Scandinavian Journal of Statistics, Vol. 35: 496–523, 2008
doi: 10.1111/j.1467-9469.2008.00593.x

Prediction Error Estimation Under Bregman Divergence for Non-Parametric Regression and Classification

CHUNMING ZHANG

Department of Statistics, University of Wisconsin-Madison

ABSTRACT. Prediction error is critical to assess model fit and evaluate model prediction. We propose the cross-validation (CV) and approximated CV methods for estimating prediction error under the Bregman divergence (BD), which embeds nearly all of the commonly used loss functions in the regression, classification procedures and machine learning literature. The approximated CV formulas are analytically derived, which facilitate fast estimation of prediction error under BD. We then study a data-driven optimal bandwidth selector for local-likelihood estimation that minimizes the overall prediction error or equivalently the covariance penalty. It is shown that the covariance penalty and CV methods converge to the same mean-prediction-error-criterion. We also propose a lower-bound scheme for computing the local logistic regression estimates and demonstrate that the algorithm monotonically enhances the target local likelihood and converges. The idea and methods are extended to the generalized varying-coefficient models and additive models.

Key words: cross-validation, exponential family, generalized varying-coefficient model, local likelihood, loss function, prediction error

1. Introduction

Prediction error is critical to assess the performance of statistical methods and select statistical models. Different loss functions are used for computing the prediction error in different machine-learning problems. In binary classification, for example, the misclassification error rate is more suitable because the class labels are not numeric. Other important margin-based loss functions have been introduced for binary classification in the machine-learning literature (Hastie *et al.*, 2001). Hence, it is important to assess the prediction error under a broad class of loss functions.

A broad and important class of loss functions is the Bregman q-class divergence. It accounts for different types of output variables and includes the quadratic loss, the deviance loss for the exponential family of distributions, the misclassification loss and other popular loss functions in machine learning; see section 2. Once a prediction error criterion is chosen, the estimates of prediction error are needed. Desirable features include computational expediency and theoretical consistency. In the traditional non-parametric regression models, residual-based cross-validation (CV) is a useful data-driven method for automatic smoothing (Wong, 1983; Rice, 1984; Hall & Johnstone, 1992; Härdle *et al.*, 1992) and can be handily computed. With the arrival of the optimism theorem (Efron, 2004), estimating the prediction error becomes estimating covariance-penalty terms. Following Efron (2004), the covariance penalty can be estimated using model-based bootstrap procedures. A viable model-free method is the CV estimation of the covariance penalty. Both methods can be shown to be asymptotically equivalent to the first-order approximation. However, both methods are extremely computationally intensive in the context of local-likelihood estimation, particularly for large sample sizes. The challenge then arises from efficient computation of the estimated prediction error based on CV.

The computational problem is resolved via the newly developed approximate formulas for the CV covariance-penalty estimates. A key component is to establish the 'leave-one-out formulas', which offer an analytic connection between the leave-one-out estimates and their 'keep-all-in' counterparts. This technical work integrates the infinitesimal perturbation idea (Pregibon, 1981) with the Sherman–Morrison–Woodbury formulas (Golub & Van Loan, 1996, p. 50). It is a natural extension of the CV formula for least-squares regression estimates and generalized linear regression estimates (Davidson & Hinkley, 1997, p. 67), and is applicable to both parametric and non-parametric models.

The applications of estimated prediction error pervade almost every facet of statistical model selection and forecasting. To be more specific, we focus on local-likelihood estimation in varying coefficient models for response variables having distributions in the exponential family. Typical examples include fitting Bernoulli distributed binary responses, and Poisson distributed count responses, among many other non-normal outcomes. As a flexible non-parametric model-fitting technique, the local-likelihood method possesses nice sampling properties. For details, see, for example, Tibshirani & Hastie (1987), Staniswalis (1989), Severini & Staniswalis (1994) and Fan *et al.* (1995). An important issue in applications is the choice of smoothing parameter. Currently, most of the existing methods deal with Gaussian type of responses; clearly there is a lack of methodology for non-Gaussian responses. The approximate CV provides a simple and fast method for this purpose. The versatility of the choice of smoothing parameters is enhanced by an appropriate choice of the divergence measure in the q-class of loss functions.

The computational cost of the approximate CV method is further reduced via a newly introduced empirical version of CV, called ECV, which is based on an empirical construction of the 'degrees of freedom', a notion which provides useful insights into the local-likelihood modelling complexity. We propose a data-driven bandwidth selection method, based on minimizing ECV, which will be shown to be asymptotically optimal in minimizing a broad q-class of prediction error. Compared with the two-stage bandwidth selector of Fan *et al.* (1998), our proposed method has a broader domain of applications and can be more easily understood and implemented.

Some specific attentions are needed for the local logistic regression with binary responses, whose distribution belongs to an important member of the exponential family. To address the numerical instability, we propose to replace the Hessian matrix by its global lower-bound (LB) matrix, which does not involve estimating parameter vectors and therefore can easily be inverted before the start of the Newton–Raphson (NR) iteration. A similar idea of LB was used in Böhning & Lindsay (1988) for some parametric fitting. We make a conscientious effort to further develop this idea for the local logistic estimation. The resulting LB method gains a number of advantages: the LB algorithm, at each iteration, updates the gradient vector but does not recalculate the Hessian matrix; thus, is as simple and stable as the local least-squares regression estimation. The LB method ensures that each iterative estimate monotonically increases the target local likelihood. In contrast, this property is not shared by the standard NR method. Hence, the LB iteration is guaranteed to converge to the true local MLE, whereas the NR is not necessarily convergent. Moreover, we develop a new and adaptive data-driven method for bandwidth selection, which can effectively guard against under- or oversmoothing.

The paper is organized as follows. Section 2 addresses the issue of estimating prediction error. Section 3 develops computationally feasible versions of the CV estimates of the prediction error. Section 4 proposes a new bandwidth selection method for binary responses, based on the LB method and the CV estimates of the prediction error. Section 5 describes the extension to generalized varying-coefficient model. Section 6 presents simulation evaluations and section 7 analyses real data. Technical conditions and proofs are relegated to the Appendix.

2. Estimating prediction error

To begin with, we assume that the response variable Y given the vector $\mathtt{X}$ of input variables has a distribution in the exponential family, taking the form

$$f_{Y\,|\,\mathtt{X}}(y;\theta(\mathtt{x}))=\exp[\{y\theta(\mathtt{x})-b(\theta(\mathtt{x}))\}/a(\psi)+c(y,\psi)], \tag{1}$$

for some known functions $a(\cdot)$, $b(\cdot)$ and $c(\cdot,\cdot)$, where $\theta(\mathtt{x})$ is called a canonical parameter and ψ is called a dispersion parameter, respectively. It is well known that $m(\mathtt{x})\equiv E(Y\,|\,\mathtt{X}=\mathtt{x})=b'(\theta(\mathtt{x}))$ and $\sigma^2(\mathtt{x})\equiv\operatorname{var}(Y\,|\,\mathtt{X}=\mathtt{x})=a(\psi)b''(\theta(\mathtt{x}))$; see Nelder & Wedderburn (1972) and McCullagh & Nelder (1989). The canonical link is $g(\cdot)=(b')^{-1}(\cdot)$, resulting in $g(m(\mathtt{x}))=\theta(\mathtt{x})$. For simplicity of notation and exposition, we will focus only on estimating the canonical parameter $\theta(\mathtt{x})$.

2.1. Bregman divergence

The prediction error depends on the divergence measure. For non-Gaussian responses, the quadratic loss function is not always adequate. For binary classification, a reasonable choice of divergence measure is the misclassification loss, $Q(Y,\hat{m})=\mathrm{I}\{Y\neq\mathrm{I}(\hat{m}>0.5)\}$, where $\mathrm{I}(\cdot)$ is an indicator function and $\hat{m}$ is an estimator. However, this measure does not differentiate the predictions $\hat{m}=0.6$ and $\hat{m}=0.9$ when $Y=1$ or 0. In the case that $Y=1$, $\hat{m}=0.9$ gives a better prediction than $\hat{m}=0.6$. The negative Bernoulli log likelihood, $Q(Y,\hat{m})=-Y\ln(\hat{m})-(1-Y)\ln(1-\hat{m})$, captures this. Other loss functions possessing similar properties include the hinge loss function, $Q(Y,\hat{m})=\max\{1-(2Y-1)\operatorname{sign}(\hat{m}-0.5),0\}$, in the support vector machine and the exponential loss function, $Q(Y,\hat{m})=\exp\{-(Y-0.5)\ln(\hat{m}/(1-\hat{m}))\}$, popularly used in AdaBoost. These four loss functions, shown in Fig. 1, belong to the margin-based loss functions written in the form, $V(Y^*F)$, for $Y^*=2Y-1$ and some function F.

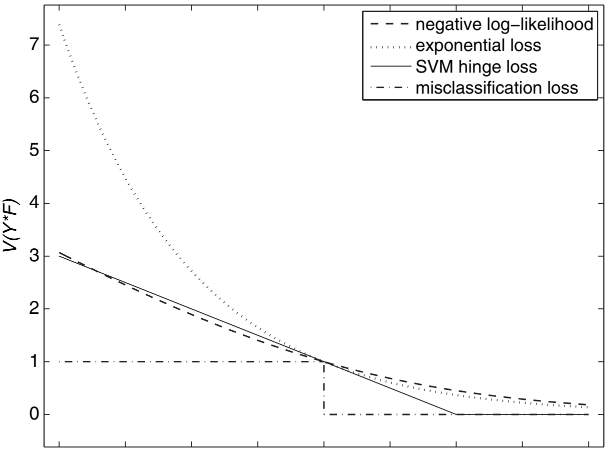

Fig. 1. Illustration of margin-based loss functions. Line types are indicated in the legend box. Each function has been re-scaled to pass through the point (0, 1).

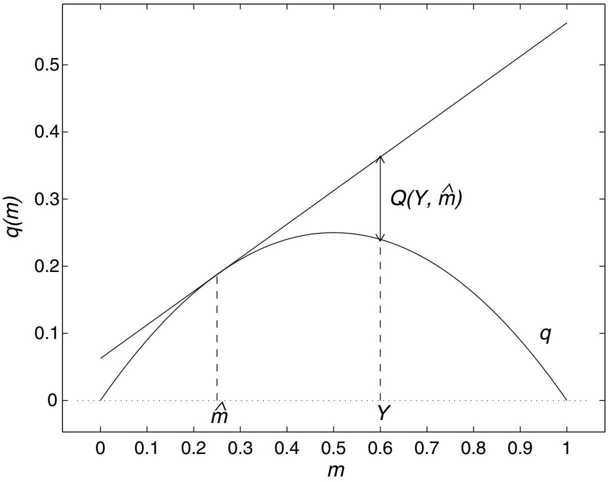

Fig. 2. Illustration of $Q(Y,\hat{m})$ as defined in (2). The concave curve is the q-function; the two dashed lines indicate locations of Y and $\hat{m}$; the solid strict line is $q(\hat{m})+q'(\hat{m})(Y-\hat{m})$; the length of the vertical line with arrows at each end is $Q(Y,\hat{m})$.

To address the versatility of loss functions, we appeal to a device introduced by Bregman (1967) and Efron (1986). For a concave function $q(\cdot)$, define a q-class of error measures Q as

$$Q(Y,\hat{m})=q(\hat{m})+q'(\hat{m})(Y-\hat{m})-q(Y). \tag{2}$$

A graphical illustration of Q associated with q is displayed in Fig. 2. Due to the concavity of q, Q is non-negative. However, since $Q(\cdot,\cdot)$ is not generally symmetric in its arguments, Q is not a 'metric' or 'distance' in the strict sense. Hence, we call Q the Bregman 'divergence' (BD).

It is easy to see that, with the flexible choice of q, the BD is suitable for a broad class of error measures. Below we present some notable examples of the Q-loss constructed from the q-function. A function $q_1(m)=am-m^2$ for some constant a yields the quadratic loss $Q_1(Y,\hat{m})=(Y-\hat{m})^2$. For the exponential family (1), the function $q_2(m)=2\{b(\theta)-m\theta\}$ with $b'(\theta)=m$ results in the deviance loss,

$$Q_2(Y,\hat{m})=2\{Y(\tilde{\theta}-\hat{\theta})-b(\tilde{\theta})+b(\hat{\theta})\}, \tag{3}$$

where $b'(\tilde{\theta})=Y$ and $b'(\hat{\theta})=\hat{m}$. For a binary response variable Y, $q(m)=\min\{m,1-m\}$ gives the misclassification loss; $q(m)=2\min\{m,1-m\}$ results in the hinge loss; $q_3(m)=2\{m(1-m)\}^{1/2}$ yields the exponential loss,

$$Q_3(Y,\hat{m})=\exp\{-(Y-0.5)\ln(\hat{m}(1-\hat{m}))\}. \tag{4}$$

2.2. *Prediction error under Bregman divergence*

Let $m_i=m(\mathtt{X}_i)$ and $\hat{m}_i$ be its estimate based on independent observations $\{(\mathtt{X}_i, Y_i)\}_{i=1}^n$. Set $\text{err}_i=Q(Y_i,\hat{m}_i)$ and $\text{Err}_i=E_o\{Q(Y_i^o,\hat{m}_i)\}$, where Y_i^o is an independent copy of Y_i and is independent of $(Y_1,\ldots,Y_n)$, and E_o refers to the expectation with respect to the probability law of Y_i^o. Note that the conditional prediction error, defined by Err_i, is not observable, whereas the apparent error, err_i, is observable. As noted in Tibshirani (1996), directly estimating the conditional prediction error is very difficult. Alternatively, estimating Err_i is equivalent to estimating the *optimism* $O_i=\text{Err}_i-\text{err}_i$.

Efron (2004) derives the optimism theorem to represent the expected optimism as the covariance penalty, namely, $E(O_i) = 2\ \mathrm{cov}(\hat{\lambda}_i, Y_i)$, where $\hat{\lambda}_i = -q'(\hat{m}_i)/2$. As a result, the predictive error can be estimated by

$$\widehat{\mathrm{Err}}_i = \mathrm{err}_i + 2\ \widehat{\mathrm{cov}}_i, \tag{5}$$

where $\widehat{\mathrm{cov}}_i$ is an estimator of the covariance penalty, $\mathrm{cov}(\hat{\lambda}_i, Y_i)$. This is an insightful generalization of AIC. Henceforth, the total prediction error $\mathrm{Err} = \sum_{i=1}^n \mathrm{Err}_i$ can be estimated by $\widehat{\mathrm{Err}} = \sum_{i=1}^n \widehat{\mathrm{Err}}_i$.

2.3. Estimation of covariance penalty

In non-parametric estimation, we write $\hat{m}_{i,h}$, $\hat{\lambda}_{i,h}$, $\widehat{\mathrm{Err}}_i(h)$ and $\widehat{\mathrm{Err}}(h)$ to stress their dependence on a smoothing parameter h. The CV estimation of Err_i is $Q(Y_i, \hat{m}_{i,h}^{-i})$, where the superscript $-i$ indicates deletion of the ith data point $(\mathtt{X}_i, Y_i)$ in the fitting process. This yields the CV estimate of the total prediction error by

$$\widehat{\mathrm{Err}}^{\mathrm{CV}}(h) = \sum_{i=1}^n Q(Y_i, \hat{m}_{i,h}^{-i}). \tag{6}$$

Naive computation of $\{\hat{m}_{i,h}^{-i}\}_{i=1}^n$ is intensive. Section 3 will devise strategies by which actual computations of the leave-one-out estimates are not needed. A distinguished feature is that our method is widely applicable to virtually all regression and classification problems. The approximated CV is particularly attractive to a wide array of large and complex problems in which a quick and crude selection of the model parameter is needed.

