國立臺灣大學理學院數學系(所)

博士論文

Department of Mathematics

College of Science

National Taiwan University

Doctoral Dissertation

廣義多參數概似模型之估計

On Estimation Methods in Generalized Multiparameter Likelihood Model

陳律閎

Lu-Hung Chen

指導教授: 鄭明燕教授

Advisor: Prof. Cheng, Ming-Yen

中華民國 99 年 1 月

January, 2010

國立臺灣大學數學系(所)

博士論文

廣義多參數概似模型之估計

陳律閎 撰

99
1

Abstract

Multiparameter likelihood models (MLMs) with multiple covariates have a wide range of applications; however, they encounter the "curse of dimensionality" problem when the dimension of the covariates is large. We develop a generalized multiparameter likelihood model that copes with multiple covariates and adapts to dynamic structural changes well. It includes some popular models, such as the partially linear and varying-coefficients models, as special cases. We discuss the backfitting and profile likelihood procedures and present a simple, effective two-step method to estimate both the parametric and the nonparametric components when the model is fixed. All these estimators of the parametric component has the $n^{-1/2}$ convergence rate, and the estimator of the nonparametric component enjoys an adaptivity property. We suggest a data-driven procedure for selecting the bandwidths, and propose an initial estimator in backfitting and profile likelihood estimation of the parametric part to ensure stability of the approach in general settings. We further develop an automatic procedure to identify constant parameters in the underlying model. We provide several simulation studies and an application to infant mortality data of China to demonstrate the performance of our proposed method.

中文摘要

能處理多個共變數(covariate)的多參數概似模型(Multiparameter Likelihood Models, MLMs)有非常廣泛的應用。然而，當共變數的維度很大時，我們會遇到"維度的詛咒(curse of dimensionality)"的問題。我們將多參數概似模型推廣成半參數模型，使之能處理較大的共變數維度同時能適應動態的結構變化。我們的模型包含了許多特例，如部份線性模型(partially linear models)、變係數模型(varying coefficients models)等。我們介紹兩種既有的方法以及提出一個簡單而且有效的兩步驟估計法來估計此模型的參數化的部份以及非參數的部份。這些估計方式在參數化的部份具有和一般參數化模型一樣的收斂速度($n^{-1/2}$)，非參數的部份則能估的和已知參數化的部份時一樣好(即具有adaptivity property)。我們也提了一個自動帶寬選擇(bandwidth selection)法，以及一個自動化的流程來決定哪些共變數應該被放在參數化的部份。我們做了一些模擬研究，並且舉了一個中國嬰兒死亡率的資料來顯示我們估計方法的性能。

致謝

首先感謝指導教授鄭明燕老師多年以來費心的指導與栽培，不管是學術上知識的教導還是其對工作、研究的嚴謹與熱誠的身教言教，均給我莫大的影響，讓我獲益良多。沒有鄭老師給我這麼好的題目以及過程中的指導，本研究的進行及論文撰寫不可能如此順利的完成，在此謹致上由衷的謝忱。

另外要感謝曾勝滄老師、戴政老師、張淑惠老師以及丘政民老師首肯擔任我的口試審查委員，給我論文撰寫以及實務研究方面之詳細指正，讓本論文不致犯錯並更加完備。此外，感謝陳祝嵩老師慷慨提供計算資源，使論文得以如期完成。

最後，謹將本論文獻給我的父母，感謝他們的養育之恩及對我永遠的支持、鼓勵。

Contents

List of Figures

List of Tables

1 Introduction

Consider statistical modeling of the relationship between a response variable and some covariates. Maximum likelihood estimation is most powerful when the joint distribution of the response variable and covariates is specified by a parametric form. But parametric approaches are at risk for model misspecification, which can result in seriously biased estimation, misinterpretation of data, and other problems. Nonparametric modeling is more flexible and allows data to present the unknown truth; however, it often comes up against the "curse of dimensionality" problem — that is, model instability when the dimension of the covariates is large. Numerous hybrids of parametric and nonparametric models, generally called semiparametric models, have been proposed to achieve a good balance between flexibility and stability in model specification. We will review related models in Section 3.

In this article we suggest a semiparametric model for a population $(\mathbf{X}, U, Y)$ in which U is a continuous variable and the conditional density function of Y given $(\mathbf{X}, U)$ is specified by

$$f\Big(Y;\ \mathbf{X},\ \boldsymbol{\theta},\ \mathbf{x}_1^{\mathrm{T}}\mathbf{a}_1(U),\ \cdots,\ \mathbf{x}_\ell^{\mathrm{T}}\mathbf{a}_\ell(U)\Big), \tag{1.1}$$

where f is a known parametric density function, $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_q)^{\mathrm{T}}$ is an unknown constant vector, $\mathbf{X} = (X_1, \cdots, X_p)^{\mathrm{T}}$ with $X_1 \equiv 1$, and $\mathbf{x}_j$ is a p_j-dimensional subvector of $\mathbf{X}$ and $\mathbf{a}_j(\cdot) = (a_{j1}(\cdot), \cdots, a_{jp_j}(\cdot))^{\mathrm{T}}$ is an unknown function, $j = 1, \cdots, \ell$. Here $1 \le \ell \le d$, where d is as defined in (1.2). Model (1.1) is a hybrid of the standard

multiparameter likelihood model (MLM) that assumes that the conditional density function of Y given $\mathbf{X}$ follows the form

$$f\Big(Y;\ a_1(\mathbf{X}),\ \cdots,\ a_d(\mathbf{X})\Big), \tag{1.2}$$

where f has d identifiable parameters and $a_1(\mathbf{X}), \cdots, a_d(\mathbf{X})$ are unknown functions; that is, Y depends on $\mathbf{X}$ through the d identifiable parameters in f being modeled as nonparametric functions of $\mathbf{X}$. Aerts and Claeskens (1997) studied a locally linear maximum likelihood estimator of MLMs when $\mathbf{X}$ is univariate, and Cheng and Peng (2007) proposed a variance reduction technique to improve the estimation. The MLM provides a general framework for specifying statistical relationship between response and covariates under a wide range of data configurations, including continuous, categorical, binary and count variables as the response and cases where the response is univariate or vector-valued. In addition, it can be easily adopted to cope with various statistical problems, such as mean regression, variance estimation, quantile regression, hazard regression, logistic regression, and longitudinal data analysis (for details, see, e.g., Aerts and Claeskens 1997; Loader 1999; Claeskens and Aerts 2000; Cheng and Peng 2007.)

With the availability of U, model (1.1) specifies ℓ of the d parameter functions in model (1.2) by some nonparametric or semiparametric form, and if $d - \ell > 0$, then the other $d-\ell$ parameter functions in (1.2) are now modeled parametrically in (1.1), with $\boldsymbol{\theta}$ comprising all of the constant parameters. Like MLM (1.2), (1.1) pro-

vides a unified approach to modeling a wide range of data settings and dealing with various inference problems. Nonetheless, (1.1) avoids the curse of dimensionality problem that (1.2) has when the dimension of $\mathbf{X}$ is large, and it allows parametric, nonparametric or semiparametric modeling of the parameter functions in (1.2). Furthermore, (1.1) broadens the application of MLMs, because it can cope with categorical covariates, which often arise in practice. Model (1.1) is a very general semiparametric model provided that there exists a continuous variable U and other covariates $\mathbf{X}$. It reduces to a partially linear model (3.1) when $\ell = 1$, $\mathbf{x}_1 = X_1 \equiv 1$ and $\boldsymbol{\theta}$ interacts with $\mathbf{X}$ through a linear function. When $\ell = 1, q = 0$ and $\mathbf{x}_1 = \mathbf{X}$, model (1.1) becomes the varying-coefficients model (3.2) of Hastie and Tibshirani (1993) with the same modifying variable U. Thus (1.1) inherits the stability, flexibility, and interpretability that varying-coefficients models enjoy. In addition, it is closely related the regression model II of Bickel, Klaassen, and Ritov and Wellner (1993, sec 4.3).

Here we propose a simple, effective, and fast two-step procedure to estimate both the constant and functional parameters in (1.1). The implementation of this model involve none of the iteration usually required by conventional approaches, such as profile likelihood and backfitting. Furthermore, we develop an Akiake Information Criterion (AIC) data-driven procedure to select the bandwidths required in the two-step estimation. The use of an AIC criterion (and modified versions) to select smoothing parameters in nonparametric regression and local likelihood modeling has

been extensively discussed and implemented (see, e.g. Hurvich, Simonoff, and Tsai 1998; Loader 1999; Schucany. For local likelihood estimation, Aerts and Claeskens (1997) considered cross-validation and plug-in bandwidths, and Fan, Farmen and Gijbels (1998) suggested a bandwidth selector based on an approximation to the integrated mean squared error. The backfitting and profile likelihood approaches can be applied to estimate the constant parameters, too. We propose a new initial estimator to ensure stability of the backfitting and profile likelihood approaches regardless of in which types of model features (e.g., location, scale, and shape) the constant parameter play roles. In general, neither profile likelihood nor the two-step estimator of the constant parameters is consistently superior to the other (see the asymptotic results and discussion in Sec. 6 and simulation results reported in Sec. 7). Nevertheless, the major strength of the two-step estimation is its simple and fast implementation and numerical stability, with no iteration required.

In practice, the real challenge is that we are often given a collection of significant covariates but do not know which of the parameter functions are constant and which are functional in (1.1); that is, we are not sure about the specification of $\boldsymbol{\theta}$ and $\mathbf{x}_1, \cdots, \mathbf{x}_\ell$. In an attempt to solve this fundamental identification problem, we suggest a stepwise procedure based on a version of the Bayes information criterion (BIC) accounted for our model. Identification of constant parameters and bandwidth selection interact with each other. We propose selecting the bandwidths first and then keeping them fixed throughout the procedure for identifying the constant

parameters. This approach indeed resolves a complex problem in an effective, fast, and stable fashion and is confirmed to have these properties by a simulation study and a real data analysis. We are not aware of any existing methods for identifying constant parameters or covariates in the parametric component of a semiparametric model, although there is an abundant literature on a different issue of variable selection for parametric models, nonparametric models, and parametric or nonparametric components in semiparametric models. For example, Irizarry (2001) derived weighted versions of the AIC and BIC and posterior probability model selection criteria for one-parameter local likelihood models. Fan and Li (2002) used profile likelihood techniques in their nonconcave penalized likelihood approach to selecting variables in the parametric part of Cox's proportional hazards model. Fan and Li (2004) incorporated profiling ideas in their construction of penalized least squares for variable selection in the parametric component of a semiparametric model for longitudinal data analysis. Bunea (2004) constructed a penalized least squares criterion for variable selection in the linear part of a partially linear model. For a generalized varying-coefficient partially linear model, Li and Liang (2008) used a nonconcave penalized likelihood to select significant variables in the parametric component and a generalized likelihood ratio test to select significant variables in the nonparametric component, assuming that the two sets of covariates in the parametric and nonparametric components are separated in advance.

In section 2 we provide some motivating examples for model (1.1) and discuss

the identifiability issue. We review the two classical estimation procedures: backfitting and profile likelihood estimation, and then present our two-step estimation procedure for both the constant and functional parameters and a new initial estimator for profile likelihood estimation of the constant parameters in Section 4, and address bandwidth selection and identification of the constant parameters are addressed in Section 5. We investigate the asymptotic properties of the backfitting, profile likelihood, and two-step estimators in Section 6. In section 7 we present three simulated examples and an analysis of a motivating example concerning infant mortality in Section 8. We defer proofs of the theoretical results to Appendixes.

2 Motivating examples and model identifiability

In applications, some of the unknown functional parameters in the MLM (1.2) may simply be unknown constants. Under such circumstances, we would pay a price in efficiency if the unknown constants were still treated as unknown functions. An example of this is an analysis of 103 annual maximum temperatures (Cheng and Peng, 2007) in which $Y|\mathbf{X}$ is modeled by an extreme value distribution, where $\mathbf{X}$ is year. The estimates of the shape and scale parameter curves are flat except in the boundary regions, which is reasonable because the two parameters are unlikely to change much within 100 years. To accommodate such situations, (1.2) needs to be restricted to the following semiparametric model:

$$f\Big(Y;\ \mathbf{X},\ \boldsymbol{\theta},\ a_1(\mathbf{X}),\ \cdots,\ a_\ell(\mathbf{X})\Big), \tag{2.1}$$

where $1 \leq \ell < d$ and $\boldsymbol{\theta}$ is a q-dimensional unknown parameter. Here ℓ out of the d parameter functions in (1.2) remain unknown functions of $\mathbf{X}$, and the other $d-\ell$ parameter functions are formulated by certain parametric forms, for example, unknown constants, with $\boldsymbol{\theta}$ comprising all of the constant parameters. The model studied by Severini and Wong (1992) is a special case of (2.1) with $d = 2$, $\ell = 1$, and $q = 1$. These authors studied profile likelihood estimation of $\boldsymbol{\theta}$, along with consistent estimators of a least favorable curve.

When the dimension of $\mathbf{X}$ is large, neither (1.2) nor (2.1) would work, because the curse of dimensionality problem. Claeskens and Aerts (2000) suggested alle-

viating this problem by restricting $a_1(\cdot), \cdots, a_d(\cdot)$ in (1.2) to additive models and estimating them using a backfitting algorithm. Alternatively, a restriction of (2.1),

$$f\Big(Y;\ \mathbf{X},\ \boldsymbol{\theta},\ \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_1,\ \cdots,\ \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_\ell\Big), \tag{2.2}$$

where $\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_\ell$ are unknown constant vectors, would cope with the curse of dimensionality problem. But (2.2) actually implies a constant impact of $\mathbf{X}$ on Y, which is somewhat implausible in practice. For example, in the analysis of infant mortality in China detailed later, the impact of type of region of residence on mortality would not be a constant along the time U, because China has changed greatly since 1950, and the difference between rural and urban regions has changed. The impact must vary with U and the pattern of the change is of interest. Although we can modify (2.2) to some other parametric models involved with U to capture the trend, for example,

$$f\Big(Y;\ \mathbf{X},\ \boldsymbol{\theta},\ \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_1 P_1(U),\ \cdots,\ \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}_\ell P_\ell(U)\Big),$$

where $P_j(U)$ is some polynomial of U, $j = 1, \cdots, \ell$. However, determining the correct forms of $P_j(\cdot)$ to catch the dynamic changes is difficult. To capture the dynamic pattern of the changes in the impact more accurately, we extend (2.2) to

$$f\Big(Y;\ \mathbf{X},\ \boldsymbol{\theta},\ \mathbf{X}^{\mathrm{T}}\mathbf{a}_1(U),\ \cdots,\ \mathbf{X}^{\mathrm{T}}\mathbf{a}_\ell(U)\Big), \tag{2.3}$$

where $\boldsymbol{\theta}$ is an unknown constant vector, and $\mathbf{a}_j(\cdot) = (a_{j1}(\cdot), \cdots, a_{jp}(\cdot))^{\mathrm{T}}$ is a vector of unspecified smooth functions, $j = 1, \cdots, \ell$. In (2.3), $\mathbf{a}_1(\cdot),\ \cdots,\ \mathbf{a}_\ell(\cdot)$ must share

the same dimension p, and all of the $a_{ij}(\cdot)$'s are assumed to be functional. This model assumption may be unnecessary in some situations. The analysis of infant mortality in China is an example; the impact of ethnic group or type of feeding on infant mortality can be formulated as an unknown constant parameter. To remove such unnecessary restrictions and make the model more versatile, we generalize (2.3) to (1.1) with all the $a_{ij}(\cdot)$'s in (2.3) that are constant absorbed by $\boldsymbol{\theta}$ in (1.1).

When $\mathbf{a}_1(\cdot), \cdots, \mathbf{a}_\ell(\cdot)$ are all constant, model (2.3) reduces to model (2.2), and $\boldsymbol{\mathcal{I}}(\boldsymbol{\gamma})$ defined in Theorem 3 becomes $\tilde{\boldsymbol{\mathcal{I}}}$, where $\tilde{\boldsymbol{\mathcal{I}}}$ is $\boldsymbol{\mathcal{I}}(\boldsymbol{\gamma})$ with $\mathbf{a}_j(U)$ replaced by $\boldsymbol{\beta}_j$. Condition (S7) in Appendix C ensures that the smallest eigenvalue of $\tilde{\boldsymbol{\mathcal{I}}}$ is greater than the positive number λ_0 in condition (S7). If $(\mathbf{X}_i,\ Y_i)$, $i = 1, \cdots, n$, is a sample from model (2.2) then, under condition (S7), the Fisher information matrix is

$$\sum_{i=1}^{n} \mathrm{diag}(\mathbf{I}_q,\ \mathbf{I}_\ell \otimes \mathbf{X}_i)\, \tilde{\boldsymbol{\mathcal{I}}}_i\, \mathrm{diag}(\mathbf{I}_q,\ \mathbf{I}_\ell \otimes \mathbf{X}_i^{\mathrm{T}}) > \lambda_0 \sum_{i=1}^{n} \mathrm{diag}(\mathbf{I}_q,\ \mathbf{I}_\ell \otimes \mathbf{X}_i)\mathrm{diag}(\mathbf{I}_q,\ \mathbf{I}_\ell \otimes \mathbf{X}_i^{\mathrm{T}})$$
$$\approx n\lambda_0 \mathrm{diag}(\mathbf{I}_q,\ \mathbf{I}_\ell \otimes E(\mathbf{X}\mathbf{X}^{\mathrm{T}})) > 0,$$

where $\tilde{\boldsymbol{\mathcal{I}}}_i$ is $\tilde{\boldsymbol{\mathcal{I}}}$ with $\mathbf{X}$ replaced by $\mathbf{X}_i$. Here $\mathbf{I}_k$ denotes a size k identity matrix, and for any matrixes $\mathbf{A}$ and $\mathbf{B}$, $diag(\mathbf{A}, \mathbf{B})$ denotes the matrix

$$\begin{pmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \end{pmatrix}.$$

The condition (S7) ensures that the Fisher information matrix of the parametric model (2.2) is positive-definite; that is, model (2.2) is identifiable. Furthermore, for any given value of U, the local version of model (2.3) is model (2.2); thus, under

condition (7), model (2.3) is identifiable for any given value of U, and so model (2.3) is identifiable. In addition, model (1.1) specifies some of the $a_{ij}(\cdot)$'s in model (2.3) as constant and thus is identifiable. Based on the foregoing arguments, we have the following lemma.

Lemma 1. *Under condition (S7) in Appendix C, both models (1.1) and (2.3) are identifiable.*

3 Reviews of related models

Many semiparametric models have been proposed and developed. Most of them focus on the regression case or the extension of generalized linear models. For example, Engle, Granger, Rice, and Weiss (1986) proposed a partially linear regression model of the form:

$$Y = \mathbf{X}^T\beta + g(\mathbf{U}) + \epsilon \tag{3.1}$$

where $\mathbf{X} = (X_1, \cdots, X_p)^T$ and $\mathbf{U} = (U_1, \cdots, U_d)^T$ are vectors of covariates, $\beta = (\beta_1, \cdots, \beta_p)^T$ is a vector of unknown parameters, $g(\cdot)$ is a unknown smooth function from $\mathbb{R}^d$ to $\mathbb{R}$, and ϵ is independent of $(\mathbf{X}, \mathbf{U})$ with mean zero and finite variance $E(\epsilon^2) = \sigma^2$. They applied this model to analyze the relationship between temperature and electricity usage. In their paper, β and $g(\cdot)$ are estimated by smoothing spline:

$$\left(\hat{\beta}, \hat{g}\right) = \arg\min_{\beta, g} \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \mathbf{X}_i^T\beta - g(\mathbf{U}_i)\right)^2 + \lambda\int g''(u)^2 du,$$

where λ is a smoothing parameter and can be automatically determined by cross-validation. Cai, Fan, Jiang, and Zhou (2007) used partially linear hazard regression to analyze multivariate survival data. They assumed that the marginal hazard function follows

$$\lambda_{ij}(t) = Y_{ij}(t)\lambda_{0j}(t)\exp\left[\beta^T X_{ij}(t) + g\left(U_{ij}(t)\right)\right],$$

where $Y_{ij}(t) = \mathbf{1}(X_{ij} \geqslant t)$ is an at-risk indicator process, $\lambda_{0j}(t)$ is an unspecified baseline hazard function, and $g(\cdot)$ is an unspecified smooth function. The coefficients

of the parametric part β is estimated by profile partial likelihood approach, and the nonparametric part $g(\cdot)$ is estimated by local partial likelihood approach. For more details and applications about the partially linear model can be found in Härdle et al. (2000).

Hastie and Tibshirani (1993) proposed the varying coefficients model of the form:

$$Y = \mathbf{X}^T \mathbf{a}(U) + \epsilon, \tag{3.2}$$

where $\mathbf{X} = (X_1, \cdots, X_p)^T$, $\mathbf{a}(U) = (a_1(U), \cdots, a_p(U))^T$ are unspecified smooth functions. They proposed a smoothing spline approach to estimate $a_j(\cdot)$, that is, find $a_j(\cdot)$, $j = 1, \cdots, p$, to minimize

$$\sum_{i=1}^{n} \left\{ y_i - \sum_{j=1}^{p} x_{ij} a_j(u_i) \right\}^2 + \sum_{j=1}^{p} \lambda_j \int a_j''(u)^2 du$$

where $\lambda_j \geqslant 0$, $j = 1, \cdots, p$ are predefined smoothing parameters. Fan and Zhang (1998) proposed to estimate $a_j(\cdot)$ by local linear smoothing. Suppose that $a_j(\cdot)$ has a continuous second order derivative. For each given u_0, we approximate $a_j(u)$ locally by a linear function $a_j(u_0) \approx a_j + b_j(u - u_0)$, $j = 1, \cdots, p$. Let $(\hat{a}_1, \hat{b}_1, \cdots, \hat{a}_p, \hat{b}_p)$ be minimizer of

$$\sum_{i=1}^{n} \left\{ y_i - \sum_{j=1}^{p} \left(a_j - b_j(U_i - u_0) \right) x_{ij} \right\}^2 K_h(U_i - u),$$

where $K_h(t) = K(t/h)/h$, $K(t)$ is a kernel function and h is bandwidth; then the local linear estimator of $a_j(u)$ is taken to be $\hat{a}_j$, $j = 1, \cdots, p$. The bandwidth h can

be automatically selected by cross-validation. Similar ideas can be found in Hoover et al. (1998). They applied the varying coefficients model to longitudinal data: let

$$Y_{ij} = X_{ij1}a_1(t_{ij}) + \cdots + X_{ijp}a_p(t_{ij}) + \epsilon_i(t_{ij})$$

for $i = 1, \cdots, n$ and $j = 1, \cdots, n_i$, where n denotes the number of subjects, n_i denotes the number of measurements for the i-th subject, $a_1(t), \cdots, a_p(t)$ are unknown smooth functions which are estimated by smoothing spline or local polynomial smoothing with smoothing parameter selected by cross-validation, and $\epsilon_i(t)$ are uncorrelated stochastic processes.

Cai et al. (2000) proposed the generalized varying coefficient models that follows:

$$g(m(U, \mathbf{X})) = \mathbf{X}^T\mathbf{a}(U),$$

where g is a link function and $m(U, \mathbf{X}) = E(Y|U, \mathbf{X})$. They applied a local maximum likelihood estimation proposed by Fan et al. (1998) to estimate $\mathbf{a}(\cdot)$. Let $f(y; m(U, \mathbf{X}))$ denote the log conditional density function of Y given $(U, \mathbf{X}^T)$. For any given u, let $(\hat{\mathbf{a}}^T, \hat{\mathbf{b}}^T)$ be the maximizer of the local likelihood function

$$L(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^{n} f\Big(y_i; g^{-1}\Big[\mathbf{X}_i^T\Big\{\mathbf{a} + \mathbf{b}(U_i - u_0)\Big\}\Big]\Big)K_h(U_i - u),$$

where $\mathbf{a} = (a_1, \cdots, a_p)^T$ and $\mathbf{b} = (b_1, \cdots, b_p)^T$. The bandwidth h can be selected by minimizing the cross-validation criteria:

$$\text{CV} = -\sum_{i=1}^{n} f\Big\{y_i; g^{-1}\Big(\mathbf{X}_i^T\hat{\mathbf{a}}^{\setminus i}(U_i)\Big)\Big\},$$

where $\hat{\mathbf{a}}^{\setminus i}(U_i)$ is the estimated value of $\mathbf{a}(U_i)$ with the i-th observation deleted.

In practice, some of the components of $\mathbf{a}(\cdot)$ in model (3.2) can be constant (or other parametric forms) while other components have unknown interactions with U. With out loss of generality, we can write the model as

$$Y = \mathbf{Z}_1^T \mathbf{a}_1(U) + \mathbf{Z}_2^T \mathbf{a}_2 + \epsilon \tag{3.3}$$

where $(\mathbf{Z}_1^T, \mathbf{Z}_2^T)^T = \mathbf{X}$. This leads to a semiparametric model model known as semivarying coefficients model. Zhang et al. (2002) proposed a two step estimation procedure: they first treat $\mathbf{a}_2$ as functionals of U and appeal to local linear smoothing to get the initial estimator of $\mathbf{a}_2(U_i)$, namely, $\tilde{\mathbf{a}}_2(U_i)$. Then, they average $\tilde{\mathbf{a}}_2(U_i)$ over $i = 1, \cdots, n$ to get the final estimator of a_2 and show that their estimator of $\mathbf{a}_2$ has $n^{-1/2}$ convergence rate when the bandwidth for the initial estimator $\tilde{\mathbf{a}}_2(U_i)$ in the first step is taken to be of order $O(n^{-1/4})$. Fan and Huang (2005) proposed a profile least-square technique to estimate $\mathbf{a}_2$. Their idea is that for any given $\mathbf{a}_2$, model (3.3) can be written as

$$\tilde{Y} = \mathbf{Z}_1^T \mathbf{a}_1(U) + \epsilon$$

which is a standard varying coefficients model, where $\tilde{Y} = Y - \mathbf{Z}_2^T \mathbf{a}_2$. Then the estimator of $\mathbf{a}_1(U)$ can be obtained by local linear smoothing, which can be written as $\tilde{\mathbf{a}}_1(U) = S\tilde{Y}$, where S is the smoothing matrix. Substituting $\tilde{\mathbf{a}}_1(U)$ for $\mathbf{a}_1(U)$ in model (3.3) we have

$$(I - S)Y = (I - S)\mathbf{Z}_2^T \mathbf{a}_2 + \epsilon,$$

and the least square estimator of $\mathbf{a}_2$ becomes

$$\hat{\mathbf{a}}_2 = \left\{\mathbf{Z}_2^T(I-S)^T(I-S)\mathbf{Z}_2\right\}^T \mathbf{Z}_2^T(I-S)^T(I-S)Y. \tag{3.4}$$

Hence we can start from an initial guess of $\mathbf{a}_2$ which is not far from its true value, and estimate $\hat{\mathbf{a}}_2$ iteratively, as shown in Section 4.2. Fan and Huang (2005) showed that the asymptotic variance of their estimator reaches the lower bound for semi-parametric models.

4 Estimation procedures

Suppose that we have a sample $(\mathbf{X}_i,\ U_i,\ Y_i)$, $i = 1,\ \cdots,\ n$, from $(\mathbf{X},\ U,\ Y)$, which obeys model (1.1). Let $\mathbf{x}_{i,j}$ be the p_j-dimensional subvector of $\mathbf{X}_i$ that corresponds to $\mathbf{x}_j$, $j = 1,\ \cdots,\ \ell$, $i = 1,\ \cdots,\ n$. We discuss existing backfitting and profile likelihood approaches, and introduce our two-step procedures for estimating both the constant and functional parameters in Sections 4.1, 4.2, and 4.3.

4.1 Backfitting estimation

The idea of backfitting is on iteration. If $\boldsymbol{\theta}$ is given, model (1.1) reduces to a nonparametric model and the functional parameters can be estimated by regular local likelihood approach as follows. For any fixed u, by Taylor's expansion, we have, for each j,

$$\mathbf{a}_j(U_i) \approx \mathbf{a}_j(u) + \dot{\mathbf{a}}_j(u)(U_i - u)$$

when U_i is in a neighborhood of u, where $\dot{\mathbf{a}}_j(u) = d\mathbf{a}_j(u)/du$. This leas to the following local log-likelihood function:

$$\sum_{i=1}^{n} K_{h_1}(U_i - u) \log f\left(Y_i;\ \mathbf{X}_i,\ \boldsymbol{\theta},\ \mathbf{x}_{i,1}^{\mathrm{T}}\left\{\mathbf{a}_1 + \mathbf{b}_1(U_i - u)\right\},\ \cdots,\ \mathbf{x}_{i,\ell}^{\mathrm{T}}\left\{\mathbf{a}_\ell + \mathbf{b}_\ell(U_i - u)\right\}\right), \tag{4.1}$$

where $K_{h_1}(\cdot) = K(\cdot/h_1)/h_1$, $K(\cdot)$ is a kernel function, and $h_1 > 0$ is a bandwidth. Note that we assume $\boldsymbol{\theta}$ in (4.1) is known. Maximizing (4.1) with respect to $\left(\mathbf{a}_1^{\mathrm{T}}, \mathbf{b}_1^{\mathrm{T}}, \cdots, \mathbf{a}_\ell^{\mathrm{T}}, \mathbf{b}_\ell^{\mathrm{T}}\right)^{\mathrm{T}}$ we get the maximizer $\left(\hat{\mathbf{a}}_1(u)^{\mathrm{T}},\ \hat{\mathbf{b}}_1(u)^{\mathrm{T}}, \cdots, \hat{\mathbf{a}}_\ell(u)^{\mathrm{T}}, \hat{\mathbf{b}}_\ell(u)^{\mathrm{T}}\right)^{\mathrm{T}}$.

The estimator of $\mathbf{a}(u)$ is taken to be $\hat{\mathbf{a}}(u) = \left(\hat{\mathbf{a}}_1(u)^{\mathrm{T}}, \cdots, \hat{\mathbf{a}}_\ell(u)^{\mathrm{T}}\right)^{\mathrm{T}}$. On the other hand, when $\mathbf{a}_1(\cdot), ..., \mathbf{a}_l(\cdot)$ are given, model (1.1) becomes the parametric model model (2.1) and the constant parameters can be estimated by maximum likelihood approach. Hence the backfitting algorithm start from an initial guess of $\boldsymbol{\theta}$, plug-in this guess to replace $\boldsymbol{\theta}$ in 4.1 and update estimates of $\mathbf{a}_j(\cdot)$, $j = 1, \cdots, l$, and then update estimates of $\boldsymbol{\theta}$ iteratively until the estimates of $\boldsymbol{\theta}$ converges. We state the details as follows.

(a) Initialize $\boldsymbol{\theta}$ by a proper guess $\hat{\boldsymbol{\theta}}_{BF}^{(0)}$. Set $k = 1$.

