Bayesian Averaging, Prediction and Nonnested Model Selection

Han Hong and Bruce Preston[1]

Previous version: January 2006

This version: August 2008

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.

JEL Classification: C14; C52
Keywords: Model selection criteria, Nonnested, Posterior odds, BIC

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 (2004a), Schorfheide (2000), Stock and Watson (2001), Timmermann (2005), Clark and McCracken (2006) and Wright (2003). These methods are 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 admits the possiblity of model misspecification .

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 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 $O_p\left(1/\sqrt{T}\right)$.

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 smaller order of magnitude than the inverse of the square root of the sample size in large samples. These findings essentially combine Schwarz' 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 Bayesian posterior odds comparison is inconsistent for selecting between nonnested 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).

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). 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 from 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 dy-

namic 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 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 discusses the implications for our results for non-likelihood-based models, and section 7 concludes.

2 Model assumptions and implications

For clarity of exposition, in the rest of the paper 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 $f(y_t, \beta)$ 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

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

which minimizes the Kullback-Leibler distance between the parametric model and the data.

Under standard assumptions, the random objective function converges to a population limit when the sample size increases without bound. It is therefore 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 $\hat{Q}(\beta)$:

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

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*

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

ASSUMPTION 3 *There exist quantities 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), (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. This distribution is useful for model selection tests but is not directly used in model selection criteria developed in section 4. In particular, satisfaction of the information matrix equality, $\Omega_T = -\mathcal{A}_\beta + o_p(1)$, is not necessary for our discussion of model selection criteria. This is especially relevant under potential model misspecification.

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

Assumption 4 typically follows from an application of the central limit theorem

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

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, where

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

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

Beyond the likelihood setting, assumption 4 can in general be easily verified for extreme estimators based on optimizating random objective functions. These include M estimator, generalized method of moment estimators and their recent information theoretic variants. It can be shown to hold for generalized method of moment estimators regardless of whether the model is correctly specified or misspecified.

The property of the final term (Qc) is immediate.

3 Bayesian predictive analysis

Consider the 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 β given the data Y_T:

$$f\left(y_{T+1}|Y_T\right) = \int f\left(y_{T+1}|Y_T, \beta\right) f\left(\beta|Y_T\right) d\beta$$

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

We will assume that $f\left(y_{T+1}|Y_T; \beta\right) = f\left(y_{T+1}|\bar{Y}_T; \beta\right)$, 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 $f\left(y_{T+1}|\bar{Y}_T, \beta\right)$. The length of the prediction interval comprises two parts: the first is due to $f\left(y_{T+1}|\bar{Y}_T, \beta\right)$ and the second is due to the uncertainty from estimating β. 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 β, and therefore will consider a fixed value $\bar{y}$ of the random component $\bar{Y}_T$, and consider

$$f\left(y_{T+1}|\bar{y}, Y_T\right) = \int f\left(y_{T+1}|\bar{y}, \beta\right) f\left(\beta|Y_T\right) d\beta, \tag{3.2}$$

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 $f\left(y_{T+1}|\bar{y}, Y_T\right)$. For example, a mean prediction can be obtained by

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

A median prediction is given by

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

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

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

This asymptotic approximation of the posterior distribution of β by a normal distribution is the key element of the results formally developed in section 4, and forms the basis of the following asymptotic approximation of the predictive distribution.

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

$$\int f(y_{T+1}|\bar{y}, \beta)(2\pi)^{-\frac{\dim(\beta)}{2}} \det(-T\mathcal{A}_\beta)^{\frac{1}{2}} e^{-\frac{1}{2}(\beta-\hat{\beta})'(-T\mathcal{A}_\beta)(\beta-\hat{\beta})} d\beta$$

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

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

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

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

If $E(y_{T+1}|\bar{y}; \beta)$ is linear in β, 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

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

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 $\hat{\beta}$, $f\left(y_{T+1}|\bar{y}, \hat{\beta}\right)$.

