Bayesian State Space Models in Macroeconometrics

Joshua C.C. Chan  
Purdue University and UTS

Rodney W. Strachan*  
University of Queensland

March 19, 2020

*School of Economics, University of Queensland, r.strachan@uq.edu.au. Rodney Strachan is a Fellow of the Rimini Centre for Economic Analysis and the Centre for Applied Macroeconomic Analysis.
ABSTRACT

State space models play an important role in macroeconometric analysis and the Bayesian approach has been shown to have many advantages. This paper outlines recent developments in state space modelling applied to macroeconomics using Bayesian methods. We outline the directions of recent research, specifically the problems being addressed and the solutions proposed. After presenting a general form for the linear Gaussian model, we discuss the interpretations and virtues of alternative estimation routines and their outputs. This discussion includes the Kalman filter and smoother, and precision based algorithms. As the advantages of using large models have become better understood, a focus has developed on dimension reduction and computational advances to cope with high-dimensional parameter spaces. We give an overview of a number of recent advances in these directions.

Many models suggested by economic theory are either non-linear or non-Gaussian, or both. We discuss work on the particle filtering approach to such models as well as other techniques that use various approximations - to either the time $t$ state and measurement equations or to the full posterior for the states - to obtain draws.

Keywords: State space model, filter, smoother, non-linear, non-Gaussian, high-dimension, dimension reduction.

JEL Classification: C11, C22, E32
1 Introduction

State space modelling using Bayesian methods has a long history (Kitagawa and Gersch (1984)). These models, also called dynamic linear models, are an incredibly powerful tool for capturing unobserved processes and the specification encompasses a wide range of potential mechanisms. As such, they play an important role in macroeconometric modelling. They have, for example, allowed us to better estimate trends and cycles (Harvey (1985), Watson (1986) and more recently Morley (2002), Harvey, Trimbur and van Dijk (2007), Stock and Watson (2007), Chan, Léon-González, and Strachan (2018), Grant and Chan (2017)), they arise rather naturally in DSGE models (e.g., see Kulish, Morley and Robinson (2017) or for an excellent overview of Bayesian estimation of the DSGE see Herbst and Schorfheide (2016)), they have been used to improve forecasting and to account for structural change (Harvey and Todd (1983), Aoki (1987), Clark (2011), D’Agostino, Gambetti, and Giannone (2013)), they help us to understand how uncertainty and volatility have evolved and how the transmission of monetary policy shocks has evolved over time (Cogley and Sargent (2001, 2005), Primiceri (2005), Carriero, Clark and Marcellino (2016, 2018)).

There exists many alternative frameworks for models that evolve, such as non-parametric Bayesian or Gaussian processes and local likelihood approaches (see, for example, Kalli and Griffin (2018) and Petrova (2019)). State space models share features with many of the non-parametric models, but with the notable distinction that they tend to define evolution of the time-varying parameters, or states, in the direction of time (with an exception to this in Koop and Potter (2010)). State space models, however, have a number of advantages such as the ease of developing economically sensible specifications for the evolution of parameters and the existence of efficient and accessible algorithms for their estimation. For these reasons, a wide range of
macroeconometric models have been framed in the state space form. Examples in-
clude single equation models such as unobserved components models, multivariate
models such as time-varying parameter vector autoregression models (TVP-VAR),
and both static and dynamic factor models.

The literature on state space modelling is quite extensive and so, in this article,
we focus upon its application in Bayesian macroeconometric analysis. Even with this
limitation, the survey will be incomplete as we attempt to focus on only a few of the
more recent contributions; although we will necessarily discuss more established ones.

We focus primarily upon the linear Gaussian state space model with some dis-
cussion of extensions later in the article. To set the specification and notation, for
t = 1, . . . , T, let \( y_t \) be an \( n \times 1 \) vector of variables of interest so that we have a
model for a multivariate process, and let \( x_t \) be an \( n \times k \) matrix of regressors. Then a
reasonably general form for the linear Gaussian state space model is:

\[
y_t = x_t \pi_t + \varepsilon_t \tag{1}
\]

\[
\pi_t - \mu = P(\pi_{t-1} - \mu) + R \eta_t \tag{2}
\]

\[
\begin{pmatrix}
\varepsilon_t \\
\eta_t
\end{pmatrix}
\sim N\left(0, \begin{bmatrix}
\Sigma_t & 0 \\
0 & \Omega_t
\end{bmatrix}\right) \quad \text{and} \quad \pi_1 \sim N(\bar{\pi}_1, V_1).
\]

The dimensions of \( \pi_t \) and \( \mu \) (\( k \times 1 \)) and the known matrix \( P \) (\( k \times k \)) are determined
by the column dimension of the regressor \( x_t \) while the known matrix \( R \) (\( k \times r \)) need
not be square, i.e., \( r \leq k \).

The equation (1) is commonly called the measurement equation or observation
equation and provides the link from the data to the states, \( \pi_t \). The distribution
chosen for \( \varepsilon_t \) and (1) for all \( t \) will in most cases define the likelihood. The state
equation or transition equation (2) governs the evolution of the states or time varying
parameters \( \pi_t \). With the distributions for \( \pi_1 \) and \( \eta_t \) in a Bayesian analysis, the state equation defines the prior. Of particular importance is the prior correlation structure this implies for the states.

In macroeconometrics, it is common to assume \( E(\varepsilon_t \eta_t') = 0 \) and that \( P \) and \( R \) are \((k \times k)\) identity matrices; \( P = R = I_k \) and so \( r = k \), although more general specifications of the state space model have been employed. Cogley and Sargent (2001), for example, extend this specification to allow \( E(\varepsilon_t \eta_t') \neq 0 \). In an unobserved components (UC) model, Morley (2010) uses \( x_t = (1, 1) \), \( \Sigma_t = 0 \ \forall t \), an AR(2) for the cycle (the first element of \( \pi_t \)) and permits the shocks to the trend (the second element of \( \pi_t \)) and cycle to be correlated.

