Bayesian Near-Boundary Analysis in Basic Macroeconomic Time Series Models*

Michiel De Pooter[1] Francesco Ravazzolo[2] Rene Segers[3] Herman K. Van Dijk[3,†]

[1] *Federal Reserve Board*
[2] *Norges Bank*
[3] *Tinbergen Institute and Econometric Institute*
Erasmus University Rotterdam

Econometric Institute Report EI 2008-13

August 21, 2008

Abstract

Several lessons learnt from a Bayesian analysis of basic macroeconomic time series models are presented for the situation where some model parameters have substantial posterior probability near the boundary of the parameter region. This feature refers to near-instability within dynamic models, to forecasting with near-random walk models and to clustering of several economic series in a small number of groups within a data panel. Two canonical models are used: a linear regression model with autocorrelation and a simple variance components model. Several well-known time series models like unit root and error correction models and further state space and panel data models are shown to be simple generalizations of these two canonical models for the purpose of posterior inference. A Bayesian model averaging procedure is presented in order to deal with models with substantial probability both near and at the boundary of the parameter region. Analytical, graphical and empirical results using U.S. macroeconomic data, in particular on GDP growth, are presented.

Keywords: Gibbs sampler, MCMC, autocorrelation, nonstationarity, reduced rank models, state space models, error correction models, random effects panel data models, Bayesian model averaging

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

---

*This paper is a substantial revision and extension of De Pooter *et al.* (2006). We are very grateful to participants of the 3rd World Conference on Computational Statistics & Data Analysis (Cyprus, 2005), the International Conference on Computing in Economics and Finance (Geneva, 2007), and the Advances in Econometrics Conference (Louisiana, 2007), for their helpful comments on earlier versions of the paper and we are in particular indebted to William Griffiths, Lennart Hoogerheide and Arnold Zellner for many useful comments. Herman Van Dijk gratefully acknowledges financial support from the Netherlands Organization for Scientific Research (NWO). The views expressed in this paper are solely the responsibility of the authors and should not be interpreted as reflecting the views of the Board of Governors of the Federal Reserve System or of any other employee of the Federal Reserve System nor the views of Norges Bank (the Central Bank of Norway).

This paper contains a number of references to Appendices. These appendices contain sampling schemes, conditional and marginal density results, probability density functions and an overview of models used in the paper. They are available on `http://people.few.eur.nl/hkvandijk/research.htm`. Also given on this website is the computer code and data which was used in the empirical applications in the paper.

†Corresponding author. Tinbergen Institute, Erasmus University Rotterdam, P.O. Box 1738, NL-3000 DR Rotterdam, The Netherlands. Tel.: +31-10-4088900, fax: +31-10-4089031. *E-mail addresses*: `michiel.d.depooter@frb.gov` (M. De Pooter), `francesco.ravazzolo@norges-bank.no` (F. Ravazzolo), `rsegers@few.eur.nl` (R. Segers), `hkvandijk@few.eur.nl` (H.K. Van Dijk).

1 Introduction

Stable economic growth with possibly temporary periodic deviations - better known as business cycles - is one of the most important economic issues for any country. A widely used macroeconomic time series to measure these characteristics of growth and cycles is real Gross Domestic Product. A commonly used model in this context is the linear autoregressive model with deterministic trend terms. Using such time series and model classes, econometric analysis -Bayesian and Non-Bayesian- leads for most industrialized nations to substantial evidence that economic growth evolves according to a trend process that is largely determined by stochastic shocks. Otherwise stated, in the autoregressive models one finds substantial empirical evidence of a characteristic root that is near the boundary of unity or at this boundary in the parameter region. There exists an enormous literature on this topic; here we only mention the well-known study of Nelson and Plosser (1982) and for empirical evidence on industrialized nations over the past century we refer to Van Dijk (2004).

There are several other examples in macro economics of the existence of substantial posterior probability near the boundary of a parameter region. In business cycle analysis one may be interested to know how much of the variation in an economic time series is due to the cycle and how much is due to the trend. This issue is relevant in the context of an adequate policy mix for stimulating long term economic growth and short term business cycle control. Using a structural time series model one may find substantial posterior probability of the cyclical component near zero and the relative weights of the trend and cyclical components are then very uncertain. Another example occurs with typical characteristics of financial time series in stock markets. These series behave close to a random walk model or, otherwise stated, close to a model with a characteristic root close to the boundary of unity. The economic issue of such a process is whether financial markets are efficient in the sense that the optimal forecast for future stock market prices is the current price. A fourth example is the club or cluster behavior in panel data models for economic growth in industrialized nations. In this context, convergence of economic growth is studied and the number of clusters may be relatively small. Substantial probability may occur at the boundary of the parameter region of the number of clubs and, as a consequence, large uncertainty exists with respect to the correct number of clubs.

To explore the issue of Bayesian near-boundary analysis in basic economic time series models, one derives their likelihoods and specifies prior information. Our approach with respect to specifying prior information is to start with uniform priors on a large region. The use of such noninformative priors means that we concentrate on the information content in and the shape of the likelihood function. Given our diffuse priors, the posterior distributions of parameters of interest may or may not exist. The latter occurs, in particular, at the boundary of the parameter region, due to nonidentifiability of some parameters of interest. We discuss how Information matrix priors or training sample priors may regularize or smooth the shape of irregular posterior distributions. In our analysis we make use of an interplay of analytical techniques, simulation methods and graphics. As a simulation technique we use the Gibbs sampling method. A brief introduction is given in Section 2. We note that graphs in the context of Bayesian analysis are becoming more and more important see, for example, Murrell (2005). In our analysis we therefore also place emphasis on presenting results in a graphical way.

In this paper we make use of two classes of canonical models. The first class of models is known as the class of single equation dynamic regression models. A first contribution

of this paper is to show that, for our purpose of a Bayesian near-boundary analysis in the parameter region, basic members of this class of models like unit root models, distributed lag models and error correction models are special cases of the well-known linear regression model with first-order autocorrelation in the disturbances. We also indicate that an error correction model with near unit root behavior is - for posterior analysis - equivalent to an instrumental variable regression model with possibly weak instruments; see also Hoogerheide and Van Dijk (2001) and Hoogerheide and Van Dijk (2008). Interpretation of the models, their posteriors and the effect of smoothness priors like the Information matrix and training sample priors is one aim. A second aim is to illustrate a Bayesian analysis of economic growth in U.S. Gross Domestic Product using different model specifications.

The second class of models deals with variance parameters as parameters of interest. We discuss how the simple regression model with heteroscedasticity can be used as an introduction to the class of Hierarchical Linear Mixed Models (HLMM). As a second contribution, we show that the latter model serves as a parent class for time-varying parameter models such as state space models and panel data models. We investigate what happens when the density of one of the variance parameters is located near the zero bound and what happens when the number of components/groups in a panel is very small. We show that the latter case also leads to a boundary issue. We note that a combination of the first and second class of basic models has recently become important in empirical analysis.

A third contribution of this paper is to show how Bayesian model averaging over models with substantial posterior probability near and at the boundary leads to better forecasting. That is to say, we do not consider the case where substantial posterior probability is near and/or at the boundary as an econometric inferential problem where model selection is appropriate in order to determine or test whether the economic process is stationary or nonstationary but as a case where model averaging is to be preferred. This may lead to improved forecasting.

The results of our analysis may be summarized as 'lessons learnt from models used' and the start of a road map for learning Bayesian near-boundary analysis. A summary is presented in Section 7. Some key lessons are: investigate the shape of the likelihood of the parameters of interest; investigate the influence of smoothness priors in case of substantial near-boundary posterior probability; learn which simulation technique may be used efficiently in which situation; apply Bayesian model averaging over models with posterior probability near a boundary and models with substantial probability at the boundary. The topic of this paper should be of interest to Bayesians who make use of basic regression models for economic time series when the focus is on the information content of the likelihood. The topic should interest non-Bayesians who are very knowledgeable about basic econometric models and want to learn how the information in the likelihood function of such models is summarized according to Bayes' rule.

There exists an excellent literature of Bayesian analysis of regression models with autocorrelation, unit root models, distributed lag models, unobserved component models and panel data models. An incomplete list of references is given as Chib (1991), Chib and Greenberg (1994, 1995) and Chib and Carlin (1999), Schotman and Van Dijk (1991a,b) and Harvey (1989). The purpose of the present paper is to extend this analysis to situations where substantial posterior probability is near or at the boundary of the parameter region. We emphasize that this paper does not result and is not intended to result in a simple message with respect to using one model, one particular prior and one simulation techniques. We do not believe in such simplistic claims but rather in a situation where

different priors, different models and simulation algorithms are suitable depending on the problem studied, the data, and the shape of the posterior of the model considered. Our purpose is to investigate a set of models and, next, to explore Bayesian model averaging.

The content of this paper is structured as follows. In Section 2 we briefly review some (artificial) examples of shapes of posterior densities that the researcher may encounter in econometric practice. We also give an introduction to Gibbs sampling. This method is very natural given our derivations of joint, conditional and marginal posteriors for the linear regression model and this model with possibly autocorrelated disturbances. In Section 3 we present some basic results for the linear regression model that will be used in subsequent sections. In Sections 4 and 5 we present our empirical analysis and present some theoretical results of near-boundary posterior probability for a number of models for economic time series. Section 6 deals with forecasting U.S. Gross Domestic Product (GDP) using Bayesian model averaging. Section 7 contains a summary of models used and lessons learnt. The Appendices contain some technical details. We note that the technical level of the paper is like that of an introductory graduate econometrics course. Matrix notation is used in order to indicate the common, linear (sub-)structure of several models.

2 Preliminaries I: Basics of Gibbs Sampling and Typical Shapes of Posterior Densities

2.1 Basics of Gibbs sampling

As discussed in, for instance, Van Dijk (1999) and Hamilton (2006), the 'simulation revolution in Bayesian econometric inference' is to a large extent due to the advent of computers with ever-increasing computational power. This allows researchers to apply alternative Bayesian simulation techniques for estimation in which extensive use is made of pseudo-random number generators. One of the most important and widely used simulation methods is Gibbs sampling, developed by Geman and Geman (1984), Tanner and Wong (1987) and Gelfand and Smith (1990). This method has become a popular tool in econometrics for analyzing a wide variety of problems; see for instance Chib and Greenberg (1996) and Geweke (1999). Judging from numerous recent articles in the literature, Gibbs sampling is still gaining more and more momentum. Recent textbooks such as Bauwens *et al.* (1999), Koop (2003), Lancaster (2004), and Geweke (2005) discuss how Gibbs sampling is used in a wide range of econometric models, in particular in models with latent variables. Mixture processes are an important class of latent variable models in econometrics; the most well-known due to Hamilton (1989). In recent papers by, for instance, Celeux *et al.* (2000), Frühwirth-Schnatter (2001), Jasra *et al.* (2005) and Geweke (2007), the issue of convergence of the Gibbs sampler in this class of models is discussed. The posterior distribution in mixture processes may be multimodal and may exhibit ridges often due to near nonidentification of parameters. A detailed analysis of this topic is beyond the scope of the present paper but we note and will make use of the distinction by Geweke (2007) between the interpretation of a model and its posterior densities on the one hand and the numerical efficiency and convergence of a simulation algorithm on the other hand.

One may characterize the Gibbs sampling algorithm as an application of the *divide-and-conquer* principle[1]. First, a K-dimensional parameter vector $\boldsymbol{\theta}$ is divided into m

[1]We are necessarily brief in our explanation of the Gibbs sampler. See e.g. Casella and George (1992), or Hoogerheide *et al.* (2008), for a more elaborate discussion.

components $\theta_1, \theta_2, ..., \theta_m$, where $m \leq K$. Second, for many posterior distributions which are intractable in terms of simulation the lower-dimensional *conditional* distributions turn out to be remarkably simple and tractable. The Gibbs sampler exploits this feature, as it samples precisely from these conditional distributions. Its usefulness is, for example, demonstrated by Chib and Greenberg (1996) and Smith and Roberts (1993).

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

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

where $\boldsymbol{\theta}_{-i}$ denotes the parameter component vector $\boldsymbol{\theta}$ without the i^{th} component, can all be sampled from. The Gibbs sampling algorithm starts with the specification of an initial set of values: $(\theta_1^{(0)}, ..., \theta_m^{(0)})$ and then generates a sequence

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

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

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

2: Generate (*the j^{th} Gibbs step*) :

$$\begin{aligned}
&\theta_1^{(j)} \text{ from } p(\theta_1|\theta_2^{(j-1)}, \ldots, \theta_m^{(j-1)}) \\
&\theta_2^{(j)} \text{ from } p(\theta_2|\theta_1^{(j)}, \theta_3^{(j-1)}, \ldots, \theta_m^{(j-1)}) \\
&\theta_3^{(j)} \text{ from } p(\theta_3|\theta_1^{(j)}, \theta_2^{(j)}, \theta_4^{(j-1)}, \ldots, \theta_m^{(j-1)}) \\
&\vdots \\
&\theta_m^{(j)} \text{ from } p(\theta_m|\theta_1^{(j)}, \ldots, \theta_{m-1}^{(j)})
\end{aligned}$$

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

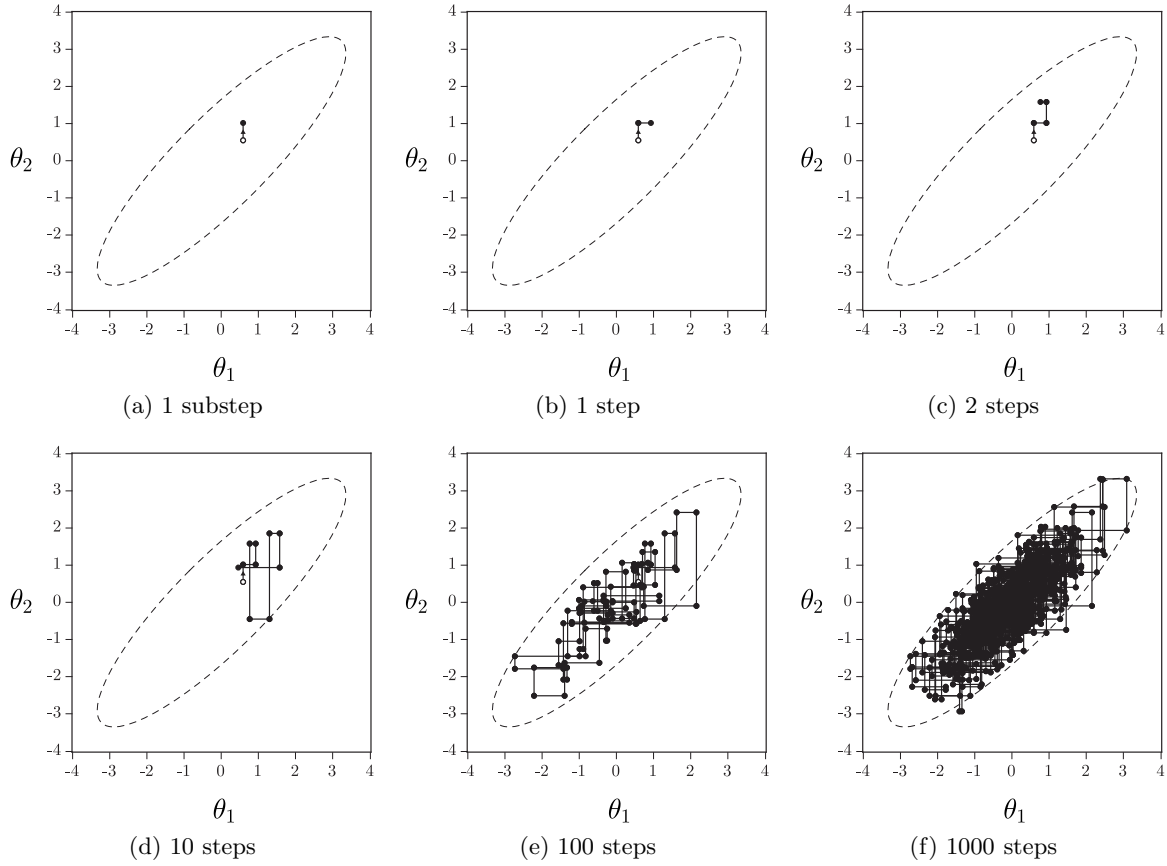
The above algorithm yields a sequence of J realizations $\boldsymbol{\theta}^{(j)} = (\theta_1^{(j)}, ..., \theta_m^{(j)})$, for $j = 1, 2, ..., J$, from a Markov chain, converging to the target distribution. We will refer to Step **2** of the algorithm as 'the Gibbs step' and for each of the models that we discuss in the subsequent sections we will always indicate what the Gibbs step looks like. Note that the components of $\boldsymbol{\theta}$ do not necessarily need to be one-dimensional. Generating draws for blocks of parameters where some of the θ_i components denote a block of parameters is also possible.

The Gibbs algorithm is illustrated in Figure 1 where we show an example path of Gibbs sampled points, when the conditional densities of $\theta_1|\theta_2$ and $\theta_2|\theta_1$ are both standard Normal and assuming a correlation coefficient of 0.75. The sample path is shown at different stages of the algorithm.

The key feature of Gibbs sampling is:

For large enough J the sequence of Gibbs draws, generated from the conditional distributions, is distributed according to the joint and marginal posterior distributions.[2]

Figure 1: **Gibbs sampling: Example steps**



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

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

Notes: Panels (a) through (f) show consecutive steps of the Gibbs sampler using two conditional posterior densities, $p(\theta_1|\theta_2)$ and $p(\theta_2|\theta_1)$ which are both standard normal with a correlation coefficient of 0.75. The open circles in Panels (a)-(f) indicate the starting vector $(\theta_1^{(0)}, \theta_2^{(0)})$.

A simple argument for the general case is as follows. Suppose θ_i and $\boldsymbol{\theta}_{-i}$ have a *joint* posterior distribution with density $p(\theta_i, \boldsymbol{\theta}_{-i})$. Thus this posterior should exist, which must be carefully verified, compare also Sections 3 and 4. Then $\boldsymbol{\theta}_{-i}$ has the *marginal* posterior distribution with density $p(\boldsymbol{\theta}_{-i})$. Denote by $\boldsymbol{\theta}_{-i}$ and $\boldsymbol{\theta}^{j-1}$ the drawing generated in the $(j-1)^{\text{th}}$ step from the marginal posterior with density $p(\boldsymbol{\theta}_{-i})$. In the j$^{\text{th}}$ step of the Gibbs sampling algorithm, $\theta_i^{(j)}$ is drawn from $p(\theta_i|\boldsymbol{\theta}_{-i}^{(j-1)})$, which is the density of the conditional distribution of θ_i given $\boldsymbol{\theta}_{-i}^{(j-1)}$. The joint density of $\theta_i^{(j)}$ and $\boldsymbol{\theta}_{-i}^{(j-1)}$ is

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

[2]Strictly speaking, sensitivity to initial conditions persists, but it becomes negligible if the sequence of Gibbs draws mixes sufficiently well.

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

Because in practice it may take some time for the Markov chain to converge, it is common to discard the first B draws, where typically $B << J$. These draws are referred to as the *burn-in draws*. Consequently, posterior results will be based only on draws $\boldsymbol{\theta}^{(B+1)}, \ldots, \boldsymbol{\theta}^{(J)}$ of the generated chain. Furthermore, the sequence of draws sometimes displays some degree of autocorrelation. When autocorrelations are significant up to the $(h-1)^{\text{th}}$ lag, one can consider using only every h^{th} draw and to discard the intermediate draws (h is known as the thinning value)[3]. An altogether different approach is to generate *multiple* Markov chains instead of just one chain and to use only the final draw from each sequence. Doing so implies that the Gibbs algorithm has to be executed a substantial number of times. When opting for this approach the researcher does not have to worry about which value to choose for h. Although the drawback of this method is that it can be very computationally intensive, it can alternatively help prevent posterior results from being (partially) determined by a particular set of starting values. We show in the next section that randomizing over $\boldsymbol{\theta}^{(0)}$ can be a worthwhile endeavor when the likelihood displays signs of multimodality.

2.2 Three typical shapes of posterior densities

To illustrate the kinds of shapes that may occur in posterior densities we work through a number of examples which are based on the model in Gelman and Meng (1991). Suppose that we have a joint posterior density of (θ_1, θ_2), which has the following form

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

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

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

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

Note that, typically, the joint density of (θ_1, θ_2) is not Normal. By choosing different parameter configurations for a, b, c_1 and c_2 we can construct joint posterior densities with rather different shapes, while the conditional densities remain Normal. In the remainder of this section we consider three types of shapes and we apply the Gibbs sampler to each

[3] The current consensus in the literature, however, seems to be to always include the information of all draws, even when these are correlated.

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

of these. Although the shapes are all in a way artificial since they are not based directly on a model and data, doing so will give us some early insights into different shapes of (joint) posterior densities and boundary issues which we discuss in detail in the remainder of this paper.

For each of the examples below the j^{th} Gibbs step consists of sequentially drawing from (5) and (6):

j^{th} **Gibbs step for the Gelman-Meng model:**

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

(i) Bell-shape

The first parameter configuration that we consider for the joint density in (4) is the following; $[a = b = c_1 = c_2 = 0]$ in which case the joint density is given by

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

Both the conditional densities and the joint density are standard Normal. The latter is depicted in Figure 2(a). Gibbs sampling simply comes down to obtaining draws by iteratively drawing from standard Normal densities[5]. A scatterplot of one thousand of such draws is shown in Figure 2(b).

The estimated conditional means and variances are equal to 0 and 1 for both parameters. These are exactly the parameters of the marginal densities which, in this case, we know to be standard Normal. In fact, for the chosen parameter configuration, the conditional and marginal densities coincide since the conditional density for θ_1 does not depend on θ_2 and vice versa. In this particular example it would therefore obviously not be necessary to use Gibbs sampling.

