Lecture III: State-space models

Dimitris Korobilis[a,b]

[a]University of Essex, UK

[b]Rimini Center for Economic Analysis, Italy

Barcelona Graduate School of Economics
29 June 2018

Outline

SS models

Popular SS models

Bayesian estimation

Flexible SS models

Machine Learning Inference

State-space methods

- State space methods are used for a wide variety of time series problems
- Very flexible specification for problems with latent (unobserved) dynamics
- E.g. time-varying parameters, dynamic factors, stochastic volatilities, and Markov switching states are all example of latent state variables which can be estimated using state-space methods
- Advantage of state space models: well-developed set of MCMC algorithms for doing Bayesian inference

The normal linear state-space model

Let y_t be observed (given) data, and F, H, Q, R be known (given) coefficient matrices.
The normal linear state-space model is of the form

$$\begin{aligned} y_t &= Hx_t + v_t \qquad (1) \\ x_t &= Fx_{t-1} + u_t \qquad (2) \end{aligned}$$

where x_t is the latent (unobserved) state that we need to estimate.

- $v_t \sim N(0, R)$ and $u_t \sim N(0, Q)$ are Gaussian (Normal)
- The first equation is the “measurement equation”
- The second equation is the “state equation”
- The system can be estimated only given an initial condition for x_0
- We assume that $x_0 \sim N(\underline{x}, \underline{V})$

The normal linear state-space model

$$\begin{aligned} y_t &= Hx_t + v_t & (3) \\ x_t &= Fx_{t-1} + u_t & (4) \\ x_0 &\sim N(\underline{x}, \underline{V}) & (5) \end{aligned}$$

Since F, H, Q, R and y_t are given (known), we can estimate x_t *recursively* (i.e. iteratively, for $t = 1, ..., T$).

The solution to estimating x_t can be given by the Kalman filter recursion

In the next we present the iterations over $t = 1$ to T, where the notation $x_{t|s}$ means "x_t given information up to time s"

Alternatively we can write: $x_{t|s} = x_t | y_1, ..., y_s$

The Kalman filter

for $t = 1$ to T

<u>Predict step:</u>

$$\begin{aligned} x_{t|t-1} &= F x_{t-1|t-1} && (6) \\ P_{t|t-1} &= F P_{t-1|t-1} F' + Q && (7) \end{aligned}$$

<u>Update step:</u>

$$\begin{aligned} v_{t|t-1} &= y_t - H x_{t|t-1} && (8) \\ K_t &= P_{t|t-1} H' (R + H P_{t|t-1} H')^{-1} && (9) \\ x_{t|t} &= x_{t|t-1} + K_t v_{t|t-1} && (10) \\ P_{t|t} &= P_{t|t-1} + K_t F P'_{t|t-1} && (11) \end{aligned}$$

where we use $x_{0|0} = \underline{x}$ and $P_{0|0} = \underline{V}$

Comments on the Kalman Filter

- We present the simple case of the state-space (SS) model:

$$y_t = Hx_t + v_t \tag{12}$$
$$x_t = Fx_{t-1} + u_t \tag{13}$$

But the formulas can be generalized to the case where:

$$y_t = H_t x_t + v_t \tag{14}$$
$$x_t = F_t x_{t-1} + u_t \tag{15}$$

and $v_t \sim N(0, R_t)$ and $u_t \sim N(0, Q_t)$

- As long as all these matrices are given, we can simply insert (each time t) in the formulas above the values of H_t, F_t, R_t, Q_t
- This fact gives us great flexibility: H_t or F_t do not have to be "parameters", rather they can be e.g. data matrices
- Keep this in mind for when we specify time-varying VARs
- SS form is very flexible and can give us solutions to many models - we will see examples below

The Kalman smoother

- Estimates from the Kalman filter can be quite “noisy”.
- Many times we want to use future data to update parameter estimates.
- In this case we are using a technique called smoothing

Example: Rauch-Tung-Striebel smoother

for t = T-1 to 1

$$x_{t|t+1} = x_{t|t} + P_{t|t}F'P_{t+1|t}^{-1}\left(x_{t+1|t+1} - Fx_{t|t}\right) \tag{16}$$

$$P_{t|t+1} = P_{t|t} - P_{t|t}F'P_{t+1|t}^{-1}FP_{t|t} \tag{17}$$

Then we set $x_1 = x_{1|2}$,..., $x_t = x_{t|t+1}$,..., $x_{T-1} = x_{T-1|T}$, and finally for the last observation: $x_T = (x_{T|T+1} =)x_{T|T}$.

