Bayesian Prediction with a Cointegrated Vector Autoregression

Mattias Villani*

Abstract

A complete procedure for calculating the joint predictive distribution of future observations based on the cointegrated vector autoregression is presented. The large degree of uncertainty in the choice of the cointegration vectors is incorporated into the analysis through a prior distribution on the cointegration vectors which allows the forecaster to realistically express his beliefs. This prior leads to a form of model averaging where the predictions from the models based on the different cointegration vectors are weighted together in an optimal way. The ideas of Litterman (1980) are adapted for the prior on the short run dynamics with a resulting prior which only depends on a few hyperparameters and is therefore easily specified. A straight forward numerical evaluation of the predictive distribution based on Gibbs sampling is proposed. The prediction procedure is applied to a seven variable system with focus on forecasting the Swedish inflation.

Keywords: Bayesian, Cointegration, Inflation forecasting, Model averaging, Predictive density.

1. Introduction

The idea of cointegration (Engle and Granger, 1987; Johansen, 1995) has become extremely popular in applied work. With the introduction of this concept, econometricians were given the possibility to incorporate theoretically motivated long run equilibriums into their otherwise relatively unrestricted models.

In most applications, many theoretical cointegration restrictions are available and the choice between them is often a very difficult task. Even though the restrictions are empirically testable, the evidence from such tests is often inconclusive; we may obtain weak support for one or several of the theoretical restrictions while

*Department of Statistics, Stockholm University, S-106 91 Stockholm, Sweden. E-mail: mattias.villani@stat.su.se.

the other restrictions are clearly rejected or we may even find strong support for more than one of them. In addition, we must also consider the possibility to ignore them all and estimate the long run relations entirely from data.

If the model is to be used for prediction then there is, of course, always the possibility to produce forecasts based from each of the models with the different theoretical restrictions imposed and from the model with empirically estimated restrictions. It seems reasonable, however, to weight these predictions with a weight proportional to the empirical support given to the restriction. In any case, it is often unsatisfying, and practically complicated, to have a whole range of predictions based on different restrictions.

The aim of this paper is to introduce a complete Bayesian approach to prediction using a cointegrated VAR model. In a Bayesian setting, the uncertainty regarding the correct cointegration restriction should be reflected in the prior distribution of the model parameters. We propose a prior distribution which allows us to work simultaneously with several plausible cointegration restrictions with an optimal way to combine the predictions. Furthermore, the suggested prior also allows parts or even all of the cointegration restrictions to be estimated freely from the data. In effect, long run restrictions are imposed only if the data support them.

Even if we impose long-run restrictions on the VAR, most parameters describe the short run dynamics of the system and these parameters are still unrestricted. Litterman (1980, 1986) improved the prediction performance of the unrestricted VAR by using a prior distribution on the coefficients that centered the VAR over the unrelated random walk model, *a priori*. A slight variant of this idea is used here to improve the estimates of the short run dynamics, and therefore to improve predictions.

An often neglected part of the prediction phase is the uncertainty in the point predictions, which in many cases is as important as the point predictions. This is certainly true in the inflation forecasting application dealt with here as an increasing number of central banks now reports their inflation targets as an interval rather than a single number. In traditional forecasting applications of the VAR with cointegration restrictions, the forecasting uncertainty is distorted by primarily two conditionings. First, the conditioning on one of the theoretical suggestions in the forecasting stage will produce a distorted view of the certainty in the forecasts. Secondly, the uncertainty attributed to the estimation of the model parameters is at best accounted for by rough, and often dubious, corrections based on asymptotic theory. The aim here is to produce uncertainty statements that fully account for all sources of uncertainty and not only those attributed to future stochastic disturbances to the system.

2. The cointegrated vector autoregression

Consider the ordinary p-dimensional vector autoregressive process with K lags

$$\mathbf{x}_t = \sum_{i=1}^{K} \mathbf{\Pi}_i \mathbf{x}_{t-i} + \mathbf{\Phi}\mathbf{d}_t + \boldsymbol{\varepsilon}_t, \tag{2.1}$$

where $\mathbf{x}_t$ contains an observation on the p time series at time t, $\mathbf{\Pi}_i$ is the matrix of coefficients describing the dynamics of the system while $\mathbf{d}_t$ contains d deterministic trend or dummy variables at time t whose effect on $\mathbf{x}_t$ is captured by $\mathbf{\Phi}$. Finally, $\boldsymbol{\varepsilon}_t$ is a vector of error terms assumed to follow the $N_p(\mathbf{0}, \mathbf{\Sigma})$ distribution with independence between time periods.

An equivalent parametrization of the VAR model, better suited for cointegration analysis, is the error correction model (ECM)

$$\mathbf{\Delta x}_t = \mathbf{\Pi x}_{t-1} + \sum_{i=1}^{K-1} \mathbf{\Gamma}_i \mathbf{\Delta x}_{t-i} + \mathbf{\Phi}\mathbf{d}_t + \boldsymbol{\varepsilon}_t, \tag{2.2}$$

where $\mathbf{\Delta x}_{t-i} = \mathbf{x}_{t-i} - \mathbf{x}_{t-i-1}$, $\mathbf{\Pi} = \sum_{i=1}^{K} \mathbf{\Pi}_i - \mathbf{I}_p$ and $\mathbf{\Gamma}_i = -\sum_{j=i+1}^{K} \mathbf{\Pi}_j$. The greatest merit of the ECM is that it separates the long-run component of the series ($\mathbf{\Pi x}_{t-1}$) from the short-run dynamics ($\sum_{i=1}^{K-1} \mathbf{\Gamma}_i \mathbf{\Delta x}_{t-i}$), a separation that both simplifies the interpretation of the non-stationary processes dealt with in this paper and suggests the two major types of prior beliefs for such processes, see section 4.

The concept of *integration* is essential to the definition of cointegration. A process is integrated of order zero, or $I(0)$, if, in rough terms, it is stationary but its cumulative sum is non-stationary, see Johansen (1995a) for a more precise definition. A process is said to be integrated of order d, or $I(d)$, if the dth difference of the process is $I(0)$.

If the processes are at most $I(1)$, and the rank of $\mathbf{\Pi}$ is equal to r, then there exist r linear combinations of the time-series that are $I(0)$ (Engle and Granger, 1987). That is, the individual processes may be $I(1)$, and therefore drift around like random walks, but, at least some of them, are tied together by long-run equilibrium relationships given by the coefficients of the r stationary linear combinations; the original processes are said to be *cointegrated*. Formally, if $rank(\mathbf{\Pi}) = r$, we can write $\mathbf{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}'$, where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are both matrices of dimension $p \times r$. By inserting this decomposition into (2.2) we obtain

$$\mathbf{\Delta x}_t = \boldsymbol{\alpha}\boldsymbol{\beta}'\mathbf{x}_{t-1} + \sum_{i=1}^{K-1} \mathbf{\Gamma}_i \mathbf{\Delta x}_{t-i} + \mathbf{\Phi}\mathbf{d}_t + \boldsymbol{\varepsilon}_t, \tag{2.3}$$

where $\boldsymbol{\beta}'\mathbf{x}_{t-1}$ can be interpreted as a new set of r stationary processes representing departures from the r equilibria while the *adjustment coefficients* in $\boldsymbol{\alpha}$ describe the adjustment back to equilibrium. The columns of $\boldsymbol{\beta}$ are the *cointegration vectors*.

Note that the decomposition of $\boldsymbol{\Pi}$ into $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ is not unique; any $r \times r$ invertible matrix $\mathbf{T}$ can be used to transform $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}'$ without affecting their product; that is, $\boldsymbol{\Pi} = \boldsymbol{\alpha}\mathbf{T}\mathbf{T}^{-1}\boldsymbol{\beta}' = \boldsymbol{\alpha}^*\boldsymbol{\beta}^{*\prime}$, where $\boldsymbol{\alpha}^* = \boldsymbol{\alpha}\mathbf{T}$ and $\boldsymbol{\beta}^{*\prime} = \mathbf{T}^{-1}\boldsymbol{\beta}'$, and only the column spaces of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are therefore determinable and not $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ themselves. Thus, without restrictions on $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ the model in (2.3) is not identified. To get a unique decomposition of $\boldsymbol{\Pi}$, restrictions on either $\boldsymbol{\alpha}$ or $\boldsymbol{\beta}$ must be imposed. The latter option is chosen here. Following Johansen (1995a), general linear identifying restrictions on the ith cointegration vector can be represented in the form

$$\boldsymbol{\beta}_i = \mathbf{H}_i\boldsymbol{\varphi}_i, \ (i = 1, ..., r) \tag{2.4}$$

where $\mathbf{H}_i$ is an $p \times (p - r_i)$ matrix determined by the r_i restrictions and $\boldsymbol{\varphi}_i$ is a vector containing the remaining $p - r_i$ free elements in $\boldsymbol{\beta}_i$. The restrictions in (2.4) only determine $\boldsymbol{\beta}_i$ up to an arbitrary constant and a normalization of the following form will be used to settle this indeterminacy

$$\boldsymbol{\beta}_i = \mathbf{h}_i + \mathbf{H}^i\boldsymbol{\psi}_i, \ (i = 1, ..., r) \tag{2.5}$$

where $\mathbf{h}_i$ zero vector with unity in the ith position, $\mathbf{H}^i$ is the same as $\mathbf{H}_i$ but with the ith row equal to a zero vector and $\boldsymbol{\psi}_i$ is a vector with the $d_i = p - r_i - 1$ free coefficients left after the identifying restrictions and the normalization. The conditions on the set of $\mathbf{h}_i$ and $\mathbf{H}^i$ to be identifying can be found in Johansen (1995b) and is automatically checked in computer programs like PcFiml (Doornik and Hendry, 1997) and Cats in Rats (Hansen and Juselius, 1995).

