SEMI-LINEAR LINEAR INDEX MODEL WHEN THE LINEAR COVARIATES AND INDICES ARE INDEPENDENT

By Yun Sam Chong, Jane-Ling Wang* and Lixing Zhu†

Wecker Associate,
University of California at Davis,
and The University of Hong Kong

Summary Most dimension reduction models are suited for continuous but not for discrete covariates. A flexible model to incorporate both discrete and continuous covariates is to assume that some covariates, a q-dimensional vector $\boldsymbol{Z}$, are related to the response variable, Y, through a linear relationship; while the remaining covariates, a p-dimensional vector $\boldsymbol{X}$, are related to Y through k indices which are $\boldsymbol{X}'\mathbf{B}$ and some unknown function g. This results in a semi-parametric model called semi-linear index model, which includes both the popular single-index model and the partial linear model as special cases. To avoid the curse of dimensionality, k should be much smaller than p, and this is often realistic as the key features of a high dimensional variable can often be extracted through a low-dimensional subspace. Two approaches to estimate the model components have been considered by Carroll, Fan, Gijbels and Wand (1997) and Yu and Ruppert (2002). Both focus on the simple case where $k = 1$, as these approaches are challenged computationally when $k > 1$. These computational challenges are partly triggered by the fact that the estimation of the parametric components is integrated with the non-parametric estimation of the link function g. Moreover, the theory for both approaches is incomplete. We show in this paper that a simple approach which separates the dimension reduction stage to estimate $\mathbf{B}$ from the remaining model components is available when the two covariates Z and X are independent. For instance, one can apply any suitable dimension reduction approach, such as the average derivative method, the projection pursuit regression or sliced inverse regression, to get an initial estimator for $\mathbf{B}$ which is consistent at the $\sqrt{n}$ rate, and then apply a suitable approach, such as the profile

*Wang's research was supported in part by NSF grants DMS-0406430 and DMS-98-03627.

†Zhu's research was supported by the grants (HKU7181/02H and HKU7060/04P) from The Research Grants Council of Hong Kong.

AMS 2000 subject classifications: Primary 62G08, 62G20; secondary 62F12

Keywords and phrases: Partial regression; single-index; nonparametric smoothing; dimension reduction; projection pursuit regression; sliced inverse regression.

approach of partial regression, to estimate the regression coefficient of Z and the link function g. All three estimates can be refined by iterating the procedure once. Such an approach is computationally simple and yields efficient estimates for both parameters at the $\sqrt{n}$ rate. We provide both theoretical proofs and empirical evidence.

1. Introduction.. One of the areas Kjell Doksum has made seminal contributions is *dimension-reduction methods*. This includes work in transformation models (Doksum (1987), Dabrowska and Doksum (1988ab)), and Doksum and Gasko (1990)), where the first two papers demonstrate that the partial likelihood method for proportional hazards model can be extended to general transformation models. The third paper explores another semiparametric model, the generalized odds-rate model, and introduces a class of rather efficient estimators for the proportionality parameter. The last paper links the binary regression model to survival models. Another line of work, also for *dimension-reduction methods* is the average derivative estimator (**ADE**) method (Doksum and Samarov(1995) and Chaudhuri, Doksum and Samarov(1997)), where the average derivative approach is shown to be a promising dimension reduction tool. All these aforementioned papers employed semiparametric models to accomplish the dimension reduction goal and to further explore inferences for the parametric components.

Our objective in this paper is to explore the dimension reduction topic through a particular semiparametric model, termed SLIM (semi-linear indices model). We show that in a simple and special situation, efficiency for the parametric estimators can easily be achieved by various dimension reduction tools. The model is motivated by the fact that many dimension-reduction methods, such as projection pursuit regression(**PPR**), average derivative estimator method(**ADE**), and sliced inverse regression(**SIR**), assume implicitly that the predictors are continuous variables and will not work well when some of the predictors are discrete. One solution to this problem is the use of a semiparametric model where it is assumed that the response variable, Y, has a parametric relationship with some q-dimensional covariates $\boldsymbol{Z}$ (some of which may be discrete), but a nonparametric relationship with other p-dimensional covariates $\boldsymbol{X}$. If $\boldsymbol{X}$ is of high dimension, additional dimension reduction is needed and the most common approach is to assume that all information of $\boldsymbol{X}$ that is related to Y is carried through a few, say k, indices. More specifically, we assume:

$$Y = \boldsymbol{Z}'\boldsymbol{\theta} + g(\boldsymbol{X}'\mathbf{B}) + e. \tag{1}$$

The function g (we call it the link function) and the $p \times k$ matrix $\mathbf{B}$ describe the dimension-reduction model through which Y and $\boldsymbol{X}$ are related, $\boldsymbol{\theta}$ is the vector of parameters describing the linear relationship between Y and $\boldsymbol{Z}$, and e is an error term. Model (1) is a semi-linear model

with k indices and will be abbreviated as *SLIM*(semi-linear indices model) hereafter. Dimension reduction is accomplished because k is usually much smaller than the dimension p of $\boldsymbol{X}$. In addition, the other covariate vector $\boldsymbol{Z}$ is related to Y through a linear relation. When the link function g is unknown, the matrix $\mathbf{B}$ can be identified only in terms of direction and not size. We thus assume hereafter that the column vectors of $\mathbf{B}$ are all of unit length and the first component is always nonnegative.

The special case $p = 1$ has vast appeal to econometricians and is called the "partial linear model" (Engle, Granger, Rice and Weiss (1986), Heckman (1986), Rice (1986), Denby (1986), Chen(1988), Speckman (1988), Severini and Staniswalis (1994), Bhattacharya and Zhao (1997), Hamilton, S. A. & Troung, Y. K. (1997), Mammen and van de Geer (1997)). Model (1) also includes another popular model when p might be larger than 1 but $k = 1$, in which case Y and $\boldsymbol{X}$ are related through a single dimension reduction direction called index and the resulting model is called the "single index model" in the economics literature (Stoker (1989), Härdle, Hall and Ichimura (1993), Chiou and Müller (1999), Stute and Zhu (2005)). In contrast to partial linear models and single index models, where hundreds of papers appeared in the literature, the results are sparse for the *SLIM* model in (1). Carroll, Fan, Gijbels, and Wand (1997) is the first work where this topic is explored, focusing on the case $k = 1$ with a single index. Their methods sometimes encounter numerical difficulties and this was noticed independently by Chong (1999) and Yu and Ruppert (2002). The latter authors circumvented the problem by assuming that (in addition to $k = 1$) the link function g lies in a known, finite-dimensional spline space, yielding a flexible parametric model. The approach taken in Chong (1999) is different and completely nonparametric, employing a local polynomial smoother to estimate g. Moreover, the number of indices k is not assumed to be 1 or even known, and is being estimated along the way.

We consider in this paper that a random sample of n observations are collected, and use $\mathbf{y} = (y_1, \ldots, y_n)'$ to denote the vector of observed responses and

$$\mathbf{Z} = \begin{bmatrix} z_{11} & \ldots & z_{1q} \\ \vdots & & \vdots \\ z_{n1} & \ldots & z_{nq} \end{bmatrix} \text{ and } \mathbf{X} = \begin{bmatrix} x_{11} & \ldots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \ldots & x_{np} \end{bmatrix}$$

to represent the observed values of $\boldsymbol{Z}$ and $\boldsymbol{X}$, with the first subscript representing the observation number and the second subscript representing the position in the array of variables. Restating equation (1) to reflect the observations we obtain,

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\theta} + g(\mathbf{XB}) + \boldsymbol{\epsilon}, \tag{2}$$

where $\boldsymbol{\epsilon}$ is the $n \times 1$ vector of observed errors.

Our goal is to show that simple and non-iterative algorithms are available when the two covariates $\boldsymbol{Z}$ and $\boldsymbol{X}$ are independent of each other, and the procedures yield efficient estimators for parameters $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$. The independence assumption could be fulfilled, for instance, in clinical trials when $\boldsymbol{Z}$ models the treatments effects and patients are assigned to treatments randomly. Specifically for $\boldsymbol{\theta}$, we show that our procedures are as efficient as the nonlinear estimate obtained as if g and $\mathbf{B}$ were both known. We also show that our procedures provide adaptive estimates for $\mathbf{B}$, in the sense that the asymptotic variance of the $\mathbf{B}$-estimator is equal to the one that assumes a known link function. Finally, we illustrate our procedures through extensive simulation studies and show that they compare favorably with the procedure in Carroll et al. (1997). We note here that the algorithms in Section 2 were first reported in the Ph.D. thesis of Chong (1999), and have not been published previously. Moreover, the asymptotic results reported here are different from and deeper than those obtained in Chong (1999). For instance, Theorem 2 in Section 2 is for a different and better estimator than the one in Theorem 2 of Chong (1999), and Theorem 3 in Section 2 is completely new.

2. Main Results.. Hereafter, we assume that $\boldsymbol{X}$ and $\boldsymbol{Z}$ are independent. Consequently, we may consider $\boldsymbol{Z}'\boldsymbol{\theta}$ of equation (1) to be part of the error term and use only the values of Y and $\boldsymbol{X}$ to obtain an estimate of $\mathbf{B}$. The theorem below shows that we can obtain a $\sqrt{n}$-consistent estimate for $\mathbf{B}$ when we apply the sliced inverse regression (**SIR**) method in Li (1991) to Y and $\boldsymbol{X}$, if the following linear condition is satisfied:

$$\text{for any } \boldsymbol{b} \in \Re^p, E(\boldsymbol{X}'\boldsymbol{b}|\boldsymbol{X}'\mathbf{B}) \text{ is linear in } \boldsymbol{X}'\boldsymbol{\beta}_1, \ldots, \boldsymbol{X}'\boldsymbol{\beta}_k. \tag{3}$$

THEOREM 1. *Under condition (3), $E(\boldsymbol{X}|Y) - E(\boldsymbol{X}) \propto \boldsymbol{\Sigma}_x \mathbf{B}\boldsymbol{a}^*$ for some $\boldsymbol{a}^* \in \Re^k$, where $\boldsymbol{\Sigma}_x$ is the covariance matrix of X, and "$\propto$" stands for "proportional to".*

The proof is similar to the one in Li (1991) because $E(\boldsymbol{X}|Y) = E(E(\boldsymbol{X}|\boldsymbol{X}'\mathbf{B}, \boldsymbol{Z}'\boldsymbol{\theta}, e)|Y) = E(E(\boldsymbol{X}|\boldsymbol{X}'\mathbf{B})|Y)$, where the last equality follows from the fact that $\boldsymbol{X}$ is independent of $\boldsymbol{Z}$ and e. Details of the proof will not be presented here as they can be found in Section 5.3 of Chong (1999).

