

# Nonnested Model Selection Criteria

Han Hong and Bruce Preston<sup>1</sup>

This Version: January 2006

---

## Abstract

This paper studies model selection for a general class of models based on minimizing random distance functions, giving particular focus to the relation between consistent model selection criteria, Bayesian posterior odds calculation and model averaging. Weak conditions are given under which the proposed model selection criteria are consistent, 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. As a corollary, in the case of non-nested models, it is shown that while traditional Bayesian methods for model selection based on Bayes factors choose models with the best fit objective functions, they do not necessarily consistently select the most parsimonious model among the best fitting models. In addition, we study the asymptotic relationships between Bayesian and frequentist predictors.

*JEL Classification:* C14; C52

*Keywords:* Model selection criteria, Nonnested, Posterior odds, BIC

---

## 1 Introduction

Models are used to summarize statistical properties of data, identify parameters of interest, and conduct policy evaluation. Of considerable import then is the proper determination of the best available model. A persistent challenge to empirical identification of the best available model is the possibility of model misspecification.

Building on a large literature in econometrics and statistics, this paper studies model selection criteria for models based on minimizing random distance functions under considerably weaker regularity conditions than in the existing literature. Conditions are given that ensure consistency regardless of whether models are nested or non-nested and regardless of whether models are misspecified or not. The proposed selection criteria provide a useful complement to model specification tests — as

---

<sup>1</sup>We thank Raffaella Giacomini, Jin Hahn, Bernard Salanié, Frank Schorfheide and Chris Sims for comments and the NSF (SES 0452143) for support. The usual caveat applies.

proposed by Newey (1985), Fan (1994) and Hansen (1982) among others — for evaluating fit and discriminating between models since hypothesis testing procedures may lead to the simultaneous acceptance or rejection of two hypotheses — see Gourieroux and Monfort (1995).

Model selection criteria (MSC) are defined by the combination of a random objective function and a penalization term. As an example, suppose  $\beta$  is estimated by minimizing a random distance function  $\hat{Q}(\beta)$  associated with some model  $f(y_t, \beta)$  that depends on observed data  $y_t$  and parameterized by the vector  $\beta$ . For example,  $\hat{Q}(\beta)$  can define an M-estimator where  $\hat{Q}(\beta) = \sum_{t=1}^T \hat{q}(y_t, \beta)$ , or  $\hat{Q}(\beta)$  can define a generalized method of moments estimator (Hansen (1982)). The proposed model selection criteria take the form

$$MSC = \hat{Q}(\beta) - \dim(f, Q) * C_T$$

implicitly defining penalty functions of the form  $\dim(f, Q) * C_T$ . The first component of the penalty function,  $\dim(f, Q)$ , rewards both parsimony in the parameterization of model  $f$  and dimensionality of the model estimation procedure, while the second component,  $C_T$ , is a sequence of constants depending on the sample size,  $T$ . Model selection is then determined by choosing the model for which  $MSC$  is greatest in value.

Following a large literature on the statistical evaluation of models – see Andrews (1999), Andrews and Lu (2001) and Gourieroux and Monfort (1995) – consistency of MSCs is defined in two senses. First, it requires that inferior models are chosen with probability approaching 0. The second requirement is that among the best models, the most parsimonious one with the most informative estimation procedure is chosen with probability converging to 1, where  $\dim(f, Q)$  characterizes the degree of parsimony of the model and information of the estimation procedure.

Consistent model selection requires an appropriate choice of penalty function. For nested models, it is well known that  $C_T$  must satisfy  $C_T = o(T)$  and  $C_T \rightarrow \infty$  to ensure consistency. Such a condition is satisfied by Bayesian information criterion, BIC, but not the Akaike information criterion, AIC (both later defined). In the context of parametric likelihood models, Sin and White (1996) demonstrate that non-nested models require the additional restriction that  $\frac{C_T}{\sqrt{T}} \rightarrow \infty$  for consistency in model selection. The first goal of this paper is to derive a related condition for a broader class of nonlikelihood based models under much weaker regularity conditions that allow for nonsmooth objective functions. Given a penalty function satisfying these requirements, the proposed model selection criterion ensures consistency regardless of whether models (both likelihood and nonlikelihood based) are nested or not and regardless of whether models are correctly specified or not.

The second goal of this paper is to relate MSC to Bayesian posterior odds calculation. Because it

adopts a penalty function that is of order  $C_T = O(\log T)$ , an immediate corollary is that the BIC (or Schwartz criterion) represents an inconsistent model selection criterion for non-nested models — violating the second requirement for consistency under the null hypothesis that two models have asymptotically equal fit. Worth noting is that when models are nonnested this null hypothesis necessarily implies both models are misspecified. Since posterior odds ratios and Bayes factors can be shown to be equivalent to BIC up to a negligible term, traditional methods for model selection in Bayesian inference similarly lead to inconsistent inference in the case of nonnested models. The second principle contribution of this paper demonstrates that these conclusions extend to so-called generalized posterior odds ratios for more general objective functions under weak regularity conditions. Generalized posterior odds ratios are constructed from Laplace-type estimators which include parametric likelihood as a special case. Hence, generalized posterior odds ratios similarly inherit the property of inconsistency as a model selection criterion between two nonnested models.

It is important to be clear about the substantive content of this result. What is demonstrated, given a set of assumptions about the model environment (which includes a particular criterion for ranking models), is what one can reasonably expect to learn from the data. And, in particular, whether a given set of models can be distinguished using a finite amount of data. As such, the finding that posterior odds ratios are inconsistent when choosing between nonnested misspecified models underscores that caution is appropriate when this criterion is adopted as a model selection device. Furthermore, the finding of inconsistency in this case does not suggest other Bayesian approaches for model selection would not be appropriate. Indeed, from a decision theoretic perspective posterior odds impose a zero-one loss function which may have questionable merit. Alternative loss functions, such as a quadratic loss, could also be considered as discussed by Schorfheide (2000).

Motivated by the well known classical result that the asymptotic distribution properties of post-selection estimation and prediction are not affected by first stage model selection so long as the model selection criterion is consistent, the connections between Bayesian and classical predictions are explored. They are shown to be equivalent up to a term of order  $o_p(1/\sqrt{T})$ . Furthermore, for predictions based on minimization of a general class of nonlinear loss functions, we demonstrate conditions under which the posterior odds place asymptotically unitary probability weight on a single model and the asymptotic distribution properties of prediction intervals are not affected by model averaging — providing an analogue to the stated classical result. What is surprising about this theoretical finding is that 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 (2001).

While the modification of standard model selection criteria is a straightforward exercise, our analysis

provides an important technical contribution by establishing all results under very weak regularity conditions. Indeed, the approach is valid for general statistical criterion functions that admit discontinuous non-smooth semi-parametric functions. Similarly, data generating processes can range from i.i.d. settings to nonlinear dynamic models.

Furthermore, it is hoped that by providing a classical interpretation of Bayesian methods, recent discussion of model comparison exercises can be clarified. This is of considerable import given the burgeoning use of Bayesian methods in the estimation of dynamic stochastic general equilibrium models in modern macroeconomics. See, *inter alia*, Fernandez-Villaverde and Rubio-Ramirez (2004a), Schorfheide (2000), Smets and Wouters (2002), Lubik and Schorfheide (2003) and Justiniano and Preston (2004) for examples of estimation in both closed 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 generalized 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 generalized nonlikelihood based model selection criterion proposed here builds most directly on recent work by Andrews (1999), Andrews and Lu (2001), Kitamura (2002) and Hong, Preston, and Shum (2003). The former papers extend model selection criteria from parametric likelihood models to unconditional moment models by using GMM J-statistics rather than likelihood functions. Kitamura (2002) develops model selection tests for nonparametric GMM models. The latter paper further replaces the J-statistic with generalized empirical likelihood statistics in the construction of the model selection criterion and tests. It therefore provides an information-theoretic analogy with model selection criteria based on standard parametric likelihood models for moment-based models. The present paper is best viewed as extending this work to allow nonnested model selection criteria to be constructed using models specified as minimizing very general statistical criteria.

The results of the present paper are also related to recent work in a likelihood context by Fernandez-Villaverde and Rubio-Ramirez (2004a). In contrast to their paper, which assumes the existence of a unique asymptotically superior model, the present analysis provides explicit consideration of the possibility that two or more possibly misspecified models are asymptotically equally good (in a sense which will be made clear) under much weaker regularity conditions. Such analysis discloses large sample statistical properties of the model selection criterion that are distinct from the case in which one model is assumed to be asymptotically better than another (with both being nonnested and misspecified). By admitting the possibility of two equally good though potentially non-nested and misspecified models the analysis is more naturally couched in classical estimation theory and

hypothesis testing.

The paper proceeds as follows. Section 2 describes the generalized model selection criterion and develops the conditions required for consistency in model selection. Section 3 establishes the equivalence of the generalized posterior odds ratio and BIC up to a term that is asymptotically negligible. An immediate corollary is the inconsistency of posterior odds ratio when comparing two nonnested models. Section 4 underscores the possibility of designing model selection tests (hypothesis testing). In the case of nonnested models generalized posterior odds ratios permit frequentist probability statements regarding the likelihood of any model under consideration. Section 5 turns to model selection based forecasting performance. It gives conditions under which that Bayesian and Classical prediction are equivalent up to a term that is of order  $o_p(1/\sqrt{T})$ . Furthermore, we discuss properties of prediction procedures resulting from averaging two or more models. Section 6 highlights the broad applicability of our results by considering examples from parametric likelihood, GMM and generalized empirical likelihood based estimation. Section 7 concludes.

## 2 Generalized Nonnested Model Selection Criteria

Consider two competing models  $f(Y; \beta)$ ,  $\beta \in \mathcal{B}$  and  $g(Y; \alpha)$ ,  $\alpha \in \Lambda$  (generalization to the case of multiple models is immediate). Suppose  $\beta$  is estimated by minimizing a random distance function

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

and  $\alpha$  estimated by minimizing a random distance function

$$\hat{L}(\alpha) \equiv L(y_t, t = 1, \dots, T; \alpha).$$

For example, as in Vuong (1989) and Sin and White (1996), such distance functions could be the log likelihood functions for two parametric models  $f(Y; \beta)$  and  $g(Y; \alpha)$

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

and

$$\hat{L}(\alpha) = \sum_{t=1}^T \log g(y_t; \alpha),$$

which minimize the Kullback-Leibler distances between the parametric models and the data under the assumption that the data is generated by an i.i.d. data generating process.

Under standard assumptions, the random objective functions converge to a population limit when the sample size increases without bound. It is therefore assumed that there exist functions  $Q(\beta)$  and  $L(\alpha)$ , uniquely maximized at  $\beta_0$  and  $\alpha_0$ , which are the uniform limit of the random sample analogs:

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

and

$$\sup_{\alpha \in \Lambda} \left| \frac{1}{T} \hat{L}(\alpha) - L(\alpha) \right| \xrightarrow{p} 0.$$

Furthermore, the direct comparison between  $Q(\beta)$  and  $L(\alpha)$  is assumed to be interpretable. For example, when both  $Q(\beta)$  and  $L(\alpha)$  are likelihood functions corresponding to parametric models, the difference between them can be interpreted as the difference in the ability of the two models to minimize the Kullback-Leibler distance between the parametric model and the data. Such an information-theoretic interpretation is also possible, for example, when both are constructed as empirical likelihoods corresponding to a different set of model and moment conditions.