By comparing (6) and (5), we see that the covariance penalty is estimated by

$$\sum_{i=1}^n \{Q(Y_i, \hat{m}_{i,h}^{-i}) - Q(Y_i, \hat{m}_{i,h})\}.$$

This can be linked with the jackknife method for estimating the covariance penalty. Hence, it is expected that the CV method is asymptotically equivalent to a bootstrap method.

2.4. Asymptotic prediction error

To gain insight into $\widehat{\mathrm{Err}}(h)$, we appeal to asymptotic theory. Simple algebra shows that

$$\mathrm{Err}_i(h) = E_o\{Q(Y_i^o, \hat{m}_{i,h})\} = Q(m_i, \hat{m}_{i,h}) + E\{Q(Y_i, m_i)\}.$$

By Taylor's expansion and (2), $Q(m_i, \hat{m}_{i,h}) \doteq -(\hat{m}_{i,h} - m_i)^2 q''(\hat{m}_{i,h})/2$. Hence,

$$\mathrm{Err}_i(h) \doteq -(\hat{m}_{i,h} - m_i)^2 q''(\hat{m}_{i,h})/2 + E\{Q(Y_i, m_i)\}.$$

Note that the last term does not depend on h and hence minimizing $\widehat{\mathrm{Err}}(h)$ is asymptotically equivalent to the minimizing mean-prediction-error-criterion,

$$\mathrm{MPEC}(h) = -2^{-1} \int E[\{\hat{m}_h(x) - m(x)\}^2 \mid \mathcal{X}] q''(m(x)) f_X(x)\, \mathrm{d}x, \tag{7}$$

with $\mathcal{X} = (\mathtt{X}_1, \ldots, \mathtt{X}_n)$ and $f_X(x)$ being the probability density of $\mathtt{X}$. This criterion differs from the mean-integrated-squared-error criterion defined by

$$\mathrm{MISE}(h) = \int E[\{\hat{m}_h(x) - m(x)\}^2 \mid \mathcal{X}]\{b''(\theta(x))\}^{-2} f_X(x)\, \mathrm{d}x, \tag{8}$$

recalling that $\hat{\theta}(x) - \theta(x) \doteq \{b''(\theta(x))\}^{-1}\{\hat{m}_h(x) - m(x)\}$.

Expression (7) reveals that asymptotically, different loss functions automatically introduce different weighting schemes in (7). This provides a useful insight into various error measures used in practice. The weighting schemes vary substantially over the choices of q. In particular, for the q_1-function yielding the quadratic loss in section 2.1, the q_2-function producing the deviance-loss and the q_3-function inducing the exponential-loss for the binary responses, they deliver, respectively,

$$\mathrm{MPEC}_1(h)=\int E[\{\hat{m}_h(x)-m(x)\}^2\mid \mathcal{X}]f_X(x)\,\mathrm{d}x,$$

$$\mathrm{MPEC}_2(h)=\int E[\{\hat{m}_h(x)-m(x)\}^2\mid \mathcal{X}]\{b''(\theta(x))\}^{-1}f_X(x)\,\mathrm{d}x,$$

$$\mathrm{MPEC}_3(h)=\int E[\{\hat{m}_h(x)-m(x)\}^2\mid \mathcal{X}]\{4[m(x)\{1-m(x)\}]^{3/2}\}^{-1}f_X(x)\,\mathrm{d}x.$$

3. Approximate cross-validation

This section aims at deriving the approximate and empirical versions of (6) for the local maximum-likelihood (ML) estimator. We focus on the univariate case in this section. The results will be extended to the generalized varying coefficient models in section 5 incorporating multivariate covariates.

Assume that the function $\theta(\cdot)$ has a $(p+1)$th continuous derivative at a point x. For X_j close to x, a Taylor expansion implies that $\theta(X_j)\doteq \mathbf{x}_j(x)^{\mathrm{T}}\boldsymbol{\beta}(x)$, in which $\mathbf{x}_j(x)=(1,(X_j-x),\ldots,(X_j-x)^p)^{\mathrm{T}}$ and $\boldsymbol{\beta}(x)=(\beta_0(x),\ldots,\beta_p(x))^{\mathrm{T}}$. Based on independent observations, $\boldsymbol{\beta}(x)$ can be estimated by maximizing the local log-likelihood,

$$\ell(\boldsymbol{\beta};x)\equiv\sum_{j=1}^{n} l(\mathbf{x}_j(x)^{\mathrm{T}}\boldsymbol{\beta};Y_j)K_h(X_j-x), \tag{9}$$

in which $l(\cdot;y)=\ln\{f_{Y\mid X}(y;\cdot)\}$ denotes the conditional log-likelihood function, $K_h(\cdot)=K(\cdot/h)/h$ for a kernel function K, and h is a bandwidth. Let $\hat{\boldsymbol{\beta}}(x)=(\hat\beta_0(x),\ldots,\hat\beta_p(x))^{\mathrm{T}}$ be the local ML estimator. Then, the local MLEs of $\theta(x)$ and $m(x)$ are given by $\hat\theta(x)=\hat\beta_0(x)$ and $\hat m(x)=b'(\hat\theta(x))$, respectively. A similar estimation procedure, based on the $n-1$ observations excluding (X_i, Y_i), leads to the local log-likelihood function $\ell^{-i}(\boldsymbol{\beta};x)$, and the corresponding local MLEs, $\hat{\boldsymbol{\beta}}^{-i}(x)$, $\hat\theta^{-i}(x)$ and $\hat m^{-i}(x)$, respectively. Note that K-fold ($K<n$) CV is comparatively less frequently used in non-parametric smoothing for selecting h.

3.1. Leave-one-out formulas

Let $\mathbf{X}(x)=(\mathbf{x}_1(x),\ldots,\mathbf{x}_n(x))^{\mathrm{T}}$, $\mathbf{W}(x;\boldsymbol{\beta})=\mathrm{diag}\{K_h(X_j-x)b''(\mathbf{x}_j(x)^{\mathrm{T}}\boldsymbol{\beta})\}$, and $S_n(x;\boldsymbol{\beta})=\mathbf{X}(x)^{\mathrm{T}}\mathbf{W}(x;\boldsymbol{\beta})\mathbf{X}(x)$. Define

$$\mathcal{H}(x;\boldsymbol{\beta})=\{\mathbf{W}(x;\boldsymbol{\beta})\}^{1/2}\mathbf{X}(x)\{S_n(x;\boldsymbol{\beta})\}^{-1}\mathbf{X}(x)^{\mathrm{T}}\{\mathbf{W}(x;\boldsymbol{\beta})\}^{1/2}.$$

This projection matrix is an extension of the hat matrix in multiple linear regression and will be useful for computing the leave-one-out estimator. Let $\mathcal{H}_{ii}(x;\boldsymbol{\beta})$ be its ith diagonal element and set $H_i=\mathcal{H}_{ii}(X_i;\hat{\boldsymbol{\beta}}(X_i))$. Below we summarize our main result.

Proposition 1
Assume condition A2 in the Appendix. Then for any $h>0$ and $i=1,\ldots,n$,

$$\hat{\boldsymbol{\beta}}^{-i}(x)-\hat{\boldsymbol{\beta}}(x)\doteq-\frac{\{S_n(x;\hat{\boldsymbol{\beta}}(x))\}^{-1}\mathbf{x}_i(x)K_h(X_i-x)\{Y_i-b'(\mathbf{x}_i(x)^{\mathrm{T}}\hat{\boldsymbol{\beta}}(x))\}}{1-\mathcal{H}_{ii}(x;\hat{\boldsymbol{\beta}}(x))}, \tag{10}$$

$$\hat{\theta}_i^{-i} - \hat{\theta}_i \doteq -\frac{H_i}{1-H_i} \cdot \frac{Y_i - \hat{m}_i}{b''(\hat{\theta}_i)}, \tag{11}$$

$$\hat{m}_i^{-i} - \hat{m}_i \doteq -\frac{H_i}{1-H_i}(Y_i - \hat{m}_i). \tag{12}$$

Note that the approximation becomes exact when the loss function is the quadratic loss; as shown in Zhang (2003), $\hat{m}_i^{-i} = \hat{m}_i - \{H_i/(1-H_i)\}(Y_i - \hat{m}_i)$. In addition, for $h \to \infty$, (10) coincides with the counterpart of generalized linear models (Davidson & Hinkley, 1997, p. 67). Furthermore, the results can easily be extended to the estimator that minimizes the local Bregman divergence, replacing $l(\mathbf{x}_j(x)^{\mathrm{T}}\boldsymbol{\beta}; Y_j)$ in (9) by $Q(Y_j, g^{-1}(\mathbf{x}_j(x)^{\mathrm{T}}\boldsymbol{\beta}))$. Using proposition 1, we can derive a simplified formula for computing the CV estimate of the overall prediction error.

Proposition 2

Assume conditions A1 and A2 in the Appendix. Then for any $h > 0$,

$$\widehat{\mathrm{Err}}^{\mathrm{CV}} \doteq \sum_{i=1}^{n} \left[Q(Y_i, \hat{m}_i) + 2^{-1} q''(\hat{m}_i)(Y_i - \hat{m}_i)^2 \{1 - 1/(1-H_i)^2\}\right]. \tag{13}$$

Proposition 2 gives an approximation formula, which avoids computing ‘leaving-one-out’ estimates, for all q-class of loss functions. In particular, for the function q_1, we have

$$\sum_{i=1}^{n} (Y_i - \hat{m}_i^{-i})^2 = \sum_{i=1}^{n} (Y_i - \hat{m}_i)^2/(1-H_i)^2.$$

For this particular loss function, the approximation is actually exact. For the function q_2 leading to the deviance loss Q_2 defined in (3), we have

$$\sum_{i=1}^{n} Q_2(Y_i, \hat{m}_i^{-i}) \doteq \sum_{i=1}^{n} \left[Q_2(Y_i, \hat{m}_i) - \frac{(Y_i - \hat{m}_i)^2}{b''(\hat{\theta}_i)} \{1 - 1/(1-H_i)^2\}\right].$$

For the exponential loss defined in (4) for binary classification, we have

$$\sum_{i=1}^{n} Q_3(Y_i, \hat{m}_i^{-i}) \doteq \sum_{i=1}^{n} \left[Q_3(Y_i, \hat{m}_i) - \frac{(Y_i - \hat{m}_i)^2}{4\{\hat{m}_i(1-\hat{m}_i)\}^{3/2}} \{1 - 1/(1-H_i)^2\}\right].$$

Remark 1. An alternative approach to smoothing in likelihood-based models is smoothing splines. For non-Gaussian responses with a univariate predictor, Xiang & Wahba (1996) selected the penalization parameter to minimize the comparative Kullback–Leibler loss, in which an approximate ‘leave-one-out’ formula is derived through a series of first-order Taylor expansions. The argument presented here for ‘leave-one-out’ formulas is somewhat more direct and can conveniently be extended to smoothing splines and other smoothing techniques dealing with multivariate predictors under the broader q-class of loss functions.

3.2. Two theoretical issues

Two theoretical issues are particularly interesting. The first one concerns the asymptotic convergence of $\hat{h}_{\mathrm{ACV}}$, the minimizer of the right-hand side of (13). Following a suitable modification to the result of Altman & MacGibbon (1998), the ratio $\hat{h}_{\mathrm{ACV}}/h_{\mathrm{AMPEC}}$ converges in probability to 1, where h_{AMPEC} is the minimizer of the asymptotic form of MPEC(h) defined in (7).

Table 1. *Comparison of the asymptotic optimal bandwidths $h_{\text{AMPEC}}(q_2)$ from (14), h_{AMISE} from (15) and $h_{\text{AMPEC}}(q_3)$ (for Bernoulli Responses), using $p=1$ and the Epanechnikov kernel*

	$h_{\text{AMPEC}}(q_2)$		h_{AMISE}		$h_{\text{AMPEC}}(q_3)$
Example	Poisson	Bernoulli	Poisson	Bernoulli	Bernoulli
1	0.070	0.106	0.079	0.108	0.107
2	0.089	0.151	0.099	0.146	0.148
3	0.127	0.184	0.136	0.188	0.186

The explicit expression of h_{AMPEC}, associated with the q-class of error measures, can be obtained by the delta method. Setting $-2^{-1}q''(m(x))\{b''(\theta(x))\}^2$ to be the weight function, h_{AMPEC} (for odd degrees p of local polynomial fitting) can be derived from Fan *et al.* (1995, p. 147):

$$h_{\text{AMPEC}}(q)=C_p(K)\left[\frac{a(\psi)\int b''(\theta(x))q''(m(x))\,\mathrm{d}x}{n\int\{\theta^{(p+1)}(x)\}^2\{b''(\theta(x))\}^2q''(m(x))f_X(x)\,\mathrm{d}x}\right]^{1/(2p+3)},$$

where $C_p(K)$ is a constant depending only on the degree and kernel of the local regression. In particular, for the q_2-function which gives the deviance loss, we have

$$h_{\text{AMPEC}}(q_2)=C_p(K)\left[\frac{a(\psi)\,|\Omega_X|}{\int\{\theta^{(p+1)}(x)\}^2b''(\theta(x))f_X(x)\,\mathrm{d}x}\right]^{1/(2p+3)}n^{-1/(2p+3)}, \tag{14}$$

where $|\Omega_X|$ measures the length of the support of f_X. Apparently, this asymptotically optimal bandwidth differs from the asymptotically optimal bandwidth,

$$h_{\text{AMISE}}=C_p(K)\left[\frac{a(\psi)\int\{b''(\theta(x))\}^{-1}\mathrm{d}x}{\int\{\theta^{(p+1)}(x)\}^2f_X(x)\,\mathrm{d}x}\right]^{1/(2p+3)}n^{-1/(2p+3)}, \tag{15}$$

determined by minimizing the asymptotic MISE(h) of $\hat{\theta}$ defined in (8), with an exception of the Gaussian family.

The second issue concerns how far away $h_{\text{AMPEC}}(q_2)$ departs from h_{AMISE}. For Poisson and Bernoulli response variables, Table 1 shows that the differences between $h_{\text{AMPEC}}(q_2)$ and h_{AMISE} are noticeable for the examples in section 6.1. To gain further insights, we will need the following definition.

Definition 1

Two functions F and G are called 'similarly ordered' if $\{F(s)-F(t)\}\{G(s)-G(t)\}\geq 0$ for all s in the domain of F and all t in the domain of G, and 'oppositely ordered' if the inequality is reversed.

The following theorem characterizes the relation between $h_{\text{AMPEC}}(q_2)$ and h_{AMISE}.

Proposition 3

Define $F(x)=\{\theta^{(p+1)}(x)\}^2b''(\theta(x))f_X(x)$ and $G(x)=\{b''(\theta(x))\}^{-1}$. Assume that p is odd.