The random walk specification for the state equation \((P = R = I_k)\)

\[
\pi_t = \pi_{t-1} + \eta_t
\]

is a popular default. Although it may not always have a sound justification as a prior, this specification does have important virtues such as inducing a greater degree of smoothness than the stationary specification. It may also more closely match our prior beliefs than a stationary specification. In finance, to obtain a realistic model of the variance it is common when specifying the state equation for stochastic volatility (we discuss this in more detail in Section 4) to assume \( P \) is a scalar and \(|P| < 1\).

However, the volatilities for macroeconomic processes are not well represented by stationary, mean reverting processes (Eisenstat and Strachan (2016)). Typical estimates of the volatility of, say, inflation, log real per capita GDP or unemployment, display deviations of long durations. They may be stationary processes but a linear autoregressive process will not adequately capture the behaviour. Eisenstat and Strachan (2016) make this argument and introduce a change-point stationary model for
the state equation which allowed estimation of the time varying mean and persistence of the states.

We structure the remaining sections of this paper as follows. In Section 2, we outline the most popular, almost universally adopted, estimation procedure - the Kalman filter - as well as another procedure that shows promise for high-dimensional systems. In Section 3, we consider approaches to deal with overparameterization. The discussion first covers techniques that allow the data to decide whether each state should vary over time. We then discuss approaches to reducing the dimension of the entire model and thereby premitting estimation of larger models. The applications in both cases are to TVP-VAR models. Section 5 considers popular and recent approaches to estimation of non-linear and non-Gaussian state space models where the Kalman filter cannot be applied. Section 6 concludes with suggestions for future research.

2 Estimation of the linear Gaussian model

Estimation of the $(Tk \times 1)$ vector of states $\pi = (\pi'_1, \ldots, \pi'_T)'$ is most commonly done using the Kalman filter and Kalman smoother. The Kalman filter, first proposed by Kalman (1960) and Kalman and Busy (1961), has been studied extensively and explanations are available in several books and articles; one excellent and very accessible reference is Durbin and Koopman (2001). To begin establishing notation, denote by $y^{1:t}$ the information available in the data up to time $t$ such that $y^{1:T}$ represents all of the information in the sample. Beginning with $V_1$ and $\pi_1$, the Kalman filter produces, at time $t$, estimates of the mean and variance of the state $\pi_t$ using the data up to $t - 1$, that is $p_t = E(\pi_t|y^{1:t-1})$ and $V_t = V(\pi_t|y^{1:t-1})$. One representation of the steps
in the Kalman filter is

\[ \nu_t = y_t - x_t \pi_t \quad F_t = x_t V_t x_t' + \Sigma_t \]
\[ K_t = V_t x_t' F_t^{-1} \quad L_t = I - K_t x_t \]
\[ p_{t+1} = p_t + K_t \nu_t \quad V_{t+1} = V_t L_t' + R \Omega_t R' \]

The contemporaneous filtering equations (see Durbin and Koopman (2001) page 68) are another representation of this filtering approach but produce estimates of the mean and variance of the state \( \pi_t \) using the data up to \( t \), that is \( p_{t|t} = E(\pi_t|y^{1:t}) \) and \( V_{t|t} = V(\pi_t|y^{1:t}) \). The Kalman smoother, by contrast, produces estimates of the mean and variance of the state \( \pi_t \) using all available data, that is \( \hat{\pi}_t = E(\pi_t|y^{1:T}) \) and \( \hat{V}_t = V(\pi_t|y^{1:T}) \) and so from a statistical perspective can be said to be efficient.

To obtain the smoothed estimates, computation begins with the end values \( r_T = 0 \) and \( N_T = 0 \) and works backwards from \( t = T \) to \( t = 1 \) through the recursions

\[ r_{t-1} = x_t' F_t \nu_t \quad N_{t-1} = x_t' F_t^{-1} x_t + L_t' N_t L_t \]
\[ \hat{\pi}_t = p_t + V_t r_{t-1} \quad \hat{V}_t = V_t - V_t N_{t-1} V_t \]

There have been substantial improvements in Kalman filter-based algorithms over the past thirty years in estimating linear Gaussian state space models. Important examples include Carter and Kohn (1994), Frühwirth-Schnatter (1994), de Jong and Shephard (1995) and Durbin and Koopman (2002).

When undertaking a study using the state space model, the statistician has the choice to report either filtered estimates - \( p_t \) and \( V_t \) (or \( p_{t|t} \) and \( V_{t|t} \)) - or smoothed estimates - \( \hat{\pi}_t \) and \( \hat{V}_t \). Each of these estimators estimates different things in both a statistical and an economic sense, and so each of these estimates will serve different purposes. When using either the filter or smoother in macroeconometrics, it is worth-
while to give some consideration to the economic meaning of the outputs in each case and to what purpose each is best suited. Estimates from the Kalman filter $p_t$ and $V_t$ may include variation beyond actual variation in the underlying state and this excess variation is due to learning (Sims (2001)). Such variation is incorporated into the estimates from the Kalman smoother.

In comments on Cogley and Sargent (2001), Sims (2001) pointed out the significance from an economic perspective of the differences in the outputs from the Kalman filter and the Kalman smoother. The difference between the filtered and smoothed estimates tells us something about the difference between the best estimates made at the time $t$ and ex-post estimates that use all available data today (or at time $T$). Another less formal way to look at this difference is to think of smoothed estimates as estimating what was actually happening at the time $t$ whereas filtered estimates replicate estimates using the information available to the econometrician at time $t$; possibly to then replicate forecasts made at that time. The latter view is not entirely satisfactory because it does not take into account that the econometrician’s information set at time $t$ would include unrevised or real-time data, not final data. In this sense it is less clear from an economic perspective how to interpret filtered estimates produced using final data.