A conventional consistent model selection criterion takes the form of a comparison between

$$\hat{Q}(\hat{\beta}) - \dim(\beta) * C_T \quad \text{and} \quad \hat{L}(\hat{\alpha}) - \dim(\alpha) * C_T,$$

where  $C_T$  is a sequence of constants that tends to  $\infty$  as  $T$  goes to  $\infty$  at a rate to be prescribed below, and  $\dim(\cdot)$  indexes the parametric dimension of the model.  $\hat{\beta}$  and  $\hat{\alpha}$  are maximands of the corresponding objective functions. The second term in the model selection criteria penalizes the dimension of the parametric model. For instance, BIC takes  $C_T = \log T$  and AIC adopts  $C_T = 2$ . In addition to penalizing the parametric dimension of the model, the dimension of the estimation procedure, such as the number of moment conditions used in a generalized empirical likelihood framework — as in Andrews and Lu (2001) — could also be penalized. The goal of the latter penalization term is to preserve the parsimony of the model and the informativeness of the estimation procedure, thereby improving the precision of the model with a finite amount of data. To emphasize this possibility, we will adopt a general notation  $\dim(f, Q)$  and  $\dim(g, L)$ , rather than  $\dim(\beta)$  and  $\dim(\alpha)$ . The first argument refers to the parametric dimension of the model, while the second argument refers to the dimension of the information used in the inference procedure regarding model parameters.

Model selection is therefore based on the comparison of the two quantities

$$\hat{Q}(\hat{\beta}) - \dim(f, Q) * C_T \quad \text{and} \quad \hat{L}(\hat{\alpha}) - \dim(g, L) * C_T.$$

Model  $f$  is selected if the former object is greater in magnitude than the latter, while model  $g$  is selected otherwise. Central then to the consistency of the proposed MSC are the asymptotic properties of  $\hat{Q}(\hat{\beta}) - \hat{L}(\hat{\alpha})$ . The remainder of this section characterizes the asymptotic properties of this difference.

Typically, the following decomposition holds for  $\hat{Q}(\beta)$  and  $\hat{L}(\alpha)$ :

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

and

$$\hat{L}(\hat{\alpha}) = \underbrace{\hat{L}(\hat{\alpha}) - \hat{L}(\alpha_0)}_{(La)} + \underbrace{\hat{L}(\alpha_0) - TL(\alpha_0)}_{(Lb)} + \underbrace{TL(\alpha_0)}_{(Lc)}.$$

Under suitable regularity conditions, the following statements are true:

$$\begin{aligned} (Qa) &= O_p(1), & (Qb) &= O_p(\sqrt{T}), & (Qc) &= O(T), \\ (La) &= O_p(1), & (Lb) &= O_p(\sqrt{T}), & (Lc) &= O(T). \end{aligned}$$

The regularity conditions under which the first part holds (( $Qa$ ) and ( $La$ )) 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 therefore can allow for 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 nonsmooth moment conditions. Even for simulation based estimation methods involving simulating smooth moment conditions, it can be difficult for researchers to insure that the simulated objective functions are smooth in model parameters. Without loss of generality, we state the conditions for model ( $f, Q$ ) only.

**ASSUMPTION 1** *The true parameter vector  $\beta_0$  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*

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

**ASSUMPTION 3** *There exist quantities  $\Delta_T$ ,  $J_T$ ,  $\Omega_T$ , 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*

$$R_T(\beta) = \hat{Q}(\beta) - \hat{Q}(\beta_0) - (\beta - \beta_0)' \Delta_T + \frac{1}{2} (\beta - \beta_0)' (T J_T) (\beta - \beta_0)$$

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

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

**THEOREM 1** *Under assumptions (1), (2) and (3),  $\hat{Q}(\hat{\beta}) - \hat{Q}(\beta_0) = O_p(1)$ .*

The asymptotic distribution of  $\hat{Q}(\hat{\beta}) - \hat{Q}(\beta_0)$  is also easy to derive in many situations. They are useful for model selection tests but are not directly used in MSCs. In particular, satisfaction of the information matrix equality that  $\Omega_T = \mathcal{A}_\beta + o_p(1)$  is not necessary for our discussion of model selection criteria. This is especially relevant under potential model misspecification.

**ASSUMPTION 4**  $\hat{Q}(\beta_0) - TQ(\beta_0) = O_p(\sqrt{T})$ .

If  $\hat{Q}(\beta)$  takes the M estimator form of  $\hat{Q}(\beta) = \sum_{t=1}^T \hat{q}(y_t; \beta)$ , a typical application of the central limit theorem yields

$$\frac{1}{\sqrt{T}} \left( \hat{Q}(\beta_0) - TQ(\beta_0) \right) \xrightarrow{d} N(0, \Sigma_Q), \quad (2.1)$$

where

$$\Sigma_Q = \lim Var \left( \frac{1}{\sqrt{T}} \left( \hat{Q}(\beta_0) - TQ(\beta_0) \right) \right).$$

For example, in the case of the log likelihood function for i.i.d. observations

$$\hat{Q}(\beta) = \sum_{t=1}^T \log f(y_t; \beta), \quad \text{and} \quad Q(\beta) = E \log f(y; \beta),$$

such convergence follows immediately from the central limit theorem:  $\Sigma = Var(\log f(y_t; \beta))$ .

In the case where  $\hat{Q}(\beta)$  takes the form of a quadratic norm such as the GMM estimator, it can be similarly shown that when  $Q(\beta_0) \neq 0$ , or when the GMM model is misspecified, (2.1) continues to hold. On the other hand, when  $Q(\beta_0) = 0$ , or when the GMM model is correctly specified,  $\hat{Q}(\beta_0) - TQ(\beta_0) = \hat{Q}(\beta_0)$  typically converges in distribution to the negative of the quadratic norm of a normal distribution, a special case of which is the  $\chi^2$  distribution when an optimal weighting matrix is being used. Regardless  $\hat{Q}(\beta_0) - TQ(\beta_0) = O_p(1)$  implies  $\hat{Q}(\beta_0) - TQ(\beta_0) = O_p(\sqrt{T})$ , therefore the statement that  $(Qb) = O_p(\sqrt{T})$  is valid in both cases.

The properties of the final terms  $(Qc)$  and  $(Lc)$  are immediate.

This decomposition forms the basis of the analysis of model selection criteria. Given our interest in connecting the proposed model selection criteria to Bayes factors, we now turn specifically to discussing the Bayesian (Schwartz) information criteria,

$$BIC = \hat{Q}(\hat{\beta}) - \hat{L}(\hat{\alpha}) - (\dim(f, Q) - \dim(g, L)) \times \log T$$

implicitly defining  $C_T = \log T$ . This will render the connection to model selection using Bayes factors or posterior odds ratios transparent, with the understanding that the results generalize immediately to other penalty functions that may be of interest. It now can be demonstrated that whether models are nested or nonnested has important consequences for the asymptotic properties of  $\hat{Q}(\hat{\beta}) - \hat{L}(\hat{\alpha})$  and consistency of the BIC as model selection criterion. The two cases of nested and nonnested models are treated in turn.

## 2.1 Nested Models

There are two cases of interest: one model is better in the sense that  $Q(\beta_0) \neq L(\alpha_0)$  and both models are equally good in the sense that  $Q(\beta_0) = L(\alpha_0)$ . Consider the former. Without loss of generality, let  $Q(\beta)$  be the larger nesting model and  $L(\alpha)$  be the smaller model nested in  $Q(\beta)$ . When  $\beta_0 \neq \alpha_0$ , it is typically the case that  $Q(\beta_0) > L(\alpha_0)$ . The goal of BIC is to select the correct model,  $Q(\beta)$ , with probability converging to 1. In this case, this will be true because

$$BIC = \underbrace{(Qa) - (La)}_{O_p(1)} + \underbrace{(Qb) - (Lb)}_{O_p(\sqrt{T})} + \underbrace{T(Q(\beta_0) - L(\alpha_0))}_{O(T)} - (\dim(f, Q) - \dim(g, L)) \times \log T,$$

will be dominated by  $+T(Q(\beta_0) - L(\alpha_0))$ , and therefore increases to  $+\infty$  with probability converging to 1 as  $T \rightarrow \infty$ . In other words, for any  $M > 0$ ,

$$P(BIC > M) \rightarrow 1.$$

Hence BIC selects the correct model with probability converging to 1 if there is one correct model. More generally, BIC selects one of the correct models with probability converging to 1.

Suppose now that both models are equally good, so that  $Q(\beta_0) = L(\alpha_0)$ . Because we are discussing nested models, this also means that  $\beta_0 = \alpha_0$  (with the obvious abuse of the notation of equality with different dimensions), and that  $\hat{Q}(\beta_0) = \hat{L}(\alpha_0)$  almost surely. In the case of likelihood models, this means that  $f(Z_t; \beta_0) = g(Z_t; \alpha_0)$  almost surely, since  $g(\cdot)$  is just a subset of  $f(\cdot)$ . The true model lies in the common subset of  $(\beta_0, \alpha_0)$  and therefore has to be the same model.

In this case the second term is identically equal to 0:

$$(Qb) - (Lb) = \hat{Q}(\beta_0) - \hat{L}(\alpha_0) - (TQ(\beta_0) - TL(\alpha_0)) \equiv 0.$$

Given that the last terms  $(Qc)$  and  $(Lc)$  disappear as a consequence of the equality  $Q(\beta_0) = L(\alpha_0)$ , the BIC comparison is reduced to

$$BIC = \underbrace{(Qa) - (La)}_{O_p(1)} - (\dim(f, Q) - \dim(g, L)) \times \log T.$$

The second term, which is of order  $O(\log T)$ , will dominate. So if  $\dim(f, Q) > \dim(g, L)$ , BIC will converge to  $-\infty$  with probability converging to 1. In other words, for any  $M > 0$ ,

$$P(BIC < -M) \longrightarrow 1.$$

Hence, given two equivalent models, in the sense of  $Q(\beta_0) = L(\alpha_0)$ , the BIC will choose the most parsimonious model (namely the one with the smallest dimension, either  $\dim(f, Q)$  or  $\dim(g, L)$ ) with probability converging to 1.

It is clear from the above arguments that instead of using  $C_T = \log T$ , we can choose any sequence of  $C_T$  such that  $C_T \longrightarrow \infty$  and  $C_T = o(T)$ .

## 2.2 Nonnested Models

Many model comparisons are performed among models that are not nested inside each other. A leading example is the choice between a nonlinear model and its linearized version. For instance, in an extensive literature on estimating consumption Euler equations, there has been debate on the appropriateness of using log-linear versus non-linear Euler equations to estimate household preference parameters. See, for instance, Carroll (2001), Paxson and Ludvigson (1999), and Attanasio and Low (forthcoming). More recently, Fernandez-Villaverde and Rubio-Ramirez (2003) and Fernandez-Villaverde and Rubio-Ramirez (2004b) show that nonlinear filtering methods render

feasible the estimation of some classes of nonlinear dynamic stochastic general equilibrium models. Being equipped with model selection criteria to compare multiple non-linear non-nested models is clearly desirable.

This section shows that in contrast to the case of nested models, the comparison of nonnested models imposes more stringent requirements on  $C_T$  for consistent model selection. Indeed, the further condition on  $C_T$  that  $C_T/\sqrt{T} \rightarrow \infty$  is required in addition to  $C_T = o(T)$ . As an example,  $C_T = \sqrt{T} \log T$  will satisfy both requirements, but  $C_T = \log T$  will not satisfy the second requirement. Let's call the model selection criteria *NIC* when we choose  $C_T = \log T \sqrt{T}$ . To our knowledge these rate conditions are first due to Sin and White (1996) in the context of smooth likelihood models.

Suppose, first, that  $Q(\beta_0)$  is greater than  $L(\alpha_0)$ , implying model  $(f, \beta)$  is better than model  $(g, \alpha)$ . Then, as before, *NIC* is dominated by  $T(Q(\beta_0) - L(\alpha_0))$ , which increases to  $+\infty$  with probability converging to 1. This is true regardless of whether we choose  $C_T = \log T$  or  $C_T = \sqrt{T} \log T$ , since both are of smaller order of magnitude than  $T$ . Hence the behavior of *NIC* when one model is better than the other is essentially the same for both nested models and nonnested models.