(b) Estimate $\mathbf{a}_j(\cdot)$ by maximizing (4.1) with $\boldsymbol{\theta}$ being replaced by $\hat{\boldsymbol{\theta}}_{BF}^{(k-1)}$ with respect to $\left(\mathbf{a}_1^T, \mathbf{b}_1^T, \cdots, \mathbf{a}_\ell^T, \mathbf{b}_l^T\right)^T$ we get the maximizer $\left(\hat{\mathbf{a}}_1^{(k)^T}, \hat{\mathbf{b}}_1^{(k)^T}, \cdots, \hat{\mathbf{a}}_\ell^{(k)^T}, \hat{\mathbf{b}}_\ell^{(k)^T}\right)^T$. The estimator of $\mathbf{a}_j(\cdot)$ in this step is taken to be $\hat{\mathbf{a}}_j^{(k)}(\cdot)$, $j = 1, \cdots, l$.

(c) Estimate $\boldsymbol{\theta}$ by maximizing

$$f\left(Y; \mathbf{X}, \boldsymbol{\theta}, \mathbf{x}_1^{\mathrm{T}}\hat{\mathbf{a}}_1^{(k)}(U), \ \cdots, \ \mathbf{x}_\ell^{\mathrm{T}}\hat{\mathbf{a}}_\ell^{(k)}(U)\right), \tag{4.2}$$

with respect to $\boldsymbol{\theta}$ we get the maximizer $\hat{\boldsymbol{\theta}}_{BF}^{(k)}$. The estimator of $\boldsymbol{\theta}$ is taken to be $\hat{\boldsymbol{\theta}}_{BF}^{(k)}$. If $\left\| \hat{\boldsymbol{\theta}}_{BF}^{(k)} - \hat{\boldsymbol{\theta}}_{BF}^{(k-1)} \right\|$ is smaller than a pre-defined tolerance, we say that $\hat{\boldsymbol{\theta}}_{BF}^{(k)}$ converges and the estimation procedure is completed. Denotes the final estimates $\hat{\boldsymbol{\theta}}_{BF} = \hat{\boldsymbol{\theta}}_{BF}^{(k)}$. Otherwise, change k to $k+1$ and go to (b). In backfitting, (4.2) is maximized by solving

$$\frac{d}{d\boldsymbol{\theta}} \sum_{i=1}^{n} \log f\left(Y_i; X_i, \boldsymbol{\theta}, \mathbf{x}_1^T\hat{\mathbf{a}}_1^{(k)}(U_i), \cdots, \mathbf{x}_\ell^T\hat{\mathbf{a}}_\ell^{(k)}(U_i)\right) = 0$$

or by minimizing

$$\left\| \frac{d}{d\boldsymbol{\theta}} \sum_{i=1}^{n} \log f \left(Y_i; X_i, \boldsymbol{\theta}, \mathbf{x}_1^T \hat{\mathbf{a}}_1^{(k)} \left(U_i \right), \cdots, \mathbf{x}_\ell^T \hat{\mathbf{a}}_\ell^{(k)} \left(U_i \right) \right) \right\|$$

to avoid singularity.

It can be shown by Theorem 1 that $\hat{\boldsymbol{\theta}}_{BF}$ has $n^{-1/2}$ convergence rate under some regularity conditions if the bandwidth in (4.1) satisfies $h_1 \propto n^{-1/4}$ (that is, $\hat{\mathbf{a}}_j$ needs to be undersmoothed). However, there are some disadvantages for backfitting. First, the bandwidth is difficult to choose automatically, especially when the initial guess of $\hat{\boldsymbol{\theta}}_{BF}^{(0)}$ is far from the true value of $\boldsymbol{\theta}$. Under this circumstance, the variations of $\hat{\mathbf{a}}_j$ may be dominated by the variations due to $\hat{\boldsymbol{\theta}}_{BF}$, which is unknown for us. Second, the estimation requires iterations and is computation intensive. Third, if the initialization $\hat{\boldsymbol{\theta}}_{BF}^{(0)}$ is far from the true value of $\boldsymbol{\theta}$, backfiiting procedure usually requires more iterations, or even fails to converge. Finally, if the design of U is sparse, estimation of $\hat{\mathbf{a}}_j$ may fail, and thus $\hat{\boldsymbol{\theta}}_{BF}$ may diverge.

4.2 Profile likelihood estimation

A profile likelihood estimator for $\boldsymbol{\theta}$ maximizes, with respect to $\boldsymbol{\theta}$, a profiled log-likelihood

$$\sum_{i=1}^{n} \log f \Big(Y_i; \mathbf{X}_i, \boldsymbol{\theta}, \mathbf{x}_{i,1}^{\mathrm{T}} \tilde{\mathbf{a}}_{1\boldsymbol{\theta}}(U_i), \cdots, \mathbf{x}_{i,\ell}^{\mathrm{T}} \tilde{\mathbf{a}}_{\ell,\boldsymbol{\theta}}(U_i) \Big),$$

where, for any given $\boldsymbol{\theta}$, $\tilde{\mathbf{a}}_{\boldsymbol{\theta}}(\cdot) = \left(\tilde{\mathbf{a}}_{1,\boldsymbol{\theta}}(\cdot)^{\mathrm{T}}, \cdots, \tilde{\mathbf{a}}_{\ell,\boldsymbol{\theta}}(\cdot)^{\mathrm{T}}\right)^{\mathrm{T}}$ is an estimator for $\mathbf{a}(\cdot)$. In practice, we need to find the minimizer of

$$\left\| \frac{\partial L_n}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}, \tilde{\mathbf{a}}_{\boldsymbol{\theta}}) + \frac{\partial L_n}{\partial \mathbf{a}}(\boldsymbol{\theta}, \tilde{\mathbf{a}}_{\boldsymbol{\theta}}) \frac{\partial}{\partial \boldsymbol{\theta}} \tilde{\mathbf{a}}_{\boldsymbol{\theta}} \right\| \tag{4.3}$$

by iteration, where L_n is the conditional log-likelihood function

$$L_n(\boldsymbol{\theta}, \mathbf{a}) = \sum_{i=1}^{n} \log f\left(Y_i;\ \mathbf{X}_i,\ \boldsymbol{\theta},\ \mathbf{x}_{i,1}^{\mathrm{T}} \mathbf{a}_1(U_i),\ \cdots,\ \mathbf{x}_{i,\ell}^{\mathrm{T}} \mathbf{a}_\ell(U_i)\right), \tag{4.4}$$

where $\mathbf{a}(\cdot) = \left(\mathbf{a}_1(\cdot)^{\mathrm{T}}, \cdots, \mathbf{a}_\ell(\cdot)^{\mathrm{T}}\right)^{\mathrm{T}}$. We describe the details of profile likelihood estimation as follows.

(a) Initialize $\boldsymbol{\theta}$ by a proper guess $\tilde{\boldsymbol{\theta}}_{PR}^{(0)}$. Set $k = 1$.

(b) Maximizing

$$\sum_{i=1}^{n} K_{h_1}(U_i - u) \log f\left(Y_i; X_i, \tilde{\boldsymbol{\theta}}_{PR}^{(k-1)}, \mathbf{x}_{i,1}^{T}\left\{\mathbf{a}_1 + \mathbf{b}_1(U_i - u)\right\}, \cdots, \mathbf{x}_{i,\ell}^{T}\left\{\mathbf{a}_\ell + \mathbf{b}_\ell(U_i - u)\right\}\right), \tag{4.5}$$

with respect to $\left(\mathbf{a}_1^T, \mathbf{b}_1^T, \cdots, \mathbf{a}_\ell^T, \mathbf{b}_\ell^T\right)^T$ we get the maximizer $\left(\tilde{\mathbf{a}}_1^{(k)^{\mathrm{T}}}, \tilde{\mathbf{b}}_1^{(k)^T}, \cdots, \tilde{\mathbf{a}}_\ell^{(k)^T}, \tilde{\mathbf{b}}_\ell^{(k)^T}\right)^T$. The estimator of $\mathbf{a}_{j,\boldsymbol{\theta}}(\cdot)$ is taken to be $\tilde{\mathbf{a}}_j^{(k)}(\cdot)$, $j = 1, \cdots, \ell$.

(c) Estimate $\boldsymbol{\theta}$ by minimizing

$$\left\| \frac{\partial L_n}{\partial \boldsymbol{\theta}}\left(\boldsymbol{\theta}, \tilde{\mathbf{a}}_{\boldsymbol{\theta}}^{(k)}\right) + \frac{\partial L_n}{\partial \mathbf{a}}\left(\boldsymbol{\theta}, \tilde{\mathbf{a}}_{\boldsymbol{\theta}}\right) \frac{\partial}{\partial \boldsymbol{\theta}} \tilde{\mathbf{a}}_{\boldsymbol{\theta}}^{(k)} \right\| \tag{4.6}$$

with respect to $\boldsymbol{\theta}$ we get the maximizer $\tilde{\boldsymbol{\theta}}_{PR}^{(k)}$. The elements of $\frac{\partial}{\partial \boldsymbol{\theta}} \tilde{\mathbf{a}}_{\boldsymbol{\theta}}^{(k)}$ can be estimated by assuming that $\mathbf{a}_{j\boldsymbol{\theta}}$ is a polynomial of $\boldsymbol{\theta}_i$, $i = 1, ..., q$, $j = 1, ..., \ell$.

The estimator of $\boldsymbol{\theta}$ in this step is taken to be $\tilde{\boldsymbol{\theta}}_{PR}^{(k)}$. If $\left\|\tilde{\boldsymbol{\theta}}_{PR}^{(k)} - \tilde{\boldsymbol{\theta}}_{PR}^{(k-1)}\right\|$ is smaller than a pre-defined tolerance, we say that $\hat{\boldsymbol{\theta}}_{PR}^{(k)}$ converges and the estimation procedure is completed. Denote the final estimates $\tilde{\boldsymbol{\theta}}_{PR} = \tilde{\boldsymbol{\theta}}_{PR}^{(k)}$. Otherwise, change k to $k+1$ and go to (b).

Let $\nu^* = \mathbf{a}'_{\boldsymbol{\theta}_0}(\cdot) = \frac{\partial}{\partial \boldsymbol{\theta}} \mathbf{a}_{\boldsymbol{\theta}}(\cdot)\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$ be an $l \times q$ matrix. If ν^* satisfies

$$E_0\left(\frac{\partial L}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}_0, \mathbf{a}_0) + \frac{\partial L}{\partial \mathbf{a}}(\boldsymbol{\theta}_0, \mathbf{a}_0)\,\nu^*\right)^T \left(\frac{\partial L}{\partial \mathbf{a}}(\boldsymbol{\theta}_0, \mathbf{a}_0)\,\nu\right) = 0$$

for all $\nu \in \Lambda$, where

$$L\big(\boldsymbol{\theta}, \mathbf{a}\big) = \log f\left(Y; \mathbf{X}, \boldsymbol{\theta}, \mathbf{x}_1^{\mathrm{T}}\mathbf{a}_1(U), \cdots, \mathbf{x}_\ell^{\mathrm{T}}\mathbf{a}_\ell(U)\right),$$

$\frac{\partial L}{\partial \mathbf{a}}(\boldsymbol{\theta}_0, \mathbf{a}_0)$ is a $1 \times l$ vector and denotes the partial derivative of $L(\boldsymbol{\theta}, \mathbf{z})$ with respect to $\mathbf{z}$ evaluated at the true values $(\boldsymbol{\theta}_0, \mathbf{a}_0)$, Λ denotes the space of $\mathbf{a}$, and E_0 is the expectation taken under the true parameters $\boldsymbol{\theta}_0$ and $\mathbf{a}_0$, then $\mathbf{a}_{\boldsymbol{\theta}}(\cdot)$ are called the least favorable curves. If the least favorable curves exist and with some regularity conditions, it can be shown in Theorem 2 that $\tilde{\boldsymbol{\theta}}_{PR}$ has $n^{-1/2}$ convergence rate if the bandwidth h used in (4.5) satisfies $h \propto n^{-1/5}$.

When the specified semiparametric model is generally like (1.1), in which $\boldsymbol{\theta}$ may involve shape or scale parameters in f, stability of the iteration relies heavily on the proper choice of the initial estimate. Under semiparametric models for the regression mean, Fan and Huang (2005) and Lam and Fan (2008) used difference-based methods to obtain a reliable initial estimate. But, difference-based methods

may not work for model (1.1), because some of the elements in $\boldsymbol{\theta}$ can be other than mean parameters. We propose a new initial estimate for the backfitting and and profile likelihood procedures as follows.

First, we derive some rough estimates of $\mathbf{a}_j(U_i)$, $i = 1, \cdots, n, j = 1, \cdots, \ell$. Consider a model obtained by replacing $\boldsymbol{\theta}$ in (1.1) with $\mathbf{a}_0(U)$, a q-dimensional unknown function of U. This model is now a fully nonparametric model and the functional parameters can be estimated by regular local likelihood approach as follows. For any given u, let $\left(\bar{\mathbf{a}}_0(u)^{\mathrm{T}}, \bar{\mathbf{b}}_0(u)^{\mathrm{T}}, \bar{\mathbf{a}}_1(u)^{\mathrm{T}}, \bar{\mathbf{b}}_1(u)^{\mathrm{T}}, \cdots, \bar{\mathbf{a}}_\ell(u)^{\mathrm{T}}, \bar{\mathbf{b}}_\ell(u)^{\mathrm{T}}\right)^{\mathrm{T}}$ be the maximizer, with respect to $\left(\mathbf{a}_0^{\mathrm{T}}, \mathbf{b}_0^{\mathrm{T}}, \mathbf{a}_1^{\mathrm{T}}, \mathbf{b}_1^{\mathrm{T}}, \cdots, \mathbf{a}_\ell^{\mathrm{T}}, \mathbf{b}_\ell^{\mathrm{T}}\right)^{\mathrm{T}}$, of the local log-likelihood function

$$\sum_{i=1}^{n} K_{h_1}(U_i - u) \log f\Big(Y_i; \mathbf{X}_i, \mathbf{a}_0 + \mathbf{b}_0(U_i - u), \mathbf{x}_{i,1}^{\mathrm{T}}\big\{\mathbf{a}_1 + \mathbf{b}_1(U_i - u)\big\}, \cdots, \mathbf{x}_{i,\ell}^{\mathrm{T}}\big\{\mathbf{a}_\ell + \mathbf{b}_\ell(U_i - u)\big\}\Big).$$

Here h_1 can be taken as the bandwidth $\hat{h}_1$ in Section 5.2, because it is selected for local likelihood estimation by assuming model (5.2). Letting $u = U_i$ in the foregoing procedure, we have $\bar{\mathbf{a}}_j(U_i)$, $j = 1, \cdots, \ell, i = 1, \cdots, n$. Then our initial estimate $\bar{\boldsymbol{\theta}}$ is the maximizer of

$$\sum_{i=1}^{n} \log f\Big(Y_i; \mathbf{X}_i, \boldsymbol{\theta}, \mathbf{x}_{i,1}^{\mathrm{T}} \bar{\mathbf{a}}_1(U_i), \cdots, \mathbf{x}_{i,\ell}^{\mathrm{T}} \bar{\mathbf{a}}_\ell(U_i)\Big).$$

During the iteration in finding the minimizer of (4.6), $\tilde{\mathbf{a}}_{\boldsymbol{\theta}}(\cdot)$ is taken to be the estimator that solves (4.5) with h_1 replaced by $\hat{h}_1$. With this choice of bandwidth, the least favorable curve is well approximated, by the nature of model (5.2). On convergence of the iteration, we obtain the profile likelihood estimator for $\boldsymbol{\theta}$. Then

we can estimate $\mathbf{a}(\cdot)$ and select the bandwidth in the same manner as described later in Sections 4.3 and 5.2 with $\hat{\boldsymbol{\theta}}$ replaced by the profile likelihood estimator for $\boldsymbol{\theta}$.

Unlike backfitting, the profile likelihood estimation does not need to undersmooth the estimates of functional parameters. However, the profile likelihood estimation requires the least favorable curve assumption and more assumptions of $\frac{\partial}{\partial \boldsymbol{\theta}} \mathbf{a}_{\boldsymbol{\theta}}(\cdot)$ to attain $\sqrt{n}$ consistency, which is not always satisfied for all models. For example, as mentioned in Fan and Wong (2000), if Y is from $N(\mu(\cdot), \sigma^2)$, then the profile likelihood estimator of σ^2 is not consistent. This restricts the application of profile likelihood estimation. Furthermore, the profile likelihood approach also suffers some drawbacks as backfitting does. First, the bandwidth h used in (4.5) is difficult to select automatically, especially when the initialization $\hat{\boldsymbol{\theta}}_{PR}^{(0)}$ is far from the true value of $\boldsymbol{\theta}$. In fact, the iteration may not converge under this situation even the bandwidth is correctly specified. Second, the profile likelihood approach requires more computation on estimating $\frac{\partial}{\partial \boldsymbol{\theta}} \mathbf{a}_{\boldsymbol{\theta}}(\cdot)$ so is even more computationally intensive. Finally, the iteration may also diverge when the design of U is sparse.

4.3 Two-step estimation

Our two-step approach first produces an estimator for the constant vector $\boldsymbol{\theta}$, then plugs this estimator into the local likelihood function to estimate the functions $\mathbf{a}_j(\cdot)$, $j = 1, \cdots, \ell$.

The estimation procedure for $\boldsymbol{\theta}$ consists of two stages. First, we treat $\boldsymbol{\theta}$ as

an unknown function of U and appeal to the local likelihood approach to get a preliminary estimator $\tilde{\boldsymbol{\theta}}(U_i)$ for $\boldsymbol{\theta}(U_i)$ for each U_i, $i = 1, \cdots, n$. Then we average $\tilde{\boldsymbol{\theta}}(U_i)$ over $i = 1, \cdots, n$ to get the final estimator for $\boldsymbol{\theta}$. The procedure is as follows. Consider the model that specifies the conditional density of Y given $\mathbf{X}$ and U as:

$$f\Big(Y;\ \mathbf{X},\ \boldsymbol{\theta}(U),\ \mathbf{x}_1^{\mathrm{T}}\mathbf{a}_1(U),\ \cdots,\ \mathbf{x}_\ell^{\mathrm{T}}\mathbf{a}_\ell(U)\Big). \tag{4.7}$$

For any fixed u, by Taylor's expansion, we have, for each j,

$$\mathbf{a}_j(U_i) \approx \mathbf{a}_j(u) + \dot{\mathbf{a}}_j(u)(U_i - u).$$

when U_i is in a neighborhood of u, where $\dot{\mathbf{a}}_j(u) = d\mathbf{a}_j(u)/du$. This leads to the following local log-likelihood function:

$$\sum_{i=1}^{n} K_h(U_i - u) \log f\Big(Y_i;\ \mathbf{X}_i,\ \boldsymbol{\theta},\ \mathbf{x}_{i,1}^{\mathrm{T}}\big\{\mathbf{a}_1 + \mathbf{b}_1(U_i - u)\big\},\ \cdots,\ \mathbf{x}_{i,\ell}^{\mathrm{T}}\big\{\mathbf{a}_\ell + \mathbf{b}_\ell(U_i - u)\big\}\Big), \tag{4.8}$$

where $K_h(\cdot) = K(\cdot/h)/h$, $K(\cdot)$ is a kernel function, and $h > 0$ is a bandwidth. Maximizing (4.8) with respect to $\big(\boldsymbol{\theta}^{\mathrm{T}}, \mathbf{a}_1^{\mathrm{T}}, \mathbf{b}_1^{\mathrm{T}}, \cdots, \mathbf{a}_\ell^{\mathrm{T}}, \mathbf{b}_\ell^{\mathrm{T}}\big)^{\mathrm{T}}$ we get the maximizer $\Big(\tilde{\boldsymbol{\theta}}(u)^{\mathrm{T}}, \tilde{\mathbf{a}}_1(u)^{\mathrm{T}},\ \tilde{\mathbf{b}}_1(u)^{\mathrm{T}}, \cdots, \tilde{\mathbf{a}}_\ell(u)^{\mathrm{T}}, \tilde{\mathbf{b}}_\ell(u)^{\mathrm{T}}\Big)^{\mathrm{T}}$. In the foregoing local likelihood estimation, $\boldsymbol{\theta}$ is fitted by a local constant vector, because $\boldsymbol{\theta}$ is constant under model (1.1) and fitting it by a local constant vector stabilizes the procedure. For $i = 1, \cdots, n$, let $u = U_i$; we get an initial estimator $\tilde{\boldsymbol{\theta}}(U_i)$ of $\boldsymbol{\theta}$. The final estimator of $\boldsymbol{\theta}$ is taken to be

$$\hat{\boldsymbol{\theta}} = n^{-1} \sum_{i=1}^{n} \tilde{\boldsymbol{\theta}}(U_i). \tag{4.9}$$

For $\hat{\boldsymbol{\theta}}$ to achieve the $n^{-1/2}$ convergence rate, we need to choose a relatively small bandwidth h so that the biases of $\tilde{\boldsymbol{\theta}}(\cdot)$ and $\tilde{\mathbf{a}}_j(\cdot)$, $j = 1, \cdots, \ell$, are dominated by $n^{-1/2}$. This ensures that estimating the constant and the functional parts simultaneously in the first step does not create extra bias for $\boldsymbol{\theta}$. Then, averaging over $\tilde{\boldsymbol{\theta}}(U_i)$, $i = 1, \cdots, n$, as in (4.9) brings the variance from the order $(nh)^{-1}$ in nonparametric estimation back to the order n^{-1} in parametric estimation. Later, we show that $\hat{\boldsymbol{\theta}}$ is root-n consistent when h is chosen properly. Like any other maximum local likelihood estimation procedure, the bandwidth h cannot be chosen too small, or otherwise one runs into problems with singularity of the design matrix. From an asymptotic standpoint, condition (S5) keeps the bandwidth h from being too small; thus, conditions (S5) and (S7) guarantee that the estimators $\tilde{\boldsymbol{\theta}}(U_1), \cdots, \tilde{\boldsymbol{\theta}}(U_n)$ exist. Furthermore, the method of Cheng and Wu (2008) can be used to modify the local likelihood function (4.8) to overcome the singularity problem caused by a small h or sparsity in the design points U_i's. This approach also can be applied to (4.10) when estimating the function $\mathbf{a}(u)$.

With $\hat{\boldsymbol{\theta}}$, we can estimate $\mathbf{a}(u)$ using the maximum local likelihood approach. Note that the estimator $\tilde{\mathbf{a}}(u) = \left(\tilde{\mathbf{a}}_1(u)^{\mathrm{T}}, \cdots, \tilde{\mathbf{a}}_\ell(u)^{\mathrm{T}}\right)^{\mathrm{T}}$ that we obtained before is too noisy and is not appropriate for this purpose, because the bandwidth h is intentionally chosen to be small to get a good estimator of $\boldsymbol{\theta}$. Thus we use another, larger bandwidth to estimate $\mathbf{a}(u)$. We replace $\boldsymbol{\theta}$ in (4.8) by $\hat{\boldsymbol{\theta}}$ to get a local log-likelihood

function for $\mathbf{a}(u)$,

$$\sum_{i=1}^{n} K_{h_1}(U_i - u) \log f\Big(Y_i;\ \mathbf{X}_i,\ \hat{\boldsymbol{\theta}},\ \mathbf{x}_{i,1}^{\mathrm{T}}\big\{\mathbf{a}_1 + \mathbf{b}_1(U_i - u)\big\},\ \cdots,\ \mathbf{x}_{i,\ell}^{\mathrm{T}}\big\{\mathbf{a}_\ell + \mathbf{b}_\ell(U_i - u)\big\}\Big), \tag{4.10}$$

where $h_1 > 0$ is a bandwidth different from h. We could use a kernel other than K at this step, but this does not matter much. Maximizing (4.10) with respect to $\big(\mathbf{a}_1^T, \mathbf{b}_1^T, \cdots, \mathbf{a}_\ell^T, \mathbf{b}_\ell^T\big)^T$, we get the maximizer $\big(\hat{\mathbf{a}}_1(u)^T, \hat{\mathbf{b}}_1(u)^T, \cdots, \hat{\mathbf{a}}_\ell(u)^T, \hat{\mathbf{b}}_\ell(u)^T\big)^T$. Our estimator of $\mathbf{a}(u)$ is taken to be $\hat{\mathbf{a}}(u) = \big(\hat{\mathbf{a}}_1(u)^{\mathrm{T}}, \cdots, \hat{\mathbf{a}}_\ell(u)^{\mathrm{T}}\big)^{\mathrm{T}}$. Because the convergence rate of $\hat{\boldsymbol{\theta}}$ is $n^{-1/2}$ (see Sec. 6), $\hat{\mathbf{a}}(u)$ would work as well as when $\boldsymbol{\theta}$ is known and is used in the local log-likelihood (4.10); that is, $\hat{\mathbf{a}}(u)$ has the adaptivity property.

In some cases, local likelihood estimation of the varying coefficients $\mathbf{a}_j(\cdot)$, $j = 1, \cdots, \ell$, may require a different amount of smoothing (see, e.g., Claeskens and Aerts 2000). Backfitting ideas can be implemented to achieve this goal, as follows: (a) Use $\hat{\mathbf{a}}(\cdot)$ as the initial estimate; (b) for each j, substitute all of the local linear coefficient functions except the jth and h_1 in (4.10) by the previous estimates and use the bandwidth for smoothing the jth functional parameter, and then maximize the resulted local likelihood to find an estimate of $\mathbf{a}_j(\cdot)$; and (c) iterate step (b) until convergence. Convergence usually is attained quickly in this case.

5 Bandwidth selection and identifying constant parameters

In reality, we do not know which of the parameters are constant and which are functional in model (1.1). This is essentially a model selection problem. The problem can be formulated in the form of successive tests of null hypotheses against multiple alternative hypotheses, and actually only one of the alternative hypotheses is the one we are looking for. Thus even if we construct a test statistics, choosing an appropriate threshold is challenging. To avoid this troublesome issue, information-criteria-based model selection procedures are often used.

There are many model selection criteria under parametric assumptions, including cross-validation (Stone 1974), the AIC (Akaike 1970), the BIC (Schwarz 1978), and nonconcave penalized likelihood (Fan and Li 2001). Of these various criteria, the AIC and BIC are likely the most commonly used in practice, because of their easy implementation. We use the concepts of the AIC and BIC to select the bandwidths h_1 and h in the estimation procedures and to identify the constant parameters in model (1.1).

5.1 Model selection criteria

5.1.1 Akaike Information Criterion (AIC)

The Kullback-Leibler information (Kullback and Leibler 1951) is a widely used distance to measure the similarity between two probability distributions. For two given probability distributions with density functions g and f, the Kullback-Leibler information is defined by

$$D(g, f) = \int g(y) \log \frac{g(y)}{f(y)} dy.$$

Let's start from the parametric case. Assume M_0 denotes the true but unknown model with density function g and M_k is a candidate model with density function $f(\cdot|\boldsymbol{\theta})$. Let M be the collection of candidate models. Our goal is to seek an M_k in M such that $D(g, f(\cdot|\theta))$ is minimized. Let $\boldsymbol{\theta}=(\theta_1, ..., \theta_{\mathcal{K}})$ for some $\mathcal{K}$. By definition,

$$D(g, f(\cdot|\boldsymbol{\theta})) = \int g(y) \log \frac{g(y)}{f(y|\boldsymbol{\theta})} dy = E_Y[\log g(Y)] - E_Y[\log f(Y|\boldsymbol{\theta})],$$

where $E_Y[\log g(Y)]$ is actually a constant. Hence, minimizing $D(g, f(\cdot|\boldsymbol{\theta}))$ is equivalent to minimize $-E_Y[\log f(Y|\boldsymbol{\theta})]$. In practice we don't know the true value of $\boldsymbol{\theta}$, denoted by $\boldsymbol{\theta}_0$, we replace it by $\hat{\boldsymbol{\theta}}$, which is the mle of $\boldsymbol{\theta}$, and use $-E_{\hat{\boldsymbol{\theta}}}\left[E_Y\left[\log f\left(Y|\hat{\boldsymbol{\theta}}\right)\right]\right]$ to estimate $-E_Y[\log f(Y|\boldsymbol{\theta}_0)]$. That is, instead of minimizing $-E_Y[\log f(Y|\boldsymbol{\theta}_0)]$, we minimize its estimate $-E_{\hat{\theta}}\left[E_Y\left[\log f\left(Y|\hat{\boldsymbol{\theta}}\right)\right]\right]$. By Taylor expansion we can easily derive that

$$-E_{\hat{\boldsymbol{\theta}}}\left[E_Y\left[\log f\left(Y|\hat{\boldsymbol{\theta}}\right)\right]\right] \simeq -E_Y\left[\log f\left(Y|\hat{\boldsymbol{\theta}}\right)\right] + tr\left(I\left(\boldsymbol{\theta}_0\right)\Sigma\right),$$

where $I(\boldsymbol{\theta}_0) = E_Y\left[-\frac{\partial^2 \log f(Y|\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$, $i, j = 1, ..., \mathcal{K}$ is a $\mathcal{K} \times \mathcal{K}$ matrix and Σ is the covariance matrix of $\hat{\boldsymbol{\theta}}$. Since the log-likelihood $\log f\left(Y|\hat{\boldsymbol{\theta}}\right)$ is naturally an unbiased estimator of $E_Y\left[\log f\left(Y|\hat{\boldsymbol{\theta}}\right)\right]$, we have

$$-E_{\hat{\boldsymbol{\theta}}}\left[E_Y\left[\log f\left(Y|\hat{\boldsymbol{\theta}}\right)\right]\right] \simeq -\log f\left(Y|\hat{\boldsymbol{\theta}}\right) + tr\left(I\left(\boldsymbol{\theta}_0\right)\Sigma\right). \tag{5.1}$$

Moreover, Akaike (1973) showed that if $f \simeq g$, then $I(\boldsymbol{\theta}_0) \simeq \Sigma^{-1}$ and thus $tr\left(I\left(\boldsymbol{\theta}_0\right)\Sigma\right) \simeq \mathcal{K}$. Multiply (5.1) by 2 and approximate $tr\left(I\left(\boldsymbol{\theta}_0\right)\Sigma\right)$ by $\mathcal{K}$ we can obtain the definition of AIC:

$$AIC = -2\log f\left(Y|\hat{\boldsymbol{\theta}}\right) + 2\mathcal{K}.$$

Now return to our model. Based on the standard AIC formula, we have the following version of AIC for model (1.1):

$$\text{AIC} = -2\sum_{i=1}^{n} \log f\Big(Y_i;\ \mathbf{X}_i,\ \hat{\boldsymbol{\theta}},\ \mathbf{x}_{i,1}^{\mathrm{T}}\hat{\mathbf{a}}_1(U_i),\ \cdots,\ \mathbf{x}_{i,\ell}^{\mathrm{T}}\hat{\mathbf{a}}_\ell(U_i)\Big) + 2\mathcal{K}\,.$$

To work out $\mathcal{K}$, we have to determine how many unknown parameters each unknown function $a_{ij}(\cdot)$ amounts to. In nonparametric modeling, when a locally polynomial approximation is used, Fan and Gijbels (1996) suggested that an unknown function amounts to

$$tr\left\{\left(G^T W_0 G\right)^{-1} G^T W_0^2 G\right\}$$

unknown parameters, where

$$G = \begin{pmatrix} 1 & U_1 - u \\ \vdots & \vdots \\ 1 & U_n - u \end{pmatrix}, \; W_0 = \text{diag}\left(K_{h_1}\left(U_1 - u\right), \cdots, K_{h_1}\left(U_n - u\right)\right).$$

To make it more easy to compute, we can look into its asymptotic version. When the sample size n is large enough, we have

$$tr\left\{\left(G^T W_0 G\right)^{-1} G^T W_0^2 G\right\} \simeq h_1^{-1}\left(\nu_0 + \nu_2/\mu_2\right),$$

where $\nu_i = \int t^i K^2(t)\, dt$, $\mu_i = \int t^i K(t)\, dt$. Further, $\nu_0 + \nu_2/\mu_2 = 1.028571$ when the Epanechnikov kernel $K(t) = 0.75\left(1 - t^2\right)_+$ is used. Thus, for our case $\mathcal{K} = q + 1.028571\left(p_1 + \cdots + p_l\right) h_1^{-1}$ if we use the Epanechnikov kernel in our estimation procedure.

