Regularization parameter selection for penalized empirical likelihood estimator

Ando, Tomohiro

Sueishi, Naoya

(Citation)
Economics Letters,178:1-4

(Issue Date)
2019-05

(Resource Type)
journal article

(Version)
Accepted Manuscript

(URL)
https://hdl.handle.net/20.500.14094/90005996

Regularization Parameter Selection for Penalized Empirical Likelihood Estimator

Tomohiro Ando

Melbourne Business School, University of Melbourne, 200 Leicester Street, Carlton, Victoria 3053, Australia.

Naoya Sueishi*

Graduate School of Economics, Kobe University, 2-1 Rokkodai-cho, Nada-ku, Kobe 657-8501, Japan.
Phone:+81-78-803-6827.

Abstract

Penalized estimation is a useful technique for variable selection when the number of candidate variables is large. A crucial issue in penalized estimation is the selection of the regularization parameter because the performance of the estimator largely depends on an appropriate choice. However, no theoretically sound selection method currently exists for the penalized estimation of moment restriction models. To address this important issue, we develop a novel information criterion, which we call the empirical likelihood information criterion, to select the regularization parameter of the penalized empirical likelihood estimator. The information criterion is derived as an estimator of the expected value of the Kullback–Leibler information criterion from an estimated model to the true data generating process.

Keywords: Information criterion, Variable selection.

*Corresponding author
Email addresses: T.Ando@mbs.edu. (Tomohiro Ando), sueishi@econ.kobe-u.ac.jp (Naoya Sueishi)

1. Introduction

Variable selection has long been an important issue in empirical analysis. Empirical researchers typically have a large number of candidate explanatory variables from which they select the appropriate ones considering restriction of the sample size and/or a preference for a parsimonious model.

Subset variable selection methods using information criteria were originally investigated by Andrews and Lu (2001) and Hong, Preston, and Shum (2003) for moment restriction models. Andrews and Lu (2001) proposed criteria that resemble the Akaike information criterion (AIC; Akaike 1973), the Bayesian information criterion (BIC; Schwarz 1978), and the Hannan-Quinn criterion (Hannan and Quinn 1979) based on the J-statistic of the GMM estimator. Hong, Preston, and Shum (2003) replaced the J-statistic with the generalized empirical likelihood (EL) statistic. Although these criteria have attractive properties, they suffer from computational burden when the number of candidate variables is large.

In the wake of the success of the penalized least squares estimator as a method of variable selection (Tibshirani 1996; Fan and Li 2001; Zou and Hastie 2005; Zou 2006; Zhang 2010), penalized GMM and EL estimators have been proposed in the econometrics literature. Caner (2009) and Shi (2016) considered the penalized GMM estimator with a Lasso-type penalty. Caner and Zhang (2014) proposed the adaptive elastic net GMM estimator. Leng and Tang (2012) and Chang, Chen, and Chen (2015) studied the penalized empirical likelihood (PEL) estimator for independent and weakly dependent observations, respectively. Chang, Tang, and Wu (2018) proposed an alternative PEL estimator that imposes a penalty for both the parameter of interest and the Lagrange multiplier of the EL estimator.

Most of the existing studies investigate the asymptotic properties of the penalized GMM and EL estimators under the assumption that the regularization parameter converges to zero at an appropriate rate. However, asymptotic theory does not guide the selection of the regularization parameter in an actual implementation. Because the performance of the penalized estimators largely depends on the choice of the regularization parameter, providing a data-dependent selection method for the regularization parameter is a crucial issue.

The purpose of this paper is to propose a new information criterion, which we call the empirical likelihood information criterion (ELIC), for selecting the regularization parameter of the PEL estimator. Our goal is to find a model that provides a good approximation rather than one that is correct. The idea behind our information-theoretic approach has its origin in Akaike (1973). We express a moment restriction model in the form of a set of probability distributions and then evaluate the goodness of the model using an estimate of the Kullback–Leibler information criterion (KLIC) from the estimated model to the true data generating process (DGP). Our information criterion is derived as a bias-corrected estimator of the KLIC.

There are a few studies related to ours. Shi (2016) proposed AIC- and BIC-type criteria for the selection of the regularization parameter of the penalized GMM estimator; these are modifications of the criteria of Andrews and Lu (2001) and Wang, Li, and Leng (2009). Leng and Tang (2012) also suggested a similar BIC-type criterion for the PEL estimator. However, the asymptotic properties of the modified BICs are unknown in the case of penalized GMM and EL estimation.

2. PEL Estimator and KLIC

Let $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ be i.i.d. random vectors from an unknown distribution μ. The parameter of interest, $\boldsymbol{\theta}_0 \in \Theta \subset \mathbb{R}^p$, is identified by a set of moment restrictions:

$$E\left[\boldsymbol{m}(\boldsymbol{y}_i, \boldsymbol{\theta}_0)\right] = \int \boldsymbol{m}(\boldsymbol{y}, \boldsymbol{\theta}_0) d\mu(\boldsymbol{y}) = \mathbf{0}, \tag{2.1}$$

where $\boldsymbol{m} : \mathbb{R}^d \times \Theta \to \mathbb{R}^r$ is a known vector-valued function with $r > p$. Some elements of $\boldsymbol{\theta}_0$ may be zero. If there is no value of $\boldsymbol{\theta}_0 \in \Theta$ that satisfies all moment restrictions, then the model is misspecified.

We employ the PEL estimator of Leng and Tang (2012) to simultaneously achieve variable selection and estimation. For a given value of a regularization parameter κ, the PEL estimator for model (2.1) is defined as

$$(\hat{\boldsymbol{\theta}}_\kappa, \hat{\boldsymbol{\lambda}}_\kappa) = \arg\min_{\boldsymbol{\theta}\in\Theta} \arg\max_{\boldsymbol{\lambda}\in\hat{\Lambda}(\boldsymbol{\theta})} \left\{ \frac{1}{n}\sum_{i=1}^{n} \log(1 - \boldsymbol{\lambda}'\boldsymbol{m}(\boldsymbol{y}_i, \boldsymbol{\theta})) + \sum_{j=1}^{p} p_\kappa(\theta_j) \right\},$$

where $\hat{\Lambda}(\boldsymbol{\theta}) = \{\boldsymbol{\lambda} \in \mathbb{R}^r : \boldsymbol{\lambda}'\boldsymbol{m}(\boldsymbol{y}_i, \boldsymbol{\theta}) < 1, i = 1, \ldots, n\}$, θ_j is the j-th element of $\boldsymbol{\theta}$, and $p_\kappa(\cdot)$ is a penalty function with the regularization parameter κ. The candidates of the penalty function include the L_1 penalty, the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li 2001), and the minimax concave penalty (Zhang 2010), for instance.

To find an optimal regularization parameter, we need a criterion to evaluate the performance of the PEL estimator. For this purpose, we introduce the KLIC for the moment restriction model. Let $\mathbf{M}$ denote the set of all probability measures on $\mathbb{R}^d$. For each $\boldsymbol{\theta} \in \Theta$, we define $\mathcal{P}_{\boldsymbol{\theta}} = \{P \in \mathbf{M} : \int \boldsymbol{m}(\boldsymbol{y}, \boldsymbol{\theta}) dP(\boldsymbol{y}) = \mathbf{0}\}$. Then, we define $\mathcal{P} = \cup_{\boldsymbol{\theta} \in \Theta} \mathcal{P}_{\boldsymbol{\theta}}$ as the moment restriction model. The KLIC from $\mathcal{P}$ to μ is defined as $K(\mu \| \mathcal{P}) = \min_{P \in \mathcal{P}} K(\mu \| P)$, where

$$K(\mu \| P) = \begin{cases} -\int \log\left(\frac{dP}{d\mu}(\boldsymbol{y})\right) d\mu(\boldsymbol{y}) & \text{if } P \ll \mu \\ \infty & \text{otherwise.} \end{cases}$$

By a duality theorem, the KLIC is equivalently characterized as

$$K(\mu \| \mathcal{P}) = \min_{\boldsymbol{\theta} \in \Theta} \int \log\left(1 - \boldsymbol{\lambda}(\boldsymbol{\theta})' \boldsymbol{m}(\boldsymbol{y}, \boldsymbol{\theta})\right) d\mu(\boldsymbol{y}),$$

where $\boldsymbol{\lambda}(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\lambda} \in \Lambda(\boldsymbol{\theta})} \int \log(1 - \boldsymbol{\lambda}' \boldsymbol{m}(\boldsymbol{y}, \boldsymbol{\theta})) d\mu(\boldsymbol{y})$ and $\Lambda(\boldsymbol{\theta}) \subset \mathbb{R}^r$ is a set of possible values of $\boldsymbol{\lambda}$ (see, for instance, Chen, Hong, and Shum 2007).