On the other hand, when both models are equally good, so that  $Q(\beta_0) = L(\alpha_0)$ , *NIC* comprises the non-vanishing term

$$(Qb) - (Lb) = \hat{Q}(\beta_0) - \hat{L}(\alpha_0) - (TQ(\beta_0) - TL(\alpha_0))$$

which is of order  $O(\sqrt{T})$ . In contrast to model selection criteria in the nested case, it is no longer true that  $\hat{Q}(\beta_0) \equiv \hat{L}(\alpha_0)$  with probability one when the two models are equally good. Hence the model selection criterion takes the form

$$NIC = \underbrace{(Qa) - (La)}_{O_p(1)} + \underbrace{(Qb) - (Lb)}_{O_p(\sqrt{T})} - (\dim(f, Q) - \dim(g, L)) \times C_T.$$

Hence for choice of penalty function  $C_T = \log T \sqrt{T}$ , or any other sequence that increases to  $\infty$  faster than  $\sqrt{T}$ , the last term dominates. So if  $\dim(f, Q) > \dim(g, L)$ , *NIC* converges to  $-\infty$  with probability converging to 1, or

$$P(NIC < -M) \rightarrow 1 \quad \text{for any } M > 0.$$

Thus, *NIC* will choose the most parsimonious model among the two models with probability converging to 1.

In contrast, if the *BIC* had been used, where  $C_T = \log T$ , then the final term fails to dominate. The second term, which is random, might instead dominate. It is immediate that model  $(f, Q)$

and model  $(g, L)$  will both be selected with strictly positive probabilities. Such model selection behavior does not have a clear interpretation when a unique ranking of models is desired. While an analyst may be content with identifying just one of the best fitting models, regardless of whether it is the most parsimonious or not, the definition of consistency we adopt from the literature seeks to uniquely rank models given a model selection criterion and a particular set of assumptions. As such, positive probability weights on multiple models fails our requirements.

While the properties of AIC and BIC for selecting among nested parametric models are well understood (e.g. see Sims (2001) and Gouriéroux and Monfort (1995)), the following theorem serves to summarize the above discussion about the NIC.

**THEOREM 2** *Suppose that assumptions (1), (2), (3) and (4) hold for models  $(f, Q)$  and  $(g, L)$ . Then*

$$NIC = \hat{Q}(\hat{\beta}) - \hat{L}(\hat{\alpha}) - (\dim(f, Q) - \dim(g, L)) \times \log T\sqrt{T}$$

*has the following properties for  $(f, Q)$  and  $(g, L)$  either nested or nonnested:*

1. *If  $Q(\beta_0) > L(\alpha_0)$  then  $P(NIC > M) \rightarrow 1$  for all  $M > 0$ ;*
2. *If  $Q(\beta_0) = L(\alpha_0)$  and  $\dim(f, Q) - \dim(g, L) > 0$  then  $P(NIC < -M) \rightarrow 1$  for all  $M > 0$ .*

### 2.3 Multiple Models

The previous theorem can be immediately extended to the case of multiple models. In the multiple model selection case, suppose there are a total of  $M$  models, of which  $k$  of them are equally good, in the sense of having the same limit objective function with magnitude  $Q(\beta_0)$ , but the other  $M - k$  models have lower valued limit objective functions. Then with probability converging to 1, both the BICs and the NICs for the  $k$  good models will be infinitely larger than the those for the  $M - k$  inferior models. In other words, with probability converging to 1, none of the  $M - k$  inferior models will ever be selected by either BIC or NIC comparison.

On the other hand, the behavior of BIC and NIC can be very different among the  $k$  best models depending on whether they are nested or nonnested. If the  $k$  best models are all nested inside each other, both BICs and NICs will select the most parsimonious model with probability converging to 1. On the other hand, if there exist two models that are not nested inside each other, then BIC will put random weights on at least two models. NIC, however, will still choose the most parsimonious model among all the  $k$  best models.

### 3 Generalized Bayes Factors and Posterior Odds

The previous section demonstrates the requirements for consistent model selection. Appropriate choice of penalty function ensures consistency in the model selection procedure, regardless of whether the models are nested or not, and regardless of whether the models are misspecified or not. We now turn to the specific case of Bayes factors or posterior odds as a model selection criterion. This is of considerable interest given the increasing use of Bayesian methods in economics, particularly the recent macroeconomics literature on estimation of dynamic stochastic general equilibrium models which makes use of the posterior odds ratio for model selection and prediction.

The central goal of this section is to show generalized Bayes factors can be constructed for non-likelihood based models and that the associated generalized posterior odds ratios are equivalent, up to an  $o_p(1)$  term, to the BIC. The contribution of this result is threefold. First, for the parametric likelihood case, this equivalence between the traditional posterior odds ratio and the BIC is demonstrated under considerably weaker conditions than done in the earlier literature, such as those in Fernandez-Villaverde and Rubio-Ramirez (2004a) and Bunke and Milhaud (1998). Second, and related, the equivalence is extended to a general class of statistical criterion functions – that does not require the objective function to be smoothly differentiable – under the same weak regularity conditions. Third, we show that because the generalized posterior odds ratio implicitly chooses a penalization function with  $C_T = \log T$ , it is inherently an inconsistent model selection criterion for comparison of nonnested models.

Consider the integral transformations of models  $(f, Q)$  and  $(g, L)$  defined as

$$e^{\hat{Q}(\beta)} \pi(\beta) / \int e^{\hat{Q}(\beta')} \pi(\beta') d\beta' \quad \text{and} \quad e^{\hat{L}(\alpha)} \gamma(\alpha) / \int e^{\hat{L}(\alpha')} \gamma(\alpha') d\alpha',$$

where  $\pi(\beta)$  and  $\gamma(\alpha)$  are the corresponding prior distributions on the two parameter spaces  $\mathcal{B}$  and  $\Gamma$ . Because these objects are properly interpreted as distributions, Chernozhukov and Hong (2003) show that estimation can proceed using simulation methods from Bayesian statistics, such as Markov Chain Monte Carlo methods, using various location measures to identify the parameters of interest. These so-called Laplace-type estimators are therefore defined analogously to Bayesian estimators but use general statistical criterion functions in place of the parametric likelihood function. By considering integral transformations of these statistical functions to give quasi-posterior distributions, this approach provides a useful alternative consistent estimation method to handle intensive computation of many classical extremum estimators – such as, among others, the GMM estimators of Hansen (1982), Powell (1986)'s censored median regression, nonlinear IV regression such as Berry, Levinsohn, and Pakes (1995) and instrumental quantile regression as in Chernozhukov and Hansen (2005).

Associated with these Laplace-type estimators is the generalized posterior odds ratio, defined as

$$\log \frac{P_Q}{P_L} \times \frac{\int e^{\hat{Q}(\beta)} \pi(\beta) d\beta}{\int e^{\hat{L}(\alpha)} \gamma(\alpha) d\alpha},$$

when the two models have prior probability weights  $P_Q$  and  $P_L = 1 - P_Q$ . As an example, consider the case when  $\hat{Q}(\beta)$  and  $\hat{L}(\alpha)$  are the log-likelihood functions for two parametric models  $f(y; \beta)$  and  $g(y; \alpha)$ . It follows that

$$\hat{Q}(\beta) = \sum_{t=1}^T \log f(y_t; \beta) = \log \prod_{t=1}^T f(y_t; \beta) \equiv \log \tilde{Q}(\beta)$$

and

$$\hat{L}(\alpha) = \sum_{t=1}^T \log g(y_t; \alpha) = \log \prod_{t=1}^T g(y_t; \alpha) \equiv \log \tilde{L}(\alpha)$$

give the log likelihoods for each model  $(f, Q)$  and  $(g, L)$ . The posterior odds can then be written in the more familiar form

$$\log \frac{\int \tilde{Q}(\beta) \pi(\beta) d\beta}{\int \tilde{L}(\alpha) \gamma(\alpha) d\alpha} \times \frac{P_Q}{P_L}.$$

### 3.1 Large Sample Properties of generalized bayes factors

The following theorem connects the properties of the pseudo Bayes posterior distribution to the extremum estimator and therefore establishes the relation between the generalized Bayes factor and the extremum estimator analog of BICs. Under the regularity conditions stated in assumptions 1 to 3, the generalized Bayes factor is asymptotically equivalent to the use of the BIC as a model selection criterion.

**THEOREM 3** *Under assumptions (1), (2) and (3), the generalized Bayes factor satisfies the following relation*

$$T^{\frac{\dim(\beta)}{2}} \int e^{\hat{Q}(\beta) - \hat{Q}(\hat{\beta})} \pi(\beta) d\beta \xrightarrow{P} \pi(\beta_0) (2\pi)^{\frac{\dim(\beta)}{2}} \det(-\mathcal{A}_\beta)^{-1/2}.$$

The formal details of the proof are relegated to an appendix. In the following we discuss the informal intuition behind the result assuming that  $\hat{Q}(\beta)$  is smoothly differentiable. Consider the expression:

$$\log \int e^{\hat{Q}(\beta)} \pi(\beta) d\beta - \hat{Q}(\hat{\beta}) = \log \int e^{\hat{Q}(\beta) - \hat{Q}(\hat{\beta})} \pi(\beta) d\beta.$$

It can be approximated up to an  $o_p(1)$  term as follows. First, as before,  $\hat{Q}(\beta) - \hat{Q}(\hat{\beta})$  can be approximated by a quadratic function centered at  $\hat{\beta}$ , in a size  $1/\sqrt{T}$  neighborhood around  $\hat{\beta}$ :

$$\hat{Q}(\beta) - \hat{Q}(\hat{\beta}) \approx \frac{1}{2} (\beta - \hat{\beta})' \frac{\partial^2 \hat{Q}_T(\hat{\beta})}{\partial \beta \partial \beta'} (\beta - \hat{\beta}).$$

Second, in this  $1/\sqrt{T}$  size neighborhood, the prior density is approximately constant around  $\beta_0$  as  $\pi(\hat{\beta}) \xrightarrow{p} \pi(\beta_0)$ . Therefore the impact of the prior density is negligible except for the value at  $\pi(\beta_0)$ . On the other hand, outside the  $1/\sqrt{T}$  size neighborhood around  $\hat{\beta}$ , the difference between  $e^{\hat{Q}(\beta)}$  and  $e^{\hat{Q}(\hat{\beta})}$  is exponentially small, and makes only asymptotically negligible construction to the overall integral.

The appendix proves formally the approximation:

$$\begin{aligned} \log \int e^{\hat{Q}(\beta)} \pi(\beta) d\beta - \hat{Q}(\hat{\beta}) &= \log \pi(\beta_0) \int e^{\frac{1}{2}(\beta - \hat{\beta})' \frac{\partial^2 \hat{Q}_T(\hat{\beta})}{\partial \beta \partial \beta'} (\beta - \hat{\beta})} d\beta + o_p(1) \\ &= \log \pi(\beta_0) \int e^{\frac{1}{2}(\beta - \hat{\beta})' T \mathcal{A}_\beta (\beta - \hat{\beta})} d\beta + o_p(1) \\ &= \log \left( \pi(\beta_0) (2\pi)^{\frac{\dim(\beta)}{2}} \det(-T \mathcal{A}_\beta)^{-\frac{1}{2}} \right) + o_p(1) \\ &= \underbrace{\log \pi(\beta_0) + \frac{\dim(\beta)}{2} \log(2\pi) - \frac{1}{2} \det(-\mathcal{A}_\beta)}_{C(\mathcal{A}_\beta, \beta)} - \frac{1}{2} \dim(\beta) \log T + o_p(1), \end{aligned}$$

where  $C(\mathcal{A}_\beta, \beta)$  can also depend on the prior weight on the model itself. The constant term  $C(\mathcal{A}_\beta, \beta)$  will eventually be dominated by  $-\dim(\beta) \log T$  whenever it becomes relevant. Therefore  $C(\mathcal{A}_\beta, \beta)$  never matters for the purpose of model selection.

Exploiting this approximation, the log posterior odds ratio can be written as

$$\hat{Q}(\hat{\beta}) - \hat{L}(\hat{\alpha}) - \left( \frac{1}{2} \dim(\beta) - \frac{1}{2} \dim(\alpha) \right) \log T + C(\mathcal{A}_\beta, \beta) - C(\mathcal{A}_\alpha, \alpha) + \log \frac{P_Q}{P_L}$$

implicitly defining a penalization term  $C_T = \log T$ . It is immediately clear that this expression is asymptotically equivalent, up to a constant that is asymptotically negligible, to BIC. As discussed previously, this choice of  $C_T$  gives a consistent model selection criterion only when comparing nested models. In the nonnested case, when the null hypothesis contends that two models are asymptotically equivalent in fit and therefore misspecified, it fails to select the most parsimonious model with probability converging to 1.