Example 1: Dynamic Factor model

$$\begin{aligned} y_t &= \Lambda f_t + \nu_t \quad (18) \\ f_t &= \Phi f_{t-1} + u_t \quad (19) \end{aligned}$$

- f_t are the latent (unobserved) dynamic factors
- Λ is the loadings matrix, Φ are (V)AR coefficients

Stock and Watson (several papers) approximate f_t using PCA, and then they use OLS to estimate Λ, Φ and the covariances (PCA is consistent estimator of the static factors).
→ Assume we know the coefficients, then we can estimate f_t with one run of the Kalman filter/smoother recursions

Example 1: Dynamic Factor model (con'd)

What to do if we have a DFM with, say, 3 lags?

$$
\begin{aligned}
y_t &= \Lambda f_t + \nu_t && (20) \\
f_t &= \Phi_1 f_{t-1} + \Phi_2 f_{t-2} + \Phi_3 f_{t-3} + u_t && (21)
\end{aligned}
$$

→ We only need to convert the second equation to a (V)AR(1) model (companion form):

$$
\begin{bmatrix} f_t \\ f_{t-1} \\ f_{t-2} \end{bmatrix} = \begin{bmatrix} \Phi_1 & \Phi_2 & \Phi_3 \\ I & 0 & 0 \\ 0 & I & 0 \end{bmatrix} \begin{bmatrix} f_{t-1} \\ f_{t-2} \\ f_{t-3} \end{bmatrix} + \begin{bmatrix} \nu_t \\ 0 \\ 0 \end{bmatrix} \tag{22}
$$

and convert the measurement equation accordingly:

$$
\begin{bmatrix} y_t \\ f_{t-1} \\ f_{t-2} \end{bmatrix} = \begin{bmatrix} \Lambda & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix} \begin{bmatrix} f_t \\ f_{t-1} \\ f_{t-2} \end{bmatrix} + \begin{bmatrix} u_t \\ 0 \\ 0 \end{bmatrix} \tag{23}
$$

Now we have a state-space form, Kalman filter can be applied!

Example 2: Stochastic volatility model 1

$$
\begin{aligned}
y_t &= \sigma_t v_t && (24) \\
\log \sigma_t^2 &= \log \sigma_{t-1}^2 + u_t && (25)
\end{aligned}
$$

This is a nonlinear state-space model, however, we need to use $\log()$ to ensure that the volatility is positive

→ σ_t^2 must always be positive so using a state equation of the form

$$\sigma_t^2 = \sigma_{t-1}^2 + u_t \qquad (26)$$

does not guarantee that σ_t^2 will positive.

This is not a problem, as long as the model can be written in a *conditionally* linear form, Kalman filter is still possible.

Example 2: Stochastic volatility model 2

$$\begin{aligned} y_t &= \sigma_t v_t && (27) \\ \log \sigma_t^2 &= \log \sigma_{t-1}^2 + u_t && (28) \end{aligned}$$

Take squares and then logarithms on the measurement equation we have:

$$\begin{aligned} y_t^2 &= \sigma_t^2 v_t^2 \Rightarrow && (29) \\ \log y_t^2 &= \log \sigma_t^2 + \log v_t^2 && (30) \end{aligned}$$

Final state-space form of the stochastic volatility model:

$$\begin{aligned} \log y_t^2 &= \log \sigma_t^2 + \log v_t^2 && (31) \\ \log \sigma_t^2 &= \log \sigma_{t-1}^2 + u_t && (32) \end{aligned}$$

Example 2: Stochastic volatility model 3

$$\begin{aligned} \log y_t^2 &= \log \sigma_t^2 + \log \nu_t^2 \qquad (33) \\ \log \sigma_t^2 &= \log \sigma_{t-1}^2 + u_t \qquad (34) \end{aligned}$$

Set $z_t = \log y_t^2$ and $h_t = \log \sigma_t^2$ and $\epsilon_t = \log \nu_t^2$. Then we have the following general state-space form:

$$\begin{aligned} z_t &= h_t + \epsilon_t \qquad (35) \\ h_t &= h_{t-1} + u_t \qquad (36) \end{aligned}$$

- This model is linear, but it is **not** Normal: $\epsilon_t = \log \nu_t^2$ follows a $\log - \chi^2$ distribution with 1 degree of freedom iff ν_t is univariate Normal (Gaussian).
- Solution of Kim, Shephard and Chib (1998, REStud): approximate ϵ_t using a mixture of Normal distributions
- Then the State-Space form above is going to be *conditionally* Normal, and Kalman filter can be applied.

Example 2: Stochastic volatility model 4

A digression:
A mixture of Normals with C components, has the following form

$$F(\bullet) \sim \pi_1 N(a_1, V_1) + \pi_2 N(a_2, V_2) + ... + \pi_C N(a_C, V_C) \qquad (37)$$

where $\sum_{c=1}^{C} \pi_c = 1$ and π_c are non-negative probabilities.

Kim, Shepard and Chib (1998) show that

$$\begin{aligned} F(\epsilon_t) &\approx 0.0073N(-10.1, 5.8) + 0.10N(-3.9, 2.6) \qquad (38) \\ &+ 0.00002N(-8.6, 5.2) + .044N(2.8, 0.2) \\ &+ 0.34N(0.6, 0.6) + 0.24N(1.8, 0.3) + 0.26N(-1, 1.3) \end{aligned}$$

i.e. the $\log - \chi^2$ can be approximated using a 7-component mixture of Normals

Example 3: TVP-VAR

Consider the VAR:

$$y_t = z_t \beta + \varepsilon_t \tag{39}$$

where z_t contains the intercept and lags (and probably deterministic and exogenous variables), $\varepsilon_t \sim N(0, \Sigma)$.

A flexible way to allow for structural instability in the VAR coefficients, is to use the model:

$$\begin{aligned} y_t &= z_t \beta_t + \varepsilon_t & (40) \\ \beta_t &= \beta_{t-1} + \eta_t & (41) \end{aligned}$$

where $\eta_t \sim N(0, Q)$.