It is known that $\boldsymbol{\theta}_0$ minimizes

$$\int \log\left(1 - \boldsymbol{\lambda}(\boldsymbol{\theta})' \boldsymbol{m}(\boldsymbol{y}, \boldsymbol{\theta})\right) d\mu(\boldsymbol{y}). \tag{2.2}$$

Minimizing the sample analog of (2.2) yields the EL estimator of Qin and Lawless (1994). It can be verified that $K(\mu \| \mathcal{P}) \geq 0$ and $K(\mu \| \mathcal{P}) = 0$ if and only if $\mu \in \mathcal{P}$. Thus, the KLIC defines a pseudo-distance between the model and the DGP. If the model is misspecified, then the minimizer of (2.2), which we denote $\boldsymbol{\theta}^*$, is called the pseudo-true parameter value of the EL estimator.

The above argument suggests to use the following quantity to evaluate the PEL estimator:

$$K(\mu \| \hat{P}_\kappa) = \int \log\left(1 - \hat{\boldsymbol{\lambda}}_\kappa' \boldsymbol{m}(\boldsymbol{y}, \hat{\boldsymbol{\theta}}_\kappa)\right) d\mu(\boldsymbol{y}), \tag{2.3}$$

where

$$\frac{d\hat{P}_\kappa}{d\mu}(\boldsymbol{y}) = \frac{1}{1 - \hat{\boldsymbol{\lambda}}_\kappa' \boldsymbol{m}(\boldsymbol{y}, \hat{\boldsymbol{\theta}}_\kappa)}.$$

The quantity (2.3) is interpreted as the KLIC from the estimated model to the DGP. We view (2.3) as a loss for using κ and find the regularization parameter that minimizes an estimate of the risk, that is, the expected value of the loss.

3. Empirical Likelihood Information Criterion

A simple estimator of the risk is

$$\frac{1}{n}\sum_{i=1}^{n}\log\left(1-\hat{\boldsymbol{\lambda}}_{\kappa}'\boldsymbol{m}(\boldsymbol{y}_i,\hat{\boldsymbol{\theta}}_{\kappa})\right). \tag{3.1}$$

However, (3.1) is a biased estimator because $(\hat{\boldsymbol{\theta}}_{\kappa},\hat{\boldsymbol{\lambda}}_{\kappa})$ is correlated with $\boldsymbol{y}_i$. We develop an information criterion by obtaining a bias-corrected estimator of the risk.

Our bias correction idea for deriving the information criterion is the same as that for deriving the AIC although Akaike (1973) used a parametric likelihood instead of an EL. Akaike (1973) derived the bias under the assumptions that (i) the parametric model is correctly specified and (ii) the parameter is estimated by the maximum likelihood estimator. Konishi and Kitagawa (1996) provided a general bias correction method that relaxes both conditions (i) and (ii) although the model must be parametric. We extend the result of Konishi and Kitagawa (1996) so that the method can be applied for the moment restriction model.

In the following discussion, we consider an asymptotic framework in which the sample size goes to infinity for each fixed κ. We do not a priori assume that κ converges to zero at a certain rate because we aim to select κ in a data-driven way. Moreover, we assume that r and p are fixed and finite. However, our information criterion does not change even if r and p increase slowly with the sample size as long as the main term of the bias does not change.

Let $\mathcal{K}$ be a set of possible values of κ. For each $\kappa\in\mathcal{K}$ and $P\in\mathbf{M}$, we define the statistical functional of the PEL estimator:

$$(\boldsymbol{T}^{P}_{\theta,\kappa}(P),\boldsymbol{T}^{P}_{\lambda,\kappa}(P))=\arg\min_{\boldsymbol{\theta}\in\Theta}\arg\max_{\boldsymbol{\lambda}\in\Lambda(\boldsymbol{\theta})}\left\{E_P\left[\log\left(1-\boldsymbol{\lambda}'\boldsymbol{m}(\boldsymbol{y}_i,\boldsymbol{\theta})\right)\right]+\sum_{j=1}^{p}p_{\kappa}(\theta_j)\right\},$$

where E_P denotes the expected value with respect to P. Moreover, we write $\boldsymbol{T}^{P}_{\kappa}(P)=(\boldsymbol{T}^{P}_{\theta,\kappa}(P)',\boldsymbol{T}^{P}_{\lambda,\kappa}(P)')'$. The PEL estimator is written as $(\hat{\boldsymbol{\theta}}_{\kappa},\hat{\boldsymbol{\lambda}}_{\kappa})=(\boldsymbol{T}^{\hat{\mu}}_{\theta,\kappa}(\hat{\mu}),\boldsymbol{T}^{\hat{\mu}}_{\lambda,\kappa}(\hat{\mu}))$, where $\hat{\mu}$ is the empirical distribution of $\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n$. For a fixed value of κ, the estimand of the PEL estimator is $(\boldsymbol{\theta}^{*}_{\kappa},\boldsymbol{\lambda}^{*}_{\kappa})=(\boldsymbol{T}^{\mu}_{\theta,\kappa}(\mu),\boldsymbol{T}^{\mu}_{\lambda,\kappa}(\mu))$.

In a usual notation of a statistical functional, there is no superscript P and the functional is simply written as $\boldsymbol{T}(\cdot)$. We use the notation $\boldsymbol{T}^{P}_{\kappa}(\cdot)$

to indicate that the functional of a sparse estimator depends on P (Öllerer, Croux, and Alfons 2015). This is because zero components in $\boldsymbol{T}_\kappa^P(P)$ are generally different from those of $\boldsymbol{T}_\kappa^Q(Q)$ when $P \neq Q$.

Let $\boldsymbol{T}_\kappa(\cdot) = \boldsymbol{T}_\kappa^\mu(\cdot)$ and $\hat{\boldsymbol{T}}_\kappa(\cdot) = \boldsymbol{T}_\kappa^{\hat\mu}(\cdot)$. We impose the following conditions.

(A.1) For all $\kappa \in \mathcal{K}$, $\hat{\boldsymbol{T}}_\kappa(\cdot) = \boldsymbol{T}_\kappa(\cdot)$ with probability approaching one.

(A.2) For all $\kappa \in \mathcal{K}$, $\boldsymbol{T}_\kappa(\hat\mu)$ has the following von Mises expansion:

$$\boldsymbol{T}_\kappa(\hat\mu) = \boldsymbol{T}_\kappa(\mu) + \frac{1}{n}\sum_{j=1}^n \boldsymbol{T}_\kappa^{(1)}(\boldsymbol{y}_j;\mu) + \frac{1}{2n^2}\sum_{j=1}^n\sum_{k=1}^n \boldsymbol{T}_\kappa^{(2)}(\boldsymbol{y}_j,\boldsymbol{y}_k;\mu) + \boldsymbol{r}_{\kappa,n}$$

with $\|\boldsymbol{r}_{\kappa,n}\| = o_p(n^{-1})$.

(A.3) $\boldsymbol{m}(\boldsymbol{y},\boldsymbol{\theta})$ is twice continuously differentiable with respect to $\boldsymbol{\theta}$.

(A.4) $p_\kappa(\theta)$ is twice continuously differentiable in θ for $|\theta| > 0$.