Fernandez-Villaverde and Rubio-Ramirez (2004a) also explore the large sample properties of the posterior odds ratio in the case of parametric likelihood. They demonstrate, assuming there exists

a unique asymptotically superior model in the sense of  $Q(\beta) > L(\alpha)$ , that the posterior odds ratio will select model  $(f, Q)$  with probability converging to 1 as  $T$  goes to infinity. Theorem 3 serves to generalize significantly this finding in three dimensions. First, the results presented here are established under very weak regularity conditions. Second, the results apply to a large class of estimation procedures based on minimizing random distance functions — not only likelihood models. These functions need not be differentiable. Third, only by considering a null hypothesis that admits the possible equivalence of the two models, in the sense that  $Q(\beta) = L(\alpha)$ , can a complete classical interpretation properly be given to the posterior odds ratio as a model selection criterion. Statistically distinguishing models is fundamental to classical hypothesis testing. Given the absence of prior information in classical estimation it is therefore necessary to entertain the possibility that two or more models are equally good in a testing framework. The results of this paper are therefore seen to be couched naturally in the classical paradigm.

#### 4 Model Selection Tests

One of the deficiencies of model selection criteria is the inability to make probability statements about the choice of best model in an hypothesis testing framework. Indeed, model selection criteria only determine which model to select asymptotically. However, Vuong (1989) demonstrates conditions under which model selection criteria can be reformulated to provide a model selection test, allowing one to make frequentist probability statements regarding the likelihood of available models. While Vuong (1989)'s test applies originally to two models, multiple model tests are considered in several recent papers, including White (2000) and Hansen (2003). We now show analogous results for generalized posterior odds ratios, therefore providing a means to evaluate statistically how large is the difference in fit between two nonnested models.

Again, focus is given to a two model comparison to simplify discussion. Let  $P_Q$  be the prior probability of model  $(f, Q)$  and  $P_L$  be the prior probability of model  $(g, L)$ . The posterior odds ratio is then defined as

$$OR = \frac{P_Q \int e^{\hat{Q}(\beta)} \pi(\beta) d\beta}{P_L \int e^{\hat{L}(\alpha)} \pi(\alpha) d\alpha}.$$

The results presented in the previous sections show that

$$\log(OR) = \hat{Q}(\hat{\beta}) - \hat{L}(\hat{\alpha}) - \underbrace{\frac{1}{2}(\dim(\beta) - \dim(\alpha)) \log T}_{O(\log T)} + \underbrace{C(\mathcal{A}_\beta, \beta) - C(\mathcal{A}_\alpha, \alpha)}_{C(\mathcal{A}_\beta, \beta, \mathcal{A}_\alpha, \alpha)} + \log \frac{P_Q}{P_L} + o_p(1).$$

The implications of using odds ratios for model selection tests are now clear from the results we obtain in the previous sections.

#### 4.1 Nonnested models

For strictly nonnested models, the null hypothesis of  $Q(\beta_0) = L(\alpha_0)$  implies that neither model can be correctly specified. Under this null hypothesis,

$$\frac{1}{\sqrt{T}} \left( \hat{Q}(\hat{\beta}) - \hat{L}(\hat{\alpha}) \right) = \frac{1}{\sqrt{T}} \left( \hat{Q}(\beta_0) - \hat{L}(\alpha_0) \right) + o_p(1).$$

Hence

$$\frac{1}{\sqrt{T}} \log(OR) = \frac{1}{\sqrt{T}} \left( \hat{Q}(\beta_0) - \hat{L}(\alpha_0) \right) + o_p(1).$$

The asymptotic distribution of  $\frac{1}{\sqrt{T}} \left( \hat{Q}(\beta_0) - \hat{L}(\alpha_0) \right)$ , which is typically normal, can easily be estimated by numerical evaluation, repeated parametric simulation, or resampling methods. But it is usually not related to the shape of the posterior distribution of  $\beta$  and  $\alpha$ .

Local power properties of Vuong (1989)'s model selection test are not affected either by the addition of the order  $\log T$  term in BIC, because this term only relates to the dimension of the parameters and is not related to the local departures from the null.

Obviously, the NIC we discussed above cannot be used to formulate a nonnested model selection test because of its deterministic behavior under the null hypothesis.

#### 4.2 Nested models

For nested models (suppose  $g$  is nested inside  $f$ ), the log of the posterior odds behaves very differently from the likelihood ratio test statistic  $\hat{Q}(\hat{\beta}) - \hat{L}(\hat{\alpha})$ .

Under the null hypothesis that  $Q(\beta_0) = L(\alpha_0)$ , which implies  $\hat{Q}(\beta_0) \equiv \hat{L}(\alpha_0)$  because of the nesting assumption, the log likelihood ratio test statistic  $\hat{Q}(\hat{\beta}) - \hat{L}(\hat{\alpha})$  has an asymptotic chi square distribution under correct specification and is asymptotically distributed as a weighted sum of chi squares under misspecification.

However, in this case with probability converging to 1,

$$\log(OR) \longrightarrow -\infty$$

at the rate of  $\log T$  because of the term  $-\frac{1}{2}(\dim(\beta) - \dim(\alpha)) \log T$ . In other words, if the conventional chi square distribution is being used, then using  $\log(OR)$  in place of the LR test statistic will

result in 100% nonrejection of the null hypothesis with probability converging to 1. The behavior of  $\log(OR)$  in this case is essentially deterministic in large sample. It is, therefore, impossible to use  $\log(OR)$  to formulate a model selection test statistic between nesting models.

## 5 Bayesian Predictive Analysis

A well known classical result can be stated as follows: the frequentist asymptotic distribution properties of post selection estimation and prediction are not affected by the model selection step as long as the model selection criteria is consistent. That Bayesian researchers often consider averaging parameter estimates or predictions obtained from several different models begs the question: does an analogous result hold for Bayesian inference? By appealing to the results of Theorem 3, we provide an answer in the affirmative, by proceeding in two steps. First, for a given model, the relation between classical and bayesian prediction is analyzed and shown to be equivalent up to the order  $o_p(1/\sqrt{T})$ . Second, the asymptotic properties of Bayesian model averaging are derived.

To simplify notation, focus is given to the case of two models, although generalization to multiple models is immediate. Let  $Y_T = (y_1, \dots, y_T)$  denote the entire data set. To focus on Bayesian interpretations, one can assume that both  $\hat{Q}$  and  $\hat{L}$  are the log likelihood functions corresponding to two different models, although other distance functions can also be used to obtain average predictions.

### 5.1 Single Model Prediction

Consider a single model  $(f, Q)$ , and the associated Bayesian inference problem of predicting  $y_{T+1}$  given a data set  $Y_T$  with observations up to  $T$ . Typically we need to calculate the predictive density of  $y_{T+1}$  given  $Y_T$ , and for this purpose need to average over the posterior distribution of the model parameters  $\beta$  given the data  $Y_T$ :

$$\begin{aligned} f(y_{T+1}|Y_T) &= \int f(y_{T+1}|Y_T, \beta) f(\beta|Y_T) d\beta \\ &= \int f(y_{T+1}|Y_T, \beta) \frac{e^{\hat{Q}(\beta)} \pi(\beta)}{\int e^{\hat{Q}(\beta)} \pi(\beta) d\beta} d\beta. \end{aligned}$$

We will assume that  $f(y_{T+1}|Y_T; \beta) = f(y_{T+1}|\bar{Y}_T; \beta)$ , where  $\bar{Y}_T$  is a finite dimensional subcomponent of  $Y$ . For example,  $\bar{Y}_T$  can be  $y_T$ , the most recent observation in  $Y_T$ . It is well understood that the first order randomness in the prediction is driven by  $f(y_{T+1}|\bar{Y}_T, \beta)$ . The length of the prediction interval comprises two parts: the first is due to  $f(y_{T+1}|\bar{Y}_T, \beta)$  and the second is due to the uncertainty from estimating  $\beta$ . While the second part will decrease to zero as the sample size

$T$  increases, the first part remains constant.

We are interested in the second order uncertainty in the prediction that is due to the estimation of the parameter  $\beta$ , and therefore will consider a fixed value  $\bar{y}$  of the random component  $\bar{Y}_T$ , and consider

$$f(y_{T+1}|\bar{y}, Y_T) = \int f(y_{T+1}|\bar{y}, \beta) f(\beta|Y_T) d\beta, \quad (5.2)$$

where  $\bar{y}$  can potentially differ from the observed realization of  $\bar{Y}_T$  in the sample. For example, one might consider out of sample predictions where  $\bar{y}$  does not take the realized value of  $\bar{Y}_T$ .

Point predictions can be constructed as functionals of the posterior predictive density  $f(y_{T+1}|\bar{y}, Y_T)$ . For example, a mean prediction can be obtained by

$$E(y_{T+1}|\bar{y}, Y_T) = \int E(y_{T+1}|\bar{y}, Y_T; \beta) f(\beta|Y_T) d\beta.$$

On the other hand, a median prediction is given by

$$\text{med}(f_{y_{T+1}}(\cdot|\bar{y}, Y)) = \inf \left\{ x : \int^x f(y_{T+1}|\bar{y}, Y) dy_{T+1} \geq \frac{1}{2} \right\}.$$

Under suitable regularity conditions,  $\hat{\beta}$  is an asymptotic sufficient statistic for the random posterior distribution and  $f(\beta|Y_T)$  is approximately normal with mean  $\hat{\beta}$  and variance  $-\frac{1}{T}(\mathcal{A}_\beta)^{-1}$ :

$$f(\beta|Y_T) \stackrel{A}{\approx} N\left(\hat{\beta}, -\frac{1}{T}(\mathcal{A}_\beta)^{-1}\right).$$

Hence  $f(y_{T+1}|\bar{y}, Y_T)$  can be shown to be approximately, up to order  $o_p\left(\frac{1}{\sqrt{T}}\right)$ ,

$$\begin{aligned} & \int f(y_{T+1}|\bar{y}, \beta) (2\pi)^{-\frac{\dim(\beta)}{2}} \det(-T\mathcal{A}_\beta)^{\frac{1}{2}} e^{-\frac{1}{2}(\beta-\hat{\beta})'(-T\mathcal{A}_\beta)(\beta-\hat{\beta})} d\beta \\ &= \int f\left(y_{T+1}|\bar{y}, \hat{\beta} + \frac{h}{\sqrt{T}}\right) (2\pi)^{-\frac{\dim(\beta)}{2}} \det(-\mathcal{A}_\beta)^{\frac{1}{2}} e^{-\frac{h'(-\mathcal{A}_\beta)h}{2}} dh. \end{aligned}$$

Researchers are most often interested in mean predictions and predictive intervals. Mean prediction is convenient to analyze because of its linearity property. For instance,

$$\begin{aligned} E(y_{T+1}|\bar{y}, Y_T) &= \int E(y_{T+1}|\bar{y}, \beta) f(\beta|Y_T) d\beta \\ &= \int E\left(y_{T+1}|\bar{y}, \hat{\beta} + \frac{h}{\sqrt{T}}\right) f(h|Y) dh. \end{aligned}$$

If  $E(y_{T+1}|\bar{y}; \beta)$  is linear in  $\beta$ , then it is easy to see from the fact that  $f(h|Y)$  is approximately normal with mean 0 and variance  $-\mathcal{A}_\beta^{-1}$  that

$$E(y_{T+1}|\bar{y}, Y_T) = E(y_{T+1}|\bar{y}; \hat{\beta}) + o_p\left(\frac{1}{\sqrt{T}}\right).$$

Hence Bayesian mean prediction is asymptotically equivalent up to order  $\frac{1}{\sqrt{T}}$  with frequentist prediction, where the predictive density is formed using the extremum estimate  $\hat{\beta}$ ,  $f(y_{T+1}|\bar{y}, \hat{\beta})$ .

