Statistical Estimation in Generalized Multiparameter Likelihood Models

Ming-Yen Cheng, Wenyang Zhang and Lu-Hung Chen

Abstract

Multiparameter likelihood models (MLM) 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 coefficient models, as special cases. When the model is fixed, a simple and effective two-step method is developed to estimate both the parametric and the nonparametric components. The proposed estimator of the parametric component has the $n^{-1/2}$ convergence rate, and the estimator of the nonparametric component enjoys an adaptivity property. A data-driven procedure is suggested to select the bandwidths involved. Also proposed is a new initial estimator in 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. A simulation study and an application to the infant mortality data of China are given to demonstrate performance of the proposed methods.

KEY WORDS: bandwidth selection, model selection, multiple covariate, partially linear models, profile likelihood, varying coefficient models, semiparametric models.

SHORT TITLE: Generalized multiparameter likelihood models.

Ming-Yen Cheng is Professor, Department of Mathematics, National Taiwan University, Taipei 106, Taiwan (Email: cheng@math.ntu.edu.tw) and Professor of Statistics, Department of Statistical Science, University College London, Gower Street, London, WC1E 6BT, UK (Email: ming-yen.cheng@ucl.ac.uk). Wenyang Zhang is Reader in Statistics, Department of Mathematical Sciences, University of Bath, Bath BA2 7AY, UK (Email: W.Zhang@bath.ac.uk). Lu-Hung Chen is Ph.D. Candidate, Department of Mathematics, National Taiwan University, Taipei 106, Taiwan (Email: d95221006@ntu.edu.tw). Cheng’s research was supported in part by the National Science Council (grant NSC95-2118-M-002-003-MY2). This project began while Wenyang Zhang visited National Taiwan University, with funding supported by the Mathematics Division of the National Center of Theoretical Sciences (Taipei). We thank the associate editor and two referees for constructive comments which lead to substantial improvement of this paper.
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. However, parametric approaches are in risk of model misspecification, which could result in seriously biased estimation, misinterpretation of data, and other difficulties. Nonparametric modeling is more flexible and allows data to present the unknown truth. However, people always come up against the “curse of dimensionality” problem, i.e. model instability when the dimension of the covariates is large. There are many hybrids of parametric and nonparametric models, generally called semiparametric models, proposed to achieve a good balance between flexibility and stability in model specification. Examples are varying coefficients models (Hastie and Tibshirani, 1993; Xia and Li, 1999; Fan and Huang, 2005; Cai, Fan, Jiang and Zhou, 2008; Li and Liang 2008), partially linear models (Engle, Granger, Rice and Weiss, 1986; Härdle, Liang and Gao, 2000; Cai, Fan, Jiang and Zhou, 2007) and additive models (Hastie and Tibshirani, 1990; Fan, Härdle and Mammen, 1998; Fan and Jiang, 2005; Jiang and Zhou, 2007).

In this paper 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

\[
f(Y; \mathbf{X}, \boldsymbol{\theta}, \mathbf{x}_1^T a_1(U), \ldots, \mathbf{x}_\ell^T a_\ell(U)),
\]

where \(f\) is a known parametric density function, \(\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)^T\) is an unknown constant vector, \(\mathbf{X} = (X_1, \ldots, X_p)^T\) with \(X_1 \equiv 1\), and \(\mathbf{x}_j\) is a \(p_j\)-dimensional subvector of \(\mathbf{X}\) and \(a_j(\cdot) = (a_{j1}(\cdot), \ldots, a_{jp_j}(\cdot))^T\) is an unknown function, \(j = 1, \ldots, \ell\). Here, \(1 \leq \ell \leq d\), where \(d\) is defined in (1.2). Model (1.1) is a hybrid of the standard MLM which assumes the conditional density function of \(Y\) given \(\mathbf{X}\) follows the form

\[
f(Y; a_1(\mathbf{X}), \ldots, a_d(\mathbf{X})),
\]

where \(f\) has \(d\) identifiable parameters and \(a_1(\mathbf{X}), \ldots, a_d(\mathbf{X})\) are unknown functions, i.e. \(Y\) depends on \(\mathbf{X}\) through the \(d\) identifiable parameters in \(f\) being modeled as
nonparametric functions of $X$. Aerts and Claeskens (1997) studied a locally linear maximum likelihood estimator of MLM when $X$ is univariate, and Cheng and Peng (2007) proposed a variance reduction technique to improve the estimation. 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. See Aerts and Claeskens (1997), Loader (1999), Claeskens and Aerts (2000) and Cheng and Peng (2007), among others, for further details.

With the availability of $U$, model (1.1) specifies $\ell$ of the $d$ parameter functions in model (1.2) by some nonparametric or semiparametric form, and if $d - \ell > 0$ the other $d - \ell$ parameter functions in (1.2) are now modeled parametrically in (1.1), with $\theta$ consisting of all the constant parameters. Like MLM (1.2), (1.1) offers a unified approach to modeling a wide range of data settings and to dealing with various inference problems. Nonetheless, (1.1) avoids the curse of dimensionality problem (1.2) has when the dimension of $X$ is large and it allows parametric or nonparametric or semiparametric modeling of the parameter functions in (1.2). Further, (1.1) broadens the application of MLM since it can cope with categorical covariates, which often arise in practice. Model (1.1) is a very general semiparametric model provided there exists a continuous variable $U$ and other covariates $X$. First, it reduces to a partially linear model when $\ell = 1$, $x_1 = X_1 \equiv 1$ and $\theta$ interacts with $X$ through a linear function. When $\ell = 1$, $q = 0$ and $x_1 = X$, model (1.1) becomes the varying coefficients model of Hastie and Tibshirani (1993) with the same modifying variable $U$. Thus, (1.1) inherits the model stability, flexibility and interpretability that varying coefficients models enjoy. In addition, it is closely related the Regression Model II in Section 4.3 of Bickel, Klaassen, Ritov and Wellner (1993).

We propose a simple, effective and fast two-step procedure to estimate both the constant and functional parameters in (1.1). Its implementation does not involve any iteration which is usually required by conventional approaches, e.g. profile likelihood
and backfitting. Further, we develop an AIC data-driven procedure to select the bandwidths required in the two-step estimation. Using an AIC criterion, or modified versions, to select smoothing parameters in nonparametric regression and local likelihood modeling has been extensively discussed and implemented, c.f. Hurvich, Simonoff and Tsai (1998), Loader (1999), Schucany (2004), among others. For local likelihood estimation, Aerts and Claeskens (1997) considered cross-validation and plug-in bandwidths, and Farmen and Gijbels (1998) suggested a bandwidth selector based on an approximation to the integrated mean squared error. A profile likelihood approach can be applied to estimate the constant parameters. A new initial estimator is proposed to ensure stability of the profile likelihood approach no matter what types of features, e.g. location, scale or shape, in the model the constant parameters are playing 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 Section 5 and simulation results reported in Section 6. Nevertheless, the major strength of the two-step estimation is its simple and fast implementation and numerical stability, as no iteration is 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), i.e. we are not sure about the specification of $\theta$ and $x_1, \ldots, x_\ell$. In an attempt to solve this fundamental identification problem, we suggest a stepwise procedure based on a version of the BIC criterion accounted for our model. Identification of constant parameters and bandwidth selection interact with each other. We propose to select the bandwidths first and then keep 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 exists 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 AIC and BIC and posterior
probability model selection criteria for one-parameter local likelihood models. Fan and Li (2002) employed 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 the two sets of covariates in the parametric and nonparametric components are separated in advance.

Section 2 provides some motivating examples for model (1.1) and discusses the identifiability issue. 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 are given in Section 3. Bandwidth selection and identification of the constant parameters are addressed in Section 4. Asymptotic properties of the two-step estimators are investigated in Section 5. Section 6 presents a simulated example and an analysis of a motivating example about infant mortality. Proofs of the theoretical results are deferred to the Appendix.

2 Motivating examples and model identifiability

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

$$f(Y; X, \theta, a_1(X), \ldots, a_\ell(X)),$$

(2.1)

where $1 \leq \ell < d$ and $\theta$ is a $q$-dimensional unknown parameter. Here, $\ell$ out of the $d$ parameter functions in (1.2) remain as unknown functions of $X$, and the other $d - \ell$ parameter functions are formulated by certain parametric forms, e.g. unknown constants, with $\theta$ consisting of all the constant parameters. The model studied by Severini and Wong (1992) is a special case of (2.1) with $d = 2$, $\ell = 1$, $q = 1$, and the covariate being univariate. Profile likelihood estimation of $\theta$, along with consistent estimators of a least favorable curve, was studied therein.

