Bayesian State Space Models
State space methods are used for a wide variety of time series problems.

They are important in and of themselves in economics (e.g. trend-cycle decompositions, structural time series models, dealing with missing observations, etc.)

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, constant \( \beta \) is unrealistic.

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 \( \varepsilon_t \) to be i.i.d. \( N(0, \Sigma) \).

In empirical macroeconomics, this is often unrealistic.

Want to have \( \text{var} (\varepsilon_t) = \Sigma_t \).

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 \( \varepsilon_t \) 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)
- \( \beta_t \) is \( k \times 1 \) vector of states (e.g. VAR coefficients)
- \( \varepsilon_t \) ind \( N(0, \Sigma_t) \)
- \( u_t \) ind \( N(0, Q_t) \).
- \( \varepsilon_t \) 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 $\delta$, $T_t$, $\Sigma_t$ and $Q_t$ (called “system matrices”) posterior simulators for $\beta_t$ for $t = 1, \ldots, T$ exist.


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: $\beta^t = (\beta_1', ..., \beta_t')'$ 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 $\beta^T$ 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 $\beta^T$:
  \[
  \beta_{t+1} | \beta_t, Q \sim N(\beta_t, Q)
  \]
- Formally:
  \[
  p(\beta^T | Q) = \prod_{t=1}^{T} p(\beta_t | \beta_{t-1}, Q)
  \]
- Hierarchical: since it depends on $Q$ which, in turn, requires its own prior.
Note $\beta_0$ enters prior for $\beta_1$.

Need prior for $\beta_0$

Standard treatments exist.

E.g. assume $\beta_0 = 0$, then:

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

Or Carter and Kohn (1994) simply assume $\beta_0$ has some prior that researcher chooses
Convenient to use Wishart priors for $\Sigma^{-1}$ and $Q^{-1}$

\[ \Sigma^{-1} \sim \mathcal{W} (S^{-1}, \nu) \]

\[ Q^{-1} \sim \mathcal{W} (Q^{-1}, \nu_Q) \]
Want MCMC algorithm which sequentially draws from
\[ p \left( \Sigma^{-1} | y^T, \beta^T, Q \right), \ p \left( Q^{-1} | y^T, \Sigma, \beta^T \right) \text{ and } p \left( \beta^T | y^T, \Sigma, Q \right). \]

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 \( p \left( \Sigma^{-1} | y^T, \beta^T, Q \right) \text{ and } p \left( Q^{-1} | y^T, \Sigma, \beta^T \right) \) using methods similar to those used in section on VAR independent Normal-Wishart model.
Conditional on $\beta^T$, measurement equation is like a VAR with known coefficients.

This leads to:

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

where

$$\bar{v} = T + \nu$$

$$\bar{S} = S + \sum_{t=1}^{T} (y_t - W_t \delta - Z_t \beta_t) (y_t - W_t \delta - Z_t \beta_t)'$$
Conditional on $\beta^T$, state equation is also like a VAR with known coefficients.

This leads to:

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

where

$$ \overline{v}_Q = T + \nu_Q $$

$$ \overline{Q} = Q + \sum_{t=1}^{T} (\beta_{t+1} - \beta_t) (\beta_{t+1} - \beta_t)' $$
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)
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).
- \( \theta \) are structural parameters (e.g. parameters for steady states, tastes, technology, policy and driving the exogenous shocks).
- \( u_t \) are structural shocks \((\mathcal{N}(0, I))\).
- \( \Gamma_j (\theta) \) are often highly nonlinear functions of \( \theta \).
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 \( B(\theta) \)
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 $\theta$ 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 for restricted VAR could be used to carry out inference on $\theta$.
- Unfortunately, restrictions in $A(\theta)$ and $B(\theta)$ are typically nonlinear and complicated.
- Parameters in $\theta$ 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”
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 $\theta$, no Gibbs sampler, etc...