(a) If F and G are oppositely ordered, then $h_{\text{AMPEC}}(q_2)\leq h_{\text{AMISE}}$. If F and G are similarly ordered, then $h_{\text{AMPEC}}(q_2)\geq h_{\text{AMISE}}$.

(b) Assume that $b''(\theta(x))$ is bounded away from 0 and ∞. Write $m_{b''}=\min_{x\in\Omega_X}b''(\theta(x))$ and $M_{b''}=\max_{x\in\Omega_X}b''(\theta(x))$. If $\theta(x)$ is a polynomial function of degree $p+1$, and f_X is a uniform density on Ω_X, then

$$\left\{\frac{4m_{b''}M_{b''}}{(m_{b''}+M_{b''})^2}\right\}^{1/(2p+3)} \leq \frac{h_{\text{AMPEC}}(q_2)}{h_{\text{AMISE}}} \leq 1,$$

in which the equalities are satisfied if and only if the exponential family is Gaussian.

3.3. Empirical cross-validation

The approximate CV criterion (13) can be further simplified. To this end, we first approximate the 'degrees of freedom' $\sum_{i=1}^{n} H_i$ (Hastie & Tibshirani, 1990). To facilitate presentation, we now define the 'equivalent kernel' $\mathcal{K}(t)$ induced by the local-polynomial fitting as the first element of the vector $S^{-1}(1, t, \ldots, t^p)^{\mathrm{T}} K(t)$, in which the matrix $S = (\mu_{i+j-2})_{1 \leq i,j \leq p+1}$ with $\mu_k = \int t^k K(t)\,\mathrm{d}t$ (see Ruppert & Wand, 1994).

Proposition 4

Assume conditions A and B in the Appendix. Then

$$\sum_{i=1}^{n} H_i = \mathcal{K}(0)\left(|\Omega_X|/h\right)\{1 + o_P(1)\},$$

where $|\Omega_X|$ denotes the length of the support of the random variable X.

Proposition 4 shows that the degrees of freedom is asymptotically independent of the design density and the conditional density. It approximates the notion of model complexity in non-parametric fitting.

Proposition 4 does not specify the constant term. To use the asymptotic formula for finite samples, we need some bias corrections. Note that when $h \to \infty$, the local polynomial fitting becomes a global polynomial fitting. Hence, its degrees of freedom should be $p+1$. This leads us to propose the following empirical formula:

$$\sum_{i=1}^{n} H_i \doteq (p+1-a) + \{\mathcal{C}n/(n-1)\}\,\mathcal{K}(0)\,|\Omega_X|/h. \tag{16}$$

In the Gaussian family, Zhang (2003) used simulations to determine the choices a and $\mathcal{C}$ (see Table 2), which uses the Epanechnikov kernel function, $K(t) = 0.75(1-t^2)_+$. Interestingly, our simulation studies in section 6 demonstrate that these choices also work well for Poisson responses. However, for Bernoulli responses, we find that for $p=1$, slightly different choices given by $a = 0.7$ and $\mathcal{C} = 1.09$ provide better approximations.

We propose the empirical version of the estimated total prediction error by replacing H_i in (13) with their empirical average,

$$\bar{H}_E = (p+1-a)/n + \{\mathcal{C}/(n-1)\}\,\mathcal{K}(0)\,|\Omega_X|/h,$$

leading to the empirical cross-validation (ECV) criterion,

Table 2. *Choices of a and $\mathcal{C}$, in the empirical formulas (16) and (28), for the pth degree local polynomial regression for Gaussian responses*

Design type	p	a	$\mathcal{C}$	Design type	p	a	$\mathcal{C}$
Fixed	0	0.55	1	Random	0	0.30	0.99
	1	0.55	1		1	0.70	1.03
	2	1.55	1		2	1.30	0.99
	3	1.55	1		3	1.70	1.03

$$\widehat{\text{Err}}^{\text{ECV}}(h)=\sum_{i=1}^{n}\left[Q(Y_i,\hat{m}_i)+2^{-1}q''(\hat{m}_i)(Y_i-\hat{m}_i)^2\left\{1-1/(1-\bar{H}_E)^2\right\}\right]. \tag{17}$$

This avoids calculating the smoother matrix $\mathcal{H}$. Yet, it turns out to work reasonably well in practice. A data-driven optimal bandwidth selector, $\hat{h}_{\text{ECV}}$, can be obtained by minimizing (17).

4. Non-parametric logistic regression

Non-parametric logistic regression plays a prominent role in classification and regression analysis. Yet, distinctive challenges arise from the local MLE and bandwidth selection. When the responses in a local neighborhood are entirely zeros or entirely ones (or nearly so), the local MLE does not exist. Müller & Schmitt (1988, p. 751) reported that the local-likelihood method suffers from a substantial proportion of 'incalculable estimates'. Fan & Chen (1999) proposed to add a ridge parameter to attenuate the problem. The numerical instability problem still exists as the ridge parameter can be very close to zero. A numerically viable solution is the lower bound method, which we now introduce.

4.1. Lower bound method for local MLE

The lower-bound method is very simple. For optimizing a concave function $\mathcal{L}$, the LB method replaces the Hessian matrix $\nabla^2\mathcal{L}(\boldsymbol{\beta})$ in the Newton–Raphson algorithm by a negative definite matrix $\mathbf{B}$, such that $\nabla^2\mathcal{L}(\boldsymbol{\beta})\geq\mathbf{B}$, for all $\boldsymbol{\beta}$. Lemma 1, shown in Böhning (1999, p. 14), indicates that the Newton–Raphson estimate, with the Hessian matrix replaced by the surrogate $\mathbf{B}$, can always enhance the target function $\mathcal{L}$.

Lemma 1

Starting from any $\boldsymbol{\beta}_0$, the LB iterative estimate, defined by $\boldsymbol{\beta}_{\text{LB}}=\boldsymbol{\beta}_0-\boldsymbol{B}^{-1}\nabla\mathcal{L}(\boldsymbol{\beta}_0)$, satisfies $\mathcal{L}(\boldsymbol{\beta}_{\text{LB}})-\mathcal{L}(\boldsymbol{\beta}_0)\geq-2^{-1}\nabla\mathcal{L}(\boldsymbol{\beta}_0)^{\text{T}}\boldsymbol{B}^{-1}\nabla\mathcal{L}(\boldsymbol{\beta}_0)\geq0.$

For local logistic regression, $\nabla^2\ell(\boldsymbol{\beta};x)=-\mathbf{X}(x)^{\text{T}}\mathbf{W}(x;\boldsymbol{\beta})\mathbf{X}(x)$. Since $\mathbf{0}\leq\mathbf{W}(x;\boldsymbol{\beta})\leq4^{-1}\mathbf{K}(x)$, where $\mathbf{K}(x)=\text{diag}\{K_h(X_j-x)\}$, the Hessian matrix $\nabla^2\ell(\boldsymbol{\beta};x)$ indeed has a lower bound, $\mathbf{B}(x)=-4^{-1}\mathbf{X}(x)^{\text{T}}\mathbf{K}(x)\mathbf{X}(x)$, and the LB-adjusted Newton–Raphson iteration for computing $\hat{\boldsymbol{\beta}}(x)$ becomes

$$\boldsymbol{\beta}_L=\boldsymbol{\beta}_{L-1}-\{\mathbf{B}(x)\}^{-1}\mathbf{X}(x)^{\text{T}}\mathbf{K}(x)\mathbf{r}(x;\boldsymbol{\beta}_{L-1}),\quad L=1,2,\ldots, \tag{18}$$

where $\mathbf{r}(x;\boldsymbol{\beta})=(r_1(x;\boldsymbol{\beta}),\ldots,r_n(x;\boldsymbol{\beta}))^{\text{T}}$ with $r_j(x;\boldsymbol{\beta})=Y_j-1/[1+\exp\{-\mathbf{x}_j(x)^{\text{T}}\boldsymbol{\beta}\}]$.

The LB method offers a number of advantages to compute $\hat{\boldsymbol{\beta}}(x)$. First, the corresponding LB matrix $\mathbf{B}(x)$ is free of the parameter vector $\boldsymbol{\beta}$, and thus can be computed in advance of the NR iteration. This in turn reduces the computational cost. Second, the LB matrix is stable, as it is the same matrix used in the least-squares local-polynomial regression estimates and does not depend on estimated local parameters. Third, since the local-likelihood function $\ell(\boldsymbol{\beta};x)$ is concave, the LB iteration is guaranteed to increase $\ell(\boldsymbol{\beta};x)$ at each step and converge to its global maximum $\hat{\boldsymbol{\beta}}(x)$.

4.2. A hybrid bandwidth selection method

For binary responses, our simulation studies show that the bandwidth choice minimizing (13) or its empirical version (17) tends to produce over-smoothed estimates. Such a problem was also encountered in Aragaki & Altman (1997) and Fan *et al.* (1998, Table 1). Because of the importance of binary responses in non-parametric regression and classification, a new bandwidth selector that specifically accommodates binary responses is needed.

We first employ the LB scheme (18) to derive a new one-step estimate of $\hat{\boldsymbol{\beta}}^{-i}(x)$, starting from $\hat{\boldsymbol{\beta}}(x)$. Define $S_n(x)=\mathbf{X}(x)^{\mathrm{T}}\mathbf{K}(x)\mathbf{X}(x)$ and $\mathcal{S}_i=\boldsymbol{e}_1^{\mathrm{T}}\{S_n(X_i)\}^{-1}\boldsymbol{e}_1K_h(0)$, where $\boldsymbol{e}_1=(1,0,\ldots,0)^{\mathrm{T}}$. The resulting leave-one-out formulas and the CV estimates of the total prediction error are displayed in proposition 5.

Proposition 5

Assume conditions A1 and A2 in the Appendix. Then for local-likelihood MLE in the Bernoulli family, for any $h>0$ and $i=1,\ldots,n$,

$$\hat{\boldsymbol{\beta}}^{-i}(x)-\hat{\boldsymbol{\beta}}(x)\doteq-\frac{4\{S_n(x)\}^{-1}\mathbf{x}_i(x)K_h(X_i-x)\{Y_i-b'(\mathbf{x}_i(x)^{\mathrm{T}}\hat{\boldsymbol{\beta}}(x))\}}{1-K_h(X_i-x)\mathbf{x}_i(x)^{\mathrm{T}}\{S_n(x)\}^{-1}\mathbf{x}_i(x)},\tag{19}$$

$$\hat{\theta}_i^{-i}-\hat{\theta}_i\doteq-4\{\mathcal{S}_i/(1-\mathcal{S}_i)\}(Y_i-\hat{m}_i),\tag{20}$$

$$\hat{m}_i^{-i}-\hat{m}_i\doteq-4b''(\hat{\theta}_i)\{\mathcal{S}_i/(1-\mathcal{S}_i)\}(Y_i-\hat{m}_i),\tag{21}$$

$$\widehat{\mathrm{Err}}^{\mathrm{CV}}\doteq\sum_{i=1}^{n}\left[Q(Y_i,\hat{m}_i)+2^{-1}q''(\hat{m}_i)(Y_i-\hat{m}_i)^2\left[1-\left\{1+\frac{4b''(\hat{\theta}_i)\mathcal{S}_i}{1-\mathcal{S}_i}\right\}^2\right]\right].\tag{22}$$

Using a bandwidth selector that minimizes (22) tends to under-smooth the binary responses. To better appreciate this, note that the second term in (22) is approximately

$$-q''(\hat{m}_i)(Y_i-\hat{m}_i)^2\{4b''(\hat{\theta}_i)\}\mathcal{S}_i,\tag{23}$$

and the second term in (13) can be approximated as

$$-q''(\hat{m}_i)(Y_i-\hat{m}_i)^2H_i.\tag{24}$$

As demonstrated in lemma 3 in the Appendix, $\mathcal{S}_i$ decreases with h and $H_i\doteq\mathcal{S}_i$. Since $0\le 4b''(\hat{\theta}_i)\le 1$ for the Bernoulli family, (23) down weighs the effects of model complexity, resulting in a smaller bandwidth.

The above discussion leads us to define a hybrid version of $\widehat{\mathrm{Err}}^{\mathrm{CV}}$ as

$$\sum_{i=1}^{n}\left[Q(Y_i,\hat{m}_i)+2^{-1}q''(\hat{m}_i)(Y_i-\hat{m}_i)^2\left[1-\left\{1+\frac{2b''(\hat{\theta}_i)\mathcal{S}_i}{1-\mathcal{S}_i}+\frac{2^{-1}H_i}{1-H_i}\right\}^2\right]\right],\tag{25}$$

which averages terms in (23) and (24) to mitigate the oversmoothing problem of criterion (13). This new criterion has some desirable properties: $2b''(\hat{\theta}_i)\mathcal{S}_i/(1-\mathcal{S}_i)+2^{-1}H_i/(1-H_i)$ is bounded below by $2^{-1}H_i/(1-H_i)$; thus, guarding against undersmoothing, and is bounded above by $\{\mathcal{S}_i/(1-\mathcal{S}_i)+H_i/(1-H_i)\}/2$, thus diminishing the influence of oversmoothing. An empirical CV criterion is to replace $\mathcal{S}_i$ and H_i in (25) by their empirical averages, which are (16) divided by n. A hybrid bandwidth selector for binary responses can be obtained by minimizing this ECV.

5. Extension to generalized varying-coefficient model

This section extends the techniques of sections 3 and 4 to a useful class of multi-predictor models. The major results are presented in propositions 6–8.

Consider multivariate predictor variables, containing a scalar U and a vector $\mathtt{X}=(X_1,\ldots,X_d)^{\mathrm{T}}$. For the response variable Y having a distribution in the exponential-family, define by $m(u,\mathtt{x})=E(Y\,|\,U=u,\mathtt{X}=\mathtt{x})$ the conditional mean regression function, where $\mathtt{x}=(x_1,\dots,x_d)^{\mathrm{T}}$. The generalized varying-coefficient model assumes that the $(d+1)$-variate canonical parameter function $\theta(u,\mathtt{x})=g(m(u,\mathtt{x}))$, with the canonical link g, takes the form

$$g(m(u,\mathtt{x}))=\theta(u,\mathtt{x})=\sum_{k=1}^{d}a_k(u)x_k=\mathtt{x}^{\mathrm{T}}\boldsymbol{A}(u), \tag{26}$$

for a vector $\boldsymbol{A}(u)=(a_1(u),\dots,a_d(u))^{\mathrm{T}}$ of unknown smooth coefficient functions.