Because $E(\boldsymbol{X}|Y) = E(E(\boldsymbol{X}|\boldsymbol{X}'\mathbf{B})|Y) = E(\boldsymbol{\Sigma}_x\mathbf{B}\boldsymbol{a}^*|Y)$ for some $\boldsymbol{a}^* \in \Re^k$, we can use **SIR**, proposed in Li (1991) and reviewed in Chen and Li (1998), to estimate $\mathbf{B}$. Variants of **SIR**, such as **SIR II** (Li (1991)), **SAVE** (Cook and Weisberg (1991)), **PHD** (Li (1992)) etc., are also feasible in case **SIR** fails. All these **SIR** based procedures are simple as they do not involve smoothing and separate the dimension reduction stage from the model fitting stage. Li (1991) and Zhu and Ng (1995) states that **SIR** yields a $\sqrt{n}$-consistent estimate for $\mathbf{B}$, when $\boldsymbol{Z}$ is not

present in model. These results continue to hold when $\boldsymbol{Z}$ is independent of $\boldsymbol{X}$. Zhu and Fang (1996) used kernel estimation, where the root n consistency also holds.

By the same token, the average derivative method, **ADE**, can also be applied under certain smoothness conditions as described in Härdle and Stoker (1989) and Samarov (1993). The resulting estimate would be $\sqrt{n}$ consistent like SIR and it has the same advantage as SIR (or its variants) that it separates the dimension reduction stage from model fitting. A relevant method, the Outer Product of Gradients estimation (OPG) proposed by Xia, Tong, Li and Zhu (2002) can also be applied. While SIR relies on the linear conditional mean design condition (3), it is much simpler to implement than **ADE** which involves the estimation of the derivative of g. These two different approaches compliment each other as dimension reduction tools.

If the additivity assumption is satisfied in the projection pursuit regression (**PPR**) model (Friedman and Stuetzle (1981)), one can also employ the **PPR**-estimators for $\mathbf{B}$, which were shown to be $\sqrt{n}$-consistent in Hall (1989). Hristache, Juditsky and Spokoiny (2001) provided a new class of $\sqrt{n}$-consistent estimators. These projection pursuit type estimators typically yield a more efficient initial estimators for $\mathbf{B}$ than **SIR** or **ADE** since **PPR** utilizes the additive model structure and attempts to estimate $\mathbf{B}$ iteratively while estimating the unknown link function. However, **ADE** and **SIR** (or its variants) have the advantage that they rely on no model assumption, separate the dimension reduction stage from model fitting, and are thus computationally simpler and more robust than the **PPR** approach.

2.1. ***Estimation of*** θ***.*** There are two ways to estimate $\boldsymbol{\theta}$:

i Procedures starting with dimension reduction: Start with a dimension-reduction procedure to obtain an estimate $\hat{\mathbf{B}}$ for $\mathbf{B}$ and then follow the steps of the partially linear model, using $\boldsymbol{X}'\hat{\mathbf{B}}$ instead of the unknown $\boldsymbol{X}'\mathbf{B}$ to estimate $\boldsymbol{\theta}$.

ii Procedures starting with initial estimation of the linear component: Because $\boldsymbol{Z}$ and $\boldsymbol{X}$ are independent, linear regression of y on $\boldsymbol{Z}$ will yield a consistent estimate of $\boldsymbol{\theta}$. The linear regression procedure is computationally simple, so we may start with this initial estimate of $\boldsymbol{\theta}$ and use it to improve the dimension-reduction step above.

When using partial linear model estimation to estimate $\boldsymbol{\theta}$, there are two common approaches based on either partial splines (Wahba (1984)) or partial regression (proposed independently by Denby (1986) and Speckman (1986)). The partial regression method is a profile approach so it is also referred to as the profile estimator in the literature. We advocate the use of the partial regression estimator even though partial spline estimators would be fine when $\boldsymbol{Z}$ and $\boldsymbol{X}$ are independent as reported in Heckman (1986). Simulation results in Chapter 6 of Chong (1999) suggest that the two procedures provide numerically equivalent estimators under the independence assumption, but the partial spline procedure might be biased when the independence assumption

is violated as demonstrated in Rice (1986). There is thus no advantage to employ the partial spline procedure in our setting.

The partial regression stage involves a smoothing method to estimate the unknown link function g. The choice of smoother is not critical; we employed the local polynomial smoother due to its appealing properties as reported in Fan (1993). This results in a linear smoother in the sense that we may construct a smoothing matrix $\mathbf{S}$ such that $\mathbf{Su}$ represents the result of smoothing a vector of generic observations, $\mathbf{u}$, using the linear smoother $\mathbf{S}$. Details are given in Appendix B. Below we use the partial regression procedure to estimate $\boldsymbol{\theta}$ and provide the algorithms for each of the approaches above.

Algorithm for Procedure 1 which begins with dimension reduction :

i Apply a dimension-reduction procedure to $\boldsymbol{X}$ and y to obtain an estimate $\hat{\mathbf{B}}$ of $\mathbf{B}$.
ii Use $\mathbf{X}\hat{\mathbf{B}}$ to obtain a smoothing matrix $\mathbf{S}$.
iii Take $\hat{\boldsymbol{\theta}} = (\mathbf{Z}'(\mathbf{I}-\mathbf{S})'(\mathbf{I}-\mathbf{S})\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{I}-\mathbf{S})'(\mathbf{I}-\mathbf{S})\mathbf{y}$ to be the estimate for $\boldsymbol{\theta}$.

Algorithm for Procedure 2 which starts with an initial estimator of the linear component:

i Apply least squares to $\boldsymbol{Z}$ and y to obtain an initial estimate $\hat{\boldsymbol{\theta}}_0$ of $\boldsymbol{\theta}$.
ii Apply a dimension-reduction procedure to $\boldsymbol{X}$ and $y - \boldsymbol{Z}'\hat{\boldsymbol{\theta}}_0$ to obtain an estimate $\hat{\mathbf{B}}$ of $\mathbf{B}$.
iii Use $\mathbf{X}\hat{\mathbf{B}}$ to obtain a smoothing matrix $\mathbf{S}$.
iv Take $\hat{\boldsymbol{\theta}}_1 = (\mathbf{Z}'(\mathbf{I}-\mathbf{S})'(\mathbf{I}-\mathbf{S})\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{I}-\mathbf{S})'(\mathbf{I}-\mathbf{S})\mathbf{y}$ to be the revised estimate for $\boldsymbol{\theta}$.

Simulation results in Section 3 suggest that Procedure 2 which uses the residuals to perform the dimension reduction step is slightly more efficient than Procedure 1. We thus present the asymptotic distribution of $\hat{\boldsymbol{\theta}}_1$ based on Procedure 2 only. Note that, following Theorem 1 or the discussions afterwards at the end of Section 1, many initial $\sqrt{n}$-consistent estimators of $\mathbf{B}$ exist. We thus make such an assumption in the following theorem.

THEOREM 2. *Under conditions (1)–(11), listed in Appendix A, and* $\|\hat{\mathbf{B}} - \mathbf{B}\| = O_P(n^{-1/2})$,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}) \Rightarrow N(0, \mathbf{E}^{-1}\sigma^2) \tag{4}$$

for the partial regression estimate in Procedure 2 using residuals, where

$$\mathbf{E} = \begin{bmatrix} E(Z_1^2) & \cdots & E(Z_1 Z_q) \\ \vdots & & \vdots \\ E(Z_1 Z_q) & \cdots & E(Z_q^2) \end{bmatrix},$$

and $E(Z_i) = 0$ *for all* i.

In other words, when we start with a $\sqrt{n}$-consistent estimate for $\boldsymbol{\beta}$, then $\hat{\boldsymbol{\theta}}_1$ is consistent for $\boldsymbol{\theta}$ with the same efficiency as an estimate that we would obtain if we knew $\boldsymbol{\beta}$ and g. This illustrates the adaptiveness of $\hat{\boldsymbol{\theta}}_1$; no iteration is required, and Many $\sqrt{n}$-consistent estimators for $\boldsymbol{\beta}$ exist.

2.2. ***Estimation of*** $\mathbf{B}$ ***and*** g*..* While our primary interest is in the estimation of $\boldsymbol{\theta}$, we may also be interested in estimating $\mathbf{B}$ and g. Although both procedures in Section 2.1 involve the estimation of $\mathbf{B}$, we will generally want a refined estimate. For instance, after obtaining $\hat{\theta}_1$, we can obtain a revised estimate $\hat{\mathbf{B}}_2$ for $\mathbf{B}$, and then an estimate for g by smoothing $Y - \boldsymbol{Z}'\hat{\boldsymbol{\theta}}_1$ on $\boldsymbol{X}'\hat{\mathbf{B}}_2$.

i Apply a dimension-reduction procedure to $\mathbf{X}$ and $Y - \mathbf{Z}\hat{\theta}_1$ to obtain a revised estimate $\hat{\mathbf{B}}_2$ of $\mathbf{B}$.
ii Use $\mathbf{X}\hat{\mathbf{B}}_2$ to obtain a smoothing matrix $\mathbf{S}$.
iii Let $\hat{g}(\mathbf{X}\hat{\mathbf{B}}_2) = \mathbf{S}(\mathbf{y} - \mathbf{Z}\hat{\boldsymbol{\theta}}_1)$ be the estimate for g.

Next, we present the asymptotic results for the indices parameters. For simplicity, we present only the single index case with $k = 1$, which involves only one-dimensional smoothing. The general case can be derived similarly by using a k-dimensional smoother with modified rates of convergence.

When $k = 1$, the matrix $\mathbf{B}$ becomes a vector and we denote it by $\boldsymbol{\beta}$. Let $\hat{\boldsymbol{\beta}}_2$ be the corresponding estimator in step 1 above. Theorem 3 below shows that it is optimal in the sense that the asymptotic variance of $\hat{\boldsymbol{\beta}}_2$ is equal to the nonlinear least squares estimator that is obtained when the link function $g(\cdot)$ is known and when the linear part $Z'\theta$ is absent in the model. That is, the impact of nonparametric estimation of $g(\cdot)$ and the linear part $Z'\theta$ is negligible asymptotically. Let

$$W = \int \Big\{X - E(\mathbf{X}|\mathbf{X}'\boldsymbol{\beta})\Big\}\Big\{\mathbf{X} - \mathbf{E}(\mathbf{X}|\mathbf{X}'\boldsymbol{\beta})\Big\}' (\mathbf{g}'(\mathbf{X}'\boldsymbol{\beta}))^{\mathbf{2}}\mathbf{f}_{\mathbf{X}}(\mathbf{X})\mathbf{d}\,\mathbf{X},$$

where $f_{\mathbf{X}}$ is the density function of the p-dimensional vector, $\mathbf{X}$, and W^- denotes the generalized inverse of W.

