Journal of Econometrics 167 (2012) 358–369

Contents lists available at SciVerse ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Bayesian averaging, prediction and nonnested model selection

Han Hong [a,*], Bruce Preston [b,c]

[a] Department of Economics, Stanford University, 579 Serra Mall, Stanford, CA 94305, USA

[b] Department of Economics, Columbia University, 420 West 118th Street, New York, NY 10027, USA

[c] Research School of Economics, Australian National University, Canberra, Australia

ARTICLE INFO

Article history:
Available online 1 October 2011

JEL classification:
C14
C52

Keywords:
Model selection criteria
Nonnested
Posterior odds
BIC

ABSTRACT

This paper studies the asymptotic relationship between Bayesian model averaging and post-selection frequentist predictors in both nested and nonnested models. We derive conditions under which their difference is of a smaller order of magnitude than the inverse of the square root of the sample size in large samples. This result depends crucially on the relation between posterior odds and frequentist model selection criteria. Weak conditions are given under which consistent model selection is feasible, regardless of whether models are nested or nonnested and regardless of whether models are correctly specified or not, in the sense that they select the best model with the least number of parameters with probability converging to 1. Under these conditions, Bayesian posterior odds and BICs are consistent for selecting among nested models, but are not consistent for selecting among nonnested models and possibly overlapping models. These findings have important bearing for applied researchers who are frequent users of model selection tools for empirical investigation of model predictions.

1. Introduction

Bayesian methods are becoming increasingly popular, both as a framework of model selection and also as a tool of forecasting—see, among others, Fernandez-Villaverde and Rubio-Ramirez (2004), Schorfheide (2000), Stock and Watson (2003), Timmermann (2006), Clark and McCracken (2009) and Wright (2008). They are also often used to summarize statistical properties of data, identify parameters of interest, and conduct policy evaluation. While empirical applications of these methods are abundant, less is understood about their theoretical sampling properties. This paper provides a starting point for understanding the relation between Bayesian forecast averaging and frequentist model selection and prediction in a general framework that incorporates both nested and nonnested models.

We study the large sample properties of Bayesian prediction and model averaging for both nested and nonnested models. We first show that, for a single model, the difference between Bayesian and frequentist predictors are of a smaller order of magnitude than the inverse of the square root of the sample size in large samples, regardless of the expected loss function used in forming the Bayesian predictors. This contrasts with the difference between MLE and Bayesian estimators, formed using a variety of loss functions, which is of the order $O_p(1/\sqrt{T})$.

For multiple models, we derive general conditions under which the Bayesian posterior odds place asymptotically unit weight on the best model with the most parsimonious parameterization. Under these conditions, the Bayesian average model forecast is equivalent to the frequentist post-selection forecast up to a term that is of a smaller order of magnitude than the inverse of the square root of the sample size in large samples. These findings essentially combine Schwarz's original contribution regarding BIC—that it is an asymptotic approximation to posterior odds—with the insights by Sin and White (1996) who demonstrate the inconsistency of BIC for selecting among nonnested models. The conditions we derive are weaker, more general, and allow for a much wider class of models.

An immediate consequence of multiple model comparison is that both the BIC and Bayesian posterior odds comparison are inconsistent in choosing a true and parsimonious model for selecting among nonnested models and some overlapping models. While this procedure will select one of the best fitting models, it does not necessarily choose the most parsimonious model with probability converging to 1 in large samples. Consistent selection among possibly nonnested models is feasible using nonnested model selection criteria in the spirit of Sin and White (1996).

These findings have a bearing for applied researchers who are frequent users of model selection tools for empirical investigation of model predictions. In addition, empirical analyses frequently find that forecasts generated from averages of a number of models typically perform better than forecasts of any one of the underlying models—see, for instance, Stock and Watson (2003).

* Corresponding author.
E-mail address: doubleh@stanford.edu (H. Hong).

0304-4076/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2011.09.021

Our theoretical findings suggest that this can be because the models under consideration are close to each other and are all misspecified. As long as the posterior weights are non-degenerate among the set of models under comparison, it is possible for model averaging to outperform each model as long as all models are misspecified. Indeed, it is shown that, for nonnested models, posterior weights will be non-degenerate, as long as the models are sufficiently close to each other. The case of nested models is more interesting. It turns out that when the two models are sufficiently far from each other or sufficiently close to each other, the posterior weights will be degenerate. However, when the two models are "just close enough" but "not too close", the posterior weights can be non-degenerate, and, as a consequence, model averaging can outperform each individual model.

Our results are of interest given the burgeoning use of Bayesian methods in the estimation of dynamic stochastic general equilibrium models in modern macroeconomics. Fernandez-Villaverde and Rubio-Ramirez (2004), Schorfheide (2000), Smets and Wouters (2003), Lubik and Schorfheide (2007) and Justiniano and Preston (2010) present examples of estimation in both closed-economy and open-economy settings. These papers all appeal to posterior odds ratios as a criterion for model selection. By giving a classical interpretation to the posterior odds ratio, the present paper intends to provide useful information regarding the conditions under which such selection procedures ensure consistency. The analysis contributes to understanding the practical limitations of standard model selection procedures given a finite amount of data.