It will be convenient to have a condensed representation of the ECM in the actual estimation phase, where all observations are shown explicitly. For this purpose, let us write the model in (2.3) for the whole sample of T time periods as

$$\boldsymbol{\Delta}\mathbf{X} = \mathbf{X}_{-1}\boldsymbol{\beta}\boldsymbol{\alpha}' + \mathbf{Z}\boldsymbol{\Gamma} + \mathbf{D}\boldsymbol{\Phi}' + \mathbf{E}, \tag{2.6}$$

where $\boldsymbol{\Delta}\mathbf{X} = (\boldsymbol{\Delta}\mathbf{x}_T, ..., \boldsymbol{\Delta}\mathbf{x}_1)'$, $\mathbf{X}_{-i} = (\mathbf{x}_{T-i}, ..., \mathbf{x}_{-i+1})'$, $\mathbf{Z} = (\boldsymbol{\Delta}\mathbf{X}_{-1}, ..., \boldsymbol{\Delta}\mathbf{X}_{-K+1})$, $\boldsymbol{\Gamma} = (\boldsymbol{\Gamma}_1, ..., \boldsymbol{\Gamma}_K)'$, $\mathbf{D} = (\mathbf{d}_T, ..., \mathbf{d}_1)'$ and $\mathbf{E} = (\boldsymbol{\varepsilon}_T, ..., \boldsymbol{\varepsilon}_1)'$.

3. An application to inflation forecasting in Sweden

The methods introduced in this paper are applied to a data set collected and analyzed by Jacobson, Jansson, Vredin and Warne (1999), hereafter referred to as JJVW, at the central bank of Sweden in an attempt to find a benchmark model

from which monetary policy can be studied and inflation forecasts generated. The data consist of quarterly observations from 1972:2 to 1996:4 on the following seven variables

$$\mathbf{x}_t = \begin{pmatrix} y_t & p_t & i_t & e_t & y_t^* & p_t^* & i_t^* \end{pmatrix}',$$

where all starred variables refer to foreign variables and the remaining variables are Swedish measures. y_t denotes 100 ln GDP, $p_t = 100 \ln$ CPI, $i_t = 100 \ln(1+I_t/100)$, where I_t is the three month treasury bills rate in percent and $e_t = 100 \ln S_t$, where S_t is the geometric sum of the nominal Swedish Krona exchange rate of Sweden's 20 most important trading partners.

To control for large devaluations of the Swedish krona and regime shifts in economic policy both in Sweden and in foreign countries, JJVW added five dummy variables to the analysis. Finally, a constant term was included in the model and thus

$$\mathbf{d}_t = \begin{pmatrix} 1 & i_{1,t} & i_{2,t} & i_{3,t} & i_{4,t} & i_{5,t} \end{pmatrix}',$$

where $i_{j,t}$ denotes the jth dummy at time t. Further details can be found in JJVW.

A battery of tests in JJVW suggested firmly that the four lags in the VAR model were sufficient. Bayesian lag length inference in the vector autoregression has been proposed in Villani (1998a) based on the fractional Bayes approach to model selection (O'Hagan, 1995). With a vague prior on the VAR parameters and a uniform prior on the lag length from $K = 0$ to $K = 8$, the following posterior probabilities were obtained: $p(K = 4|\mathbf{x}^{(T)}) = 0.904$, $p(K = 5|\mathbf{x}^{(T)}) = 0.0931$, $p(K = 6|\mathbf{x}^{(T)}) = 0.0029$, where $\mathbf{x}^{(T)}$ denotes data up to time T, and approximately zero for other lag lengths. Thus, this Bayesian analysis supports the use of the VAR with four lags.

The common trends representation of the cointegrated VAR (Stock and Watson, 1988; Warne, 1993) was used by JJVW in an attempt to identify the number of cointegration relationships. Basically this idea stems from the fact that if the VAR process has r cointegration relations, then it can be shown (Stock and Watson, 1988) that the individual series are driven by $p - r$ underlying, unobservable, random walks called *common trends*. Usually such trends are assumed to be general economic factors like technology or money supply, or simply real and nominal shocks, respectively. In JJVW it is argued that four common trends are expected: real and nominal stochastic trend both in Sweden and abroad; the usual empirical tests and descriptive statistics gave some, but far from conclusive, support to this theoretically motivated choice. To be able to compare our results with the ones in JJVW, we will use exactly the same model with $K = 4$ and $r = 3$.

JJVW discusses different, *a priori* plausible, structures of $\boldsymbol{\beta}$ and several fully specified cointegration vectors are put forward as candidates for long run equilibrium relationships. A short summary is given in Table 3.1.

Relation	Cointegration vector	Economic reasoning
$e_t + p_t^* - p_t$	$\mathbf{b}_1 = (0, -1, 0, 1, 0, 1, 0)'$	Goods market equilibrium: real exchange rate is $I(0)$, PPP.
i_t	$\mathbf{b}_2 = (0, 0, 1, 0, 0, 0, 0)'$	$i_t = \bar{i}_t + E(\Delta p_t)$, where $\bar{i}_t$ is the real interest rate and both $\bar{i}_t$ and $E(\Delta p_t)$ are assumed to be $I(0)$.
i_t^*	$\mathbf{b}_3 = (0, 0, 0, 0, 0, 0, 1)'$	See i_t.
$i_t - i_t^*$	$\mathbf{b}_4 = (0, 0, 1, 0, 0, 0, -1)'$	Financial markets equilibrium: $i_t = i_t^* + E(\Delta e_t)$, $E(\Delta e_t)$ is $I(0)$.

Table 3.1: Plausible long-run relationships and their cointegration vectors for the Swedish monetary data

4. Prior distribution

4.1. The prior on $\mathbf{\Phi}$ and $\mathbf{\Sigma}$

Assuming that very little information on $\mathbf{\Phi}$ and $\mathbf{\Sigma}$ is available, then the following default prior seems acceptable

$$p(\mathbf{\Phi}, \mathbf{\Sigma}) \propto |\mathbf{\Sigma}|^{-(p+1)/2},$$

which is the limiting case of a Wishart prior on $\mathbf{\Sigma}^{-1}$ with the degrees of freedom approaching zero (Geisser, 1965), and is therefore a prior that adds very little prior information into the analysis.

4.2. The prior on the short run dynamics

Let $\mathbf{\Gamma}$ be distributed independently of $\mathbf{\Phi}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ *a priori,* and assume that $\operatorname{vec} \mathbf{\Gamma}$ follows a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Psi}$. In a successful attempt to improve the predictive ability of the unrestricted VAR, Litterman (Litterman, 1980, 1986; Doan et al. 1984) endowed the normal prior on the parameters with more structure. Some of his ideas will here be adapted to fit the short-run dynamics of the cointegrated VAR, see Stark (1998) for a similar approach. The following statements provide a starting point for the prior on $\mathbf{\Gamma}$.

- The coefficients in $\mathbf{\Gamma}$ should be more or less centered around zero, thus $\boldsymbol{\mu} = \mathbf{0}$.
- The elements in $\mathbf{\Gamma}$ can for convenience be assumed independent *a priori.*
- Any given element in $\mathbf{\Gamma}_{i+1}$ is more likely to be zero than the corresponding element in $\mathbf{\Gamma}_i$.

- The beliefs about a coefficient that describes the dynamics *within* a variable may be different from the beliefs about a coefficient that describes the dynamics *between* variables.

