Ann Inst Stat Math (2010) 62:413–438
DOI 10.1007/s10463-008-0186-0

Optimal tuning parameter estimation in maximum penalized likelihood method

Masao Ueki · Kaoru Fueda

Received: 20 August 2007 / Revised: 16 April 2008 / Published online: 13 July 2008

Abstract In maximum penalized or regularized methods, it is important to select a tuning parameter appropriately. This paper proposes a direct plug-in method for tuning parameter selection. The tuning parameters selected using a generalized information criterion (Konishi and Kitagawa, *Biometrika*, *83*, 875–890, 1996) and cross-validation (Stone, *Journal of the Royal Statistical Society, Series B*, *58*, 267–288, 1974) are shown to be asymptotically equivalent to those selected using the proposed method, from the perspective of estimation of an optimal tuning parameter. Because of its directness, the proposed method is superior to the two selection methods mentioned above in terms of computational cost. Some numerical examples which contain the penalized spline generalized linear model regressions are provided.

Keywords Cross-validation · Direct plug-in method · Generalized information criterion · Kullback–Leibler information · Maximum penalized likelihood method · Penalized spline · Ridge regression · Tuning parameter estimation

1 Introduction

Maximum penalized or regularized methods are widely used tools to stabilize estimators, which are introduced in spline smoothing (Green and Silverman 1994).

M. Ueki (✉)
The Institute of Statistical Mathematics, 4-6-7, Minami-Azabu, Minato-ku,
Tokyo 106-8569, Japan
e-mail: uekim@ism.ac.jp

K. Fueda
Graduate School of Environmental Science, Okayama University, Naka 3-1-1,
Tsushima, Okayama 700-8530, Japan
e-mail: fueda@ems.okayama-u.ac.jp

They include the ridge estimator, which is introduced to avoid multi-collinearity in least-squares multiple regression estimation, but with a rather different motivation from them. The efficiency of the penalization methods depends strongly on setting the tuning parameter that controls the extent of penalization. Therefore, it is important to select the tuning parameter appropriately. In the description in this paper, we consider penalization in a maximum likelihood framework: the maximum penalized likelihood estimator (MPLE; Good and Gaskins 1971; Green and Silverman 1994). Recently, concerning the MPLE, Fan and Li (2001) proposed a maximum penalized likelihood approach for automatic variable selection, similarly to the Lasso (Tibshirani 1996).

Several methods are useful to select the tuning parameter. They are established through proposal of an appropriate selection criterion, where the selection is done by minimizing them with respect to the tuning parameter, as typified by Marrows' C_p, cross-validation, generalized cross-validation and other methods in the least-squares regression context. Information criteria, including AIC-type criteria (Akaike 1974; Takeuchi 1976; Shimodaira 2003), cross-validation, BIC-type criteria and others, play a central role in the likelihood framework. One AIC-type criterion for the MPLE is a generalized information criterion (GIC; Konishi and Kitagawa 1996), which forms the empirical log-likelihood with the correction term for the bias, derived analytically with the influence function. The GIC can evaluate the models not only with MPLE but also with a robust estimator, maximum weighted likelihood estimator, etc. Imoto and Konishi (2003) and Nonaka and Konishi (2005) used GIC for selection of smoothing parameters in nonlinear regression models estimated by regularization. Alternatively, cross-validation (CV; Stone 1974) is applicable to choose the value of a tuning parameter in the maximum penalized likelihood method. The CV requires no analytic calculations as in the GIC, although the computational cost for the CV is higher than the GIC. This paper analyzes the properties of the tuning parameters selected using the GIC and CV.

On the other hand, for BIC-type criteria, the penalization is interpreted as a Bayesian inference with a corresponding prior density. Konishi et al. (2004) extended the BIC to evaluate models with MPLE, particularly for radial basis function networks.

Existing methods of selecting tuning parameter are based on minimizing each criterion. Finding the minimum is usually accomplished through a sequential search or numerical optimizations; thereby, it is often extensive in its requisite computations. To overcome computational problems, we derive an *optimal tuning parameter* in expected log-likelihood, or equivalently, in Kullback–Leibler information, with an asymptotic theory under model misspecification. In normal noise regression models, it coincides with Wand's (1999) optimal smoothing parameter in penalized spline regression. We propose a *direct plug-in tuning parameter* by replacing unknown quantities in the optimal tuning parameter by suitable consistent estimatorss. The direct plug-in tuning parameter often comes to be simple under the parametric assumption that the model includes the true distribution. We call it the *parametric direct plug-in tuning parameter*. The direct plug-in tuning parameter saves the computational cost relative to existing methods because of the directness with analytical calculations.

This paper is organized as follows. Section 2 describes the MPLE. Section 3 proposes the direct plug-in method, through defining an optimal tuning parameter with respect to the Kullback–Leibler information. Section 4 shows that the tuning parameters selected by the GIC and CV are asymptotically equivalent to that with the direct plug-in method, in the sense that the tuning parameter selection is a point estimation. Section 5 discusses the behavior of Kullback–Leibler information when using our direct plug-in method or existing methods, GIC and CV, for tuning parameter estimation. Some numerical examples are described in Sect. 6. Section 7 presents simulations for the penalized spline generalized linear model regression.

2 Maximum penalized likelihood estimator

This section describes the maximum penalized likelihood estimator (MPLE). First assume n i.i.d. observations $X_n = (X_1, \ldots, X_n)$ from an unknown true distribution $G(x)$. Then, let $\{f(\cdot;\boldsymbol{\theta}) : \boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T \in \Theta\}$ be a model with a parameter vector $\boldsymbol{\theta}$, where Θ is an open subset in $\mathbb{R}^p$. Using a penalty term $k(\boldsymbol{\theta})$, the penalized log-likelihood and the MPLE are given, respectively, as

$$\sum_{\alpha=1}^{n} \log f(X_\alpha;\boldsymbol{\theta}) - \lambda k(\boldsymbol{\theta}) \quad \text{and} \quad \hat{\boldsymbol{\theta}}_\lambda = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left\{ \sum_{\alpha=1}^{n} \log f(X_\alpha;\boldsymbol{\theta}) - \lambda k(\boldsymbol{\theta}) \right\}, \quad (1)$$

where λ is a scalar tuning parameter controlling the extent of the penalization. The MPLE is used widely for stabilization of the estimators, such as the MLE. It is important to discuss the selection of the tuning parameter λ because the performance of MPLE depends strongly on the value of the tuning parameter λ.

The above described MPLE includes the ridge-type estimator, which is frequently used in practice. Consider a multiple regression model $y = \beta_1 + x_2\beta_2 + \cdots + x_p\beta_p + \epsilon$, where ϵ is a normal random variable with mean 0 and variance σ^2. Also, $\beta_2, \ldots, \beta_p$ are coefficients of covariates $x_2, \ldots, x_p$; β_1 is an intercept. Let $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$. Presuming that we have n independent observations $Y_1, X_{1,2}, \ldots, X_{1,p}), \ldots, (Y_n, X_{n,2}, \ldots, X_{n,p})$, let $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ be an n-dimensional observation vector and let X be the $n \times p$ design matrix. The ridge-type estimator of $\boldsymbol{\beta}$ is written as $\hat{\boldsymbol{\beta}}_\lambda = (X^T X + \lambda I_p)^{-1} X^T \mathbf{Y}$, where I_p is a $p \times p$ identity matrix, and λ is a scalar tuning parameter, which controls the penalization. In a view of likelihood framework, $\hat{\boldsymbol{\beta}}_\lambda$ maximizes the penalized log-likelihood

$$\log \left\{ \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} ||\mathbf{Y} - X\boldsymbol{\beta}||^2 \right) \right\} - \lambda \frac{||\boldsymbol{\beta}||^2}{2\sigma^2},$$

with a penalty $||\boldsymbol{\beta}||^2/(2\sigma^2)$. Here $||\cdot||$ denotes the Euclidean norm. The ridge-type estimator could stabilize the ordinary least-squares estimator.

3 Direct plug-in tuning parameter

In this section, we define an optimal tuning parameter and propose the direct plug-in method. We define the optimality with the goodness of the resultant MPLE $\hat{\boldsymbol{\theta}}_\lambda$ in the sense of expected log-likelihood:

$$\eta(\boldsymbol{\theta}) = \eta(\boldsymbol{\theta}; G) = \int \log f(x; \boldsymbol{\theta}) \mathrm{d}G(x). \tag{2}$$

It equals the Kullback–Leibler information ignoring the constant term, which is independent of the model. The model is better if its expected log-likelihood value is larger. Therefore, we define the *optimal tuning parameter* λ_{opt} as

$$\lambda_{\text{opt}} = \underset{\lambda}{\operatorname{argmax}}\, E\{\eta(\hat{\boldsymbol{\theta}}_\lambda)\}.$$

Using the notation in (2) and assuming that the penalty term $k(\boldsymbol{\theta})$ in (1) does not depend on X_n, the MPLE is rewritten as $\hat{\boldsymbol{\theta}}_\lambda = \operatorname{argmax}_{\boldsymbol{\theta}}\{n\eta(\boldsymbol{\theta}; \hat{G}) - \lambda k(\boldsymbol{\theta})\}$, where $\hat{G}$ is the empirical distribution function corresponding to the observations X_n. Analogously, we define

$$\boldsymbol{\theta}_\lambda = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\{n\eta(\boldsymbol{\theta}; G) - \lambda k(\boldsymbol{\theta})\}.$$

We also define $\boldsymbol{\theta}_0 = \operatorname{argmax}_{\boldsymbol{\theta}}\{n\eta(\boldsymbol{\theta}; G)\}$, which is the nearest point to the G in the model $\{f(\cdot; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$. Here, $\boldsymbol{\theta}_0$ coincides with the true parameter vector if the model includes the true distribution G. We designate it as the model under a parametric assumption. We assume that both the MPLE and MLE converge to the $\boldsymbol{\theta}_0$ as the sample size increases. We further assume that the penalty function $k(\cdot)$ is independent of n. Alternatively, Konishi and Kitagawa (1996) deal with a penalized log-likelihood function $\sum_{\alpha=1}^{n}\{\log f(X_\alpha; \boldsymbol{\theta}) - \lambda k(\boldsymbol{\theta})\}$ instead of (1). In their assumption, the MPLE does not necessarily converge to the $\boldsymbol{\theta}_0$ as the sample size increases.

The following result reveals the effect of tuning parameter λ for the mean expected log-likelihood.

Theorem 1 *Assume that the assumptions provided in Appendix hold. Let $\hat{\boldsymbol{\theta}}_\lambda$ be the MPLE defined by Eq. (1) and $\hat{\boldsymbol{\theta}}$ be the MLE. Then the difference of the mean expected log-likelihood between the models with $\hat{\boldsymbol{\theta}}_\lambda$ and $\hat{\boldsymbol{\theta}}$ is given as*

$$E\{\eta(\hat{\boldsymbol{\theta}}_\lambda) - \eta(\hat{\boldsymbol{\theta}})\} = -\frac{1}{2n^2}\{a(\boldsymbol{\theta}_0)\lambda^2 - 2b(\boldsymbol{\theta}_0)\lambda\} + o(n^{-2}), \tag{3}$$

where

$$a(\boldsymbol{\theta}) = \frac{\partial k(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\{J(\boldsymbol{\theta})\}^{-1}\frac{\partial k(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

and

$$b(\boldsymbol{\theta}) = tr\left\{\frac{\partial^2 k(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} V(\boldsymbol{\theta})\right\} + q(\boldsymbol{\theta})^T d(\boldsymbol{\theta}) + 2q(\boldsymbol{\theta})^T e(\boldsymbol{\theta}).$$

Here, $q(\boldsymbol{\theta}) = \{J(\boldsymbol{\theta})\}^{-1}\partial k(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$, $J(\boldsymbol{\theta}) = -\partial^2\eta(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T$,

$$I(\boldsymbol{\theta}) = \int \frac{\partial \log f(x;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} \frac{\partial \log f(x;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^T} \mathrm{d}G(x), \quad V(\boldsymbol{\theta}) = \{J(\boldsymbol{\theta})\}^{-1} I(\boldsymbol{\theta}) \{J(\boldsymbol{\theta})\}^{-1},$$

$$d_i(\boldsymbol{\theta}) = \sum_{j,k=1}^{p} \frac{\partial^3\eta(\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j\partial\theta_k} V_{jk}(\boldsymbol{\theta})$$

and

$$e_i(\boldsymbol{\theta}) = \sum_{j,k=1}^{p} J^{jk}(\boldsymbol{\theta}) \int \frac{\partial^2 \log f(x;\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j} \frac{\partial \log f(x;\boldsymbol{\theta})}{\partial\theta_k} \mathrm{d}G(x),$$

where $J^{ij}(\boldsymbol{\theta})$ *denotes the* (i, j) *element of the matrix* $\{J(\boldsymbol{\theta})\}^{-1}$.