And even if $E(y_{T+1}|\bar{y}; \beta)$ is not linear in β, as long as it is sufficiently smooth in β, it is still possible to use a first-order Taylor expansion of $E\left(y_{T+1}|\bar{y}; \hat{\beta} + \frac{h}{\sqrt{T}}\right)$ 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},

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

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 non-smooth 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 $\rho(\cdot)$, which is assumed to be convex:

$$\hat{\lambda}(\bar{y}, Y_T) = \arg\min_{\lambda} E\left(\rho(y_{T+1}, \lambda) | \bar{y}, Y_T\right)$$

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

where the predictive density $f(y_{T+1} | \bar{y}, Y_T)$ is defined in equation (3.2). For example, to construct a one-sided τth predictive interval, we can take

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

The following focuses on this loss function.

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

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

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

$$\bar{\rho}(\bar{y}, \lambda; \beta) = \int \rho(y_{T+1}, \lambda) f(y_{T+1} | \bar{y}, \beta) \, dy_{T+1}$$

$$= E\left(\rho(y_{T+1}, \lambda) | \bar{y}, \beta\right).$$

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

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

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

THEOREM 2 *Under assumptions (1), (2) and (3), and assuming that for each $\bar{y}$, $\bar{\rho}(\bar{y}, \lambda; \beta)$ is three times continuously differentiable in β and λ in a small neighborhood of β_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 λ and β, 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 β is smoothly differentiable around the percentiles 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 $O_p\left(1/\sqrt{T}\right)$. 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 $o_p\left(1/\sqrt{T}\right)$. 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, rather than to define the estimator itself. Note also that we consider only prediction at a fixed value and have not yet considered out of sample variation. The recent work of Hansen (2007) discovered important negative correlation between in-sample and out-of-sample variations of predictive criterion functions.

4   Posterior odds and consistent model selection

Bayesian averaging is a popular method for making predictions in the context of multiple models. The weights used in Bayesian averaging are calculated through posterior odds ratios. To generalize the results in the previous section to multiple models, it is important to understand first the relation between posterior odds ratio calculation and consistent frequentist model selection criteria. This is of independent 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.

4.1   Large sample properties of Bayes factors

The posterior distribution is an integral transformation of the model (f, Q), defined as

$$e^{\hat{Q}(\beta)} \pi(\beta) \left/ \int e^{\hat{Q}(\beta')} \pi(\beta')\, d\beta'. \right.$$

Associated with the posterior distribution is the Bayes factor

$$P_Q \times \int e^{\hat{Q}(\beta)} \pi(\beta)\, d\beta,$$

where P_Q is a prior probability weight of model (f, Q).

The following theorem connects the properties of the Bayes posterior distribution to the extremum estimator and establishes the relation between the Bayes factor and the extremum estimator analog of BICs. Under the regularity conditions stated in assumptions 1 to 3, the Bayes factor is asymptotically equivalent to using BIC as a model selection criterion. While the relation between Bayes factor and BIC comes from the original contribution by Schwartz, the conditions in the following theorem are considerably weaker and more general.

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

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

The formal details of the proof are relegated to an appendix. When $\hat{Q}(\beta)$ is smoothly differentiable, intuition can be gleaned by considering the expression:

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

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

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

Second, in this $1/\sqrt{T}$ size neighborhood, the prior density is approximately constant around β_0 as $\pi\left(\hat{\beta}\right) \xrightarrow{p} \pi(\beta_0)$. The impact of the prior density is negligible except for the value at $\pi(\beta_0)$. 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:

$$\log \int e^{\hat{Q}(\beta)} \pi(\beta) \, d\beta - \hat{Q}\left(\hat{\beta}\right) = \log \pi(\beta_0) \int e^{\frac{1}{2}(\beta - \hat{\beta})' \frac{\partial^2 \hat{Q}_T(\hat{\beta})}{\partial \beta \partial \beta'}(\beta - \hat{\beta})} d\beta + o_p(1)$$

$$= \log \pi(\beta_0) \int e^{\frac{1}{2}(\beta - \hat{\beta})' T \mathcal{A}_\beta(\beta - \hat{\beta})} d\beta + o_p(1)$$

$$= \log\left(\pi(\beta_0)(2\pi)^{\frac{\dim(\beta)}{2}} \det\left(-T\mathcal{A}_\beta\right)^{-\frac{1}{2}}\right) + o_p(1)$$

$$= \underbrace{\log \pi(\beta_0) + \frac{\dim(\beta)}{2} \log(2\pi) - \frac{1}{2} \det\left(-\mathcal{A}_\beta\right)}_{C(\mathcal{A}_\beta, \beta)} - \frac{1}{2} \dim(\beta) \log T + o_p(1),$$

where $C(\mathcal{A}_\beta, \beta)$ can also depend on the prior weight on the model itself.

