State Space Methods

Bank of Korea Global Initiative Program

September 2014

Introduction

- State space methods are used for a wide variety of time series problems
- They are important in and of themselves
- Also time-varying parameter VARs (TVP-VARs) and stochastic volatility are state space models
- Advantage of state space models: well-developed set of MCMC algorithms for doing Bayesian inference

- Remember: our general notation for a VAR was:

$$y_t = Z_t \beta + \varepsilon$$

- In many macroeconomic applications, unrealistic to assume constant β
- This leads to TVP-VAR:

$$y_t = Z_t \beta_t + \varepsilon_t$$

- where

$$\beta_{t+1} = \beta_t + u_t$$

- This is a state space model.
- In VAR assume ε_t to be i.i.d. $N(0, \Sigma)$
- In empirical macroeconomics, this is often unrealistic.
- Want to have $var(\varepsilon_t) = \Sigma_t$
- This also leads to state space models.

The Normal Linear State Space Model

- Fairly general version of Normal linear state space model:
- Measurement equation:

$$y_t = W_t \delta + Z_t \beta_t + \varepsilon_t$$

- State equation:

$$\beta_{t+1} = T_t \beta_t + u_t$$

- y_t and ε_t defined as for VAR
- W_t is known $M \times p_0$ matrix (e.g. lagged dependent variables or explanatory variables with constant coefficients)
- Z_t is known $M \times K$ matrix (e.g. lagged dependent variables or explanatory variables with time varying coefficients)
- β_t is $k \times 1$ vector of states (e.g. VAR coefficients)
- ε_t ind $N(0, \Sigma_t)$
- u_t ind $N(0, Q_t)$.
- ε_t and u_s are independent for all s and t.
- T_t is a $k \times k$ matrix (usually fixed, but sometimes not).

- Key idea: for given values for δ, T_t, Σ_t and Q_t (called "system matrices") posterior simulators for β_t for $t = 1, .., T$ exist.
- E.g. Carter and Kohn (1994, Btka), Fruhwirth-Schnatter (1994, JTSA), DeJong and Shephard (1995, Btka) and Durbin and Koopman (2002, Btka).
- I will not present details of these (standard) algorithms
- These algorithms involve use of methods called Kalman filtering and smoothing
- Filtering = estimating a state at time t using data up to time t
- Smoothing = estimating a state at time t using data up to time T
- Recently other algorithms have been proposed in several papers by Joshua Chan (Australian National University) and Bill McCausland (University of Montreal)

- Notation: $\beta^t = \left(\beta_1', .., \beta_t'\right)'$ stacks all the states up to time t (and similar superscript t convention for other things)
- Gibbs sampler: $p\left(\beta^T | y^T, \delta, T^T, \Sigma^T, Q^T\right)$ drawn use such an algorithm
- $p\left(\delta | y^T, \beta^T, T^T, \Sigma^T, Q^T\right)$, $p\left(T^T | y^T, \beta^T, \delta, \Sigma^T, Q^T\right)$, $p\left(\Sigma^T | y^T, \beta^T, \delta, T^T, Q^T\right)$ and $p\left(Q^T | y^T, \beta^T, \delta, T^T, \Sigma^T\right)$ depend on precise form of model (typically simple since, conditional on β^T have a Normal linear model)
- Typically restricted versions of this general model used
- TVP-VAR of Primiceri (2005, ReStud) has $\delta = 0$, $T_t = I$ and $Q_t = Q$

Example of an MCMC Algorithm

- Special case $\delta = 0$, $T_t = I, \Sigma_t = \Sigma$ and $Q_t = Q$
- Homoskedastic TVP-VAR of Cogley and Sargent (2001, NBER)
- Need prior for all parameters
- But state equation implies hierarchical prior for β^T:

$$\beta_{t+1} | \beta_t, Q \sim N\left(\beta_t, Q\right)$$

- Formally:

$$p\left(\beta^T | Q\right) = \prod_{t=1}^{T} p\left(\beta_t | \beta_{t-1}, Q\right)$$

- Hierarchical: since it depends on Q which, in turn, requires its own prior.

- Note β_0 enters prior for β_1.
- Need prior for β_0
- Standard treatments exist.
- E.g. assume $\beta_0 = 0$, then:

$$\beta_1 | Q \sim N(0, Q)$$

- Or Carter and Kohn (1994) simply assume β_0 has some prior that researcher chooses

- Convenient to use Wishart priors for Σ^{-1} and Q^{-1}
- $$\Sigma^{-1} \sim W\left(\underline{S}^{-1}, \underline{\nu}\right)$$
- $$Q^{-1} \sim W\left(\underline{Q}^{-1}, \underline{\nu}_Q\right)$$

- Want MCMC algorithm which sequentially draws from $p\left(\Sigma^{-1}|y^{T},\beta^{T},Q\right)$, $p\left(Q^{-1}|y^{T},\Sigma,\beta^{T}\right)$ and $p\left(\beta^{T}|y^{T},\Sigma,Q\right)$.
- For $p\left(\beta^{T}|y^{T},\Sigma,Q\right)$ use standard algorithm for state space models (e.g. Carter and Kohn, 1994)
- Can derive $p\left(\Sigma^{-1}|y^{T},\beta^{T},Q\right)$ and $p\left(Q^{-1}|y^{T},\Sigma,\beta^{T}\right)$ using methods similar to those used in section on VAR independent Normal-Wishart model.

- Conditional on β^T, measurement equation is like a VAR with known coefficients.
- This leads to:

$$\Sigma^{-1}|y^T, \beta^T \sim W\left(\overline{S}^{-1}, \overline{\nu}\right)$$

- where

$$\overline{\nu} = T + \underline{\nu}$$

-

$$\overline{S} = \underline{S} + \sum_{t=1}^{T}\left(y_t - W_t\delta - Z_t\beta_t\right)\left(y_t - W_t\delta - Z_t\beta_t\right)'$$

- Conditional on β^T, state equation is also like a VAR with known coefficients.
- This leads to:

$$Q^{-1}|y^T, \beta^T \sim W\left(\overline{Q}^{-1}, \overline{\nu}_Q\right)$$

- where

$$\overline{\nu}_Q = T + \underline{\nu}_Q$$

-

$$\overline{Q} = \underline{Q} + \sum_{t=1}^{T} \left(\beta_{t+1} - \beta_t\right)\left(\beta_{t+1} - \beta_t\right)'.$$

DSGE Models as State Space Models

- DSGE = Dynamic, stochastic general equilibrium models popular in modern macroeconomics and commonly used in policy circles (e.g. central banks).
- I will not explain the macro theory, other than to note they are:
- Derived from microeconomic principles (based on agents and firms decision problems), dynamic (studying how economy evolves over time) and general equilibrium.
- Solution (using linear approximation methods) is a linear state space model
- Note: recent work with second order approximations yields nonlinear state space model
- Survey: An and Schorfheide (2007, Econometric Reviews)
- Computer code: http://www.dynare.org/ or some authors post code (e.g. code for Del Negro and Schorfheide 2008, JME on web)