The use of the Kalman filter and smoother to estimate the state space model has become so commonplace that the estimation routine and model are sometimes referred to synonymously. It is important to distinguish the model from the estimation, however, as there do exist alternative approaches to estimation. More recently, a new class of algorithms has been proposed for estimating, primarily, linear Gaussian state space models. These are the precision-based algorithms, or precision samplers, and they have been proposed a number of times. For example, Rue (2001), Chan and Jeliazkov (2009) and McCausland, Miller, and Pelletier (2011) all propose sim-
ilar approaches. These estimators do not use a filtering approach, rather they draw all states in one step in much the same way one would draw a vector of regression coefficients from its posterior distribution.

In outlining the precision sampler, we follow the approach of Chan and Jeliazkov (2009). Stacking all states into one large \((T_k \times 1)\) vector \(\pi = (\pi_1', \ldots, \pi_T')'\), the conditional (upon the parameters) posterior for this vector is normal with \((T_k \times 1)\) mean vector \(\overline{\pi}\) and covariance matrix \(\overline{V}_\pi\). To see this, stack (1) over time as

\[
y = X\pi + \varepsilon \quad \varepsilon \sim N(0, \Sigma)
\]

where \(y = (y_1', \ldots, y_T')'\) and \(\varepsilon = (\varepsilon_1', \ldots, \varepsilon_T')'\) with dimensions \((nT \times 1)\) and \((nT \times 1)\) respectively, and

\[
X = \begin{bmatrix} x_1 & 0 & \cdots & 0 \\ 0 & x_2 & 0 & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & x_T \end{bmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix} \Sigma_1 & 0 & \cdots & 0 \\ 0 & \Sigma_2 & 0 & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & \Sigma_T \end{bmatrix}.
\]
Then stack the state equation (2) over time to obtain

\[
D \pi = M \eta \quad \text{\( \eta \sim N(0, \Omega) \)}
\]

\[
D = \begin{bmatrix}
I_k & 0 & 0 & \cdots & 0 \\
-P & I_k & 0 & \cdots & 0 \\
0 & -P & I_k & \cdots & 0 \\
\vdots & \ddots & \ddots & \ddots & \vdots \\
0 & 0 & \cdots & -P & I_k \\
\end{bmatrix}, \quad \eta = \begin{bmatrix}
\eta_1 \\
\eta_2 \\
\vdots \\
\eta_{T-1} \\
\eta_T \\
\end{bmatrix}
\]

\[
\Omega = \begin{bmatrix}
V_1 & 0 & \cdots & 0 \\
0 & \Omega_2 & 0 & \cdots \\
\vdots & \ddots & \ddots & \ddots \\
0 & 0 & \cdots & \Omega_T \\
\end{bmatrix}, \quad \bar{\pi} = \begin{bmatrix}
\bar{\pi}_1 \\
0 \\
\vdots \\
0 \\
\end{bmatrix}
\]

and \( M = (I_T \otimes R) \). Assuming \( r = k \) and \( R = I_k \), Chan and Jeliazkov (2009) show the resulting posterior mean (\( \bar{\pi} \)) and covariance matrix are (\( \bar{\Sigma} \)) defined by

\[
\bar{\Sigma} = [D' \Omega^{-1} D + X' \Sigma^{-1} X]^{-1}
\]

and the full vector of states has posterior \( \pi \sim N(\pi, \bar{\Sigma}) = p(\pi | y) \).

The precision algorithms take advantage of the sparse and banded structure of many of the very large matrices that enter into the posterior for the states. For example, since each of \( D, \Omega, X \) and \( \Sigma \) are banded - that is, they mostly consist of zeros and the relatively few non-zero elements are arranged along the main diagonal - the matrices \( D' \Omega^{-1} D \) and \( X' \Sigma^{-1} X \) are also banded. As a result, the \((Tk \times Tk)\)
inverse covariance matrix of the precision matrix

\[ P_{\pi} = V_{\pi}^{-1} = D'\Omega^{-1}D + X'\Sigma^{-1}X \]

is banded as well. These features, plus the fact that the matrix \( P_{\pi} \) is symmetric, imply a much lower cost of storage and fewer computations are needed. Importantly, we need not compute the inverse \( V_{\pi} = P_{\pi}^{-1} \); doing so would involve \( \mathcal{O}(T^3) \) operations. Taking into account this structure and using efficient functions in standard software packages (e.g., Matlab), the algorithms avoid many redundant computations and reducing the order of operations to \( \mathcal{O}(T) \). Having \( P_{\pi} \), we first take the Cholesky \( P_{\pi} = C'C \) where \( C \) is an upper-triangular matrix, and write the equations defining the mean vector as

\[ C'\pi = D'\Omega^{-1}\pi + X'\Sigma^{-1}y. \]

Second, we define \( \gamma = C\pi \) and solve by forward substitution the following system of equations for \( \gamma : \)

\[ C'\gamma = D'\Omega^{-1}\pi + X'\Sigma^{-1}y. \]

Finally, having \( \gamma \), we then solve the system of equations \( C\pi = \gamma \) for \( \pi \). Similarly, we can use the same type of operations to obtain a draw of the \((Tk \times 1)\) vector of states, \( \pi \). That is, we draw a \((Tk \times 1)\) vector of independent standard Normal random variables, \( z \sim N(0, I_{Tk}) \) and solve \( C\pi = z \) for \( \pi \). The resulting vector \( \pi \) is a draw from \( N(\bar{\pi}, V_{\pi}) \). These operations result in a lower order of computations to produce each estimated state than the Kalman smoother, but the algorithm produces the same outputs as does the Kalman smoother. The significance of this precision based approach is that, with a lower cost of computation, macroeconometricians are able to consider estimation of larger models. As the precision sampler saves on
computations, and this saving increases with dimension $n$, they enable estimation of higher dimensional systems for less additional cost. In a recent paper, Chan et al. (2020) take advantage of this in estimating a high dimensional time varying parameter (TVP) vector autoregressive (VAR) model. In their specification, $P = R = I_k$ and $\Omega_t = \Omega$ has reduced rank.

