Interweaving Strategy to Improve the Mixing Properties of MCMC Estimation for Time-Varying Parameter Models∗

Stef De Visscher1,2 and Gerdie Everaert1

1Ghent University, Faculty of Economics and Business Administration, Sint-Pietersplein 6, B-9000 Gent, Belgium.
2Research Foundation – Flanders (FWO)

October 30, 2017

Abstract

Time-varying parameter models have become increasingly popular in macroeconometric research. Their flexibility, however, entails the risk of over-parameterization, such that model specification, through deciding whether parameters are fixed or vary over time, may be of key importance. Frühwirth-Schnatter and Wagner (2010) show how to extend Bayesian variable selection in a standard regression framework to the state space models required to estimate time-varying parameters. A crucial aspect of their MCMC approach is the use of a non-centered reparameterization of the model. In this paper we look at the implications of this alternative representation on the mixing properties of the Markov chain. We first document that convergence is adversely affected when the variance of the innovations to the time-varying parameters grows large relative to that of the error terms in the observation equation. We next show that the MCMC mixing efficiency can be boosted considerably by interweaving the non-centered with the original centered parameterization, while still being able to carry out state space model selection. Our simulations further show that interweaving often even improves on the best performer out of the original centered and non-centered parameterization such that our strategy may also be fruitful outside a model selection framework.

Keywords: time-varying parameters, model selection, MCMC, mixing properties, interweaving

∗Correspondence: Stef De Visscher, Faculty of Economics and Business Administration, Sint-Pietersplein 6, B-9000 Gent, Belgium. Email: stef.devisscher@ugent.be
1 Introduction

In recent years, a lot of macroeconometric research has focused on models with time-varying parameters (TVP) to account for and study structural changes in the relationships between economic variables. Among many examples are dynamic interactions between the macroeconomy and the yield curve (Diebold et al., 2006), changes in the data generating process of inflation (Stock and Watson, 2007), shifts in the monetary transmission mechanism (Koop et al., 2009) and variations in Okun’s Law coefficient over time (Huang and Lin, 2008). TVPs have also been introduced in vector autoregressions (see e.g. Primiceri, 2005; Baumeister and Peersman, 2013). Although TVP models often lead to better forecasting (see e.g. Groen et al., 2013, for inflation forecasting), their flexibility entails the risk of over-parameterization. Hence, model specification, through deciding which components to include and whether they are fixed or vary over time, is of key importance. Moreover, in many applications the question whether a parameter is stable or not is an explicit test of whether restrictions implied by economic theory hold (see e.g. Koop et al., 2010; Everaert et al., 2017).

TVP models are typically cast into a state space representation. Let $y_t$ denote a scalar dependent variable and $X_t$ a $(1 \times m)$ vector of explanatory variables, then a general state space representation of the TVP model is given by

$$
\begin{align*}
    y_t &= X_t \beta_t + \epsilon_t, \\
    B(L) \beta_t &= \omega_t,
\end{align*}
$$

for $t = 1, \ldots, T$ and where $\beta_t$ is an $(m \times 1)$ state vector including the TVPs, $B(L)$ an $(m \times m)$ lag polynomial and $\Theta$ an $(m \times m)$ covariance matrix. Analyzing whether the parameters in $\beta_t$ truly vary over time implies testing whether the diagonal elements of $\Theta$ are zero against the alternative that they are positive. From a classical statistical point of view, such a test is non-regular as the null hypothesis of a zero variance lies on the boundary of the parameter space. A Bayesian approach is, in general, better suited to deal with this testing problem. The modern approach to Bayesian model selection is to jointly sample model indicators and parameters using Markov Chain Monte Carlo (MCMC) methods. While initially designed for identifying non-zero effects in standard regression models, Frühwirth-Schnatter and Wagner (2010) (henceforth FS-W) generalize this stochastic variable selection approach to state space models. Their MCMC algorithm is innovative in two respects. First, they replace the standard inverse gamma ($\text{IG}$) prior for the state innovation variances in $\Theta$ by a normal ($\mathcal{N}$) distribution centered at zero for the square root of these variances. It is well known that the hyperparameters of the $\text{IG}$ distribution strongly influence the posterior density when the true variance parameter is close to zero. FS-W show
that the $N$ prior is less influential and hence more suitable under model specification uncertainty. Second, they consider the following non-centered parameterization (NCP) of the centered parameterization (CP) in (1)-(2)

$$
y_t = X_t \beta_0 + (X_t \circ \tilde{\beta}_t) \sqrt{\theta} + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma^2_{\varepsilon}),$$  

(3)

$$
B(L) \tilde{\beta}_t = \tilde{\omega}_t, \quad \tilde{\omega}_t \sim N(0, I_m),$$  

(4)

where $\tilde{\beta}_t$ is restricted to start at zero ($\tilde{\beta}_0 = 0$) and with $\circ$ denoting the element-wise (Hadamard) product and $\sqrt{\theta} = \text{vec}(\sqrt{\Theta})$ an $(m \times 1)$ coefficient vector. This NCP splits the time-varying coefficients $\beta_t$ in a fixed part $\beta_0$ and a time-varying part $\tilde{\beta}_t$ such that model selection boils down to deciding whether $(X_t \circ \tilde{\beta}_t)$ is a relevant variable. Hence, the advantage of the NCP over the CP is that the decision on whether the variances in $\Theta$ are zero or not is reformulated into a standard variable selection problem of whether the elements in $\sqrt{\theta}$ are zero in the regression model formulation of (3). Recent applications of this stochastic model specification search can be found in, among others, Grassi and Proietti (2014); Berger et al. (2016); Everaert et al. (2017); Chan and Grant (2017).

