

Information Criteria for Moment Restriction Models

Naoya Sueishi*

University of Wisconsin[†]

This version: November 2009

Abstract

This paper proposes model and moment selection criteria for moment restriction models. Our selection methods are based on the Cressie-Read discrepancy. We propose information criteria based on the empirical Cressie-Read (ECR) estimator. Our criteria are natural extensions of the Akaike information criterion (AIC) and the Takeuchi information criterion (TIC). In the spirit of Akaike (1973), we derive asymptotically unbiased estimators for the expected value of the Cressie-Read discrepancy from the fitted model to the true data generating process. The resulting criteria are penalized ECR statistics. In a Monte Carlo study, we examine the properties of our criteria based on the empirical likelihood and exponential tilting estimators.

Keywords: Akaike information criterion; Cressie-Read information criterion; Empirical likelihood; Exponential tilting; Model selection; Takeuchi information criterion.

JEL Classification: C52.

*I am grateful to Bruce Hansen and Jack Porter for many valuable discussions and suggestions. The comments from Ryo Okui greatly improved the paper. I also thank Masato Kagihara, Qing-feng Liu, Enrique Pinzon and seminar participants at Fukuoka University for their helpful comments.

[†]Department of Economics, 1180 Observatory Drive, University of Wisconsin, Madison, WI 53706. Email: sueishi@wisc.edu

1 Introduction

Model selection has a long history. In particular, parametric likelihood-based selection methods have been discussed extensively in statistical literature. Examples include Akaike information criterion (AIC; Akaike (1973)), Takeuchi information criterion (TIC; Takeuchi (1976)), Bayesian information criterion (BIC; Schwarz (1978)), final prediction error (FPE; Akaike (1970), Shibata (1984)) and generalized information criterion (GIC; Konishi and Kitagawa (1996)). However, in many applications, econometric models are specified through moment restrictions rather than parametric density functions. Researchers are typically faced with two or more different moment restrictions, from which they need to choose the proper one.

This paper proposes information criteria for moment restriction models based on the empirical Cressie-Read (ECR) estimator (Newey and Smith (2004), Schennach (2007)). The ECR estimator nests the empirical likelihood (EL) estimator (Qin and Lawless (1994)) and the exponential tilting (ET) estimator (Kitamura and Stutzer (1997), Imbens, Spady, and Johnson (1998)). Recently, several model selection and hypothesis testing methods have been advocated based on the generalized method of moments (GMM) estimator and the generalized empirical likelihood (GEL) estimator including Andrews (1999), Andrews and Lu (2001), Ramalho and Smith (2002), Hong, Preston, and Shum (2003), Chen, Hong, and Shum (2007) and Hall, Inoue, Jana, and Shin (2007). Among them, Andrews and Lu (2001) and Hong, Preston, and Shum (2003) are highly relevant to this work.

Andrews and Lu (2001) propose model and moment selection procedures using the GMM estimator. Let (b, c) denote a pair of model and moment selection vector. Let $|b|$ and $|c|$ denote the number of parameters and the number of moments. Let $J_n(b, c)$ denote the J test statistic using (b, c) . Andrews and Lu (2001) propose the selection criteria of the form:

$$J_n(b, c) - \kappa_n(|c| - |b|)$$

where $|c| - |b|$ is the number of over-identifying restrictions and κ_n is a non-decreasing sequence. The term $\kappa_n(|c| - |b|)$ can be interpreted as a bonus (penalty) for using additional moments (parameters). Their selection rule picks the pair (b, c) which minimizes the criterion. Hong, Preston, and Shum (2003) propose selection criteria using the GEL estimator. They replace the J -statistic with the GEL statistic. In both procedures, the sequence κ_n must be specified by researchers. For example, $\kappa_n = 2$ yields an AIC-like criterion, whereas $\kappa_n = \log n$ yields a BIC-like criterion.

The main concern of the above papers is to choose the “correct” model. Given a vector of moment restrictions, they assume that some of which are correct and some of which are incorrect. Andrews and Lu (2001) and Hong, Preston, and Shum (2003) show that their BIC-like criteria asymptotically select the correct model and all correct moment restrictions.

This paper is distinct from Andrews and Lu (2001) and Hong, Preston, and Shum (2003) in the following two senses. First, we do not assume that the true data generating process (DGP) is included in the set of candidate models. Since models are simplifications of reality, all models are potentially misspecified in some sense. Therefore, we view models as approximations to the

true DGP and address the issue of selecting a model that provides the best approximation to the true DGP rather than selecting the “correct” model. Second, we theoretically derive the penalty (or bonus) term of the criteria based on the Akaike’s principle, while the penalty is specified in an ad hoc manner in Andrews and Lu (2001) and Hong, Preston, and Shum (2003). We also show that the AIC-like criterion of Hong, Preston, and Shum (2003) is asymptotically equivalent to the cross-validation criterion.

The closeness of the model to the DGP is measured by the Cressie-Read discrepancy (Cressie and Read (1984)). We introduce a new information criterion, which we call the Cressie-Read information criterion (CRIC). The CRIC is a generalization of the well-known Kullback-Leibler information criterion (KLIC). The use of the CRIC provides analogous model selection procedures with those of parametric likelihood-based models. In the spirit of Akaike (1973) and Takeuchi (1976), we derive asymptotically unbiased estimators of the expected value of the Cressie-Read discrepancy from the fitted model to the DGP. We show that the resulting selection criteria are penalized empirical Cressie-Read statistics (Baggerly (1998)), which are proper analogues of the AIC and the TIC.

We also discuss the robustness of our criteria. Especially, we compare the EL-based criteria and the ET-based criteria. Recently, Schennach (2007) has suggested that the KLIC is not a proper criterion under misspecification. In the presence of misspecification, we cannot define the KLIC for moment restriction models if the functions defining the moment restrictions are unbounded. Hence, in spite of the popularity of the KLIC in likelihood-based models, the KLIC is not necessarily a good criterion for moment restriction models. On the other hand, the entropy-based criterion does not suffer from such a problem. We do not need to restrict our candidate models to be bounded. Since the EL and the ET estimators solve the empirical version of Kullback-Leibler discrepancy and entropy minimization problems, this implies that the ET-based criteria are more robust to misspecification than the EL-based criteria.

To evaluate the properties of our criteria, we conduct Monte Carlo experiments under two different designs. In the first design, we consider an instruments selection problem given that the model is fixed. In the second design, we consider a problem of selecting model and instruments simultaneously when all candidate models are misspecified. We evaluate the criteria by the performance of post-selection estimators. The results of the simulation suggest that our criteria can be useful alternatives to existing criteria especially when all candidate models are potentially misspecified.

The rest of this paper is organized as follows. Section 2 introduces the CRIC for moment restriction models. Section 3 investigates the properties of the ECR estimator under misspecification. Information criteria for model selection are formulated in Section 4. In Section 5, we report the results of Monte Carlo simulation. Section 6 concludes. Proofs are given in the Appendix.

2 Cressie-Read information criterion

In this section we define and formulate the Cressie-Read information criterion (CRIC) for moment restriction models. Throughout the paper, we assume that y_1, \dots, y_n are independent and identically distributed random variables from an unknown true density $f(y)$.

2.1 Definition of the CRIC

We first introduce the CRIC for parametric models. Let $g(y)$ be a density function. The Cressie-Read (CR) discrepancy from g to the true density f is defined as

$$CR(f, g) = \int f(y)h\left(\frac{g(y)}{f(y)}\right) dy$$

where

$$h(x) = \frac{x^{\alpha+1} - 1}{\alpha(\alpha + 1)}, \quad -\infty < \alpha < \infty.$$

The degenerated cases $\alpha = -1$ and $\alpha = 0$ are handled by taking limits. Different values of α yield different discrepancies. For instance, $\alpha = -1$ and $\alpha = 0$ yield the Kullback-Leibler (KL) discrepancy ($h(x) = -\log x$) and the entropy ($h(x) = x \log x$).

Suppose that a model is given by a set of parametric density functions $\{g(y, \beta) : \beta \in \mathcal{B}\}$, where \mathcal{B} is a parameter space. We define the CRIC from the parametric model to the true DGP as

$$CRIC(f, g_\beta) = \min_{\beta \in \mathcal{B}} \int f(y)h\left(\frac{g(y, \beta)}{f(y)}\right) dy.$$

Hence, the CRIC nests the Kullback-Leibler information criterion (KLIC), which is defined as

$$KLIC(f, g_\beta) = \min_{\beta \in \mathcal{B}} \int f(y) \log\left(\frac{f(y)}{g(y, \beta)}\right) dy.$$

If the parametric model is correctly specified, that is, if there exists $\beta_0 \in \mathcal{B}$ such that $g(y, \beta_0) = f(y)$, then $CRIC(f, g_\beta) = 0$. Otherwise, $CRIC(f, g_\beta) > 0$. If the model is misspecified, the value β^* which minimizes the CR discrepancy is called pseudo-true value. Hence, the pseudo-true value is

$$\beta^* = \arg \min_{\beta \in \mathcal{B}} \int f(y)h\left(\frac{g(y, \beta)}{f(y)}\right) dy.$$

Note that the pseudo-true value depends on the CR parameter α if the model is misspecified.

It is well known that the maximum likelihood estimator (MLE)

$$\hat{\beta} = \arg \max_{\beta \in \mathcal{B}} \frac{1}{n} \sum_{i=1}^n \log g(y_i, \beta)$$

solves the sample version of the KLIC problem. White (1982) shows that the MLE is \sqrt{n} -consistent for the pseudo-true value of the KLIC and is asymptotically normally distributed. Thus the MLE is a consistent estimator for the best approximation to the true density f among the parametric class $\{g(y, \beta) : \beta \in \mathcal{B}\}$ in terms of the KL discrepancy.

2.2 CRIC for moment restriction models

We can apply the same idea to moment restriction models. The model we consider is

$$E[m(y, \theta_0)] = 0, \quad (2.1)$$

where $m : \mathbb{R}^{d_y} \times \Theta \rightarrow \mathbb{R}^l$ is a known function up to parameters $\theta \in \Theta \subset \mathbb{R}^p$. The expectation is taken with respect to the true DGP $f(y)$. The moment restriction model is said to be correctly specified if there exists θ_0 such that (2.1) is satisfied.

For each $\theta \in \Theta$, define $\mathcal{M}_\theta = \{g(y) : \int m(y, \theta)g(y)dy = 0\}$, a set of density functions which are compatible with the moment restriction. Then we can define $\mathcal{M} = \cup_{\theta} \mathcal{M}_\theta$ as the moment restriction model. We say that the model \mathcal{M} is misspecified if $f \notin \mathcal{M}$. It can be easily verified that this is equivalent to the following definition:

Definition 2.1

A model \mathcal{M} is said to be misspecified if $\inf_{\theta \in \Theta} \|E[m(y, \theta)]\| > 0$.

If the parameter vector is just-identified, then we can always find θ such that $E[m(y, \theta)] = 0$. Therefore, we exclusively consider the over-identified moment restrictions in the following discussion.