Estimation Strategy for DSGE

- Most linearized DSGE models written as:

$$\Gamma_0(\theta) z_t = \Gamma_1(\theta) E_t(z_{t+1}) + \Gamma_2(\theta) z_{t-1} + \Gamma_3(\theta) u_t$$

- z_t is vector containing both observed variables (e.g. output growth, inflation, interest rates) and unobserved variables (e.g. technology shocks, monetary policy shocks).
- Note, theory usually written in terms of z_t as deviation of variable from steady state (an issue I will ignore here to keep exposition simple)
- θ are structural parameters (e.g. parameters for steady states, tastes, technology, policy and driving the exogenous shocks).
- u_t are structural shocks ($N(0, I)$).
- $\Gamma_j(\theta)$ are often highly nonlinear functions of θ

Solving the DSGE Model

- Methods exist to solve linear rational expectations models such as the DSGE
- If unique equilibrium exists can be written as:

$$z_t = A(\theta) z_{t-1} + B(\theta) u_t$$

- Looks like a VAR, but....
- Some elements of z_t typically unobserved
- and highly nonlinear restrictions involved in $A(\theta)$ and $B(\theta)$

Write DSGE Model as State Space Model

- Let y_t be elements of z_t which are observed.
- Measurement equation:

$$y_t = Cz_t$$

  where C is matrix which picks out observed elements of z_t
- Equation on previous slide is state equation in states z_t
- Thus we have state space model
- Special case since measurement equation has no errors (although measurement errors sometimes added) and state equation has some states which are observed.
- But state space algorithms described earlier in this lecture still work
- Remember, before I said: "for given values for system matrices, posterior simulators for the states exist"
- If θ were known, DSGE model provides system matrices in Normal linear state space model

Estimating the Structural Parameters

- If $A(\theta)$ and $B(\theta)$ involved simple linear restrictions, then methods similar to those for the restricted VAR (see Lecture 2) could be used to carry out inference on θ.
- Unfortunately, restrictions in $A(\theta)$ and $B(\theta)$ are typically nonlinear and complicated
- Parameters in θ are structural so we are likely to have prior information about them
- Example from Del Negro and Schorfheide (2008, JME):
- "Household-level data on wages and hours worked could be used to form a prior for a labor supply elasticity"
- "Product level data on price changes could be the basis for a price-stickiness prior"

Estimating the Structural Parameters (cont.)

- Prior for structural parameters, $p(\theta)$, can be formed from other sources of information (e.g. micro studies, economic theory, etc.)
- Here: prior times likelihood is a mess
- Thus, no analytical posterior for θ, no Gibbs sampler, etc...
- Solution: Metropolis-Hastings algorithm

The Metropolis-Hastings Algorithm

- For now, I leave DSGE and state space models and return to our general notation:
- θ is a vector of parameters and $p(y|\theta)$, $p(\theta)$ and $p(\theta|y)$ are the likelihood, prior and posterior, respectively.
- Metropolis-Hastings algorithm takes draws from a convenient *candidate generating density*.
- Let θ^* indicate a draw taken from this density which we denote as $q\left(\theta^{(s-1)};\theta\right)$.
- Notation: θ^* is a draw taken of the random variable θ whose density depends on $\theta^{(s-1)}$.

- We are drawing the wrong distribution, $q\left(\theta^{(s-1)};\theta\right)$, instead of $p(\theta|y)$
- We have to correct for this somehow.
- Metropolis-Hastings algorithm corrects for this via an acceptance probability
- Takes candidate draws, but only some of these candidate draws are accepted.

- The Metropolis-Hastings algorithm takes following form:
- *Step 1:* Choose a starting value, $\theta^{(0)}$.
- *Step 2:* Take a candidate draw, θ^* from the candidate generating density, $q\left(\theta^{(s-1)};\theta\right)$.
- *Step 3:* Calculate an acceptance probability, $\alpha\left(\theta^{(s-1)},\theta^*\right)$.
- *Step 4:* Set $\theta^{(s)}=\theta^*$ with probability $\alpha\left(\theta^{(s-1)},\theta^*\right)$ and set $\theta^{(s)}=\theta^{(s-1)}$ with probability $1-\alpha\left(\theta^{(s-1)},\theta^*\right)$.
- *Step 5:* Repeat Steps 1, 2 and 3 S times.
- *Step 6:* Take the average of the S draws $g\left(\theta^{(1)}\right),...,g\left(\theta^{(S)}\right)$.

- These steps will yield an estimate of $E\left[g(\theta)|y\right]$ for any function of interest.
- Note: As with Gibbs sampling, Metropolis-Hastings algorithm requires the choice of a starting value, $\theta^{(0)}$. To make sure that the effect of this starting value has vanished, wise to discard S_0 initial draws.
- Intuition for acceptance probability, $\alpha\left(\theta^{(s-1)},\theta^*\right)$, given in textbook (pages 93-94).

$$\alpha\left(\theta^{(s-1)},\theta^*\right) = \min\left[\frac{p(\theta=\theta^*|y)q\left(\theta^*;\theta=\theta^{(s-1)}\right)}{p\left(\theta=\theta^{(s-1)}|y\right)q\left(\theta^{(s-1)};\theta=\theta^*\right)},1\right]$$

Choosing a Candidate Generating Density

- Independence Chain Metropolis-Hastings Algorithm
- Uses a candidate generating density which is independent across draws.
- That is, $q\left(\theta^{(s-1)};\theta\right)=q^{*}(\theta)$ and the candidate generating density does not depend on $\theta^{(s-1)}$.
- Useful in cases where a convenient approximation exists to the posterior. This convenient approximation can be used as a candidate generating density.
- Acceptance probability simplifies to:

$$\alpha\left(\theta^{(s-1)},\theta^{*}\right)=\min\left[\frac{p\left(\theta=\theta^{*}|y\right)q^{*}\left(\theta=\theta^{(s-1)}\right)}{p\left(\theta=\theta^{(s-1)}|y\right)q^{*}\left(\theta=\theta^{*}\right)},1\right].$$

- Not popular with DSGE models since convenient approximation unlikely to exist

- Random Walk Chain Metropolis-Hastings Algorithm
- Popular with DSGE – useful when you cannot find a good approximating density for the posterior.
- No attempt made to approximate posterior, rather candidate generating density is chosen to wander widely, taking draws proportionately in various regions of the posterior.
- Generates candidate draws according to:

$$\theta^{*} = \theta^{(s-1)} + w$$

where w is called the *increment random variable*.