Given knowledge of Σ and Q, we can estimate β_t using the Kalman filter/smoother.

Covariances and other parameters in a SS framework

$$\begin{aligned} y_t &= Hx_t + v_t && (42) \\ x_t &= Fx_{t-1} + u_t && (43) \end{aligned}$$

with $v_t \sim N(0, R)$ and $u_t \sim N(0, Q)$ are Gaussian (Normal)

- In this system, the state x_t is unknown, but so are the parameters H, F, R, Q
- It would be far from realistic to assume that these are known (unless imposed by the model)
- For instance, in time-varying parameter models we assume that $F = I$ (see previous slide)
- When estimating linear Gaussian SS models, the Gibbs sampler can be used
- For nonlinear SS formulations (e.g. DSGE models), more complex computational methods are needed

Estimation using the Gibbs sampler

$$y_t = Hx_t + v_t \tag{44}$$

$$x_t = Fx_{t-1} + u_t \tag{45}$$

- It turns out that the system can be estimated using the conditioning arguments of the Gibbs sampler
- Draw x_t using the Kalman filter and smoother given values of H, F, R, Q
- Draw H, F, R, Q given the value of x_t, using standard formulas from regression and (V)AR models.
- The only adaption we need to make to the Kalman filter (and smoother) is that once we obtain $x_{t|t}$ (or $x_{t|t+1}$ if using smoother) is to draw from

$$x_t^{[s]} \sim N(x_{t|t}, P_{t|t}) \tag{46}$$

Estimation using the Gibbs sampler

- Additionally, instead of using the Rauch-Tung-Striebel smoother, people typically use the Carter and Kohn (1994) "Simulation Smoother"
- This simply takes into account the fact that we now use random draws of x_t, instead of working with just means and variances
- I provide in the MATLAB examples a functions "$carter_kohn$" that provides this procedure
- In detail you need to run the Kalman filter iterations, then sample $x_T \sim N(x_{T|T}, P_{T|T})$
- Then run the backward recursions (for $t = T - 1$ to 1)

$$
\begin{aligned}
x_{t|t+1} &= x_{t|t} + P_{t|t}F'P_{t+1|t}^{-1}(x_{t+1} - Fx_{t|t}) && (47) \\
P_{t|t+1} &= P_{t|t} - P_{t|t}F'P_{t+1|t}^{-1}FP_{t|t} && (48)
\end{aligned}
$$

where we draw $x_t \sim N(x_{t|t+1}, t|t+1)$ (note x_{t+1} enters formula for $x_{t|t+1}$)

Example: Gibbs sampler for the TVP-VAR

The TVP-VAR with homoskedastic covariance is

$$\begin{aligned} y_t &= z_t \beta_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, \Sigma) && (49) \\ \beta_t &= \beta_{t-1} + \eta_t, \quad \eta_t \sim N(0, Q) && (50) \end{aligned}$$

Assume you specify initial condition $\beta_0 \sim N(0, 4I)$ and priors $\Sigma^{-1} \sim W(\underline{\nu}, \underline{S}^{-1})$ and $Q^{-1} \sim W(\underline{\nu}_Q, \underline{S}_Q^{-1})$ then sample from

1. $\beta_t|\Sigma, Q, data \sim N(\beta_{t|t+1}, P_{t|t+1})$ (from the Kalman filter/smoother)
2. $\Sigma^{-1}|\beta_t, Q, data \sim$ $W(T + \underline{\nu}, (\underline{S} + \sum_{t=1}^{T}(y_t - z_t\beta_t)'(y_t - z_t\beta_t))^{-1})$
3. $Q|\Sigma^{-1}, \beta_t, data \sim$ $W(T - 1 + \underline{\nu}_Q, (\underline{S}_Q + \sum_{t=1}^{T-1}(\beta_t - \beta_{t-1})'(\beta_t - \beta_{t-1}))^{-1})$

Examples of other SS models

- I have already said that the SS form is very powerful and can allow you to consider several extensions
- I am going to give here some more difficult examples, and these are only some of all the possibilities you can use
- There is so much theory of state-space models, and so many different estimation methods
- I can't cover these in one lecture, so what I am going to do here is demonstrate how different authors have exploited the linear state-space form
- This can show you how to do the same and come up with models that will allow you to write good research papers
- While there several such examples (e.g. Siem-Jan Koopman has a vast work on different SS models using maximum likelihood), it is true that the Bayesian approach makes estimation easy

Unobserved Components Stochastic Volatility (UCSV)

Stock and Watson (2007) propose the following model for inflation dynamics:

$$\pi_t = \tau_t + \eta_t \tag{51}$$

$$\tau_t = \tau_{t-1} + \epsilon_t \tag{52}$$

- This is also called a "local-level" model in the SS literature
- τ_t is a slowly changing component (trend); η_t is the random component (disturbance) of π_t
- Stock and Watson (2007) extend this model by assuming $\eta_t \sim N(0, \sigma^2_{\eta,t})$ and $\epsilon_t \sim N(0, \sigma^2_{\epsilon,t})$
- That not only the level of inflation has stochastic volatility $(\sigma^2_{\eta,t})$, but also trend inflation $(\sigma^2_{\epsilon,t})$
- It is interesting to examine estimation of this model since it involves **three** state variables, $\tau_t, \sigma^2_{\eta,t}, \sigma^2_{\epsilon,t}$