(ii) Ridges

The second parameter configuration that we examine is $[a = c_1 = c_2 = 0]$[6]. The joint density is now given by

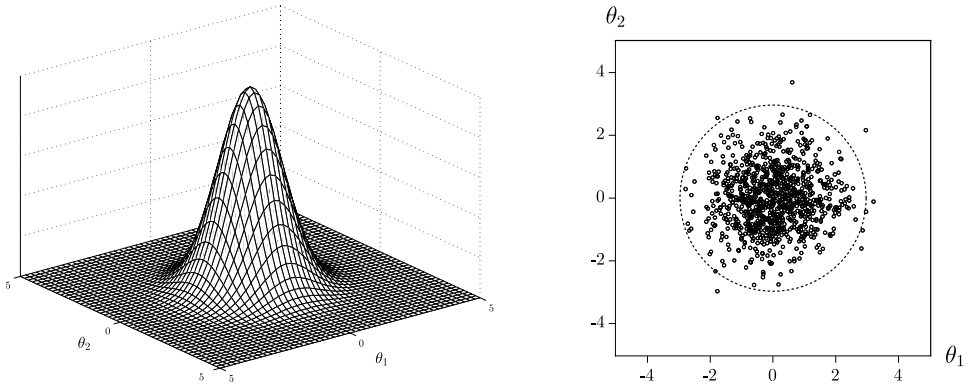
$$\begin{aligned} p(\theta_1,\theta_2) &\propto \exp\left[-\tfrac{1}{2}(\theta_1^2 - 2b\theta_1\theta_2 + \theta_2^2)\right] \\ &\propto \exp\left[-\tfrac{1}{2}\begin{bmatrix}\theta_1 & \theta_2\end{bmatrix}\begin{bmatrix}1 & b\\ b & 1\end{bmatrix}\begin{bmatrix}\theta_1\\ \theta_2\end{bmatrix}\right] \end{aligned} \tag{8}$$

It is apparent from Figure 3 that the shape of this density depends on the value of b. When b tends to 1 a ridge along the line $\theta_1 = \theta_2$ appears in the shape of the posterior. The scatterplots of Gibbs draws for this example in Figure 3 reveal that Gibbs sampling tends to become less efficient in such a case. Ridges may occur in econometric models where the Information matrix tends to become singular, that is when $b \to 1$; see the next section for examples. We emphasize that the posterior in Figures 3(e) and 3(f) is defined on a bounded region with bounds -75 and +75. This posterior is constant along the diagonal and it is a continuous function defined on a bounded region and thus a proper density.

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

[6]Note that this parameter vector violates the earlier stated parameter restrictions when $b \geq 1$.

Figure 2: **Gelman-Meng: Bell-shape**



(a) Joint posterior density (b) Gibbs draws

Notes: Panel (a) shows the Gelman-Meng joint posterior density for θ_1 and θ_2 given in (4) for parameter values $[a = b = c_1 = c_2 = 0]$ whereas Panel (b) shows the scatterplot of 1,000 draws from the Gibbs sampler together with the 99% highest probability density region.

(iii) Bimodality

The third and final configuration we consider is $[a = 1, b = 0]$ and large, but not necessarily equal, values for c_1 and c_2.[7] Here we select $c_1 = c_2 = 10$ which gives

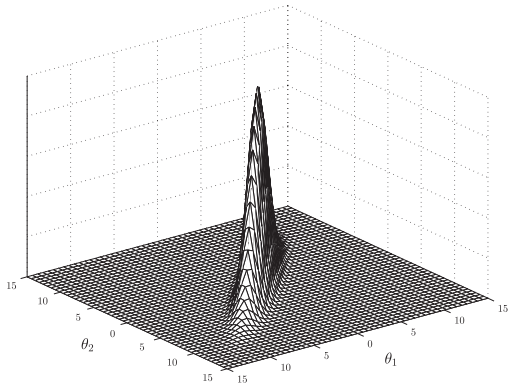
$$\begin{aligned} p(\theta_1, \theta_2) &\propto \exp\left[-\tfrac{1}{2}[\theta_1^2\theta_2^2 + \theta_1^2 + \theta_2^2 - 20\theta_1 - 20\theta_2]\right] \\ &\propto \exp\left[-\tfrac{1}{2}\begin{bmatrix} \theta_1 - 10 & \theta_2 - 10 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} \theta_1 - 10 \\ \theta_2 - 10 \end{bmatrix} + \theta_1^2\theta_2^2\right] \end{aligned} \tag{9}$$

At first sight the scatterplot of one thousand Gibbs draws, shown in Figure 4(b), seems perfectly reasonable and posterior means and variances can easily be computed. However, when inspecting the joint density as depicted in Figure 4(a) we immediately see that the joint density $p(\theta_1, \theta_2)$ is bimodal and that the Gibbs sampler has sampled from one mode but not from the other. Apparently it tends to get stuck in one of the two modes[8]. This is because the modes are too far apart with an insufficient amount of probability mass in between the two modes for the sampler to regularly jump from one to the other. Admittedly, substantially increasing the number of draws substantially will eventually lead to a switch. However, one cannot be certain when this will occur. The scatterplot shows that with a single run, one thousand draws is clearly not enough. However, although not shown here, also one million draws is still an insufficient number to witness a switch. Therefore, the Gibbs output only provides the researcher with information about a subset of the full domain of $p(\theta_1, \theta_2)$ and posterior results are thus incomplete. One option to try and at least signal the bimodality of the likelihood is to execute the Gibbs sampler several times with widely dispersed initial values. However, we do note that even when doing so the issue of determining how much probability mass is located in each of the modes remains nontrivial. Although the example we discuss here is a rather extreme case, it

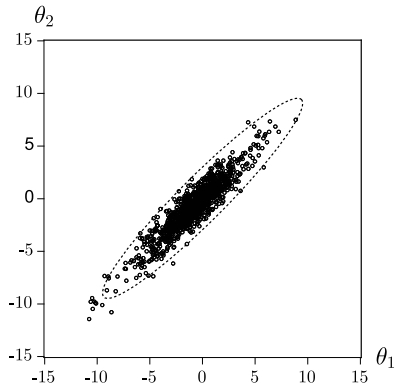
[7] See also Hoogerheide *et al.* (2007) for a further analysis of the three types of shapes (elliptical shapes, bimodality, ridges) for posterior densities in the IV model.

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

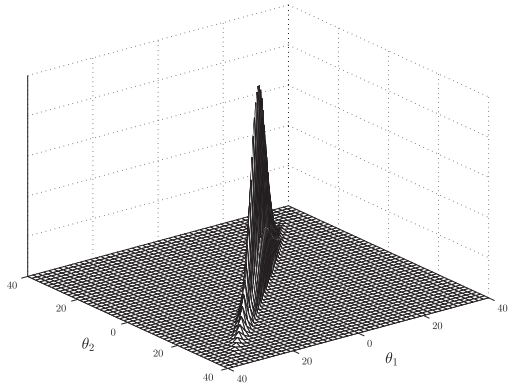
Figure 3: **Gelman-Meng: Ridges**



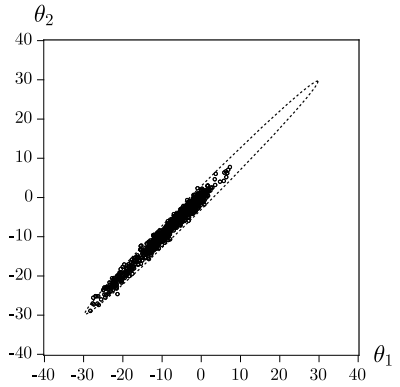
(a) Joint posterior density, $b = 0.95$



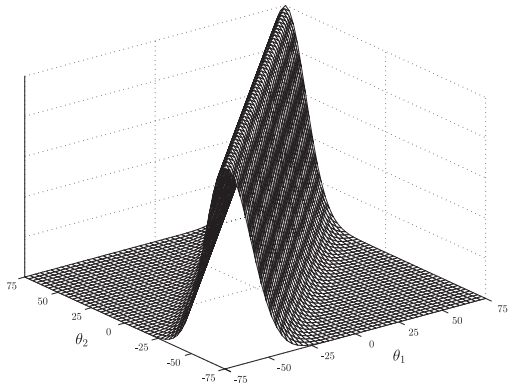
(b) Gibbs draws, $b = 0.95$



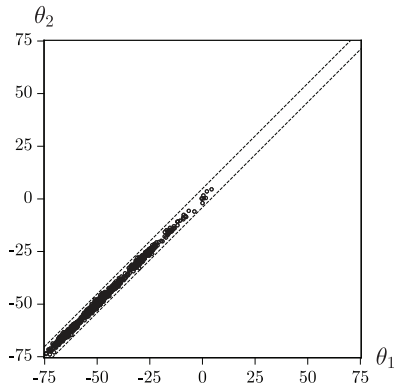
(c) Joint posterior density, $b = 0.995$



(d) Gibbs draws, $b = 0.995$



(e) Joint posterior density, $b = 1.0$



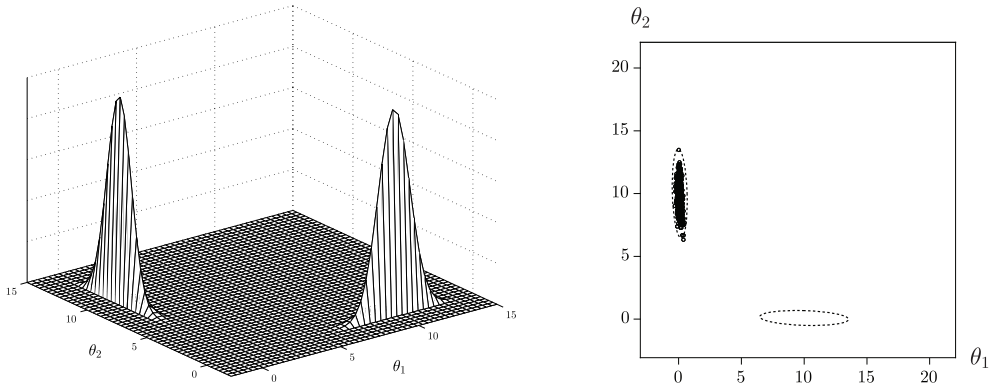
(f) Gibbs draws, $b = 1.0$

Notes: Panel (a) shows the Gelman-Meng joint posterior density for θ_1 and θ_2 given in (4) for parameter values [$a = c_1 = c_2 = 0$ and $b = 0.95$] whereas panel (b) shows the scatterplot of 1,000 draws from the Gibbs sampler together with the 99% highest probability density region. Panels (c) and (d) show the same figures for $b = 0.995$ and (e) and (f) for $b = 1.0$.

should be clear that multimodality can result in very slow converge for the Gibbs sampler. Multimodality may occur in reduced rank models when one is close to the boundary of the parameter region.

Summarizing, the above examples of a bell-shaped, a ridge-shaped, and a bimodal-

Figure 4: **Gelman-Meng: Bimodality**



(a) Joint posterior density (b) Gibbs draws

Notes: Panel (a) shows the Gelman-Meng joint posterior density for θ_1 and θ_2 given in (4) with parameter values [$a = 1, b = 0$ and $c_1 = c_2 = 10$] whereas panel (b) shows the scatterplot of 1,000 draws from the Gibbs sampler together with the 99% highest probability density region.

shaped density, indicate that it is essential to scrutinize a proposed model and the shape of its posterior distribution before moving on to drawing posterior inference on its parameters through a simulation method. Doing so may not always be straightforward, however, especially in large dimensional spaces.

3 Preliminaries II: Joint, Conditional and Marginal Posterior and Predictive Densities for the Linear Regression Model

3.1 Linear regression model

We start our model analysis by considering the basic linear regression model where the variation of a dependent variable y_t is explained 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):

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

The goal is to draw inference on the $(K\times1)$ vector of regression parameters $\beta = (\beta_1\ \beta_2 \dots \beta_K)'$[9] and the scalar variance parameter σ_ε^2. In matrix notation, this model is given by

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

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

In the following we provide basic results for the joint, conditional and marginal posterior densities of the linear regression model in (11) which are useful for simulation purposes.

[9] A simple case is when $x_t \equiv 1$ and $\beta = \mu$ in which case one only estimates the mean and variance of y.

More details can be found in, e.g., Zellner (1971), Koop (2003) and Geweke (2005). For an expert reader we suggest to consider only the summary tables and diagrams in Appendix B.

Joint density

We start by specifying the likelihood for the linear regression model in (11) as:

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

Combining the likelihood with a noninformative or Jeffreys' prior[10]

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

gives the joint posterior density

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

where we define D as the data information set, i.e. $D \equiv (y, X)$.

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

$$(y-X\beta)'(y-X\beta) = (y-X\hat{\beta})'(y-X\hat{\beta}) + (\beta-\hat{\beta})'X'X(\beta-\hat{\beta}) \tag{15}$$

with $\hat{\beta} = (X'X)^{-1}X'y$, the OLS estimator of β. One can now rewrite (14) as

$$p(\beta,\sigma_\varepsilon^2|D) \propto (\sigma_\varepsilon^2)^{-\frac{(T+2)}{2}} \exp\left[-\frac{1}{2\sigma_\varepsilon^2}\left[(y-X\hat{\beta})'(y-X\hat{\beta}) + (\beta-\hat{\beta})'X'X(\beta-\hat{\beta})\right]\right] \tag{16}$$

The density (16) is known as the *Normal Inverted-Gamma* density of $(\beta,\sigma_\varepsilon^2)$, see Raiffa and Schlaifer (1961, p. 310) and Zellner (1971, Chapter 3).

Conditional densities

The only part of the posterior in (16) which is relevant for determining the posterior density of β conditional on a value for σ_ε^2 is the part that depends on β. The first part, $(y-X\hat{\beta})'(y-X\hat{\beta})$, only depends on the data D and does therefore not enter the conditional density of β. From the probability density functions given in Appendix C, we can recognize, for a given value of σ_ε^2, a multivariate Normal density for β which has mean vector $M = \hat{\beta}$ and variance matrix $S = \sigma_\varepsilon^2[X'X]^{-1}$, see equation (C-4). Similarly, the conditional density of σ_ε^2, for a given parameter vector β, follows from equation (C-3) and is Inverted Gamma

[10] A noninformative prior for the regression parameters can simply be specified as $p(\beta) \propto 1$. For a variance parameter a noninformative 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. If one has prior information it is strictly advisable to include this in the analysis (see the discussion in Lancaster, 2004 and Geweke, 2005). Specifying conjugate priors is, however, not always an easy task, especially when one is faced with a large dimensional parameter region. Here we focus on noninformative priors since we are concerned with what we can learn about the model parameters through the data likelihood.

with location parameter $m = \frac{1}{2}(y - X\beta)'(y - X\beta)$ and $\nu = \frac{1}{2}T$ degrees of freedom. Summarizing, the conditional posterior densities are

$$p(\beta|D, \sigma_\varepsilon^2) \sim \mathcal{N}\left(\hat{\beta}, \sigma_\varepsilon^2[X'X]^{-1}\right) \tag{17}$$

$$p(\sigma_\varepsilon^2|D, \beta) \sim \mathcal{IG}\left(\frac{1}{2}(y - X\beta)'(y - X\beta), \frac{1}{2}T\right) \tag{18}$$

Gibbs sampling for the basic linear regression model consists of iteratively drawing from the conditional densities $p(\beta|D, \sigma_\varepsilon^2)$ and $p(\sigma_\varepsilon^2|D, \beta)$. A scheme of derivations for Gibbs sampler results is presented in the top part of Figure 5. The j^{th} Gibbs step consists of

j^{th} Gibbs step for the linear regression model:
- generate $\beta^{(j)}|\sigma_\varepsilon^{2(j-1)}$ from $p(\beta|D, \sigma_\varepsilon^2) \sim \mathcal{N}\left(\hat{\beta}, \sigma_\varepsilon^{2(j-1)}[X'X]^{-1}\right)$
- generate $\sigma_\varepsilon^{2(j)}|\beta^{(j)}$ from $p(\sigma_\varepsilon^2|D, \beta) \sim \mathcal{IG}\left(\frac{1}{2}(y - X\beta^{(j)})'(y - X\beta^{(j)}), \frac{1}{2}T\right)$

Marginal densities

Ultimately, we are interested in learning about the properties of the marginal densities of β and σ_ε^2. In this model it is straightforward to derive these. Using the results of Appendix C, the marginal posterior densities are given as

$$p(\beta|D) \sim t\left(\hat{\beta}, \hat{s}^2[X'X]^{-1}, T - K\right) \tag{19}$$

$$p(\sigma_\varepsilon^2|D) \sim \mathcal{IG}\left(\frac{1}{2}(y - X\hat{\beta})'(y - X\hat{\beta}), \frac{1}{2}(T - K)\right) \tag{20}$$

where $\hat{s}^2 = (y - X\hat{\beta})'(y - X\hat{\beta})/(T - K)$. A scheme for the derivations for the joint and marginal posterior densities of the linear regression model is given in Appendix B, Figure B-1. Since in this case one can directly simulate from the marginal densities without having to rely on the Gibbs sampler to obtain posterior results, we present direct sampling results in Appendix B, Figure B-2.

We emphasize that the derivation of conditional and marginal densities does not change if we were to replace $x_t\beta$ by ρy_{t-1} in (10) using a uniform prior. That is, within a noninformative Bayesian analysis one can go from a static analysis with $x_t\beta$ to a dynamic model: the posterior of regression parameters remains Student-t while in the frequentist world one cannot go from a static analysis to a simple dynamic analysis without a change in the properties of the OLS estimators. The same argument also hold for predictive densities which we discuss next.

Predictive densities

Suppose one is interested in constructing forecasts of future values of y_t in the linear regression model. A vector of Q future values, $\widetilde{y} = [y_{T+1}\ y_{T+2}\ ...\ y_{T+Q}]'$, is then assumed to be generated by the following model:

$$\widetilde{y} = \widetilde{X}\beta + \widetilde{\varepsilon} \tag{21}$$

where $\widetilde{X}$ is a $Q \times K$ matrix of given values for the independent variables in the Q future periods and $\widetilde{\varepsilon}$ is a $Q \times 1$ vector of future errors which are assumed to be i.i.d. Normal with

zero mean and variance-covariance matrix $\sigma_\varepsilon^2 \mathrm{I}_Q$. The marginal predictive density for $\widetilde{y}$ can be derived by integrating the joint density $p(\widetilde{y}, \beta, \sigma_\varepsilon^2 | \widetilde{X}, D)$ with respect to β and σ_ε^2:

$$p(\widetilde{y}|\widetilde{X}, D) = \int_{\sigma_\varepsilon^2} \int_\beta p(\widetilde{y}, \beta, \sigma_\varepsilon^2 | \widetilde{X}, D) d\beta \, d\sigma_\varepsilon^2 \tag{22}$$

The joint density is specified as follows:

$$p(\widetilde{y}, \beta, \sigma_\varepsilon^2 | \widetilde{X}, D) = p(\widetilde{y}|\beta, \sigma_\varepsilon^2, \widetilde{X}) p(\beta, \sigma_\varepsilon^2 | D) \tag{23}$$

where $p(\beta, \sigma_\varepsilon^2|D)$ is the posterior density in (14) and $p(\widetilde{y}|\beta, \sigma_\varepsilon^2, \widetilde{X})$ is the conditional predictive density of $\widetilde{y}$ in (21) which is given as:

$$p(\widetilde{y}|\beta, \sigma_\varepsilon^2, \widetilde{X}) \propto (\sigma_\varepsilon^2)^{-\frac{1}{2}Q} \exp\left[-\frac{1}{2\sigma_\varepsilon^2} (\widetilde{y} - \widetilde{X}\beta)'(\widetilde{y} - \widetilde{X}\beta) \right] \tag{24}$$

which is a kernel of a multivariate Normal variable with mean $\widetilde{X}\beta$ and covariance matrix $\sigma_\varepsilon^2 \mathrm{I}_Q$. Scheme 1 shows a Gibbs sampling scheme for predictive analysis.

For each draw of $(\beta, \sigma_\varepsilon^2)$ one can draw $\widetilde{y}$ from (24). The draws that are obtained in this way are draws from the predictive density in (22). The joint density in (23) is a combination of (14) and (24) and becomes:

$$p(\widetilde{y}, \beta, \sigma_\varepsilon^2 | \widetilde{X}, D) \propto (\sigma_\varepsilon^2)^{-\frac{1}{2}(T+Q+2)} \exp\left[-\tfrac{1}{2\sigma_\varepsilon^2} \Big((\widetilde{y} - \widetilde{X}\beta)'(\widetilde{y} - \widetilde{X}\beta) + (y - X\beta)'(y - X\beta) \Big) \right] \tag{25}$$

The first step to *analytically* obtain the marginal predictive density follows from integrating with respect to σ_ε^2 which results in:

$$p(\widetilde{y}, \beta | \widetilde{X}, D) \propto [(\widetilde{y} - \widetilde{X}\beta)'(\widetilde{y} - \widetilde{X}\beta) + (y - X\beta)'(y - X\beta)]^{-\frac{1}{2}(T+Q)}$$

The second step is to complete the squares on β and to integrate with respect to the K elements of β which gives:

$$p(\widetilde{y}|\widetilde{X}, D) \propto [(T - K) + (\widetilde{y} - \widetilde{X}\widehat{\beta})' H (\widetilde{y} - \widetilde{X}\widehat{\beta})]^{-\frac{1}{2}(T+Q-K)} \tag{26}$$

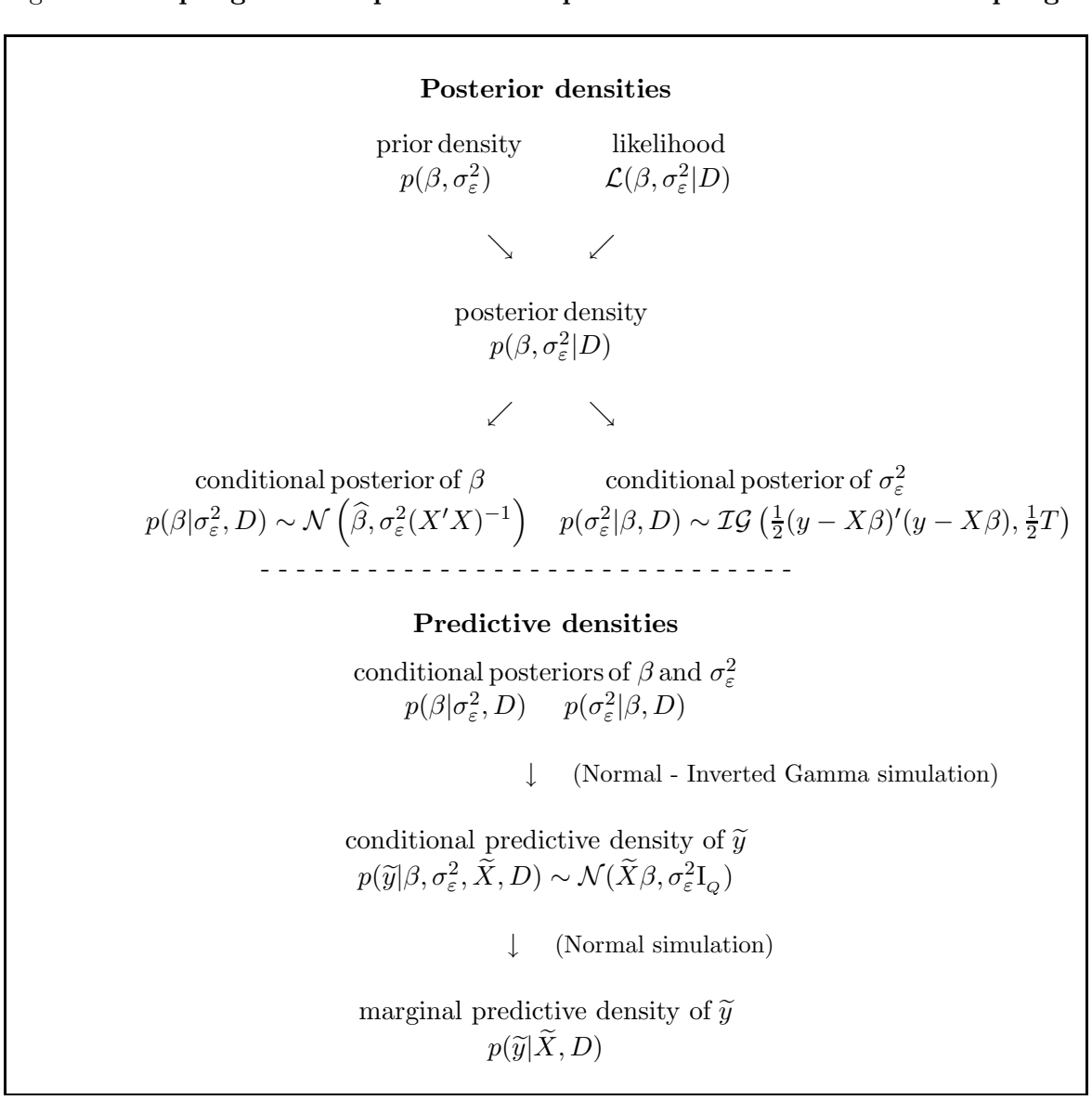
where $H = (I + \widetilde{X}(X'X)^{-1}\widetilde{X}')/\hat{s}^2$. Equation (26) indicates that $\widetilde{y}$ has a multivariate Student-t distribution with mean $\widetilde{X}\widehat{\beta}$, scale matrix H^{-1} and $(T - K)$ degrees of freedom. By means of (26) one can draw *directly* from the predictive density. Schemes 2 and 3, both listed in Appendix B, summarize the derivations of distributions that are needed in a direct sampling and Gibbs sampling scheme.