The NCP is convenient for model selection, but may adversely affect convergence of the Markov chain. Differences in mixing properties between a CP and a NCP have been documented in the literature (see e.g. Pitt and Shephard, 1999; Papaspiliopoulos et al., 2003, 2007) and shown to depend on the characteristics of the underlying data generating process. Pitt and Shephard (1999) show that when sampling the unconditional mean in a first-order autoregressive (AR) plus noise model, a Markov chain based on the CP mixes much better compared to using the NCP when the unobserved state is highly persistent. A similar result can be found in Kim et al. (1998), who show that the NCP of a stochastic volatility model leads to very low MCMC sampling efficiency when persistence is high. More in general, the typical pattern is that if one parameterization leads to fast convergence, the other is found to be slow. Kastner and Frühwirth-Schnatter (2014) show that the NCP leads to more efficient sampling relative to using a CP in a stochastic volatility model when the error variance in the state equation is considerably smaller than that in the observation equation and persistence in the latent state is low. Similarly, Yu and Meng (2011) show that in a missing data formulation, the relative convergence rate of the CP and NCP depends on the fraction of missing data. When the amount of unobserved information is low, the NCP is expected to mix well, while convergence should be faster in the CP if there is a lot of missing information.

In practice, one can in principle decide which of the two competing parameterizations to use, depending on their relative mixing properties. However, when a test for time variation of the states is needed, the
CP is not an option as the stochastic model specification search of FS-W is based on the NCP. Building on the insights of Yu and Meng (2011), Kastner and Frühwirth-Schnatter (2014) and Simpson et al. (2017), this paper shows how an interweaving (IW) strategy can be used to boost sampling efficiency in TVP models while still being able to do specification search in the way proposed by FS-W. First, by taking advantage of the contrasting features of the two alternative parameterizations, the IW strategy has the capacity to reduce the dependence in the Markov chain. Yu and Meng (2011) show that IW minimally leads to an algorithm that is better than the worst of the two constituent parameterizations, but often it improves on the best performer and it may even converge when both the CP and the NCP fail to do so. Second, model specification search is possible from the non-centered block in the interwoven Gibbs sampler.

Our Monte Carlo (MC) simulation results for the local level model example considered by FS-W reveal that mixing in the NCP deteriorates when the importance of the state variances increases relative to the innovation variance in the observation equation. In contrast, mixing in the CP deteriorates when the relative importance of the state variances decreases. We further show that sampling efficiency can be boosted considerably by using an IW strategy, which even improves on the best performer out of the original CP and NCP under all parameter settings. Hence, IW provides the possibility to use the NCP for model selection while avoiding poor mixing properties of the MCMC algorithm. Moreover, even if one is not interested in testing for time-variation, such that the CP can be used, interweaving with the NCP may still be fruitful to improve sampling efficiency.

The remainder of this paper is organized as follows. Section 2 outlines the MCMC sampling schemes for the CP and NCP of the local level model example considered by FS-W and shows how to combine them in an IW strategy. Section 3 presents MC simulation results. Section 4 applies our IW strategy to a trend-cycle decomposition of U.S. GDP and to analyzing the stability of Okun’s Law coefficient for the U.S. Section 5 concludes.

2 MCMC sampling schemes for alternative parameterizations

In this section we present the MCMC sampling schemes for the CP and NCP and how these can be combined in an IW strategy. In order to keep the analogy with FS-W, we outline the various sampling
schemes using the following dynamic linear trend model:

\[ y_t = u_t + \varepsilon_t, \quad \varepsilon_t \sim N\left(0, \sigma^2_\varepsilon\right), \]  
\[ u_t = u_{t-1} + a_{t-1} + \omega_{1t}, \quad \omega_{1t} \sim N\left(0, \theta_1\right), \]  
\[ a_t = a_{t-1} + \omega_{2t}, \quad \omega_{2t} \sim N\left(0, \theta_2\right), \]

where \( u_t \) is a random walk with stochastic drift \( a_t \) with unknown initial values \( u_0 \) and \( a_0 \) respectively. This specification can be obtained from the model in (1)-(2) by setting \( X_t = (1, 0), B(L) = (1, 1; 0, 1) \) and relabeling \( \beta_{1t} \) and \( \beta_{2t} \) to \( u_t \) and \( a_t \), respectively.

### 2.1 Non-centered parameterization (NCP)

A key aspect of the stochastic model specification search proposed by FS-W is to rewrite the model in (5)-(7) into its non-centered counterpart:

\[ y_t = \mu_0 + a_0 t + \sqrt{\theta_1} \tilde{\mu}_t + \sqrt{\theta_2} \tilde{A}_t + \varepsilon_t, \quad \varepsilon_t \sim N\left(0, \sigma^2_\varepsilon\right), \]  
\[ \tilde{\mu}_t = \tilde{\mu}_{t-1} + \tilde{\omega}_{1t}, \quad \tilde{\omega}_{1t} \sim N\left(0, 1\right), \]  
\[ \tilde{a}_t = \tilde{a}_{t-1} + \tilde{\omega}_{2t}, \quad \tilde{\omega}_{2t} \sim N\left(0, 1\right), \]  
\[ \tilde{A}_t = \tilde{A}_{t-1} + \tilde{a}_{t-1}, \]

with \( \tilde{\mu}_t \) and \( \tilde{a}_t \) two independent random walk processes restricted to start at \( \tilde{\mu}_0 = \tilde{a}_0 = \tilde{A}_0 = 0 \). For further use below, let \( y = \{y_t\}_{T}^T, \tilde{\mu} = \{\tilde{\mu}_t\}_{T}^T, \tilde{a} = \{\tilde{a}_t\}_{T}^T \), and \( \tilde{A} = \{\tilde{A}_t\}_{T}^T \) denote the data and unobserved states stacked over time, while \( \theta_i \) (for \( i = 1, 2 \)) refers to one of the innovation variances \( \theta_1 \) or \( \theta_2 \).

