

# 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.)
- They can be used to deal with unit root issues and ARMA
- Also time-varying parameter (TVP) models can be used to deal with parameter change/structural breaks/regime change
- Dynamic factor models are state space models
- Stochastic volatility are state space models
- Advantage of state space models: well-developed set of MCMC algorithms for doing Bayesian inference

# The Local Level Model

- Explain basic ideas in simplest state space model: the local level model
- For  $t = 1, \dots, T$  have

$$y_t = \alpha_t + \varepsilon_t$$

- $\varepsilon_t$  is i.i.d.  $N(0, h^{-1})$ .
- $\alpha_t$  which is not observed (called a *state*) and follows random walk for  $t = 1, \dots, T - 1$ :

$$\alpha_{t+1} = \alpha_t + u_t$$

- $u_t$  is i.i.d.  $N(0, Q)$
- $\varepsilon_t$  and  $u_s$  are independent of one another for all  $s$  and  $t$ .
- First equation: measurement (observation) equation, second state equation
- $\alpha_1$  is *initial condition*.

# Relationship to Other Models

- Can write

$$\Delta y_t = \varepsilon_t - \varepsilon_{t-1} + u_{t-1}$$

- $\Delta y_t$  is stationary ( $I(0)$ ) whereas  $y_t$  has unit root ( $I(1)$ )

- Can write

$$\alpha_t = \alpha_1 + \sum_{j=1}^{t-1} u_j$$

- this is a trend (stochastic trend)
- local level model decomposes  $y_t$ , into a trend component,  $\alpha_t$ , and an error or irregular component,  $\varepsilon_t$ .
- Test of whether  $Q = 0$  is one way of testing for a unit root.
- These results illustrate how all usual univariate time series things: ARIMA modelling, unit root testing, etc. can be done in state space framework

# Relationship to Other Models

- $\alpha_t$  is the mean (or level) of  $y_t$ .
- Mean is varying over time, hence terminology *local level model*
- Measurement equation can be interpreted as simple example of regression model involving only an intercept.
- But the intercept varies over time: time varying parameter model
- Extensions of local level model used to investigate parameter change in various contexts.

# The Likelihood Function of the Local Level Model

- Define  $y = (y_1, \dots, y_T)'$  and  $\varepsilon = (\varepsilon_1, \dots, \varepsilon_T)'$  then local level model:

$$y = I_T \alpha + \varepsilon$$

- This is a regression model with explanatory variables  $I_T$  and coefficients  $\alpha = (\alpha_1, \dots, \alpha_T)'$
- Likelihood function has standard form for the Normal linear regression model
- Note relation to Fat Data:  $T$  observations and  $T$  explanatory variables
- Here hierarchical prior is provided by state equation

- State equation gives us:

$$\alpha_{t+1} | \alpha_t, Q \sim N(\alpha_t, q)$$

- Or

$$p(\alpha | Q) = \prod_{t=1}^T p(\alpha_{t+1} | \alpha_t, Q)$$

- This is a hierarchical prior: since it depends on  $Q$  which, in turn, requires its own prior.
- The fact that it is a Normal prior means we can use standard results for Normal linear regression model

# Posterior for Local Level Model

- I will not repeat exact formula here
- See Topic 1 slides or page 187 of my textbook for natural conjugate case
- But the formulae will depend on  $Q$
- Textbook discusses (pages 188-190) discusses one estimation method, see below for MCMC method
- An issue arises:  $\alpha$  is  $T \times 1$  which can be very large (dimension of states even larger in general state space models)
- Remember: if regression had  $k$  explanatory variables, posterior involved manipulations (inverting, etc.)  $k \times k$  matrices
- If  $k = T$  or more, this rapidly gets demanding (or impossible)
- For state space models, special methods based on Kalman filtering used to avoid such manipulations
- Will discuss below, but remember that state space models basically just regression models with a particular hierarchical prior



# Filtering versus Smoothing in the Local Level Model

- Notation: superscripts for all observations up to a specific time
- E.g.  $y^T = (y_1, \dots, y_T)'$  is all observations in the sample
- $\alpha^t = (\alpha_1, \dots, \alpha_t)'$  is all states up to the current period ( $t$ )
- Filtering = using  $y^t$
- $E(\alpha_t | y^t)$  is the filtered estimate of the state
- $E(y_{t+1} | y^t)$  is estimate of  $y_{t+1}$  (unknown at time  $t$ )
- Used for real time forecasting
- Smoothing = using  $y^T$
- $E(\alpha_t | y^T)$  is smoothed estimate of state
- E.g. estimate of trend inflation using the full sample of data

