arXiv:math/0506028v1 [math.ST] 2 Jun 2005

Assessing Prediction Error of Nonparametric Regression and Classification under Bregman Divergence

Jianqing Fan
Department of Operations Research
and Financial Engineering
Princeton University
Princeton, NJ 08544
jqfan@princeton.edu

Chunming Zhang
Department of Statistics
University of Wisconsin
Madison, WI 53706-1685
cmzhang@stat.wisc.edu

February 1, 2008

Abstract

Prediction error is critical to assessing the performance of statistical methods and selecting statistical models. We propose the cross-validation and approximated cross-validation methods for estimating prediction error under a broad q-class of Bregman divergence for error measures which embeds nearly all of the commonly used loss functions in regression, classification procedures and machine learning literature. The approximated cross-validation formulas are analytically derived, which facilitate fast estimation of prediction error under the Bregman divergence. We then study a data-driven optimal bandwidth selector for the local-likelihood estimation that minimizes the overall prediction error or equivalently the covariance penalty. It is shown that the covariance penalty and cross-validation 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 it is as simple and stable as the local least-squares regression estimation. The algorithm monotonically enhances the target local-likelihood and converges. The idea and methods are extended to the generalized varying-coefficient models and semiparametric models.

Key words and Phrases: Cross-validation; Exponential family; Generalized varying-coefficient model; Local likelihood; Loss function; Prediction error.

Short title: Assessing Prediction Error under Bregman Divergence

1 Introduction

Assessing prediction error lies at the heart of statistical model selection and forecasting. Depending on the needs of statistical learning and model selection, the quadratic loss is not always appropriate. In binary classification, for example, the misclassification error rate is more suitable. The corresponding loss, however, does not differentiate the predictive powers of two procedures which forecast the class label being 1 with probabilities 60% and 90% respectively. When the true class label is 1, the second procedure is more accurate, whereas when the true class label is 0, the first procedure is more accurate. The quantification of the predicative accuracy requires an appropriate introduction of loss functions. An example of this is the negative Bernoulli log-likelihood loss function. Other important margin-based loss functions have been introduced for binary classification in the machine learning literature (Hastie, Tibshirani and Friedman 2001). Hence, it is important to assess the prediction error under a broader 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 exponential family of distributions, 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 nonparametric regression models, the residual-based cross-validation (CV) is a useful data-driven method for the automatic smoothing (Wong 1983; Rice 1984; Härdle, Hall, and Marron 1992; Hall and Johnstone 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 cross-validated 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 the large sample sizes. The challenge then arises from efficient computation of the estimated prediction error based on the cross-validation.

The computational problem is resolved via the newly developed approximate formulas for the cross-validated 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 formula (Golub and Van Loan 1996, p. 50). It is a natural extension of the cross-validation formula for least-squares regression estimates, and is applicable to both parametric and nonparametric models.

The applications of estimated prediction error pervade almost every facet of statistical model selection and forecasting. To be more specific, we focus on the local-likelihood estimation in varying coefficient models for response variables having distributions in the exponential family. Typical examples include fitting the Bernoulli distributed binary responses, and the Poisson distributed count responses, among many other non-normal outcomes. As a flexible nonparametric model-fitting technique, the local-likelihood method possesses nice sampling properties. For details, see, for example, Tibshirani and Hastie (1987), Staniswalis (1989), Severini and Staniswalis (1994), and Fan, Heckman, and Wand (1995). An important issue in application is the choice of smoothing parameter. Currently, most of the existing methods deal with the Gaussian type of responses; clearly there is a lack of methodology for non-Gaussian responses. The approximate cross-validation 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 that provides useful insights into the local-likelihood modeling 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, Farmen, and Gijbels (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 and 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 over-smoothing.

The paper is organized as follows. Section 2 addresses the issue of estimating prediction error. Sections 3 develops computationally feasible versions of the cross-validated estimates of the prediction error. Section 4 proposes a new bandwidth selection method for binary responses, based on the LB method and the cross-validated estimates of the prediction error. Sections 5–6 extend our results to generalized varying-coefficient models and semiparametric models respectively. Section 7 presents simulation evaluations and Section 8 analyzes real data. Technical conditions and proofs are relegated to the Appendix.

2 Estimating Prediction Error

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

$$f_{Y|\mathtt{X}}(y;\theta(\mathtt{x})) = \exp[\{y\theta(\mathtt{x}) - b(\theta(\mathtt{x}))\}/a(\psi) + c(y,\psi)], \tag{2.1}$$

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

$$m(\mathtt{x}) \equiv E(Y|\mathtt{X}=\mathtt{x}) = b'(\theta(\mathtt{x})), \quad \text{and} \quad \sigma^2(\mathtt{x}) \equiv \mathrm{var}(Y|\mathtt{X}=\mathtt{x}) = a(\psi)b''(\theta(\mathtt{x})).$$

See Nelder and Wedderburn (1972) and McCullagh and 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 $\theta(\mathtt{x})$. The results can easily be generalized to other link functions.

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, $Q(Y,\widehat{m}) = \mathrm{I}\{Y \neq \mathrm{I}(\widehat{m} > .5)\}$, where $\mathrm{I}(\cdot)$ is an indicator function and $\widehat{m}$ is an estimator. However, this measure does not differentiate the predictions $\widehat{m} = .6$ and $\widehat{m} = .9$ when $Y = 1$ or 0. In the case that $Y = 1$, $\widehat{m} = .9$ gives a better prediction than $\widehat{m} = .6$. The negative Bernoulli log-likelihood, $Q(Y,\widehat{m}) = -Y\ln(\widehat{m}) - (1-Y)\ln(1-\widehat{m})$, captures this. Other loss functions possessing similar properties include the hinge loss function, $Q(Y,\widehat{m}) = \max\{1-(2Y-1)\mathrm{sign}(\widehat{m}-.5), 0\}$, in the support vector machine and the exponential loss function, $Q(Y,\widehat{m}) = \exp\{-(Y-.5)\ln(\widehat{m}/(1-\widehat{m}))\}$, popularly used in AdaBoost. These four loss functions, shown in Figure 1, belong to the margin-based loss functions written in the form, $V(Y^*F)$, for $Y^* = 2Y-1$ and some function F.

[*Put Figure 1 about here*]

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

$$Q(Y,\widehat{m}) = q(\widehat{m}) + q'(\widehat{m})(Y-\widehat{m}) - q(Y). \tag{2.2}$$

A graphical illustration of Q associated with q is displayed in Figure 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).

[*Put Figure 2 about here*]

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 $Q_1(Y, \widehat{m}) = (Y - \widehat{m})^2$.
- For the exponential family (2.1), the function $q_2(m) = 2\{b(\theta) - m\theta\}$ with $b'(\theta) = m$ results in the deviance loss,

$$Q_2(Y, \widehat{m}) = 2\{Y(\widetilde{\theta} - \widehat{\theta}) - b(\widetilde{\theta}) + b(\widehat{\theta})\}, \tag{2.3}$$

  where $b'(\widetilde{\theta}) = Y$ and $b'(\widehat{\theta}) = \widehat{m}$.

- For a binary response variable Y, the function $q(m) = \min\{m, (1-m)\}$ gives the misclassification loss; the function $q(m) = .5\min\{m, (1-m)\}$ results in the hinge loss; the function $q_3(m) = 2\{m(1-m)\}^{1/2}$ yields the exponential loss,

$$Q_3(Y, \widehat{m}) = \exp\{-(Y - .5)\ln(\widehat{m}/(1-\widehat{m}))\}. \tag{2.4}$$

2.2 Prediction Error under Bregman Divergence

Let $m_i = m(\mathtt{X}_i)$ and $\widehat{m}_i$ be its estimate based on independent observations $\{(\mathtt{X}_i, Y_i)\}_{i=1}^n$. Set

$$\mathrm{err}_i = Q(Y_i, \widehat{m}_i) \quad \text{and} \quad \mathrm{Err}_i = E_o\{Q(Y_i^o, \widehat{m}_i)\},$$

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 difference $O_i = \mathrm{Err}_i - \mathrm{err}_i$, called *optimism*.

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

$$\widehat{\mathrm{Err}}_i = \mathrm{err}_i + 2\,\widehat{\mathrm{cov}}_i, \tag{2.5}$$

where $\widehat{\mathrm{cov}}_i$ is an estimator of the covariance penalty, $\mathrm{cov}(\widehat{\lambda}_i, Y_i)$. 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 nonparametric estimation, we write $\widehat{m}_{i,h}$ and $\widehat{\lambda}_{i,h}$ to stress their dependence on a smoothing parameter h. According to (2.5), the total prediction error is given by

$$\widehat{\mathrm{Err}}(h) = \sum_{i=1}^{n} Q(Y_i, \widehat{m}_{i,h}) + \sum_{i=1}^{n} 2\,\widehat{\mathrm{cov}}(\widehat{\lambda}_{i,h}, Y_i). \tag{2.6}$$

Estimating the covariance-penalty terms in (2.6) is not trivial. Three approaches are potentially applicable to the estimation of the covariance penalty: model-based bootstrap developed in Efron (2004), the data-perturbation (DP) method proposed by Shen, Huang and Ye (2004), and model-free cross-validation. All three are computationally intensive in the context of local-likelihood estimation (the DP method also needs to select a perturbation size). In contrast, the third method allows us to develop approximate formulas to significantly gain computational expedience. For this reason, we focus on the cross-validation method.

The cross-validated estimation of Err_i is $Q(Y_i, \widehat{m}_{i,h}^{-i})$, where the superscript $-i$ indicates the deletion of the ith data point $(\mathtt{X}_i, Y_i)$ in the fitting process. This yields the cross-validated estimate of the total prediction error by

$$\widehat{\mathrm{Err}}^{\mathrm{CV}}(h) = \sum_{i=1}^{n} Q(Y_i, \widehat{m}_{i,h}^{-i}). \tag{2.7}$$

Naive computation of $\{\widehat{m}_{i,h}^{-i}\}$ 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 (2.7) with (2.6), the covariance penalty in (2.7) is estimated by

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

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

2.4 Asymptotic Prediction Error

To gain insight on $\widehat{\mathrm{Err}}(h)$, we appeal to the asymptotic theory. Simple algebra shows that $\mathrm{Err}_i(h) = Q(m_i, \widehat{m}_{i,h}) + E\{Q(Y_i, m_i)\}$. By Taylor's expansion and (2.2),

$$Q(m_i, \widehat{m}_{i,h}) \doteq -(\widehat{m}_{i,h} - m_i)^2 q''(\widehat{m}_{i,h})/2.$$

Hence,

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

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

$$\mathrm{MPEC}(h) = -2^{-1} \int E[\{\widehat{m}_h(x) - m(x)\}^2 | \mathcal{X}] q''(m(x)) f_X(x) dx, \tag{2.8}$$

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

$$\mathrm{MISE}(h) = \int E[\{\widehat{m}_h(x) - m(x)\}^2 | \mathcal{X}] \{b''(\theta(x))\}^{-2} f_X(x) dx, \tag{2.9}$$

recalling that

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

Expression (2.8) reveals that asymptotically, different loss functions automatically introduce different weighting schemes in (2.8). 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, we have

$$\mathrm{MPEC}_1(h) = \int E[\{\widehat{m}_h(x) - m(x)\}^2 | \mathcal{X}] f_X(x) dx.$$

For the q_2-function producing the deviance-loss, we have

$$\mathrm{MPEC}_2(h) = \int E[\{\widehat{m}_h(x) - m(x)\}^2 | \mathcal{X}] \{b''(\theta(x))\}^{-1} f_X(x) dx.$$

For the q_3-function inducing the exponential-loss for the binary responses, we have

$$\mathrm{MPEC}_3(h) = \int E[\{\widehat{m}_h(x) - m(x)\}^2 | \mathcal{X}] \frac{f_X(x)}{4[m(x)\{1 - m(x)\}]^{3/2}} dx.$$

3 Approximate Cross-Validation

This section aims at deriving the approximate and empirical versions of (2.7) for the local maximum likelihood estimator. We focus on the univariate problem 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, the Taylor expansion implies that

$$\theta(X_j) \doteq \mathbf{x}_j(x)^T \boldsymbol{\beta}(x),$$

in which $\mathbf{x}_j(x) = (1, (X_j - x), \ldots, (X_j - x)^p)^T$ and $\boldsymbol{\beta}(x) = (\beta_0(x), \ldots, \beta_p(x))^T$. Based on the independent observations, the local parameters can be estimated by maximizing the local log-likelihood,

$$\ell(\boldsymbol{\beta}; x) \equiv \sum_{j=1}^{n} l(\mathbf{x}_j(x)^T \boldsymbol{\beta}; Y_j) K_h(X_j - x), \tag{3.1}$$

in which $l(\cdot; y) = \ln\{f_{Y|X}(y; \cdot)\}$ denotes the conditional log-likelihood function, $K_h(\cdot) = K(\cdot/h)/h$ for a kernel function K and h is a bandwidth. Let $\widehat{\boldsymbol{\beta}}(x) = (\widehat{\beta}_0(x), \ldots, \widehat{\beta}_p(x))^T$ be the local maximum likelihood estimator. Then, the local MLEs of $\theta(x)$ and $m(x)$ are given by $\widehat{\theta}(x) = \widehat{\beta}_0(x)$ and $\widehat{m}(x) = b'(\widehat{\theta}(x))$, 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, $\widehat{\boldsymbol{\beta}}^{-i}(x)$, $\widehat{\theta}^{-i}(x)$, and $\widehat{m}^{-i}(x)$, respectively.

3.1 Weighted Local-Likelihood

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

$$\ell_{i,\delta}(\boldsymbol{\beta}; x) = \sum_{j=1}^{n} \delta_{ij}\, l(\mathbf{x}_j(x)^T \boldsymbol{\beta}; Y_j) K_h(X_j - x), \tag{3.2}$$

with the weight $\delta_{ii} = \delta$ and the rest weights $\delta_{ij} = 1$. Let $\widehat{\boldsymbol{\beta}}_{i,\delta}(x)$ be the maximizer. Note that when $\delta = 1$, this estimator is the local maximum likelihood estimator and when $\delta = 0$, it is the leave-one-out estimator.