THEOREM 3. *Under conditions 1 – 11 stated in Appendix A and in addition,* $h = O(n^{-1/5})$*, we have, for any unit-vector* $u \neq \boldsymbol{\beta}$*,*

$$n^{1/2}u'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \Longrightarrow N(0, u'\sigma^2(W^-)u)$$

2.3. ***Iterated estimate of*** θ*..* While the estimator for $\boldsymbol{\theta}$ in Section 2.1 is already asymptotically efficient, it might be improved in the finite sample case by iterating the algorithm. For instance, following the steps of the previous section and after obtaining estimates for $\mathbf{B}$ and g, one can use partial regression to obtain the revised estimate for $\boldsymbol{\theta}$.

- Let the partial regression estimate $\hat{\boldsymbol{\theta}}_2 = (\mathbf{Z}'(\mathbf{I}-\mathbf{S})'(\mathbf{I}-\mathbf{S})\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{I}-\mathbf{S})'(\mathbf{I}-\mathbf{S})\mathbf{y}$ be the revised estimate for $\boldsymbol{\theta}$.

The simulation studies in Section 3 indicate some gain by adding one iteration.

3. Simulations. In this section we check the numerical performance of the procedures in Section 2 and compare it to the GPLSIM algorithm from Carroll, Fan, Gijbels, and Wand (1997). We consider two different models, that is, a linear model,

$$Y = 2 + \boldsymbol{X}'\boldsymbol{\beta} + Z\theta + 0.3e, \tag{5}$$

and a quadratic model

$$Y = (2 + \boldsymbol{X}'\boldsymbol{\beta})^2 + Z\theta + 0.3e. \tag{6}$$

In each model, θ is a scalar with value 1. The variable Z will be a single binary variable with values 0 and 1, and $Z = 1$ with probability 1/2. The e's are standard normal, and $\boldsymbol{\beta}$ are $(0.75, 0.5, -0.25, -0.25, 0.25)'$. The $\boldsymbol{X}$'s are standard multivariate normal, with mean $(0,0,0,0,0)'$ and covariance $\mathbf{I}_5$. Thus, in these simulations the assumption on the distribution of $\boldsymbol{X}$ is satisfied for both projection pursuit regression and sliced inverse regression, and we focus on these two dimension reduction methods. The average derivative method can also be used at additional computational cost. Although the simulations shown here have $\boldsymbol{X}$ with components independent from one another, we also ran simulations that have a correlation structure on $\boldsymbol{X}$. The results for those simulations are not much different from those shown here.

We ran $N = 100$ simulations each on the linear model and the quadratic model, and the sample size is $n = 100$ in both cases. The link function is estimated with a local linear smoother as defined in Appendix B, and the bandwidths are chosen by a cross-validation method. We advocate the generalized cross-validation procedure due to its computational advantage over the least squares cross validation method. Simulation results not reported here, which can be found in Chapter 6 of Chong (1999), show that the two-cross validation methods yielded very similar results. We thus only report the findings based on generalized cross-validation (Craven and Wahba (1979)), defined in equation (41) of Appendix B.

The performance of the two types of partial regression estimators, with and without an initial estimate of $\boldsymbol{\theta}$, are compared using two types of dimensions reduction tools, the **PPR** and **SIR** with 2, 5, 20, and 20 elements per slice. The results of estimating $\boldsymbol{\theta}$ for the linear and quadratic model are reported in Table 1 and Table 2 respectively.

We find that, as expected, PPR generally outperforms SIR, but only slightly. With only one iteration, the estimators in section 2.3 are nearly as efficient as the one with $\boldsymbol{\beta}$ known regardless

TABLE 1
Estimates of in the linear model (5)

		W/O initial $\hat{\theta}$	One iter w/o initial $\hat{\theta}$	With initial $\hat{\theta}$	One iter with initial $\hat{\theta}$
PPR	Bias	-0.0411	-0.0029	-0.0013	-0.0007
	SD	0.0623	0.0581	0.0603	0.0592
	MSE	0.00557	0.00338	0.00363	0.0035
SIR_2	Bias	-0.0459	-0.005	-0.0058	-0.0024
	SD	0.0708	0.0606	0.0618	0.061
	MSE	0.00712	0.0037	0.00385	0.00373
SIR_5	Bias	-0.0447	-0.0047	-0.0008	-0.0007
	SD	0.0621	0.0606	0.064	0.0623
	MSE	0.00585	0.00369	0.0041	0.00388
SIR_{10}	Bias	-0.0423	-0.0034	-0.0023	-0.0003
	SD	0.0655	0.06	0.0624	0.0603
	MSE	0.00608	0.00361	0.0039	0.00364
SIR_{20}	Bias	-0.0441	-0.0032	0.0006	-0.001
	SD	0.065	0.0612	0.065	0.0599
	MSE	0.00617	0.00376	0.00423	0.00358
given	Bias	0.0025	0.0025		
	SD	0.0564	0.0564		
	MSE	0.00318	0.00318		

TABLE 2
Estimates of in the quadratic model (6)

		W/O initial $\hat{\theta}$	One iter w/o initial $\hat{\theta}$	With initial $\hat{\theta}$	One iter with initial $\hat{\theta}$
PPR	Bias	-0.0466	-0.0043	-0.005	-0.0009
	SD	0.0625	0.0586	0.0786	0.059
	MSE	0.00608	0.00345	0.0062	0.00348
SIR_2	Bias	-0.0385	-0.0002	-0.002	0.0043
	SD	0.0915	0.0767	0.1011	0.077
	MSE	0.00985	0.00588	0.0102	0.00595
SIR_5	Bias	-0.0391	-0.004	-0.0017	-0.0002
	SD	0.0731	0.0731	0.0974	0.0707
	MSE	0.00688	0.00536	0.00949	0.005
SIR_{10}	Bias	-0.0425	-0.0022	-0.0049	0.0001
	SD	0.086	0.0815	0.1021	0.0808
	MSE	0.0092	0.00665	0.0105	0.00653
SIR_{20}	Bias	-0.04	-0.0068	-0.0062	-0.0078
	SD	0.0853	0.0887	0.0993	0.0885
	MSE	0.00887	0.00791	0.0099	0.00789
given	Bias	0.0006	0.0006		
	SD	0.0571	0.0571		
	MSE	0.00326	0.00326		

of which dimension reduction method has been employed. Iteration helps the estimator without an initial estimate of $\boldsymbol{\theta}$ much more than the one with an initial estimate. This suggests also that further iteration will not improve the estimation of $\boldsymbol{\theta}$. We also compared the performance of the dimension reduction estimators in Section 2.2, but due to space limitation, the results are not reported here. Details of additional simulations can be found in Chong (1999).

In both simulations we tried to compare our approach with the GPLSIM algorithm in Carroll et. al (1997) but were unable to obtain any meaningful results for their procedure due to computational difficulties, triggered possibly by the relatively high dimension of $\boldsymbol{X}$. The minimization in the GPLSIM is now for a five-dimensional $\boldsymbol{\beta}$ and a scalar θ, whereas it is for a three-dimensional $\boldsymbol{\beta}$ and a scalar θ in the simulation model (7) in that paper. We thus instead adopt the simulation model presented in that article. The simulation has $n = 200$ with $N = 100$ simulations based on the model

$$Y_i = \sin(\pi \frac{\boldsymbol{\beta}^T \boldsymbol{X}_i - A}{B - A}) + \theta Z_i + \epsilon_i, \tag{7}$$

with $A = \sqrt{3}/2 - 1.645/\sqrt{12}$, $B = \sqrt{3}/2 + 1.645/\sqrt{12}$, $\boldsymbol{X}_i$ distributed as a uniform variable on the cube $[0,1]^3$, $Z_i = 0$ for i odd and $Z_i = 1$ for i even, and $\epsilon_i \sim N(0, \sigma^2 = 0.01)$. The parameters are $\boldsymbol{\beta} = (1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3})^T$ and $\theta = 0.3$.

Since the design is nearly symmetric, we do not use the **SIR** procedure here (Li (1991)) and only **PPR** was employed to estimate $\boldsymbol{\beta}$. Again, when implementing the program for GPLSIM encountered difficulties. The initial estimates for both $\boldsymbol{\beta}$ and θ seem crucial, so we decided to use our estimates of $\boldsymbol{\beta}$ and θ in Sections 2.1 and 2.2 as the initial estimates for the GPLSIM procedure, and then iterate our procedure once to make both procedures comparable as GPLSIM utilizes the same initial estimates. We used only procedure 1 in this simulation since the results in Tables 1 and 2 show no benefit using the initial estimator for $\boldsymbol{\theta}$ in procedure 2 if we iterate once for both estimates of $\boldsymbol{\beta}$ and θ using our procedures. The results for the three procedures, (a) our estimator in Section 2.1, (b) GPLSIM using our estimator as the initial estimator in its iterated algorithm, and (c) one iteration of our procedure as described in Section 2.3, are reported in the last three rows of Table 3. For comparison with the optimal procedure, we also include in the first row the nonlinear least squares procedure available in S-PLUS, assuming that everything is known except $\boldsymbol{\beta}$ and θ. Relative efficiencies with respect to this optimal procedure are reported for all three procedures in the last column of Table 3. To save computing time, we used the same bandwidth (based on generalized cross -validation) for the iteration as for the first partial regression step. The GPLSIM procedure uses a plug-in method for estimating the bandwidth. The results in Table 3 suggest that the iterated **PPR** estimators outperform those from GPLSIM slightly. Note that both approaches utilize the PPR estimators in the second row

as initial estimators. Our procedures are computationally much more stable and simpler than GPLSIM.

TABLE 3
Comparison of our procedures with GPLSIM using the model in Carroll et. al (1997), where the initial estimators of GLPSIM are our from our Procedure 1 in Section 2.1

Estimate of θ	Mean	SD	MSE	RE
NLS	0.3007	0.0106	0.000114	1
PPR-PR	0.2945	0.0229	0.000556	4.89
PPR-PR, iterated	0.3	0.0165	0.000273	2.4
GPLSIM	0.3053	0.0165	0.000302	2.66

4. Conclusions. We have demonstrated that when $\boldsymbol{X}$ and $\boldsymbol{Z}$ are independent, the estimation of the dimension-reduction direction $\boldsymbol{\beta}$ is straightforward and much simpler algorithms than those in the literature are available. Consequently, the problem to estimate the linear parameter, $\boldsymbol{\theta}$, is equivalent to the one in partially linear model in the sense that the same efficiency can be attained as in the partial linear model which assumes a known $\boldsymbol{\beta}$. In addition, we show that the indices, $\mathbf{B}$, in the semiparametric index components can also be estimated optimally. The theoretical results presented in Theorems 2 and 3 here improve upon those in Carroll et. al (1977), where the asymptotic distributions of both estimates for $\mathbf{B}$ and $\boldsymbol{\theta}$ were derived under the additional stringent assumption that those estimators are already known to be $\sqrt{n}$ -consistent. We show that such an assumption can be dispensed with when X and Z are independent.