The CRIC for the moment restriction model \mathcal{M} is defined as $\min_{g \in \mathcal{M}} CR(f, g)$. Chen, Hong, and Shum (2007) obtain the KLIC for moment restriction models. Similar to the KLIC problem in Chen, Hong, and Shum (2007), the CRIC problem is also characterized in two steps. For fixed θ , we solve

$$\begin{aligned} v(\theta) \equiv \min_{g \in \mathcal{M}_\theta} CR(f, g) &= \min_{g(\cdot)} \int f(y)h\left(\frac{g(y)}{f(y)}\right) dy \quad \text{s.t.} \\ &\int g(y)dy = 1 \quad \text{and} \quad \int m(y, \theta)g(y)dy = 0. \end{aligned} \quad (2.2)$$

The infinite dimensional primal problem (2.2) has the following finite dimensional dual problem:

$$v^*(\theta) \equiv \max_{\mu \in \mathbb{R}} \max_{\bar{\lambda} \in \mathbb{R}^l} \left[\mu - \int f(y)h^*(\mu + \bar{\lambda}'m(y, \theta))dy \right],$$

where h^* is the convex conjugate of h :

$$h^*(x) = \frac{(\alpha x)^{(\alpha+1)/\alpha}}{\alpha + 1} - \frac{1}{\alpha(\alpha + 1)}.$$

The Fenchel duality theorem implies $v(\theta) = v^*(\theta)$. See Borwein and Lewis (1991) and Kitamura (2006) for details. Define $\lambda = \bar{\lambda}/\mu$. Concentrating μ out and ignoring a constant which does not depend on the model, we have

$$CRIC(f, g_\theta) = \min_{\theta \in \Theta} \max_{\lambda \in \Lambda} \int f(y)\rho(\lambda' m(y, \theta))dy \quad (2.3)$$

where

$$\rho(\xi) = -\frac{1}{\alpha + 1} (1 + \alpha\xi)^{(\alpha+1)/\alpha}.$$

Table 1: Discrepancy and Estimator (adopted from Schennach (2007))

α	Discrepancy	Estimator	$h(x)$	$\rho(\xi)$
-1	Kullback-Leibler	EL	$-\log x$	$\log(1 - \xi)$
0	Entropy	ET	$x \log x$	$-\exp(\xi)$
α	Cressie-Read	ECR	$\frac{x^{\alpha+1}-1}{\alpha(\alpha+1)}$	$-\frac{1}{\alpha+1}(1 + \alpha\xi)^{(\alpha+1)/\alpha}$

See Table 1 for the relationship between $h(x)$ and $\rho(\xi)$.

Strictly speaking, the CRIC defined by (2.3) is not the same as $\min_{g \in \mathcal{M}} CR(f, g)$, since we ignore a constant term. However, for the purpose of comparing models, it is enough to know the relative ranking. Thus, hereafter, we refer to (2.3) as the CRIC. Also, without loss of generality, we can normalize $\rho(\xi)$ so that $\rho(0) = 0$.

The value (θ^*, λ^*) which solves

$$\min_{\theta \in \Theta} \max_{\lambda \in \Lambda} \int f(y) \rho(\lambda' m(y, \theta)) dy \quad (2.4)$$

is called the pseudo-true value since it gives us the best approximating model in terms of the CR discrepancy. The pseudo-true value depends on the parameter α . The implied density that is closest to f is given by

$$g(y, \theta^*) = \frac{f(y) \rho_1(\lambda^{*'} m(y, \theta^*))}{\int f(y) \rho_1(\lambda^{*'} m(y, \theta^*)) dy}, \quad (2.5)$$

where $\rho_1(\xi) = d\rho(\xi)/d\xi$. If the model is misspecified then $\lambda^* \neq 0$ because $\lambda^* = 0$ and (2.5) imply that $g(y, \theta^*) = f(y)$.

The estimator which solves the empirical problem of (2.4)

$$\min_{\theta \in \Theta} \max_{\lambda \in \Lambda} \frac{1}{n} \sum_{i=1}^n \rho(\lambda' m(y_i, \theta)) \quad (2.6)$$

is the empirical Cressie-Read (ECR) estimator, which nests the empirical likelihood ($\alpha = -1$) and exponential tilting ($\alpha = 0$) estimators. The representation (2.6) is the GEL representation of the ECR estimator (Smith (1997)). Although the ECR family and the GEL family share many estimators in common, they do not completely coincide. The reason that we restrict our attention to the ECR estimator is that the ECR estimator has a best approximation interpretation while the GEL estimator may not. In the next section, we investigate the properties of the ECR estimator under misspecification.

2.3 Deficiency of the KLIC

As mentioned above, different choices of α yield different measures of discrepancy and different pseudo-true values. Then, which value of α is preferable for model selection? Although the KLIC ($\alpha = -1$) is a popular measure of discrepancy in the parametric likelihood case, Schennach

(2007) suggests that the KLIC has a serious problem if it is applied to the moment restriction models. If $\|m(y, \theta)\|$ is unbounded, we cannot properly define the pseudo-true value for the KLIC. The key problem is the following. If $\lambda \neq 0$ and $\|m(y, \theta)\|$ is unbounded, then $1 - \lambda' m(y_i, \theta)$ may take on negative values, and hence $\log(1 - \lambda' m(y_i, \theta))$ is ill-defined. Therefore if $\|m(y, \theta)\|$ is unbounded, then it must be the case that $\lambda = 0$. However the pseudo-true value λ^* cannot be zero by misspecification assumption. It is common that $\|m(y, \theta)\|$ is unbounded even though $E[m(y, \theta)]$ is bounded. For instance, in a linear instrumental variable model, $m(y, \theta) = z(\tilde{y} - x'\theta)$ is unbounded if the data has unbounded support. This fact suggests that KLIC might not be a good measure of discrepancy under possible misspecification.

We can circumvent this difficulty by using the entropy-based criterion ($\alpha = 0$). In contrast to the KLIC, the entropy-based criterion does not have a “singular” point. Now $\lambda' m(y, \theta)$ can take on any value. Therefore we do not need to put a restrictive assumption $\sup_{\theta} \sup_y \|m(y, \theta)\| < \infty$ to define the pseudo-true value. Another good feature of the entropy is that it can be interpreted as the KLIC from f to the model g . Since KL discrepancy is not symmetric with respect to its argument, $KL(f, g)$ is different from $KL(g, f)$. These facts suggest that the entropy is a good measure of discrepancy for model selection.

3 ECR estimator under misspecification

Before discussing model selection, we formally investigate the large sample properties of the ECR estimator under misspecification. Similar results are also obtained by Chen, Hong, and Shum (2007) and Schennach (2007) in the case of the EL estimator. We extend their result to the ECR family.

Hall and Inoue (2003) study the large sample behavior of the two-step GMM estimator in misspecified models. Compared to the GMM estimator, an advantageous point of using the ECR estimator for model selection is that we do not have to choose a weight matrix. In misspecified models, the pseudo-true value of the two-step GMM estimator depends on the weight matrix. That means that the goodness of the model is determined by the choice of the weight matrix. Since there is no theoretical guidance for the choice of the weight matrix under misspecification, this induces arbitrariness in model selection. By using the ECR estimator, we can avoid such arbitrariness.

Let $\gamma = (\theta', \lambda')'$. Let $Q(\theta, \lambda) = E[\rho(\lambda' m(y_i, \theta))]$ and let $Q_n(\theta, \lambda) = n^{-1} \sum_{i=1}^n \rho(\lambda' m(y_i, \theta))$. Let $M_i(\theta) = \partial m(y_i, \theta) / \partial \theta'$. Define

$$\phi(y_i, \gamma) = \begin{pmatrix} \rho_1(\lambda' m(y_i, \theta)) M_i(\theta)' \lambda \\ \rho_1(\lambda' m(y_i, \theta)) m(y_i, \theta) \end{pmatrix}.$$

The ECR estimator $\hat{\gamma}_{\text{ECR}} = (\hat{\theta}'_{\text{ECR}}, \hat{\lambda}'_{\text{ECR}})'$, which solves (2.6), satisfies the first order condition:

$$\frac{1}{n} \sum_{i=1}^n \phi(y_i, \hat{\gamma}_{\text{ECR}}) = 0. \quad (3.1)$$

Notice that (3.1) is a just-identified system. Hence a standard just-identified GMM theory applies. This idea is also used in Schennach (2007).

Now we make a set of assumptions. Since our objective is not finding the weakest possible conditions, we list high level assumptions.

Assumption 3.1

1. The data $\{y_i\}_{i=1}^n$ are i.i.d. with support $\mathcal{Y} \subset \mathbb{R}^{d_y}$.
2. $\Theta \subset \mathbb{R}^p$ and $\Lambda \subset \mathbb{R}^l$ are compact.
3. $E[m(y_i, \theta)m(y_i, \theta)']$ is positive definite uniformly over $\theta \in \Theta$.
4. $Q(\theta, \lambda)$ is continuous in $\theta \in \Theta$ and $\lambda \in \Lambda$.
5. The problem of $\min_{\theta \in \Theta} \max_{\lambda \in \Lambda} Q(\theta, \lambda)$ has a unique saddle point solution (θ^*, λ^*) with $\theta^* \in \text{int}(\Theta)$ and $\lambda^* \in \text{int}(\Lambda)$.
6. $\sup_{\theta \in \Theta} \sup_{\lambda \in \Lambda} |Q_n(\theta, \lambda) - Q(\theta, \lambda)| = o_p(1)$.
7. $\partial\phi(y_i, \gamma)/\partial\gamma'$ is continuous in $\gamma \in \mathcal{N}$ and $E[\sup_{\gamma \in \mathcal{N}} \|\partial\phi(y_i, \gamma)/\partial\gamma'\|] < \infty$, where \mathcal{N} is a neighborhood of (θ^*, λ^*)
8. $E[\partial\phi(y_i, \gamma^*)/\partial\gamma']$ is of full rank.
9. $E[\phi(y_i, \gamma^*)\phi(y_i, \gamma^*)']$ exists.

Condition 6 is a key condition when misspecification exists. A well-known sufficient condition for condition 6 is a dominance condition: $E[\sup_{\theta \in \Theta} \sup_{\lambda \in \Lambda} \rho(\lambda' m(y_i, \theta))] < \infty$ (see e.g., Newey and McFadden (1994)). However as is discussed in Schennach (2007), in the EL case, the dominance condition is violated if the model is misspecified and $\|m(y, \theta)\|$ is unbounded. Schennach (2007) shows that if $\|m(y, \theta)\|$ is unbounded then the EL estimator ceases to be \sqrt{n} -consistent for any $\theta \in \Theta$. On the other hand, condition 6 can be satisfied under a mild assumption in the case of ET estimator. Therefore, the ET estimator is more robust to misspecification than the EL estimator.