Roughly speaking, condition (A.1) is satisfied if the zero component in $\hat{\boldsymbol{\theta}}_\kappa$ coincides with the zero components in $\boldsymbol{\theta}_\kappa^*$ with probability approaching one. This condition is satisfied for commonly used penalty functions, such as the L_1 penalty and the SCAD penalty, if $n^{-1}\sum_{i=1}^n \log(1-\boldsymbol{\lambda}'\boldsymbol{m}(\boldsymbol{y}_i,\boldsymbol{\theta}))$ converges in probability to $E[\log(1-\boldsymbol{\lambda}'\boldsymbol{m}(\boldsymbol{y}_i,\boldsymbol{\theta}))]$ uniformly over $\boldsymbol{\theta} \in \Theta$ and $\boldsymbol{\lambda} \in \Lambda(\boldsymbol{\theta})$.

In condition (A.2), $\boldsymbol{T}_\kappa^{(1)}(\boldsymbol{y};\mu)$ and $\boldsymbol{T}_\kappa^{(2)}(\boldsymbol{y},\boldsymbol{z};\mu)$ are given by the Gâteaux derivatives of $\boldsymbol{T}_\kappa(\cdot)$:

$$\boldsymbol{T}_\kappa^{(1)}(\boldsymbol{y};\mu) = \left.\frac{d}{dt}\boldsymbol{T}_\kappa((1-t)\mu + t\delta_{\boldsymbol{y}})\right|_{t=0}$$

and

$$\boldsymbol{T}_\kappa^{(2)}(\boldsymbol{y},\boldsymbol{z};\mu) = \left.\frac{d^2}{dtds}\boldsymbol{T}_\kappa((1-s-t)\mu + t\delta_{\boldsymbol{y}} + s\delta_{\boldsymbol{z}}))\right|_{t=0,s=0},$$

where $\delta_{\boldsymbol{y}}$ is the point mass at $\boldsymbol{y}$. Notice that $\boldsymbol{T}_\kappa^{(1)}(\boldsymbol{y};\mu)$ is the influence function of $\boldsymbol{T}_\kappa(\cdot)$ at μ. The explicit forms of $\boldsymbol{T}_\kappa^{(1)}(\boldsymbol{y};\mu)$ and $\boldsymbol{T}_\kappa^{(2)}(\boldsymbol{y},\boldsymbol{z};\mu)$ are given, for instance, by Gatto and Ronchetti (1996) and La Vecchia, Ronchetti, and Trojani (2012) for the M-estimator.

Let $\boldsymbol{\gamma} = (\boldsymbol{\theta}',\boldsymbol{\lambda}')'$ and $\boldsymbol{\phi}(\boldsymbol{y},\gamma) = \partial\log(1-\boldsymbol{\lambda}'\boldsymbol{m}(\boldsymbol{y},\boldsymbol{\theta}))/\partial\boldsymbol{\gamma}$. We obtain the following theorem.

Theorem 3.1. *Suppose that conditions (A.1)–(A.4) hold. Then, for each* $\kappa \in \mathcal{K}$*, we obtain*

$$\frac{1}{n}\sum_{i=1}^{n}\log\left(1-\hat{\boldsymbol{\lambda}}_{\kappa}'\boldsymbol{m}(\boldsymbol{y}_i,\hat{\boldsymbol{\theta}}_{\kappa})\right)-\int\log\left(1-\hat{\boldsymbol{\lambda}}_{\kappa}'\boldsymbol{m}(\boldsymbol{y},\hat{\boldsymbol{\theta}}_{\kappa})\right)d\mu(\boldsymbol{y})=\frac{1}{n}b_{\kappa}+o_p\left(\frac{1}{n}\right),$$

where b_κ *satisfies*

$$E[b_\kappa]=\mathrm{tr}\left(\int \boldsymbol{T}_{\kappa}^{(1)}(\boldsymbol{y};\mu)\boldsymbol{\phi}(\boldsymbol{y},\boldsymbol{T}_{\kappa}(\mu))'d\mu(\boldsymbol{y})\right). \tag{3.2}$$

The proof is given in the supplemental material.

The ELIC is obtained by estimating $E[b_\kappa]$. We estimate $E[b_\kappa]$ by a sample analog of (3.2):

$$\hat{b}_\kappa=\mathrm{tr}\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{T}}_{\kappa}^{(1)}(\boldsymbol{y}_i;\hat{\mu})\boldsymbol{\phi}(\boldsymbol{y}_i,\hat{\boldsymbol{T}}_{\kappa}(\hat{\mu}))'\right).$$

To obtain $\hat{\boldsymbol{T}}_{\kappa}^{(1)}(\boldsymbol{y};\hat{\mu})$, we invoke the fact that the nonzero components of the PEL estimator can be written in the form of the M-estimator. Let q_κ be the number of nonzero components in $\hat{\boldsymbol{\theta}}_\kappa$ and write $\hat{\boldsymbol{\theta}}_\kappa=(\hat{\boldsymbol{\theta}}_{1\kappa}',\hat{\boldsymbol{\theta}}_{2\kappa}')'=(\hat{\boldsymbol{\theta}}_{1\kappa}',\boldsymbol{0}')'\in\mathbb{R}^{q_\kappa}\times\mathbb{R}^{p-q_\kappa}$. Let $\boldsymbol{\theta}_{1\kappa}$ be the $q_\kappa\times 1$ sub-vector of $\boldsymbol{\theta}$ whose elements are estimated to be nonzero. Moreover, let $\boldsymbol{m}(\boldsymbol{y}_i,\boldsymbol{\theta}_{1\kappa})=\boldsymbol{m}(\boldsymbol{y}_i,(\boldsymbol{\theta}_{1\kappa}',\boldsymbol{0}')')$ and $M_\kappa(\boldsymbol{y}_i,\boldsymbol{\theta}_{1\kappa})=\partial\boldsymbol{m}(\boldsymbol{y}_i,\boldsymbol{\theta}_{1\kappa})/\partial\boldsymbol{\theta}_{1\kappa}'$. Then, $\hat{\boldsymbol{\gamma}}_{1\kappa}=(\hat{\boldsymbol{\theta}}_{1\kappa}',\hat{\boldsymbol{\lambda}}_{\kappa}')'$ solves the first-order condition:

$$\boldsymbol{0}=\sum_{i=1}^{n}\boldsymbol{\psi}_{\kappa}(\boldsymbol{y}_i,\hat{\boldsymbol{\gamma}}_{1\kappa})=\sum_{i=1}^{n}\begin{pmatrix}-\frac{M_\kappa(\boldsymbol{y}_i,\hat{\boldsymbol{\theta}}_{1\kappa})'\hat{\boldsymbol{\lambda}}_\kappa}{1-\hat{\boldsymbol{\lambda}}_{\kappa}'\boldsymbol{m}(\boldsymbol{y}_i,\hat{\boldsymbol{\theta}}_{1\kappa})}+\boldsymbol{r}(\hat{\boldsymbol{\theta}}_{1\kappa})\\ -\frac{\boldsymbol{m}(\boldsymbol{y}_i,\hat{\boldsymbol{\theta}}_{1\kappa})}{1-\hat{\boldsymbol{\lambda}}_{\kappa}'\boldsymbol{m}(\boldsymbol{y}_i,\hat{\boldsymbol{\theta}}_{1\kappa})}\end{pmatrix},$$

where $\boldsymbol{r}(\hat{\boldsymbol{\theta}}_{1\kappa})=(p_{\kappa}'(\hat{\theta}_{1\kappa 1}),\ldots,p_{\kappa}'(\hat{\theta}_{1\kappa q_\kappa}))'$. Thus, the elements of $\hat{\boldsymbol{T}}_{\kappa}^{(1)}(\boldsymbol{y};\hat{\mu})$ that corresponds to $\hat{\boldsymbol{\gamma}}_{1\kappa}$ are

$$\left[-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\boldsymbol{\psi}_{\kappa}(\boldsymbol{y}_i,\hat{\boldsymbol{\gamma}}_{1\kappa})}{\partial\boldsymbol{\gamma}_{1\kappa}'}\right]^{-1}\boldsymbol{\psi}_{\kappa}(\boldsymbol{y},\hat{\boldsymbol{\gamma}}_{1\kappa}).$$

