Introduction to Bayesian Econometrics Course

Norges Bank

May, 2007

Overheads for Lecture on

Forecasting in Dynamic Factor Models

Using Bayesian Model Averaging

Gary Koop, University of Strathclyde

1 Introduction

- Based on a paper "Forecasting in Dynamic Factor Models using Bayesian Model Averaging" (coauthored with Simon Potter, *Econometric Reviews*, 2004).

- Uses methods for Normal linear regression model

- Shows how Bayesian model averaging done using forecasting

- Introduces a new posterior simulation algorithm: Markov Chain Monte Carlo Model Composition (MC^3)

- Note on terminology: Gibbs sampler is the most popular of a class of algorithms called Markov Chain Monte Carlo (MCMC) algorithms

2 Motivation for Application

- Many recent papers based on a model where information in a large number of variables is used to explain a single (or a few) dependent variables.

- Information in numerous explanatory variables extracted using factor analysis.

- Forecasting: e.g. Stock and Watson (2002) using diffusion indexes, etc.

- Structural modelling (e.g. identification of monetary shocks): e.g. the FAVAR of Bernanke, Boivin and Eliasz (2002), Reichlin and coauthors, etc.

- Theoretical econometric work: Bai and Ng (2002), Reichlin and coauthors, West (2002), etc.

- Generic problem: lots and lots of potential explanatory variables, you know some are probably important but do not know which ones. Bayesian model averaging is ideally suited for such a situation.

- Our paper: considers various ways of implementing Bayesian model averaging in forecasting problems when there are hundreds of potential explanatory variables.

3 Dynamic Factor Models

$$y_{t+h} = \alpha(L)\, y_t + \gamma(L)\, w_t + \varepsilon_t$$

- $\alpha(L)$ and $\gamma(L)$ are polynomials in the lag operator
- w_t is a k_w-vector where k_w is huge e.g. $k_w = 215$ in Stock and Watson (2002, JBES)
- Standard sequential testing/model selection criteria can lead you astray.
- Hence, dynamic factor model:

$$y_{t+h} = \alpha(L)\, y_t + \beta(L)\, f_t + \varepsilon_t$$

where f_t is q-vector of factors extracted from w_t (e.g. using principal components)

- $q << k_w$ is usually "small" (but should it be?)

- Model selection/pre-test problems can still be substantive when using models with non-sequential factors.

- Note that our variant of the dynamic factor model is a regression model (with lags of dependent variable and factors as explanatory variables)

4 Bayesian Model Averaging

- Researcher often has many possible models and the common strategy (for virtually all non-Bayesians and many Bayesians) is to select one model.

- The problems associated with presentation of results from a single model selected on the basis of a sequence of hypothesis tests have long been recognized in the statistical literature (the so-called pretest problem).

- Intuitive idea: each time a hypothesis test is carried out, the possibility exists that a mistake will be made (i.e. the researcher will reject the better model for a not so good one). This possibility multiplies sequentially with each hypothesis test.

- Even if procedure does lead to the selection of the "best" model, standard decision theory implies that

it is rarely desirable to present results for this model while ignoring all evidence from the not quite so good model(s).

- Response to these problems: Bayesian model averaging (BMA)

- Suppose the researcher is entertaining R possible models, denoted by $M_1, ..., M_R$, to forecast y^*.

- Models and parameters are random variables and rules of probability imply:

$$E\left(y^*|Data\right) = \sum_{r=1}^{R} p\left(M_r|Data\right) E\left(y^*|Data, M_r\right) \qquad (*)$$

- Overall point forecast $E\left(y^{*}|Data\right)$, is weighted average of point estimates in every model $E\left(y^{*}|Data, M_r\right)$. Weights in weighted average are the posterior model probabilities, $p\left(M_r|Data\right)$.

- Can use same idea with entire predictive distribution:

$$p\left(y^{*}|Data\right) = \sum_{r=1}^{R} p\left(M_r|Data\right) p\left(y^{*}|Data, M_r\right)$$

5 Bayesian Model Averaging in the Normal Linear Regression Model

- Researcher often faced with the situation where numerous potential explanatory variables exists. Many of these explanatory variables are probably irrelevant, but you do not know which ones.