On the other hand, even if  $E(y_{T+1}|\bar{y}; \beta)$  is not linear in  $\beta$ , as long as it is sufficiently smooth in  $\beta$ , it is still possible to use a first order Taylor expansion of  $E(y_{T+1}|\bar{y}; \hat{\beta} + \frac{h}{\sqrt{T}})$  around  $\hat{\beta}$  to prove that the same result holds. Given the generic notation of  $y_{T+1}$ , these arguments also apply without change to more general functions of  $y_{T+1}$ , in the sense that for a general function  $t(\cdot)$  of  $y_{T+1}$ ,

$$E(t(y_{T+1})|\bar{y}, Y_T) = E(t(y_{T+1})|\bar{y}, \hat{\beta}) + o_p\left(\frac{1}{\sqrt{T}}\right).$$

These results can be generalized to nonlinear predictions that can be expressed as nonlinear functions of the predictive density  $f(y_{T+1}|\bar{y}, Y_T)$ . For example, median prediction and predictive intervals can readily be constructed. In the following, we will formulate a general prediction as one that is defined through minimizing posterior expected nonlinear and nonsmooth loss functions.

A general class of nonlinear predictors  $\hat{\lambda}(\bar{y}, Y_T)$  can be defined as the solution to minimizing an expected loss function  $\rho(\cdot)$ :

$$\begin{aligned} \hat{\lambda}(\bar{y}, Y_T) &= \arg \min_{\lambda} E(\rho(y_{T+1}, \lambda) | \bar{y}, Y_T) \\ &= \arg \min_{\lambda} \int \rho(y_{T+1}, \lambda) f(y_{T+1}|\bar{y}, Y_T) dy_{T+1}, \end{aligned}$$

where the predictive density  $f(y_{T+1}|\bar{y}, Y_T)$  is defined in equation (5.2). For example, if we are interested in constructing a one-sided  $\tau$ th predictive interval, then we can take

$$\rho(y_{T+1}, \lambda) \equiv \rho_\tau(y_{T+1} - \lambda) = (\tau - 1(y_{T+1} \leq \lambda))(y_{T+1} - \lambda).$$

Unless stated otherwise, in the following focus is given to this loss function.

We are interested in comparing  $\hat{\lambda}(\bar{y}, Y)$  to the nonlinear frequentist prediction, defined as

$$\tilde{\lambda}(\bar{y}, Y_T) = \arg \min_{\lambda} \int \rho(y_{T+1}, \lambda) f(y_{T+1}|\bar{y}, \hat{\beta}) dy_{T+1}.$$

Also define the infeasible loss function where the uncertainty from estimation of the unknown parameters is absent as

$$\begin{aligned}\bar{\rho}(\bar{y}, \lambda; \beta) &= \int \rho(y_{T+1}, \lambda) f(y_{T+1}|\bar{y}, \beta) dy_{T+1} \\ &= E(\rho(y_{T+1}, \lambda) | \bar{y}, \beta).\end{aligned}$$

Then the Bayesian predictor and the frequentist predictor can be written as

$$\hat{\lambda}(\bar{y}, Y_T) = \arg \min_{\lambda} \int \bar{\rho}(\bar{y}, \lambda; \beta) f(\beta|Y_T) d\beta,$$

and

$$\tilde{\lambda}(\bar{y}, Y_T) = \arg \min_{\lambda} \bar{\rho}(\bar{y}, \lambda; \hat{\beta}).$$

The next theorem establishes the asymptotic relation between  $\hat{\lambda}(\bar{y}, Y_T)$  and  $\tilde{\lambda}(\bar{y}, Y_T)$ .

**THEOREM 4** *Under assumptions (1), (2) and (3), and assuming that for each  $\bar{y}$ ,  $\bar{\rho}(\bar{y}, \lambda; \beta)$  is three times continuous differentiable in  $\beta$  and  $\lambda$  in a small neighborhood of  $\beta_0$  and  $\lambda_0 \equiv \arg \min_{\lambda} \bar{\rho}(\bar{y}, \lambda, \beta_0)$ , with uniformly integrable derivatives, then*

$$\sqrt{T} \left( \hat{\lambda}(\bar{y}, Y_T) - \tilde{\lambda}(\bar{y}, Y_T) \right) \xrightarrow{p} 0.$$

Note that the condition of this theorem only requires the integrated loss function  $\bar{\rho}(\bar{y}, \lambda; \beta)$  to be smoothly differentiable in  $\lambda$  and  $\beta$ , and does not impose smoothness conditions directly on  $\rho(y_{T+1}, \lambda)$ . Therefore the results cover predictive intervals as long as the predictive density given  $\beta$  is smoothly differentiable around the percentiles that are to be predicted.

Different loss functions can be used to construct a variety of Bayesian point estimators from the posterior density. Unless the loss function is convex and symmetric around 0, the corresponding Bayes estimator is typically different from the frequentist maximum likelihood estimator at the order of  $O_p(1/\sqrt{T})$ . In contrast, we found that when different loss functions are used to define different predictors, Bayesian and frequentist predictors coincide with each other up to the order  $o_p(1/\sqrt{T})$ . This is probably a more relevant result concerning loss functions because researchers are typically more interested in using loss functions to define the properties of predictions than to define the estimator itself.

## 5.2 Model averaging

Now consider prediction using the average of two or more models. It is well known that the frequentist asymptotic distribution properties of post selection estimation and prediction are not affected by the model selection step as long as the model selection criterion is consistent. Here we demonstrate conditions under which the property of consistent model selection is possessed by the Bayesian prediction procedure (in the sense of asymptotically placing unitary probability weight on a single model).

With two models  $(f, Q)$  and  $(g, L)$ , we can write the predictive density as

$$f(y_{T+1}|\bar{y}, Y_T) = \frac{BF_Q}{BF_Q + BF_L} f(y_{T+1}|\bar{y}, Y_T, Q) + \frac{BF_L}{BF_Q + BF_L} f(y_{T+1}|\bar{y}, Y_T, L)$$

where the posterior probability weights on each model are defined as

$$BF_Q = P_Q \int e^{\hat{Q}(\beta)} \pi(\beta) d\beta \quad \text{and} \quad BF_L = P_L \int e^{\hat{L}(\alpha)} \gamma(\alpha) d\alpha.$$

In the case of likelihood models  $e^{\hat{Q}(\beta)} = f(Y_T|\beta)$  and  $e^{\hat{L}(\alpha)} = f(Y_T|\alpha)$ . In particular, note

$$f(Y_T) = BF_Q + BF_L$$

is the marginal density of the data  $Y_T$ . The model specific predictive densities are respectively,

$$\begin{aligned} f(y_{T+1}|\bar{y}, Y_T, Q) &= \frac{\int e^{\hat{Q}(\beta)} \pi(\beta) f(y_{T+1}|\bar{y}, \beta) d\beta}{\int e^{\hat{Q}(\beta)} \pi(\beta) d\beta} \\ &= \int f(y_{T+1}|\bar{y}, \beta) f(\beta|Y_T, Q) d\beta \end{aligned}$$

and  $f(y_{T+1}|\bar{y}, Y_T, L) = \int f(y_{T+1}|\bar{y}, \alpha) f(\alpha|Y_T, L) d\alpha$ . As before, a general class of predictions can be defined by

$$\hat{\lambda}(\bar{y}, Y_T) = \arg \min_{\lambda} \int \rho(y_{T+1}, \lambda) f(y_{T+1}|\bar{y}, Y_T) dy_{T+1}.$$

The following theorem establishes the asymptotic properties of the Bayesian predictor formed by averaging the two models.

**THEOREM 5** *Suppose the assumptions stated in Theorem 4 hold for both models  $(f, Q)$  and  $(g, L)$ . Also assume that one of the following two conditions holds:*

1.  $Q(\beta_0) > L(\alpha_0)$ ,  $(f, Q)$  and  $(g, L)$  can be either nested or nonnested.

2.  $Q(\beta_0) = L(\alpha_0)$  but  $(f, Q)$  is nested inside  $(L, g)$ . In addition,  $\dim(\alpha) - \dim(\beta) > 1$ .

Then  $\sqrt{T} \left( \hat{\lambda}(\bar{y}, Y_T) - \tilde{\lambda}(\bar{y}, Y_T) \right) \xrightarrow{p} 0$ , where  $\tilde{\lambda}(\bar{y}, Y_T) = \arg \min_{\lambda} \bar{\rho}(\bar{y}, \lambda; \hat{\beta})$  is formed from  $(f, Q)$ .

It is also clear from previous discussions that if  $Q(\beta_0) = L(\alpha_0)$  and  $(f, Q)$  and  $(g, L)$  are not nested, then the weight on neither  $BF_Q$  nor  $BF_L$  will converge to 1, and the behavior of  $f(y_{T+1}|\bar{y}, Y_T)$  will be a random average between the two models.

These theoretical results contrast with the conventional wisdom regarding forecasting that averages (or some combination) of available forecasting models perform better than any one model taken in isolation. Recent studies by Stock and Watson (2001), Stock and Watson (2003) and Wright (2003) adduce evidence consistent with this view. The former papers show that a range of combination forecasts outperform benchmark autoregressive models when forecasting output growth of seven industrialized economies and similarly that the predictive content of asset prices in forecasting inflation and output growth when models are averaged improves upon univariate benchmark models. The latter paper shows that Bayesian model averaging can improve upon the random walk model of exchange rate forecasts. Understanding the fundamental cause of this disjunction between theoretical and empirical findings is left for future work. However, these empirical findings might suggest that many of the models under consideration might be locally close so that a local analysis based on the assumption that the distance between two models shrinks to zero as the sample size increases might provide a frequentist justification for the advantage of model averaging.

## 6 Examples

Many estimation procedures based on a variety of objective functions, including parametric likelihood models, generalized method of moment models, and generalized empirical likelihood models, are special cases of the results presented above. The following gives some specific examples of generalized posterior odds ratios constructed from these three random distance functions. Only the latter implicitly delivers a penalization term for the dimension of the information used in the estimation procedure, though all penalize the parametric dimension of the model. However, all examples have the common property that they fail to deliver consistent model selection when considering nonnested models given the results of Section 3.

To recapitulate the parametric likelihood case previously discussed, consider the model  $f(y_t, \beta)$ . The objective function  $\hat{Q}(\beta)$  is given by the associated log-likelihood function  $\hat{Q}(\beta) = \sum_{t=1}^T \log f(y_t, \beta)$ .

Applying the results of section 3 it is immediate that the volume

$$P_Q \int e^{\hat{Q}(\beta)} \pi(\beta) d\beta = e^{\hat{Q}(\hat{\beta})} e^{C(\mathcal{A}_{\beta}, \beta)} T^{-\frac{1}{2} \dim(\beta)} \times e^{o_p(1)},$$

shrinks at a rate related to the number of parameters of the model. Hence only model parameterization is being penalized.

It is worth underscoring that while penalization for parameterization is a natural choice for parametric likelihood models, in general it is not obvious this is necessarily the most desirable form of penalty function outside the likelihood framework. In the context of generalized Bayesian inference, there may be grounds to consider alternative penalty functions that also penalize the dimension of the estimation procedure. For instance, Andrews and Lu (2001) and Hong, Preston, and Shum (2003) consider such penalty functions that involve both the number of parameters and the number of moment conditions. The use of additional moments in GMM and GEL contexts is desirable on efficiency grounds. The following discussion explores such estimation procedures in conjunction with the Laplace-type estimators and associated generalized Bayes factors of section 3.

Andrews and Lu (2001) proposed consistent model and moment selection criteria for GMM estimation. Interestingly, such selection criteria, which award the addition of moment conditions and penalize the addition of parameters, can not be achieved using a generalized bayes factor constructed from the GMM objective function:

$$\hat{Q}(\beta) = T g_T(\beta)' W_T g_T(\beta), \quad \text{where} \quad g_T(\beta) = \frac{1}{T} \sum_{t=1}^T m(y_t, \beta),$$

for  $m(y_t, \beta)$  a vector of moment conditions and  $W_T \xrightarrow{p} W$  positive definite.