# The Kalman Filter

- I will not derive or state exact formulae, just the main ideas
- Good reference: Durbin and Koopman, Time Series Analysis by State Space Methods
- Formulae below depend on  $Q$  and  $h$ , for now assume it is known
- Can prove

$$\alpha_t | y^{t-1} \sim N(a_{t|t-1}, P_{t|t-1})$$

$$\alpha_t | y^t \sim N(a_{t|t}, P_{t|t})$$

- Kalman filter involves simple formulae linking  $a_{t|t-1}, P_{t|t-1}, a_{t|t}, P_{t|t}$
- Also formula for predictive density  $p(y_{t+1} | y^t)$  which can be used for real time forecasting
- Formula for likelihood function (used for maximum likelihood estimation)

# Kalman Filter Recursions

- Start with initial condition,  $a_{1|1}, P_{1|1}$  (Bayesians assume prior)
- Calculate  $a_{2|1}, P_{2|1}$  using Kalman filtering formulae
- Calculate  $a_{2|2}, P_{2|2}$
- ...
- Calculate  $a_{t|t-1}, P_{t|t-1}$
- Calculate  $a_{t|t}, P_{t|t}$
- etc.

# Kalman Filter Recursions

- Each calculation on previous slide only depended on the last one
- New observation added, only need to update using this
- Simplifies computation: no need for manipulations involving  $T \times T$  matrices
- At every point in time get filtered estimate of state, predictive density, etc.
- Run the Kalman filter from  $t = 1, \dots, T$

# State Smoothing

- Smoothing uses full sample,  $y^T$
- Suitable for estimation (e.g. estimating trend inflation)
- Standard recursive formulae exist with same “update one observation at a time”
- Can prove

$$\alpha_t | y^T \sim N(a_{t|T}, P_{t|T})$$

- First run Kalman filter from  $t = 1, \dots, T$
- Then state smoother from  $t = T, \dots, 1$
- Set of simple recursive formulae for  $a_{t|T}$  and  $P_{t|T}$

# Summary of Estimation in Local Level Model

- Local level model has parameters  $\alpha^T$ ,  $Q$  and  $h$
- Kalman filter and state smoother provides formula for  $p(\alpha^T | y^T, Q, h)$  and  $p(\alpha^T | y^t, Q, h)$
- And  $p(y^{t+1} | y^t, Q, h)$  for forecasting
- Bayesian can complete the Gibbs sampler with  $p(Q | y^T, h, \alpha^T)$  and  $p(h | y^T, Q, \alpha^T)$
- Exact forms depend on prior, but simple based on Normal linear regression model

# The Normal Linear State Space Model

- 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 regression model
- Illustrate as though for a regression or AR model, but much more general
- General theory has  $y_t$  being  $M \times 1$  vector
- Usual for macroeconomics: VARs have  $M$  variables, DSGE models involve  $M$  variables
- But my applications will be for single equation:  $M = 1$

# The Normal Linear State Space Model

- $W_t$  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. regression or AR 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  $t = 1, \dots, T$  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
- I have outlined general form for the local level model above
- Recently other algorithms have been proposed in several papers by Joshua Chan (Australian National University) and Bill McCausland (University of Montreal)
- These do not use Kalman filter, but exploit special band structure of large  $T \times T$  matrices to invert key matrices directly

- Notation:  $\beta^t = (\beta'_1, \dots, \beta'_t)'$  stacks all the states up to time  $t$  (and similar superscript  $t$  convention for other things)
- Gibbs sampler:  $p(\beta^T | y^T, \delta, T^T, \Sigma^T, Q^T)$  drawn use such an algorithm
- $p(\delta | y^T, \beta^T, T^T, \Sigma^T, Q^T)$ ,  $p(T^T | y^T, \beta^T, \delta, \Sigma^T, Q^T)$ ,  $p(\Sigma^T | y^T, \beta^T, \delta, T^T, Q^T)$  and  $p(Q^T | y^T, \beta^T, \delta, T^T, \Sigma^T)$  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  $Q_t = Q$
- Computer tutorial 4 considers a time-varying parameter AR model
- $Z_t$  contains lags of dependent variable,  $\delta = 0$ ,  $T_t = I$  and  $Q_t$  is a diagonal matrix