Proposition 3.1

Under Assumption 3.1, $\hat{\theta}_{ECR} \xrightarrow{p} \theta^*$, $\hat{\lambda}_{ECR} \xrightarrow{p} \lambda^*$, and

$$\sqrt{n} \begin{pmatrix} \hat{\theta}_{ECR} - \theta^* \\ \hat{\lambda}_{ECR} - \lambda^* \end{pmatrix} \xrightarrow{d} N(0, H^{-1} S H^{-1}),$$

where

$$H = E \left[\begin{pmatrix} \rho_{1i} \frac{\partial M_i' \lambda^*}{\partial \theta'} + \rho_{2i} M_i' \lambda^* \lambda^{*'} M_i & \rho_{1i} M_i' + \rho_{2i} M_i' \lambda^* m(y_i, \theta^*)' \\ \rho_{1i} M_i + \rho_{2i} m(y_i, \theta^*) \lambda^{*'} M_i & \rho_{2i} m(y_i, \theta^*) m(y_i, \theta^*)' \end{pmatrix} \right]$$

$$S = E \left[\begin{pmatrix} \rho_{1i}^2 M_i' \lambda^* \lambda^{*'} M_i & \rho_{1i}^2 M_i' \lambda^* m(y_i, \theta^*)' \\ \rho_{1i}^2 m(y_i, \theta^*) \lambda^{*'} M_i & \rho_{1i}^2 m(y_i, \theta^*) m(y_i, \theta^*)' \end{pmatrix} \right],$$

with $\rho_{1i} = \rho_1(\lambda^{*'} m(y_i, \theta^*))$, $\rho_{2i} = \rho_2(\lambda^{*'} m(y_i, \theta^*))$ and $M_i = M_i(\theta^*)$.

4 Information Criteria for Model Selection

Although the CRIC is an ideal measure, we cannot compute the CRIC in practice, since we do not know the true density function. Now we propose information criteria based on the ECR estimator. Following Akaike (1973) and Takeuchi (1976), we derive our selection criteria based on the asymptotically unbiased estimators for the expected value of the CRIC of the fitted model.

4.1 AIC and TIC

To understand our idea, it will be helpful to briefly review the AIC and TIC. In Section 2 we saw that the MLE minimized the sample version of the KL discrepancy between the model and the DGP. Now we study the actually attained population KLIC from the fitted model to the DGP. Let $\hat{\beta}$ be the MLE calculated from n observations y_1, \dots, y_n . Let $\hat{g}(y) = g(y, \hat{\beta})$ be the fitted density function. Then the KLIC of the fitted model is

$$\begin{aligned} KLIC(f, \hat{g}) &= \int f(y) \log f(y) dy - \int f(y) \log g(y, \hat{\beta}) dy \\ &\equiv C_f - \ell(\hat{\beta}). \end{aligned}$$

A constant C_f does not depend on the model. Therefore only $\ell(\hat{\beta})$ is relevant to assess the goodness of the model. Notice that $\ell(\hat{\beta})$ is a random variable. The expected value of $\ell(\hat{\beta})$ is

$$E[\ell(\hat{\beta})] = E \left[\int f(y) \log g(y, \hat{\beta}) dy \right]$$

where the expectation is with respect to the MLE. The AIC procedure estimates $E[\ell(\hat{\beta})]$ and picks the model which attain the highest $E[\ell(\hat{\beta})]$ or the lowest expected value of the KLIC.

A simple estimator of $E[\ell(\hat{\beta})]$ is $n^{-1} \sum_{i=1}^n \log g(y_i, \hat{\beta})$. However, this simple estimator tends to be upper-biased. The bias stems from the fact that we use the same data to estimate both the parameter and the expected value. Let p be the dimension of β . Under the assumption that the model is correctly specified, Akaike (1973) shows that the bias of the simple estimator is asymptotically p/n , or

$$E[\ell(\hat{\beta})] = E \left[\frac{1}{n} \sum_{i=1}^n \log g(y_i, \hat{\beta}) \right] - \frac{p}{n} + o(n^{-1}).$$

This implies the following criterion:

$$\text{AIC} = -2 \sum_{i=1}^n \log g(y_i, \hat{\beta}) + 2p. \quad (4.1)$$

The AIC selection procedure picks the model which minimizes (4.1).

Takeuchi (1976) derives a general information criterion without assuming correct specification. Takeuchi (1976) shows

$$E[\ell(\hat{\beta})] = E \left[\frac{1}{n} \sum_{i=1}^n \log g(y_i, \hat{\beta}) \right] - \frac{1}{n} \text{tr}(Q^{-1}\Omega) + o(n^{-1}),$$

where

$$\begin{aligned} Q &= -E \left[\frac{\partial^2 g(y_i, \beta^*)}{\partial \beta \partial \beta'} \right] \\ \Omega &= E \left[\left(\frac{\partial g(y_i, \beta^*)}{\partial \beta} \right) \left(\frac{\partial g(y, \beta^*)}{\partial \beta'} \right) \right]. \end{aligned}$$

The expectation is taken with respect to the true density f . This leads to the following information criterion:

$$\text{TIC} = -2 \sum_{i=1}^n \log g(y_i, \hat{\beta}) + 2 \text{tr}(\widehat{Q^{-1}\Omega}), \quad (4.2)$$

where $\widehat{Q^{-1}\Omega}$ is an estimate of $Q^{-1}\Omega$. Notice that if the model is correctly specified, then $Q = \Omega$ by the information equality. Hence, $\text{tr}(Q^{-1}\Omega) = p$ and (4.2) is reduced to (4.1). Stone (1977) and Shibata (1989) show that the TIC is asymptotically equivalent to the cross-validation.

4.2 ECR-based model selection criteria

Now we evaluate the CRIC of the fitted moment restriction model. Since the ECR estimator solves the sample version of the saddle point problem, it will be natural to construct our information criterion based on the ECR estimator. Given the ECR estimator $\hat{\theta}_{\text{ECR}}$, the fitted density of the model which is closest to f is

$$g(y, \hat{\theta}_{\text{ECR}}) = \frac{f(y)\rho_1(\hat{\lambda}'m(y, \hat{\theta}_{\text{ECR}}))}{\int f(y)\rho_1(\hat{\lambda}'m(y, \hat{\theta}_{\text{ECR}}))dy},$$

where

$$\hat{\lambda} = \arg \max_{\lambda \in \Lambda} \int f(y)\rho(\lambda'm(y, \hat{\theta}_{\text{ECR}}))dy.$$

Note that $\hat{\lambda}$ is not the same as the ECR estimator $\hat{\lambda}_{\text{ECR}}$, which solves

$$\max_{\lambda \in \Lambda} \frac{1}{n} \sum_{i=1}^n \rho(\lambda'm(y_i, \hat{\theta}_{\text{ECR}})).$$

Therefore, given $\hat{\theta}_{\text{ECR}}$, $\hat{\lambda}$ gives us the best fitted model in terms of the population CRIC, while $\hat{\lambda}_{\text{ECR}}$ gives us the best fitted model in terms of the empirical CRIC evaluated at observed sample points.

The CRIC of the fitted model is

$$\text{CRIC}(f, \hat{g}) = \int f(y)\rho(\hat{\lambda}'m(y, \hat{\theta}_{\text{ECR}}))dy,$$

and its expected value is

$$E[\text{CRIC}(f, \hat{g})] = E \left[\int f(y)\rho(\hat{\lambda}'m(y, \hat{\theta}_{\text{ECR}}))dy \right], \quad (4.3)$$

where the expectation is with respect to $(\hat{\theta}_{\text{ECR}}, \hat{\lambda})$. Our selection rule estimates (4.3) and selects the model which attains the minimum expected CRIC. In other words, we regard $\text{CRIC}(f, g)$

as our loss function and choose the model which attains the minimum estimated risk. Notice that $\hat{\lambda}$ is not observable since it depends on the unknown true density. Therefore we replace $\hat{\lambda}$ with $\hat{\lambda}_{\text{ECR}}$ to estimate (4.3). As in the case of AIC and TIC, a simple estimator $n^{-1} \sum_{i=1}^n \rho(\hat{\lambda}'_{\text{ECR}} m(y_i, \hat{\theta}_{\text{ECR}}))$ is a biased estimator for (4.3). Now there are two sources of bias: using same data to estimate parameters and expectation and using $\hat{\lambda}_{\text{ECR}}$ instead of $\hat{\lambda}$. We evaluate the bias of the order $O(n^{-1})$ and obtain a bias-corrected estimator. To this end, we utilize the next lemma.

Lemma 4.1

Suppose that Assumption 3.1 holds. Let $\hat{\lambda}$ solve

$$\max_{\lambda \in \Lambda} \int f(y) \rho(\lambda' m(y, \hat{\theta}_{\text{ECR}})) dy.$$

Then

$$\sqrt{n} \begin{pmatrix} \hat{\theta}_{\text{ECR}} - \theta^* \\ \hat{\lambda} - \lambda^* \end{pmatrix} \xrightarrow{d} N(0, V) = N \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} V_{11} & V_{11}A' \\ AV_{11} & AV_{11}A' \end{pmatrix} \right),$$

where V_{11} is the asymptotic variance of $\hat{\theta}_{\text{ECR}}$, and

$$A = -E[\rho_{2i} m(y_i, \theta^*) m(y_i, \theta^*)']^{-1} (E[\rho_{1i} M_i] + E[\rho_{2i} m(y_i, \theta^*) \lambda^* M_i]).$$

Having the lemma in hand, we obtain the following theorem.

Theorem 4.1

Suppose that Assumption 3.1 holds. Then an asymptotically unbiased estimator for (4.3) is

$$\frac{1}{n} \sum_{i=1}^n \rho(\hat{\lambda}'_{\text{ECR}} m(y_i, \hat{\theta}_{\text{ECR}})) + \frac{1}{2n} \text{tr}(H^{-1}S) + \frac{1}{2n} \text{tr}(HV),$$

where H , S and V are given in Proposition 3.1 and Lemma 4.1. Moreover, if the model is correctly specified, then we have

$$\frac{1}{n} \sum_{i=1}^n \rho(\hat{\lambda}'_{\text{ECR}} m(y_i, \hat{\theta}_{\text{ECR}})) - \frac{1}{2n} (l - 2p),$$

where l is the number of moments and p is the number of parameters.

The theorem implies the following two information criteria:

$$\text{ECR-AIC1} = 2 \sum_{i=1}^n \rho(\hat{\lambda}'_{\text{ECR}} m(y_i, \hat{\theta}_{\text{ECR}})) - (l - 2p) \tag{4.4}$$

$$\text{ECR-TIC1} = 2 \sum_{i=1}^n \rho(\hat{\lambda}'_{\text{ECR}} m(y_i, \hat{\theta}_{\text{ECR}})) + \text{tr}(\widehat{H^{-1}S}) + \text{tr}(\widehat{HV}), \tag{4.5}$$

where $\widehat{H^{-1}S}$ and \widehat{HV} are estimators of $H^{-1}S$ and HV . For instance, we can estimate $H^{-1}S$ by

$$\widehat{H^{-1}S} = \left[\frac{1}{n} \sum_{i=1}^n \frac{\partial \phi(y_i, \hat{\gamma}_{\text{ECR}})}{\partial \gamma} \right]^{-1} \frac{1}{n} \sum_{i=1}^n \phi(y_i, \hat{\gamma}_{\text{ECR}}) \phi(y_i, \hat{\gamma}_{\text{ECR}})'$$

Also, V can be estimated by

$$\hat{V} = \begin{pmatrix} \hat{V}_{11} & \hat{V}_{11}\hat{A}' \\ \hat{A}\hat{V}_{11} & \hat{A}\hat{V}_{11}\hat{A}' \end{pmatrix},$$

where \hat{V}_{11} is the upper diagonal element of $H^{-1}\widehat{SH}^{-1}$, which is obtained similarly as $\widehat{H^{-1}S}$.