Moreover, the elements of $\hat{\boldsymbol{T}}_{\kappa}^{(1)}(\boldsymbol{y};\hat{\mu})$ that correspond to $\hat{\boldsymbol{\theta}}_{2\kappa}$ are zero because the distribution of $\hat{\boldsymbol{\theta}}_{2\kappa}$ degenerates. Thus, we obtain the following information criterion:

$$\mathrm{ELIC}(\kappa)=2\sum_{i=1}^{n}\log\left(1-\hat{\boldsymbol{\lambda}}_{\kappa}'\boldsymbol{m}(\boldsymbol{y}_i,\hat{\boldsymbol{\theta}}_{\kappa})\right)-2\hat{b}_\kappa,$$

where

$$\hat{b}_\kappa = \mathrm{tr}\left(\left[-\sum_{i=1}^{n} \frac{\partial \boldsymbol{\psi}_\kappa(\boldsymbol{y}_i, \hat{\boldsymbol{\gamma}}_{1\kappa}')}{\partial \boldsymbol{\gamma}_{1\kappa}'} \right]^{-1} \sum_{i=1}^{n} \boldsymbol{\psi}_\kappa(\boldsymbol{y}_i, \hat{\boldsymbol{\gamma}}_{1\kappa}) \boldsymbol{\phi}_\kappa(\boldsymbol{y}_i, \hat{\boldsymbol{\gamma}}_{1\kappa})' \right)$$

with

$$\boldsymbol{\phi}_\kappa(\boldsymbol{y}_i, \hat{\boldsymbol{\gamma}}_{1\kappa}) = \begin{pmatrix} -\frac{M_\kappa(\boldsymbol{y}_i, \hat{\boldsymbol{\theta}}_{1\kappa})' \hat{\boldsymbol{\lambda}}_\kappa}{1 - \hat{\boldsymbol{\lambda}}_\kappa' \boldsymbol{m}(\boldsymbol{y}_i, \hat{\boldsymbol{\theta}}_{1\kappa})} \\ -\frac{\boldsymbol{m}(\boldsymbol{y}_i, \hat{\boldsymbol{\theta}}_{1\kappa})}{1 - \hat{\boldsymbol{\lambda}}_\kappa' \boldsymbol{m}(\boldsymbol{y}_i, \hat{\boldsymbol{\theta}}_{1\kappa})} \end{pmatrix}.$$

The second term of the ELIC is called the penalty term, which effectively balances between goodness-of-fit and model parsimony. We select the value of κ that minimizes the ELIC.

Although our main focus is on the selection of the regularization parameter, the ELIC can also be used for a classical subset selection problem. When $\kappa = 0$, we have

$$E[b_0] = \mathrm{tr}\left(E\left[-\frac{\partial \boldsymbol{\phi}(\boldsymbol{y}_i, \boldsymbol{\gamma}^*)}{\partial \boldsymbol{\gamma}'} \right]^{-1} E[\boldsymbol{\phi}(\boldsymbol{y}_i, \boldsymbol{\gamma}^*) \boldsymbol{\phi}(\boldsymbol{y}_i, \boldsymbol{\gamma}^*)'] \right),$$

where $\boldsymbol{\gamma}^* = (\boldsymbol{\theta}^{*\prime}, \boldsymbol{\lambda}(\boldsymbol{\theta}^*)')'$. Furthermore, if the moment restriction model is correctly specified, then we have $E[b_0] = r - p$. Thus, the ELIC coincides with the AIC of Hong, Preston, and Shum (2003), which can be used for the subset variable selection for the EL estimation.

If we approximate the bias of the PEL estimator by $r - q_\kappa$, then we obtain

$$\mathrm{AIC}(\kappa) = 2 \sum_{i=1}^{n} \log\left(1 - \hat{\boldsymbol{\lambda}}_\kappa' \boldsymbol{m}(\boldsymbol{y}_i, \hat{\boldsymbol{\theta}}_\kappa) \right) - 2(r - q_\kappa).$$

Zou, Hastie, and Tibshirani (2007) showed that the number of nonzero coefficients is an unbiased estimator for the number of degrees of freedom of the Lasso; our result is similar to theirs.

For regression models, it is known that the AIC tends to select overfitting models when the sample size is small. As suggested by an anonymous reviewer, a remedy for this problem is to use the AIC_C of Hurvich and Tsai (1989), which is a bias-corrected version of the AIC. However, they established the bias correction method only for linear regression and autoregressive models. As far as we know, a bias correction method has not been established even for general parametric likelihood models. A theoretical derivation of the AIC$_C$ for the PEL estimator will further complicates the problem because our AIC depends not only on the number of parameters but also on the number of moment restrictions as well as the size of the regularization parameter. We would like to explore this issue in a future project.

Acknowledgements

We thank an anonymous referee for his comment. We also thank seminar participants at Otaru University of Commerce, Hiroshima University, and Kyoto University and participants of The 2nd Annual International Conference on Applied Econometrics in Hawaii for their comments. This research was supported by JSPS KAKENHI Grant Number 15K03396.

Appendix

Supplementary material, which includes results of Monte Carlo study and the proof of Theorem 3.1, is available online.

References

Akaike, H. (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. K., and B. Lu (2001): "Consistent Model and Moment Selection Procedure for GMM Estimation with Application to Dynamic Panel Data Models," *Journal of Econometrics*, 101, 123–164.

Caner, M. (2009): "Lasso-Type GMM Estimator," *Econometric Theory*, 25, 270–290.

Caner, M., and H. H. Zhang (2014): "Adaptive Elastic Net for Generalized Methods of Moments," *Journal of Business & Economic Statistics*, 32, 30–47.

Chang, J., S. X. Chen, and X. Chen (2015): "High Dimensional Generalized Empirical Likelihood for Moment Restrictions with Dependent Data," *Journal of Econometrics*, 185, 283–304.

Chang, J., C. Y. Tang, and T. T. Wu (2018): "A New Scope of Penalized Empirical Likelihood with High-Dimensional Estimating Equations," *Annals of Statistics*, 46, 3185–3216.

Chen, X., H. Hong, and M. Shum (2007): "Nonparametric Likelihood Ratio Model Selection Tests Between Parametric Likelihood and Moment Condition Models," *Journal of Econometrics*, 141, 109–140.

Fan, J., and R. Li (2001): "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360.

Gatto, R., and E. Ronchetti (1996): "General Saddlepoint Approximations of Marginal Densities and Tail Probabilities," *Journal of the American Statistical Association*, 91, 666–673.

Hannan, E. J., and B. G. Quinn (1979): "The determination of the Order of an Autoregression," *Journal of the Royal Statistical Society, Series B*, 41, 190–195.

Hong, H., B. Preston, and M. Shum (2003): "Generalized Empirical Likelihood-Based Model Selection Criteria for Moment Condition Models," *Econometric Theory*, 19, 923–943.

Hurvich, C. M., and C.-L. Tsai (1989): "Regression and Time Series Model Selection in Small Samples," *Biometrika*, 76, 297–307.

Konishi, S., and G. Kitagawa (1996): "Generalised Information Criteria in Model Selection," *Biometrika*, 83, 875–890.

La Vecchia, D., E. Ronchetti, and F. Trojani (2012): "Higher-Order Infinitesimal Robustness," *Journal of the American Statistical Association*, 107, 1546–1557.

Leng, C., and C. Y. Tang (2012): "Penalized Empirical Likelihood and Growing Dimensional General Estimating Equations," *Biometrika*, 99, 703–716.

Öllerer, V., C. Croux, and A. Alfons (2015): "The Influence Function of Penalized Regression Estimators," *Statistics: A Journal of Theoretical and Applied Statistics*, 49, 741–765.

Qin, J., and J. Lawless (1994): "Empirical Likelihood and General Estimating Equations," *Annals of Statistics*, 22, 300–325.