The weighted local MLE is usually found via the Newton-Raphson iteration,

$$\boldsymbol{\beta}_L = \boldsymbol{\beta}_{L-1} - \{\nabla^2 \ell_{i,\delta}(\boldsymbol{\beta}_{L-1}; x)\}^{-1} \nabla \ell_{i,\delta}(\boldsymbol{\beta}_{L-1}; x), \quad L = 1, 2, \ldots, \tag{3.3}$$

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.) When the initial estimator $\boldsymbol{\beta}_0$ is good enough, the one-step ($L = 1$) estimator is as efficient as the fully iterated estimator (Fan and Chen 1999).

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 $\widehat{\boldsymbol{\beta}}(x)$ as the initial value. Namely, $\widehat{\boldsymbol{\beta}}^{-i}(x)$ is approximated by (3.3) with $\boldsymbol{\beta}_0 = \widehat{\boldsymbol{\beta}}(x)$ and $\delta = 0$. With this idea and other explicit formulas and approximation, we can derive the approximate leave-one-out formulas.

3.2 Leave-One-Out Formulas

Let $\mathbf{X}(x) = (\mathbf{x}_1(x), \ldots, \mathbf{x}_n(x))^T$,

$$\mathbf{W}(x; \boldsymbol{\beta}) = \text{diag}\{K_h(X_j - x) b''(\mathbf{x}_j(x)^T \boldsymbol{\beta})\}, \tag{3.4}$$

and $S_n(x;\boldsymbol{\beta}) = \mathbf{X}(x)^T\mathbf{W}(x;\boldsymbol{\beta})\mathbf{X}(x)$. Define

$$\mathcal{H}(x;\boldsymbol{\beta}) = \{\mathbf{W}(x;\boldsymbol{\beta})\}^{1/2}\mathbf{X}(x)\{S_n(x;\boldsymbol{\beta})\}^{-1}\mathbf{X}(x)^T\{\mathbf{W}(x;\boldsymbol{\beta})\}^{1/2}. \tag{3.5}$$

This projection matrix is an extension of the hat matrix in the multiple regression and will be useful for computing the leave-one-out estimator. Let $\mathcal{H}_{ii}(x;\boldsymbol{\beta})$ be its i-th diagonal element and $H_i = \mathcal{H}_{ii}(X_i;\widehat{\boldsymbol{\beta}}(X_i))$. Then, our main results can be summarized as follows.

Proposition 1 *Assume condition* (*A2*) *in the Appendix. Then for* $i = 1, \ldots, n$,

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}^{-i}(x) - \widehat{\boldsymbol{\beta}}(x) &\doteq -\frac{\{S_n(x;\widehat{\boldsymbol{\beta}}(x))\}^{-1}\mathbf{x}_i(x)K_h(X_i - x)\{Y_i - b'(\mathbf{x}_i(x)^T\widehat{\boldsymbol{\beta}}(x))\}}{1 - \mathcal{H}_{ii}(x;\widehat{\boldsymbol{\beta}}(x))}, && (3.6)\\
\widehat{\theta}_i^{-i} - \widehat{\theta}_i &\doteq -\frac{H_i}{1 - H_i}(Y_i - \widehat{m}_i)/b''(\widehat{\theta}_i), && (3.7)\\
\widehat{m}_i^{-i} - \widehat{m}_i &\doteq -\frac{H_i}{1 - H_i}(Y_i - \widehat{m}_i). && (3.8)
\end{aligned}$$

Note that the approximation becomes exact when the loss function is the quadratic loss. In fact, Zhang (2003) shows an explicit delete-one-out formula,

$$\widehat{m}_i^{-i} = \widehat{m}_i - \frac{H_i}{1 - H_i}(Y_i - \widehat{m}_i).$$

In addition, (3.6)–(3.8) hold for $h \to \infty$, namely, the parametric maximum likelihood estimator. Even the results for this specific case appear new. Furthermore, the results can easily be extended to the estimator that minimizes the local Bregman divergence, replacing $l(\mathbf{x}_j(x)^T\boldsymbol{\beta}; Y_j)$ in (3.1) by $Q(Y_j, g^{-1}(\mathbf{x}_j(x)^T\boldsymbol{\beta}))$.

Using Proposition 1, we can derive a simplified formula for computing the cross-validated estimate of the overall prediction error.

Proposition 2 *Assume conditions* (*A1*) *and* (*A2*) *in the Appendix. Then*

$$\widehat{\mathrm{Err}}^{\mathrm{CV}} \doteq \sum_{i=1}^{n}\Big[Q(Y_i, \widehat{m}_i) + 2^{-1}q''(\widehat{m}_i)(Y_i - \widehat{m}_i)^2\big\{1 - 1/(1 - H_i)^2\big\}\Big]. \tag{3.9}$$

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

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

For this particular loss function, the approximation is actually exact. For the function q_2 leading to the deviance loss Q_2 defined in (2.3), we have

$$\sum_{i=1}^{n} Q_2(Y_i, \widehat{m}_i^{-i}) \doteq \sum_{i=1}^{n}\left[Q_2(Y_i, \widehat{m}_i) - \frac{(Y_i - \widehat{m}_i)^2}{b''(\widehat{\theta}_i)}\big\{1 - 1/(1 - H_i)^2\big\}\right]. \tag{3.10}$$

For the exponential loss defined in (2.4) for binary classification, we observe

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

3.3 Two Theoretical Issues

Two theoretical issues are particularly interesting. The first one concerns the asymptotic convergence of $\widehat{h}_{\text{ACV}}$, the minimizer of the right hand side of (3.9). Following a suitable modification to the result of Altman and MacGibbon (1998), the ratio $\widehat{h}_{\text{ACV}}/h_{\text{AMPEC}}$ converges in probability to 1, where h_{AMPEC} is the minimizer of the asymptotic form of MPEC(h) defined in (2.8).

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):

$$h_{\text{AMPEC}}(q) = C_p(K)\left[\frac{a(\psi)\int b''(\theta(x))q''(m(x))dx}{\int\{\theta^{(p+1)}(x)\}^2\{b''(\theta(x))\}^2 q''(m(x))f_X(x)dx}\right]^{1/(2p+3)} n^{-1/(2p+3)}, \tag{3.11}$$

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

$$h_{\text{AMPEC}}(q_2) = C_p(K)\left[\frac{a(\psi)|\Omega_X|}{\int\{\theta^{(p+1)}(x)\}^2 b''(\theta(x))f_X(x)dx}\right]^{1/(2p+3)} n^{-1/(2p+3)}, \tag{3.12}$$

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

$$h_{\text{AMISE}} = C_p(K)\left[\frac{a(\psi)\int\{b''(\theta(x))\}^{-1}dx}{\int\{\theta^{(p+1)}(x)\}^2 f_X(x)dx}\right]^{1/(2p+3)} n^{-1/(2p+3)}, \tag{3.13}$$

determined by minimizing the asymptotic MISE(h) of $\widehat{\theta}$ defined in (2.9), with an exception of the Gaussian family.

[*Put Table 1 about here*]

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

Definition 1 *Two functions F and G are called similarly ordered if $\{F(x_1) - F(x_2)\}\{G(x_1) - G(x_2)\} \ge 0$ for all x_1 in the domain of F and all x_2 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 $F(x) = \{\theta^{(p+1)}(x)\}^2 b''(\theta(x))f_X(x)$ and $G(x) = \{b''(\theta(x))\}^{-1}$. Assume that p is odd.*

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

(*b*) *Assume that $b''(\theta(x))$ is bounded away from zero and infinity. 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*

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

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

3.4 Empirical Cross-Validation

The approximate CV criterion (3.9) can be further simplified. To this end, we first approximate the "degrees of freedom" $\sum_{i=1}^n H_i$ (Hastie and Tibshirani 1990). To facilitate presentations, we now define the "equivalent kernel" $\mathcal{K}(t)$ induced by the local-polynomial fitting as the first element of the vector $S^{-1}(1, t, \ldots, t^p)^T K(t)$, in which the matrix $S = (\mu_{i+j-2})_{1\le i,j\le p+1}$ with $\mu_k = \int t^k K(t)\,dt$. See Ruppert and Wand (1994).

Proposition 4 *Assume conditions* (**A**) *and* (**B**) *in the Appendix. If $n \to \infty$, $h \to 0$, and $nh \to \infty$, we have*

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

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 nonparametric fitting.

[*Put Table 2 about here*]

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:

$$\sum_{i=1}^n H_i \doteq (p + 1 - \mathsf{a}) + \mathcal{C}n/(n-1)\mathcal{K}(0)|\Omega_X|/h. \tag{3.14}$$

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, $K(t) = .75(1-t^2)_+$. Interestingly, our simulation studies in Section 7 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 $\mathsf{a} = .7$ and $\mathcal{C} = 1.09$ provide better approximations.

We propose the empirical version of the estimated total prediction error by replacing H_i in (3.9) with their empirical average, $\overline{H}_E = (p+1-\mathtt{a})/n + \mathcal{C}/(n-1)\mathcal{K}(0)|\Omega_X|/h$, leading to the empirical cross-validation (ECV) criterion,

$$\widehat{\text{Err}}^{\text{ECV}}(h) = \sum_{i=1}^{n} \Big[Q(Y_i, \widehat{m}_i) + 2^{-1} q''(\widehat{m}_i)(Y_i - \widehat{m}_i)^2 \big\{1 - 1/(1-\overline{H}_E)^2\big\}\Big]. \tag{3.15}$$

This avoids calculating the smoother matrix $\mathcal{H}$. Yet, it turns out to work reasonably well in practice. A data-driven optimal bandwidth selector, $\widehat{h}_{\text{ECV}}$, can be obtained by minimizing (3.15).

4 Nonparametric Logistic Regression

Nonparametric 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 and Schmitt (1988, p. 751) reported that the local-likelihood method suffers from a substantial proportion of "incalculable estimates". Fan and 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 $\mathcal{L}$, it replaces the Hessian matrix $\nabla^2\mathcal{L}(\boldsymbol{\beta})$ in the Newton-Raphson algorithm by a negative definite matrix $\mathbf{B}$, such that

$$\nabla^2\mathcal{L}(\boldsymbol{\beta}) \geq \mathbf{B}, \quad \text{for all } \boldsymbol{\beta}.$$

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 $\boldsymbol{\beta}_{\text{LB}} = \boldsymbol{\beta}_0 - \mathbf{B}^{-1}\nabla\mathcal{L}(\boldsymbol{\beta}_0)$, leads to a monotonic increase of $\mathcal{L}(\cdot)$, that is, $\mathcal{L}(\boldsymbol{\beta}_{\text{LB}}) - \mathcal{L}(\boldsymbol{\beta}_0) \geq -2^{-1}\nabla\mathcal{L}(\boldsymbol{\beta}_0)^T\mathbf{B}^{-1}\nabla\mathcal{L}(\boldsymbol{\beta}_0) \geq 0$.*

For the local logistic regression, $\nabla^2\ell(\boldsymbol{\beta}; x) = -\mathbf{X}(x)^T\mathbf{W}(x;\boldsymbol{\beta})\mathbf{X}(x)$. Since $\mathbf{0} \leq \mathbf{W}(x;\boldsymbol{\beta}) \leq 4^{-1}\mathbf{K}(x)$, 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,

$$\mathbf{B}(x) = -4^{-1}\mathbf{X}(x)^T\mathbf{K}(x)\mathbf{X}(x), \tag{4.1}$$

and the LB-adjusted Newton-Raphson iteration for computing $\widehat{\boldsymbol{\beta}}(x)$ becomes

$$\boldsymbol{\beta}_L = \boldsymbol{\beta}_{L-1} - \{\mathbf{B}(x)\}^{-1}\mathbf{X}(x)^T\mathbf{K}(x)\mathbf{r}(x;\boldsymbol{\beta}_{L-1}), \quad L = 1, 2, \ldots, \tag{4.2}$$

where $\mathbf{r}(x;\boldsymbol{\beta}) = (Y_1 - m_1(x;\boldsymbol{\beta}), \ldots, Y_n - m_n(x;\boldsymbol{\beta}))^T$ with $m_j(x;\boldsymbol{\beta}) = 1/[1+\exp\{-\mathbf{x}_j(x)^T\boldsymbol{\beta}\}]$.

The LB method offers a number of advantages to compute $\widehat{\boldsymbol{\beta}}(x)$. Firstly, 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. Secondly, 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. Thirdly, 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 $\widehat{\boldsymbol{\beta}}(x)$.

4.2 A Hybrid Bandwidth Selection Method

For binary responses, our simulation studies show that the bandwidth choice minimizing (3.9) or its empirical version (3.15) tends to produce over-smoothed estimates. Such a problem was also encountered in Aragaki and Altman (1997) and Fan, Farmen and Gijbels (1998, Table 1). Because of the importance of binary responses in nonparametric regression and classification, a new bandwidth selector that specifically accommodates the binary responses is needed.

We first employ the LB scheme (4.2) to derive a new one-step estimate of $\widehat{\boldsymbol{\beta}}^{-i}(x)$, starting from $\widehat{\boldsymbol{\beta}}(x)$. Define $S_n(x) = \mathbf{X}(x)^T\mathbf{K}(x)\mathbf{X}(x)$ and $\mathcal{S}_i = \boldsymbol{e}_1^T\{S_n(X_i)\}^{-1}\boldsymbol{e}_1 K_h(0)$, where $\boldsymbol{e}_1 = (1,0,\ldots,0)^T$. The resulting leave-one-out formulas and the cross-validated estimates of the total prediction error are displayed in Proposition 5.