Solution: Metropolis-Hastings algorithm (see Topic 2)
• 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
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:

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

\[
h_{t+1} = \mu + \phi (h_t - \mu) + \eta_t
\]

- \( \varepsilon_t \) is i.i.d. \( N(0, 1) \) and \( \eta_t \) is i.i.d. \( N\left(0, \sigma^2_\eta\right) \). \( \varepsilon_t \) and \( \eta_s \) 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 $\mu$ is the unconditional mean of $h_t$.
• Initial conditions: if $|\phi| < 1$ (stationary) then:

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

• if $\phi = 1$, $\mu$ drops out of the model and However, when $\phi = 1$, need a prior such as $h_0 \sim N (h, V_h)$
• e.g. Primiceri (2005) chooses $V_h$ using training sample
MCMC Algorithm for Stochastic Volatility Model

- MCMC algorithm involves sequentially drawing from $p\left(h^T | y^T, \mu, \phi, \sigma^2_\eta\right)$, $p\left(\phi | y^T, \mu, \sigma^2_\eta, h^T\right)$, $p\left(\mu | y^T, \phi, \sigma^2_\eta, h^T\right)$ and $p\left(\sigma^2_\eta | y^T, \mu, \phi, h^T\right)$

- Last three standard forms based on results from Normal linear regression model and will not present here.

- Several algorithms exist for $p\left(h^T | y^T, \mu, \phi, \sigma^2_\eta\right)$

- 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:

\[ y_t^* = h_t + \epsilon_t^* \]

where \( y_t^* = \ln(y_t^2) \) and \( \epsilon_t^* = \ln(\epsilon_t^2) \).

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. \( \epsilon_t^* = \ln(\epsilon_t^2) \))

Idea: use mixture of different Normal distributions to approximate distribution of \( \epsilon_t^* \).
Mixtures of Normal distributions are very flexible and have been used widely in many fields to approximate unknown or inconvenient distributions.

\[ p(\varepsilon_t^*) \approx \sum_{i=1}^{7} q_i f_{\mathcal{N}}(\varepsilon_t^* | m_i, \nu_i^2) \]

where \( f_{\mathcal{N}}(\varepsilon_t^* | m_i, \nu_i^2) \) is the p.d.f. of a \( \mathcal{N}(m_i, \nu_i^2) \)

since \( \varepsilon_t \) is \( \mathcal{N}(0, 1) \), \( \varepsilon_t^* \) involves no unknown parameters

Thus, \( q_i, m_i, \nu_i^2 \) for \( i = 1, \ldots, 7 \) 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, \( s_t \in \{1, 2, \ldots, 7\} \)

\[
\begin{align*}
\varepsilon_t^* | s_t = i & \sim N(m_i, \nu_i^2) \\
\Pr(s_t = i) & = q_i
\end{align*}
\]

MCMC algorithm does not draw from \( p\left( h^T | y^T, \mu, \phi, \sigma^2_{\eta} \right) \), but from \( p\left( h^T | y^T, \mu, \phi, \sigma^2_{\eta}, s^T \right) \).

But, conditional on \( s^T \), knows which of the Normals \( \varepsilon_t^* \) comes from.

Result is a Normal linear state space model and familiar algorithm can be used.

Finally, need \( p\left( s^T | y^T, \mu, \phi, \sigma^2_{\eta}, h^T \right) \) but this has simple form (see Kim, Shephard and Chib, 1998)
Multivariate Stochastic Volatility

- $y_t$ is now $M \times 1$ vector and $\varepsilon_t$ is i.i.d. $N(0, \Sigma_t)$.
- Many ways of allowing $\Sigma_t$ to be time-varying
- But must worry about overparameterization problems
- $\Sigma_t$ for $t = 1, \ldots, T$ 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 \( \eta_{it} \) 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 \( \Sigma_t \) to be diagonal often will be a bad idea.
Multivariate Stochastic Volatility Model 2

- Cogley and Sargent (2005, RED)

\[ \Sigma_t = L^{-1} D_t L^{-1}' \]

- \( D_t \) 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 $\text{cov} (\varepsilon_{1t}, \varepsilon_{2t}) = d_{1t} L_{21}$ 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.
Primiceri (2005, ReStud):

\[ \Sigma_t = L_t^{-1} D_t L_t^{-1}' \]

- \( L_t \) is same as Cogley-Sargent’s \( L \) but is now time varying.
- Does not restrict \( \Sigma_t \) 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

$$l_t = \left( L_{21,t}, L_{31,t}, L_{32,t}, \ldots, L_{p(p-1),t} \right)'$$

$$l_{t+1} = l_t + \zeta_t$$

$\zeta_t$ is i.i.d. $N(0, D_\zeta)$ and $D_\zeta$ 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 $D_\zeta$ is not diagonal have to be careful (no longer Normal state space model)
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
If time permits, I will go through this application involving state space methods

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:

\[ y_t = \phi + x_{t-1}' \beta + \sum_{j=1}^{p} \gamma_j y_{t-j} + \epsilon_t \]

- \( y_t \) 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.
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: $\hat{\theta} =$ estimate using data through $\tau$.

