows directly from recursively substituting lagged values of μ_t. Because the dynamic processes for ε_t and η_t are now very different one can distinguish σ_ε^2 from σ_η^2 and therefore identify both variance components.

Gibbs Sampling

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

$$\begin{aligned} y_t &= x_t \beta_t + \varepsilon_t, \qquad \text{with} \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2) \qquad \text{and} \qquad t = 1, \ldots, T \qquad (62) \\ \beta_t &= \beta_{t-1} + \eta_t, \qquad \text{with} \qquad \eta_t \sim \mathcal{N}(0, \Sigma_\eta) \qquad \text{and} \qquad \mathcal{E}[\varepsilon_s \eta_{k,t}] = 0 \qquad (63) \end{aligned}$$

with x_t a $(1 \times K)$ vector of explanatory variables, β_t the $(K \times 1)$ state vector and Σ_η a $(K \times K)$ diagonal covariance matrix with diagonal elements $\sigma_{\eta,k}^2$ for $k = 1, \ldots, K$. We use the 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

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

where $p(\beta|\sigma_\eta^2)$ 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[23]

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

with β_k the vector of the k^{th} state variable and where KFS represents the Kalman Filter Sampler from 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

$$\begin{aligned} \Delta \mathrm{M}_t &= \beta_{0,t} + \beta_{1,t}\Delta i_{t-1} + \beta_{2,t}\mathrm{INF}_{t-1} + \beta_{3,t}\mathrm{SURP}_{t-1} + \beta_{4,t}\Delta \mathrm{M}_{t-1} + \varepsilon_t \qquad (64) \\ \beta_{k,t} &= \beta_{k,t-1} + \eta_{k,t} \quad \text{for} \quad k = 0, \ldots, 4 \qquad (65) \\ \varepsilon_t &\sim \mathcal{N}(0, \sigma_{\varepsilon}^{2}), \quad \eta_{k,t} \sim \mathcal{N}(0, \sigma_{\eta,k}^{2}) \quad \text{and} \quad \mathcal{E}[\varepsilon_t \eta_{k,s}] = 0. \qquad (66) \end{aligned}$$

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 that study with two more years of data (1962:I-1963:IV) but more importantly, we apply 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 13 shows the time-series of the posterior means for the state variables β_k, $k = 0, \ldots, 4$.

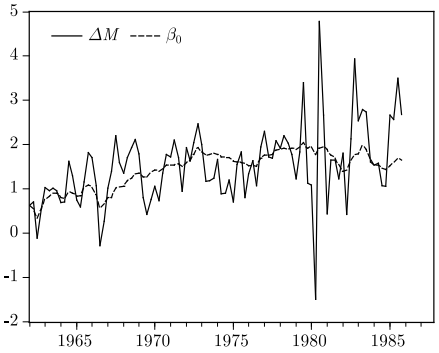
Table 2: Posterior results for the State-Space Model

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)

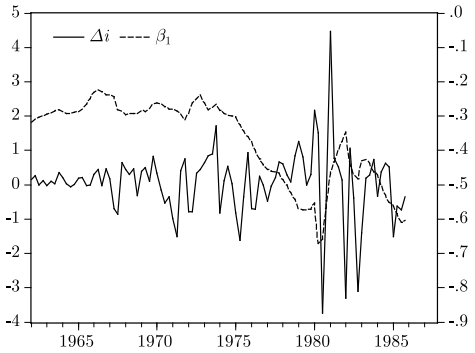
Notes: The table shows posterior means and standard deviations (in between brackets) for the variance components of model (64)-(66). Quarterly data for the period 1962:I-1985:IV for M1 growth, changes in the 3-Month Treasury Bill rate, CPI inflation rate and detrended full employment budget surplus were 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 13 are clear evidence that the coefficients in (64) show substantial variation across time. 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 varies over time. Especially the change in parameters around the Volcker period is striking and very similar for all parameters.