Schwarz, G. (1978): "Estimating the Dimension of a Model," *Annals of Statistics*, 6, 461–464.

Shi, Z. (2016): "Estimation of Sparse Structral Parameter with Many Endogenous Variables," *Econometric Reviews*, 35, 1582–1608.

Tibshirani, R. (1996): "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society. Series B*, 58, 267–288.

Wang, H., B. Li, and C. Leng (2009): "Shrinkage Tuning Parameter Selection with a Diverging Number of Parameters," *Journal of the Royal Statistical Society: Series B*, 71, 671–683.

Zhang, C.-H. (2010): "Nearly Unbiased Variable Selection under Minimax Concave Penalty," *Annals of Statistics*, 38, 894–942.

Zou, H. (2006): "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429.

Zou, H., and T. Hastie (2005): "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society, Series B*, 67, 301–320.

Zou, H., T. Hastie, and R. Tibshirani (2007): "On the "Degrees of Freedom" of the LASSO," *Annals of Statistics*, 35, 2173–2192.

Supplement to "Regularization Parameter Selection for Penalized Empirical Likelihood Estimator"

Tomohiro Ando Naoya Sueishi

November 7, 2018

Abstract

This supplement contains results of Monte Carlo study and the proof of the theorem in the main paper.

A Monte Carlo Study

We compared the ELIC, the AIC, and the BIC, where the BIC is given by

$$\mathrm{BIC}(\kappa) = 2\sum_{i=1}^{n} \log\left(1 - \hat{\boldsymbol{\lambda}}_{\kappa}' \boldsymbol{m}(\boldsymbol{y}_i, \hat{\boldsymbol{\theta}}_{\kappa})\right) - \log n(r - q_{\kappa}).$$

Following the proof of Theorem 1 in Zhang, Li, and Tsai (2010), we can show that the BIC asymptotically identifies the true model if the model is correctly specified and p is fixed. However, the BIC does not have a theoretical justification when p increases with the sample size.

Leng and Tang (2012) suggested the following modified BIC:

$$\mathrm{BIC}_{\mathrm{M}}(\kappa) = 2\sum_{i=1}^{n} \log\left(1 - \hat{\boldsymbol{\lambda}}_{\kappa}' \boldsymbol{m}(\boldsymbol{y}_i, \hat{\boldsymbol{\theta}}_{\kappa})\right) + C_n(\log n)q_{\kappa}.$$

Although the value of C_n cannot be uniquely determined, they recommend $C_n = \max\{\log\log p, 1\}$. We do not report the result of $\mathrm{BIC}_{\mathrm{M}}$ because there is little difference between the BIC and the $\mathrm{BIC}_{\mathrm{M}}$ when p is less than 20.

Calculation of the ELIC necessitates first and second derivatives of the objective function of the PEL estimator. In our simulation, we used a numerical derivative calculation, which is often employed in the literature (see, for instance, Ando 2007).

We conducted three simulations. Case 1 considered a correctly specified model. We assessed the performance of the criteria in terms of the detection of model sparsity and mean squared error (MSE) of the estimator. Case 2 considered the nonparametric IV estimation problem, which is the case where candidate models are misspecified. Case 3 compared the AIC and the ELIC as bias estimators. Cases 1 and 3 employed the SCAD penalty:

$$p_\kappa(u) = \begin{cases} \kappa|u| & |u| \le \kappa \\ -(u^2 - 2a\kappa|u| + \kappa^2)/[2(a-1)] & \kappa < |u| \le a\kappa \\ (a+1)\kappa^2/2 & |u| > a\kappa. \end{cases}$$

Following the suggestion of Fan and Li (2001), we set $a = 3.7$. Case 2 employed the L_2 penalty.

Case 1: Variable selection DGP is specified by the following equations:

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i, \quad \boldsymbol{x}_i = \boldsymbol{z}_i + \boldsymbol{\eta}_i,$$

where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$ and $\boldsymbol{x}_i = (x_{i1}, \dots, x_{ip})'$ are the p-dimensional vector of coefficients and explanatory variables, respectively. As we will see below, p changes with n. The vector $\boldsymbol{z}_i = (z_{i1}, \dots, z_{ip})'$ is independent of ε_i and was generated from a p-dimensional normal distribution $N(0, \Sigma_z)$. Moreover, the vector $(\boldsymbol{\eta}_i', \varepsilon_i)'$ was generated from a $(p+1)$-dimensional normal distribution $N(0, \Sigma_{\eta\varepsilon})$, where

$$\Sigma_{\eta\varepsilon} = \begin{pmatrix} 1 & 0 & \cdots & 0 & \rho_{\eta_1\varepsilon}\sigma \\ 0 & 1 & \ddots & \vdots & \rho_{\eta_2\varepsilon}\sigma \\ \vdots & \ddots & \ddots & 0 & \vdots \\ 0 & \cdots & 0 & 1 & \rho_{\eta_p\varepsilon}\sigma \\ \rho_{\eta_1\varepsilon}\sigma & \rho_{\eta_2\varepsilon}\sigma & \cdots & \rho_{\eta_p\varepsilon}\sigma & \sigma^2 \end{pmatrix}.$$

Note that $\rho_{\eta_k\varepsilon}$ specifies the magnitude of endogeneity.

Let $\sigma_{z,jk}$ be the (j,k)-th element of Σ_z. We consider two different settings. We first consider the following setting for Σ_z and $\Sigma_{\eta\varepsilon}$.

$$\begin{aligned} \text{DGP 1}: \quad & \boldsymbol{\beta} = (0.2, 0.12, 0, 0, -0.2, 0, \dots, 0)', \\ & \sigma_{z,jk} = 1 \ (j = k) \text{ and } \sigma_{z,jk} = 0.1 \ (j \ne k), \\ & \rho_{\eta_k\varepsilon} = 0.5 \ (k = 1, \dots, 4) \text{ and } \rho_{\eta_k\varepsilon} = 0 \ (k = 5, \dots, p). \end{aligned}$$

There are four endogenous predictors among p variables because the value of $\rho_{\eta_k\varepsilon}$ is nonzero for $k = 1, \dots, 4$. Moreover, there are three nonzero coefficients in $\boldsymbol{\beta}$. We let the variance σ^2 vary at 0.1, 0.5 and 1.0.

In the second setting, the number of nonzero coefficients increases from 3 to 8.

$$\begin{aligned} \text{DGP 2}: \quad & \boldsymbol{\beta} = (0.2, 0.12, 0.1, 0, -0.2, 0, 0.2, 0.15, 0.5, 0.5, 0, 0, \dots, 0)', \\ & \sigma_{z,jk} = 1 \ (j = k), \ \sigma_{z,jk} = 0.1 \ (j \ne k), \text{ and } \rho_{\eta_k\varepsilon} = 0.8 \ (k = 1, \dots, p). \end{aligned}$$

All p variables are endogenous in this setting.

We estimated the model by using the following moment restrictions:

$$E\left[\boldsymbol{m}(\boldsymbol{y}_i, \boldsymbol{\beta})\right] = E[\boldsymbol{a}(\boldsymbol{z}_i)(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})] = \boldsymbol{0},$$

where $\boldsymbol{a}(\boldsymbol{z}_i) = (z_{i1}, \ldots, z_{ip}, z_{i1}^2, \ldots, z_{ip}^2)'$ is the vector of instrumental variables. Thus, the number of moment restrictions, r, is $2p$. Similar to Chang, Chen, and Chen (2015), the dimension p was tied to the sample size n. We set $p = \lfloor 6n^{2/15} \rfloor$, where $\lfloor \cdot \rfloor$ is the operator that truncates the decimal numbers. The sample size was set to be $n = 100$, 200, 500 and 1000. The candidate values of κ were prepared as $10^{-k/4}$ for $k = 0, 1, ..., 20$.