Unobserved Components Stochastic Volatility (UCSV) 2

The full model is

$$
\begin{aligned}
\pi_t &= \tau_t + \eta_t && (53)\\
\tau_t &= \tau_{t-1} + \epsilon_t && (54)\\
\log\sigma^2_{\eta,t} &= \log\sigma^2_{\eta,t-1} + \nu_{\eta,t} && (55)\\
\log\sigma^2_{\epsilon,t} &= \log\sigma^2_{\epsilon,t-1} + \nu_{\epsilon,t} && (56)
\end{aligned}
$$

Note that SW (2007) define $\nu_t = (\nu_{\eta,t}, \nu_{\epsilon,t}) \sim N(0, \gamma I_2)$ and fix $\gamma = 0.2$, but they could have easily estimated this parameter They show that this UCSV model forecats inflation well, and they prefer MCMC methods because estimation is much simplified

Unobserved Components Stochastic Volatility (UCSV): MCMC estimation

1. Conditional on $\sigma_{\eta,t}, \sigma_{\epsilon,t}$ estimate τ_t by applying the Kalman filter and simulation smoother (Carter and Kohn, 1994, algorithm) on the SS model:

$$\begin{aligned} \pi_t &= \tau_t + \eta_t && (57) \\ \tau_t &= \tau_{t-1} + \epsilon_t && (58) \end{aligned}$$

2. Conditional on $\tau_t, \sigma_{\epsilon,t}$ estimate $\sigma_{\eta,t}$ using the SS model

$$\begin{aligned} \log((\pi_t - \tau_t)^2) &= \log\sigma^2_{\eta,t} + \log(\eta^2_t) && (59) \\ \log\sigma^2_{\eta,t} &= \log\sigma^2_{\eta,t-1} + \nu_{\eta,t} && (60) \end{aligned}$$

3. Conditional on $\tau_t, \sigma_{\eta,t}$ estimate $\sigma_{\epsilon,t}$ using the SS model

$$\begin{aligned} \log((\tau_t - \tau_{t-1})^2) &= \log\sigma^2_{\epsilon,t} + \log(\epsilon^2_t) && (61) \\ \log\sigma^2_{\epsilon,t} &= \log\sigma^2_{\epsilon,t-1} + \nu_{\epsilon,t} && (62) \end{aligned}$$

Flexible forms of time-variation in coefficients: Cooley and Prescott (1976)

Cooley and Prescott (1976) define the following time-varying parameter regression model:

$$\begin{aligned} y_t &= x_t \beta_t + \epsilon_t && (63) \\ \beta_t &= \beta_t^p + \nu_t && (64) \\ \beta_t^p &= \beta_{t-1}^p + u_t && (65) \end{aligned}$$

- β_t is the time-varying regression coefficient, assume $\epsilon_t \sim N(0, \sigma^2)$ for simplicity
- β_t^p is the permanent and persistent component of the regression coefficients (follows random walk)
- ν_t is the transitory and random componne of the regression coefficients ($\nu_t \sim N(0, R)$)
- This is a "hierarhical state-space" model; will explain estimation when I teach the time-varying panel VAR tomorrow

Flexible forms of time-variation in coefficients: Belmonte, Koop and Korobilis (2014)

Belmonte, Koop and Korobilis (2014) define the following time-varying parameter regression model:

$$\begin{aligned} y_t &= x_t \beta_t + \epsilon_t & (66) \\ \beta_t &= \beta_{t-1} + u_t, \beta_0 \sim N(b_0, V_0) & (67) \end{aligned}$$

and they note that an equivalent form is

$$\begin{aligned} y_t &= x_t c + x_t c_t + \epsilon_t & (68) \\ c_t &= c_{t-1} + u_t, c_0 \sim N(0, 0) \equiv 0 & (69) \end{aligned}$$

where it holds $\beta_t = c + c_t$, thus we decompose regression coefficients into constant component, and the time-varying component

Flexible forms of time-variation in coefficients: Belmonte, Koop and Korobilis (2014) 2

•This means that now the initial condition β_0 becomes a regression coefficient, c, and we can put any prior on it

• BKK (2014) explore shrinkage priors on c and c_t which allow to shrink the model to:

- Constant parameter regression when $c_t = 0$
- Regularized TVP regression when $c = 0$
- Regularized regression when both $c = c_t = 0$
- Unrestricted TVP regression when both c, c_t not zero

Flexible forms of time-variation in coefficients: Giordani and Kohn (2008)

Giordani and Kohn (2008) specify a flexible dynamic mixture specification which allows to model structural breaks

$$
\begin{aligned}
y_t &= x_t\beta_t + \epsilon_t \qquad (70)\\
\beta_t &= \beta_{t-1} + K_t u_t, \qquad (71)
\end{aligned}
$$

- K_t can be a Bernoulli variable (0/1) or a multinomial one (1,2,3,4) etc. Gerlach, Carter and Kohn (2000) show how to sample K_t efficiently
- I provide code for this model, but not estimation details (see JASA paper of GCK (2000), which has lots of equations!)
- Giordani and Kohn (2008) use K_t to take values 0/1
- If $K_t = 0$, then $\beta_t = \beta_{t-1}$, i.e. constant coefficient
- If $K_t = 1$, then $\beta_t = \beta_{t-1} + u_t$, i.e. time-varying coefficient