We first describe the local-likelihood estimation of $\boldsymbol{A}(u)$, based on the independent observations $\{(U_j,\mathtt{X}_j,Y_j)_{j=1}^n\}$. Assume that the $a_k(\cdot)$'s are $(p+1)$ times continuously differentiable at a fitting point u. Put $\boldsymbol{A}^{(\ell)}(u)=(a_1^{(\ell)}(u),\dots,a_d^{(\ell)}(u))^{\mathrm{T}}$. Denote by $\boldsymbol{\beta}(u)=(\boldsymbol{A}(u)^{\mathrm{T}},\dots,\boldsymbol{A}^{(p)}(u)^{\mathrm{T}}/p!)^{\mathrm{T}}$ the $d(p+1)$ by 1 vector of coefficient functions along with their derivatives, $\mathbf{u}_j(u)=(1,(U_j-u),\dots,(U_j-u)^p)^{\mathrm{T}}$, and $\mathbf{I}_d$ a $d\times d$ identity matrix. For observed covariates U_j close to the point u,

$$\boldsymbol{A}(U_j)\doteq\boldsymbol{A}(u)+(U_j-u)\boldsymbol{A}^{(1)}(u)+\cdots+(U_j-u)^p\boldsymbol{A}^{(p)}(u)/p!=\{\mathbf{u}_j(u)\otimes\mathbf{I}_d\}^{\mathrm{T}}\boldsymbol{\beta}(u),$$

in which the symbol $\otimes$ denotes the Kronecker product, and thus from (26), $\theta(U_j,\mathtt{X}_j)\doteq\{\mathbf{u}_j(u)\otimes\mathtt{X}_j\}^{\mathrm{T}}\boldsymbol{\beta}(u)$. The local-likelihood MLE $\hat{\boldsymbol{\beta}}(u)$ maximizes the local log-likelihood function,

$$\ell(\boldsymbol{\beta};u)=\sum_{j=1}^{n}l(\{\mathbf{u}_j(u)\otimes\mathtt{X}_j\}^{\mathrm{T}}\boldsymbol{\beta};\,Y_j)K_h(U_j-u).$$

The first d entries of $\hat{\boldsymbol{\beta}}(u)$ supply the local MLEs $\hat{\boldsymbol{A}}(u)$ of $\boldsymbol{A}(u)$, and the local MLEs of $\theta(u,\mathtt{x})$ and $m(u,\mathtt{x})$ are given by $\hat{\theta}(u,\mathtt{x})=\mathtt{x}^{\mathrm{T}}\hat{\boldsymbol{A}}(u)$ and $\hat{m}(u,\mathtt{x})=b'(\hat{\theta}(u,\mathtt{x}))$, respectively. A similar estimation procedure, applied to $n-1$ observations excluding $(U_i,\mathtt{X}_i,Y_i)$, leads to the local log-likelihood function, $\ell^{-i}(\boldsymbol{\beta};u)$, and the corresponding local MLEs, $\hat{\boldsymbol{\beta}}^{-i}(u)$, $\hat{\theta}^{-i}(u,\mathtt{x})$ and $\hat{m}^{-i}(u,\mathtt{x})$, respectively.

5.1. Leave-one-out formulas

To derive the leave-one-out formulas in the case of multivariate covariates, we need some additional notation. Let

$$\mathbf{X}^*(u)=(\mathbf{u}_1(u)\otimes\mathtt{X}_1,\dots,\mathbf{u}_n(u)\otimes\mathtt{X}_n)^{\mathrm{T}},$$
$$\mathbf{W}^*(u;\boldsymbol{\beta})=\mathrm{diag}\{K_h(U_j-u)b''(\{\mathbf{u}_j(u)\otimes\mathtt{X}_j\}^{\mathrm{T}}\boldsymbol{\beta})\},$$
$$S_n^*(u;\boldsymbol{\beta})=\mathbf{X}^*(u)^{\mathrm{T}}\mathbf{W}^*(u;\boldsymbol{\beta})\mathbf{X}^*(u),$$

and define a projection matrix as

$$\mathcal{H}^*(u;\boldsymbol{\beta})=\{\mathbf{W}^*(u;\boldsymbol{\beta})\}^{1/2}\mathbf{X}^*(u)\{S_n^*(u;\boldsymbol{\beta})\}^{-1}\mathbf{X}^*(u)^{\mathrm{T}}\{\mathbf{W}^*(u;\boldsymbol{\beta})\}^{1/2}.$$

Let $\mathcal{H}_{ii}^*(u;\boldsymbol{\beta})$ be its ith diagonal entry and $H_i^*=\mathcal{H}_{ii}^*(U_i;\hat{\boldsymbol{\beta}}(U_i))$. Propositions 6 and 7 present the leave-one-out formulas and CV estimate of the total prediction error.

Proposition 6

Assume condition A2 in the Appendix. Then for any $h>0$ and $i=1,\dots,n$,

$$\hat{\boldsymbol{\beta}}^{-i}(u)-\hat{\boldsymbol{\beta}}(u)\doteq-\frac{\{S_n^*(u;\hat{\boldsymbol{\beta}}(u))\}^{-1}\{\mathbf{u}_i(u)\otimes\mathtt{X}_i\}K_h(U_i-u)\{Y_i-b'(\{\mathbf{u}_i(u)\otimes\mathtt{X}_i\}^{\mathrm{T}}\hat{\boldsymbol{\beta}}(u))\}}{1-\mathcal{H}_{ii}^*(u;\hat{\boldsymbol{\beta}}(u))},$$
$$\hat{\theta}_i^{-i}-\hat{\theta}_i\doteq-\frac{H_i^*}{1-H_i^*}\cdot\frac{Y_i-\hat{m}_i}{b''(\hat{\theta}_i)},$$
$$\hat{m}_i^{-i}-\hat{m}_i\doteq-\frac{H_i^*}{1-H_i^*}(Y_i-\hat{m}_i).$$

Proposition 7
Assume conditions A1 and A2 in the Appendix. Then for any $h>0$,

$$\widehat{\mathrm{Err}}^{\mathrm{CV}} \doteq \sum_{i=1}^{n}[Q(Y_i,\hat{m}_i)+2^{-1}q''(\hat{m}_i)(Y_i-\hat{m}_i)^2\{1-1/(1-H_i^*)^2\}]. \tag{27}$$

5.2. *Empirical cross-validation*

In the generalized varying-coefficient model, the asymptotic expression of the degrees of freedom $\sum_{i=1}^{n} H_i^*$ is given below.

Proposition 8
Assume conditions A and C in the Appendix. Then

$$\sum_{i=1}^{n} H_i^* = d\mathcal{K}(0)\left(|\Omega_U|/h\right)\{1+o_P(1)\}.$$

As $h\to\infty$, the total number of model parameters becomes $d(p+1)$ and this motivates us to propose the empirical formula for degrees of freedom:

$$\sum_{i=1}^{n} H_i^* \doteq d[(p+1-a)+\{\mathcal{C}n/(n-d)\}\mathcal{K}(0)\,|\Omega_U|\,/h]. \tag{28}$$

The empirical version of the estimated total prediction error is to replace H_i^* in (27) by $d[(p+1-a)/n+\{\mathcal{C}/(n-d)\}\mathcal{K}(0)\,|\Omega_U|\,/h]$. Call $\widehat{\mathrm{Err}}^{\mathrm{ECV}}(h)$ the empirical version of the CV criterion. Compared with the bandwidth selector in Cai *et al.* (2000), the $\widehat{\mathrm{Err}}^{\mathrm{ECV}}(h)$-minimizing bandwidth selector, $\hat{h}_{\mathrm{ECV}}$, is much easier to obtain.

5.3. *Binary responses*

For Bernoulli responses, the LB method in section 4 continues to be applicable for obtaining $\hat{\boldsymbol{\beta}}(u)$ and $\hat{\boldsymbol{\beta}}^{-i}(u)$. For the local logistic regression, $\nabla^2\ell(\boldsymbol{\beta};u)$ has a lower bound, $\mathbf{B}(u)=-4^{-1}\mathbf{X}^*(u)^{\mathrm{T}}\mathbf{K}^*(u)\mathbf{X}^*(u)$, where $\mathbf{K}^*(u)=\mathrm{diag}\{K_h(U_j-u)\}$. Similar to (18), the LB-adjusted NR iteration for $\hat{\boldsymbol{\beta}}(u)$ proceeds as follows,

$$\boldsymbol{\beta}_L=\boldsymbol{\beta}_{L-1}-\{\mathbf{B}(u)\}^{-1}\mathbf{X}^*(u)^{\mathrm{T}}\mathbf{K}^*(u)\mathbf{r}^*(u;\boldsymbol{\beta}_{L-1}),\quad L=1,2,\ldots,$$

where $\mathbf{r}^*(u;\boldsymbol{\beta})=(r_1^*(u;\boldsymbol{\beta}),\ldots,r_n^*(u;\boldsymbol{\beta}))^{\mathrm{T}}$ with

$$r_j^*(u;\boldsymbol{\beta})=Y_j-1/(1+\exp[-\{\mathbf{u}_j(u)\otimes\mathtt{X}_j\}^{\mathrm{T}}\boldsymbol{\beta}]).$$

The leave-one-out formulas and the CV estimates of the prediction error are similar to those in proposition 5, with $S_n(x)$ replaced by

$$S_n^*(u)=\mathbf{X}^*(u)^{\mathrm{T}}\mathbf{K}^*(u)\mathbf{X}^*(u)$$

and $\mathcal{S}_i$ by

$$\mathcal{S}_i^*=(\boldsymbol{e}_1\otimes\mathtt{X}_i)^{\mathrm{T}}\{S_n^*(U_i)\}^{-1}(\boldsymbol{e}_1\otimes\mathtt{X}_i)K_h(0).$$

In the spirit of (25), the hybrid selection criterion for bandwidth is

$$\sum_{i=1}^{n}\left[Q(Y_i,\hat{m}_i)+2^{-1}q''(\hat{m}_i)(Y_i-\hat{m}_i)^2\left[1-\left\{1+\frac{2b''(\hat{\theta}_i)\mathcal{S}_i^*}{1-\mathcal{S}_i^*}+\frac{2^{-1}H_i^*}{1-H_i^*}\right\}^2\right]\right]. \tag{29}$$

The ECV criterion can be obtained similarly via replacing S_i^* and H_i^* by their empirical averages, which are (28) divided by n.

6. Simulations

For Bernoulli responses, we apply the hybrid bandwidth selector to local logistic regression. Throughout our simulations, we use the q_2-function associated with the deviance loss for bandwidth selection, combined with the local-linear likelihood method and the Epanechnikov kernel. Unless specifically mentioned otherwise, the sample size is $n=400$.

6.1. Generalized non-parametric regression model

For simplicity, we assume that the predictor variable X has the uniform probability density on the interval $(0, 1)$. The bandwidth $\hat{h}_{\text{ECV}}$ is searched over an interval, $[h_{\min}, 0.5]$, at a geometric grid of 30 points. We take $h_{\min}=3h_0$ for Poisson regression, whereas for logistic regression, we take $h_{\min}=5h_0$ in example 1 and $h_{\min}=0.1$ in examples 2 and 3, where $h_0=\max[5/n, \max_{2\le j\le n}\{X_{(j)}-X_{(j-1)}\}]$, with order statistics $X_{(1)}\le\cdots\le X_{(n)}$.

Poisson regression: We first consider the response variable Y which, conditional on $X=x$, follows a Poisson distribution with parameter $\lambda(x)$. The function $\theta(x)=\ln\{\lambda(x)\}$ is given in the test examples,

Example 1: $\theta(x)=3.5[\exp\{-(4x-1)^2\}+\exp\{-(4x-3)^2\}]-1.5$,
Example 2: $\theta(x)=\sin\{2(4x-2)\}+1.0$,
Example 3: $\theta(x)=2-0.5(4x-2)^2$.

As an illustration, we first generate from (X, Y) one sample of independent observations $\{(X_j, Y_j)_{j=1}^n\}$. Figure 3(A) plots the degrees of freedom as a function of h. It is clearly seen that the actual values (denoted by dots) are well approximated by the empirical values (denoted by circles) given by (16). To see the performance of $\hat{h}_{\text{ECV}}$, Fig. 3(B) gives boxplots of the relative error, $\{\hat{h}_{\text{ECV}}-h_{\text{AMPEC}}(q_2)\}/h_{\text{AMPEC}}(q_2)$ and $\{\hat{h}_{\text{ECV}}-h_{\text{AMISE}}\}/h_{\text{AMISE}}$, based on 100 random samples; refer to Table 1 for values of $h_{\text{AMPEC}}(q_2)$ and h_{AMISE}. We observe that $\hat{h}_{\text{ECV}}$ is closer to $h_{\text{AMPEC}}(q_2)$ than to h_{AMISE}; this is in accordance with the discussion of section 3.2. In Fig. 3(C), we simulate another 100 random samples and for each set obtain $\hat{h}_{\text{ECV}}$ to estimate

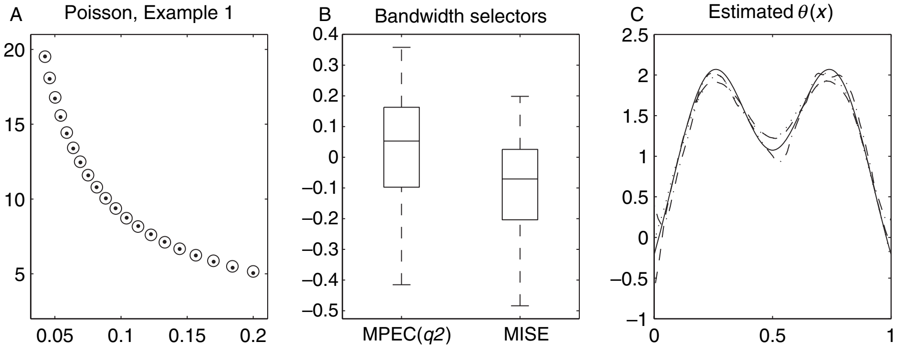

Fig. 3. Local-likelihood non-parametric regression for Poisson responses. (A) Plot of $\sum_{i=1}^n H_i$ versus h. Dots denote the actual values, centres of circles stand for the empirical values given by (16), for local-linear smoother with $a=0.70$ and $C=1.03$. (B) Boxplots of $\{\hat{h}_{\text{ECV}}-h_{\text{AMPEC}}(q_2)\}/h_{\text{AMPEC}}(q_2)$ and $\{\hat{h}_{\text{ECV}}-h_{\text{AMISE}}\}/h_{\text{AMISE}}$. (C) Estimated curves from three typical samples are presented corresponding to the 25th (the dotted curve), the 50th (the dashed curve), and the 75th (the dash-dotted curve) percentiles among the ASE-ranked values. The solid curves denote the true functions.

$\theta(x)$. We present the estimated curves from three typical samples. The typical samples are selected in such a way that their ASE values, in which $\mathrm{ASE} = n^{-1}\sum_{j=1}^{n}\{\hat{\theta}(X_j) - \theta(X_j)\}^2$, are equal to the 25th (dotted line), 50th (dashed line) and 75th (dash-dotted line) percentiles in the 100 replications. Inspection of these fitted curves suggests that the bandwidth selector based on minimizing the CV deviance does not exhibit undersmoothing in the local-likelihood regression estimation. Similar plots for examples 2 and 3 are omitted.

Logistic regression: We now consider the Bernoulli response variable Y with canonical parameter, $\theta(x) = \mathrm{logit}\{P(Y = 1 \mid X = x)\}$, chosen according to

Example 1: $\theta(x) = 7[\exp\{-(4x-1)^2\} + \exp\{-(4x-3)^2\}] - 5.5$,
Example 2: $\theta(x) = 2.5\sin(2\pi x)$,
Example 3: $\theta(x) = 2 - (4x-2)^2$.

In Fig. 4, we conduct the simulation experiments serving a similar purpose to Fig. 3. Plots in the middle panel support the convergence of the hybrid bandwidth selector $\hat{h}_{\mathrm{ECV}}$ to $h_{\mathrm{AMPEC}}(q_2)$, without suffering from the under- or oversmoothing problem.