5.1.2 Bayesian Information Criterion (BIC)

Assume $m(M_k)$ denotes the prior distribution of some model M_k and $\pi(\boldsymbol{\theta}|M_k)$ be the prior distribution of $\boldsymbol{\theta}$ given the model M_k. We can then obtain the posterior distribution of model M_k:

$$P(M_k|Y) = \int \frac{f(Y|M_k, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta}|M_k)\, m(M_k)}{h(Y)} d\boldsymbol{\theta},$$

where $h(Y)$ denotes the marginal density of Y, which is irrelevant of the model. A reasonable choice is to choose the model which maximizes the posterior probability. Further, Schwarz (1978) showed that under some regularity conditions of f,

maximizing the posterior is asymptotically equivalent to minimizing BIC:

$$BIC = -2 \log f\left(Y|\hat{\boldsymbol{\theta}}\right) + \log(n)\,\mathcal{K}.$$

In parametric cases, the major advantage of BIC is its consistency property. That is, if the true model is contained in the candidate set, BIC will select the true model with probability 1 as the sample size $n \to \infty$. If the true model is not selected, BIC tends to select simpler models since it penalize more on model complexity. Hence the prediction error may be larger (compared to the result of AIC) if the true model is not selected. That is, BIC does not serve the efficiency property. See Burnham and Aderson (2003) for more details.

Since the only difference between AIC and BIC is the penalty on model complexity, we can also obtain the version of BIC for model (1.1)

$$\text{BIC} = -2 \sum_{i=1}^{n} \log f\left(Y_i;\ \mathbf{X}_i,\ \hat{\boldsymbol{\theta}},\ \mathbf{x}_{i,1}^{\mathrm{T}}\hat{\mathbf{a}}_1(U_i),\ \cdots,\ \mathbf{x}_{i,\ell}^{\mathrm{T}}\hat{\mathbf{a}}_\ell(U_i)\right) + \mathcal{K}\log(n)\,.$$

The AIC and BIC formulas can be applied to any models of the form (1.1), which can have different q, ℓ or $\mathbf{x}_j$, and model (5.2). In the latter case, $\mathcal{K} = h_1^{-1}(\nu_0 + \nu_2/\mu_2)(q + p_1 + \cdots + p_\ell)$ and $\hat{\boldsymbol{\theta}}$, $\hat{\mathbf{a}}_j(\cdot)$, $j = 1, \cdots, \ell$, are replaced by $\bar{\mathbf{a}}_j(\cdot)$, $j = 0, 1, \cdots, \ell$, in the formulas.

5.2 Bandwidth selection

Suppose that (1.1) is the true underlying model, and it is used to analyze the data. The choice of bandwidths h and h_1 determines the performance of the two-

step estimators described in Section 4.3. Compared with choosing h_1, choosing h is relatively simple, because h is used to get undersmoothed estimators of the functional parameters. It follows from Theorem 3 that we get a good estimator of $\boldsymbol{\theta}$ by letting h be of order $n^{-\alpha}$ for any $\alpha \in (1/4, 1)$. In practice, we may take $\alpha = 1/4 + \delta$ for a small $\delta > 0$ to avoid difficulties in the maximization of (4.8) caused by design sparsity. Proper selection of h_1 is crucial for $\hat{\mathbf{a}}(\cdot)$ to perform well. We propose first obtaining a reasonable choice of h_1, using the relationship between the optimal rate of h_1 and a suitable rate of h to determine h, and, finally, selecting h_1.

To get a reasonable choice of h_1, we compute the version of AIC for different values of h_1 under the model:

$$f\Big(Y;\ \mathbf{X},\ \mathbf{a}_0(U),\ \mathbf{x}_1^{\mathrm{T}}\mathbf{a}_1(U),\ \cdots,\ \mathbf{x}_\ell^{\mathrm{T}}\mathbf{a}_\ell(U)\Big). \tag{5.2}$$

This yields an AIC function of h_1 only; h is not involved, because there are no constant parameters in (5.2). Then $\hat{h}_1$, the minimizer of the AIC function of h_1, is a rough approximation to the optimal value of h_1 in the two-step estimator $\hat{\mathbf{a}}(\cdot)$ for $\mathbf{a}(\cdot)$ in (1.1). The reason for this is that when data generated from (1.1) are modeled by (5.2), the true value of $\mathbf{a}_0(\cdot)$ is the constant vector $\boldsymbol{\theta}$, so the curve estimate of $\mathbf{a}_0(\cdot)$ is roughly flat for a wide range of h_1, and the AIC criterion for (5.2) measures mainly the performance of the estimators of $\mathbf{a}_j(\cdot)$, $j = 1, \cdots, \ell$, while h_1 varies. Here we use AIC to select the bandwidth since it is an approximation of the expected Kullback-Leibler information. When the model is fixed, the Kullback-Leibler information

measures the distance between the probability distribution of the true underlying model and the probability distribution with the estimated parameters under the specified model. Hence, the bandwidth that minimizes AIC generates the best estimates of the parameters under a specified model in the sense of the Kullback-Leibler information.

As discussed earlier, selection of the bandwidth h in the two-step estimation of $\boldsymbol{\theta}$ is not a major issue. Any bandwidth h will do as long as it is relatively small but not too small. In the light of Theorem 2, which suggests that the optimal rate of h_1 is $n^{-1/5}$, and the discussion on choice of h earlier we take $\hat{h} = n^{-0.051}\hat{h}_1$. Note that value of $n^{-0.051}$ falls in the narrow range (0.5559, 0.7907) for $n \in [10^2, 10^5]$.

We choose the bandwidth $\hat{h}_1$ for estimating the functions $\mathbf{a}_0(\cdot), \mathbf{a}_1(\cdot), \cdots, \mathbf{a}_\ell(\cdot)$ in model (5.2), which specifies the constant vector $\boldsymbol{\theta}$ in the true model (1.1) as functional. We refine our data-driven selection of h_1, which is required in the two-step estimation of $\mathbf{a}(\cdot)$ in the true model (1.1). Based $\hat{h}$, we obtain the two-step estimator $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$ in (1.1): (a) Plug-in $\hat{\boldsymbol{\theta}}$ into (4.10) to find $\hat{\mathbf{a}}(\cdot)$, and (b) compute the AIC criterion for model (1.1) for a range of h_1. We denote the minimizer of this AIC function by $\tilde{h}_1$.

5.3 Identifying constant parameters

We propose a procedure to identify which parameters are constant and which are functional in model (1.1) based on the BIC criterion. This model selection prob-

lem interacts with the bandwidth selection problem; the BIC formula depends on the bandwidth h_1. In fact, choosing the bandwidth and the constant parameters simultaneously is almost impossible, because either a complex model or a small bandwidth can result in a small bias and a large variance, and either a simple model or a large bandwidth can result in a large bias and a small variance. Thus a complex model with a large bandwidth would have same effects as a simple model with a small bandwidth. A sensible solution is to first choose the bandwidths h_1 and h and then identify the constant parameters.

We start with a model $\mathcal{M}_0$ of the form (5.2), and then determine which parameters in $\mathcal{M}_0$ are functional and which are constant. The choice of $\mathbf{a}_0(U)$ and $\mathbf{x}_1, \cdots, \mathbf{x}_\ell$, and knowledge of how they determine the dependence of Y on $\mathbf{X}$ and U in $\mathcal{M}_0$, should come from the basic assumptions on the model; in practice, they are determined by the analyst. Because of the curse of dimensionality issue, we need to impose some basic assumptions on the model based on some knowledge about the data that we are analyzing, which usually is available from the background of the data or people working in the area where the data arise. Because all of the unknown parameters in $\mathcal{M}_0$ are functions, as in Section 5.2, we choose the bandwidth h_1 for estimating the unknown functions by minimizing the version of the AIC for model $\mathcal{M}_0$. For simplicity of notation, we denote this bandwidth by $\hat{h}_1$ and again let $\hat{h} = \hat{h}_1 n^{-0.051}$. Then we fix at these two bandwidths throughout the model selection procedure.

Ideally, we could compute the BICs for all possible combinations, and the chosen combination would be the one with the smallest BIC value. Unfortunately, however, this approach would immediately become computationally impossible when κ (the number of parameters that can be either functional or constant) is not very small because there are 2^{κ} possible combinations. We propose the following iterative procedure to reduce the computational burden. We start with $\mathcal{M}_0$ as the candidate model and at the $\mathcal{L}$th step of the iteration we examine whether one of the functional parameters in the candidate model $\mathcal{M}_{\mathcal{L}}$ can be further reduced to a constant.

(a): Set $\mathcal{L} = 0$. Based on model $\mathcal{M}_0$, compute local likelihood estimates of all of the unknown parameter functions using bandwidth $\hat{h}_1$.

(b): If $\mathcal{L} = \kappa$ (i.e., all of the κ parameters are reduced to constants in $\mathcal{M}_{\mathcal{L}}$) then $\mathcal{M}_{\mathcal{L}}$ is the chosen model, and model selection is completed. Otherwise, for each of the unknown functions in the candidate model $\mathcal{M}_{\mathcal{L}}$, say $a_{ij}(\cdot)$, that could be reduced to a constant, calculate

$$S_{ij} = \sum_{k=1}^{n} \Big(\hat{a}_{ij}(U_k) - \bar{a}_{ij} \Big)^2, \quad \bar{a}_{ij} = n^{-1} \sum_{k=1}^{n} \hat{a}_{ij}(U_k).$$

Changing the function $a_{ij}(\cdot)$ in $\mathcal{M}_{\mathcal{L}}$ that has the smallest S_{ij} to a constant parameter results in a new model, $\mathcal{M}_{\mathcal{L}+1}$.

(c): Based on the new model $\mathcal{M}_{\mathcal{L}+1}$ and the bandwidths $\hat{h}_1$ and $\hat{h}$, compute the estimates of the unknown functions and constants. Compute the BIC of $\mathcal{M}_{\mathcal{L}+1}$ and compare it with that of $\mathcal{M}_{\mathcal{L}}$. If $\mathcal{M}_{\mathcal{L}}$ has a smaller BIC, then $\mathcal{M}_{\mathcal{L}}$ is the

chosen model and the model selection is completed. Otherwise, $\mathcal{M}_{\mathcal{L}+1}$ becomes the candidate model; thus we denote the new constant parameter in (b) as $\theta_{\mathcal{L}+1}$ and change $\mathcal{L}$ to $\mathcal{L}+1$ then go to (b).

The foregoing iterative process continues until $\mathcal{M}_{\mathcal{L}}$ has a smaller BIC than $\mathcal{M}_{\mathcal{L}+1}$ for some $\mathcal{L} < \kappa$ (i.e., $\mathcal{M}_{\mathcal{L}}$ is the chosen model) or until $\mathcal{L} = \kappa$ (i.e., the chosen model has all of the considered parameters constant). Apparently, the final chosen model can be written in the form of (1.1).

6 Asymptotic properties

In this section we investigate asymptotic distributions of the backfitting estimators given in Section 4.1, the profile likelihood estimators given in Section 4.2, and the two-step estimators given in Section 4.3.

For simplicity of notation, the theory presented here concerns the case with $\mathbf{x}_j = \mathbf{X}$, $j = 1, \cdots, \ell$. The established theory straightforwardly carries over to the general case where $\mathbf{x}_1, \cdots, \mathbf{x}_\ell$, are different. Let $\pi(u)$ be the density of U and let $\ddot{\mathbf{a}}_j(u)$ be the second derivative of $\mathbf{a}_j(u)$, $j = 1, \cdots, \ell$. Write

$$\mathbf{z} = (z_1, \cdots, z_\ell)^{\mathrm{T}},\ z_j = \mathbf{X}^{\mathrm{T}}\mathbf{a}_j(u),\ j = 1, \cdots, \ell,\ \mathbf{D} = \mathbf{I}_\ell \otimes (\mathbf{X}^{\mathrm{T}},\ \mathbf{0}_{1\times p})^{\mathrm{T}},\ \mathbf{D}_c = \mathbf{I}_\ell \otimes (\mathbf{0}_{1\times p},\ \mathbf{X}^{\mathrm{T}})^{\mathrm{T}}.$$

Theorem 1 and 2 gives the asymptotic distribution of $\hat{\boldsymbol{\theta}}_{BF}$ and $\hat{\boldsymbol{\theta}}_{PR}$. Note that the backfitting and profiling procedure produce estimators with the same asymptotic distribution. The backfitting procedure requires that undersmoothing be used to estimate $\hat{\mathbf{a}}(U)$, whereas the profiling procedure does not.

Theorem 1. *Assume that the regularity conditions (S2)–(S4) and (BF1)–(BF5) stated in Appendix A and C hold, and that the bandwidth h satisfies $nh^4 \to 0$ and not $h \propto n^{-1/5}$, then we have*

$$n^{1/2}\left(\hat{\boldsymbol{\theta}}_{BF} - \boldsymbol{\theta}\right) \xrightarrow{D} N\left(\mathbf{0}_{q\times 1},\ \mathcal{G}^{-1}(\boldsymbol{\theta}_0)\Sigma_1\mathcal{G}^{-1}(\boldsymbol{\theta}_0)\right) \quad \textit{when } n \longrightarrow \infty,$$

where

$$\mathcal{G}(\boldsymbol{\theta}) = \frac{d}{d\boldsymbol{\theta}} E\left[\frac{\partial}{\partial \boldsymbol{\theta}} L\left(\boldsymbol{\theta}, \mathbf{a}_{0\boldsymbol{\theta}_0}(U)\right)\right],$$

$$\Sigma_1 = cov\left[\frac{\partial}{\partial\boldsymbol{\theta}}L\left(\boldsymbol{\theta}_0, \mathbf{a}_{0\boldsymbol{\theta}_0}\right) + \frac{\partial}{\partial\mathbf{a}}L\left(\boldsymbol{\theta}_0, \mathbf{a}_{0\boldsymbol{\theta}_0}\right)\frac{\partial}{\partial\boldsymbol{\theta}}\mathbf{a}_{0\boldsymbol{\theta}_0}(U)\right]$$

Theorem 2. *Assume the regularity conditions (PR1)–(PR4) stated in Appendix B and allow that $h \propto n^{-1/5}$, then we have*

$$n^{1/2}\left(\hat{\boldsymbol{\theta}}_{PR} - \boldsymbol{\theta}\right) \xrightarrow{D} N\left(\mathbf{0}_{q\times 1},\ \mathcal{G}^{-1}(\boldsymbol{\theta}_0)\Sigma_1\mathcal{G}^{-1}(\boldsymbol{\theta}_0)\right) \quad \textit{when } n \longrightarrow \infty,$$

The asymptotic distributions of the proposed 2-step estimators are discussed by Cheng et al. (2009). We state them in Theorem 3 and 4 as follows. Theorem 3 gives the asymptotic distribution of $\hat{\boldsymbol{\theta}}$ and shows that $\hat{\boldsymbol{\theta}}$ is asymptotically unbiased as an estimator of the constant parameter $\boldsymbol{\theta}$, provided that the bandwidth h is of an order smaller than that of optimal bandwidths used in univariate smoothing.

Theorem 3. *Under the regularity conditions (S1)–(S7) stated in the Appendix, if $h = o(n^{-1/4})$ and $nh/\log^2 n \longrightarrow \infty$, then we have*

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{D} N(\mathbf{0}_{q\times 1},\ \boldsymbol{\Delta}) \quad \textit{when } n \longrightarrow \infty,$$

where

$$\boldsymbol{\Delta} = (\mathbf{I}_q,\ \mathbf{0}_{q\times 2p\ell})E\big\{\mathbf{V}_c(U)^{-1}\mathbf{V}_0(U)\mathbf{V}_c(U)^{-1}\big\}(\mathbf{I}_q,\ \mathbf{0}_{q\times 2p\ell})^T,$$

$$\mathbf{V}_0(u) = E\big\{\mathbf{H}\boldsymbol{\mathcal{I}}(\boldsymbol{\gamma})\mathbf{H}^T\big|U = u\big\}, \quad \mathbf{V}_c(u) = \mathbf{V}_0(u) + E\big\{\mu_2\mathbf{H}_c\boldsymbol{\mathcal{I}}(\boldsymbol{\gamma})\mathbf{H}_c^T\big|U = u\big\},$$

$$\mathbf{H} = diag\,(\mathbf{I}_q, \mathbf{D}), \quad \mathbf{H}_c = diag\,(\mathbf{0}_{q\times q}, \mathbf{D}_c), \ \boldsymbol{\mathcal{I}}(\boldsymbol{\gamma}) = -E\left\{\dot{\mathbf{g}}(Y;\ \mathbf{X},\ \boldsymbol{\gamma})|\mathbf{X},\ U\right\},$$

$$\mathbf{g}(Y;\ \mathbf{X},\ \boldsymbol{\gamma}) = \frac{\partial \log f(Y;\ \mathbf{X},\ \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}, \quad \dot{\mathbf{g}}(Y;\ \mathbf{X},\ \boldsymbol{\gamma}) = \frac{\partial \mathbf{g}(Y;\ \mathbf{X},\ \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}, \quad \boldsymbol{\gamma} = (\boldsymbol{\theta}^T,\ \mathbf{z}^T)^T.$$

In general, neither the profile likelihood nor the two-step estimator of the constant parameter $\boldsymbol{\theta}$ is consistently superior to the other in their asymptotic performance. The profile likelihood estimator may have a smaller asymptotic variance than the two-step estimator when both are asymptotically normal (Severini and Wong 1992), but on the other hand, there are situations for which Theorem 1 holds but the profile likelihood estimator does not work because it requires existence of the least favorable curves (see the example discussed in Fan and Wong 2000).

Theorem 4. *Under the regularity conditions stated in the Appendix, if* $h_1 \longrightarrow 0$ *and* $nh_1/\log^2 n \longrightarrow \infty$*, then we have*

$$(nh_1)^{1/2}\big\{\hat{\mathbf{a}}(u) - \mathbf{a}(u) + \boldsymbol{\mathcal{B}}\big\} \xrightarrow{D} N(\mathbf{0}_{p\ell\times 1}, \boldsymbol{\Sigma}) \quad \textit{when } n \longrightarrow \infty,$$

where

$$\boldsymbol{\mathcal{B}} = 2^{-1}\mu_2 h_1^2 \mathbf{I}_\ell \otimes \{(1,\ 0) \otimes \mathbf{I}_p\}\mathbf{G}_c^{-1}\boldsymbol{\Gamma}\,,$$

$$\boldsymbol{\Gamma} = E\Big\{\mathbf{D}\boldsymbol{\mathcal{I}}_1(\mathbf{z})\big(\ddot{\mathbf{a}}_1(u),\ \cdots,\ \ddot{\mathbf{a}}_\ell(u)\big)^T\mathbf{X}\Big|U = u\Big\},$$

$$\boldsymbol{\Sigma} = \mathbf{I}_\ell \otimes \{(1,\ 0) \otimes \mathbf{I}_p\}\mathbf{G}_c^{-1}\mathbf{G}\mathbf{G}_c^{-1}\pi(u)^{-1}\mathbf{I}_\ell \otimes \{(1,\ 0)^T \otimes \mathbf{I}_p\}\,,$$

$$\mathbf{G} = E\left\{\nu_0\mathbf{D}\boldsymbol{\mathcal{I}}_1(\mathbf{z})\mathbf{D}^T + \nu_2\mathbf{D}_c\boldsymbol{\mathcal{I}}_1(\mathbf{z})\mathbf{D}_c^T\big|\,U = u\right\},$$

$$\mathbf{G}_c = E\left\{\mathbf{D}\boldsymbol{\mathcal{I}}_1(\mathbf{z})\mathbf{D}^T + \mu_2\mathbf{D}_c\boldsymbol{\mathcal{I}}_1(\mathbf{z})\mathbf{D}_c^T\big|\,U = u\right\},$$

$$\boldsymbol{\mathcal{I}}_1(\mathbf{z}) = -E\left\{ \frac{\partial^2 \log f(Y;\ \mathbf{X},\ \boldsymbol{\theta},\ \mathbf{z})}{\partial \mathbf{z} \partial \mathbf{z}^T} \middle| \mathbf{X}, U \right\} \Bigg|_{U=u}.$$

Theorem 4 says that our estimator $\hat{\mathbf{a}}(\cdot)$ has the adaptivity property; it has the same asymptotic distribution as the estimator of $\mathbf{a}(\cdot)$ obtained by maximizing (4.10), with $\hat{\boldsymbol{\theta}}$ replaced by the true value of $\boldsymbol{\theta}$. In addition, the optimal bandwidth h_1 is of order $n^{-1/5}$, and the optimal convergence rate of $\hat{\mathbf{a}}(\cdot)$ is $n^{-2/5}$. We defer the proofs of these two theorems to the Appendix.

7 Simulation study and data analysis

7.1 Logistic Regression

Consider the following logistic regression model

$$\log\left(\frac{P\left(Y=1|X=x,U=u\right)}{1-P\left(Y=1|X=x,U=u\right)}\right)=a_1\left(u\right)x_1+a_2\left(u\right)x_2+a_3x_3+a_4x_4 \quad (7.1)$$

where $a_1(\cdot)$ and $a_2(\cdot)$ are unknown functional parameters, a_3, a_4 are unknown constant parameters, and X and U are independent. Further, assume $X \sim \mathrm{N}(0, I)$, $U \sim \mathrm{Uniform}(0,1)$, $a_1(u) = \sin(2\pi u)$, $a_2(u) = \cos(2\pi u)$, $a_3 = 2$ and $a_4 = 1$. The sample sizes were set to be 500 and 1000. For each sample size we repeated the experiment 300 times.

The kernel function K was set to be the Epanechnikov kernel. The bandwidths h and h_1 were respectively taken to be the data-driven AIC bandwidths $\hat{h}$ and $\tilde{h}_1$ given in Section 5.2, with model 5.2 specifying the conditional distribution

$$\log\left(\frac{P\left(Y=1|X=x,U=u\right)}{1-P\left(Y=1|X=x,U=u\right)}\right)=a_1\left(u\right)x_1+a_2\left(u\right)x_2+a_3\left(u\right)x_3+a_4\left(u\right)x_4 \quad (7.2)$$

We use the mean integrated absolute error (MIAE) to access the accuracy of an estimator. The MAIE of an estimator of an unknown constant is defined as its mean absolute error. The MIAE of a estimator $\hat{a}(\cdot)$ of an unknown function $a(\cdot)$ is defined as

$$\mathrm{MIAE} = E\left(\mathrm{IAE}\right), \text{ where IAE}= \int \left|\hat{a}\left(u\right)-a\left(u\right)\right| du.$$

The proposed two-step estimation method and the profile likelihood estimation using our suggested initial value were employed to estimate $a_1(\cdot)$, $a_2(\cdot)$, a_3, and a_4 for the 300 random samples. Table 1 compares the performances of our two-step estimation and the profile likelihood estimation under different sample sizes. The result suggests that both the two-step and the profile likelihood estimation methods do work well. The profile likelihood method is slightly better than the two-step method in estimating the constant parameters, while they preform equally well in estimating the functional parameters.

Table 1: MIAEs of different estimation methods for logistic regression.

	Two-step				Profile Likelihood			
Sample size	$a_1(\cdot)$	$a_2(\cdot)$	a_3	a_4	$a_1(\cdot)$	$a_2(\cdot)$	a_3	a_4
1000	0.1775	0.1753	0.1887	0.1174	0.1781	0.1748	0.1650	0.1121
500	0.2488	0.2465	0.2123	0.1368	0.2509	0.2514	0.2014	0.1406

To give a visible picture of how well the two-step estimators of the functional parameters work, the pointwise 10%, 50% and 90% quantiles of the 300 estimates of $a_1(\cdot)$ and $a_2(\cdot)$ are plotted in Fig. 1 when the sample size $n = 1000$ and 2 when $n = 500$. The solid lines are the true curves. Further, we single out the samples with median total IAE performance, i.e. the one that yields the median of the IAEs for

$a_1(\cdot)$ and $a_2(\cdot)$. In these samples, the estimates of a_3 and a_4 are respectively 2.280 and 1.094 when the sample size is 1000 and 1.9590 and 0.8112 when the sample size is 500. The dotted lines are the estimates, based on this sample, when a_3 and a_4 are treated unknown. The dashed lines are the estimates, based on the same sample, when a_3 and a_4 are treated known and replaced by their true values in the local likelihood function 4.10. From Fig. 1 and 2, we can see that the proposed method works quite well. Also, the estimators of the unknown functional parameters work as well as when the unknown constant parameters are replaced by their true values. This means the proposed estimators for the functional parameters do have the adaptivity property.

Suppose we do not know which of the four parameters are constant and which are functional. The BIC model selection procedure proposed in Section 3.3, with the start model M_0 specified by (7.2), was applied to the simulated samples. When the sample size is 1000, 276 of the 300 samples specify the true model, 7 samples pick a_1, a_2, and a_3 as constant parameters, 9 samples take a_1, a_2, and a_4 as constant parameters, and the remaining 8 samples determine all the parameters as constant parameters. When the sample size is 500, 251 of the 300 samples specify the true model, 19 samples select a_3 as constant parameter, 2 samples prefer a_4 as constant parameter, 8 samples pick a_1, a_3, and a_4 as constant parameters, 9 samples take a_2, a_3, and a_4 as constant parameters, 8 samples determine all the parameters as constant parameters, and the remaining 3 samples determine all the parameters as

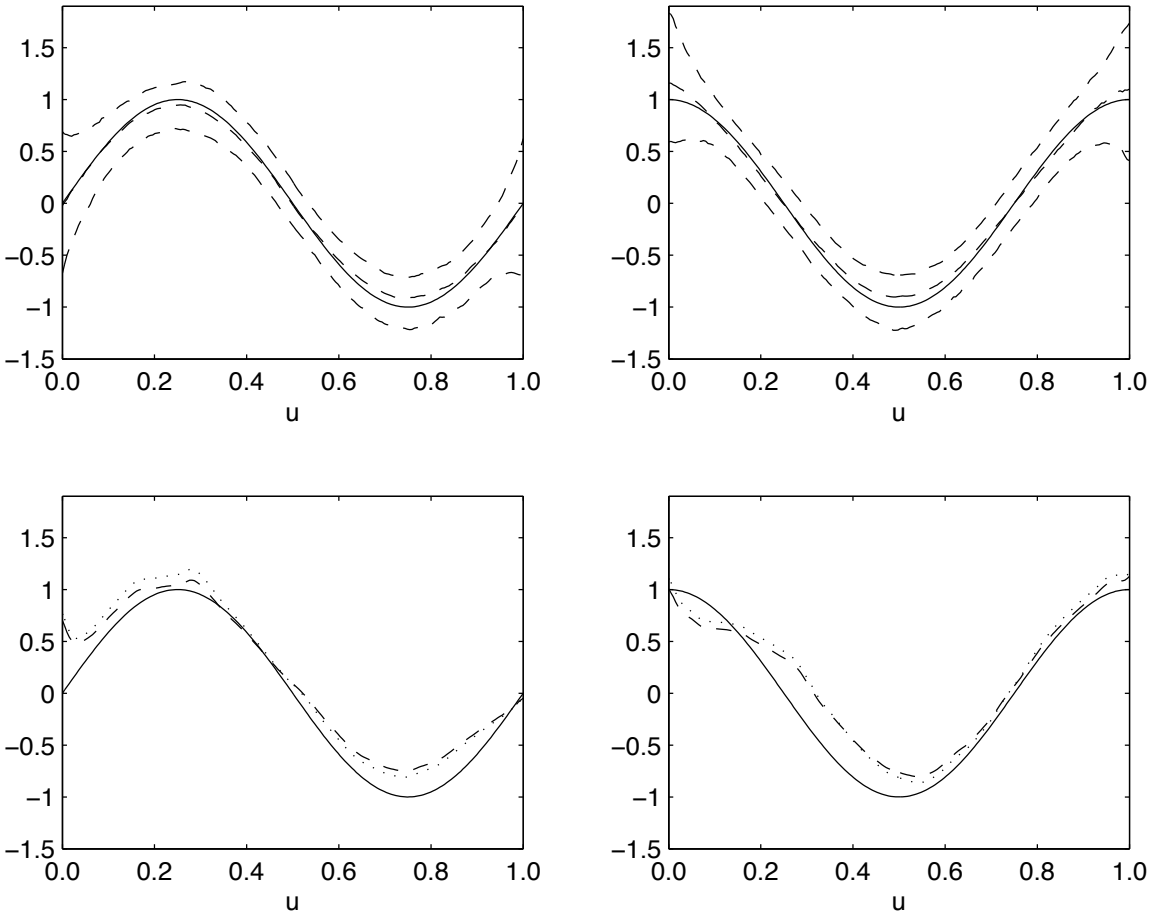

Figure 1: Functional parameters in logistic regression when the sample size is 1000. The left and right columns depict results for $a_1(\cdot)$ and $a_2(\cdot)$, respectively. In the upper row, the long-dash lines are the pointwise 10%, 50% and 90% quantiles of the 300 estimates. The bottom row plots the estimates based on the sample with median total ISE performance when the constant coefficients a_3 and a_4 are treated unknown (dotted) or known (dashed). The solid lines represent the true functions.