Since $J(\boldsymbol{\theta}_0)$ is positive definite, the first term in the right hand side of Eq. (3) attains a maximum at $\lambda = b(\boldsymbol{\theta}_0)/a(\boldsymbol{\theta}_0)$, asymptotically, unless $\partial k(\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta} = 0$. Consequently, the optimal tuning parameter λ_{opt} is given as

$$\lambda_{\mathrm{opt}} = \frac{b(\boldsymbol{\theta}_0)}{a(\boldsymbol{\theta}_0)} + o(1). \tag{4}$$

Then, the MPLE $\hat{\boldsymbol{\theta}}_{\lambda_{\mathrm{opt}}}$ optimally dominates the MLE $\hat{\boldsymbol{\theta}}$ in mean Kullback–Leibler information, asymptotically. When $\lambda = b(\boldsymbol{\theta}_0)/a(\boldsymbol{\theta}_0) + o(1)$, the mean expected log-likelihood attains the asymptotic maximum of $n^{-2}b(\boldsymbol{\theta}_0)^2/\{2a(\boldsymbol{\theta}_0)\} > 0$ with error $o(n^{-2})$.

It is a direct consequence from (4) that

$$\hat{\lambda}_{\mathrm{DPI}} = \frac{\hat{b}(\tilde{\boldsymbol{\theta}})}{\hat{a}(\tilde{\boldsymbol{\theta}})} \tag{5}$$

is a point estimator of λ_{opt}, where $\hat{a}(\cdot)$ and $\hat{b}(\cdot)$ are suitable consistent estimators of $a(\cdot)$ and $b(\cdot)$, and $\tilde{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}_0$ such as the MLE $\hat{\boldsymbol{\theta}}$. We call $\hat{\lambda}_{\mathrm{DPI}}$ a *direct plug-in tuning parameter*. The $\hat{\lambda}_{\mathrm{DPI}}$ is obtained directly, whereas the existing methods for selecting λ relies on some sequential search, or numerical optimization. Consequently, the suggested method reduces computational costs drastically, with some analytical effort. For instance, $\hat{\lambda}_{\mathrm{DPI}}$ can be calculated by substituting the empirical distribution $\hat{G}(x)$ instead of the unknown distribution $G(x)$ in (4). Alternatively, the calculation of $\hat{\lambda}_{\mathrm{DPI}}$ is often easier under the parametric assumption that the true distribution G is included in the model, where the true parameter is $\boldsymbol{\theta}_0$. Then we use the specific functional forms of $a(\boldsymbol{\theta})$ and $b(\boldsymbol{\theta})$ directly. The quantities under the parametric assumption are simpler than those under misspecification because $J(\boldsymbol{\theta}) = I(\boldsymbol{\theta})$ and $V(\boldsymbol{\theta}) = \{J(\boldsymbol{\theta})\}^{-1}$ hold. Accordingly, we obtain the corresponding estimator by substituting $\tilde{\boldsymbol{\theta}}$ into $a(\boldsymbol{\theta})$ and $b(\boldsymbol{\theta})$ under the parametric assumption as

$$\hat{\lambda}_{\mathrm{PDPI}} = \frac{b(\tilde{\boldsymbol{\theta}})}{a(\tilde{\boldsymbol{\theta}})}. \tag{6}$$

We designate $\hat{\lambda}_{\mathrm{PDPI}}$ as the *parametric direct plug-in tuning parameter*. The method (5) is a consistent estimator of λ_{opt}. Therefore, the mean and mean squared error of (5), respectively, equal $\lambda_{\mathrm{opt}} + o(1)$ and $o(1)$. Implementation of (6) is often easier than that of (5). In the following, we present some examples of $\hat{\lambda}_{\mathrm{PDPI}}$.

3.1 Examples of $\hat{\lambda}_{\mathrm{PDPI}}$

Example 1 We deal with the ridge-type estimator described in Sect. 2, where we note that the i.i.d. framework is extended to regression settings (see, e.g., Fan and Li 2001). We consider a quadratic polynomial regression model in which σ^2 is known,

$$f(y|x; \boldsymbol{\beta}) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{(y-\beta_1-\beta_2 x-\beta_3 x^2)^2}{2\sigma^2}\right\},$$

with a parameter vector $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)^T \in \mathbb{R}^3$. Let $(Y_1, X_1), \ldots, (Y_n, X_n)$ be the n observation pairs. The penalty function is $k(\boldsymbol{\beta}) = ||\boldsymbol{\beta}||^2/(2\sigma^2)$. For the true parameter vector $\boldsymbol{\beta}_0$, the expected log-likelihood is $\eta(\hat{\boldsymbol{\beta}}; G) = -||X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)||^2/(2\sigma^2) - n/(2\sigma^2) - (n/2)\log(2\pi\sigma^2)$, where X is the $n \times 3$ design matrix whose (i, α)-element is X_α^{i-1}. The optimal tuning parameter $\lambda_{\mathrm{opt}} = b(\boldsymbol{\beta}_0)/a(\boldsymbol{\beta}_0) + o(1)$ in (4) is calculated as $a(\boldsymbol{\beta}) = \boldsymbol{\beta}^T(X^TX/n)^{-1}\boldsymbol{\beta}/\sigma^2$ and $b(\boldsymbol{\beta}) = \mathrm{tr}\{(X^TX/n)^{-1}\}$. Accordingly, we have

$$\hat{\lambda}_{\mathrm{PDPI}} = \frac{\sigma^2\mathrm{tr}\{(X^TX)^{-1}\}}{\hat{\boldsymbol{\beta}}^T(X^TX)^{-1}\hat{\boldsymbol{\beta}}},$$

where $\hat{\boldsymbol{\beta}}$ is the ordinary least-squares estimator. The optimal tuning parameter in this example is a special case of Wand's (1999) $\lambda_{\mathrm{AMASE},1}$ when $m = X\boldsymbol{\beta}$, $D = I_p$ in his notation.

Example 2 The model used here is the normal distribution model, $\{N(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+\}$ with a penalty $k(\mu, \sigma^2) = \mu^2/(2\sigma^2)$. The MPLE for (μ, σ^2) is the maximizer of

$$n\,\eta(\mu, \sigma^2; \hat{G}) + k(\mu, \sigma^2) = \frac{n}{2}\log(2\pi\sigma^2) + \sum_{\alpha=1}^{n}\frac{(X_\alpha-\mu)^2}{2\sigma^2} + \lambda\frac{\mu^2}{2\sigma^2}.$$

They are given explicitly as $\hat{\mu}_\lambda = \bar{X}/(1+\lambda/n)$ and $\hat{\sigma}_\lambda^2 = \sum(X_\alpha - \hat{\mu}_\lambda)^2/n + \lambda\,\hat{\mu}_\lambda^2/n$. For the true parameters (μ_0, σ_0^2), the expected log-likelihood is $\eta(\hat{\mu}, \hat{\sigma}^2; G) = -(\hat{\mu} - \mu_0)^2/(2\hat{\sigma}^2) - \sigma_0^2/(2\hat{\sigma}^2) - \log(2\pi\hat{\sigma}^2)/2$. The optimal tuning parameter λ_{opt} is calculated using $a(\mu, \sigma^2) = \mu^2/\sigma^2 + \mu^4/(2\sigma^4)$ and $b(\mu, \sigma^2) = 1 + 3\mu^2/\sigma^2$. Accordingly,

$$\hat{\lambda}_{\mathrm{PDPI}} = \frac{\hat{\mu}^2/\hat{\sigma}^2 + \hat{\mu}^4/(2\hat{\sigma}^4)}{1 + 3\hat{\mu}^2/\hat{\sigma}^2},$$

where $(\hat{\mu}, \hat{\sigma})$ is the MLE.

Example 3 Here we consider a re-parameterized exponential density: $f\{x; \mu(\theta)\} = \mu(\theta)e^{-\mu(\theta)x}$, where $\mu(\theta) = e^{\theta}$, $\theta \in \mathbb{R}$ and $x \geq 0$. The penalty function is $k(\theta) = \theta^2/2$. For the true parameter θ_0, the expected log-likelihood is $\eta(\hat{\theta}; G) = \hat{\theta} - e^{\hat{\theta}}/e^{\theta_0}$. The optimal tuning parameter λ_{opt} is calculated as $a(\theta) = \theta^2$ and $b(\theta) = 1 + \theta$. Accordingly,

$$\hat{\lambda}_{\text{PDPI}} = \frac{1 - \hat{\theta}}{\hat{\theta}^2},$$

where $\hat{\theta}$ is the MLE.

3.2 A numerical differentiation approach in estimation of $b(\cdot)$

It is often cumbersome to implement (5) since estimation of $b(\boldsymbol{\theta}_0)$ in (4) requires third-order tensor computation. Here we propose a method to avoid the difficulties by using a numerical differentiation as used in DiCiccio and Efron (1992, 1996) and DiCiccio and Monti (2001). The motivation comes from the facts that

$$\begin{aligned}\boldsymbol{q}(\boldsymbol{\theta})^T \boldsymbol{d}(\boldsymbol{\theta}) &= \sum_{i,j,k} q_i(\boldsymbol{\theta}) \frac{\partial^3 \eta(\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j \partial\theta_k} V_{jk}(\boldsymbol{\theta}) = \sum_{j,k} \frac{\mathrm{d}}{\mathrm{d}\epsilon} \frac{\partial^2 \eta\{\boldsymbol{\theta} + \epsilon \boldsymbol{q}(\boldsymbol{\theta})\}}{\partial\theta_j \partial\theta_k} \bigg|_{\epsilon=0} V_{jk}(\boldsymbol{\theta}) \\ &= -\frac{\mathrm{d}}{\mathrm{d}\epsilon} \mathrm{tr}\left[J\left\{\boldsymbol{\theta} + \epsilon \boldsymbol{q}(\boldsymbol{\theta})\right\} V(\boldsymbol{\theta})\right] \bigg|_{\epsilon=0}, \qquad (7)\end{aligned}$$

and, similarly,

$$\begin{aligned}2\boldsymbol{q}(\boldsymbol{\theta})^T \boldsymbol{e}(\boldsymbol{\theta}) &= \sum_{j,k} J^{jk} \frac{\mathrm{d}}{\mathrm{d}\epsilon} I_{jk}\{\boldsymbol{\theta} + \epsilon \boldsymbol{q}(\boldsymbol{\theta})\} \bigg|_{\epsilon=0} \\ &= \frac{\mathrm{d}}{\mathrm{d}\epsilon} \mathrm{tr}\left[\{J(\boldsymbol{\theta})\}^{-1} I\left\{\boldsymbol{\theta} + \epsilon \boldsymbol{q}(\boldsymbol{\theta})\right\}\right] \bigg|_{\epsilon=0}. \qquad (8)\end{aligned}$$

The right-hand sides in both (7) and (8) enable us to use a numerical differentiation approach, in which we use suitable consistent estimators $\hat{I}(\cdot)$ and $\hat{J}(\cdot)$ of $I(\cdot)$ and $J(\cdot)$, respectively. Lastly, we substitute a consistent estimator $\tilde{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}_0$ into $\boldsymbol{\theta}$ in both (7) and (8). The method is advantageous to avoid the difficulties caused by the third-order tensor quantities in $b(\cdot)$. See also Sect. 7 for practical implementations.