Let γ_{ij}^k denote the element at the ith row and jth column of $\mathbf{\Gamma}_k$, that is, the coefficient that describes how the difference of the ith series $(\Delta x_{i,t})$ is affected by changes in difference of the jth series lagged k time periods $(\Delta x_{j,t-k})$. A plausible structure on $\mathbf{\Psi}$, which is diagonal by the second statement above, satisfying the stated requirements is

$$Std(\gamma_{ij}^k) = \begin{cases} \frac{\lambda}{k} & \text{Own lags } (i = j) \\ \frac{\lambda\theta}{k}\frac{\sigma_i}{\sigma_j} & \text{Foreign lags } (i \neq j) \end{cases}$$

where $Std(\cdot)$ denotes the standard deviation, $\lambda = Std(\gamma_{ii}^1)$, for all i, and, if it is regarded as more probable *a priori* that a series is affected by its own lags than that it is affected by the lags of another series, then $\theta \in (0, 1)$. σ_i is the square root of the ith diagonal element of $\mathbf{\Sigma}$ whose presence adjusts for the differing variability in the time series. Note that $Var(\gamma_{ij}^k)$ is a decreasing function of the lag length implying that longer lags are more likely to have coefficients equal to zero. Other damping rates than k^{-1} on the standard deviation scale can, of course, be used if it is considered more appropriate.

If beliefs about one or more coefficients in $\mathbf{\Gamma}$ do not conform to the structure described above, then the corresponding elements in $\mathbf{\Psi}$ may simply be modified to fit the genuine beliefs of the investigator. For example, for quarterly data we may expect that the coefficients of the fourth lag is more likely to be different for zero when compared to, for example, the third lag. The elements in $\mathbf{\Psi}$ corresponding to coefficients of the fourth lag should then be increased.

4.3. The prior on the long run structure

Typically, there are some cointegration vectors which the investigator holds as particularly plausible *a priori*. Such vectors will here be termed *candidate vectors.* We have r cointegration vectors to determine and the candidate vectors, which can of course appear simultaneously if $r > 1$, can therefore be mixed to form a set of Q *matrix candidates* for $\boldsymbol{\beta}$. Not all r-subsets of the vector candidates will be plausible *a priori,* and therefore discarded as matrix candidates, but there is also the possibility that only some columns of a matrix candidate are specified and the remaining ones are estimated from data. For the Swedish monetary data in section 3, the four vector candidates were given in Table 3.1. Plausible matrix candidates are $\boldsymbol{\beta}_1 = (\mathbf{b}_1, \cdot, \cdot)$, $\boldsymbol{\beta}_2 = (\mathbf{b}_4, \cdot, \cdot)$, $\boldsymbol{\beta}_3 = (\mathbf{b}_2, \mathbf{b}_3, \cdot)$, $\boldsymbol{\beta}_4 = (\mathbf{b}_1, \mathbf{b}_4, \cdot)$, $\boldsymbol{\beta}_5 = (\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3)$ and $\boldsymbol{\beta}_6 = (\cdot, \cdot, \cdot)$, where a dot denotes an unspecified column of

$\boldsymbol{\beta}$. Two things are worth noting. First, the matrix consisting of only unspecified columns, representing complete ignorance regarding the nature of the cointegration relationships, is included in the set of matrix candidates. Secondly, some possible matrix candidates have been considered improbable, for example matrices implying that only one of the interest rates is stationary, which seems to be a plausible assumption.

The matrix candidates define submodels of the ECM. The usual approach is to determine, by hypothesis tests, the support given by the data to these submodels and then condition on the most plausible one in the subsequent analysis and prediction. At the very best, several models (matrix candidates) are entertained and different scenarios (e.g. predictions) are presented. In a Bayesian analysis we need only to assign priors to all uncertain aspects of the problem. Since there is uncertainty regarding the appropriate matrix candidate we simply assign prior probabilities $p(\boldsymbol{\beta}_1), ..., p(\boldsymbol{\beta}_Q)$ to them and then proceed to calculate the posterior distribution. As we will see in section 6, this leads to a procedure sometimes referred to as *Bayesian model averaging* (Draper, 1995) where each submodel contributes to the final inference with a weight proportional to the support in the data of this submodel. Bayesian model averaging allows us to treat each model separately. Thus we focus on the inference of one of the subset models and in section 6 we show that there is a simple way to combine the subset models into an overall inference tool.

By imposing fully specified vector candidates to some columns of $\boldsymbol{\beta}$, we are effectively fixing the corresponding restriction matrices, $\mathbf{h}_i$ and $\mathbf{H}^i$ for those columns. Of course, restricting whole columns of $\boldsymbol{\beta}$ to known numbers will produce over-identifying restrictions.

If a matrix candidate fully specifies $\boldsymbol{\beta}$ then a uniform prior for $\boldsymbol{\alpha}$ could be used as a default prior without much controversy. For matrix candidates with one or more unspecified columns, we have to assign priors to both $\boldsymbol{\alpha}$ and the remaining free parameters of $\boldsymbol{\beta}$. A uniform prior for all these parameters has been used by Geweke (1996) and Bauwens and Lubrano (1996). Another prior has been developed by Villani (1999b). To write down this prior some notation is needed first. Define a_{ii} as the ith diagonal element of $(\mathbf{H}_i'\mathbf{H}_i)^{-1}$ and let $\mathbf{C}_i$ equal $(\mathbf{H}_i'\mathbf{H}_i)^{-1}$ with the ith row and column deleted. Finally, let $C_q(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ denote the q-variate Cauchy with location $\boldsymbol{\mu}$ and scale (precision) matrix $\boldsymbol{\Lambda}$. The prior in Villani (1999b) is then of the form

$$\boldsymbol{\psi}_i \in C_{d_i}(\mathbf{0}, a_{ii}^{-1}\mathbf{C}_i),$$

with independence between the $\boldsymbol{\psi}_i$. The motivation of this prior is that it assigns equal probability to each possible r-dimensional space spanned by the columns of $\boldsymbol{\beta}$ (the cointegration spaces), which constitutes the distinguishable 'outcomes' of

$\boldsymbol{\beta}$. Villani (1999b) suggests that a uniform prior on $\boldsymbol{\alpha}$ can be used for convenience in the calculations.

There is ongoing research about the appropriate prior for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ and the final word has certainly not been said on this issue. We wish to concentrate on other aspects of the analysis here, however, and we note that in situations where the data are at least moderately informative relative to the prior then the choice of prior will matter little.

5. Numerical evaluation of the posterior distribution by Gibbs sampling

The posterior is analytically intractable and a numerical evaluation is called for. The popular Gibbs sampler provides an automatic numerical method for routine applications. We will only give an outline of the algorithm and we refer to Smith and Roberts (1993) for a more complete introduction. Briefly, the posterior distribution of the parameters in the cointegrated ECM does not belong to a standard family of distribution from which we can generate samples. The posterior distribution of each parameter matrix (i.e. one of $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\Gamma}$ and $\boldsymbol{\Phi}$) conditional on the other parameter matrices, the so called full conditional posterior, are all easy to sample from, however. The Gibbs sampler exploits this fact and produces a sample from the posterior distribution by iterating through the full conditional posteriors. Although this sample consists of dependent observations, estimates of functions of the parameters (e.g. forecasts) based on the Gibbs samples do converge to the right values.

To implement the Gibbs sampler, the full conditional posteriors of the cointegrated ECM are needed. The necessary results are derived in Geweke (1996) and Villani (1999b) for a uniform prior on the short run dynamics. The more general case of a normal prior for $\boldsymbol{\Gamma}$ gives the following full conditional posterior

$$\operatorname{vec} \boldsymbol{\Gamma} | \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \mathbf{x}^{(T)} \in \mathbf{N}_{p^2(k-1)}(\bar{\boldsymbol{\gamma}}, \boldsymbol{\Omega}^{-1}),$$

where $\mathbf{x}^{(T)}$ denotes data up to time T, $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1} \otimes \mathbf{Z}'\mathbf{Z} + \boldsymbol{\Psi}^{-1}$, $\bar{\boldsymbol{\gamma}} = \boldsymbol{\Omega}^{-1}(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{Z}'\mathbf{Z})\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\gamma}} = \operatorname{vec}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(\boldsymbol{\Delta}\mathbf{X} - \mathbf{X}_{-1}\boldsymbol{\beta}\boldsymbol{\alpha}' - \mathbf{D}\boldsymbol{\Phi})$. This follows from the treatment of the multivariate regression in Zellner (1971) and the details will not be presented here.

A complication arises in the full conditional posterior of $\boldsymbol{\Sigma}$, since $\operatorname{diag}(\boldsymbol{\Sigma})$ is not only present in the likelihood but also in the prior covariance matrix of $\boldsymbol{\Gamma}$. The most frequently used solution to this dilemma is to set $\sigma_j^2 = s_j^2$ in the prior covariance matrix of $\boldsymbol{\Gamma}$, where s_j^2 is the ith diagonal element of $\hat{\boldsymbol{\Sigma}}$, the ML estimate of $\boldsymbol{\Sigma}$.