- Acceptance probability simplifies to:

$$\alpha\left(\theta^{(s-1)}, \theta^{*}\right) = \min\left[\frac{p\left(\theta = \theta^{*}|y\right)}{p\left(\theta = \theta^{(s-1)}|y\right)}, 1\right]$$

- Choice of density for w determines form of candidate generating density.
- Common choice is Normal:

$$q\left(\theta^{(s-1)}; \theta\right) = f_N(\theta|\theta^{(s-1)}, \Sigma).$$

- Researcher must select Σ. Should be selected so that the acceptance probability tends to be neither too high nor too low.
- There is no general rule which gives the optimal acceptance rate. A rule of thumb is that the acceptance probability should be roughly 0.5.
- A common approach sets $\Sigma = c\Omega$ where c is a scalar and Ω is an estimate of posterior covariance matrix of θ (e.g. the inverse of the Hessian evaluated at the posterior mode)

- Popular (e.g. DYNARE) to use random walk Metropolis-Hastings with DSGE models.
- Note acceptance probability depends only on posterior = prior times likelihood
- DSGE Prior chosen as discussed above
- Algorithms for Normal linear state space models evaluate likelihood function

Metropolis-within-Gibbs

- Remember: the Gibbs sampler involved sequentially drawing from $p\left(\theta_{(1)}|y,\theta_{(2)}\right)$ and $p\left(\theta_{(2)}|y,\theta_{(1)}\right)$.
- Using a Metropolis-Hastings algorithm for either (or both) of the posterior conditionals used in the Gibbs sampler, $p\left(\theta_{(1)}|y,\theta_{(2)}\right)$ and $p\left(\theta_{(2)}|y,\theta_{(1)}\right)$, is perfectly acceptable.
- This statement is also true if the Gibbs sampler involves more than two blocks.
- Such *Metropolis-within-Gibbs* algorithms are common since many models have posteriors where most of the conditionals are easy to draw from, but one or two conditionals do not have convenient form.

Nonlinear State Space Models

- Normal linear state space model useful for empirical macroeconomists
- E.g. trend-cycle decompositions, TVP-VARs, linearized DSGE models, etc.
- Some models have y_t being a nonlinear function of the states (e.g. DSGE models which have not been linearized)
- Increasing number of Bayesian tools for nonlinear state space models (e.g. the particle filter)
- Here we will focus on stochastic volatility

Univariate Stochastic Volatility

- Begin with y_t being a scalar (common in finance)
- Stochastic volatility model:

$$y_t = \exp\left(\frac{h_t}{2}\right)\varepsilon_t$$

-

$$h_{t+1} = \mu + \phi\left(h_t - \mu\right) + \eta_t$$

- ε_t is i.i.d. $N\left(0,1\right)$ and η_t is i.i.d. $N\left(0,\sigma_\eta^2\right)$. ε_t and η_s are independent.
- This is state space model with states being h_t, but measurement equation is not a linear function of h_t

- h_t is log of the variance of y_t (log volatility)
- Since variances must be positive, common to work with log-variances
- Note μ is the unconditional mean of h_t.
- Initial conditions: if $|\phi| < 1$ (stationary) then:

$$h_0 \sim N\left(\mu, \frac{\sigma_\eta^2}{1-\phi^2}\right)$$

- if $\phi = 1$, μ drops out of the model and However, when $\phi = 1$, need a prior such as $h_0 \sim N\left(\underline{h}, \underline{V}_h\right)$
- e.g. Primiceri (2005) chooses $\underline{V}_h$ using training sample

MCMC Algorithm for Stochastic Volatility Model

- MCMC algorithm involves sequentially drawing from $p\left(h^T|y^T,\mu,\phi,\sigma_\eta^2\right)$, $p\left(\phi|y^T,\mu,\sigma_\eta^2,h^T\right)$, $p\left(\mu|y^T,\phi,\sigma_\eta^2,h^T\right)$ and $p\left(\sigma_\eta^2|y^T,\mu,\phi,h^T\right)$
- Last three standard forms based on results from Normal linear regression model and will not present here.
- Several algorithms exist for $p\left(h^T|y^T,\mu,\phi,\sigma_\eta^2\right)$
- Here we describe a popular one from Kim, Shephard and Chib (1998, ReStud)
- For complete details, see their paper. Here we outline ideas.

- Square and log the measurement equation:

$$y_t^* = h_t + \varepsilon_t^*$$

- where $y_t^* = \ln\left(y_t^2\right)$ and $\varepsilon_t^* = \ln\left(\varepsilon_t^2\right)$.
- Now the measurement equation is linear so maybe we can use algorithm for Normal linear state space model?
- No, since error is no longer Normal (i.e. $\varepsilon_t^* = \ln\left(\varepsilon_t^2\right)$)
- Idea: use mixture of different Normal distributions to approximate distribution of ε_t^*.

- Mixtures of Normal distributions are very flexible and have been used widely in many fields to approximate unknown or inconvenient distributions.
- $$p\left(\varepsilon_t^*\right) \approx \sum_{i=1}^{7} q_i f_N\left(\varepsilon_t^* | m_i, v_i^2\right)$$
- where $f_N\left(\varepsilon_t^* | m_i, v_i^2\right)$ is the p.d.f. of a $N\left(m_i, v_i^2\right)$
- since ε_t is $N(0,1)$, ε_t^* involves no unknown parameters
- Thus, q_i, m_i, v_i^2 for $i = 1, .., 7$ are not parameters, but numbers (see Table 4 of Kim, Shephard and Chib, 1998).

- Mixture of Normals can also be written in terms of component indicator variables, $s_t \in \{1, 2, .., 7\}$
- $$\begin{array}{l} \varepsilon_t^* | s_t = i \sim N\left(m_i, v_i^2\right) \\ \Pr\left(s_t = i\right) = q_i \end{array}$$
- MCMC algorithm does not draw from $p\left(h^T | y^T, \mu, \phi, \sigma_\eta^2\right)$, but from $p\left(h^T | y^T, \mu, \phi, \sigma_\eta^2, s^T\right)$.
- But, conditional on s^T, knows which of the Normals ε_t^* comes from.
- Result is a Normal linear state space model and familiar algorithm can be used.
- Finally, need $p\left(s^T | y^T, \mu, \phi, \sigma_\eta^2, h^T\right)$ but this has simple form (see Kim, Shephard and Chib , 1998)

Learning Bayesian Computation

- Remember: this course has a lecture format
- But I have made up some computer question sheets that you you may wish to work through on your own
- Computer session 3 (on the course website) has questions relating to an influential state space model used in
- Stock and Watson (2007) "Why Has U.S. Inflation Become Harder to Forecast?," Journal of Money, Credit and Banking.
- MATLAB computer code which answers the questions is also available on the website