A better model has a smaller value of the information criterion. The ECR-AIC1 has an intuitive form. If the model is correctly specified, then an additional use of moments is rewarded because the first term of (4.4) tends to be large as we increases l even if the moment restriction is correct. On the other hand, an additional use of parameters is penalized for fixed l so that we pick up a parsimonious model.

The ECR-TIC1 is more robust to misspecification than the ECR-AIC1. However, it does not necessarily mean that we should always use the ECR-TIC1 rather than the ECR-AIC1. Shibata (1989) discusses the disadvantage of the TIC in the parametric case. The same argument applies to our case. The ECR-TIC1 requires estimation of $H^{-1}S$ and HV . The estimation error of $\text{tr}(\widehat{H^{-1}S})$ and $\text{tr}(\widehat{HV})$ is large when l and p are large relative to the sample size, which makes the selection procedure unstable. The approximation by (4.4) will be good if the approximating model is good. If the model deviated greatly from the true DGP, then we would not consider it as a candidate model. Therefore, in practice, we conjecture that $l - 2p$ will be a good bias approximation even when the model is misspecified. If one believes that all candidate models provide rather poor approximations and has large sample size, then the ECR-TIC1 will perform better than the ECR-AIC1.

As special cases of the ECR-AIC1, we obtain EL- and ET-based criteria:

$$\text{EL-AIC1} = 2 \sum_{i=1}^n \log \left(1 - \hat{\lambda}'_{\text{EL}} m(y_i, \hat{\theta}_{\text{EL}}) \right) - (l - 2p) \quad (4.6)$$

$$\text{ET-AIC1} = 2 \sum_{i=1}^n \left(1 - \exp \left(\hat{\lambda}'_{\text{ET}} m(y_i, \hat{\theta}_{\text{ET}}) \right) \right) - (l - 2p). \quad (4.7)$$

As discussed in Section 3, the EL estimator has a serious drawback if the model is misspecified. We can apply the EL-based criteria only to bounded moment restrictions. On the other hand, the ET estimator is robust to misspecification, and hence the ET-based criteria can be applicable to a wide class of settings.

By a simple application of Sin and White (1996), it can be easily shown that if the numbers of candidate models and moments are finite and fixed, then our ECR-based criteria asymptotically select the pair of model and moment that is closest to the true DGP in terms of the CR discrepancy. Hence, if there is only one pair which attains the minimum CR discrepancy, then we can select the best pair with probability approaching one. In this sense, it is possible to argue that our criteria are consistent. In contrast, if there are multiple pairs of model and moment which attain the minimum CR discrepancy, then our criteria select one of them with probability approaching one, and there is no guarantee that the selected pair maximizes the number of over-identifying restrictions. Therefore, if we adopt the definition of Andrews and Lu (2001) for consistency, our criteria are not consistent. However, if all models are misspecified, there is no

theoretical justification for maximizing the number of over-identification restrictions. Thus the lack of consistency is not a serious problem.

4.3 New interpretation of existing criteria

We conclude this section by considering the relationship between our criteria and the AIC-like criterion proposed by Hong, Preston, and Shum (2003). Their AIC-like criterion can be interpreted as an estimator of

$$E \left[\int f(y) \rho \left(\hat{\lambda}'_{\text{ECR}} m(y, \hat{\theta}_{\text{ECR}}) \right) dy \right]. \quad (4.8)$$

Here, $\hat{\lambda}_{\text{ECR}}$ is the ECR estimator of λ^* . Although (4.8) is not the expected CRIC, this is also a well-defined risk function. By a simple modification of the proof of Theorem 4.1, an asymptotically unbiased estimator for (4.8) is obtained by

$$\frac{1}{n} \sum_{i=1}^n \rho \left(\hat{\lambda}'_{\text{ECR}} m(y_i, \hat{\theta}_{\text{ECR}}) \right) + \frac{1}{n} \text{tr}(H^{-1}S).$$

If the model is correctly specified, we have $\text{tr}(H^{-1}S) = p - l$ (see the proof of Theorem 4.1). Therefore, we obtain the following information criteria:

$$\text{ECR-AIC2} = 2 \sum_{i=1}^n \rho \left(\hat{\lambda}'_{\text{ECR}} m(y_i, \hat{\theta}_{\text{ECR}}) \right) - 2(l - p) \quad (4.9)$$

$$\text{ECR-TIC2} = 2 \sum_{i=1}^n \rho \left(\hat{\lambda}'_{\text{ECR}} m(y_i, \hat{\theta}_{\text{ECR}}) \right) + 2\text{tr}(\widehat{H^{-1}S}). \quad (4.10)$$

These criteria are parallel to (4.1) and (4.2). Also, (4.9) is the same as the AIC-like criterion in Hong, Preston, and Shum (2003). Since the ECR statistic is asymptotically equivalent to the J -statistic in correct specification, (4.9) is also equivalent to the MMSC-AIC in Andrews and Lu (2001) in large sample. This result gives a new interpretation to their information criteria, though the penalty term is determined in ad hoc manner in Andrews and Lu (2001) and Hong, Preston, and Shum (2003).

Also, an equivalence result between the TIC and the cross-validation also holds in the semi-parametric case. Let $(\hat{\theta}_{(-i)}, \hat{\lambda}_{(-i)})$ be the leave-one-out estimator. Then the cross-validation criterion is

$$\text{CV} = \frac{1}{n} \sum_{i=1}^n \rho \left(\hat{\lambda}'_{(-i)} m(y_i, \hat{\theta}_{(-i)}) \right).$$

We have the following result.

Proposition 4.1

The ECR-TIC2 is asymptotically equivalent to the CV.

Hence, if the model is correctly specified, then the AIC-like criteria of Hong, Preston, and Shum (2003) is asymptotically equivalent to the cross-validation. The reason that the CV is equivalent to (4.10) and not (4.5) is intuitively clear. The CV only corrects the bias caused by using the

same data to estimate the parameter and the expected value and does not correct the bias caused by using $\hat{\lambda}_{\text{ECR}}$ instead of $\hat{\lambda}$.

It is hard to determine which is a better criterion between (4.3) and (4.8). Some people may believe that (4.9) and (4.10) are the proper analogues of the AIC and TIC rather than (4.4) and (4.5). In my opinion, however, $\hat{\lambda}_{\text{ECR}}$ is just a by-product of estimating θ^* and is not of particular concern in itself. In model selection point of view, we are interested in the closest model with respect to the population CRIC, not with respect to the empirical CRIC evaluated at particular points. Since $\hat{\lambda}_{\text{ECR}}$ is associated with the empirical CRIC, it might better to avoid using $\hat{\lambda}_{\text{ECR}}$ for the criterion of the model selection.

5 Monte Carlo simulation

We report the results of Monte Carlo experiments to evaluate the performance of our selection criteria. We consider ECR-AIC1, ECR-TIC1, ECR-AIC2, ECR-TIC2 and BIC-like criteria proposed by Hong, Preston, and Shum (2003). For each type of criterion, we calculate the EL- and ET-based criteria. For instance, BIC-like criteria are given by

$$\begin{aligned} \text{EL-BIC} &= 2 \sum_{i=1}^n \log \left(1 - \hat{\lambda}'_{\text{EL}} m(y_i, \hat{\theta}_{\text{EL}}) \right) - (l - p) \log n \\ \text{ET-BIC} &= 2 \sum_{i=1}^n \left(1 - \exp \left(\hat{\lambda}'_{\text{ET}} m(y_i, \hat{\theta}_{\text{ET}}) \right) \right) - (l - p) \log n. \end{aligned}$$

Therefore we compare ten criteria in total.

We consider two different designs in the following discussion. In the first design, we select a set of instruments given that the model is fixed. In the second design, we consider a problem of selecting model and instruments simultaneously when the true model is potentially infinite dimensional.

5.1 Design 1

The first design is similar to that of Hong, Preston, and Shum (2003). The true DGP is given by

$$\begin{aligned} y_i &= 1 + x_i + 0.5u_i, \\ x_i &= \sum_{j=1}^{100} \theta_j \eta_{ji} + 0.5u_i \end{aligned}$$

for $i = 1, \dots, n$, where u_i and $\{\eta_{ji}\}_{j=1}^{100}$ are i.i.d. $N(0, 1)$ random variables. The coefficients θ_j are specified as $\theta_j = \sqrt{2\alpha} j^{-\alpha-1/2}$. By changing the values of α , we can control the degree of decline of θ_j .

We estimate the intercept and the slope coefficient of the first equation by using instrumental variables. We consider eight candidate instruments: $z_{1i} = \eta_{1i}$, $z_{2i} = \eta_{2i}$, $z_{3i} = \eta_{3i} + 0.01u_i$, $z_{4i} = \eta_{4i} + 0.01u_i$, $z_{5i} = \eta_{5i} + 0.1u_i$, $z_{6i} = \eta_{6i} + 0.1u_i$, $z_{7i} = \eta_{7i} + 0.15u_i$ and $z_{8i} = \eta_{8i} + 0.15u_i$.

Only z_1 and z_2 are valid instruments, since other instruments are correlated with the error term. The instrument becomes increasingly bad as the number increases.

We consider six sets of instruments.

M1 constant, z_1, z_2

M2 constant, z_1, z_2, z_3, z_4

M3 constant, z_1, z_2, z_3, z_5

M4 constant, $z_1, z_2, z_3, z_4, z_5, z_6$

M5 constant, $z_1, z_2, z_3, z_4, z_5, z_6, z_7$

M6 constant, $z_1, z_2, z_3, z_4, z_5, z_6, z_7, z_8$

All sets of instruments except for M1 are invalid. The sets of instruments are determined so that there is a clear difference in the RMSE between the estimators that use different sets of instruments. However, in contrast to Hong, Preston, and Shum (2003), we determine instruments so that invalid moments are not severely bad especially when the sample size is small. Therefore it is rather difficult to distinguish between good instruments and bad instruments.

We present the relative performance of the criteria by the selection probability and the performance of post-selection EL and ET estimators. To assess the performance of post-selection estimators, we calculate the mean and RMSE of the post-selection estimators of the slope coefficient. We also report the rejection rate of a 5 % t -test that the slope coefficient is equal to the true value of unity. In terms of these measures, M1 and M2 are good sets of instruments because estimators based on them are more efficient than those based on M3-M6 when the sample size is sufficiently large. As we will see below, the post-selection estimators based on M2 perform better than the estimators based on M1 in terms of the RMSE even though M2 contains invalid instruments. Also the sized distortion caused by using invalid instruments is trivial for M2. On the other hand, M3-M6 are bad sets of instruments because they distort the inference.

Tables 2 reports the selection probabilities of the ten criteria for $\alpha = 0.2$. We show the results of 1,000 replications for four sample sizes $n = 100, 250, 500$ and 1,000. From Table 2, we see that the difference between EL-based criteria and ET-based criteria is small. Also, the AIC and TIC type criteria perform very similarly for all sample sizes.