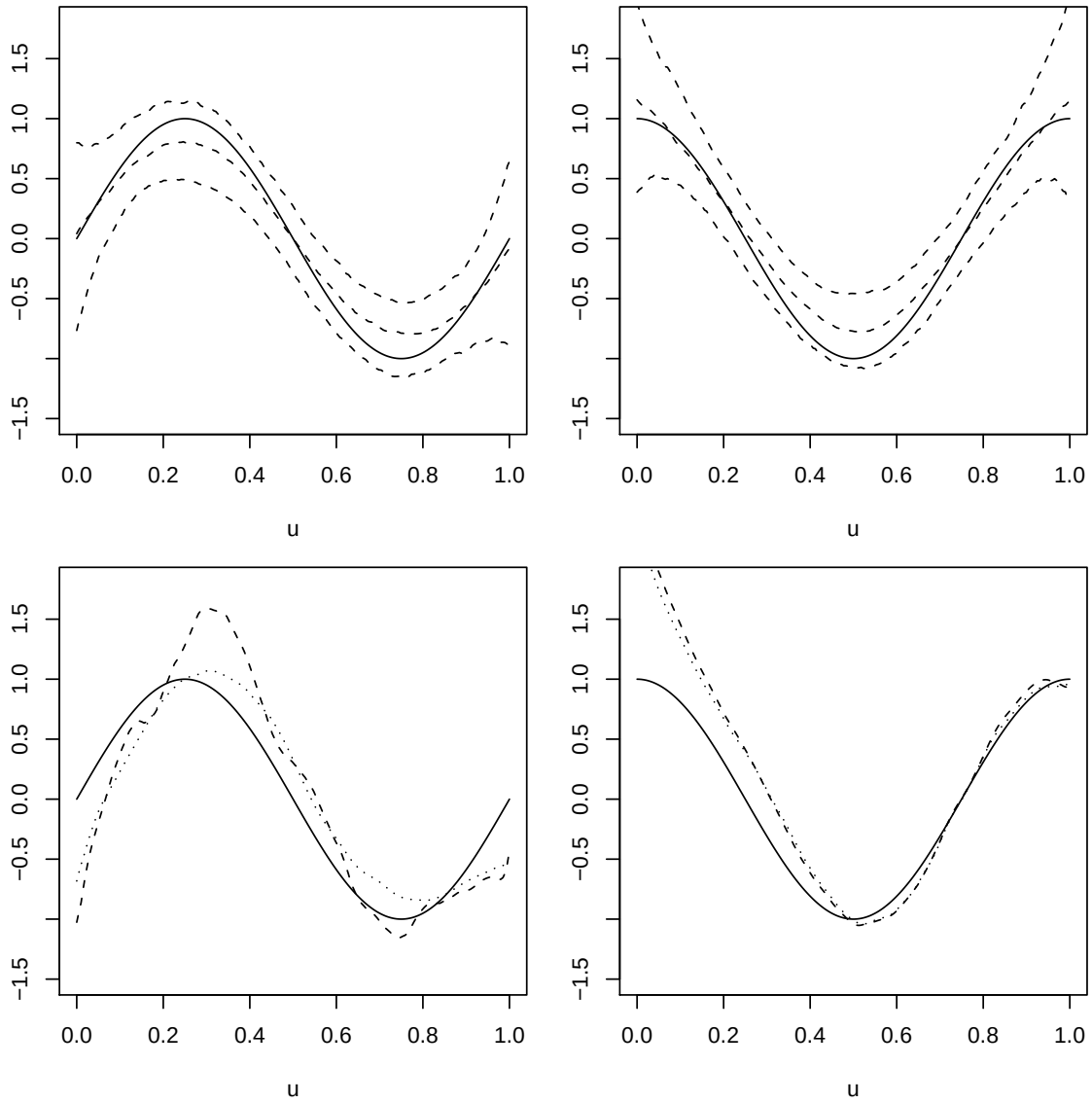

Figure 2: Functional parameters in logistic regression when the sample size is 500. The left and right columns depict results for $a_1(\cdot)$ and $a_2(\cdot)$, respectively. In the upper row, the long-dash lines are the pointwise 10%, 50% and 90% quantiles of the 300 estimates. The bottom row plots the estimates based on the sample with median total ISE performance when the constant coefficients a_3 and a_4 are treated unknown (dotted) or known (dashed). The solid lines represent the true functions.

functional parameters. The left panels of Fig. 3 and 4 show boxplots of the 300 predicted values of $a_3 \equiv 2$ and $a_4 \equiv 1$, and the right panels of Fig. 3 and 4 depict the point cloud of the predicted value against the true value of $a_1(U_0)$ and $a_2(U_0)$ for the 300 samples.

Other than the BIC criterion, we can also apply the AIC criterion to build the model selection procedure. In this case, when sample size is 1000, 242 of the 300 samples specify the true model, and the remaining 58 samples determine all the parameters as functional parameters. When the sample size is 500, 211 of the 300 samples specify the true model, and the remaining 89 samples select all the parameters as functional parameters. The left panels of Fig. 5 and 6 show boxplots of the 300 predicted values of $a_3 \equiv 2$ and $a_4 \equiv 1$, and the right panels of Fig. 5 and 6 depict the point cloud of the predicted value against the true value of $a_1(U_0)$ and $a_2(U_0)$ for the 300 samples. Note that although the AIC criterion does not select correct model as many times as the BIC criterion does, it generates less prediction error. This is because that the AIC criterion tends to select more complex models (for example, model with all a_j as functions for $j = 1, \cdots, 4$), while the BIC criterion tends to select simpler models (for example, model with all a_j as constants for $j = 1, \cdots, 4$) due to their penalties to model complexity. When the model is mis-specified, specifying the functional parameters $a_3(\cdot)$ and $a_3(\cdot)$ as constants would result in a large bias and inconsistency in post-model selection inference, while misspecifying the constant parameter a_1 and a_2 as functionals is only a minor

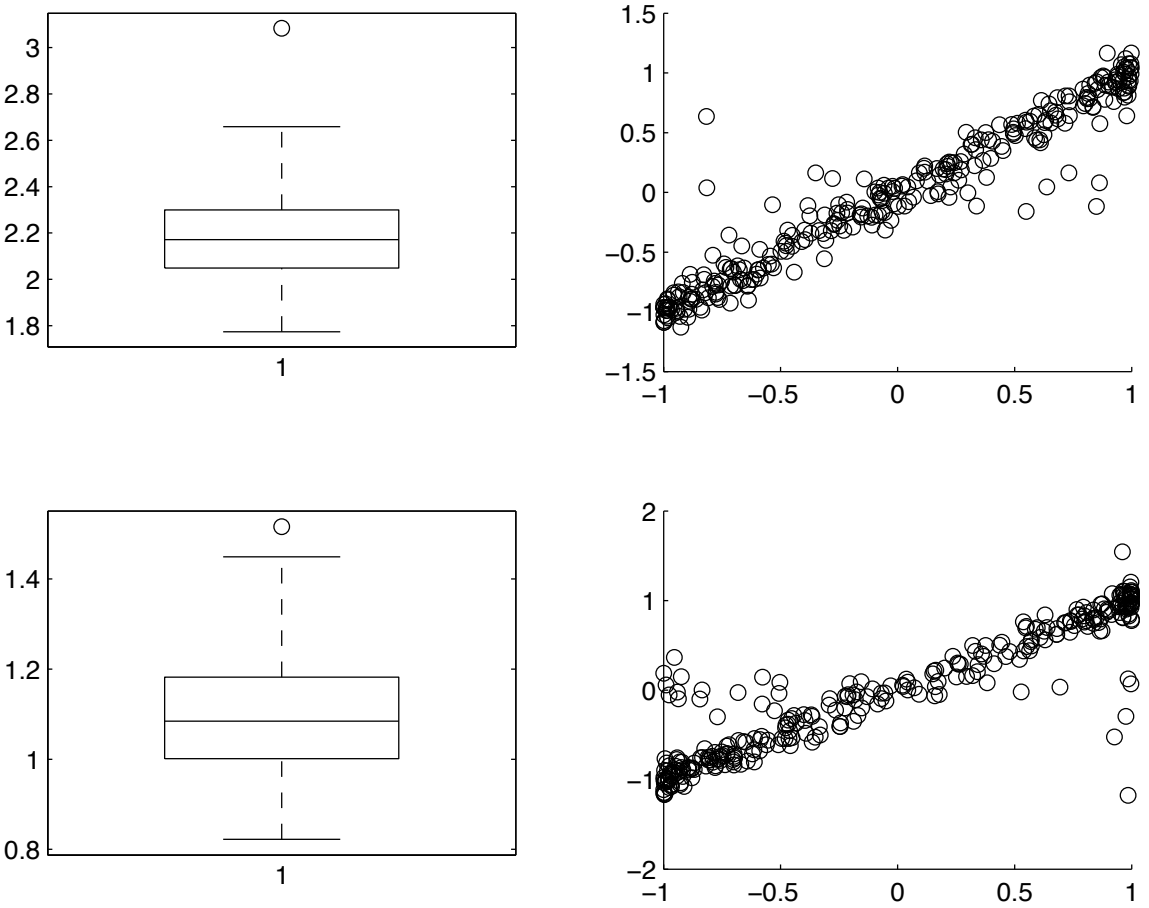

Figure 3: Parameter predictions in the logistic example when the sample size is 1000. Left: boxplots of the predicted values of $a_3 \equiv 2$ and $a_4 \equiv 1$ based on the selected model for the 300 samples. Right: scatterplots of the predicted value against the true value of $a_1(U_0)$ and $a_2(U_0)$ for the 300 samples.

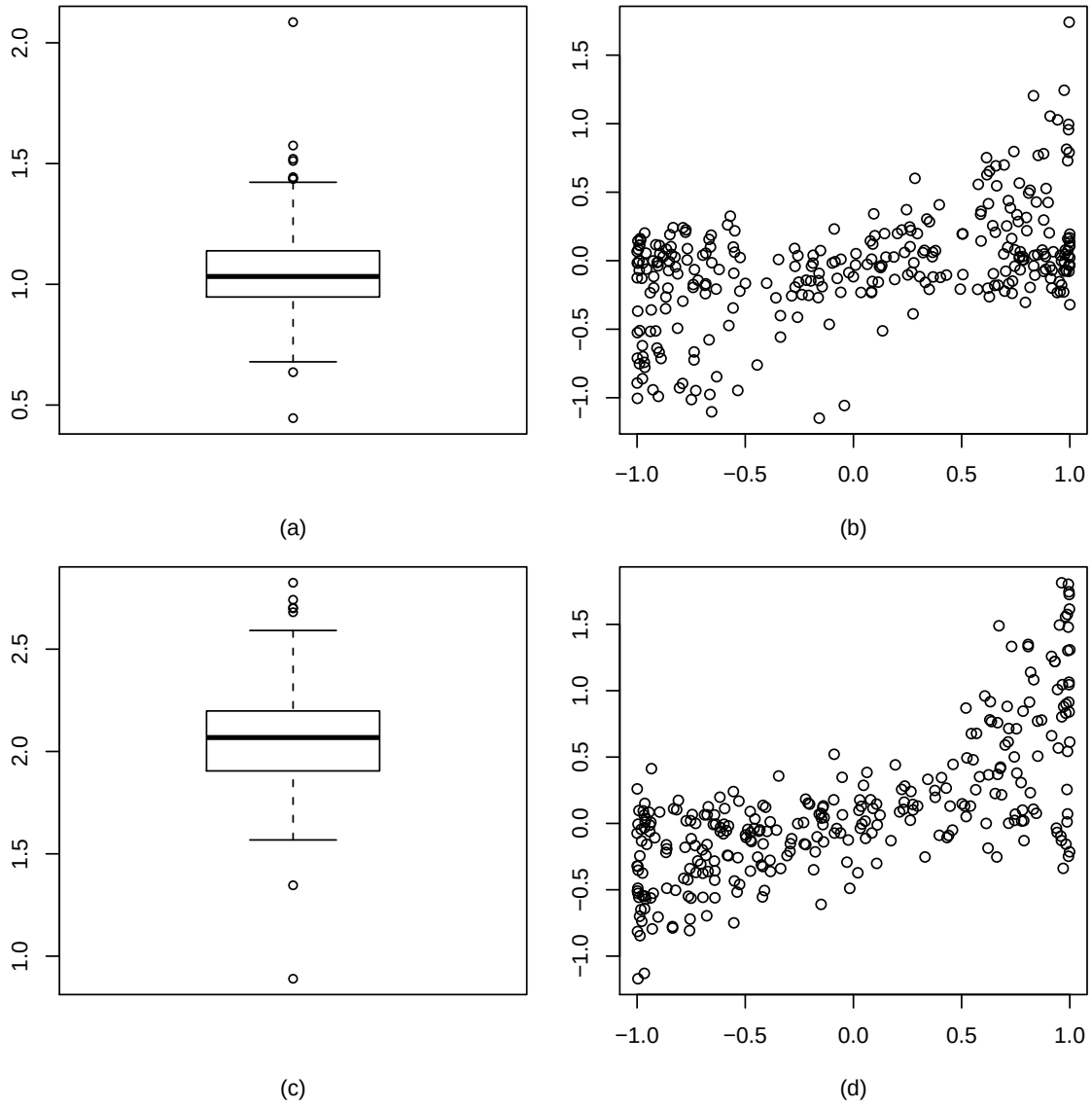

Figure 4: Parameter predictions in the logistic example when the sample size is 500. Left: boxplots of the predicted values of $a_3 \equiv 2$ and $a_4 \equiv 1$ based on the selected model for the 300 samples. Right: scatterplots of the predicted value against the true value of $a_1(U_0)$ and $a_2(U_0)$ for the 300 samples.

problem because the nonparametric estimates under the wrong model still look flat. The results of our simulation may support empirically that the consistency property of BIC and the efficiency property of AIC in parametric models also hold in our model.

7.2 Weibull model

Suppose that, conditional on $X = x$ and $U = u$, Y has a Weibull distribution with density function

$$f\big(y; x, \theta, a(u)x\big) = \frac{\theta}{\big\{a(u)x\big\}^{\theta}}\, y^{\theta-1} \exp\Big[- \big\{y/a(u)x\big\}^{\theta}\Big],\ y > 0, \tag{7.3}$$

where the constant $\theta > 0$ is the shape parameter and is taken to be 2, the function $a(\cdot)$ is the scale parameter and is set to be a quadratic function $a(u) = \beta_0+\beta_1 u+\beta_2 u^2$, $U \sim$ Uniform$(0, 1)$, $X \sim$ Uniform$(1, 2)$, and X and U are independent. This example is motivated by some real applications. For example, in reliability data analysis, Meeker and Escobar (1997), Nelson (1984) and Wang and Kececioglu (2000) studied the low-cycle fatigue life data for a strain-controlled test on 26 cylindrical specimens of a nickel-base superalloy to estimate the curve giving the number of cycles at which 0.1% of the population of such specimens would fail, as a function of the pseudostress U. They assumed that the logarithm of the number of cycles condition on the pseudostress follows a weibull distribution with a constant shape parameter (independent of the pseudostress) and a functional scale parameter. The scale pa-

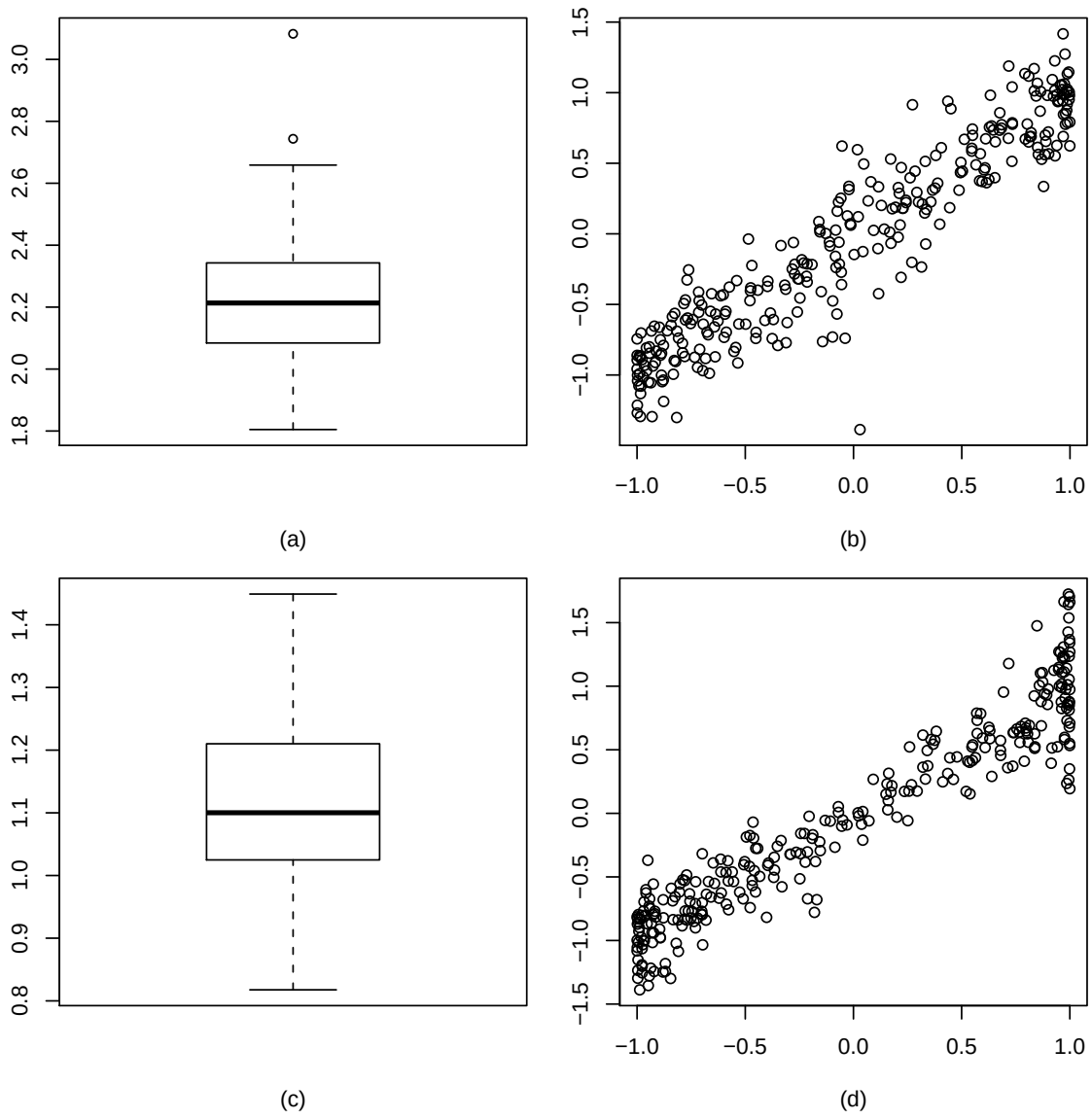

Figure 5: Parameter predictions in the logistic example by the AIC criterion when the sample size is 1000. Left: boxplots of the predicted values of $a_3 \equiv 2$ and $a_4 \equiv 1$ based on the selected model for the 300 samples. Right: scatterplots of the predicted value against the true value of $a_1(U_0)$ and $a_2(U_0)$ for the 300 samples.

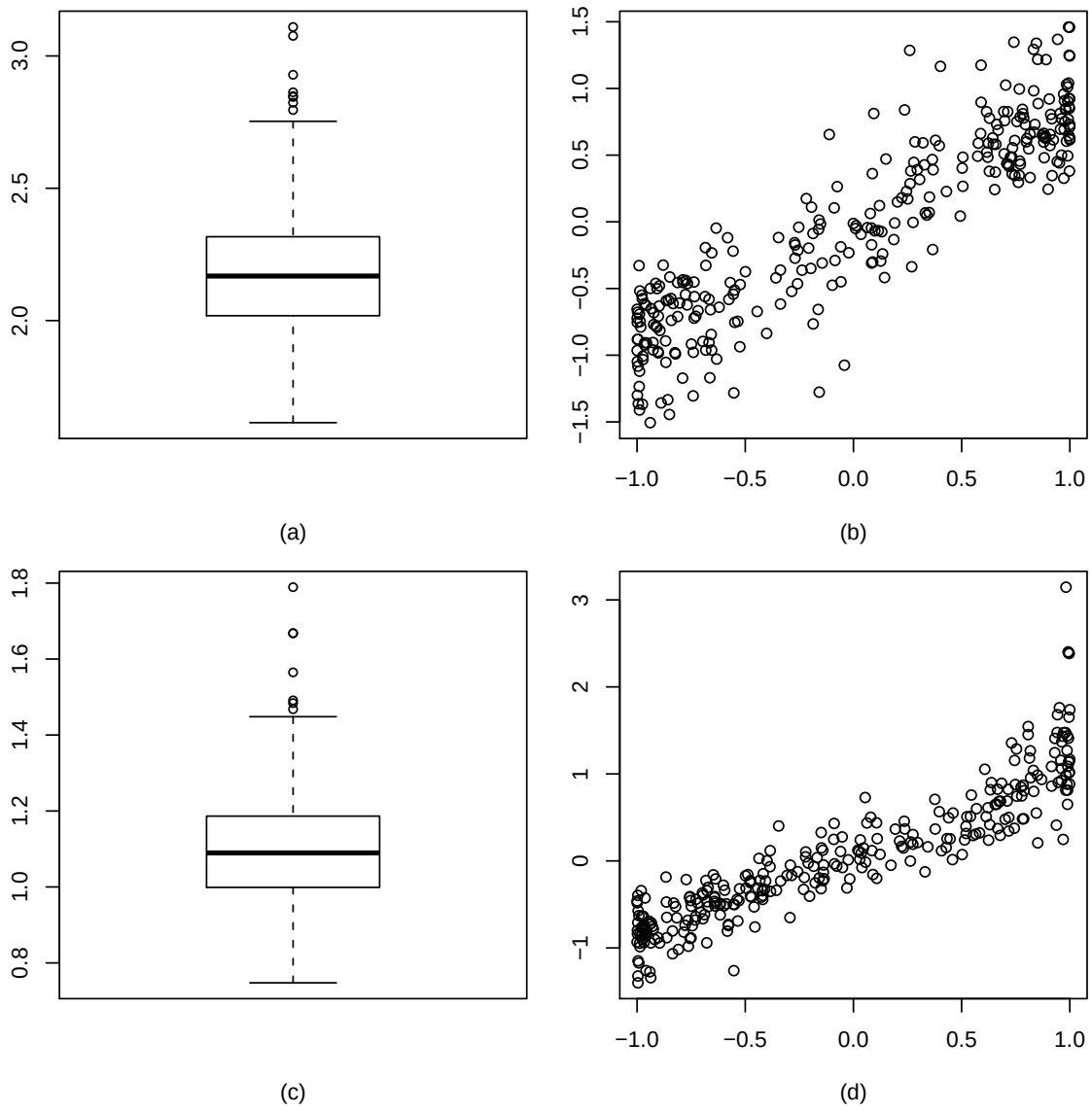

Figure 6: Parameter predictions in the logistic example by the AIC criterion when the sample size is 500. Left: boxplots of the predicted values of $a_3 \equiv 2$ and $a_4 \equiv 1$ based on the selected model for the 300 samples. Right: scatterplots of the predicted value against the true value of $a_1(U_0)$ and $a_2(U_0)$ for the 300 samples.

rameter was often assumed to be a linear, quadratic, or log-linear function of the pseudostress. In our implementation, we set $\beta_0 = 2$, $\beta_1 = -1.6$, and $\beta_2 = 3.6$. The sample sizes are taken to be 250, 500, and 1000; for each size we simulated 300 samples from this model and applied our estimation and model selection procedures to the samples.

In the two-step estimation procedure, we used the bandwidths $\hat{h}$ and $\tilde{h}_1$ in Section 5.2, with model (5.2) specifying the conditional density

$$f\big(y; x, a_0(u), a(u)x\big) = \frac{a_0(u)}{\big\{a(u)x\big\}^{a_0(u)}}\, y^{a_0(u)-1} \exp\Big[- \big\{y/a(u)x\big\}^{a_0(u)}\Big],\ y > 0. \quad (7.4)$$

The kernel function K was taken to be the Epanechnikov kernel. The MIAEs for θ and $a(\cdot)$ are 0.0175 and 0.1090 with sample size 1000, 0.0318 and 0.1369 with sample size 500, and 0.0623 and 0.1774 with sample size 250. The bias and standard deviation for θ are 0.00149 and 0.0496 with sample size 1000, in agreement with the theory that $\hat{\theta}$ is asymptotically unbiased (see Theorem 1). The bias and standard deviation for θ with sample size 500 and 250 are reported in Table 3 and 4. The left panel of Fig. 7 plots the pointwise 10%, 50%, and 90% quantiles of the 300 curve estimates of $a(\cdot)$ with sample sizes 1000, 500, and 250 from top to bottom. Both estimators of θ and $a(\cdot)$ are quite accurate. In addition, the constant parameter θ is estimated with a higher level of accuracy than the functional parameter $a(\cdot)$. This coincides with our theory that $\hat{\theta}$ has a faster rate of convergence than $\hat{a}(\cdot)$. The right panel of Fig. 7 plots the estimates of $a(\cdot)$ based on the sample with median

IAE performance when θ is treated as unknown (dotted line) and known (dashed line). The estimates are close to each other, indicating that our estimator of $a(\cdot)$ has the adaptivity property.

We also applied the profile likelihood method described in Section 4.2 to the same 300 samples. The MIAEs, biases and standard deviations for θ and $a(\cdot)$ with sample sizes $n = 1000$, 500, and 250 are summarized in Table 2, 3 and 4, respectively. Note that the profile likelihood estimators diverge in some samples with sample size 250 which may due to design sparsity. In this example, θ is the scale parameter and $a(\cdot)$ determines the shape parameter in the conditional Weibull distribution. The two-step method performs slightly better than the profile likelihood method in estimating both the constant (scale) parameter and the functional (shape) parameter.

We also fitted parametric models to the same 300 examples. Later, quadratic model denotes the case if $a(\cdot)$ is correctly specified as a quadratic function, cubic model denotes the case if $a(\cdot)$ is mis-specified as a cubic function, and linear model denotes the case if $a(\cdot)$ is assumed to be a linear function. Table 2 – 4 list the performances of different methods.

Suppose that it is not known which of the two parameters are constant and which are functional. For each of the 300 samples simulated from (7.3), we used the model selection procedure in Section 5.3, with the start model $\mathcal{M}_0$ given by model (7.4), to select the constant parameters. When the sample size is 1000, 295 samples specify the true model, and for all of the other 4 samples, the model with both θ and

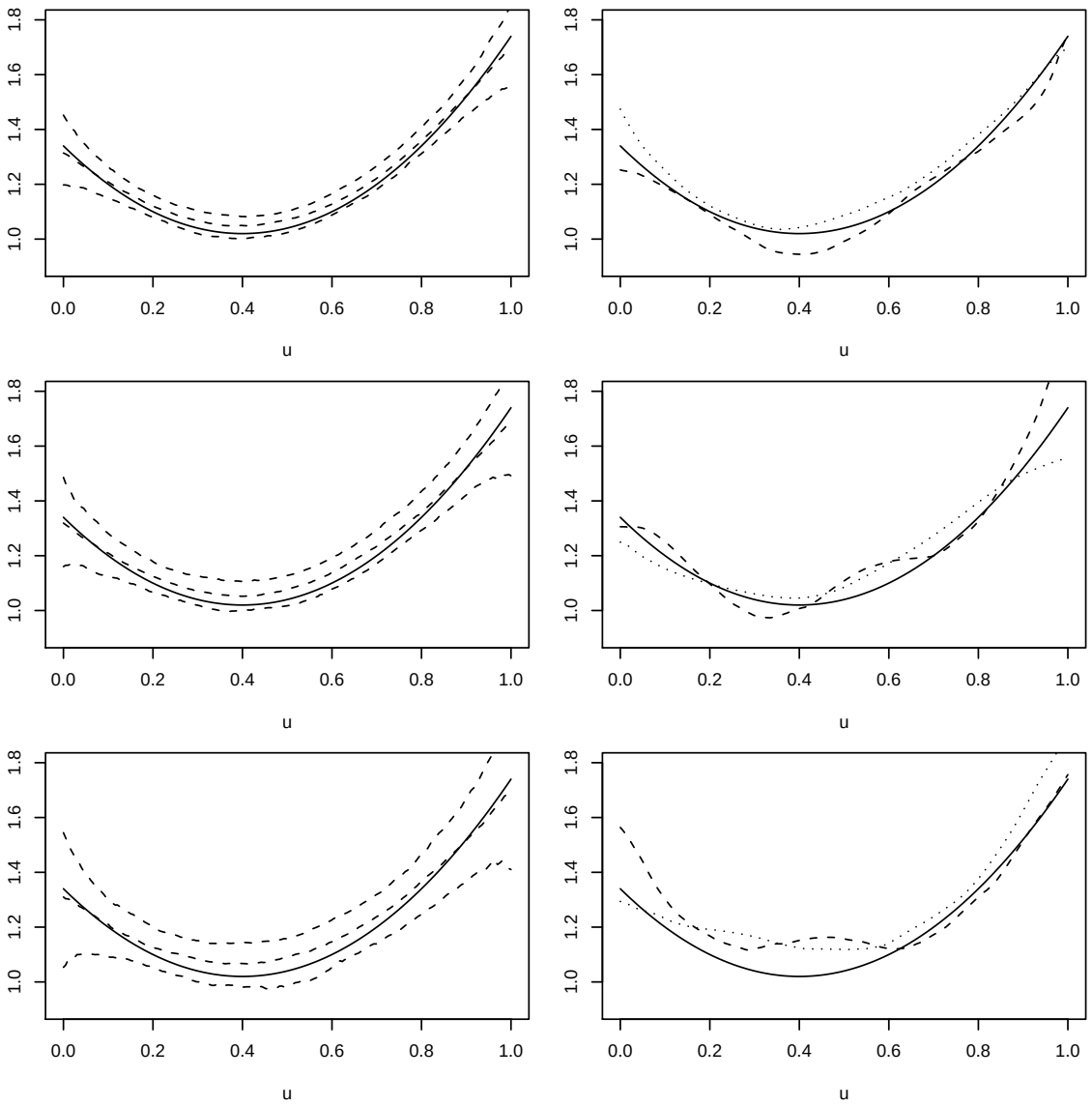

Figure 7: Estimates of the functional parameter in the Weibull example. Left panel: Pointwise 10%, 50%, and 90% quantiles (long-dashed lines) of the 300 estimates of $a(\cdot)$ (solid line) for sample size 1000, 500, and 250 from top to bottom. Right panel: Estimates of $a(\cdot)$ (solid line) based on the sample with median IAE performance with the constant parameter θ treated as unknown (dotted) or known (dashed line).

Table 2: Performances of different estimation methods on the Weibull example when the sample size is 1000.

	MAE of θ	bias of θ	std of θ	MIAE of $a(\cdot)$
Two-step	0.0175	0.00149	0.0496	0.1090
Profile likelihood	0.0181	0.00155	0.0504	0.1121
Linear model	0.0266	0.00248	0.0501	0.2001
Quadratic model	0.0168	0.00111	0.0488	0.0542
Cubic model	0.0180	-0.00156	0.0524	0.1080

Table 3: Performances of different estimation methods on the Weibull example when the sample size is 500.