To evaluate the performance of each criterion, we calculated the percentages of correctly identified nonzero coefficients (C) and correctly identified zero coefficients (ZC), which were defined as follows:

$$\mathrm{C} = \frac{\sum_{j:\beta_j \neq 0} I(\hat{\beta}_j \neq 0)}{\sum_{j=1}^{p} I(\beta_j \neq 0)} \quad \text{and} \quad \mathrm{ZC} = \frac{\sum_{j:\beta_j = 0} I(\hat{\beta}_j = 0)}{\sum_{j=1}^{p} I(\beta_j = 0)},$$

where $I(\cdot)$ is the indicator function. If the true model is identified, then C = ZC = 1. In addition to C and ZC, we calculated the MSE, which is given as

$$\mathrm{MSE} = \frac{1}{n} \sum_{i=1}^{n} \left\{ \boldsymbol{x}_i'(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \right\}^2 .$$

Tables 1 and 2 report the average values of C, ZC, and MSE over 200 repetitions. With regard to sparsity, as predicted from theory, the BIC performed better than did the AIC or the ELIC. In each case, the percentage of correctly specified ZC was close to one for the BIC, whereas it is slightly less than one for the AIC and the ELIC. In particular, the ELIC tends to select overfitted models. However, the difference between the three criteria diminishes as the sample size increases. Although the BIC performs well in identifying ZC, it tends to select underfitted models compared to the AIC and the ELIC. The percentage of correctly specified coefficients for the BIC is rather smaller. As a result, the MSE of the BIC is larger than that of either the AIC or the ELIC.

Case 2: Nonparametric IV model Application of our selection criterion is not limited to sparse estimation. Our ELIC can be employed in many situations where regularization is necessary. A potential application is nonparametric IV estimation. The instability of nonparametric IV estimators unless properly regularized is well known because of the ill-posed inverse problem (Carrasco, Florens, and Renault 2007).

DGP is specified by

$$y_i = g(x_i) + \varepsilon_i, \quad x_i = z_i + \eta_i,$$

where ε_i, η_i, and z_i were generated as

$$\begin{pmatrix} \varepsilon_i \\ \eta_i \\ z_i \end{pmatrix} \sim N \left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0.5\sigma & 0 \\ 0.5\sigma & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right).$$

Table 1: Correctly specified model (DGP1). C: the percentage of correctly identified nonzero coefficients; ZC: the percentage of correctly identified zero coefficients; MSE: the average of the mean squared errors.

DGP 1

$\sigma^2 = 0.1$	MSE			C			ZC		
n	ELIC	AIC	BIC	ELIC	AIC	BIC	ELIC	AIC	BIC
100	0.0468	0.0506	0.0687	0.91	0.86	0.72	0.82	0.93	0.96
200	0.0411	0.0410	0.0494	0.93	0.92	0.86	0.89	0.93	0.94
500	0.0329	0.0325	0.0354	0.96	0.97	0.94	0.93	0.93	0.94
1000	0.0280	0.0281	0.0314	0.98	0.98	0.96	0.94	0.94	0.94
$\sigma^2 = 0.5$									
100	0.1154	0.1271	0.1448	0.70	0.48	0.29	0.87	0.96	0.98
200	0.0864	0.0937	0.1294	0.81	0.72	0.40	0.95	0.97	0.99
500	0.0772	0.0790	0.0927	0.82	0.81	0.68	0.98	0.98	0.99
1000	0.0753	0.0764	0.0806	0.83	0.82	0.76	0.99	0.99	1.00
$\sigma^2 = 1.0$									
100	0.1630	0.1499	0.1687	0.63	0.47	0.24	0.89	0.94	0.97
200	0.1191	0.1312	0.1532	0.69	0.51	0.24	0.93	0.94	0.98
500	0.0958	0.1000	0.1492	0.74	0.65	0.25	0.94	0.94	0.98
1000	0.0935	0.0974	0.1236	0.74	0.70	0.46	0.94	0.95	0.99

Table 2: Correctly specified model (DGP2). C: the percentage of correctly identified nonzero coefficients; ZC: the percentage of correctly identified zero coefficients; MSE: the average of the mean squared errors.

DGP 2

$\sigma^2 = 0.1$		MSE			C			ZC	
n	ELIC	AIC	BIC	ELIC	AIC	BIC	ELIC	AIC	BIC
100	0.0909	0.1159	0.2186	0.94	0.90	0.78	0.70	0.82	0.89
200	0.0443	0.0449	0.0741	0.99	0.99	0.95	0.73	0.87	0.92
500	0.0327	0.0305	0.0311	1.00	1.00	1.00	0.75	0.89	0.97
1000	0.0285	0.0280	0.0281	0.99	1.00	1.00	0.87	0.97	1.00
$\sigma^2 = 0.5$									
100	0.2488	0.2551	0.3664	0.78	0.75	0.62	0.73	0.85	0.91
200	0.2240	0.2013	0.3054	0.79	0.76	0.60	0.80	0.89	0.94
500	0.1635	0.1560	0.1655	0.83	0.84	0.80	0.94	0.93	0.95
1000	0.1511	0.1468	0.1562	0.83	0.85	0.82	0.99	0.98	0.99
$\sigma^2 = 1.0$									
100	0.4159	0.4720	0.5382	0.68	0.53	0.46	0.69	0.83	0.90
200	0.2956	0.2479	0.3944	0.69	0.77	0.53	0.85	0.85	0.97
500	0.1741	0.1741	0.1977	0.80	0.80	0.75	0.98	0.98	0.98
1000	0.1690	0.1692	0.1693	0.81	0.81	0.80	0.99	0.99	0.99

We considered three different specifications for $g(x)$:

$$\begin{aligned} \text{DGP 1}: &\quad g(x) = -0.1\log(|x|) + 0.3\sin(\pi x). \\ \text{DGP 2}: &\quad g(x) = 0.5\sin(x/3). \\ \text{DGP 3}: &\quad g(x) = -0.02x. \end{aligned}$$

To approximate unknown true functions, we fit the linear combination of the set of $p = 15$ basis functions. We used cubic B-spline basis functions for $\boldsymbol{b}(x_i)$. The moment restrictions are

$$E\left[\boldsymbol{m}(\boldsymbol{y}_i, \boldsymbol{\beta})\right] = E\left[\boldsymbol{a}(z_i)(y_i - \boldsymbol{b}(x_i)'\boldsymbol{\beta})\right] = \mathbf{0}.$$

The vector of instrumental variables was set as $\boldsymbol{a}(z_i) = (\boldsymbol{a}_1(z_i)', \boldsymbol{a}_2(z_i)', \boldsymbol{a}_3(z_i)')'$ with

$$\begin{aligned} \boldsymbol{a}_1(z_i) &= (|z_i|, z_i, z_i^2, \ldots, z_i^6)' \\ \boldsymbol{a}_2(z_i) &= (\sin(\pi z_i), \sin(2\pi z_i), \ldots, \sin(6\pi z_i))' \\ \boldsymbol{a}_3(z_i) &= (\cos(z_i), \cos(z_i/2), \ldots, \cos(z_i/6))'. \end{aligned}$$

Our estimator can be viewed as an EL analog of the penalized series estimator of Newey and Powell (2003) and Blundell, Chen, and Kristensen (2007). Notice that the moment restriction models are misspecified because of the approximation error.

The candidate values of κ were prepared as 0 and $10^{1-k/2.5}$ for $k = 0, 1, ..., 20$. As noted in Section 3.3, both the AIC and the BIC select $\kappa = 0$ because their penalty term does not change with κ. By contrast, our proposed ELIC can incorporate penalty amounts.

The results are summarized in Table 3. We report the averages and the standard deviations of MSE over 200 repetitions. The standard deviations illustrate that the MSE is not particularly volatile even though the number of repetitions is rather small. As a benchmark, we report the MSE of the estimator that uses the infeasible best regularization parameter. The result of the best estimator is labeled Oracle. Moreover, we report the MSE of the nonpenalized EL estimator ($\kappa = 0$). We observe that the PEL estimator with the ELIC performed better than the nonpenalized EL estimator in all cases. Furthermore, the MSE of the ELIC converges towards Oracle as the sample size increases. These results suggest that our ELIC can be used not only for sparse estimation but also in a wide variety of other settings.