This NCP is interesting in a number of respects. First, conditional on \( \tilde{\mu} \) and \( \tilde{A} \) the parameters \( \phi = (\mu_0, a_0, \sqrt{\theta_1}, \sqrt{\theta_2}, \sigma^2_\varepsilon) \) in the observation equation (8) can be sampled using a standard Bayesian linear regression. This allows us to estimate \( \sqrt{\theta_i} \) with a \( N \) prior centered at zero, instead of using the standard \( IG \) prior for \( \theta_i \). This is of particular importance as the scale and shape parameters of the \( IG \) prior tend to have a strong influence on the posterior distribution of \( \theta_i \) when the true value is close to zero. Since zero is not in the support of the \( IG \) distribution, using it as a prior pushes the posterior density away from zero. Using the \( N \) distribution as prior is convenient as for both \( \theta_i = 0 \) and \( \theta_i > 0 \) the posterior of \( \sqrt{\theta_i} \) is symmetric around zero. FS-W demonstrate that when \( \theta_i = 0 \), the posterior distribution of \( \sqrt{\theta_i} \) is not pushed away from zero under a \( N \) prior. Second, the signs of both \( \sqrt{\theta_1} \) and \( \tilde{\mu} \) can be changed without changing their product, and similarly for \( \sqrt{\theta_2} \) and \( \tilde{A} \). This lack of identification offers a first piece of information on whether time variation is relevant or not. For truly time-varying parameters,
the innovation variance \( \theta_i \) will be positive resulting in a posterior distribution for \( \sqrt{\theta_i} \) that is bimodal with modes \( \pm \sqrt{\theta_i} \). For time-invariant parameters, \( \theta_i \) is zero such that \( \sqrt{\theta_i} \) becomes unimodal at zero.

Third, as the square root of the innovation variance \( \sqrt{\theta_i} \) enters equation (8) as a regression coefficient, it is feasible to apply the stochastic variable selection approach of George and McCulloch (1993) to decided whether \( \tilde{\mu} \) and \( \tilde{A} \) are relevant model components.

The MCMC algorithm used to estimate the model in (8)-(11) revolves around cycling through two successive blocks. In the first block, the states \( \tilde{\alpha} = (\tilde{\mu}, \tilde{a}, \tilde{A}) \) are sampled from \( g(\tilde{\alpha}|y, \phi) \) conditional on the initial values and parameters collected in \( \phi = (\mu_0, a_0, \sqrt{\theta_1}, \sqrt{\theta_2})' \). In the second block, \( \phi \) is sampled from \( g(\phi|y, \tilde{\alpha}) \) conditional on the states \( \tilde{\alpha} \). More specifically, starting from a set of initial values \( \tilde{\alpha}(0) \) and \( \phi(0) \), the MCMC algorithm iterates over the following two blocks:

1. **Sample \( \tilde{\alpha}^{(j)} \) from \( g(\tilde{\alpha}|y, \phi^{(j-1)}) \)**

   Conditional on \( \phi \), the non-centered time-varying parameters \( \tilde{\alpha}_t \) are estimated with the Kalman filter and sampled using the simulation smoother of Durbin and Koopman (2002). In matrix form, the conditional state space representation is given by the observation equation

   \[
   y_t - \mu_0 - a_0 t = \begin{bmatrix} \sqrt{\theta_1} & 0 & \sqrt{\theta_2} \end{bmatrix} \begin{bmatrix} \tilde{\mu}_t \\ \tilde{a}_t \\ \tilde{A}_t \end{bmatrix} + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma^2_\varepsilon), \quad (12)
   \]

   with the transition of the unobserved states described by

   \[
   \begin{bmatrix} \tilde{\mu}_{t+1} \\ \tilde{a}_{t+1} \\ \tilde{A}_{t+1} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} \tilde{\mu}_t \\ \tilde{a}_t \\ \tilde{A}_t \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} \tilde{\omega}_{1t} \\ \tilde{\omega}_{2t} \end{bmatrix}, \quad \begin{bmatrix} \tilde{\omega}_{1t} \\ \tilde{\omega}_{2t} \end{bmatrix} \sim \mathcal{N} \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right), \quad (13)
   \]

   where the initial conditions are restricted to \( \tilde{\mu}_0 = \tilde{a}_0 = \tilde{A}_0 = 0 \).

2. **Sample \( \phi^{(j)} \) from \( g(\phi|y, \tilde{\alpha}^{(j)}) \)**

   Conditional on \( \tilde{\alpha}_t \), observation equation (8) is a standard linear regression model

   \[
   y_t = x_t \phi + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma^2_\varepsilon), \quad (14)
   \]

   where \( x_t = (1, t, \tilde{\mu}_t, \tilde{A}_t) \). Under the conjugate normal-inverse gamma prior

   \[
   \phi \sim \mathcal{N}(b_0, B_0 \sigma^2_\varepsilon), \quad \sigma^2_\varepsilon \sim \mathcal{IG}(c_0, C_0), \quad (15)
   \]

   we can sample the variance \( \sigma^2_\varepsilon \) of the observation errors from \( \mathcal{IG}(c_T, C_T) \) and, conditional on \( \sigma^2_\varepsilon \),
the parameters $\phi$ from $\mathcal{N}(b_T, B_T \sigma_\varepsilon^2)$, with the posterior moments defined as:

\begin{align*}
  b_T &= B_T (x'y + B_0^{-1}b_0), \\
  B_T &= (x'x + B_0^{-1})^{-1}, \\
  c_T &= c_0 + T/2, \\
  C_T &= C_0 + 0.5 \left( y'y + b_0'B_0^{-1}b_0 - b'_T B_T^{-1}b_T \right).
\end{align*}

(2*) Perform random sign switch

To re-enforce the fact that the sign of the standard deviations $\sqrt{\theta}$ and the states $\tilde{\alpha}$ are not separately identified, we perform a random sign switch, i.e. $\sqrt{\theta_1}$ and $\tilde{\mu}$ are left unchanged with probability 0.5 while with the same probability they are replaced by $-\sqrt{\theta_1}$ and $-\tilde{\mu}$. A similar sign switch is performed for $\sqrt{\theta_2}$ and $\tilde{A}$.

2.2 Centered parameterization

To align with the notation of the NCP, we split $u_t$ defined in (6) as $u_t = \mu_t + A_t$ such that the CP in (5)-(7) can be rewritten as

\begin{align*}
  y_t &= \mu_t + A_t + \varepsilon_t, \\
  \mu_t &= \mu_{t-1} + \omega_{1t}, \\
  a_t &= a_{t-1} + \omega_{2t}, \\
  A_t &= A_{t-1} + a_{t-1},
\end{align*}

where $A_0 = 0$ and with $\mu_0$ and $a_0$ unrestricted.