4.2 Consistent model selection

Now introduce a competing model, $g(Y; \alpha)$, $\alpha \in \Lambda$ (generalizing the notations to multiple models is immediate), where α is estimated by the random log-likelihood function:

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

Similarly to model (f, Q), we assume that the decomposition of

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

satisfies the stochastic order properties

$$(La) = O_p(1), \qquad (Lb) = O_p\left(\sqrt{T}\right), \qquad (Lc) = O(T).$$

We also assume that assumptions (1), (2), (3) and (4), and subsequently theorems 1, 2 and 3 hold for model (g, L).

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

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

where C_T is a sequence of constants that tends to ∞ as T goes to ∞, at a rate to be prescribed below, and $\dim()$ 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 $C_T = 2$.

While the discussion so far has focused on parametric likelihood models, all results can be applied without modification to other non-likelihood-based models, including method of moment models. They are also applicable to non-log-likelihood random criterion functions that measure other notions of distance between the model and the data generating process, such as the Cressie-Read discrepancy statistic in generalized empirical likelihood methods.

For non-likelihood-based models, in addition to penalizing the parametric dimension of the model, the dimension of the estimation procedure, such as the number of moment conditions could also be penalized as in Andrews and Lu (2001). 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.

For the specific case of the Bayesian (Schwartz) information criteria,

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

implicitly defining $C_T = \log T$. It now will be demonstrated that whether models are nested or nonnested has important consequences for the asymptotic properties of $\hat{Q}\left(\hat{\beta}\right) - \hat{L}\left(\hat{\alpha}\right)$ and consistency of the BIC as model selection criterion. The two cases of nested and nonnested models are treated in turn.

4.2.1   Nested Models

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

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

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

$$P\left(BIC > M\right) \longrightarrow 1.$$

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 $Q\left(\beta_0\right) = L\left(\alpha_0\right)$. 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 $\hat{Q}\left(\beta_0\right) = \hat{L}\left(\alpha_0\right)$ almost surely. In the case of likelihood models, this means that $f\left(Z_t; \beta_0\right) = g\left(Z_t; \alpha_0\right)$ almost surely, since $g\left(\cdot\right)$ is just a subset of $f\left(\cdot\right)$. The true model lies in the common subset of $\left(\beta_0, \alpha_0\right)$ and therefore has to be the same model.

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

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

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

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

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

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

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

4.2.2 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.

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 $C_T/\sqrt{T} \longrightarrow \infty$ is required in addition to $C_T = o\left(T\right)$. 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, that $Q\left(\beta_0\right)$ is greater than $L\left(\alpha_0\right)$, implying model $\left(f,\beta\right)$ is better than model $\left(g,\alpha\right)$. Then, as before, NIC is dominated by $T\left(Q\left(\beta_0\right) - L\left(\alpha_0\right)\right)$, 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.

However, when both models are equally good, so that $Q\left(\beta_0\right) = L\left(\alpha_0\right)$, 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 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 ∞ 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

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

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 Gourieroux and Monfort (1995)), the following theorem serves to summarize the above discussion about the NIC.

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

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

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

1. If $Q\left(\beta_0\right) > L\left(\alpha_0\right)$ then $P\left(NIC > M\right) \longrightarrow 1$ for all $M > 0$;

2. If $Q\left(\beta_0\right) = L\left(\alpha_0\right)$ and $\dim\left(f, Q\right) - \dim\left(g, L\right) > 0$ then $P\left(NIC > M\right) \longrightarrow 1$ for all $M > 0$.

4.3 Posterior odds and BIC

The posterior odds ratio between the two models is defined as

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

when the two models have prior probability weights P_Q and $P_L = 1 - P_Q$. Exploiting the approximation in section 4.1, up to a term of order $o_p\left(1\right)$, the log posterior odds ratio can be written as

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

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 $Q\left(\beta\right) > L\left(\alpha\right)$, that the posterior odds ratio will select model $\left(f, Q\right)$ with probability converging to 1 as T goes to infinity. Theorem 3 serves to generalize this finding. First, the results presented here are established under very weak regularity conditions. Second, only by considering a null hypothesis that admits the possible equivalence of the two models, in the sense that $Q\left(\beta\right) = L\left(\alpha\right)$, 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 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.