Multivariate Stochastic Volatility

- y_t is now $M \times 1$ vector and ε_t is i.i.d. $N(0, \Sigma_t)$.
- Many ways of allowing Σ_t to be time-varying
- But must worry about overparameterization problems
- Σ_t for $t = 1, .., T$ contains $\frac{TM(M+1)}{2}$ unknown parameters
- Here we discuss three particular approaches popular in macroeconomics
- To focus on multivariate stochastic volatility, use model:

$$y_t = \varepsilon_t$$

Multivariate Stochastic Volatility Model 1

- $$\Sigma_t = D_t$$
- where D_t is a diagonal matrix with diagonal elements d_{it}
- d_{it} has standard univariate stochastic volatility specification
- $d_{it} = \exp(h_{it})$ and

$$h_{i,t+1} = \mu_i + \phi_i (h_{it} - \mu_i) + \eta_{it}$$

- if η_{it} are independent (across both i and t) then Kim, Shephard and Chib (1998) MCMC algorithm can be used one equation at a time.
- But many interesting macroeconomic features (e.g. impulse responses) depend on error covariances so assuming Σ_t to be diagonal often will be a bad idea.

Multivariate Stochastic Volatility Model 2

- Cogley and Sargent (2005, RED)
- $$\Sigma_t = L^{-1} D_t L^{-1\prime}$$
- D_t is as in Model 1 (diagonal matrix with diagonal elements being variances)
- L is a lower triangular matrix with ones on the diagonal.
- E.g. $M = 3$

$$L = \begin{bmatrix} 1 & 0 & 0 \\ L_{21} & 1 & 0 \\ L_{31} & L_{32} & 1 \end{bmatrix}$$

- We can transform model as:

$$Ly_t = L\varepsilon_t$$

- $\varepsilon_t^* = L\varepsilon_t$ will now have a diagonal covariance matrix – can use algorithm for Model 1.
- MCMC algorithm: $p\left(h^T|y^T, L\right)$ can use Kim, Shephard and Chib (1998) algorithm one equation at a time.
- $p\left(L|y^T, h^T\right)$ results similar to those from a series of M regression equations with independent Normal errors.
- See Cogley and Sargent (2005) for details.

- Cogley-Sargent model allows the covariance between errors to change over time, but in restricted fashion.
- E.g. $M = 2$ then $cov(\varepsilon_{1t}, \varepsilon_{2t}) = d_{1t}L_{21}$ which varies proportionally with the error variance of the first equation.
- Impulse response analysis: a shock to i^{th} variable has an effect on j^{th} variable which is constant over time
- In many macroeconomic applications this is too restrictive.

Multivariate Stochastic Volatility Model 3

- Primiceri (2005, ReStud):

$$\Sigma_t = L_t^{-1} D_t L_t^{-1\prime}$$

- L_t is same as Cogley-Sargent's L but is now time varying.
- Does not restrict Σ_t in any way.
- MCMC algorithm same as for Cogley-Sargent except for L_t

- How does L_t evolve?
- Stack unrestricted elements by rows into a $\frac{M(M-1)}{2}$ vector as $l_t = \left(L_{21,t}, L_{31,t}, L_{32,t}, .., L_{p(p-1),t}\right)'$.
- $$l_{t+1} = l_t + \zeta_t$$
- ζ_t is i.i.d. $N\left(0, D_\zeta\right)$ and D_ζ is a diagonal matrix.
- Can transform model so that algorithm for Normal linear state space model can draw l_t
- See Primiceri (2005) for details
- Note: if D_ζ is not diagonal have to be careful (no longer Normal state space model)

Sequential Monte Carlo and Particle Filtering Methods for State Space Models

- In this course, we have focussed on MCMC Methods such as the Gibbs sampler
- However, there is a new set of methods that is growing in popularity
- Some argue these will be the dominant computational tools of the future, particularly for nonlinear state space models
- If time permits, I will offer a brief introduction
- The following website contains a variety of materials (including some nice videos)
- http://www.stats.ox.ac.uk/~doucet/smc_resources.html
- If you love computers, I note that these methods (unlike MCMC methods) can typically be parallelized
- This means you can use the massive computing power in graphical processing units (GPUs)
- See the manuscript "Massively Parallel Sequential Monte Carlo for Bayesian Inference" by Durham and Geweke

Aside on Model Averaging and Model Selection

- I have said much about estimation and forecasting, but little about model comparison
- Remember Bayesians often use marginal likelihoods (which are the basic building block in posterior model probabilities)
- Marginal likelihoods can be used for model selection:
- Select model with highest marginal likelihood
- Or model averaging:
- Retain all models but present forecasts/results which are weighted average with weights proportional to marginal likelihoods
- Marginal likelihoods can be hard to calculate using MCMC methods
- An advantage of sequential Monte Carlo methods is that marginal likelihood at each point in time is produced
- But there is an increasing interest in various methods which do model averaging and selection in a quick manner
- To give an example of how this is done and provide an application of state space modelling, I next will present an empirical example

A Macroeconomic Application: Inflation Forecasting using Dynamic Model Averaging

- Based on the paper: Koop and Korobilis (2012, IER)
- Macroeconomists typically have many time series variables
- But even with all this information forecasting of macroeconomic variables like inflation, GDP growth, etc. can be very hard
- Sometimes hard to beat very simple forecasting procedures (e.g. random walk)
- Imagine a regression of inflation on many predictors
- Such a regression might fit well in practice, but forecast poorly

- Why? There are many reasons, but three stand out:
- Regressions with many predictors can over-fit (over-parameterization problems)
- Marginal effects of predictors change over time (parameter change/structural breaks)
- The relevant forecasting model may change (model change)
- We use an approach called Dynamic Model Averaging (DMA) in an attempt to address these problems

The Generalized Phillips Curve

- Phillips curve: inflation depends on unemployment rate
- Generalized Phillips curve: Inflation dependent on lagged inflation, unemployment and other predictors
- Many papers use generalized Phillips curve models for inflation forecasting
- Regression-based methods based on:

$$y_t = \phi + x'_{t-1}\beta + \sum_{j=1}^{p} \gamma_j y_{t-j} + \varepsilon_t$$

- y_t is inflation and x_{t-1} are lags of other predictors
- To make things concrete, following is our list of predictors (other papers use similar)

- UNEMP: unemployment rate.
- CONS: the percentage change in real personal consumption expenditures.
- INV: the percentage change in private residential fixed investment.
- GDP: the percentage change in real GDP.
- HSTARTS: the log of housing starts (total new privately owned housing units).
- EMPLOY: the percentage change in employment (All Employees: Total Private Industries, seasonally adjusted).
- PMI: the change in the Institute of Supply Management (Manufacturing): Purchasing Manager's Composite Index.