APPENDIX A: *Proofs.*

Without loss of generality and for simplicity, we will focus on the single-index model with $k = 1$, although the proof can be extended to multiple-indices models. We will use a vector $\boldsymbol{\beta}$ instead of the matrix $\mathbf{B}$ to describe the relationship between Y and $\boldsymbol{X}$ when $k = 1$. This is a partially linear single-index model, given by the equation

$$Y = \boldsymbol{Z}'\boldsymbol{\theta} + g(\boldsymbol{X}'\boldsymbol{\beta}) + e. \tag{8}$$

We first present the assumptions for the theorems.

1. $E(e) = 0$, $Var(e) = \sigma^2 < \infty$.
2. $E(\boldsymbol{Z}) = 0$, $E(\|\boldsymbol{Z}\|^2) < \infty$.
3. $h = \text{const} \cdot n^{-a}$, where $0 < a < \frac{1}{3}$.
4. g is twice differentiable, with the second derivative bounded and continuous.

5. The density function, $f_X : \Re^p \to \Re$, of the p-dimensional random vector $\boldsymbol{X}$ is twice differentiable with the second derivative bounded and continuous.
6. $f_{\mathbf{X}}$ is bounded away from zero.
7. K is Lipschitz continuous on the real line.
8. K has support $[-1, 1]$.
9. $K(u) \geq 0$ for all u and $\int_{-1}^{1} K(u)du = 1$.
10. $\int_{-1}^{1} uK(u)du = 0$.
11. $\int_{-1}^{1} u^2K(u)du = M_K \neq 0$.

Remark: Assumptions 1 and 2 are necessary conditions for the asymptotic normality of an estimator. Assumption 3 is commonly used in nonparametric estimation. Assumptions 4 and 5 are also common conditions. Assumption 6 and 5 imply that the distribution of $\boldsymbol{X}$ has bounded support and the $f_{\mathbf{X}}$ is bounded from above. With assumption (4), we can also conclude that g is bounded from above. These conditions are used to avoid the boundary effect when a nonparametric smoother is employed to construct an estimator of a nonparametric regression function. All conditions on the kernel function are commonly used in the literature. Therefore, the imposed conditions are mild.

For the clarity of the proof of Theorem 2 , we divide the tedious proof into five Lemmas. Suppose we observe the data $(Y_j, \boldsymbol{X}_j, \boldsymbol{Z}_j)$, $j = 1, \ldots, n$, where $\boldsymbol{X} \in \Re^p$ and $\boldsymbol{Z} \in \Re^q$. Consequently, $\boldsymbol{\beta} \in \Re^p$ and $\boldsymbol{\theta} \in \Re^q$. For $i_1 = 0, 1$ and $i_2 = 0, 1, 2$, and any $\boldsymbol{\beta}^* \in \Re^p$ define

$$\xi_j^{i_1,i_2}(\mathbf{x}, \boldsymbol{\beta}^*) = Y_j^{i_1} K_h((\boldsymbol{X}_j - \mathbf{x})'\boldsymbol{\beta}^*)((\boldsymbol{X}_j - \mathbf{x})'\boldsymbol{\beta}^*)^{i_2} \tag{9}$$

and

$$\alpha_n^{i_1,i_2}(\mathbf{x}, \boldsymbol{\beta}^*) = n^{-1} \sum_{j=1}^{n} (\xi_j^{i_1,i_2}(\mathbf{x}, \boldsymbol{\beta}^*) - E(\xi_j^{i_1,i_2}(\mathbf{x}, \boldsymbol{\beta}^*))). \tag{10}$$

LEMMA 1. *Under conditions (1)–(8) for $i_1 = 0, 1$ and $i_2 = 0, 1, 2$ and $\|\boldsymbol{\beta}^* - \boldsymbol{\beta}\| = O(n^{-\frac{1}{2}})$,*

$$\sup_{x \in \Re^p} \sup_{\beta^* : \|\beta^* - \beta\| = O(n^{-1/2})} \sqrt{nh} |\alpha_n^{i_1,i_2}(\mathbf{x}, \boldsymbol{\beta}^*) - \alpha_n^{i_1,i_2}(\mathbf{x}, \boldsymbol{\beta})| \xrightarrow{P} 0. \tag{11}$$

Proof. In order to use arguments such as those provided in the proof of Theorem II. 37 in Pollard (1984, pages 34-35), we first show that for any $\epsilon > 0$, $\mathbf{x} \in \Re^p$, $\boldsymbol{\beta}^* \in \Re^p$, and $\|\boldsymbol{\beta}^* - \boldsymbol{\beta}\| = O(n^{-\frac{1}{2}})$,

$$P\left(\sqrt{nh} |\alpha_n^{i_1,i_2}(\mathbf{x}, \boldsymbol{\beta}^*) - \alpha_n^{i_1,i_2}(\mathbf{x}, \boldsymbol{\beta})| > \frac{\epsilon}{2}\right) \leq \frac{1}{2}. \tag{12}$$

This is in preparation to apply the symmetrization approach, see, e.g. Pollard (1984,pages 14-16). The left-hand side of (12) is equal to

$$P\left(|\alpha_n^{i_1,i_2}(\mathbf{x},\boldsymbol{\beta}^*)-\alpha_n^{i_1,i_2}(\mathbf{x},\boldsymbol{\beta})|>\frac{\epsilon}{2\sqrt{nh}}\right), \tag{13}$$

which is less than or equal to

$$\frac{4nh}{\epsilon^2}E\{[\alpha_n^{i_1,i_2}(\mathbf{x},\boldsymbol{\beta}^*)-\alpha_n^{i_1,i_2}(\mathbf{x},\boldsymbol{\beta})]^2\} \tag{14}$$

by Chebychev's inequality. We now prove that this value is less than or equal to $1/2$. Recalling the definition of $\alpha_n^{i_1,i_2}$ and the independence of $\xi_j^{i_1,i_2}$, an elementary calculation yields that

$$\begin{aligned}
&\frac{4nh}{\epsilon^2}E\{[\alpha_n^{i_1,i_2}(\mathbf{x},\boldsymbol{\beta}^*)-\alpha_n^{i_1,i_2}(\mathbf{x},\boldsymbol{\beta})]^2\}\\
&=\frac{4h}{\epsilon^2}Var[\xi^{i_1,i_2}(\mathbf{x},\boldsymbol{\beta}^*)-\xi^{i_1,i_2}(\mathbf{x},\boldsymbol{\beta})]\\
&=\frac{4h}{\epsilon^2}E\{Y^{2i_1}[K_h((\boldsymbol{X}-\mathbf{x})'\boldsymbol{\beta}^*)((\boldsymbol{X}-\mathbf{x})'\boldsymbol{\beta}^*)^{i_2}-K_h((\boldsymbol{X}-\mathbf{x})'\boldsymbol{\beta})((\boldsymbol{X}-\mathbf{x})'\boldsymbol{\beta})^{i_2}]^2\}. \qquad (15)
\end{aligned}$$

If $i_1=0$, then $Y^{2i_1}=1$ Let M_{f_X} be the upper bound of $f_{\mathbf{X}}$ and $T=\{\boldsymbol{t}\in\Re^p: K((\boldsymbol{t}-\mathbf{x})'\boldsymbol{\beta}^*/h)>0$ or $K((\boldsymbol{t}-\mathbf{x})'\boldsymbol{\beta}/h)>0\}$. Then (15) is equal to

$$\begin{aligned}
&\frac{4h}{\epsilon^2}\int_{t\in T}\left[\frac{1}{h}K((\boldsymbol{t}-\mathbf{x})'\boldsymbol{\beta}^*/h)((\boldsymbol{t}-\mathbf{x})'\boldsymbol{\beta}^*)^{i_2}-\frac{1}{h}K((\boldsymbol{t}-\mathbf{x})'\boldsymbol{\beta}/h)((\boldsymbol{t}-\mathbf{x})'\boldsymbol{\beta})^{i_2}\right]^2 f_{\mathbf{X}}(\boldsymbol{t})d\boldsymbol{t}\\
&=\frac{4}{h\epsilon^2}\int_{t\in T}[K((\boldsymbol{t}-\mathbf{x})'\boldsymbol{\beta}^*/h)((\boldsymbol{t}-\mathbf{x})'\boldsymbol{\beta}^*)^{i_2}-K((\boldsymbol{t}-\mathbf{x})'\boldsymbol{\beta}/h)((\boldsymbol{t}-\mathbf{x})'\boldsymbol{\beta})^{i_2}]^2 f_{\mathbf{X}}(\boldsymbol{t})d\boldsymbol{t}\\
&=\frac{4h^{2i_2}}{h\epsilon^2}\int_{x+hu\in T}[K(\boldsymbol{u}'\boldsymbol{\beta}^*)(\boldsymbol{u}'\boldsymbol{\beta}^*)^{i_2}-K(\boldsymbol{u}'\boldsymbol{\beta})(\boldsymbol{u}'\boldsymbol{\beta})^{i_2}]^2 f_{\mathbf{X}}(\mathbf{x}+h\boldsymbol{u})d(h\boldsymbol{u}). \qquad (16)
\end{aligned}$$

Let $U=\{\boldsymbol{u}\in\Re^p:\mathbf{x}+h\boldsymbol{u}\in T\}=\{\boldsymbol{u}\in\Re^p: K(\boldsymbol{u}'\boldsymbol{\beta}^*)>0$ or $K(\boldsymbol{u}'\boldsymbol{\beta})>0\}$ be a compact set. Then (16) is equal to

$$\begin{aligned}
&\frac{4h^{2i_2}h^p}{h\epsilon^2}\int_U[K(\boldsymbol{u}'\boldsymbol{\beta}^*)(\boldsymbol{u}'\boldsymbol{\beta}^*)^{i_2}-K(\boldsymbol{u}'\boldsymbol{\beta})(\boldsymbol{u}'\boldsymbol{\beta})^{i_2}]^2 f_{\mathbf{X}}(\mathbf{x}+h\boldsymbol{u})d\boldsymbol{u}\\
&\le\frac{4h^{2i_2}h^p}{h\epsilon^2}CM_{f_X}\int_U(\boldsymbol{u}'(\boldsymbol{\beta}^*-\boldsymbol{\beta}))^2 d\boldsymbol{u}\\
&=\frac{4h^{2i_2}h^p}{h\epsilon^2}CM_{f_X}O(h^{-p})O(h^{-2}n^{-1})=\frac{h^{2i_2}}{\epsilon^2}O\left(\frac{1}{nh^3}\right).
\end{aligned}$$