4.4 Multiple Models

The previous theorem can be immediately extended to the case of multiple models. 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.

However, 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. In contrast, if there exist two models that are not nested inside each other, then BIC will put random weights on at least two models. But NIC will still choose the most parsimonious model among all the k best models.

5   Bayesian model averaging and post-selection prediction

The results in the previous section can now be applied to study 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

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

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,

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

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

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

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

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 2 hold for both models* (f, Q) *and* (g, L). *Also assume that one of the following two conditions holds:*

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

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

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

It is clear from previous discussion that if $Q\left(\beta_0\right) = L\left(\alpha_0\right)$ 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 $f\left(y_{T+1}|\bar{y}, Y_T\right)$ 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, we offer the following remarks based on our findings

5.1 Models that are locally close

For forecasts generated from averages of a number of models to perform better than forecasts of any one of the underlying models, the Bayesian posterior weights have to be asymptotically non-degenerate among at least two of the models that are being averaged. From the previous analysis,

we have shown that this is possible among nonnested models when two of the models are "equally good" $(Q(\beta_0) = L(\alpha_0))$.

Our analysis also suggests that among nested models, unless their dimensions are essentially the same $(\dim(f, Q) = \dim(g, L))$, it is most likely that posterior weights will concentrate on one of the models, either the "best" model or the most parsimonious model. It turns out that it is still possible for the posterior weights to be non-degenerate among nested model when the two models are "just close enough" but "not too close". To see this, suppose that $Q(\beta_0) - L(\alpha_0) = h(T)$, where $h(T)$ is a function of the sample size T. Recall the decomposition for nested models:

$$\underbrace{(Qa) - (La)}_{O_p(1)} + \underbrace{T(Q(\beta_0) - L(\alpha_0))}_{T \times O(h(T))} - (\dim(f, Q) - \dim(g, L)) \times C_T.$$

As long as $T \times h(T)$ is approximately the same order as C_T, the penalization term will not dominate the second term, and it is then possible for posterior weights to be non-degenerate if the last two terms happen to approximately offset each other.

6   Generalized nonnested model selection criteria

While results in the previous sections are stated with the parametric log-likelihood function in mind, they apply without modification to a wide class of random non-likelihood-based objective functions assuming direct comparison between $Q(\beta)$ and $L(\alpha)$ is interpretable. Such interpretation is possible, for example, when both are constructed as information-theoretic empirical likelihoods corresponding to a different set of model and moment conditions.

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 χ^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.

For non-likelihood Q and L, the integral transformations

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

can be 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 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 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).

The following provides two specific examples of generalized posterior odds ratios constructed from non-likelihood random distance functions. They implicitly deliver a penalization term for the dimension of the information used in the estimation procedure, though all penalize the parametric dimension of the model.

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.

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:

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

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

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

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

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 proposes an information-theoretic alternative to efficient two-step method of moment estimators that minimizes a likelihood distance between the data and the model moments. A GEL estimator is defined as the saddle point of a GEL function,

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

For example, in the case of exponential tilting

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

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 mimics the model and moment selection criteria of Andrews and Lu (2001) and others. We define such a GEL Bayes factor as

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

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

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

indicates that λ 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 parameter space λ_0 and β_0 such that $T\hat{Q}\left(\hat{\beta}, \hat{\lambda}\right) - Q\left(\beta_0, \lambda_0\right) = O_p\left(1\right)$, where $Q\left(\beta, \lambda\right)$ is the uniform probably limit of $\hat{Q}\left(\beta, \lambda\right)/T$. Then*

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

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

7 Conclusion

This paper exploits the connections between Bayesian and classical predictions. 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. 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 model selection criteria can 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 model comparison methods 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.

It is also worth noting 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., 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.

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.

CLARK, T. E., AND M. W. MCCRACKEN (2006): "Combining Forecasts from Nested Models," working paper, Federal Reserve Bank of Kansas.