- TBILL: three month Treasury bill (secondary market) rate.
- SPREAD: the spread between the 10 year and 3 month Treasury bill rates.
- DJIA: the percentage change in the Dow Jones Industrial Average.
- MONEY: the percentage change in the money supply (M1).
- INFEXP: University of Michigan measure of inflation expectations.
- COMPRICE: the change in the commodities price index (NAPM commodities price index).
- VENDOR: the change in the NAPM vendor deliveries index.

Forecasting With Generalized Phillips Curve

- Write more compactly as:

$$y_t = z_t \theta + \varepsilon_t$$

- z_t contains all predictors, lagged inflation, an intercept
- Note $z_t =$ information available for forecasting y_t
- When forecasting h periods ahead will contain variables dated $t - h$ or earlier

- Consider forecasting $y_{\tau+1}$.
- Recursive forecasting methods: $\widehat{\theta} =$ estimate using data through τ.
- So $\widehat{\theta}$ will change (a bit) with τ, but can change too slowly
- Rolling forecasts use: $\widehat{\theta}$ an estimate using data from $\tau - \tau_0$ through τ.
- Better at capturing parameter change, but need to choose τ_0
- Recursive and rolling forecasts might be imperfect solutions
- Why not use a model which formally models the parameter change as well?

Time Varying Parameter (TVP) Models

- TVP models gaining popularity in empirical macroeconomics

$$
\begin{aligned}
y_t &= z_t \theta_t + \varepsilon_t \\
\theta_t &= \theta_{t-1} + \eta_t
\end{aligned}
$$

- $\varepsilon_t \overset{ind}{\sim} N(0, H_t)$
- $\eta_t \overset{ind}{\sim} N(0, Q_t)$
- State space methods described above can be used to estimate them

- Why not use TVP model to forecast inflation?
- Advantage: models parameter change in a formal manner
- Disadvantage: same predictors used at all points in time.
- If number of predictors large, over-fit, over-parameterization problems
- In our empirical work, we show very poor forecast performance

Digression on Bayesian Model Averaging (BMA)

- Return to basic regression model

$$y_t = z_t\theta + \varepsilon_t$$

- Suppose z_t contains m variables and m is large (over-fitting, over-parameterization)
- Model selection: Select a single model based on a series of hypothesis tests
- Problem: 2^m potential models: serious pre-test problems
- An increasingly common response: BMA
- Average over all models (with data-based weights: marginal likelihoods or information criteria)

- Problem: Computational burden of dealing with 2^m models
- Solution: Sophisticated algorithms which simulate from model space (MC-cubed)
- Note: MC-cubed can also be used to select a single best model in an automatic and computational efficient way
- BMA is increasingly popular with cross-sectional data sets (e.g. cross-country growth regressions)
- Question: can we use similar ideas in a time series context?

Dynamic Model Averaging (DMA)

- Define K models which have $z_t^{(k)}$ for $k = 1, .., K$, as predictors
- $z_t^{(k)}$ is subset of z_t.
- Set of models:

$$\begin{aligned} y_t &= z_t^{(k)} \theta_t^{(k)} + \varepsilon_t^{(k)} \\ \theta_{t+1}^{(k)} &= \theta_t^{(k)} + \eta_t^{(k)} \end{aligned}$$

- $\varepsilon_t^{(k)}$ is $N\left(0, H_t^{(k)}\right)$
- $\eta_t^{(k)}$ is $N\left(0, Q_t^{(k)}\right)$
- Let $L_t \in \{1, 2, .., K\}$ denote which model applies at t

- Why not just forecast using BMA over these TVP models at every point in time?
- Different weights in averaging at every point in time.
- Or why not just select a single TVP forecasting model at every point in time?
- Different forecasting models selected at each point in time.
- If K is large (e.g. $K = 2^m$), this is computationally infeasible.
- With cross-sectional BMA have to work with model space $K = 2^m$ which is computationally burdensome
- In present time series context, forecasting through time τ involves $2^{m\tau}$ models.
- Also, Bayesian inference in TVP model requires MCMC (unlike cross-sectional regression). Computationally burdensome.
- Even clever algorithms like MC-cubed are not good enough to handle this.

- Another strategy has been used to deal with similar problems in different contexts (e.g. multiple structural breaks): Markov switching
- Markov transition matrix, P,
- Elements $p_{ij} = \Pr\left(L_t = i | L_{t-1} = j\right)$ for $i, j = 1, .., K$.
- "If j is the forecasting model at $t-1$, we switch to forecasting model i at time t with probability p_{ij}"
- Bayesian inference is theoretically straightforward, but computationally infeasible
- P is $K \times K$: an enormous matrix.
- Even if computation were possible, imprecise estimation of so many parameters

Solution: DMA

- Adopt approach used by Raftery et al (2010 Technometrics) in an engineering application
- Involves two approximations
- First approximation means we do not need MCMC in each TVP model (only need run a standard Kalman filtering and smoothing)
- See paper for details. Idea: replace $Q_t^{(k)}$ and $H_t^{(k)}$ by estimates

- Sketch of some Kalman filtering ideas (where y^{t-1} are observations through $t-1$)

$$\theta_{t-1}|y^{t-1} \sim N\left(\widehat{\theta}_{t-1}, \Sigma_{t-1|t-1}\right)$$

- Textbook formula for $\widehat{\theta}_{t-1}$ and $\Sigma_{t-1|t-1}$
- Then update

$$\theta_t|y^{t-1} \sim N\left(\widehat{\theta}_{t-1}, \Sigma_{t|t-1}\right)$$

- $$\Sigma_{t|t-1} = \Sigma_{t-1|t-1} + Q_t$$

- Get rid of Q_t by approximating:

$$\Sigma_{t|t-1} = \frac{1}{\lambda}\Sigma_{t-1|t-1}$$

- $0 < \lambda \leq 1$ is forgetting factor

- Forgetting factors like this have long been used in state space literature
- Implies that observations j periods in the past have weight λ^j.
- Or effective window size of $\frac{1}{1-\lambda}$.
- Choose value of λ near one
- $\lambda = 0.99$: observations five years ago $\approx 80\%$ as much weight as last period's observation.
- $\lambda = 0.95$: observations five years ago $\approx 35\%$ as much weight as last period's observations.
- We focus on $\lambda \in [0.95, 1.00]$.
- If $\lambda = 1$ no time variation in parameters (standard recursive forecasting)

Back to Model Averaging/Selection

- Goal for forecasting at time t given data available at time $t-1$ is $\pi_{t|t-1,k} \equiv \Pr\left(L_t = k | y^{t-1}\right)$
- Can average across $k = 1, .., K$ forecasts using $\pi_{t|t-1,k}$ as weights (DMA)
- E.g. point forecasts ($\widehat{\theta}_{t-1}^{(k)}$ from Kalman filter in model k):

$$E\left(y_t | y^{t-1}\right) = \sum_{k=1}^{K} \pi_{t|t-1,k} z_t^{(k)} \widehat{\theta}_{t-1}^{(k)}$$