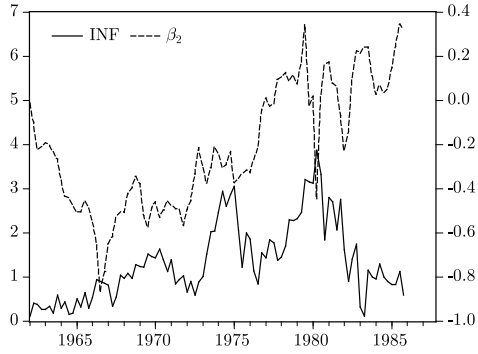
Figure 13: Time-varying parameters in the State-Space Model



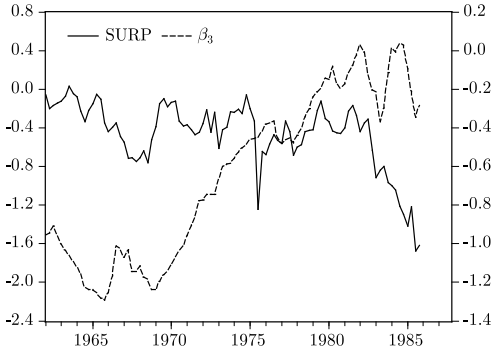
(a) ΔM and β_0



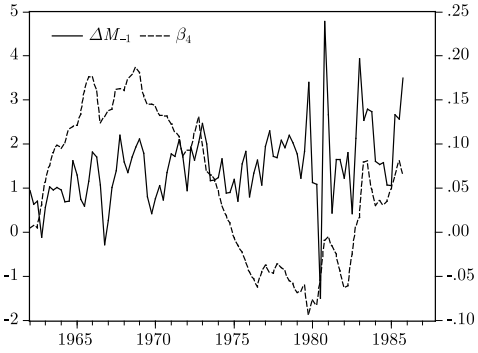
(b) Δi and β_1



(c) INF and β_2



(d) SURP and β_3



(e) ΔM_{-1} and β_4

Notes: The graphs show the posterior means for the time-varying parameters in the model (64)-(66). Panel (a) shows ΔM and β_0 whereas panel (b)-(e) show β_k for $k = 1, \ldots, 4$ with the accompanying exogenous variables. In each graphs, the scale for β_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

$$\begin{aligned} y_{i,t} &= \mu_i + \varepsilon_{it}, \qquad \text{with} \qquad \varepsilon_{i,t} \sim N(0, \sigma_\varepsilon^2) \qquad \text{and} \qquad t = 1, \ldots, T,\ i = 1, \ldots, N, \quad (67) \\ \mu_i &= \theta + \eta_i, \qquad \text{with} \qquad \eta_t \sim N(0, \sigma_\eta^2) \quad (68) \end{aligned}$$

with $\Theta = (\mu, \theta, \sigma_\eta^2, \sigma_\varepsilon^2)$ 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 because 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 emerges.

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 is defined such that its element (i, t) contains observation $y_{i,t}$. 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 Y_t (the t^{th} row of Y). Furthermore, define the overall sum of squares as

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

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	$p\left(\sigma_\varepsilon^2 \vert y, \mu, \theta, \sigma_\eta^2\right) \sim \mathcal{IG}\left(\frac{1}{2}E^{(j)'}E^{(j)}, \frac{1}{2}(TN-2)\right)$
- 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 (67)-(68) 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. *, **, *** indicate that zero is not contained in 90%, 95%, 99% highest posterior density region (HPD) region, respectively. Only posterior standard deviations for Australia are given. An Inverted Gamma density with parameters $r = 10^{-5}$ and $y = 1$ was used as prior distribution for the variance components.