4 Tuning parameters selected using the GIC and CV

The previous section presented a description of a direct plug-in method for tuning parameter selection. Although the method is useful, it sometimes requires complicated analytic calculations. In such cases, it is valuable to implement some existing methods, e.g., a generalized information criterion (GIC; Konishi and Kitagawa 1996) applied to MPLE, BIC-type criteria, cross-validation (CV; Stone 1974), and others. In tuning parameter selection for the MPLE, it remains a noticeable problem that seems not to have been investigated: 'How well does the selection method perform using these selection criteria?' This section analyzes the properties of λ selected with GIC and CV through the result (4).

First, we consider the GIC. According to the fact that the MPLE is an M-estimator (Konishi and Kitagawa 1996), the GIC applied to the selection of λ is given as

$$\mathrm{GIC}(\lambda) = -2\sum_{\alpha=1}^{n} \log f(X_\alpha; \hat{\boldsymbol{\theta}}_\lambda) + 2\mathrm{tr}\left[\{\hat{R}_\lambda(\hat{\boldsymbol{\theta}}_\lambda)\}^{-1}\hat{Q}_\lambda(\hat{\boldsymbol{\theta}}_\lambda)\right],$$

where

$$\hat{R}_\lambda(\boldsymbol{\theta}) = -\frac{1}{n}\sum_{\alpha=1}^{n} \frac{\partial^2\{\log f(X_\alpha;\boldsymbol{\theta}) - \frac{\lambda}{n}k(\boldsymbol{\theta})\}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T},$$

$$\hat{Q}_\lambda(\boldsymbol{\theta}) = \frac{1}{n}\sum_{\alpha=1}^{n} \frac{\partial\{\log f(X_\alpha;\boldsymbol{\theta}) - \frac{\lambda}{n}k(\boldsymbol{\theta})\}}{\partial\boldsymbol{\theta}}\frac{\partial \log f(X_\alpha;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^T}.$$

Define $\hat{\lambda}_{\mathrm{GIC}} = \mathrm{argmin}_\lambda\, \mathrm{GIC}(\lambda)$.

Theorem 2 *Suppose that the assumptions provided in Appendix hold. Then $\hat{\lambda}_{\mathrm{GIC}}$ is a consistent estimator of λ_{opt}.*

Second, we consider the CV, which uses the leave-one-out method:

$$\mathrm{CV}(\lambda) = -\sum_{\alpha=1}^{n} \log f(X_\alpha; \hat{\boldsymbol{\theta}}_\lambda^{(-\alpha)}).$$

In that equation, $\hat{\boldsymbol{\theta}}_\lambda^{(-\alpha)}$ is the MPLE with the observations $(X_1, \ldots, X_{\alpha-1}, X_{\alpha+1}, \ldots, X_n)$. Define $\hat{\lambda}_{\mathrm{CV}} = \mathrm{argmin}_\lambda\, \mathrm{CV}(\lambda)$.

Theorem 3 *Suppose that the assumptions provided in Appendix hold. Then $\hat{\lambda}_{\mathrm{CV}}$ is a consistent estimator of λ_{opt}.*

Comparing Theorems 2 and 3 implies an asymptotic equivalency between $\hat{\lambda}_{\mathrm{GIC}}$ and $\hat{\lambda}_{\mathrm{CV}}$.

5 Behaviors of mean Kullback–Leibler information with $\hat{\lambda}_{\mathrm{DPI}}$, $\hat{\lambda}_{\mathrm{GIC}}$ and $\hat{\lambda}_{\mathrm{CV}}$

The goal of selecting the tuning parameter is not to estimate the optimal tuning parameter, but to improve the performance of estimation. Therefore, it is more important to analyze the behavior of mean Kullback–Leibler information. Here we study the behaviors with the direct plug-in tuning parameter and the tuning parameters selected by the GIC and CV.

Theorem 4 *Assume that the assumptions provided in Appendix hold. Then the expected log-likelihoods with $\hat{\lambda}_{\mathrm{DPI}}$, $\hat{\lambda}_{\mathrm{GIC}}$ and $\hat{\lambda}_{\mathrm{CV}}$ are given as*

$$\begin{aligned}
\eta(\hat{\boldsymbol{\theta}}_{\hat{\lambda}_{\mathrm{DPI}}}) &= \eta(\hat{\boldsymbol{\theta}}) - \frac{\lambda_{\mathrm{opt}}}{n^2}\sum_{i,j}^{p} q_i c_{j,n}^{(1)} \kappa_{ij} + o_p(n^{-3/2}),\\
\eta(\hat{\boldsymbol{\theta}}_{\hat{\lambda}_{\mathrm{GIC}}}) &= \eta(\hat{\boldsymbol{\theta}}_{\hat{\lambda}_{\mathrm{DPI}}}) + o_p(n^{-3/2}) \quad \textit{and}\\
\eta(\hat{\boldsymbol{\theta}}_{\hat{\lambda}_{\mathrm{CV}}}) &= \eta(\hat{\boldsymbol{\theta}}_{\hat{\lambda}_{\mathrm{DPI}}}) + o_p(n^{-3/2}),
\end{aligned}$$

respectively, where $c_{j,n}^{(1)}$ and κ_{ij} are defined in the Appendix.

It is noteworthy that $\hat{\lambda}_{\mathrm{DPI}}$, $\hat{\lambda}_{\mathrm{GIC}}$ and $\hat{\lambda}_{\mathrm{CV}}$ possess equivalent performance asymptotically.

6 Numerical results

In this section, we provide some numerical results for Examples 1–3 in Sect. 3. We compare the proposed method, the direct plug-in tuning parameter, with the tuning parameters selected by the GIC and CV. The settings for simulations are as follows:

Example 1 (continued). Artificial data are from the model with a true parameter $\boldsymbol{\beta}_0 = (-3/4, 1, 1)^T$ and $\sigma^2 = 1$, in which we employ the fixed design that the n covariates are spaced equally on $[-1, 1]$. In selecting λ with GIC and CV, the candidates of λ's are $(0, 0.1, 0.2, 0.3, \ldots, 99.9, 100)$.

Example 2 (continued). The artificial data are from the model with a true parameter $(\mu_0, \sigma_0^2) = (2, 2^2)$. In selecting λ with GIC and CV, the candidates of λ's are (0, 0.1, 0.2, 0.3, ..., 19.9, 20).

Example 3 (continued). The artificial data are from the model with a true parameter $\theta_0 = 2$. In selecting λ with GIC and CV, the candidates of λ's are $(-1, -0.99, -0.98, -0.97, \ldots$, 2.99,3).

We repeat the numerical experiments 1,000 times for two cases: $n = 20$ and $n = 50$. Table 1 shows the summaries of the tuning parameters selected by each method in 1,000 simulations. Figures 1, 2, and 3 show the boxplots for estimators $\hat{\lambda}_{\mathrm{PDPI}}$, $\hat{\lambda}_{\mathrm{GIC}}$, and $\hat{\lambda}_{\mathrm{CV}}$, together with the dotted horizontal line, which indicates the optimal tuning parameter λ_{opt} given by Eq. (4), where only Fig. 1 displays the results for $\log(\lambda + 0.1)$ because of the large variability relative to Examples 2 and 3. The simulation results illustrate that $\hat{\lambda}_{\mathrm{GIC}}$, $\hat{\lambda}_{\mathrm{CV}}$ and $\hat{\lambda}_{\mathrm{PDPI}}$ converge to λ_{opt} as n increases. Notably, there are large difference between the mean value of DPI and those of GIC and CV in Example 1 when $n = 20$. This observation was caused by some cases that the criteria diverged as the tuning parameter gets larger. This might mean that the sample size $n = 20$ is somewhat fewer than the sample size enough to get stable selections with GIC and CV in Example 1. Apparently, the proposed method $\hat{\lambda}_{\mathrm{PDPI}}$ performs better than the other methods because of the smaller mean squared errors (Table 1) and the minor variations (Figs. 1, 2, 3). In all simulation results, $\hat{\lambda}_{\mathrm{GIC}}$ is more stable than $\hat{\lambda}_{\mathrm{CV}}$.

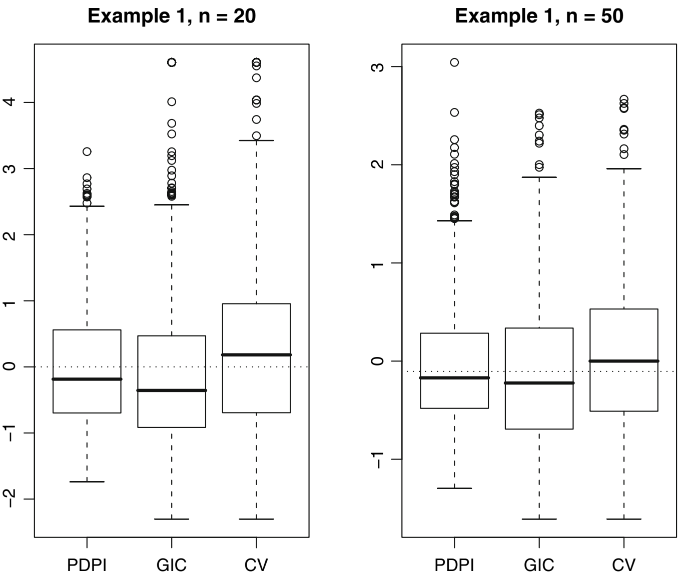

Fig. 1 Example 1. *Boxplots* for three estimates of λ_{opt}, for $n = 20$ and $n = 50$, in 1,000 simulations, where $\log(\lambda + 0.1)$ are displayed. The *dotted horizontal line* indicates $\log(\lambda_{\text{opt}} + 0.1)$

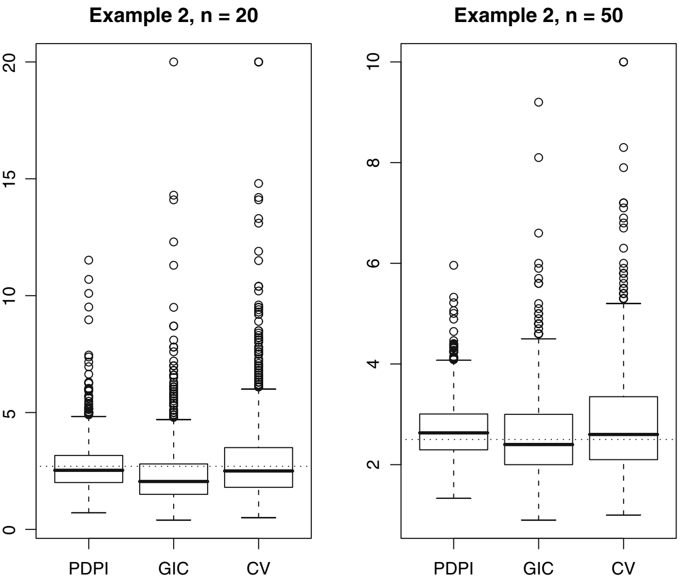

Fig. 2 Example 2. *Boxplots* for three estimates of λ_{opt}, for $n = 20$ and $n = 50$, in 1,000 simulations. The *dotted horizontal line* indicates λ_{opt}

Table 1 Summaries of numerical experiments in Examples 1–3

Example	n	λ_{opt}	b/a	PDPI Mean	MSE	GIC mean	MSE	CV mean	MSE
1	20	0.9	0.77	1.45	4.63	3.56	218.21	5.85	381.84
	50	0.8	0.78	1.11	1.69	1.09	1.60	1.33	2.28
2	20	2.7	2.67	2.72	1.25	2.41	2.46	3.05	4.49
	50	2.5	2.67	2.70	0.39	2.54	0.72	2.84	1.20
3	20	0.76	0.75	0.76	0.01	0.64	0.09	0.75	0.13
	50	0.75	0.75	0.75	0.01	0.70	0.04	0.76	0.06

In the 1,000 simulations, for $n = 20$ and $n = 50$, λ_{opt}, b/a in (4) and the mean and MSE with respect to λ_{opt}, for $\hat{\lambda}_{\text{PDPI}}$, $\hat{\lambda}_{\text{GIC}}$ and $\hat{\lambda}_{\text{CV}}$, are shown