The BIC shows quite different features from other criteria. The BIC selects the worst set of instruments (M6) with high probability when the sample size is small ($n = 100$). Even for $n = 1,000$, the BIC selects a poor set of instruments (M4) with a certain probability. The effect of increasing sample size is enormous for the BIC. For the EL-BIC, the probability of selecting M2 increases from 0.279 to 0.722 as the sample size increases from 500 to 1,000. In contrast, the selection probabilities of AIC1 and TIC1 are less sensitive to the sample size.

Although the probability of selecting good sets of instruments (M1 or M2) are almost same in both AIC1 and AIC2 in large sample, the probabilities of selecting M1 and M2 are quite different. For the AIC1, the probability of selecting M2 increases as the sample size increases, but it increases more slowly compared to that of the AIC2. For $n = 1,000$, the EL-AIC2 selects M2 with probability 0.838, while the EL-AIC1 selects M2 with probability 0.615.

Next we evaluate the criteria by the performance of post-selection inference. The results are summarized in Tables 3. Following Hong, Preston, and Shum (2003), we report the mean and RMSE of the post-selection estimators of the slope coefficient. We also present the rejection rates of a 5 % t -test which tests the slope coefficient is equal to the true value of unity. The standard errors of the estimators are obtained under the correct specification assumption. The results based on each of six sets of instruments are labeled M1-M6.

Table 3 shows that M2 outperforms other sets of instruments in terms of the RMSE. The rejection rates of the t -test based on M1 and M2 are close to the nominal size. In contrast, the rejection rates based on M4-M6 are quite large. Therefore, the cost of using M4-M6 can be huge for post-selection inference.

As will be anticipated from the results of Table 2, the AIC and TIC perform similarly. Also, there is no clear distinction between EL and ET estimators. The RMSE based on the AIC1 is almost the same as that based on the AIC2, though the AIC1 works slightly better than the AIC2 in terms of the rejection rate. The BIC performs poorly. The rejection rate based on the EL-BIC is 0.252 for $n = 500$. Thus the post-selection inference based on the BIC is quite misleading.

As is shown by Hong, Preston, and Shum (2003), the BIC type criteria work well when there is a clear distinction between good instruments and bad instruments. However, in reality, the difference will be rather unclear. In such a case, it is safe to use AIC or TIC type criteria because using bad instruments is more harmful than using less instruments when the sample size is modest. The AIC and TIC type criteria perform reasonably well in our design. The choice between AIC1 and AIC2 depends on the preference of the researcher. The AIC1 is conservative compared to the AIC2 in the sense that the AIC1 tends to select valid instruments with high probability. However, AIC1 tends to select less number of instruments even when additional instrument can improve the efficiency.

We also study the performance of the criteria for different values of α and heteroskedastic errors. We do not report the results since they are similar to the results given above. The results are also not sensitive to the distributions of $\{\eta_{ji}\}$ and u_i .

Table 2: Selection probabilities (Design 1)

n	Empirical likelihood						Exponential tilting					
	M1	M2	M3	M4	M5	M6	M1	M2	M3	M4	M5	M6
	AIC1											
100	0.255	0.303	0.186	0.137	0.071	0.048	0.274	0.300	0.170	0.138	0.070	0.048
250	0.310	0.458	0.129	0.093	0.008	0.002	0.326	0.425	0.140	0.089	0.018	0.002
500	0.348	0.573	0.060	0.019	0.000	0.000	0.366	0.547	0.006	0.027	0.000	0.000
1000	0.374	0.615	0.011	0.000	0.000	0.000	0.382	0.609	0.009	0.000	0.000	0.000
	TIC1											
100	0.266	0.305	0.180	0.132	0.073	0.044	0.281	0.297	0.178	0.133	0.063	0.048
250	0.316	0.457	0.125	0.093	0.007	0.002	0.325	0.417	0.143	0.095	0.016	0.004
500	0.351	0.571	0.060	0.018	0.000	0.000	0.363	0.552	0.058	0.027	0.000	0.000
1000	0.378	0.612	0.010	0.000	0.000	0.000	0.379	0.612	0.009	0.000	0.000	0.000
	AIC2											
100	0.051	0.243	0.135	0.223	0.159	0.189	0.051	0.228	0.120	0.226	0.134	0.241
250	0.098	0.451	0.111	0.231	0.061	0.048	0.111	0.430	0.120	0.229	0.078	0.032
500	0.124	0.694	0.072	0.099	0.010	0.001	0.111	0.696	0.008	0.106	0.007	0.000
1000	0.141	0.838	0.016	0.005	0.000	0.000	0.149	0.828	0.018	0.005	0.000	0.000
	TIC2											
100	0.059	0.260	0.135	0.224	0.152	0.170	0.095	0.254	0.149	0.194	0.126	0.182
250	0.102	0.448	0.112	0.232	0.063	0.043	0.119	0.439	0.129	0.218	0.067	0.028
500	0.126	0.691	0.072	0.100	0.010	0.001	0.114	0.695	0.075	0.109	0.006	0.001
1000	0.142	0.836	0.017	0.005	0.000	0.000	0.150	0.827	0.018	0.005	0.000	0.000
	BIC											
100	0.003	0.050	0.013	0.168	0.191	0.575	0.003	0.033	0.009	0.139	0.164	0.652
250	0.001	0.101	0.013	0.276	0.020	0.403	0.002	0.106	0.012	0.270	0.208	0.402
500	0.001	0.279	0.022	0.399	0.163	0.136	0.001	0.281	0.013	0.430	0.145	0.130
1000	0.003	0.722	0.013	0.238	0.021	0.003	0.000	0.707	0.015	0.261	0.012	0.005

Table 3: Post-selection results (Design 1)

	$n = 100$			$n = 250$			$n = 500$			$n = 1000$		
	Mean	RMSE	Rej.	Mean	RMSE	Rej.	Mean	RMSE	Rej.	Mean	RMSE	Rej.
	Empirical likelihood											
M1	0.996	0.070	0.054	0.998	0.044	0.060	0.998	0.031	0.053	0.999	0.021	0.055
M2	1.000	0.063	0.078	1.002	0.038	0.055	1.002	0.028	0.063	1.004	0.019	0.064
M3	1.013	0.064	0.100	1.015	0.043	0.109	1.015	0.031	0.119	1.017	0.026	0.177
M4	1.024	0.064	0.147	1.027	0.045	0.193	1.027	0.037	0.241	1.028	0.034	0.401
M5	1.039	0.070	0.209	1.041	0.054	0.320	1.040	0.048	0.448	1.041	0.046	0.701
M6	1.050	0.077	0.274	1.052	0.063	0.432	1.052	0.057	0.646	1.054	0.057	0.872
AIC1	1.001	0.068	0.094	1.001	0.044	0.080	1.000	0.030	0.071	1.001	0.021	0.067
TIC1	1.001	0.069	0.093	1.001	0.044	0.080	0.999	0.030	0.071	1.001	0.021	0.067
AIC2	1.010	0.068	0.133	1.007	0.044	0.118	1.002	0.031	0.088	1.002	0.020	0.071
TIC2	1.010	0.068	0.129	1.007	0.044	0.116	1.002	0.031	0.088	1.002	0.020	0.071
BIC	1.032	0.070	0.207	1.031	0.052	0.261	1.020	0.039	0.252	1.008	0.023	0.139
	Exponential tilting											
M1	0.996	0.070	0.055	0.995	0.045	0.059	0.999	0.031	0.051	0.999	0.022	0.048
M2	1.000	0.064	0.087	1.000	0.040	0.074	1.003	0.027	0.063	1.002	0.020	0.061
M3	1.014	0.066	0.111	1.013	0.042	0.104	1.015	0.032	0.113	1.016	0.025	0.163
M4	1.025	0.065	0.156	1.024	0.045	0.171	1.027	0.037	0.251	1.027	0.032	0.380
M5	1.039	0.071	0.241	1.038	0.053	0.286	1.040	0.048	0.461	1.041	0.045	0.697
M6	1.050	0.078	0.305	1.051	0.062	0.414	1.051	0.058	0.647	1.053	0.056	0.892
AIC1	1.002	0.070	0.108	0.998	0.045	0.085	1.001	0.030	0.068	1.000	0.022	0.065
TIC1	1.002	0.070	0.110	0.999	0.045	0.086	1.001	0.030	0.068	1.000	0.022	0.065
AIC2	1.013	0.070	0.165	1.004	0.045	0.114	1.004	0.030	0.082	1.001	0.021	0.067
TIC2	1.011	0.070	0.151	1.004	0.044	0.108	1.003	0.030	0.083	1.001	0.021	0.067
BIC	1.013	0.072	0.244	1.028	0.052	0.254	1.020	0.039	0.256	1.008	0.024	0.148

5.2 Design 2

Next we consider the problem of selecting model and instruments simultaneously. Our design is similar to that of Newey and Powell (2003). The true DGP is given by a set of equations

$$\begin{aligned} y_i &= \mu(x_i) + e_i, & \mu(x_i) &= \log(|x_i - 1| + 1)\text{sgn}(x_i - 1) \\ x_i &= z_i + v_i \end{aligned}$$

for $i = 1, \dots, n$. The errors e_i and v_i and the instrument z_i are generated by

$$\begin{pmatrix} e_i \\ v_i \\ z_i \end{pmatrix} \sim N \left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 & 0 \\ 0.5 & 1 & 0 \\ 0 & 0 & 3 \end{pmatrix} \right).$$

Notice that $E[e_i|z_i] = 0$ but $E[e_i|x_i] \neq 0$.

We assume that the econometrician knows that the error term e_i is correlated with the explanatory variable x_i . The goal is to estimate $\mu(\cdot)$ nonparametrically by using splines. Since e_i is correlated with x_i , this is a nonparametric instrumental variable estimation problem. Our estimation method can be interpreted as an information theoretic alternative to the nonparametric two stage least squares estimator of Newey and Powell (2003).

To approximate the function $\mu(\cdot)$, we consider four candidate sets of regressors: $x_{1i} = (1, x_i)'$, $x_{2i} = (1, x_i, x_i^2)'$, $x_{3i} = (1, x_i, x_i^2, [\max\{x_i - 1, 0\}]^2)'$ and $x_{4i} = (1, x_i, x_i^2, [\max\{x_i - 1, 0\}]^2, [\max\{x_i + 1, 0\}]^2)'$. Since the true DGP is infinite dimensional, all approximating models are misspecified. Also, we consider three candidate sets of instruments: $z_{1i} = (1, z_i, z_i^2, z_i^3, [\max\{z_i - 2, 0\}]^3, [\max\{z_i - 1, 0\}]^3)'$, $z_{2i} = (1, z_i, z_i^2, z_i^3, [\max\{z_i - 2, 0\}]^3, [\max\{z_i - 1, 0\}]^3, [\max\{z_i + 1, 0\}]^3)'$ and $z_3 = (1, z_i, z_i^2, z_i^3, [\max\{z_i - 2, 0\}]^3, [\max\{z_i - 1, 0\}]^3, [\max\{z_i + 1, 0\}]^3, [\max\{z_i + 2, 0\}]^3)'$. Thus we consider 12 pairs of model and instruments in total. The moment restriction models we estimate can be written as

$$E[z_{ki}(y_i - x'_{ji}\theta_j)] = 0$$

for $j = 1, \dots, 4$ and $k = 1, 2, 3$.