3 Approaches to overparameterization

Macroeconometricians have come to understand the importance, for a range of reasons, of allowing model parameters to evolve over time. One important example of this is in VAR models with time varying parameters and, in particular, with heteroskedasticity. These models have lead to an improved understanding of policy and the macroeconomic environment (see, among many others, Cogley and Sargent (2005), Cogley, Morozov and Sargent (2005), Primiceri (2005), Koop, Léon-González and Strachan (2009) and Canova and Ferroni (2012)). A natural candidate specification to permit evolution of the parameters, and one that is frequently used for time varying parameter models, is the state space model. The TVP-VAR in state space form has proven very successful.

Over recent years, there has been an increasing interest in estimating higher dimensional state space models with the TVP-VAR being an important example. Alongside the literature on time varying parameters has been work pointing out empirical and theoretical issues that arise from using small VARs, that is VARs with only a few variables (Carriero, Kapetanios and Marcellino (2011), Giannone, Lenza, Momferatou and Onorante (2014), Koop (2013) and Koop and Korobilis (2013)). Mapping from a dynamic stochastic general equilibrium (DSGE) model to a VAR involves approximations to the implied specification (such as log linearization or truncating the
lag length of the $\text{VAR}(\infty)$ (Giacomini (2013)) or encounters the issue of nonfundamentalness (resulting in the errors in the DSGE not mapping to the structural errors of the VAR) (Fernández-Villaverde and Rubio-Ramírez (2007)). Moving to a larger VAR will often mitigate these concerns. In the case where we map from a DSGE, using a larger VAR is an effort to better match the information set of the econometrician to that of agents acting in the model. Theoretical issues such as these have naturally led to consideration of how to estimate larger models.

Larger models also have empirical advantages. For example, in forecasting there has been growing evidence that large VAR models perform better than factor models (Banbura, Giannone and Reichlin (2010), Carriero, Clark and Marcellino (2019), Carriero, Kapetanios and Marcellino (2009), Giannone, Lenza, Momferatou and Omran (2014), Koop (2011)).

These two streams of literature have naturally led to work developing higher dimensional and time varying models. As the number of variables included in the VAR increases, however, such models quickly become overparametrized or overfitting leading to poor estimation and inference. Allowing for time-variation, with large TVP-VAR models, significantly exacerbates this issue. To mitigate the overparameterization issues, researchers have sought ways to reduce the dimensions of such models to improve estimation.

One approach to achieving parsimony is to allow the data to select parameters to be set to zero: an application of a technique commonly called \textit{variable selection}. A second approach is to allow the distribution of the parameter to be shrunk towards zero, so that the parameter can effectively be treated as if it were zero. This is the \textit{shrinkage} approach. A third approach treats the entire model structure (rather than individual parameters) to reduce the dimension of the model. We leave discussion of this last approach to the end of this section.
Variable selection (see, for example, George and McCulloch (1993) and Smith and Kohn (1996)) involves equating the value of a binary indicator, $\gamma \in \{0, 1\}$, to the restriction that a parameter is set to zero or not. This gives a prior distribution for the parameter that is a discrete-continuous mixture: the prior has a spike at zero with a continuous distribution elsewhere and for this reason is commonly called the *spike-and-slab* prior. The aim in using this prior is to empirically determine whether a parameter is exactly zero, or is drawn from a continuous distribution. Shrinkage priors, on the other hand, do not set parameters to zero but place mass around zero and are continuous everywhere. The advantage of shrinkage priors is that they are generally computationally more efficient. The idea in this approach is that parameters that are really zero will have posteriors that are shrunk towards zero. When this occurs, the parameter is treated as *effectively* zero. The researcher must decide between the computational advantages of the shrinkage priors (with no dimension reduction) and the actual dimension reduction (but computational challenges) of using spike-and-slab priors for variable selection.

Gerlach, Carter and Kohn (2000) and Frühwirth-Schnatter and Wagner (2010), use related approaches to let the data decide which parameters are time varying. The more general and earlier approach is that of Gerlach et al. (2000) in which the time-variation is allowed to switch on and off where necessary. To do this, they developed a dynamic mixture approach that could allow time variation of the states at some points in time but to turn off time variation at other times. This approach uses an indicator variable, $\gamma_t$, that could turn off the shocks to the states. The state equation is specified as $\tilde{\alpha}_t = \tilde{\alpha}_{t-1} + \gamma_t \eta_t$ where $\gamma_t \in \{0, 1\}$. A value of $\gamma_t = 0$ implies the state does not vary from time $t - 1$ to $t$. One way to interpret this is that the shock at time $t$ was zero, $\eta_t = 0$. Another interpretation is that the variance of the shock is zero, $E(\eta_t^2) = 0$. We will return to this latter interpretation shortly.
As the indicator’s value can change over time, computation can prove to be slow or challenging. An important contribution of Gerlach \textit{et al.} (2000) was the development of an efficient algorithm for sampling the indicators by first integrating out the states. The model proposed by Gerlach \textit{et al.} (2000) permits more general specifications than the one we have discussed here and the approach has been applied in a number of macroeconometric studies including Koop \textit{et al.} (2009), Chan, Koop, Léon-González and Strachan (2012), Koop and Korobilis (2012), Groen, Paap and Ravazzolo (2013), Korobilis (2013) and Clark and Ravazzolo (2015).