6. Prediction

6.1. The predictive distribution

The predictive distribution of a set of future observations $\mathbf{x}_{T+1}, ..., \mathbf{x}_{T+h}$ conditional on data up to time T, denoted by $p(\mathbf{x}_{T+1}, ..., \mathbf{x}_{T+h}|\mathbf{x}^{(T)})$, is the single measure needed for a Bayesian treatment of the prediction problem. Note that the predictive distribution is *not* conditional on the parameters and the uncertainty in estimation is therefore fully accounted for. The full predictive distribution cannot be obtained analytically, but Thompson and Miller (1986) have proposed a double simulation procedure for univariate autoregressions which is easily extended to the multivariate case with parameters generated from the posterior distribution by Gibbs sampling. The plan is as follows, where $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Gamma}, \boldsymbol{\Phi}, \boldsymbol{\Sigma})$.

1. Use the Gibbs sampler to generate a sequence of n_1 observations from the posterior distribution of $\boldsymbol{\theta}$.

2. For each $\boldsymbol{\theta}$ in this sequence, simulate n_2 prediction paths from the following decomposition of the conditional predictive distribution

$$p(\mathbf{x}_{T+1}, ..., \mathbf{x}_{T+h}|\boldsymbol{\theta}, \mathbf{x}^{(T)}) = p(\mathbf{x}_{T+1}|\boldsymbol{\theta}, \mathbf{x}^{(T)}) \cdots p(\mathbf{x}_{T+h}|\boldsymbol{\theta}, \mathbf{x}^{(T+h-1)}).$$

   Thus, generate $\mathbf{x}_{T+1}$ from $p(\mathbf{x}_{T+1}|\boldsymbol{\theta}, \mathbf{x}^{(T)})$ and then continue to generate $\mathbf{x}_{T+2}$ from $p(\mathbf{x}_{T+2}|\boldsymbol{\theta}, \mathbf{x}^{(T+1)})$ by conditioning on data up to time T and the previously generated $\mathbf{x}_{T+1}$ and so on.

Note that $p(\mathbf{x}_{T+i}|\boldsymbol{\theta}, \mathbf{x}^{(T+i-1)})$ is a normal distribution for $i > 0$. The sample of $n_1 n_2$ prediction paths from $p(\mathbf{x}_{T+1}, ..., \mathbf{x}_{T+h}|\mathbf{x}^{(T)})$ can be used in a vast number of ways to gain a rich understanding of the future behavior of the processes. For example, we can focus on the joint predictive distribution over future time periods for a given variable or the joint predictive distribution of a subset of the variables in the system for a given future time period.

Even though the double simulation technique is manageable even in larger systems like the one we deal with in section 7, it is often sufficient to calculate the mean and covariance matrix of the marginal predictive distribution since the predictive distribution is often fairly close to normal. The mean and variance can be computed by simple Gibbs sampling without the need for double simulation.

6.2. The mean of the marginal predictive distribution

The mean of the marginal predictive distribution can be written

$$E(\mathbf{x}_{T+h}|\mathbf{x}^{(T)}) = E_{\boldsymbol{\theta}|\mathbf{x}^{(T)}}\left[E(\mathbf{x}_{T+h}|\boldsymbol{\theta}, \mathbf{x}^{(T)})\right], \tag{6.1}$$

where $E_{\boldsymbol{\theta}|\mathbf{x}^{(T)}}(\cdot)$ and $E(\cdot)$ denotes expectation with respect to posterior of $\boldsymbol{\theta}$ and the distribution of $\mathbf{x}_t$ conditional on $\boldsymbol{\theta}$, respectively. The representation in (6.1) is convenient since it is well-known from previous work (see, for example, Lütkepohl, 1991) that

$$E(\mathbf{x}_{T+h}|\boldsymbol{\theta}, \mathbf{x}^{(T)}) = \mathbf{J}\tilde{\boldsymbol{\Pi}}^h \mathbf{Y}_T + \mathbf{J}\sum_{i=0}^{h-1} \tilde{\boldsymbol{\Pi}}^i \boldsymbol{\Phi}\mathbf{d}_{T+h-i}, \tag{6.2}$$

where $\mathbf{Y}_T = (\mathbf{x}'_T, \mathbf{x}'_{T-1}...., \mathbf{x}'_{T-K+1})'$, $\mathbf{J} = (\mathbf{I}_p, \mathbf{0}, ..., \mathbf{0})$ and

$$\tilde{\boldsymbol{\Pi}} = \begin{bmatrix} \boldsymbol{\Pi}_1 & \boldsymbol{\Pi}_2 & \cdots & \boldsymbol{\Pi}_{K-1} & \boldsymbol{\Pi}_K \\ \mathbf{I}_p & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_p & & \mathbf{0} & \mathbf{0} \\ \vdots & & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I}_p & \mathbf{0} \end{bmatrix}.$$

The mean of the predictive distribution in (6.1) can be calculated numerically by a simple arithmetic average of the conditional expectation, $E(\mathbf{x}_{T+h}|\boldsymbol{\theta}, \mathbf{x}^{(T)})$, over the Gibbs samples. To calculate the conditional expectation in (6.2) the coefficients of the VAR are needed, which can be obtained from the ECM from the following relationships

$$\begin{aligned} \boldsymbol{\Pi}_1 &= \boldsymbol{\Gamma}_1 + \boldsymbol{\alpha}\boldsymbol{\beta}' + \mathbf{I}_p \\ \boldsymbol{\Pi}_j &= \boldsymbol{\Gamma}_j - \boldsymbol{\Gamma}_{j-1}, \ j = 2, ..., K-1 \\ \boldsymbol{\Pi}_K &= -\boldsymbol{\Gamma}_{K-1}. \end{aligned}$$

6.3. The variance of the marginal predictive distribution

Using the well-known conditional variance formula, the covariance matrix of the predictive distribution can be expressed as

$$V(\mathbf{x}_{T+h}|\mathbf{x}^{(T)}) = E_{\boldsymbol{\theta}|\mathbf{x}^{(T)}}\left[V(\mathbf{x}_{T+h}|\boldsymbol{\theta}, \mathbf{x}^{(T)})\right] + V_{\boldsymbol{\theta}|\mathbf{x}^{(T)}}\left[E(\mathbf{x}_{T+h}|\boldsymbol{\theta}, \mathbf{x}^{(T)})\right] \tag{6.3}$$

The first term in (6.3) can be estimated as an arithmetic average of

$$V(\mathbf{x}_{T+h}|\boldsymbol{\theta}, \mathbf{x}^{(T)}) = \sum_{i=0}^{h-1} \boldsymbol{\Theta}_i \boldsymbol{\Sigma} \boldsymbol{\Theta}'_i,$$

where $\boldsymbol{\Theta}_i = \mathbf{J}\tilde{\boldsymbol{\Pi}}^i\mathbf{J}'$, over the Gibbs draws from the posterior. The second term in (6.3) is simply the posterior variance of the conditional expectation and can

therefore be estimated by the sample variance of the conditional expectations in each Gibbs draw.

The most common classical approach is to consider only the first term in (6.3) with the posterior distribution replaced by the Dirac delta function at the ML estimate of $\boldsymbol{\theta}$. This only takes the future disturbances into account when forming prediction intervals and is therefore called the *error based* prediction variance. A likelihood based approach to account for the uncertainty regarding the unknown parameters of the model, which in a Bayesian setting corresponds (approximately) to the second term in (6.3), is only available by using an asymptotic approximation, an approximation which is often very crude, especially if unit roots are present in the system (Lütkepohl, 1991; Doornik and Hendry, 1995). Intervals based on this approximation will here be termed *asymptotic* prediction intervals.