When the dimension of $X$ is large, neither (1.2) nor (2.1) would work due to the curse of dimensionality problem. Claeskens and Aerts (2000) suggested to alleviate this problem by restricting $a_1(\cdot), \ldots, a_d(\cdot)$ in (1.2) to additive models and estimate them by a backfitting algorithm. Alternatively, the following restriction of (2.1)

$$f(Y; X, \theta, X^T \beta_1, \ldots, X^T \beta_\ell),$$

(2.2)

where $\beta_1, \ldots, \beta_\ell$ are unknown constant vectors, would cope with the curse of dimensionality problem. However, (2.2) actually implies that the impact of $X$ on $Y$ is constant, 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 $U$ as China has changed very much since 1950 and the difference between rural and urban regions has been changing. The impact must vary with $U$ and the pattern of the change is of interest. To fit the data more accurately and to catch the dynamic pattern of the changes in the impact, we extend (2.2) to

$$f(Y; X, \theta, X^T a_1(U), \ldots, X^T a_\ell(U)),$$

(2.3)

where $\theta$ is an unknown constant vector, and $a_j(\cdot) = (a_{j1}(\cdot), \ldots, a_{jp}(\cdot))^T$, $j = 1, \ldots, \ell$. In (2.3), $a_1(\cdot), \ldots, a_\ell(\cdot)$ have to share the same dimension $p$ and all 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 restriction and to 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 \( \theta \) in (1.1).

When \( a_1(\cdot), \ldots, a_\ell(\cdot) \) are all constant, model (2.3) reduces to model (2.2) and \( \mathcal{I}(\gamma) \) defined in Theorem 1 becomes \( \tilde{\mathcal{I}} \), where \( \tilde{\mathcal{I}}(\gamma) \) with \( a_j(U) \) replaced by \( \beta_j \).

Condition (7) in the Appendix ensures the smallest eigenvalue of \( \tilde{\mathcal{I}} \) is greater than the positive number \( \lambda_0 \) in Condition (7). If \((X_i, Y_i), i = 1, \ldots, n,\) is a sample from model (2.2) then, under Condition (7), the Fisher information matrix is

\[
\sum_{i=1}^{n} \text{diag}(I_q, I_\ell \otimes X_i) \tilde{\mathcal{I}}_i \text{diag}(I_q, I_\ell \otimes X_i^T) > \lambda_0 \sum_{i=1}^{n} \text{diag}(I_q, I_\ell \otimes X_i) \text{diag}(I_q, I_\ell \otimes X_i^T)
\]

\[
\approx n\lambda_0 \text{diag}(I_q, I_\ell \otimes \mathbb{E}(XX^T)) > 0,
\]

where \( \tilde{\mathcal{I}}_i \) is \( \tilde{\mathcal{I}} \) with \( X \) replaced by \( X_i \). Here, \( I_k \) denotes a size \( k \) identity matrix and, for any matrices \( A \) and \( B \), \( \text{diag}(A, B) \) denotes the matrix

\[
\begin{pmatrix}
A & 0 \\
0 & B
\end{pmatrix}.
\]

Hence, Condition (7) ensures the Fisher information matrix of the parametric model (2.2) is positive-definite, i.e. model (2.2) is identifiable. Further, for any given value of \( U \), the local version of model (2.3) is model (2.2). Therefore, under Condition (7), model (2.3) is identifiable for any given value of \( U \), and so model (2.3) is identifiable.

Also, model (1.1) specifies some of the \( a_{ij}(\cdot) \)'s in model (2.3) as constant and therefore is identifiable. Based on the above arguments, we have the following lemma.

**Lemma 1.** Under Condition (7) in the Appendix, both models (1.1) and (2.3) are identifiable.

### 3 Estimation procedures

Suppose we have a sample \((X_i, U_i, Y_i), i = 1, \ldots, n,\) from \((X, U, Y)\) which obeys model (1.1). Let \( x_{i,j} \) be the \( p_j \)-dimensional subvector of \( X_i \) that corresponds to \( x_j, j = 1, \ldots, \ell, i = 1, \ldots, n. \) We introduce our two-step and profile likelihood procedures for estimating both the constant and functional parameters in Sections 3.1 and 3.2.
3.1 Two-step estimation

Our two-step approach produces an estimator for the constant vector $\theta$ first, and then this estimator is plugged-in into the local likelihood function to estimate the functions $a_j(\cdot)$, $j = 1, \ldots, \ell$.

The estimation procedure for $\theta$ consists of two stages. First, we treat $\theta$ as an unknown function of $U$ and appeal to the local likelihood approach to get a preliminary estimator $\hat{\theta}(U_i)$ for $\theta(U_i)$ for each $U_i$, $i = 1, \ldots, n$. Then, we average $\hat{\theta}(U_i)$ over $i = 1, \ldots, n$ to get the final estimator for $\theta$. The procedure is detailed as follows. The conditional log-likelihood function is

$$L_0(\theta, a) = \sum_{i=1}^{n} \log f(Y_i; X_i, \theta, x_{i1}^T a_1(U_i), \ldots, x_{i\ell}^T a_\ell(U_i)),$$

(3.1)

where $a(\cdot) = (a_1(\cdot)^T, \ldots, a_\ell(\cdot)^T)^T$. Given $u$, by Taylor’s expansion, we have for each $j$

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

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

$$\sum_{i=1}^{n} K_h(U_i - u) \log f(Y_i; X_i, \theta, x_{i1}^T \{a_1 + b_1(U_i - u)\}, \ldots, x_{i\ell}^T \{a_\ell + b_\ell(U_i - u)\}),$$

(3.2)

where $K_h(\cdot) = K(\cdot/h)/h$, $K(\cdot)$ is a kernel function, and $h > 0$ is a bandwidth. Maximizing (3.2) with respect to $(\theta^T, a_1^T, b_1^T, \ldots, a_\ell^T, b_\ell^T)^T$ we get the maximizer $(\hat{\theta}(u)^T, \hat{a}_1(u)^T, \hat{b}_1(u)^T, \ldots, \hat{a}_\ell(u)^T, \hat{b}_\ell(u)^T)^T$. In the above local likelihood estimation, $\theta$ is fitted by a local constant vector since $\theta$ is constant under model (1.1) and fitting it by a local constant vector stabilizes the procedure. For $i = 1, \ldots, n$, let $u = U_i$ and we get an initial estimator $\hat{\theta}(U_i)$ of $\theta$. The final estimator of $\theta$ is taken to be

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

(3.3)

For $\hat{\theta}$ to achieve the $n^{-1/2}$ convergence rate, we have to choose a relatively small bandwidth $h$ so that the biases of $\hat{\theta}(\cdot)$ and $\hat{a}_j(\cdot)$, $j = 1, \ldots, \ell$, 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 $\theta$. Then, averaging over $\hat{\theta}(U_i), i = 1, \ldots, n$,
as in (3.3) brings the variance from the order \((nh)^{-1}\) in nonparametric estimation back to the order \(n^{-1}\) in parametric estimation. We will show later that \(\hat{\theta}\) is root-

\(n\) consistent when \(h\) is properly chosen. Like any other maximum local likelihood estimation procedure, the bandwidth \(h\) cannot be chosen too small; otherwise one runs into problems with singularity of the design matrix. From the asymptotic point of view, Condition (5) prohibits the bandwidth \(h\) from being too small. Thus, Conditions (5) and (7) guarantee the estimators \(\hat{\theta}(U_1), \ldots, \hat{\theta}(U_n)\) exist. Further, the method of Cheng, Wu and Yang (2008) can be employed to modify the local likelihood function (3.2) to overcome the singularity problem caused by a small \(h\) or sparsity in the design points \(U_i\)'s. This approach can also be applied to (3.4) when estimating the function \(a(u)\).