With the objective function  $\hat{Q}(\beta)$  and associated prior  $P_Q$ , the volume

$$P_Q \int e^{\hat{Q}(\beta)} \pi(\beta) d\beta = e^{\hat{Q}(\hat{\beta})} e^{C(\mathcal{A}_{\beta}, \beta)} T^{-\frac{1}{2} \dim(\beta)} \times e^{o_p(1)},$$

again only shrinks at a rate related to the number of parameters and not the number of moment conditions. Generalized Bayes factors using the quadratic GMM objective functions do not encompass the model selection criteria proposed by Andrews and Lu (2001). This is not surprising given the lack of likelihood interpretation of conventional two step GMM estimators based on a quadratic norm of the moments.

The recent literature on generalized empirical likelihood (GEL) estimators proposed an information theoretic alternative to efficient two step method of moment estimators that minimizes a likelihood

distance between the data and the moment models. A GEL estimator is defined as the saddle point of a GEL function,

$$\left(\hat{\beta}, \hat{\lambda}\right) = \arg \max_{\beta \in \mathcal{B}} \arg \min_{\lambda \in \Lambda} \hat{Q}(\beta, \lambda).$$

For example, in the case of exponential tilting,

$$\hat{Q}(\beta, \lambda) = \sum_{t=1}^T \rho(\lambda' m(y_t, \beta)) \quad \text{where} \quad \rho(x) = e^x.$$

Given the connection between GEL and parametric likelihood models, it is interesting to ask whether one can define a generalized bayes factor based on the GELs that mimicks the model and moment selection criteria of Andrews and Lu (2001) and others. We define such a GEL Bayes factor as

$$\text{GELBF} = \int \frac{1}{\int e^{-\hat{Q}(\beta, \lambda)} \phi(\lambda) d\lambda} \pi(\beta) d\beta$$

where  $\phi(\lambda)$  is a prior density on the lagrange multiplier  $\lambda$ . Intuitively, a large volume of the integral

$$\int e^{-\hat{Q}(\beta, \lambda)} \phi(\lambda) d\lambda$$

indicates that  $\lambda$  tends to be large, and therefore that the GMM model  $(f, Q)$  is more likely to be incorrect, or misspecified. Hence, we use its inverse to indicate the strength of the moments involved in the GMM model. The asymptotic equivalence between this GEL bayes factor and a model and moment selection criteria is given in the following proposition:

**Proposition 1** *Suppose assumptions (1), (2) and (3) hold, assume also that there are interior points in the paramter space  $\lambda_0$  and  $\beta_0$  such that  $T\hat{Q}(\hat{\beta}, \hat{\lambda}) - Q(\beta_0, \lambda_0) = O_p(1)$ , where  $Q(\beta, \lambda)$  is the uniform probably limit of  $\hat{Q}(\beta, \lambda)/T$ . Then*

$$\log \text{GELBF} = \hat{Q}(\hat{\beta}, \hat{\lambda}) + \frac{1}{2} (\dim(\lambda) - \dim(\beta)) \log T + O_p(1).$$

As a consequence of this proposition, asymptotically the GEL Bayes factor puts all weight on models with the best fit  $Q(\beta_0, \lambda_0)$ , and among models with equal best fit the one with the largest number of overidentifying moment conditions  $\dim(\lambda) - \dim(\beta)$ . Model selection behaviors are undetermined among models with equal fit and equal number of overidentifying moment conditions.

## 7 Conclusion

This paper analyzes the relation between consistent model selection criteria and Bayesian procedures for model averaging and selection for a large class of models based on minimizing nonsmooth random distance functions. Importantly, the model selection criteria are consistent regardless of whether models are nested or nonnested and regardless of whether models are correctly specified or not.

Consistency is defined as requiring that inferior models are chosen with probability converging to zero and that among the set of equally best fitting models, the model with the highest degree of parsimony is chosen. For model selection criteria of the form

$$MSC = \hat{Q}(\beta) - \dim(f, Q) * C_T$$

consistency requires the latter penalty function to have the properties  $C_T = o(T)$  and  $C_T \rightarrow \infty$  in the case of nested models. In the case of non-nested models the additional restriction that  $\frac{C_T}{\sqrt{T}} \rightarrow \infty$  is also required for consistency in the model selection procedure.

Because it adopts a penalty function that fails to satisfy this latter property, the BIC represents an inconsistent model selection criterion for nonnested models. As posterior odds ratios can be shown to be equivalent to BIC up to a negligible term, traditional methods for model selection in Bayesian inference similarly lead to inconsistency in comparison of nonnested models. These results are established under very weak regularity conditions and extended to generalized posterior odds ratios constructed from a class of Laplace-type estimators.

Finally, the connections between Bayesian and classical predictions are explored. For predictions based on minimization of a general class of nonlinear loss functions, we demonstrated conditions under which the asymptotic distribution properties of prediction intervals are not affected by model averaging and the posterior odds place asymptotically unitary probability weight on a single model. This establishes an analogue to the well-known classical result: that asymptotical distribution properties of post-selection estimation and prediction are not affected by first stage model selection so long as the model selection criterion is consistent.

Of course, when confronted with multiple misspecified models it is clear that there is no unique approach to model selection. We consider one possible approach in the spirit of AIC, BIC and related variants that have been proposed to tackle model selection in particular statistical environments. In so doing, one particular criterion for ranking models is being evaluated, and the conditions under which a unique ranking obtains determined. Importantly, while the number of parameters might appropriately capture model parsimony in the nested case, this may not necessarily be true in the

nonnested case. There may be better ways to capture model complexity in this instance. Exploring possible criteria is left for future work.

This paper should not be interpreted as claiming that this criterion is in any way optimal, or necessarily superior to other alternatives. Importantly, the proposed model selection criterion should not be used to the exclusion of standard measures of model fit. For example, there are a number of single model specification tests available — see, among others, Newey (1985), Fan (1994) and Hansen (1982) — which should be applied in assessing model fit. Even the best model chosen by NIC may be rejected by such specification tests indicating that all models are far from the true data generating process (see Sims (2003)). Indeed, as emphasized by Gourieroux and Monfort (1995), hypothesis testing procedures may lead to the simultaneous acceptance or rejection of two nonnested hypothesis. The former may reflect a lack a data while the latter may suggest the testing framework is misspecified. Our NIC, therefore, provides a complementary approach for discriminating between models.

In analyzing the connections between BIC and our NIC it is also worth noting that from a decision theoretic perspective the former implicitly defines a zero-loss function. This form of loss function may well have questionable merit in certain contexts, and Bayesian inference often advocates the use of more general classes of loss function for model evaluation – see Schorfheide (2000) for a recent discussion and promotion of such an approach. For instance, one potential criticism of ranking models based only on statistical fit and parameter parsimony, is that such criterion could well give rise to perverse policy recommendations if the selected model fails to capture important components for the transmission mechanism in the case of monetary policy. However, this will in general be true for any proposed criterion for ranking models, whether decision theoretic or purely statistical, and only serves to emphasize that analysis of this kind is never meant to supersede sound economic reasoning.

What this paper has attempted to treat carefully, given a plausible set of models, finite sample and a particular set of assumptions, is whether models can be statistically distinguished. It remains as further work to analyze the statistical properties of general classes of decision theoretic approaches to model selection and whether said approaches resolve the inconsistency of posterior odds ratios under our assumptions.

## References

- ANDREWS, D. (1994): “Empirical Process Methods in Econometrics,” in *Handbook of Econometrics, Vol. 4*, ed. by R. Engle, and D. McFadden, pp. 2248–2292. North Holland.

- ANDREWS, D. (1999): “Consistent Moment Selection Procedures for Generalized Method of Moments Estimation,” *Econometrica*, 67, 543–564.
- ANDREWS, D., AND B. LU (2001): “Consistent model and moment selection procedures for GMM estimation with application to dynamic panel data models,” *Journal of Econometrics*, 101, 123–164.
- ATTANASIO, O., AND H. LOW (forthcoming): “Estimating Euler Equations,” *Review of Economic Dynamics*.
- BERRY, S., J. LEVINSOHN, AND A. PAKES (1995): “Automobile Prices in Market Equilibrium,” *Econometrica*, 63, 841–890.
- BUNKE, O., AND X. MILHAUD (1998): “Asymptotic Behavior of Bayes Estimates under Possibly Incorrect Models,” *The Annals of Statistics*, 26(2), 617–644.
- CARROLL, C. D. (2001): “Death to the Log-Linearized Consumption Euler Equation! (And Very Poor Health to the Second-Order Approximation),” *Advances in Macroeconomics*.
- CHERNOZHUKOV, V., AND C. HANSEN (2005): “An IV Model of Quantile Treatment Effects,” *Econometrica*, 73(1), 245–261.
- CHERNOZHUKOV, V., AND H. HONG (2003): “A MCMC Approach to Classical Estimation,” *Journal of Econometrics*, 115(2), 293–346.
- FAN, Y. (1994): “Testing the Goodness of Fit of a Parametric Density Function by Kernel Method,” *Econometric Theory*, 10, 316–356.
- FERNANDEZ-VILLAYERDE, J., AND J. F. RUBIO-RAMIREZ (2003): “Estimating Nonlinear Dynamic Equilibrium Economies: Linear versus Nonlinear Likelihood,” unpublished, University of Pennsylvania.
- (2004a): “Comparing Dynamic Equilibrium Models to Data: A Bayesian Approach,” *Journal of Econometrics*, 123, 153–187.
- (2004b): “Estimating Nonlinear Dynamic Equilibrium Economies: A Likelihood Approach,” University of Pennsylvania, PIER Working Paper 04-001.
- GOURIEROUX, C., AND A. MONFORT (1995): *Statistics and Econometric Models*. Cambridge University Press.
- HANSEN, L. (1982): “Large Sample Properties of Generalized Method of Moments Estimators,” *Econometrica*, 50(4), 1029–1054.
- HANSEN, P. R. (2003): “A Test for Superior Predictive Ability,” manuscript, Stanford University.
- HONG, H., B. PRESTON, AND M. SHUM (2003): “Generalized Empirical Likelihood-Based Model Selection Criteria for Moment Condition Models,” *Econometric Theory*, 19, 923–943.

- JUSTINIANO, A., AND B. PRESTON (2004): “New Open Economy Macroeconomics and Imperfect Pass-through: An Empirical Analysis,” unpublished, Columbia University and International Monetary Fund.
- KITAMURA, Y. (2002): “A Likelihood-based Approach to the Analysis of a Class of Nested and Non-nested Models,” Department of Economics, Yale University.
- LUBIK, T. A., AND F. SCHORFHEIDE (2003): “Do Central Banks Respond to Exchange Rate Movements? A Structural Investigation,” unpublished, Johns Hopkins University and University of Pennsylvania.
- NEWBY, W. (1985): “Maximum Likelihood Specification Testing and Conditional Moment Tests,” *Econometrica*, pp. 1047–1070.
- NEWBY, W., AND D. MCFADDEN (1994): “Large Sample Estimation and Hypothesis Testing,” in *Handbook of Econometrics, Vol. 4*, ed. by R. Engle, and D. McFadden, pp. 2113–2241. North Holland.
- PAKES, A., AND D. POLLARD (1989): “Simulation and the Asymptotics of Optimization Estimators,” *Econometrica*, 57(5), 1027–1057.
- PAXSON, C. H., AND S. LUDVIGSON (1999): “Approximation Bias in Euler Equation Estimation,” NBER Working Paper No. T0236.
- POLLARD, D. (1991): “Asymptotics for Least Absolute Deviation Regression Estimator,” *Econometric Theory*, 7, 186–199.
- POWELL, J. L. (1986): “Censored Regression Quantiles,” *Journal of Econometrics*, 32, 143–155.
- SCHORFHEIDE, F. (2000): “Loss Function-Based Evaluation of DSGE Models,” *Journal of Applied Econometrics*, 15, 645–670.
- SIMS, C. (2001): “Time Series Regression, Schwartz Criterion,” Lecture Note, Princeton University.
- (2003): “Probability Models for Monetary Policy Decisions,” unpublished, Princeton University.
- SIN, C.-Y., AND H. WHITE (1996): “Information Criteria for Selecting Possibly Misspecified Parametric Models,” *Journal of Econometrics*, 71, 207–225.
- SMETS, F., AND R. WOUTERS (2002): “An Estimated Dynamics Stochastic General Equilibrium Model of the Economy,” National Bank of Belgium, Working Paper No. 35.
- STOCK, J. H., AND M. W. WATSON (2001): “Forecasting Output and Inflation: The Role of Asset Prices,” NBER Working Paper 8180.
- (2003): “Combination Forecasts of Output Growth in a Seven-country Data Set,” unpublished, Princeton University.
- VUONG, Q. (1989): “Likelihood-ratio tests for model selection and non-nested hypotheses,” *Econometrica*, pp. 307–333.