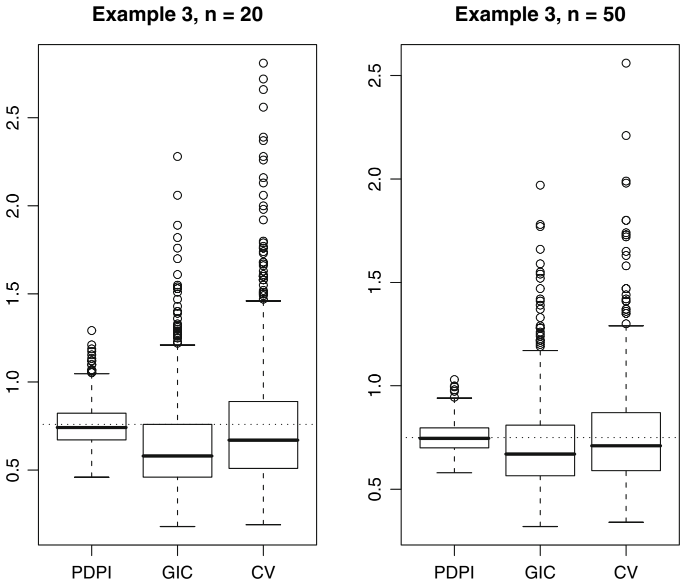

Fig. 3 Example 3. *Boxplots* for three estimates of λ_{opt}, for $n = 20$ and $n = 50$, in 1,000 simulations. The *dotted horizontal line* indicates λ_{opt}

7 Tuning parameter selections in penalized spline regressions

This section demonstrates the direct plug-in tuning parameters in more practical cases of smoothing problems, "penalized spline generalized linear model regressions". See Wand (1999) for a penalized spline least-squares regression. The degree p penalized spline uses a transformation of the univariate covariate x into a $(p+K+1)$-dimensional vector $\boldsymbol{\phi}(x) = \{1, x, x^2, \ldots, x^p, (x-\kappa_1)_+^p, (x-\kappa_2)_+^p, \ldots, (x-\kappa_K)_+^p\}$, where $(y)_+ = \max\{0, y\}$, and κ_j's are called the knots. Suppose that we have n pairs of observations $(Y_\alpha, x_\alpha)_{\alpha=1,\ldots,n}$ and define X be an $n \times (p + K + 1)$ design matrix whose αth row is $\boldsymbol{\phi}(x_\alpha)$. Now we consider a penalized spline generalized linear model regression. The method models the conditional distribution of Y given x by $f\{y; \mu(x)\}$, where f is a

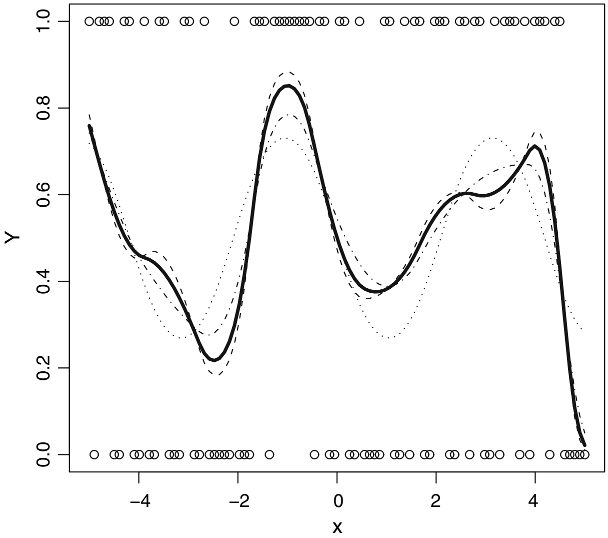

Fig. 4 Example 4. One random sample example ($n = 100$) of the penalized spline with λ using PDPI (*bold*), GIC (*dashed*) and CV (*long dashed*) together with the true curve $\mu(x)$ (*dotted*)

member of an exponential family,

$$f(y; \mu) = f(y; \mu, \psi) = \exp\left[\frac{y\xi(\mu) - u\{\xi(\mu)\}}{\psi} + v(y, \psi)\right],$$

where $\xi(\mu) = u'^{-1}(\mu)$ represents the natural parameter, $u(\cdot)$ is the cumulant function, v is the normalizing factor, ψ is an unknown scale parameter and $\mu(x) = E(Y|X = x)$ is the conditional expectation of Y given x. The expectation function $\mu(x)$ is approximated by a transformed linear combination $h\{\boldsymbol{\phi}(x)\boldsymbol{\beta}\}$ by a link function $h(\cdot)$, where $\boldsymbol{\beta}$ is a $(p + K + 1)$-dimensional coefficients vector. We take ξ^{-1} as a link function h (canonical link) hereafter. The MLE for $\boldsymbol{\beta}$ often results in an over-smoothed estimation. A penalized method then works successively. To estimate $\boldsymbol{\beta}$, we maximize the following penalized log-likelihood criterion about $\boldsymbol{\beta}$,

$$\sum_{\alpha=1}^{n} \log f[Y_\alpha; \xi^{-1}\{\boldsymbol{\psi}(x_\alpha)\boldsymbol{\beta}\}] - \frac{\lambda}{2}\boldsymbol{\beta}^T D\boldsymbol{\beta}, \tag{9}$$

where $D = \mathrm{diag}(\mathbf{0}_{p+1}, \mathbf{1}_K)$ is a $(p+K+1)$-diagonal matrix representing a roughness penalty, in which $\mathbf{0}_k$ and $\mathbf{1}_k$ denote k-dimensional vectors of, respectively, zeros and ones. Denoting the above MPLE by $\hat{\boldsymbol{\beta}}_\lambda$ with a tuning parameter λ, we obtain an estimate of $\mu(x)$ by $\hat{\mu}_\lambda(x) = \xi^{-1}\{\boldsymbol{\phi}(x)\hat{\beta}_\lambda\}$. One can see that the first term of (9) is the criterion in the familiar generalized linear model regression with design matrix X and a coefficient vector $\boldsymbol{\beta}$.

In what follows, we apply our DPI method. Notably, the estimation of $b(\boldsymbol{\beta}_0)$ in the DPI method requires a complicated third-order tensor computation. Thus, we use a numerical differentiation approach presented in Sect. 3.2. From the fact that there appears no y in the expression of $\partial^2 \log f[y; \xi^{-1}\{\boldsymbol{\psi}(x)\boldsymbol{\beta}\}]/\partial\beta_i\partial\beta_j = -u''\{\boldsymbol{\psi}(x)\boldsymbol{\beta}\}\psi_i(x)\psi_j(x)/\psi$, we can find that $\partial^k\hat{\eta}(\cdot)/\partial\beta_{i_1}\ldots\partial\beta_{i_k} = \partial^k\eta(\cdot)/\partial\beta_{i_1}\ldots\partial\beta_{i_k}$ for $k \geq 2$, where "hat" means that the empirical distribution is substituted, and that $\boldsymbol{e}(\boldsymbol{\beta}_0)$ is zero. Consequently, we apply the numerical differentiation to compute $\hat{\boldsymbol{q}}(\hat{\boldsymbol{\beta}})^T\hat{\boldsymbol{d}}(\hat{\boldsymbol{\beta}})$, where $\hat{\boldsymbol{\beta}}$ is the MLE. In addition we use the parametric assumption, then $I = J$. Then (7) becomes

$$\hat{\boldsymbol{q}}(\hat{\boldsymbol{\beta}})^T\hat{\boldsymbol{d}}(\hat{\boldsymbol{\beta}}) = -\frac{\mathrm{d}}{\mathrm{d}\epsilon}\mathrm{tr}\left[J\{\hat{\boldsymbol{\beta}} + \epsilon\boldsymbol{q}(\hat{\boldsymbol{\beta}})\}\{J(\hat{\boldsymbol{\beta}})\}^{-1}\right]\Bigg|_{\epsilon=0}. \tag{10}$$

In this setting, $J(\hat{\boldsymbol{\beta}}) = X^T\hat{W}X/n$ where $\hat{W}$ is a diagonal matrix whose (α, α)-element is $u''\{\boldsymbol{\psi}(x_\alpha)\hat{\boldsymbol{\beta}}\}/\psi$,

$$a(\hat{\boldsymbol{\beta}}) = \hat{\boldsymbol{\beta}}^T D(X^T\hat{W}X/n)^{-1}D\hat{\boldsymbol{\beta}},$$

and

$$b(\hat{\boldsymbol{\beta}}) = \mathrm{tr}\{D(X^T\hat{W}X/n)^{-1}\} - \frac{\mathrm{d}}{\mathrm{d}\epsilon}\mathrm{tr}\left[J\{\hat{\boldsymbol{\beta}} + \epsilon\boldsymbol{q}(\hat{\boldsymbol{\beta}})\}\{J(\hat{\boldsymbol{\beta}})\}^{-1}\right]\Bigg|_{\epsilon=0}.$$

Using these quantities, $\hat{\lambda}_{\mathrm{PDPI}} = b(\hat{\boldsymbol{\beta}})/a(\hat{\boldsymbol{\beta}})$. We examine the following simulations or $\hat{\lambda}_{\mathrm{PDPI}}$, where we employ the Fisher scoring algorithm to compute the MPLE (e.g., Green and Silverman 1994; Konishi et al. 2004). We employ "grad" function of "numDeriv" package (Gilbert 2006) in the statistical software R to carry out the numerical differentiation. Throughout our simulations, we use n equally spaced design points in $[-5, 5]$, i.e., $x_\alpha = -5 + 10(\alpha - 1)/(n - 1)$, $\alpha = 1, \ldots, n$, eight fixed knots $\kappa_j = -5 + 10j/9$, $(j = 1, \ldots, 8)$ and the degree $p = 2$. In addition, we compare the selections using GIC and CV, in which the candidates of λ are $(0, 1, 4, 9, 16, \ldots, 50^2)/100$. We took the simulations 500 times for two cases of $n = 50$ and 100. We also set $\hat{\lambda}_{\mathrm{PDPI}}$ be zero if $\hat{\lambda}_{\mathrm{PDPI}}$ takes negative value.

Example 4 (Penalized spline Bernoulli regression) We consider a Bernoulli regression model: $f(y; \mu) = \mu^y(1 - \mu)^y$, $u(\xi) = \log(1 + e^\xi)$, $\psi = 1$ and $v(y, \psi) = 0$. We take the data from the model $f\{y_\alpha; \mu(x_\alpha)\}$ with $\mu(x) = 1/[1 + \exp\{\sin(1.5x)\}]$. Figure 4 illustrates one random sample example ($n = 100$) of the penalized spline with λ using PDPI (bold), GIC (dashed) and CV (long dashed) together with the true curve $\mu(x)$ (dotted). One can imagine how the penalized spline method with tuning parameter estimators behaves. Figure 5 shows the boxplot of the estimated λ's by each method, where $\log(\lambda + 0.1)$ is shown. Figure 6 displays values of the true expected log-likelihoods $\eta(\hat{\boldsymbol{\beta}}_{\hat{\lambda}})$ for each estimate $\hat{\boldsymbol{\beta}}_{\hat{\lambda}}$ where $\hat{\lambda}$ represents the selected λ. The average expected log-likelihoods are -22.1, -21.5 and -24.9 for PDPI, GIC and CV when $n = 50$, and -44.1, -44.2 and -47.6 for PDPI, GIC and CV when $n = 100$.