- Selecting a single model could be misleading: *model uncertainty* is ignored and sequential testing procedures would involve many tests.

- BMA is a sensible alternative.

- Consider a set of possible linear regression models: All potential explanatory variables are stacked in a $T \times K$ matrix X and set of models given by:

$$y = X_r \beta_r + \varepsilon$$

X_r is a $N \times k_r$ matrix containing some (or all) columns of X.

- In our case X_r are various factors and lagged dependent variables

- Note: minor extension to allow some explanatory variables to be common to all models.

- Since there are 2^K possible subsets of X, there are 2^K possible choices for X_r and, thus, $R = 2^K$.

- If K is at all large, then the number of possible models is astronomical.

- E.g. 30 potential explanatory variables and 2^{30} models. If the computer could analyze each model in 0.001 of a second, it would take almost two years to analyze all the models!

- In such cases, directly doing Bayesian model averaging by explicitly calculating every term in (*) is impossible.

- MC^3 algorithms have been developed to surmount this problem.

5.1 Results for a Single Model

- Remember: *When comparing models using posterior odds ratios, it is acceptable to use noninformative priors over parameters which are common to all models. However, informative, proper priors should be used over all other parameters.*

- Can use noninformative prior for h:

$$p(h) \propto \frac{1}{h},$$

and for the intercept:

$$p(\alpha) \propto 1.$$

- But we need informative prior for β. One commonly-used *benchmark prior* called the g-prior.

- This is a natural conjugate Normal-Gamma prior with:

$$\beta_r|h \sim N\left(0_{k_r}, h^{-1}\left[g_r X_r' X_r\right]^{-1}\right).$$

- See textbook for motivation for the g-prior. It depends only on a scalar prior hyperparameter g_r.

- $g_r = 0$ corresponds to a perfectly noninformative prior. The value $g_r = 1$ implies that prior and data information are weighted equally in the posterior covariance matrix. One strategy a researcher could follow is to try a range of values for g_r between 0 and 1.

- Another common strategy is to choose g_r based on some measure such as an information criterion.

- We use various benchmark values suggested in the literature (see, e.g., Fernandez, Ley and Steel, 2001)

- $g_r = \frac{1}{T}$ — asymptotic relationship with Schwarz criteria

- $g_r = \frac{1}{\ln(T)^3}$ — asymptotic relationship with Hannan-Quinn

- $g_r = \frac{1}{K^2}$ — risk inflation criterion of George and Foster

- Empirical Bayes: Choose value for g_r which maximizes marginal likelihood

- Lecture 2 showed analytical posterior results exist with this prior. Hence, $E\left(y^*|y, M_r\right), p\left(y^*|y, M_r\right)$ and $p\left(M_r|y\right)$ can be evaluated easily.

6 Bayesian Computation: MC-cubed

- MCMC algorithms take draws from the parameter space, MC-cubed algorithm draw from model space (since Bayesians treat parameters and models as random variables same ideas hold)

- A chain of models is drawn $M^{(s)}$ for $s = 1, .., S$.

- BMA involving $g(y^*)$ (i.e. any function of your forecast) can be approximated by $\widehat{g}_{S_1}$ where

$$\widehat{g}_{S_1} = \frac{1}{S_1} \sum_{s=S_0+1}^{S} E\left[g(y^*)|y, M^{(s)}\right].$$

- As with Gibbs sampler, $\widehat{g}_{S_1}$ converges to $E[g(y^*)|y]$ as S_1 goes to infinity (where $S_1 = S - S_0$).

Details of MC^3

- Candidate model, M^*, is drawn (with equal probability) from the set of models containing:

1. $M^{(s-1)}$
2. All models which add one extra explanatory variable to $M^{(s-1)}$
3. All models which delete one variable from $M^{(s-1)}$

- $M^{(s)}$ is set equal to M^* with probability A, else $M^{(s)} = M^{(s-1)}$
- If A is chosen correctly, models drawn in this way will do BMA correctly
- A is acceptance probability (with simple formula, see page 273)

6.1 Practical Issues with MC-cubed

- Problem 1: MC^3 can provide a highly correlated sample from model space — can be very slow and inefficient.