Frühwirth-Schnatter and Wagner (2010) frame the question of which parameters should vary over time as a variable selection problem. The set up resembles that of Gerlach \textit{et al.} (2000) with the addition of a transformation to the non-centered parameterization, although it is less general in that it does not allow an indicator for each point in time. In Frühwirth-Schnatter and Wagner (2010), the parameter that is to be set to zero if $\gamma = 0$ is the variance of the state equation in which case the time variation is turned off.

To demonstrate this approach, consider a scalar state and a standard form of the measurement and state equations as discussed above with respect to the approach of Gerlach \textit{et al.} (2000), but with a single time-invariant indicator

\begin{align*}
y_t &= \tilde{\alpha}_t + \varepsilon_t \\
\tilde{\alpha}_t &= \tilde{\alpha}_{t-1} + \gamma \eta_t
\end{align*}

where $\gamma \in \{0, 1\}$, $\eta_t \sim N (0, \tilde{\sigma}^2)$ (and so we could roughly write $\gamma \eta_t \sim N (0, \gamma^2 \tilde{\sigma}^2)$) and $\tilde{\alpha}_1 = \alpha \sim N (0, \omega^2)$. By subtracting the initial state from all states and dividing this by $\gamma \tilde{\sigma}$ (we ignore the obvious problem of dividing by zero when $\gamma = 0$) to preserve the conceptual link between Gerlach \textit{et al.} (2000) and Frühwirth-Schnatter and
Wagner (2010)) we have the non-centered parameterization, \( \tilde{\alpha}_t = \alpha + \gamma \tilde{\sigma} \alpha_t \) and the model becomes

\[
\begin{align*}
    y_t &= \alpha + \gamma \tilde{\sigma} \alpha_t + \varepsilon_t \\
    \alpha_t &= \alpha_{t-1} + z_t
\end{align*}
\]

where \( z_t \sim N(0, 1) \) and \( \alpha_1 = 0 \). The constraint \( \tilde{\sigma} > 0 \) is treated as an identification issue in Frühwirth-Schnatter and Wagner (2010) although here we provide a different interpretation and parameterize this identification issue. To see this, introduce an unidentified bivariate random variable \( \iota \in \{-1, 1\} \) where \( P(\iota = 1) = 1 - P(\iota = -1) \) and insert this into the measurement equation as \( \tilde{\sigma} \alpha_t = \tilde{\sigma} i \iota \alpha_t = \sigma \alpha_t \) where \( \sigma = i \tilde{\sigma} \).

The state equation is now \( y_t = \tilde{\alpha}_0 + \gamma \sigma \alpha_t + \varepsilon_t \). Although \( \alpha_t \) has been transformed we do not change the notation as it has the same distribution before and after the transformation. This transformation using the unidentified \( \iota \) is a form of parameter expansion (see discussions in Liu, Rubin and Wu (1998) and Liu and Wu (1999)) that is commonly used to improve computation. An important requirement for a parameter expansion to improve computatoin is that the prior distribution of the unidentified parameter (\( \iota \)) is independent of the parameters in the model, which it is in this case.

The state equation error standard deviation, \( \sigma \), is now just a regression coefficient on the state with support \( \mathbb{R} \). If the state is not varying over time then we would expect to see this coefficient restricted to \( \sigma = 0 \), restricting the ‘variable’ \( \alpha_t \) out of the equation. To implement variable selection the binary indicator \( \gamma \in \{0, 1\} \) in the measurement equation is given a prior, \( P(\gamma = 1) = P(\gamma = 0) = \frac{1}{2} \) say. Thus we see the link to Gerlach et al. (2000) as \( y_t = \tilde{\alpha}_0 + \gamma \sigma \alpha_t + \varepsilon_t \). If \( \gamma = 0 \) then \( \sigma = 0 \) and the state is constant, whereas if \( \gamma = 1 \) then \( \sigma \neq 0 \), and the state may vary.
over time. It is for this reason Bitto and Frühwirth-Schnatter (2019) refer to this as a variance selection problem. The posterior probability $P(\gamma = 0|y)$ tells us the empirical support for the restriction that the state is time invariant.

There have been a number of applications and extensions of the non-centered parameterization for state space models in macroeconometrics. Chan (2018), for example, uses this specification with the Savage-Dickey density ratio to compute the Bayes factor for the restriction to time-invariance of parameter in various models of inflation. Towards the end of this section we discuss an extension and generalization of the specification in Frühwirth-Schnatter and Wagner (2010) by Chan, Eisenstat and Strachan (2020).

Despite its strong theoretical appeal, variable selection tends to be more computationally difficult than the approximation to variable selection through shrinkage. For this reason there have been many approaches inducing shrinkage including Belmonte, Koop, and Korobilis (2014), Kalli and Griffin (2014), Uribe and Lopes (2017), Rockova and McAlinn (2017), Huber, Koop and Onorante (2020) and Bitto and Frühwirth-Schnatter (2019). In the following discussion we consider a few implementations of shrinkage.

An important contribution of Frühwirth-Schnatter and Wagner (2010) is the argument for using a normal prior on the parameter $\sigma$ rather than the traditional default of the inverted gamma on $\tilde{\sigma}^2$. The difference, which is important for questions of parsimony, is in how these two priors treat values near zero. The more commonly used inverted gamma distribution on $\tilde{\sigma}^2$ prevents shrinkage as it is bounded away from zero. The normal prior on $\sigma$, which implies a gamma prior on $\tilde{\sigma}^2$, is more innocuous in that it allows weight around zero. In an inflation forecasting exercise, Belmonte et al. (2014) remove the variable selection parameter ($\gamma$) and use shrinkage priors on the prior variance of the initial state ($\tilde{a}_0$) and the state variance ($\tilde{\sigma}$) in the non-
centered specification. The prior uses a Lasso prior for the standard deviation. Bitto and Frühwirth-Schnatter (2019) take a similar approach to Belmonte et al. (2014) but use a Normal-gamma prior for shrinkage which encompasses the Lasso prior as a special case. The prior in Bitto and Frühwirth-Schnatter (2019) can produce a more pronounced peak at zero while allowing considerable mass away from zero, effectively bringing the continuous shrinkage approach closer to the variable selection or spike-and-slab approach. Giannone, Lenza and Primiceri (2018) explore the support for sparse (as could result from spike-and-slab priors) versus dense (models with many parameters shrunk towards zero) time invariant models. They find support for dense models over sparse models. A paper that brings together selection and shrinkage is Huber et al. (2020) using sparsification (see Barbieri and Berger (2004)) where sparsification involves setting parameters with low posterior probabilities of being non-zero to zero. In a TVP-VAR, and in contrast to the results of Giannone et al. (2018), they find strong evidence of sparsity and that large sparse models forecast well.