We emphasize again that in a Bayesian noninformative framework all these derivations carry over directly to a dynamic model with lagged endogenous variables.

4 Single Equation Dynamic Regression Models

For this class of models the near-boundary issue refers to near-instability of dynamic models. This is an important boundary issue in the sense that it has substantial implications for efficient forecasting. The purpose of this section is three-fold. We start with derivations of posteriors of parameters of interest for different dynamic model specifications (and the construction of corresponding Gibbs samplers). We show here the uniformity of the

Figure 5: **Sampling scheme: posterior and predictive results for Gibbs sampling**



Notes: The figure presents results for Gibbs sampling schemes to obtain posterior and predictive results.

derivations for different model structures. We also discuss interpretation of the deterministic terms in autoregressive models with a focus on the issue of near-boundary analysis. The key feature in this context is: under what conditions do the dynamic economic processes under consideration return to a deterministic mean or trend and/or when does there exist a random walk or stochastic trend? Is there a substantial probability mass in the stationary region and/or on the boundary of a random walk or stochastic trend model? These are boundary issues which have important implications for forecasting. Finally, we present empirical illustrations using some major U.S. macroeconomic and financial series.

4.1 Posterior analysis and Gibbs samplers

4.1.1 Linear regression with autocorrelation

We are now ready to analyze the extension of the model in (10) by allowing the error terms to have first-order autocorrelation[11]. That is:

$$y_t = x_t\beta + \nu_t, \qquad t = 1, ..., T \tag{27}$$

$$\nu_t = \rho\nu_{t-1} + \varepsilon_t, \qquad \text{with} \qquad \varepsilon_t \sim \text{i.i.d}\, \mathcal{N}(0, \sigma_\varepsilon^2) \tag{28}$$

where ρ is the parameter that determines the strength of the autocorrelation. For expository purposes with respect to the derivation of the conditional and marginal posterior densities we distinguish between two cases: one where the domain of ρ is not restricted and one where it is. We emphasize that for economic purposes the domain of this parameter is in most cases restricted to the interval $-1 \leq \rho \leq 1$. We note that later we will distinguish between the cases where ρ is 1 and where ρ is in the bounded interval of $(0, 1)$. The domain for the remaining parameters is given by $-\infty < \beta < \infty$ and $0 < \sigma_\varepsilon^2 < \infty$. When $\rho = 0$, the autocorrelation model coincides with the basic linear regression model since ν_t reduces to a white noise series. As we will see later, difficulties occur when there is a constant term and ρ has substantial posterior probability mass at the edges of its domain. By substituting (28) in (27) and rewriting the resulting expression in matrix notation, we obtain

$$y - \rho y_{-1} = X\beta - X_{-1}\beta\rho + \varepsilon, \qquad \text{with} \qquad \varepsilon \sim \text{i.i.d}\, \mathcal{N}(0, \sigma_\varepsilon^2 \mathrm{I}_T) \tag{29}$$

where y_{-1} and X_{-1} denote the one-period lagged values of y and X. This reformulation shows that the autocorrelation model is nonlinear in its parameters β and ρ. This problem of inference on a product (or ratio) of parameters is a classic issue. A detailed analysis is, however, beyond the scope of the present paper. For an early example see Press (1969) and we refer to Fieller (1954) and Van Dijk (2003) for more references. Although this issue of nonlinearity hampers parameter estimation and inference when using frequentist estimation approaches, obtaining posterior results using Gibbs sampling is straightforward as we will show below, in the case where ρ is unrestricted and no such deterministic terms as a constant or trend occur in the equation. We first turn to deriving the joint, conditional and marginal densities. It will become apparent that the autocorrelation model serves as a template for several other well-known econometric models.

Joint, posterior and marginal densities

The combination of the likelihood of the autocorrelation model with the noninformative prior in (13) and a uniform prior on ρ on a large region, and, further, assuming that the initial observations are fixed nonrandom quantities, gives the following joint posterior density,

$$p(\beta, \rho, \sigma_\varepsilon^2 | D) \propto (\sigma_\varepsilon^2)^{-\frac{1}{2}(T+2)} \exp\left[-\frac{1}{2\sigma_\varepsilon^2}\big(y - \rho y_{-1} - X\beta + X_{-1}\beta\rho\big)'\big(y - \rho y_{-1} - X\beta + X_{-1}\beta\rho\big)\right] \tag{30}$$

where D once again represents the known data (y, X). In case the domain of the parameter ρ is bounded, we make use of an indicator function $\mathrm{I}(\beta, \rho)$ which is 1 on the domain

[11] For a more general discussion on Bayesian inference in dynamic econometric models, we refer to Chib (1993) and Chib and Greenberg (1994).

specified (which is usually $(-\infty < \beta < \infty), (-1 \leq \rho \leq 1)$ and 0 elsewhere). Thus, we obtain a truncated posterior density defined on the region indicated. We now derive the expression for the conditional densities $p(\beta|\rho, \sigma_\varepsilon^2, D)$, $p(\rho|\beta, \sigma_\varepsilon^2, D)$ and $p(\sigma_\varepsilon^2|\beta, \rho, D)$ and the marginal densities $p(\beta|D)$ and $p(\rho|D)$. For analytical convenience, we start with the derivations for the case where ρ is not restricted.

To facilitate the derivation of the conditional densities it is useful to rewrite the model in (29) in two different ways. In each case we condition on one of the two types of regression coefficients. First, we rewrite (29) *conditional on values of* ρ,

$$y^* = X^*\beta + \varepsilon \qquad \text{where} \qquad \begin{cases} y^* = y^*(\rho) \equiv y - \rho y_{-1} \\ X^* = X^*(\rho) \equiv X - \rho X_{-1} \end{cases} \tag{31}$$

Second, we rewrite (29) *conditional on values of* β which then becomes

$$\tilde{y} = \rho\tilde{y}_{-1} + \varepsilon \qquad \text{where} \qquad \begin{cases} \tilde{y} = \tilde{y}(\beta) \equiv y - X\beta \\ \tilde{y}_{-1} = \tilde{y}_{-1}(\beta) \equiv y_{-1} - X_{-1}\beta \end{cases} \tag{32}$$

To derive the conditional density for β we use (31) to rewrite the joint posterior density. Doing so gives us the joint density of the basic linear regression model again so we can re-use all our earlier derivations. It therefore follows immediately that the conditional density for β is multivariate Normal with mean $m = \widehat{\beta}^* \equiv (X^{*\prime}X^*)^{-1}X^{*\prime}y^*$ and variance matrix $S = S_\beta \equiv \sigma_\varepsilon^2(X^{*\prime}X^*)^{-1}$. Similarly, using (32) we obtain the result that the conditional density for the unrestricted parameter ρ is Normal with mean $m = \hat{\rho} \equiv (\tilde{y}_{-1}{}'\tilde{y}_{-1})^{-1}\tilde{y}_{-1}{}'\tilde{y}$ and variance $s^2 = \sigma_\rho^2 \equiv \sigma_\varepsilon^2(\tilde{y}_{-1}{}'\tilde{y}_{-1})^{-1}$. The conditional density for σ_ε^2 is again Inverted Gamma with parameter $m = \frac{1}{2}\varepsilon'\varepsilon \equiv \frac{1}{2}\big(y - \rho y_{-1} - X\beta + X_{-1}\beta\rho\big)'\big(y - \rho y_{-1} - X\beta + X_{-1}\beta\rho\big)$ and $\nu = \frac{1}{2}T$ degrees of freedom. Summarizing, we have

$$\begin{aligned} p(\beta|\rho, \sigma_\varepsilon^2, D) &\sim \mathcal{N}\left(\widehat{\beta}, S_\beta\right) \\ p(\rho|\beta, \sigma_\varepsilon^2, D) &\sim \mathcal{N}\left(\widehat{\rho}, \sigma_\rho^2\right) \\ p(\sigma_\varepsilon^2|\beta, \rho, D) &\sim \mathcal{IG}\left(\frac{1}{2}\varepsilon'\varepsilon, \frac{1}{2}T\right) \end{aligned}$$

Whereas in the basic regression model Gibbs sampling was unnecessary because the marginal densities could be derived analytically, here we do not have analytical results and therefore we need Gibbs sampling. This is due to the fact that the marginal densities of β, ρ and σ_ε^2 are not a member of any known class of densities. We show this as follows. After integrating out σ_ε^2 from the joint density we get

$$p(\beta, \rho|D) \propto \left[\big(y - \rho y_{-1} - X\beta + X_{-1}\beta\rho\big)'\big(y - \rho y_{-1} - X\beta + X_{-1}\beta\rho\big)\right]^{-\frac{T}{2}} \tag{33}$$

We can rewrite this joint density in two different ways

$$p(\beta, \rho|D) \propto \left[\tilde{y}'M_{\tilde{y}_{-1}}\tilde{y} + (\rho - \hat{\rho})'\tilde{y}_{-1}{}'\tilde{y}_{-1}(\rho - \hat{\rho})\right]^{-\frac{T}{2}} \tag{34}$$

$$p(\beta, \rho|D) \propto \left[y^{*\prime}M_{X^*}y^* + (\beta - \beta^*)'X^{*\prime}X^*(\beta - \beta^*)\right]^{-\frac{T}{2}} \tag{35}$$

where $M_{\tilde{y}_{-1}}$ and M_{X^*} are idempotent residual maker matrices of $\tilde{y}_{-1}$ and X^* respectively[12].

[12]The general residual maker matrix is given as $M_A = \mathrm{I}_T - A(A'A)^{-1}A'$.

Integrating out ρ from (34) and β from (35) gives the marginal densities

$$p(\beta|D) \propto \left[(y - X\beta)' M_{y_{-1} - X_{-1}\beta}(y - X\beta)\right]^{-\frac{T-1}{2}} \left[(y_{-1} - X_{-1}\beta)'(y_{-1} - X_{-1}\beta)\right]^{-\frac{1}{2}} \quad (36)$$

$$p(\rho|D) \propto \left[(y - \rho y_{-1})' M_{X - \rho X_{-1}}(y - \rho y_{-1})\right]^{-\frac{T-K}{2}} \left[(X - \rho X_{-1})'(X - \rho X_{-1})\right]^{-\frac{1}{2}} \quad (37)$$

In case the parameter ρ is restricted to the interval $[-1, 1]$ we proceed as follows. Equations (30), (34) and (35) are now changed with inclusion of the indicator function $\mathrm{I}(\beta, \rho)$. The right hand side of equation (36) now contains the function $c(\beta)$ given as $c(\beta) = \Phi\left(\frac{1-\hat{\rho}}{\sigma_\rho}\right) - \Phi\left(\frac{-1-\hat{\rho}}{\sigma_\rho}\right)$ where Φ stands for the standard Normal distribution function. The conditional normal density of ρ given β, σ_ε^2 and D is in this case a truncated normal density and the right hand side of equation (37) is now changed with the inclusion of the indicator function $\mathrm{I}(\rho)$ which is defined as 1 on the interval $[-1, 1]$ and 0 elsewhere.

Both densities in (36) and (37) - and their truncated variants - do not belong to a known class of density functions which means that we need Gibbs sampling to obtain posterior results. Despite the fact that the marginal densities of β and ρ can not be determined analytically, applying the Gibbs sampler is, however, a straightforward exercise, conditional upon the fact that all variables in the data matrix X have some nontrivial data variability.

Fisher Information matrix

The Fisher Information matrix can provide information as to whether problems are likely to occur when ρ approaches the edges of its domain, in the sense that the joint posterior density becomes improper. The Fisher Information matrix is defined as minus the expectation of the matrix of second order derivatives of the log likelihood with respect to the parameter vector $\boldsymbol{\theta} = (\beta, \rho, \sigma_\varepsilon^2)$, i.e. $\mathcal{I} = -E[\frac{\delta^2 \ln L(\boldsymbol{\theta}|D)}{\delta\boldsymbol{\theta}\delta\boldsymbol{\theta}'}]$. For the linear model with autocorrelation the Information matrix is given by[13]

$$\mathcal{I} = -E\begin{bmatrix} \frac{\delta^2 \ln L}{\delta\rho^2} & \frac{\delta^2 \ln L}{\delta\rho\delta\beta'} & \frac{\delta^2 \ln L}{\delta\rho\delta\sigma_\varepsilon^2} \\ \frac{\delta^2 \ln L}{\delta\beta\delta\rho} & \frac{\delta^2 \ln L}{\delta\beta\delta\beta'} & \frac{\delta^2 \ln L}{\delta\beta\delta\sigma_\varepsilon^2} \\ \frac{\delta^2 \ln L}{\sigma_\varepsilon^2\delta\rho} & \frac{\delta^2 \ln L}{\delta\sigma_\varepsilon^2\delta\beta'} & \frac{\delta^2 \ln L}{\delta\sigma_\varepsilon^4} \end{bmatrix} = \begin{bmatrix} \frac{T}{1-\rho^2} & 0 & 0 \\ 0 & \frac{(X-\rho X_{-1})'(X-\rho X_{-1})}{\sigma_\varepsilon^2} & 0 \\ 0 & 0 & \frac{T}{2\sigma_\varepsilon^4} \end{bmatrix} \quad (38)$$

The inverse of the Information matrix shows that even when $|\rho| = 1$ none of the variances 'explode'. In the next sections we will see that this not always needs to be the case. More general, under the condition that all variables in X have some variability, there are no issues in terms of impropriety of the joint posterior density when ρ reaches the edges of its domain.

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

Gibbs sampling for the unrestricted case of ρ

j^{th} **Gibbs step for the linear regression model with autocorrelation:**			
- generate	$\beta^{(j)}\|\rho^{(j-1)}, \sigma_\varepsilon^{2(j-1)}$	from	$p(\beta\|\rho, \sigma_\varepsilon^2, D) \sim \mathcal{N}\left(\widehat{\beta^*}^{(j-1)}, S_\beta^{(j-1)}\right)$
- generate	$\rho^{(j)}\|\beta^{(j)}, \sigma_\varepsilon^{2(j-1)}$	from	$p(\rho\|\beta, \sigma_\varepsilon^2, D) \sim \mathcal{N}\left(\hat{\rho}^{(j)}, \sigma_\rho^{2(j-1)}\right)$
- generate	$\sigma_\varepsilon^{2(j)}\|\beta^{(j)}, \rho^{(j)}$	from	$p(\sigma_\varepsilon^2\|\beta, \rho, D) \sim \mathcal{IG}\left(\frac{1}{2}\varepsilon^{(j)\prime}\varepsilon^{(j)}, \frac{1}{2}T\right)$

We can see from the conditional densities given earlier that the Gibbs sampler has no difficulties with the nonlinearities in the likelihood. This is due to the fact that *conditionally* on one regression parameter, the model for the other regression parameter is the basic linear regression model as shown in (31) and (32). In fact, the joint posterior density of ρ and any element of β, or the other way around, resembles the density shown in Figure 2(a). Therefore, the Gibbs sampler is a very convenient tool for drawing inference on the parameters in these types of models.

We distinguish between a Gibbs step when there is no truncation for ρ (then all draws are accepted) and the case of a truncated domain for ρ. In the latter case, a simple solution for the Gibbs step is to ignore drawings outside the bounded region ($-1 < \rho < 1$). A more efficient algorithm has been developed by Geweke (1991, 1996).

4.1.2 Distributed lag models: Koyck model

A further extension of the basic linear regression model is the univariate distributed lag model[14]. This model has proven to be one of the workhorses of econometric modelling practice since it offers the econometrician a straightforward tool to investigate the dependence of a variable on its own history or on the history 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, governed by a single unknown parameter. The resulting model is known as the geometric distributed lag model or simply as the Koyck model. Below, we discuss the boundary issue that can occur in this model which results in a parameter (near) nonidentification issue.

The Koyck model is given by

$$y_t = \beta w_t + \nu_t, \qquad t = 1, ..., T \tag{39}$$

$$w_t = (1-\rho)\sum_{i=0}^{\infty} \rho^i x_{t-i} \tag{40}$$

$$\nu_t = \rho\nu_{t-1} + \varepsilon_t, \qquad \text{with} \qquad \varepsilon_t \sim \text{i.i.d}\, \mathcal{N}(0, \sigma_\varepsilon^2) \tag{41}$$

where we allow for first-order autocorrelation in the error term. It is assumed that $0 \leq \rho \leq 1$, $-\infty < \beta < \infty$ and $0 < \sigma_\varepsilon^2 < \infty$. Note that the effect of lagged values of the (here single) explanatory variable x_t is determined solely by ρ and that this parameter is assumed to be equal to the first-order autocorrelation parameter. In marketing econometrics the parameter ρ is usually referred to as the 'retention' parameter.

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

We assume that ν_t is serially correlated. One may also assume that ν_t is i.i.d. Then the transformed model has an MA(1) error. Another closely related model that also gives a boundary problem is the so called partial adjustment model[15]. This model is given as

$$\begin{aligned} y_t^* &= \beta w_t + \nu_t && (42) \\ y_t - y_{t-1} &= (1-\rho)(y_t^* - y_{t-1}) + \varepsilon_t && (43) \end{aligned}$$

where y_t is observed but y_t^* unobserved.

For the Koyck model, substituting (41) in (39) gives a similar type of expression as we found for the linear model with autocorrelation. In particular, with matrix notation one obtains

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

Equation (40) puts additional structure on the term $w - \rho w_{-1}$. More specifically, it holds that $w - \rho w_{-1} = (1-\rho)x$ which gives

$$y = \rho y_{-1} + \beta(1-\rho)x + \varepsilon, \qquad \text{with} \qquad \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2 \mathrm{I}_T) \tag{44}$$

The result in (44) shows that the Koyck model is nested in the autocorrelation model and that therefore all earlier derivations hold here as well. The main difference, however, is that contrary to the autocorrelation model, the specific structure that is placed on the exogenous variable will result in a boundary issue when ρ is near 1. We can understand why this is the case by realizing that β will be near nonidentification for values of ρ close to 1. This means that y effectively becomes a random walk and that exogenous variables no longer have any influence on y. When $\rho = 1$, then β is not identified and the model reduces to a random walk. We will analyze the joint, conditional and marginal densities to give insights in the consequences of the nonidentification of β when applying the Gibbs sampler.

Joint, posterior and marginal densities

Derivations for the joint and conditional densities are very similar to before. Therefore we only report the joint and conditional densities for the case of the bounded domain of ρ. The joint density, after integrating out σ_ε^2, is specified as

$$p(\beta, \rho | D) \propto \left[\big(y - \rho y_{-1} - \beta(1-\rho)x\big)' \big(y - \rho y_{-1} - \beta(1-\rho)x\big) \right]^{-\frac{T}{2}} \mathrm{I}(\beta, \rho)$$

where $\mathrm{I}(\beta, \rho)$ is an indicator function which is 1 on the region bounded by $0 \leq \rho \leq 1, -\infty < \beta < \infty$ and 0 elsewhere. The conditional densities - given that ρ is an element of the interval (0,1) - are given as

$$\begin{aligned} p(\beta | \rho, \sigma_\varepsilon^2, D) &\sim \mathcal{N}\left(\beta^*, \sigma_\beta^2\right) \\ p(\rho | \beta, \sigma_\varepsilon^2, D) &\sim \mathcal{TN}\left(\widehat{\rho}, \sigma_\rho^2\right) \\ p(\sigma_\varepsilon^2 | \beta, \rho, D) &\sim \mathcal{IG}\left(\frac{1}{2}\varepsilon'\varepsilon, \frac{1}{2}T\right) \end{aligned}$$

[15] We are indebted to William Griffiths for pointing this out.

where $\mathcal{TN}$ indicates a Truncated Normal density. The parameters in the conditional densities are specified as

$$\begin{aligned}\beta^* &= (x^{*\prime}x^*)^{-1}x^{*\prime}y^* = \left[(1-\rho)^2 x'x\right]^{-1} x'(y-\rho y_{-1}) && (45)\\ \sigma_\beta^2 &= \sigma_\varepsilon^2 (x^{*\prime}x^*)^{-1} = \sigma_\varepsilon^2 \left[(1-\rho)^2 x'x\right]^{-1} && (46)\end{aligned}$$

and

$$\begin{aligned}\hat{\rho} &= (\tilde{y}_{-1}'\tilde{y}_{-1})^{-1}\tilde{y}_{-1}'\tilde{y} = \left[(y_{-1}-\beta x)'(y_{-1}-\beta x)\right]^{-1}(y_{-1}-\beta x)'(y-\beta x) && (47)\\ \sigma_\rho^2 &= \sigma_\varepsilon^2 (\tilde{y}_{-1}'\tilde{y}_{-1})^{-1} = \sigma_\varepsilon^2 \left[(y_{-1}-\beta x)'(y_{-1}-\beta x)\right]^{-1} && (48)\end{aligned}$$

and $\varepsilon'\varepsilon = \big(y-\rho y_{-1}-\beta(1-\rho)x\big)'\big(y-\rho y_{-1}-\beta(1-\rho)x\big)$. The density for ρ is truncated to the unit interval which is indicated by the density notation $\mathcal{TN}$.

At first sight, it may seem straightforward to apply the Gibbs sampler to the Koyck model. However, upon closer inspection of the conditional density parameters it becomes clear that a problem can occur for values of ρ close to 1. Suppose that a value near 1 is drawn for ρ. The conditional variance of β given this draw will be close to infinity, see (46), which means that any large value is likely to be drawn for β. If the next draw for β is indeed large then the conditional variance of ρ goes to zero, see (48). As a result the next draw for ρ is again going to be close to 1, see (47). This means that the Gibbs Markov chain will converge very slowly. Convergence is not achieved for the case $\rho = 1$ since this is an absorbing state of the Markov chain. The extent of this problem depends on how much probability mass there actually exists close to $\rho = 1$ and at $\rho = 1$.