If $i_1 = 1$, then (15) is equal to, together with the independence of X, Z and e,

$$\begin{aligned}
&\frac{4h}{\epsilon^2} E\left\{ Y^2 \Big[K_h((\boldsymbol{X}-\mathbf{x})'\boldsymbol{\beta}^*)((\boldsymbol{X}-\mathbf{x})'\boldsymbol{\beta}^*)^{i_2} - K_h((\boldsymbol{X}-\mathbf{x})'\boldsymbol{\beta})((\boldsymbol{X}-\mathbf{x})'\boldsymbol{\beta})^{i_2} \Big]^2 \right\} \\
&= \frac{8h}{\epsilon^2} E\left\{ g^2(\boldsymbol{X}'\boldsymbol{\beta}) \Big[K_h((\boldsymbol{X}-\mathbf{x})'\boldsymbol{\beta}^*)((\boldsymbol{X}-\mathbf{x})'\boldsymbol{\beta}^*)^{i_2} - K_h((\boldsymbol{X}-\mathbf{x})'\boldsymbol{\beta})((\boldsymbol{X}-\mathbf{x})'\boldsymbol{\beta})^{i_2} \Big]^2 \right\} \\
&\quad + \frac{8h}{\epsilon^2} \Big[E(\boldsymbol{Z}'\boldsymbol{\theta})^2 + E(e^2) \Big] \cdot \\
&\qquad E\left\{ \Big[K_h((\boldsymbol{X}-\mathbf{x})'\boldsymbol{\beta}^*)((\boldsymbol{X}-\mathbf{x})'\boldsymbol{\beta}^*)^{i_2} - K_h((\boldsymbol{X}-\mathbf{x})'\boldsymbol{\beta})((\boldsymbol{X}-\mathbf{x})'\boldsymbol{\beta})^{i_2} \Big]^2 \right\}.
\end{aligned} \tag{17}$$

The second term of (17) has the same order as (15) because $E(\boldsymbol{Z}'\boldsymbol{\theta})^2 + E(e^2)$ is bounded. Similar to (16), the first term of (17) can be bounded as follows:

$$\begin{aligned}
&\frac{8}{h\epsilon^2} \int_{t\in T} g^2(\boldsymbol{t}'\boldsymbol{\beta}) [K_h((\boldsymbol{t}-\mathbf{x})'\boldsymbol{\beta}^*)((\boldsymbol{t}-\mathbf{x})'\boldsymbol{\beta}^*/h)^{i_2} - K_h((\boldsymbol{t}-\mathbf{x})'\boldsymbol{\beta}/h)((\boldsymbol{t}-\mathbf{x})'\boldsymbol{\beta})^{i_2}]^2 \times \\
&\qquad f_{\mathbf{X}}(\boldsymbol{t}) d\boldsymbol{t} \\
&= \frac{h^{2i_2}}{\epsilon^2} O\left(\frac{1}{nh^3} . \right)
\end{aligned} \tag{18}$$

Therefore, the left-hand side of (12) has order $O(1/(nh^3\epsilon^2))$, so that it is less than or equal to $\frac{1}{2}$ when n is large enough. Thus (12) is proved. This inequality ensures the use of symmetrization arguments. The next step is to show that conclusion (11) is the maximum value of an empirical process indexed by a VC class of functions. Hereafter, we will suppress the i_1, i_2 superscripts.

Let $\mathcal{F}_n = \{ f_{n,x,\beta^*}(\cdot,\cdot) : \|\mathbf{x}\| \le C \text{ and } \|\boldsymbol{\beta}^*\| \le A \}$ be a class of functions indexed by $\mathbf{x}$ and $\boldsymbol{\beta}^*$ consisting of

$$f^{i_1,i_2}_{n,x,\beta^*}(y,\boldsymbol{t}) = y^{i_1} [K((\boldsymbol{t}-\mathbf{x})'\boldsymbol{\beta}^*/h)((\boldsymbol{t}-\mathbf{x})'\boldsymbol{\beta}^*)^{i_2} - K((\boldsymbol{t}-\mathbf{x})'\boldsymbol{\beta}/h)((\boldsymbol{t}-\mathbf{x})'\boldsymbol{\beta})^{i_2}].$$

Therefore, the left-hand side of (11) is equal to $\sqrt{nh} \sup_{f\in\mathcal{F}_n} \big| \sum_{j=1}^n f(Y_j, \boldsymbol{X}_j) \big|$. Note that

$$f^{i_1,i_2}_{n,x,\beta^*}(Y_j, \boldsymbol{X}_j) = h \cdot (\xi_j^{i_1,i_2}(\mathbf{x}, \boldsymbol{\beta}^*) - \xi_j^{i_1,i_2}(\mathbf{x}, \boldsymbol{\beta}))$$

and

$$A = \|\boldsymbol{\beta}\| + O(n^{-1/2}) \text{ since } \|\boldsymbol{\beta}^* - \boldsymbol{\beta}\| = O(n^{-1/2}).$$

For any fixed n, let $\epsilon_n = \frac{\epsilon}{8}\sqrt{\frac{h}{n}}$. Suppose there is a set F° consisting of functions $\{f_1^\circ, \ldots, f_m^\circ\}$, each in $\mathcal{F}_n$, such that

$$\min_{i\in 1,\ldots,m} n^{-1} \sum_{j=1}^n |f(Y_j, \boldsymbol{X}_j) - f_i^\circ(Y_j, \boldsymbol{X}_j)| < \epsilon_n \text{ for every } f \in \mathcal{F}_n. \tag{19}$$

Let $N_1(\epsilon_n, P_n, \mathcal{F}_n)$ be the minimum m for all sets that satisfy (19). Let f^* denote the function out of $(f_1^\circ, \ldots, f_{N_1}^\circ)$ that achieves the minimum in the expression (19). We now show that $\mathcal{F}_n$ is a VC class of functions, that is, for any n, and some w such that

$$N_1(\epsilon_n, P_n, \mathcal{F}_n) \leq (const.) \cdot n^w.$$

For each set satisfying (19) and for each f_i° there is a pair $(\boldsymbol{s}_i, \boldsymbol{\beta}_i)$ such that $f_i^\circ(y, \boldsymbol{t}) \equiv f_{n,s_i,\beta_i}(y, \boldsymbol{t})$. Then

$$\begin{aligned}
&|f_{n,x,\beta^*}^{i_1,i_2}(Y, \boldsymbol{X}) - f_{n,s_i,\beta_i}^{i_1,i_2}(Y, \boldsymbol{X})| \\
&= |Y^{i_1}[K((\boldsymbol{X}-\mathbf{x})'\boldsymbol{\beta}^*/h)((\boldsymbol{X}-\mathbf{x})'\boldsymbol{\beta}^*)^{i_2} - K((\boldsymbol{X}-\mathbf{x})'\boldsymbol{\beta}/h)((\boldsymbol{X}-\mathbf{x})'\boldsymbol{\beta})^{i_2}] \\
&\quad - Y^{i_1}[K((\boldsymbol{X}-\boldsymbol{s}_i)'\boldsymbol{\beta}_i/h)((\boldsymbol{X}-\boldsymbol{s}_i)'\boldsymbol{\beta}_i)^{i_2} - K((\boldsymbol{X}-\boldsymbol{s}_i)'\boldsymbol{\beta}/h)((\boldsymbol{X}-\boldsymbol{s}_i)'\boldsymbol{\beta})^{i_2}]| \\
&= |Y^{i_1}[K((\boldsymbol{X}-\mathbf{x})'\boldsymbol{\beta}^*/h)((\boldsymbol{X}-\mathbf{x})'\boldsymbol{\beta}^*)^{i_2} - K((\boldsymbol{X}-\boldsymbol{s}_i)'\boldsymbol{\beta}_i/h)((\boldsymbol{X}-\boldsymbol{s}_i)'\boldsymbol{\beta}_i)^{i_2} \\
&\quad + K((\boldsymbol{X}-\boldsymbol{s}_i)'\boldsymbol{\beta}/h)((\boldsymbol{X}-\boldsymbol{s}_i)'\boldsymbol{\beta})^{i_2} - K((\boldsymbol{X}-\mathbf{x})'\boldsymbol{\beta}/h)((\boldsymbol{X}-\mathbf{x})'\boldsymbol{\beta})^{i_2}]| \\
&\leq \frac{|Y^{i_1}|h^{i_2}}{h}(\text{const})(|(\boldsymbol{X}-\mathbf{x})'\boldsymbol{\beta}^* - (\boldsymbol{X}-\boldsymbol{s}_i)'\boldsymbol{\beta}_i| + |(\boldsymbol{X}-\boldsymbol{s}_i)'\boldsymbol{\beta} - (\boldsymbol{X}-\mathbf{x})'\boldsymbol{\beta}|) \\
&= \frac{|Y^{i_1}|h^{i_2}}{h}(\text{const})(|\boldsymbol{X}'(\boldsymbol{\beta}^* - \boldsymbol{\beta}_i) - \mathbf{x}'\boldsymbol{\beta}^* + \boldsymbol{s}_i'\boldsymbol{\beta}_i| + |(\mathbf{x}-\boldsymbol{s}_i)'\boldsymbol{\beta}|) \\
&\leq \frac{|Y^{i_1}|h^{i_2}}{h}(\text{const})(|\boldsymbol{X}'(\boldsymbol{\beta}^* - \boldsymbol{\beta}_i)| + |(\boldsymbol{s}_i-\mathbf{x})'\boldsymbol{\beta}^*| + |\boldsymbol{s}_i'(\boldsymbol{\beta}_i - \boldsymbol{\beta}^*)| + |(\mathbf{x}-\boldsymbol{s}_i)'\boldsymbol{\beta}|) \\
&\leq \frac{|Y^{i_1}|h^{i_2}}{h}(\text{const})(|\boldsymbol{X}'(\boldsymbol{\beta}^* - \boldsymbol{\beta}_i)| + \|\boldsymbol{s}_i - \mathbf{x}\|A + C\|\boldsymbol{\beta}_i - \boldsymbol{\beta}^*\| + \|\mathbf{x}-\boldsymbol{s}_i\|\|\boldsymbol{\beta}\|).
\end{aligned}$$