- Can forecast with model j at time t if $\pi_{t|t-1,j}$ is highest (Dynamic model selection: DMS)
- Raftery et al (2010) propose another forgetting factor to approximate $\pi_{t|t-1,k}$

- Complete details in Raftery et al's paper.
- Basic idea is that can use similar state space updating formulae for models as is done with states
- Then use similar forgetting factor to get approximation

$$\pi_{t|t-1,k} = \frac{\pi^{\alpha}_{t-1|t-1,k}}{\sum_{l=1}^{K} \pi^{\alpha}_{t-1|t-1,l}}$$

- $0 < \alpha \leq 1$ is forgetting factor with similar interpretation to λ
- Focus on $\alpha \in [0.95, 1.00]$

- Interpretation of forgetting factor α
- Easy to show:

$$\pi_{t|t-1,k} = \prod_{i=1}^{t-1} \left[p_k \left(y_{t-i} | y^{t-i-1} \right) \right]^{\alpha^i}$$

- $p_k \left(y_t | y^{t-1} \right)$ is predictive density for model k evaluated at y_t (measure of forecast performance of model k)
- Model k will receive more weight at time t if it has forecast well in the recent past
- Interpretation of "recent past" is controlled by the forgetting factor, α
- $\alpha = 0.99$: forecast performance five years ago receives 80% as much weight as forecast performance last period
- $\alpha = 0.95$: forecast performance five years ago receives only about 35% as much weight.
- $\alpha = 1$: can show $\pi_{t|t-1,k}$ is proportional to the marginal likelihood using data through time $t-1$ (standard BMA)

Summary So Far

- We want to do DMA or DMS
- These use TVP models which allow marginal effects to change over time
- These allow for forecasting model to switch over time
- So can switch from one parsimonious forecasting model to another (avoid over-parametization)
- But a full formal Bayesian analysis is computationally infeasible
- Sensible approximations make it computationally feasible.
- State space updating formula must be run K times, instead of (roughly speaking) K^T MCMC algorithms

Forecasting US Inflation

- Data from 1960Q1 through 2008Q4
- Real time data (forecasting at time τ using data as known at time τ)
- Two measure of inflation based on PCE deflator (core inflation) and GDP deflator
- 14 predictors listed previously (all variables transformed to be approximately stationary)
- All models include an intercept and two lags of the dependent variable
- 3 forecast horizons: $h = 1,4,8$

Is DMA Parsimonious?

- Even though 14 potential predictors, most probability is attached to very parsimonious models with only a few predictors.
- $Size_k$ = number of predictors in model k
- ($Size_k$ does not include the intercept plus two lags of the dependent variable)
- Figure 1 plots

$$E\left(Size_t\right) = \sum_{k=1}^{K} \pi_{t|t-1,k} Size_k$$

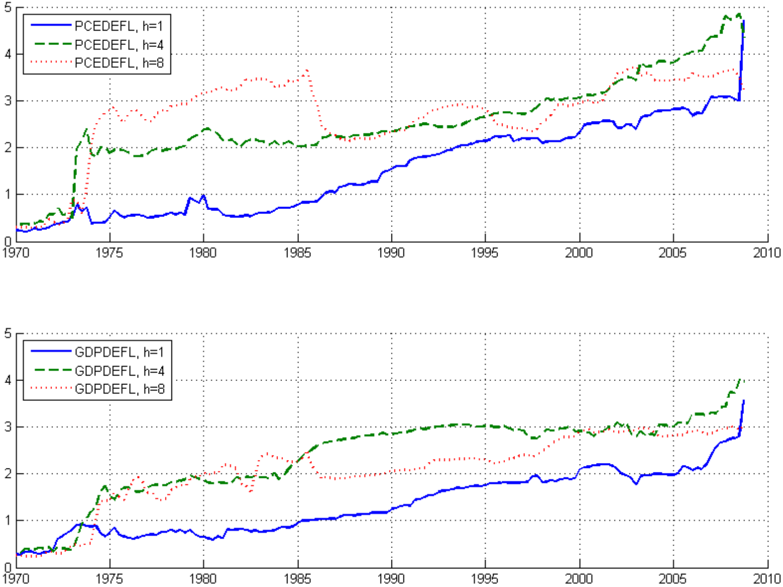

Figure 1: Expected Number of Predictors

Which Variables are Good Predictors for Inflation?

- Posterior inclusion probabilities for j^{th} predictor $=$

$$\sum_{k \in J} \pi_{t|t-1,k}$$

- where $k \in J$ indicates models which include j^{th} predictor
- See Figure 2, 3 and 4 for 2 measures of inflation and 3 forecast horizons
- Any predictor where the inclusion probability is never above 0.5 is excluded from the appropriate figure.
- Lots of evidence of predictor change in all cases.
- DMA/DMS will pick this up automatically

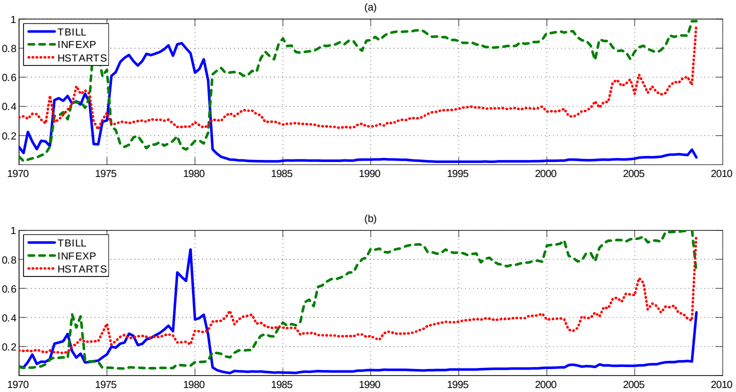

Figure 2: Posterior Probability of Inclusion of Predictors, $h = 1$. GDP deflator inflation top, PCE deflator inflation bottom

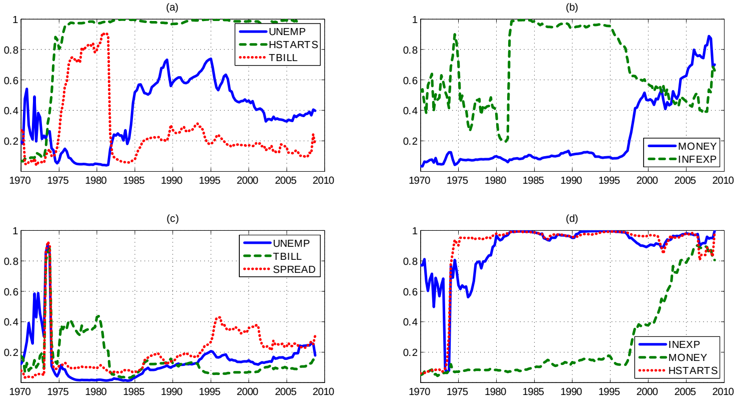

Figure 3: Posterior Probability of Inclusion of Predictors, $h = 4$. GDP deflator inflation top, PCE deflator inflation bottom