# Example of an MCMC Algorithm

- Special case  $\delta = 0$ ,  $T_t = I$ ,  $\Sigma_t = h$  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
- $h$  is error precision in measurement equation, just use Gamma prior for it as in Normal linear regression model

- Common to use Wishart prior for  $Q^{-1}$
- 

$$Q^{-1} \sim W(\underline{Q}^{-1}, \underline{\nu}_Q)$$

- Remember regression models had parameters  $\beta$  and  $\sigma^2$
- There proved convenient to work with  $h = \frac{1}{\sigma^2}$
- With  $Q$  proves convenient to work with  $Q^{-1}$
- In regression  $h$  typically had Gamma distribution
- With state equations (more than one equation)  $Q^{-1}$  will typically have Wishart distribution
- Wishart is matrix generalization of Gamma
- Details see appendix to textbook.
- If  $\Sigma^{-1}$  is  $W(C, c)$  then “Mean” is  $cC$  and  $c$  is degrees of freedom.
- Note: easy to take random draws from Wishart.

- Want MCMC algorithm which sequentially draws from  $p(h^{-1}|y^T, \beta^T, Q)$ ,  $p(Q^{-1}|y^T, h, \beta^T)$  and  $p(\beta^T|y^T, h, Q)$ .
- For  $p(\beta^T|y^T, h, Q)$  use standard algorithm for state space models (e.g. Carter and Kohn, 1994)
- Can derive  $p(h|y^T, \beta^T, Q)$  using Normal linear regression model results
- That is, conditional on  $\beta^T$ , measurement equation is just a regression with known coefficients.

- $p(Q^{-1}|y^T, h, \beta^T)$  use multiple equation extension of Normal linear regression model
- Conditional on  $\beta^T$ , state equation is also like a series of regression equations
- This leads to:

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

- where

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

- 

$$\bar{Q} = \underline{Q} + \sum_{t=1}^T (\beta_{t+1} - \beta_t) (\beta_{t+1} - \beta_t)'$$



# 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)

# 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)
- $\theta$  are structural parameters (e.g. parameters for steady states, tastes, technology, policy, etc.).
- $u_t$  are structural shocks ( $N(0, I)$ ).
- $\Gamma_j(\theta)$  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  $B(\theta)$

# 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 often 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 linear methods similar to regressions 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”

## 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  $\theta$ , no Gibbs sampler, etc...
- Solution: Metropolis-Hastings algorithm (see my textbook chapter 5, section 5)

- 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, dynamic factor 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



# Stochastic Volatility

- Popular in finance, but increasingly macroeconomists realize importance of allowing for time-varying volatility
- Note: multivariate stochastic volatility in VARs is very popular (also nonlinear state space model, simple extension of univariate case)
- 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.  $N\left(0, \sigma_\eta^2\right)$ .  $\varepsilon_t$  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:

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

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

# MCMC Algorithm for Stochastic Volatility Model

- MCMC algorithm involves sequentially drawing from  $p(h^T | y^T, \mu, \phi, \sigma_\eta^2)$ ,  $p(\phi | y^T, \mu, \sigma_\eta^2, h^T)$ ,  $p(\mu | y^T, \phi, \sigma_\eta^2, h^T)$  and  $p(\sigma_\eta^2 | y^T, \mu, \phi, h^T)$
- Last three standard forms based on results from Normal linear regression model and will not present here.
- Several algorithms exist for  $p(h^T | y^T, \mu, \phi, \sigma_\eta^2)$
- 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  $y_t^* = \ln(y_t^2)$  and  $\varepsilon_t^* = \ln(\varepsilon_t^2)$ .
- 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.  $\varepsilon_t^* = \ln(\varepsilon_t^2)$ )
- Idea: use mixture of different Normal distributions to approximate distribution of  $\varepsilon_t^*$ .

- Mixtures of Normal distributions are very flexible and have been used widely in many fields to approximate unknown or inconvenient distributions.