K has a compact support, so for large n, $n^{-1}\sum_{j=1}^n |Y_j^{i_1}|\|\boldsymbol{X}_j\|$ and $n^{-1}\sum_{j=1}^n |Y_j^{i_1}|$ are bounded by a constant with probability one. For all $\mathbf{x}$ with $\|\mathbf{x}\| < C$ and all $\boldsymbol{\beta}^*$ with $\|\boldsymbol{\beta}^*\| < A$,

$$n^{-1}\sum_{j=1}^n |f_{n,x,\beta^*}^{i_1,i_2}(Y_j, \boldsymbol{X}_j) - f_{n,s_i,\beta_i}^{i_1,i_2}(Y_j, \boldsymbol{X}_j)| \leq (\text{const})\frac{h^{i_2}}{h}(\|\boldsymbol{\beta}^* - \boldsymbol{\beta}_i\| + \|\boldsymbol{s}_i - \mathbf{x}\|). \tag{20}$$

That is, for any two functions in $\mathcal{F}_n$, the distance only relates to the distances of β and x and h^{i_2}/h. Let $\epsilon_n = \frac{\epsilon}{8}\sqrt{\frac{h}{n}}$.

$$N_1(\epsilon_n, P_n, \mathcal{F}_n) \leq (\text{const}) \cdot \frac{1}{\epsilon^2}\left(\frac{n}{h^3}\right)^p = (\text{const}) \cdot n^w,$$

where we let $w = p(1+3a)$. Let the constant be $M/2$. This means that $\mathcal{F}_n$ is a VC-class of functions.

Invoking similar arguments that are used to prove Theorem II.37 (Pollard,1984, p.35) or those

of Zhu (1993), we have

$$\begin{aligned}
& P(\sup_x \sup_{\beta^*} \sqrt{nh}|\alpha_n(\mathbf{x},\boldsymbol{\beta}^*) - \alpha_n(\mathbf{x},\boldsymbol{\beta})| > \epsilon) \\
\leq\ & Mn^w E\left(\max_{i\in\{1,\ldots,N_1\}} \exp\left(\frac{-\frac{n}{2}(\frac{\epsilon^2}{8^2}\cdot\frac{h}{n})}{n^{-1}\sum_{j=1}^n (f_{n,s_i,\beta_i}(Y_j,\boldsymbol{X}_j))^2} \right) \right) \\
=\ & Mn^w E\left(\max_{i\in\{1,\ldots,N_1\}} \exp\left(\frac{-\epsilon^2/128}{n^{-1}h^{-1}\sum_{j=1}^n (f_{n,s_i,\beta_i}(Y_j,\boldsymbol{X}_j))^2} \right) \right) \\
\leq\ & Mn^w \left(\sup_x \sup_{\beta^*} \exp\left(\frac{-\epsilon^2/128}{n^{-1}h\sum_{j=1}^n (\xi_j(\mathbf{x},\boldsymbol{\beta}^*) - \xi_j(\mathbf{x},\boldsymbol{\beta}))^2} \right) \right) \\
=\ & Mn^w \exp\left(\frac{-\epsilon^2/128}{O(1/nh^3)} \right) \to 0.
\end{aligned}$$

because $n^{-1}h\sum_{j=1}^n (\xi_j(\mathbf{x},\boldsymbol{\beta}^*) - \xi_j(\mathbf{x},\boldsymbol{\beta}))^2$ has the same order as $hE((\xi(\mathbf{x},\boldsymbol{\beta}^*) - \xi(\mathbf{x},\boldsymbol{\beta}))^2)$, which has the same order as (15). Therefore, $P(\sup_x \sup_{\beta^*} \sqrt{nh}|\alpha_n(\mathbf{x},\boldsymbol{\beta}^*) - \alpha_n(\mathbf{x},\boldsymbol{\beta})| > \epsilon) \to 0$, for any given ϵ and relation (11) holds. □

LEMMA 2. *Under conditions (1)–(8), for $i_1 = 0, 1$ and $i_2 = 0, 1, 2$ and $\|\boldsymbol{\beta}^* - \boldsymbol{\beta}\| = O(n^{-\frac{1}{2}})$,*

$$\begin{aligned}
&\sup_{x\in\Re^p} \sup_{\beta^*:\|\beta^*-\beta\|=O(n^{-1/2})} |n^{-1}\sum_{j=1}^n [Y_j^{i_1} K_h((\boldsymbol{X}_j - \mathbf{x})'\boldsymbol{\beta}^*)((\boldsymbol{X}_j - \mathbf{x})'\boldsymbol{\beta}^*)^{i_2} \\
&\quad - E(Y_j^{i_1} K_h((\boldsymbol{X}_j - \mathbf{x})'\boldsymbol{\beta}^*)((\boldsymbol{X}_j - \mathbf{x})'\boldsymbol{\beta}^*)^{i_2})]| = O_P(1/\sqrt{nh}). \qquad (21)
\end{aligned}$$

Proof. Similar arguments as used in the proof of Lemma 1 apply. Readers are referred to Chong (1999) for details. □

LEMMA 3. *Under conditions (1)–(8), for $i_2 = 0, 1, 2$ and $\|\boldsymbol{\beta}^* - \boldsymbol{\beta}\| = O(n^{-\frac{1}{2}})$,*

$$\begin{aligned}
&\sup_{x\in\Re^p} \sup_{\beta^*:\|\beta^*-\beta\|=O(n^{-1/2})} |n^{-1}\sum_{j=1}^n \boldsymbol{Z}_j'(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) K_h((\boldsymbol{X}_j - \mathbf{x})'\boldsymbol{\beta}^*)((\boldsymbol{X}_j - \mathbf{x})'\boldsymbol{\beta}^*)^{i_2} \\
&\quad = O_P(1/(n\sqrt{h})). \qquad (22)
\end{aligned}$$

Proof. Note that $\boldsymbol{Z}_j'(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})=\sum_{i=1}^q Z_{ji}(\hat{\theta}_i-\theta_i)$. The left-hand side of (22) is equal to

$$\begin{aligned}
&\sup_x \sup_{\beta^*} |n^{-1}\sum_{j=1}^n [\sum_{i=1}^q Z_{ji}(\hat{\theta}_i-\theta_i)] K_h((\boldsymbol{X}_j-\mathbf{x})'\boldsymbol{\beta}^*)((\boldsymbol{X}_j-\mathbf{x})'\boldsymbol{\beta}^*)^{i_2}| \\
&\leq \sum_{i=1}^q \sup_x \sup_{\beta^*} |n^{-1}\sum_{j=1}^n Z_{ji}(\hat{\theta}_i-\theta_i) K_h((\boldsymbol{X}_j-\mathbf{x})'\boldsymbol{\beta}^*)((\boldsymbol{X}_j-\mathbf{x})'\boldsymbol{\beta}^*)^{i_2}| \\
&\leq \sum_{i=1}^q |\hat{\theta}_i-\theta_i| \sup_x \sup_{\beta^*} |n^{-1}\sum_{j=1}^n Z_{ji} K_h((\boldsymbol{X}_j-\mathbf{x})'\boldsymbol{\beta}^*)((\boldsymbol{X}_j-\mathbf{x})'\boldsymbol{\beta}^*)^{i_2}|. \qquad (23)
\end{aligned}$$

Since $q \ll n$, the order of (23) is the same as the order of a single term of the first summation. Without loss of generality, then, we may take $q=1$. We have $|\hat{\theta}-\theta|=O_P(n^{-1/2})$ because $\hat{\theta}$ is obtained through a least squares regression of Y and Z, so we want to show that

$$\sup_x \sup_{\beta^*} |n^{-1}\sum_{j=1}^n Z_j K_h((\boldsymbol{X}_j-\mathbf{x})'\boldsymbol{\beta}^*)((\boldsymbol{X}_j-\mathbf{x})'\boldsymbol{\beta}^*)^{i_2}| = O_P(1/\sqrt{nh}). \qquad (24)$$

Note that Z_j's are independent of the X_j's. Similar arguments as in the proof of Lemma 1 can be applied again. For details see Chong (1999). □

LEMMA 4. *Under conditions (4)–(11), letting* $E_{i_1,i_2}=E(Y^{i_1}K_h((\boldsymbol{X}-\mathbf{x})'\boldsymbol{\beta})((\boldsymbol{X}-\mathbf{x})'\boldsymbol{\beta})^{i_2})$, *we have*

$$E_{i_1,i_2}=\begin{cases} f(\mathbf{x}'\boldsymbol{\beta})+O(h^2), & i_1=0,i_2=0 \\ \Big(E(Z'\theta)+g(\mathbf{x}'\boldsymbol{\beta})\Big)f(\mathbf{x}'\boldsymbol{\beta})+O(h^2), & i_1=1,i_2=0 \\ O(h^2), & i_1=0,i_2=1 \\ O(h^2), & i_1=1,i_2=1 \\ O(h^2), & i_1=0,i_2=2 \end{cases} \qquad (25)$$

Also, uniformly over $x\in\Re^p$, $\|\beta^*-\beta\|=O(n^{-1/2})$,

$$\begin{aligned}
|E(K_h((\boldsymbol{X}-\mathbf{x})'\boldsymbol{\beta}^*))-f(\mathbf{x}'\boldsymbol{\beta})| &= O(h^2+n^{-1/2}), \\
|E(YK_h((\boldsymbol{X}-\mathbf{x})'\boldsymbol{\beta}^*))-\Big(E(Z'\theta)+g(\mathbf{x}'\boldsymbol{\beta})\Big)f(\mathbf{x}'\boldsymbol{\beta})| &= O(h^2+n^{-1/2}), \\
|E(K_h((\boldsymbol{X}-\mathbf{x})'\boldsymbol{\beta}^*)(\boldsymbol{X}-\mathbf{x})'\boldsymbol{\beta}^*)| &= O(h^2), \\
|E(YK_h((\boldsymbol{X}-\mathbf{x})'\boldsymbol{\beta}^*)(\boldsymbol{X}-\mathbf{x})'\boldsymbol{\beta}^*)| &= O(h^2), \text{ and} \\
|E(K_h((\boldsymbol{X}-\mathbf{x})'\boldsymbol{\beta}^*)((\boldsymbol{X}-\mathbf{x})'\boldsymbol{\beta}^*)^2)| &= O(h^2). \qquad (26)
\end{aligned}$$