With \(\hat{\theta}\), we can estimate \(a(u)\) using the maximum local likelihood approach again. Note that the estimator \(\hat{a}(u) = (\hat{a}_1(u)^T, \ldots, \hat{a}_\ell(u)^T)^T\) we obtained before is too noisy and is not appropriate for this purpose since the bandwidth \(h\) is intentionally chosen small in order to get a good estimator of \(\theta\). Hence, we will use another, larger bandwidth to estimate \(a(u)\). We replace \(\theta\) in (3.2) by \(\hat{\theta}\) to get a local log-likelihood function for \(a(u):\)

\[
\sum_{i=1}^{n} K_{h_1}(U_i - u) \log f(Y_i; X_i, \hat{\theta}, x_i^T\{a_1 + b_1(U_i - u)\}, \ldots, x_i^T\{a_\ell + b_\ell(U_i - u)\}) \tag{3.4}
\]

where \(h_1 > 0\) is a bandwidth different from \(h\). We could use a kernel other than \(K\) at this step but that does not matter much. Maximizing (3.4) with respect to \((a_1^T, b_1^T, \ldots, a_\ell^T, b_\ell^T)^T\) we get the maximizer \((\hat{a}_1(u)^T, \hat{b}_1(u)^T, \ldots, \hat{a}_\ell(u)^T, \hat{b}_\ell(u)^T)^T\). Our estimator of \(a(u)\) is taken to be \(\hat{a}(u) = (\hat{a}_1(u)^T, \ldots, \hat{a}_\ell(u)^T)^T\). As the convergence rate of \(\hat{\theta}\) is \(n^{-1/2}\), see Section 5, \(\hat{a}(u)\) would work as well as when \(\theta\) is known in the local log-likelihood (3.4). That is, \(\hat{a}(u)\) has the adaptivity property.

In some cases, local likelihood estimation of the varying coefficients \(a_j(\cdot), j = 1, \ldots, \ell\), may require a different amount of smoothing, see, for example, Claeskens and Aerts (2000). Backfitting ideas can be implemented to achieve this goal: (i) use \(\hat{a}(\cdot)\) as the initial estimate; (ii) for each \(j\), substitute all the local linear coefficient functions except the \(j\)th and \(h_1\) in (3.4) respectively by the previous estimates and the bandwidth for smoothing the \(j\)th functional parameter, and then maximize the resulted local likelihood to find an estimate of \(a_j(\cdot)\); (iii) iterate step (ii) until
convergence. The convergence is usually attained quickly.

3.2 Profile likelihood estimation

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

$$
\sum_{i=1}^{n} \log f(Y_i; X_i, \theta, x_{i1}^T \tilde{a}_1(U_i), \ldots, x_{i\ell}^T \tilde{a}_\ell(U_i)),
$$

where, for any given $\theta$, $\tilde{a}_\theta(\cdot) = (\tilde{a}_1(\cdot)^T, \ldots, \tilde{a}_\ell(\cdot)^T)^T$ is an estimator for $a(\cdot)$. In practice, one needs to find the minimizer of

$$
\left| \frac{\partial L_0}{\partial \theta}(\theta, \tilde{a}_\theta) + \frac{\partial L_0}{\partial a}(\theta, \tilde{a}_\theta) \frac{\partial}{\partial \theta} \tilde{a}_\theta \right|
$$

by iteration, where $L_0$ is given in (3.1). When the specified semiparametric model is general like (1.1), in which $\theta$ may involve shape or scale parameters in $f$, stability of the iteration heavily relies on proper choice of the initial estimate. Under semiparametric models for the regression mean, Fan and Huang (2005) and Lam and Fan (2008) employed difference-based methods to obtain a reliable initial estimate. However, difference-based methods may not work for model (1.1) since some of the elements in $\theta$ can be other than mean parameters. We propose a new initial estimate in the following.

First, we derive some rough estimates of $a_j(U_i), i = 1, \ldots, n, j = 1, \ldots, \ell$. Consider a model obtained by replacing $\theta$ in (1.1) with $a_0(U), a_0(U) = a_0(U) + b_0(U), a_1(U) + b_1(U), \ldots, a_\ell(U) + b_\ell(U)$, of the local log-likelihood function

$$
\sum_{i=1}^{n} K_{h_1}(U_i - u) \log f(Y_i; X_i, a_0 + b_0(U_i - u), \ldots, x_{i\ell}(a_\ell + b_\ell(U_i - u))).
$$

Here, $h_1$ can be taken as the bandwidth $\hat{h}_1$ in Section 4.2 since it is selected for local likelihood estimation by assuming model (3.6). Letting $u = U_i$ in the above
procedure, we have \( \tilde{a}_j(U_i), j = 1, \cdots, \ell, i = 1, \cdots, n \). Then, our initial estimate \( \tilde{\theta} \) is the maximizer of

\[
\sum_{i=1}^{n} \log f(Y_i; X_i, \theta, x_i^T \tilde{a}_1(U_i), \cdots, x_i^T \tilde{a}_\ell(U_i)).
\]

During the iteration in finding the minimizer of (3.5), \( \tilde{a}_{\theta}(\cdot) \) is taken as the estimator that solves (3.4) with \( \hat{\theta} \) and \( \hat{h}_1 \) replaced by \( \theta \) and \( \hat{h}_1 \) respectively. With this choice of bandwidth the least favorable curve is approximated well, by the nature of model (3.6). Upon convergence of the iteration, we obtain the profile likelihood estimator for \( \theta \). Then we can estimate \( a(\cdot) \) and select the bandwidth in the same manner as described in Sections 3.1 and 4.2 with \( \hat{\theta} \) replaced by the profile likelihood estimator for \( \theta \).

## 4 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. There are many model selection criteria under parametric assumptions, e.g. cross-validation (Stone, 1974), AIC (Akaike, 1970), BIC (Schwarz, 1978), and nonconcave penalized likelihood (Fan and Li, 2001). Among these various criteria, due to their easy implementation, AIC and BIC are probably most commonly used in practice. We appeal to the AIC and BIC ideas to select the bandwidths \( h_1 \) and \( h \) in the estimation procedures and to identify the constant parameters in model (1.1).

### 4.1 Model selection criteria

The standard AIC and BIC formulae contain the number of unknown parameters in the model, so does the versions of AIC and BIC for model (1.1). To work out that number for (1.1), we have to find out how many unknown parameters each unknown function \( a_{ij}(\cdot) \) amounts to. In nonparametric modeling, when a locally linear approximation is used, Fan and Gijbels (1996) suggested that an unknown
function amounts to
\[ \text{tr}\{(G_0^T W_0 G_0)^{-1} G_0^T W_0^2 G_0\} \]
unknown parameters, where
\[ G_0 = \begin{pmatrix} 1 & U_1 - u \\ \vdots & \vdots \\ 1 & U_n - u \end{pmatrix}, \quad W_0 = \text{diag}(K_{h_1}(U_1 - u), \cdots, K_{h_1}(U_n - u)). \]

To simplify the calculation, we look into its asymptotic version: when \( n \) is large enough,
\[ \text{tr}\{(G_0^T W_0 G_0)^{-1} G_0^T W_0^2 G_0\} \approx h_1^{-1}(\nu_0 + \nu_2/\mu_2), \]
where \( \nu_i = \int t^i K^2(t)dt, \mu_i = \int t^i K(t)dt \). Thus, we take
\[ K = q + h_1^{-1}(\nu_0 + \nu_2/\mu_2)(p_1 + \cdots + p_\ell), \]
as the number of parameters involved in our estimation procedure for model (1.1).
When the Epanechnikov kernel \( K(t) = 0.75(1 - t^2)_+ \) is used, \( \nu_0 + \nu_2/\mu_2 = 1.028571 \).
The AIC and BIC formulae given later apply to any models of the form (1.1), which can have different \( q, \ell \) or \( x_j \), and model (3.6); in the latter case, \( K = h_1^{-1}(\nu_0 + \nu_2/\mu_2)(q + p_1 + \cdots + p_\ell) \) and \( \hat{\theta}, \hat{a}_j(\cdot), j = 1, \cdots, \ell \), are replaced by \( \bar{a}_j(\cdot), j = 0, 1, \cdots, \ell \), in the formulae.