WHITE, H. (2000): “A Reality Check for Data Snooping,” *Econometrica*, 68, 1097–1126.

WRIGHT, J. H. (2003): “Bayesian Model Averaging and Exchange Rate Forecasts,” Board of Governors of the Federal Reserve System, International Finance Discussion Papers, No. 779.

## A Proof of Theorem 1

It has been shown (Pakes and Pollard (1989), Newey and McFadden (1994) and Andrews (1994)) that under assumption (1),

$$\sqrt{T} \left( \hat{\beta} - \beta_0 \right) = -\mathcal{A}_\beta \frac{\Delta_T}{\sqrt{T}} + o_p(1).$$

Combined with assumption (3), which states that

$$\hat{Q} \left( \hat{\beta} \right) - \hat{Q} \left( \beta_0 \right) = \sqrt{T} \left( \hat{\beta} - \beta_0 \right)' \frac{\Delta_T}{\sqrt{T}} - \frac{1}{2} \sqrt{T} \left( \hat{\beta} - \beta_0 \right)' (J_T) \sqrt{T} \left( \hat{\beta} - \beta_0 \right) + o_p(1),$$

because of  $R_T \left( \hat{\beta} \right) = o_p(1)$ , we can write

$$\hat{Q} \left( \hat{\beta} \right) - \hat{Q} \left( \beta_0 \right) = -\frac{1}{2} \frac{\Delta_T'}{\sqrt{T}} \mathcal{A}_\beta \frac{\Delta_T}{\sqrt{T}} + o_p(1).$$

The conclusion then follows from the assumption that  $\frac{\Delta_T}{\sqrt{T}} = O_p(1)$ . ■

## B Proof of Theorem 3

Define  $\tilde{\beta}_T = \beta_0 - \frac{1}{T} \mathcal{A}_\beta^{-1} \Delta_T$ , and define  $h = \sqrt{T} \left( \beta - \tilde{\beta}_T \right)$ . Then through a change of variables, we can write

$$T^{\frac{\dim(\beta)}{2}} \int e^{\hat{Q}(\beta)} \pi(\beta) d\beta = \int e^{\hat{Q}\left(\frac{h}{\sqrt{T}} + \tilde{\beta}_T\right)} \pi\left(\frac{h}{\sqrt{T}} + \tilde{\beta}_T\right) dh.$$

Chernozhukov and Hong (2003) has shown that (see equation A5 of p326):

$$\begin{aligned} & \int e^{\hat{Q}\left(\frac{h}{\sqrt{T}} + \tilde{\beta}_T\right) - \hat{Q}(\beta_0) + \frac{1}{2T} \Delta_T' \mathcal{A}_\beta^{-1} \Delta_T} \pi\left(\frac{h}{\sqrt{T}} + \tilde{\beta}_T\right) dh \\ & \xrightarrow{p} \pi(\beta_0) (2\pi)^{\frac{\dim(\beta)}{2}} \det(-\mathcal{A}_\beta)^{-1/2}. \end{aligned}$$

Therefore the proof for theorem 3 will be completed if one can show that

$$\hat{Q} \left( \hat{\beta} \right) - \left( \hat{Q} \left( \beta_0 \right) - \frac{1}{2T} \Delta_T' \mathcal{A}_\beta^{-1} \Delta_T \right) \xrightarrow{p} 0,$$

where  $\hat{\beta}$  is the conventional  $M$  estimator, defined as (see Pakes and Pollard (1989)),

$$\hat{Q}(\hat{\beta}) = \inf_{\beta} \hat{Q}(\beta) + o_p(T^{-1/2}).$$

It has been shown (Pakes and Pollard (1989), Newey and McFadden (1994) and Andrews (1994) that under assumptions (1), (2) and (3),

$$\sqrt{T}(\hat{\beta} - \beta_0) = -\mathcal{A}_{\beta} \frac{\Delta_T}{\sqrt{T}} + o_p(1).$$

Substituting this into assumption (3), together with  $\Delta_T/\sqrt{T} = O_p(1)$  and  $J_T = O_p(1)$  we immediately see that

$$R(\hat{\beta}) = \hat{Q}(\hat{\beta}) - \hat{Q}(\beta_0) + \frac{\Delta'_T}{\sqrt{T}} \mathcal{A}_{\beta} \frac{\Delta_T}{\sqrt{T}} - \frac{1}{2} \frac{\Delta'_T}{\sqrt{T}} \mathcal{A}_{\beta} \frac{\Delta_T}{\sqrt{T}} + o_p(1).$$

Therefore the desired conclusion follows from the assumption that  $R(\hat{\beta}) = o_p(1)$ . ■

### C Proof of Theorem 4

Define  $\hat{\psi}(\bar{y}, Y_T) = \sqrt{T}(\hat{\lambda}(\bar{y}, Y_T) - \tilde{\lambda}(\bar{y}, Y_T))$  so that  $\hat{\lambda}(\bar{y}, Y_T) = \tilde{\lambda}(\bar{y}, Y_T) + \hat{\psi}(\bar{y}, Y_T)/\sqrt{T}$ . By definition,  $\hat{\psi}(\bar{y}, Y_T)$  minimizes the following loss function with respect to  $\psi$ ,

$$\int \bar{\rho}\left(\bar{y}, \tilde{\lambda}(\bar{y}, Y_T) + \frac{\psi}{\sqrt{T}}; \beta\right) f(\beta|Y_T) d\beta.$$

Also define  $h \equiv \sqrt{T}(\beta - \hat{\beta})$  as the localized parameter space around  $\hat{\beta}$ . The implied density for localized parameter  $h$  is given by

$$\xi(h) = \left(\frac{1}{\sqrt{T}}\right)^{\dim(\beta)} f\left(\hat{\beta} + \frac{h}{\sqrt{T}}|Y_T\right).$$

Then  $\hat{\psi}(\bar{y}, Y_T)$  also minimizes the equivalent loss function of

$$Q_T(\psi) = \int \bar{\rho}\left(\bar{y}, \tilde{\lambda} + \frac{\psi}{\sqrt{T}}; \hat{\beta} + \frac{h}{\sqrt{T}}\right) \xi(h) dh$$

where we are using the shorthand notations  $\hat{\lambda} = \hat{\lambda}(\bar{y}, Y_T)$  and  $\tilde{\lambda} = \tilde{\lambda}(\bar{y}, Y_T)$ . For a given  $\psi$ , we are interested in the asymptotic behavior of  $Q_T(\psi)$  as  $T \rightarrow \infty$ . Define

$$\bar{Q}_T(\psi) = \bar{\rho}\left(\bar{y}, \tilde{\lambda} + \frac{\psi}{\sqrt{T}}; \hat{\beta}\right) + \int \bar{\rho}\left(\bar{y}, \tilde{\lambda}; \hat{\beta} + \frac{h}{\sqrt{T}}\right) \xi(h) dh - \bar{\rho}\left(\bar{y}, \tilde{\lambda}; \hat{\beta}\right). \quad (\text{C.3})$$

Essentially,  $\bar{Q}_T(\psi)$  is a first order approximation to  $Q_T(\psi)$ . Under the assumptions stated in Theorem 4, it can be shown that for each  $\psi$ ,<sup>2</sup>

$$T [Q_T(\psi) - \bar{Q}_T(\psi)] \xrightarrow{p} 0.$$

Because  $Q_T(\psi)$  and  $\bar{Q}_T(\psi)$  are both convex in  $\psi$ , and since  $\bar{Q}_T(\psi)$  is uniquely minimized at  $\psi \equiv 0$ , the convexity lemma (e.g. Pollard (1991)) is used to deliver the desired result that

$$\hat{\psi} = \sqrt{T} \left( \hat{\lambda}(\bar{y}, Y_T) - \tilde{\lambda}(\bar{y}, Y_T) \right) \xrightarrow{p} 0.$$

An alternative proof can be based on a standard Taylor expansion of the first order conditions that define  $\hat{\lambda}(\bar{y}, Y_T)$  and  $\tilde{\lambda}(\bar{y}, Y_T)$ . This is straightforward but the notations will be more complicated. End of proof of theorem 4. ■

## D Proof of Theorem 5

It is clear from Theorem 3 and its following discussions that under either one of the stated conditions,

$$w_Q \equiv \frac{BF_Q}{BF_Q + BF_L} \xrightarrow{p} 1 \quad \text{as } T \rightarrow \infty.$$

It is because from Theorem 3, for constants

$$C_Q = P_Q \pi(\beta_0) (2\pi)^{\dim(\beta)/2} \det(-\mathcal{A}_\beta)^{-1/2}$$

and

$$C_L = P_L \pi(\alpha_0) (2\pi)^{\dim(\alpha)/2} \det(-\mathcal{A}_\alpha)^{-1/2}$$

we can write

$$\begin{aligned} 1 - w_Q &= \frac{C_L e^{\hat{L}(\hat{\alpha})} T^{-\dim(\alpha)/2} (1 + o_p(1))}{C_Q e^{\hat{Q}(\hat{\beta})} T^{-\dim(\beta)/2} (1 + o_p(1)) + C_L e^{\hat{L}(\hat{\alpha})} T^{-\dim(\alpha)/2} (1 + o_p(1))} \\ &= \frac{(1 + o_p(1)) \frac{C_L}{C_Q} e^{\hat{L}(\hat{\alpha}) - \hat{Q}(\hat{\beta})} T^{\frac{d_\beta - d_\alpha}{2}}}{(1 + o_p(1)) \left( 1 + \frac{C_L}{C_Q} e^{\hat{L}(\hat{\alpha}) - \hat{Q}(\hat{\beta})} T^{\frac{d_\beta - d_\alpha}{2}} \right)}. \end{aligned}$$

---

<sup>2</sup>As  $T \rightarrow \infty$ ,  $\xi(h)$  converges in a strong total variation norm in probability to  $\phi(h; -\mathcal{A}_\beta^{-1})$ , the multivariate normal density with mean 0 and variance  $-\mathcal{A}_\beta^{-1}$ . In fact, the proof of Theorem 3 shows that

$$\int h^\alpha \xi(h) dh = O_p(1),$$

for all  $\alpha \geq 0$ . This, combined with the stated assumption that the differentiability of  $\bar{\rho}(\bar{Y}, \lambda; \beta)$  with respect to  $\lambda$  and  $\beta$ , implies the stated convergence in probability.

While  $w_Q \xrightarrow{p} 1$  under the stated conditions, the specific rate of convergence depends on the specific condition stated in Theorem 3. Under condition 1, It is clear that  $\exists \delta > 0$  such that with probability converging to 1, for all  $T$  large enough,

$$1 - w_Q < e^{-T\delta}. \quad (\text{D.4})$$

On the other hand, when condition 2 holds, we know that  $\hat{L}(\hat{\alpha}) - \hat{Q}(\hat{\beta})$  converges to a quadratic norm of a normal distribution. It can then be shown that

$$T^{\frac{d\alpha-d\beta}{2}} (1 - w_Q) \xrightarrow{d} \frac{C_L}{C_Q} \exp(\bar{\chi}^2), \quad (\text{D.5})$$

where  $\bar{\chi}^2$  is typically distributed as the quadratic form of a random vector such as described in Vuong (1989).