6.2. Generalized varying-coefficient model

We consider examples of the generalized varying-coefficient model (26). We take $h_{\min} = 3h_0$ for Poisson regression, where $h_0 = \max[5/n, \max_{2\le j\le n}\{U_{(j)} - U_{(j-1)}\}]$, and $h_{\min} = 0.1$ for logistic regression.

Poisson regression: We consider a variable Y, given values $(u, \mathtt{x})$ of the covariates $(U, \mathtt{X})$, following a Poisson distribution with parameter $\lambda(u, \mathtt{x})$, where the varying-coefficient functions in $\ln\{\lambda(u, \mathtt{x})\}$ are specified as $d = 3$, $a_1(u) = 5.5 + 0.1\exp(2u-1)$, $a_2(u) = 0.8u(1-u)$ and $a_3(u) = 0.2\sin^2(2\pi u)$. We assume that U is a uniform random variable on the interval [0, 1] and is independent of $\mathtt{X} = (X_1, X_2, X_3)^{\mathrm{T}}$, with $X_1 \equiv 1$, where (X_2, X_3) follows a zero-mean and unit-variance bivariate normal distribution with correlation coefficient $1/\sqrt{2}$. In Fig. 5, plot (A) reveals that the actual degrees of freedom are well captured by the empirical formula (28). To evaluate the performance of $\hat{h}_{\mathrm{ECV}}$, we generate 100 random samples of size 400. Figure 5(B)–(D) plots the estimated curves of $a_1(u)$, $a_2(u)$ and $a_3(u)$ from three typical samples. The typical samples are selected so that their ASE values, in which $\mathrm{ASE} = n^{-1}\sum_{j=1}^{n}\{\hat{\theta}(U_j, \mathtt{X}_j) - \theta(U_j, \mathtt{X}_j)\}^2$, correspond to the 25th (dotted line), 50th (dashed line) and 75th (dash-dotted

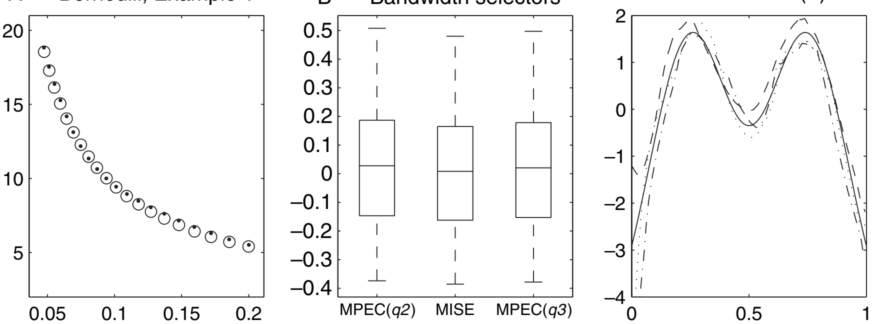

Fig. 4. Local-likelihood non-parametric regression for Bernoulli responses. Notes are similar to those of Fig. 3. Here $\hat{h}_{\mathrm{ECV}}$ minimizes the empirical version of (25); the formula (16) uses $a = 0.70$ and $C = 1.09$ for H_i and $a = 0.70$ and $C = 1.03$ for $\mathcal{S}_i$.

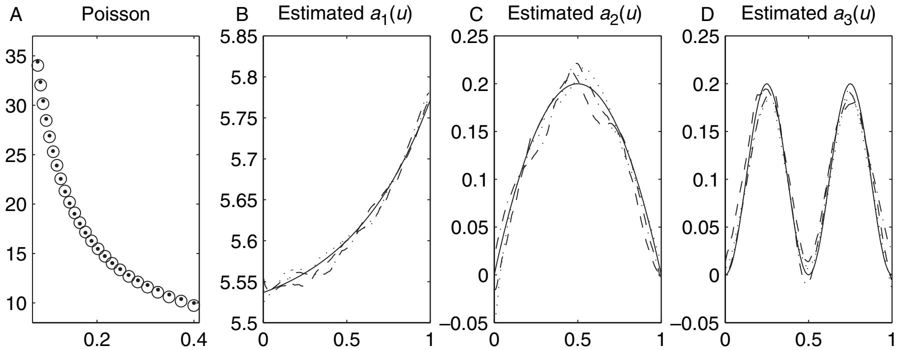

Fig. 5. Local-likelihood varying coefficient regression for Poisson responses. (A) Plot of $\sum_{i=1}^{n} H_i^*$ versus h. Dots denote the actual values, centres of circles stand for the empirical values given by (28), for local-linear smoother with $a=0.70$ and $\mathcal{C}=1.03$. (B)–(D) Estimated curves from three typical samples are presented corresponding to the 25th (the dotted curve), the 50th (the dashed curve), and the 75th (the dash-dotted curve) percentiles among the ASE-ranked values. The solid curves denote the true functions.

line) percentiles in the 100 replications. These plots provide convincing evidences that $\hat{h}_{\mathrm{ECV}}$, when applied to multiple smooth curves (possessing comparable degrees of smoothness) simultaneously, performs competitively well with that to fitting a single smooth curve.

Logistic regression: Consider the varying-coefficient logistic regression model for Bernoulli responses, where varying-coefficient functions in $\mathrm{logit}\{P(Y=1 \mid U=u, \mathtt{X}=\mathtt{x})\}$ are specified as $d=3$, $a_1(u)=\exp(2u-1)-1.5$, $a_2(u)=0.8\{8u(1-u)-1\}$ and $a_3(u)=0.9\{2\sin(\pi u)-1\}$. We assume that $X_1=1$; X_2 and X_3 are uncorrelated standard normal variables, and are independent of $U \sim U(0,1)$. Figure 6 depicts plots whose captions are similar to those for Fig. 5. Compared with previous examples of univariate logistic regression and varying-coefficient Poisson regression, the current model fitting for binary responses is considerably more challenging. Despite the increased difficulty, the LB local-likelihood logistic regression estimates, using the hybrid bandwidth selector $\hat{h}_{\mathrm{ECV}}$, captures the major features of the model structure with reasonably good details.

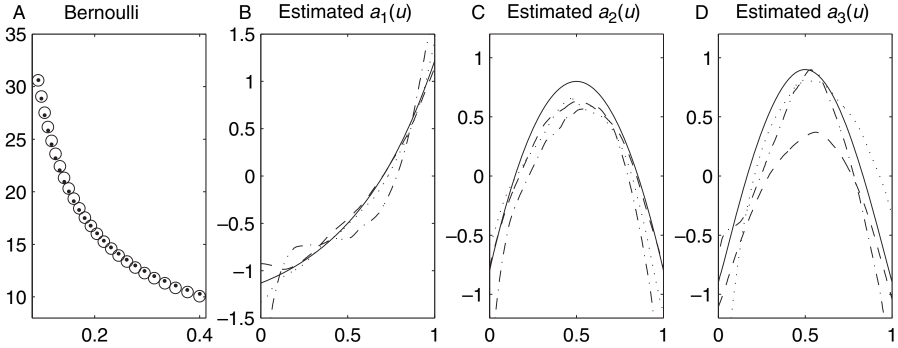

Fig. 6. Local-likelihood varying coefficient regression for Bernoulli responses. Notes are similar to those of Fig. 5. Here $\hat{h}_{\mathrm{ECV}}$ minimizes the empirical version of (29); the formula (28) uses $a=0.70$ and $\mathcal{C}=1.09$ for H_i^* and $a=0.70$ and $\mathcal{C}=1.03$ for $\mathcal{S}_i^*$.

6.3. Generalized additive model

Estimating multi-variate non-parametric regression functions is a challenging task. An efficient technique which overcomes the ‘curse-of-dimensionality’ is the generalized additive modelling (Hastie & Tibshirani, 1990). It assumes that

$$\theta(\mathtt{X}) = g(E(Y \mid \mathtt{X})) = \alpha + \sum_{j=1}^{d} f_j(X_j)$$

for a parameter α and univariate smooth functions $f_1, \ldots, f_d$. To ensure identifiability, the conditions $E\{f_j(X_j)\} = 0, j = 1, \ldots, d$, are usually imposed. The unknown parameter and functions can be estimated via iterative back-fitting local-likelihood estimation, in which only univariate smoothing is needed. Thus, the bandwidth selection method in sections 3 and 4 can be adopted. Consider independent covariates $X_1 \sim U(0,1)$, $X_2 \sim U(0,1)$ and $X_3 \sim U(-1,1)$. For Poisson regression with $\alpha = 4$,

$$f_1(X_1) = \sin(2\pi X_1)/5 - E\{\sin(2\pi X_1)/5\},$$

$$f_2(X_2) = \{2 - (4X_2 - 2)^2\}/5 - E[\{2 - (4X_2 - 2)^2\}/5]$$

and

$$f_3(X_3) = X_3^2/5 - E(X_3^2/5),$$

Fig. 7 plots local-likelihood estimates of f_j based on 100 replications of size 200. Similar plots in Fig. 8 are for logistic regression with $\alpha = 0$,

$$f_1(X_1) = 2\sin(2\pi X_1) - E\{2\sin(2\pi X_1)\},$$

$$f_2(X_2) = \{2 - (4X_2 - 2)^2\} - E\{2 - (4X_2 - 2)^2\}$$

and

$$f_3(X_3) = X_3^2 - E(X_3^2).$$

The results further support the effectiveness of the proposed $\hat{h}_{\mathrm{ECV}}$ when applied to multivariate modelling.

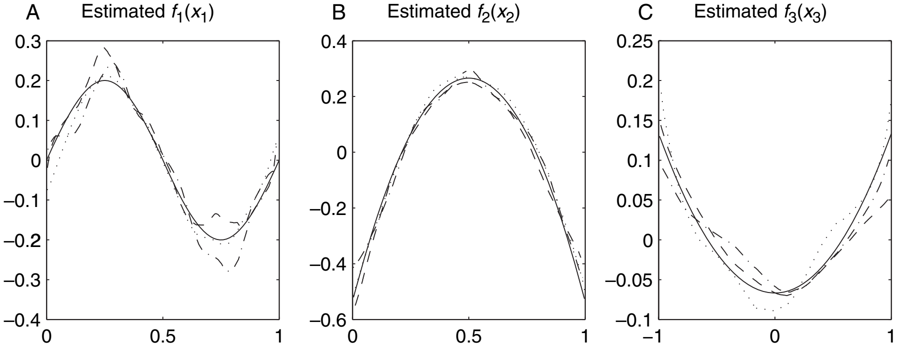

Fig. 7. Generalized additive regression for Poisson responses. Notes are similar to those of Fig. 5.

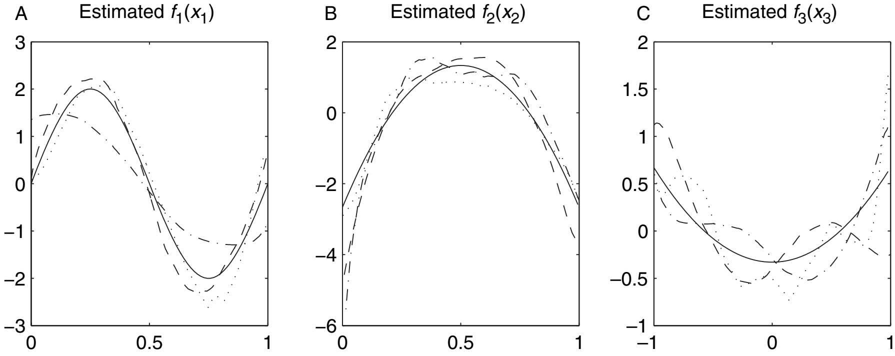

Fig. 8. Generalized additive regression for Bernoulli responses. Notes are similar to those of Fig. 5.

7. Real data applications

Example 7.1 (Bank employee data)

We apply the hybrid bandwidth selection method for binary responses to analyse an employee (year 1995) dataset (example 11.3 of Albright *et al.*, 1999) of the Fifth National Bank of Springfield. The bank, whose name has been changed, was charged in court with that its female employees received substantially smaller salaries than its male employees. For each of its 208 employees, the dataset consists of eight variables, including JobGrade, a categorical variable for the current job level, with possible values 1–6 (6 is highest); YrHired, year employee was hired; YrBorn, year employee was born; Gender: a categorical variable with values 'Female' and 'Male'; YrsPrior, number of years of work experience at another bank prior to working at Fifth National.

To understand how the probability of promotion to high levels of managerial job (and thus high salary) is associated with gender and years of work experience, and how this association changes with respect to age, we fit a varying-coefficient logistic model,

$$\operatorname{logit}\{P(Y=1 \mid U=u, X_1=x_1, X_2=x_2)\}=a_0(u)+a_1(u)x_1+a_2(u)x_2, \tag{30}$$

with Y the indicator of JobGrade at least 4, U the covariate Age, X_1 the indicator of being Female, and X_2 the covariate WorkExp (calculated as $95-\text{YrHired}+\text{YrsPrior}$). Following Fan & Peng (2004), outliers have been deleted, with the remaining 199 data for analysis. For this medium-sized data, use of the bandwidth selector $\hat{h}_{\mathrm{ACV}}$ which minimizes (29) seems to be more natural than $\hat{h}_{\mathrm{ECV}}$.

Our preliminary study shows a monotone decreasing pattern in the fitted curve of $a_2(u)$. This is no surprise; the covariates Age and WorkExp are highly correlated, as can be seen from the scatter plot in Fig. 9(A). Such high correlation may cause some identifiability problem, thus in model (30), we replace X_2 with a de-correlated variable, $X_2-E(X_2|U)$, which is known to be uncorrelated with any measurable function of U. The projection part, $E(X_2|U=u)$, can easily be estimated by a univariate local linear regression fit. Likewise, its bandwidth parameter can simply be chosen to minimize the approximate CV function (for Gaussian family), illustrated in Fig. 9(B).