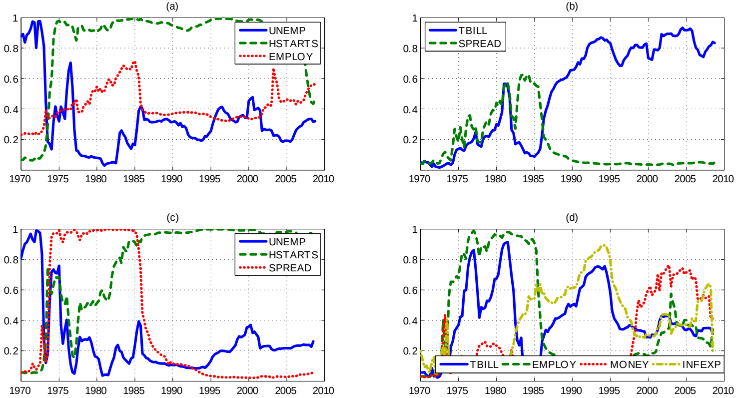

Figure 4: Posterior Probability of Inclusion of Predictors, $h = 8$. GDP deflator inflation top, PCE deflator inflation bottom

Forecast Performance

- Pseudo-out-of-sample forecasting exercise
- forecast evaluation begins in 1970Q1
- Measures of forecast performance using point forecasts
- Mean squared forecast error (MSFE) and mean absolute forecast error (MAFE).
- Forecast metric involving entire predictive distribution: the sum of log predictive likelihoods.
- Predictive likelihood = Predictive density for y_t (given data through time $t-1$) evaluated at the actual outcome.

Forecasting Methods

- DMA with $\alpha = \lambda = 0.99$.
- DMS with $\alpha = \lambda = 0.99$.
- DMA with $\alpha = \lambda = 0.95$.
- DMS with $\alpha = \lambda = 0.95$.
- DMA, with constant coefficiants ($\lambda = 1$, $\alpha = 0.99$)
- BMA as a special case of DMA (i.e. we set $\lambda = \alpha = 1$).
- TVP-AR(2)-X: Traditional TVP model .
- TVP-AR(2) model (as preceding but excluding predictors)

- Traditional g-prior BMA
- UC-SV: Unobserved components with stochastic volatility model of Stock and Watson (2007).
- Recursive OLS using AR(p)
- As preceding, but adding the predictors.
- Rolling OLS using AR(p) (window of 40 quarters)
- As preceding, but adding the predictors
- Random walk
- Note: in recursive and rolling OLS forecasts p selected at each point in time using BIC

Discussion of Log Predictive Likelihoods

- Preferred method of Bayesian forecast comparison
- Some variant of DMA or DMS always forecast best.
- DMS with $\alpha = \lambda = 0.95$ good for both measures of inflation at all horizons.
- Conventional BMA forecasts poorly.
- TVP-AR(2) and UC-SV have substantially lower predictive likelihoods than the DMA or DMS approaches.
- Of the non-DMA approaches, UC-SV approach of Stock and Watson (2007) consistently is the best performer.
- TVP model with all predictors tends to forecast poorly
- Shrinkage provided by DMA or DMS is of great value in forecasting.
- DMS tends to forecast a bit better than DMA

Discussion of MSFE and MAFE

- Patterns noted with predictive likelihoods mainly still hold (although DMA does better relative to DMS)
- Simple forecasting methods (AR(2) or random walk model) are inferior to DMA and DMS
- Rolling OLS using all predictors forecast bests among OLS-based methods.
- DMS and DMA with $\alpha = \lambda = 0.95$ always lead to lower MSFEs and MAFEs than rolling OLS with all the predictors.
- In some cases rolling OLS with all predictors leads to lower MSFEs and MAFEs than other implementations of DMA or DMS.
- In general: DMA and DMS look to be safe options. Usually they do best, but where not they do not go too far wrong
- Unlike other methods which might perform well in some cases, but very poorly in others

Forecast results: GDP deflator inflation, $h = 1$

	MAFE	MSFE	log(PL)
DMA ($\alpha = \lambda = 0.99$)	0.248	0.306	-0.292
DMS ($\alpha = \lambda = 0.99$)	0.256	0.318	-0.277
DMA ($\alpha = \lambda = 0.95$)	0.248	0.310	-0.378
DMS ($\alpha = \lambda = 0.95$)	0.235	0.297	-0.237
DMA ($\lambda = 1$, $\alpha = 0.99$)	0.249	0.306	-0.300
BMA (DMA with $\alpha = \lambda = 1$)	0.256	0.316	-0.320
TVP-AR(2) ($\lambda = 0.99$)	0.260	0.327	-0.344
TVP-AR(2)-X ($\lambda = 0.99$)	0.309	0.424	-0.423
BMA-MCMC ($g = \frac{1}{T}$)	0.234	0.303	-0.369
UC-SV ($\gamma = 0.2$)	0.256	0.332	-0.320
Recursive OLS - AR(BIC)	0.251	0.326	-
Recursive OLS - All Preds	0.265	0.334	-
Rolling OLS - AR(2)	0.251	0.325	-
Rolling OLS - All Preds	0.252	0.327	-
Random Walk	0.262	0.349	-

Forecast results: GDP deflator inflation, $h = 4$

	MAFE	MSFE	log(PL)
DMA ($\alpha = \lambda = 0.99$)	0.269	0.349	-0.421
DMS ($\alpha = \lambda = 0.99$)	0.277	0.361	-0.406
DMA ($\alpha = \lambda = 0.95$)	0.255	0.334	-0.455
DMS ($\alpha = \lambda = 0.95$)	0.249	0.316	-0.307
DMA ($\lambda = 1$, $\alpha = 0.99$)	0.277	0.355	-0.445
BMA (DMA with $\alpha = \lambda = 1$)	0.282	0.363	-0.463
TVP-AR(2) ($\lambda = 0.99$)	0.320	0.401	-0.480
TVP-AR(2)-X ($\lambda = 0.99$)	0.336	0.453	-0.508
BMA-MCMC ($g = \frac{1}{T}$)	0.285	0.364	-0.503
UC-SV ($\gamma = 0.2$)	0.311	0.396	-0.473
Recursive OLS - AR(BIC)	0.344	0.433	-
Recursive OLS - All Preds	0.302	0.376	-
Rolling OLS - AR(2)	0.328	0.425	-
Rolling OLS - All Preds	0.273	0.349	-
Random Walk	0.333	0.435	-