Another approach to computing high dimensional state space models is to change the model specification in such a way that results in significant computational and memory savings. We discuss three such approaches. The first of these makes use of techniques developed over fifty years ago when computing power was very limited. In a series of papers Koop and Korobilis (2011, 2012, 2013) (among others) apply this technique in a dynamic model averaging approach suggested by Raftery, Karny and Ettler (2010). The updating equation for the filtered estimate of the state covariance matrix, $V_{t+1}$, in the contemporaneous filtering equations representation is

$$V_{t+1} = V_{t|t} + R\Omega_t R'.$$

Raftery et al. (2010) introduce the forgetting factor $\lambda$, where $0 < \lambda \leq 1$, and replace
the above updating equation with

\[ V_{t+1} = \frac{1}{\lambda} V_{t|t}. \]

If we assume \( \Omega_t = (\lambda^{-1} - 1) V_{t|t} \), then this is the correct updating equation for a properly defined state space model. Using forgetting factors to alter the model specification in this way significantly reduces the computational burden. This computational advantage comes from not needing to compute or simulate \( \Omega_t \). Koop and Korobilis (2013) clearly demonstrate this advantage when they compute a TVP-VAR with four lags for 25 variables, implying a state vector with dimension \( k = 2,525 \). They demonstrate that this is particularly useful in a forecasting exercise.

Koop and Korobilis (2013) avoid a full exploration of the posterior distribution which is appropriate for a forecasting exercise. However, there are important questions that do require knowledge of the full posterior. For example, a simple measures of uncertainty are the highest posterior density or the percentile range (e.g., inter-quartile range) for objects of interest. Chan, Eisenstat and Strachan (2020) present an approach that provides a full MCMC exploration of the posterior for a high-dimensional TVP-VAR with stochastic volatility. This approach also imposes a change in the model specification but one that is suggested by the data. That is, they impose rank reduction on the state equation covariance matrix \( \Omega \) such that they estimate a singular state space model. This feature of the state covariance matrix was noted in Cogley and Sargent (2005). This simple and empirically supported restriction results in a reduction of the dimension of the model by around 90% to 99%, thereby making the model computable. Chan et al. (2020) improve computation by implementing a number of techniques: the rank restriction, they estimate the VAR in structural rather than reduced form, generalize the non-centered parameterization.
of Frühwirth-Schnatter and Wagner (2010) to the matrix case, employ parameter expansions and use the precision sampling approach of Chan and Jeliazkov (2009). They apply this to a study of the evolution of the transmission of news and non-news shocks in a system with 15 variables.

The approach in Chan et al. (2020) works by reducing the dimension of the parameter space by reducing the sources of variation in the states. Koop, Korobilis and Pettenuzzo (2019) achieve a reduction in the dimension of the parameter space by using compressed regression to reduce the dimension of the data. Bayesian compressed regressions, introduced by Guhaniyogi and Dunson (2015), randomly project the data into a lower dimension. Importantly, the draws of these projections are not informed by the data. To demonstrate this in the linear regression model for a \((T \times 1)\) vector \(y\) with a \(T \times k\) matrix of regressors \(X\), the compression is applied as follows

\[
y = X\beta + \varepsilon = X\Phi\gamma + \varepsilon = X^{\Phi}\gamma + \varepsilon
\]

where \(X^{\Phi} = X\Phi\), \(\gamma\) is an \((q \times 1)\) vector and \(\Phi\) is the \((k \times q)\) random projection matrix. Each element of \(\Phi\) depends upon a single scalar parameter, \(\nu\), which is drawn independently of the data. The \((i, j)^{th}\) element of \(\Phi\), \(\Phi_{ij}\), takes one of three discrete values determined by the probability of each value. Specifically, \(\Phi_{ij} \in \{-\frac{1}{\sqrt{\nu}}, 0, \frac{1}{\sqrt{\nu}}\}\) with probabilities \(\nu^2\), \(2(1 - \nu)\nu\) and \((1 - \nu)^2\) respectively. The unknown parameter \(\nu\) might be drawn, for example, from a uniform distribution bounded away from zero and one for numerical stability. When choosing \(q\), Guhaniyogi and Dunson (2015) suggest using Bayesian Model Averaging. Koop et al. (2019) apply this to the TVP-VAR by combining it with dynamic model averaging. By using the structural form of the VAR, they apply compression to each equation one at a time and the computational result is impressive. They are able to apply this to a system with \(n = 129\) variables.
and 13 lags. The state vector for a full TVP-VAR with these specifications would have dimension over $k > 210,000$. The approach is demonstrated to be empirically useful in a forecasting exercise in which it performs well.

In this section we have discussed various approaches to dealing with parameter proliferation which often occurs in state space models used in macroeconomics. The range of approaches involve shrinkage, variable selection or alterations to the structure of the model to either reduce its dimension or permit easier computation. As models grow and the advantages of large models become more apparent and better understood, this is likely to remain a very active area of research.