Flexible forms of time-variation in coefficients: Koop and Potter (2007)

Koop and Potter (2007) specify a structural break model inspired from time-varying parameter models

$$\begin{aligned} y_t &= x_t \beta_{s_t} + \epsilon_t && (72) \\ \beta_{s_t} &= \beta_{s_{t-1}} + u_{s_t}, && (73) \end{aligned}$$

- s_t is a Markov Switching variable which indexes regimes $(s_t \in 1, 2, 3, ..., K$
- The prior on s_t determines the number of breaks: Geometric prior gives K low (e.g. 3-4), Poisson prior makes $K \to T$
- In the extreme case where $K = T$ regimes are identified, then we have TVP model
- In the extreme case where $K = 0$ regimes are identified, then constant parameters

Flexible forms of time-variation in coefficients: Koop and Potter (2010)

Koop and Potter (2010) specify a very flexible time-varying parameter model

$$\begin{aligned} y_t &= x_t \beta_t + \epsilon_t && (74) \\ \beta_t &= \beta_{t-1} + d(z_t, z_{t-1}) u_t, && (75) \end{aligned}$$

- In this specification $d(\bullet)$ is a distance function
- z_t is an index variable (e.g. time t, or exogenous variable)
- This specification can nest a vast amount of popular specifications

Flexible forms of time-variation in coefficients: Koop and Potter (2010) 2

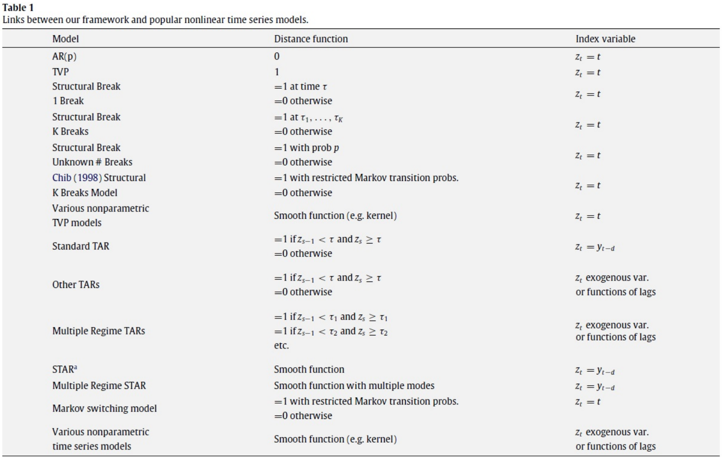

Table 1
Links between our framework and popular nonlinear time series models.

Model	Distance function	Index variable
AR(p)	0	$z_t = t$
TVP	1	$z_t = t$
Structural Break 1 Break	=1 at time τ =0 otherwise	$z_t = t$
Structural Break K Breaks	=1 at $\tau_1, \ldots, \tau_K$ =0 otherwise	$z_t = t$
Structural Break Unknown # Breaks	=1 with prob p =0 otherwise	$z_t = t$
Chib (1998) Structural K Breaks Model	=1 with restricted Markov transition probs. =0 otherwise	$z_t = t$
Various nonparametric TVP models	Smooth function (e.g. kernel)	$z_t = t$
Standard TAR	=1 if $z_{s-1} < \tau$ and $z_s \geq \tau$ =0 otherwise	$z_t = y_{t-d}$
Other TARs	=1 if $z_{s-1} < \tau$ and $z_s \geq \tau$ =0 otherwise	z_t exogenous var. or functions of lags
Multiple Regime TARs	=1 if $z_{s-1} < \tau_1$ and $z_s \geq \tau_1$ =1 if $z_{s-1} < \tau_2$ and $z_s \geq \tau_2$ etc.	z_t exogenous var. or functions of lags
STAR[a]	Smooth function	$z_t = y_{t-d}$
Multiple Regime STAR	Smooth function with multiple modes	$z_t = y_{t-d}$
Markov switching model	=1 with restricted Markov transition probs. =0 otherwise	$z_t = t$
Various nonparametric time series models	Smooth function (e.g. kernel)	z_t exogenous var. or functions of lags

Evaluation of these extenstions

- There are so many such examples in macro/finance
- Common place is that conditional on some parameters, which might be key in each specification, we have a linear SS model to deal with
- E.g. in Giordani and Kohn (2008) conditional on knowing value of K_t, we have a linear SS model
- Similar with Koop and Potter (2007, 2010) where we condition on s_t or $d(z_t)$
- In other models we might have two or more SS models so we need to write the correct SS for each state variable, conditional on others (Stock and Watson, 2007, model; Cooley and Prescott, 1976, model)

Variational Bayes methods

- Variational approximations is a body of deterministic techniques for making approximate inference for parameters in complex statistical models
- The name “variational approximations” has its roots in the mathematical topic known as variational calculus.
    - ♣ Variational calculus is concerned with the problem of optimizing a functional over a class of functions on which that functional depends.
- Variational approximations are useful for both maximum likelihood and Bayesian inferences
- Variational Bayes methods have grown in popularity as a way of approximating posterior densities which are difficult to analyze using MCMC methods
- See Blei, Kucukelbir and McAuliffe (2017), Ormerod and Wand (2010) and Wand (2017) for recent surveys