This CP differs from the NCP in two respects. First, sampling the centered states $\alpha = (\mu, a, A)$ in one block means that the initial conditions $\mu_0$ and $a_0$ are drawn together with the time-varying part of the states. Second, we now sample innovation variances $\theta_i$ instead of standard deviations $\sqrt{\theta_i}$, which is done in a separate block. More specifically, starting from a set of initial values $\alpha^{(0)}, \sigma_\varepsilon^{(0)}$ and $\theta^{(0)}$, the MCMC algorithm iterates over the following three blocks:

(1) Sample $\alpha^{(j)}$ from $g(\alpha|y, \theta^{(j-1)}, \sigma_\varepsilon^{(j-1)})$

First, apply the Kalman filter and the simulation smoother to draw $\alpha$ using a state space model.
This is based on the general result that the innovation variance $\theta_i$ is replaced by a $\mathcal{N}$ prior on the standard deviation $\sqrt{\theta_i}$. This is to avoid that the prior pushes the states $\alpha_t$ towards varying over time. Since below we will interweave the CP with the NCP, it is important that we align the prior on $\theta_i$ in the CP with the $\mathcal{N}$ prior on $\sqrt{\theta_i}$ in the NCP. Following Kastner and Frühwirth-Schnatter (2014), we use a Gamma ($\mathcal{G}$) prior $\theta_i \sim V_0 \chi^2_1 = \mathcal{G}(\frac{1}{2}, 2V_0)$, defined using the shape and scale parameterization, and where $V_0 = B_0 \theta_i^2$ is the prior variance on $\sqrt{\theta_i}$ in the NCP.\footnote{This is based on the general result that $X \sim \mathcal{N}(0, \sigma^2)$ implies $X^2 \sim \sigma^2 \chi^2_1 = \mathcal{G}(\frac{1}{2}, 2\sigma^2)$.}

Since the $\mathcal{G}$ prior is non-conjugate, we add a Metropolis–Hastings (MH) step to update $\theta_i$. Following Kastner and Frühwirth-Schnatter (2014), we use the auxiliary conjugate prior $p_{\text{aux}}(\theta_i) \propto \sqrt{\theta_i}^{-1}$, to denote the density of an improper conjugate $\mathcal{IG}(-\frac{1}{2}, 0)$ prior, and obtain a proposal density for $p(\theta_i)$ as

$$
\theta_1|\mu_t \sim \mathcal{IG}(c_T, C_T^{\mu}), \quad \theta_2|a_t \sim \mathcal{IG}(c_T, C_T^{a}),
$$

where $c_T = T/2 - 0.5$, $C_T^{\mu} = (\Delta \mu_t^2 \Delta \mu_t) / 2$ and $C_T^{a} = (\Delta a_t^2 \Delta a_t) / 2$. We take a candidate draw $\theta_{i,\text{new}}^\text{aux}$ from their proposal densities and accept it with a probability of $\min(1, R)$, where

$$
R = \frac{p(\theta_{i,\text{new}}^\text{aux})}{p(\theta_{i,\text{old}}^\text{aux})} \times \frac{p_{\text{aux}}(\theta_{i,\text{old}})}{p_{\text{aux}}(\theta_{i,\text{new}})} = \exp \left\{ \frac{\theta_{i,\text{old}} - \theta_{i,\text{new}}}{2V_0} \right\},
$$
for \( i = 1, 2 \) and with \( \theta_{i,\text{old}} \) denoting the last available draw for \( \theta_i \) in the Markov chain.

### 2.3 Interweaving the CP and NCP

It has already been documented that convergence of the MCMC algorithm can be slow, depending on the data generating process, the parameter values and whether a CP or NCP is used. In general, the rate of convergence when using the NCP deteriorates compared to using the CP as the amount of information in the unobserved states increases relative to the variance of the errors in the observation equation (see e.g. Yu and Meng, 2011; Kastner and Frühwirth-Schnatter, 2014). Translated to our local level model, we expect that increasing the state innovation variances in \( \theta \) relative to the observation error variance \( \sigma_\varepsilon^2 \) will slow down convergence of the MCMC algorithm based on the NCP.

The persistence in the Markov chain can be reduced by interweaving the CP and NCP. The basic idea is that the innovation variances \( \theta_1 \) and \( \theta_2 \) are sampled twice during each sweep of the MCMC algorithm, once using the CP and once using the NCP. Yu and Meng (2011) show that by taking advantage of the contrasting features of the alternative parameterizations, the IW strategy can outperform both approaches in terms of sampling efficiency. Intuitively, the CP will perform well in the parameter region where the NCP shows poor mixing (and vice versa). Applying the IW technique to our local level model yields the following sampling algorithm:

1. Sample \( \alpha^{(j)} \) from \( g(\alpha|y, \theta^{(j-1)}, \sigma_\varepsilon^{2(j-1)}) \) using the CP

2. Sample \( \theta^{(j)} \) from \( g(\theta|y, \alpha^{(j)}) \) using the CP

2*) Transform the centered \( \alpha^{(j)} \) to the non-centered \( \tilde{\alpha}^{(j)} \)

3. Sample \( \phi^{(j)} \) from \( g(\phi|y, \tilde{\alpha}^{(j)}) \) using the NCP

3*) Perform a random sign switch of the standard deviations \( \sqrt{\theta_1} \) and \( \sqrt{\theta_2} \) and the states \( \tilde{\alpha} \) in the NCP

Steps 1, 2, 3 and 3* are as outlined in the MCMC for the CP and NCP. In step 2*, we move to the NCP using the standardizations:

\[
\tilde{\mu}_t = \frac{\mu_t - \mu_0}{\sqrt{\theta_1}}, \quad \tilde{a}_{it} = \frac{a_{it} - a_0}{\sqrt{\theta_2}}, \quad \tilde{A}_t = \frac{A_t - a_0 t}{\sqrt{\theta_2}}.
\]

Note that interweaving does not come down to simply alternating between the two parameterizations but rather involves replacing the first step of the NCP-based sampling scheme by the CP-based algorithm. Interweaving, thus, ‘plugs in’ the CP scheme into the NCP sampling algorithm.
3 Monte Carlo Experiment