To evaluate the performance of the criteria, we calculate the RMSE of the post-selection estimators. The RMSE is calculated over the realization of x_i . As a benchmark, we also calculate the RMSE based on the best pair of model and instruments in each experiment. We show the results of 1,000 replications for three sample sizes $n = 500, 1,000$ and $2,000$. We report the mean of the RMSE over repetition. We also report the ratio of the mean of RMSE based on the each criterion to the mean of RMSE based on the best pair.

The results are summarized in Table 4. The result based on the best pair is labeled Oracle. Again, the AIC1 and TIC1 perform quite similarly for both EL and ET cases. However, the TIC2 performs better than the AIC2, especially when the number of observations is large. This result suggests the possibility that the TIC works better than the AIC when all models are misspecified and the sample size is sufficiently large. The AIC1 outperforms the AIC2 for all sample sizes. The BIC is inferior to other criteria. It may not be a good idea to use the BIC type criteria

Table 4: Post-selection results (Design 2)

	$n = 500$		$n = 1000$		$n = 2000$	
	RMSE	Ratio	RMSE	Ratio	RMSE	Ratio
Empirical likelihood						
AIC1	0.1599	1.282	0.1166	1.201	0.0895	1.158
AIC2	0.1615	1.295	0.1204	1.240	0.0965	1.248
TIC1	0.1614	1.294	0.1172	1.207	0.0901	1.165
TIC2	0.1621	1.300	0.1178	1.213	0.0901	1.165
BIC	0.1906	1.529	0.1405	1.447	0.0992	1.283
Oracle	0.1247	1.000	0.0971	1.000	0.0773	1.000
Exponential tilting						
AIC1	0.1639	1.312	0.1173	1.219	0.0886	1.163
AIC2	0.1649	1.319	0.1211	1.259	0.0944	1.239
TIC1	0.1640	1.312	0.1173	1.220	0.0890	1.168
TIC2	0.1634	1.307	0.1185	1.232	0.0884	1.161
BIC2	0.1983	1.587	0.1431	1.487	0.0987	1.296
Oracle	0.1250	1.000	0.0962	1.000	0.0762	1.000

when all models are potentially misspecified and/or the true model is infinite dimensional, since there is no theoretical justification for maximizing the number of over-identification in such a case.

In sum, AIC1 and TIC1 perform reasonably well in our setting. The RMSE based on the AIC1 and TIC1 gets close to the RMSE based the Oracle as the sample size increases. To my knowledge, there is no prior work in the literature which addresses the issue of selecting model and instruments simultaneously in the context of nonparametric instrumental variable estimation. The results of the Monte Carlo study suggests that our criteria could be a good reference point. Further investigation will be required to analyze the difference between AIC1 and AIC2.

6 Conclusion

This paper proposes model and moment selection criteria based on the ECR estimator. We obtain the criteria from the information-theoretic point view. Our method extends the AIC and the TIC to the moment restriction model selection. Our main contributions are the following: (i) We proposed information criteria which are applicable even when all candidate models are potentially misspecified; (ii) We derived the penalty term of the criteria based on the Akaike's principle, whereas the penalty is determined in an ad hoc way in prior works.

Also, we give a new interpretation to the AIC-like criterion of Hong, Preston, and Shum

(2003). Under correct specification, we show that their criterion is asymptotically equivalent to the cross-validation criterion. This finding will justify the usage of AIC-like criterion even though the criterion lacks consistency.

In Monte Carlo experiments, we compare our criteria with the BIC type criteria of Hong, Preston, and Shum (2003). We study the performance of the EL- and ET-based criteria as the special cases of the ECR-based criteria. In the first experiment, we consider an instruments selection problem. In the second experiment, we consider a model and instruments selection problem when all models are misspecified. We analyze the selection probabilities and the performance of post-selection estimators. The results of the experiments suggest that our criteria can be useful alternatives to existing criteria especially when all candidate models are potentially misspecified.

This paper addresses only model selection and does not discuss how to conduct post-selection inference. Leeb and Pötscher (2005) argue that model selection has an important impact on subsequent inference. Even if the model selection procedure is consistent, it might be misleading to rely on the usual asymptotic theory. We hope to study this important issue for future research.

A Appendix

Proof of Proposition 3.1. For the proof of consistency, we closely follow Christoffersen, Hahn, and Inoue (2001) and Chen, Hong, and Shum (2007).

Condition 3 implies that $\lambda(\theta)$ is continuous with respect to θ , where

$$\lambda(\theta) = \arg \max_{\lambda \in \Lambda} E[\rho(\lambda' m(y_i, \theta))].$$

Let $L = E[\rho(\lambda^{*'} m(y_i, \theta^*))]$. The saddle point property implies that

$$E[\rho(\lambda(\theta)' m(y_i, \theta))] > L$$

for all $\theta \neq \theta^*$. Let $\hat{\lambda}_{\text{ECR}}(\theta) = \arg \max_{\lambda \in \Lambda} n^{-1} \sum_{i=1}^n \rho(\lambda' m(y_i, \theta))$. Then by definition of $\hat{\lambda}_{\text{ECR}}(\theta)$,

$$\frac{1}{n} \sum_{i=1}^n \rho(\lambda(\theta)' m(y_i, \theta)) \leq \frac{1}{n} \sum_{i=1}^n \rho(\hat{\lambda}_{\text{ECR}}(\theta)' m(y_i, \theta)). \quad (\text{A.1})$$

Let $\Gamma(\theta, \delta)$ be an open ball of radius δ around θ . By condition 6, for all $\delta > 0$, there exists $h > 0$ such that

$$P\left(\inf_{\theta \in \Theta \setminus \Gamma(\theta^*, \delta)} \frac{1}{n} \sum_{i=1}^n \rho(\lambda(\theta)' m(y_i, \theta)) < L + h\right) \rightarrow 0 \quad (\text{A.2})$$

as $n \rightarrow \infty$. Hence, by (A.1) and (A.2), we have

$$P\left(\inf_{\theta \in \Theta \setminus \Gamma(\theta^*, \delta)} \frac{1}{n} \sum_{i=1}^n \rho(\hat{\lambda}_{\text{ECR}}(\theta)' m(y_i, \theta)) < L + h\right) \rightarrow 0. \quad (\text{A.3})$$

Moreover, by condition 6, $\hat{\lambda}_{\text{ECR}}(\theta^*) \xrightarrow{p} \lambda(\theta^*) = \lambda^*$, and hence

$$P\left(\frac{1}{n} \sum_{i=1}^n \rho\left(\hat{\lambda}_{\text{ECR}}(\theta^*)' m(y_i, \theta^*)\right) > L + h\right) \rightarrow 0. \quad (\text{A.4})$$

Combination of (A.3) and (A.4) implies consistency of $\hat{\theta}_{\text{ECR}}$.

Using condition 6 again, by consistency of $\hat{\theta}_{\text{ECR}}$, we have

$$\frac{1}{n} \sum_{i=1}^n \rho\left(\lambda' m(y_i, \hat{\theta}_{\text{ECR}})\right) \xrightarrow{p} E[\rho(\lambda' m(y_i, \theta^*))]$$

uniformly in $\lambda \in \Lambda$, and thus $\hat{\lambda}_{\text{ECR}} = \hat{\lambda}_{\text{ECR}}(\hat{\theta}_{\text{ECR}}) \xrightarrow{p} \lambda^*$.

Next we show the asymptotic normality. Following Schennach (2007), we view the ECR estimator as a just-identified GMM estimator. The ECR estimator $\hat{\gamma}_{\text{ECR}} \equiv (\hat{\theta}'_{\text{ECR}}, \hat{\lambda}'_{\text{ECR}})'$ satisfies the following first order condition:

$$\frac{1}{n} \sum_{i=1}^n \phi(y_i, \hat{\gamma}_{\text{ECR}}) = 0. \quad (\text{A.5})$$

Since the above equation is just-identified, a standard theory of GMM (see e.g., Newey and McFadden (1994)) applies. Hence, by conditions 7-9, we have

$$\sqrt{n} \begin{pmatrix} \hat{\theta}_{\text{ECR}} - \theta^* \\ \hat{\lambda}_{\text{ECR}} - \lambda^* \end{pmatrix} \xrightarrow{d} N(0, H^{-1} S H^{-1}),$$

where

$$H = E\left[\frac{\partial \phi(y_i, \gamma^*)}{\partial \gamma'}\right] \quad \text{and} \quad S = E[\phi(y_i, \gamma^*) \phi(y_i, \gamma^*)'].$$

This completes the proof. \square

Proof of Lemma 4.1. Since $\lambda(\theta)$ is continuous in θ , $\hat{\lambda} = \lambda(\hat{\theta}_{\text{ECR}}) \xrightarrow{p} \lambda(\theta^*) = \lambda^*$ by consistency of $\hat{\theta}_{\text{ECR}}$. The first order condition is

$$\int f(y) \rho_1\left(\hat{\lambda}' m(y, \hat{\theta}_{\text{ECR}})\right) m(y, \hat{\theta}_{\text{ECR}}) dy = 0. \quad (\text{A.6})$$

Expanding (A.6) around $(\theta^{*'}, \lambda^{*'})'$ yields

$$\begin{aligned} 0 &= \left[\int f(y) \rho_2(\lambda^{*'} m(y, \theta^*)) m(y, \theta^*) m(y, \theta^*)' dy \right] (\hat{\lambda} - \lambda^*) \\ &\quad + \left[\int f(y) \rho_1(\lambda^{*'} m(y, \theta^*)) \frac{\partial m(y, \theta^*)}{\partial \theta'} dy + \int f(y) \rho_2(\lambda^{*'} m(y, \theta^*)) m(y, \theta^*) \lambda^{*'} \frac{\partial m(y, \theta^*)}{\partial \theta'} \right] \\ &\quad \times (\hat{\theta}_{\text{ECR}} - \theta^*) + o_p(n^{-1/2}). \end{aligned}$$

Hence we have

$$\begin{aligned} \sqrt{n} (\hat{\lambda} - \lambda^*) &= -E[\rho_{2i} m(y_i, \theta^*) m(y_i, \theta^*)']^{-1} (E[\rho_{1i} M_i] + E[\rho_{2i} m(y_i, \theta^*) \lambda^{*'} M_i]) \\ &\quad \times \sqrt{n} (\hat{\theta}_{\text{ECR}} - \theta^*) + o_p(1). \end{aligned}$$

and the result follows. \square

Proof of Theorem 4.1. Our proof is similar to that of Burnham and Anderson (2002), which derives the AIC and TIC in likelihood models.