Example 5 (Penalized spline Poisson regression) We consider a Poisson regression model: $f(y; \mu) = e^{-\mu}\mu^y/y!$, $u(\xi) = e^\xi$, $\psi = 1$ and $v(y, \psi) = -\log(y!)$. We

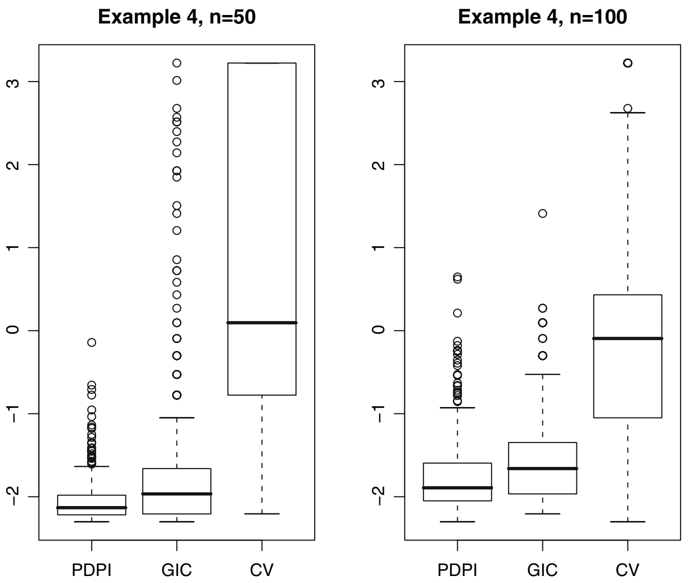

Fig. 5 Example 4. *Boxplots* for three estimates of λ_{opt}, for $n = 50$ and $n = 100$, in 500 simulations, where $\log(\lambda + 0.1)$ are displayed

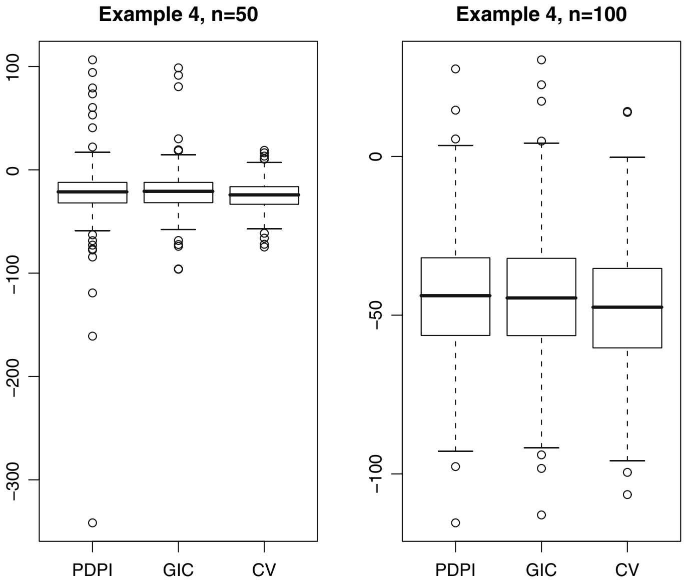

Fig. 6 Example 4. *Boxplots* for the resulting expected log-likelihoods with three estimates of λ_{opt}, for $n = 50$ and $n = 100$, in 500 simulations

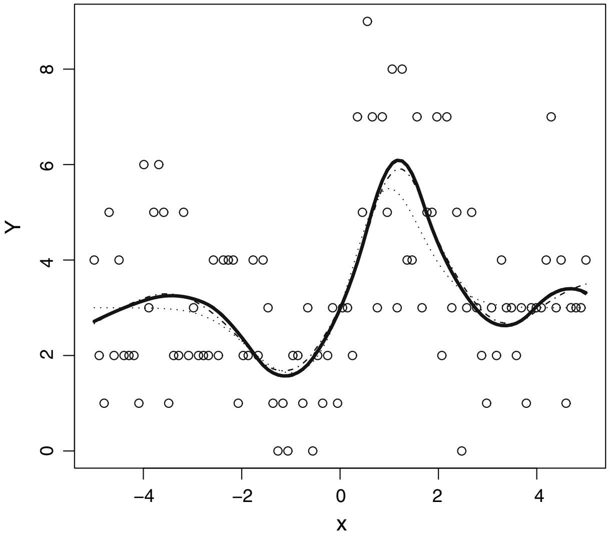

Fig. 7 Example 5. One random sample ($n = 100$) example of the penalized spline with λ using PDPI (*bold*), GIC (*dashed*) and CV (*long dashed*) together with the true curve $\mu(x)$ (*dotted*)

take the data from the model $f\{y_\alpha; \mu(x_\alpha)\}$ with $\mu(x) = \exp\{x \exp(-x^2/2) + \log 3\}$. Figure 7 illustrates one random sample example ($n = 100$) of the penalized spline with λ using PDPI (bold), GIC (dashed) and CV (long dashed) together with the true curve $\mu(x)$ (dotted). Figure 8 shows the boxplot of the estimated λ's by each method, where $\log(\lambda + 0.1)$ is shown. Figure 9 displays values of the true expected log-likelihoods $\eta(\hat{\boldsymbol{\beta}}_{\hat{\lambda}})$ for each estimate $\hat{\boldsymbol{\beta}}_{\hat{\lambda}}$ where $\hat{\lambda}$ represents the selected λ (in which, model independent constant is removed). The average expected log-likelihoods are 33.7, 33.7 and 32.1 for PDPI, GIC and CV when $n = 50$, and 62.6, 62.7 and 61.7 for PDPI, GIC and CV when $n = 100$.

Figures 5 and 8 tell that estimated λ's behave as in the previous Examples 1–3. It is important to emphasize the expected log-likelihoods for comparison of each method's performance, in which large expected log-likelihoods is preferable. Figures 6 and 9 state that the PDPI method is useful. Moreover the performance with the PDPI method is near to that using the GIC.

8 Concluding remarks

In this paper, we propose a direct plug-in method for selecting a tuning parameter in maximum penalized likelihood methods. The use of our asymptotic result for a tuning parameter reduces the computational cost in seeking the minimum that the existing methods require. It also gives a guide for the selections using some existing methods, even if it is subtle that the regularity conditions for the direct plug-in method hold.

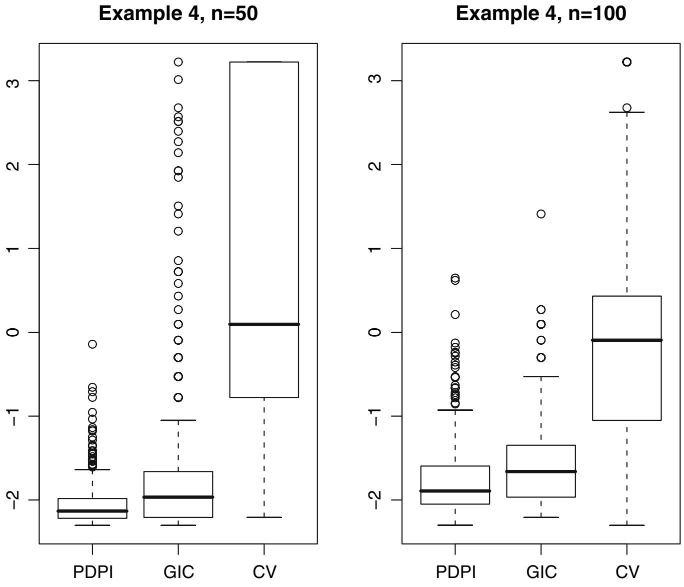

Fig. 8 Example 5. *Boxplots* for three estimates of λ_{opt}, for $n = 50$ and $n = 100$, in 500 simulations, where $\log(\lambda + 0.1)$ are displayed

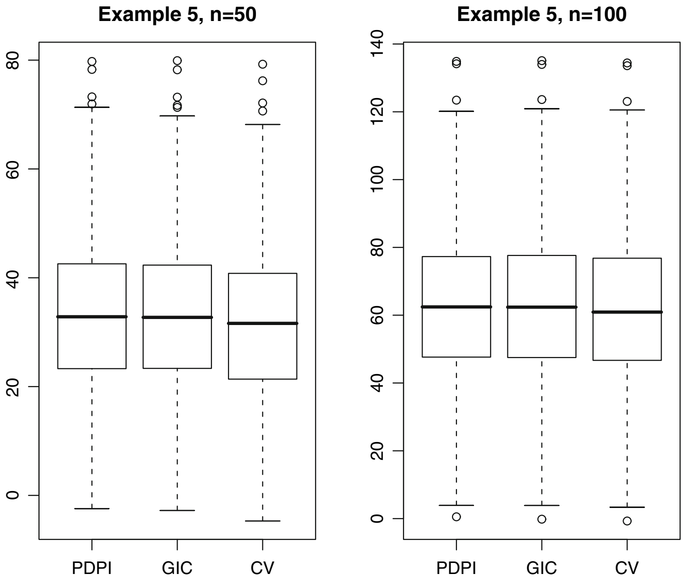

Fig. 9 Example 5. *Boxplots* for the resulting expected log-likelihoods with three estimates of λ_{opt}, for $n = 50$ and $n = 100$, in 500 simulations

Appendix A: Proofs

Assumptions

Here, we provide the following regularity conditions resembling those described in White (1982). The model $f(\cdot;\boldsymbol{\theta})$ is sufficiently smooth. It has the same support as that of the true distribution function, $G(x)$, for all $\boldsymbol{\theta}$. The parameter vectors $\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_\lambda$ and $\hat{\boldsymbol{\theta}}_\lambda$ lie in the interior for all λ. In addition, $J(\boldsymbol{\theta})$, the Hessian matrix of $-\eta(\boldsymbol{\theta})$, is positive definite at $\boldsymbol{\theta}_0$. Finally, the penalty function $k(\boldsymbol{\theta})$ is sufficiently smooth.

Preparations and useful results

The arguments to be described work mainly with the functions $S(x_1,\ldots,x_j)$, which depend on the data $X_1, X_2, \ldots$ only through the implied arguments $x_1,\ldots,x_j$, such as

$$\int S(x_1,\ldots,x_j)\mathrm{d}G(x_r)=0 \quad \text{for} \quad 1\le r\le j. \tag{11}$$

We assign a superscript (j) to a function S satisfying (11) as $S^{(j)}(x_1,\ldots,x_j)$. Then we note that $n^{-1}E\left\{\sum_{\alpha=1}^n S^{(1)}(X_\alpha)\right\}=0, n^{-2}E\left\{\sum_{\alpha,\beta=1}^n S^{(2)}(X_\alpha,X_\beta)\right\}=O(n^{-1})$ and, for $j\ge 3$, $n^{-j}E\left\{\sum_{\alpha_1,\ldots,\alpha_j=1}^n S^{(j)}(X_{\alpha_1},\ldots,X_{\alpha_j})\right\}=o(n^{-1})$, for i.i.d. samples $X_n=(X_1,\ldots,X_n)$ from a distribution $G(x)$.

Before proving the theorems, we prepare some notations:

$$\ell_{i_1\ldots i_m}(x;\boldsymbol{\theta})=\frac{\partial^m \log f(x;\boldsymbol{\theta})}{\partial\theta_{i_1}\ldots\partial\theta_{i_m}}, \quad \gamma_{i_1\ldots i_m}(\boldsymbol{\theta})=\frac{\partial^m k(\boldsymbol{\theta})}{\partial\theta_{i_1}\ldots\partial\theta_{i_m}},$$
$$\kappa_{i_1\ldots i_m}(\boldsymbol{\theta})=\frac{\partial^m \eta(\boldsymbol{\theta})}{\partial\theta_{i_1}\ldots\partial\theta_{i_m}}, \quad L_{i_1\ldots i_m}(\boldsymbol{\theta})=\frac{1}{n}\sum_{\alpha=1}^n \ell_{i_1\ldots i_m}(X_\alpha;\boldsymbol{\theta}). \tag{12}$$

Here note that $\kappa_{ij}(\boldsymbol{\theta})=-J_{ij}(\boldsymbol{\theta})$. For brevity, we omit $\boldsymbol{\theta}_0$ in $A(\boldsymbol{\theta}_0)$, where A is a function evaluated at $\boldsymbol{\theta}_0$; that is, we write $A(\boldsymbol{\theta}_0)$ as A.

Under the regularity conditions, the MLE $\hat{\boldsymbol{\theta}}$ satisfies that

$$\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0=\frac{1}{n}\boldsymbol{c}_n^{(1)}+\frac{1}{2n^2}\boldsymbol{c}_n^{(2)}+o_p(n^{-1}), \tag{13}$$

where $\boldsymbol{c}_n^{(1)}=\sum_{\alpha=1}^n \boldsymbol{T}^{(1)}(X_\alpha;\boldsymbol{\theta}_0)$ and $\boldsymbol{c}_n^{(2)}=\sum_{\alpha,\beta=1}^n \boldsymbol{T}^{(2)}(X_\alpha,X_\beta;\boldsymbol{\theta}_0)$. The vector valued functions $\boldsymbol{T}^{(1)}(x;\boldsymbol{\theta})$ and $\boldsymbol{T}^{(2)}(x_1,x_2;\boldsymbol{\theta})$, which satisfy (11), are called the first and second compact derivatives, respectively (Konishi and Kitagawa 1996, 2003). The specification of their exact expressions is given in Appendix.