- Remember (from Lecture 1) that

$$p(M_r|y) \propto p\left(y|M_r\right) p\left(M_r\right)$$

- $p\left(M_r\right)$ is prior model probability which must be chosen

- Why not try "noninformative prior" which treats all models equally?

$$p\left(M_r\right) = \frac{1}{R}$$

- Problem 2: Work of Ed George (see his website at University of Pennsylvania) show that with many correlated regressors, "noninformative" priors over model space can potentially put most of the prior probability in small regions of model space.

- Solution to Problems 1 and 2: Orthogonalize regressors

- But that is exactly what dynamic factor models do.

- Original model with all K potential explanatory variables (ignoring variables common to all models):

$$y = X\beta + \varepsilon$$

can be written as:

$$y = Z\alpha + \varepsilon$$

where

$$Z = XW$$

$$\alpha = W^{-1}\beta$$

and columns of Z are orthogonal

- Here we choose W to be the matrix of eigenvectors of $X'X$ and, thus, Z is the usual matrix of factors

7 The Data

- The same as that used in Stock and Watson (2002, NBER working paper)

- 162 U.S. quarterly time series from 1959Q1 through 2001Q1

- Stock and Watson transformations to stationarity used for all variables (to avoid worrying about unit root issues).

- Focus on forecasting GDP and CPI

- Given transformations, GDP growth and growth in inflation.

8 Empirical Issues

We compare Bayesian model averaging to:

1. Bayesian model selection: Using output from BMA algorithm, select the single model with highest $p\left(M_r|y\right)$

2. Conventional model selection: Model with first q factors where q maximizes marginal likelihood

3. AR(p)

- Lag lengths: We choose AR(p) with best forecasting performance (in root mean squared error sense).

- p = 2 for both variables (results using p = 4 qualitatively similar)

- Factor models all include AR(p) component plus factors calculated using p lags of explanatory variables.

- Note: we put the AR(p) base case in the most advantageous position

- Two priors over model space (more in paper):

  1. the noninformative prior

  2. 99.9% prior which in noninformative over first q factors (where 99.9% of the variation in X is included in the first q factors)

- Four priors over parameter space (three benchmark values for g_r plus empirical Bayes)

9 Empirical Results Using the Entire Sample

- Summary:

- Factor models perform very well relative to AR(2)

- BMA and Bayesian model selection allow for substantial improvements in marginal likelihoods (information criteria) over conventional procedures

- Lots of factors have explanatory power (e.g. OLS regression of GDP growth on first 100 factors yields 46 coefficients with t-stats >1 and 15 with t-stats>2)

- Non-sequential factors often selected (e.g. with inflation and 99.9% prior, model with highest marginal likelihood contains factors ranked 1, 2, 3, 5, 6 and 11)

10 Forecasting Setup

- Forecast horizons: h=1, 4 and 8

- Forecasting from 1970Q1 through 2001Q1-h

- Dynamic factor forecasting models at time τ all based on

$$y_{\tau+h} = \gamma_0 + \gamma_1 y_\tau + \gamma_1 y_{\tau-1} + Z_\tau \alpha + \varepsilon_\tau$$

where Z_τ contains factors constructed using data through τ (based on 2 lags).

- Models differ in which factors are included in Z_τ.

- Forecasts evaluated using root mean squared error:

$$RMSE = \sqrt{\sum \left[y_{\tau+h} - E\left(y_{\tau+h} | \Omega_\tau \right) \right]}$$

where Ω_τ denotes data through time τ

- RMSE's presented as % of AR(2) RMSE

11 Forecasting Results

- See tables 3, 4 and 5

- Summary: Forecasting results less favorable than in-sample results for BMA (or any factor model).

- At h=4 or 8 no clear improvements over an AR(2) for either series. Following discussion relates to h=1.

- For h=1, factor models do beat AR(2) and BMA with the 99.9% prior does beat conventional methods.

- BMA with noninformative prior over model space forecasts poorly (too many factors included).