Proposition 5 *Assume conditions* (A1) *and* (A2) *in the Appendix. Then for the local-likelihood MLE in the Bernoulli family,*

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}^{-i}_{LB}(x) - \widehat{\boldsymbol{\beta}}(x) &\doteq -\frac{4\{S_n(x)\}^{-1}\mathbf{x}_i(x)K_h(X_i-x)\{Y_i - b'(\mathbf{x}_i(x)^T\widehat{\boldsymbol{\beta}}(x))\}}{1-K_h(X_i-x)\mathbf{x}_i(x)^T\{S_n(x)\}^{-1}\mathbf{x}_i(x)}, && (4.3)\\
\widehat{\theta}_i^{-i} - \widehat{\theta}_i &\doteq -\frac{4\mathcal{S}_i}{1-\mathcal{S}_i}(Y_i - \widehat{m}_i), && (4.4)\\
\widehat{m}_i^{-i} - \widehat{m}_i &\doteq -\frac{4b''(\widehat{\theta}_i)\mathcal{S}_i}{1-\mathcal{S}_i}(Y_i - \widehat{m}_i), && (4.5)\\
\widehat{\mathrm{Err}}^{\mathrm{CV}} &\doteq \sum_{i=1}^{n}\Big[Q(Y_i,\widehat{m}_i) + 2^{-1}q''(\widehat{m}_i)(Y_i-\widehat{m}_i)^2\big[1-\{1+4b''(\widehat{\theta}_i)\mathcal{S}_i/(1-\mathcal{S}_i)\}^2\big]\Big]. && (4.6)
\end{aligned}$$

Direct use of a bandwidth selector that minimizes (4.6) tends to under-smooth the binary responses. To better appreciate this, note that the second term in (4.6) is approximately

$$-\,q''(\widehat{m}_i)(Y_i-\widehat{m}_i)^2\{4b''(\widehat{\theta}_i)\}\mathcal{S}_i, \tag{4.7}$$

and the second in (3.9) can be approximated as

$$-\,q''(\widehat{m}_i)(Y_i-\widehat{m}_i)^2 H_i. \tag{4.8}$$

As demonstrated in Lemma 3 in the Appendix, $\mathcal{S}_i$ decreases with h and $H_i \doteq \mathcal{S}_i$. Since $0 \le 4b''(\widehat{\theta}_i) \le 1$ for the Bernoulli family, (4.7) 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

$$\sum_{i=1}^{n} \Big[Q(Y_i, \widehat{m}_i) + 2^{-1} q''(\widehat{m}_i)(Y_i - \widehat{m}_i)^2 \big[1 - \{1 + 2b''(\widehat{\theta}_i)\mathcal{S}_i/(1-\mathcal{S}_i) + 2^{-1}H_i/(1-H_i)\}^2\big]\Big], \qquad (4.9)$$

which averages terms in (4.7) and (4.8) to mitigate the oversmoothing problem of criterion (3.9). This new criterion has some desirable properties: $2b''(\widehat{\theta}_i)\mathcal{S}_i/(1-\mathcal{S}_i) + 2^{-1}H_i/(1-H_i)$ is bounded below by $2^{-1}H_i/(1-H_i)$, thus guarding against under-smoothing, and is bounded above by $\{\mathcal{S}_i/(1-\mathcal{S}_i) + H_i/(1-H_i)\}/2$, thus diminishing the influence of over-smoothing. An empirical cross-validation criterion is to replace $\mathcal{S}_i$ and H_i in (4.9) by their empirical averages, which are (3.14) 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 $\mathtt{X} = (X_1, \ldots, X_d)^T$. 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 $\mathtt{x} = (x_1, \ldots, x_d)^T$. 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

$$g(m(u, \mathtt{x})) = \theta(u, \mathtt{x}) = \sum_{k=1}^{d} a_k(u) x_k = \mathtt{x}^T \boldsymbol{A}(u). \qquad (5.1)$$

for a vector $\boldsymbol{A}(u) = (a_1(u), \ldots, a_d(u))^T$ 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 $a_k(\cdot)$'s are $(p+1)$-times continuously differentiable at a fitting point u. Put $\boldsymbol{A}^{(\ell)}(u) = (a_1^{(\ell)}(u), \ldots, a_d^{(\ell)}(u))^T$. Denote by $\boldsymbol{\beta}(u) = (\boldsymbol{A}(u)^T, \boldsymbol{A}^{(1)}(u)^T, \ldots, \boldsymbol{A}^{(p)}(u)^T/p!)^T$ the $d(p+1)$ by 1 vector of coefficient functions along with their derivatives, $\mathbf{u}_j(u) = (1, (U_j - u), \ldots, (U_j - u)^p)^T$, and $\mathbf{I}_d$ a $d \times d$ identity matrix. For observed covariates U_j close to the point u,

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

in which the symbol $\otimes$ denotes the Kronecker product, and thus from (5.1),

$$\theta(U_j, \mathtt{X}_j) \doteq \{\mathbf{u}_j(u) \otimes \mathtt{X}_j\}^T \boldsymbol{\beta}(u).$$

The local-likelihood MLE $\widehat{\boldsymbol{\beta}}(u)$ maximizes the local log-likelihood function:

$$\ell(\boldsymbol{\beta}; u) = \sum_{j=1}^{n} l(\{\mathbf{u}_j(u) \otimes \mathtt{X}_j\}^T \boldsymbol{\beta}; Y_j) K_h(U_j - u). \tag{5.2}$$

The first d entries of $\widehat{\boldsymbol{\beta}}(u)$ supply the local MLEs $\widehat{\boldsymbol{A}}(u)$ of $\boldsymbol{A}(u)$, and the local MLEs of $\theta(u, \mathtt{x})$ and $m(u, \mathtt{x})$ are given by $\widehat{\theta}(u, \mathtt{x}) = \mathtt{x}^T \widehat{\boldsymbol{A}}(u)$ and $\widehat{m}(u, \mathtt{x}) = b'(\widehat{\theta}(u, \mathtt{x}))$, 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, $\widehat{\boldsymbol{\beta}}^{-i}(u)$, $\widehat{\theta}^{-i}(u, \mathtt{x})$, and $\widehat{m}^{-i}(u, \mathtt{x})$ respectively.

5.1 Leave-One-Out Formulas

To derive the leave-one-out formulas in the case of multivariate covariates, we need some additional notations. Let $\mathbf{X}^*(u) = (\mathbf{u}_1(u) \otimes \mathtt{X}_1, \ldots, \mathbf{u}_n(u) \otimes \mathtt{X}_n)^T$, $\mathbf{W}^*(u; \boldsymbol{\beta}) = \operatorname{diag}\{K_h(U_j - u) b''(\{\mathbf{u}_j(u) \otimes \mathtt{X}_j\}^T \boldsymbol{\beta})\}$, and $S_n^*(u; \boldsymbol{\beta}) = \mathbf{X}^*(u)^T \mathbf{W}^*(u; \boldsymbol{\beta}) \mathbf{X}^*(u)$. Define a projection matrix as

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

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

Proposition 6 *Assume condition* (*A*2) *in the Appendix. Then for* $i = 1, \ldots, n$,

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

Proposition 7 *Assume conditions* (*A*1) *and* (*A*2) *in the Appendix. Then*

$$\widehat{\mathrm{Err}}^{\mathrm{CV}} \doteq \sum_{i=1}^{n} \Big[Q(Y_i, \widehat{m}_i) + 2^{-1} q''(\widehat{m}_i)(Y_i - \widehat{m}_i)^2 \big\{1 - 1/(1 - H_i^*)^2\big\} \Big]. \tag{5.3}$$

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. If* $n \to \infty$, $h \to 0$, *and* $nh \to \infty$, *we have*

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

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:

$$\sum_{i=1}^{n} H_i^* \doteq d\{(p+1-\mathsf{a}) + \mathcal{C}n/(n-d)\mathcal{K}(0)|\Omega_U|/h\}. \tag{5.4}$$

The empirical version of the estimated total prediction error is to replace H_i^* in (5.3) by $d\{(p+1-\mathsf{a})/n + \mathcal{C}/(n-d)\mathcal{K}(0)|\Omega_U|/h\}$. Call $\widehat{\mathrm{Err}}^{\mathrm{ECV}}(h)$ the empirical version of the cross-validation criterion. Compared with the bandwidth selector in Cai, Fan and Li (2000), the $\widehat{\mathrm{Err}}^{\mathrm{ECV}}(h)$-minimizing bandwidth selector, $\widehat{h}_{\mathrm{ECV}}$, is much easier to obtain.

5.3 Binary Responses

For Bernoulli responses, the LB method in Section 4 continues to be applicable for obtaining $\widehat{\boldsymbol{\beta}}(u)$ and $\widehat{\boldsymbol{\beta}}^{-i}(u)$. For the local logistic regression, $\nabla^2\ell(\boldsymbol{\beta}; u)$ has a lower bound, $\mathbf{B}(u) = -4^{-1}\mathbf{X}^*(u)^T\mathbf{K}^*(u)\mathbf{X}^*(u)$, where $\mathbf{K}^*(u) = \mathrm{diag}\{K_h(U_j - u)\}$. Similar to (4.2), the LB-adjusted NR iteration for $\widehat{\boldsymbol{\beta}}(u)$ proceeds as follows,

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

where $\mathbf{r}^*(u; \boldsymbol{\beta}) = (Y_1 - m_1^*(u; \boldsymbol{\beta}), \ldots, Y_n - m_n^*(u; \boldsymbol{\beta}))^T$ with $m_j^*(u; \boldsymbol{\beta}) = 1/(1 + \exp[-\{\mathbf{u}_j(u) \otimes \mathtt{X}_j\}^T\boldsymbol{\beta}])$. The leave-one-out formulas and the cross-validated estimates of the prediction error are similar to those in Proposition 5, with $S_n(x)$ replaced by $S_n^*(u) = \mathbf{X}^*(u)^T\mathbf{K}^*(u)\mathbf{X}^*(u)$ and $\mathcal{S}_i$ by $\mathcal{S}_i^* = (\boldsymbol{e}_1 \otimes \mathtt{X}_i)^T\{S_n^*(U_i)\}^{-1}(\boldsymbol{e}_1 \otimes \mathtt{X}_i)K_h(0)$. In the spirit of (4.9), the hybrid selection criterion for bandwidth is

$$\sum_{i=1}^{n} \Big[Q(Y_i, \widehat{m}_i) + 2^{-1}q''(\widehat{m}_i)(Y_i - \widehat{m}_i)^2\big[1 - \{1 + 2b''(\widehat{\theta}_i)\mathcal{S}_i^*/(1 - \mathcal{S}_i^*) + 2^{-1}H_i^*/(1 - H_i^*)\}^2\big]\Big]. \tag{5.5}$$

The ECV criterion can be obtained similarly via replacing $\mathcal{S}_i^*$ and H_i^* by their empirical averages, which are (5.4) divided by n.

6 Extension to Semiparametric Model

A further extension of model (5.1) is to allow part of the covariates independent of U, resulting in

$$\theta(u, \mathtt{x}, \mathtt{z}) = \mathtt{x}^T\boldsymbol{A}(u) + \mathtt{z}^T\boldsymbol{\beta}, \tag{6.1}$$

in which $(u, \mathtt{x}, \mathtt{z})$ lists values of all covariates $(U, \mathtt{X}, \mathtt{Z})$. This model keeps the flexibility that $\mathtt{Z}$ does not interact with U. The challenge is how to choose a bandwidth to efficiently estimate both the parametric and nonparametric components. In this section, we propose a two-stage bandwidth selection method which is applicable to general semiparametric models. This is an attempt to answer an important question raised by Bickel and Kwon (2001) on the bandwidth selection for semiparametric models.

The parameters in model (6.1) can be estimated via the profile likelihood method. For each given $\boldsymbol{\beta}$, applying the local-likelihood method with a bandwidth h, we obtain an estimate $\widehat{\boldsymbol{A}}(u; \boldsymbol{\beta}, h)$, depending on h and $\boldsymbol{\beta}$. Substituting it into (6.1), we obtain a pseudo parametric model:

$$\theta(u, \mathtt{x}, \mathtt{z}) \doteq \mathtt{x}^T \widehat{\boldsymbol{A}}(u; \boldsymbol{\beta}, h) + \mathtt{z}^T \boldsymbol{\beta}. \tag{6.2}$$

Regarding (6.2) as a parametric model with parameter $\boldsymbol{\beta}$, by using the maximum likelihood estimation method, we obtain the profile likelihood estimators $\widehat{\boldsymbol{\beta}}(h)$ and $\widehat{\boldsymbol{A}}(u; \widehat{\boldsymbol{\beta}}(h), h)$. This estimator is semiparametrically efficient.

We now outline a two-stage method for choosing the bandwidth. The idea is also applicable to other semiparametric problems. Starting from a very small bandwidth h_0, we obtain a semiparametric estimator $\widehat{\boldsymbol{\beta}}(h_0)$ (see a justification below). To avoid difficulty of implementation, the nearest type of bandwidth can be used. This estimator is usually root-n consistent. Thus, $\boldsymbol{\beta}$ in model (6.1) can be regarded as known and hence model (6.1) becomes a varying coefficient model. Applying a bandwidth selection method for varying coefficient models, such as the approximate cross-validation method in the previous section, we obtain a bandwidth $\widehat{h}$. Using this $\widehat{h}$, we obtain the profile likelihood estimator $\widehat{\boldsymbol{\beta}}(\widehat{h})$ and $\widehat{\boldsymbol{A}}(u; \widehat{\boldsymbol{\beta}}(\widehat{h}), \widehat{h})$. This is a two-stage method for choosing the bandwidth for a semiparametric model.

To illustrate the idea, we specifically consider the partially linear model:

$$Y_i = a(U_i) + \mathtt{Z}_i^T \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \ldots, n. \tag{6.3}$$

Assume that the data have already been sorted according to U_i. Let h_0 be the nearest two-point bandwidth so that

$$\widehat{a}(U_i; \boldsymbol{\beta}, h_0) = 2^{-1}(Y_i - \mathtt{Z}_i^T \boldsymbol{\beta} + Y_{i-1} - \mathtt{Z}_{i-1}^T \boldsymbol{\beta}).$$

Substituting this into (6.3) and rearranging the equation, we have

$$Y_i - Y_{i-1} \doteq (\mathtt{Z}_i - \mathtt{Z}_{i-1})^T \boldsymbol{\beta} + 2\varepsilon_i. \tag{6.4}$$

Applying the least-squares method, we obtain an estimator $\widehat{\boldsymbol{\beta}}(h_0)$.