So $\hat{\theta}$ will change (a bit) with $\tau$, but can change too slowly

Rolling forecasts use: $\hat{\theta}$ an estimate using data from $\tau - \tau_0$ through $\tau$.

Better at capturing parameter change, but need to choose $\tau_0$

Recursive and rolling forecasts might be imperfect solutions

Why not use a model which formally models the parameter change as well?
TVP models gaining popularity in empirical macroeconomics

\[
y_t = z_t \theta_t + \epsilon_t \\
\theta_t = \theta_{t-1} + \eta_t
\]

\[\epsilon_t \overset{ind}{\sim} N(0, H_t)\]
\[\eta_t \overset{ind}{\sim} N(0, Q_t)\]

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
Define $K$ models which have $z_t^{(k)}$ for $k = 1, \ldots, K$, as predictors.

$z_t^{(k)}$ is subset of $z_t$.

Set of models:

$$y_t = z_t^{(k)} \theta_t^{(k)} + \epsilon_t^{(k)}$$

$$\theta_{t+1} = \theta_t^{(k)} + \eta_t^{(k)}$$

$\epsilon_t^{(k)}$ is $N\left(0, H_t^{(k)}\right)$

$\eta_t^{(k)}$ is $N\left(0, Q_t^{(k)}\right)$

Let $L_t \in \{1, 2, \ldots, K\}$ 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 $\tau$ 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 $p_{ij} = \Pr (L_t = i | L_{t-1} = j)$ for $i, j = 1, \ldots, K$.

“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 $t - 1$)

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

Textbook formula for $\hat{\theta}_{t-1}$ and $\Sigma_{t-1|t-1}$

Then update

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

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

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 \( \lambda^j \).
• Or effective window size of \( \frac{1}{1-\lambda} \).
• Choose value of \( \lambda \) 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).
Goal for forecasting at time $t$ given data available at time $t-1$ is
$$\pi_{t|t-1,k} \equiv \Pr(L_t = k|y^{t-1})$$

Can average across $k = 1, \ldots, K$ forecasts using $\pi_{t|t-1,k}$ as weights (DMA)

E.g. point forecasts ($\hat{\theta}_{t-1}^{(k)}$ from Kalman filter in model $k$):

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

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

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

0 < $\alpha$ \leq 1 is forgetting factor with similar interpretation to $\lambda$

Focus on $\alpha \in [0.95, 1.00]$
Interpretation of forgetting factor $\alpha$

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, $\alpha$

$\alpha = 0.99$: 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
Data from 1960Q1 through 2008Q4
Real time data (forecasting at time $\tau$ using data as known at time $\tau$)
Two measures 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

$$E (Size_t) = \sum_{k=1}^{K} \pi_{t|t-1,k} Size_k$$
Figure 1: Expected Number of Predictors
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

- recursive 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 coefficients ($\lambda = 1, \alpha = 0.99$)
- 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
- 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
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.
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, $h = 1$

<table>
<thead>
<tr>
<th>Method</th>
<th>MAFE</th>
<th>MSFE</th>
<th>log(PL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DMA ($\alpha = \lambda = 0.99$)</td>
<td>0.248</td>
<td>0.306</td>
<td>-0.292</td>
</tr>
<tr>
<td>DMS ($\alpha = \lambda = 0.99$)</td>
<td>0.256</td>
<td>0.318</td>
<td>-0.277</td>
</tr>
<tr>
<td>DMA ($\alpha = \lambda = 0.95$)</td>
<td>0.248</td>
<td>0.310</td>
<td>-0.378</td>
</tr>
<tr>
<td>DMS ($\alpha = \lambda = 0.95$)</td>
<td>0.235</td>
<td>0.297</td>
<td>-0.237</td>
</tr>
<tr>
<td>DMA ($\lambda = 1, \alpha = 0.99$)</td>
<td>0.249</td>
<td>0.306</td>
<td>-0.300</td>
</tr>
<tr>
<td>BMA (DMA with $\alpha = \lambda = 1$)</td>
<td>0.256</td>
<td>0.316</td>
<td>-0.320</td>
</tr>
<tr>
<td>TVP-AR(2) ($\lambda = 0.99$)</td>
<td>0.260</td>
<td>0.327</td>
<td>-0.344</td>
</tr>
<tr>
<td>TVP-AR(2)-X ($\lambda = 0.99$)</td>
<td>0.309</td>
<td>0.424</td>
<td>-0.423</td>
</tr>
<tr>
<td>BMA-MCMC ($g = \frac{1}{T}$)</td>
<td>0.234</td>
<td>0.303</td>
<td>-0.369</td>
</tr>
<tr>
<td>UC-SV ($\gamma = 0.2$)</td>
<td>0.256</td>
<td>0.332</td>
<td>-0.320</td>
</tr>
<tr>
<td>Recursive OLS - AR(BIC)</td>
<td>0.251</td>
<td>0.326</td>
<td>-</td>
</tr>
<tr>
<td>Recursive OLS - All Preds</td>
<td>0.265</td>
<td>0.334</td>
<td>-</td>
</tr>
<tr>
<td>Rolling OLS - AR(2)</td>
<td>0.251</td>
<td>0.325</td>
<td>-</td>
</tr>
<tr>
<td>Rolling OLS - All Preds</td>
<td>0.252</td>
<td>0.327</td>
<td>-</td>
</tr>
<tr>
<td>Random Walk</td>
<td>0.262</td>
<td>0.349</td>
<td>-</td>
</tr>
</tbody>
</table>
Forecast results: GDP deflator inflation, $h = 4$