In this section, we set up a MC experiment to illustrate the mixing performance of the MCMC algorithm based on the standard CP and the NCP of the local level model in (5)-(7) and compare it to that of an IW strategy. We generate data for 121 combinations of \((\theta_1, \theta_2)\) while fixing \(\sigma^2_\varepsilon = 1\). More specifically, we vary \(\theta_1\) form 0 to 0.1 with steps of 0.01 and \(\theta_2\) form 0 to 0.01 with steps of 0.001. We expect the mixing properties of the NCP to deteriorate as \(\theta_1\) and/or \(\theta_2\) grow large relative to \(\sigma^2_\varepsilon\), and the other way around for the CP. We generate 1,000 datasets of length \(T = 250\) for every possible \((\theta_1, \theta_2)\) combination and apply the three sampling schemes (CP, NCP, IW) using 11,000 MCMC draws, of which the first 1,000 are discarded as burn-in. Throughout all simulations we use a vague normal prior \(\mathcal{N}(0, 10^2\sigma^2_\varepsilon)\) for the parameters \(\phi\) in the NCP and the corresponding gamma prior \(\mathcal{G}(\frac{1}{2}, 2\cdot10^2\sigma^2_\varepsilon)\) for the innovation variances \(\theta\) in the CP. For \(\sigma^2_\varepsilon\) we use an uninformative inverse gamma prior \(\mathcal{I}\mathcal{G}(c_0, C_0)\) where we set \(c_0 = 0.001\) and \(C_0 = 0.001\). We initialize the sampler with the true values of the states and parameters.

Before presenting the full MC results, we illustrate the persistence in the Markov chain by plotting the trajectory of draws for \(\sqrt{\theta_1}\) and \(\sqrt{\theta_2}\) for two contrasting parameter settings using a single time series randomly selected from the pool of 1,000 samples generated in the MC simulation. Figure 1 shows that the NCP has better mixing properties than the CP-based MCMC scheme for relative low innovation variances \((\theta_1 = 0.01; \theta_2 = 0.001)\). The opposite result image emerges from Figure 2, where for higher innovation variances \((\theta_1 = 0.1; \theta_2 = 0.01)\) the CP shows much better mixing behavior than the NCP. Interestingly, the mixing properties of the IW algorithm are adequate and not fundamentally different over the two parameter settings.

We next turn to the full MC simulation results. To formally evaluate the mixing behavior of the three alternative algorithms, we calculate the inefficiency factor (IF). This is an estimate for the integrated autocorrelation time \(\tau = 1 + 2\sum_{t=1}^d \hat{c}(t)\) with \(\hat{c}(t)\) the estimated autocorrelation function at lag \(t\) and \(d\) the lag for which \(\hat{c}(t) < 0.01\). The IF can be interpreted as the factor by which the squared MC standard error increases due to the dependence in the Markov chain. The IF equals one in the ideal situation of complete independence, while higher values imply a reduction in the “effective” number of draws such that more MCMC iterations are needed to attain the same MC standard error.

A surface plot of the average IF over the 1,000 generated datasets for every \((\theta_1, \theta_2)\) combination can be found in Figure 3. The difference in sampling efficiency over the three algorithms is remarkable. The CP-based sampling algorithm tends to mix better for larger values of \(\theta_1\) and \(\theta_2\). For the highest considered parameter values the IF of \(\sqrt{\theta_1}\) is around 450, while going up to 600 for lower parameter combinations.
Figure 1: MC simulation: Trajectories of draws from the CP, NCP and IW scheme for the standard deviations $\sqrt{\theta_1}$ and $\sqrt{\theta_2}$ of the state innovations when setting $\theta_1 = 0.01$ and $\theta_2 = 0.001$ in the data generating process.

Notes: In the CP we sample $\theta$ instead of $\sqrt{\theta}$ but we plot the trajectory of the square root of $\theta$ for direct comparability with the other schemes. For the NCP and IW we plot the absolute value of $\sqrt{\theta}$ as the graph would otherwise be blurred by the sign switches.

Figure 2: MC simulation: Trajectories of draws from the CP, NCP and IW scheme for the standard deviations $\sqrt{\theta_1}$ and $\sqrt{\theta_2}$ of the state innovations when setting $\theta_1 = 0.1$ and $\theta_2 = 0.01$ in the data generating process.

Notes: See Figure 1.

The NCP in contrast performs poorly for larger values of $\theta_1$ and/or $\theta_2$, with the IF amounting to 250 for sampling $\sqrt{\theta_1}$ and even going up to 1250 for sampling $\sqrt{\theta_2}$. When $\theta_1$ and $\theta_2$ are close to zero, the IF is close to 1 indicating almost perfect mixing behavior. Turning to the efficiency of the IW sampler, Figure 3 shows that although the IF tends to increase in $\theta_1$ and $\theta_2$, it is always much lower than that of the CP.
and NCP algorithms. The highest IF is around 90 for $\sqrt{\theta_1}$ and about 140 for $\sqrt{\theta_2}$. Our MC experiment thus shows that the IW strategy offers a huge increase in sampling efficiency, especially compared to the NCP-based algorithm in cases where the variance of the unobserved states is high.

The average IF plotted in Figure 3 hides a lot of variation in the sampling properties over repeated samples. Figure 4 therefore displays box plots of the IFs for a number of $(\theta_1, \theta_2)$ combinations. For conciseness we only show the four most extreme parameter pairs. The pattern emerging from Figure 4 is that when the mixing of a sampling scheme is poor on average the dispersion of the IFs over repeated samples is also high. This implies that the mixing performance of the CP and NCP schemes, next to the underlying data generating process, also to a large extent depends on the data sample at hand. For the NCP in the case where $\theta_1 = 0.1$ and $\theta_2 = 0.01$, for instance, the IF for sampling $\sqrt{\theta_2}$ can run from around 100 up to 2,900. As efficient sampling schemes show much less variation in their IF, the IW strategy offers not only the best but also the most stable mixing performance across repeated samples and parameter combinations.