	MAE of θ	bias of θ	std of θ	MIAE of $a(\cdot)$
Two-step	0.0318	0.00169	0.0711	0.1367
Profile likelihood	0.0344	0.00195	0.0721	0.1400
Linear model	0.0339	0.00221	0.0702	0.1989
Quadratic model	0.0299	0.00152	0.0698	0.0704
Cubic model	0.0320	0.00177	0.0780	0.1360

Table 4: Performances of different estimation methods on the Weibull example when the sample size is 250.

	MAE of θ	bias of θ	std of θ	MIAE of $a(\cdot)$
Two-step	0.0623	0.0149	0.1201	0.1774
Profile likelihood	1.2555	1.8400	21.8001	18.6441
Linear model	0.0666	0.0183	0.0911	0.1802
Quadratic model	0.0414	0.0100	0.1989	0.0780
Cubic model	0.0517	0.0126	0.1075	0.1511

a as functions of u was selected. This indicated that our model selection criterion has a high success rate of 98%. When the sample size is 500, 271 samples specify the true model, and the other 22 samples take the model with both θ and a as functions. When the sample size is 250, 160 samples specify the true model, 62 samples prefer the model with both θ and a as constants, and the other 78 samples opt the model with both θ and a as functions. This indicated that our model selection criterion has a high success rate when the sample size is moderately large (98% when the sample size is 1000, and 90% when the sample size is 500.)

To further examine the performance of the model selection procedure, for each sample sizes, we used the selected model and the corresponding parameter estimates to predict the true values of the parameters θ and $a(U_0)$ associated with a future

observation (Y_0, X_0, U_0) for each of the 300 samples. The mean absolute prediction errors (MAPE) for θ and $a(U_0)$ are reported in Table 5. The left panel of Fig. 8 shows boxplots of the 300 predicted values of $\theta \equiv 2$ with sample sizes 1000, 500, and 250 from top to bottom, and the right panel of Fig. 8 depicts the point clouds of the predicted value against the true value of $a(U_0)$ for the 300 samples. We can see that the predictions are both quite satisfactory even though the model selection procedure may misspecifies the model when the sample size is moderately large. Thus we can conclude from this example that our proposed estimation and model selection procedures work together to provide a powerful tool for multiparameter likelihood modeling even when there is little knowledge regarding whether or not some of the parameters are constant.

We also implement a different model selection procedure with the BIC criterion being substituted by the AIC criterion. When the sample size is 1000, 230 samples specify the true model, and the other 70 samples pick the model with both θ and a as functions of u. When the sample size is 500, 212 samples specify the true model, and the other 88 samples choose the model with both θ and a as functions. When the sample size is 250, 166 samples specify the true model, 143 samples prefer the model with both θ and a as functions, and the other 4 samples opt the model with both θ and a as constants. Table 5 summarizes the results of the model selection procedures based on BIC and AIC criterion. The left panel of Fig. 9 shows boxplots of the 300 predicted values of $\theta \equiv 2$ with sample sizes 1000, 500, and 250 from top to bottom,

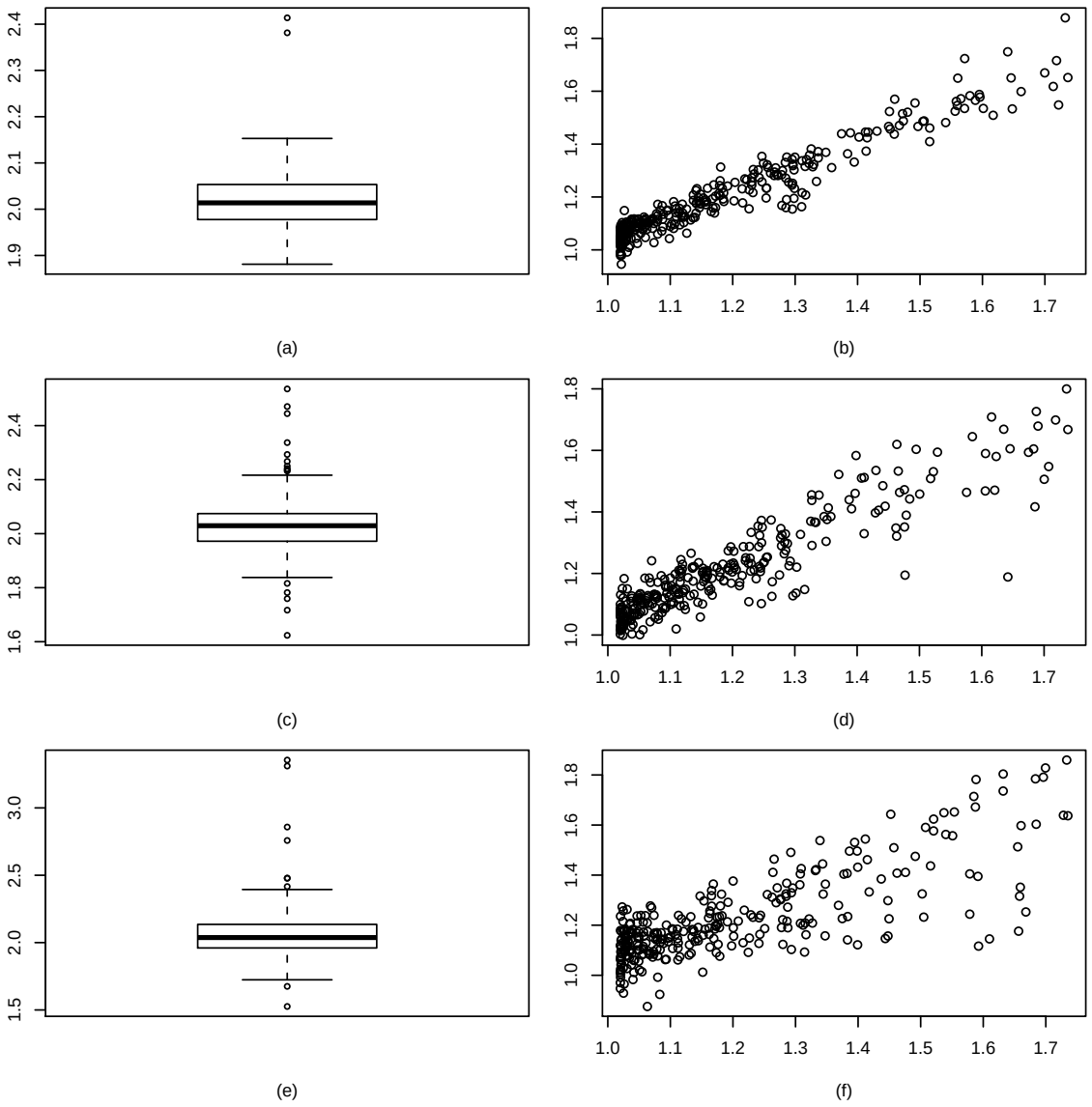

Figure 8: Parameter predictions in the Weibull example with the BIC criterion. Left panel: Boxplots of the predicted values of $\theta \equiv 2$ based on the selected model for the 300 samples of sample sizes 1000, 500, and 250, from top to bottom. Right panel: Scatterplots of the predicted value against the true value of $a(U_0)$ for the 300 samples.

and the right panel of Fig. 9 depicts the point clouds of the predicted value against the true value of $a(U_0)$ for the 300 samples. As in the case of logistic regression, we can see that the AIC criterion tends to select more complex models (for example, model with both θ and a as functions) and thus generates less prediction error, while the BIC criterion tends to select simpler models (for example, model with both θ and a as constants) and thus generate larger prediction error, although the BIC criterion selects the correct model more times than the AIC criterion does,

Table 5: Performances of model selection procedures of the Weibull example.

		1000	500	250
BIC	True model selected	295	271	160
	MAPE of θ	0.0416	0.0671	0.1122
	MAPE of $a(\cdot)$	0.1453	0.1610	0.1788
AIC	True model selected	230	212	166
	MAPE of θ	0.0507	0.0704	0.1065
	MAPE of $a(\cdot)$	0.1489	0.1609	0.1692

In model (7.3), the Weibull conditional distribution has $d = 2$ parameters, of which $\ell = 1$ follow a nonparametric form and the other $d-\ell = 1$ follow a parametric form. A more complex model involves changing the conditional distribution to $\text{Weibull}\big(\theta + a_1(u)x, a(u)x\big)$, where x, u, θ and $a(\cdot)$ are the same as in (7.3), and

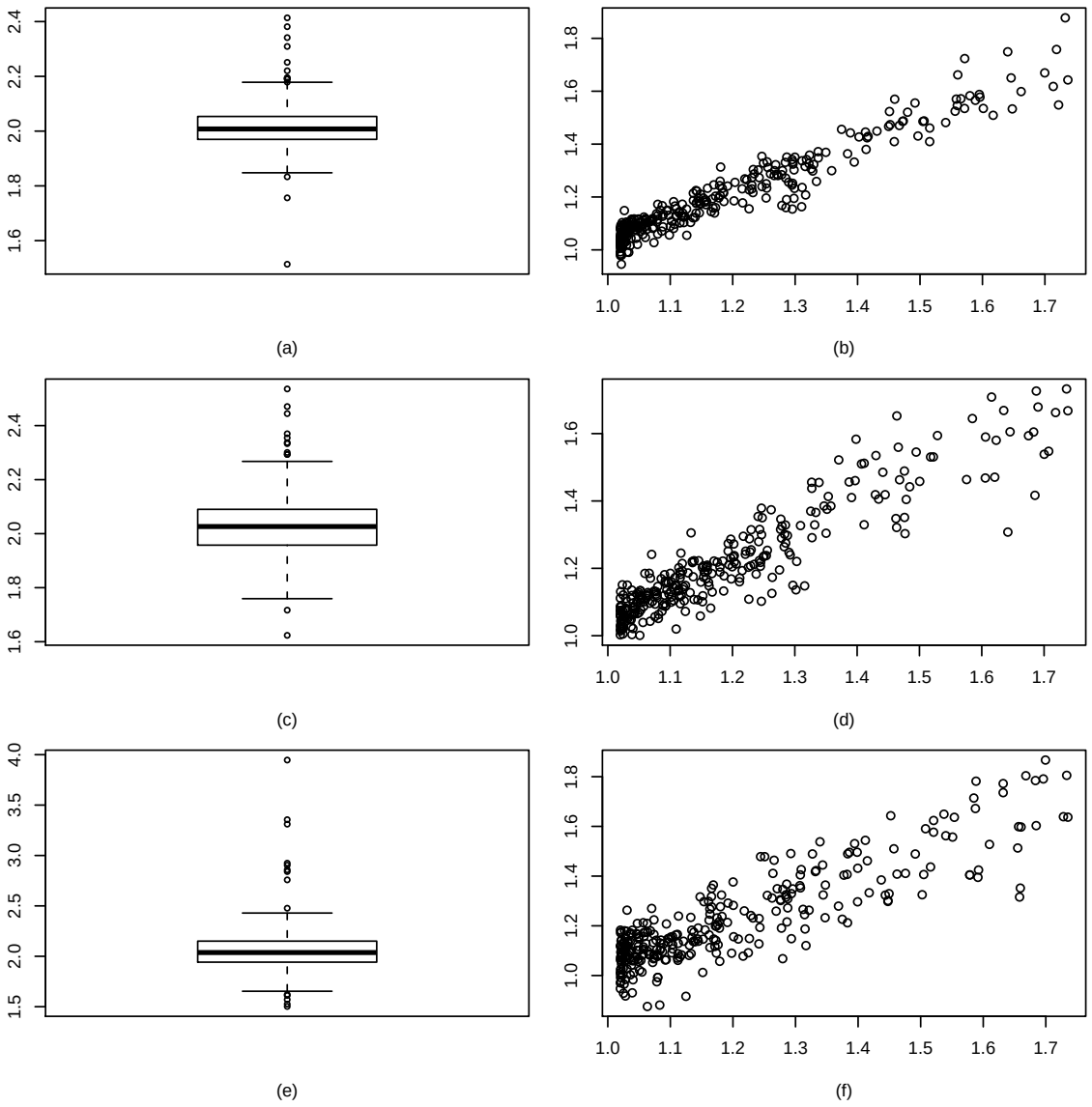

Figure 9: Parameter predictions in the Weibull example with the AIC criterion. Left panel: Boxplots of the predicted values of $\theta \equiv 2$ based on the selected model for the 300 samples of sample sizes 1000, 500, and 250, from top to bottom. Right panel: Scatterplots of the predicted value against the true value of $a(U_0)$ for the 300 samples.

$a_1(u) = 0.1 + 0.1\cos(2\pi u)$. In this case, $\ell = d = 2$, with the shape parameter modeled semiparametrically and the scale parameter modeled nonparametrically.

7.3 Hazard Regression

Suppose that we have a random sample of n subjects with J failure types in each subject. Consider the following semiparametric varying-coefficients hazard regression model

$$\lambda_{ij}(t) = Y_{ij}(t)\lambda_{0j}(t)\exp\{\boldsymbol{\theta}^T W_{ij}(t) + \mathbf{a}(U_{ij}(t))^T Z_{ij}(t)\}, \tag{7.5}$$

where $i = 1, \cdots, n$ indicates subject and $j = 1, \cdots, J$ denotes the jth failure type in the ith subject, $W_{ij}(\cdot) = (W_{ij1}(\cdot), \cdots, W_{ijq}(\cdot))^T$ is a vector of covariates that has parametric effect on the logarithm of the hazard, $Z_{ij}(\cdot) = (Z_{ij1}(\cdot), \cdots, Z_{ijp}(\cdot))^T$ is a vector of covariates that may interact with $U_{ij}(\cdot)$, $Y_{ij}(t) = I(X_{ij} \geq t)$ is an indicator, $X_{ij} = \min(T_{ij}, C_{ij})$ is the observed time with the failure time T_{ij} and the censoring time C_{ij}, $\lambda_{0j}(\cdot)$ is an unspecified baseline hazard function, and $\mathbf{a}(\cdot) = (a_1(\cdot), \cdots, a_p(\cdot))^T$ is a vector of unspecified smooth coefficient function. The marginal hazard function $\lambda_{ij}(t)$ is defined as

$$\lambda_{ij}(t) = \lim_{h \downarrow 0} \frac{1}{h} P\left(T_{ij} \leq t + h | T_{ij} \geq t, \mathcal{F}_{t,ij}\right),$$

where $\mathcal{F}_{t,ij}$ represents the failure, censoring and covariate information up to time t for the (i, j) failure type as well as the covariate information of the other failure types

in the ith subject up to time t. The censoring time is assumed to be independent of the failure time conditional on the covariates (i.e. independent censoring).

The estimation is usually carried out by maximizing the partial likelihood of model (7.5):

$$L(\boldsymbol{\theta}, \mathbf{a}) = \prod_{j=1}^{J} \prod_{i=1}^{n} \left\{ \frac{\exp\{\boldsymbol{\theta}^T W_{ij}(X_{ij}) + \mathbf{a}(U_{ij}(X_{ij}))^T Z_{ij}(X_{ij})\}}{\sum_{l \in \mathcal{R}_j(X_{ij})} \exp\{\boldsymbol{\theta}^T W_{lj}(X_{ij}) + \mathbf{a}(U_{lj}(X_{ij}))^T Z_{lj}(X_{ij})\}} \right\}^{\Delta_{ij}}, \quad (7.6)$$

where $\mathcal{R}_j(t) = \{i : X_{ij} \geq t\}$ denotes the set of the individuals at risk just prior to time t for failure type j, and Δ_{ij} is an indicator which equals 1 if X_{ij} is a failure time and 0 otherwise. By applying Taylor expansion on $\mathbf{a}$ we can obtain the local log-partial likelihood as we do in (4.8):

$$\sum_{j=1}^{J} \sum_{i=1}^{n} K_h(U_{ij}(X_{ij}) - u_0) \Delta_{ij} \left\{ \boldsymbol{\theta}^T W_{ij}(X_{ij}) + \boldsymbol{\gamma}^T V_{ij}(X_{ij}) - R^*_{ij}(\boldsymbol{\theta}, \boldsymbol{\gamma}) \right\} \quad (7.7)$$

where $V_{ij}(v) = \left\{ Z_{ij}(v)^T, Z_{ij}(v)^T (U_{ij}(v) - u_0) \right\}^T$, $\boldsymbol{\gamma} = (\mathbf{a}^T, \mathbf{b}^T)^T$, and

$$R^*_{ij}(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \log \left(\sum_{l \in \mathcal{R}_j(X_{ij})} \exp\{\boldsymbol{\theta}^T W_{lj}(X_{ij}) + \boldsymbol{\gamma}^T V_{lj}(X_{ij})\} K_h(U_{ij}(X_{ij}) - u_0) \right)$$

With the local log-partial likelihood (7.7) we can construct the two-step estimation proposed in Section 4.3; together with the partial likelihood (7.6) we can build the profile likelihood estimation introduced in Section 4.2.

In this example we set $n = 250$, $J = 2$, the baselines $\lambda_{01} = 1$ and $\lambda_{02} = 2$. The true parameter was set as $\boldsymbol{\theta}_0 = (0.8, 0.6, 1)^T$, and the coefficient function was set as $a(u) = 2 - 3\cos((u - 0.5)\pi/2)$. We assumed that $Z_{ij} \sim N(0, 1)$, $W_{ij1} \sim$ Bernoulli(0.5), $W_{ij2} \sim N(0, 1)$, $W_{ij3} \sim U(0, 1)$, and the covariate U_{ij} was set as W_{ij3}.

The censoring time distribution was generated from an exponential distribution with mean chosen to produce a 41% of censoring. The failure times were generated by the extension of the model of Clayton and Cuzick (1985). The number of replication is 300. The proposed two-step estimation method was employed to estimate θ_1, θ_2, θ_3, and $a(\cdot)$. The MIAEs are 0.09001 for θ_1, 0.0508 for θ_2, 0.1555 for θ_3, and 0.1362 for $a(\cdot)$. We compared our results with the profile likelihood estimation proposed by Cai et. al. (2008). The MIAEs are 0.0897, 0.0504, 0.1559 and 0.1367. This suggests that profile likelihood is slightly better than the two-step method in estimating the linear part, while they preform equally well in estimating the nuisance parameters.

8 Analysis for infant mortality in China

The data for this analysis come from a National Survey of Fertility and Contraceptive Prevalence, often referred to as the "Two per Thousand Fertility Survey," conducted by China's State Family Planning Commission between July 1, 1988 and July 15, 1988. The survey, representing a sample of 2 per 1,000 persons in mainland China, targeted ever-married resident women age 15-57 years. All provinces in the Chinese mainland took part in the survey. The sample for this study is restricted to births occurring after 1949, that is, after the founding of the People's Republic of China. Thus we have a total of 118,346 births (61,286 boys and 57,060 girls), contributed by 35,652 women. Of these births, 6,909 infants died before their first birthday, yielding an infant mortality rate of 58.4 per thousand.

The response variable Y was taken to be the binary variable, death or survival within the first year. Thus births occurring within 12 months before the survey were excluded, and the remaining 114,337 births were used for the logistic regression analysis. Selection of relevant independent variables was guided by previous studies on the determinants of infant mortality and constrained by those that were included in the survey. Thus we used the following variables: year of birth (U); reproductive patterns [mother's age at the birth of the child (X_2), first child (X_3) and previous birth interval (X_4)]; and socioeconomic variables [urban-rural residence (X_5), mother's education (X_6), geographic region of residence (X_7), and ethnicity (X_9)].

We also included other control variables, such as sex of the child (X_8) and breast-feeding (X_{10}) during the first year of life. Table 7 lists all the variables included in the study together with their descriptive statistics.

We categorized the mother's age into two categories: between 15 to 35 (appropriate age) and otherwise (inappropriate age). To see if the impacts of the independent variables vary of time, we can compare the parametric model without involving with U:

$$\log\left(\frac{P(Y=1|\mathbf{X}=x)}{1-P(Y=1|\mathbf{X}=x)}\right) \;=\; a_1+\sum_{i=2}^{10} a_i\, x_i\,,$$

with the nonparametric model $\mathcal{M}_0$ specifying

$$\log\left(\frac{P(Y=1|\mathbf{X}=x,\ U=u)}{1-P(Y=1|\mathbf{X}=x,\ U=u)}\right) \;=\; a_1(u)+\sum_{i=2}^{10} a_i(u)\, x_i\,.$$

However, when constructing a test that involves the nonparametric forms, the standard chi-squared approximation (e.g. Pearson's chi-square test or tests based on deviance) fails because the effective number of parameters tends to infinity. Cai et al. (2000) suggested a bootstrap approach to facilitate model testing. In our case, bootstrap is computationally impractical due to our large sample size. Here we considered an alternative parametric model $\mathcal{M}_0'$

$$\log\left(\frac{P(Y=1|\mathbf{X}=x,U=u)}{1-P(Y=1|\mathbf{X}=x,U=u)}\right) \;=\; \left(a_{1,u}+\sum_{i=2}^{10} a_{i,u}\, x_i\right) I(U=u)\,.$$

That is, we treated data in different years as independent and fitted a separate model for each year. The difference between $\mathcal{M}_0$ and $\mathcal{M}_0'$ is that the impacts of the

dependent variables are assumed to be smooth in $\mathcal{M}_0$, while to be independent in $\mathcal{M}_0'$. When comparing the parametric model with $\mathcal{M}_0'$, the p-value of the Pearson's chi-square test is 4×10^{-24}, which is definitely significant in all significant levels. This demonstrate that the impacts of the independent variables do vary of time. Although we construct a test to test $\mathcal{M}_0$ against $\mathcal{M}_0'$, the deviance of $\mathcal{M}_0$ is smaller than that of $\mathcal{M}_0'$, suggesting that $\mathcal{M}_0$ may be a better fit. The estimated impacts of the covariates with model $\mathcal{M}_0$ and $\mathcal{M}_0'$ are presented in Fig. 10 and 11, respectively.

Then we used the model selection procedure outlined in Section 5.3 to determine which effect of the covariates are constant (invariant to time U) and which are functional in the logistic regression. We started with the nonparametric model $\mathcal{M}_0$ and used the bandwidth selection procedure described in Section 5.2 to select the bandwidths h and h_1. The selected bandwidth are $\hat{h} = 10.80\%$ and $\hat{h}_1 = 19.55\%$ of the time range. After that we start with model $\mathcal{M}_0$ as the candidate model and iteratively examine whether one of the functional parameters in the candidate model can be further reduced to a constant. Our model selection procedure suggests that the impacts of the mother's age (X_2), mother's education (X_6), ethnicity (X_9), child's sex (X_8), and type of feeding (X_{10}) are constant; thus we used model (1.1) with the assumption that

$$\begin{aligned} \log\left(\frac{P(Y=1|\mathbf{X}=x,\ U=u)}{1-P(Y=1|\mathbf{X}=x,\ U=u)}\right) &= a_1(u) + a_2x_2 + a_3(u)x_3 + a_4(u)x_4 + a_5(u)x_5 \\ &\quad + a_6x_6 + a_7(u)x_7 + a_8x_8 + a_9x_9 + a_{10}x_{10} \end{aligned}$$

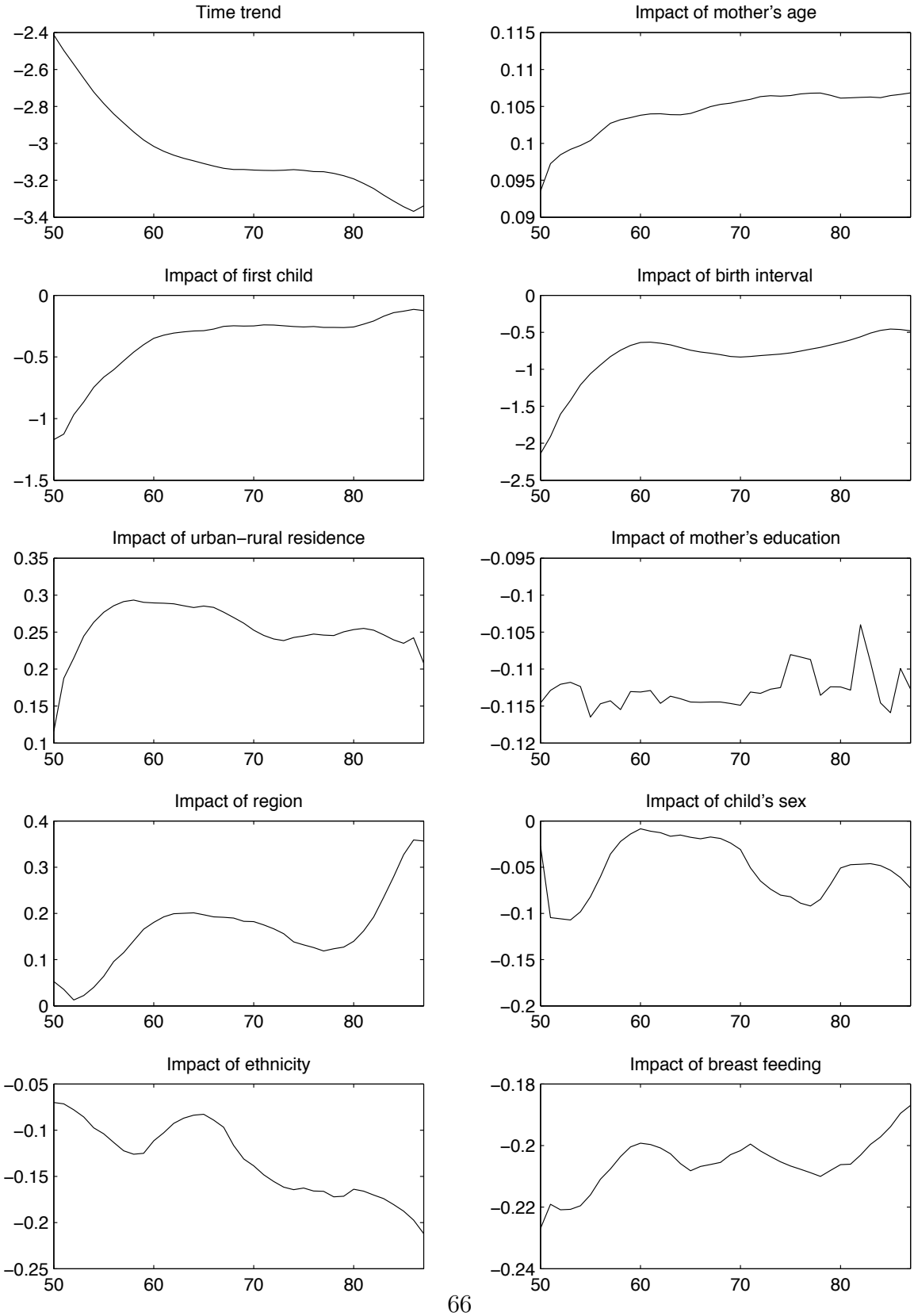

Figure 10: *Impacts of covariates on infant mortality with model* $\mathcal{M}_0$.

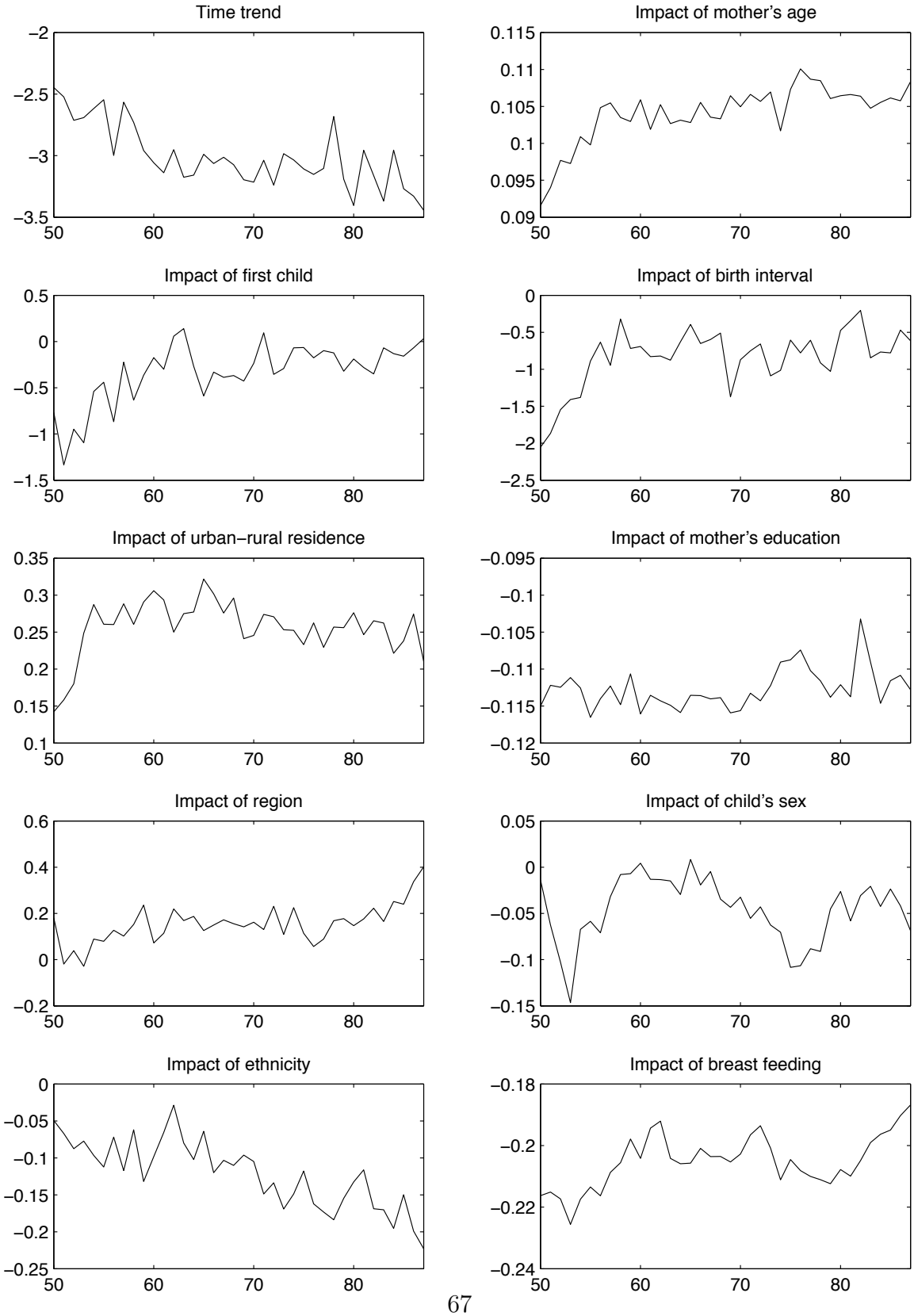

Figure 11: *Impacts of covariates on infant mortality with model* $\mathcal{M}_0'$.

fits the data. We denoted this model as $\mathcal{M}$ and used the proposed two-step estimation method to estimate the effects of the sociodemographic variables on infant death with this model.