To see why such a crude parametric estimator $\widehat{\boldsymbol{\beta}}(h_0)$ is root-n consistent, let us take the difference of (6.3). Under the mild conditions, $a(U_i) - a(U_{i-1}) = O_P(n^{-1})$. Hence, the difference of (6.3) yields

$$Y_i - Y_{i-1} = (\mathtt{Z}_i - \mathtt{Z}_{i-1})^T \boldsymbol{\beta} + \varepsilon_i - \varepsilon_{i-1} + O_P(n^{-1}).$$

Hence, the least-squares estimator $\widehat{\boldsymbol{\beta}}$, which is the same as $\widehat{\boldsymbol{\beta}}(h_0)$, is root-n consistent. See Yachew (1997) for a proof. This example shows that even if a very crude bandwidth h_0 is chosen, the parametric component $\widehat{\boldsymbol{\beta}}(h_0)$ is still root-n consistent.

The two-stage bandwidth selector is to apply a data-driven bandwidth selector to the following univariate nonparametric regression problem:

$$Y_i = a(U_i) + \mathsf{Z}_i^T \widehat{\boldsymbol{\beta}}(h_0) + \varepsilon_i,$$

and to use the resulting bandwidth for the original semiparametric problem. Such an idea was implemented in Fan and Huang (2005). They reported that the resulting semiparametric and nonparametric estimators are efficient.

7 Simulations

The purpose of the simulations is three-fold: to assess the accuracy of the empirical formulas (3.14) and (5.4), the performance of the bandwidth selector $\widehat{h}_{\mathrm{ECV}}$, and the behavior of the proposed bandwidth selector for local-likelihood estimation. 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 $n = 400$. A complete copy of Matlab codes is available upon request.

7.1 Generalized Nonparametric Regression Model

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

[*Put Figure 3 about here*]

Poisson regression: We first consider the response variable Y which, conditional on $X = x$, follows a Poisson distribution with parameter $\lambda(x)$:

$$P(Y = y|X = x) = \{\lambda(x)\}^y \exp\{-\lambda(x)\}/y!, \quad y = 0, 1, \ldots.$$

The function $\theta(x) = \ln\{\lambda(x)\}$ is given in the test examples,

$$\begin{aligned}
&\text{Example 1:} && \theta(x) = 3.5[\exp\{-(4x-1)^2\} + \exp\{-(4x-3)^2\}] - 1.5,\\
&\text{Example 2:} && \theta(x) = \sin\{2(4x-2)\} + 1.0,\\
&\text{Example 3:} && \theta(x) = 2 - .5(4x-2)^2.
\end{aligned}$$

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 (3.14). To see the performance of $\widehat{h}_{\text{ECV}}$, Figure 3(b) gives boxplots of the relative error, $\{\widehat{h}_{\text{ECV}} - h_{\text{AMPEC}}(q_2)\}/h_{\text{AMPEC}}(q_2)$ and $\{\widehat{h}_{\text{ECV}} - h_{\text{AMISE}}\}/h_{\text{AMISE}}$, based on 100 random samples; refer to Table 1 for values of $h_{\text{AMPEC}}(q_2)$ and h_{AMISE}. We observe that $\widehat{h}_{\text{ECV}}$ is closer to $h_{\text{AMPEC}}(q_2)$ than to h_{AMISE}; this is in accordance with the discussion of Section 3.3. In Figure 3(c), we simulate another 100 random samples and for each set obtain $\widehat{h}_{\text{ECV}}$ to estimate $\theta(x)$. We present the estimated curves from three typical samples. The typical samples are selected in such a way that their ASE values, in which ASE $= n^{-1}\sum_{j=1}^n\{\widehat{\theta}(X_j) - \theta(X_j)\}^2$, 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 cross-validated deviance does not exhibit undersmoothing in the local-likelihood regression estimation. In Figure 3, similar results are also displayed in the middle panel [Figures 3(d)–(f)] for Example 2, and in the bottom panel [Figures 3(g)–(i)] for Example 3.

[*Put Figure 4 about here*]

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

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

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

7.2 Generalized Varying-Coefficient Model

We consider examples of the generalized varying-coefficient model (5.1). We take $h_{\min} = 3\,h_0$ for Poisson regression, where $h_0 = \max[5/n, \max_{2\le j\le n}\{U_{(j)} - U_{(j-1)}\}]$, and $h_{\min} = .1$ for logistic regression.

[*Put Figure 5 about here*]

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 below:

$$\text{Example 1:}\qquad d=2,\ a_1(u)=5.5+.1\ \exp(2u-1),\ a_2(u)=.8u(1-u),$$
$$\text{Example 2:}\qquad d=3,\ a_1(u)=5.5+.1\ \exp(2u-1),\ a_2(u)=.8u(1-u),\ a_3(u)=.2\sin^2(2\pi u).$$

We assume that U is a uniform random variable on the interval $[0,1]$ and is independent of $\mathtt{X}=(X_1,X_2,X_3)^T$, 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 Figure 5, plot (a) reveals that the actual degrees of freedom are well captured by the empirical formula (5.4). To evaluate the performance of $\widehat{h}_{\mathrm{ECV}}$, we generate 100 random samples of size 400. Figure 5(b)–(c) plot the estimated curves of $a_1(u)$ and $a_2(u)$ from three typical samples. The typical samples are selected so that their ASE values, in which $\mathrm{ASE}=n^{-1}\sum_{j=1}^{n}\{\widehat{\theta}(U_j,\mathtt{X}_j)-\theta(U_j,\mathtt{X}_j)\}^2$, correspond to the 25th (dotted line), 50th (dashed line), and 75th (dash-dotted line) percentiles in the 100 replications. The corresponding results for Example 2 are given in Figure 5(a')–(d'). These plots provide convincing evidences that $\widehat{h}_{\mathrm{ECV}}$, when applied to recovering multiple smooth curves simultaneously, performs competitively well with that to fitting a single smooth curve.

[*Put Figure 6 about here*]

Logistic regression: We now consider the varying-coefficient logistic regression model for Bernoulli responses, in which varying-coefficient functions in $\mathrm{logit}\{P(Y=1|U=u,\mathtt{X}=\mathtt{x})\}$ are specified as

$$\text{Example 1:}\qquad d=2,\ a_1(u)=1.3\{\exp(2u-1)-1.5\},\ a_2(u)=1.2\{8u(1-u)-1\},$$
$$\text{Example 2:}\qquad d=3,\ a_1(u)=\exp(2u-1)-1.5,\ a_2(u)=.8\{8u(1-u)-1\},\ a_3(u)=.9\{2\sin(\pi u)-1\}.$$

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 Figure 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, through the use of the hybrid bandwidth selector $\widehat{h}_{\mathrm{ECV}}$, captures the major features of the model structure with reasonably good details.

8 Real Data Applications

In this section, we apply the hybrid bandwidth selection method for binary responses to analyze an employee dataset (Example 11.3 of Albright, et al. 1999) of the Fifth National Bank of Springfield, based on year 1995 data. 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.

The data set was carefully analyzed by Fan and Peng (2004). After adjusting for covariates such as age, years of work experience, and education level, they did not find stark evidence of discrimination. However, they pointed out that 77.29% (R^2) of the salary variation can be explained by the job grade alone and the question becomes whether it was harder for females to be promoted, after adjusting for confounding variables such as age and years of work experience. They did not carry out the analysis further.

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 regression model,

$$\operatorname{logit}\{P(Y=1|U=u, X_1=x_1, X_2=x_2)\} = a_0(u) + a_1(u)x_1 + a_2(u)x_2, \tag{8.1}$$

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 $95-$YrHired$+$YrsPrior). Following Fan and Peng (2004), outliers have been deleted, with the remaining 199 data for analysis. For this medium-lengthed data, use of the bandwidth selector $\widehat{h}_{\mathrm{ACV}}$ which minimizes (5.5) seems to be more natural than $\widehat{h}_{\mathrm{ECV}}$.

[*Put Figure 7 about here*]

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 viewed from the scatter plot in Figure 7(a). Such high correlation may cause some identifiability problem, thus in model (8.1), 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 cross-validation function (for Gaussian family), illustrated in Figure 7(b).

After the de-correlation step, we now refit model (8.1). The bottom panel of Figure 7 depicts the estimated varying coefficient functions of $a_0(u)$, $a_1(u)$ and $a_2(u)$, plus/minus the pointwise 1.96

times of their estimated standard error. The selected bandwidth is 16.9 [see Figure 7(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 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.

[*Put Figure 8 about here*]

To see whether the choice of smoothing variable U makes a difference in drawing the above conclusion, we fit again model (8.1) 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, Figure 8 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 Figure 7, 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 (8.1), 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.

References

Albright, S. C., Winston, W. L., and Zappe, C. J. (1999), *Data Analysis and Decision Making with Microsoft Excel*, Duxbury Press, Pacific Grove, California.

Altman, N., and MacGibbon, B. (1998), "Consistent Bandwidth Selection for Kernel Binary Regression," *J. Statist. Plann. Inference*, 70, 121–137.

Aragaki, A., and Altman, N. S. (1997), "Local Polynomial Regression for Binary Response," in *Computer Science and Statistics: Proceedings of the 29th Symposium on the Interface.*

Bickel, P.J. and Kwon, J. (2001), "Inference for Semiparametric Models: Some Current Frontiers" (with discussion), *Statistica Sinica*, 11, 863–960.

Böhning, D., and Lindsay, B. G. (1988), "Monotonicity of Quadratic Approximation Algorithms," *Ann. Inst. Statist. Math.*, 40, 641–663.

Brègman, 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. and Li, R. (2000), "Efficient Estimation and Inferences for Varying-Coefficient Models," *Jour. Ameri. Statist. Assoc.*, 95, 888–902.

Efron, B. (2004), "The Estimation of Prediction Error: Covariance Penalties and Cross-Validation" (with discussion), *J. Amer. Statist. Assoc.*, 99, 619–642.

Fan, J., and Chen, J. (1999), "One-Step Local Quasi-Likelihood Estimation," *J. R. Statist. Soc.*, Ser. B, 61, 927–943.

Fan, J., Heckman, N., and 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., and Gijbels, I. (1998), "Local Maximum Likelihood Estimation and Inference," *J. R. Statist. Soc.*, Ser. B, 60, 591–608.

Fan, J. and Huang, T. (2005), "Profile Likelihood Inferences on Semiparametric Varying-Coefficient Partially Linear Models," *Bernoulli*, to appear.

Fan, J. and Peng, H. (2004), "On Non-Concave Penalized Likelihood with Diverging Number of Parameters," *The Annals of Statistics*, 32, 928–961.

Golub, G. H., and Van Loan, C. F. (1996), *Matrix Computations* (Third edition), Baltimore, MD: Johns Hopkins University Press.

Hall, P. and 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., and Pólya, G. (1988), *Inequalities* (Second edition), Cambridge, England: Cambridge University Press.

Härdle, W., Hall, P., and Marron, J. S. (1992), "Regression Smoothing Parameters That Are Not Far From Their Optimum," *J. Amer. Statist. Assoc.*, 87, 227–233.

Hastie, T. J., and Tibshirani, R. J. (1990), *Generalized Additive Models*, London: Chapman and Hall.

Hastie, T. J., Tibshirani, R. and Friedman, J. (2001), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer-Verlag, New York.

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models* (2nd ed.), London: Chapman and Hall.

Mitrinović, D. S., Pečarić, J. E., and Fink, A. M. (1993), *Classical and New Inequalities in Analysis*, Kluwer Academic Publishers Group, Dordrecht.

Müller, H.-G., and Schmitt, T. (1988), "Kernel and Probit Estimates in Quantal Bioassay," *J. Amer. Statist. Assoc.*, 83, 750–759.

Nelder, J. A., and Wedderburn, R. W. M. (1972), "Generalized Linear Models," *J. R. Statist. Soc.*, Ser. A, 135, 370–384.

Pregibon, D. (1981), "Logistic Regression Diagnostics," *Ann. Statist.*, 9, 705–24.

Rice, J. (1984), "Bandwidth Choice for Nonparametric Regression," *Ann. Statist.*, 20, 712–736.

Ruppert, D. and Wand, M.P. (1994), "Multivariate Weighted Least Squares Regression," *Ann. Statist.*, 22, 1346–1370.

Severini, T. A., and Staniswalis, J. G. (1994), "Quasi-Likelihood Estimation in Semiparametric Models," *J. Amer. Statist. Assoc.*, 89, 501–511.

Shen, X., Huang, H.-C. and Ye, J. (2004), "Adaptive Model Selection and Assessment for Exponential-Family Models," *Technometrics*, 46, 306-317.

Staniswalis, J. G. (1989), "The Kernel Estimate of a Regression Function in Likelihood-Based Models," *J. Amer. Statist. Assoc.*, 84, 276–283.

Tibshirani, R., and 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.

Yatchew, A. (1997), "An Elementary Estimator for the Partial Linear Model," *Economics Letters*, 57, 135–143.

Zhang, C. M. (2003), "Calibrating the Degrees of Freedom for Automatic Data Smoothing and Effective Curve Checking," *J. Amer. Statist. Assoc.*, 98, 609–628.

Appendix: Proofs of Main Results

We first impose some technical conditions. They 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.

Condition (B):

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

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

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 the continuous $(p+1)$-th derivative in Ω_U.

(C3) For the use of canonical links, the matrix $\Gamma(u) = E\{b''(\theta(u, \mathtt{X}))\mathtt{X}\mathtt{X}^T | U = u\}$ is positive definite for each $u \in \Omega_U$ and is Lipschitz continuous.

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

Before proving the main results of the paper, we need the following lemma.