Forecast results: GDP deflator inflation, $h = 8$

	MAFE	MSFE	log(PL)
DMA ($\alpha = \lambda = 0.99$)	0.333	0.413	-0.583
DMS ($\alpha = \lambda = 0.99$)	0.338	0.423	-0.578
DMA ($\alpha = \lambda = 0.95$)	0.293	0.379	-0.570
DMS ($\alpha = \lambda = 0.95$)	0.295	0.385	-0.424
DMA ($\lambda = 1$, $\alpha = 0.99$)	0.346	0.423	-0.626
BMA (DMA with $\alpha = \lambda = 1$)	0.364	0.449	-0.690
TVP-AR(2) ($\lambda = 0.99$)	0.398	0.502	-0.662
TVP-AR(2)-X ($\lambda = 0.99$)	0.410	0.532	-0.701
BMA-MCMC ($g = \frac{1}{T}$)	0.319	0.401	-0.667
UC-SV ($\gamma = 0.2$)	0.350	0.465	-0.613
Recursive OLS - AR(BIC)	0.436	0.516	-
Recursive OLS - All Preds	0.369	0.441	-
Rolling OLS - AR(2)	0.380	0.464	-
Rolling OLS - All Preds	0.325	0.398	-
Random Walk	0.428	0.598	-

Forecast results: core inflation, $h = 1$

	MAFE	MSFE	log(PL)
DMA ($\alpha = \lambda = 0.99$)	0.253	0.322	-0.451
DMS ($\alpha = \lambda = 0.99$)	0.259	0.326	-0.430
DMA ($\alpha = \lambda = 0.95$)	0.267	0.334	-0.519
DMS ($\alpha = \lambda = 0.95$)	0.236	0.295	-0.348
DMA ($\lambda = 1$, $\alpha = 0.99$)	0.250	0.317	-0.444
BMA (DMA with $\alpha = \lambda = 1$)	0.259	0.331	-0.464
TVP-AR(2) ($\lambda = 0.99$)	0.280	0.361	-0.488
TVP-AR(2)-X ($\lambda = 0.99$)	0.347	0.492	-0.645
BMA-MCMC ($g = \frac{1}{T}$)	0.269	0.352	-0.489
UC-SV ($\gamma = 0.2$)	0.269	0.341	-0.474
Recursive OLS - AR(BIC)	0.310	0.439	-
Recursive OLS - All Preds	0.303	0.421	-
Rolling OLS - AR(2)	0.316	0.430	-
Rolling OLS - All Preds	0.289	0.414	-
Random Walk	0.294	0.414	-

Forecast results: core inflation, $h = 4$

	MAFE	MSFE	log(PL)
DMA ($\alpha = \lambda = 0.99$)	0.311	0.406	-0.622
DMS ($\alpha = \lambda = 0.99$)	0.330	0.431	-0.631
DMA ($\alpha = \lambda = 0.95$)	0.290	0.382	-0.652
DMS ($\alpha = \lambda = 0.95$)	0.288	0.353	-0.499
DMA ($\lambda = 1$, $\alpha = 0.99$)	0.315	0.412	-0.636
BMA (DMA with $\alpha = \lambda = 1$)	0.325	0.429	-0.668
TVP-AR(2) ($\lambda = 0.99$)	0.355	0.459	-0.668
TVP-AR(2)-X ($\lambda = 0.99$)	0.378	0.556	-0.764
BMA-MCMC ($g = \frac{1}{T}$)	0.307	0.414	-0.633
UC-SV ($\gamma = 0.2$)	0.340	0.443	-0.651
Recursive OLS - AR(BIC)	0.390	0.513	-
Recursive OLS - All Preds	0.325	0.442	-
Rolling OLS - AR(2)	0.378	0.510	-
Rolling OLS - All Preds	0.313	0.422	-
Random Walk	0.407	0.551	-

Forecast results: core inflation, $h = 8$

	h=8		
	MAFE	MSFE	log(PL)
DMA ($\alpha = \lambda = 0.99$)	0.357	0.448	-0.699
DMS ($\alpha = \lambda = 0.99$)	0.369	0.469	-0.699
DMA ($\alpha = \lambda = 0.95$)	0.317	0.403	-0.673
DMS ($\alpha = \lambda = 0.95$)	0.293	0.371	-0.518
DMA ($\lambda = 1$, $\alpha = 0.99$)	0.366	0.458	-0.733
BMA (DMA with $\alpha = \lambda = 1$)	0.397	0.490	-0.779
TVP-AR(2) ($\lambda = 0.99$)	0.450	0.573	-0.837
TVP-AR(2)-X ($\lambda = 0.99$)	0.432	0.574	-0.841
BMA-MCMC ($g = \frac{1}{T}$)	0.357	0.454	-0.788
UC-SV ($\gamma = 0.2$)	0.406	0.528	-0.774
Recursive OLS - AR(BIC)	0.463	0.574	-
Recursive OLS - All Preds	0.378	0.481	-
Rolling OLS - AR(2)	0.428	0.540	-
Rolling OLS - All Preds	0.338	0.436	-
Random Walk	0.531	0.698	-

Comparison to Greenbook Forecasts

- Greenbook forecasts are for GDP deflator inflation
- Published with a lag so different data span
- $h = 1$ forecasts from 1970Q1 through 2004Q1
- $h = 4$ forecasts from 1974Q1 through 2004Q4
- Greenbook forecasts are point forecasts (no predictive likelihoods)
- Table presents MAFEs relative to simple random walk
- For $h = 1$ DMA beats Greenbook
- When $h = 4$ we beat Greenbook forecasts when $\alpha = \lambda = 0.95$

Comparison of DMA with Greenbook forecasts: MAFE

	$h = 1$	$h = 4$
Greenbook forecasts	0.91	0.84
DMA $\alpha = \lambda = 0.99$	0.80	0.94
DMA $\alpha = \lambda = 0.95$	0.77	0.83

Conclusions

- When forecasting in the presence of change/breaks/turbulence want an approach which:
- Allows for forecasting model to change over time
- Allows for marginal effects of predictors to change over time
- Automatically does the shrinkage necessary to reduce risk of over-parameterizations/over-fitting
- In theory, DMA and DMS should satisfy these criteria
- In practice, we find DMA and DMS to forecast well in an exercise involving US inflation.

Summary

- MCMC algorithms such as the Gibbs sampler are modular in nature (sequentially draw from blocks)
- By combining simple blocks together you can end up with very flexible models
- This is strategy pursued here.
- For state space models there are a standard set of algorithms which can be combined together in various ways to produce quite sophisticated models
- Our MCMC algorithms for complicated models all combine simpler algorithms.
- E.g. Primiceri's complicated model involves blocks which use Carter and Kohn's algorithm and blocks which use Kim, Shephard and Chib's algorithm (and even the latter relies upon Carter and Kohn)