Variational Bayes methods

- There are various ways to perform variational Bayes (henceforth VB) inference, but the simplest, and most popular is the “density transform approach”
- Idea: Approximate (complex) posterior density by some other density for which inference is more tractable
- This is the main idea in much older Importance Sampling or Laplace Approximations algorithms
- The key feature is that in VB inference we try to minimize the Kullback–Leibler divergence between the true posterior and the proposal distribution

Basics of VB calculus

Consider data vector y and parameters $\theta \in \Theta$. Posterior is then

$$p(\theta|y) = \frac{p(y,\theta)}{p(y)} \tag{76}$$

Let q be an arbitrary function in Θ. Then the log of marginal likelihood $p(y)$ is

$$\begin{aligned}
\log p(y) &= \log p(y) \int q(\theta) d\theta = \int \log p(y) q(\theta) d\theta \\
&= \int q(\theta) \log \left\{ \frac{p(y,\theta)/q(\theta)}{p(\theta|y)/q(\theta)} \right\} d\theta \\
&= \int q(\theta) \log \left\{ \frac{p(y,\theta)}{q(\theta)} \right\} d\theta + \int q(\theta) \log \left\{ \frac{q(\theta)}{p(\theta|y)} \right\} d\theta \\
&\geqslant \int q(\theta) \log \left\{ \frac{p(y,\theta)}{q(\theta)} \right\} d\theta
\end{aligned}$$

Basics of VB calculus

$$\log p(y) \geqslant \int q(\theta) \log \left\{ \frac{p(y,\theta)}{q(\theta)} \right\} d\theta \tag{77}$$

- The integral $\int q(\theta) \log \left\{ \frac{q(\theta)}{p(\theta|y)} \right\} d\theta$ is known as Kullback-Leibler divergence between $q(\theta)$ and $p(\theta|y)$
- Equality in formula above arises only if $q(\theta) = p(\theta|y)$ almost everywhere, i.e. zero KL divergence
- It follows that $p(y) \geqslant p(y;q)$ where $p(y;q) = \exp\left(\int q(\theta) \log \left\{ \frac{p(y,\theta)}{q(\theta)} \right\} d\theta \right)$ is called the q-dependent lower bound of the marginal likelihood
- VB inference: minimize the KL distance, OR maximize the lower bound $p(y;q)$ of the marginal likelihood

Further assumptions in VB inference

VB is obviously an optimization problem. Two further assumptions simplify derivations

1. $q(\theta)$ factorizes into $\prod_{i=1}^{M} q_i(\theta_i)$ for some partition $\{\theta_1, ..., \theta_M\}$ of model parameters θ
   - Restriction is also known as mean field approximation and has its roots in Physics
2. q is a member of a parametric family of density functions
   - Many times it is convenient if q belongs to the exponential family of distributions (Normal, Gamma, Beta, Dirichlet, Geometric, Chi-squared etc)

♣ Note: One more condition that makes Bayesian posterior inference easier in general (regardless of whether using VB or MCMC) is to use conjugate priors.

Example: VB inference in state-space models

- Let me demonstrate VB inference in a time-varying parameter (state-space) model
- Consider the standard time-varying parameter model

$$\begin{aligned} y_t &= X_t \beta_t + \varepsilon_t && (78) \\ \beta_t &= \beta_{t-1} + \eta_t && (79) \end{aligned}$$

  and for simplicity assume $\varepsilon \sim N\left(0, \sigma^2\right)$ and $\eta_{j,t} \sim N(0, r_{j,t})$.
- Joint posterior for this model is of the form

$$p\left(\beta_{1:T}, \sigma^2, r_{1:T} | y_{1:T}\right) \propto \prod_{t=1}^{T} p(\beta_t | \beta_{t-1}, r_t) p(y_t | \beta_t, \sigma^2) p(r_t) p(\sigma^2)$$

- Replace with variational distribution

$$q(\beta_{1:T}, \sigma^2, r_{1:T}) = q(\beta_{1:T}) q(\sigma^2) q(r_{1:T})$$

Example: VB inference in state-space models

- In the state-space setting we need minor adaptation of previous formulas
- Assume s state variable (β_t) and θ all other parameters

$$p\left(y\right) \geqslant \exp\left[\int q\left(s,\theta\right)\log\left\{\frac{p\left(y,s,\theta\right)}{q\left(s,\theta\right)}\right\}d\theta\right] \equiv \mathcal{F}\left(q(s,\theta)\right) \tag{80}$$

- Given the mean field approximation $q(s,\theta) = q(s)q(\theta)$, we can maximize $\mathcal{F}\left(q(s,\theta)\right)$ iteratively:

$$q\left(s\right) \propto \exp\left[\int q(\theta)\log p\left(y,s|\theta\right)d\theta\right], \tag{81}$$

$$q\left(\theta\right) \propto p(\theta)\exp\left[\int q(s)\log p\left(y,\theta|s\right)ds\right]. \tag{82}$$

which resembles the Expectation-Maximization (EM) algorithm of Dempster, Laird and Rubin (1977).