The estimate of the impact of the mother's age (X_2) is 0.1041, indicating that women of inappropriate age are at high risk for infant mortality. The estimate of the effect of the mother's education (X_6) is -0.1088, which means that educated women are at lower risk for infant mortality. This difference between educated women and noneducated women is understandable, because well-educated women generally have readier access to information on nutrition and health care and are better at implementing medical advice. The estimated effect of the child's sex (X_8) is -0.0492, indicating a lower risk of mortality in male infants. Traditional Chinese culture always favors boys. As in much of the developing world, Chinese girls receive far less attention and resources than boys. The estimate of impact of ethnicity (X_9) is -0.1371, which tells us that the Han have a lower risk of infant mortality compared with people from Chinese minority groups. The estimate of the impact of breast-feeding (X_{10}) is -0.2058, suggesting the superiority of breast-feeding over other kinds of feeding.

The estimated impacts of the other functional factors (U, X_3, X_4, X_5, X_7) are presented in Fig. 12. The confidence bands were constructed using $\hat{a}_j(\cdot) \pm 1.96SE$, where SE is the standard error computed by a sandwich method (see Cai, Fan, and Li 2000). Fig. 12(a) shows that infant mortality started high in 1950, increasing

from 1950, and reached its highest level in 1959. After 1959, the mortality dropped steadily; however, the pace of the decline slowed down after 1970.

Fig. 12(b) clearly shows that mortality of the first birth is lower than that of the others; the impact of first child on mortality is a negative curve. The interpretation of this finding is that the first child has the advantage of having no previous sibling to compete with for the parents' attention and resources. Cultural factors also may contribute to the lower mortality of first births in China. In China, the birth of the first child is a very important event for a family, and the first child usually receives much more attention and care than others. Moreover, Chinese grandmothers generally play a very important role in taking care of their grandchildren, especially in rural areas. Their involvement, advice, and supervision can overcome some of the disadvantages that first births encounter as a result of physiological difficulties in delivery and the mothers' lack of previous childbearing and child care experience. It also appears that although the impact of first child on mortality is always negative, its absolute value decreased sharply from 1950 to 1960 sharply, then only slightly thereafter.

Fig. 12(c) shows that the impact of birth interval on infant mortality also is a negative curve, meaning that a longer birth interval would enhance an infant's chance of survival. This finding is in accordance with those from previous studies. Like the impact of the first child, the absolute value of the impact of birth interval dropped sharply between 1950 and 1960, remained unchanged until 1967, and then

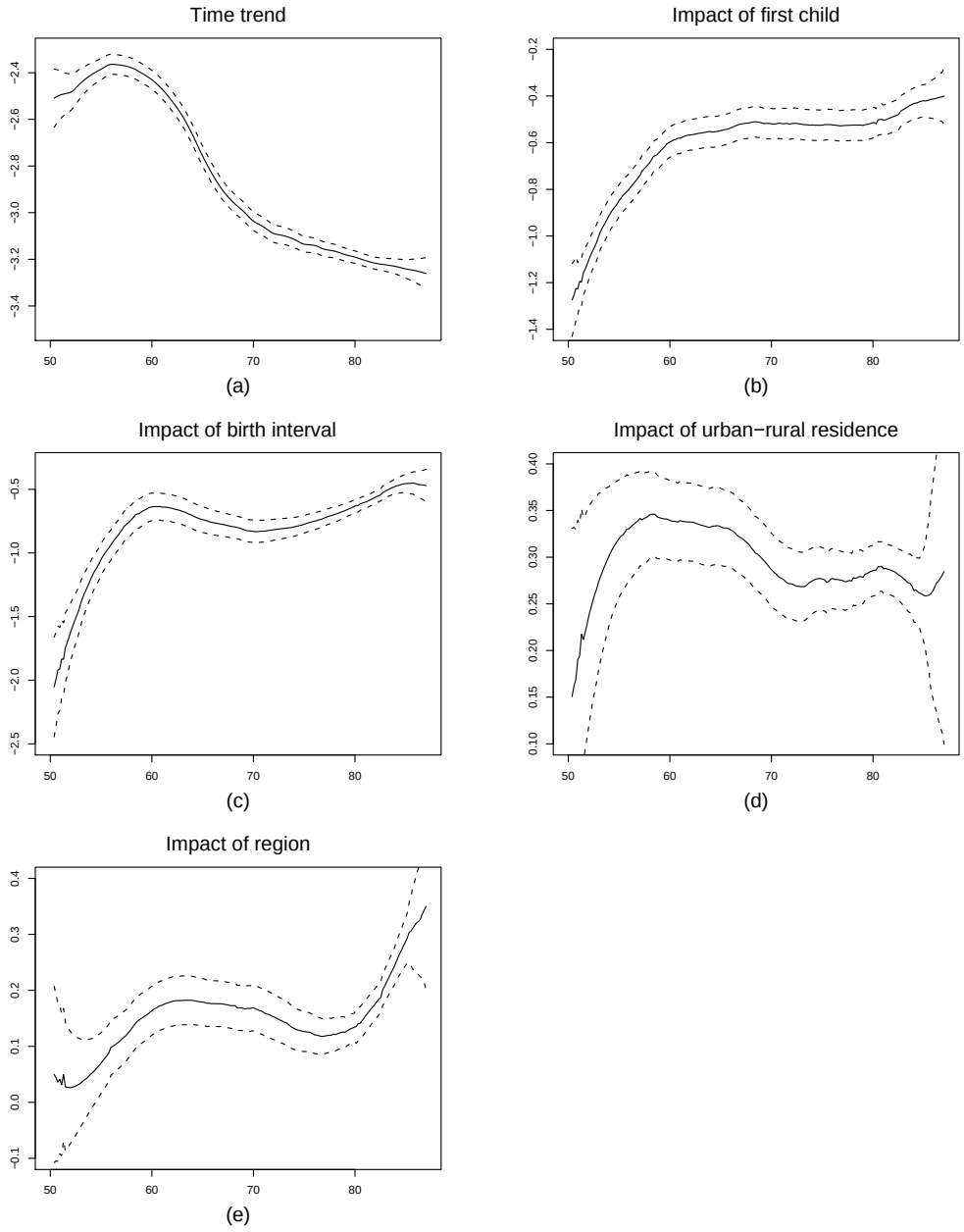

Figure 12: *Impacts of covariates on infant mortality.* Effects of year of birth (U), first child (X_3), previous birth interval (X_4), urban-rural residence (X_5), and geographic region of residence (X_7) against time. The solid curves represents estimates of impacts of the covariates; the dashed curves, the 95% bands of the estimates.

dropped again thereafter.

Fig. 12(d) shows that infant mortality was always higher in rural areas than in urban areas. The impact of rural residence on infant mortality rose sharply from 1950 to 1958, then dropped sharply until 1972, remained steady until 1981, and then again dropped sharply. This suggests a decreasing difference between rural and urban residence from 1958 to 1987.

We took the three cities of Beijing, Shanghai and Tianjin, as a reference. Fig. 12(e) suggests lower infant mortality in these three cities than in other places, with the difference increasing between 1952 to 1961, then decreasing until 1978, and then increasing again thereafter. The interpretation of this finding is the Chinese government invested in these three cities much more heavily than in other places. Indeed, the three cities received priority on almost everything for quite a long time. Before 1980, many goods (including some important medicines and nutritional foods) could be bought only in these three cities, and the three cities had the best hospitals, health care, and environmental sanitation.

Again, instead of testing $\mathcal{M}$ against $\mathcal{M}_0$, we tested their parametric alterna-

tives, $H_0 : \mathcal{M}'$ against $H_1 : \mathcal{M}'_0$, where $\mathcal{M}'$ is specified by

$$\begin{aligned}\log\left(\frac{P(Y=1|\mathbf{X}=x,U=u)}{1-P(Y=1|\mathbf{X}=x,U=u)}\right) &= a_{1,u}I(U=u)+a_2x_2+a_{3,u}x_3I(U=u)\\ &\quad +a_{4,u}x_4I(U=u)+a_{5,u}x_5I(U=u)+a_6x_6\\ &\quad +a_{7,u}x_7I(U=u)+a_8x_8+a_9x_9+a_{10}x_{10}\end{aligned}$$

The p-value of the Pearson's chi-square test was 0.112>0.1, which may not be significant and cannot reject H_0. However, the AIC for model $\mathcal{M}'$ is 48680.93, while the AIC for model $\mathcal{M}'_0$ is 48842.28. This suggests that model $\mathcal{M}'$ provides a better fit than model $\mathcal{M}'_0$. Further, if we assumed that the degree of freedoms of $\mathcal{M}$ and $\mathcal{M}_0$ equal to their alternatives $\mathcal{M}'$ and $\mathcal{M}'_0$, the p-value of the Pearson's chi-square test $H_0 : \mathcal{M}$ against $H_1 : \mathcal{M}_0$ became 0.02, which is much more significant. This coincided with the result of our model selection procedure. Table 6 compares the estimated impacts of the constant parameters $(X_2, X_6, X_8, X_9, X_{10})$ with model $\mathcal{M}$ and $\mathcal{M}'$. The estimated impacts of the other functional factors (U, X_3, X_4, X_5, X_7) with model $\mathcal{M}'$ are presented in Fig. 13.

We also used the profile likelihood estimation method with our proposed initialization to estimate the effects of the independent variables with the selected model $\mathcal{M}$. The estimate of the impact of the mother's age (X_2) is 0.1044, the estimate of the effect of the mother's education (X_6) is -0.1087, the estimate of the effect of the child's sex (X_8) is -0.0486, the estimate of the impact of ethnicity (X_9) is -0.1444,

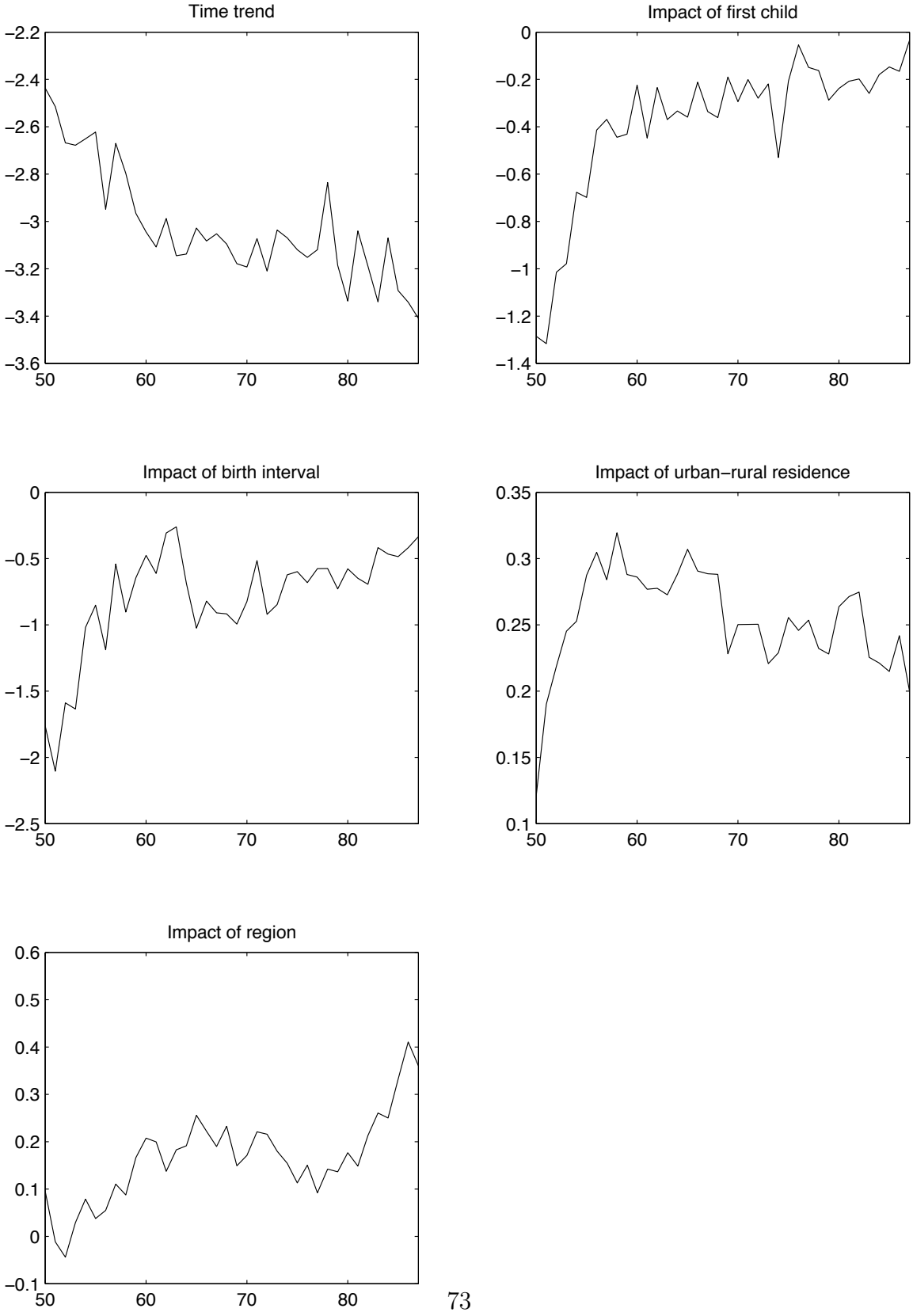

Figure 13: *Impacts of covariates on infant mortality with model $\mathcal{M}$'.*

Table 6: Estimated impacts of the constant parameters with model $\mathcal{M}$ and $\mathcal{M}'$

Model	X_2	X_6	X_8	X_9	X_{10}
$\mathcal{M}$	0.1041	-0.1088	-0.0492	-0.1371	-0.2058
$\mathcal{M}'$	0.0818	-0.1394	-0.0651	-0.1889	-0.3468

and the estimate of the impact of breast-feeding (X_{10}) is -0.2050. All these results are close to those of our proposed two-step estimation method.

Table 7: List of Covariates with Descriptive Statistics

Variable/category	mean	% of sample
Year of Birth (U, 1950-1987)	1972.2	-
Demographic Variables		
Maternal age (X_2)	25.7	
First Birth (X_3)		
yes	-	36.2
no	-	63.8
Previous birth interval (X_4, in months)	21.5	-
Socioeconomic Variables		
Place of Residence (X_5)		
rural	-	80.6
urban	-	19.4
Educational Attainment (X_6)		
illiterate/semiliterate	-	55.6
primary school+	-	44.4
Geographic Region of Residence (X_7)		
Beijing/Shanghai/Tianjin	-	6.8
Others	-	93.2

Ethnicity (X_9)		
Han	-	86.9
minority	-	13.1
Other Controls		
Sex of child (X_8)		
girl	-	48.2
boy	-	51.8
Breastfed (X_{10})		
yes	-	88.7
no	-	11.3
Total Number of Births		114,337

9 Conclusion and Future Works

In this article we propose a generalized multiparameter likelihood model together with an efficient two-step estimation procedure. The model is very a general semi-parametric model, which includes some popular models, such as the partially linear, varying-coefficients, and semi-varying generalized linear models, as special cases. We also discuss some possible alternative approaches for estimating the model, including backfitting and profile likelihood. We suggest a data-driven procedure for selecting the bandwidths, and develop an automatic procedure to identify constant parameters in the underlying model. Theoretical properties and simulation results show that our estimators of both constant parameters and functional parameters are accurate. Although in some cases, the profile likelihood approach may has better performance than our proposed two-step estimation, it requires more constraints which are not always satisfied. Further, profile likelihood performs poor when the sample size is small. The simulation results also suggest that our model selection procedure is effective when the sample size is moderately large.

Here we assume that the covariates involved is known. However, this is often not the case in practice. In the future, we are going to develop a data-driven variable selection method to decide which parameters should be included in the model. Inferences under this model is also of interest. Further, our approaches rely heavily on kernel smoothing, hence design sparsity is an important issue. Existing methods

in the regression case such as ridging (Seifert and Gasser, 2000) and interpolation (Hall and Turlach, 1997) may not be directly applicable. Finally, our approaches may be applied to different settings such as hazzard regression, longitudinal data analysis, etc..

REFERENCES

Aerts, M., and Claeskens, G. (1997), "Local polynomial estimators in multiparameter likelihood models," *Journal of the American Statistical Association*, 92, 1536-1545.

Akaike, H. (1970), "Statistical predictor identification," *Annals of the Institute of Statistical Mathematics*, 22, 203–217.

Bickel, P.J., Klaassen, C.A.J., Ritov, Y., and Wellner, J.A. (1993), *Efficient and adaptive estimation for semiparametric models*, Baltimore: The Johns Hopkins University Press.

Bunea, F. (2004), "Consistent covariate selection and post model selection inference in semiparametric regression," *The Annals of Statistics*, 32, 898–927.

Burnham, K. and Aderson, D. (2003), *Model Selection and Multimodel Inference: A Practical Information-Theoretical Approach* (2nd Edition), New York: Springer-Verlag.

Cai, J., Fan, J., Jiang, J., and Zhou, H. (2007), "Partially linear hazard regression for multivariate survival data," *Journal of the American Statistical Association*, 102, 538–551.

Cai, J., Fan, J., Jiang, J., and Zhou, H. (2008), "Partially linear hazard regression with varying-coefficients for multivariate survival data," *Journal of the Royal Statistical Society Series B*, 70, 141–158

Cai, Z., Fan, J., and Li, R. Z. (2000), "Efficient estimation and inferences for varying-coefficient models," *Journal of the American Statistical Association*, 95, 888-902.

Chen, X., Linton, O. and Van Keilegom, I. (2003), "Estimation of semiparameric models when the criterion function is not smooth," *Econometrica*, 71, 1591–1608.

Cheng, M.-Y., and Peng, L. (2007), "Variance reduction in multiparameter likelihood models," *Journal of the American Statistical Association*, 102, 293–304.

Cheng, M.-Y., Zhang, W. and Chen, L.-H. (2009), "Statistical Estimation in Generalized Multiparameter Likelihood Models," *Journal of the American Statistical Association*, 104, 1179–1191.

Cheng, M.-Y., and Wu, J.S. (2008), "Adapting to design sparsity in univariate and bivariate local linear regression," Manuscript.

Claeskens, G., and Aerts, M. (2000), "On local estimating equations in additive multiparameter models," *Statistics and Probability Letters*, 49, 139–148.

Clayton, D. and Cuzick, L. (1985), "Multivariate generalizations of the proportional hazards model (with discussion)," *Journal of the Royal Statistical Society A*, 148, 82–117.

Engle, R.F., Granger, C.W.J., Rice, J., and Weiss, A. (1986), "Semiparametric estimates of the relation between weather and electricity sales," *Journal of the American Statistical Association*, 81, 310–320.

Fan, J., and Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications*, London: Chapman and Hall.

Fan, J., Farmen, M., and Gijbels, I. (1998), "Local maximum likelihood estimation and inference," *Journal of the Royal Statistical Society Series B*, 60, 591 - 608.

Fan, J., Härdle, W., and Mammen, E. (1998), "Direct estimation of additive and linear components for high dimensional data," *The Annals of Statistics*, 26, 943–971.

Fan, J., and Huang, T. (2005), "Profile Likelihood Inferences on semiparametric varying-coefficient partially linear models," *Bernoulli*, 11, 1031–1057.

Fan, J., and Jiang, J. (2005), "Nonparametric inference for additive models," *Journal of the American Statistical Association*, 100, 890–907.

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

_____ (2002), "Variable selection for Cox's proportional hazards model and frailty model," *The Annals of Statistics*, 30, 74–99.

_____ (2004), "New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis," *Journal of the American Statistical Association*, 99, 710–723.

Fan, J., and Wong, W. H. (2000), "On Profile Likelihood: Comment," *Journal of the American Statistical Association*, 95, 468–471.

Fan, J. and Zhang, W. (1999), "Statistical estimation in varying coefficient models," *The Annals of Statistics*, 27, 1491–1518.

Hall, P., and Turlach, B.A. (1997), "Interpolation methods for adapting to sparse design in nonparametric regression," *Journal of the American Statistical Association*, 92, 466 – 472.

Härdle, W. Liang, H., and Gao, J. (2000), *Partially Linear Models*, Heidelberg: Physica-Verlag.

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

Hastie, T., and Tibshirani, R. (1993), "Varying-coefficient models," *Journal of the Royal Statistical Society Series B*, 55, 757–796.

Hoover, D. R., Rice, J. A., Wu, C. O. and Yang, L.-P. (1998). "Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data," *Biometrika*, 85, 80–822

Hurvich, C.M., Simonoff, J.S., and Tsai, C.-L. (1998), "Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion," *Journal of the Royal Statistical Society Series B*, 60, 271–293.

Irizarry, R.A. (2001), "Information and posterior probability criteria for model selection in local likelihood estimation," *Journal of the American Statistical Association*, 96, 303–315.

Keilegom, I. V. and Carroll, J. R. (2007) "Backfitting Versus Profiling in General Criterion Functions," *Statistica Sinica*, 17, 797–816.

Kullback, S. and Leibler, R. A. (1951), "On information and sufficiency," *Annals of Mathematical Statistics*, 22, 79-86.

Lam, C., and Fan, J. (2008), "Profile-kernel likelihood inference with diverging number of parameters," *The Annals of Statistics*, 36, 2232–2260.

Li, R., and Liang, H. (2008), "Variable selection in semiparametric regression modeling," *The Annals of Statistics*, 36, 261–286.

Loader, C. (1999), *Local regression and likelihood*, New York: Springer-Verlag.

Mack, Y. P., and Silverman, B. W. (1982), "Weak and strong uniform consistency of kernel regression estimates," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandle Gebiete*, **61**, 405–415.

Meeker, W. Q. and Escobar, L. A. (1998), *Statistical Methods for Reliability Data*, New York: John Wiley & Sons.

Murphy, S.A., and van der Vaart, A.W. (2000), "On profile likelihood" (with discussion), *Journal of the American Statistical Association*, 95, 449–465.

Nelson, W. (1984), "Fitting of fatigue curves with nonconstant standard deviation to data with runouts," *Journal of Testing and Evaluation*, 12, 140–146.

Schucany, W.R. (2004), "Kernel smoothers: an overview of curve estimators for the first graduate course in nonparametric statistics," *Statistical Science*, 19, 663–675.

Schwarz, G. (1978), "Estimating the dimension of a model," *The Annals of Statistics*, 6, 461–464.

Severini, T., and Wong, W. H. (1992), "Profile likelihood and conditionally parametric models," *The Annals of Statistics*, 20, 1768–1802.

Seifert, B., and Gasser, Th. (2000), "Data adaptive ridging in local polynomial regression," *Journal of Computational & Graphical Statistics,* 9, 338–360.

Stone, C. J. (1985). "Additive Regression and Other Nonparametric Models", *The Annals of Statistics*, 13, 689–705.

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

Wang, W. and Kececioglu, D. B. (2000). "Fitting the Weibull log-linear model to accelerated life-test data," *IEEE Transactions on Reliability*, 49, 217–223.

Xia, Y., and Li, W. K. (1999), "On the estimation and testing of functional-coefficient linear models," *Statistica Sinica*, 9, 735–58.

Zhang, W. and Lee, S. Y. (2000), "Variable bandwidth selection in varying-coefficient models," *Journal of Multivariate Analysis*, 74, 116–134

Zhang, W., Lee, S. Y. and Song, X. (2002). "Local polynomial fitting in semi-varying coefficient models," *Journal of Multivariate Analysis*, 82 166–188.

Appendices

A Proofs for Backfitting

Here we follow the proofs of Van Keilegom and Carroll (2007) and extend their results to the current setup. We first introduce two theorems in Chen, Linton and Van Keilegom (2003) (CLV hereafter) which we will make use later:

Theorem CLV1. *Let* $M(\boldsymbol{\theta}, \mathbf{a}) = E\left\{\frac{d}{d\boldsymbol{\theta}} L(\boldsymbol{\theta}, \mathbf{a})\right\}$ *and* $M_n(\boldsymbol{\theta}, \mathbf{a}) = \frac{d}{d\boldsymbol{\theta}} L_n(\boldsymbol{\theta}, \mathbf{a})$. *Suppose that* $\boldsymbol{\theta}_0$ *satisfies* $M(\boldsymbol{\theta}_0, \mathbf{a}_0) = 0$, *and that*

(1.1) $\|M(\hat{\boldsymbol{\theta}}, \hat{\mathbf{a}})\| \leqslant \inf_{\boldsymbol{\theta}} \|M(\boldsymbol{\theta}, \hat{\mathbf{a}})\| + o_p(1)$.

(1.2) For all $\delta > 0$, *there exists* $\epsilon > 0$ *such that* $\inf_{\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| > \delta} \|M(\boldsymbol{\theta}, \mathbf{a}_0)\| \geq \epsilon > 0$.

(1.3) For all $\boldsymbol{\theta}$, $M(\boldsymbol{\theta}, \mathbf{a})$ *is continuous in* $\mathbf{a}$ *at* $\mathbf{a} = \mathbf{a}_0$.

(1.4) $\|\hat{\mathbf{a}} - \mathbf{a}_0\| = o_p(1)$.

(1.5) For all positive sequences δ_n *with* $\delta_n = o(1)$,

$$\sup_{\boldsymbol{\theta}, \|\mathbf{a} - \mathbf{a}_0\| \leq \delta_n} \|M_n(\boldsymbol{\theta}, \mathbf{a}) - M(\boldsymbol{\theta}, \mathbf{a})\| = o_p(1).$$

Then, $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 = o_p(1)$.

Theorem CLV2. *Let* $\Gamma_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \mathbf{a}) = \frac{\partial}{\partial \boldsymbol{\theta}} M(\boldsymbol{\theta}, \mathbf{a})$ *and*

$$\Gamma_{\mathbf{a}}(\boldsymbol{\theta}, \mathbf{a})[\xi] = \lim_{\tau \to 0} \frac{1}{\tau} \{M(\boldsymbol{\theta}, \mathbf{a} + \tau\xi) - M(\boldsymbol{\theta}, \mathbf{a})\}$$

be the Gâteaux-derivative of $M(\boldsymbol{\theta}, \mathbf{a})$ *in the direction of* ξ. *Suppose that* $\boldsymbol{\theta}_0$ *satisfies* $M(\boldsymbol{\theta}_0, \mathbf{a}_0) = 0$, $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| = o_p(1)$, *and that*

(2.1) $\|M_n(\hat{\boldsymbol{\theta}}, \hat{\mathbf{a}})\| = \inf_{\boldsymbol{\theta}} \|M_n(\boldsymbol{\theta}, \hat{\mathbf{a}})\| + o_p(1/\sqrt{n})$.

(2.2) *(i)* $\Gamma_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \mathbf{a}_0)$ *exists for all* $\boldsymbol{\theta}$, *and is continuous at* $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

(ii) The matrix $\Gamma_{\boldsymbol{\theta}} \equiv \Gamma_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, \mathbf{a}_0)$ *is of full rank.*

(2.3) $\Gamma_{\mathbf{a}}(\boldsymbol{\theta}, \mathbf{a}_0)[\mathbf{a} - \mathbf{a}_0]$ *exists in all directions* $[\mathbf{a} - \mathbf{a}_0]$. *For all positive sequences* δ_n *with* $\delta_n = o(1)$,

(i) $\|M(\boldsymbol{\theta}, \mathbf{a}) - M(\boldsymbol{\theta}, \mathbf{a}_0) - \Gamma_{\mathbf{a}}(\boldsymbol{\theta}, \mathbf{a}_0)[\mathbf{a} - \mathbf{a}_0]\| \leq c\|\mathbf{a} - \mathbf{a}_0\|^2$ *for some constant* $c \geq 0$.

(ii) $\|\Gamma_{\mathbf{a}}(\boldsymbol{\theta}, \mathbf{a}_0)[\mathbf{a} - \mathbf{a}_0] - \Gamma_{\mathbf{a}}(\boldsymbol{\theta}_0, \mathbf{a}_0)[\mathbf{a} - \mathbf{a}_0]\| \leq o(1)\delta_n$.

(2.4) $\|\hat{\mathbf{a}} - \mathbf{a}_0\| = o_p(n^{-1/4})$.

(2.5) For all positive sequences δ_n *with* $\delta_n = o(1)$,

$$\sup_{\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq \delta_n, \|\mathbf{a} - \mathbf{a}_0\| \leq \delta_n} \|M_n(\boldsymbol{\theta}, \mathbf{a}) - M(\boldsymbol{\theta}, \mathbf{a}) - M_n(\boldsymbol{\theta}_0, \mathbf{a}_0)\| = o_p(n^{-1/2}).$$

(2.6) For some finite matrix V, $\sqrt{n}\{M_n(\boldsymbol{\theta}_0, \mathbf{a}_0) + \Gamma_{\mathbf{a}}(\boldsymbol{\theta}_0, \mathbf{a}_0)[\hat{\mathbf{a}} - \mathbf{a}]\} \xrightarrow{D} N(0, V)$.

Then, $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \xrightarrow{D} N(0, \Omega)$ *where* $\Omega = \mathcal{G}^{-1}(\boldsymbol{\theta}_0) V \mathcal{G}^{-1}(\boldsymbol{\theta}_0)$.

We utilize the two theorems mentioned before to prove theorem 1. Denote

$$M_{BF}(\boldsymbol{\theta}, \mathbf{a}) = \mathrm{E}\left\{ \frac{d}{d\boldsymbol{\theta}} L\left(\boldsymbol{\theta}, \mathbf{a}_{\boldsymbol{\theta}}\right) \right\}$$

and define

$$\begin{aligned}\Gamma_{BF,\mathbf{a}}(\boldsymbol{\theta},\mathbf{a})[\xi] &= \lim_{\tau\to 0}\frac{1}{\tau}\left\{M_{BF}(\boldsymbol{\theta},\mathbf{a}+\tau\xi)-M_{BF}(\boldsymbol{\theta},\mathbf{a})\right\}\\ &= \mathrm{E}\left\{\frac{\partial}{\partial\mathbf{a}}\mathrm{E}\left[\frac{d}{d\boldsymbol{\theta}}L\left(\boldsymbol{\theta},\mathbf{a}_{\boldsymbol{\theta}}\right)|U\right]\xi_{\boldsymbol{\theta}}\right\}\end{aligned}$$

for some function $\xi_{\boldsymbol{\theta}}$. We impose the following conditions:

(BF1) The bandwidth h satisfies $nh^4 \to 0$ as $n\to\infty$.

(BF2) (i) $\|\hat{\mathbf{a}}-\mathbf{a}_0\|_\infty = o_P(n^{-1/4})$.

(ii) $\sup\|\hat{\mathbf{a}}_{\hat{\boldsymbol{\theta}}}-\hat{\mathbf{a}}_{\boldsymbol{\theta}_0}\| = o_P(1)\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|$.

(iii) $\sup_u|n^{-1}\sum_{i=1}^n K_h(U_i-u)\frac{\partial}{\partial\mathbf{a}}\log f\big(Y_i;X_i,\boldsymbol{\theta}_0,\hat{\mathbf{a}}_{\boldsymbol{\theta}_0}(u)\big)| = o_P(n^{-1/2})$.