- 

$$p(\varepsilon_t^*) \approx \sum_{i=1}^7 q_i f_N(\varepsilon_t^* | m_i, v_i^2)$$

- where  $f_N(\varepsilon_t^* | m_i, v_i^2)$  is the p.d.f. of a  $N(m_i, v_i^2)$
- since  $\varepsilon_t$  is  $N(0, 1)$ ,  $\varepsilon_t^*$  involves no unknown parameters
- Thus,  $q_i, m_i, v_i^2$  for  $i = 1, \dots, 7$  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,  $s_t \in \{1, 2, \dots, 7\}$

- 

$$\varepsilon_t^* | s_t = i \sim N(m_i, v_i^2)$$

$$\Pr(s_t = i) = q_i$$

- MCMC algorithm does not draw from  $p(h^T | y^T, \mu, \phi, \sigma_\eta^2)$ , but from  $p(h^T | y^T, \mu, \phi, \sigma_\eta^2, s^T)$ .
- 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  $p(s^T | y^T, \mu, \phi, \sigma_\eta^2, h^T)$  but this has simple form (see Kim, Shephard and Chib, 1998)

# Summary and Other Directions

- This completes discussion of general ideas underlying state space models and few key models
- Computer tutorial 4 considers time-varying parameter AR model
- Suitable for modelling parameter change (structural breaks/regime change, etc.)
- Computer tutorial 5 considers the popular unobserved components stochastic volatility model
- State space methods growing in popularity in many other contexts
- SSVS and Lasso methods used with state space models
- Frühwirth-Schnatter and Wagner (2010). “Stochastic model specification search for Gaussian and partial non-Gaussian state space models,” *Journal of Econometrics*.
- Dynamic mixture models used to model structural breaks, outliers, nonlinearities, etc.
- Giordani, Kohn and van Dijk (2007, *JoE*).

# A Macroeconomic Application: Inflation Forecasting using Dynamic Model Averaging

- I will end this course with application which involves time series regression, state space models, model averaging and forecasting as way of summarizing major themes of this course
- Based on the paper: Koop and Korobilis (2012, International Economic Review)
- 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:  $\hat{\theta}$  = 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?

# Time Varying Parameter (TVP) Models

- TVP models gaining popularity in empirical macroeconomics

$$y_t = z_t \theta_t + \varepsilon_t$$
$$\theta_t = \theta_{t-1} + \eta_t$$

- $\varepsilon_t \stackrel{ind}{\sim} N(0, H_t)$
- $\eta_t \stackrel{ind}{\sim} N(0, Q_t)$
- 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



# Dynamic Model Averaging (DMA)

- Define  $K$  models which have  $z_t^{(k)}$  for  $k = 1, \dots, K$ , 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  $N(0, H_t^{(k)})$
- $\eta_t^{(k)}$  is  $N(0, Q_t^{(k)})$
- Let  $L_t \in \{1, 2, \dots, K\}$  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  $i, j = 1, \dots, K$ .
- "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

- 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)

# Back to Model Averaging/Selection

- Goal for forecasting at time  $t$  given data available at time  $t - 1$  is  $\pi_{t|t-1,k} \equiv \Pr(L_t = k | y^{t-1})$
- Can average across  $k = 1, \dots, K$  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

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

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



- Interpretation of forgetting factor  $\alpha$
- Easy to show:

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

- $p_k(y_t|y^{t-1})$  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

# Forecasting US Inflation

- Data from 1960Q1 through 2008Q4
- Real time data (forecasting at time  $\tau$  using data as known at time  $\tau$ )
- 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

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

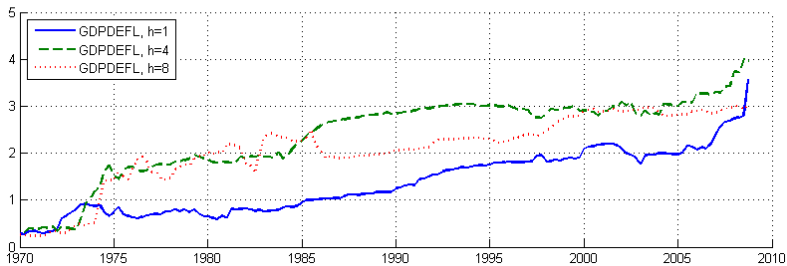
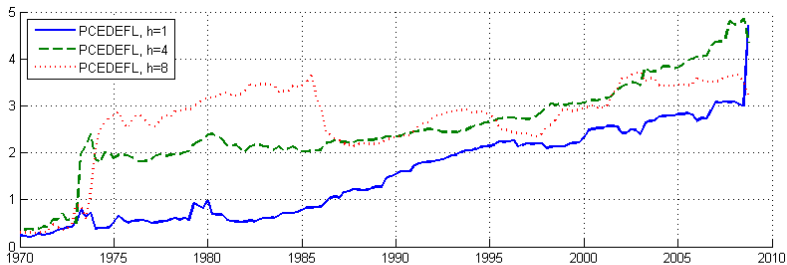