Lemma 1 *Assume that the conditions given in Appendix hold. Then,*

$$\hat{\boldsymbol{\theta}}_\lambda - \hat{\boldsymbol{\theta}} = -\frac{\lambda}{n}\left(\boldsymbol{q} + \frac{1}{n}\boldsymbol{A}_n^{(1)} + \frac{1}{n^2}\boldsymbol{A}_n^{(2)}\right) + \frac{\lambda^2}{n^2}\boldsymbol{B} + o_p(n^{-2}),$$

where $\boldsymbol{A}_n^{(1)} = \sum_{\alpha=1}^n \boldsymbol{A}^{(1)}(X_\alpha; \boldsymbol{\theta}_0)$, $\boldsymbol{A}_n^{(2)} = \sum_{\alpha,\beta=1}^n \boldsymbol{A}^{(2)}(X_\alpha, X_\beta; \boldsymbol{\theta}_0)$, *and* $\boldsymbol{A}^{(1)}(x; \boldsymbol{\theta})$ *and* $\boldsymbol{A}^{(2)}(x_1, x_2; \boldsymbol{\theta})$ *are the functions which satisfy* (11)*. In addition,* $\boldsymbol{q} = \boldsymbol{q}(\boldsymbol{\theta}_0)$ *is defined in Theorem* 1 *and* $\boldsymbol{B} = \boldsymbol{B}(\boldsymbol{\theta}_0)$ *is a quantity that is independent of the observations.*

The proof of the above lemma and the specifications of $\boldsymbol{A}^{(1)}$ and $\boldsymbol{A}^{(2)}$ are given in Appendix.

Proof of Theorem 1

Taylor expansion around $\boldsymbol{\theta}_0$ and (13) yields

$$\kappa_i(\hat{\boldsymbol{\theta}}) = \sum_j^p \left(\frac{1}{n}c_{j,n}^{(1)} + \frac{1}{2n^2}c_{j,n}^{(2)}\right)\kappa_{ij}(\boldsymbol{\theta}_0) + \frac{1}{2}\sum_{j,k}^p \frac{1}{n^2}c_{j,n}^{(1)}c_{k,n}^{(1)}\kappa_{ijk}(\boldsymbol{\theta}_0) + o_p(n^{-1}), \tag{14}$$

where we use the fact that $\kappa_i(\boldsymbol{\theta}_0) = 0$ for $i = 1, \ldots, p$ by definition of $\boldsymbol{\theta}_0$. Applying Lemma 1 and (14), we have

$$\begin{aligned}\eta(\hat{\boldsymbol{\theta}}_\lambda) &= \eta(\hat{\boldsymbol{\theta}}) + \sum_i^p \left\{-\frac{\lambda}{n}\left(q_i + \frac{1}{n}A_{i,n}^{(1)} + \frac{1}{n^2}A_{i,n}^{(2)}\right) + \frac{\lambda^2}{n^2}B_i\right\}\kappa_i(\hat{\boldsymbol{\theta}}) \\ &\quad + \frac{1}{2}\sum_{i,j}^p \frac{\lambda^2}{n^2}q_i q_j \kappa_{ij}(\hat{\boldsymbol{\theta}}) + o_p(n^{-2}) \\ &= \eta(\hat{\boldsymbol{\theta}}) - \frac{\lambda}{n}\sum_{i,j}^p q_i \left\{\left(\frac{1}{n}c_{j,n}^{(1)} + \frac{1}{2n^2}c_{j,n}^{(2)}\right)\kappa_{ij} + \frac{1}{2n^2}c_{j,n}^{(1)}\sum_k^p c_{k,n}^{(1)}\kappa_{ijk}\right\} \\ &\quad - \sum_{i,j}^p \frac{\lambda}{n^3}A_{i,n}^{(1)}c_{j,n}^{(1)}\kappa_{ij} + \frac{1}{2}\sum_{i,j}^p \frac{\lambda^2}{n^2}q_i q_j \kappa_{ij} + o_p(n^{-2}).\end{aligned} \tag{15}$$

Taking expectation for (15) over X_n, we obtain

$$\begin{aligned}E\{\eta(\hat{\boldsymbol{\theta}}_\lambda)\} = E\{\eta(\hat{\boldsymbol{\theta}})\} + \frac{\lambda}{n^2}\sum_i^p \Bigg[&\frac{1}{2}\gamma_i E_X\{T_i^{(2)}(X, X)\} \\ &-\frac{1}{2}q_i \sum_{j,k}^p E_X\{T_j^{(1)}(X)T_k^{(1)}(X)\}\kappa_{ijk}\end{aligned}$$

$$+\sum_{j}^{p} E_X\{A_i^{(1)}(X)T_j^{(1)}(X)\}J_{ij}\Bigg] - \frac{\lambda^2}{2n^2}a(\boldsymbol{\theta}_0) + o(n^{-2}), \qquad (16)$$

where we adopt $-\kappa_{ij}(\boldsymbol{\theta}) = J_{ij}(\boldsymbol{\theta})$, $\gamma_i = \sum_{j=1}^{p} J_{ij}q_j$ and (11). By (29) and (30), the second term in the right-hand side of (16) reduces to

$$\frac{\lambda}{n^2}\sum_{i,j}^{p}\left[\gamma_i \sum_{k,l}^{p} E_X\{\ell_{jk}(X)\ell_l(X)\}J^{ij}J^{kl} + E_X\{A_i^{(1)}(X)T_j^{(1)}(X)\}J_{ij}\right]. \qquad (17)$$

Moreover, according to (28) and (29),

$$\begin{aligned}&\sum_{i,j}^{p} E_X\{A_i^{(1)}(X)T_j^{(1)}(X)\}J_{ij}\\ &\quad = \sum_{i,j}^{p}\left(\gamma_{ij}V_{ij} + q_i\sum_{k}^{p}\left[E_X\{\ell_{ij}(X)\ell_k(X)\}J^{jk} + \kappa_{ijk}V_{jk}\right]\right).\end{aligned}$$

Consequently, (17) equals $\lambda b(\boldsymbol{\theta}_0)/n^2$, which proves the assertion.

Proof of Theorem 2

Using the notation (12), the GIC is rewritten as

$$\frac{1}{2n}\mathrm{GIC}(\lambda) = -L(\hat{\boldsymbol{\theta}}_\lambda) + \sum_{i,j}^{p}\{\hat{R}_\lambda(\hat{\boldsymbol{\theta}}_\lambda)\}_{ij}^{-1}\left\{\hat{Q}_{ij,0}(\hat{\boldsymbol{\theta}}_\lambda) - \frac{\lambda}{n}\gamma_i L_j(\hat{\boldsymbol{\theta}}_\lambda)\right\}, \qquad (18)$$

where $\{\hat{R}_\lambda(\boldsymbol{\theta})\}_{ij}^{-1}$ and $\hat{Q}_{ij,0}(\boldsymbol{\theta})$ are the (i, j) element of the matrices $\{\hat{R}_\lambda(\boldsymbol{\theta})\}^{-1}$ and $\hat{Q}_\lambda(\boldsymbol{\theta})|_{\lambda=0}$, respectively. Noting that $L_i(\hat{\boldsymbol{\theta}}) = 0$ $(i = 1, \ldots, p)$ and Lemma 1, the first term of GIC is expanded around $\hat{\boldsymbol{\theta}}$ as

$$\begin{aligned}-L(\hat{\boldsymbol{\theta}}_\lambda) &= -L(\hat{\boldsymbol{\theta}}) - \frac{1}{2}\sum_{i,j}^{p}(\hat{\theta}_{i,\lambda} - \hat{\theta}_i)(\hat{\theta}_{j,\lambda} - \hat{\theta}_j)L_{ij}(\hat{\boldsymbol{\theta}}) + o_p(n^{-2})\\ &= -L(\hat{\boldsymbol{\theta}}) + \frac{\lambda^2}{2n^2}a + o_p(n^{-2}), \qquad (19)\end{aligned}$$

where $\kappa_{ij}(\boldsymbol{\theta}) = -J_{ij}(\boldsymbol{\theta})$ is used. On the other hand, it follows from the notation (12) and the Taylor expansion for $\hat{R}_\lambda$ around $\lambda = 0$ that

$$\begin{aligned}\{\hat{R}_\lambda(\hat{\boldsymbol{\theta}}_\lambda)\}^{-1}_{ij} &= \{\hat{R}_0(\hat{\boldsymbol{\theta}}_\lambda)\}^{-1}_{ij} - \frac{\lambda}{n}\sum_{k,l}^{p}\{\hat{R}_0(\hat{\boldsymbol{\theta}}_\lambda)\}^{-1}_{ik}\gamma_{kl}(\hat{\boldsymbol{\theta}}_\lambda)\{\hat{R}_0(\hat{\boldsymbol{\theta}}_\lambda)\}^{-1}_{jl} + o_p(n^{-1}) \\ &= \{\hat{R}_0(\hat{\boldsymbol{\theta}}_\lambda)\}^{-1}_{ij} - \frac{\lambda}{n}\sum_{k,l}^{p}\gamma_{kl}J^{ik}J^{jl} + o_p(n^{-1}).\end{aligned}$$

Therefore, the second term of GIC is

$$\begin{aligned}\sum_{i,j}^{p}\ &\Bigg[\{\hat{R}_0(\hat{\boldsymbol{\theta}}_\lambda)\}^{-1}_{ij}\left\{\hat{Q}_{ij,0}(\hat{\boldsymbol{\theta}}_\lambda) - \frac{\lambda}{n}\gamma_i L_j(\hat{\boldsymbol{\theta}}_\lambda)\right\} \\ &\quad - \frac{\lambda}{n}\sum_{k,l}^{p}\gamma_{kl}J^{ik}J^{jl}\hat{Q}_{ij,0}(\hat{\boldsymbol{\theta}}_\lambda)\Bigg] + o_p(n^{-1}).\end{aligned} \tag{20}$$

Observing the result of Lemma 1, $\hat{\boldsymbol{\theta}}_\lambda = \hat{\boldsymbol{\theta}} + O_p(n^{-1})$, it holds that $L_j(\hat{\boldsymbol{\theta}}_\lambda) = O_p(n^{-1})$ and $\hat{Q}_{ij,0}(\hat{\boldsymbol{\theta}}_\lambda) = \hat{Q}_{ij,0}(\hat{\boldsymbol{\theta}}) + O_p(n^{-1})$. In addition, by

$$\{\hat{R}_0(\hat{\boldsymbol{\theta}}_\lambda)\}^{-1}_{ij} = \{\hat{R}_0(\hat{\boldsymbol{\theta}})\}^{-1}_{ij} - \sum_{k,l,m}^{p}\frac{\lambda}{n}q_k\{\hat{R}_0(\hat{\boldsymbol{\theta}})\}^{-1}_{il}\{\hat{R}_0(\hat{\boldsymbol{\theta}})\}^{-1}_{jm}L_{klm}(\hat{\boldsymbol{\theta}}) + o_p(n^{-1})$$

and $\hat{Q}_{ij,0}(\hat{\boldsymbol{\theta}}_\lambda) = \hat{Q}_{ij,0}(\hat{\boldsymbol{\theta}}) - 2\sum_k^p \frac{\lambda}{n}q_k\partial\hat{Q}_{ij,0}(\hat{\boldsymbol{\theta}})/\partial\theta_k + o_p(n^{-1})$. Therefore,

$$\begin{aligned}\sum_{i,j}^{p}\{\hat{R}_0(\hat{\boldsymbol{\theta}}_\lambda)\}^{-1}_{ij}\hat{Q}_{ij,0}(\hat{\boldsymbol{\theta}}_\lambda) = C_0 - \frac{\lambda}{n}\sum_{i,j,k}^{p}q_k\left(\sum_{l,m}^{p}J^{il}J^{jm}L_{klm}\hat{Q}_{ij,0} + 2M_{ik,j}J^{ij}\right) \\ +o_p(n^{-1}),\end{aligned}$$

where C_0 is a constant that is independent of λ and $M_{ik,j} = \frac{1}{n}\sum_{\alpha=1}^{n}\ell_{ik}(X_\alpha;\boldsymbol{\theta}_0)\ell_j(X_\alpha;\boldsymbol{\theta}_0)$. Combining the results described above, (20) equals

$$\begin{aligned}C_0 - \frac{\lambda}{n}&\left\{\sum_{i,j,k}^{p}q_k\left(\sum_{l,m}^{p}J^{il}J^{jm}L_{klm}\hat{Q}_{ij,0} + 2M_{ik,j}J^{ij}\right)\right. \\ &\left.+ \sum_{i,j,k,l}^{p}\gamma_{kl}J^{ik}J^{jl}\hat{Q}_{ij,0}\right\} + o_p(n^{-1}).\end{aligned}$$

Substituting (19) and (20) into (18) and observing $a > 0$,

$$\hat{\lambda}_{\text{GIC}} = \frac{1}{a} \left\{ \sum_{i,j,k,l}^{p} \gamma_{kl} J^{ik} J^{jl} \hat{Q}_{ij,0} + \sum_{i,j,k}^{p} q_k \left(\sum_{l,m}^{p} J^{il} J^{jm} L_{klm} \hat{Q}_{ij,0} + 2M_{ik,j} J^{ij} \right) \right\} + o_p(1). \tag{21}$$

Thus, we have that

$$E(\hat{\lambda}_{\text{GIC}}) = \lambda_{\text{opt}} + o(1) \quad \text{and} \quad E(\hat{\lambda}^2_{\text{GIC}}) = \lambda^2_{\text{opt}} + o(1),$$

which prove the assertion.