Let $\hat{\theta}_{\text{ECR}}$ be the ECR estimator obtained by using n observations y_1, \dots, y_n . Let $\tilde{y}_1, \dots, \tilde{y}_n$ be an i.i.d. sample from $f(y)$ which is independent of y_1, \dots, y_n . Let $\tilde{\theta}_{\text{ECR}}$ be the ECR estimator calculated by using $\tilde{y}_1, \dots, \tilde{y}_n$. Also let

$$\tilde{\lambda} = \arg \max_{\lambda \in \Lambda} \int f(y) \rho(\lambda' m(y, \tilde{\theta}_{\text{ECR}})) dy.$$

Then we have

$$\begin{aligned} E[CRIC(f, \hat{g})] &= E \left[\int f(y) \rho(\hat{\lambda}' m(y, \hat{\theta}_{\text{ECR}})) dy \right] \\ &= E_{\tilde{y}} E_y \left[\frac{1}{n} \sum_{i=1}^n \rho(\hat{\lambda}' m(\tilde{y}_i, \hat{\theta}_{\text{ECR}})) \right] \\ &= E_y E_{\tilde{y}} \left[\frac{1}{n} \sum_{i=1}^n \rho(\tilde{\lambda}' m(y_i, \tilde{\theta}_{\text{ECR}})) \right] \\ &= E \left[\frac{1}{n} \sum_{i=1}^n \rho(\tilde{\lambda}' m(y_i, \tilde{\theta}_{\text{ECR}})) \right], \end{aligned}$$

where E_y and $E_{\tilde{y}}$ denote the expectations with respect to y_1, \dots, y_n and $\tilde{y}_1, \dots, \tilde{y}_n$, respectively.

The third equality follows by symmetry.

Expanding $n^{-1} \sum_{i=1}^n \rho(\tilde{\lambda}' m(y_i, \tilde{\theta}_{\text{ECR}}))$ around $(\hat{\theta}'_{\text{ECR}}, \hat{\lambda}'_{\text{ECR}})'$ yields

$$\begin{aligned} &\frac{1}{n} \sum_{i=1}^n \rho(\tilde{\lambda}' m(y_i, \tilde{\theta}_{\text{ECR}})) \\ &= \frac{1}{n} \sum_{i=1}^n \rho(\hat{\lambda}'_{\text{ECR}} m(y_i, \hat{\theta}_{\text{ECR}})) + \left[\frac{1}{n} \sum_{i=1}^n \phi(y_i, \hat{\gamma}_{\text{ECR}}) \right]' \begin{pmatrix} \tilde{\theta}_{\text{ECR}} - \hat{\theta}_{\text{ECR}} \\ \tilde{\lambda} - \hat{\lambda}_{\text{ECR}} \end{pmatrix} \\ &\quad + \frac{1}{2} \begin{pmatrix} \tilde{\theta}_{\text{ECR}} - \hat{\theta}_{\text{ECR}} \\ \tilde{\lambda} - \hat{\lambda}_{\text{ECR}} \end{pmatrix}' \frac{1}{n} \sum_{i=1}^n \frac{\partial \phi(y_i, \bar{\gamma})}{\partial \gamma'} \begin{pmatrix} \tilde{\theta}_{\text{ECR}} - \hat{\theta}_{\text{ECR}} \\ \tilde{\lambda} - \hat{\lambda}_{\text{ECR}} \end{pmatrix} \end{aligned} \quad (\text{A.7})$$

for some $\bar{\gamma} = (\bar{\theta}', \bar{\lambda}')'$ between $(\hat{\theta}'_{\text{ECR}}, \hat{\lambda}'_{\text{ECR}})'$ and $(\tilde{\theta}'_{\text{ECR}}, \tilde{\lambda}')'$. By the first order condition, the second term of (A.7) is 0. Also, we have

$$\begin{pmatrix} \tilde{\theta}_{\text{ECR}} - \hat{\theta}_{\text{ECR}} \\ \tilde{\lambda} - \hat{\lambda}_{\text{ECR}} \end{pmatrix} = \begin{pmatrix} \tilde{\theta}_{\text{ECR}} - \theta^* \\ \tilde{\lambda} - \lambda^* \end{pmatrix} - \begin{pmatrix} \hat{\theta}_{\text{ECR}} - \theta^* \\ \hat{\lambda}_{\text{ECR}} - \lambda^* \end{pmatrix}.$$

Hence, (A.7) reduces to

$$\begin{aligned} &\frac{1}{n} \sum_{i=1}^n \rho(\tilde{\lambda}' m(y_i, \tilde{\theta}_{\text{ECR}})) \\ &= \frac{1}{n} \sum_{i=1}^n \rho(\hat{\lambda}'_{\text{ECR}} m(y_i, \hat{\theta}_{\text{ECR}})) \\ &\quad + \frac{1}{2} \begin{pmatrix} \tilde{\theta}_{\text{ECR}} - \theta^* \\ \tilde{\lambda} - \lambda^* \end{pmatrix}' \hat{H} \begin{pmatrix} \tilde{\theta}_{\text{ECR}} - \theta^* \\ \tilde{\lambda} - \lambda^* \end{pmatrix} + \frac{1}{2} \begin{pmatrix} \hat{\theta}_{\text{ECR}} - \theta^* \\ \hat{\lambda}_{\text{ECR}} - \lambda^* \end{pmatrix}' \hat{H} \begin{pmatrix} \hat{\theta}_{\text{ECR}} - \theta^* \\ \hat{\lambda}_{\text{ECR}} - \lambda^* \end{pmatrix} \\ &\quad - \begin{pmatrix} \hat{\theta}_{\text{ECR}} - \theta^* \\ \hat{\lambda}_{\text{ECR}} - \lambda^* \end{pmatrix}' \hat{H} \begin{pmatrix} \tilde{\theta}_{\text{ECR}} - \theta^* \\ \tilde{\lambda} - \lambda^* \end{pmatrix}, \end{aligned} \quad (\text{A.8})$$

where

$$\hat{H} \equiv \frac{1}{n} \sum_{i=1}^n \frac{\partial \phi(y_i, \bar{\gamma})}{\partial \gamma'} \xrightarrow{p} H.$$

By Proposition 3.1 and Lemma 4.1,

$$\begin{aligned} \sqrt{n} \begin{pmatrix} \hat{\theta}_{\text{ECR}} - \theta^* \\ \hat{\lambda}_{\text{ECR}} - \lambda^* \end{pmatrix} &\xrightarrow{d} Z_1 \sim N(0, H^{-1} S H^{-1}) \\ \sqrt{n} \begin{pmatrix} \tilde{\theta}_{\text{ECR}} - \theta^* \\ \tilde{\lambda} - \lambda^* \end{pmatrix} &\xrightarrow{d} Z_2 \sim N(0, V). \end{aligned}$$

Note that Z_1 and Z_2 are independent. Hence for large n ,

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \rho(\tilde{\lambda}' m(y_i, \tilde{\theta}_{\text{ECR}})) &= \frac{1}{n} \sum_{i=1}^n \rho(\hat{\lambda}'_{\text{ECR}} m(y_i, \hat{\theta}_{\text{ECR}})) \\ &\quad + \frac{1}{2n} Z_1' H Z_1 + \frac{1}{2n} Z_2' H Z_2 + \frac{1}{n} Z_1' H Z_2 + o_p(n^{-1}). \end{aligned}$$

Taking the expectations,

$$\begin{aligned} &E[CRIC(f, \hat{g})] \\ &\simeq E \left[\frac{1}{n} \sum_{i=1}^n \rho(\hat{\lambda}'_{\text{ECR}} m(y_i, \hat{\theta}_{\text{ECR}})) \right] + \frac{1}{2n} \text{tr}(E[H Z_1 Z_1']) + \frac{1}{2n} \text{tr}(E[H Z_2 Z_2']) \\ &= E \left[\frac{1}{n} \sum_{i=1}^n \rho(\hat{\lambda}'_{\text{ECR}} m(y_i, \hat{\theta}_{\text{ECR}})) \right] + \frac{1}{2n} \text{tr}(H V) + \frac{1}{2n} \text{tr}(H^{-1} S), \end{aligned}$$

and thus we obtain the first result.

Next, we consider the correctly specified case. Let $m_i = m(y_i, \theta^*)$ for notational simplicity. Since $\rho_1(0) = \rho_2(0) = -1$, if the model is correctly specified, then

$$H^{-1} = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}$$

with

$$\begin{aligned} B_{11} &= (E[M'_i] E[m_i m'_i]^{-1} E[M_i])^{-1} \\ B_{12} &= (E[M'_i] E[m_i m'_i]^{-1} E[M_i])^{-1} E[M'_i] E[m_i m'_i]^{-1} \\ B_{21} &= E[m_i m'_i]^{-1} E[M_i] (E[M'_i] E[m_i m'_i]^{-1} E[M_i])^{-1} \\ B_{22} &= -E[m_i m'_i]^{-1} + E[m_i m'_i]^{-1} E[M_i] (E[M'_i] E[m_i m'_i]^{-1} E[M_i])^{-1} E[M'_i] E[m_i m'_i]^{-1}. \end{aligned}$$

Also,

$$S = \begin{pmatrix} 0 & 0 \\ 0 & E[m_i m'_i] \end{pmatrix}.$$

Hence we have

$$\begin{aligned}
\text{tr}(H^{-1}S) &= \text{tr} \left\{ \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & E[m_i m'_i] \end{pmatrix} \right\} \\
&= \text{tr}(B_{22}E[m_i m'_i]) \\
&= -\text{tr}(I_l) + \text{tr}(E[m_i m'_i]^{-1}E[M_i](E[M'_i]E[m_i m'_i]^{-1}E[M_i])^{-1}E[M'_i]) \\
&= p - l.
\end{aligned}$$

Also, under correct specification,

$$\sqrt{n} \begin{pmatrix} \tilde{\theta}_{\text{ECR}} - \theta^* \\ \tilde{\lambda} - \lambda^* \end{pmatrix} \xrightarrow{d} N \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix} \right),$$

where

$$\begin{aligned}
V_{11} &= (E[M'_i]E[m_i m'_i]^{-1}E[M_i])^{-1} \\
V_{12} &= -(E[M'_i]E[m_i m'_i]^{-1}E[M_i])^{-1}E[M'_i]E[m_i m'_i]^{-1} \\
V_{21} &= -E[m_i m'_i]^{-1}E[M_i](E[M'_i]E[m_i m'_i]^{-1}E[M_i])^{-1} \\
V_{22} &= E[m_i m'_i]^{-1}E[M_i](E[M'_i]E[m_i m'_i]^{-1}E[M_i])^{-1}E[M'_i]E[m_i m'_i]^{-1}.
\end{aligned}$$

Hence,

$$\begin{aligned}
\text{tr}(HV) &= \text{tr} \left\{ \begin{pmatrix} 0 & -E[M'_i] \\ -E[M_i] & -E[m_i m'_i] \end{pmatrix} \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix} \right\} \\
&= \text{tr} \begin{pmatrix} -E[M'_i]V_{21} & -E[M'_i]V_{22} \\ -E[M_i]V_{11} - E[m_i m'_i]V_{21} & -E[M_i]V_{12} - E[m_i m'_i]V_{22} \end{pmatrix} \\
&= \text{tr} (E[M'_i]E[m_i m'_i]^{-1}E[M_i](E[M'_i]E[m_i m'_i]^{-1}E[M_i])^{-1}) \\
&\quad + \text{tr} (E[M_i](E[M'_i]E[m_i m'_i]^{-1}E[M_i])^{-1}E[M'_i]E[m_i m'_i]^{-1}) \\
&\quad - \text{tr} (E[m_i m'_i]E[m_i m'_i]^{-1}E[M_i](E[M'_i]E[m_i m'_i]^{-1}E[M_i])^{-1}E[M'_i]E[m_i m'_i]^{-1}) \\
&= p.
\end{aligned}$$