### 4.2 Bandwidth selection

Suppose (1.1) is the true model and is used to analyze the data. Choice of the bandwidths \( h \) and \( h_1 \) determines performance of the two-step estimators in Section 3.1. Compared to choosing \( h_1 \), selecting \( h \) is relatively simple since \( h \) is employed to get undersmoothed estimators of the functional parameters. It follows from Theorem 1 that we get a good estimator of \( \theta \) by letting \( h \) be of an 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 (3.2) caused by design sparsity. Proper selection of \( h_1 \) is crucial for \( \hat{a}(\cdot) \) to perform well. We propose to obtain a reasonable choice of \( h_1 \) first, then utilize the relationship between the optimal rate of \( h_1 \) and a suitable rate of \( h \) to determine \( h \), and finally select \( h_1 \).
Define the AIC criterion for model (1.1) as
\[
AIC = -2 \sum_{i=1}^{n} \log f(Y_i; X_i, \hat{\theta}, x_i^T \hat{a}_1(U_i), \ldots, x_i^T \hat{a}_\ell(U_i)) + 2k.
\]
To get a reasonable choice of \(h_1\), we compute the version of AIC for model (3.6) for different values of \(h_1\). This yields an AIC function of \(h_1\) only and \(h\) is not involved since there is no constant parameters in (3.6). 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 \(\hat{a}(\cdot)\) for \(a(\cdot)\) in (1.1). The reason is that, when modeling data from (1.1) using (3.6), the true value of \(a_0(\cdot)\) is the constant vector \(\theta\), so the curve estimate of \(a_0(\cdot)\) would be roughly flat for a wide range of \(h_1\) and the AIC criterion for (3.6) mainly measures performance of the estimators of \(a_j(\cdot), j = 1, \ldots, \ell\), while \(h_1\) varies.

Selection of the bandwidth \(h\) in the two-step estimation of \(\theta\) is not a major issue as we explained before. 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 \(n \in [10^2, 10^5]\).

The bandwidth \(\hat{h}_1\) is chosen for estimating the functions \(a_0(\cdot), a_1(\cdot), \ldots, a_\ell(\cdot)\) in model (3.6), which specifies the constant vector \(\theta\) in the true model (1.1) as functional. Now, we refine our data-driven selection of \(h_1\), which is required in the two-step estimation of \(a(\cdot)\) in the true model (1.1). Based \(\hat{h}\) we obtain the two-step estimator \(\hat{\theta}\) for \(\theta\) in (1.1). Plug-in \(\hat{\theta}\) into (3.4) to find \(\hat{a}(\cdot)\) and compute the AIC criterion for model (1.1) for a range of \(h_1\). Denote the minimizer of this AIC function as \(\tilde{h}_1\).

### 4.3 Identifying constant parameters

Define the version of BIC for model (1.1) as
\[
BIC = -2 \sum_{i=1}^{n} \log f(Y_i; X_i, \hat{\theta}, x_i^T \hat{a}_1(U_i), \ldots, x_i^T \hat{a}_\ell(U_i)) + K \log(n).
\]
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 problem interacts with the bandwidth selection problem: the BIC formula depends on the bandwidth \(h_1\). In fact, it is almost impossible to choose the bandwidth and the
constant parameters simultaneously. This is 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 choose the bandwidths \( h_1 \) and \( h \) first and then identify the constant parameters, as suggested in the following.

We start with a model \( M_0 \) which is of the form (3.6), and then decide which parameters in \( M_0 \) are functional and which are constant. The choice of \( a_0(U) \) and \( x_1, \ldots, x_\ell \), and how they determine the dependence of \( Y \) on \( X \) and \( U \) in \( M_0 \) should come from the basic assumptions on the model; in practice they are determined by the analyst. Due to the curse of dimensionality issue, we have to impose some basic assumptions on the model based on some knowledge about the data we analyze, which is usually available from the background of the data or people working with the area where the data arise. Since all the unknown parameters in \( M_0 \) are functions, as in Section 4.2, we choose the bandwidth \( h_1 \) for estimating the unknown functions by minimizing the version of the AIC criterion for model \( M_0 \). For simplicity of notation, denote this bandwidth as \( \hat{h}_1 \) and let \( \hat{h} = \hat{h}_1 n^{-0.051} \) again. 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 is the one with the smallest BIC value. Unfortunately, this approach would immediately become computationally impossible when \( \kappa \), number of parameters that can be either functional or constant, is not very small as there are \( 2^\kappa \) possible combinations. We propose the following iterative procedure to reduce the computational burden. We start with \( M_0 \) as the candidate model, and at the \( L \)-th step of the iteration we examine whether one of the functional parameters in the candidate model \( M_L \) can be further reduced to a constant.

(a): Set \( L = 0 \). Based on model \( M_0 \), compute local likelihood estimates of all the unknown parameter functions using bandwidth \( \hat{h}_1 \).

(b): If \( L = 2^\kappa - 1 \) (all the \( m \) parameters are reduced to constants in \( M_L \)) then \( M_L \) is the chosen model and the model selection is completed. Otherwise, for each of the unknown functions in the candidate model \( M_L \), say \( a_{ij}(\cdot) \), that could be
reduced to a constant calculate
\[ S_{ij} = \sum_{k=1}^{n} (\hat{a}_{ij}(U_k) - \bar{a}_{ij})^2, \quad \bar{a}_{ij} = n^{-1} \sum_{k=1}^{n} \hat{a}_{ij}(U_k). \]

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

(c): Based on the new model \( \mathcal{M}_{L+1} \) and the bandwidths \( \hat{h}_1 \) and \( \hat{h} \), compute the estimates of the unknown functions and constants. Compute the BIC of \( \mathcal{M}_{L+1} \), and compare it with that of \( \mathcal{M}_L \). If \( \mathcal{M}_L \) has a smaller BIC, then \( \mathcal{M}_L \) is the chosen model and the model selection is completed. Otherwise, \( \mathcal{M}_{L+1} \) becomes the candidate model, so we denote the new constant parameter in (b) as \( \theta_{L+1} \) and change \( L \) to \( L + 1 \) then go to (b).

The above iterative process continues until \( \mathcal{M}_L \) has a smaller BIC than \( \mathcal{M}_{L+1} \) for some \( L < 2^\kappa - 1 \) (\( \mathcal{M}_L \) is the chosen model) or until \( L = 2^\kappa - 1 \) (the chosen model has all the considered parameters constant). Apparently, the final chosen model can be written in the form of (1.1).

5 Asymptotic properties

This section investigates asymptotic distributions of the two-step estimators given in Section 3.1. Theory for profile likelihood estimation of \( \theta \) can be developed under another set of regularity conditions. The derivation is straightforward given the existing literature (Severini and Wong, 1992; Murphy and van der Vaart, 2000; Fan and Huang, 2005) and hence is omitted.

For simplicity of notation, the theory presented here concerns the case where \( x_j = X, j = 1, \ldots, \ell \). The established theory straightforwardly carries over to the general case where \( x_1, \ldots, x_\ell \), are different. Let \( \pi(u) \) be the density of \( U \) and \( \tilde{a}_j(u) \) be the second derivative of \( a_j(u) \), \( j = 1, \ldots, \ell \). Write
\[
\mathbf{z} = (z_1, \ldots, z_\ell)^T, \quad z_j = X^T a_j(u), \quad j = 1, \ldots, \ell, \quad \mathbf{D} = I_\ell \otimes (X^T, 0_{1 \times p})^T, \quad \mathbf{D}_c = I_\ell \otimes (0_{1 \times p}, \mathbf{X}^T)^T.
\]

Theorem 1 gives the asymptotic distribution of \( \hat{\theta} \) and shows that \( \hat{\theta} \) is asymptotically unbiased as an estimator of the constant parameter \( \theta \), provided that the bandwidth \( h \) is of an order smaller than that of optimal bandwidths used in univariate smoothing.
Theorem 1. Under the regularity conditions stated in the Appendix, if \( h = o(n^{-1/4}) \) and \( nh/\log^2 n \longrightarrow \infty \), we have

\[
n^{1/2}(\hat{\theta} - \theta) \overset{D}{\longrightarrow} N(0_{q \times 1}, \Delta) \quad \text{when } n \longrightarrow \infty,
\]

where

\[
\Delta = (I_q, 0_{q \times 2p})E\{V_c(U)^{-1}V_0(U)V_c(U)^{-1}\}(I_q, 0_{q \times 2p})^T,
\]

\[
V_0(u) = E\{H\mathcal{I}(\gamma)H^T|U = u\}, \quad V_c(u) = V_0(u) + E\{\mu_2 H_c\mathcal{I}(\gamma)H_c^T|U = u\},
\]

\[
H = \text{diag}(I_q, D), \quad H_c = \text{diag}(0_{q \times q}, D_c), \quad \mathcal{I}(\gamma) = -E \{\hat{g}(Y; X, \gamma)|X, U\},
\]

\[
g(Y; X, \gamma) = \frac{\partial \log f(Y; X, \gamma)}{\gamma}, \quad \hat{g}(Y; X, \gamma) = \frac{\partial g(Y; X, \gamma)}{\gamma}, \quad \gamma = (\theta^T, Z^T)^T.
\]