The paper proceeds as follows. Section 2 describes model assumptions and derives their implications on the large sample behavior of the likelihood function. Section 3 demonstrates the asymptotic equivalence between Bayesian and frequentist predictors for a single model under weak conditions. The rest of the paper generalizes this result to multiple models. Section 4 first derives weak conditions under which the generalized posterior odds ratio is equivalent to BIC up to a term that is asymptotically negligible, and under which alternative model selection criteria are feasible to select consistently between both nested and nonnested models. Section 5 makes use of the asymptotic equivalence between posterior odds ratio and BIC to derive the relation between Bayesian model averaging and frequentist post-selection prediction. Finally, Section 6 generalizes the implications for our results for Bayesian type model selection methods for non-likelihood-based objective functions as considered by Kim (2005), and Section 7 concludes.

2. Model assumptions and implications

For clarity of exposition, we identify a model with the likelihood function that is being used to estimate model parameters. All results extend to general random distance functions that satisfy the stochastic equicontinuity assumptions stated below.

A parameter β is often estimated by maximizing a random log-likelihood function $\hat{Q}(\beta)$ associated with some model $h(y_t, \beta)$ that depends on observed data y_t and parameterized by the vector β:

$$\hat{Q}(\beta) \equiv Q(y_t, t = 1 \ldots, T; \beta).$$

For example, under i.i.d. sampling of the data, as in Vuong (1989) and Sin and White (1996), the log-likelihood function takes the form of

$$\hat{Q}(\beta) = \sum_{t=1}^{T} \log h(y_t; \beta),$$

which minimizes the Kullback–Leibler distance between the parametric model and the data. Objective functions other than the log-likelihood function will also be discussed in the subsequent sections of this paper. See also the examples in Chernozhukov and Hong (2003).

Under standard assumptions, the random objective function converges to a population limit when the sample size increases without bound. It is assumed that there exists a function $Q(\beta)$, uniquely maximized at β_0, which is the uniform limit of the random sample analog

$$\sup_{\beta \in \mathcal{B}} \left| \frac{1}{T} \hat{Q}(\beta) - Q(\beta) \right| \xrightarrow{p} 0.$$

Typically, the following decomposition holds for $\hat{Q}(\hat{\beta})$, where $\hat{\beta} = \arg\sup_{\beta \in \mathcal{B}} \hat{Q}(\beta)$:

$$\hat{Q}(\hat{\beta}) = \underbrace{\hat{Q}(\hat{\beta}) - \hat{Q}(\beta_0)}_{(Qa)} + \underbrace{\hat{Q}(\beta_0) - TQ(\beta_0)}_{(Qb)} + \underbrace{TQ(\beta_0)}_{(Qc)}.$$

Under suitable regularity conditions, the following are true:

$$(Qa) = O_p(1), \qquad (Qb) = O_p\left(\sqrt{T}\right), \qquad (Qc) = O(T).$$

The regularity conditions under which the first equality holds are formally given below. They are the same as those in Chernozhukov and Hong (2003). They do not require the objective function to be smoothly differentiable, and permit complex nonlinear or simulation-based estimation methods. In particular, conditions that require smoothness of the objective function are typically violated in simulation-based estimation methods and in percentile-based non-smooth moment conditions. Even for simulation-based estimation methods, it can be difficult for researchers to insure that the simulated objective functions are smooth in model parameters.

Assumption 1. The true parameter vector β_0 belongs to the interior of a compact convex subset $\mathcal{B}$ of $R^{\dim(\beta)}$.

Assumption 2. For any $\delta > 0$, there exists $\epsilon > 0$, such that

$$\liminf_{T \to \infty} P\left\{ \sup_{|\beta - \beta_0| \geq \delta} \frac{1}{T} (\hat{Q}(\beta) - \hat{Q}(\beta_0)) \leq -\epsilon \right\} = 1.$$

Assumption 3. There exist quantities Δ_T, J_T, Ω_T, where $J_T \xrightarrow{p} -\mathcal{A}_\beta$, $\Omega_T = O(1)$,

$$\frac{1}{\sqrt{T}} \Omega_T^{-1/2} \Delta_T \xrightarrow{d} N(0, I),$$

such that if we write

$$R_T(\beta) = \hat{Q}(\beta) - \hat{Q}(\beta_0) - (\beta - \beta_0)' \Delta_T + \frac{1}{2} (\beta - \beta_0)' (TJ_T)(\beta - \beta_0)$$

then it holds that for any sequence of $\delta_T \to 0$

$$\sup_{|\beta - \beta_0| \leq \delta_T} \frac{R_T(\beta)}{1 + T|\beta - \beta_0|^2} = o_p(1).$$

Theorem 1. *Under Assumptions 1–3,* $\hat{Q}(\hat{\beta}) - \hat{Q}(\beta_0) = O_p(1)$.

Given Pakes and Pollard (1989), Newey and McFadden (1994) and Andrews (1994), the result of Theorem 1 is rather straightforward. Its proof is incorporated in the beginning of the proof for Theorem 2 and is provided in the Appendices A–D.

The asymptotic distribution of $\hat{Q}(\hat{\beta}) - \hat{Q}(\beta_0)$ is also easy to derive in many situations. This distribution is useful for model selection tests but is not directly used in model selection criteria