(BF3) (i) $L(\boldsymbol{\theta},\mathbf{a})$ is differentiable with respect to $\boldsymbol{\theta}$ and $\mathbf{a}$.

(ii) $\frac{\partial}{\partial\mathbf{a}}\mathrm{E}\left[\frac{\partial}{\partial\boldsymbol{\theta}}L\left(\boldsymbol{\theta},\mathbf{a}_{0\boldsymbol{\theta}}\right)|U\right]$ and $\frac{\partial}{\partial\boldsymbol{\theta}}\mathrm{E}\left[\frac{\partial}{\partial\mathbf{a}}L\left(\boldsymbol{\theta},\mathbf{a}_{0\boldsymbol{\theta}}\right)|U\right]$ exist for all $\boldsymbol{\theta}$, and they are equal.

(iii) $\mathrm{E}\left\{|\frac{\partial}{\partial\mathbf{a}}L(\boldsymbol{\theta}_0,\mathbf{a})|^2\right\}<\infty$ for all $\mathbf{a}$.

(iv) $\frac{\partial^{j+k+l}}{\partial\boldsymbol{\theta}^j\partial u^k\partial\mathbf{a}^l}\mathrm{E}\left\{\frac{\partial}{\partial\boldsymbol{\theta}}L(\boldsymbol{\theta},\mathbf{a})|U=u\right\}$ and $\frac{\partial^{j+k+l}}{\partial\boldsymbol{\theta}^j\partial u^k\partial\mathbf{a}^l}\mathrm{E}\left\{\frac{\partial}{\partial\mathbf{a}}L(\boldsymbol{\theta},\mathbf{a})|U=u\right\}$ exist for $0\le j+k+l\le 2$ and are bounded.

(v) $\mathcal{G}(\boldsymbol{\theta})$ exists for $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}_0$, is continuous at $\boldsymbol{\theta}_0$ and $\mathcal{G}(\boldsymbol{\theta}_0)$ is of full rank.

(BF4) $\int_0^\infty\sqrt{\log N(\epsilon^{1/s_l},\hat{\mathcal{A}},\|\cdot\|_\infty)}d\epsilon<\infty$ for $l=1,\cdots,q$, where $\hat{\mathcal{A}}=\{\mathbf{a}_{\boldsymbol{\theta}}(\cdot):\mathbf{a}\in\mathcal{A},\boldsymbol{\theta}\in\Theta\}$, $N(\epsilon,\mathcal{A},\|\cdot\|)$ is the minimal number of balls $\{\boldsymbol{\eta}:\|\boldsymbol{\eta}-\boldsymbol{\theta}\|<\epsilon\}$ of radius ϵ needed to cover $\mathcal{A}$.

(BF5) (i) For all $\delta > 0$, there exists a $\epsilon > 0$ such that $\inf_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|>\delta} \|M_{BF}(\boldsymbol{\theta}, \mathbf{a}_0)\| \geq \epsilon$

(ii) For all $\boldsymbol{\theta}$, $M_{BF}(\boldsymbol{\theta}, \mathbf{a})$ is continuous in $\mathbf{a}$ at $\mathbf{a}_0$ (with respect to the $\|\cdot\|_\infty$ norm).

(iii) $\Gamma_{BF,\mathbf{a}}(\boldsymbol{\theta}, \mathbf{a}_0)[\mathbf{a} - \mathbf{a}_0]$ exists.

For the proofs below, we restrict our attention to the case $q = 1$. The general case $q \geq 1$ can be obtained in a similar way, but requires more complex notations.

Lemma A.1. Assume (BF1)–(BF5) and (S2)–(S4) listed in Appendix C hold. Then,

$$n^{-1}\sum_{i=1}^{n} \mathrm{E}_U\left(\frac{K_h(U_i-U)}{\pi(U)}\frac{\partial}{\partial\boldsymbol{\theta}}\mathbf{a}_0\boldsymbol{\theta}_0(U)\left[\frac{\partial}{\partial\mathbf{a}}\log f\left(Y_i; X_i, \boldsymbol{\theta}_0, \mathbf{a}_0(U_i)\right) - \frac{\partial}{\partial\mathbf{a}}\log f\left(Y_i; X_i, \boldsymbol{\theta}_0, \hat{\mathbf{a}}_0(U)\right)\right]\right)$$
$$= \mathrm{E}_{U_1,U_2,Y}\Bigg(\frac{K_h(U_1-U_2)}{\pi(U_2)}\frac{\partial}{\partial\boldsymbol{\theta}}\mathbf{a}_0\boldsymbol{\theta}_0(U2)\Bigg[\frac{\partial}{\partial\mathbf{a}}\log f\{Y_i; X_i, \boldsymbol{\theta}_0, \mathbf{a}_0(U_1)\}$$
$$-\frac{\partial}{\partial\mathbf{a}}\log f\{Y_i; X_i, \boldsymbol{\theta}_0, \hat{\mathbf{a}}_{\boldsymbol{\theta}_0}(U_2)\}\Bigg]\Bigg) + o_P(n^{-1/2}),$$

where the expectations are taken conditionally on (U_i, Y_i).

Proof: This result immediately from Keilegom and Carroll (2007).

Lemma A.2. Assume (BF1)–(BF5) and (S2)–(S4) listed in Appendix C hold. Then,

$$\Gamma_{BF,\mathbf{a}}(\boldsymbol{\theta}_0, \mathbf{a}_0)[\hat{\mathbf{a}} - \mathbf{a}_0] = n^{-1}\sum_{i=1}^{n}\frac{\partial}{\partial\mathbf{a}}\log f\left(Y_i; X_i, \boldsymbol{\theta}_0, \mathbf{a}_0(U_i)\right)\frac{\partial}{\partial\boldsymbol{\theta}}\mathbf{a}_0\boldsymbol{\theta}_0(U_i) + o_P(n^{-1/2}) \tag{A.1}$$

Proof:

$$
\begin{aligned}
&\Gamma_{BF,\mathbf{a}}(\boldsymbol{\theta}_0, \mathbf{a}_0)[\hat{\mathbf{a}} - \mathbf{a}_0] \\
&= \lim_{\tau \to 0} \frac{1}{\tau} \mathrm{E} \left\{ \frac{\partial}{\partial \boldsymbol{\theta}} L \left[\boldsymbol{\theta}_0, \{\mathbf{a}_0 + \tau(\hat{\mathbf{a}} - \mathbf{a}_0)_{\boldsymbol{\theta}_0}\} \right] - \frac{\partial}{\partial \boldsymbol{\theta}} L \left[\boldsymbol{\theta}, \mathbf{a}_{0\boldsymbol{\theta}_0} \right] \right\} \\
&= \mathrm{E} \left\{ \frac{\partial}{\partial \mathbf{a}} \mathrm{E} \left(\frac{\partial}{\partial \boldsymbol{\theta}} L \left[\boldsymbol{\theta}_0, \mathbf{a}_{0\boldsymbol{\theta}} \right] | U \right) (\hat{\mathbf{a}} - \mathbf{a}_0)_{\boldsymbol{\theta}_0} \right\} \qquad \text{(A.2)} \\
&= \mathrm{E} \left\{ \frac{\partial}{\partial \boldsymbol{\theta}} \mathrm{E} \left(\frac{\partial}{\partial \mathbf{a}} L \left[\boldsymbol{\theta}_0, \mathbf{a}_{0\boldsymbol{\theta}} \right] | U \right) (\hat{\mathbf{a}} - \mathbf{a}_0)_{\boldsymbol{\theta}_0} \right\} \\
&= -\mathrm{E} \left\{ \frac{\partial}{\partial \mathbf{a}} \mathrm{E} \left(\frac{\partial}{\partial \mathbf{a}} L \left[\boldsymbol{\theta}_0, \mathbf{a}_{0\boldsymbol{\theta}} \right] | U \right) (\hat{\mathbf{a}} - \mathbf{a}_0)_{\boldsymbol{\theta}_0} \frac{\partial}{\partial \boldsymbol{\theta}} \mathbf{a}_{0\boldsymbol{\theta}_0} \right\},
\end{aligned}
$$

since $\mathrm{E}\left(\frac{\partial}{\partial \mathbf{a}} L \left[\boldsymbol{\theta}, \mathbf{a}_{0\boldsymbol{\theta}}\right] | U\right) = 0$ for all $\boldsymbol{\theta}$. Let $g(U) = \frac{\partial}{\partial \boldsymbol{\theta}} \mathbf{a}_{0\boldsymbol{\theta}_0}(U)$ and $\mathcal{H}(Y, \mathbf{a}) = \frac{\partial}{\partial \mathbf{a}} \log f(Y; X, \boldsymbol{\theta}_0, \mathbf{a}(U))$, then the right hand side of (A.1) equals

$$
n^{-1} \sum_{i=1}^{n} \mathrm{E}_U \left(\frac{K_h(U_i - U)}{\pi(U)} g(U) [\mathcal{H}\{Y_i, \mathbf{a}_0(U_i)\} - \mathcal{H}\{Y_i, \hat{\mathbf{a}}_{\boldsymbol{\theta}_0}(U_i)\}] \right) + o_P(n^{-1/2}),
$$

since $n^{-1} \sum_{i=1}^{n} K_h(U_i - u) \mathcal{H}\{Y_i, \hat{\mathbf{a}}_{\boldsymbol{\theta}_0}(u) = o_P(n^{-1/2})$ uniformly in u by assumption (BF2)(iii). Using Lemma A.1 the latter expression can be written as

$$
\begin{aligned}
&\mathrm{E}_{U_1, U_2, Y} \left(\frac{K_h(U_1 - U_2)}{\pi(U_2)} g(U_2) [\mathcal{H}\{Y, \mathbf{a}_0(U_1)\} - \mathcal{H}\{Y, \hat{\mathbf{a}}_{\boldsymbol{\theta}_0}(U_2)\}] \right) + o_P(n^{-1/2}) \\
&= \mathrm{E}_{U_1, U_2} \left(\frac{K_h(U_1 - U_2)}{\pi(U_2)} g(U_2) [\kappa\{U_1, \mathbf{a}_0(U_1)\} - \kappa\{U_1, \hat{\mathbf{a}}_{\boldsymbol{\theta}_0}(U_2)\}] \right) + o_P(n^{-1/2}),
\end{aligned}
$$

where $\kappa(U, \mathbf{a}) = \mathrm{E}[\mathcal{H}(Y, \mathbf{a}) | U]$. Using Taylor and assumptions (BF1), (BF3)(iv),

(S2) and (S3) this can be written as

$$\begin{aligned}
\mathrm{E}_{U_2}\left(\frac{\mathrm{E}_{U_1}\{K_h(U_1-U_2)\}}{\pi(U_2)}g(U_2)[\kappa\{U_2,\mathbf{a}_0(U_2)\}-\kappa\{U_2,\hat{\mathbf{a}}_{\boldsymbol{\theta}_0}(U_2)\}]\right)& \\
+\mathrm{E}_{U_2}\left(\frac{\mathrm{E}_{U_1}\{(U_1-U_2)K_h(U_1-U_2)\}}{\pi(U_2)}g(U_2)\right.& \\
\left.\frac{d}{du}[\kappa\{u,\mathbf{a}_0(u)\}-\kappa\{u,\hat{\mathbf{a}}_{\boldsymbol{\theta}_0}(U_2)\}]|_{u=U_2}\right)+o_P(n^{-1/2})& \\
=\mathrm{E}\left(g(U)\kappa\{U,\mathbf{a}_0(U)\}-\kappa\{U,\hat{\mathbf{a}}_{\boldsymbol{\theta}_0}(U)\}\right)+o_P(n^{-1/2})& \\
=-\mathrm{E}\left[g(U)\frac{\partial}{\partial\mathbf{a}}\mathrm{E}[\mathcal{H}\{\mathbf{a}(U)\}|U]\{\hat{\mathbf{a}}_{\boldsymbol{\theta}_0}(U)-\mathbf{a}_0(U)\}\right]+o_P(n^{-1/2}),&
\end{aligned}$$

since $\sup\|\hat{\mathbf{a}}_{\boldsymbol{\theta}_0}-\mathbf{a}_0\|=o_P(n^{-1/4})$. The later expression equals $\Gamma_{BF,\mathbf{a}}(\boldsymbol{\theta}_0,\mathbf{a}_0)[\hat{\mathbf{a}}-\mathbf{a}_0]+o_P(n^{-1/2})$ by (A.2). Hence, the result follows.

Proof of Theorem 1. We make use of Theorem CLV2, which states primitive conditions under which $\hat{\boldsymbol{\theta}}_{BF}$ is asymptotically normal. First of all, we need to show that $\hat{\boldsymbol{\theta}}_{BF}-\boldsymbol{\theta}_0=o_P(1)$. For this, we verify the conditions of Theorem CLV1. Condition (1.1) holds by definition of $\hat{\boldsymbol{\theta}}_{BF}$, while conditions (1.2)–(1.4) are guaranteed by assumptions (BF2) and (BF5). Finally, condition (1.5) is weaker than condition (2.5) of Theorem 2 of CLV, which we verify below. Next, we verify conditions (2.1)–(2.6) of Theorem 2 in CLV. Condition (2.1) is also valid by construction of the estimator $\hat{\boldsymbol{\theta}}_{BF}$, while condition (2.2) follows from assumption (BF3)(v). Since $\Gamma_{BF,\mathbf{a}}(\boldsymbol{\theta},\mathbf{a}_0)[\mathbf{a}-\mathbf{a}_0]=\mathrm{E}\left\{\frac{\partial}{\partial\mathbf{a}}d(U,\mathbf{a}_0)[\mathbf{a}_{\boldsymbol{\theta}}(U)-\mathbf{a}_0(U)]\right\}$, where

$d(U, \mathbf{a}) = \mathrm{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}} L(\boldsymbol{\theta}, \mathbf{a}_{\boldsymbol{\theta}})|U\right]$, we have

$$\begin{aligned} & M_{BF}(\boldsymbol{\theta}, \mathbf{a}) - M_{BF}(\boldsymbol{\theta}, \mathbf{a}_0) - \Gamma_{BF,\mathbf{a}}(\boldsymbol{\theta}, \mathbf{a}_0)[\mathbf{a} - \mathbf{a}_0] \\ & = \mathrm{E}\left\{ d(U, \mathbf{a}) - d(U, \mathbf{a}_0) - \frac{\partial}{\partial \mathbf{a}} d(U, \mathbf{a}_0)[\mathbf{a}_{\boldsymbol{\theta}}(U) - \mathbf{a}_0(U)] \right\} \\ & = \frac{1}{2}\mathrm{E}\left\{ \frac{\partial^2}{\partial \mathbf{a}^2} d(U, \xi)[\mathbf{a}_{\boldsymbol{\theta}}(U) - \mathbf{a}_0(U)]^2 \right\}, \end{aligned} \tag{A.3}$$

where $\xi(U)$ is in between $\mathbf{a}_{\boldsymbol{\theta}}(U)$ and $\mathbf{a}_0(U)$. Hence the norm of (A.3) is bounded by a constant times $\|\mathbf{a} - \mathbf{a}_0\|_\infty^2$. This shows the first part of (2.3). For the second part, it follows from the proof of Theorem 2 in CLV that it suffices to show that

$$\|\Gamma_{BF,\mathbf{a}}(\hat{\boldsymbol{\theta}}, \mathbf{a}_0)[\hat{\mathbf{a}} - \mathbf{a}_0] - \Gamma_{BF,\mathbf{a}}(\boldsymbol{\theta}_0, \mathbf{a}_0)[\hat{\mathbf{a}} - \mathbf{a}_0]\| = o_p(1)\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|,$$

and this follows from (BF2) (iii), (BF3) (iv) and (S2). Next, (2.4) follows from (BF2) (i), while (2.5) is guaranteed by (BF4). It remains to verify (2.6). Since $\Gamma_{BF,\mathbf{a}}(\boldsymbol{\theta}_0, \mathbf{a}_0)[\hat{\mathbf{a}} - \mathbf{a}_0]$ and $M_{nBF}(\boldsymbol{\theta}_0, \mathbf{a}_0)$, where $M_{nBF}(\boldsymbol{\theta}_0, \mathbf{a}_0)$ is sample version of $M_{BF}(\boldsymbol{\theta}_0, \mathbf{a}_0)$, are sums of i.i.d. terms plus negligible terms of lower order (see Lemma A.2.), this follows immediately. The asymptotic normality of $\hat{\boldsymbol{\theta}}_{BF}$ now follows.

B Proofs for Profile Likelihood Estimation

Denote

$$M_{PR}\left(\boldsymbol{\theta}, \mathbf{a}, \mathbf{a}'\right) = \mathrm{E}\left\{\frac{\partial}{\partial \boldsymbol{\theta}} L\left(\boldsymbol{\theta}, \mathbf{a}_{\boldsymbol{\theta}}\right) + \frac{\partial}{\partial \mathbf{a}} L\left(\boldsymbol{\theta}, \mathbf{a}_{\boldsymbol{\theta}}\right) \mathbf{a}'_{\boldsymbol{\theta}}\right\}$$

and

$$\Gamma_{PR,\mathbf{a},\mathbf{a}'}(\boldsymbol{\theta}, \mathbf{a}, \mathbf{a}')[\xi, \zeta] = \lim_{\tau \to 0} \frac{1}{\tau} \left\{M_{PR}(\boldsymbol{\theta}, \mathbf{a} + \tau\xi, \mathbf{a}' + \tau\zeta) - M_{PR}(\boldsymbol{\theta}, \mathbf{a}, \mathbf{a}')\right\}.$$

The assumptions we need to impose are the following:

Regularity Conditions

(PR1) $\mathbf{a}_0$ is partially differentiable with respect to the components of $\boldsymbol{\theta}$, $\|\tilde{\mathbf{a}}_{\boldsymbol{\theta}} - \mathbf{a}_0\|_\infty = o_P(n^{-1/4})$, and $\|\frac{\partial}{\partial \boldsymbol{\theta}} \tilde{\mathbf{a}}_{\boldsymbol{\theta}} - \frac{\partial}{\partial \boldsymbol{\theta}} \mathbf{a}_{0\boldsymbol{\theta}}\|_\infty = o_P(n^{-1/4})$

(PR2) (i) $L(\boldsymbol{\theta}, \mathbf{a})$ is differentiable with respect to $\boldsymbol{\theta}$ and $\mathbf{a}$.

(ii) $\frac{\partial}{\partial \mathbf{a}} \mathrm{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}} L\left(\boldsymbol{\theta}, \mathbf{a}_{0\boldsymbol{\theta}}\right) | U\right]$ and $\frac{\partial}{\partial \boldsymbol{\theta}} \mathrm{E}\left[\frac{\partial}{\partial \mathbf{a}} L\left(\boldsymbol{\theta}, \mathbf{a}_{0\boldsymbol{\theta}}\right) | U\right]$ exist for all $\boldsymbol{\theta}$, and they are equal.

(iii) $\frac{\partial}{\partial \mathbf{a}^2} \mathrm{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}} L\left(\boldsymbol{\theta}, \mathbf{a}_{0\boldsymbol{\theta}}\right) | U\right]$ and $\frac{\partial}{\partial \mathbf{a}^2} \mathrm{E}\left[\frac{\partial}{\partial \mathbf{a}} L\left(\boldsymbol{\theta}, \mathbf{a}_{0\boldsymbol{\theta}}\right) | U\right]$ exist for all $\boldsymbol{\theta}$ and $\mathbf{a}$ and are bounded within the support of U.

(iv) $\mathcal{G}(\boldsymbol{\theta})$ exists for $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}_0$, is continuous at $\boldsymbol{\theta}_0$ and $\mathcal{G}(\boldsymbol{\theta}_0)$ is of full rank.

(PR3) $\int_0^\infty \sqrt{\log N(\epsilon^{1/s_l}, \hat{\mathcal{A}}, \|\cdot\|_\infty)} d\epsilon < \infty$ for $l = 1, \cdots, q$, where $\hat{\mathcal{A}} = \{\mathbf{a}_{\boldsymbol{\theta}}(\cdot) : \mathbf{a} \in \mathcal{A}, \boldsymbol{\theta} \in \Theta\}$, $N(\epsilon, \mathcal{A}, \|\cdot\|)$ is the minimal number of balls $\{\boldsymbol{\eta} : \|\boldsymbol{\eta} - \boldsymbol{\theta}\| < \epsilon\}$.

(PR4) (i) For all $\delta > 0$, there exists a $\epsilon > 0$ such that $\inf_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|>\delta} \|M_{PR}\left(\boldsymbol{\theta}, \mathbf{a}_0, \frac{\partial}{\partial \boldsymbol{\theta}}\mathbf{a}_{0\boldsymbol{\theta}}\right)\| \geq \epsilon$

(ii) For all $\boldsymbol{\theta}$, $M_{PR}(\boldsymbol{\theta}, \mathbf{a}, \mathbf{a}')$ is continuous in $(\mathbf{a}, \mathbf{a}')$ at $\left(\mathbf{a}_0, \frac{\partial}{\partial \boldsymbol{\theta}}\mathbf{a}_{0\boldsymbol{\theta}}\right)$ (with respect to the $\|\cdot\|_\infty$ norm).

Lemma B.1. Assume (PR1)–PR(4). Then, for any $\mathbf{a}(\cdot)$, $\mathbf{a}'(\cdot)$ and $\boldsymbol{\theta}$, $\Gamma_{PR,\mathbf{a},\mathbf{a}'}(\boldsymbol{\theta}, \mathbf{a}, \mathbf{a}')[\xi, \zeta] = 0$.

Proof:

$$
\begin{aligned}
&\Gamma_{PR,\mathbf{a},\mathbf{a}'}(\boldsymbol{\theta}, \mathbf{a}, \mathbf{a}')[\xi, \zeta] \\
&= \lim_{\tau\to 0} \frac{1}{\tau} \mathrm{E}\left\{ \frac{\partial}{\partial \boldsymbol{\theta}} L\left[\boldsymbol{\theta}, (\mathbf{a}_0 + \tau\xi)_{\boldsymbol{\theta}}\right] - \frac{\partial}{\partial \boldsymbol{\theta}} L\left[\boldsymbol{\theta}, \mathbf{a}_{0\boldsymbol{\theta}_0}\right] \right\} \\
&+ \lim_{\tau\to 0} \frac{1}{\tau} \mathrm{E}\left\{ \frac{\partial}{\partial \mathbf{a}} L\left[\boldsymbol{\theta}, (\mathbf{a}_0 + \tau\xi)_{\boldsymbol{\theta}}\right] - \frac{\partial}{\partial \mathbf{a}} L\left[\boldsymbol{\theta}, \mathbf{a}_{0\boldsymbol{\theta}}\right] (\mathbf{a}_0 + \tau\zeta)_{\boldsymbol{\theta}} \right\} \\
&+ \lim_{\tau\to 0} \frac{1}{\tau} \mathrm{E}\left\{ \frac{\partial}{\partial \mathbf{a}} L\left[\boldsymbol{\theta}, \mathbf{a}_{0\boldsymbol{\theta}}\right] \tau\zeta_{\boldsymbol{\theta}} \right\}
\end{aligned}
\tag{B.1}
$$

The third term of (B.1) equals $\mathrm{E}\left\{\mathrm{E}\left(\frac{\partial}{\partial \mathbf{a}} L[\boldsymbol{\theta}, \mathbf{a}_{0\boldsymbol{\theta}}]|U\right)\zeta_{\boldsymbol{\theta}}\right\} = 0$, since $\mathrm{E}\left(\frac{\partial}{\partial \mathbf{a}} L[\boldsymbol{\theta}, \mathbf{a}_{0\boldsymbol{\theta}}]|U\right) = 0$. The first term of (B.1) can be written as

$$
\mathrm{E}\left\{ \left(\frac{\partial}{\partial \mathbf{a}} \mathrm{E}\left(\frac{\partial}{\partial \boldsymbol{\theta}} L\left[\boldsymbol{\theta}, \mathbf{a}_{0\boldsymbol{\theta}}\right] |U \right) \xi_{\boldsymbol{\theta}} \right) \right\},
$$

and the second term equals

$$
\begin{aligned}
&\mathrm{E}\left\{ \left(\frac{\partial}{\partial \mathbf{a}} \mathrm{E}\left(\frac{\partial}{\partial \mathbf{a}} L\left[\boldsymbol{\theta}, \mathbf{a}_{0\boldsymbol{\theta}}\right] |U \right) \xi_{\boldsymbol{\theta}} \right) \frac{\partial}{\partial \boldsymbol{\theta}} \mathbf{a}_{0\boldsymbol{\theta}} \right\} \\
&= -\mathrm{E}\left\{ \left(\frac{\partial}{\partial \boldsymbol{\theta}} \mathrm{E}\left(\frac{\partial}{\partial \mathbf{a}} L\left[\boldsymbol{\theta}, \mathbf{a}_{0\boldsymbol{\theta}}\right] |U \right) \xi_{\boldsymbol{\theta}} \right) \right\},
\end{aligned}
$$

since

$$
\frac{\partial}{\partial \boldsymbol{\theta}} \mathrm{E}\left(\frac{\partial}{\partial \mathbf{a}} L\left[\boldsymbol{\theta}, \mathbf{a}_{0\boldsymbol{\theta}}\right] |U \right) + \frac{\partial}{\partial \mathbf{a}} \mathrm{E}\left(\frac{\partial}{\partial \mathbf{a}} L\left[\boldsymbol{\theta}, \mathbf{a}_{0\boldsymbol{\theta}}\right] |U \right) \frac{\partial}{\partial \boldsymbol{\theta}} \mathbf{a}_{0\boldsymbol{\theta}} = 0
$$

by $\mathrm{E}\left(\frac{\partial}{\partial \mathbf{a}} L[\boldsymbol{\theta}, \mathbf{a}_{0\boldsymbol{\theta}}]|U\right) = 0$. The result now follows by (PR2).

Proof of Theorem 2. In a manner similar to the backfitting procedure, we proceed by checking the primitive conditions of Theorem CLV2 introduced in the previous section. The verification of the conditions in that theorem is much the same as for the backfitting procedure, except for conditions (2.3) and (2.5). From the proof of Lemma B.1. that $\Gamma_{PR,\mathbf{a},\mathbf{a}'}\left(\boldsymbol{\theta}, \mathbf{a}_0, \frac{\partial}{\partial \boldsymbol{\theta}}\mathbf{a}_{0\boldsymbol{\theta}}[\mathbf{a}-\mathbf{a}_0, \eta - \frac{\partial}{\partial \boldsymbol{\theta}}\mathbf{a}_{0\boldsymbol{\theta}}]\right) = \mathrm{E}\left\{\frac{\partial}{\partial \mathbf{a}} d_1(U, \mathbf{a}_0)[\mathbf{a}_{\boldsymbol{\theta}}(U) - \mathbf{a}_0(U)]\right\} + \mathrm{E}\left\{\frac{\partial}{\partial \mathbf{a}} d_2(U, \mathbf{a}_0)[\mathbf{a}_{\boldsymbol{\theta}}(U) - \mathbf{a}_0(U)]\frac{\partial}{\partial \boldsymbol{\theta}}\mathbf{a}_0(U)\right\}$, where $d_1(U, \mathbf{a}) = \mathrm{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}} L\{\boldsymbol{\theta}, \mathbf{a}_{\boldsymbol{\theta}}(U)\}|U\right]$ and $d_2(U, \mathbf{a}) = \mathrm{E}\left[\frac{\partial}{\partial \mathbf{a}} L\{\boldsymbol{\theta}, \mathbf{a}_{\boldsymbol{\theta}}(U)\}|U\right]$, we have

$$
\begin{aligned}
&M_{PR}(\boldsymbol{\theta}, \mathbf{a}, \eta) - M_{PR}(\boldsymbol{\theta}, \mathbf{a}_0, \frac{\partial}{\partial \boldsymbol{\theta}}\mathbf{a}_0) - \Gamma_{PR,\mathbf{a},\eta}\left(\boldsymbol{\theta}, \mathbf{a}_0, \frac{\partial}{\partial \boldsymbol{\theta}}\mathbf{a}_{0\boldsymbol{\theta}}[\mathbf{a}-\mathbf{a}_0, \eta - \frac{\partial}{\partial \boldsymbol{\theta}}\mathbf{a}_{0\boldsymbol{\theta}}\right) \\
&= \mathrm{E}\left\{d_1(U, \mathbf{a}) - d_1(U, \mathbf{a}_0) - \frac{\partial}{\partial \mathbf{a}} d_1(U, \mathbf{a}_0)(\mathbf{a}_{\boldsymbol{\theta}} - \mathbf{a}_0)(U)\right\} \\
&+ \mathrm{E}\left\{d_2(U, \mathbf{a}) - d_2(U, \mathbf{a}_0) - \frac{\partial}{\partial \mathbf{a}} d_2(U, \mathbf{a}_0)(\mathbf{a} - \mathbf{a}_0)(U)\eta(U)\right\} \\
&+ \mathrm{E}\left\{d_2(U, \mathbf{a}_0)\left[\eta - \frac{\partial}{\partial \boldsymbol{\theta}}\mathbf{a}_{0\boldsymbol{\theta}}\right](U)\right\} \\
&+ \mathrm{E}\left\{\frac{\partial}{\partial \mathbf{a}} d_2(U, \mathbf{a}_0)(\mathbf{a}_{\boldsymbol{\theta}} - \mathbf{a}_0)(U)_+\left[\eta - \frac{\partial}{\partial \boldsymbol{\theta}}\mathbf{a}_{0\boldsymbol{\theta}}\right](U)\right\} \\
&= \frac{1}{2}\mathrm{E}\left\{\frac{\partial^2}{\partial \boldsymbol{\theta}^2} d_1(U, \zeta_1)(\mathbf{a}_{\boldsymbol{\theta}} - \mathbf{a}_0)^2(U)\right\} \\
&+ \frac{1}{2}\mathrm{E}\left\{\frac{\partial^2}{\partial \boldsymbol{\theta}^2} d_2(U, \zeta_2)(\mathbf{a}_{\boldsymbol{\theta}} - \mathbf{a}_0)^2(U)\frac{\partial}{\partial \boldsymbol{\theta}}\mathbf{a}_{0\boldsymbol{\theta}}(U)\right\} \\
&+ \mathrm{E}\left\{\frac{\partial^2}{\partial \boldsymbol{\theta}^2} d_2(U, \mathbf{a}_0(\mathbf{a}_{\boldsymbol{\theta}} - \mathbf{a}_0)(U)\left[\eta - \frac{\partial}{\partial \boldsymbol{\theta}}\mathbf{a}_{0\boldsymbol{\theta}}\right]\right\}
\end{aligned}
\tag{B.2}
$$

since $d_2(U, \mathbf{a}_0) = 0$, where $\zeta_1(U)$ and $\zeta_2(U)$ are in between $\mathbf{a}_{\boldsymbol{\theta}}(U)$ and $\mathbf{a}_0(U)$. Hence the norm of (B.2) is bounded by a constant times $\|(\mathbf{a} - \mathbf{a}_0, \eta - \frac{\partial}{\partial \boldsymbol{\theta}} \mathbf{a}_{0\boldsymbol{\theta}})\|_\infty^2$. This shows the first part of (2.3), while the second part follows by Lemma B.1. Finally, (2.5) is guaranteed by assumption (PR3). The result now follows.