**Figure 3**: MC simulation: Average inefficiency factor for CP, NCP and IW sampling schemes.

![Figure 3](image)

**Notes**: In the CP we sample $\theta$ instead of $\sqrt{\theta}$ but we calculate the IF from the square root of $\theta$ for direct comparability with the other schemes. For the NCP and IW we calculate the IF from the absolute value of $\sqrt{\theta}$ as the results would otherwise be distorted by the sign switches.

4 Applications

In this section we apply our IW strategy to a trend-cycle decomposition of U.S. GDP and to analyzing the stability of Okun’s Law coefficient for the U.S.
Figure 4: MC simulation: Boxplot inefficiency factors for CP, NCP and IW sampling schemes.

Notes: See Figure 3.

Trend-cycle decomposition of U.S. GDP

Decomposing real GDP into its trend (potential output) and cyclical component (output gap) has received a substantial amount of attention in the empirical macroeconomic literature. A popular approach is to use a state space model in which identification stems from assuming that the trend is a random walk plus drift and the cycle is a stationary AR process (see e.g. Watson, 1986; Kuttner, 1994, for early contributions). Although easy to implement, a major challenge in practice is that the obtained trend–cycle decomposition crucially depends on the specification of the model. Morley et al. (2003) show that the cyclical component becomes negligibly small when it is allowed to be correlated with the trend component, such that almost all shocks are permanent and cycles are small and noisy. Perron and Wada (2009) argue that this result is an artifact of restricting the drift in the trend component to be constant. They show that the U.S. cycle is much more substantial/persistent and accords well with the NBER chronology when allowing for a break in trend growth occurring in 1973:1. The trend now turns out to be non-stochastic as, apart from a break in trend growth in 1973:1, permanent shocks to the level are relatively unimportant. This finding is robust to modeling the drift as a random walk such that shocks to trend growth occur every period rather than exhibiting only a single deterministic break. Endogenizing the break date, Luo and Startz (2014) also find strong evidence for a structural break in the trend growth rate of the U.S. real GDP which now, most likely, takes place around 2006:1. However, they further argue that although the variance of permanent shocks to the level of trend output is small, the data does not definitively settle the question as the posterior density shows non-trivial probability mass at higher variance values.

In this section we decompose U.S. log real GDP, $y_t$, in a trend and cyclical component by augmenting
the local level model in (5)-(7) with a stationary cycle. We model potential output, $y^p_t$, as a random walk with time-varying drift, $g_t$, while the cyclical component, $y^c_t$, is assumed to be a stationary $AR(2)$ process to allow for a typical hump-shaped pattern. More specifically the CP of the model is:

$$y_t = y^p_t + y^c_t + \varepsilon_t,$$

$$y^p_t = g_{t-1} + y^p_{t-1} + \omega^p_t, \quad \omega^p_t \sim N(0, \theta^p),$$

$$g_t = g_{t-1} + \omega^g_t, \quad \omega^g_t \sim N(0, \theta^g),$$

$$y^c_t = \rho_1 y^c_{t-1} + \rho_2 y^c_{t-2} + \omega^c_t, \quad \omega^c_t \sim N(0, \theta^c),$$

where $\varepsilon_t$ captures measurement error and non-persistent shocks. Our main interest is to test whether shocks to the level, $\omega^p_t$, and to the growth rate, $\omega^g_t$, of potential output are relevant or not. To this end, we use the NCP given by

$$y_t = y^p_0 + g_0 t + \sqrt{\theta_p} \tilde{y}^p_t + \sqrt{\theta_g} \tilde{G}_t + y^c_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma^2),$$

$$\tilde{y}^p_t = \tilde{y}^p_{t-1} + \tilde{\omega}^p_t, \quad \tilde{\omega}^p_t \sim N(0, 1),$$

$$\tilde{G}_t = \tilde{G}_{t-1} + \tilde{g}_{t-1},$$

where $\tilde{y}^p_0 = \tilde{g}_0 = \tilde{G}_0 = 0$ and with $y^c_t$ as defined in (31).

We estimate the model with the CP, NCP and IW algorithm over the period 1947-2016 using quarterly log real GDP taken from the FRED database. We use an $IG$ prior distribution for the variance parameters: $\sigma^2$ and $\theta^c$ where the shape and scale parameters are set to a low value of 0.001 to attain noninformativeness (see e.g. Gelman, 2006). For the $AR(2)$ parameters of the business cycle we opt to put a strict prior on the sum of the parameters, i.e. $(\rho_1 + \rho_2) \sim N(0.90, 0.02^2)$, to assure that the business cycle has the desired persistence at the quarterly frequency and a much less informative prior on the second lag $\rho_2 \sim N(-0.35, 1000)$, with the prior belief of $-0.35$ in line with the typical hump-shaped pattern of the output gap. Stationarity of the business cycle is ensured by using a Metropolis-Hastings step as outlined in Chib and Greenberg (1994) when sampling the $AR$ parameters. We further choose an uninformative normal prior $N(0, 100^2 \sigma^2)$ centered around zero for $\sqrt{\theta^p}$ and $\sqrt{\theta^g}$ in the NCP with corresponding gamma prior $G(\frac{1}{2}, 2 \cdot 100^2 \sigma^2)$ for $\theta^p$ and $\theta^g$ in the CP. We generate 15,000 MCMC draws of which the first 5,000 are discarded as burn-in.

Figure 5 plots the trajectories of the MCMC draws for $\sqrt{\theta^p}$ and $\sqrt{\theta^g}$ obtained from the three different sampling schemes. Looking at the trace plots of $\sqrt{\theta^p}$, the NCP-based sampler mixes relatively well while
the CP leads to much more persistence in the draws. The picture is very different for $\sqrt{\theta_g}$ where the NCP-based chain now shows very poor mixing while the CP is much more efficient. In line with the findings in our MC simulation, the IW sampler successfully brings down the persistence in Markov chain for both $\sqrt{\theta_p}$ and $\sqrt{\theta_g}$. Histograms representing the posterior distributions of $\sqrt{\theta_p}$ and $\sqrt{\theta_g}$ are plotted in Figure 6. The poor mixing properties of the two plain schemes negatively affect the shape of the obtained posterior distributions, which become multi-modal for $\sqrt{\theta_p}$ in the CP and for $\sqrt{\theta_g}$ in the NCP. The IW sampling scheme, in contrast, results in well-behaved posteriors for both parameters. Hence, we focus on the outcome of the IW scheme for the economic interpretation of the results below.