<table>
<thead>
<tr>
<th>Method</th>
<th>MAFE</th>
<th>MSFE</th>
<th>log(PL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DMA ($\alpha = \lambda = 0.99$)</td>
<td>0.269</td>
<td>0.349</td>
<td>-0.421</td>
</tr>
<tr>
<td>DMS ($\alpha = \lambda = 0.99$)</td>
<td>0.277</td>
<td>0.361</td>
<td>-0.406</td>
</tr>
<tr>
<td>DMA ($\alpha = \lambda = 0.95$)</td>
<td>0.255</td>
<td>0.334</td>
<td>-0.455</td>
</tr>
<tr>
<td>DMS ($\alpha = \lambda = 0.95$)</td>
<td>0.249</td>
<td>0.316</td>
<td>-0.307</td>
</tr>
<tr>
<td>DMA ($\lambda = 1, \alpha = 0.99$)</td>
<td>0.277</td>
<td>0.355</td>
<td>-0.445</td>
</tr>
<tr>
<td>BMA (DMA with $\alpha = \lambda = 1$)</td>
<td>0.282</td>
<td>0.363</td>
<td>-0.463</td>
</tr>
<tr>
<td>TVP-AR(2) ($\lambda = 0.99$)</td>
<td>0.320</td>
<td>0.401</td>
<td>-0.480</td>
</tr>
<tr>
<td>TVP-AR(2)-X ($\lambda = 0.99$)</td>
<td>0.336</td>
<td>0.453</td>
<td>-0.508</td>
</tr>
<tr>
<td>BMA-MCMC ($g = \frac{1}{T}$)</td>
<td>0.285</td>
<td>0.364</td>
<td>-0.503</td>
</tr>
<tr>
<td>UC-SV ($\gamma = 0.2$)</td>
<td>0.311</td>
<td>0.396</td>
<td>-0.473</td>
</tr>
<tr>
<td>Recursive OLS - AR(BIC)</td>
<td>0.344</td>
<td>0.433</td>
<td></td>
</tr>
<tr>
<td>Recursive OLS - All Preds</td>
<td>0.302</td>
<td>0.376</td>
<td></td>
</tr>
<tr>
<td>Rolling OLS - AR(2)</td>
<td>0.328</td>
<td>0.425</td>
<td></td>
</tr>
<tr>
<td>Rolling OLS - All Preds</td>
<td>0.273</td>
<td>0.349</td>
<td></td>
</tr>
<tr>
<td>Random Walk</td>
<td>0.333</td>
<td>0.435</td>
<td></td>
</tr>
</tbody>
</table>
Forecast results: GDP deflator inflation, $h = 8$