C Proofs for 2-Step Estimation

We impose the following technical conditions.

Regularity Conditions

(S1) Let X_j be the jth component of $\mathbf{X}$. We assume that $EX_j^{2s} < \infty$, $j = 1, \cdots, p$, for some $s > 2$.

(S2) Assume that $\mathbf{a}_j(\cdot)$ is twice continuously differentiable with a non vanishing second derivative, $\ddot{\mathbf{a}}_j(\cdot)$, $j = 1, \cdots, \ell$.

(S3) The marginal density $\pi(\cdot)$ of U has a continuous second derivative, has a compact support, and is bounded below.

(S4) The kernel function $K(\cdot)$ is a bounded, symmetric density function, has a compact support, and satisfies a Lipschitz condition.

(S5) As $n \to \infty$, $h \to 0$, $nh^\gamma / \log h \to \infty$, $h_1 \to 0$, $nh_1^\gamma / \log h_1 \to \infty$, for any $\gamma > s/(s-2)$ with s given in Condition (S1).

(S6) Assume that $f(y;\ \mathbf{X},\ \boldsymbol{\theta},\ \mathbf{z}) > 0$, and $f(y;\ \mathbf{X},\ \boldsymbol{\theta},\ \mathbf{z})$ has a continuous, bounded third derivative with respect to $(\boldsymbol{\theta},\ \mathbf{z})$.

(S7) There exists a positive constant λ_0 such that the smallest eigenvalue of $\boldsymbol{\mathcal{I}}(\boldsymbol{\gamma})$ is greater than λ_0. Also, assume that $E(\mathbf{X}\mathbf{X}^T)$ is positive definite.

Lemma C.1. Let $(Z_1, W_1), ..., (Z_n, W_n)$ be iid observations from a bivariate random vector (Z, W). Assume further that $E|W|^s < \infty$ and $\sup_x \int |y|^s \zeta(x,\ y)dy < \infty$, where ζ denotes the joint density of (Z, W). Let K be a bounded positive function with a bounded support, satisfying a Lipschitz condition. Then

$$\sup_{x\in\mathcal{D}} \Big| n^{-1} \sum_{i=1}^{n} \Big\{ K_h(Z_i - x)W_i - E[K_h(Z_i - x)W_i] \Big\} \Big| = O_P\Big(\{nh/\log(1/h)\}^{-1/2} \Big)$$

provided that $n^{2\varepsilon-1}h \longrightarrow \infty$ for some $\varepsilon < 1 - s^{-1}$ and $\mathcal{D}$ is a compact set.

Proof: This follows immediately from the result of Mack and Silverman (1982).

Proof of Theorem 3. By abuse of notation, from here on, we use $\mathbf{a}_j$, $\mathbf{b}_j$, and $\mathbf{V}_c$ to denote the true value of $\mathbf{a}_j(u)$, $\dot{\mathbf{a}}_j(u)$, and $\mathbf{V}_c(u)$ for a generic point u. Define

$$\tilde{\mathbf{z}}_i = (\boldsymbol{\theta}^{\mathrm{T}},\ \mathbf{c}_i^{\mathrm{T}})^{\mathrm{T}}, \quad \mathbf{c}_i = \Big(\mathbf{X}_i^{\mathrm{T}} \Big\{ \mathbf{a}_1 + \mathbf{b}_1(U_i - u) \Big\},\ \cdots,\ \mathbf{X}_i^{\mathrm{T}} \Big\{ \mathbf{a}_\ell + \mathbf{b}_\ell(U_i - u) \Big\} \Big)^{\mathrm{T}},$$

$$\boldsymbol{\xi} = \big(\boldsymbol{\theta}^{\mathrm{T}},\ \mathbf{a}_1^{\mathrm{T}},\ \mathbf{b}_1^{\mathrm{T}},\ \cdots,\ \mathbf{a}_\ell^{\mathrm{T}},\ \mathbf{b}_\ell^{\mathrm{T}} \big)^{\mathrm{T}}, \quad \boldsymbol{\gamma}_i = \big(\boldsymbol{\theta}^{\mathrm{T}},\ \mathbf{X}_i^{\mathrm{T}}\mathbf{a}_1(U_i),\ \cdots,\ \mathbf{X}_i^{\mathrm{T}}\mathbf{a}_\ell(U_i) \big)^{\mathrm{T}},$$

$$\mathbf{H}_i = \mathrm{diag}\left(\mathbf{I}_q,\ \mathbf{I}_\ell \otimes (\mathbf{X}_i^{\mathrm{T}},\ (U_i - u)\mathbf{X}_i^{\mathrm{T}})^{\mathrm{T}} \right), \quad \mathbf{B} = \mathrm{diag}\left(\mathbf{I}_q,\ \mathbf{I}_\ell \otimes \Big\{ \mathrm{diag}(1,\ h) \otimes \mathbf{I}_p \Big\} \right).$$

We first prove that $\tilde{\boldsymbol{\xi}} \equiv \Big(\tilde{\boldsymbol{\theta}}(u)^{\mathrm{T}}, \tilde{\mathbf{a}}_1(u)^{\mathrm{T}}, \tilde{\mathbf{b}}_1(u)^{\mathrm{T}}, \cdots, \tilde{\mathbf{a}}_\ell(u)^{\mathrm{T}}, \tilde{\mathbf{b}}_\ell(u)^{\mathrm{T}} \Big)^{\mathrm{T}}$, the maximizer of L given in (4.8), is a consistent estimator of $\boldsymbol{\xi}$.

Note that, given the sample, $\tilde{\mathbf{z}}_i$ is a function of $\boldsymbol{\xi}$ and $\tilde{\mathbf{z}}_i$ can be written as $\tilde{\mathbf{z}}_i(\boldsymbol{\xi})$. To prove $\tilde{\boldsymbol{\xi}}$ is consistent, we first prove

$$P\Big(\sum_{i=1}^{n} K_h(U_i - u) \log\left\{ f\left(Y_i;\ \mathbf{X}_i,\ \tilde{\mathbf{z}}_i(\boldsymbol{\xi}')\right) / f\left(Y_i;\ \mathbf{X}_i,\ \tilde{\mathbf{z}}_i(\boldsymbol{\xi})\right) \right\} < 0 \Big) \longrightarrow 1, \quad \text{as } n \longrightarrow \infty, \tag{C.1}$$

for any $\boldsymbol{\xi}' \neq \boldsymbol{\xi}$. By the law of large numbers, to prove (C.1), we need only prove

$$E\Big[K_h(U_1-u)\log\left\{f\left(Y_1;\ \mathbf{X}_1,\ \tilde{\mathbf{z}}_1(\boldsymbol{\xi}')\right)/f\left(Y_1;\ \mathbf{X}_1,\ \tilde{\mathbf{z}}_1(\boldsymbol{\xi})\right)\right\}\Big] < 0. \tag{C.2}$$

It is easy to see that

$$\begin{aligned}
&E\Big[K_h(U_1-u)\log\left\{f\left(Y_1;\ \mathbf{X}_1,\ \tilde{\mathbf{z}}_1(\boldsymbol{\xi}')\right)/f\left(Y_1;\ \mathbf{X}_1,\ \tilde{\mathbf{z}}_1(\boldsymbol{\xi})\right)\right\}\Big] \\
&= \pi(u)E\Big[\log\left\{f\left(Y_1;\ \mathbf{X}_1,\ \tilde{\mathbf{z}}_1(\boldsymbol{\xi}')\right)/f\left(Y_1;\ \mathbf{X}_1,\ \tilde{\mathbf{z}}_1(\boldsymbol{\xi})\right)\right\}|U_1=u\Big] + o(1).
\end{aligned}$$

Because the log function is strictly concave, Jensen's inequality shows that

$$\begin{aligned}
&E\Big[\log\left\{f\left(Y_1;\ \mathbf{X}_1,\ \tilde{\mathbf{z}}_1(\boldsymbol{\xi}')\right)/f\left(Y_1;\ \mathbf{X}_1,\ \tilde{\mathbf{z}}_1(\boldsymbol{\xi})\right)\right\}|U_1=u\Big] \\
&< \log\Big(E\Big[f\left(Y_1;\ \mathbf{X}_1,\ \tilde{\mathbf{z}}_1(\boldsymbol{\xi}')\right)/f\left(Y_1;\ \mathbf{X}_1,\ \tilde{\mathbf{z}}_1(\boldsymbol{\xi})\right)|U_1=u\Big]\Big) = 0.
\end{aligned}$$

Thus (C.2) holds, and this implies that (C.1) holds.

Let $\tilde{\xi}_j$ and ξ_j be the jth components of $\tilde{\boldsymbol{\xi}}$ and $\boldsymbol{\xi}$, for any $\varepsilon > 0$,

$$P(\|\tilde{\boldsymbol{\xi}}-\boldsymbol{\xi}\| > \varepsilon) \leq \sum_{j=1}^{q+2p\ell} P(|\tilde{\xi}_j - \xi_j| > \varepsilon_1),$$

where $\varepsilon_1 = \varepsilon(q+2p\ell)^{-1/2}$.

Note that conditions (6) and (7) imply that $\tilde{\boldsymbol{\xi}}$ is the unique root of the function

$$L(\boldsymbol{\xi}'') = \sum_{i=1}^{n} K_h(U_i-u)\log\left\{f\left(Y_i;\ \mathbf{X}_i,\ \tilde{\mathbf{z}}_i(\boldsymbol{\xi}'')\right)\right\}.$$

Thus for any fixed j, $j = 1, \cdots, q+2p\ell$, letting $\boldsymbol{\xi}_{-\varepsilon}$ be $\boldsymbol{\xi}$ with the jth component replaced by $\xi_j - \varepsilon_1$ and letting $\boldsymbol{\xi}_{\varepsilon}$ be $\boldsymbol{\xi}$ with the jth component replaced by $\xi_j + \varepsilon_1$,

we have

$$\begin{aligned} P(|\tilde{\xi}_j - \xi_j| \le \varepsilon_1) &\ge P\Big(L(\boldsymbol{\xi}) > L(\boldsymbol{\xi}_{-\varepsilon}),\ L(\boldsymbol{\xi}) > L(\boldsymbol{\xi}_{\varepsilon})\Big) \\ &= 1 - P\Big(L(\boldsymbol{\xi}) \le L(\boldsymbol{\xi}_{-\varepsilon}) \text{ or } L(\boldsymbol{\xi}) \le L(\boldsymbol{\xi}_{\varepsilon})\Big) \\ &\ge 1 - P\Big(L(\boldsymbol{\xi}) \le L(\boldsymbol{\xi}_{-\varepsilon})\Big) - P\Big(L(\boldsymbol{\xi}) \le L(\boldsymbol{\xi}_{\varepsilon})\Big). \end{aligned}$$

By (C.1), we have

$$P\Big(L(\boldsymbol{\xi}) \le L(\boldsymbol{\xi}_{-\varepsilon})\Big) \longrightarrow 0, \quad P\Big(L(\boldsymbol{\xi}) \le L(\boldsymbol{\xi}_{\varepsilon})\Big) \longrightarrow 0.$$

Thus,

$$P(|\tilde{\xi}_j - \xi_j| \le \varepsilon_1) \longrightarrow 1,$$

which leads to $P(\|\tilde{\boldsymbol{\xi}} - \boldsymbol{\xi}\| > \varepsilon) \longrightarrow 0$; that is, $\tilde{\boldsymbol{\xi}}$ is consistent.

By Taylor's expansion and conditions (2) and (3), we have

$$\begin{aligned} \tilde{\mathbf{z}}_i - \boldsymbol{\gamma}_i &= -\big(\mathbf{0}_{1\times q},\ \mathbf{X}_i^{\mathrm{T}}\big\{\mathbf{a}_1(U_i) - \mathbf{a}_1 - \mathbf{b}_1(U_i - u)\big\},\ \cdots,\ \mathbf{X}_i^{\mathrm{T}}\big\{\mathbf{a}_\ell(U_i) - \mathbf{a}_\ell - \mathbf{b}_\ell(U_i - u)\big\}\big)^{\mathrm{T}} \\ &= -2^{-1}\Big(\mathbf{0}_{1\times q},\ \mathbf{X}_i^{\mathrm{T}}\ddot{\mathbf{a}}_1(u),\ \cdots,\ \mathbf{X}_i^{\mathrm{T}}\ddot{\mathbf{a}}_\ell(u)\Big)^{\mathrm{T}}(U_i - u)^2 + o_P(h^2) \end{aligned}$$

uniformly in i. Together with condition (6), the foregoing equality leads to

$$\begin{aligned} \mathbf{B}^{-1}\frac{\partial L}{\partial \boldsymbol{\xi}} &= \sum_{i=1}^{n} K_h(U_i - u)\mathbf{B}^{-1}\frac{\partial \tilde{\mathbf{z}}_i}{\partial \boldsymbol{\xi}}\frac{\partial \log f(Y_i;\ \mathbf{X}_i,\ \tilde{\mathbf{z}}_i)}{\partial \tilde{\mathbf{z}}_i} = \sum_{i=1}^{n} K_h(U_i - u)\mathbf{B}^{-1}\mathbf{H}_i\mathbf{g}(Y_i;\ \mathbf{X}_i,\ \tilde{\mathbf{z}}_i) \\ &= \sum_{i=1}^{n} K_h(U_i - u)\mathbf{B}^{-1}\mathbf{H}_i\mathbf{g}(Y_i;\ \mathbf{X}_i,\ \boldsymbol{\gamma}_i) \\ &\quad -2^{-1}\sum_{i=1}^{n} K_h(U_i - u)\mathbf{B}^{-1}\mathbf{H}_i\dot{\mathbf{g}}(Y_i;\ \mathbf{X}_i,\ \boldsymbol{\gamma}_i)\Big(\mathbf{0}_{1\times q},\ \mathbf{X}_i^{\mathrm{T}}\ddot{\mathbf{a}}_1(u),\ \cdots,\ \mathbf{X}_i^{\mathrm{T}}\ddot{\mathbf{a}}_\ell(u)\Big)^{\mathrm{T}} \\ &\qquad \times(U_i - u)^2\big\{1 + o_P(1)\big\} \\ &\triangleq \mathbf{A}_1 + \mathbf{A}_2\big\{1 + o_P(1)\big\}. \end{aligned}$$

Let

$$\boldsymbol{\Omega}(u) = E\left\{ \mathbf{H}\boldsymbol{\mathcal{I}}(\boldsymbol{\gamma})\big(\mathbf{0}_{1\times q},\ \mathbf{X}^{\mathrm{T}}\ddot{\mathbf{a}}_1(u),\ \cdots,\ \mathbf{X}^{\mathrm{T}}\ddot{\mathbf{a}}_\ell(u)\big)^{\mathrm{T}} \middle| U = u \right\}.$$

By Lemma C.1 and conditions (1) and (3)–(6), we have

$$\frac{1}{n}\mathbf{A}_2 = 2^{-1}\pi(u)\boldsymbol{\Omega}(u)\mu_2 h^2\big\{1 + o_P(1)\big\}.$$

It is easy to see that

$$n^{-1/2}h^{1/2}\mathbf{A}_1 \xrightarrow{D} N\left(\mathbf{0}_{(2p\ell+q)\times 1},\ \mathbf{V}(u)\pi(u)\right),$$

where $\mathbf{V}(u) = E\left\{ \nu_0 \mathbf{H}\boldsymbol{\mathcal{I}}(\boldsymbol{\gamma})\mathbf{H}^{\mathrm{T}} + \nu_2 \mathbf{H}_c\boldsymbol{\mathcal{I}}(\boldsymbol{\gamma})\mathbf{H}_c^{\mathrm{T}} \middle| U = u \right\}$. By Lemma C.1 and conditions (1)–(6),

$$\begin{aligned}
\mathbf{B}^{-1}\frac{\partial^2 L}{\partial\boldsymbol{\xi}\partial\boldsymbol{\xi}^{\mathrm{T}}}\mathbf{B}^{-1} &= \sum_{i=1}^{n} K_h(U_i - u)\mathbf{B}^{-1}\frac{\partial\tilde{\mathbf{z}}_i}{\partial\boldsymbol{\xi}}\dot{\mathbf{g}}(Y_i;\ \mathbf{X}_i,\ \boldsymbol{\gamma}_i)\left(\frac{\partial\tilde{\mathbf{z}}_i}{\partial\boldsymbol{\xi}}\right)^{\mathrm{T}}\mathbf{B}^{-1} \\
&= \sum_{i=1}^{n} K_h(U_i - u)\mathbf{B}^{-1}\mathbf{H}_i\dot{\mathbf{g}}(Y_i;\ \mathbf{X}_i,\ \boldsymbol{\gamma}_i)\mathbf{H}_i^{\mathrm{T}}\mathbf{B}^{-1} \\
&= n\mathbf{V}_c(u)\,\pi(u)\big\{1 + o_P(1)\big\}.
\end{aligned}$$

Thus,

$$(nh)^{1/2}\mathbf{B}\left(\frac{\partial^2 L}{\partial\boldsymbol{\xi}\partial\boldsymbol{\xi}^{\mathrm{T}}}\right)^{-1}\mathbf{B}\mathbf{A}_1 \xrightarrow{D} N\left(\mathbf{0}_{(2p\ell+q)\times 1},\ \mathbf{V}_c(u)^{-1}\mathbf{V}(u)\mathbf{V}_c(u)^{-1}\,\pi(u)^{-1}\right),$$

$$\mathbf{B}\left(\frac{\partial^2 L}{\partial\boldsymbol{\xi}\partial\boldsymbol{\xi}^{\mathrm{T}}}\right)^{-1}\mathbf{B}\mathbf{A}_2 = 2^{-1}\mu_2 h^2\mathbf{V}_c(u)^{-1}\boldsymbol{\Omega}(u)\left\{1 + o_P(1)\right\}.$$

By Taylor's expansion and the consistency of $\tilde{\boldsymbol{\xi}}$, we have

$$\tilde{\boldsymbol{\xi}} - \boldsymbol{\xi} = -\left(\frac{\partial^2 L}{\partial\boldsymbol{\xi}\partial\boldsymbol{\xi}^{\mathrm{T}}}\right)^{-1}\frac{\partial L}{\partial\boldsymbol{\xi}}\left\{1 + o_P(1)\right\}.$$

This leads to

$$\tilde{\boldsymbol{\theta}}(u) - \boldsymbol{\theta} = -(\mathbf{I}_q, \mathbf{0}_{q\times(2p\ell)})\left\{\mathbf{B}\left(\frac{\partial^2 L}{\partial\boldsymbol{\xi}\partial\boldsymbol{\xi}^{\mathrm{T}}}\right)^{-1}\mathbf{B}\mathbf{A}_1 + \mathbf{B}\left(\frac{\partial^2 L}{\partial\boldsymbol{\xi}\partial\boldsymbol{\xi}^{\mathrm{T}}}\right)^{-1}\mathbf{B}\mathbf{A}_2\right\}\big\{1 + o_P(1)\big\}.$$

Let L_j, $\mathbf{A}_{2,j}$ and $\boldsymbol{\xi}_j$ be L, $\mathbf{A}_2$ and $\boldsymbol{\xi}$ but with u replaced by U_j. By Lemma C. 1 and conditions (1)–(6),

$$\frac{1}{n}\sum_{j=1}^{n}\mathbf{B}\left(\frac{\partial^2 L_j}{\partial\boldsymbol{\xi}_j\partial\boldsymbol{\xi}_j^{\mathrm{T}}}\right)^{-1}\mathbf{B}\mathbf{A}_{2,j} = 2^{-1}\mu_2 h^2 E\big\{\mathbf{V}_c(U)^{-1}\boldsymbol{\Omega}(U)\big\}\big\{1 + o_P(1)\big\}.$$

Let $\mathbf{A}_{1,j}$, $\mathbf{V}_{cj}$, and $\mathbf{H}_{i,j}$ be $\mathbf{A}_1$, $\mathbf{V}_c$, and $\mathbf{H}_i$ but with u replaced by U_j. By Lemma C.1, we have

$$\begin{aligned}
n(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) &= (\mathbf{I}_q, \mathbf{0}_{q\times(2p\ell)})\sum_{j=1}^{n}\mathbf{B}\left(\frac{\partial^2 L_j}{\partial\boldsymbol{\xi}_j\partial\boldsymbol{\xi}_j^{\mathrm{T}}}\right)^{-1}\mathbf{B}A_{1,j} + O_P(nh^2) \\
&= (\mathbf{I}_q, \mathbf{0}_{q\times(2p\ell)})\sum_{j=1}^{n}\left\{\sum_{k=1}^{n}K_h(U_k - U_j)\mathbf{B}^{-1}\mathbf{H}_{k,j}\dot{\mathbf{g}}(Y_k;\ \mathbf{X}_k,\ \boldsymbol{\gamma}_k)\mathbf{H}_{k,j}^{\mathrm{T}}\mathbf{B}^{-1}\right\}^{-1} \\
&\qquad\qquad \times\sum_{i=1}^{n}K_h(U_i - U_j)\mathbf{B}^{-1}\mathbf{H}_{i,j}\mathbf{g}(Y_i;\ \mathbf{X}_i,\ \boldsymbol{\gamma}_i) + O_P(nh^2) \\
&= n^{-1}(\mathbf{I}_q, \mathbf{0}_{q\times(2p\ell)})\sum_{i=1}^{n}\sum_{j=1}^{n}K_h(U_i - U_j)\mathbf{V}_{cj}^{-1}\pi(U_j)^{-1}\mathbf{B}^{-1}\mathbf{H}_{i,j}\mathbf{g}(Y_i;\ \mathbf{X}_i,\ \boldsymbol{\gamma}_i) \\
&\quad \times\big\{1 + o_P(1)\big\} + O_P(nh^2).
\end{aligned}$$

By tedious calculation, we have

$$\begin{aligned}
&n^{-3/2}\sum_{i=1}^{n}\sum_{j=1}^{n}K_h(U_i - U_j)\mathbf{V}_{cj}^{-1}\pi(U_j)^{-1}\mathbf{B}^{-1}\mathbf{H}_{i,j}\mathbf{g}(Y_i;\ \mathbf{X}_i,\ \boldsymbol{\gamma}_i) \\
&\quad\xrightarrow{D} N\left(\mathbf{0}_{(2p\ell+q)\times 1},\ E\big\{\mathbf{V}_c(U)^{-1}\mathbf{V}_0(U)\mathbf{V}_c(U)^{-1}\big\}\right),
\end{aligned}$$

which implies that if $h = o(n^{-1/4})$, then

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{D} N(\mathbf{0}_{q\times 1},\ \boldsymbol{\Delta}).$$

Proof of Theorem 4. Let

$$L_1 = \sum_{i=1}^{n} K_{h_1}(U_i - u) \log f(Y_i;\ \mathbf{X}_i,\ \boldsymbol{\theta},\ \mathbf{c}_i),$$

$$\boldsymbol{\eta} = (\mathbf{a}_1^{\mathrm{T}},\ \mathbf{b}_1^{\mathrm{T}},\ \cdots,\ \mathbf{a}_\ell^{\mathrm{T}},\ \mathbf{b}_\ell^{\mathrm{T}})^{\mathrm{T}}, \quad \mathbf{B}_1 = \mathbf{I}_\ell \otimes \{\mathrm{diag}(1,\ h_1) \otimes \mathbf{I}_p\},$$

$$\mathbf{m}(Y_i;\ \mathbf{X}_i,\ \boldsymbol{\theta},\ \mathbf{z}_i) = \frac{\partial \log f(Y_i;\ \mathbf{X}_i,\ \boldsymbol{\theta},\ \mathbf{z}_i)}{\partial \mathbf{z}_i}, \quad \mathbf{z}_i = (\mathbf{X}_i^{\mathrm{T}}\mathbf{a}_1(U_i),\ \cdots,\ \mathbf{X}_i^{\mathrm{T}}\mathbf{a}_\ell(U_i))^{\mathrm{T}},$$

$$\dot{\mathbf{m}}(Y_i;\ \mathbf{X}_i,\ \boldsymbol{\theta},\ \mathbf{z}_i) = \frac{\partial \mathbf{m}(Y_i;\ \mathbf{X}_i,\ \boldsymbol{\theta},\ \mathbf{z}_i)}{\partial \mathbf{z}_i}, \quad \mathbf{D}_i = \mathbf{I}_\ell \otimes (\mathbf{X}_i^{\mathrm{T}},\ (U_i - u)\mathbf{X}_i^{\mathrm{T}})^{\mathrm{T}}.$$

Let $\tilde{\boldsymbol{\eta}}$ be the maximizer of L_1 with respect to $\boldsymbol{\eta}$. Using the same argument as in the proof of Theorem 1, we can show that $\tilde{\boldsymbol{\eta}}$ is a consistent estimator of $\boldsymbol{\eta}$.

By simple calculation and conditions (2) and (6), we have

$$\begin{aligned}
\mathbf{B}_1^{-1}\frac{\partial L_1}{\partial \boldsymbol{\eta}} &= \sum_{i=1}^{n} K_{h_1}(U_i - u)\mathbf{B}_1^{-1}\frac{\partial \mathbf{c}_i}{\partial \boldsymbol{\eta}}\frac{\partial \log f(Y_i;\ \mathbf{X}_i,\ \boldsymbol{\theta},\ \mathbf{c}_i)}{\partial \mathbf{c}_i} \\
&= \sum_{i=1}^{n} K_{h_1}(U_i - u)\mathbf{B}_1^{-1}\mathbf{D}_i\frac{\partial \log f(Y_i;\ \mathbf{X}_i,\ \boldsymbol{\theta},\ \mathbf{c}_i)}{\partial \mathbf{c}_i} \\
&= \sum_{i=1}^{n} K_{h_1}(U_i - u)\mathbf{B}_1^{-1}\mathbf{D}_i\mathbf{m}(Y_i;\ \mathbf{X}_i,\ \boldsymbol{\theta},\ \mathbf{z}_i) \\
&\quad -2^{-1}\sum_{i=1}^{n} K_{h_1}(U_i - u)\mathbf{B}_1^{-1}\mathbf{D}_i\dot{\mathbf{m}}(Y_i;\ \mathbf{X}_i,\ \boldsymbol{\theta},\ \mathbf{z}_i)\Big(\ddot{\mathbf{a}}_1(u),\ \cdots,\ \ddot{\mathbf{a}}_\ell(u)\Big)^{\mathrm{T}}\mathbf{X}_i(U_i - u)^2 \\
&\quad \times\big\{1 + o_P(1)\big\} \\
&\stackrel{\triangle}{=} \mathbf{J}_1 + \mathbf{J}_2\big\{1 + o_P(1)\big\}.
\end{aligned}$$

By Lemma C.1 and conditions (1) and (3)–(6), it is easy to see that

$$\frac{1}{n}\mathbf{J}_2 = 2^{-1}\boldsymbol{\Gamma}\mu_2 h_1^2 \pi(u)\big\{1 + o_P(1)\big\}.$$

By the central limit theorem,

$$n^{-1/2}h_1^{1/2}\mathbf{J}_1 \xrightarrow{D} N\Big(\mathbf{0}_{(2p\ell)\times 1},\ \mathbf{G}\pi(u)\Big).$$

By Lemma C.1 and conditions (1)–(6),

$$\begin{aligned}
\mathbf{B}_1^{-1}\frac{\partial^2 L_1}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^{\mathrm{T}}}\mathbf{B}_1^{-1} &= \sum_{i=1}^{n} K_{h_1}(U_i-u)\mathbf{B}_1^{-1}\frac{\partial \mathbf{c}_i}{\partial\boldsymbol{\eta}}\dot{\mathbf{m}}(Y_i;\ \mathbf{X}_i,\ \boldsymbol{\theta},\ \mathbf{c}_i)\left(\frac{\partial \mathbf{c}_i}{\partial\boldsymbol{\eta}}\right)^{\mathrm{T}}\mathbf{B}_1^{-1} \\
&= \sum_{i=1}^{n} K_{h_1}(U_i-u)\mathbf{B}_1^{-1}\mathbf{D}_i\dot{\mathbf{m}}(Y_i;\ \mathbf{X}_i,\ \boldsymbol{\theta},\ \mathbf{z}_i)\mathbf{D}_i^{\mathrm{T}}\mathbf{B}_1^{-1} \\
&= n\mathbf{G}_c\pi(u)\big\{1+o_P(1)\big\}.
\end{aligned}$$

Thus,

$$(nh_1)^{1/2}\mathbf{B}_1\left(\frac{\partial^2 L_1}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^{\mathrm{T}}}\right)^{-1}\mathbf{B}_1\mathbf{J}_1 \xrightarrow{D} N\left(\mathbf{0}_{(2p\ell)\times 1},\ \mathbf{G}_c^{-1}\mathbf{G}\mathbf{G}_c^{-1}\pi(u)^{-1}\right),$$

$$\mathbf{B}_1\left(\frac{\partial^2 L_1}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^{\mathrm{T}}}\right)^{-1}\mathbf{B}_1\mathbf{J}_2 = 2^{-1}\mu_2 h_1^2\mathbf{G}_c^{-1}\boldsymbol{\Gamma}\left\{1+o_P(1)\right\}.$$

By Taylor's expansion and the consistency of $\tilde{\boldsymbol{\eta}}$, we have

$$\tilde{\boldsymbol{\eta}}-\boldsymbol{\eta} = -\left(\frac{\partial^2 L_1}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^{\mathrm{T}}}\right)^{-1}\frac{\partial L_1}{\partial\boldsymbol{\eta}}\left\{1+o_P(1)\right\}.$$

Thus,

$$(nh_1)^{1/2}\left\{\mathbf{B}_1\left(\tilde{\boldsymbol{\eta}}-\boldsymbol{\eta}\right)+2^{-1}\mu_2 h_1^2\mathbf{G}_c^{-1}\boldsymbol{\Gamma}\right\} \xrightarrow{D} N\left(\mathbf{0}_{(2p\ell)\times 1},\ \mathbf{G}_c^{-1}\mathbf{G}\mathbf{G}_c^{-1}\pi(u)^{-1}\right).$$

Because $\hat{\boldsymbol{\theta}}$ has the $n^{-1/2}$ convergence rate, the maximizer of L_1 with respect to $\boldsymbol{\eta}$ would behave exactly the same as the maximizer of (4.10) asymptotically. Thus,

$$(nh_1)^{1/2}(\hat{\mathbf{a}}-\mathbf{a}+\boldsymbol{\mathcal{B}}) \xrightarrow{D} N\Big(\mathbf{0}_{(p\ell)\times 1},\ \boldsymbol{\Sigma}\Big).$$