Example: VB inference in state-space models

- Therefore, in the time-varying parameter regression we assume the following priors

$$\begin{aligned} r_{j,t}^{-1} &\sim \mathrm{Gamma}\left(\underline{c}_0, \underline{d}_0\right), j = 1, ..., p, && (83) \\ \sigma^{-2} &\sim \mathrm{Gamma}\left(\underline{a}_0, \underline{b}_0\right), && (84) \\ \beta_0 &\sim N\left(\underline{\beta}_0, \underline{P}_0\right) && (85) \end{aligned}$$

- The result is a Variational Bayes Kalman Filter presented in next slide; for derivations see Koop and Korobilis (2018)
- It turns out that this is numerically quite similar to Kalman filter forgetting factor used in Koop and Korobilis (2012), even though it is derived using different principles

Example: VB inference in state-space models

Algorithm 1 *Variational Bayes Kalman Filter (VBKF)*

```
for $t = 1$ to $T$ do
    $s = 1$
    while $\|\beta_{t|t}^{(s)} - \beta_{t|t}^{(s-1)}\| \to 0$ do
        $\beta_{t|t-1}^{(s)} = \beta_{t-1}$
        $P_{t|t-1}^{(s)} = P_{t-1} + diag(q_t)^{(s)}$
        $K_t^{(s)} = P_{t|t-1}^{(s)} x_t' \left( x_t P_{t|t-1}^{(s)} x_t' + (\sigma^2)^{(s)} \right)^{-1}$
        $\beta_{t|t}^{(s)} = \beta_{t|t-1}^{(s)} + K_t^{(s)} \left( y_t - x_t \beta_{t|t-1}^{(s)} \right)$        ← Posterior mean of $\beta_t$
        $P_{t|t}^{(s)} = \left( I_p - K_t^{(s)} x_t \right) P_{t|t-1}^{(s)}$

        $D^{(s)} = P_{t|t}^{(s)} + \beta_{t|t}^{(s)} \beta_{t|t}^{(s)\prime} + \left( P_{t-1}^{(s)} + \beta_{t-1}^{(s)} \beta_{t-1}^{(s)\prime} \right) \left( I_p - 2\widetilde{F}_t^{(s)} \right)'$
        for $j = 1$ to $p$ do
            $c_{j,t}^{(s)} = \underline{c}_0 + 1/2$
            $d_{j,t}^{(s)} = \underline{d}_0 + D_{jj}^{(s)}/2$
            $r_{j,t}^{(s)} = d_{j,t}^{(s-1)} / c_{j,t}^{(s-1)}$        ← Posterior mean of $r_{j,t}$
        end for
        $a^{(s)} = \underline{a}_0 + T$
        $b^{(s)} = \underline{b}_0 + \sum_{t=1}^{T} \left[ x_t \left( \beta_{t|t}^{(s)} \beta_{t|t}^{(s)\prime} + P_{t|t}^{(s)} \right) x_t' - 2 x_t \beta_{t|t}^{(s)} y_t + y_t y_t' \right]$
        $(\sigma^2)^{(s)} = b^{(s-1)} / a^{(s-1)}$        ← Posterior mean of $\sigma^2$
        $s = s + 1$
    end while
end for
```

Simplifying inference by transformations

- Approximate algorithms can be very powerful, but getting there can be like climbing a mountain
- For example, deriving GAMP algorithm involves pages over pages of proofs
- Of course the final algorithm is computationally trivial and involves a few lines of code, but for harder problems we might not be able to come up easily with approximations
- Sometimes we can have huge computational savings by transforming our model
- Many times it has to do purely with the way we program, e.g. VAR in SUR form has matrices with many zeros so we can have savings if we use sparse matrix calculations in MATLAB
- But other times simply rewrite the likelihood in an equivalent form, that breaks the estimation problem in small pieces

Variable Elimination in Regression

- Variable elimination or marginalization is a machine learning procedure used in graphical models that, loosely speaking, allows (via certain rules) to break a high-dimensional inference problem into a series of smaller problems.
- Assume that we work with a regression with p predictors and we are interested in j-th predictor (e.g. a certain policy parameter)
- Rewrite regression as

$$y = x_j \beta_j + x_{(-j)} \beta_{(-j)} + \varepsilon, \tag{86}$$

- How can we obtain coefficient β_j after removing effects of $\beta_{(-j)}$?

Variable Elimination in Regression

- Solution due to Frisch and Waugh (1933, Econometrica)
- Define the annihilator matrix $M_j = I_T - x_j \left(x_j' x_j\right)^{-1} x_j'$, then the OLS estimate of β_j is

$$\widehat{\beta}_j = \left(x_j' x_j\right)^{-1} x_j' \left(y - x_{(-j)} \widehat{\beta}_{(-j)}\right) \tag{87}$$

where the sub-vector $\widehat{\beta}_{(-j)}$ is the solution of the following regression

$$\widehat{\beta}_{(-j)} = \left(x_{(-j)}^{\dagger\prime} x_{(-j)}^{\dagger}\right)^{-1} x_{(-j)}^{\dagger\prime} y^{\dagger} \tag{88}$$

with $x_{(-j)}^{\dagger} = M_j x_{(-j)}$ and $y^{\dagger} = M_j y$ denoting the projections of $x_{(-j)}$ and y on a space orthogonal to x_j