Figure 1: Expected Number of Predictors

# 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

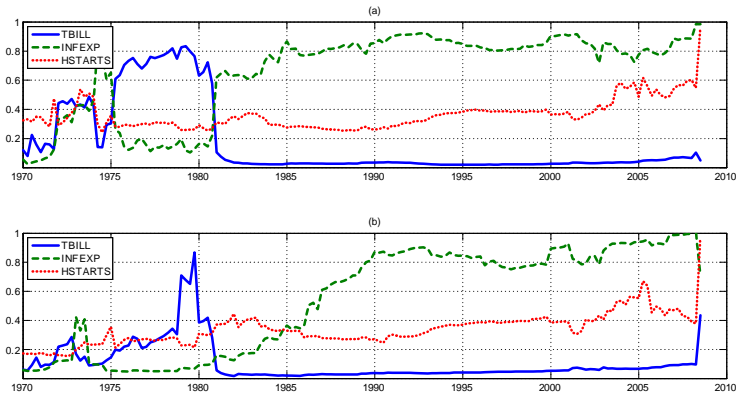


Figure 2: Posterior Probability of Inclusion of Predictors,  $h = 1$ . GDP deflator inflation top, PCE deflator inflation bottom

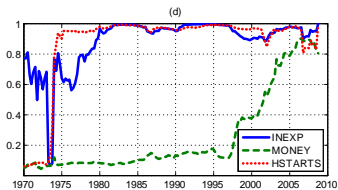
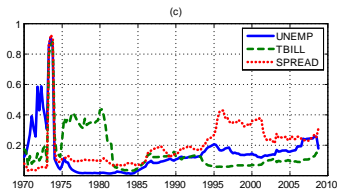
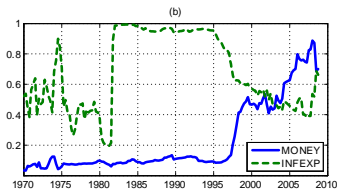
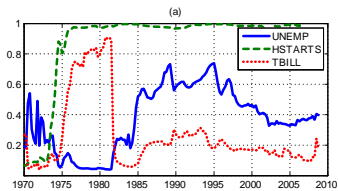


Figure 3: Posterior Probability of Inclusion of Predictors,  $h = 4$ . GDP deflator inflation top, PCE deflator inflation bottom



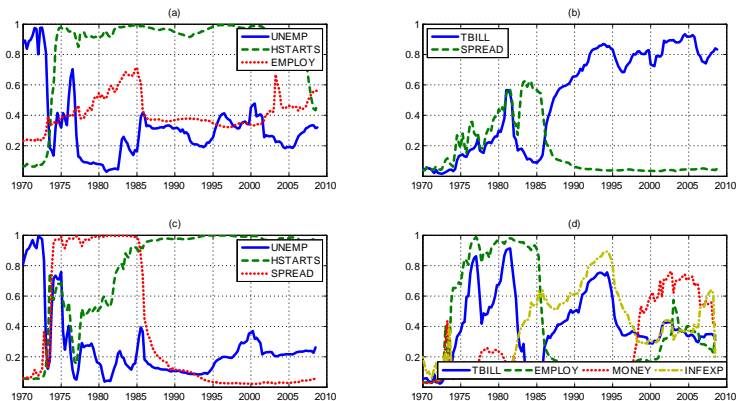


Figure 4: Posterior Probability of Inclusion of Predictors,  $h = 8$ . GDP deflator inflation top, PCE deflator inflation bottom

# Forecast Performance

- 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.

- 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
- 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	-

# 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.