4 Non-linear and non-Gaussian models

An issue that arises in a range of models is that the states $\pi_t$ evolve nonlinearly, enter the measurement equation non-linearly and/or $\varepsilon_t$ or $\eta_t$ are non-Gaussian. To allow for these possibilities, we write the measurement and state equations in a more general form as

$$y_t = F(\pi_t, \varepsilon_t) \text{ for } t > 1$$

$$\pi_t = H(\pi_{t-1}, \eta_t) \text{ for } t > 1$$

$$\pi_1 \sim f(\pi_1)$$
or, slightly more generally again,

\[ y_t \sim p(y_t | \pi_t) \text{ for } t > 1 \]  \hspace{1cm} (6)

\[ \pi_t \sim f(\pi_t | \pi_{t-1}) \text{ for } t > 1 \]  \hspace{1cm} (7)

\[ \pi_1 \sim f(\pi_1). \]  \hspace{1cm} (8)

In these cases the Kalman filter and smoother cannot be employed or require extensions to enable inference. A number of alternative approaches have been proposed and in this section we discuss three such approaches. The first involves respecifying the model into a linear Gaussian form such that the Kalman filter can be used. Next we outline an alternative filtering algorithm that allows for very general forms for the state space model, the particle filter. Lastly, we discuss another approach that draws upon the precision sampler approach discussed in Section 2.

In some cases, it is possible to respecify a non-linear and non-Gaussian model into a linear Gaussian form to permit application of standard estimation procedures. An example is the stochastic volatility model. A simple specification (e.g., Jacquier, Polson and Rossi (1994) or Kim, Shephard and Chib (1998)) of the measurement and state equations for stochastic volatility is

\[ y_t = e^{h_t/2} \varepsilon_t, \]

\[ h_t = \mu + \rho h_{t-1} + \eta_t, \]

where \( \varepsilon_t \sim iidN(0, 1) \). The measurement equation is non-linear in the state, \( h_t \), but Gaussian. Kim et al. (1998) apply a sequence of transformations to obtain a linear and (conditionally) Gaussian form. They begin by squaring both sides and then taking the log. The actual transformation of \( y_t \) is \( y_t^* = \ln (y_t^2 + c) \) where \( c > 0 \)
is a small number added to induce stability. The resulting measurement equation is now

\[ y_t^* = h_2 + \varepsilon_t^* \]

where \( \varepsilon_t^* = \ln \left( \varepsilon_t^2 \right) \sim \ln \left( \chi^2_1 \right) \).

The error \( \varepsilon_t^* \) is distributed as \( \ln \left( \chi^2_1 \right) \) which is clearly non-Gaussian. The next step is to approximate the distribution \( p(\varepsilon_t^*) \) by a mixture of normals, \( p(\varepsilon_t^*) \simeq \sum_{i=1}^{7} w_i \cdot N(\mu_i, \sigma_i^2) \).

By drawing from each of the seven normals with probability \( w_i \) the overall draws are approximately draws from \( p(\varepsilon_t^*) \). To implement this, a random latent state is introduced for each point in time \( t \), \( s_t = i \) for \( i = 1, \ldots, 7 \) and \( \Pr(\ s_t = i \) \) = \( w_i \). Conditional upon the drawn state \( s_t = i \), \( \varepsilon_t^* \sim N(\mu_i, \sigma_i^2) \) and the model is now linear and Gaussian which permits the use of the Kalman filter or smoother. Introducing the latent state or auxiliary variable \( s_t \) results in a significant simplification of the sampling algorithm.

This approach - the auxiliary mixture sampling - has been applied widely. Examples include Frühwirth-Schnatter and Wagner (2006), Omori, Chib, Shephard and Nakajima (2007), Frühwirth-Schnatter, Frühwirth and Rue (2009), Barra, Borowska and Koopman (2018) and Mao and Zhang (2018).

In cases in which the measurement or state equations are non-linear, and such a transformation to a linear form is not possible or practical, then the Extended Kalman Filter (EKF) (Anderson and Moore (1979), Jazwinski (1970)) has often proven useful. The EKF essentially involves taking a locally linear approximations to the non-linear functions \( F \) and \( H \). This linearization uses a first order Taylor series expansion of \( F \) and \( H \) to develop linear approximations to these functions which are then used in Gaussian approximations to the predictive distributions, \( p(\pi_t|y^{1:t-1}) \) and \( p(\pi_t|y^{1:t}) \).

As with any approach, a number if issues have been identified with the EKF (Creal (2012)). For example, if the true predictive distributions are not well approximated by Gaussian distributions (e.g., if they are bi-modal or heavily skewed) then the EKF
will not perform well (Arulampalam, Maskell, Gordon, and Clapp (2002)). In the case of non-linear measurement and state equations with possibly non-Gaussian shocks, the particle filter approach can be used and have some advantages over the EKF in that there is no need for functional approximation.

Particle filters or sequential Monte Carlo samplers (Del Moral, Doucet and Jasra (2006) and Doucet, de Freitas and Gordon (2001)) are a powerful tool for estimating a very wide range of processes and so have proven useful in many recent applications. The price to pay for this power and flexibility is computational burden, but this cost has been decreasing or, equivalently the computational power has been increasing. This latter phenomenon, along with advances in algorithms, have made particle filters increasingly viable. Accessible tutorials on particle filters are given in Arulampalam \textit{et al.} (2002) or, more recently, Doucet and Johansen (2011) and a recent survey by Creal (2012).

