Wageningen Summer School in Econometrics

The Bayesian Approach in Theory and Practice

September 2008

Slides for Lecture on

Bayesian Model Averaging in the Linear Regression Model

Gary Koop, University of Strathclyde

1 Summary

- Readings: Chapter 9 of textbook.

- I will discuss Bayesian model averaging (BMA) in the linear regression model.

- I will introduce a new tool of posterior computation called *Markov Chain Monte Carlo Model Composition* or MC^3.

2 Overview

- 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 has 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 learn about a parameter of interest, ϕ.

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

$$E(\phi|Data) = \sum_{r=1}^{R} p(M_r|Data) E(\phi|Data, M_r)$$

- Overall point estimate $E(\phi|Data)$, is weighted average of point estimates in every model $E(\phi|Data, M_r)$. Weights in weighted average are the posterior model probabilities, $p(M_r|Data)$.

- If we define regression models by whether each of our K explanatory variables is included/excluded then we have 2^K models! This can be enormous. This motivates use of MC^3 algorithms.

- Formally: let ϕ be vector of parameters which has a common interpretation in all models.

$$p(\phi|y) = \sum_{r=1}^{R} p(\phi|y, M_r)\, p(M_r|y)\,.$$

Or, if $g(\phi)$ is a function of ϕ:

$$E[g(\phi)|y] = \sum_{r=1}^{R} E[g(\phi)|y, M_r]\, p(M_r|y)\,. \qquad (*)$$

3 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 $N \times K$ matrix X and set of models given by:

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

where ι_N is a $N \times 1$ vector of ones , X_r is a $N \times k_r$ matrix containing some (or all) columns of X.

- Note intercept is usually assumed 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!

- Directly doing Bayesian model averaging by explicitly calculating every term in (*) is usually impossible.

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

3.1 Prior Information

- 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.

- There exist results of the form "if g_r takes a certain form, then the log of the resulting Bayes factor approximates a certain information criteria".

- Chapter 2 and 3 showed analytical posterior results exist with Normal-Gamma prior. Hence, $E(\beta_r|y, M_r)$ and $p(M_r|y)$ can be evaluated easily.

3.2 Bayesian Computation: MC-cubed

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

- The most common MC^3 algorithm for sampling from model space is like a Metropolis-Hasting algorithm.

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

- Candidate model is drawn and then accepted with a certain probability.

- If a candidate model is not accepted, then the chain remains at the current model (i.e. $M^{(s)} = M^{(s-1)}$).

- In other words, it is exactly like a traditional Metropolis-Hastings algorithm except that models are being drawn instead of parameters.

- For details about how candidates are drawn and what acceptance probability is see textbook, page 273.

- Posterior results based on the sequence of models generated from the MC^3 algorithm can be calculated by averaging over draws in the standard MCMC manner

- Equation (*) can be approximated by $\widehat{g}_{S_1}$ where

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

- As with our previous MCMC algorithms, $\widehat{g}_{S_1}$ converges to $E\left[g(\phi)|y\right]$ as S_1 goes to infinity (where $S_1 = S - S_0$).

3.3 Empirical Illustration: Model Uncertainty in Growth Regressions

- Which variables influence economic growth?

- GDP growth is dependent variable and there are many potential explanatory variables.

- In this empirical illustration, with $N = 72$ countries and contains $K = 41$ potential explanatory variables.

- Col labelled "BMA Post. Prob." can be interpreted as the probability that the corresponding explanatory variable should be included.