When $\rho = 1$, it follows directly from the joint posterior $p(\beta, \rho|D)$ that $p(\beta|\rho = 1, D)$ is constant. Thus when $\rho = 1$ the conditional density of β is uniform on the interval $(-\infty < \beta < \infty)$ and as a consequence it is improper. The conditional density of ρ is just the value of the truncated normal in the point $\rho = 1$. The economic issue is that we cannot learn (draw inference) on the parameter when $\rho = 1$, which is basically very natural for a random walk model.

To understand the behavior of the Gibbs sampler further we examine the marginal densities in detail. Given $0 < \rho < 1$, the marginal densities for β and ρ are as follows:

$$\begin{aligned}p(\beta|D) &\propto \left[(y-\beta x)' M_{y_{-1}-\beta x}(y-\beta x)\right]^{-\frac{T-1}{2}} \left[(y_{-1}-\beta x)'(y_{-1}-\beta x)\right]^{-\frac{1}{2}} c(\beta) && (49)\\ p(\rho|D) &\propto \left[(y-\rho y_{-1})' M_{(1-\rho)x}(y-\rho y_{-1})\right]^{-\frac{T-1}{2}} \left[x'x\right]^{-\frac{1}{2}} (1-\rho)^{-1} && (50)\end{aligned}$$

where $c(\beta)$ is similar as in the previous section but now such that ρ is defined on the bounded interval (0,1).

Focusing on the density for ρ, we can recognize it to be a Student-t type density, but with an additional factor $(1-\rho)^{-1}$. It is exactly this factor that is causing the behavior of the Gibbs sampler. The reason is that the joint density $p(\beta, \rho|D)$ is improper at $\rho = 1$ for $-\infty < \beta < \infty$. Graphically, this means that the joint density has a 'wall', similar to the ridge that was depicted in Figure 3(e). The marginal density for ρ will tend to infinity when ρ tends to 1.

To reiterate what we said before, the extent of the problem - given the specification of the model - depends on the data at hand. If the likelihood assigns virtually no probability mass to the region close to $\rho = 1$ then the marginal for β will be virtually indistinguishable from a Student-t density. Furthermore, the marginal density for ρ will still tend to infinitely close to $\rho = 1$, but if this event happens to be far out in the tail of the distribution then

this should not pose a serious problem. We shall show an example of this data feature in the empirical analysis relating to U.S. inflation and growth of real GDP. If, on the other hand, substantial probability mass is near $\rho = 1$ then measures should be taken to prevent the Gibbs sampler from reaching that part of the domain of ρ or, alternatively, to try and regularize the likelihood. Choosing an appropriate prior density can do the trick.

Fisher Information matrix

Analyzing the Information matrix gives similar insights in the irregularity in the joint density close to and equal to $\rho = 1$ and furthermore, it provides us with a direction for a possible solution to tackle this irregularity. The Information matrix follows directly from (38) by substituting in $X - \rho X_{-1} = (1 - \rho)x$. Therefore

$$\mathcal{I} = \begin{bmatrix} \frac{T}{1-\rho^2} & 0 & 0 \\ 0 & \frac{(1-\rho)^2 x'x}{\sigma_\varepsilon^2} & 0 \\ 0 & 0 & \frac{T}{2\sigma_\varepsilon^4} \end{bmatrix} \tag{51}$$

The Information matrix again shows that when ρ is close to 1, the variance of ρ is zero (the inverse of the first diagonal element) whereas the variance of β goes to infinity (the inverse of the second diagonal element). When $\rho = 1$, then the determinant of Information matrix is zero.

Gibbs sampling when $0 < \rho < 1$

The Gibbs j^{th} step is given by

j^{th} Gibbs step for the distributed lag model:

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

When ρ is 1 it follows that Gibbs sampling is inappropriate.

Potential solutions: truncation of parameter region, Information matrix prior or training sample prior

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

In terms of applying the first solution, one can truncate the domain of ρ and check whether there is probability mass near 1. Imposing an upper bound can be achieved by selecting for example a local uniform prior. The goal would be to only allow draws for ρ that are at least η away from 1 with $\eta > 0$ to prevent a wall in the joint posterior density.

Choosing a specific value for η would necessarily be a subjective choice. However, once a value for η is agreed upon one can apply the Gibbs sampler. Alternatively, one can use a Metropolis-Hastings type step in which only draws that fall below $1-\eta$ are accepted. For an example of this method, see e.g. Geman and Reynolds (1992).

As for the second solution, one can try and regularize the likelihood in the neighborhood of $\rho = 1$ such that it becomes a proper density. This can be achieved by using a prior that is chosen in such the way that it eliminates the factor $(1-\rho)^{-1}$. From the Information matrix in (51) we can construct the following Jeffreys' type prior for β given ρ and σ_ε^2,[16]

$$p(\beta|\rho,\sigma_\varepsilon^2) \propto \frac{(1-\rho)^2}{\sigma_\varepsilon^2} \qquad \text{for} \qquad 0<\rho<1 \tag{52}$$

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

The third solution is an alternative way of regularizing the posterior density. One can use a training sample[17] to specify a weakly informative prior for β. Schotman and Van Dijk (1991a) specify the following prior

$$p(\beta|\rho,\sigma_\varepsilon^2) \propto \mathcal{N}\left(y_0, \tfrac{\sigma_\varepsilon^2}{(1-\rho)^2}\right) \qquad \text{for} \qquad 0<\rho<1 \tag{53}$$

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

We conclude that for solutions (ii) and (iii) one has to - in most cases - replace the simple Gibbs procedures by other Monte Carlo integration methods. This is a topic outside the scope of the present paper. Further solutions, which we do not discuss here in the detail, are to reparameterize the model in such a way that the Gibbs sampler can be used without any problems for the reformulated model. However, one still has to translate the posterior results back to the original model. Without imposing some sort of prior, similar problems will still occur only now at a different stage in the analysis. For examples of reparametrization see for instance Gilks *et al.* (2000). Finally, modified versions of the Gibbs sampler such as the Collapsed Gibbs sampler (see Liu, 1994), where some parameters can be temporarily ignored when running the Gibbs sampler (in this case ρ) can be useful in this context as well. In the empirical application in Section 5 we only use the truncation prior approach.

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

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

4.1.3 Autoregressive models and error correction models with deterministic components

We present the issue of near-boundary analysis in the context of an autoregressive model with deterministic components. The simplest example is a first-order autoregressive model with an additive constant, given as

$$y_t = c + \rho y_{t-1} + \varepsilon_t \tag{54}$$

This model can be respecified as an error correction model (ECM) around a constant mean using a restriction on c. We start with rewriting (54) as

$$y_t = \mu(1-\rho) + \rho y_{t-1} + \varepsilon_t \tag{55}$$

where c is now restricted as $c = \mu(1-\rho)$. We can rewrite the latter equation as a mean reversion model, see Schotman and Van Dijk (1991a,b),

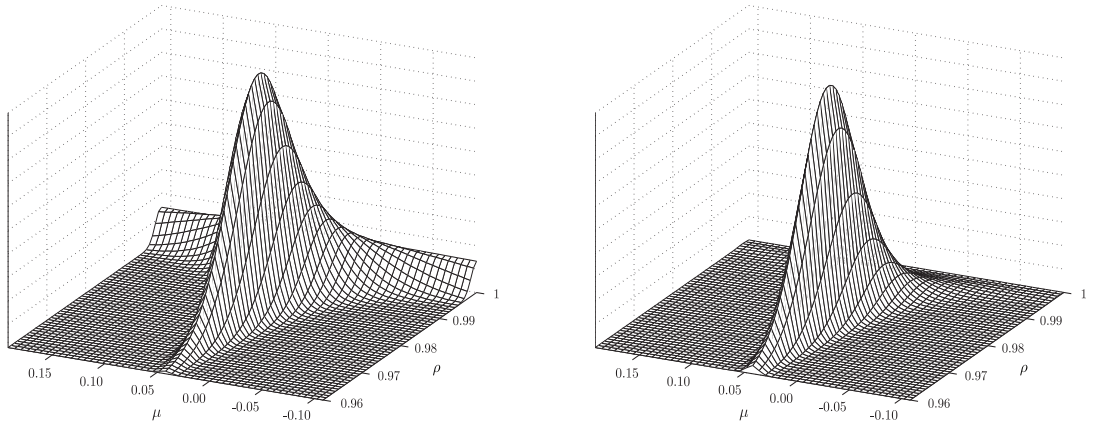
$$\Delta y_t = (\rho - 1)(y_{t-1} - \mu) + \varepsilon_t \tag{56}$$

Here one can see the expected 'return to the long-term unconditional mean (μ) of the series' when $0 < \rho < 1$. That is, when y_{t-1} is greater than μ and $0 < \rho < 1$, then the conditional expected change in y_t, given previous observations, is negative while in the opposite case the expected change in positive. Furthermore, when ρ tends to 1 then in the ECM specification (55) one has a smooth transition from stationarity to a random walk model. In other words one approaches the boundary in a continuous way. On the contrary in equation (54) one has a transition from stationarity to a random walk with drift: one hits the boundary with a 'jump'. The models are much farther apart then the ones in the ECM setup. Note that the constant term c in (54) does not have a direct interpretation in terms of being the mean of the process while in (55) the constant μ is the long term unconditional mean of the series given $0 < \rho < 1$.

Similar as for the Koyck model, imposing this particular ECM structure introduces a boundary issue when there is substantial posterior probability near $\rho = 1$. In the ECM model for y_t, the interpretation of μ depends on whether the series y is stationary ($\rho < 1$) or whether it has a unit root (i.e. $\rho = 1$). In the latter case, the mean of y does not exist and μ is thus nonidentified. Therefore, even when y is a weakly stationary process, any value for μ along the real line is likely to be drawn in the Gibbs sampler when ρ is sampled close to 1. This will not only make it very difficult to pinpoint the posterior mean of μ but it also causes the sequence of draws for ρ to have difficulties moving away from $\rho = 1$. Of course, ρ close to 1 can be an indication that one should model first differences of y instead of y itself which will circumvent the entire issue altogether. However, for series such as interest rate levels there is no clear economic interpretation why these should be I(1) processes and one is left with dealing with the boundary issue nonetheless.

For series that are near unit root, substantial probability mass will lie close to $\rho = 1$ and at $\rho = 1$ so that the impropriety of the joint posterior poses a serious issue. As an example we depicted the joint density for the unit root model for a series of monthly data on the 10-year U.S. Treasury Bond yield in Figure 6(a). A time-series plot of this series is given in Figure 7(b). Figure 6(a) clearly shows the pronounced wall feature close to and at $\rho = 1$. In order to resolve the impropriety of the joint density a local uniform prior or truncation of the domain for ρ could be used. Another possibility would be to use a regularizing prior like the Schotman and Van Dijk (1991a) prior. The joint density that results from

Figure 6: **Joint posterior density in the unit root model**



(a) Uniform prior (b) Schotman and Van Dijk (1991a) prior

Notes: Panel (a) shows the joint posterior density $p(\rho, \mu|y)$ when we use a uniform prior as in (13) whereas panel (b) shows the same posterior density however now with the prior proposed by Schotman and Van Dijk (1991a) as given in (53). In both panels we use the end-of-month 10-year U.S. Treasury Bond constant maturity yield for the period January 1960-July 2007 as the data vector y.

combining the data likelihood with this particular prior is shown in Figure 6(b). The joint density no longer has a wall feature close to $\rho = 1$ although it still flattens out somewhat near the edge of the domain. We note that this posterior may also be interpreted as the exact likelihood including the initial observation. For details see Schotman and Van Dijk (1991a).

We emphasize that the autoregressive model with an additive constant, equation (54), can be treated like the linear regression model of Section 3. Direct sampling or a simple Gibbs procedure is possible. The model with the ECM interpretation can be written as in the autoregressive form of Sections 4.1.1 and 4.1.2 and deriving the corresponding Gibbs sampling formulas is left to the interested reader. We also refer to that subsection for the convergence issues of the Gibbs sampler.

Next, we treat the autoregressive model with additive linear trend. We start with a distributed lag model of order two,

$$y_t = c + \beta t + \rho_1 y_{t-1} + \rho_2 y_{t-2} + \varepsilon_t \qquad \text{with} \qquad \varepsilon_t \sim \text{i.i.d}\, \mathcal{N}(0, \sigma_\varepsilon^2) \tag{57}$$

where t captures a linear increasing trend. We can rewrite this model as an error correction model as follows. Consider

$$\left(1 - \rho_1 L - \rho_2 L^2\right)(y_t - \mu - \delta t) = \varepsilon_t \qquad \text{with} \qquad \varepsilon_t \sim \text{i.i.d}\, \mathcal{N}(0, \sigma_\varepsilon^2) \tag{58}$$

using $c = \mu(1 - \rho_1 - \rho_2) + \delta(\rho_1 + 2\rho_2)$ and $\beta = (1 - \rho_1 - \rho_2)\delta$ and where L is the lag operator; $Ly_t = y_{t-1}$. Applying this operator to equation (58) we obtain

$$y_t = (1 - \rho_1 - \rho_2)\mu + \delta\left(t - \rho_1(t-1) - \rho_2(t-2)\right) + \rho_1 y_{t-1} + \rho_2 y_{t-2} + \varepsilon_t \tag{59}$$

This equation can be rewritten further as

$$\Delta y_t = \delta + (\rho_1 + \rho_2 - 1)\left(y_{t-1} - \mu - \delta(t-1)\right) - \rho_2\left(\Delta y_{t-1} - \delta\right) + \varepsilon_t \tag{60}$$

which shows that y_t is mean-reverting towards a linear trend when $\rho_1 + \rho_2 < 1$, otherwise we have a random walk with drift (if $\rho_1+\rho_2 = 1$). This is our ECM model with linear trend. Similar as in Section 4.1.1, the derivation of the conditional densities can be simplified if we rewrite (60) by conditioning on one of the two types of regression coefficients. Like for the linear model with autocorrelation, the idea is that given $\rho = [\rho_1, \rho_2]'$ one has a linear model in $\beta = [\mu, \delta]'$ whereas given β one has a linear model in ρ. First, we rewrite (58) conditional on values for ρ:

$$y_t^* = X_t^*\beta + \varepsilon_t \quad \text{where} \quad \begin{cases} y_t^* = y_t^*(\rho) \equiv y_t - \rho_1 y_{t-1} - \rho_2 y_{t-2} \\ X_t^* = X_t^*(\rho) \equiv [1 - \rho_1 - \rho_2, t - \rho_1(t-1) - \rho_2(t-2)] \end{cases} \tag{61}$$

$$\tilde{y}_t = \rho\tilde{y}_{t-} + \varepsilon_t \quad \text{where} \quad \begin{cases} \tilde{y}_t = \tilde{y}_t(\beta) \equiv y - \mu - \delta t \\ \tilde{y}_{t-} = \tilde{y}_{t-}(\beta) \equiv [y_{t-1} - \mu - \delta(t-1), y_{t-2} - \mu - \delta(t-2)] \end{cases} \tag{62}$$

Posterior densities and predictive densities can now be derived in the same fashion as for the linear and autoregressive models from Section 3 and 4.1.2. We note that we use the restriction $\rho_1 + \rho_2 < 1$ in the Gibbs sampling scheme, since in the unrestricted case the posterior is improper.

We can summarize this section by stating that what we did was to make a distinction between additive deterministic terms in autoregressive models and 'interpretable' deterministic terms in error correction models. The interpretation of mean or trend reversion is important in economics. In addition, the resulting forecasts from these models can be quite different. Given our interest in near-boundary analysis, the smooth transition to the boundary (from stationarity to unit roots) is relevant for model comparison, model averaging and forecasting. This is a well-known topic in the literature, see Sims and Uhlig (1991) and Schotman and Van Dijk (1991a,b, 1993) among others. In this paper we do, however, not focus on computing posterior model probabilities for model selection for the choice or test for stationarity and unit root cases. Our interest in these models is primarily from a forecasting perspective.

4.2 Illustrative empirical analysis using macroeconomic series

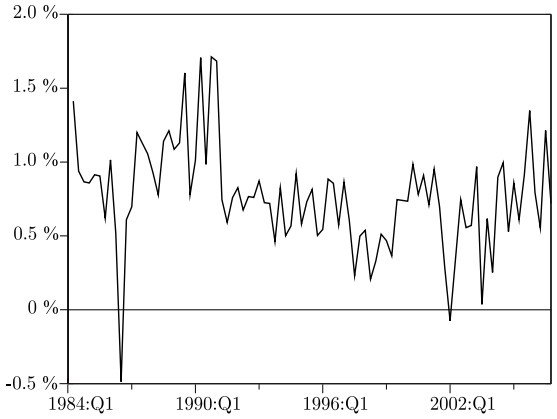
4.2.1 Possible unit root models in inflation, interest rates and dividend yield

Before we apply the models discussed in this section to analyze our main macroeconomic series of interest, U.S. GDP growth, we apply the autoregressive model of the previous paragraph to three time series which are of important economic relevance: U.S. inflation, the 10-year U.S. Treasury Bond yield and the Standard & Poor's 500 Index Dividend Yield. Figure 7 shows time series plots of each of the series. We note that this section is by no means meant to be a full attempt at modelling these series empirically, it is purely for illustrative purposes. We aim at analyzing the posterior mean, a unit root and, next, how to deal with the latter in a model averaging procedure for forecasting purposes.

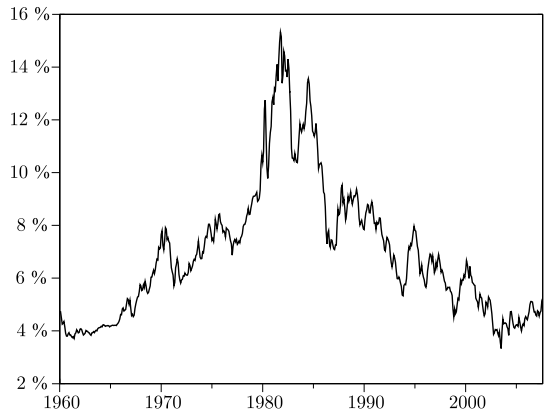
The first series we analyze is inflation. We collected quarterly U.S. CPI figures from the Federal Reserve of Philadelphia database. We then construct inflation, π_t, as the quarterly differences in log price levels, $\pi_t = \ln(CPI_t) - \ln(CPI_{t-1})$. The data sample runs from 1984:Q1 to 2006:Q3, for a total of 91 observations. The model we estimate is specified as in (55):

$$\pi_t - \mu = \rho(\pi_{t-1} - \mu) + \varepsilon_t, \qquad \text{with} \qquad \varepsilon_t \sim \text{i.i.d}\,\mathcal{N}(0, \sigma_\varepsilon^2) \tag{63}$$

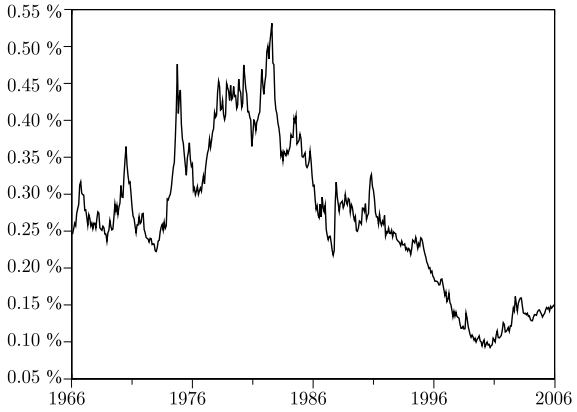
Figure 7: **Macroeconomic and financial series**



(a) inflation



(b) 10-year Treasury Bond yield



(c) Dividend Yield

Notes: Shown in this figure are in Panel a) quarterly changes in log U.S. price levels (CPI), Panel (b) end-of-month levels of the U.S. 10-year Treasury Bond constant maturity yield and Panel (c) monthly dividend yield on the Standard & Poor's 500 Index.

This specification allows us to analyze first-order autocorrelation in inflation growth. Figure 8 and Table 1 show posterior results based on 10,000 draws from the Gibbs sampler[18]. The first column of Table 1 shows that first-order autocorrelation seems to be an important feature of inflation as it cannot be rejected at any reasonable level of credibility. We note again that the posterior is improper but that the Gibbs sampler does not detect this (i.e. it does not reach the absorbing state in our finite set of random drawings) since we are so far away from the boundary. A simple truncation of the region for ρ seems a practical solution in this case. Figure 8(b) confirms that the value of $\rho = 1$ will only occur with extremely low probability.

The second series we consider is the U.S. 10-year Treasury Bond constant maturity

[18]The Gibbs sampler is applied with a burn-in period of 4,000 draws and a thinning value of two.

yield. Data for this series was collected from the FRED database and the sample spans the period January 1960 to July 2007 for a total of 571 monthly observations. The model

Table 1: **Posterior results for inflation, Treasury Bond and dividend yield series**

parameters	**CPI**	**TB**$_U$	**TB**$_R$	**TB**$_{RW}$	**DY**$_U$	**DY**$_R$	**DY**$_{RW}$
c	–	–	–	0.0005	–	–	−0.0002
	–	–	–	[0.0122]	–	–	[0.0006]
μ	0.7706	$\pm\infty$	6.9739	–	$\pm\infty$	0.2480	–
	[0.0675]	–	[1.1956]	–	–	[0.0586]	–
ρ	0.4487	0.9992	0.9895	1	0.9993	0.9893	1
	[0.0994]	[0.0025]	[0.0014]	–	[0.0025]	[0.0019]	–
σ_ε^2	0.1083	0.0856	0.0854	0.0855	0.0177∗	0.0176∗	0.0177∗
	[0.0166]	[0.0051]	[0.0051]	[0.0050]	[0.0012]	[0.0011]	[0.0011]
predictions							
RMSPE - 3m	0.3031	–	0.6922	0.7015	–	0.0213	0.0211
RMSPE - 6m	0.3184	–	0.9569	0.9844	–	0.0307	0.0303
RMSPE - 12m	0.3441	–	1.4099	1.4914	–	0.0454	0.0443

Notes: The table presents posterior means, posterior standard deviations (in between brackets), and the Root Mean Square Prediction Error (RMSPE) for models on U.S. quarterly inflation growth [CPI], U.S. 10-year Treasury Bond constant maturity yield [TB] and Standard & Poor's 500 Index dividend yield [DY]. The subscripts for TB and DY distinguish between using an unconstrained parameter space (TB$_U$, DY$_U$), imposing stationarity with $\rho < 1$ (TB$_R$, DY$_R$) and imposing a random walk with drift (TB$_{RW}$, DY$_{RW}$). A star (*) indicates that the number has been multiplied by a factor 100. The RMSPE results are for 3, 6 and 12 months ahead predictions for the sample 1985:Q1 - 2007:Q2.