Proof. Recall that f is the density function of $\mathbf{X}'\boldsymbol{\beta}$. By our assumptions, f and g and their first two derivatives are bounded and continuous. We will use M_f, $M_{f'}$, $M_{f''}$, M_g, $M_{g'}$, and $M_{g''}$ to denote the bounds. We will use the Taylor expansions

$$f(\mathbf{x}'\boldsymbol{\beta}+a)=f(\mathbf{x}'\boldsymbol{\beta})+af'(v_1), \qquad (27)$$

with v_1 between $\mathbf{x}'\boldsymbol{\beta}$ and $\mathbf{x}'\boldsymbol{\beta}+a$,

$$f(\mathbf{x}'\boldsymbol{\beta}+a)=f(\mathbf{x}'\boldsymbol{\beta})+af'(\mathbf{x}'\boldsymbol{\beta})+\frac{a^2}{2}f''(v_2), \tag{28}$$

with v_2 between $\mathbf{x}'\boldsymbol{\beta}$ and $\mathbf{x}'\boldsymbol{\beta}+a$,

$$g(\mathbf{x}'\boldsymbol{\beta}+a)f(\mathbf{x}'\boldsymbol{\beta}+a)=g(\mathbf{x}'\boldsymbol{\beta})f(\mathbf{x}'\boldsymbol{\beta})+a[g'(v_3)f(v_3)+g(v_3)f'(v_3)], \tag{29}$$

with v_3 between $\mathbf{x}'\boldsymbol{\beta}$ and $\mathbf{x}'\boldsymbol{\beta}+a$, and

$$\begin{aligned} g(\mathbf{x}'\boldsymbol{\beta}+a)f(\mathbf{x}'\boldsymbol{\beta}+a) &= g(\mathbf{x}'\boldsymbol{\beta})f(\mathbf{x}'\boldsymbol{\beta})+a[g'(\mathbf{x}'\boldsymbol{\beta})f(\mathbf{x}'\boldsymbol{\beta})+g(\mathbf{x}'\boldsymbol{\beta})f'(\mathbf{x}'\boldsymbol{\beta})] \\ &\quad +\frac{a^2}{2}[g''(v_4)f(v_4)+g'(v_4)f'(v_4)+g(v_4)f''(v_4)], \end{aligned} \tag{30}$$

with v_4 between $\mathbf{x}'\boldsymbol{\beta}$ and $\mathbf{x}'\boldsymbol{\beta}+a$. Here, we only present the calculation for the case with $i_1=0$ and $i_2=0$, the others can be computed similarly. It is clear that

$$\begin{aligned} &E(K_h((\boldsymbol{X}-\mathbf{x})'\boldsymbol{\beta}))=E(h^{-1}K((\boldsymbol{X}-\mathbf{x})'\boldsymbol{\beta}/h)) \\ &= h^{-1}\int_{x'\beta-h}^{x'\beta+h}K((t-\mathbf{x}'\boldsymbol{\beta})/h)f(t)dt \\ &= \int_{-1}^{1}K(u)f(hu+\mathbf{x}'\boldsymbol{\beta})du \\ &= \int_{-1}^{1}K(u)(f(\mathbf{x}'\boldsymbol{\beta})+huf'(\mathbf{x}'\boldsymbol{\beta})+\frac{h^2u^2}{2}f''(v_2(u)))du \\ &= f(\mathbf{x}'\boldsymbol{\beta})\int_{-1}^{1}K(u)du+hf'(\mathbf{x}'\boldsymbol{\beta})\int_{-1}^{1}uK(u)du+\frac{h^2}{2}\int_{-1}^{1}f''(v_2(u))u^2K(u)du \\ &= f(\mathbf{x}'\boldsymbol{\beta})+O(h^2), \end{aligned}$$

because

$$|\frac{h^2}{2}\int_{-1}^{1}f''(v_2(u))u^2K(u)du|\le\frac{h^2}{2}\int_{-1}^{1}M_{f''}u^2K(u)du\le\frac{h^2}{2}M_{f''}M_K=O(h^2).$$

□

Now define $\hat{g}(\mathbf{x}'\boldsymbol{\beta}^*)=\mathbf{S}_{\mathbf{X}'\boldsymbol{\beta}^*}\boldsymbol{Y}$, where the subscript on $\mathbf{S}$ denotes the variables, $\mathbf{x}_1'\boldsymbol{\beta}^*,\cdots,\mathbf{x}_1'\boldsymbol{\beta}^*$, on which the smoothing matrix is based as defined in Appendix B.

LEMMA 5. *Under conditions (1)–(11),*

$$\sup_{x\in\Re^p}\sup_{\beta^*:\|\beta^*-\beta\|=O(n^{-1/2})}|\hat{g}(\mathbf{x}'\boldsymbol{\beta}^*)-E(Z'\boldsymbol{\theta})-g(\mathbf{x}'\boldsymbol{\beta})|=O_P\left(\frac{1}{\sqrt{nh}}+h^2\right). \tag{31}$$

Proof. Let $S_{n,i_2}(\mathbf{x},\boldsymbol{\beta}^*) = n^{-1}\sum_{j=1}^n K_h((\boldsymbol{X}_j-\mathbf{x})'\boldsymbol{\beta}^*)((\boldsymbol{X}_j-\mathbf{x})'\boldsymbol{\beta}^*)^{i_2}$. Then

$$\begin{aligned} S_{n,0}(\mathbf{x},\boldsymbol{\beta}^*) &= \alpha_n^{0,0}(\mathbf{x},\boldsymbol{\beta}^*) + E(K_h((\boldsymbol{X}-\mathbf{x})'\boldsymbol{\beta}^*)) \\ &= O_P(1/\sqrt{nh}) + f(\mathbf{x}'\boldsymbol{\beta}) + O(h^2+1/\sqrt{n}), \end{aligned} \tag{32}$$

$$\begin{aligned} S_{n,1}(\mathbf{x},\boldsymbol{\beta}^*) &= \alpha_n^{0,1}(\mathbf{x},\boldsymbol{\beta}^*) + E(K_h((\boldsymbol{X}-\mathbf{x})'\boldsymbol{\beta}^*)(\boldsymbol{X}-\mathbf{x})'\boldsymbol{\beta}^*) \\ &= O_P(1/\sqrt{nh}) + O(h^2), \end{aligned} \tag{33}$$

and

$$\begin{aligned} S_{n,2}(\mathbf{x},\boldsymbol{\beta}^*) &= \alpha_n^{0,2}(\mathbf{x},\boldsymbol{\beta}^*) + E(K_h((\boldsymbol{X}-\mathbf{x})'\boldsymbol{\beta}^*)((\boldsymbol{X}-\mathbf{x})'\boldsymbol{\beta}^*)^2) \\ &= O_P(1/\sqrt{nh}) + O(h^2), \end{aligned} \tag{34}$$

by Lemmas 2 and 4. Also,

$$\begin{aligned} & n^{-1}\sum_{j=1}^n \xi_j^{1,0}(\mathbf{x},\boldsymbol{\beta}^*) \\ =\ & \alpha_n^{1,0}(\mathbf{x},\boldsymbol{\beta}^*) + E(YK((\boldsymbol{X}-\mathbf{x})'\boldsymbol{\beta}^*)) \\ =\ & O_P(1/\sqrt{nh}) + \Big(E(Z'\beta) + g(\mathbf{x}'\boldsymbol{\beta})\Big)f(\mathbf{x}'\boldsymbol{\beta}) + O(h^2+\sqrt{n}), \end{aligned} \tag{35}$$

and

$$\begin{aligned} n^{-1}\sum_{j=1}^n \xi_j^{1,1}(\mathbf{x},\boldsymbol{\beta}^*) &= \alpha_n^{1,1}(\mathbf{x},\boldsymbol{\beta}^*) + E(YK((\boldsymbol{X}-\mathbf{x})'\boldsymbol{\beta}^*)(\boldsymbol{X}_j-\mathbf{x})'\boldsymbol{\beta}^*) \\ &= O_P(1/\sqrt{nh}) + O(h^2), \end{aligned} \tag{36}$$

also by Lemmas 2 and 4. The bounds for the above five equations are uniform over $\mathbf{x}$ and $\boldsymbol{\beta}^*$. We have

$$\hat{g}(\mathbf{x}'\beta^*) = \frac{n^{-1}\sum_{i=1}^n \xi_i^{1,0}(\mathbf{x},\boldsymbol{\beta}^*)S_{n,2}(\mathbf{x},\boldsymbol{\beta}^*) - n^{-1}\sum_{i=1}^n \xi_i^{1,1}(\mathbf{x},\boldsymbol{\beta}^*)S_{n,1}(\mathbf{x},\boldsymbol{\beta}^*)}{S_{n,0}(\mathbf{x},\boldsymbol{\beta}^*)S_{n,2}(\mathbf{x},\boldsymbol{\beta}^*) - S_{n,1}^2(\mathbf{x},\boldsymbol{\beta}^*)}, \tag{37}$$

so

$$\begin{aligned} & \hat{g}(\mathbf{x}'\boldsymbol{\beta}^*) - E(Z'\boldsymbol{\theta}) - g(\mathbf{x}'\boldsymbol{\beta}) \\ =\ & \frac{(\Big(E(Z'\boldsymbol{\theta}) + g(\mathbf{x}'\boldsymbol{\beta})\Big)f(\mathbf{x}'\boldsymbol{\beta}) + O_P(h^2+\frac{1}{\sqrt{nh}}))S_{n,2}(\mathbf{x},\boldsymbol{\beta}^*) - O_P(h^2+\frac{1}{\sqrt{nh}})S_{n,1}(\mathbf{x},\boldsymbol{\beta}^*)}{(f(\mathbf{x}'\boldsymbol{\beta}) + O_P(h^2+\frac{1}{\sqrt{nh}}))S_{n,2}(\mathbf{x},\boldsymbol{\beta}^*) - S_{n,1}^2(\mathbf{x},\boldsymbol{\beta}^*)} \\ & - \frac{\Big(E(Z'\boldsymbol{\theta}) + g(\mathbf{x}'\boldsymbol{\beta})\Big)((f(\mathbf{x}'\boldsymbol{\beta}) + O_P(h^2+\frac{1}{\sqrt{nh}}))S_{n,2}(\mathbf{x},\boldsymbol{\beta}^*) - S_{n,1}^2(\mathbf{x},\boldsymbol{\beta}^*)}{(f(\mathbf{x}'\boldsymbol{\beta}) + O_P(h^2+\frac{1}{\sqrt{nh}}))S_{n,2}(\mathbf{x},\boldsymbol{\beta}^*) - S_{n,1}^2(\mathbf{x},\boldsymbol{\beta}^*)} \\ =\ & \frac{(O_P(h^2+\frac{1}{\sqrt{nh}}))^2}{O_P(h^2+\frac{1}{\sqrt{nh}})} = O_P\left(h^2+\frac{1}{\sqrt{nh}}\right), \end{aligned}$$

with bounds uniform over $\mathbf{x}$ and $\boldsymbol{\beta}^*$, showing (31). □