Therefore, we have

$$E[CRIC(f, \hat{g})] \simeq E \left[\frac{1}{n} \sum_{i=1}^n \rho \left(\hat{\lambda}'_{\text{ECR}} m(y_i, \hat{\theta}_{\text{ECR}}) \right) \right] - \frac{1}{2n}(l - 2p),$$

and the result follows. \square

Proof of Proposition 4.1. Let $\hat{\gamma}_{(-j)}$ denote the ECR estimator which is calculated based on $n - 1$ observations excluding y_j . That is

$$\hat{\gamma}_{(-j)} = \left(\hat{\theta}'_{(-j)}, \hat{\lambda}'_{(-j)} \right)' = \arg \min_{\theta \in \Theta} \arg \max_{\lambda \in \Lambda} \sum_{i \neq j} \rho(\lambda' m(y_i, \theta)).$$

Then, by the first order condition, we have

$$0 = \sum_{i \neq j} \phi(y_i, \hat{\gamma}_{(-j)}) = \sum_{i=1}^n \phi(y_i, \hat{\gamma}_{(-j)}) - \phi(y_j, \hat{\gamma}_{(-j)})$$

and hence

$$\phi(y_j, \hat{\gamma}_{(-j)}) = \sum_{i=1}^n \phi(y_i, \hat{\gamma}_{(-j)}). \quad (\text{A.9})$$

Let $\hat{\gamma}_{\text{ECR}}$ be the ECR estimator calculated by using all observation. By expanding the right hand side of (A.9) around $\hat{\gamma}_{\text{ECR}}$,

$$\begin{aligned} \frac{1}{n} \phi(y_j, \hat{\gamma}_{(-j)}) &= \frac{1}{n} \sum_{i=1}^n \phi(y_i, \hat{\gamma}_{\text{ECR}}) + \frac{1}{n} \sum_{i=1}^n \frac{\partial \phi(y_i, \hat{\gamma}_{\text{ECR}})}{\partial \gamma'} (\hat{\gamma}_{(-j)} - \hat{\gamma}_{\text{ECR}}) \\ &\quad + O_p(\|\hat{\gamma}_{(-j)} - \hat{\gamma}_{\text{ECR}}\|^2). \end{aligned}$$

Thus we obtain

$$\begin{aligned} \hat{\gamma}_{(-j)} - \hat{\gamma}_{\text{ECR}} &= \left[\frac{1}{n} \sum_{i=1}^n \frac{\partial \phi(y_i, \hat{\gamma}_{\text{ECR}})}{\partial \gamma'} \right]^{-1} \frac{1}{n} \phi(y_j, \hat{\gamma}_{(-j)}) + O_p(\|\hat{\gamma}_{(-j)} - \hat{\gamma}_{\text{ECR}}\|^2) \\ &= \left[\frac{1}{n} \sum_{i=1}^n \frac{\partial \phi(y_i, \hat{\gamma}_{\text{ECR}})}{\partial \gamma'} \right]^{-1} \frac{1}{n} \phi(y_j, \hat{\gamma}_{\text{ECR}}) + O_p(n^{-3/2}), \end{aligned}$$

where the last equality follows from the fact that $\|\hat{\gamma}_{(-j)} - \hat{\gamma}_{\text{ECR}}\| = O_p(n^{-1})$ and hence the difference between $n^{-1}\phi(y_j, \hat{\gamma}_{(-j)})$ and $n^{-1}\phi(y_j, \hat{\gamma}_{\text{ECR}})$ is negligible. Therefore, we have

$$\begin{aligned} &\frac{1}{n} \sum_{i=1}^n \rho(\hat{\lambda}'_{(-i)} m(y_i, \hat{\theta}_{(-i)})) \\ &= \frac{1}{n} \sum_{i=1}^n \rho(\hat{\lambda}'_{\text{ECR}} m(y_i, \hat{\theta}_{\text{ECR}})) + \frac{1}{n} \sum_{i=1}^n \phi(y_i, \hat{\gamma}_{\text{ECR}})' (\hat{\gamma}_{(-i)} - \hat{\gamma}_{\text{ECR}}) + O_p(n^{-3/2}) \\ &= \frac{1}{n} \sum_{i=1}^n \rho(\hat{\lambda}'_{\text{ECR}} m(y_i, \hat{\theta}_{\text{ECR}})) + \frac{1}{n} \sum_{i=1}^n \phi(y_i, \hat{\gamma}_{\text{ECR}})' \left[\frac{1}{n} \sum_{i=1}^n \frac{\partial \phi(y_i, \hat{\gamma}_{\text{ECR}})}{\partial \gamma'} \right]^{-1} \frac{1}{n} \phi(y_i, \hat{\gamma}_{\text{ECR}}) \\ &\quad + O_p(n^{-2/3}) \\ &= \frac{1}{n} \sum_{i=1}^n \rho(\hat{\lambda}'_{\text{ECR}} m(y_i, \hat{\theta}_{\text{ECR}})) + \frac{1}{n} \text{tr} \left\{ \left[\frac{1}{n} \sum_{i=1}^n \frac{\partial \phi(y_i, \hat{\gamma}_{\text{ECR}})}{\partial \gamma'} \right]^{-1} \frac{1}{n} \sum_{i=1}^n \phi(y_i, \hat{\gamma}_{\text{ECR}}) \phi(y_i, \hat{\gamma}_{\text{ECR}})' \right\} \\ &\quad + O_p(n^{-2/3}) \\ &= \frac{1}{n} \sum_{i=1}^n \rho(\hat{\lambda}'_{\text{ECR}} m(y_i, \hat{\theta}_{\text{ECR}})) + \frac{1}{n} \text{tr}(\widehat{H^{-1}S}) + O_p(n^{-2/3}). \end{aligned}$$

This completes the proof. \square

References

- AKAIKE, H. (1970): ‘‘Statistical Predictor Identification,’’ *Annals of the Institute of Statistical Mathematics*, 22, 203–217.
- (1973): ‘‘Information Theory and an Extension of the Maximum Likelihood Principle,’’ in *Second International Symposium on Information Theory*, ed. by B. Petroc, and F. Csake, pp. 267–281. Akademiai Kiado.

- ANDREWS, D. W. (1999): “Consistent Moment Selection Procedures for Generalized Method of Moments Estimation,” *Econometrica*, 67, 543–564.
- ANDREWS, D. W., AND B. LU (2001): “Consistent Models and Moment Selection Procedures for GMM Estimation with Application to Dynamic Panel Data Models,” *Journal of Econometrics*, 101, 123–164.
- BAGGERLY, K. A. (1998): “Empirical Likelihood as a Goodness-of-Fit Measure,” *Biometrika*, 85, 535–547.
- BORWEIN, J. M., AND A. S. LEWIS (1991): “Duality Relationships for Entropy-like Minimization Problems,” *SIAM Journal of Control and Optimization*, 29, 325–338.
- BURNHAM, K. P., AND D. R. ANDERSON (2002): *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer.
- CHEN, X., H. HONG, AND M. SHUM (2007): “Nonparametric Likelihood Ratio Model Selection Tests Between Parametric Likelihood and Moment Conditional Models,” *Journal of Econometrics*, 141, 109–140.
- CHRISTOFFERSEN, P., J. HAHN, AND A. INOUE (2001): “Testing, Comparing and Combining Value-at-Risk Measure,” *Journal of Empirical Finance*, 8, 325–342.
- CRESSIE, N., AND T. R. C. READ (1984): “Multinomial Goodness-of-Fit Tests,” *Journal of the Royal Statistical Society. Series B*, 46, 440–464.
- HALL, A. R., AND A. INOUE (2003): “The Large Sample Behaviour of the Generalized Method of Moments Estimator in Misspecified Models,” *Journal of Econometrics*, 114, 361–394.
- HALL, A. R., A. INOUE, K. JANA, AND C. SHIN (2007): “Information in Generalized Method of Moments Estimation and Entropy-based Moment Selection,” *Journal of Econometrics*, 138, 488–512.
- HONG, H., B. PRESTON, AND M. SHUM (2003): “Generalized Empirical Likelihood-Based Model Selection Criteria for Moment Condition Models,” *Econometric Theory*, 19, 923–943.
- IMBENS, G. W., R. H. SPADY, AND P. JOHNSON (1998): “Information Theoretic Approaches to Inference in Moment Condition Models,” *Econometrica*, 66, 333–357.
- KITAMURA, Y. (2006): “Empirical Likelihood Methods in Econometrics: Theory and Practice,” Working Paper, Yale University.
- KITAMURA, Y., AND M. STUTZER (1997): “An Information-Theoretic Alternative to Generalized Method of Moments Estimation,” *Econometrica*, 65, 861–874.
- KONISHI, S., AND G. KITAGAWA (1996): “Generalized Information Criteria in Model Selection,” *Biometrika*, 83, 875–890.

- LEEB, H., AND B. M. POTSCHER (2005): “Model Selection and Inference: Facts and Fiction,” *Econometric Theory*, 21, 21–59.
- NEWKEY, W. K., AND D. L. MCFADDEN (1994): “Large Sample Estimation and Hypothesis Testing,” in *Handbook of Econometrics*, ed. by R. F. Engle, and D. L. McFadden, vol. IV, pp. 2111–2245. Amsterdam: Elsevier.
- NEWKEY, W. K., AND J. L. POWELL (2003): “Instrumental Variable Estimation of Nonparametric Models,” *Econometrica*, 71, 1565–1578.
- NEWKEY, W. K., AND R. J. SMITH (2004): “Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators,” *Econometrica*, 72, 219–255.
- QIN, J., AND J. LAWLESS (1994): “Empirical Likelihood and General Estimating Equations,” *Annals of Statistics*, 22, 300–325.
- RAMALHO, J. J., AND R. J. SMITH (2002): “Generalized Empirical Likelihood Non-nested Tests,” *Journal of Econometrics*, 107, 99–125.
- SCHENNACH, S. M. (2007): “Point Estimation with Exponentially Tilted Empirical Likelihood,” *Annals of Statistics*, 35, 634–672.
- SCHWARZ, G. (1978): “Estimating the Dimension of a Model,” *Annals of Statistics*, 6, 461–464.
- SHIBATA, R. (1984): “Approximate Efficiency of a Selection Procedure for the Number of Regression Variables,” *Biometrika*, 71, 43–49.
- (1989): “Statistical Aspects of Model Selection,” in *From Data to Model*, ed. by J. C. Willems, pp. 215–240. Springer-Verlag New York.
- SIN, C., AND H. WHITE (1996): “Information Criteria for Selecting Possibly Misspecified Parametric Models,” *Journal of Econometrics*, 71, 207–225.
- SMITH, R. J. (1997): “Alternative Semi-Parametric Likelihood Approaches to Generalised Method of Moments Estimation,” *Economic Journal*, 107, 503–519.
- STONE, M. (1977): “An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike’s Criterion,” *Journal of the Royal Statistical Society. Series B*, 39, 44–47.
- TAKEUCHI, K. (1976): “Distribution of information Statistics and Criteria for Adequacy of Models,” *Mathematical Science*, 153, 12–18, in Japanese.
- WHITE, H. (1982): “Maximum Likelihood Estimation for Misspecified Models,” *Econometrica*, 68, 1097–1126.