- Discuss prior sensitivity and the related issue of shrinkage

Table 3a: RMSE relative to AR(2), percentage, GDPQ				
	$g = \frac{1}{T}$	$g = \frac{1}{[\ln(T)]^3}$	$g = \frac{1}{K^2}$	Optimal g
Bayesian Model Averaging (equal prior weights to all models)				
$h = 1$	171.1	172.7	102.6	173.7
$h = 4$	188.2	190.9	99.4	186.1
$h = 8$	246.3	248.3	131.0	248.5
Bayesian Model Selection (equal prior weights to all models)				
$h = 1$	181.7	184.0	103.7	176.9
$h = 4$	194.1	194.1	109.5	188.6
$h = 8$	254.2	252.5	123.4	249.2
Bayesian Model Averaging (99.9% prior)				
$h = 1$	94.1	94.1	94.2	93.0
$h = 4$	100.1	100.2	100.1	99.6
$h = 8$	99.0	99.1	99.1	99.2
Bayesian Model Selection (99.9% prior)				
$h = 1$	96.5	96.1	95.6	93.1
$h = 4$	101.5	101.8	101.4	99.7
$h = 8$	100.5	100.7	100.5	101.8
Model with First q Factors Selected				
$h = 1$	94.9	94.8	94.3	94.6
$h = 4$	99.4	99.4	97.9	100.7
$h = 8$	100.4	100.4	100.5	100.5

Table 3b: RMSE relative to AR(2), PUNEW				
	$g = \frac{1}{T}$	$g = \frac{1}{[\ln(T)]^3}$	$g = \frac{1}{K^2}$	Optimal g
Bayesian Model Averaging (equal prior weights to all models)				
$h = 1$	120.6	121.9	92.4	121.4
$h = 4$	141.7	142.6	104.0	143.5
$h = 8$	150.9	154.8	102.9	158.5
Bayesian Model Selection (equal prior weights to all models)				
$h = 1$	131.5	131.0	95.2	130.6
$h = 4$	144.6	142.7	106.3	145.1
$h = 8$	159.8	162.1	106.2	163.3
Bayesian Model Averaging (99.9% prior)				
$h = 1$	91.2	91.3	91.4	88.2
$h = 4$	100.8	100.8	100.8	101.1
$h = 8$	100.9	100.9	100.9	100.7
Bayesian Model Selection (99.9% prior)				
$h = 1$	93.5	94.6	94.7	90.0
$h = 4$	102.4	100.9	100.8	103.4
$h = 8$	100.6	100.6	100.6	101.4
Model with First q Factors Selected				
$h = 1$	92.7	93.4	94.1	89.2
$h = 4$	99.6	99.6	101.3	101.3
$h = 8$	100.0	100.0	100.0	100.0

Table 4a: Percentage of Predictive Means within 2 Standard Deviations of Actual Value, GDPQ				
	$g = \frac{1}{T}$	$g = \frac{1}{[\ln(T)]^3}$	$g = \frac{1}{K^2}$	Optimal g
Bayesian Model Averaging (equal prior weights to all models)				
$h = 1$	58.3	57.5	96.1	52.0
$h = 4$	33.1	28.2	95.2	37.1
$h = 8$	37.5	32.5	91.7	33.3
Bayesian Model Selection (equal prior weights to all models)				
$h = 1$	39.4	33.1	93.7	40.9
$h = 4$	21.8	20.2	87.1	29.0
$h = 8$	21.7	19.2	85.8	23.3
Bayesian Model Averaging (99.9% prior)				
$h = 1$	96.1	96.1	96.1	96.1
$h = 4$	91.9	91.9	91.9	91.9
$h = 8$	95.0	95.0	95.0	95.0
Bayesian Model Selection (99.9% prior)				
$h = 1$	95.3	95.3	95.3	95.3
$h = 4$	91.1	91.1	91.9	91.9
$h = 8$	95.0	95.0	95.0	94.2
Model with First q Factors Selected				
$h = 1$	96.1	96.1	95.3	95.2
$h = 4$	91.1	91.1	91.9	91.1
$h = 8$	93.3	93.3	94.2	91.7