Variable Elimination in Regression

- The previous example shows a very interesting result that can be used to design efficient algorithms for inference
- I will demonstrate this in a Bayesian setting by using Korobilis and Pettenuzzo (forthcoming, Journal of Econometrics) as an example
- Define the $\mathsf{T} \times 1$ vector $\mathsf{q_j} = \mathsf{x_j}/\|\mathsf{x_j}\|$, and generate randomly a matrix $\mathsf{Q_j}$ that is normalized as $\mathsf{Q_j Q_j'} = \mathrm{I} - \mathsf{q_j q_j'}$.
- This means that the matrix $\mathsf{Q} = [\mathsf{q_j}, \mathsf{Q_j}]$ is orthogonal
- Multiplying both sides of (86) by Q' gives

Variable Elimination in Regression

$$\begin{aligned}
Q'y &= Q'x_j\beta_j + Q'x_{(-j)}\beta_{(-j)} + Q'\varepsilon \Rightarrow && (89)\\
\begin{bmatrix} q_j'y \\ Q_j'y \end{bmatrix} &= \begin{bmatrix} q_j'x_j \\ Q_j'x_j \end{bmatrix}\beta_j + \begin{bmatrix} q_j'x_{(-j)} \\ Q_j'x_{(-j)} \end{bmatrix}\beta_{(-j)} + Q'\varepsilon \Rightarrow && (90)\\
\begin{bmatrix} y^* \\ y^+ \end{bmatrix} &= \begin{bmatrix} \|x_j\| \\ 0 \end{bmatrix}\beta_j + \begin{bmatrix} x^*_{(-j)} \\ x^+_{(-j)} \end{bmatrix}\beta_{(-j)} + \widetilde{\varepsilon}, && (91)
\end{aligned}$$

where $y^* = q_j'y$, $y^+ = Q_j'y$, $x^*_{(-j)} = q_j'x_{(-j)}$, $x^+_{(-j)} = Q_j'x_{(-j)}$ and $\widetilde{\varepsilon} = Q'\varepsilon$.

- In the derivation above we have used the fact that $Q_j'x_j = Q_j'q_j\|x_j\| = 0$ because Q_j and q_j are orthogonal
- Additionally, $var(\widetilde{\varepsilon}) = \sigma^2 Q'Q = \sigma^2 = var(\varepsilon)$ because by construction $Q'Q = I$

Variable Elimination in Regression

$$\begin{bmatrix} y^{*} \\ y^{+} \end{bmatrix} = \begin{bmatrix} \|x_{j}\| \\ 0 \end{bmatrix} \beta_{j} + \begin{bmatrix} x^{*}_{(-j)} \\ x^{+}_{(-j)} \end{bmatrix} \beta_{(-j)} + \widetilde{\varepsilon}, \qquad (92)$$

- Hence regression in (86) can be written in the form (92)
- Equivalent OLS estimation of this model proceeds in two steps:
  1. Estimate $\beta_{(-j)}, \sigma^{2}$ by regressing y^{+} to $x^{+}_{(-j)}$
  2. Obtain β_{j} by regressing y^{*} on $\|x_{j}\|$ conditional on $\beta_{(-j)}, \sigma^{2}$ being know
- The second stage regression has known variance. Korobilis and Pettenuzzo (forthcoming) take advantage of this fact in order to derive *analytical* expression for the marginal posterior $p(\beta_{j}|y)$ for a wide range of shrinkage priors
- These shrinkage priors would normally need slow Gibbs sampler, but instead we can get quickly analytical results for each β_{j} and also trivially parallelize (*parfor* all j)!

Estimation of time-varying parameters using shrinkage methods

- Time-varying parameter models are a natural extension of the linear regression model
- Standard form used in economics is

$$
\begin{aligned}
y_t &= x_t \beta_t + \varepsilon_t, && (93)\\
\beta_t &= \beta_{t-1} + u_t, && (94)\\
\beta_0 &\sim N\left(\underline{\beta}, \underline{V}\right) && (95)
\end{aligned}
$$

- Why do we really need (94)?
- Are there alternative ways of estimating this model?

Estimation of time-varying parameters using shrinkage methods

Take only equation (93) and write in static regression form

$$y = Z\beta + \varepsilon, \tag{96}$$

where $y = \left(y_1', ..., y_T'\right)'$ and $\varepsilon = \left(\varepsilon_1', ..., \varepsilon_T'\right)'$ are $T \times 1$ vectors, $\beta = (\beta_1, ..., \beta_T)$ is a $Tp \times 1$ vector of regression coefficients, and we define

$$Z = \begin{bmatrix} x_1 & 0 & \cdots & 0 \\ 0 & x_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & x_T \end{bmatrix},$$

the $T \times Tp$ right-hand side matrix of predictors.

Estimation of time-varying parameters using shrinkage methods

$$y = Z\beta + \varepsilon, \tag{97}$$

- Z has more predictors (Tp) than observations (T)
- OLS is not possible, hence economists add equation (94)
- This equation can be thought of as a hierarchical prior of the form $p\left(\beta_t|\beta_{t-1}\right) \sim N\left(\beta_{t-1}, Q\right)$
- Under this prior, posterior of β_t can be updated recursively using (Kalman) simulation smoother
- But we could simply estimate (97) directly, with any other hierarchical shrinkage prior!
- E.g. Take only equation (97) and use LASSO prior on β!
- **Estimating TVP models with large p is hard due to computational limitations of Kalman filter, but this is not the case when estimating (97) with LASSO!**