Case 3: Bias estimation This simulation investigates how well the bias correction term of the ELIC and the AIC estimates the true bias. We again considered nonparametric IV estimation with the true function $g(x) = -0.7x$. The error variance, the number of basis functions, and the vector of instrumental variables were set as $\sigma^2 = 0.5$, $p = 10$, and $\boldsymbol{a}(z_i) = (\log|z_i|, (\log|z_i|)^2, \ldots, (\log|z_i|)^{12})'$, respectively. Moreover, the SCAD penalty was employed.

Figure 1 plots the true bias (dotted line), the estimate of the ELIC (solid line), and the estimate of the AIC (dashed line) for $\kappa = 10^{-1}$ and 10^{-3}. The vertical axis is the size of the bias, and the horizontal axis is the sample size. The sample size was set to be $n = 100, 200, 400$, and 800. We report the averages of the estimates over 10,000 repetitions.

Table 3: Nonparametric IV model (DGP1). MSE over 200 repetitions are reported. Oracle: the MSE of the estimator that uses the infeasible best regularization parameter; SD: standard deviation of the MSE.

DGP 1:

$\sigma^2 = 0.1$	n	ELIC	$\kappa = 0$	Oracle
	100	0.01405	0.01605	0.01026
	SE	0.00044	0.00049	0.00032
	200	0.00846	0.00949	0.00703
	SE	0.00026	0.00029	0.00021
	500	0.00546	0.00563	0.00503
	SE	0.00014	0.00015	0.00013
	1000	0.00418	0.00427	0.00396
	SE	0.00011	0.00011	0.00010

$\sigma^2 = 0.2$	n	ELIC	$\kappa = 0$	Oracle
	100	0.03038	0.03696	0.01818
	SE	0.00091	0.00090	0.00043
	200	0.01901	0.02131	0.01323
	SE	0.00042	0.00046	0.00026
	500	0.01127	0.01238	0.00947
	SE	0.00032	0.00032	0.00025
	1000	0.00912	0.00947	0.00843
	SE	0.00016	0.00016	0.00016

Table 3: (Continued) Nonparametric IV model (DGP2).

DGP 2:

$\sigma^2 = 0.1$	n	ELIC	$\kappa = 0$	Oracle
	100	0.01554	0.01797	0.01100
	SE	0.00036	0.00039	0.00019
	200	0.00861	0.00962	0.00634
	SE	0.00019	0.00021	0.00011
	500	0.00446	0.00509	0.00336
	SE	0.00010	0.00010	0.00006
	1000	0.00271	0.00309	0.00209
	SE	0.00006	0.00006	0.00004

$\sigma^2 = 0.2$	n	ELIC	$\kappa = 0$	Oracle
	100	0.03180	0.03597	0.01506
	SE	0.00107	0.00102	0.00039
	200	0.01452	0.01888	0.00771
	SE	0.00046	0.00046	0.00013
	500	0.01060	0.01218	0.00535
	SE	0.00027	0.00029	0.00013
	1000	0.00760	0.00815	0.00470
	SE	0.00017	0.00017	0.00008

Table 3: (Continued) Nonparametric IV model (DGP3).

DGP 3:

$\sigma^2 = 0.1$	n	ELIC	$\kappa = 0$	Oracle
	100	0.01243	0.01618	0.00562
	SE	0.00046	0.00050	0.00018
	200	0.00755	0.00919	0.00401
	SE	0.00023	0.00021	0.00013
	500	0.00383	0.00442	0.00284
	SE	0.00009	0.00010	0.00008
	1000	0.00243	0.00269	0.00206
	SE	0.00007	0.00007	0.00006

$\sigma^2 = 0.2$	n	ELIC	$\kappa = 0$	Oracle
	100	0.02651	0.03078	0.01008
	SE	0.00080	0.00074	0.00027
	200	0.01613	0.01970	0.00884
	SE	0.00044	0.00048	0.00027
	500	0.01057	0.01138	0.00705
	SE	0.00032	0.00032	0.00022
	1000	0.00722	0.00750	0.00553
	SE	0.00018	0.00018	0.00011

Because the exact value of the true bias is unknown even in the simulation, it was evaluated numerically. We obtained the bias through the following iterative calculations. In the k-th step of the iteration, we first generated a set of n observations $\{\boldsymbol{y}_1^{(k)}, \dots, \boldsymbol{y}_n^{(k)}\}$ and obtained the parameter estimate $(\hat{\boldsymbol{\theta}}_\kappa^{(k)}, \hat{\boldsymbol{\lambda}}_\kappa^{(k)})$. Next, we generated m observations $\{\tilde{\boldsymbol{y}}_1^{(k)}, \dots, \tilde{\boldsymbol{y}}_m^{(k)}\}$ that are independent of $\{\boldsymbol{y}_1^{(k)}, \dots, \boldsymbol{y}_n^{(k)}\}$ and calculated

$$\frac{1}{n}\sum_{i=1}^{n} \log\left(1 - \hat{\boldsymbol{\lambda}}_\kappa^{(k)\prime} \boldsymbol{m}(\boldsymbol{y}_i^{(k)}, \hat{\boldsymbol{\theta}}_\kappa^{(k)})\right) - \frac{1}{m}\sum_{j=1}^{m} \log\left(1 - \hat{\boldsymbol{\lambda}}_\kappa^{(k)\prime} \boldsymbol{m}(\tilde{\boldsymbol{y}}_j^{(k)}, \hat{\boldsymbol{\theta}}_\kappa^{(k)})\right). \tag{A.1}$$

We set $m = 2,000,000$. The bias was approximated by the average of (A.1) over $k = 1, \dots, 10,000$.

We observe from Figure 1 that the simple estimator, which is given by (3.1) in the main paper, has a significant bias as an estimate of the risk, and the bias decreases as the sample size increases. Moreover, the true bias is larger for smaller κ. This is because $\kappa = 10^{-3}$ is too small to stabilize the estimator, and hence, the value of the risk tends to be large.

Overall, the ELIC can capture the behavior of the true bias. In contrast, the AIC fails to estimate the bias, particularly when κ is small. This result suggests that the ELIC outperforms the AIC when the performance of the estimator is sensitive to the choice of the regularization parameter.

B Proof of Theorem 3.1

Our proof is similar to that for Theorem of 2.1 of Konishi and Kitagawa (1996). By conditions (A.1) and (A.2) in the main paper, the PEL estimator can be expanded as

$$\hat{\boldsymbol{T}}_\kappa(\hat{\mu}) = \boldsymbol{T}_\kappa(\mu) + \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{T}_\kappa^{(1)}(\boldsymbol{y}_j; \mu) + \frac{1}{2n^2}\sum_{j=1}^{n}\sum_{k=1}^{n} \boldsymbol{T}_\kappa^{(2)}(\boldsymbol{y}_j, \boldsymbol{y}_k; \mu) + o_p\left(\frac{1}{n}\right). \tag{B.1}$$