**Figure 5:** Trend-cycle decomposition of U.S. GDP: Trajectories of draws for $\sqrt{\theta_p}$ and $\sqrt{\theta_g}$ from the CP, NCP and IW sampling schemes (10,000 draws)

|       | $\rho_1$ | $\rho_2$ | $|\sqrt{\theta_p}|$ | $|\sqrt{\theta_g}|$ | $\theta_c$ | $\sigma^2_\epsilon$ |
|-------|----------|----------|----------------|----------------|-----------|-------------------|
| Mean  | 1.4266   | -0.5068  | 0.0010         | 0.0004         | 0.00096   | 0.00007          |
| Median| 1.4298   | -0.5095  | 0.0008         | 0.0004         | 0.00096   | 0.00006          |
| S.D.  | 0.0835   | 0.0822   | 0.0007         | 0.0002         | 0.00001   | 0.00001          |

Notes: See Figure 1.

Due to the unidentified sign of $\sqrt{\theta}$, the posterior distribution of $\sqrt{\theta}$ is always centered at zero but will be bimodal when $\theta > 0$ and unimodal for $\theta = 0$. Figure 6 demonstrates clear-cut bimodality for $\sqrt{\theta_g}$ but not for $\sqrt{\theta_p}$. This implies that the dynamics in potential output are driven by permanent shifts in the growth rate rather than by level shocks. This result is in line with the evidence presented by Perron and Wada (2009) and Luo and Startz (2014) that shocks to the level of trend GDP are relatively unimportant once structural breaks in the growth rate are allow for. Summary results for the posterior distributions of the fixed parameters are presented in Table 1.
Figure 6: Trend-cycle decomposition of U.S. GDP: Posterior distributions for $\sqrt{\theta_p}$ and $\sqrt{\theta_g}$ from the CP, NCP and IW sampling schemes (10 000 draws)

Since in the CP-based scheme $\theta$ is estimated instead of $\sqrt{\theta}$ we take the square root of the trajectory of $\theta$ and randomly multiply the values by 1 and -1 for direct comparability with the other two schemes.

Figure 7 plots potential output (left panel) next to trend growth (middle panel) and the business cycle (right panel). Our estimate of the business cycle is able to reproduce all of the NBER recession periods and is highly comparable to the cycle reported by Luo and Startz (2014). In line with the finding of Antolin-Diaz et al. (2016) that U.S. growth exhibits a downward pattern, our measure of quarterly trend growth has decreased from around 1% in the beginning of the sample to below 0.5% at the end. The observation that this decrease is mainly due to marked drops in the 1970s and 2000s is consistent with the breaks found in 1973:1 by Perron and Wada (2009) and in 2006:1 by Luo and Startz (2014). Hence, a broken linear trend seems to be a fairly adequate approximation of U.S. potential GDP.

Stability of Okun’s Law in the U.S.

As a second illustration, we analyze the stability of Okun’s Law for the U.S. The “difference version” stipulates that there is a negative relation between the change in real economic activity and the change in the unemployment rate. Given that this is an empirically observed relationship rather than a theoretically derived result, its usefulness and especially its stability have been questioned. Daly and Hobijn (2010), for instance, report that the increase of the U.S. unemployment rate in 2009 was far above of what was predicted by Okun’s law. Although Sögner and Stiassny (2002) and Ball et al. (2013) argue that the relationship is strong and stable in the U.S., others show that it changes over time as a result of omitted
variables (Prachowny, 1993), institutional changes (Huang and Lin, 2008; Owyang and Sekhposyan, 2012) and/or due to a different reaction of unemployment in expansions and recessions (Crespo-Cuaresma, 2003; Silvapulle et al., 2004; Holmes and Silverstone, 2006; Knotek, 2007).

In order to assess the stability of Okun’s law we use a TVP linear regression model based on Sögnert and Stiassny (2002)

\[ \Delta u_t = \alpha_t + \beta_t \Delta y_t + \gamma_t \Delta y_{t-1} + \varepsilon_t, \]

where \( \Delta u_t \) is the change in the unemployment rate and \( \Delta y_t \) is the change in log real GDP. Lagged output growth \( \Delta y_{t-1} \) is added to allow for a delayed reaction of unemployment to economic activity, such that \( \beta_t \) measures the effect on impact and \( \gamma_t = \beta_t + \gamma_t^* \) the total effect. The intercept \( \alpha_t \) represents the change in unemployment when output growth is stable at zero. We model each of the three parameters as random walks, which are defined in the CP as

\[ \alpha_t = \alpha_{t-1} + \omega^\alpha_t, \quad \omega^\alpha_t \sim N(0, \theta^\alpha), \]

\[ \beta_t = \beta_{t-1} + \omega^\beta_t, \quad \omega^\beta_t \sim N(0, \theta^\beta), \]

\[ \gamma_t = \gamma_{t-1} + \omega^\gamma_t, \quad \omega^\gamma_t \sim N(0, \theta^\gamma), \]

and can be rewritten in the NCP as:

\[ \alpha_t = \alpha_0 + \sqrt{\theta^\alpha} \tilde{\alpha}_t, \quad \tilde{\alpha}_t = \tilde{\alpha}_{t-1} + \tilde{\omega}^\alpha_t, \quad \tilde{\alpha}_0 = 0, \quad \tilde{\omega}^\alpha_t \sim N(0, 1), \]

\[ \beta_t = \beta_0 + \sqrt{\theta^\beta} \tilde{\beta}_t, \quad \tilde{\beta}_t = \tilde{\beta}_{t-1} + \tilde{\omega}^\beta_t, \quad \tilde{\beta}_0 = 0, \quad \tilde{\omega}^\beta_t \sim N(0, 1), \]