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

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

\[
(nh_1)^{1/2}\{\hat{\alpha}(u) - \alpha(u) + \mathcal{B}\} \overset{D}{\longrightarrow} N(0_{p \times 1}, \Sigma) \quad \text{when } n \longrightarrow \infty,
\]

where

\[
\mathcal{B} = 2^{-1} \mu_2 h_1^2 I_\ell \otimes \{(1, 0) \otimes I_p\} G_c^{-1} \Gamma,
\]

\[
\Gamma = E\{D\mathcal{I}_1(z)(\bar{a}_1(u), \ldots, \bar{a}_\ell(u))^T|X = u\},
\]

\[
\Sigma = I_\ell \otimes \{(1, 0) \otimes I_p\} G_c^{-1} G G_c^{-1} \Gamma (u)^{-1} I_\ell \otimes \{(1, 0)^T \otimes I_p\},
\]

\[
G = E \left\{ \nu_0 D\mathcal{I}_1(z) + \nu_2 D_c\mathcal{I}_1(z)D_c^T|U = u \right\},
\]

\[
G_c = E \left\{ D\mathcal{I}_1(z) + \mu_2 D_c\mathcal{I}_1(z)D_c^T|U = u \right\},
\]

\[
\mathcal{I}_1(z) = -E \left\{ \left. \frac{\partial^2 \log f(Y; X, \theta, z)}{\partial z \partial z^T} \right| X, U \right\}_{U=u}.
\]
Theorem 2 tells that our estimator \( \hat{a}(\cdot) \) has the adaptivity property: it has the same asymptotic distribution as the estimator of \( a(\cdot) \) that is obtained by maximizing (3.4) with \( \hat{\theta} \) replaced by the true value of \( \theta \). In addition, the optimal bandwidth \( h_1 \) is of the order \( n^{-1/5} \) and the optimal convergence rate of \( \hat{a}(\cdot) \) is \( n^{-2/5} \). We leave the proof of these two theorems to the Appendix.

6 Simulation study and data analysis

6.1 A simulated example

We use a simulated example to demonstrate how well the proposed methods work. Suppose that, conditionally on \( X = x \) and \( U = u \), \( Y \) has a Weibull distribution with density function

\[
f(y; x, \theta, a(u)x) = \frac{\theta}{\{a(u)x\}^{\theta}} y^{\theta-1} \exp \left[ - \left\{ \frac{y}{a(u)x} \right\}^\theta \right], \ y > 0, \tag{6.1}
\]

where the constant \( \theta > 0 \) is taken to be 1, the function \( a(\cdot) \) is set to be \( a(u) = 2 + \sin(2\pi u) \), \( U \sim \text{Uniform}(0,1) \), \( X \sim \text{Uniform}(1,2) \) and \( X \) and \( U \) are independent. We simulated 300 samples of size 1000 from this model and applied our estimation and model selection procedures to the samples.

In the two-step estimation procedure, the bandwidths \( \hat{h} \) and \( \tilde{h}_1 \) in Section 4.2 were employed, with model (3.6) specifying the conditional density

\[
f(y; x, a_0(u), a_1(u)x) = \frac{a_0(u)}{\{a(u)x\}^{a_0(u)-1}} y^{a_0(u)-1} \exp \left[ - \left\{ \frac{y}{a(u)x} \right\}^{a_0(u)} \right], \ y > 0. \tag{6.2}
\]

The kernel function \( K \) was taken to be the Epanechnikov kernel. We use the mean integrated absolute error (MIAE) to assess the accuracy of an estimator. The MIAE of an estimator of an unknown constant is defined as its mean absolute error. The MIAE of an estimator \( \hat{a}(\cdot) \) of an unknown function \( a(\cdot) \) is defined as

\[
\text{MIAE} = E(\text{IAE}), \quad \text{where} \quad \text{IAE} = \int |\hat{a}(u) - a(u)| \, du.
\]

The MIAE for \( \theta \) and \( a(\cdot) \) are respectively 0.0154 and 0.1067. The bias and standard deviation for \( \theta \) are respectively -0.0016 and 0.0194, agreeing with the theory that \( \hat{\theta} \).
is asymptotically unbiased, see Theorem 1. Panel (a) of Fig. 1 plots the pointwise 10%, 50% and 90% quantiles of the 300 curve estimates of $a(\cdot)$. Both estimators of $\theta$ and $a(\cdot)$ are quite accurate. Also, the constant parameter $\theta$ 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 $\hat{a}(\cdot)$ does. Panel (b) of Fig. 1 plots the estimates of $a(\cdot)$ based on the sample with median IAE performance when $\theta$ is treated as unknown (dotted line) and known (dashed line). They are close to each other, indicating that our estimator of $a(\cdot)$ has the adaptivity property.

Figure 1: *Estimates of the functional parameter in the Weibull example.* Panel (a) plots the pointwise 10%, 50% and 90% quantiles (long-dash) of the 300 estimates of $a(\cdot)$ (solid). Panel (b) plots the estimates of $a(\cdot)$ (solid) based on the sample with median IAE performance when the constant parameter $\theta$ is treated as unknown (dotted) or known (dashed).

The profile likelihood method in Section 3.2 was applied to the same 300 samples. The MIAEs for $\theta$ and $a(\cdot)$ are respectively 0.0168 and 0.1077. The bias and standard deviation for $\theta$ are respectively 0.0011 and 0.0206. In this example, $\theta$ is the shape parameter and $a(\cdot)$ determines the scale parameter in the conditional Weibull distribution. The two-step method performs slightly better than the profile likelihood method in estimating the constant (shape) parameter, while the two perform equally well in estimating the functional (scale) parameter.

Suppose it is not known which of the two parameters are constant and which are functional. For each of the 300 samples simulated from (6.1), the model selection
procedure in Section 4.3, with the start model $\mathcal{M}_0$ given by model (6.2), was applied to select the constant parameters. The correct model was selected 296 times, and for all the other 4 samples the model with both $\theta$ and $a$ as functions of $u$ was selected. Thus, our model selection criterion has a high success rate of 98%. Furthermore, it never makes the mistake specifying the functional parameter $a(\cdot)$ as constant, which would result in a large bias and inconsistency in post model selection inference. Comparatively, misspecifying the constant parameter $\theta$ as functional, which our model selection rule does 2% of the time, is a minor problem as the nonparametric estimates under the wrong model would still look flat. Thus, it is a good sign that when our model selection method misspecifies the model, the fitted model still resembles the true model. As an evidence, the MIAE of estimating $\theta$ after model selection is 0.0167 and that of estimating $a(\cdot)$ is 0.1068, both of them are only slightly larger than the corresponding MIAEs with the correct model specified.

To further examine performance of the model selection procedure, for each of the 300 samples, we use the selected model and the corresponding parameter estimates to predict the true values of the parameters $\theta$ and $a(U_0)$ associated with a future observation $(Y_0, X_0, U_0)$. The mean absolute prediction errors for $\theta$ and $a(U_0)$ are 0.0170 and 0.1130 respectively. In Fig. 2, panel (a) is a boxplot of the 300 predicted values of $\theta \equiv 1$, and panel (b) depicts the point cloud 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 misspecifies the model 12% of the time. Thus, we can conclude from this example that our proposed estimation and model selection procedures work together as a powerful tool for multiparameter likelihood modeling even when there is little knowledge about whether or not some of the parameters are constant.