6.4. Predicting the first difference of the process

Our main concern in the analysis of the Swedish monetary data described in section 3 is the prediction of the future path of inflation; we are thus interested in predicting Δp_t rather than the level p_t. In general, suppose we are interested in predicting $\Delta \mathbf{x}_{T+h}$ given data up to time T. The mean of the predictive distribution of $\Delta \mathbf{x}_{T+h}$ is, of course, simply $E(\Delta \mathbf{x}_{T+h}|\mathbf{x}^{(T)}) = E(\mathbf{x}_{T+h}|\mathbf{x}^{(T)}) - E(\mathbf{x}_{T+h-1}|\mathbf{x}^{(T)})$. The variance of this distribution is still given by the conditional variance formula (6.3) with $\mathbf{x}_{T+h}$ replaced by $\Delta \mathbf{x}_{T+h}$. Everything remains unchanged except that we need an expression for

$$V(\Delta \mathbf{x}_{T+h}|\mathbf{x}^{(T)}, \boldsymbol{\theta}) = V(\mathbf{x}_{T+h}|\mathbf{x}^{(T)}, \boldsymbol{\theta}) + V(\mathbf{x}_{T+h-1}|\mathbf{x}^{(T)}, \boldsymbol{\theta}) - 2C(\mathbf{x}_{T+h}, \mathbf{x}_{T+h-1}|\mathbf{x}^{(T)}, \boldsymbol{\theta}), \tag{6.4}$$

which requires $C(\mathbf{x}_{T+h}, \mathbf{x}_{T+h-1}|\mathbf{x}^{(T)}, \boldsymbol{\theta})$, the covariance between $\mathbf{x}_{T+h}$ and $\mathbf{x}_{T+h-1}$, to be calculated. From Lütkepohl (1991, p. 32) we know that

$$V(\mathbf{Y}_{T+h}|\mathbf{Y}^{(T)}, \boldsymbol{\theta}) = \sum_{i=0}^{h-1} \tilde{\boldsymbol{\Pi}}^i \mathbf{J}' \boldsymbol{\Sigma} \mathbf{J} \tilde{\boldsymbol{\Pi}}^{i\prime},$$

where, as before, $\mathbf{Y}_{T+h} = (\mathbf{x}'_{T+h}, \mathbf{x}'_{T+h-1}...., \mathbf{x}'_{T+h-K+1})'$. The components needed for $V(\Delta \mathbf{x}_{T+h}|\mathbf{x}^{(T)}, \boldsymbol{\theta})$ in (6.4) are given by the $2p \times 2p$ submatrix in the upper left corner of $V(\mathbf{Y}_{T+h}|\mathbf{Y}^{(T)}, \boldsymbol{\theta})$.

Note the special case $h = 1$, then $V(\Delta \mathbf{x}_{T+1}|\mathbf{x}^{(T)}, \boldsymbol{\theta}) = V(\mathbf{x}_{T+1}|\mathbf{x}^{(T)}, \boldsymbol{\theta}) = \boldsymbol{\Sigma}$, since $\mathbf{x}_T$ is known.

6.5. Accounting for the uncertainty of the appropriate matrix candidate

The discussion so far in this section has implicitly been conditional on a given matrix candidate for $\boldsymbol{\beta}$ and it would therefore have been more appropriate with

the notation $p(\mathbf{x}_{T+h}|\mathbf{x}^{(T)}, \boldsymbol{\beta}_i)$, where $\boldsymbol{\beta}_i$ is the ith matrix candidate. The final goal is, of course, the overall unconditional predictive distribution of $\mathbf{x}_{T+h}$

$$p(\mathbf{x}_{T+h}|\mathbf{x}^{(T)}) = \sum_{i=1}^{Q} p(\boldsymbol{\beta}_i|\mathbf{x}^{(T)}) p(\mathbf{x}_{T+h}|\mathbf{x}^{(T)}, \boldsymbol{\beta}_i), \tag{6.5}$$

where $p(\boldsymbol{\beta}_i|\mathbf{x}^{(T)})$ is the posterior probability that $\boldsymbol{\beta} = \boldsymbol{\beta}_i$. From Bayes theorem,

$$p(\boldsymbol{\beta}_i|\mathbf{x}^{(T)}) \propto p(\boldsymbol{\beta}_i) m_i(\mathbf{x}^{(T)}), \tag{6.6}$$

where

$$m_i(\mathbf{x}^{(T)}) = \int p(\mathbf{x}^{(T)}|\boldsymbol{\theta}_i) p(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i,$$

is the *marginal likelihood* of the ith submodel, where $\boldsymbol{\theta}_i = (\boldsymbol{\alpha}, \boldsymbol{\psi}^{(i)}, \boldsymbol{\Gamma}, \boldsymbol{\Phi}, \boldsymbol{\Sigma})$, $\boldsymbol{\psi}^{(i)}$ contains the free parameters of $\boldsymbol{\beta}$ under model i, $p(\mathbf{x}|\boldsymbol{\theta})$ denotes the likelihood function and $p(\boldsymbol{\theta})$ the prior. The proportionality constant in (6.6) is determined by the condition $\sum_{i=1}^{Q} p(\boldsymbol{\beta}_i|\mathbf{x}^{(T)}) = 1$. Thus, a Bayesian gives credence to model (matrix candidate) i, for a given set of $p(\boldsymbol{\beta}_i)$, if the expected (with respect to the prior) likelihood of the data under model i is large.

The exact calculation of $m_i(\mathbf{x}^{(T)})$ can be performed numerically by, for example, importance sampling, see Villani (1999c). We are mainly interested in weights to form an overall prediction, however, and the following approximation will often be precise enough for that purpose (Draper, 1995)

$$\ln m_i(\mathbf{x}^{(T)}) \approx \frac{k_i}{2} \ln 2\pi + \ln p(\mathbf{x}|\hat{\boldsymbol{\theta}}_i) - \frac{k_i}{2} \ln T, \tag{6.7}$$

where k_i is the number of free parameters in the ith model and $p(\mathbf{x}|\hat{\boldsymbol{\theta}}_i)$ is the maximum of the likelihood function under the ith model.

A sample of n observations from the overall predictive distribution, $p(\mathbf{x}_{T+h}|\mathbf{x}^{(T)})$, can be obtained by generating $n \cdot p(\boldsymbol{\beta}_i|\mathbf{x}^{(T)})$ observations from $p(\mathbf{x}_{T+h}|\mathbf{x}^{(T)}, \boldsymbol{\beta}_i)$, for $i = 1, ..., Q$, via the algorithm in section 6.1.

The mean of the overall predictive distribution is

$$E(\mathbf{x}_{T+h}|\mathbf{x}^{(T)}) = \sum_{i=1}^{Q} p(\boldsymbol{\beta}_i|\mathbf{x}^{(T)}) E(\mathbf{x}_{T+h}|\mathbf{x}^{(T)}, \boldsymbol{\beta}_i),$$

and the variance is

$$\begin{aligned} V(\mathbf{x}_{T+h}|\mathbf{x}^{(T)}) &= \sum_{i=1}^{Q} p(\boldsymbol{\beta}_i|\mathbf{x}^{(T)}) V(\mathbf{x}_{T+h}|\mathbf{x}^{(T)}, \boldsymbol{\beta}_i) \\ &\quad + \sum_{i=1}^{Q} p(\boldsymbol{\beta}_i|\mathbf{x}^{(T)}) \left[E(\mathbf{x}_{T+h}|\mathbf{x}^{(T)}, \boldsymbol{\beta}_i) - E(\mathbf{x}_{T+h}|\mathbf{x}^{(T)}) \right]^2, \end{aligned} \tag{6.8}$$

which follows directly from the conditional variance formula. It should be noted, however, that even if each $p(\mathbf{x}_{T+h}|\mathbf{x}^{(T)}, \boldsymbol{\beta}_i)$ is normal then $p(\mathbf{x}_{T+h}|\mathbf{x}^{(T)})$ is a discrete mixture of normal distributions, which may be badly summarized by its first two moments.

The usual approach is to choose one of the matrix candidates, say $\boldsymbol{\beta}_i$, and then condition on this choice in the prediction phase. It should be clear, however, that unless $p(\mathbf{x}_{T+h}|\mathbf{x}^{(T)}, \boldsymbol{\beta}_i) = p(\mathbf{x}_{T+h}|\mathbf{x}^{(T)})$ then this conditioning is producing distorted prediction statements. If $V(\mathbf{x}_{T+h}|\mathbf{x}^{(T)}, \boldsymbol{\beta}_i)$ are approximately equal for all $\boldsymbol{\beta}_i$ with posterior weights significantly different from zero, then the second sum in (6.8) measures the uncertainty that is neglected by conditioning on only one of the matrix candidates (Draper, 1995).

The uncertainty regarding the lag length and the cointegration rank can be accounted for in the predictive distribution in exactly the same way by simply averaging with respect to the posterior distribution of K and r. As seen in section 3, however, the posterior distribution of K is very concentrated over $K = 4$ and little is probably lost by conditioning on four lags in the VAR. Judging from the inconclusive evidence regarding the cointegration rank in JJVW, there is probably something to be gained from weighing the predictive distribution with respect to this posterior, but unfortunately, a reliable procedure for calculating this posterior distribution does not seem to exist yet.

7. Empirical example

The Swedish monetary data described in section 3 and, more fully, in JJVW will be used to illustrate the Bayesian procedure introduced in this paper. The focus will be on forecasting the Swedish inflation and the 12 last quarters of the data set (1993:4 to 1996:4) will be reserved (i.e. excluded in the estimation phase) for evaluation of the inflation forecasts. The data will be analyzed conditional on $r = 3$ and $K = 4$, for reasons explained in section 3.

The posterior weights of the six possible matrix candidates were computed using the approximation in (6.7). The results are shown in Table 7.1 which presents the six candidates, the log-likelihood, the p-value of the likelihood ratio test and the approximate posterior weights under the uniform prior, $p(\boldsymbol{\beta}_i) = 1/6$ for $i = 1, ..., 6$.