Proof of Theorem 3

Lemma 1 implies that

$$\begin{aligned}(\hat{\boldsymbol{\theta}}_{\lambda}^{(-\alpha)} - \hat{\boldsymbol{\theta}}^{(-\alpha)}) - (\hat{\boldsymbol{\theta}}_{\lambda} - \hat{\boldsymbol{\theta}}) &= -\frac{\lambda}{n-1} \left[\boldsymbol{q} + \frac{1}{n-1} \{ \boldsymbol{A}_n^{(1)} - \boldsymbol{A}^{(1)}(X_\alpha) \} \right] \\ &\quad + \frac{\lambda}{n} \left(\boldsymbol{q} + \frac{1}{n} \boldsymbol{A}_n^{(1)} \right) + o_p(n^{-2}) \\ &= -\frac{\lambda}{n^2} \{ \boldsymbol{q} - \boldsymbol{A}^{(1)}(X_\alpha) \} + o_p(n^{-2}). \end{aligned} \tag{22}$$

Analogously, (13) implies that

$$\begin{aligned}\hat{\boldsymbol{\theta}}^{(-\alpha)} - \hat{\boldsymbol{\theta}} &= (\hat{\boldsymbol{\theta}}^{(-\alpha)} - \boldsymbol{\theta}_0) - (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \\ &= \frac{1}{n} \left\{ \frac{1}{n} \boldsymbol{c}_n^{(1)} - \boldsymbol{T}^{(1)}(X_\alpha) \right\} + \boldsymbol{C}_2 + o_p(n^{-2}), \end{aligned} \tag{23}$$

where $\boldsymbol{C}_2$ is a constant vector independent of λ, of order $O_p(n^{-2})$. Combining (22) and (23),

$$\hat{\boldsymbol{\theta}}_{\lambda}^{(-\alpha)} - \hat{\boldsymbol{\theta}}_{\lambda} = -\frac{\lambda}{n^2} \{ \boldsymbol{q} - \boldsymbol{A}^{(1)}(X_\alpha) \} + \frac{1}{n} \left\{ \frac{1}{n} \boldsymbol{c}_n^{(1)} - \boldsymbol{T}^{(1)}(X_\alpha) \right\} + \boldsymbol{C}_2 + o_p(n^{-2}). \tag{24}$$

On the other hand, Lemma 1 yields that

$$\ell_i(X_\alpha; \hat{\boldsymbol{\theta}}_{\lambda}) = \ell_i(X_\alpha; \hat{\boldsymbol{\theta}}) - \frac{\lambda}{n} \sum_{j}^{p} \left(q_j + \frac{1}{n} A_{j,n}^{(1)} \right) \ell_{ij}(X_\alpha; \hat{\boldsymbol{\theta}}) + o_p(n^{-1}). \tag{25}$$

Summarizing (24), (25) and (19), the following is obtained.

$$\begin{aligned}
\mathrm{CV}(\lambda) &= -\sum_{\alpha=1}^{n} \ell(X_\alpha; \hat{\boldsymbol{\theta}}_{\lambda}^{(-\alpha)}) \\
&= -nL(\hat{\boldsymbol{\theta}}_\lambda) - \sum_{\alpha=1}^{n}\sum_{i}^{p}(\hat{\theta}_{i,\lambda}^{(-\alpha)} - \hat{\theta}_{i,\lambda})\,\ell_i(X_\alpha; \hat{\boldsymbol{\theta}}_\lambda) + o_p(n^{-1}) \\
&= -nL(\hat{\boldsymbol{\theta}}_\lambda) \\
&\quad -\sum_{\alpha=1}^{n}\sum_{i}^{p}\left[-\frac{\lambda}{n^2}\{q_i - A_i^{(1)}(X_\alpha)\} + \frac{1}{n}\left\{\frac{1}{n}c_{i,n}^{(1)} - T_i^{(1)}(X_\alpha)\right\} + C_{i,2}\right] \\
&\quad \times \left\{\ell_i(X_\alpha; \hat{\boldsymbol{\theta}}) - \frac{\lambda}{n}\sum_{j}^{p}\left(q_j + \frac{1}{n}A_{j,n}^{(1)}\right)\ell_{ij}(X_\alpha; \hat{\boldsymbol{\theta}})\right\} + o_p(n^{-1}) \\
&= -nL(\hat{\boldsymbol{\theta}}_\lambda) + \sum_{\alpha=1}^{n}\frac{\lambda}{n^2}\sum_{i}^{p}\{q_i - A_i^{(1)}(X_\alpha)\}\,\ell_i(X_\alpha; \hat{\boldsymbol{\theta}}) \\
&\quad +\sum_{\alpha=1}^{n}\sum_{i,j}^{p}\frac{1}{n}\left\{\frac{1}{n}c_{i,n}^{(1)} - T_i^{(1)}(X_\alpha)\right\} \\
&\quad \frac{\lambda}{n}\left(q_j + \frac{1}{n}A_{j,n}^{(1)}\right)\ell_{ij}(X_\alpha; \hat{\boldsymbol{\theta}}) + C_3 + o_p(n^{-1}) \\
&= -nL(\hat{\boldsymbol{\theta}}) + \frac{\lambda^2}{2n}a - \frac{\lambda}{n^2}\sum_{\alpha=1}^{n}\sum_{i}^{p}A_i^{(1)}(X_\alpha)\ell_i(X_\alpha) \\
&\quad -\frac{\lambda}{n^2}\sum_{\alpha=1}^{n}\sum_{i,j}^{p}T_i^{(1)}(X_\alpha)q_j\ell_{ij}(X_\alpha) + C_3 + o_p(n^{-1}),
\end{aligned}$$

where C_3 is a constant independent of λ. Because $a > 0$,

$$\begin{aligned}
\hat{\lambda}_{\mathrm{CV}} &= \frac{1}{na}\sum_{\alpha=1}^{n}\sum_{i}^{p}\left\{A_i^{(1)}(X_\alpha)\ell_i(X_\alpha) + \sum_{j}^{p}q_jT_i^{(1)}(X_\alpha)\ell_{ij}(X_\alpha)\right\} + o_p(1) \\
&= \frac{1}{na}\sum_{\alpha=1}^{n}\sum_{i,j}^{p}\left[\left\{T_i^{(1)}(X_\alpha)\gamma_{ij} + 2q_i\ell_{ij}(X_\alpha) + q_i\sum_{k}^{p}T_k^{(1)}(X_\alpha)\kappa_{ijk}\right\}T_j^{(1)}(X_\alpha)\right. \\
&\quad \left. + q_iJ_{ij}T_j^{(1)}(X_\alpha)\right] + o_p(1), \qquad (26)
\end{aligned}$$

where we use (28). The same argument in the proof of Theorem 2 proves the assertion.

Proof of Theorem 4

Using the fact that $\hat{\lambda}_{\mathrm{DPI}} = \lambda_{\mathrm{opt}} + o_p(1)$, Eq. (15) where $\lambda = \hat{\lambda}_{\mathrm{DPI}}$ becomes

$$\begin{aligned}
\eta(\hat{\boldsymbol{\theta}}_{\hat{\lambda}_{\mathrm{DPI}}}) = \eta(\hat{\boldsymbol{\theta}}) - \frac{\lambda_{\mathrm{opt}} + o_p(1)}{n} \\
\times \sum_{i,j}^{p} q_i \left\{ \left(\frac{1}{n} c_{j,n}^{(1)} + \frac{1}{2n^2} c_{j,n}^{(2)} \right) \kappa_{ij} + \frac{1}{2n^2} c_{j,n}^{(1)} \sum_{k}^{p} c_{k,n}^{(1)} \kappa_{ijk} \right\} \\
- \sum_{i,j}^{p} \frac{\lambda_{\mathrm{opt}} + o_p(1)}{n^3} A_{i,n}^{(1)} c_{j,n}^{(1)} \kappa_{ij} \\
+ \frac{1}{2} \sum_{i,j}^{p} \frac{\{\lambda_{\mathrm{opt}} + o_p(1)\}^2}{n^2} q_i q_j \kappa_{ij} + o_p(n^{-2}).
\end{aligned} \tag{27}$$

Using the fact that the quantities which possess super-script of (j) are $O_p(n^{j/2})$ for $j = 1, 2$, we have the first assertion. Similarly, the second and third assertions in Theorem 4 follow from the consistency results in Theorems 2 and 3.

Proof of Lemma 1

Let $\boldsymbol{\xi} = \hat{\boldsymbol{\theta}}_\lambda - \hat{\boldsymbol{\theta}}$. The Taylor expansion around $\hat{\boldsymbol{\theta}}$ yields that, for $i = 1, \ldots, p$,

$$\begin{aligned}
0 &= -L_i(\hat{\boldsymbol{\theta}}_\lambda) + \frac{\lambda}{n} \gamma_i(\hat{\boldsymbol{\theta}}_\lambda) \\
&= \sum_{j}^{p} \xi_j \left\{ J_{ij} - \frac{1}{n} \nu_{ij,n}^{(1)}(\hat{\boldsymbol{\theta}}) \right\} - \frac{1}{2} \sum_{j,k}^{p} \xi_j \xi_k L_{ijk}(\hat{\boldsymbol{\theta}}) + \frac{\lambda}{n} \gamma_i(\hat{\boldsymbol{\theta}}_\lambda) + o_p(n^{-2}),
\end{aligned}$$

where $\nu_{ij,n}^{(1)}(\boldsymbol{\theta}) = \sum_{\alpha=1}^{n} \nu_{ij,n}^{(1)}(X_\alpha; \boldsymbol{\theta})$ and $\nu_{ij,n}^{(1)}(x; \boldsymbol{\theta}) = \ell_{ij}(x; \boldsymbol{\theta}) + J_{ij}$, where $\nu_{ij,n}^{(1)}(\boldsymbol{\theta})$ satisfies (11). Furthermore,

$$\begin{aligned}
\sum_{j}^{p} \xi_j J_{ij} &= \frac{1}{n} \sum_{j}^{p} \xi_j \nu_{ij,n}^{(1)}(\hat{\boldsymbol{\theta}}) + \frac{1}{2} \sum_{j,k}^{p} \xi_j \xi_k L_{ijk}(\hat{\boldsymbol{\theta}}) \\
&\quad - \frac{\lambda}{n} \left\{ \gamma_i(\hat{\boldsymbol{\theta}}) + \sum_{j}^{p} \xi_j \gamma_{ij}(\hat{\boldsymbol{\theta}}) \right\} + o_p(n^{-2}) \\
&= -\frac{\lambda}{n} \gamma_i(\hat{\boldsymbol{\theta}}) + \frac{1}{n} \sum_{j,k}^{p} J^{jk} \left\{ -\frac{\lambda}{n} \gamma_k(\hat{\boldsymbol{\theta}}) + \frac{1}{n} \sum_{l}^{p} \xi_l \nu_{kl,n}^{(1)}(\hat{\boldsymbol{\theta}}) \right\} \nu_{ij,n}^{(1)}(\hat{\boldsymbol{\theta}}) \\
&\quad + \frac{1}{2} \sum_{j,k}^{p} \xi_j \xi_k L_{ijk}(\hat{\boldsymbol{\theta}}) - \frac{\lambda}{n} \sum_{j}^{p} \xi_j \gamma_{ij}(\hat{\boldsymbol{\theta}}) + o_p(n^{-2})
\end{aligned}$$