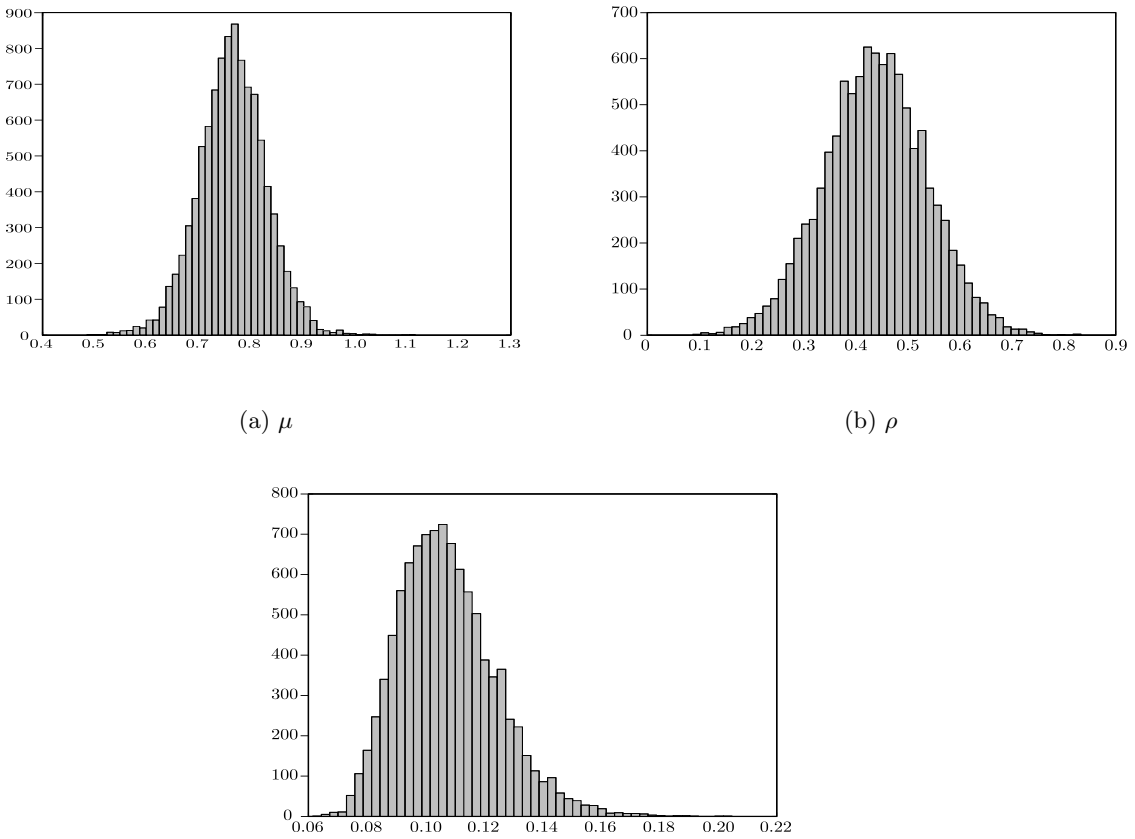
we apply to analyze unit root behavior in the 10-year yield y_t is again the ECM model in (55). Learning about the value of ρ is crucial for several reasons. For example, ρ determines whether long run forecasts will converge to a long-term mean ($\rho < 1$) or whether they will display explosive behavior ($\rho > 1$). We analyze three specifications of the ECM model. In the first application we use a noninformative prior on ρ and we refer to this approach as TB$_U$. We note that with this case of a flat prior on a large region it can be shown, using the results of sections 4.1.1 and 4.1.2, that the posterior of (μ, ρ) is improper. Yet a naive application of the Gibbs sampler will produce results. We emphasize that this is not a limitation/fault of the Gibbs sampler. It is simply the consequence of a poor search of the shape of the posterior and it is simply a wrong application of Gibbs. However, this may occur and has occurred in practice. We use TB$_R$ to indicate the second approach for which we estimate the model under stationarity by truncating the parameter space to $\rho < 1$. Finally, in the third approach we consider model (54) and impose a unit root ($\rho = 1$) and analyze a random walk with drift model, TB$_{RW}$.

Figure 9 and Table 1 show that the improper posterior density of ρ, which has an asymptote at 1, has substantial probability mass close to 1 and even some mass beyond 1. These results suggest that the 10-year Treasury yield may be nonstationary which is difficult to interpret economically, however. The second application therefore uses a truncation prior to restrict ρ to be lower than 1. Panel (b) in Figure 9 shows the posterior density of ρ. Posterior densities for the other model parameters look similar in both cases.

To assess the restrictiveness or importance of restricting ρ to be smaller than 1 we perform a small forecasting exercise[19]. In particular, we make 3, 6 and 12 month ahead forecasts (with the forecast taken to be the mean of the predictive density) starting in

[19]Because of the posterior being improper we do not report forecast results for TB$_U$ and DY$_U$ in Table 1.

Figure 8: **Posterior density histograms for CPI**



(a) μ

(b) ρ

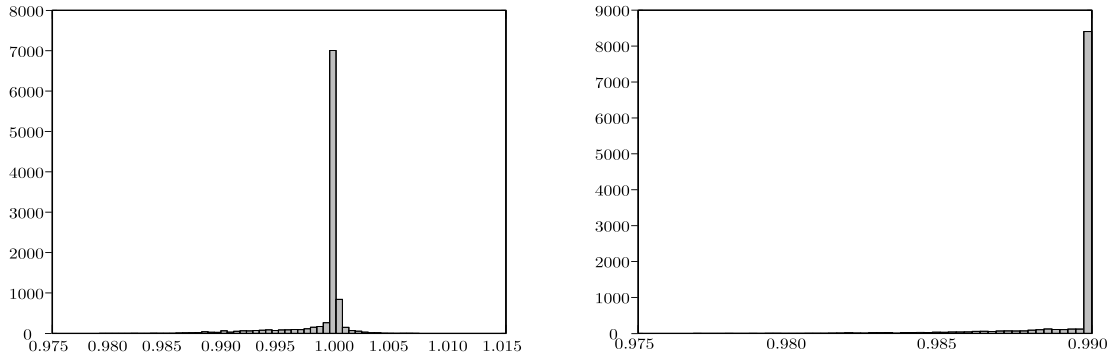
(c) σ_ε^2

Notes: Shown in this figure are posterior density histograms for the model in (63). Panel (a) shows the density for μ, (b) for ρ and (c) for σ_ε^2. The densities are based on 10,000 draws from the Gibbs sampler.

January 1985. We construct forecasts using an expanding window approach. In particular, we use the sample from January 1960 to December 1984 to construct the forecast for January 1985. We then expand the estimation sample to include the realized January 1985 yield value in order to construct the forecast for February 1985 and we keep expanding the sample in the same way until the final forecast, July 2007, which is based on the estimation sample January 1960 to June 2007. The bottom rows of Table 1 report results. We assess forecasting accuracy by means of the Root Mean Squared Prediction Error (RMSPE). The results seem to indicate that imposing the truncation is relevant since the RMSPE is reduced for all horizons.

The third and final series that we examine is the dividend yield on the Standard and Poor's 500 Index which is a commonly used predictor variable for forecasting stock returns, see e.g. Keim and Stambaugh (1986), Campbell and Shiller (1988) and Fama and French (1988). We use monthly data from January 1966 to December 2006 (492 observations). Figure 7(c) shows that the dividend yield has a quite different pattern in the initial 20 years compared to the 1990s and beyond. Whether or not the dividend yield displays unit root-type behavior is a widely discussed topic in the literature, see e.g. Cochrane (2006).

Figure 9: **Posterior density histograms for 10-year Treasury bond**

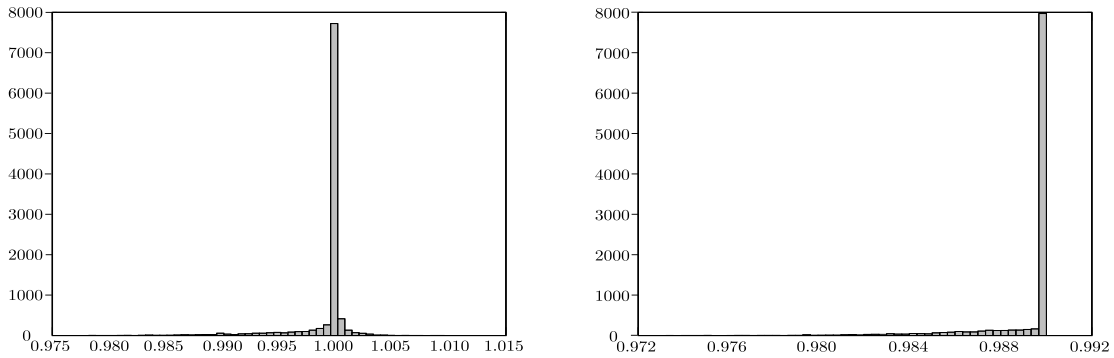


(a) ρ - unrestricted flat prior

(b) ρ - truncation prior

Notes: Shown in this figure are posterior density histograms for the model in (55) applied to the 10-year U.S. Treasury bond yield. Panel (a) shows the density of ρ with an uninformative prior whereas panel (b) shows the density of ρ using a truncation prior to ensure that $\rho < 1$. The densities are based on 10,000 draws from the Gibbs sampler.

Figure 10: **Posterior density histograms for Dividend Yield**



(a) ρ - unrestricted flat prior

(b) ρ - truncation prior

Notes: Shown in this figure are posterior density histograms for the model in (55) applied to the S&P 500 dividend yield. Panel (a) shows the density of ρ with an uninformative prior whereas panel (b) shows the density of ρ using a truncation prior to ensure that $\rho < 1$. The densities are based on 10,000 draws from the Gibbs sampler

As in the previous example we analyze the model with an uninformative prior (indicated by DY_U), a truncation prior (DY_R) and imposing a random walk with drift structure (DY_{RW}). Results are similar to those for the 10-year yield and are shown in Figure 10. A substantial part of the improper posterior of ρ_1 (with asymptote at 1) lies again beyond 1. However, restricting ρ to be lower than 1 now actually slightly worsens forecast accuracy for all horizons. The random walk with drift specification gives the most accurate forecasts albeit that the differences are very small. The results in Table 1 indicate that the data

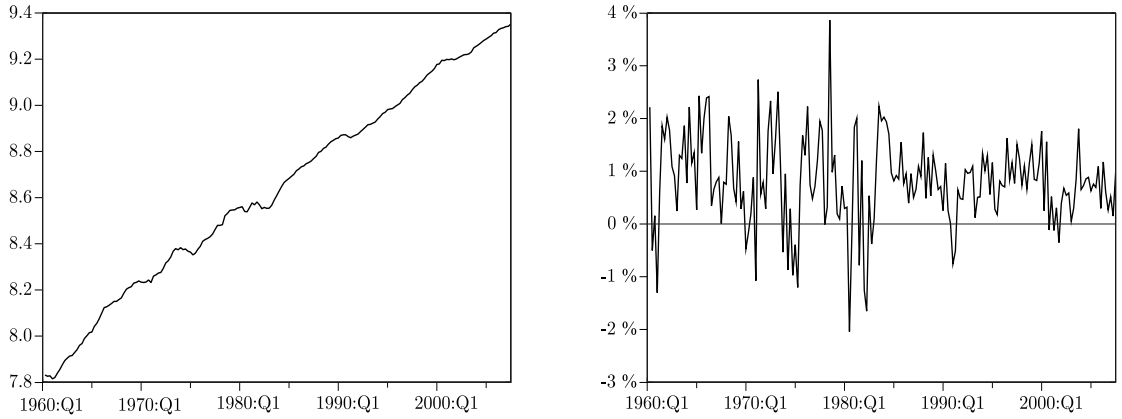
does not provide a clear answer as to whether the dividend yield series is stationary or nonstationary.

As mentioned earlier the results of this section are for illustrative purposes only. A more detailed analysis for the Treasury bond and dividend yield series should involve Bayesian diagnostic checking on the (mis-)appropriateness of using Gibbs sampling results from improper posteriors and using a misspecified model like the autoregressive model with a possible unit root in a forecasting exercise. Further, Figures 7(b) and 7(c) suggest a time varying trend and this data feature should be modelled. However, a detailed Bayesian misspecification and/or empirical analysis for these series is beyond the scope of the present paper and is therefore left to the interested reader.

4.2.2 U.S. real GDP growth

We develop our empirical analysis mainly on one key macroeconomic series: U.S. Gross Domestic Product (GDP) growth. We apply the models from the previous paragraphs (as well as those in Section 5.1.2) to model this series to illustrate boundary issues and to show how to approach these. We collected real GDP (seasonally adjusted) figures from the U.S. Department of Commerce, Bureau of Economic Analysis. Figure 11(a) plots the log quarterly GDP level for our sample 1960:Q1 to 2007:Q2 (190 observations) and shows that GDP

Figure 11: **U.S. GDP**



Notes: Shown in this figure are in Panel (a) quarterly log levels of real U.S. GDP whereas Panel (b) shows the quarterly GDP growth rate (in % terms). The sample is 1960:Q1 - 2007:Q2.

has followed an upward sloping pattern but with fluctuations around this trend. The quarterly growth rate of log GDP, $g_t = \ln GDP_t - \ln GDP_{t-1}$, shown in Figure 11(b), underlines these fluctuations with periods of positive changes followed by periods of negative changes, clearly indicating business cycles. The sample average growth rate is positive but with a high level of variation which is mainly due to different cycles (for more details we refer to Harvey *et al.*, 2007). We apply the various linear models we discussed earlier to model and assess empirical facts on stochastic shocks over the full sample so as to assess these models' suitability in a out-of-sample forecasting exercise. In the forecast exercise we use an initial in-sample period from 1960:Q1 to 1992:Q2 to obtain initial parameter estimates and we forecast the GDP growth figure for 1992:Q3. We then expand the estimation sample with

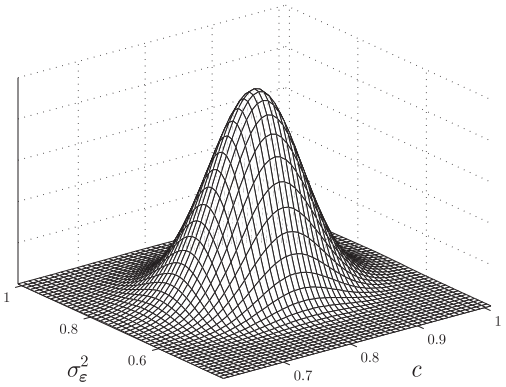
the value in 1992:Q3, re-estimating the parameters, and we forecast the next value for 1992:Q4. We continue this procedure up to the last value and we end up with a total of 60 forecasts. Also here we remark that this section is not meant as a full-fledged empirical analysis of U.S. GDP. We mainly want to analyze and compare the various linear models for illustrative purposes. It is also for this reason that we ignore the 'great moderation' that occurred in fluctuations of real GDP since 1985 in our analysis, see Stock and Watson (2002).

Random walk with drift for GDP growth

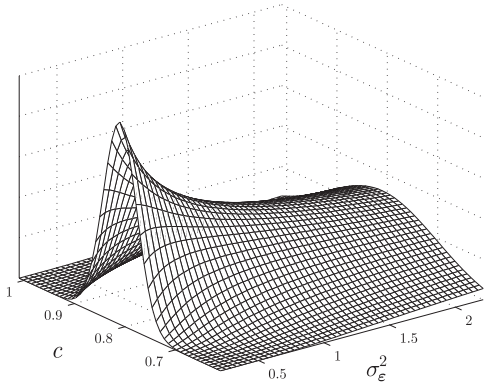
The first model that we put forth to describe U.S. GDP growth is the univariate constant model in equation (10):

$$g_t = c + \varepsilon_t \qquad \text{with} \qquad \varepsilon_t \sim \text{i.i.d. } \mathcal{N}(0, \sigma_\varepsilon^2) \tag{64}$$

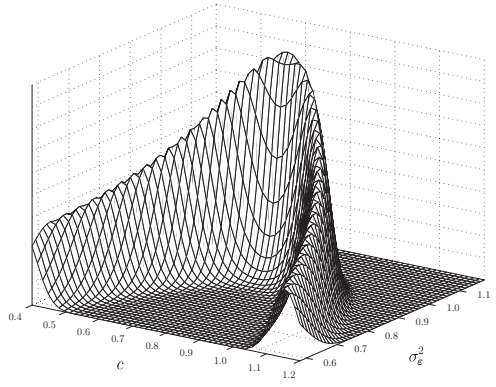
Figure 12: **Joint and conditional posterior densities**



(a) $p(c, \sigma_\varepsilon^2|D)$



(b) $p(c|\sigma_\varepsilon^2, D)$



(c) $p(\sigma_\varepsilon^2|c, D)$

Notes: Panel (a) shows the joint posterior of c and σ_ε^2 whereas panels (b) and (c) show the conditional posterior density of c for given values of σ_ε^2 and the data D; $p(c|\sigma_\varepsilon^2, D)$ and the conditional density of σ_ε^2 for given values of c and the data; $p(\sigma_\varepsilon^2|c, D)$ respectively. We apply the linear regression model (64) to quarterly U.S. GDP growth.

We refer to this model by the label 'random walk with drift' in tables and graphs. This model allows us to infer the average growth rate of GDP although with the drawback

that other fluctuations cannot be explained. Section 3 provides details on how to compute posterior densities and predictive densities using the Gibbs sampler. We note that in this case direct sampling is also a good strategy for efficient computation.

Figure 12 shows the posterior densities $p(c|\sigma_\varepsilon^2, D)$ and $p(\sigma_\varepsilon^2|c, D)$ and clearly reveals the conditional Normal and conditional Inverted Gamma densities. Also shown is the joint posterior density.

Table 2 and Figure 13, which summarize posterior results, show that the growth rate is on average positive, well distributed around the mean value of 3.25% (in annual terms) with a small variance. However, the model explains relatively little as the residual variance is quite close to the unconditional variance of GDP growth which equals 0.71%. In addition, the posterior standard deviation of σ_ε^2 is quite substantial as well.

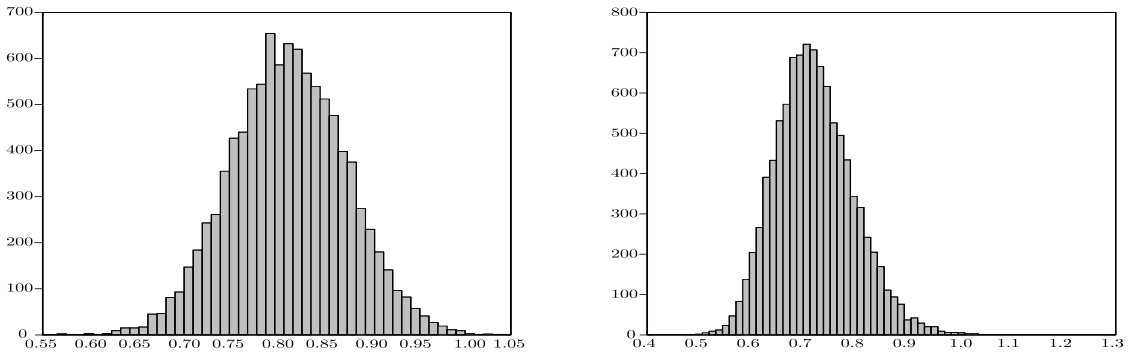
Table 2: **Posterior results for U.S. GDP**

	growth			**levels**
parameters	**RW+drift**	**simple ECM**	**ECM**	**trend**
μ	-	3.2516$*$	16.8940	–
	-	[0.3308]$*$	[3.8574]	–
c	3.2500*	–	–	1.5417$*$
	[0.2460]*	–	–	[0.5512]$*$
δ	-	–	−0.0087	–
	-	–	[0.0157]	–
β	-	–	–	0.0004
	-	–	–	[0.0001]
ρ_1	-	0.2719	1.1624	1.2246
	-	[0.0715]	[0.0901]	[0.0706]
ρ_2	-	–	−0.2596	−0.2727
	-	–	[0.0715]	[0.0701]
σ_ε^2	0.7214	0.6750	0.6711	0.6490$**$
	[0.0749]	[0.0712]	[0.0704]	[0.0682]

Notes: The table presents posterior means and posterior standard deviations (in between brackets) for models for log real U.S. GDP level/growth. The considered models are (i) the random walk with drift model (for GDP *growth*), (ii) the simple ECM model in (55) (for GDP *growth*), (iii) the ECM model in (60) (for GDP *growth*) and (iv) the trend model in (57) (for GDP *level*). Note that the interpretation of the parameter μ is different in (55) and (60). In the simple ECM μ represents the unconditional mean of the series whereas for the ECM in (60) it represents the constant of the linear trend line. The difference in interpretation also applies to c. For the random walk with drift model it represents the drift whereas for the trend model it is the constant of the trend line. One star (*) indicates that the number has been annualized (multiplied by a factor of 4), two stars (**) is for a multiplication factor of 10^4.

Moving to Table 3 which contains the MSPE results of our forecast exercise, we do find that the model has higher forecasting power than the benchmark random walk (RW) model. The RMSPE of the constant mean model is lower for all horizons compared to the random walk.

Figure 13: **GDP: random walk with drift**



(a) c (b) σ_ε^2

Notes: Shown in this figure are posterior density histograms for the model in (64). The left hand side panel shows the density for c whereas the right hand side panel is for σ_ε^2. The densities are based on 10,000 draws from the Gibbs sampler. Note that the left hand side panel shows the density of *quarterly* growth whereas the results in Table 2 are stated in terms of *annualized* growth.

Table 3: **Forecast accuracy results for U.S. GDP growth**

	RW [1]	**RW+drift** [2]	**simple ECM** [3]	**trend** [4]	**ECM** [5]	**TVP** [6]	**BMA** [2],[3]	**BMA** [4],[5],[6]	**BMA** [2]-[6]
RMSPE - 1Q	0.6198	0.4738	0.4761	0.4713	0.4724	0.4726	0.4700	0.4802	0.4865
RMSPE - 2Q	0.5462	0.5299	0.5178	0.5050	0.5220	0.5318	0.4742	0.4991	0.4991
RMSPE - 4Q	0.6118	0.5257	0.5251	0.5054	0.5202	0.5223	0.4888	0.5131	0.5474

Notes: The table presents Root Mean Squared Prediction Error (RMSPE) results for U.S. GDP growth for various models. The models used are [1] the random walk model, [2] the random walk with drift model, [3] the ECM model in (55), [4] the model with additive trend (57), [5] the error correction model in (60), [6] the time-varying parameter model (91)-(92) and three Bayesian model averaging (BMA) schemes: (i) averaging over the random walk with drift and simple ECM model, (ii) averaging over the trend, ECM and time-varying parameter model and (iii) averaging over all models. The out-of-sample period is 1992:Q3-2007:Q2 (60 forecasts) and an expanding window approach is used for each model. For the BMA models we use the initial 30 forecasts from 1985:Q1-1992:Q2 as training period.