In model (6.1), the Weibull conditional distribution has $d = 2$ parameters, of which $\ell = 1$ follows a nonparametric form and the other $d - \ell = 1$ follows a parametric form. A more complex model is to change the conditional distribution to Weibull($\theta + a_1(u)x, a(u)x$), where $x$, $u$, $\theta$ and $a(\cdot)$ are the same as in (6.1), and $a_1(u) = 0.1 + 0.1 \cos(2\pi u)$. In this case, $\ell = d = 2$, with the shape parameter modeled semiparametrically and the scale parameter modeled nonparametrically.
Figure 2: *Parameter predictions in the Weibull example.* Panel (a) is a boxplot of the predicted values of $\theta \equiv 1$ based on the selected model for the 300 samples. Panel (b) is a scatterplot of the predicted value against the true value of $a(U_0)$ for the 300 samples.

6.2 The analysis for infant mortality in China

The data for this analysis come from a 10 percent subsample of the National Survey of Fertility and Contraceptive Prevalence, often referred to as the ‘Two per Thousand Fertility Survey’, which was conducted by China’s State Family Planning Commission from 1 July 1988 to 15 July 1988. The survey, representing a sample of two per 1,000 persons in Mainland China, targeted ever-married resident women aged 15-57. All provinces in the Chinese mainland took part in the survey. The sample for this study is restricted to births after 1949, i.e. after the founding of the People’s Republic of China. So 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 died before their first birthday, yielding the infant mortality rate 58.4 per thousand.

The response variable $Y$ is taken to be the binary variable: death or survival within the first year (thus births within 12 months before the survey are excluded and the remaining 114,337 births are used for the logistic regression analysis). Selection of relevant independent variables is guided by previous studies on the determinants of infant mortality and constrained by those that were included in the survey. Therefore we use the following variables: year of birth ($U$), reproductive patterns: the age of mother at birth of child ($X_2$), first child ($X_3$) and previous birth interval ($X_4$); socio-
economic variables: urban-rural residence ($X_5$), mother’s education ($X_6$), geographic region of residence ($X_7$), and ethnicity ($X_9$). Other control variables such as sex of child ($X_8$), and breastfeeding ($X_{10}$) in the first year of life were also included.

We categorized the mother’s age to two categories: between 15 to 35 (appropriate age) and otherwise (inappropriate age). The model selection procedure in Section 4.3, with the start model $M_0$ specifying

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

and the bandwidths $\hat{h} = 10.80\%$ and $\hat{h}_1 = 19.55\%$ of the time range, was employed to select which impacts of the covariates are constant and which are functional in the logistic regression. It suggests that the impacts of the mother’s age ($X_2$), mother’s education ($X_6$), ethnic ($X_9$), sex of child ($X_8$) and type of feeding ($X_{10}$) are constant. So, we use model (1.1) with the assumption that

$$
\log \left( \frac{P(Y = 1|X = x, U = u)}{1 - P(Y = 1|X = x, U = u)} \right) = a_1(u) + a_2 x_2 + a_3(u) x_3 + a_4(u) x_4 + a_5(u) x_5 + a_6 x_6 + a_7(u) x_7 + a_8 x_8 + a_9 x_9 + a_{10} x_{10}
$$

fits the data. The proposed two-step estimation method was then used to estimate the impacts of the sociodemographic variables on infant death. The kernel function was taken to be the Epanechnikov kernel. The bandwidth in the estimation of constant parameters was $\hat{h} = 10.80\%$ of the range of time, and the bandwidth for functional coefficients was $\tilde{h}_1 = 18.92\%$ of the range.

The estimate of the impact of the mother’s age ($X_2$) is 0.1041, which indicates that inappropriate age women would have higher risk to have their infants dead. The estimate of the impact of the mother’s education ($X_6$) is $-0.1088$, which means educated women have lower risk to have their infants dead. This difference between educated women and non-educated women is understandable, because well-educated women have easier access to information on nutrition and health care and are better at implementing medical advices. The impact of sex of the child ($X_8$) is $-0.0492$ which indicates that boy infants have lower risk of death than girls. Traditional Chinese culture always favors boys. Like in many of the developing world, girls in China receive far less attention and resources than boys. The estimate of impact of
ethnicity \( (X_9) \) is \(-0.1371\), this tells us that the Han have less risk of infant death than people from minorities. The estimate of the impact of breastfeeding \( (X_{10}) \) is \(-0.2058\), which suggests that breastfeeding is better than other kinds of feeding.

The estimates of the impacts of the other factors are presented in Fig. 3. The confidence bands are 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). From panel (a) of Fig. 3, we can see that infant mortality started high in 1950, went up from 1950 and reached the highest in 1959. After 1959, the mortality went down steadily, however, the pace of the decline slowed down after 1970.

From panel (b) of Fig. 3, we can see clearly that mortality of the first birth is lower than that of the others as the impact of first child on mortality is a negative curve. The interpretation for this finding is that the first child has the advantage of having no previous sibling to compete with for parents’ attention and resources. Cultural factors may also 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 enjoys much more attention and care than others. Also, 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 because of physiological difficulties in delivery and their mother’s lack of previous childbearing and care experience. It also appears that although the impact of first child on mortality is always negative, its absolute value is decreasing from 1950 to 1960 sharply, then slightly decreasing.

From panel (c) of Fig. 3, we find the impact of birth interval on infant mortality is also a negative curve, which means longer birth interval would enhance infant’s chance of survival which is in line with the findings in other literature. Like the impact of first child, the absolute value of the impact of birth interval is going down sharply from 1950 to 1960, then stays unchanged until 1967, and then down again.

From panel (d) of Fig. 3, we can see that infant mortality in rural areas is always higher than in urban areas. The impact of rural residence on infant mortality is going up sharply from 1950 to 1958, then going down sharply until 1972, then going steadily until 1981, then going down sharply. This suggests that the difference between rural
Figure 3: Impacts of covariates on infant mortality. Panels (a)–(e) respectively plot impacts 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 are estimates of impacts of the covariates. The dashed curves are the 95% bands of the estimates.
and urban residence is getting narrow from 1958 to 1987.

We take the three municipal cities, Beijing, Shanghai and Tianjin, as reference. Panel (e) of Fig. 3 suggests that infant mortality in the three municipal cities is lower than in other places, and the difference is increasing from 1952 to 1961, then going down until 1978, and then going up again. The interpretation for this finding is the Chinese government invested in the three municipal cities much more than in other places. Indeed, the three municipal cities had priority on almost everything for quite a long time. Before 1980, there were many goods, including some important medicine and nutritional foods, which could be bought only in these three municipal cities. And these three municipal cities had the best hospitals, the best health care and the best environmental sanitation.

**APPENDIX: PROOFS**

We impose the following technical conditions.

**Regularity Conditions**

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

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

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

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

5. 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 (1).

6. Assume $f(y; X, \theta, z) > 0$, and $f(y; X, \theta, z)$ has a continuous, bounded third derivative with respect to $(\theta, z)$.
Lemma 1. Let \((Z_1, W_1), \ldots, (Z_n, W_n)\) be i.i.d observations from a bivariate random vector \((Z, W)\). Assume further that \(E|W|^s < \infty\) and \(\sup_x |y| \zeta(x, y) dy < \infty\), where \(\zeta\) denotes the joint density of \((Z, W)\). Let \(K\) be a bounded positive function with a bounded support, satisfying a Lipschitz condition. Then

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

provided that \(n^{2s-1} \xrightarrow{} \infty\) for some \(\varepsilon < 1 - s^{-1}\) and \(D\) is a compact set.

**Proof:** This follows immediately from the result obtained by Mack and Silverman (1982).

**Proof of Theorem 1.** By abuse of notation, from now on, we use \(a_j, b_j\) and \(V_c\) to denote the true value of \(a_j(u), b_j(u)\) and \(V_c(u)\), respectively, for a generic point \(u\). Define

\[
\tilde{z}_i = (\theta^T, c_i^T)^T, \quad c_i = (X_i^T \{a_1 + b_1(U_i - u)\}, \ldots, X_i^T \{a_{\ell} + b_{\ell}(U_i - u)\})^T, \\
\xi = (\theta^T, a_1^T, b_1^T, \ldots, a_{\ell}^T, b_{\ell}^T)^T, \quad \gamma_i = (\theta^T, X_i^T a_1(U_i), \ldots, X_i^T a_{\ell}(U_i))^T, \\
H_i = \text{diag} \left( I_q, I_{\ell} \otimes (X_i^T, (U_i - u)X_i^T)^T \right), \quad B = \text{diag} \left( I_q, I_{\ell} \otimes \{\text{diag}(1, h) \otimes I_p\} \right).
\]

We first prove that \(\tilde{\xi} \equiv (\tilde{\theta}(u)^T, \tilde{a}_1(u)^T, \tilde{b}_1(u)^T, \ldots, \tilde{a}_{\ell}(u)^T, \tilde{b}_{\ell}(u)^T)^T\), the maximizer of \(L\) given in (3.2), is a consistent estimator of \(\xi\).