Lemma 2 *Define* $\mathbf{V}_i(\delta) = \mathrm{diag}\{\delta_{i1}, \ldots, \delta_{in}\}$. *For* $\ell_{i,\delta}(\boldsymbol{\beta}; x)$ *defined in (3.2),*

$$\begin{aligned} \nabla \ell_{i,\delta}(\boldsymbol{\beta}; x) &= \mathbf{X}(x)^T \mathbf{V}_i(\delta) \mathbf{W}(x; \boldsymbol{\beta}) \mathbf{z}(x; \boldsymbol{\beta}) / a(\psi), \qquad (A.1) \\ \nabla^2 \ell_{i,\delta}(\boldsymbol{\beta}; x) &= -\mathbf{X}(x)^T \mathbf{V}_i(\delta) \mathbf{W}(x; \boldsymbol{\beta}) \mathbf{X}(x) / a(\psi), \qquad (A.2) \end{aligned}$$

in which

$$\begin{aligned} \mathbf{X}(x)^T \mathbf{V}_i(\delta) \mathbf{W}(x; \boldsymbol{\beta}) \mathbf{z}(x; \boldsymbol{\beta}) &= \mathbf{X}(x)^T \mathbf{W}(x; \boldsymbol{\beta}) \mathbf{z}(x; \boldsymbol{\beta}) - (1-\delta) \mathbf{x}_i(x) K_h(X_i - x)\{Y_i - m_i(x; \boldsymbol{\beta})\}, (A.3) \\ \mathbf{X}(x)^T \mathbf{V}_i(\delta) \mathbf{W}(x; \boldsymbol{\beta}) \mathbf{X}(x) &= \mathbf{X}(x)^T \mathbf{W}(x; \boldsymbol{\beta}) \mathbf{X}(x) - (1-\delta) w_i(x; \boldsymbol{\beta}) \mathbf{x}_i(x) \mathbf{x}_i(x)^T. \qquad (A.4) \end{aligned}$$

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

$$\begin{aligned} \nabla \ell_{i,\delta}(\boldsymbol{\beta}; x) &= \frac{\partial \boldsymbol{\theta}(x; \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \frac{\partial \ell_{i,\delta}(\boldsymbol{\beta}; x)}{\partial \boldsymbol{\theta}(x; \boldsymbol{\beta})} = \mathbf{X}(x)^T \frac{\partial \ell_{i,\delta}(\boldsymbol{\beta}; x)}{\partial \boldsymbol{\theta}(x; \boldsymbol{\beta})}, \qquad (A.5) \\ \nabla^2 \ell_{i,\delta}(\boldsymbol{\beta}; x) &= \mathbf{X}(x)^T \frac{\partial^2 \ell_{i,\delta}(\boldsymbol{\beta}; x)}{\partial \boldsymbol{\theta}(x; \boldsymbol{\beta}) \partial \boldsymbol{\theta}(x; \boldsymbol{\beta})^T} \mathbf{X}(x). \qquad (A.6) \end{aligned}$$

Since

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

it is easy to check that

$$\frac{\partial\ell_{i,\delta}(\boldsymbol{\beta};x)}{\partial\theta_j(x;\boldsymbol{\beta})}=\delta_{ij}\{Y_j-b'(\theta_j(x;\boldsymbol{\beta}))\}K_h(X_j-x)/a(\psi). \tag{A.7}$$

This combined with (A.5) leads to (A.1).

Following (A.7), we see that

$$\frac{\partial^2\ell_{i,\delta}(\boldsymbol{\beta};x)}{\partial\theta_j(x;\boldsymbol{\beta})\partial\theta_k(x;\boldsymbol{\beta})}=0,\ j\neq k,\quad\text{and}\quad\frac{\partial^2\ell_{i,\delta}(\boldsymbol{\beta};x)}{\{\partial\theta_j(x;\boldsymbol{\beta})\}^2}=-\delta_{ij}b''(\theta_j(x;\boldsymbol{\beta}))K_h(X_j-x)/a(\psi). \tag{A.8}$$

This along with (A.6) indicates (A.2).

(A.3) and (A.4) can be obtained by decomposing the identity matrix $\mathbf{I}$ into $\mathbf{V}_i(\delta)$ and $\mathbf{I}-\mathbf{V}_i(\delta)$.

Proof of Proposition 1

From (A.1) and (A.2), (3.3) can be rewritten as

$$\boldsymbol{\beta}_L=\boldsymbol{\beta}_{L-1}+\{\mathbf{X}(x)^T\mathbf{V}_i(\delta)\mathbf{W}(x;\boldsymbol{\beta}_{L-1})\mathbf{X}(x)\}^{-1}\{\mathbf{X}(x)^T\mathbf{V}_i(\delta)\mathbf{W}(x;\boldsymbol{\beta}_{L-1})\mathbf{z}(x;\boldsymbol{\beta}_{L-1})\}. \tag{A.9}$$

Setting $\delta=0$ in (A.9), the one-step estimate of $\widehat{\boldsymbol{\beta}}^{-i}(x)$, which starts from $\boldsymbol{\beta}_0=\widehat{\boldsymbol{\beta}}(x)$, is given by

$$\widehat{\boldsymbol{\beta}}(x)+\{\mathbf{X}(x)^T\mathbf{V}_i(0)\mathbf{W}(x;\widehat{\boldsymbol{\beta}}(x))\mathbf{X}(x)\}^{-1}\{\mathbf{X}(x)^T\mathbf{V}_i(0)\mathbf{W}(x;\widehat{\boldsymbol{\beta}}(x))\widehat{\mathbf{z}}(x)\}. \tag{A.10}$$

Using the definition of $\widehat{\boldsymbol{\beta}}(x)$ (satisfying $\nabla\ell_{i,\delta}(\boldsymbol{\beta};x)=0$ with $\delta=1$), along with (A.3) and (A.4), the above one-step estimate of $\widehat{\boldsymbol{\beta}}^{-i}(x)$ equals

$$\widehat{\boldsymbol{\beta}}(x)-\{\widehat{S}_n(x)-\widehat{w}_i(x)\mathbf{x}_i(x)\mathbf{x}_i(x)^T\}^{-1}\mathbf{x}_i(x)K_h(X_i-x)\{Y_i-\widehat{m}_i(x)\}. \tag{A.11}$$

According to the Sherman-Morrison-Woodbury formula (Golub and Van Loan 1996, p. 50),

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

Thus

$$\begin{aligned}\{\widehat{S}_n(x)-\widehat{w}_i(x)\mathbf{x}_i(x)\mathbf{x}_i(x)^T\}^{-1}\mathbf{x}_i(x)&=\{\widehat{S}_n(x)\}^{-1}\mathbf{x}_i(x)\\&+\frac{\widehat{w}_i(x)\{\widehat{S}_n(x)\}^{-1}\mathbf{x}_i(x)\mathbf{x}_i(x)^T\{\widehat{S}_n(x)\}^{-1}\mathbf{x}_i(x)}{1-\widehat{w}_i(x)\mathbf{x}_i(x)^T\{\widehat{S}_n(x)\}^{-1}\mathbf{x}_i(x)}=\frac{\{\widehat{S}_n(x)\}^{-1}\mathbf{x}_i(x)}{1-\mathcal{H}_{ii}(x;\widehat{\boldsymbol{\beta}}(x))},\end{aligned}$$

by which (A.11) becomes

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

This expression approximates $\widehat{\boldsymbol{\beta}}^{-i}(x)$ and thus leads to (3.6).

Note that $\widehat{\theta}_i = \widehat{\theta}_i(X_i)$, $\widehat{m}_i = \widehat{m}_i(X_i)$, $\widehat{\theta}_i^{-i} = \widehat{\theta}_i^{-i}(X_i)$, $\widehat{m}_i^{-i} = \widehat{m}_i^{-i}(X_i)$, and

$$H_i = \boldsymbol{e}_1^T\{\widehat{S}_n(X_i)\}^{-1}\boldsymbol{e}_1 K_h(0) b''(\widehat{\theta}_i). \tag{A.12}$$

Applying (3.6), we have

$$\begin{aligned}\widehat{\theta}_i^{-i} - \widehat{\theta}_i &= \boldsymbol{e}_1^T\{\widehat{\boldsymbol{\beta}}^{-i}(X_i) - \widehat{\boldsymbol{\beta}}(X_i)\} \\ &\doteq -\frac{\boldsymbol{e}_1^T\{\widehat{S}_n(X_i)\}^{-1}\boldsymbol{e}_1 K_h(0)(Y_i - \widehat{m}_i)}{1 - \mathcal{H}_{ii}(X_i; \widehat{\boldsymbol{\beta}}(X_i))} = -\frac{H_i}{1 - H_i}(Y_i - \widehat{m}_i)/b''(\widehat{\theta}_i),\end{aligned}$$

leading to (3.7). This, together with a first-order Taylor's expansion and the continuity of b'', yields

$$\widehat{m}_i^{-i} - \widehat{m}_i = b'(\widehat{\theta}_i^{-i}) - b'(\widehat{\theta}_i) \doteq (\widehat{\theta}_i^{-i} - \widehat{\theta}_i) b''(\widehat{\theta}_i),$$

and thus (3.8).

Proof of Proposition 2

By a first-order Taylor expansion, we have that

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

These, applied to an identity given in the Lemma of Efron (2004, Section 4),

$$Q(Y_i, \widehat{m}_i^{-i}) - Q(Y_i, \widehat{m}_i) = 2(\widehat{\lambda}_i - \widehat{\lambda}_i^{-i})(Y_i - \widehat{m}_i^{-i}) - Q(\widehat{m}_i^{-i}, \widehat{m}_i),$$

lead to

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

Summing over i and using (3.8) and (2.7), we complete the proof.

Proof of Proposition 3

From (3.12) and (3.13), we see that

$$\frac{h_{\text{AMPEC}}(q_2)}{h_{\text{AMISE}}} = \left[\frac{|\Omega_X| \int_{\Omega_X} F(x)G(x)dx}{\int_{\Omega_X} F(x)dx \int_{\Omega_X} G(x)dx}\right]^{1/(2p+3)}. \tag{A.13}$$

For part (a), it suffices to consider oppositely ordered F and G. In this case, by the Tchebychef's inequality (Hardy, Littlewood, and Pólya, 1988, p. 43 and 168), we obtain

$$|\Omega_X| \int_{\Omega_X} F(x)G(x)dx \le \int_{\Omega_X} F(x)dx \int_{\Omega_X} G(x)dx.$$

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

$$\frac{|\Omega_X| \int_{\Omega_X} F(x)G(x)dx}{\int_{\Omega_X} F(x)dx \int_{\Omega_X} G(x)dx} \leq 1,$$

which along with (A.13) indicates that $h_{\text{AMPEC}}(q_2) \leq h_{\text{AMISE}}$.

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 (A.13) becomes

$$\frac{h_{\text{AMPEC}}(q_2)}{h_{\text{AMISE}}} = \left[\frac{|\Omega_X|^2}{\int_{\Omega_X} b''(\theta(x))dx \int_{\Omega_X} \{b''(\theta(x))\}^{-1} dx}\right]^{1/(2p+3)}. \tag{A.14}$$

Incorporating the Grüss integral inequality (Mitrinović, Pečarić, and Fink 1993),

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

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

$$\int_{\Omega_X} b''(\theta(x))dx \int_{\Omega_X} \{b''(\theta(x))\}^{-1} dx \leq \frac{(m_{b''} + M_{b''})^2}{4 m_{b''} M_{b''}} |\Omega_X|^2.$$

This applied to (A.14) gives the lower bound.

Proof of Proposition 4

Define $\mathbf{H} = \text{diag}\{1, h, \ldots, h^p\}$. From (A.12), we have

$$\begin{aligned} H_i &= \boldsymbol{e}_1^T \{\widehat{S}_n(X_i)/b''(\widehat{\theta}(X_i))\}^{-1} \boldsymbol{e}_1 K_h(0) \\ &= n^{-1} \boldsymbol{e}_1^T \mathbf{H}^{-1} \{n^{-1}\mathbf{H}^{-1}\widehat{S}_n(X_i)/b''(\widehat{\theta}(X_i))\mathbf{H}^{-1}\}^{-1} \mathbf{H}^{-1} \boldsymbol{e}_1 K_h(0) \\ &= (nh)^{-1} \boldsymbol{e}_1^T \{n^{-1}\mathbf{H}^{-1}\widehat{S}_n(X_i)/b''(\widehat{\theta}(X_i))\mathbf{H}^{-1}\}^{-1} \boldsymbol{e}_1 K(0), \end{aligned} \tag{A.15}$$

where $\widehat{S}_n(x) = \sum_{j=1}^n \mathbf{x}_j(x)\mathbf{x}_j(x)^T K_h(X_j - x) b''(\mathbf{x}_j(x)^T \widehat{\boldsymbol{\beta}}(x))$. By Taylor's expansion and the continuity assumptions on b'' and f_X, it follows that for $x \in \Omega_X$,

$$n^{-1}\mathbf{H}^{-1}\widehat{S}_n(x)/b''(\widehat{\theta}(x))\mathbf{H}^{-1} = f_X(x)S + o_P(1).$$

Combining this expression with (A.15), it can be shown that

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

which will finish the proof.

Lemma 3 *Assume that the kernel function K is non-negative, symmetric and uni-modal. Then for $i = 1, \ldots, n$, $\mathcal{S}_i$ is a decreasing function of $h > 0$ for which $\mathcal{S}_i$ is well-defined.*

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

Proof of Proposition 5

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

$$\widehat{\boldsymbol{\beta}}(x) + 4\{\mathbf{X}(x)^T \mathbf{V}_i(0) \mathbf{K}(x) \mathbf{X}(x)\}^{-1} \{\mathbf{X}(x)^T \mathbf{V}_i(0) \mathbf{W}(x; \widehat{\boldsymbol{\beta}}(x)) \widehat{\mathbf{z}}(x)\}, \tag{A.16}$$

i.e., $\widehat{\boldsymbol{\beta}}(x) - 4\{S_n(x) - K_h(X_i - x)\mathbf{x}_i(x)\mathbf{x}_i(x)^T\}^{-1} \mathbf{x}_i(x) K_h(X_i - x)\{Y_i - \widehat{m}_i(x)\}$. Again, using the Sherman-Morrison-Woodbury formula (Golub and Van Loan 1996, p. 50),

$$\{S_n(x) - K_h(X_i - x)\mathbf{x}_i(x)\mathbf{x}_i(x)^T\}^{-1} = \{S_n(x)\}^{-1} + \frac{K_h(X_i - x)\{S_n(x)\}^{-1}\mathbf{x}_i(x)\mathbf{x}_i(x)^T\{S_n(x)\}^{-1}}{1 - K_h(X_i - x)\mathbf{x}_i(x)^T\{S_n(x)\}^{-1}\mathbf{x}_i(x)},$$

and thus

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

by which (A.16) becomes

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

This expression approximates $\widehat{\boldsymbol{\beta}}^{-i}(x)$ and thus leads to (4.3).