Simple error correction model for GDP growth

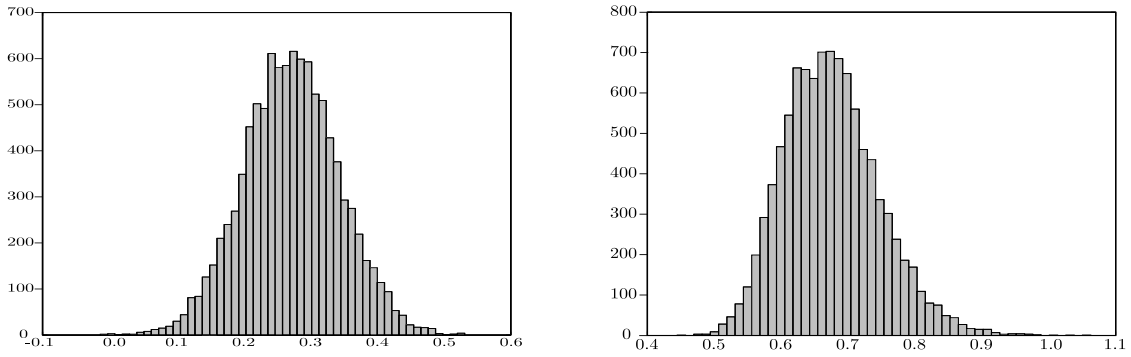
The second specification that we propose is the error correction model in (55). Allowing for autocorrelation in GDP growth may be beneficial. The model is given by

$$g_t - \mu = \rho(g_{t-1} - \mu) + \varepsilon_t, \qquad \text{with} \qquad \varepsilon_t \sim \text{i.i.d}\, \mathcal{N}(0, \sigma_\varepsilon^2) \tag{65}$$

and we refer to it as 'simple ECM'. Section 4 gives estimation details for this model. Posterior results are shown in the second column of Table 2. It can be seen for μ that whereas its posterior mean is again positive and very close to the posterior mean of c in the previous model, the explanatory power of the model increases. The posterior mean of ρ is 0.27 and Figure 14 shows that nearly all probability mass is to the right of 0. The latter observation provides evidence of the presence of autocorrelation in growth. Here we see again an example where the posterior is improper. However, the data keep the parameter

ρ away from 1. We use as potential solution a truncated density where the truncation is a binding constraint in practice. Finally, the density of σ_ε^2 has both a lower posterior mean and posterior variance which suggests that this model explains more variation in GDP growth than the random walk with drift model. Similarly, allowing for first-order autocorrelation in growth improves forecast accuracy as shown in Table 3 although the improvement is only marginal (and only for longer horizons). A possible explanation is that both models have a nearly identical posterior mean for μ.

Figure 14: **GDP: simple ECM for growth**



(a) ρ (b) σ_ε^2

Notes: Shown in this figure are posterior density histograms for the model in (65). Panel (a) shows the density for ρ and (b) for σ_ε^2. The densities are based on 10,000 draws from the Gibbs sampler.

Autoregressive model with additive linear trend for GDP level

In addition to using models for GDP *growth* we also analyze models that draw inference on the *level* of GDP. In particular we consider an autoregressive distributed lag model for GDP level. Within these types of models one attempts to explain the level of GDP by using the information in lagged GDP levels and a linearly increasing variable. The model we consider here is that of (57) and is given as:

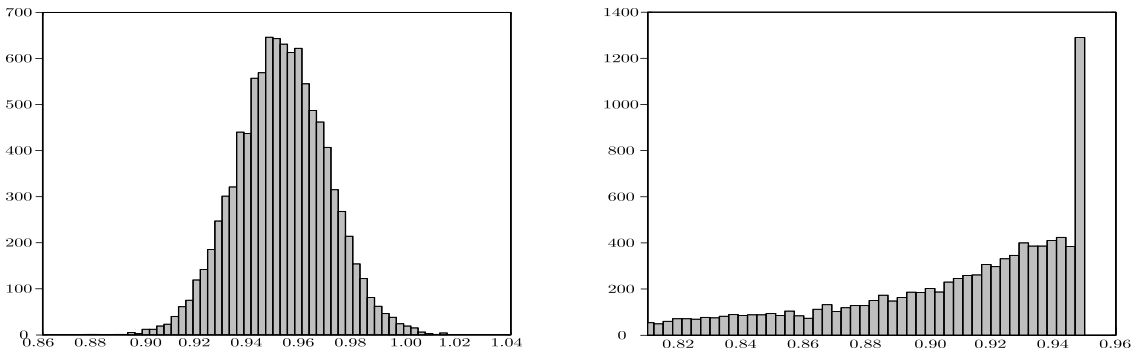
$$\ln \text{GDP}_t = c + \beta t + \rho_1 \ln \text{GDP}_{t-1} + \rho_2 \ln \text{GDP}_{t-2} + \varepsilon_t \qquad \text{with} \quad \varepsilon_t \sim \text{i.i.d.}\, \mathcal{N}(0, \sigma_\varepsilon^2) \quad (66)$$

which we refer to as 'trend'. The results in Section 3 can be used to draw inference on the parameters in this model.

Table 2 shows that the constant (or long term growth rate) is positive. However, its value is substantially lower than when we model the growth rate directly and cannot be directly interpreted as being the average growth rate. Part of the long term growth rate is now being explained by the trend (long term trend) which, as we may expect, has a positive posterior mean and a small standard deviation. The behavior of the lag parameters of past GDP levels is also interesting to examine. The first lag has a distribution which lies almost entirely beyond 1, implying compounding GDP growth. However, the parameter of the second lag is always negative, which captures the cycles in GDP. The residual variance is very small compared to the previous two models.

Panel (a) of Figure 15 shows the posterior density histogram of the persistence in $\ln GDP$, as measured by $\rho_1 + \rho_2$. We note that we do not truncate the parameter space

Figure 15: **GDP: autoregressive model with linear trend and ECM**



(a) $\rho_1 + \rho_2$ (b) $\rho_1 + \rho_2$

Notes: Shown in this figure is the posterior density histograms of the persistence parameter $\rho_1 + \rho_2$ for the trend model (66) in Panel (a) and the error correction model (60) in Panel (b). The densities are based on 10,000 draws from the Gibbs sampler.

of ρ_1 and/or ρ_2 here. We emphasize that even with a flat prior on a large region, the posteriors of the parameters in (63) are proper. Whereas this unrestricted model does well in terms of in-sample fit, the fact that there is some probability mass for $\rho_1 + \rho_2 > 1$ (as shown in Figure 15) implies that the model can potentially display explosive behavior when forecasting future values. In fact, Table 3 shows that the forecast accuracy of this model is high, but not superior to that of the random walk with drift model.

Error correction model for GDP level

As a final model for GDP levels we analyze the error correction model given in (60). Panel (b) in Figure 15 shows the persistence in GDP. Here we do truncate the parameter space such that $\rho_1 + \rho_2 < 0.95$. Not doing so results in ρ_1 and ρ_2 often adding up to values close to 1 in the Gibbs sampling algorithm which makes inference on the remaining parameters μ and δ difficult. The ECM with truncation provides accurate forecasts, even though these are marginally worse than the forecasts of the simple ECM and the trend models.

Because the forecast performance of all models is very similar it is unclear whether one should decide to model the level of GDP or its growth rate directly to forecast GDP growth. In Section 6 we propose a simple procedure, that of model averaging, to alleviate this decision problem. As we show in Table 3, the model averaging of a random walk with drift and a simple error correction model gives the most accurate forecasts.

4.3 Remark on equivalence between boundary problems in dynamic regression models and instrumental variable models

The final class of models that we discuss in the current section is the class of multivariate models. The issues involved here are similar to those surrounding univariate unit root models, i.e. nonidentifiability of parameters. This results in the Information matrix being singular, or alternatively, in the Hessian having a reduced rank. This reduced rank problem

can occur in several well-known models, such as for example cointegration models, Vector Autoregressive (VAR) and Simultaneous Equation Models (SEM) which in turn are closely linked to Instrumental Variables (IV) models.

To show which role nonidentifiability plays in these models we give an example by means of a just identified IV model and in particular we focus on the Incomplete Simultaneous Equation Model (INSEM). Our analysis, which is necessarily brief, is based on Van Dijk (2003) and Hoogerheide *et al.* (2007) and we refer to these studies for a more in-depth analysis. Consider the INSEM model as it is specified in Zellner *et al.* (1988)[20]

$$y = x\beta + \varepsilon \tag{67}$$

$$x = z\pi + \nu \tag{68}$$

$$[\varepsilon \ \nu]' \sim \mathcal{N}\left([0 \ 0]', \Sigma\right) \qquad \text{with} \qquad \Sigma = \begin{bmatrix} \sigma_\varepsilon^2 & \sigma_{\varepsilon,\nu} \\ \sigma_{\varepsilon,\nu} & \sigma_\nu^2 \end{bmatrix} \tag{69}$$

with y, x and z all having dimensions $(T \times 1)$ and β and π being scalar parameters. In this model, (67) is to be interpreted as the structural relation of interest, x is an endogenous variable and z is the (weakly exogenous) instrument. Similarly, β is the structural parameter of interest and $|\pi|$ measures the quality of the instrument. Furthermore, the correlation parameter $\rho = \frac{\sigma_{\varepsilon,\nu}}{\sqrt{\sigma_\varepsilon^2 \sigma_\nu^2}}$ measures the degree of endogeneity of x in the equation for y. Equations (67)-(69) comprise the *structural form* of the INSEM. By substituting (68) in (67) we can derive the *reduced form* which is given by

$$y = z\pi\beta + \xi \tag{70}$$

$$x = z\pi + \nu \tag{71}$$

with $\xi = \varepsilon + \nu\beta$ and $[\varepsilon \ \nu]'$ as in (69). We can interpret the reduced form model as a multivariate regression model which is nonlinear in the parameters β and ρ as in (44). As was the case in the unit root model this nonlinearity can lead to a nonidentifiability problem. In particular, when we assume a flat prior on the parameters β and π, then when $\pi = 0$ (this is usually labeled as the case of no identification or case of irrelevant instruments) the joint posterior density is improper because it is flat and nonzero in the direction of β. In fact, the joint density looks very similar to that in Figure 6(a) in the sense that it has a wall at $\pi = 0$. Therefore, β is not identified when $\pi = 0$ whereas it is for any $\pi \neq 0$. In a multivariate setting where y, x and z are all matrices and β and π are matrices as well, the identification problem of (part of the elements) of β occurs when $\pi = 0$ or when π is of reduced rank. The above problem is known as *local nonidentification* and is discussed in detail in Kleibergen and Van Dijk (1998).

As a result of the local nonidentification problem, the *marginal* density of π is non-integrable because of infinite probability mass near $\pi = 0$ (see Kleibergen and Van Dijk, 1998). Whether or not the impropriety of the joint density will be revealed in the output from the Gibbs sampler is unclear. Slow convergence of the Gibbs sampler due to the fact that $\pi = 0$ is acting as an absorbing state could be an indication. Examples of bimodal posterior densities on bounded intervals are given in Hoogerheide *et al.* (2007). A possible solution to circumvent the local nonidentification problem in INSEM models would again be the specification of sensible prior densities. However, it can be an arduous task to find conjugate priors, mainly since these will have to curtail multiple parameters all at the same time.

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

5 Variance Components and State Space Models

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

5.1 Heteroscedasticity and Hierarchical Linear Mixed Models

Before we specify the basic setup of the HLMM we first discuss two preliminary models, focusing on variances of disturbances. The models serve to identify the following two issues: (i) a *degrees of freedom bound* which refers to a sufficient number of observations or a sufficient number of cross-sectional groups, and (ii) an *identification* issue or *labeling* issue with respect to the different variance components.

5.1.1 Linear regression model with a small number of observations or naive heteroscedasticity

T = small, degrees of freedom bound

In Section 3 we analyzed the basic linear regression model. Now we revisit this model which we simplify using $x_t = 1$ and $\beta = \mu$

$$y_t = \mu + \varepsilon_t, \qquad t = 1, ..., T, \qquad \text{with} \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2) \tag{72}$$

We emphasize that the number of observations T may refer to the number of observation in a time-series and to individuals or groups of individuals in a cross-section. For notational convenience, we use the same symbol T here for time series and cross-section observations. If we use a *uniform* prior on *both* μ and σ_ε^2,

$$p(\mu, \sigma_\varepsilon^2) \propto 1 \tag{73}$$

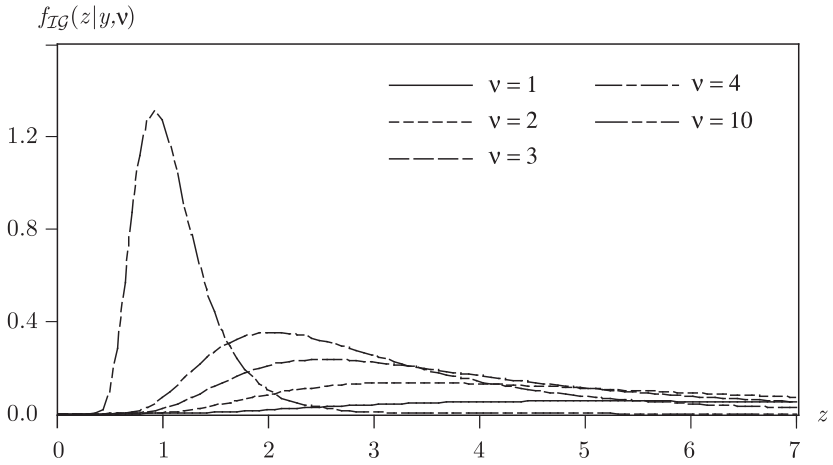
then, from the results of Table B-1 in Appendix B we can derive the marginal densities of μ and σ_ε^2. Note, however, that because the prior is now that of (73) instead of (13), we have two degrees of freedom less. The marginal densities are given by

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

with y the full vector of observations; $y = [y_1 \ \dots \ y_T]'$, $\hat{\mu} = \frac{1}{T}\iota_T' y$ and $s^2 = y' M_{\iota_T} y$. The degrees of freedom for the general linear regression model with the prior in (73) is $T-K-2$. For the model in (72) we have $K = 1$. From the parameters of the marginal densities and

the conditions given in Appendix C it is clear that in order for these Student-t and Inverted Gamma densities to exist one needs more than 3 observations, i.e. $T > 3$. Analogous results can be derived for the existence of higher order moments. For illustration, Figure 16 shows that the right tail of an Inverted Gamma density tends to zero at a rate that is too small when the number of degrees of freedom is too small. For instance, for $T = 4$ the zeroth moment exists but the higher moments do not[21].

Figure 16: **Inverted Gamma density**



Notes: The graph shows the Inverted Gamma density function, as specified in (C-3), for $m = 10$ and for a varying number of degrees of freedom, ν.

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

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

Only focusing on these *conditional* densities shows that $T = 3$ is a already sufficient for the Gibbs sampler to function properly. However, it follows from our analysis that the *marginal* densities for μ and σ_ε^2 do not exist. Thus, we have a simple case where the Gibbs sampler can be applied as a simulation method, whereas the joint and marginal densities do not exist (similar as the ridge example in Section 2.2) , see also the discussion in for example Koop (2003). Therefore, the generated Gibbs sequences cannot be interpreted. We emphasize that for the usual number of time series observations this degrees of freedom restriction is obviously of no significance. However, for the case of the number of groups in a panel it may become restrictive. In Section 5.5 we give an example using a panel data model.

Naive heteroscedasticity and a degrees of freedom bound

Consider model (72) in which each observation y_t, for $t = 1, \ldots, T$, is allowed to have its own variance parameter σ_t^2. When we use a uniform prior on μ as well as on each of the

[21] A Jeffreys' type 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 $T > 1$.

σ_t^2 components then the posterior of $(\mu, \sigma_1^2, \ldots, \sigma_T^2)$ is given as:

$$p(\mu, \sigma_1^2, \ldots, \sigma_T^2|y) \propto \prod_{t=1}^{T} \left(\sigma_t^2\right)^{-\frac{1}{2}} \exp\left[-\frac{1}{2\sigma_t^2}(y_t - \mu)^2\right] \tag{74}$$

The posterior is unbounded since for some t, one may have $y_t = \mu$ and/or $\sigma_t^2 = 0$, then $p(\mu, \sigma_1^2, \ldots, \sigma_T^2|y) \to \infty$. One solution would be to partition the observations into groups, assuming that per group the variance is constant whereas it is allowed to be different across groups. Each partition needs to be chosen in such a way that it contains a sufficient number of observations. Our main point, although trivial as it may seem, is that the degrees of freedom restriction implies that one needs multiple observations to draw inference on variance components. This becomes particularly relevant in dynamic panels with groups of observations, see Section 5.5. We note that Geweke (1993) uses a weakly informative Inverted Gamma density which makes the posterior more regular. The linear regression model with naive heteroscedasticity serves as introduction to the class of Hierarchical Linear Mixed Models, since the latter class is not so well-known in economics.

5.1.2 Hierarchical Linear Mixed Model

An example of a canonical model with multiple variance components is the class of Hierarchical Linear Mixed Models (HLMM). Instead of T variances, one has two variance components with the additional issue of two stochastic processes. We introduce this class through the following hierarchical model with two variance components

$$y_t = \mu_t + \varepsilon_t, \qquad \text{with} \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2) \qquad \text{for} \qquad t = 1, \ldots, T \tag{75}$$

$$\mu_t = \theta + \eta_t, \qquad \text{with} \qquad \eta_t \sim \mathcal{N}(0, \sigma_\eta^2) \qquad \text{and} \qquad E[\varepsilon_t \eta_s] = 0 \tag{76}$$

with $\mu = (\mu_1, \ldots, \mu_T)'$ a vector containing the time-varying mean of y and θ the mean of the distribution of μ_t which, for any t, is Normal with variance σ_η^2.

This model serves as a parent model for more elaborate models such as state space models or panel data models. Before moving on to introducing and discussing these models, we analyze the base model by distinguishing between two cases. The first case assumes that σ_ε^2 is known with T being small whereas the second case is the opposite: σ_ε^2 is assumed unknown and T is large. Each case helps to gain a better understanding of the existence conditions for joint, conditional and marginal posterior distributions of the HLMM class of models. Note that unless stated otherwise, we assume a uniform prior for each of the variance components.

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

Because σ_ε^2 is given and equal to 1, the only unknown variance component is σ_η^2. The requirement on a minimum number of degrees of freedom as discussed in the previous paragraph is of importance here. Sensible posterior results can only be obtained when there is a sufficient number of observations. The conditional densities $p(\theta|\sigma_\eta^2)$ and $p(\sigma_\eta^2|\theta)$ can be derived by *substituting* (76) in (75)

$$y_t = \theta + \varepsilon_t + \eta_t \qquad \text{with} \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2),\ \eta_t \sim \mathcal{N}(0, \sigma_\eta^2) \tag{77}$$

The joint posterior density is

$$p(\theta, \sigma_\eta^2|\sigma_\varepsilon^2 = 1, y) \propto (\sigma_\eta^2 + 1)^{-\frac{1}{2}(T-1)} \exp\left[-\frac{1}{2(\sigma_\eta^2 + 1)}(y - \iota_T\theta)'(y - \iota_T\theta)\right] \tag{78}$$

By using the transformation of random variables: $\sigma_\eta^{2*} \equiv \sigma_\eta^2 + 1$, one is back in the situation of the beginning of Section 5.1. A degrees of freedom bound is necessary for obtaining existence of posterior moments. Gibbs sampling is possible but not sensible when the number of observations (or the number of groups) is less than or equal to 3. This is essentially the same result as is more formally derived by Hobert and Casella (1996). We refer the interested reader to that reference for details but we emphasize here that substitution of (76) in (75) yields a simplified derivation of the Hobert and Casella result on the degrees of freedom restriction. As before, the Gibbs sampler may seem to 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[22].

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

By taking T large enough, one no longer needs to worry about the marginal posterior densities possibly being nonexistent. However, making the first variance component, σ_ε^2, unknown as well introduces a new issue. More specifically, one now has to deal with an identification or labeling issue in the sense that it is not possible to distinguish the two variance components from each other. Why this is the case can be made clear as follows. Note first that since T is assumed to be large enough, the marginal densities of σ_ε^2 and σ_η^2 will exist. However, respecifying the model in (77) to

$$y = \iota_T \theta + \varepsilon + \eta \qquad \text{with} \qquad \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2 \mathrm{I}_T) \qquad \text{and} \qquad \eta \sim \mathcal{N}(0, \sigma_\eta^2 \mathrm{I}_T) \tag{79}$$

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

$$p(\sigma_\eta^2, \sigma_\varepsilon^2 | y) \propto (\sigma_\eta^2 + \sigma_\varepsilon^2)^{-\frac{1}{2}(T-1)} \exp\left(-\frac{1}{2} \frac{(y - \iota_T \hat{\theta})'(y - \iota_T \hat{\theta})}{\sigma_\eta^2 + \sigma_\varepsilon^2} \right) \tag{80}$$

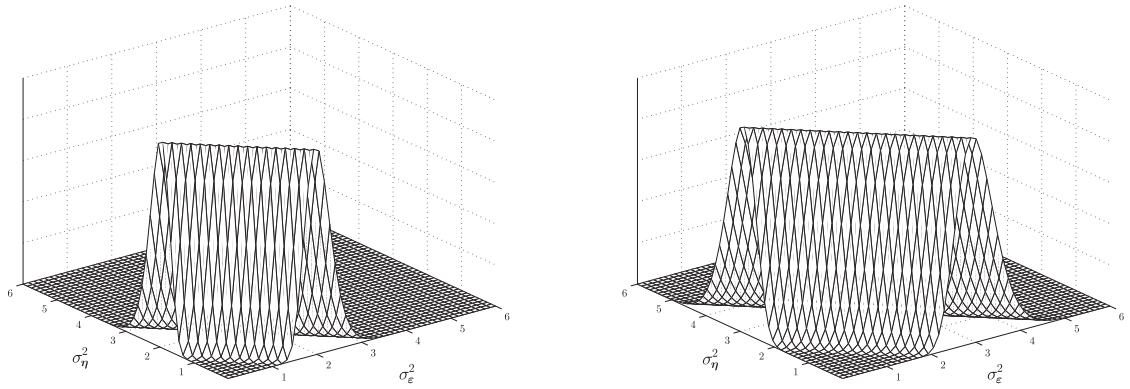
with $\hat{\theta} = \bar{y}$ being the average of y. Clearly, only the total variance is identified, *not* the individual components. Furthermore, the roles of σ_ε^2 and σ_η^2 are interchangeable (this explains the use of the term *labeling* issue). This holds true for any value of the *signal-to-noise* ratio which is defined as $\mathrm{SN} = \sigma_\eta^2 / \sigma_\varepsilon^2$. Figure 17 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 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 (75)-(76) 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[23]. The Gibbs sampler can therefore be used to obtain posterior results.

[22] Note that Hobert and Casella (1996) assume an independent Jeffreys' prior as a result of which the Inverted Gamma density of σ_η^2 has one degree of freedom since their sample consists of $T = 2$ groups/observations.