$$
\begin{aligned}
&= -\frac{\lambda}{n}\gamma_i(\hat{\boldsymbol{\theta}}) - \frac{\lambda}{n^2}\sum_{j,k}^{p} J^{jk}\gamma_k(\hat{\boldsymbol{\theta}})\nu_{ij,n}^{(1)}(\hat{\boldsymbol{\theta}}) \\
&\quad - \sum_{j,k,l,m} J^{jk}J^{lm}\frac{\lambda}{n^3}\gamma_m(\hat{\boldsymbol{\theta}})\nu_{kl,n}^{(1)}(\hat{\boldsymbol{\theta}})\nu_{ij,n}^{(1)}(\hat{\boldsymbol{\theta}}) \\
&\quad + \frac{\lambda^2}{2n^2}\sum_{j,k,l,m}^{p} J^{jl}J^{km}\gamma_l(\hat{\boldsymbol{\theta}})\gamma_m(\hat{\boldsymbol{\theta}})L_{ijk}(\hat{\boldsymbol{\theta}}) \\
&\quad + \frac{\lambda^2}{n^2}\sum_{j,k}^{p} J^{jk}\gamma_k(\hat{\boldsymbol{\theta}})\gamma_{ij}(\hat{\boldsymbol{\theta}}) + o_p(n^{-2}) \\
&\equiv -\frac{\lambda}{n}I_{i,1} - \frac{\lambda}{n}I_{i,2} - \frac{\lambda}{n}I_{i,3} + \frac{\lambda^2}{2n^2}I_{i,4} + \frac{\lambda^2}{n^2}I_{i,5} + o_p(n^{-2}).
\end{aligned}
$$

In what follows, we shall compute $I_{i,1}, \ldots, I_{i,5}$ specifically. By (13),

$$
I_{i,1} = \gamma_i + \sum_{j}^{p}\left(\frac{1}{n}c_{j,n}^{(1)} + \frac{1}{2n^2}c_{j,n}^{(2)}\right)\gamma_{ij} + \frac{1}{2n^2}\sum_{j,k}^{p} c_{j,n}^{(1)}c_{k,n}^{(1)}\gamma_{ijk} + o_p(n^{-1}).
$$

Put $\nu_{ijk,n}^{(1)}(\boldsymbol{\theta}) = \sum_{\alpha=1}^{n}\nu_{ijk}^{(1)}(X_\alpha;\boldsymbol{\theta})$ and $\nu_{ijk}^{(1)}(x;\boldsymbol{\theta}) = \ell_{ijk}(x;\boldsymbol{\theta}) - \kappa_{ijk}$, where $\nu_{ijk,n}^{(1)}(\boldsymbol{\theta})$ satisfies (11). Then $I_{i,2}$ equals

$$
\begin{aligned}
&\sum_{j,k}^{p} J^{jk}\left(\gamma_k + \sum_{l}^{p}\frac{1}{n}c_{l,n}^{(1)}\gamma_{kl}\right)\left\{\frac{1}{n}\nu_{ij,n}^{(1)} + \sum_{l}^{p}\left(\frac{1}{n}c_{l,n}^{(1)} + \frac{1}{2n^2}c_{l,n}^{(2)}\right)L_{ijl}\right. \\
&\qquad \left. + \frac{1}{2n^2}\sum_{l,m}^{p} c_{l,n}^{(1)}c_{m,n}^{(1)}L_{ijlm}\right\} + o_p(n^{-1}) \\
&\quad = \sum_{j,k}^{p} J^{jk}\left(\gamma_k + \sum_{l}^{p}\frac{1}{n}c_{l,n}^{(1)}\gamma_{kl}\right) \\
&\qquad \left\{\frac{1}{n}\nu_{ij,n}^{(1)} + \sum_{l}^{p}\left(\frac{1}{n}c_{l,n}^{(1)} + \frac{1}{2n^2}c_{l,n}^{(2)}\right)\left(\frac{1}{n}\nu_{ijl,n}^{(1)} + \kappa_{ijl}\right)\right. \\
&\qquad \left. + \frac{1}{2n^2}\sum_{l,m}^{p} c_{l,n}^{(1)}c_{m,n}^{(1)}\kappa_{ijlm}\right\} + o_p(n^{-1}).
\end{aligned}
$$

Similarly,

$$
I_{i,3} = \sum_{j,k,l,m}^{p} J^{jk}J^{lm}\gamma_m\frac{1}{n^2}S_{kl,n}^{(1)}S_{ij,n}^{(1)} + o_p(n^{-1}),
$$

where $S^{(1)}_{ij,n} = \nu^{(1)}_{ij,n} + \sum_k^p c^{(1)}_{k,n}\kappa_{ijk}$ which satisfies (11), and

$$I_{i,4} = \sum_{j,k,l,m}^{p} J^{jl}J^{km}\gamma_l\gamma_m\kappa_{ijk} + o_p(1), \quad I_{i,5} = \sum_{j,k}^{p} J^{jk}\gamma_k\gamma_{ij} + o_p(1).$$

The Lemma is obtained by summarizing each term that has a superscript of (1) or (2) or double (1)'s in $I_{i,1}$, $I_{i,2}$ and $I_{i,3}$:

$$\sum_j^p A^{(1)}_{j,n}J_{ij} = \sum_j^p \left(c^{(1)}_{j,n}\gamma_{ij} + q_j\nu^{(1)}_{ij,n} + q_j\sum_k^p c^{(1)}_{k,n}\kappa_{ijk} \right),$$

and

$$\begin{aligned}\sum_j^p A^{(2)}_{j,n}J_{ij} &= \frac{1}{2}\sum_j^p c^{(2)}_{j,n}\gamma_{ij} + \frac{1}{2}\sum_{j,k}^p c^{(1)}_{j,n}c^{(1)}_{k,n}\gamma_{ijk} \\ &\quad + \sum_{j,l}^p q_j\left(c^{(1)}_{l,n}\nu^{(1)}_{ijl,n} + \frac{1}{2}c^{(2)}_{l,n}\kappa_{ijl} + \frac{1}{2}c^{(1)}_{l,n}\sum_m^p c^{(1)}_{m,n}\kappa_{ijlm} \right) \\ &\quad + \sum_{j,l,m}^p J^{jm}\gamma_{lm}c^{(1)}_{l,n}\left(\nu^{(1)}_{ij,n} + \sum_k^p c^{(1)}_{k,n}\kappa_{ijk} \right) + \sum_{j,k,l}^p J^{jk}q_l S^{(1)}_{kl,n}S^{(1)}_{ij,n},\end{aligned}$$

where $q_i = \sum_j^p J^{ij}\gamma_j$ is used. On the other hand, $I_{i,4}$ and $I_{i,5}$ derive that $\sum_j^p B_j J_{ij} = \frac{1}{2}\sum_{j,k}^p q_j q_k \kappa_{ijk} + \sum_j^p q_j\gamma_{ij}$.

It is straightforward from the above to specify the functions $\boldsymbol{A}^{(1)}(x;\boldsymbol{\theta})$ and $\boldsymbol{A}^{(2)}(x;\boldsymbol{\theta})$ satisfying (11): $\boldsymbol{A}^{(1)}(x;\boldsymbol{\theta})$ satisfies

$$\begin{aligned}&\sum_i^p A^{(1)}_i(x;\boldsymbol{\theta})J_{ij}(\boldsymbol{\theta}) \\ &= \sum_i^p \left\{ T^{(1)}_i(x;\boldsymbol{\theta})\gamma_{ij}(\boldsymbol{\theta}) + q_i(\boldsymbol{\theta})\nu^{(1)}_{ij}(x;\boldsymbol{\theta}) + q_i(\boldsymbol{\theta})\sum_k^p T^{(1)}_k(x;\boldsymbol{\theta})\kappa_{ijk}(\boldsymbol{\theta}) \right\}.\end{aligned} \tag{28}$$

Here, $\boldsymbol{A}^{(2)}(x;\boldsymbol{\theta})$ is derived similarly.

Expressions of $c^{(1)}_n$ and $c^{(2)}_n$

Analogous arguments in the proof of Lemma 1 show that

$$c^{(1)}_{i,n} = \sum_j^p J^{ij}L^{(1)}_i \quad \text{and} \quad c^{(2)}_{i,n} = \sum_{j,k,l}^p c^{(1)}_{k,n}\left(2\nu^{(1)}_{jk,n} + \sum_l^p c^{(1)}_{l,n}\kappa_{jkl} \right) J^{ij}, \tag{29}$$

where $L_i^{(1)} = nL_i(\boldsymbol{\theta}_0) = \sum_{\alpha=1}^{n} \ell_i(X_\alpha; \boldsymbol{\theta}_0)$ which satisfies (11). The relations (29) are also obtained from the compact derivatives (Konishi and Kitagawa 1996) for the M-estimator applied to MLE. By (29), we have

$$E_X\{T_i^{(2)}(X, X)\} = \sum_{j,k,l}^{p} \left[2J^{kl} E_X\{\ell_l(X)\ell_{jk}(X)\} + V_{kl}\kappa_{jkl}\right] J^{ij}, \qquad (30)$$

where $E_X\{T_k^{(1)}(X)T_l^{(1)}(X)\} = V_{kl}$ is used.

Acknowledgments We are deeply grateful to the associate editor and the editor who give us many helpful suggestions and editorial encouragements.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716–723.

DiCiccio, T. J., Efron, B. (1992). More accurate confidence intervals in exponential families. *Biometrika, 79*, 231–245.

DiCiccio, T. J., Efron, B. (1996). Bootstrap confidence intervals (with discussion). *Statistical Science, 11*, 189–228.

DiCiccio, T. J., Monti, A. C. (2001). Accurate confidence limits for scalar functions of vector M-estimands. *Biometrika, 89*, 437–451.

Fan, J., Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association, 96*, 1348–1360.

Gilbert. P. numDeriv: Accurate Numerical Derivatives. R package Version 2006.4-1. http://www.bank-banque-canada.ca/pgilbert.

Good, I. J., Gaskins, R. A. (1971). Nonparametric roughness penalties for probability densities. *Biometrika, 58*, 255–277.

Green, P. J., Silverman, B. W. (1994). *Nonparametric regression and generalized linear models, a roughness penalty approach*. London: Chapman and Hall.

Imoto, S., Konishi, S. (2003). Selection of smoothing parameters in B-spline nonparametric regression models using information criteria. *Annals of the Institute of Statistical Mathematics, 55*, 671–687.

Konishi, S., Kitagawa, G. (1996). Generalised information criteria in model selection. *Biometrika, 83*, 875–890.

Konishi, S., Kitagawa, G. (2003). Asymptotic theory for information criteria in model selection – Functional approach. *Journal of Statistical Planning and Inference, 114*, 45–61.

Konishi, S., Ando, T., Imoto, S. (2004). Bayesian information criteria and smoothing parameter selection in radial basis function networks. *Biometrika, 91*, 27–43.

Nonaka, Y., Konishi, S. (2005). Nonlinear regression modeling using regularized local likelihood method. *Annals of the Institute of Statistical Mathematics, 57*, 617–635.

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference, 90*, 227–244.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society, Series B, 36*, 111–147.

Takeuchi, K. (1976). Distribution of informational statistics and a criterion of model fitting. *Suri-Kagaku (Mathematical Sciences), 153*, 12–18 (in Japanese).

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B, 58*, 267–288.

Wand, M. P. (1999). On the optimal amount of smoothing in penalised spline regression. *Biometrika, 86*, 936–940.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrika, 50*, 1–26.