The sad conclusion from Table 7.1, one that is also reached in JJVW, is that the data give almost all support to the unrestricted matrix $\boldsymbol{\beta}_6$; that is, none of the theoretically motivated long run equilibria are supported by data. Since only the unrestricted candidate would matter significantly in an overall analysis (see equation 6.5), there is no need to produce predictions based on any other matrix candidate. JJVW nevertheless continue to use both $\boldsymbol{\beta}_5$ and $\boldsymbol{\beta}_6$ in the forecasting

Matrix candidate	Loglike	p-value	$p(\boldsymbol{\beta}_i \mid \mathbf{x}^{(T)})$
$\boldsymbol{\beta}_1 = (\mathbf{b}_1, \cdot, \cdot)$	242.26	$1.6 \cdot 10^{-3}$	0.037
$\boldsymbol{\beta}_2 = (\mathbf{b}_4, \cdot, \cdot)$	237.55	$2.1 \cdot 10^{-5}$	$3.2 \cdot 10^{-4}$
$\boldsymbol{\beta}_3 = (\mathbf{b}_2, \mathbf{b}_3, \cdot)$	217.88	$3.0 \cdot 10^{-11}$	$2.2 \cdot 10^{-10}$
$\boldsymbol{\beta}_4 = (\mathbf{b}_1, \mathbf{b}_4, \cdot)$	226.33	$5.4 \cdot 10^{-8}$	$1.0 \cdot 10^{-6}$
$\boldsymbol{\beta}_5 = (\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3)$	195.82	$5.1 \cdot 10^{-18}$	$1.3 \cdot 10^{-17}$
$\boldsymbol{\beta}_6 = (\cdot, \cdot, \cdot)$	250.96	–	0.963

Table 7.1: Restrictions on $\boldsymbol{\beta}$

stage under the motivation that the model with $\boldsymbol{\beta}_5$ is theoretically reasonable and produces good forecasts. Stark (1998) observes the same feature in his forecasting model. We will also analyze both $\boldsymbol{\beta}_5$ and $\boldsymbol{\beta}_6$ for comparability reasons, but it should be remembered that a strict Bayesian analysis would only analyze the model with $\boldsymbol{\beta} = \boldsymbol{\beta}_5$ if its posterior weight is meaningfully different from zero and such a posterior weight requires an extremely, and probably unreasonably, large prior probability on $\boldsymbol{\beta}_5$ (in Table 7.1, 1/6 was used). We adopt the same terminology as JJVW and call $\boldsymbol{\beta}_5$ and $\boldsymbol{\beta}_6$ the *theoretical* and *empirical* cointegration vectors, respectively.

Our next task is to determine appropriate values for the two parameters λ and θ in the prior for the short-run dynamics. To pin-point exact values for these parameters may be difficult even for an experienced analyst and it would be desirable if relatively small changes in λ and θ did not lead to large changes in the prediction paths. To check this, the square root of the mean squared error (RMSE) of the dynamic (see below) predictions was calculated for a range of values for both λ and θ. The results are presented in Table 7.2. It is evident that RMSE is very insensitive to changes in θ and, with the exception of $\lambda = 0.1$, the RMSE changes also very little when λ varies. Like the RMSE's, the prediction paths did not change significantly when we changed λ and θ, the exception again being $\lambda = 0.1$ which produced a somewhat different prediction path (results available from the author by request). Commonly used values for λ ranges from 0.1 to 0.4 and around 0.2 for θ, see Lütkepohl (1991), Litterman (1986), Doan et al. (1992) and Karlsson and Kadiyala (1996). Since the exact choice λ and θ is of lesser importance, it seems to be safe to use $\lambda = 0.3$ and $\theta = 0.2$ in our application.

Within the two models based on empirical and theoretical cointegration vectors, JJVW produced inflation predictions using both unrestricted estimation of the short run dynamics and from a restricted version of the models where a sequence of tests was used to set insignificant coefficients in $\boldsymbol{\Gamma}$ to zero. This is of

$\lambda\backslash\theta$	0.1	0.2	0.3	0.4	0.5
0.1	0.725	0.748	0.736	0.733	0.751
0.2	0.604	0.594	0.549	0.575	0.542
0.3	0.562	0.512	0.527	0.518	0.492
0.4	0.545	0.506	0.514	0.507	0.520
0.5	0.511	0.496	0.512	0.519	0.562

Table 7.2: Root Mean Squared Errors (RMSE) of the forecasts from 1994:1 to 1996:4 for the empirical cointegration vectors for different priors on the short run dynamics.

	1994-96		1997-98		
Forecasting model	Dynamic	Recursive	Dynamic	Recursive	Average
Bayes, theo.	0.415	0.422	0.598	0.503	0.485
Bayes uniform, emp.	0.512	0.405	0.499	0.535	0.488
ML restr., emp.	0.509	0.440	0.509	0.669	0.532
Bayes Cauchy, emp.	0.670	0.403	0.459	0.629	0.540
Random walk	0.645	0.578	0.530	0.663	0.604
ML restr., theo.	0.535	0.486	1.149	0.531	0.675
ML unrestr., emp.	0.836	0.481	0.553	0.905	0.694
ML unrest., theo.	0.703	0.499	1.237	0.653	0.773

Table 7.3: Root Mean Squared Error (RMSE) of the predictions for the quarterly Swedish inflation during 1994-1998.

course the usual classical way to circumvent the explicit use of a shrinkage prior, like the one suggested here, and reflects the widespread belief that unrestricted VAR models are often overparametrized. Not only is the classical significance testing much cruder than the Bayesian shrinkage approach, but, to the best of our knowledge, in the current state of classical (ML) cointegration estimation, $\boldsymbol{\beta}$ must be assumed known in order for a reestimation of the model after exclusion restrictions to be possible (see, for example, PcFiml in Doornik and Hendry, 1997). Thus, $\boldsymbol{\beta}$ cannot be reestimated after the imposition of restrictions and the assumption that $\boldsymbol{\beta}$ is known will make the prediction intervals too narrow. This is in contrast to our Bayesian approach where the simplification of the model by shrinkage priors and the estimation of cointegration relations are handled simultaneously.

Both dynamic (*ex ante*) and recursive 1-step ahead inflation forecasts were calculated. The dynamic forecasts for periods $t+1$ to $t+h$ only use data for estimation of model parameters up to time t, while the recursive forecasts are

based on an updating procedure where the model parameters are reestimated in each time period using the maximum sample length available prior to forecasting.

Initial experimentation with the double simulation procedure described in section 6.1 produced predictive distributions of Δp_t which were very close to normal. Thus, in the following only the mean and the variance of the distributions will be reported. All reported predictions are based on 100000 iterations of the Gibbs sampling algorithm, although the convergence of the predictions to their right values was obtained already after approximately 50000 iterations.

The evaluation data for the period 1994:1 to 1996:4 have been consumed to some extent in the determination of the prior parameters λ and θ in the Bayesian approach and perhaps even more in the significance testing for the restricted models. The data set has therefore been updated by eight more quarters to include 1998:4 as the last data period. By using exactly the same models as before (but using data up to 1996:4 for estimation of parameters in the dynamic forecasts and data up to 1996:4, 1997:1,... for the recursive forecasts) a pure evaluation period from 1997:1 to 1998:4 is obtained.

The prediction performance of eight forecasting models during the two evaluations periods are displayed in Table 7.3. The last column of this table is an average over the four preceding columns and represents overall performance. The use of an unweighted average, which gives more weight to the eight last observations, is motivated since the second evaluation period is more 'pure' than the first. In Table 7.3, the models are ranked from the best to the worst with respect to overall RMSE. Thus, the three Bayesian procedures are ranked first, second and fourth best.

The prediction paths and their uncertainty are displayed in Figure 7.1 (dynamic forecasts) and 7.2 (recursive forecasts). Each subgraph in these figures shows the actual inflation and the prediction path with corresponding 50% and 90% prediction intervals. The two evaluation periods are separated by a dotted line. Note that the four ML-based predictions have three prediction bands. The most narrow interval is the 50% error based (see section 6.3) interval which ignores the uncertainty in the estimation of the model parameters. The second most narrow interval is the 90% error based interval while the largest interval is the 90% *asymptotic* (see section 6.3) interval which (approximately) accounts for parameter uncertainty.