[23] The density shown in Figure 3(e) on the other hand has a ridge on the domain $[0, \infty) \times [0, \infty)$ which makes it nonintegrable on this domain. For expository purposes we restricted the domain in this figure and therefore for the figure the posterior is a proper density.

Figure 17: **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 (80) with a signal-to-noise (SN) ratio of 1 and 0.5 respectively. For both panels y was simulated from (75)-(76) 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.

The joint posterior density of $(\sigma_\varepsilon^2, \sigma_\eta^2)$ is given in (80). The conditional density of σ_ε^2, given some draw or value for σ_η^2 is given by

$$p(\sigma_\varepsilon^2|\sigma_\eta^2 = \bar{\sigma}_\eta^2, y) \propto (\bar{\sigma}_\eta^2 + \sigma_\varepsilon^2)^{-\frac{1}{2}(T-1)} \exp\left[-\frac{1}{2}\frac{(y - \iota\hat{\theta})'(y - \iota\hat{\theta})}{\bar{\sigma}_\eta^2 + \sigma_\varepsilon^2}\right] \tag{81}$$

and in a similar way one obtains $p(\sigma_\eta^2|\sigma_\varepsilon^2 = \bar{\sigma}_\varepsilon^2, y)$. Thus we are back in the situation of Section 5.1.1. However, now these conditional posterior densities are identical and the role of the two variance components is interchangeable. The dynamic processes in (75) and (76) have an identical structure. The result is an identification or labeling issue since it is possible to relabel σ_ε^2 as σ_η^2 and everything remains the same.

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

$$p(\sigma_\eta^2, \sigma_\varepsilon^2|y) \propto \frac{1}{\sigma_\varepsilon^2}\frac{1}{\sigma_\eta^2}\left(\frac{1}{\sigma_\eta^2 + \sigma_\varepsilon^2}\right)^{\frac{1}{2}(T-1)} \exp\left(-\frac{1}{2}\frac{(y - \iota_T\hat{\theta})'(y - \iota_T\hat{\theta})}{\sigma_\eta^2 + \sigma_\varepsilon^2}\right) \tag{82}$$

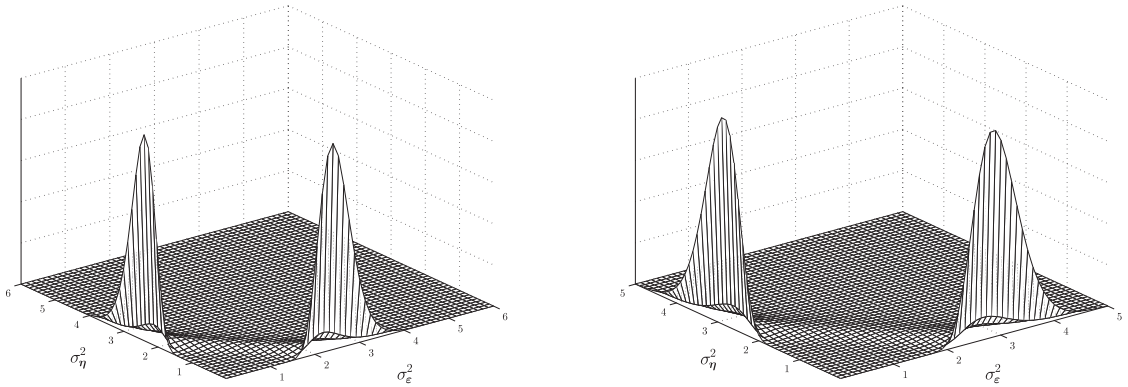
Figure 18 shows that the Jeffreys'-type prior causes the joint density to shoot off to infinity for either $\sigma_\varepsilon^2 \to 0$ or $\sigma_\eta^2 \to 0$ [24]. Therefore, the joint posterior is now improper and the Gibbs sampler will not converge[25]. The Jeffreys' Information matrix prior in this case is given as

$$p(\sigma_\eta^2, \sigma_\varepsilon^2) \propto \sqrt{c(\sigma_\eta^2, \sigma_\varepsilon^2, T)}\left|\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}\right|^{\frac{1}{2}} \tag{83}$$

[24] Although Figure 18 is similar in shape as Figure 4 the two figures have a very different interpretation. Whereas Figure 4 shows a density that has two well-defined modes (albeit far apart) the density in Figure 18 is only well behaved in the domain $(\delta, \infty) \times (\delta, \infty)$ for a δ that is sufficiently far away from zero. The latter density goes to infinity when either of the variance components tends to zero.

[25] For an early Bayesian paper in this field we refer to Griffiths *et al.* (1979).

Figure 18: **Joint posterior density of σ_ε^2 and σ_η^2 with a Jeffreys'-type prior**



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

Notes: Panels (a) and (b) show the joint density in (82) with a signal-to-noise (SN) ratio of 1 and 0.5 respectively. For both panels y was simulated from (75)-(76) 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.

where $c(\sigma_\eta^2, \sigma_\varepsilon^2, T) = \frac{1}{2}(T+3)(\sigma_\eta^2 + \sigma_\varepsilon^2)^{-2}$. Clearly the determinant in the right hand side of (83) is zero. This is in a certain sense obvious since the gradients of the log likelihood are the same for σ_ε^2 and σ_η^2 and the ridge in Figure 17 shows the constancy of the likelihood function. We note that in our evaluation of the likelihood we marginalized in an analytical way (by substituting the state equation (76) into the measurement equation (75)) with respect to the unobserved component μ_t and thus we make use of the EM method to evaluate the Information matrix. In Hobert and Casella (1996), Theorem 1, a number of conditions are stated that ensure propriety of the posterior density in HLMM models. Our derivation is a simplified version which we achieve by substituting the state equation into the measurement equation.

Summarizing, our analysis indicates the following. A uniform prior yields a proper posterior on a bounded region for $(\sigma_\eta^2, \sigma_\varepsilon^2)$, compare also Gelman (2006). However, there exists an identification or labelling issue for σ_η^2 and σ_ε^2, see Frühwirth-Schnatter (2001). A weakly informative prior like an (independent) Jeffreys'-type prior is not appropriate and leads to improper posteriors. A natural conjugate informative prior has the possible disadvantage of cutting away posterior probability mass observed near zero.

Solutions: structural time series or cross sectional information

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

Another possibility is to extend the basic HLMM in such a way that one can distinguish σ_ε^2 from σ_η^2. Two possible directions can be taken here. The first direction is to change the dynamics of μ by changing the specification of the model in (75)-(76) to that of a state space model. The variance components can then be identified from the imposed additional model structure. The second direction is to use a second source of information. Including additional information via more dependent variables in a panel data model enables one to identify σ_η^2 from the cross-sectional observations. We discuss both types of models in the following sections.

5.2 State space model: a random walk for μ_t

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 process for the state variable μ_t gives

$$\begin{aligned} y_t &= \mu_t + \varepsilon_t, \qquad \text{with} \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2) \qquad \text{and} \qquad t = 1, \ldots, T \qquad (84) \\ \mu_t &= \mu_{t-1} + \eta_t, \qquad \text{with} \qquad \eta_t \sim \mathcal{N}(0, \sigma_\eta^2) \qquad \text{and} \qquad E[\varepsilon_t \eta_s] = 0 \qquad (85) \end{aligned}$$

This model, which is generally known as the *local level* model or *time-varying parameter* model, see Harvey (1989), is a basic specification of a state-space model and has been studied extensively in the literature, see e.g. Koop and Van Dijk (2000).

More elaborate state space models are easily obtained by including explanatory variables in the measurement equation (84) and state equation (85), 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 given 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), Frühwirth-Schnatter (1994), De Jong and Shephard (1995) and Durbin and Koopman (2001).

The specification in (85) implies that μ_t is a random walk process which follows from recursively substituting μ_{t-1}, μ_{t-2} etc. Due to the additional structure of the state space model one can now distinguish σ_ε^2 from σ_η^2 and therefore identify both variance components.

We explain the Gibbs sampler by solving (85) in a recursive way which yields

$$y_t = \mu_0 + \sum_{i=1}^{t} \eta_i + \varepsilon_t \qquad (86)$$

For simplicity, we assume that the initial value μ_0 equals zero so that we can obtain the posterior density of $(\sigma_\eta^2, \sigma_\varepsilon^2)$ as

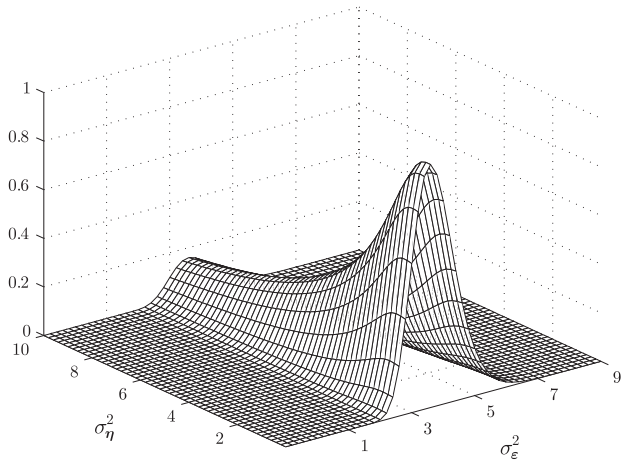
$$p(\sigma_\eta^2, \sigma_\varepsilon^2 | y, \mu_0 = 0) \propto |\sigma_\eta^2 V + \sigma_\varepsilon^2 \mathrm{I}_T|^{-\frac{1}{2}} \exp\left[-\frac{1}{2} y'(\sigma_\eta^2 V + \sigma_\varepsilon^2 \mathrm{I}_T)^{-1} y\right] \qquad (87)$$

where $V = CC'$ and C is the so-called random walk generating matrix which is defined as

$$C = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 1 & 1 & & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 1 & 1 & \ldots & 1 \end{bmatrix}$$

Figure 19 depicts the joint posterior density of σ_ε^2 and σ_η^2 for 100 data points generated for the case where the "true" σ_ε^2 is chosen as 4 and the "true" σ_η^2 is chosen as 1.

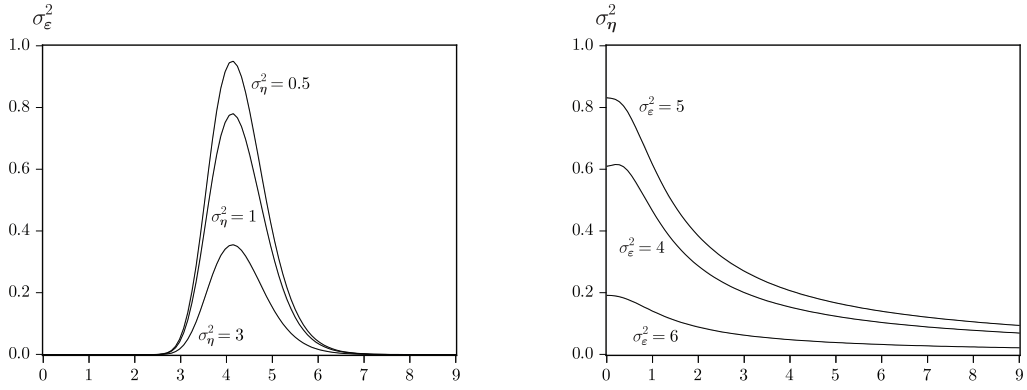
Figure 19: **Joint posterior density of σ_ε^2 and σ_η^2**



Notes: The graph shows the joint density (87) for 100 data points generated for the case where the "true" σ_ε^2 is chosen as 4 and the "true" σ_η^2 is chosen as 1.

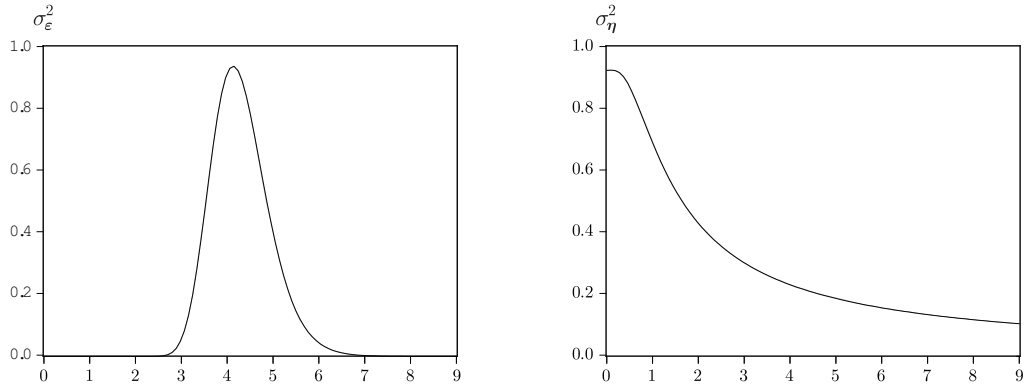
In a similar way as indicated before, we can derive the conditional posterior densities of $\sigma_\eta^2|\sigma_\varepsilon^2 = \bar{\sigma}_\varepsilon^2$ and $\sigma_\varepsilon^2|\sigma_\eta^2 = \bar{\sigma}_\eta^2$. Both conditional densities are again Inverted Gamma-type which are translated by a constant. It follows that Gibbs sampling is now proper because the variance components are distinguishable. Details are omitted in order to save space. Figures 20 and 21 show the conditional densities for some draws of the variance parameters.

Figure 20: **Conditional posterior densities of σ_ε^2 and σ_η^2**



Notes: Panels (a) and (b) show the conditional posterior densities for σ_ε^2 and σ_η^2 for 100 data points generated for the case where the "true" σ_ε^2 is chosen as 4 and the "true" σ_η^2 is chosen as 1.

Figure 21: **Marginal densities of σ_ε^2 and σ_η^2**



(a) σ_ε^2 (b) σ_η^2

Notes: Panels (a) and (b) show the marginal densities for σ_ε^2 and σ_η^2 for 100 data points generated for the case where the "true" σ_ε^2 is chosen as 4 and the "true" σ_η^2 is chosen as 1.

5.3 State space model: Gibbs sampling with explanatory variables

We explain the Gibbs step in a state space model by means of a model that is slightly more complicated than the above local level model:

$$y_t = x_t\beta_t + \varepsilon_t, \quad \text{with} \quad \varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2) \quad \text{and} \quad t = 1, \ldots, T \tag{88}$$

$$\beta_t = \beta_{t-1} + \eta_t, \quad \text{with} \quad \eta_t \sim \mathcal{N}(0, \Sigma_\eta) \quad \text{and} \quad E[\varepsilon_s \eta_{k,t}] = 0 \tag{89}$$

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

$$p(y|\boldsymbol{\beta}, \sigma_\varepsilon^2, \Sigma_\eta) = p(y|\boldsymbol{\beta}, \sigma_\varepsilon^2) p(\boldsymbol{\beta}|\Sigma_\eta) \tag{90}$$

where $\boldsymbol{\beta}$ is the $T \times K$ matrix of latent states. Furthermore, we use $\boldsymbol{\beta}_k$ to denote the k^{th} column of $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_t$ to denote the t^{th} row of $\boldsymbol{\beta}$. $p(\boldsymbol{\beta}|\sigma_\eta^2)$ has to be factorized further down to its individual elements $p(\beta_{k,t}|\beta_{k,t-1})$. It is straightforward to show that the Gibbs step in this case is given by[26]

j^{th} Gibbs step for the state space model with explanatory variables:

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

where KFS represents the Kalman filter sampler using one of the above mentioned algorithms.

[26] 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).

5.4 Empirical application: U.S. GDP growth

In our empirical analysis to infer and forecast U.S. GDP growth we also use the local level/time-varying parameter (TVP) model given in (84) and (85). Specifying a random walk for the state variable μ_t gives the TVP model[27]

$$g_t = c_t + \varepsilon_t, \quad \text{with } \varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2) \text{ and } t = 1, \ldots, T \tag{91}$$

$$c_t = c_{t-1} + \eta_t, \text{ with } \eta_t \sim \mathcal{N}(0, \sigma_\eta^2) \text{ and } E[\varepsilon_t \eta_s] = 0 \tag{92}$$

This model can be interpreted as capturing time varying growth. As a special case of this model we have the random walk with drift model when there is no time-variation in c_t.

When we apply the TVP model to our U.S. GDP growth series we obtain that the posterior mean of σ_ε^2 equals 0.7030 compared to 0.7214 for the random walk model with drift. This suggests that allowing for GDP growth to vary over time does indeed improve in-sample fit. However, the increased in-sample explanatory power of the TVP model does not help in out-of-sample forecasting as shown in Table 3.

5.5 Panel data model

The attractive feature of panel data models is that by using time series observations in addition to 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

$$y_{i,t} = \mu_i + \varepsilon_{it}, \quad \text{with} \quad \varepsilon_{i,t} \sim \mathcal{N}(0, \sigma_\varepsilon^2) \quad \text{and} \quad t = 1, \ldots, T, \; i = 1, \ldots, N, \tag{93}$$

$$\mu_i = \theta + \eta_i, \quad \text{with} \quad \eta_i \sim \mathcal{N}(0, \sigma_\eta^2) \tag{94}$$

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

[27] We also analyzed a more general formulation of the state equation by estimating $c_t = \alpha_0 + \alpha_1 c_{t-1} + \eta_t$ but this resulted in posterior densities which were very tightly centered around 0 (for α_0) and 1 (for α_1) so we settled for the random walk specification.

Gibbs sampling

As for the state space model, the likelihood for a Random Effects panel model can be factorized as

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

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

j^{th} Gibbs step for the panel data model:

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

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

$$M_i = \frac{\sigma_\eta^{2(j-1)}}{\sigma_\eta^{2(j-1)} + (1/T)\sigma_\varepsilon^{2(j-1)}}\iota_T' y^t + \frac{\sigma_\varepsilon^{2(j-1)}}{T\sigma_\eta^{2(j-1)} + \sigma_\varepsilon^{2(j-1)}}\theta_i^{(j)} \tag{95}$$

The expression in (95) shows that draws for μ_i are based on a weighted average of the information in the cross section (through $\theta_i^{(j)}$) and the information in the time-series (through z_t) and that the weights are determined by the two variance components. See also Gelfand *et al.* (1990) for more details.

Empirical application: cross-country GDP growth

We use the Gibbs sampler to analyze the random effects model for a panel of OECD annual real per capita GDP 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 setup of the panel model that we consider here is very limited. For example, we assume that growth rates are independent across countries and that there is no autocorrelation in growth rates. Nevertheless, it may serve as a good starting-point from which to consider more elaborate models.

Table 4 shows posterior results for the full panel (final column) that includes all individual countries (as a single group). In the table we only report posterior standard deviations for Australia since those for the other countries are qualitatively similar. The mean growth rate θ of the 17 countries is estimated at 1.90%. Interestingly, some part of the variation in the data is due to cross-country differences in growth, which is reflected

Table 4: **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]
	$\hat{\sigma}_{\varepsilon}^2$	50.716	45.286	39.833	37.138	47.215	38.042
		[4.251]	[3.246]	[2.568]	[2.182]	[2.146]	[1.321]
	$\hat{\sigma}_{\eta}^2$	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 table shows posterior means and posterior standard deviations (in between brackets) for the random effects panel model (93)-(94) 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 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 U.S. dollars converted at Geary-Khamis purchasing power parities, see Maddison (1995) for a full description. We applied a log transformation to remove the exponential trend in GDP levels across time. Posterior results are based on 100,000 draws after a burn-in of $B = 10,000$ draws and selecting every $h = 10^{\text{th}}$ draw. *** indicates that zero is not contained in the 99% highest posterior density (HPD) region, ** indicates that zero is contained in the 99% but not in the 90% and 95% HPD region and * that zero is contained in the 99% and 95% but not in the 90% HPD region. Only posterior standard deviations for Australia are given as those for the other countries are all very similar.

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[28]. These results, which are shown in the first five columns of Table 4 corroborate the analytical result from Section 5.1 which for a panel model translates to a minimum required number of individuals in a group. The results for $N = 3$ show that, compared to

[28]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 results for larger N, the posterior mean and standard deviation for σ_η^2 are very large. Especially the standard deviation of 35.602 seems to indicate that the second moment does not exist. In fact, we know that with $N = 3$ neither posterior mean nor posterior standard deviation indeed exists. Including at least one additional country helps to identify the mean but still not the variance of σ_η^2. From $N = 6$ onwards the variance seems to be more reasonable, although the values are still comparatively large.

We re-emphasize that this panel data model is used for illustrative purposes only. For a more detailed analysis of cross-country growth analysis over a long period we refer to, e.g. Barro (1991), Sala-i-Martin (1994), Quah (1997) and Zellner and Palm (2004).

6 Forecasting U.S. GDP growth using model averaging

Since the seminal article of Bates and Granger (1969) several papers have shown that combinations of forecasts can outperform individual forecasts in terms of symmetric loss functions. For example, Stock and Watson (2004) find that using forecast combinations to predict output growth in seven countries generally perform better than forecasts based on single models. Marcellino (2004) has extended this analysis to a large European data set and reaches broadly the same conclusion. Bayesian approaches have been widely used to construct forecast combinations, see for example Leamer (1978), Hodges (1987), Draper (1995), Min and Zellner (1993), and Strachan and Van Dijk (2007). In general one derives the posterior probability for any individual model and combines these. The predictive density accounts then for model uncertainty by averaging over the probabilities of individual models. Sala-i-Martin *et al.* (2004) propose a Bayesian model averaging method based on frequentist estimates of individual models and apply their method to long term GDP growth. Ravazzolo *et al.* (2007) extend this approach to a complete Bayesian estimation of a linear regression framework to combine individual models.

Suppose we are considering J predictive models to forecast GDP growth with. The predictive density of $g_{T+1} = \ln GDP_{T+1} - \ln GDP_T$ given the data up to time T, $D = (y_T, X_T)$, is computed by inferring the following linear regression

$$\widehat{g}_{T+1} = c + \sum_{j=1}^{j} \beta_j \widehat{g}_{j,T+1} + \varepsilon_{T+1} \quad \varepsilon_{T+1} \sim \text{ i.i.d.} \mathcal{N}(0, \sigma_\varepsilon^2) \tag{96}$$

where $\widehat{g}_{i,T+1}$ is the forecast given by model $j = 1, .., J$, which can be chosen as the mean of the conditional predictive density for model m_j given D; $p(g_{T+1}|D, m_j)$, for $j = 1, ..., J$. We use this Bayesian averaging scheme to combine the growth forecasts from the models analyzed in Section 4.2.2 and Section 5.4. We estimate the model in (96) and compute forecasts by applying uninformative priors and using the derivations in Section 3. For point estimates and forecasts we use the posterior mean[29].