Using the definition of  $\bar{\rho}(\bar{y}, \lambda; \beta)$ , and define  $\bar{\rho}(\bar{y}, \lambda; \alpha)$  similarly, we can write  $\hat{\lambda}(\bar{y}, Y_T)$  as the minimizer with respect to  $\lambda$  of

$$w_Q \int \bar{\rho}(\bar{y}, \lambda; \beta) f(\beta|Y_T, Q) d\beta + (1 - w_Q) \int \bar{\rho}(\bar{y}, \lambda; \alpha) f(\alpha|Y_T, L) d\alpha.$$

As before, define  $\hat{\psi}(\bar{y}, Y_T) = \sqrt{T} \left( \hat{\lambda}(\bar{y}, Y_T) - \tilde{\lambda}(\bar{y}, Y_T) \right)$  and define  $h \equiv \sqrt{T} \left( \beta - \hat{\beta} \right)$ . Then  $\hat{\psi}(\bar{y}, Y_T)$  equivalently minimizes, with respect to  $\psi$ ,

$$Q_T(\psi) = w_Q Q_T^1(\psi) + (1 - w_Q) Q_T^2(\psi)$$

where, with  $\tilde{\lambda} \equiv \tilde{\lambda}(\bar{y}, Y_T)$ ,

$$Q_T^1(\psi) = \int \bar{\rho} \left( \bar{y}, \tilde{\lambda} + \frac{\psi}{\sqrt{T}}; \hat{\beta} + \frac{h}{\sqrt{T}} \right) \xi(h) dh,$$

and

$$Q_T^2(\psi) = \int \bar{\rho} \left( \bar{y}, \tilde{\lambda} + \frac{\psi}{\sqrt{T}}; \alpha \right) f(\alpha|Y_T, L) d\alpha.$$

Now recall the definition of  $\bar{Q}_T(\psi)$  in equation (C.3) in the proof of theorem 4. Also define

$$\tilde{Q}_T(\psi) = w_Q \bar{Q}_T(\psi) + (1 - w_Q) \bar{\rho}(\bar{y}, \tilde{\lambda}; \hat{\alpha}).$$

We are going to show that with this definition

$$T \left( Q_T(\psi) - \tilde{Q}_T(\psi) \right) \xrightarrow{p} 0. \quad (\text{D.6})$$

If (D.6) holds, it then follows again from the convexity lemma of Pollard (1991) and the fact that  $\tilde{Q}_T(\psi)$  is uniquely optimized at  $\psi = 0$  that

$$\hat{\psi}(\bar{y}, Y_T) = \sqrt{T} \left( \hat{\lambda}(\bar{y}, Y_T) - \tilde{\lambda}(\bar{y}, Y_T) \right) \xrightarrow{p} 0.$$

Finally, we will verify (D.6). With the definition of  $\tilde{Q}_T(\psi)$ , we can write

$$T \left( Q_T(\psi) - \tilde{Q}_T(\psi) \right) = T w_Q \left( Q_T^1(\psi) - \bar{Q}_T(\psi) \right) + T(1 - w_Q) \left( Q_T^2(\psi) - \bar{\rho}(\bar{y}, \tilde{\lambda}; \hat{\alpha}) \right).$$

Because  $w_Q \xrightarrow{p} 1$ , it follows from Theorem 4 that  $w_Q T \left( Q_T^1(\psi) - \bar{Q}_T(\psi) \right) \xrightarrow{p} 0$ . As the sample size increases,  $f(\alpha|Y_T, L)$  tends to concentrate on  $\hat{\alpha}$ , therefore it can also be shown that

$$Q_T^2(\psi) - \bar{\rho}(\bar{y}, \tilde{\lambda}; \hat{\alpha}) \xrightarrow{p} 0.$$

Now if either condition 1 holds or if condition 2 holds and  $d_\alpha - d_\beta > 2$ , then because of (D.4) and (D.5),  $T(1 - w_Q) \xrightarrow{p} 0$  and the second term vanishes in probability. Finally, in the last case where  $d_\alpha = d_\beta + 2$  under condition 2, we know from equation (D.5) that

$$T(1 - w_Q) \xrightarrow{d} \frac{C_L}{C_Q} \exp(\bar{\chi}^2) = O_p(1). \quad (\text{D.7})$$

Hence it is also true that  $T(1 - w_Q) \left( Q_T^2(\psi) - \bar{\rho}(\bar{y}, \tilde{\lambda}; \hat{\alpha}) \right) = o_p(1)$ . Therefore (D.6) holds.  $\blacksquare$

Remark: It also follows from the same arguments as above that the results of Theorem 5 does not hold when  $d_\alpha = d_\beta + 1$ . In fact, in this case, we can redefine

$$\tilde{Q}_T(\psi) = w_Q \bar{Q}_T(\psi) + (1 - w_Q) \left[ \bar{\rho}(\bar{y}, \tilde{\lambda}; \hat{\alpha}) + \frac{\partial}{\partial \lambda} \bar{\rho}(\bar{y}, \tilde{\lambda}; \hat{\alpha}) \frac{\psi}{\sqrt{T}} \right].$$

We can then follow the same logic as before to show that

$$T \left( Q_T(\psi) - \tilde{Q}_T(\psi) \right) \xrightarrow{p} 0.$$

Note that in the definition of  $\bar{Q}_T(\psi)$  in equation (C.3) of Theorem 4, using a second order Taylor expansion we can replace  $\bar{\rho}(\bar{y}, \tilde{\lambda} + \frac{\psi}{\sqrt{T}}; \hat{\beta})$  by

$$\frac{1}{T} \psi' \frac{1}{2} \bar{\rho}_{\lambda\lambda}(\bar{y}, \tilde{\lambda}; \hat{\beta}) \psi + o_p\left(\frac{1}{T}\right).$$

As such we can write

$$\begin{aligned} & T \left( Q_T(\psi) - w_Q \left( \int \bar{\rho} \left( \bar{y}, \tilde{\lambda}; \hat{\beta} + \frac{h}{\sqrt{T}} \right) \xi(h) dh - \bar{\rho}(\bar{y}, \tilde{\lambda}; \hat{\beta}) \right) - (1 - w_Q) \bar{\rho}(\bar{y}, \tilde{\lambda}; \hat{\alpha}) \right) \\ &= \frac{1}{2} \psi' \bar{\rho}_{\lambda\lambda}(\tilde{\lambda}; \hat{\beta}) \psi + \sqrt{T} (1 - w_Q) \frac{\partial}{\partial \lambda} \bar{\rho}(\bar{y}, \tilde{\lambda}; \hat{\alpha}) \psi + o_p(1). \end{aligned}$$

It follows from both (D.7) and the convergence of  $\frac{\partial}{\partial \lambda} \bar{\rho}(\bar{y}, \tilde{\lambda}; \hat{\alpha}) \xrightarrow{p} \frac{\partial}{\partial \lambda} \bar{\rho}(\bar{y}, \lambda_0; \alpha_0)$  that

$$\sqrt{T} (1 - w_Q) \frac{\partial}{\partial \lambda} \bar{\rho}(\bar{y}, \tilde{\lambda}; \hat{\alpha}) \xrightarrow{d} \frac{C_L}{C_Q} \exp(\bar{\chi}^2) \frac{\partial}{\partial \lambda} \bar{\rho}(\bar{y}, \lambda_0; \alpha_0)'$$

where  $\lambda_0 = \arg \min_{\lambda} \bar{\rho}(\bar{y}, \lambda; \beta_0)$ . Hence again with convexity arguments for uniform convergence we can show that

$$\begin{aligned} \hat{\psi}(\bar{y}, Y_T) &\xrightarrow{d} \arg \min_{\psi} \frac{1}{2} \psi' \bar{\rho}_{\lambda\lambda}(\lambda_0; \beta_0) \psi + \frac{C_L}{C_Q} \exp(\bar{\chi}^2) \frac{\partial}{\partial \lambda} \bar{\rho}(\bar{y}, \lambda_0; \alpha_0)' \psi \\ &= -\bar{\rho}_{\lambda\lambda}(\lambda_0; \beta_0)^{-1} \frac{C_L}{C_Q} \exp(\bar{\chi}^2) \frac{\partial}{\partial \lambda} \bar{\rho}(\bar{y}, \lambda_0; \alpha_0)'. \end{aligned}$$

In other words, if  $d_{\beta} = d_{\alpha} - 1$ ,  $\sqrt{T} \left( \hat{\lambda}(\bar{y}, Y_T) - \tilde{\lambda}(\bar{y}, Y_T) \right)$  converges in distribution to a nondegenerate random variable.

## E Proof of proposition 1

(Sketch:) Define  $\lambda(\beta) = \arg \max_{\lambda \in \Lambda} -Q(\beta, \lambda)$  and  $\hat{\lambda}(\beta) = \arg \max_{\lambda \in \Lambda} -\hat{Q}(\beta, \lambda)$ . For given  $\beta$ , using the same reasonings as in theorem 3, it can be shown that

$$T^{\frac{\dim(\lambda)}{2}} \int e^{\hat{Q}(\beta, \hat{\lambda}(\beta)) - \hat{Q}(\beta, \lambda)} \phi(\lambda) d\lambda \xrightarrow{p} \phi(\lambda(\beta)) (2\pi)^{\frac{\dim(\lambda)}{2}} \det(A_{\lambda}(\beta))^{-1/2},$$

where,

$$A_{\lambda}(\beta) = \frac{\partial^2 Q(\beta, \lambda)}{\partial \lambda \partial \lambda'} \Big|_{\lambda=\lambda(\beta)}.$$

This convergence can be shown to be uniform in  $\beta$  under suitable regularity conditions. Taking this as given, write

$$\text{GELBF} = T^{\frac{\dim(\lambda)}{2}} \int e^{\hat{Q}(\beta, \hat{\lambda}(\beta))} \left[ \phi(\lambda(\beta)) (2\pi)^{\frac{\dim(\lambda)}{2}} \det(A_{\lambda}(\beta))^{-1/2} \right]^{-1} e^{o_p(1)} \pi(\beta) d\beta.$$

To simplify notation, define

$$\bar{\pi}(\beta) = \left[ \phi(\lambda(\beta)) (2\pi)^{\frac{\dim(\lambda)}{2}} \det(A_{\lambda}(\beta))^{-1/2} \right]^{-1} \pi(\beta),$$

and rewrite the second-to-last expression as

$$\text{GELBF} = T^{\frac{\dim(\lambda)}{2}} e^{o_p(1)} \int e^{\hat{Q}(\beta, \hat{\lambda}(\beta))} \bar{\pi}(\beta) d\beta.$$

Applying the results from theorem 3 yields

$$\text{GELBF} = T^{\frac{\dim(\lambda)}{2}} e^{\rho_p(1)} e^{\hat{Q}(\hat{\beta}, \hat{\lambda}(\hat{\beta}))} T^{-\frac{\dim(\beta)}{2}} \bar{\pi}(\beta_0) \det(-\mathcal{A}_\beta)^{-\frac{1}{2}} e^{\rho_p(1)},$$

where now, because of the envelope theorem  $\left. \frac{\partial \bar{Q}(\beta, \lambda)}{\partial \lambda} \right|_{\lambda=\lambda(\beta)} = 0$  for all  $\beta$ ,

$$\mathcal{A}_\beta = - \left. \frac{\partial^2 Q(\beta_0, \lambda)}{\partial \beta \partial \beta'} \right|_{\lambda=\lambda(\beta_0)}.$$

In other words,

$$\log \text{GELBF} = \hat{Q}(\hat{\beta}, \hat{\lambda}(\hat{\beta})) + \frac{1}{2} (\dim(\lambda) - \dim(\beta)) \log T + C(\beta),$$

where  $C(\beta)$  is a constant related to  $\mathcal{A}_\beta$  and  $\bar{\pi}(\beta_0)$ .