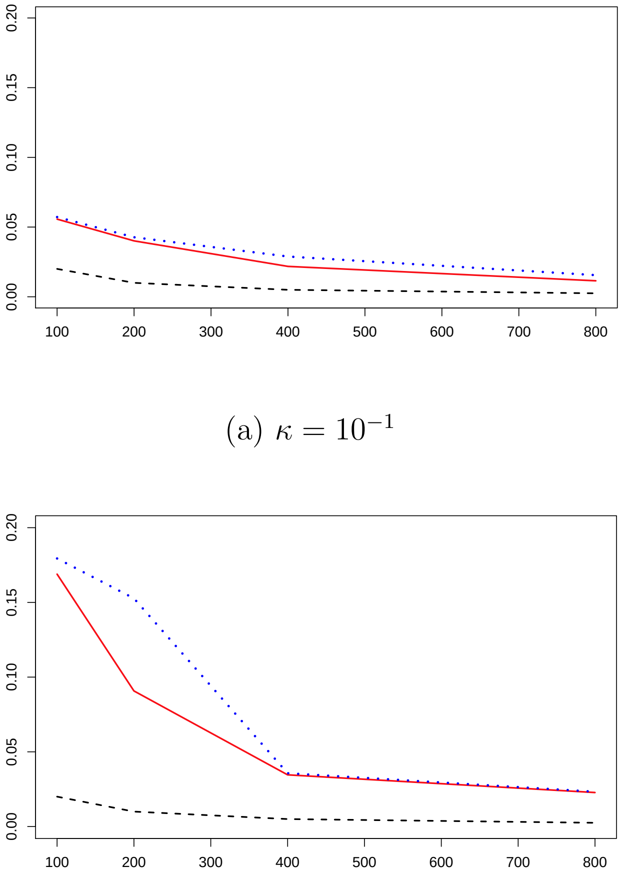
Here, we can take $\boldsymbol{T}_\kappa^{(1)}(\boldsymbol{y}; \mu)$ and $\boldsymbol{T}_\kappa^{(2)}(\boldsymbol{y}, \boldsymbol{z}; \mu)$ so that they satisfy

$$\int \boldsymbol{T}_\kappa^{(1)}(\boldsymbol{y}; \mu) d\mu(\boldsymbol{y}) = \boldsymbol{0} \tag{B.2}$$

and

$$\int \boldsymbol{T}_\kappa^{(2)}(\boldsymbol{y}, \boldsymbol{z}; \mu) d\mu(\boldsymbol{y}) = \int \boldsymbol{T}_\kappa^{(2)}(\boldsymbol{y}, \boldsymbol{z}; \mu) d\mu(\boldsymbol{z}) = \boldsymbol{0}. \tag{B.3}$$

By the theory of U-statistics, we see that the second and third terms of the right-hand side of (B.1) are of order $O_p(n^{-1/2})$ and $O_p(n^{-1})$, respectively.



(a) $\kappa = 10^{-1}$

(b) $\kappa = 10^{-3}$

Figure 1: The true bias (dotted line) and asymptotic bias estimates of ELIC (solid line) and AIC (dashed line).

Expanding $n^{-1}\sum_{i=1}^{n}\log(1-\hat{\boldsymbol{\lambda}}_\kappa'\boldsymbol{m}(\boldsymbol{y}_i;\hat{\boldsymbol{\theta}}_\kappa))$ around $\boldsymbol{T}_\kappa(\mu)$ yields

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\log(1-\hat{\boldsymbol{\lambda}}_\kappa'\boldsymbol{m}(\boldsymbol{y}_i;\hat{\boldsymbol{\theta}}_\kappa))\\
&=\frac{1}{n}\sum_{i=1}^{n}\log(1-\boldsymbol{T}_{\lambda,\kappa}(\mu)'\boldsymbol{m}(\boldsymbol{y}_i,\boldsymbol{T}_{\theta,\kappa}(\mu)))+\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{T}_\kappa^{(1)}(\boldsymbol{y}_j;\mu)'\boldsymbol{\phi}(\boldsymbol{y}_i,\boldsymbol{T}_\kappa(\mu))\\
&+\frac{1}{2n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\boldsymbol{T}_\kappa^{(2)}(\boldsymbol{y}_j,\boldsymbol{y}_k;\mu)'\boldsymbol{\phi}(\boldsymbol{y}_i,\boldsymbol{T}_\kappa(\mu))\\
&+\frac{1}{2n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\boldsymbol{T}_\kappa^{(1)}(\boldsymbol{y}_j;\mu)'\frac{\partial\boldsymbol{\phi}(\boldsymbol{y}_i,\boldsymbol{T}_\kappa(\mu))}{\partial\boldsymbol{\gamma}'}\boldsymbol{T}_\kappa^{(1)}(\boldsymbol{y}_k;\mu)+o_p\left(\frac{1}{n}\right).
\end{aligned}
$$

Similarly, we have

$$
\begin{aligned}
&\int\log(1-\hat{\boldsymbol{\lambda}}_\kappa'\boldsymbol{m}(\boldsymbol{y},\hat{\boldsymbol{\theta}}_\kappa))d\mu(\boldsymbol{y})\\
&=\int\log(1-\boldsymbol{T}_{\lambda,\kappa}(\mu)'\boldsymbol{m}(\boldsymbol{y},\boldsymbol{T}_{\theta,\kappa}(\mu))d\mu(\boldsymbol{y})+\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{T}_\kappa^{(1)}(\boldsymbol{y}_j;\mu)'\int\boldsymbol{\phi}(\boldsymbol{y},\boldsymbol{T}_\kappa(\mu))d\mu(\boldsymbol{y})\\
&+\frac{1}{2n^2}\sum_{j=1}^{n}\sum_{k=1}^{n}\boldsymbol{T}_\kappa^{(2)}(\boldsymbol{y}_j,\boldsymbol{y}_k;\mu)'\int\boldsymbol{\phi}(\boldsymbol{y},\boldsymbol{T}_\kappa(\mu))d\mu(\boldsymbol{y})\\
&+\frac{1}{2n^2}\sum_{j=1}^{n}\sum_{k=1}^{n}\boldsymbol{T}_\kappa^{(1)}(\boldsymbol{y}_j;\mu)'\left[\int\frac{\partial\boldsymbol{\phi}(\boldsymbol{y},\boldsymbol{T}_\kappa(\mu))}{\partial\boldsymbol{\gamma}'}d\mu(\boldsymbol{y})\right]\boldsymbol{T}_\kappa^{(1)}(\boldsymbol{y}_k;\mu)+o_p\left(\frac{1}{n}\right).
\end{aligned}
$$

Thus, using (B.2) and (B.3), we obtain

$$
\frac{1}{n}\sum_{i=1}^{n}\log\left(1-\hat{\boldsymbol{\lambda}}_\kappa'\boldsymbol{m}(\boldsymbol{y}_i;\hat{\boldsymbol{\theta}}_\kappa)\right)-\int\log\left(1-\hat{\boldsymbol{\lambda}}_\kappa'\boldsymbol{m}(\boldsymbol{y},\hat{\boldsymbol{\theta}}_\kappa)\right)d\mu(\boldsymbol{y})=\frac{1}{n}b_\kappa+o_p\left(\frac{1}{n}\right),
$$

where b_κ satisfies

$$
E[b_\kappa]=\mathrm{tr}\left(\int\boldsymbol{T}_\kappa^{(1)}(\boldsymbol{y};\mu)\boldsymbol{\phi}(\boldsymbol{y},\boldsymbol{T}_\kappa(\mu))'d\mu(\boldsymbol{y})\right).
$$

□

References

Ando, T. (2007): "Bayesian Predictive Information Criterion for the Evaluation of Hierarchical Bayesian and Empirical Bayes Models," *Biometrika*, 94, 443–458.

Blundell, R., X. Chen, and D. Kristensen (2007): "Semi-Nonparametric IV Estimation of Shape-Invariant Engel Curves," *Econometrica*, 75, 1613–1669.

Carrasco, M., J.-P. Florens, and E. Renault (2007): "Linear Inverse Problems in Structural Econometrics Estimation Based on Spectral Decomposition and Regularization," in *Handbook of Econometrics*, ed. by J. Heckman, and E. Leamer, vol. 6. Elsevier.

Chang, J., S. X. Chen, and X. Chen (2015): "High Dimensional Generalized Empirical Likelihood for Moment Restrictions with Dependent Data," *Journal of Econometrics*, 185, 283–304.

Fan, J., and R. Li (2001): "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360.

Konishi, S., and G. Kitagawa (1996): "Generalised Information Criteria in Model Selection," *Biometrika*, 83, 875–890.

Leng, C., and C. Y. Tang (2012): "Penalized Empirical Likelihood and Growing Dimensional General Estimating Equations," *Biometrika*, 99, 703–716.

Newey, W. K., and J. L. Powell (2003): "Instrumental Variable Estimation of Nonparametric Models," *Econometrica*, 71, 1565–1578.

Zhang, Y., R. Li, and C.-L. Tsai (2010): "Regularization Parameter Selection via Generalized Information Criterion," *Journal of the American Statistical Association*, 105, 312–323.