The set of models we discussed earlier can be grouped in two classes: (i) specifications that model GDP growth with a constant intercept as the random walk plus drift and the simple error correction model and (ii) specifications that model GDP growth with trends as the additive linear trend model for levels, the error correction model and the local level model. We showed in Table 3 that the five models provide similar out-of-sample results and a practitioner may therefore find it difficult to choose which individual model to use. As

[29] An alternative approach, often associated with the general definition of Bayesian model averaging, is to restrict $c = 0$ and to approximate β_j with the posterior probability based on the marginal likelihood for model j. We refer to Ravazzolo *et al.* (2007) for a comparison of these and alternative methods.

an alternative, we propose applying the Bayesian averaging approach using three schemes. In the first scheme we average models which only have constant terms (μ,c), in the second scheme we average models with trends and as third and final scheme we average over all the five models. In the empirical application we compute additional forecasts from individual models for the period 1985Q1-1992:Q2, for a total of 30 observations, and we use these to compute the first forecast for 1992:Q3 for the averaging schemes in (96). We use an expanding window to update the model averaging coefficients.

The results of the three schemes as presented in the final columns of Table 3. The table provides evidence that model averaging is an appealing strategy for forecasting. All the three schemes produce MSPEs which are similar to the best individual model. The first scheme in fact even outperforms the best individual model which is the additive linear trend model for levels. The combination of models with only constant terms gives the most accurate forecasts even if the two individual models used in this strategy are less precise than the other models. A possible explanation is that having to determine only two averaging weights results in smaller estimation error than in the other two schemes, where three and five weights need to be derived respectively. For the first scheme we find that the estimated weights of the random walk with drift and the simple ECM model do show quite some variation over time. A promising extension over the averaging scheme in (96) would therefore be to make the weights time varying but we consider this to be beyond the scope of this paper. To summarize, Table 3 indicates that model averaging is a safe and accurate strategy, but that its success is likely to depend on how it is implemented.

7 Final Remarks: Models used and Lessons Learnt

Using a set of basic economic time series models, focusing on dynamic processes and time varying structures, we presented the results of a Bayesian analysis for the case where substantial posterior probability is near and/or at the boundary of the parameter region. As canonical models we considered the dynamic regression model with autocorrelation in the disturbances and the Hierarchical Linear Mixed Model in a variance components setup. We have indicated how several empirically relevant model structures are simple extensions of these two canonical ones. A summary of models used and their key features is shown in Table A-1. The results of our Bayesian analysis may be summarized as 'lessons learnt'. We describe these as follows:

Single equation dynamic regression models and no boundary issues

When the model and data information are such that boundary issues do not occur even under diffuse priors, then posteriors are - at least - locally proper and basic simulation techniques like Gibbs sampling can successfully be used for computing posterior and predictive results. The role of deterministic trend terms should be carefully evaluated.

Single equation dynamic regression models and the presence of boundary issues

In many macroeconomic processes, the information in the data is weak and the mass of the likelihood function may be close to the boundary of the parameter region. Examples are nearly nonstationary processes or nearly nonidentified processes as exhibited by inflation, interest rates, dividend yield and GDP processes. The single equation dynamic regression model serves in this case as a workhorse for unit root models, distributed lag models,

and error correction models. A flat prior leads to improper posteriors. Regularization or smoothness priors like the Information matrix prior may be fruitfully used to obtain proper posteriors. Gibbs sampling may fail and the use of more indirect sampling methods like Importance Sampling and/or Metropolis Hastings is to be recommended.

Time varying parameters models and the presence of boundary issues

A simple regression model with a time varying variance explains the structure of Hierarchical Linear Mixed Models (HLMM). The latter class serves as workhorse for state space models and panel data models. Boundary issues occur due to a lack of degrees of freedom or nonidentification of the variance component. The degrees of freedom problem may occur in panel data problems when a small number of groups is considered. Here we concentrate on right tail behavior of the posterior density. When there exists a substantial probability near the zero variance bound, uniform priors are still recommended, since conditional conjugate priors will cut off relevant data mass near zero, see also the recommendation by Gelman (2006). Informative dynamic structure in time series is used in state space models and a sufficient number of units are used in the cross-sectional models to regularize the shape of the posterior. This is shown theoretically and empirically.

Given these 'lessons', a practical path for an empirical economic researcher is to investigate the shape of the posterior distribution of the parameters of interest and to classify this shape in two categories. As long as this shape is approximately elliptical and much probability mass is in the interior of the parameter region, then applying Gibbs sampling is straightforward and yields accurate results. When the posterior distribution has strong nonelliptical contours and substantial mass is at the boundary of the parameter region then warning signals for the researcher should appear. It depends on the specification of the model and the information in the data in which situation a researcher is located. A second advice for the empirical researcher is to apply Bayesian model averaging in cases where substantial posterior probability is at the boundary of the parameter region, compare Section 6.

Some directions on how to continue further in nonstandard cases are left for future research. One could think of a reparameterization of the model, the use of subjective informative priors, and the use of predictive priors, see, for example, Geweke (2005).

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.

Bates, J. M. and C. W. J. Granger (1969), Combination of Forecasts, *Operational Research Quarterly*, 20, 451–468.

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

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

Campbell, J. and R. Shiller (1988), Stock Prices, Earnings, and Expected Dividends, *Journal of Finance*, 43, 661–676.

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.

Celeux, G., M. Hurn, and C. Robert (2000), Computational and Inferential Difficulties with Mixture Posterior Distributions, *Journal of the American Statistical Association*, 95, 957–970.

Chib, S. (1991), Bayes Regression With Autocorrelated Errors: A Gibbs Sampling Approach, *Journal of Econometrics*, 58, 275–294.

Chib, S. (1993), Bayes Estimation of Regressions with Autoregressive Errors: A Gibbs Sampling Approach, *Journal of Econometrics*, 58, 275–294.

Chib, S. and B. P. Carlin (1999), On MCMC Sampling in Hierarchical Longitudinal Models, *Statistics and Computing*, 9, 17–26.

Chib, S. and E. Greenberg (1994), Bayes Inference in Regression Models with ARMA (p,q) Errors, *Journal of Econometrics*, 64, 183–206.

Chib, S. and E. Greenberg (1995), Understanding the Metropolis-Hasting Algorithm, *The American Statistician*, 49, 327–335.

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

Cochrane, J. H. (2006), The Dog That Did Not Bark: A Defense of Return Predictability, *forthcoming in Review of Financial Studies*.

Draper, D. (1995), Assessment and Propagation of Model Uncertainty, *Journal of the Royal Statistical Society Series B*, 56, 45–98.

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

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

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

Fama, E. and K. French (1988), Dividend Yields and Expected Stock Returns, *Journal of Financial Economics*, 22, 3–25.

Fieller, E. C. (1954), Some Problems in Interval Estimation, *Journal of the Royal Statistical Society. Series B*, 16, 175–185.

Frühwirth-Schnatter, S. (1994), Applied State Space Modelling of Non-Gaussian Time Series Using Integration-based Kalman Filtering, *Statistics and Computing*, 4, 259–269.

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.

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

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

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

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

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

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

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

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

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

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

Geweke, J. (2007), Interpretation and Inference in Mixture Models: Simple MCMC works, *Computational Statistics and Data Analysis*, 51, 3529–3550.

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

Griffiths, W. E., R. G. Drynan, and S. Prakash (1979), Bayesian Estimation of a Random Coefficients Model, *Journal of Econometrics*, 10, 201–220.

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

Hamilton, J. D. (1989), A New Approach to the Economic Analysis of Nonstationary Time-Series Subject to Changes in Regime, *Econometrica*, 57, 357–384.

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

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

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

Harvey, A. C., T. M. Trimbur, and H. K. Van Dijk (2007), Bayes Estimates of the Cyclical Component in Twentieth Century U.S. Gross Domestic Product, in G. L. Mazzi and G. Savio (eds.), *Growth and Cycle in the Eurozone*, Palgrave MacMillan, New York, pp. 76–89.

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.

Hodges, J. (1987), Uncertainty, Policy Analysis and Statistics, *Statistical Science*, 2, 259–291.

Hoogerheide, L. F., J. F. Kaashoek, and H. K. Van Dijk (2007), 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, *Journal of Econometrics*, 139, 154–180.

Hoogerheide, L. F. and H. K. Van Dijk (2001), Comparison of the Anderson-Rubin Test for Overidentification and the Johansen Test for Cointegration, *Econometric Institute Report 2001-4*.

Hoogerheide, L. F. and H. K. Van Dijk (2008), Possible Ill-behaved Posteriors in Econometric Models, *Tinbergen Institute Discussion Paper 08-36/4*.

Hoogerheide, L. F., H. K. Van Dijk, and R. D. Van Oest (2008), Simulation Methods for Bayesian Econometric Inference, in *Handbook of Computational Economics and Statistics*, Elsevier (*forthcoming*).

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

Jasra, A., C. Holmes, and D. Stephens (2005), Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling, *Statistical Science*, 20, 50–67.

Keim, D. and R. Stambaugh (1986), Predicting Returns in the Stock and Bond Markets, *Journal of Financial Economics*, 17, 357–424.

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

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

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

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

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

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

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

Leamer, E. (1978), *Specification Searches*, New York: Wiley.

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.

Marcellino, M. (2004), Forecasting Pooling for Short Time Series of Macroeconomic Variables, *Oxford Bulletin of Economic and Statistics*, 66, 91–112.

Min, C. and A. Zellner (1993), Bayesian and Non-Bayesian Methods for Combining Models and Forecasts with Applications to Forecasting International Growth Rates, *Journal of Econometrics*, 56, 89–118.

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

Nelson, C. and C. Plosser (1982), Trends and Random Walks in Macroeconomic Time Series: Some Evidence and Implications, *Journal of Monetary Economics*, 10, 139–162.

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

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

Press, S. J. (1969), The t-Ratio Distribution, *Journal of American Statistical Association*, 64, 242–252.

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.

Ravazzolo, F., H. K. van Dijk, and M. Verbeek (2007), Predictive Gains from Forecast Combination Using Time-varying Model Weight, *Econometric Institute Report 2007-26*.

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 (1991a), A Bayesian Analysis of the Unit Root in Real Exchange Rates, *Journal of Econometrics*, 49, 195–238.

Schotman, P. and H. K. Van Dijk (1991b), On Bayesian Roots to Unit Roots, *Journal of Applied Econometrics*, 6, 387–401.

Schotman, P. and H. K. Van Dijk (1993), Posterior Analysis of Possibly Integrated Time Series with an Application to Real GNP, in P. Caines, J. Geweke, and M. Taqqu (eds.), *New Directions in Time Series Analysis part II*, Springer Verlag, Heidelberg, pp. 341–361.

Sims, C. A. and H. Uhlig (1991), Understanding Unit Rooters: A Helicopter Tour, *Econometrica*, 59, 1591–1599.

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.

Stock, J. H. and M. W. Watson (2002), Has the Business Cycle Changed and Why?, *NBER Macroeconomics Annual*.

Stock, J. H. and M. W. Watson (2004), Combination Forecasts of Output Growth in a Seven-country Data Set, *Journal of Forecasting*, 23, 405–430.

Strachan, R. and H. K. Van Dijk (2007), Bayesian Model Averaging in Vector Autoregressive Processes with an Investigation of Stability of the US Great Ratios and Risk of a Liquidity Trap in the USA, UK and Japan, *Econometric Institute Report 2007-09*.

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

Sala-i-Martin, X., G. Doppelhoffer, and R. Miller (2004), Determinants of long-term growth: A Bayesian averaging of classical estimates (BACE) approach, *American Economic Review*, 94, 813–835.

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

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

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

Van Dijk, H. K. (2004), Twentieth Century Shocks, Trends and Cycles in Industrialized Nations, *De Economist*, 152, 211–232.

Zellner, A. (1971), *An Introduction to Bayesian Inference in Econometrics*, Wiley, New York.

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.

Zellner, A. and F. C. Palm (2004), *The Structural Econometric Time Series Analysis Approach*, Cambridge University Press.

A Summary of models used and lessons learnt

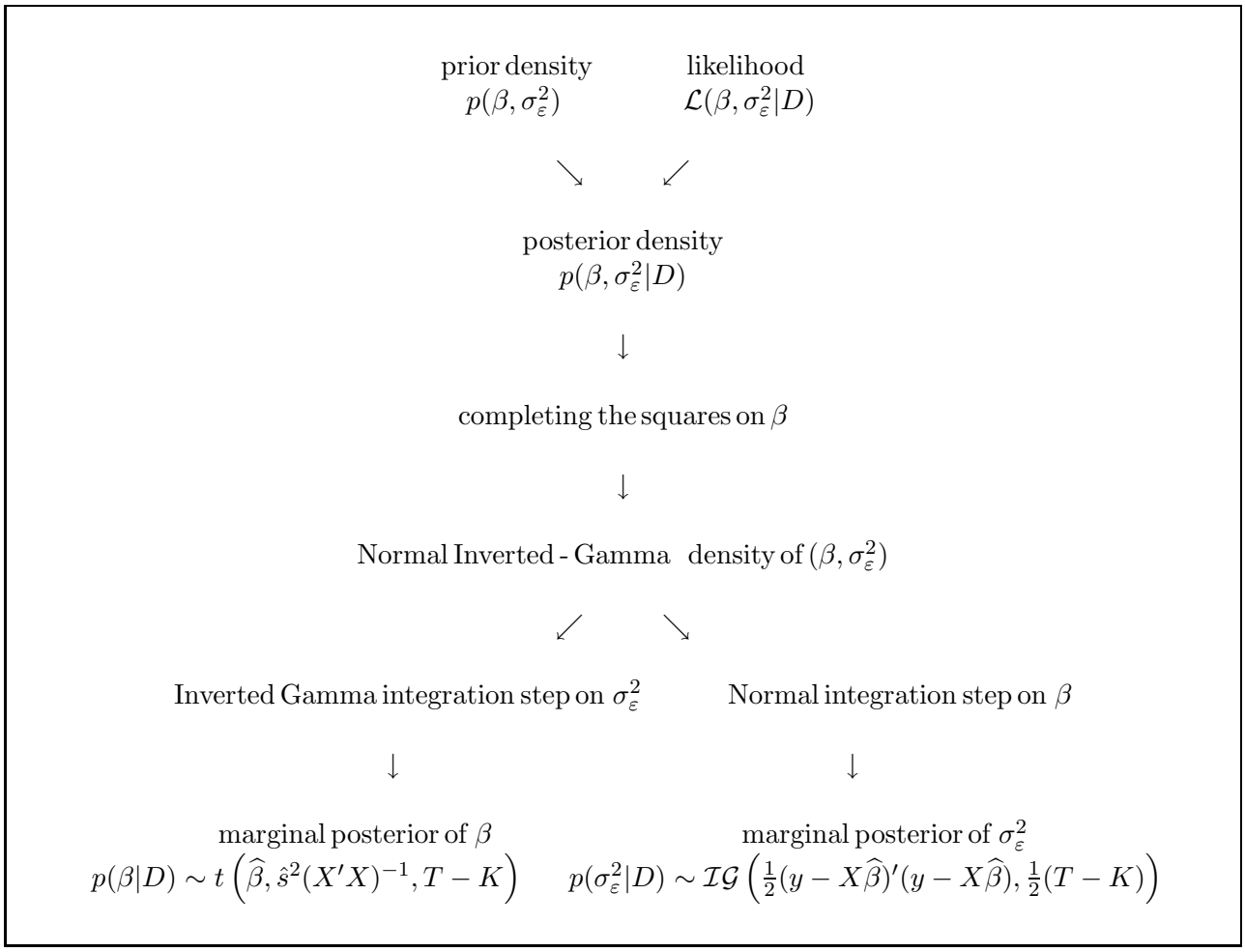
Table A-1: **Summary of models and their key features**

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

B Summary results for linear regressions models

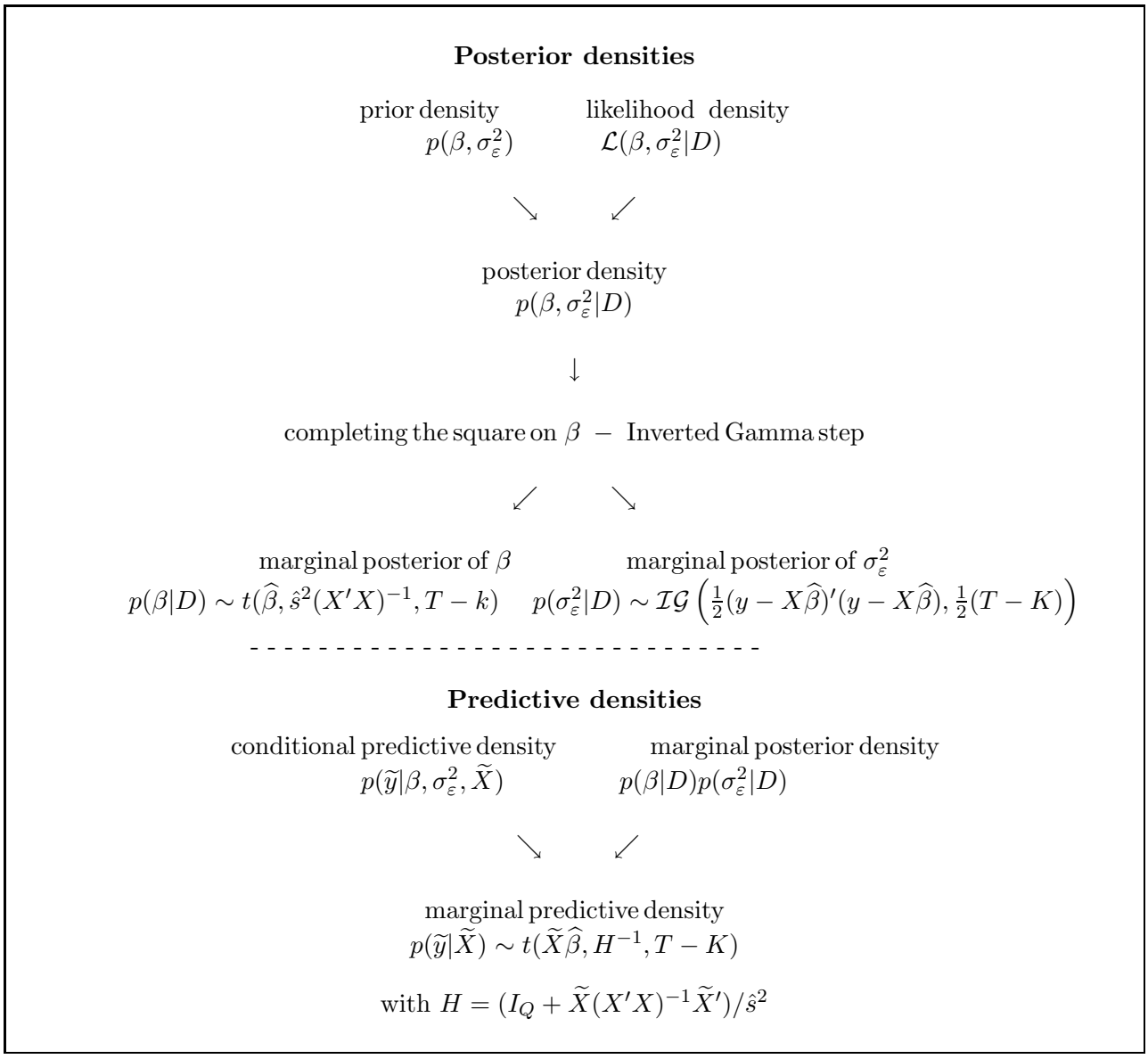
The diagrams below trace the main steps to derive the posterior and the predictive densities of a linear model by direct sampling and by the Gibbs algorithm. Results of the multivariate linear regression model are reported, but the scheme can be adapted to other linear forms. The tables list the probability density functions of the posteriors for the parameters and of the predictive density for different linear models.

Figure B-1: **Sampling scheme: joint and marginal posterior density results**



Notes: The figure presents joint and marginal density results for direct sampling.

Figure B-2: **Sampling scheme: marginal and predictive density results**



Notes: The figure presents marginal and predictive density results for direct sampling.

Table B-1: **Overview of distributions for *direct* sampling**

model	**marginal density regression parameters**	**marginal density residual variance**	**predictive density**
constant	univariate Student-t $t\left(\widehat{\mu}, \hat{s}^2/T, T-1\right)$	Inverted Gamma $\mathcal{IG}\left(\frac{1}{2}(y-\iota\widehat{\mu})'(y-\iota\widehat{\mu}), \frac{1}{2}(T-1)\right)$	multivariate Student-t $t\left(\widehat{\mu}, \hat{s}^2(\mathrm{I}_Q + Q/T), T-1\right)$
multiple regression	multivariate Student-t $t\left(\widehat{\beta}, \hat{s}^2(X'X)^{-1}, T-K\right)$	Inverted Gamma $\mathcal{IG}\left(\frac{1}{2}(y-X\widehat{\beta})'(y-X\widehat{\beta}), \frac{1}{2}(T-K)\right)$	Multivariate Student-t $t\left(\widetilde{X}\widehat{\beta}, H^{-1}, T-K\right)$

Table B-2: **Overview of distributions for *Gibbs* sampling**

model	**posterior density regression parameters**	**posterior density residual variance**	**predictive density**
constant	univariate normal $\mathcal{N}\left(\widehat{\mu}, \sigma_\varepsilon^2/T\right)$	Inverted Gamma $\mathcal{IG}\left(\frac{1}{2}(y-\iota\mu)'(y-\iota\mu), \frac{1}{2}T\right)$	multivariate normal $\mathcal{N}\left(\widetilde{\mu}, \sigma_\varepsilon^2\mathrm{I}_Q\right)$
multiple regression	multivariate normal $\mathcal{N}\left(\widehat{\beta}, \sigma_\varepsilon^2(X'X)^{-1}\right)$	Inverted Gamma $\mathcal{IG}\left(\frac{1}{2}(y-X\beta)'(y-X\beta), \frac{1}{2}T\right)$	multivariate normal $\mathcal{N}(\widetilde{X}\beta, \sigma_\varepsilon^2\mathrm{I}_Q)$

C Probability Density Functions

In this appendix we specify several univariate and multivariate probability density functions 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 typically indicate vectors or matrices. More properties of the below densities and concise derivations of moments and moment conditions can be found in for example Raiffa and Schlaifer (1961) or Poirier (1995).

C.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

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

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

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

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

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

with $B(\frac{1}{2},\frac{1}{2}\nu)$ the Beta 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 m and ν, i.e. $Z \sim \mathcal{IG}(m, \nu)$, then the density of Z and its first two moments are given by

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

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

with $\Gamma(\nu)$ the Gamma function defined as $\Gamma(\nu) \equiv \int_0^\infty z^{\nu-1} \exp(-z) dz$.

C.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

$$f_{\mathcal{N}}^{(N)}(z|M,S) \equiv (2\pi)^{-\frac{1}{2}N}|S|^{-\frac{1}{2}} \exp\left[-\tfrac{1}{2}(z-M)'S^{-1}(z-M)\right] \quad \text{for} \quad \begin{array}{l} -\infty < z < \infty \\ -\infty < M(i) < \infty \ [i = 1, \ldots, N] \\ z'Sz > 0 \ \forall \ z \neq 0 \end{array} \tag{C-4}$$

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

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

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

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