FAN, Y. (1994): "Testing the Goodness of Fit of a Parametric Density Function by Kernel Me thod," *Econometric Theory*, 10, 316–356.

FERNANDEZ-VILLAVERDE, J., AND J. F. RUBIO-RAMIREZ (2003): "Estimating Nonlinear Dynamic Equilibrium Economies: Linear versus Nonlinear Likelihood," unpublished, University of Pennsylvannia.

———— (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 Pennsylvannia, PIER Working Paper 04-001.

GOURIEROUX, C., AND A. MONFORT (1995): *Statistics and Econometric Models.* Cambridge University Press, Cambridge, UK.

HANSEN, L. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50(4), 1029–1054.

HANSEN, P. (2007): "In Sample and Out-of-Sample Fit," working paper,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 amd Imperfect Pass-through: An Emprical Analysis," unpublished, Columbia University and International Monetary Fund.

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.

NEWEY, W. (1985): "Maximum Likelihood Specification Testing and Conditional Moment Tests," *Econometrica*, pp. 1047–1070.

NEWEY, 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.

TIMMERMANN, A. (2005): "Forecast Combinations," in *Handbook of Economic Forecasting*, ed. by C. W. G. Graham Elliott, and A. Timmermann. North Holland.

VUONG, Q. (1989): "Likelihood-ratio tests for model selection and non-nested hypotheses," *Econometrica*, pp. 307–333.

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

A Proof of Theorem 1

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

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

Combined with assumption 3, this implies that

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

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

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

The conclusion follows from the assumptions that $\frac{\Delta_T}{\sqrt{T}} = O_p\left(1\right)$ and that $-\mathcal{A}_\beta$ is positive definite. ∎

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

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

Chernozhukov and Hong (2003) has shown that (equation A5 of p326) under the same set of assumptions:

$$\int e^{\hat{Q}\left(\frac{h}{\sqrt{T}}+\tilde{\beta}_T\right)-\hat{Q}(\beta_0)+\frac{1}{2T}\Delta_T'\mathcal{A}_\beta^{-1}\Delta_T}\pi\left(\frac{h}{\sqrt{T}}+\tilde{\beta}_T\right)dh$$

$$\xrightarrow{p}\pi\left(\beta_0\right)\left(2\pi\right)^{\frac{\dim(\beta)}{2}}\det\left(-\mathcal{A}_\beta\right)^{-1/2}.$$

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)),

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

This is indeed the conclusion of Theorem 3. ∎

C Proof of Theorem 2

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

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

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

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

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

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

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

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

Essentially, $\bar{Q}_T\left(\psi\right)$ is a first order approximation to $Q_T\left(\psi\right)$. Under the assumptions stated in Theorem 2, it can be shown that for each ψ, [2]

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

---

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

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

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

Because $Q_T(\psi)$ and $\bar{Q}_T(\psi)$ are both convex in ψ, 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 2. ∎

D Proof of Theorem 5

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

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

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

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

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,

$$1 - w_Q < e^{-T\delta}. \tag{D.4}$$

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

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

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 λ 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 ψ,

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

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 2. Also define

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

We are going to show that with this definition

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

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

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

Because $w_Q \xrightarrow{p} 1$, it follows from the proof of Theorem 2 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

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

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

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

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

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

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

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 2, using a second order Taylor expansion we can replace $\bar{\rho}\left(\bar{y}, \tilde{\lambda} + \frac{\psi}{\sqrt{T}}; \hat{\beta}\right)$ by

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

As such we can write

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

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

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

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

$$\hat{\psi}(\bar{y}, Y_T) \xrightarrow{d} \arg\min_\psi \frac{1}{2}\psi' \bar{\rho}_{\lambda\lambda}(\lambda_0; \beta_0)\psi + \frac{C_L}{C_Q}\exp\left(\bar{\chi}^2\right)\frac{\partial}{\partial\lambda}\bar{\rho}(\bar{y}, \lambda_0; \alpha_0)'\psi$$

$$= -\bar{\rho}_{\lambda\lambda}(\lambda_0; \beta_0)^{-1}\frac{C_L}{C_Q}\exp\left(\bar{\chi}^2\right)\frac{\partial}{\partial\lambda}\bar{\rho}(\bar{y}, \lambda_0; \alpha_0)'.$$

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 non degenerate random variable.