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

$$M_i = \frac{\sigma_\eta^{2(j-1)}}{\sigma_\eta^{2(j-1)} + (1/T)\sigma_\varepsilon^{2(j-1)}} \iota_T Y_t + \frac{\sigma_\varepsilon^{2(j-1)}}{T\sigma_\eta^{2(j-1)} + \sigma_\varepsilon^{2(j-1)}} \theta_i^{(j)} \tag{69}$$

The expression in (69) 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 Y_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 rather limited. Nevertheless, it serves 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, as theory predicts, those for the other countries are qualitatively similar. To generate 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, a significant 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[24]. 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 neither the posterior mean nor the standard deviation exist for σ_η^2. 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 Concluding Remarks

Using a set of well-known econometric models that are widely used in practice, we presented a road map for effective application of Gibbs sampling in the context of a Bayesian analysis. The models considered range from the Cochran-Orcutt model for serial correlation in a regression set up, via a Koyk model for Distributed Lag analysis to Unit Root, State-Space and Panel Data models.

We recommend that every applied researcher investigates the shape of the criterion function, usually the posterior and/or the predictive density. The issues that one may face "en route" can be classified as follows. For equation system or level parameters it is important whether there exists substantial probability mass near the boundary of the parameter region. This may affect the convergence of the Gibbs sampler in an adverse manner. Such a situation occurs in near unit root and near non-identified processes. In

[24]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 case of variance components models, where we focus on State-Space and Panel Data models, it is important to investigate the degrees of freedom problem. That is, enough grouping of observations in the time and cross-section domain should take place. Also, a labeling issue may occur with respect to different variances.

The issues listed may seriously hamper the convergence of the Gibbs sampler. However, some simple solutions are presented that allow again good use of Gibbs. The information matrix prior is effective in cases where the likelihood function exhibits singularities. Inequality restrictions and a dynamic recursive structure like the Kalman filter help in the case of variance components problems.

The analysis is illustrated using several data sets that refer to growth of Gross Domestic Product of several countries, to financial data such US money growth, Treasure Bill rate and to the well-known Lydia Pinkham sales and advertising data.

We end this paper with two remarks. In terms of methods we note that several other solutions help the Gibbs sampler in terms of convergence. Reparametrization of the model and informative 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 we refer to Geweke (2005) for a more detailed analysis.

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 (2000), *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. (1995), Marginal Likelihood from the Gibbs output, *Journal of the American Statistical Association*, 90, 1313–1321.

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

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

Frühwirth-Schnatter, S. (2001), Markov Chain Monte Carlo Estimation of Classical and Dynamic Switching and Mixture Models, *Journal of American Statistical Association*, 96, 194–209.

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.

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. (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-Interscience.

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

Greene, W. H. (2000), *Econometric Analysis*, Prentice Hall, New York.

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

Griliches, Z. and M. Intriligator (eds.) (1983-1986), *Handbook of Econometrics*, vol. I-III, Elsevier Science, Amsterdam: North Holland.

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

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

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

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

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.

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

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.

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.

McFadden, D. and R. Engle (eds.) (1994), *Handbook of Econometrics*, vol. IV, Elsevier Science, Amsterdam: North Holland.

Morgan, M. (1990), *The History of Econometric Ideas*, Cambridge University Press, Cambridge, United Kingdom.

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

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.

Redner, R. and H. F. Walker (1984), Mixture Densities, Maximum Likelihood and The EM Algorithm, *SIAM Review*, 26, 195–239.

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. (1991), *Tools for Statistical Inference*, Springer-Verlag, New York.

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

Tinbergen, J. (1939a), Statistical Testing of Business Cycle Theories, I, A Method and Its Application to Investment Activity, in *League of Nations, Geneva*.

Tinbergen, J. (1939b), Statistical Testing of Business Cycle Theories, II, Business Cycles in the United States of America, in *League of Nations, Geneva*.

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.

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

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

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

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

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

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

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 about the mean are given by

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

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

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

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

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

$$\begin{array}{rcl} \boldsymbol{\mu}_1 & = & M \\ \boldsymbol{\mu}_2 & = & S \end{array}$$