The striking feature of Figure 7.1 and 7.2 is the smoothness of the Bayesian prediction paths compared to their ML counterparts. Bayesian predictions from a diffuse prior on the short run dynamics (not shown here) gave essentially the same shaky prediction paths as the ML predictions, so the smoothness of the Bayesian predictions is caused by the shrinkage prior on the short run parameters. It is our opinion that the difficulty to produce forecasts that closely tracks the inflation

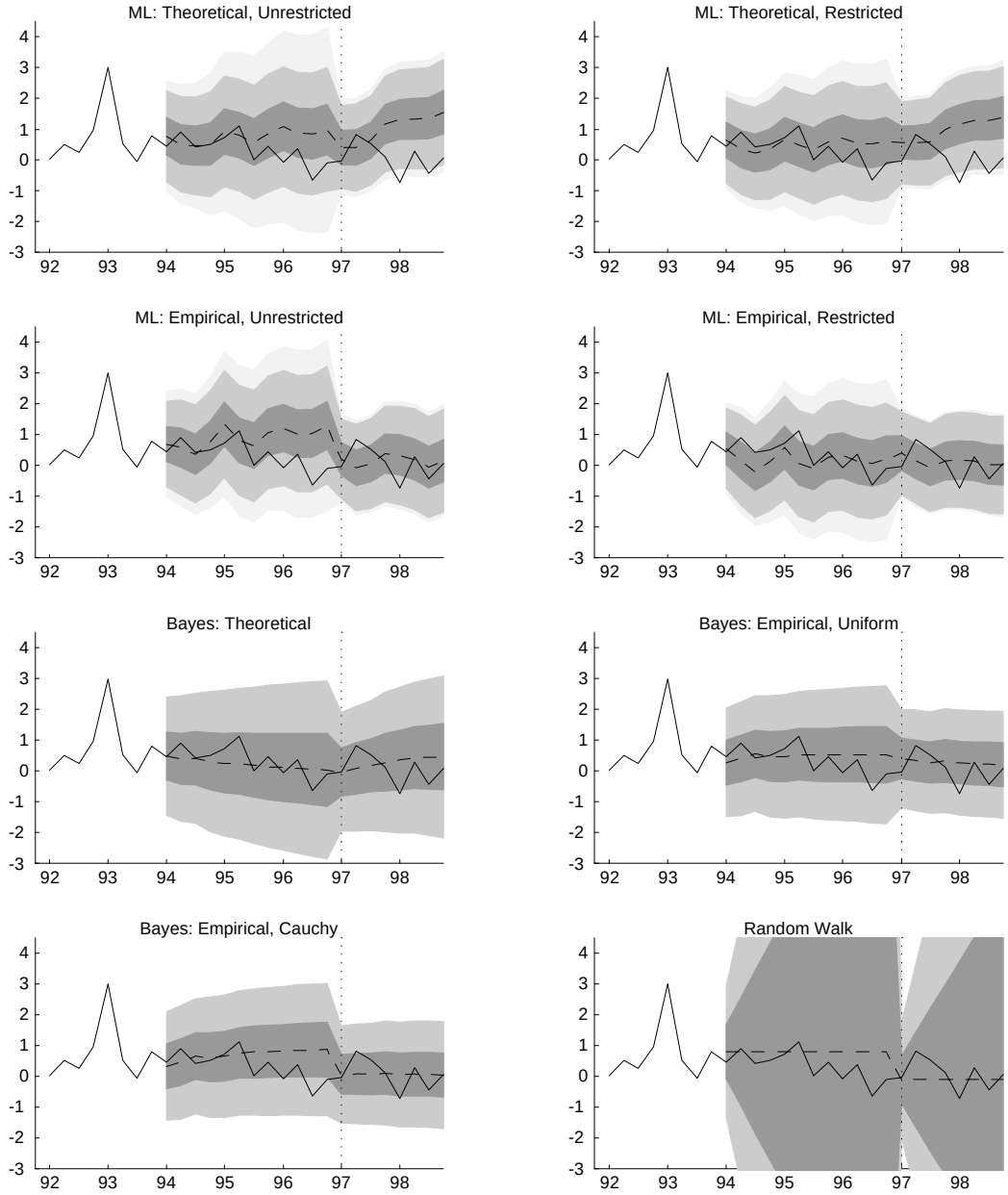

Figure 7.1: Dynamic predictions of Swedish inflation (——) 1994-98. Point prediction (- - -) with 50% (dark band) and 90% (lighter band) prediction intervals. Two different 90% bands of the ML predictions are displayed, one ignoring parameter uncertainty (medium light band) and another partially accounting for it (lightest band). The 50% band for the ML predictions ignores parameter uncertainty. The prediction bands for the random walk increase linearly and have been cut to fit in the graph.

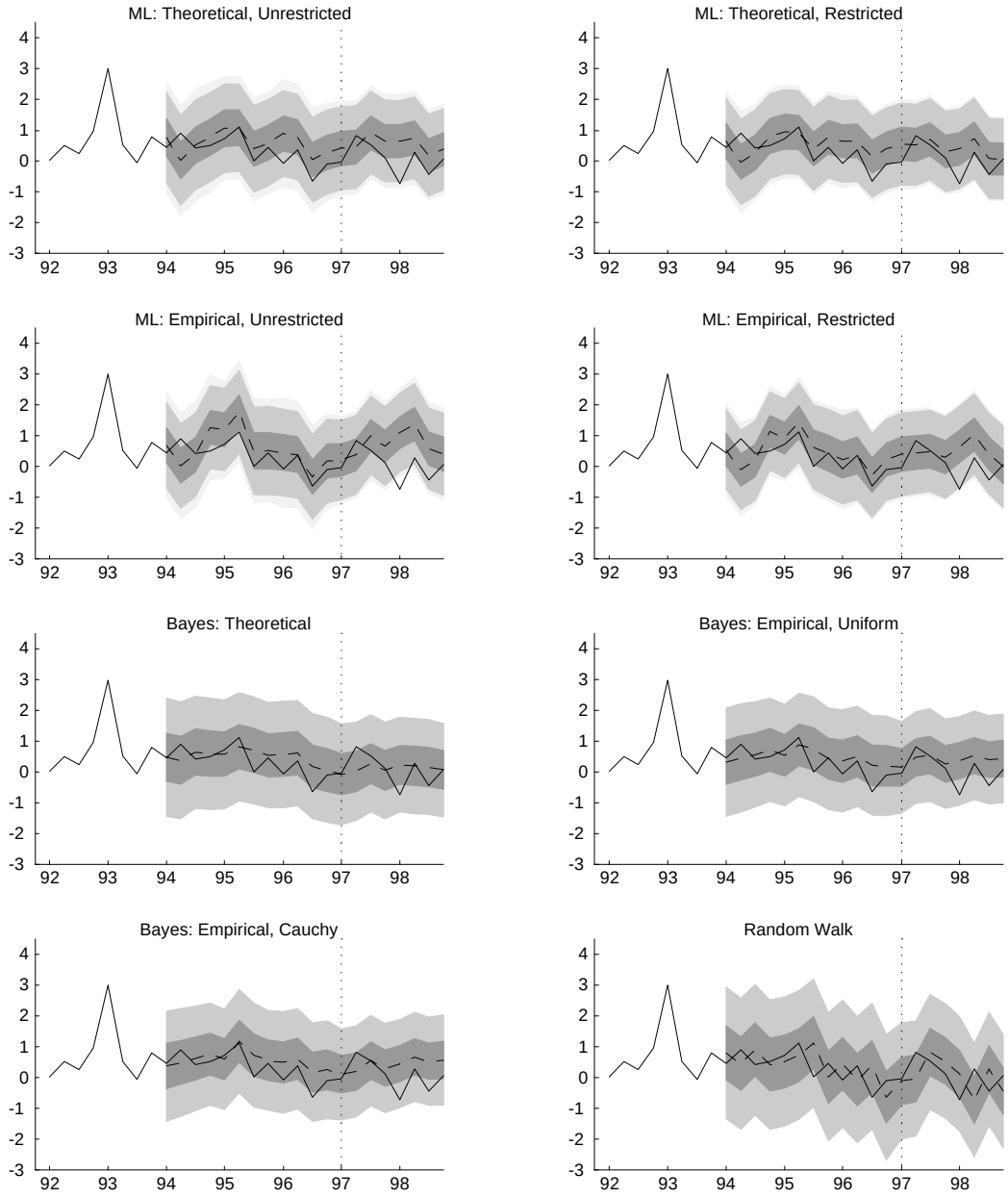

Figure 7.2: Recursive predictions of Swedish inflation (——) 1994-98. Point prediction (- - -) with 50% (dark band) and 90% (lighter band) prediction intervals. Two different 90% bands of the ML predictions are displayed, one ignoring parameter uncertainty (medium light band) and another partially accounting for it (lightest band). The 50% band for the ML predictions ignores parameter uncertainty.