After the de-correlation step, we now refit model (30). The bottom panel of Fig. 9 depicts the estimated functions of $a_0(u)$, $a_1(u)$ and $a_2(u)$, ± 1.96 times their estimated standard error. The selected bandwidth is 16.9 (see Fig. 9(C)). Both the intercept term and (de-correlated) WorkExp have the statistically significant effects on the probability of promotion. As an employee gets older, the probability of getting promoted keeps increasing until around 40

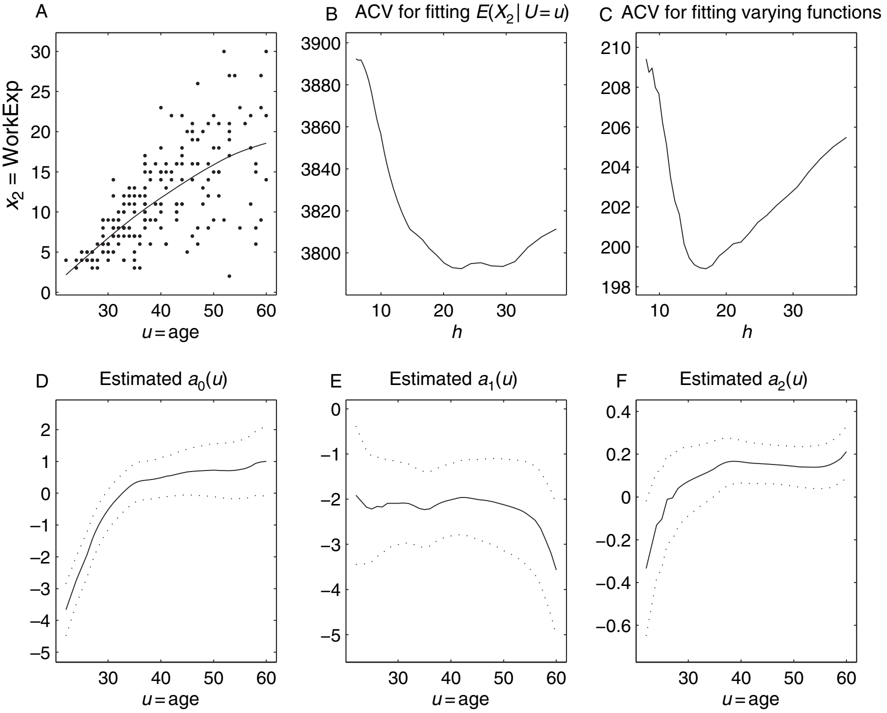

Fig. 9. Applications to the job grade data set modelled by (30). (A) Scatter plot of work experience versus age along with a local linear fit. (B) Plot of the approximate cross-validation (CV) function for the local linear fit in (A). (C) Plot of the approximate CV function, defined in (29), for fitting varying coefficient functions. (D)–(F) Estimated $a_0(u)$, $a_1(u)$ and $a_2(u)$, respectively, where the dotted curves are the estimated functions ± 1.96 times the estimated standard errors.

years of age and levels off after that. It is interesting to note that the fitted coefficient function of $a_1(u)$ for gender is below zero within the entire age span. This may be interpreted as the evidence of discrimination against female employees being promoted and lends support to the plaintiff.

To see whether the choice of smoothing variable U makes a difference in drawing the above conclusion, we fit again model (30) with U given by the covariate WorkExp and X_2 by the de-correlated Age (due to the same reason of monotonicity as in the previous analysis). Again, the result (omitted here) shows that gender has an adverse effect and the evidence for discrimination continues to be strong. Indeed, the estimated varying-function of $a_1(u)$ is qualitatively the same as that in Fig. 9, as far as the evidence of discrimination is concerned.

We would like to make a final remark on the de-correlation procedure: This step does not alter (30), particularly the function $a_1(\cdot)$. If this step is not taken, then the estimate of $a_1(u)$ from either choice of U continues to be below zero and thus does not alter our previous interpretation of the gender effect.

Example 7.2 (Boston Housing data)

The data set contains the response MEDV, the median value of owner-occupied homes (in \$1000's) in 506 US census tracts of the Boston metropolitan area in 1970, along with several explanatory variables which might affect housing values (see Harrison & Rubinfeld, 1978). The covariates CRIM (per capita crime rate by town), ZN (proportion of residential land

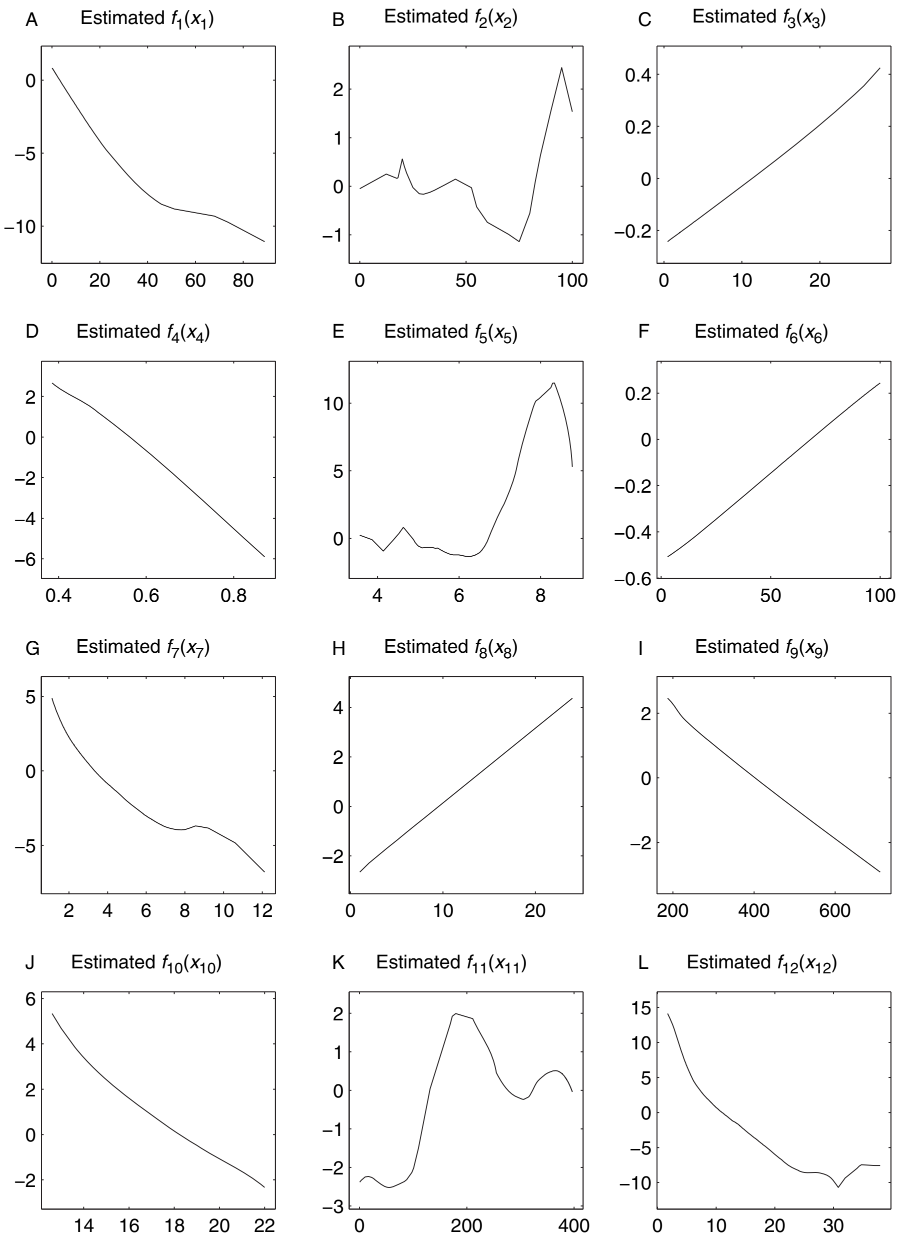

Fig. 10. Applications to the Boston housing data modelled by (31).

zoned for lots over 25,000 sq. ft.), INDUS (proportion of non-retail business acres per town), NOX (nitric oxides concentration (parts per 10 million)), RM (average number of rooms per dwelling), AGE (proportion of owner-occupied units built prior to 1940), DIS (weighted distances to five Boston employment centres), RAD (index of accessibility to radial highways), TAX (full-value property-tax rate per \$10,000), PTRATIO (pupil–teacher ratio by town), B ($1000(\mathrm{Bk}-0.63)^2$ where Bk is the proportion of blacks by town) and LSTAT (% lower status of the population) are denoted by $X_1, \ldots, X_{12}$, respectively.

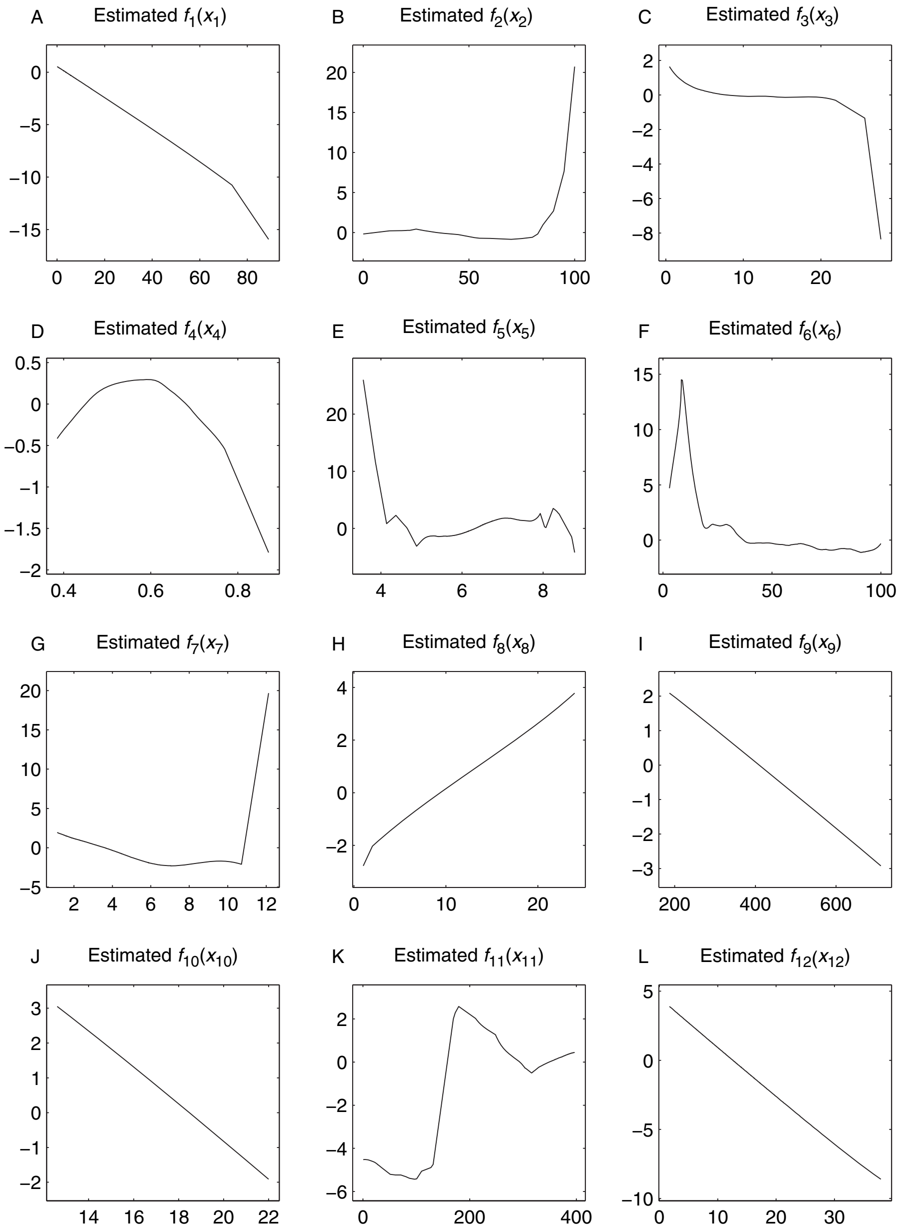

Fig. 11. Applications to the Boston housing data modelled by (32).

First, we intend to understand the association between the median value of owner-occupied homes, denoted by Y, and the 12 covariates. The additive model

$$E(Y \mid X_1 = x_1, \ldots, X_{12} = x_{12}) = \alpha + \sum_{j=1}^{12} f_j(x_j) \tag{31}$$

is fitted to the data set, and the proposed $\hat{h}_{\mathrm{ECV}}$ is applied. Figure 10 depicts the estimated curves $f_j(\cdot)$. The trends in panels (D) and (J), for example, suggest that the housing price tends to be lower in the tracts with more crowded schools, and decreases with the level of air pollution.

Second, to predict whether the median housing price can be categorized as either ‘high’ or ‘low’ (compared with the average of MEDV), the additive logistic model

$$\operatorname{logit}\{P(Y^* = 1 \mid X_1 = x_1, \ldots, X_{12} = x_{12})\} = \alpha + \sum_{j=1}^{12} f_j(x_j) \tag{32}$$

is fitted to the data set, where Y^* equals 1 if Y exceeds the average of MEDV and 0 otherwise. The estimated component curves are illustrated in Fig. 11. The effects of most covariates on housing price agree well with those in Fig. 10.

8. Discussion

In this paper, we aim to develop effective methods for estimating prediction error under a broad q-class of loss functions, with applications to non-parametric regression and classification.

A number of extensions could be further made. First, a comparison with alternative approaches and other smoothing techniques could be carefully carried out. Second, the current paper focuses on non-parametric estimators with a fixed number of explanatory variables. For the analysis of high-dimensional data, like spatio-temporal fMRI brain images, functional data objects and gene expression profiles, it would be interesting to investigate the prediction error estimation for penalized estimators in the presence of a diverging number of covariates. These issues will be explored in future work.

Acknowledgements

The research is supported in part by National Science Foundation grants DMS-03-53941, DMS-07-05209 and Wisconsin Alumni Research Foundation. The author thanks the Editor, the Associate Editor and two anonymous referees for their helpful comments and suggestions on an earlier version of this paper. The author also appreciates Professor Jianqing Fan’s comments.

References

Albright, S. C., Winston, W. L. & Zappe, C. J. (1999). *Data analysis and decision making with Microsoft Excel*. Duxbury Press, Pacific Grove, California.

Altman, N. & MacGibbon, B. (1998). Consistent bandwidth selection for kernel binary regression. *J. Statist. Plan. Inference* **70**, 121–137.

Aragaki, A. & Altman, N. S. (1997). Local polynomial regression for binary response. In *Computing science and statistics: Proceedings of the 29th Symposium on the Interface*. Houston, Texas.

Böhning, D. & Lindsay, B. G. (1988). Monotonicity of quadratic approximation algorithms. *Ann. Inst. Statist. Math.* **40**, 641–663.

Bregman, L. M. (1967). A relaxation method of finding a common point of convex sets and its application to the solution of problems in convex programming. *U.S.S.R. Comput. Math. and Math. Phys.* **7**, 620–631.

Cai, Z., Fan, J. & Li, R. (2000). Efficient estimation and inferences for varying-coefficient models. *J. Amer. Statist. Assoc.* **95**, 888–902.

Davidson, A. C. & Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge University Press, Cambridge.

Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *J. Amer. Statist. Assoc.* **81**, 461–470.

Efron, B. (2004). The estimation of prediction error: covariance penalties and cross-validation (with discussion). *J. Amer. Statist. Assoc.* **99**, 619–642.

Fan, J. & Chen, J. (1999). One-step local quasi-likelihood estimation. *J. Roy. Statist. Soc. Ser. B* **61**, 927–943.

Fan, J., Heckman, N. & Wand, M. P. (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *J. Amer. Statist. Assoc.* **90**, 141–150.

Fan, J., Farmen, M. & Gijbels, I. (1998). Local maximum likelihood estimation and inference. *J. Roy. Statist. Soc. Ser. B* **60**, 591–608.

Fan, J. & Peng, H. (2004). On non-concave penalized likelihood with diverging number of parameters. *Ann. Statist.* **32**, 928–961.

Golub, G. H. & Van Loan, C. F. (1996). *Matrix computations*. 3rd edn. Johns Hopkins University Press, Baltimore.

Hall, P. & Johnstone, I. (1992). Empirical functionals and efficient smoothing parameter selection (with discussion). *J. Royal. Statist. Soc. B* **54**, 475–530.

Hardy, G. H., Littlewood, J. E. & Pólya, G. (1988). *Inequalities*, 2nd edn. Cambridge University Press, Cambridge, UK.