Applying (4.3), we have

$$\begin{aligned} \widehat{\theta}_i^{-i} - \widehat{\theta}_i &= \boldsymbol{e}_1^T\{\widehat{\boldsymbol{\beta}}^{-i}(X_i) - \widehat{\boldsymbol{\beta}}(X_i)\} \\ &\doteq -\frac{4\boldsymbol{e}_1^T\{S_n(X_i)\}^{-1}\boldsymbol{e}_1 K_h(0)(Y_i - \widehat{m}_i)}{1 - \mathcal{S}_i} = -\frac{4\mathcal{S}_i}{1 - \mathcal{S}_i}(Y_i - \widehat{m}_i), \end{aligned}$$

leading to (4.4). Proofs of (4.5) and (4.6) are similar to those of Proposition 2.

Proofs of Propositions 6–7

The technical arguments are similar to the proofs of Propositions 1–2 and thus details are omitted.

Proof of Proposition 8

Recalling the definition of H_i^* in Section 5.2, we have that

$$\begin{aligned} H_i^* &= (\boldsymbol{e}_1 \otimes \mathtt{X}_i)^T \{S_n^*(U_i; \widehat{\boldsymbol{\beta}}(U_i))\}^{-1} (\boldsymbol{e}_1 \otimes \mathtt{X}_i) \ \{K_h(0) b''(\widehat{\theta}(U_i, \mathtt{X}_i))\} \\ &= (nh)^{-1} (\boldsymbol{e}_1 \otimes \mathtt{X}_i)^T \{n^{-1} (\mathbf{H} \otimes \mathbf{I}_d)^{-1} \widehat{S}_n^*(U_i) (\mathbf{H} \otimes \mathbf{I}_d)^{-1}\}^{-1} (\boldsymbol{e}_1 \otimes \mathtt{X}_i) K(0) b''(\widehat{\theta}(U_i, \mathtt{X}_i)), \quad \text{(A.17)} \end{aligned}$$

where

$$\begin{aligned} \widehat{S}_n^*(u) &= \sum_{j=1}^n \{\mathbf{u}_j(u) \otimes \mathtt{X}_j\} \{\mathbf{u}_j(u) \otimes \mathtt{X}_j\}^T K_h(U_j - u) b''(\{\mathbf{u}_j(u) \otimes \mathtt{X}_j\}^T \widehat{\boldsymbol{\beta}}(u)) \\ &= \sum_{j=1}^n \big[\{\mathbf{u}_j(u) \mathbf{u}_j(u)^T\} \otimes (\mathtt{X}_j \mathtt{X}_j^T)\big] K_h(U_j - u) b''(\{\mathbf{u}_j(u) \otimes \mathtt{X}_j\}^T \widehat{\boldsymbol{\beta}}(u)). \end{aligned}$$

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

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

This expression applied to (A.17) further implies that

$$\begin{aligned} \sum_{i=1}^n H_i^* &= \sum_{i=1}^n \frac{1}{nh f_U(U_i)} (\boldsymbol{e}_1 \otimes \mathtt{X}_i)^T \{S^{-1} \otimes \Gamma(U_i)^{-1}\} (\boldsymbol{e}_1 \otimes \mathtt{X}_i) K(0) b''(\widehat{\theta}(U_i, \mathtt{X}_i)) \{1 + o_P(1)\} \\ &= \sum_{i=1}^n \frac{1}{nh f_U(U_i)} (\boldsymbol{e}_1^T S^{-1} \boldsymbol{e}_1) K(0) \big\{\mathtt{X}_i^T \Gamma(U_i)^{-1} \mathtt{X}_i b''(\widehat{\theta}(U_i, \mathtt{X}_i))\big\} \{1 + o_P(1)\} \\ &= \sum_{i=1}^n \frac{\mathcal{K}(0)}{nh f_U(U_i)} \big\{\mathtt{X}_i^T \Gamma(U_i)^{-1} \mathtt{X}_i b''(\theta(U_i, \mathtt{X}_i))\big\} \{1 + o_P(1)\}. \qquad \text{(A.18)} \end{aligned}$$

For (A.18), a direct calculation gives that

$$\begin{aligned} E\big\{\mathtt{X}^T \Gamma(U)^{-1} \mathtt{X} b''(\theta(U, \mathtt{X}))/f_U(U)\big\} &= \operatorname{tr}\big[E\big\{\Gamma(U)^{-1} \mathtt{X}\mathtt{X}^T b''(\theta(U, \mathtt{X}))/f_U(U)\big\}\big] \\ &= \operatorname{tr}\big\{E\big[\Gamma(U)^{-1} E\{b''(\theta(U, \mathtt{X})) \mathtt{X}\mathtt{X}^T | U\}/f_U(U)\big]\big\} \\ &= \operatorname{tr}\big[E\big\{\Gamma(U)^{-1} \Gamma(U)/f_U(U)\big\}\big] \\ &= \operatorname{tr}(|\Omega_U| \mathbf{I}_d) = d|\Omega_U|. \end{aligned}$$

This completes the proof.

Table 1: *The Asymptotic Optimal Bandwidths $h_{\text{AMPEC}}(q_2)$ Calculated From (3.12) and h_{AMISE} From (3.13), with $p = 1$, the Epanechnikov Kernel, and Examples Given in Section 7.1*

Exponential family	*Example*	$h_{\text{AMPEC}}(q_2)$	h_{AMISE}
Poisson	1	.070	.079
	2	.089	.099
	3	.127	.136
Bernoulli	1	.106	.108
	2	.151	.146
	3	.184	.188

Table 2: *Choices of* `a` *and $\mathcal{C}$, in the Empirical Formulas (3.14) and (5.4), 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

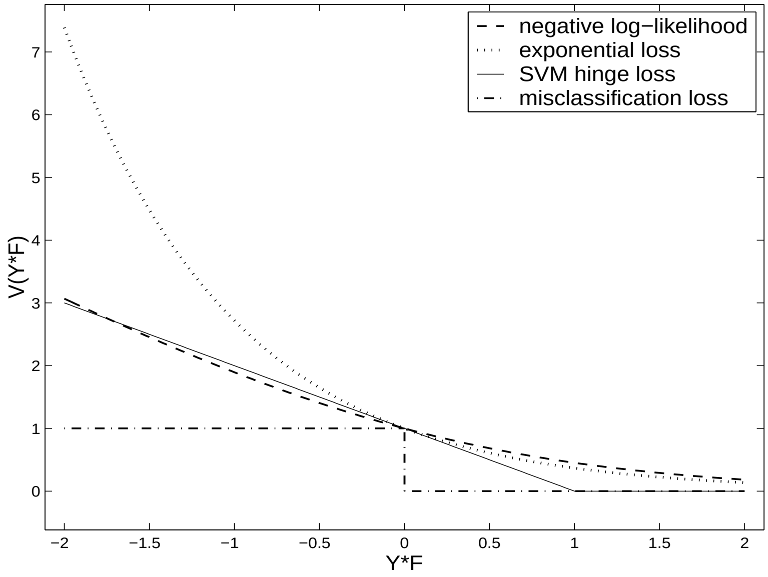


Figure 1: *Illustration of Margin-Based Loss Functions. Line types are indicated in the legend box. Each function has been re-scaled to pass the point* $(0,1)$*.*

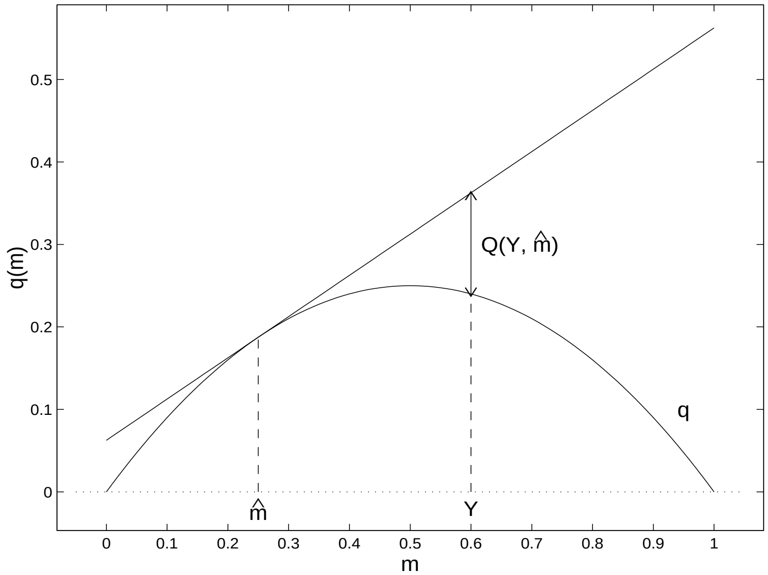


Figure 2: *Illustration of the Bregman Divergence* $Q(Y, \widehat{m})$ *as Defined in (2.2). The concave curve is the q-function; the two dashed lines give locations of* Y *and* $\widehat{m}$*; the solid strict line is* $q(\widehat{m}) + q'(\widehat{m})(Y - \widehat{m})$*; the vertical line with arrows at each end is* $Q(Y, \widehat{m})$*.*

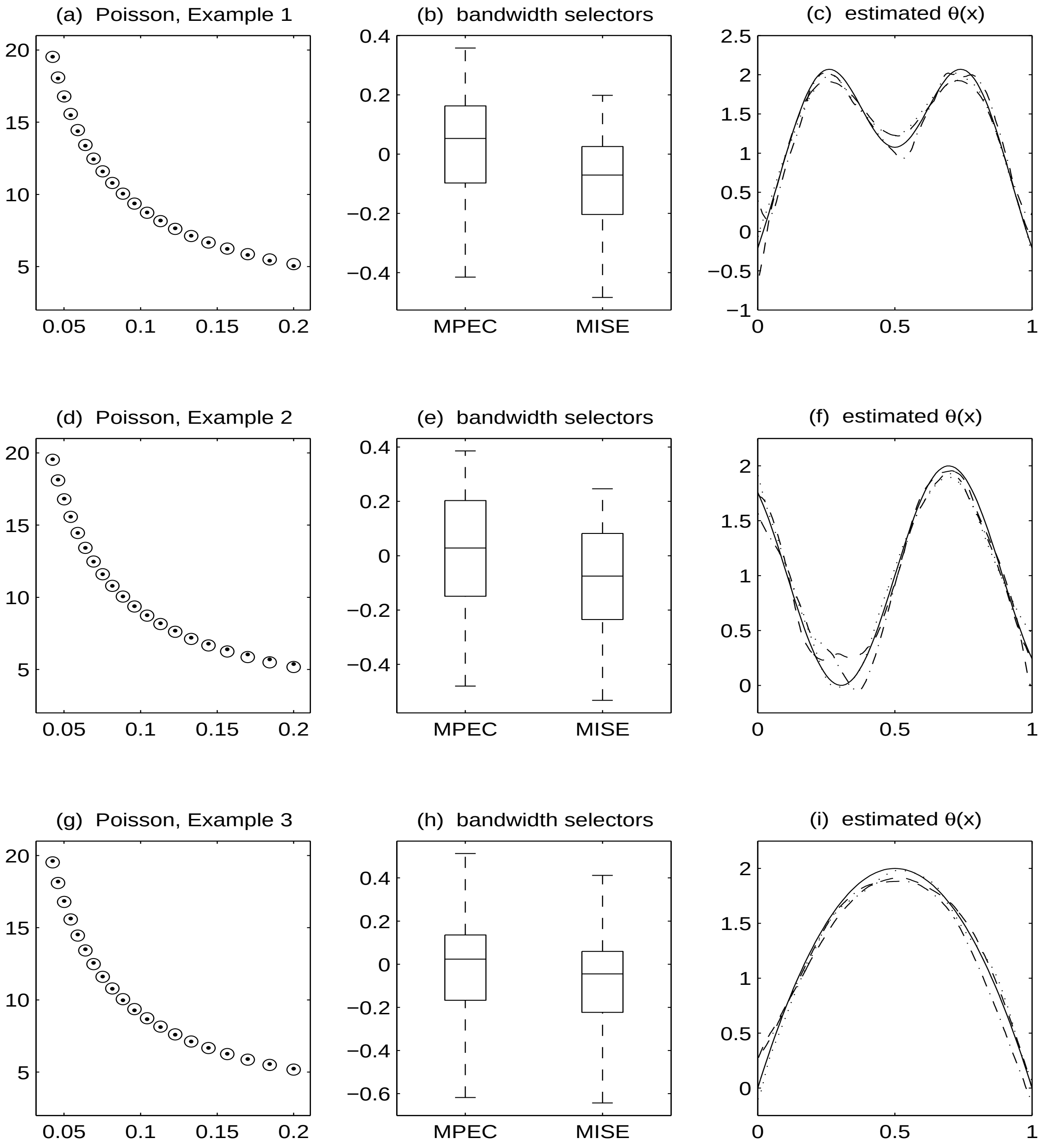


Figure 3: *Evaluation of Local-Likelihood Nonparametric Regression for Poisson Responses. Left panel: plots of* $\sum_{i=1}^{n} H_i$ *versus* h*. Dots denote the actual values, centers of circles stand for the empirical values given by (3.14), for local-linear smoother with* $\mathtt{a} = .70$ *and* $\mathcal{C} = 1.03$*. Middle panels [figures (b), (e), and (h)]: boxplots of* $\{\widehat{h}_{\mathrm{ECV}} - h_{\mathrm{AMPEC}}(q_2)\}/h_{\mathrm{AMPEC}}(q_2)$ *and* $\{\widehat{h}_{\mathrm{ECV}} - h_{\mathrm{AMISE}}\}/h_{\mathrm{AMISE}}$*. Panels (c), (f), and (i): estimated curves from three typical samples are presented corresponding to the* 25^{th} *(the dotted curve), the* 50^{th} *(the dashed curve), and the* 75^{th} *(the dash-dotted curve) percentiles among the ASE-ranked values. The solid curves denote the true functions.*