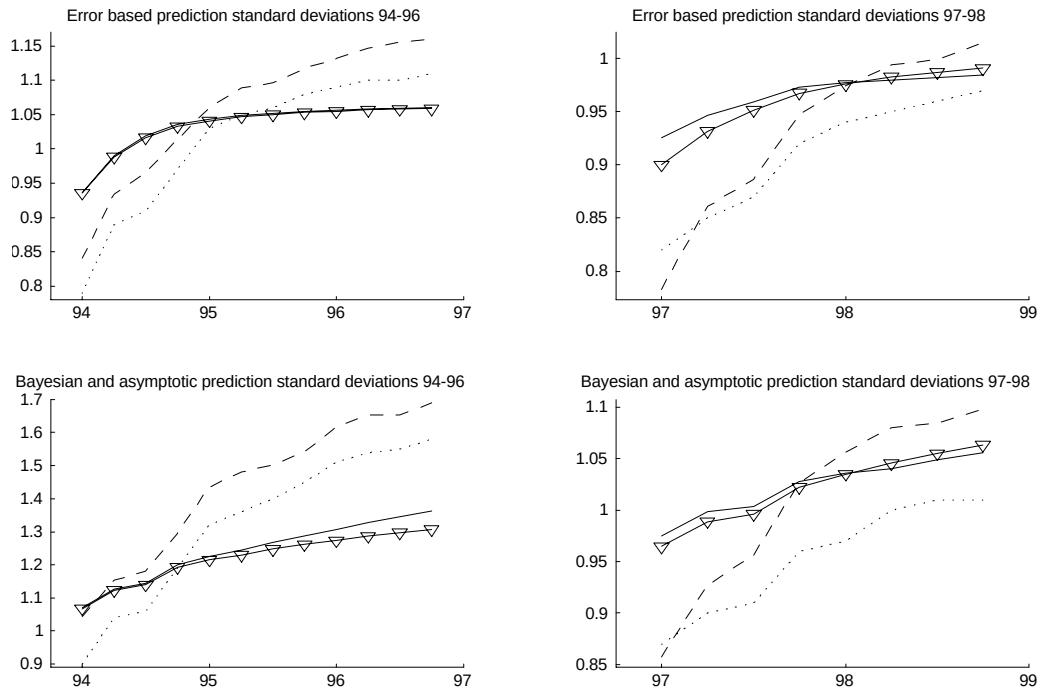

Figure 7.3: Standard deviation of the dynamic predictions based on the empirical cointegration vectors. ML unrestricted (- - -), ML restricted ($\cdot\cdot\cdot$), Bayes Cauchy ($-\triangledown-$) and Bayes uniform (——).

(see ML methods, especially dynamic predictions) makes the smoother Bayesian predictions more convincing.

From Table 7.3 and Figure 7.1 and 7.2, it is apparent that the random walk does a rather good job in predicting future inflation. This is hardly surprising given that the inflation was rather constant during the evaluation periods. Even if the no-change forecast implied by the random walk turns out to be good prediction rule, the intervals of the dynamic predictions in figure 7.1 (which increase linearly and have been cut to fit the graph) clearly shows that the random walk is not a good forecasting model for inflation, however; the prediction intervals are simply too large for any reasonable forecaster.

In accordance with the conclusions in JJVW and Stark (1998), the point predictions from the models with the, unsupported, theoretical cointegration vectors do seem to be of about equal quality as the point predictions from the models based on the empirical cointegration vectors. The prediction intervals are larger for the theoretical cointegration vectors, however, caused by an inferior statistical fit to the data.

Equally important as the point predictions themselves are their standard deviations. Figure 7.3 displays the standard deviations of the predictions (both error based and asymptotic) from the models with empirical cointegration vectors as a function of the forecasting horizon. The Bayesian error based standard deviation has been defined as the square root of the first term in (6.3). It should be noted that the Bayesian error based standard deviation is only presented for the purpose of comparison with the ML methods error based standard deviation; there is no reason to ignore parameter uncertainty in a Bayesian approach. The standard deviations from the two Bayesian methods are close to each other while there are large differences between the unrestricted and restricted ML method. The Bayesian standard deviations are always larger than for the unrestricted ML method for short horizons, but then becomes smaller for larger horizons. The same is true for the restricted ML method during the first evaluation period while in the second period, it produces smaller standard deviation for all horizons.

8. Conclusions

The Bayesian approach to prediction in the cointegrated vector autoregression presented here offers several advantages over traditional approaches. Firstly, whole predictive densities for the future observations can be obtained and these densities can have any distributional form. Secondly, the predictive density affords a straight forward *probability* interpretation, which seems to be the way people *de facto*, but falsely, interpret frequentist prediction intervals. Thirdly, predictive distributions are not conditional on the parameters of the model. Fourthly, the shrinkage prior on the short run dynamics avoids the dubious sequential testing procedures which are often used to simplify the model by exclusion restrictions. Fifthly, the uncertainty about the long run restrictions, the number of lags in the model and the cointegration rank is reflected in the predictive distribution and intuitive weighting formulas can be used to average the predictive distributions obtained by conditioning on the unknown quantities. As expected, these advantages come at the cost of an increased computational burden. This disadvantage is becoming less and less important with each innovation in computing technology, however, and our empirical example demonstrates that the proposed procedure is manageable even in large systems with many variables and lags.

References

[1] Bauwens, L. and Lubrano, M. (1996). Identification restrictions and posterior densities in cointegrated Gaussian VAR systems. In *Advances in Econometrics*, Volume **11**, Part B, JAI Press, 3-28.

[2] Doan, T. A., Litterman, R. B. and Sims, C. A. (1984). Forecasting and conditional projection using realistic prior distributions. *Econometric Reviews*, **3**, 1-144.

[3] Doan, T. A. (1992). *RATS User's Manual Version 4*, Evanston: Estima.

[4] Doornik, J. A. and Hendry, D. F. (1997). *Modelling Dynamic Systems Using PcFiml 9.0 for Windows*, London: International Thomson Business Press.

[5] Dickey, J. M. (1967). Matric-variate generalizations of the multivariate t distribution and the inverted multivariate t distribution. *Ann. Math. Statist.*, **38**, 511-18.

[6] Draper, D. (1995). Assessment and propagation of model uncertainty. *J. R. Statist. Soc.* **B57**, 45-97.

[7] Engle, R. F. and Granger, C. W. J. (1987). Co-integration and error correction: Representation, estimation and testing. *Econometrica*, **55**, 251-76.

[8] Geisser, S. (1965). A Bayes approach for combining correlated estimates. *J. Amer. Statist. Assoc.*, **60**, 602-607.

[9] Geweke, J. (1996). Bayesian Reduced Rank Regression in Econometrics. *Journal of Econometrics*, **75**, 121-146.

[10] Hansen, H. and Juselius, K. (1995). *CATS in RATS - Cointegration Analysis of Time Series*, Evanston: Estima.

[11] Jacobson, T., Jansson, P., Vredin, A. and Warne, A. (1999). Monetary policy analysis and inflation targeting in a small open economy: A VAR approach, *Sveriges Riksbank Working Paper Series* No. **77**.

[12] Johansen, S. (1995a). *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models*. Oxford: Oxford University Press.

[13] Johansen, S. (1995b). Identifying restrictions on linear equations: with applications to simultaneous equations and cointegration, *Journal of econometrics*, **69**, 111-32.

[14] Kleibergen, F. and van Dijk, H. K. (1994). On the shape of the likelihood/posterior in cointegration models, *Econometric Theory*. **10**, 514-51.

[15] Litterman, R. B. (1980). A Bayesian procedure for forecasting with vector autoregressions, mimeo, Massachusetts Institute of Technology.

[16] Litterman, R. B. (1986). Forecasting with Bayesian vector autoregressions - Five years of experience, Journal of Business and Economic Statistics. **4**, 25-38.

[17] Lütkepohl H. (1991) *Introduction to Multiple Time Series Analysis*. New York: Springer-Verlag.

[18] Smith, A. F. M. and Roberts, G. O. (1993). Bayesian Computation via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods (with discussion). *J. Roy. Statist. Soc.* B, **55**, 3-23.

[19] Stark, T. (1998). A Bayesian vector error corrections model of the U.S. economy, Working Paper, 98-12, Economic Research Division, Federal Reserve Bank of Philadelphia.

[20] Stock, J. H. and Watson, M. W. (1988). Testing for common trends, *J. Amer. Statist. Assoc.*, **83**, 1097-107.

[21] Thomson, P. A. and Miller, R. B. (1986). Sampling the future: A Bayesian approach to forecasting from univariate models, *Journal of Business and Economic Statistics*. **4**, 427-36.

[22] Villani, M. (1999a). Fractional Bayesian lag length inference in multivariate autoregressive processes, *Journal of Time Series Analysis*, under revision.

[23] Villani, M. (1999b). Reference priors and conditional posteriors for cointegration models, Unpublished manuscript.

[24] Villani, M. (1999c). A Bayesian approach to restrictions on the cointegration space, Unpublished manuscript.

[25] Warne, A. (1993). A common trends model: Identification, Estimation and Inference. Seminar paper No. **555**, Institute for International Economic Studies, Stockholm University.

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