Härdle, W., Hall, P. & Marron, J. S. (1992). Regression smoothing parameters that are not far from their optimum. *J. Amer. Statist. Assoc.* **87**, 227–233.

Harrison, D. & Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *J. Environ. Econ. Manage.* **5**, 81–102.

Hastie, T. J. & Tibshirani, R. J. (1990). *Generalized additive models*. Chapman and Hall, London.

Hastie, T. J., Tibshirani, R. & Friedman, J. (2001). *The elements of statistical learning: data mining, inference, and prediction*. Springer-Verlag, New York.

McCullagh, P. & Nelder, J. A. (1989). *Generalized linear models*, 2nd edn. Chapman and Hall, London.

Mitrinović, D. S., Pečarić, J. E. & Fink, A. M. (1993). *Classical and new inequalities in analysis*. Kluwer Academic Publishers Group, Dordrecht.

Müller, H.-G. & Schmitt, T. (1988). Kernel and probit estimates in quantal bioassay. *J. Amer. Statist. Assoc.* **83**, 750–759.

Nelder, J. A. & Wedderburn, R. W. M. (1972). Generalized linear models. *J. Roy. Statist. Soc. Ser. A* **135**, 370–384.

Pregibon, D. (1981). Logistic regression diagnostics. *Ann. Statist.* **9**, 705–724.

Rice, J. (1984). Bandwidth choice for nonparametric regression. *Ann. Statist.* **20**, 712–736.

Ruppert, D. & Wand, M. P. (1994). Multivariate locally weighted least squares regression. *Ann. Statist.* **22**, 1346–1370.

Severini, T. A. & Staniswalis, J. G. (1994). Quasi-likelihood estimation in semiparametric models. *J. Amer. Statist. Assoc.* **89**, 501–511.

Staniswalis, J. G. (1989). The kernel estimate of a regression function in likelihood-based models. *J. Amer. Statist. Assoc.* **84**, 276–283.

Tibshirani, R. & Hastie, T. (1987). Local likelihood estimation. *J. Amer. Statist. Assoc.* **82**, 559–567.

Tibshirani, R. (1996). Bias, variance and prediction error for classification rules. *Technical report*. Statistics Department, University of Toronto.

Wong, W. H. (1983). On the consistency of cross-validation in kernel nonparametric regression. *Ann. Statist.* **11**, 1136–1141.

Xiang, D. & Wahba, G. (1996). A generalized approximate cross validation for smoothing splines with non-Gaussian data. *Statist. Sinica* **6**, 675–692.

Zhang, C. M. (2003). Calibrating the degrees of freedom for automatic data smoothing and effective curve checking. *J. Amer. Statist. Assoc.* **98**, 609–628.

Received March 2007, in final form December 2007

Chunming Zhang, Department of Statistics, University of Wisconsin, Madison, WI 53706, USA.
E-mail: cmzhang@stat.wisc.edu

Appendix

We first impose some technical assumptions, which are not the weakest possible.

Condition A

A1. The function q is concave and $q''(\cdot)$ is continuous.

A2. $b''(\cdot)$ is continuous and bounded away from zero.

A3. The kernel function K is a symmetric probability density function with bounded support, and is Lipschitz continuous.

A4. $(nh)^{-1}\ln(1/h)\to 0$ as $n\to\infty$.

Condition B

B1. The design variable X has a bounded support Ω_X and the density function f_X is Lipschitz continuous and bounded away from 0.

B2. $\theta(x)$ has continuous $(p+1)$th derivative in Ω_X.

B3. $n\to\infty$, $h\to 0$, $nh\to\infty$ and $n^{2\varepsilon-1}h\to\infty$ for some $\varepsilon<1-s^{-1}$ and some $s>0$.

Condition C

C1. The covariate U has a bounded support Ω_U and its density function f_U is Lipschitz continuous and bounded away from 0.

C2. $a_j(u)$, $j=1,\ldots,d$, has continuous $(p+1)$th derivative in Ω_U.

C3. The matrix $\Gamma(u)=E\{b''(\theta(u,\mathtt{X}))\mathtt{X}\mathtt{X}^T \mid U=u\}$ is positive definite for each $u\in\Omega_U$ and is Lipschitz continuous.

C4. There exists some $s>0$ such that $E(\|\mathtt{X}\|^{2s})<\infty$. Also, $n\to\infty$, $h\to 0$, $nh\to\infty$ and $n^{2\varepsilon-1}h\to\infty$ for some $\varepsilon<1-s^{-1}$.

Notation

Throughout our derivations, we simplify notation by writing $\theta_j(x;\boldsymbol{\beta})=\mathbf{x}_j(x)^{\mathrm{T}}\boldsymbol{\beta}$, $m_j(x;\boldsymbol{\beta})=b'(\theta_j(x;\boldsymbol{\beta}))$, $Z_j(x;\boldsymbol{\beta})=\{Y_j-m_j(x;\boldsymbol{\beta})\}/b''(\theta_j(x;\boldsymbol{\beta}))$, $\mathbf{z}(x;\boldsymbol{\beta})=(Z_1(x;\boldsymbol{\beta}),\ldots,Z_n(x;\boldsymbol{\beta}))^{\mathrm{T}}$ and $w_j(x;\boldsymbol{\beta})=K_h(X_j-x)b''(\theta_j(x;\boldsymbol{\beta}))$; their corresponding quantities evaluated at $\hat{\boldsymbol{\beta}}(x)$ are denoted by $\hat{\theta}_j(x)$, $\hat{m}_j(x)$, $\hat{Z}_j(x)$, $\hat{\mathbf{z}}(x)$ and $\hat{w}_j(x)$. Similarly, define $\hat{S}_n(x)=S_n(x;\hat{\boldsymbol{\beta}}(x))$.

Weighted local likelihood (illustrated for section 3)

To compute approximately $\hat{\boldsymbol{\beta}}^{-i}(x)$ from $\hat{\boldsymbol{\beta}}(x)$, we apply the 'infinitesimal perturbation' idea developed in Pregibon (1981). For fixed i, we introduce the weighted local log likelihood,

$$\ell_{i,\delta}(\boldsymbol{\beta};x)=\sum_{j=1}^{n}\delta_{ij}l(\mathbf{x}_j(x)^{\mathrm{T}}\boldsymbol{\beta};\,Y_j)K_h(X_j-x),\tag{33}$$

with the weight $\delta_{ii}=\delta$ and other weights $\delta_{ij}=1$. Let $\hat{\boldsymbol{\beta}}_{i,\delta}(x)$ be the maximizer, which is the local ML estimator when $\delta=1$ and the leave-one-out estimator when $\delta=0$. The weighted local MLE can be found via the Newton–Raphson iteration,

$$\boldsymbol{\beta}_L=\boldsymbol{\beta}_{L-1}-\{\nabla^2\ell_{i,\delta}(\boldsymbol{\beta}_{L-1};x)\}^{-1}\nabla\ell_{i,\delta}(\boldsymbol{\beta}_{L-1};x),\quad L=1,2,\ldots,\tag{34}$$

where $\nabla\ell$ denotes the gradient vector and $\nabla^2\ell$ the Hessian matrix. (Explicit expressions of $\nabla\ell$ and $\nabla^2\ell$ are given in lemma 2.) The key ingredient for calculating the leave-one-out estimator is to approximate it by its one-step estimator using the 'keep-all-in' estimator $\hat{\boldsymbol{\beta}}(x)$ as the initial value. Before proving the main results, we need lemma 2 below.

Lemma 2

For $\ell_{i,\delta}(\boldsymbol{\beta};x)$ defined in (33),

$$\nabla\ell_{i,\delta}(\boldsymbol{\beta};x)=\mathbf{X}(x)^{\mathrm{T}}\mathbf{V}_i(\delta)\mathbf{W}(x;\boldsymbol{\beta})\mathbf{z}(x;\boldsymbol{\beta})/a(\psi),\tag{35}$$

$$\nabla^2\ell_{i,\delta}(\boldsymbol{\beta};x)=-\mathbf{X}(x)^{\mathrm{T}}\mathbf{V}_i(\delta)\mathbf{W}(x;\boldsymbol{\beta})\mathbf{X}(x)/a(\psi),\tag{36}$$

in which $\mathbf{V}_i(\delta)=\mathrm{diag}\{\delta_{i1},\ldots,\delta_{in}\}$ and $r_i(x;\boldsymbol{\beta})=Y_i-m_i(x;\boldsymbol{\beta})$,

$$\mathbf{X}(x)^{\mathrm{T}}\mathbf{V}_i(\delta)\mathbf{W}(x;\boldsymbol{\beta})\mathbf{z}(x;\boldsymbol{\beta}) = \mathbf{X}(x)^{\mathrm{T}}\mathbf{W}(x;\boldsymbol{\beta})\mathbf{z}(x;\boldsymbol{\beta}) - (1-\delta)\mathbf{x}_i(x)K_h(X_i - x)r_i(x;\boldsymbol{\beta}), \tag{37}$$

$$\mathbf{X}(x)^{\mathrm{T}}\mathbf{V}_i(\delta)\mathbf{W}(x;\boldsymbol{\beta})\mathbf{X}(x) = \mathbf{X}(x)^{\mathrm{T}}\mathbf{W}(x;\boldsymbol{\beta})\mathbf{X}(x) - (1-\delta)w_i(x;\boldsymbol{\beta})\mathbf{x}_i(x)\mathbf{x}_i(x)^{\mathrm{T}}. \tag{38}$$

Proof. Defining a vector $\boldsymbol{\theta}(x;\boldsymbol{\beta}) = (\theta_1(x;\boldsymbol{\beta}),\ldots,\theta_n(x;\boldsymbol{\beta}))^{\mathrm{T}}$, we have that

$$\nabla \ell_{i,\delta}(\boldsymbol{\beta};x) = \frac{\partial \boldsymbol{\theta}(x;\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \frac{\partial \ell_{i,\delta}(\boldsymbol{\beta};x)}{\partial \boldsymbol{\theta}(x;\boldsymbol{\beta})} = \mathbf{X}(x)^{\mathrm{T}} \frac{\partial \ell_{i,\delta}(\boldsymbol{\beta};x)}{\partial \boldsymbol{\theta}(x;\boldsymbol{\beta})}, \tag{39}$$

$$\nabla^2 \ell_{i,\delta}(\boldsymbol{\beta};x) = \mathbf{X}(x)^{\mathrm{T}} \frac{\partial^2 \ell_{i,\delta}(\boldsymbol{\beta};x)}{\partial \boldsymbol{\theta}(x;\boldsymbol{\beta})\partial \boldsymbol{\theta}(x;\boldsymbol{\beta})^{\mathrm{T}}}\mathbf{X}(x). \tag{40}$$

Since

$$\ell_{i,\delta}(\boldsymbol{\beta};x) = \sum_{j=1}^{n} \delta_{ij}[\{Y_j\mathbf{x}_j(x)^{\mathrm{T}}\boldsymbol{\beta} - b(\mathbf{x}_j(x)^{\mathrm{T}}\boldsymbol{\beta})\}/a(\psi) + c(Y_j,\psi)]K_h(X_j - x),$$

it is easy to check that (35) can be derived from (39) and

$$\partial \ell_{i,\delta}(\boldsymbol{\beta};x)/\partial \theta_j(x;\boldsymbol{\beta}) = \delta_{ij}\{Y_j - b'(\theta_j(x;\boldsymbol{\beta}))\}K_h(X_j - x)/a(\psi). \tag{41}$$

Following (41), we see that $\partial^2 \ell_{i,\delta}(\boldsymbol{\beta};x)/\{\partial\theta_j(x;\boldsymbol{\beta})\partial\theta_k(x;\boldsymbol{\beta})\} = 0$ for $j \neq k$, and $\partial^2 \ell_{i,\delta}(\boldsymbol{\beta};x)/\{\partial\theta_j(x;\boldsymbol{\beta})\}^2 = -\delta_{ij}b''(\theta_j(x;\boldsymbol{\beta}))/a(\psi)K_h(X_j - x)$. This along with (40) gives (36). Equations (37)–(38) follow from decomposing an identity matrix $\mathbf{I}$ into $\mathbf{V}_i(\delta)$ and $\mathbf{I} - \mathbf{V}_i(\delta)$.

Proof of Proposition 1. From (35) and (36), (34) can be rewritten as

$$\boldsymbol{\beta}_L = \boldsymbol{\beta}_{L-1} + \{\mathbf{X}(x)^{\mathrm{T}}\mathbf{V}_i(\delta)\mathbf{W}(x;\boldsymbol{\beta}_{L-1})\mathbf{X}(x)\}^{-1}\mathbf{X}(x)^{\mathrm{T}}\mathbf{V}_i(\delta)\mathbf{W}(x;\boldsymbol{\beta}_{L-1})\mathbf{z}(x;\boldsymbol{\beta}_{L-1}). \tag{42}$$

Setting $\delta = 0$ in (42), the one-step estimate of $\hat{\boldsymbol{\beta}}^{-i}(x)$ starting from $\boldsymbol{\beta}_0 = \hat{\boldsymbol{\beta}}(x)$ is

$$\hat{\boldsymbol{\beta}}(x) + \{\mathbf{X}(x)^{\mathrm{T}}\mathbf{V}_i(0)\mathbf{W}(x;\hat{\boldsymbol{\beta}}(x))\mathbf{X}(x)\}^{-1}\{\mathbf{X}(x)^{\mathrm{T}}\mathbf{V}_i(0)\mathbf{W}(x;\hat{\boldsymbol{\beta}}(x))\hat{\mathbf{z}}(x)\}.$$

Using the definition of $\hat{\boldsymbol{\beta}}(x)$ (satisfying $\nabla \ell_{i,\delta}(\boldsymbol{\beta};x) = 0$ with $\delta = 1$), along with (37) and (38), the above one-step estimate of $\hat{\boldsymbol{\beta}}^{-i}(x)$ equals

$$\hat{\boldsymbol{\beta}}(x) - \{\hat{S}_n(x) - \hat{w}_i(x)\mathbf{x}_i(x)\mathbf{x}_i(x)^{\mathrm{T}}\}^{-1}\mathbf{x}_i(x)K_h(X_i - x)\{Y_i - \hat{m}_i(x)\}. \tag{43}$$

By the Sherman–Morrison–Woodbury formula (Golub & Van Loan, 1996, p. 50),

$$\{\hat{S}_n(x) - \hat{w}_i(x)\mathbf{x}_i(x)\mathbf{x}_i(x)^{\mathrm{T}}\}^{-1} = \{\hat{S}_n(x)\}^{-1} + \frac{\hat{w}_i(x)\{\hat{S}_n(x)\}^{-1}\mathbf{x}_i(x)\mathbf{x}_i(x)^{\mathrm{T}}\{\hat{S}_n(x)\}^{-1}}{1 - \hat{w}_i(x)\mathbf{x}_i(x)^{\mathrm{T}}\{\hat{S}_n(x)\}^{-1}\mathbf{x}_i(x)}.$$

Thus

$$\{\hat{S}_n(x) - \hat{w}_i(x)\mathbf{x}_i(x)\mathbf{x}_i(x)^{\mathrm{T}}\}^{-1}\mathbf{x}_i(x) = \{\hat{S}_n(x)\}^{-1}\mathbf{x}_i(x)/\{1 - \mathcal{H}_{ii}(x;\hat{\boldsymbol{\beta}}(x))\},$$

by which (43) becomes

$$\hat{\boldsymbol{\beta}}(x) - \{\hat{S}_n(x)\}^{-1}\mathbf{x}_i(x)K_h(X_i - x)\{Y_i - \hat{m}_i(x)\}/\{1 - \mathcal{H}_{ii}(x;\hat{\boldsymbol{\beta}}(x))\}.$$

This expression approximates $\hat{\boldsymbol{\beta}}^{-i}(x)$ and thus leads to (10). To show (11) and (12), note that $\hat{\theta}_i = \hat{\theta}_i(X_i)$, $\hat{m}_i = \hat{m}_i(X_i)$, $\hat{\theta}_i^{-i} = \hat{\theta}_i^{-i}(X_i)$, $\hat{m}_i^{-i} = \hat{m}_i^{-i}(X_i)$ and

$$H_i = \boldsymbol{e}_1^{\mathrm{T}}\{\hat{S}_n(X_i)\}^{-1}\boldsymbol{e}_1K_h(0)b''(\hat{\theta}_i). \tag{44}$$

Applying (10) gives

$$\hat{\theta}_i^{-i} - \hat{\theta}_i = \boldsymbol{e}_1^{\mathrm{T}}\{\hat{\boldsymbol{\beta}}^{-i}(X_i) - \hat{\boldsymbol{\beta}}(X_i)\} \doteq -\boldsymbol{e}_1^{\mathrm{T}}\{\hat{S}_n(X_i)\}^{-1}\boldsymbol{e}_1K_h(0)(Y_i - \hat{m}_i)\{1 - \mathcal{H}_{ii}(X_i;\hat{\boldsymbol{\beta}}(X_i))\},$$

leading to (11). This and a first-order Taylor's expansion and the continuity of b'' yield $\hat{m}_i^{-i} - \hat{m}_i = b'(\hat{\theta}_i^{-i}) - b'(\hat{\theta}_i) \doteq (\hat{\theta}_i^{-i} - \hat{\theta}_i) b''(\hat{\theta}_i)$ and thus (12).