Noticing that, given the sample, \(\tilde{z}_i\) is a function of \(\xi\) and we can write \(\tilde{z}_i\) as \(\tilde{z}_i(\xi)\).

To prove \(\tilde{\xi}\) is consistent, we first prove

\[
P \left( \frac{1}{n} \sum_{i=1}^n K_h(U_i - u) \log \left\{ f(Y_i; X_i, \tilde{z}_i(\xi')) / f(Y_i; X_i, \tilde{z}_i(\xi)) \right\} < 0 \right) \xrightarrow{} 1, \quad \text{as } n \xrightarrow{} \infty, \quad (A.1)
\]

for any \(\xi' \neq \xi\). By the Law of Large Numbers, to prove (A.1), we only need to prove

\[
E \left[ K_h(U_1 - u) \log \left\{ f(Y_1; X_1, \tilde{z}_1(\xi')) / f(Y_1; X_1, \tilde{z}_1(\xi)) \right\} \right] < 0. \quad (A.2)
\]

It is easy to see

\[
E \left[ K_h(U_1 - u) \log \left\{ f(Y_1; X_1, \tilde{z}_1(\xi')) / f(Y_1; X_1, \tilde{z}_1(\xi)) \right\} \right] \\
= \pi(u) E \left[ \log \left\{ f(Y_1; X_1, \tilde{z}_1(\xi')) / f(Y_1; X_1, \tilde{z}_1(\xi)) \right\} | U_1 = u \right] + o(1).
\]
Since the log function is strictly concave, Jensen’s inequality shows that
\[
E\left[\log\left\{f(Y_1; X_1, \tilde{z}_1(\xi')) / f(Y_1; X_1, \tilde{z}_1(\xi))\right\} \mid U_1 = u\right] < \log\left(E\left[f(Y_1; X_1, \tilde{z}_1(\xi')) / f(Y_1; X_1, \tilde{z}_1(\xi)) \mid U_1 = u\right]\right) = 0.
\]

So, (A.2) does hold and this implies that (A.1) holds.

Let \(\tilde{\xi}_j\) and \(\xi_j\) be the \(j\)th components of \(\tilde{\xi}\) and \(\xi\) respectively, for any \(\varepsilon > 0\),
\[
P(\|\tilde{\xi} - \xi\| > \varepsilon) \leq \sum_{j=1}^{q+2p\ell} P(|\tilde{\xi}_j - \xi_j| > \varepsilon_1),
\]
where \(\varepsilon_1 = \varepsilon(q + 2p\ell)^{-1/2}\).

Notice that conditions (6) and (7) imply that \(\tilde{\xi}\) is the unique root of the function
\[
L(\xi''') = \sum_{i=1}^{n} K_h(U_i - u) \log\left\{f(Y_i; X_i, \tilde{z}_i(\xi'''))\right\}.
\]

So, for any fixed \(j, j = 1, \cdots, q + 2p\ell\), letting \(\xi_{-\varepsilon}\) be \(\xi\) with the \(j\)th component replaced by \(\xi_j - \varepsilon_1\), \(\xi_{\varepsilon}\) be \(\xi\) with the \(j\)th component replaced by \(\xi_j + \varepsilon_1\), we have
\[
P(|\tilde{\xi}_j - \xi_j| \leq \varepsilon_1) \geq P\left(L(\xi) > L(\xi_{-\varepsilon}), L(\xi) > L(\xi_{\varepsilon})\right)
\]
\[= 1 - P\left(L(\xi) \leq L(\xi_{-\varepsilon}) \text{ or } L(\xi) \leq L(\xi_{\varepsilon})\right)
\]
\[\geq 1 - P\left(L(\xi) \leq L(\xi_{-\varepsilon})\right) - P\left(L(\xi) \leq L(\xi_{\varepsilon})\right).
\]

By (A.1), we have
\[
P\left(L(\xi) \leq L(\xi_{-\varepsilon})\right) \rightarrow 0, \quad P\left(L(\xi) \leq L(\xi_{\varepsilon})\right) \rightarrow 0.
\]

So,
\[
P(|\tilde{\xi}_j - \xi_j| \leq \varepsilon_1) \rightarrow 1,
\]
which leads to \(P(\|\tilde{\xi} - \xi\| > \varepsilon) \rightarrow 0\), i.e. \(\tilde{\xi}\) is consistent.

By Taylor’s expansion and Conditions (2)-(3), we have
\[
\tilde{z}_i - \gamma_i = -\left(0_{1 \times q}, \begin{array}{c} X_i^T \{a_1(U_i) - a_1(U_i - u)\}, \cdots, X_i^T \{a_\ell(U_i) - a_\ell(U_i - u)\} \end{array} \right)^T
\]
\[= -2^{-1}\left(0_{1 \times q}, X_i^T \hat{a}_1(u), \cdots, X_i^T \hat{a}_\ell(u) \right)^T (U_i - u)^2 + o_P(h^2)
\]

26
uniformly in $i$. Together with Condition (6), the above equality leads to

$$B^{-1} \frac{\partial L}{\partial \xi} = \sum_{i=1}^{n} K_h(U_i - u) B^{-1} \frac{\partial z_i}{\partial \xi} \frac{\partial \log f(Y_i; X_i, z_i)}{\partial z_i} = \sum_{i=1}^{n} K_h(U_i - u) B^{-1} H_i g(Y_i; X_i, \gamma_i)$$

$$= \sum_{i=1}^{n} K_h(U_i - u) B^{-1} H_i g(Y_i; X_i, \gamma_i)$$

$$- 2^{-1} \sum_{i=1}^{n} K_h(U_i - u) B^{-1} H_i g(Y_i; X_i, \gamma_i) \left( 0_{1 \times q}, X_i \tilde{a}_1(u), \ldots, X_i \tilde{a}_\ell(u) \right)^T$$

$$\times (U_i - u)^2 \{1 + o_P(1)\}$$

$$\triangleq A_1 + A_2 \{1 + o_P(1)\}.$$ 

Let

$$\Omega(u) = E \left\{ H \mathcal{T} (\gamma) (0_{1 \times q}, X^T \tilde{a}_1(u), \ldots, X^T \tilde{a}_\ell(u))^T \bigg| U = u \right\}.$$ 

By Lemma 1 and Conditions (1), (3)–(6), we have

$$\frac{1}{n} A_2 = 2^{-1} \pi(u) \Omega(u) \mu_2 h^2 \{1 + o_P(1)\}.$$ 

It is easy to see that

$$n^{-1/2} h^{1/2} A_1 \xrightarrow{D} N \left( 0_{(2p\ell + q) \times 1}, \ V(u) \pi(u) \right),$$ 

where $V(u) = E \left\{ \nu_0 H \mathcal{T}(\gamma) H^T + \nu_2 H \mathcal{J}(\gamma) H_c^T \bigg| U = u \right\}$. By Lemma 1 and Conditions (1)–(6),

$$B^{-1} \frac{\partial^2 L}{\partial \xi \partial \xi^T} B^{-1} = \sum_{i=1}^{n} K_h(U_i - u) B^{-1} \frac{\partial z_i}{\partial \xi} \frac{\partial \log f(Y_i; X_i, \gamma_i)}{\partial z_i} \left( \frac{\partial z_i}{\partial \xi} \right)^T B^{-1}$$

$$= \sum_{i=1}^{n} K_h(U_i - u) B^{-1} H_i g(Y_i; X_i, \gamma_i) H_i^T B^{-1}$$

$$= n V_c(u) \pi(u) \{1 + o_P(1)\}.$$ 

So,

$$(nh)^{1/2} B \left( \frac{\partial^2 L}{\partial \xi \partial \xi^T} \right)^{-1} B A_1 \xrightarrow{D} N \left( 0_{(2p\ell + q) \times 1}, \ V_c(u)^{-1} V(u) V_c(u)^{-1} \pi(u)^{-1} \right),$$

$$B \left( \frac{\partial^2 L}{\partial \xi \partial \xi^T} \right)^{-1} B A_2 = 2^{-1} \mu_2 h^2 V_c(u)^{-1} \Omega(u) \{1 + o_P(1)\}.$$ 