\[ \gamma_t = \gamma_0 + \sqrt{\theta^\gamma} \tilde{\gamma}_t, \quad \tilde{\gamma}_t = \tilde{\gamma}_{t-1} + \tilde{\omega}^\gamma_t, \quad \tilde{\gamma}_0 = 0, \quad \tilde{\omega}^\gamma_t \sim N(0, 1). \]
We estimate this model using U.S. quarterly data over the period 1948-2016 with the unemployment rate $u_t$ and log real GDP $y_t$ taken from the FRED database. As before, we use uninformative prior distributions, i.e. $IG(0.001, 0.001)$ for $\sigma^2_\varepsilon$ and $N(0, 100^2\sigma^2_\varepsilon)$ for $\sqrt{\sigma_\alpha}$, $\sqrt{\sigma_\beta}$ and $\sqrt{\sigma_\gamma}$ in the NCP with corresponding $G(\frac{1}{2}, 2 \cdot 100^2\sigma^2_\varepsilon)$ for $\theta_\alpha$, $\theta_\beta$ and $\theta_\gamma$ in the CP. We generate 15,000 MCMC draws of which the first 5,000 are discarded as burn-in.

Figure 8 plots the MCMC trajectories when sampling the innovation standard deviations $\sqrt{\theta}$ using the three different schemes. The CP is clearly the worst performer, while the NCP and IW samplers both show much better mixing properties. The reason for the relatively bad performance of the CP is that the state innovation variances $\theta$ are relatively small compared to the error variance $\sigma^2_\varepsilon$, as indicated by the posterior results reported in Table 2. The posterior distributions of the standard deviations $\sqrt{\theta}$ are presented in Figure 9. The poor mixing properties of the CP disfigures the posterior distributions, especially for $\sqrt{\theta_\beta}$, while for the NCP and IW these are all well shaped. This highlights that even if one is not interested in testing for time variation, such that the CP can be used, interweaving with the NCP may still be fruitful to improve sampling efficiency.

Figure 8: Stability of Okun’s Law in the U.S.: Trajectories of draws for $\sqrt{\theta_\alpha}$, $\sqrt{\theta_\beta}$ and $\sqrt{\theta_\gamma}$ from the CP, NCP and IW sampling schemes (10,000 draws)

Notes: See Figure 1.
Table 2: Stability of Okun’s Law in the U.S.: Summary results for the posterior distributions of the fixed parameters (IW sampling scheme, 10,000 draws)

<table>
<thead>
<tr>
<th></th>
<th>$\sqrt{\theta_\alpha}$</th>
<th>$\sqrt{\theta_\beta}$</th>
<th>$\sqrt{\theta_\gamma}$</th>
<th>$\sigma^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.0203</td>
<td>0.0038</td>
<td>0.0115</td>
<td>0.0493</td>
</tr>
<tr>
<td>Median</td>
<td>0.0190</td>
<td>0.0029</td>
<td>0.0106</td>
<td>0.0491</td>
</tr>
<tr>
<td>S.D.</td>
<td>0.0088</td>
<td>0.0033</td>
<td>0.0068</td>
<td>0.0047</td>
</tr>
</tbody>
</table>

Figure 9: Stability of Okun’s Law in the U.S.: Posterior distributions for $\sqrt{\theta_\alpha}$, $\sqrt{\theta_\beta}$ and $\sqrt{\theta_\gamma}$ from the CP, NCP and IW sampling schemes (10,000 draws)

Notes: See Figure 6.

The evolution of the TVPs is plotted in Figure 10. The contemporaneous impact of a change in economic activity on unemployment $\beta_t$ seems to be stable over time, a finding that is supported by the unimodal posterior distribution of $\sqrt{\theta_\beta}$ in Figure 9. The plot of the total impact $\gamma_t$ reveals some variation over time, especially towards the end of the sample. This is confirmed by the bimodality in the posterior distribution of $\sqrt{\theta_\gamma}$. In line with Knotek (2007), our estimation results suggest that unemployment changes have become a bit more sensitive to output growth over the last decades. The data are, however, not very informative in this respect as the posterior also has non-negligible probability mass at zero when using posteriors from the adequately mixing NCP or IW schemes. Our results further indicate that the main source of instability in Okun’s Law can be attributed to the intercept $\alpha_t$, showing a gradual decline from the 50s onwards followed by a steep drop around the time of the Great Recession. In line with
Figure 10: Stability of Okun’s Law in the U.S.: Median estimates for the TVPs together with their 90% credible interval (IW sampling scheme, 10,000 draws).

Knotek (2007) and Owyang and Sekhposyan (2012), this implies that nowadays a stable unemployment rate goes together with less economic growth compared to the 1950s and 1960s. Note that the instability in $\alpha_t$ is found to be an important feature of the model as highlighted by the clear-cut bimodality in the posterior distribution of $\sqrt{\theta_{\alpha}}$. This finding is also in line with Prachowny (1993) who argues that other factors like hours worked, capacity utilization, labor supply and productivity growth are omitted variables in Okun’s Law. In our TVP model, these omitted variables are (partly) captured by the time-varying intercept $\alpha_t$.

5 Conclusion

In this paper we have demonstrated that the mixing properties of MCMC estimation of TVP models crucially depend on the underlying data generating process and on the way the state space model is parametrized. When the unobserved state innovation variances are small compared to the variance of error terms in the observation equation, the NCP mixes well while the CP performs poorly. The reverse image emerges when the state innovation variances grow in relative importance. In practice, one can in principle decide which of the two competing parameterizations to use, depending on their relative mixing properties. However, when a test for time variation of the states is needed, the CP is not an option as the stochastic model specification search of FS-W is based on the NCP. In this paper we show how to boost the MCMC efficiency by interweaving the NCP with the CP, while still being able to do model selection in the way proposed by FS-W. Taking advantage of the contrasting features of the two alternative parameterizations, the IW strategy substantially reduces the dependence in the Markov chain. Our MC simulation results show that sampling efficiency increases considerably when using the IW strategy and even improves on the best performer out of the original CP and NCP under all parameter settings. Hence, IW provides the possibility to use the NCP for model selection while avoiding poor mixing properties of
the MCMC algorithm. Moreover, even if one is not interested in testing for time variation, such that the CP can be used, interweaving with the NCP may still be fruitful to improve sampling efficiency.
References