Proof of Proposition 2. Recall $\hat{\lambda}_i = -q'(\hat{m}_i)/2$ defined in section 2.2. By a first-order Taylor expansion, we have that $\hat{\lambda}_i - \hat{\lambda}_i^{-i} = 2^{-1}\{q'(\hat{m}_i^{-i}) - q'(\hat{m}_i)\} \doteq 2^{-1} q''(\hat{m}_i)(\hat{m}_i^{-i} - \hat{m}_i)$ and $Q(\hat{m}_i^{-i}, \hat{m}_i) \doteq -2^{-1} q''(\hat{m}_i)(\hat{m}_i^{-i} - \hat{m}_i)^2$. These, applied to an identity given in a lemma of Efron (2004, section 4), $Q(Y_i, \hat{m}_i^{-i}) - Q(Y_i, \hat{m}_i) = 2(\hat{\lambda}_i - \hat{\lambda}_i^{-i})(Y_i - \hat{m}_i^{-i}) - Q(\hat{m}_i^{-i}, \hat{m}_i)$, lead to

$$\begin{aligned} Q(Y_i, \hat{m}_i^{-i}) - Q(Y_i, \hat{m}_i) &\doteq q''(\hat{m}_i)(\hat{m}_i^{-i} - \hat{m}_i)(Y_i - \hat{m}_i^{-i}) + 2^{-1} q''(\hat{m}_i)(\hat{m}_i^{-i} - \hat{m}_i)^2 \\ &= 2^{-1} q''(\hat{m}_i)\{(Y_i - \hat{m}_i)^2 - (Y_i - \hat{m}_i^{-i})^2\}. \end{aligned}$$

Summing over i and using (12) and (6), we complete the proof.

Proof of Proposition 3. From (14) and (15), we see that

$$\frac{h_{\text{AMPEC}}(q_2)}{h_{\text{AMISE}}} = \left[|\Omega_X| \int_{\Omega_X} F(x)G(x)\,\mathrm{d}x \Big/ \left\{ \int_{\Omega_X} F(x)\,\mathrm{d}x \int_{\Omega_X} G(y)\,\mathrm{d}y \right\} \right]^{1/(2p+3)}. \tag{45}$$

To show part (a), it suffices to consider oppositely ordered F and G. In this case, by the Tchebychef's inequality (Hardy *et al.*, 1988, p. 43 and 168), we obtain

$$|\Omega_X| \int_{\Omega_X} F(x)G(x)\,\mathrm{d}x \le \int_{\Omega_X} F(x)\,\mathrm{d}x \int_{\Omega_X} G(y)\,\mathrm{d}y.$$

Since $F \ge 0$ and $G \ge 0$, it follows that

$$|\Omega_X| \int_{\Omega_X} F(x)G(x)\,\mathrm{d}x \Big/ \left\{ \int_{\Omega_X} F(x)\,\mathrm{d}x \int_{\Omega_X} G(y)\,\mathrm{d}y \right\} \le 1,$$

which along with (45) indicates that $h_{\text{AMPEC}}(q_2) \le h_{\text{AMISE}}$.

To verify part (b), it can be seen that under its assumptions, for a constant $C > 0$, $F(x) = C/|\Omega_X| b''(\theta(x))$ is oppositely ordered with $G(x) = \{b''(\theta(x))\}^{-1}$, and thus the conclusion of part (a) immediately indicates the upper bound 1. To show the lower bound, we first observe that (45) becomes

$$\frac{h_{\text{AMPEC}}(q_2)}{h_{\text{AMISE}}} = \left[\frac{|\Omega_X|^2}{\int_{\Omega_X} b''(\theta(x))\,\mathrm{d}x \int_{\Omega_X} \{b''(\theta(y))\}^{-1} \mathrm{d}y} \right]^{1/(2p+3)}. \tag{46}$$

Incorporating the Grüss integral inequality (Mitrinović *et al.*, 1993),

$$\left| \frac{1}{|\Omega_X|} \int_{\Omega_X} F(x)G(x)\,\mathrm{d}x - \frac{1}{|\Omega_X|^2} \int_{\Omega_X} F(x)\,\mathrm{d}x \int_{\Omega_X} G(y)\,\mathrm{d}y \right| \le \frac{1}{4}(M_F - m_F)(M_G - m_G),$$

where $M_F = \max_{x \in \Omega_X} F(x)$, $m_F = \min_{x \in \Omega_X} F(x)$ and M_G and m_G are similarly defined, we deduce

$$\int_{\Omega_X} b''(\theta(x))\,\mathrm{d}x \int_{\Omega_X} \{b''(\theta(y))\}^{-1} \mathrm{d}y \le (m_{b''} + M_{b''})^2 / (4 m_{b''} M_{b''}) |\Omega_X|^2.$$

This applied to (46) gives the lower bound.

Proof of Proposition 4. Define $\mathbf{H}=\mathrm{diag}\{1,h,\ldots,h^p\}$. From (44), we have that

$$H_i=(nh)^{-1}\boldsymbol{e}_1^{\mathrm{T}}\{n^{-1}\mathbf{H}^{-1}\hat{S}_n(X_i)/b''(\hat{\theta}(X_i))\mathbf{H}^{-1}\}^{-1}\boldsymbol{e}_1K(0), \tag{47}$$

where $\hat{S}_n(x)=\sum_{j=1}^n\mathbf{x}_j(x)\mathbf{x}_j(x)^{\mathrm{T}}K_h(X_j-x)b''(\mathbf{x}_j(x)^{\mathrm{T}}\hat{\boldsymbol{\beta}}(x))$. By Taylor's expansion and the assumptions on b'' and f_X, it follows that uniformly in $x\in\Omega_X$, $n^{-1}\mathbf{H}^{-1}\hat{S}_n(x)/b''(\hat{\theta}(x))\mathbf{H}^{-1}=f_X(x)S+o_P(1)$. Combining this expression with (47), it can be shown that

$$\sum_{i=1}^n H_i=\sum_{i=1}^n\frac{1}{nhf_X(X_i)}\boldsymbol{e}_1^{\mathrm{T}}S^{-1}\boldsymbol{e}_1K(0)\{1+o_P(1)\}=\frac{\mathcal{K}(0)}{nh}\sum_{i=1}^n\frac{1}{f_X(X_i)}\{1+o_P(1)\},$$

which will finish the proof.

Lemma 3

Assume that the kernel function K is non-negative, symmetric and uni-modal. Then for $i=1,\ldots,n$, $\mathcal{S}_i$ decreases in $h>0$ for which $\mathcal{S}_i$ is well-defined.

Proof. Consider the matrices $A_i(h)=\mathbf{X}(X_i)^{\mathrm{T}}\mathrm{diag}\{K(|X_j-X_i|/h)\}_{j=1}^n\mathbf{X}(X_i)$, $i=1,\ldots,n$. If K is non-negative and uni-modal, then $0<h_1<h_2$ implies that $A_i(h_1)\leq A_i(h_2)$ or, equivalently, $\{A_i(h_1)\}^{-1}\geq\{A_i(h_2)\}^{-1}$. We complete the proof by noting $\mathcal{S}_i=\boldsymbol{e}_1^{\mathrm{T}}\{A_i(h)\}^{-1}\boldsymbol{e}_1K(0)$, since K is symmetric.

Proof of Proposition 5. The one-step estimate of $\hat{\boldsymbol{\beta}}^{-i}(x)$, starting from $\boldsymbol{\beta}_0=\hat{\boldsymbol{\beta}}(x)$, is given by

$$\hat{\boldsymbol{\beta}}(x)+4\{\mathbf{X}(x)^{\mathrm{T}}\mathbf{V}_i(0)\mathbf{K}(x)\mathbf{X}(x)\}^{-1}\{\mathbf{X}(x)^{\mathrm{T}}\mathbf{V}_i(0)\mathbf{W}(x;\hat{\boldsymbol{\beta}}(x))\hat{\mathbf{z}}(x)\}, \tag{48}$$

i.e.

$$\hat{\boldsymbol{\beta}}(x)-4\{S_n(x)-K_h(X_i-x)\mathbf{x}_i(x)\mathbf{x}_i(x)^{\mathrm{T}}\}^{-1}\mathbf{x}_i(x)K_h(X_i-x)\{Y_i-\hat{m}_i(x)\}.$$

Again, using the Sherman–Morrison–Woodbury formula,

$$\begin{aligned}\{S_n(x)-K_h(X_i-x)\mathbf{x}_i(x)\mathbf{x}_i(x)^{\mathrm{T}}\}^{-1}=\{S_n(x)\}^{-1}&+K_h(X_i-x)\{S_n(x)\}^{-1}\mathbf{x}_i(x)\mathbf{x}_i(x)^{\mathrm{T}}\\&\times\{S_n(x)\}^{-1}/\tau_i(x),\end{aligned}$$

where $\tau_i(x)=1-K_h(X_i-x)\mathbf{x}_i(x)^{\mathrm{T}}\{S_n(x)\}^{-1}\mathbf{x}_i(x)$ and thus

$$\{S_n(x)-K_h(X_i-x)\mathbf{x}_i(x)\mathbf{x}_i(x)^{\mathrm{T}}\}^{-1}\mathbf{x}_i(x)=\{S_n(x)\}^{-1}\mathbf{x}_i(x)/\tau_i(x),$$

by which (48) becomes $\hat{\boldsymbol{\beta}}(x)-4\{S_n(x)\}^{-1}\mathbf{x}_i(x)K_h(X_i-x)\{Y_i-\hat{m}_i(x)\}/\tau_i(x)$. This expression approximates $\hat{\boldsymbol{\beta}}^{-i}(x)$ and thus leads to (19). Applying (19), we have

$$\theta_i^{-i}-\hat{\theta}_i=\boldsymbol{e}_1^{\mathrm{T}}\{\hat{\boldsymbol{\beta}}^{-i}(X_i)-\hat{\boldsymbol{\beta}}(X_i)\}\doteq-4\boldsymbol{e}_1^{\mathrm{T}}\{S_n(X_i)\}^{-1}\boldsymbol{e}_1K_h(0)(Y_i-\hat{m}_i)/(1-\mathcal{S}_i),$$

which leads to (20). Proofs of (21) and (22) are similar to those of proposition 2.

Proofs of Propositions 6 and 7. The technical arguments are similar to those of propositions 1 and 2 and are omitted.

Proof of Proposition 8. Recalling the definition of H_i^* in section 5.2, we have that for $\boldsymbol{\tau}_i = \mathbf{e}_1 \otimes \mathtt{X}_i$,

$$H_i^* = (nh)^{-1} \boldsymbol{\tau}_i^{\mathrm{T}} \{ n^{-1} (\mathbf{H} \otimes \mathbf{I}_d)^{-1} \hat{S}_n^*(U_i) (\mathbf{H} \otimes \mathbf{I}_d)^{-1} \}^{-1} \boldsymbol{\tau}_i \times K(0) b''(\hat{\theta}(U_i, \mathtt{X}_i)), \tag{49}$$

where

$$\hat{S}_n^*(u) = \sum_{j=1}^{n} \left[\{ \mathbf{u}_j(u) \mathbf{u}_j(u)^{\mathrm{T}} \} \otimes (\mathtt{X}_j \mathtt{X}_j^{\mathrm{T}}) \right] K_h(U_j - u) b'' \left(\{ \mathbf{u}_j(u) \otimes \mathtt{X}_j \}^{\mathrm{T}} \hat{\boldsymbol{\beta}}(u) \right).$$

It can be shown that uniformly in $u \in \Omega_U$,

$$\begin{aligned} & n^{-1} (\mathbf{H} \otimes \mathbf{I}_d)^{-1} \hat{S}_n^*(u) (\mathbf{H} \otimes \mathbf{I}_d)^{-1} \\ &= n^{-1} \sum_{j=1}^{n} \left[\{ \mathbf{H}^{-1} \mathbf{u}_j(u) \mathbf{u}_j(u)^{\mathrm{T}} \mathbf{H}^{-1} \} \otimes (\mathtt{X}_j \mathtt{X}_j^{\mathrm{T}}) \right] K_h(U_j - u) b''(\theta(u, \mathtt{X}_j)) + o_P(1) \\ &= f_U(u) [S \otimes E\{ b''(\theta(u, \mathtt{X})) \mathtt{X} \mathtt{X}^{\mathrm{T}} \mid U = u \}] + o_P(1) = f_U(u) \{ S \otimes \Gamma(u) \} + o_P(1). \end{aligned}$$

This expression applied to (49) further implies that

$$\sum_{i=1}^{n} H_i^* = \sum_{i=1}^{n} \frac{\mathcal{K}(0)}{nhf_U(U_i)} \left\{ \mathtt{X}_i^{\mathrm{T}} \Gamma(U_i)^{-1} \mathtt{X}_i b''(\theta(U_i, \mathtt{X}_i)) \right\} \{ 1 + o_P(1) \}. \tag{50}$$

For (50), a direct calculation yields $E\{ \mathtt{X}^{\mathrm{T}} \Gamma(U)^{-1} \mathtt{X} b''(\theta(U, \mathtt{X})) / f_U(U) \} = d |\Omega_U|$.