By Taylor’s expansion and the consistency of $\hat{\xi}$, we have

$$\tilde{\xi} - \xi = - \left( \frac{\partial^2 L}{\partial \xi \partial \xi^T} \right)^{-1} \frac{\partial L}{\partial \xi} \{1 + o_P(1)\}.$$
This leads to
\[
\tilde{\theta}(u) - \theta = -(I_q, 0_{q \times (2p \ell)}) \left\{ B \left( \frac{\partial^2 L}{\partial \xi \partial \xi^T} \right)^{-1} B A_1 + B \left( \frac{\partial^2 L}{\partial \xi \partial \xi^T} \right)^{-1} B A_2 \right\} \{1 + o_P(1)\}.
\]

Let \( L_j, A_{2,j} \) and \( \xi_j \) be \( L, A_2 \) and \( \xi \) respectively but with \( u \) replaced by \( U_j \). By Lemma 1 and Conditions (1)–(6),
\[
\frac{1}{n} \sum_{j=1}^{n} B \left( \frac{\partial^2 L_j}{\partial \xi_j \partial \xi_j^T} \right)^{-1} B A_{2,j} = 2^{-1} \mu_2 h^2 E\{V_c(U)^{-1} \Omega(U)\}\{1 + o_P(1)\}.
\]

Let \( A_{1,j}, V_{c,j} \) and \( H_{i,j} \) be \( A_1, V_c \) and \( H_i \) respectively but with \( u \) replaced by \( U_j \). By Lemma 1, we have
\[
n(\hat{\theta} - \theta) = (I_q, 0_{q \times (2p \ell)}) \sum_{j=1}^{n} B \left( \frac{\partial^2 L_j}{\partial \xi_j \partial \xi_j^T} \right)^{-1} B A_{1,j} + O_P(nh^2)
\]
\[
= (I_q, 0_{q \times (2p \ell)}) \sum_{j=1}^{n} \left\{ \sum_{k=1}^{n} K_h(U_k - U_j) B^{-1} H_{k,j} \dot{g}(Y_k; X_k; \gamma_k) H_{k,j}^T B^{-1} \right\}^{-1}
\times \sum_{i=1}^{n} K_h(U_i - U_j) B^{-1} H_{i,j} g(Y_i; X_i; \gamma_i) + O_P(nh^2)
\]
\[
= n^{-1}(I_q, 0_{q \times (2p \ell)}) \sum_{i=1}^{n} \sum_{j=1}^{n} K_h(U_i - U_j) V_{c,j}^{-1} \pi(U_j)^{-1} B^{-1} H_{i,j} g(Y_i; X_i; \gamma_i)
\times \{1 + o_P(1)\} + O_P(nh^2).
\]

By tedious calculation, we have
\[
n^{-3/2} \sum_{i=1}^{n} \sum_{j=1}^{n} K_h(U_i - U_j) V_{c,j}^{-1} \pi(U_j)^{-1} B^{-1} H_{i,j} g(Y_i; X_i; \gamma_i)
\overset{D}{\rightarrow} N\left( 0_{(2p \ell + q) \times 1}, E\{V_c(U)^{-1} V_0(U) V_c(U)^{-1}\} \right),
\]
which implies that, if \( h = o(n^{-1/4}) \),
\[
n^{1/2}(\hat{\theta} - \theta) \overset{D}{\rightarrow} N(0_{q \times 1}, \Delta).
\]

**Proof of Theorem 2.** Let
\[
L_1 = \sum_{i=1}^{n} K_{h_1}(U_i - u) \log f(Y_i; X_i; \theta, c_i),
\]
\[
\eta = (a^T_1, b^T_1, \cdots, a^T_\ell, b^T_\ell)^T, \quad B_1 = I_\ell \otimes \{\text{diag}(1, h_1) \otimes I_p\},
\]
\[
28
\]
\[ m(Y_i; X_i, \theta, z_i) = \frac{\partial \log f(Y_i; X_i, \theta, z_i)}{\partial z_i}, \quad z_i = (X_i^T a_1(U_i), \ldots, X_i^T a_\ell(U_i))^T, \]

\[ \hat{m}(Y_i; X_i, \theta, z_i) = \frac{\partial m(Y_i; X_i, \theta, z_i)}{\partial z_i}, \quad D_i = I_\ell \otimes (X_i^T, (U_i - u)X_i^T)^T. \]

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

By simple calculation and Conditions (2) and (6) we have

\[
B_1^{-1} \frac{\partial L_1}{\partial \eta} = \sum_{i=1}^{n} K_{h_1}(U_i - u) B_1^{-1} \frac{\partial c_i \partial \log f(Y_i; X_i, \theta, c_i)}{\partial \eta} \frac{\partial c_i}{\partial \eta} \\
= \sum_{i=1}^{n} K_{h_1}(U_i - u) B_1^{-1} D_i \frac{\partial \log f(Y_i; X_i, \theta, c_i)}{\partial c_i} \\
= \sum_{i=1}^{n} K_{h_1}(U_i - u) B_1^{-1} D_i \hat{m}(Y_i; X_i, \theta, z_i) \\
- 2^{-1} \sum_{i=1}^{n} K_{h_1}(U_i - u) B_1^{-1} D_i \hat{m}(Y_i; X_i, \theta, z_i) (\bar{a}_1(u), \ldots, \bar{a}_\ell(u))^T \rightleftharpoons \frac{1}{n} \{ 1 + o_P(1) \}.
\]

\[ \triangleq J_1 + J_2 \{ 1 + o_P(1) \}. \]

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

\[ \frac{1}{n} J_2 = 2^{-1} \Gamma \mu_2 h_1^2 \pi(u) \{ 1 + o_P(1) \}. \]

By the Central Limit Theorem,

\[ n^{-1/2} h_1^{1/2} J_1 \overset{D}{\rightarrow} N \left( 0_{(2\rho \ell) \times 1}, \ G\pi(u) \right). \]

By Lemma 1 and Conditions (1)–(6),

\[
B_1^{-1} \frac{\partial^2 L_1}{\partial \eta \partial \eta^T} B_1^{-1} = \sum_{i=1}^{n} K_{h_1}(U_i - u) B_1^{-1} \frac{\partial c_i \hat{m}(Y_i; X_i, \theta, c_i)}{\partial \eta} \hat{m}(Y_i; X_i, \theta, c_i) \left( \frac{\partial c_i}{\partial \eta} \right)^T B_1^{-1} \\
= \sum_{i=1}^{n} K_{h_1}(U_i - u) B_1^{-1} D_i \hat{m}(Y_i; X_i, \theta, z_i) D_i^T B_1^{-1} \\
= n G \pi(u) \{ 1 + o_P(1) \}. 
\]

So,

\[
(n h_1)^{1/2} B_1 \left( \frac{\partial^2 L_1}{\partial \eta \partial \eta^T} \right)^{-1} B_1 J_1 \overset{D}{\rightarrow} N \left( 0_{(2\rho \ell) \times 1}, \ G\pi^{-1}(u) \{ 1 + o_P(1) \} \right), \\
B_1 \left( \frac{\partial^2 L_1}{\partial \eta \partial \eta^T} \right)^{-1} B_1 J_2 = 2^{-1} \mu_2 h_1^2 G \pi^{-1} \Gamma \{ 1 + o_P(1) \}. 
\]
By Taylor’s expansion and the consistency of $\tilde{\eta}$, we have

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

So,

$$\left( n h_1 \right)^{1/2} \left\{ B_1 (\tilde{\eta} - \eta) + 2^{-1} \mu_2 h_1^2 G_e^{-1} \Gamma \right\} \overset{D}{\longrightarrow} N \left( 0_{(2p\ell)\times 1}, \ G_e^{-1}GG_e^{-1} \pi (u)^{-1} \right).$$

Because $\hat{\theta}$ has the $n^{-1/2}$ convergence rate, the maximizer of $L_1$ with respect to $\eta$ would behave exactly the same as the maximizer of (3.4) asymptotically. Hence,

$$\left( n h_1 \right)^{1/2} (\tilde{a} - a + B) \overset{D}{\longrightarrow} N \left( 0_{(p\ell)\times 1}, \ \Sigma \right).$$

REFERENCES