Table 11.1: Bayesian Model Averaging Results			
Explanatory Variable	BMA Post. Prob.	Posterior Mean	Posterior St. Dev
Prim Sch Enrol	0.207	0.004	0.010
Life expect	0.935	0.001	3.4×10^{-4}
GDP in 1960	0.998	-0.016	0.003
Fr GDP Mining	0.460	0.019	0.023
Deg. of Cap.	0.452	0.001	0.001
# Yrs Open Ec.	0.515	0.007	0.008
% Pop. Engl	0.068	-4.3×10^{-4}	0.002
% Pop. For.	0.067	2.9×10^{-4}	0.001
Exch Rate Dist	0.081	-4.0×10^{-6}	1.7×10^{-5}
Equip Inve	0.927	0.161	0.068
Non-equip Inv	0.427	0.024	0.032
Sd. of Bl Mk Pr	0.049	-6.3×10^{-7}	3.9×10^{-6}
Out Orient	0.039	-7.1×10^{-5}	5.9×10^{-4}
Bl Mkt Prem	0.181	-0.001	0.003
Area	0.031	-5.0×10^{-9}	1.1×10^{-7}
Lat Am	0.207	-0.002	0.004
Sub-Sah Afr	0.736	-0.011	0.008
Higher Ed Enr	0.043	-0.001	0.010
Public Ed Sh	0.032	0.001	0.025
Rev's+Coups	0.030	-3.7×10^{-6}	0.001
War	0.076	-2.8×10^{-4}	0.001

Table 11.1 (continued): Bayesian Model Averaging Results			
Explanatory Variable	BMA Post. Prob.	Posterior Mean	Posterior St. Dev
Political Rights	0.094	-1.5×10^{-4}	0.001
Civil Liberties	0.127	-2.9×10^{-4}	0.001
Latitude	0.041	9.1×10^{-7}	3.1×10^{-5}
Age	0.083	-3.9×10^{-6}	1.6×10^{-5}
British Colony	0.037	-6.6×10^{-5}	0.001
Frac Buddhist	0.201	0.003	0.006
Fract Catholic	0.126	-2.9×10^{-4}	0.003
Frac Confucian	0.989	0.056	0.014
Ethno. Frac.	0.056	3.2×10^{-4}	0.002
French Colony	0.050	2.0×10^{-4}	0.001
Fraction Hindu	0.120	-0.003	0.011
Fraction Jewish	0.035	-2.3×10^{-4}	0.003
Fraction Muslim	0.651	0.009	0.008
Primary Exports	0.098	-9.6×10^{-4}	0.004
Frac Protestant	0.451	-0.006	0.007
Rule of Law	0.489	0.007	0.008
Spanish Colony	0.057	2.2×10^{-4}	1.5×10^{-3}
Pop Growth	0.036	0.005	0.046
Workers to Pop	0.046	-3.0×10^{-4}	0.002
Labor	0.072	6.7×10^{-9}	3.7×10^{-8}

Table 11.2: Posterior Model Probabilities for Top 10 Models		
	$p(M_r\|y)$ Analytical	$p(M_r\|y)$ MC^3 estimate
1	0.0089	0.0088
2	0.0078	0.0080
3	0.0052	0.0052
4	0.0035	0.0035
5	0.0032	0.0035
6	0.0029	0.0029
7	0.0028	0.0028
8	0.0028	0.0025
9	0.0028	0.0025
10	0.0024	0.0023

- Compare Table 11.1 to Table 11.3 which presents posterior results for the single most preferred model.

- Posterior model probability of preferred model is only 0.0089.

- A comparison of Tables 11.1 and 11.3 indicates that, by choosing a single model, the researcher can be seriously misled.

- E.g. consider the marginal effect of each explanatory variable on growth. For some explanatory variables, the posterior means are quite different. For instance, the marginal effect of the Non-equipment investment variable is estimated at 0.024 in Table 11.1, but is almost three times as big in Table 11.3.

- Posterior standard deviations in Table 11.3 are all much smaller than those in Table 11.1. Intuitively, selecting a single model ignores uncertainty about whether that model is the correct one. By ignoring model uncertainty, Table 11.3 provides a misleadingly over-confident view of the accuracy of results.

Table 11.3: Posterior Results for the Preferred Model			
Explanatory Variable	BMA Post. Prob.	Posterior Mean	Posterior St. Dev
Life expect	0.935	0.001	1.9×10^{-4}
GDP in 1960	0.998	-0.017	0.002
Deg of Cap	0.452	0.003	7.9×10^{-4}
Eq Inv	0.927	0.159	0.039
Non-eq Inv	0.427	0.064	0.019
Sub-Sah Afr	0.736	-0.013	0.003
Fr Confucian	0.989	0.058	0.011
Fr Muslim	0.651	0.010	0.003
Fr Protestant	0.451	-0.011	0.004
Rule of Law	0.489	0.017	0.004