Table 4b: Percentage of Predictive Means within 2 Standard Deviations of Actual Value, PUNEW				
	$g = \frac{1}{T}$	$g = \frac{1}{[\ln(T)]^3}$	$g = \frac{1}{K^2}$	Optimal g
Bayesian Model Averaging (equal prior weights to all models)				
$h = 1$	63.0	58.3	92.1	58.3
$h = 4$	44.4	40.3	90.3	41.9
$h = 8$	31.7	32.5	89.2	46.7
Bayesian Model Selection (equal prior weights to all models)				
$h = 1$	26.8	22.8	86.6	36.2
$h = 4$	33.9	33.1	85.5	34.7
$h = 8$	25.8	23.3	83.3	35.0
Bayesian Model Averaging (99.9% prior)				
$h = 1$	88.2	88.2	87.4	89.2
$h = 4$	87.9	87.9	87.9	87.1
$h = 8$	89.2	87.5	87.5	87.5
Bayesian Model Selection (99.9% prior)				
$h = 1$	86.6	86.6	86.6	87.4
$h = 4$	87.9	87.9	87.9	86.3
$h = 8$	87.5	87.5	87.5	87.5
Model with First q Factors Selected				
$h = 1$	87.2	87.2	85.8	88.8
$h = 4$	87.1	87.1	87.9	87.1
$h = 8$	86.7	86.7	86.7	86.7

Table 5a: Number of Factors in Model (average over all τ), GDPQ				
	$g = \frac{1}{T}$	$g = \frac{1}{[\ln(T)]^3}$	$g = \frac{1}{K^2}$	Optimal g
$E\left(\sum_{j=1}^{K} \gamma_j \vert Data\right)$ for Bayesian Model Averaging (equal prior weights to all models)				
$h = 1$	47.8	50.6	3.9	63.7
$h = 4$	67.2	70.5	3.9	64.0
$h = 8$	63.5	67.0	3.1	66.0
$\sum_{j=1}^{K} \gamma_j$ for Selected Model for Bayesian Model Selection (equal prior weights to all models)				
$h = 1$	55.0	60.8	1.6	66.9
$h = 4$	70.9	72.3	2.0	62.9
$h = 8$	65.2	67.6	1.1	66.4
$E\left(\sum_{j=1}^{K} \gamma_j \vert Data\right)$ for Bayesian Model Averaging (99.9% prior)				
$h = 1$	3.4	3.4	3.4	5.2
$h = 4$	3.4	3.4	3.4	5.2
$h = 8$	1.9	1.9	1.9	4.0
$\sum_{j=1}^{K} \gamma_j$ for Selected Model for Bayesian Model Selection (99.9% prior)				
$h = 1$	2.3	2.3	2.3	4.7
$h = 4$	3.1	3.0	3.0	4.2
$h = 8$	0.4	0.4	0.4	2.6

Table 5b: Number of Factors in Model (average over all τ), PUNEW				
	$g = \frac{1}{T}$	$g = \frac{1}{[\ln(T)]^3}$	$g = \frac{1}{K^2}$	Optimal g
$E\left(\sum_{j=1}^{K} \gamma_j \vert Data\right)$ for Bayesian Model Averaging (equal prior weights to all models)				
$h = 1$	48.7	52.8	4.5	59.3
$h = 4$	62.5	67.8	3.0	65.9
$h = 8$	65.9	69.1	3.0	63.9
$\sum_{j=1}^{K} \gamma_j$ for Selected Model for Bayesian Model Selection (equal prior weights to all models)				
$h = 1$	68.1	72.6	2.3	62.6
$h = 4$	64.8	68.9	0.4	67.5
$h = 8$	66.3	70.5	0.4	64.9
$E\left(\sum_{j=1}^{K} \gamma_j \vert Data\right)$ for Bayesian Model Averaging (99.9% prior)				
$h = 1$	4.4	4.4	4.4	5.8
$h = 4$	1.7	1.7	1.6	3.5
$h = 8$	1.4	1.4	1.4	1.8
$\sum_{j=1}^{K} \gamma_j$ for Selected Model for Bayesian Model Selection (99.9% prior)				
$h = 1$	4.0	4.0	4.0	5.2
$h = 4$	0.1	0.04	0.1	1.7
$h = 8$	0.4	0.4	0.4	1.0