The particle filter uses importance sampling and resampling to draw from an approximation to a sequence of probability distributions. Unlike the EKF which uses an approximation of \( F \) and \( H \), the particle filter approximates the distributions \( p \) and \( f \). This has advantages in cases where the local linearization and Gaussian approximation poorly match the true distribution. Another useful feature of this technique is that the probability distributions need only be known up to the normalizing constant and are approximated by a ‘cloud’ of weighted random samples. The weights \( w_t \) are attached to each sampled value to approximate the density of the states. Sequential Monte Carlo methods are employed to propagate these samples, or particles, over time. An empirically and theoretically important problem with sequential importance sampling is that the mass is degenerate. That is, as the algorithm progresses, one weight converges to one while the other weights collapse to zero. A solution to this issue using a Bootstrap filter was proposed by Gordon, Salmon and Smith (1993).
These authors propose a resampling step to obtain equally weighted particles. With independent propagation the degeneracy problem is avoided.

To give a feel for the flexibility of the particle filtering approach, recall the the state space model in a very general form (6)-(8), one that is potentially non-Gaussian and non-linear. Applying the particle filter at step $t$ involves drawing a number of potential states $\pi_t$, keeping those that are most likely or probable, and then propagating forward to $\pi_{t+1}$. A simple version of the Bootstrap filter of Gordon et al. (1993) begins with $m$ draws of the initial state $\pi_1$.

1. **Initialize:**

   - For $i = 1, \ldots, m$ draw $\pi^{(i)}_1 \sim f(\pi_1)$ and set $w^{(i)}_1 = \frac{1}{m}$.

   For $t = 2, \ldots, T$

2. **Importance sampling:**

   - For $i = 1, \ldots, m$ draw $\tilde{\pi}^{(i)}_t \sim f(\pi_t|\pi^{(i)}_{t-1})$ and set $\pi^{(i)}_{1:t} = \left(\frac{\pi^{(i)}_1}{\pi^{(i)}_{t-1}}, \pi^{(i)}_t\right)$

   - For $i = 1, \ldots, m$ compute and normalize the weights $w^{(i)}_t \propto p(y^i_t|\pi_t^{(i)})$.

3. **Resampling:**

   - Resample $m$ particles $\{\pi^{(i)}_t\}_{i=1}^m$ with replacement from $\left\{\pi^{(i)}_t, w^{(i)}_t\right\}_{i=1}^m$

   - Return to Step 2.

From the notation used to demonstrate the particle filter, it is clear that there is no theoretical constraint on the dimension of the state $\pi_t$. In practice, however, the approximation error grows exponential in the dimension of the state (Rebeschini and van Handel (2015)). There is generally much greater variation in the weights with
higher dimension state vectors and so one particle will take almost all of the weight; leading again to a degeneracy issue. There have been a number of papers developing methods to address degeneracy in high dimensional filtering methods (see, for example, Djurić and Bugallo (2013), Naesseth, Lindsten and Schön (2015), Rebeschini and van Handel (2015), Beskos, Crisan, Jasra, Kamatani and Zhou (2017), Robert and Künsch (2017) and Wigren, Murray and Lindsten (2018)). However, there remain few applications of such techniques to what could be regarded as high-dimensional states in macroeconomics.

Above we mentioned the advantage of approximating the distributions rather than the functions in the model and in Section 2 we discussed the computational advantages of the precision sampler (in which we draw all states at once) over the filtering approach. This computational gain becomes more important if we take into account that there is clearly a need for modelling high-dimensional systems in macroeconomics. Chan and Strachan (2012) develop an approach to drawing states from a possibly non-linear and non-Gaussian state space model of (moderately) high dimension. Their application is to evaluation of credit shocks near the zero lower bound for interest rates. For this application, they estimate a TVP-VAR-SV where the non-linearity and non-Gaussianity are induced by the truncated support of interest rates at zero. Using second order expansions and Gaussian approximations to the likelihood $p(y|\pi)$ and the density for $p(\pi)$ implied by the state equations, they develop an approximation to the posterior for the states, $p(\pi|y)$. They then develop three sampling schemes for drawing the states from the posterior: a Metropolis-Hastings algorithm, an Accept-Reject Metropolis-Hastings algorithm and finally a collapsed sampler using the cross-entropy method (e.g., Chan and Kroese (2012)) to develop the proposal density. This general approach of estimating non-linear state space models has been used in various applications such as Chan, Koop and Potter (2013 and

5 Summary and conclusion

In this paper, we have provided a review of recent developments in state space modelling in macroeconomics taking a Bayesian approach. For the purpose of estimating a linear Gaussian state space model, the Kalman filter remains the most popular choice for analysts. The precision based samplers show promise as an alternative to the Kalman smoother, particularly when economists prefer larger models for empirical or theoretical reasons, however, a return of dynamic model averaging using forgetting factors shows considerable potential for estimation of large models. These latter models have been shown to perform very well in forecasting exercises. While we found only application of the Kalman filter using forgetting factors, there does not appear to be any impediment to applying the smoother with this approach and doing so could resolve some interpretational issues as well as broaden its application. As the rationale for large models has strengthened, the focus on computational costs has increased leading to approaches that save on computation by dimension reduction - such as turning off time-variation - or better algorithms, or both. Advances in the theory around particle filters has improved inference in non-linear and non-Gaussian models but challenges remain when applying this approach in high-dimensions. A precision based approximation to the posterior for the states, however, shows promise in some applications.

Recent interest has turned to large models, including time-varying parameter models where state space models dominate. There have been considerable advances into computation, specification and inference in these high-dimensional models and this
research continues. Challenges, and so new directions for state space modelling, include estimation and inference in TVP-VARMAs and treatment of identification in large factor models; both of which have theoretical justifications when the underlying model is assumed to be a DSGE model. This takes us beyond using these models only for forecasting and evidence on time variation. Once VARs and VECMs were established as workhorse models, research turned to consider restrictions on these models as a way to establish support for economic theories or features implied by economic theories. Similarly, evidence on time-variation in the support for restrictions (see, for example, Koop, Léon-González and Strachan (2010, 2011)) offers further opportunities for new applied work.
6 References


Groen, J., Paap, R., Ravazzolo, F. 2013. Real-time inflation forecasting in a changing