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

<table>
<thead>
<tr>
<th>Model Description</th>
<th>MAFE</th>
<th>MSFE</th>
<th>log(PL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DMA ($\alpha = 0.99$)</td>
<td>0.253</td>
<td>0.322</td>
<td>-0.451</td>
</tr>
<tr>
<td>DMS ($\alpha = 0.99$)</td>
<td>0.259</td>
<td>0.326</td>
<td>-0.430</td>
</tr>
<tr>
<td>DMA ($\alpha = 0.95$)</td>
<td>0.267</td>
<td>0.334</td>
<td>-0.519</td>
</tr>
<tr>
<td>DMS ($\alpha = 0.95$)</td>
<td>0.236</td>
<td>0.295</td>
<td>-0.348</td>
</tr>
<tr>
<td>DMA ($\lambda = 1, \alpha = 0.99$)</td>
<td>0.255</td>
<td>0.317</td>
<td>-0.444</td>
</tr>
<tr>
<td>BMA (DMA with $\alpha = \lambda = 1$)</td>
<td>0.259</td>
<td>0.331</td>
<td>-0.464</td>
</tr>
<tr>
<td>TVP-AR(2) ($\lambda = 0.99$)</td>
<td>0.280</td>
<td>0.361</td>
<td>-0.488</td>
</tr>
<tr>
<td>TVP-AR(2)-X ($\lambda = 0.99$)</td>
<td>0.347</td>
<td>0.492</td>
<td>-0.645</td>
</tr>
<tr>
<td>BMA-MCMC ($g = \frac{1}{T}$)</td>
<td>0.269</td>
<td>0.352</td>
<td>-0.489</td>
</tr>
<tr>
<td>UC-SV ($\gamma = 0.2$)</td>
<td>0.269</td>
<td>0.341</td>
<td>-0.474</td>
</tr>
<tr>
<td>Recursive OLS - AR(BIC)</td>
<td>0.310</td>
<td>0.439</td>
<td>-</td>
</tr>
<tr>
<td>Recursive OLS - All Preds</td>
<td>0.303</td>
<td>0.421</td>
<td>-</td>
</tr>
<tr>
<td>Rolling OLS - AR(2)</td>
<td>0.316</td>
<td>0.430</td>
<td>-</td>
</tr>
<tr>
<td>Rolling OLS - All Preds</td>
<td>0.289</td>
<td>0.414</td>
<td>-</td>
</tr>
<tr>
<td>Random Walk</td>
<td>0.294</td>
<td>0.414</td>
<td>-</td>
</tr>
</tbody>
</table>
### Forecast results: core inflation, $h = 4$

<table>
<thead>
<tr>
<th>Model Description</th>
<th>MAFE</th>
<th>MSFE</th>
<th>log(PL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DMA ($\alpha = \lambda = 0.99$)</td>
<td>0.311</td>
<td>0.406</td>
<td>-0.622</td>
</tr>
<tr>
<td>DMS ($\alpha = \lambda = 0.99$)</td>
<td>0.330</td>
<td>0.431</td>
<td>-0.631</td>
</tr>
<tr>
<td>DMA ($\alpha = \lambda = 0.95$)</td>
<td>0.290</td>
<td>0.382</td>
<td>-0.652</td>
</tr>
<tr>
<td>DMS ($\alpha = \lambda = 0.95$)</td>
<td>0.288</td>
<td>0.353</td>
<td>-0.499</td>
</tr>
<tr>
<td>DMA ($\lambda = 1, \alpha = 0.99$)</td>
<td>0.315</td>
<td>0.412</td>
<td>-0.636</td>
</tr>
<tr>
<td>BMA (DMA with $\alpha = \lambda = 1$)</td>
<td>0.325</td>
<td>0.429</td>
<td>-0.668</td>
</tr>
<tr>
<td>TVP-AR(2) ($\lambda = 0.99$)</td>
<td>0.355</td>
<td>0.459</td>
<td>-0.668</td>
</tr>
<tr>
<td>TVP-AR(2)-X ($\lambda = 0.99$)</td>
<td>0.378</td>
<td>0.556</td>
<td>-0.764</td>
</tr>
<tr>
<td>BMA-MCMC ($g = \frac{1}{T}$)</td>
<td>0.307</td>
<td>0.414</td>
<td>-0.633</td>
</tr>
<tr>
<td>UC-SV ($\gamma = 0.2$)</td>
<td>0.340</td>
<td>0.443</td>
<td>-0.651</td>
</tr>
<tr>
<td>Recursive OLS - AR(BIC)</td>
<td>0.390</td>
<td>0.513</td>
<td>-</td>
</tr>
<tr>
<td>Recursive OLS - All Preds</td>
<td>0.325</td>
<td>0.442</td>
<td>-</td>
</tr>
<tr>
<td>Rolling OLS - AR(2)</td>
<td>0.378</td>
<td>0.510</td>
<td>-</td>
</tr>
<tr>
<td>Rolling OLS - All Preds</td>
<td>0.313</td>
<td>0.422</td>
<td>-</td>
</tr>
<tr>
<td>Random Walk</td>
<td>0.407</td>
<td>0.551</td>
<td>-</td>
</tr>
</tbody>
</table>
Forecast results: core inflation, $h = 8$

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

- 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 overparameterization/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.
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)