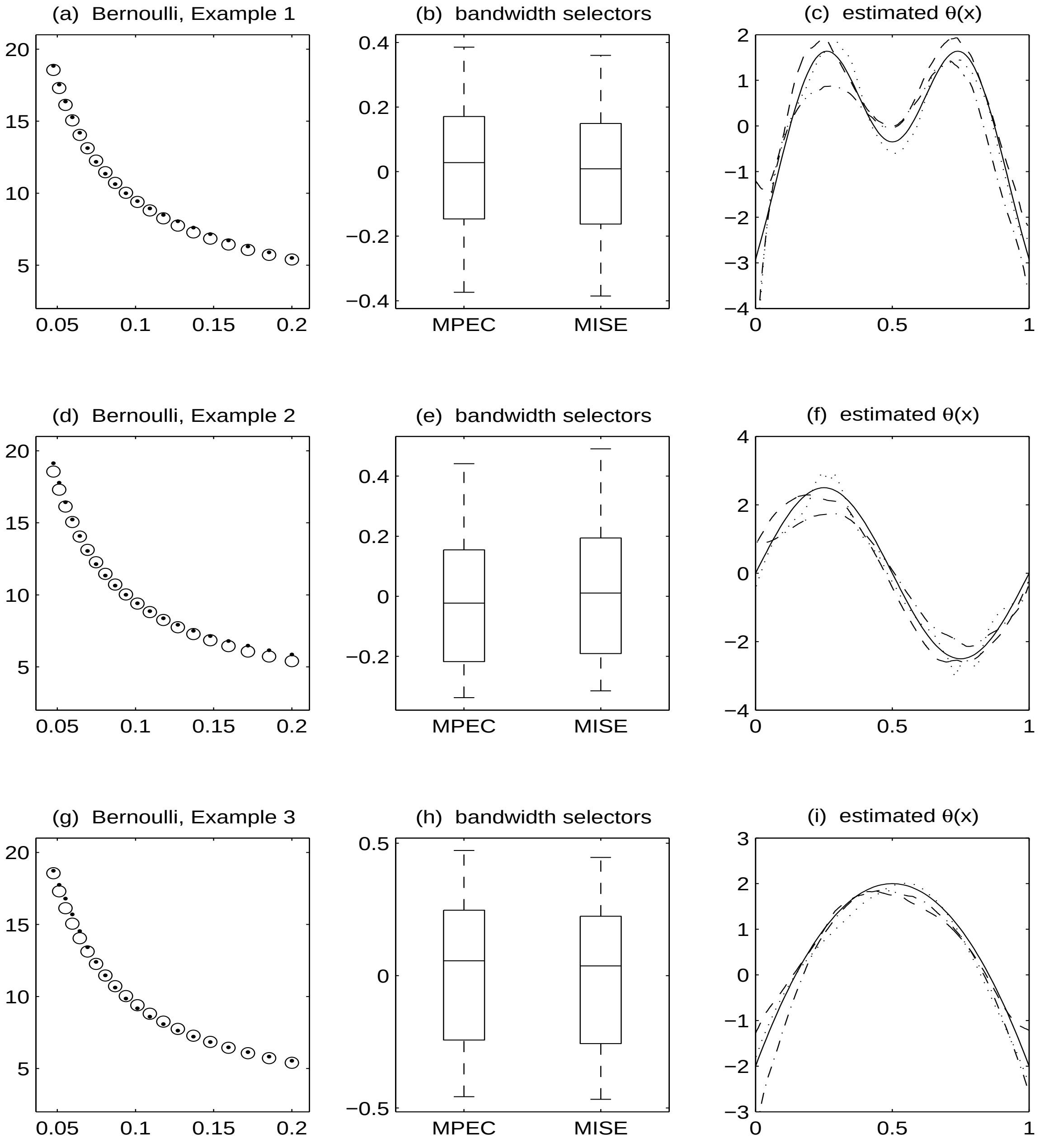


Figure 4: *Evaluation of Local-Likelihood Nonparametric Regression for Bernoulli Responses. Captions are similar to those for Figure 3. Here $\widehat{h}_{\mathrm{ECV}}$ minimizes the empirical version of (4.9); the empirical formula (3.14) uses $\mathsf{a} = .70$ and $\mathcal{C} = 1.09$ for H_i and $\mathsf{a} = .70$ and $\mathcal{C} = 1.03$ for $\mathcal{S}_i$.*

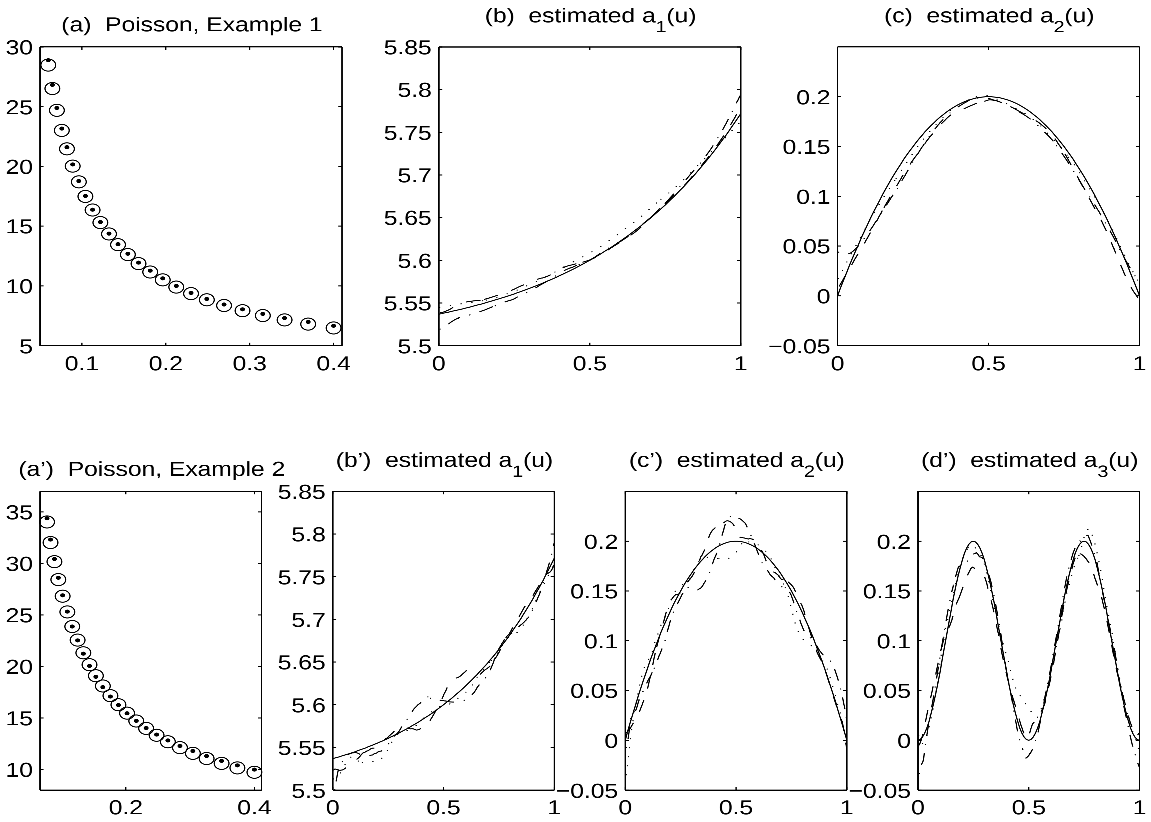


Figure 5: *Evaluation of Local-Likelihood Varying Coefficient Regression for Poisson Responses. Panels (a) and (a'): plots of* $\sum_{i=1}^{n} H_i^*$ *versus* h*. Dots denote the actual values, centers of circles stand for the empirical values given by (5.4), for local-linear smoother with* $\mathsf{a} = .70$ *and* $\mathcal{C} = 1.03$*. Panels (b)-(c) for Example 1 and panels (b')-(d') for Example 2: estimated curves from three typical samples are presented corresponding to the* 25^{th} *(the dotted curve), the* 50^{th} *(the dashed curve), and the* 75^{th} *(the dash-dotted curve) percentiles among the ASE-ranked values. The solid curves denote the true functions.*

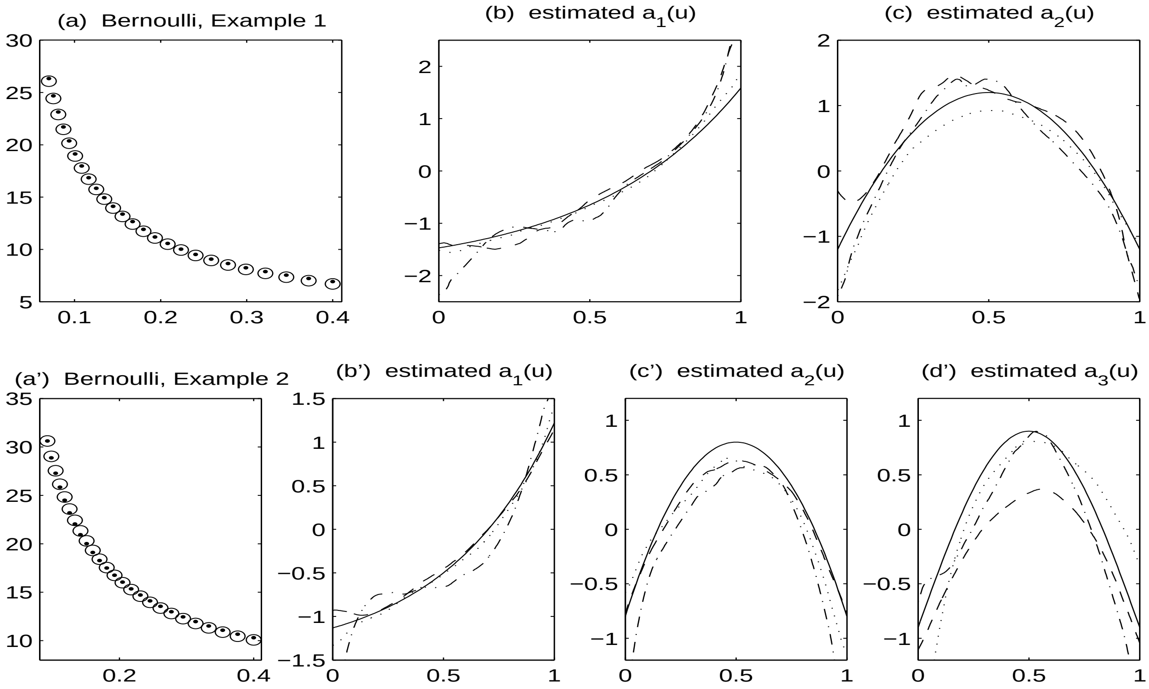


Figure 6: *Evaluation of Local-Likelihood Varying Coefficient Regression for Bernoulli Responses. Captions are similar to those for Figure 5. Here* $\widehat{h}_{\text{ECV}}$ *minimizes the empirical version of (5.5); the empirical formula (5.4) uses* $\mathtt{a} = .70$ *and* $\mathcal{C} = 1.09$ *for* H_i^* *and* $\mathtt{a} = .70$ *and* $\mathcal{C} = 1.03$ *for* $\mathcal{S}_i^*$.

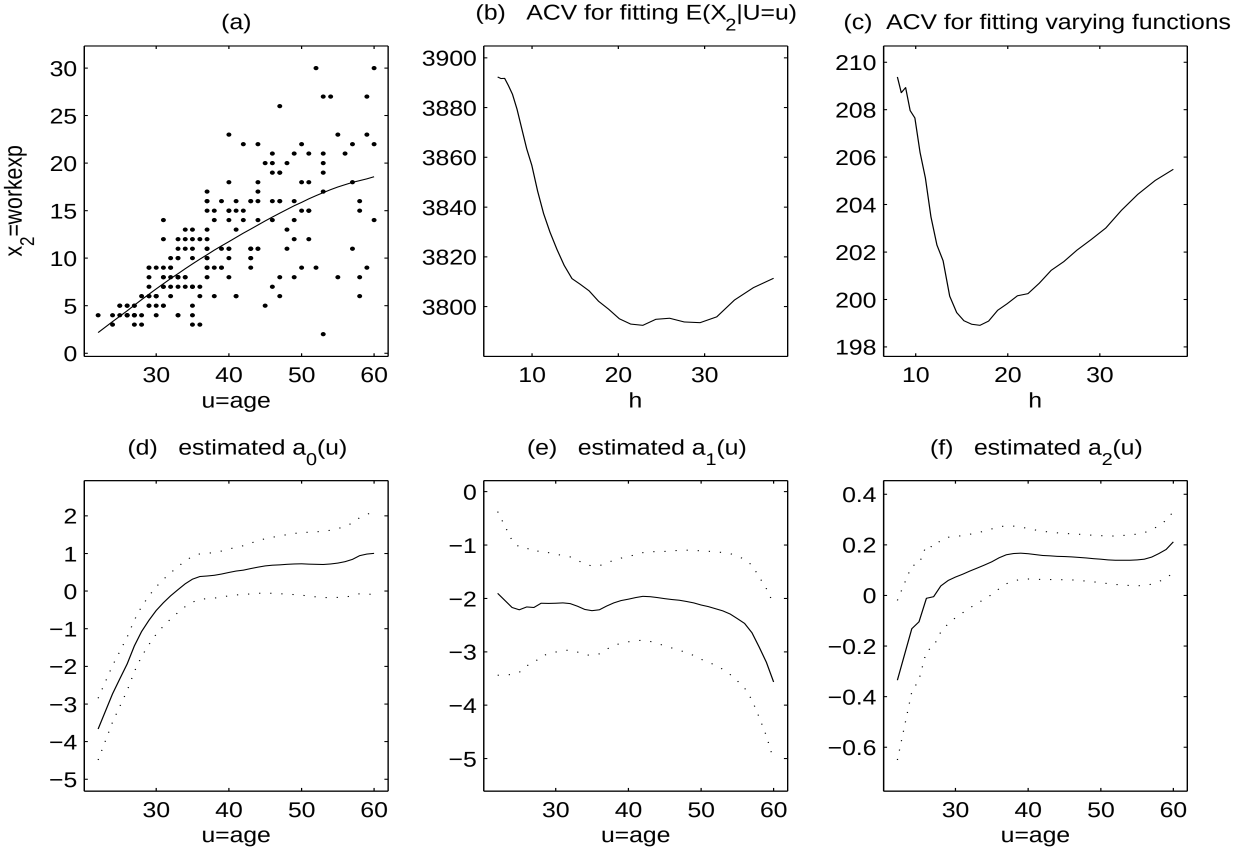


Figure 7: *Applications to the Job Grade Data Set Modeled by (8.1). (a) scatter plot of work experience versus age along with a local linear fit; (b) plot of the approximate CV function against bandwidth for the local linear fit in (a); (c) plot of the approximate CV function, defined in (5.5), against bandwidth for fitting varying coefficient functions; (d) estimated $a_0(u)$; (e) estimated $a_1(u)$; (f) estimated $a_2(u)$. The dotted curves are the estimated functions plus/minus* 1.96 *times of the estimated standard errors.*

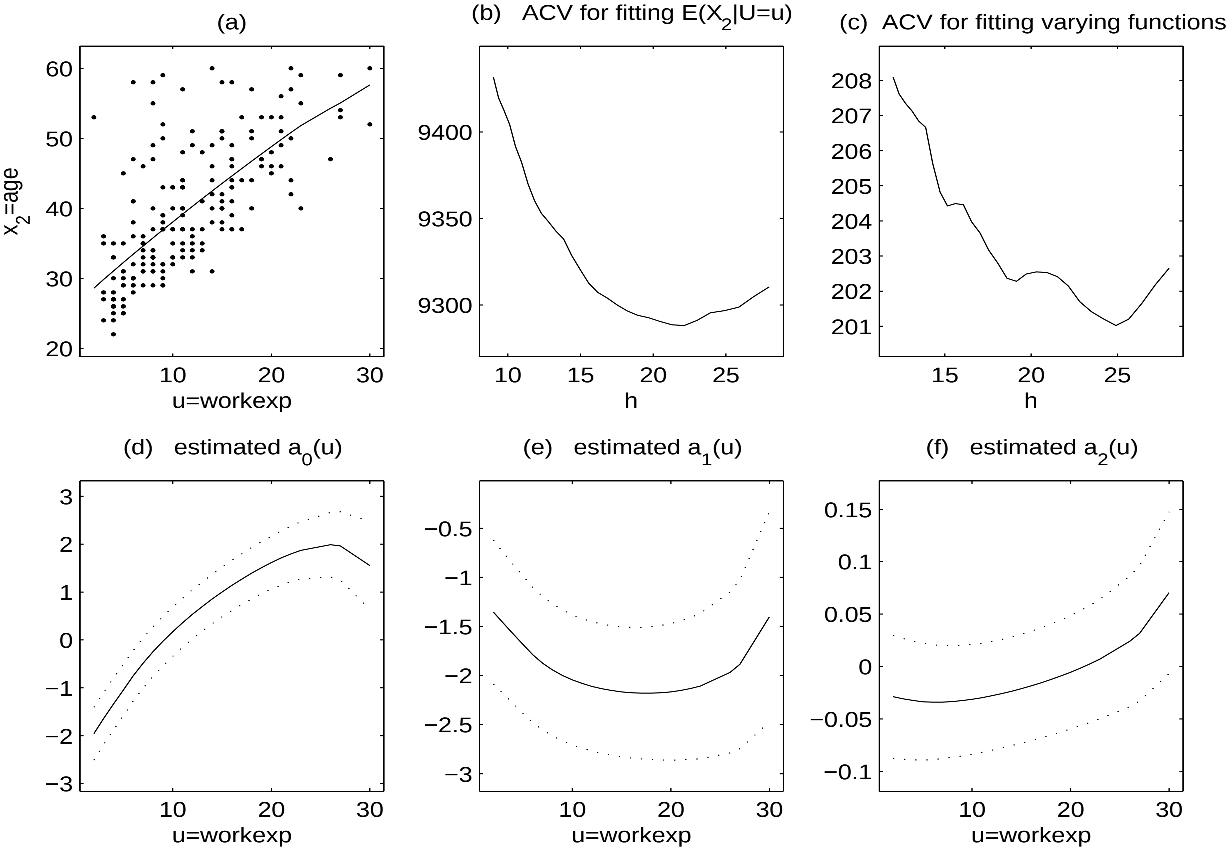
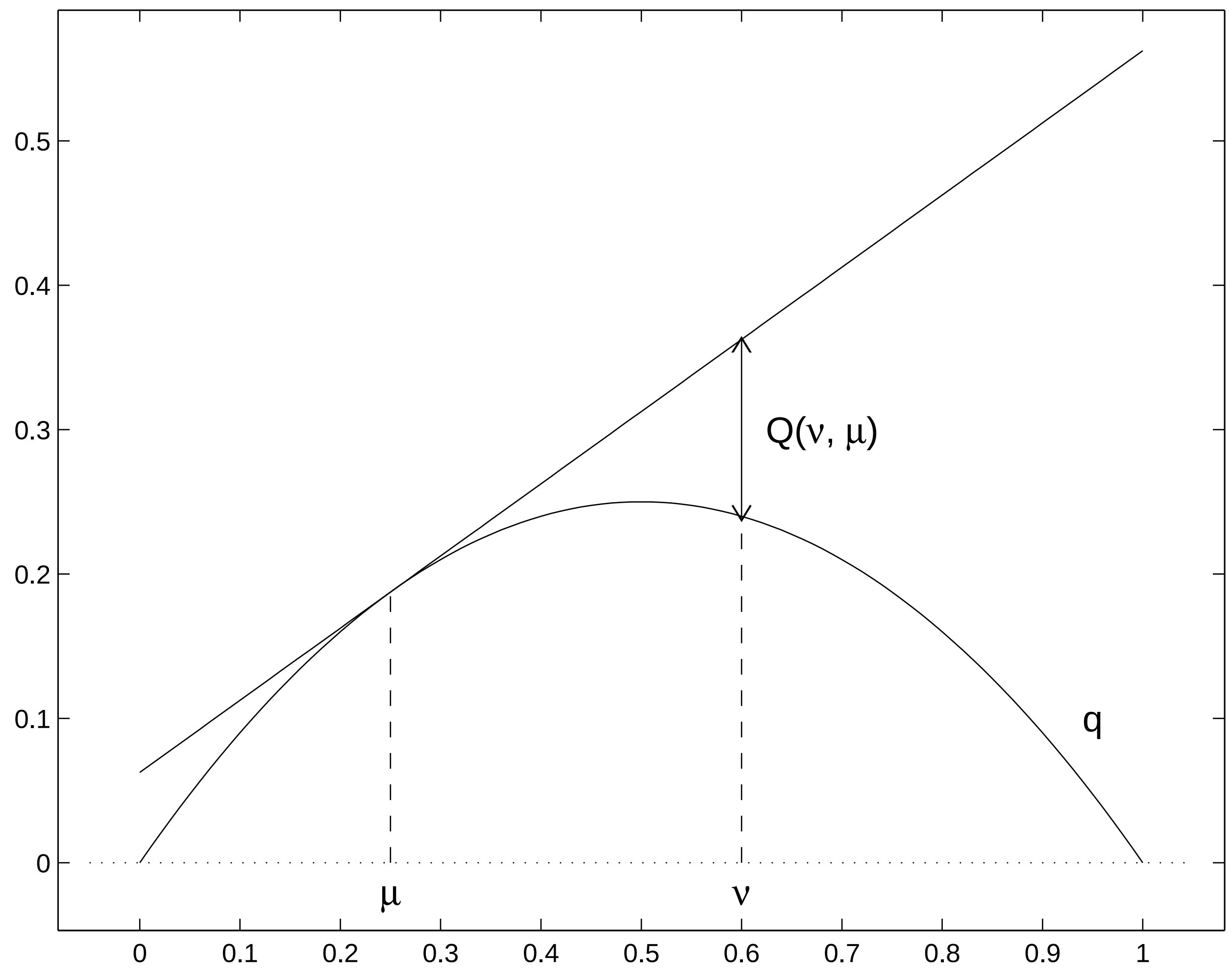
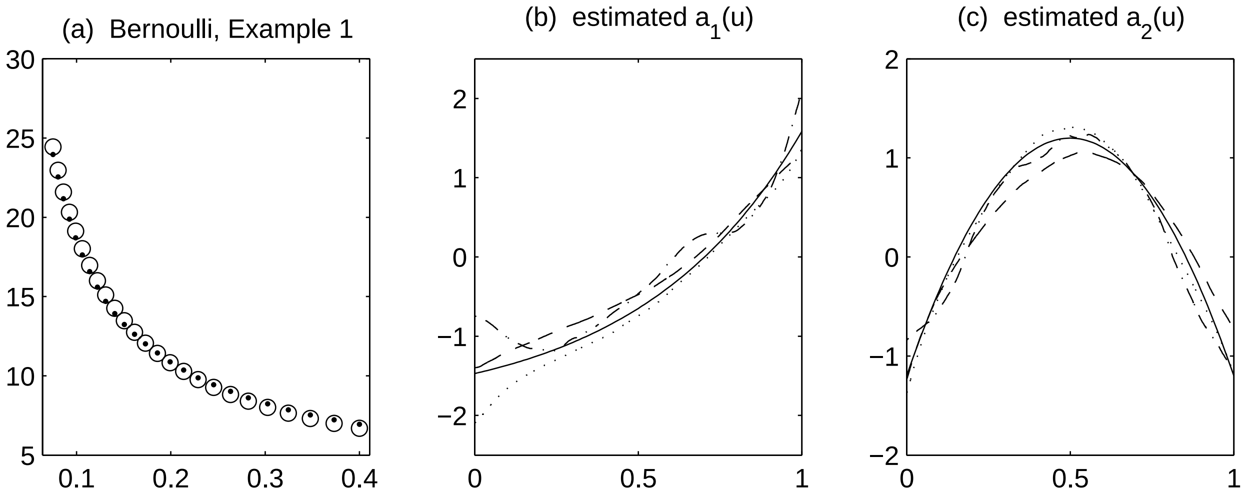


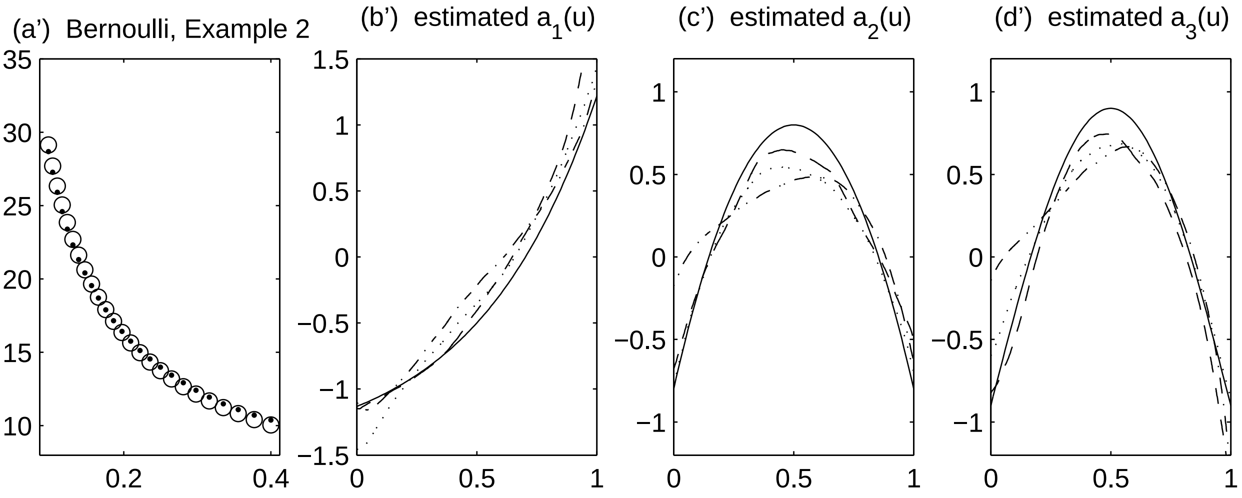
Figure 8: *Applications to the Job Grade Data Set Modeled by (8.1). Captions are similar to those for Figure 7, except that U is* WorkExp *and X_2 is the de-correlated* Age*.*



0.5
0.4
0.3
0.2
0.1
0
Q(ν, μ)
q
μ
ν
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1



(a) Bernoulli, Example 1
30
25
20
15
10
5
0.1
0.2
0.3
0.4
(b) estimated $a_1(u)$
2
1
0
−1
−2
0
0.5
1
(c) estimated $a_2(u)$
2
1
0
−1
−2
0
0.5
1


(a') Bernoulli, Example 2
35
30
25
20
15
10
0.2
0.4
(b') estimated $a_1(u)$
1.5
1
0.5
0
−0.5
−1
−1.5
0
0.5
1
(c') estimated $a_2(u)$
1
0.5
0
−0.5
−1
0
0.5
1
(d') estimated $a_3(u)$
1
0.5
0
−0.5
−1
0
0.5
1