Proof of Theorem 2. Define $\mathbf{E}_n = \widehat{Cov}(\mathbf{Z})$ the sample covariance matrix of $\mathbf{Z}$ and $\widehat{Cov}(\mathbf{Z}, Y)$ the sample covariance between $\mathbf{Z}$ and Y. Further, let

$$\mathbf{g} = \begin{bmatrix} g(\boldsymbol{X}_1'\boldsymbol{\beta}) \\ \vdots \\ g(\boldsymbol{X}_n'\boldsymbol{\beta}) \end{bmatrix}, \quad \text{and} \quad \hat{\mathbf{g}} = \begin{bmatrix} \hat{g}(\mathbf{x}_1'\hat{\boldsymbol{\beta}}) \\ \vdots \\ \hat{g}(\mathbf{x}_n'\hat{\boldsymbol{\beta}}) \end{bmatrix}.$$

Then

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= (\mathbf{E}_n)^{-1}\widehat{Cov}(\mathbf{Z}, ((\mathbf{I}-\mathbf{S})\boldsymbol{Y}) \\ &= (\mathbf{E}_n)^{-1}(\mathbf{Z}-\bar{Z})'(\boldsymbol{Y}-\bar{\boldsymbol{Y}} - (\hat{\mathbf{g}}-\mathbf{g}) \\ &= \mathbf{Z}'(\mathbf{g}+\mathbf{Z}\boldsymbol{\theta}+\mathbf{e}-\hat{\mathbf{g}}) \\ &= \boldsymbol{\theta} + (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{e} + (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{g}-\hat{\mathbf{g}}) \end{aligned} \tag{38}$$

Because $E(\mathbf{e}) = \mathbf{0}$ and e and $\boldsymbol{Z}$ are independent, $E((\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{e}) = \mathbf{0}$. It is easy to prove that, by the Weak Law of Large Numbers, $Var((\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{e}) \to \mathbf{E}_n^{-1}\sigma^2$ in probability. $n\mathbf{E}_n^{-1} \xrightarrow{P} \mathbf{E}^{-1}$, so $\sqrt{n}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{e} \xrightarrow{d} N(0, \mathbf{E}^{-1}\sigma^2)$ by the Central Limit Theorem.

Now we need to show that $\sqrt{n}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{g}-\hat{\mathbf{g}}) \xrightarrow{P} \mathbf{0}$. As stated above, $n(\mathbf{Z}'\mathbf{Z})^{-1} = n\mathbf{E}_n^{-1} \xrightarrow{P} \mathbf{E}^{-1} = O(1)$, it remains to show that $\frac{1}{\sqrt{n}}\mathbf{Z}'(\mathbf{g}-\hat{\mathbf{g}}) \xrightarrow{P} \mathbf{0}$. Towards this, we will show that

$$\sup_{\beta^*} \| \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \boldsymbol{Z}_j(\hat{g}(\boldsymbol{X}_j'\boldsymbol{\beta}^*) - g(\boldsymbol{X}_j'\boldsymbol{\beta}))\| \xrightarrow{P} 0. \tag{39}$$

Using an argument as in Lemma 3, we will let $q = 0$ without loss of generality. We will thus write Z in place of $\boldsymbol{Z}$, and the norm becomes an absolute value.

Note that Z is independent of $(\hat{g}(\boldsymbol{X}_j'\boldsymbol{\beta}^*) - g(\boldsymbol{X}_j'\boldsymbol{\beta}))$. Using arguments similar to those found in the first three lemmas, we have

$$\begin{aligned} & P(\sup_{\beta^*} |n^{-1/2} \sum_{j=1}^{n} Z_j(\hat{g}(\boldsymbol{X}_j'\boldsymbol{\beta}^*) - g(\boldsymbol{X}_j'\boldsymbol{\beta}))| > \epsilon) \\ \leq\ & 4E\left[Dn^w \sup_{\beta^*} \exp\left(\frac{-\epsilon^2/128}{n^{-1}\sum_{j=1}^{n}(Z_j(\hat{g}(\boldsymbol{X}_j'\boldsymbol{\beta}^*) - g(\boldsymbol{X}_j'\boldsymbol{\beta})))^2} \right) \right]. \end{aligned} \tag{40}$$

Lemma 5 implies that

$$\sup_{x \in \Re^p} \sup_{\beta^*: \|\beta^* - \beta\| = O(n^{-1/2})} |\hat{g}(\mathbf{x}'\boldsymbol{\beta}^*) - g(\mathbf{x}'\boldsymbol{\beta})| = O_P\left(\frac{1}{\sqrt{nh}} + h^2 \right),$$

so $n^{-1}\sum_{j=1}^{n}|Z_j|^2 = O_P(1)$, and

$$n^{-1}\sum_{j=1}^{n}|Z_j(\hat{g}(\boldsymbol{X}_j'\boldsymbol{\beta}^*) - g(\boldsymbol{X}_j'\boldsymbol{\beta}))|^2 = O_P(((1/\sqrt{nh}) + h^2)^2).$$

Therefore, the probability in (40) goes to zero, which implies (39) . The proof of Theorem 2 is now completed. □

Proof of Theorem 3: Let

$$g(u|\boldsymbol{\beta}) = E(Y - Z'\boldsymbol{\theta}|X'\boldsymbol{\beta} = u).$$

Here and below, $\boldsymbol{\beta}$ is always a unit p-vector. $g(\cdot)$ is estimated by local polynomial smoother. Let $\boldsymbol{X} = (X_1, \cdots, X_n)'$, $\boldsymbol{Y} = (Y_1, \cdots, Y_n)'$ and $\boldsymbol{Z} = (Z_1, \cdots, Z_n)'$. The estimator is defined as

$$\hat{g}(X'\boldsymbol{\beta}|\boldsymbol{\beta}) = \mathbf{S}_{X'\boldsymbol{\beta}}(\boldsymbol{Y} - \boldsymbol{Z}\hat{\boldsymbol{\theta}}),$$

where $\mathbf{S}_{X'\boldsymbol{\beta}}$ is the smoothing matrix based on the variables $X_1'\boldsymbol{\beta}, \cdots, X_n'\boldsymbol{\beta}$ similar to the situation right before Lemma 5 . For the convenience of notations, we define $\tilde{Y} = Y - Z'\boldsymbol{\theta}$ and $\tilde{g}(u|\boldsymbol{\beta}) = \mathbf{S}_{X'\boldsymbol{\beta}}\tilde{\boldsymbol{Y}}$. Since $g(u|\boldsymbol{\beta}) = g(\boldsymbol{\beta}'X)$ we may estimate $\boldsymbol{\beta}$ by selecting the orientation $\boldsymbol{\beta}^*$ which minimizes a measure of the distance $g(\cdot|\boldsymbol{\beta}^*) - g$. To this end, define

$$\hat{D}(\boldsymbol{\beta}^*, h) = \sum_{i=1}^{n}[Y_i - Z_i'\hat{\boldsymbol{\theta}} - \hat{g}(X_i'\boldsymbol{\beta}^*|\boldsymbol{\beta}^*)]^2 = (\boldsymbol{Y} - \boldsymbol{Z}\hat{\boldsymbol{\theta}})'(I - \mathbf{S}_{X'\boldsymbol{\beta}^*})'(I - \mathbf{S}_{X'\boldsymbol{\beta}^*})(\boldsymbol{Y} - \boldsymbol{Z}\hat{\boldsymbol{\theta}}).$$

Note that our initial estimator $\tilde{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is root-n consistent. Therefore, the minimization only needs to be taken over $\boldsymbol{\beta}^*$ such that $|\boldsymbol{\beta}^* - \tilde{\boldsymbol{\beta}}| = O(1/\sqrt{n})$, that is, $|\boldsymbol{\beta}^* - \boldsymbol{\beta}| = O(1/\sqrt{n})$. We then define the minimizer $\hat{\boldsymbol{\beta}}$ as the estimator of $\boldsymbol{\beta}$.

It is clear that

$$\begin{aligned}
\hat{D}(\boldsymbol{\beta}^*, h) &= \tilde{\boldsymbol{Y}}'(I - \mathbf{S}_{X'\boldsymbol{\beta}^*})'(I - \mathbf{S}_{X'\boldsymbol{\beta}^*})\tilde{\boldsymbol{Y}} \\
&+ (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})'\boldsymbol{Z}'(I - \mathbf{S}_{X'\boldsymbol{\beta}^*})'(I - \mathbf{S}_{X'\boldsymbol{\beta}^*})\boldsymbol{Z}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \\
&- (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})'\boldsymbol{Z}'(I - \mathbf{S}_{X'\boldsymbol{\beta}^*})'(I - \mathbf{S}_{X'\boldsymbol{\beta}^*})\tilde{\boldsymbol{Y}} \\
&- \tilde{\boldsymbol{Y}}'(I - \mathbf{S}_{X'\boldsymbol{\beta}^*})'(I - \mathbf{S}_{X'\boldsymbol{\beta}^*})\boldsymbol{Z}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \\
&=: \tilde{D}(\boldsymbol{\beta}^*, h) + I_{n1}(\boldsymbol{\beta}^*, h) + I_{n2}(\boldsymbol{\beta}^*, h) + I_{n3}(\boldsymbol{\beta}^*, h).
\end{aligned}$$

Invoking the arguments used to prove Theorem of Härdle, Hall and Ichimura (1993), we have

$$\tilde{D}(\boldsymbol{\beta}^*, h) = \tilde{D}(\boldsymbol{\beta}^*) + T(h) + R_1(\boldsymbol{\beta}^*, h) + R_2(h)$$

where

$$\begin{aligned}\tilde{D}(\boldsymbol{\beta}^*) &= \sum(\tilde{Y}_i - g(Z_i'\boldsymbol{\beta}^*|\boldsymbol{\beta}^*))^2 \\ T(h) &= \sum(\hat{g}(Z_i'\boldsymbol{\beta}|\boldsymbol{\beta}) - g(Z_i'\boldsymbol{\beta}))^2\end{aligned}$$

and uniformly over β^* and h such that $\|\boldsymbol{\beta}^* - \beta\| = O(1/\sqrt{n}), h = O(n^{-1/5})$

$\|R_1(\boldsymbol{\beta}^*, h)\| = o_p(n^{1/5}) \qquad \|R_2(h)\| = o_p(1).$

Furthermore, from their arguments, we have for some constants A_1 and A_2,

$$\begin{aligned}\tilde{D}(\boldsymbol{\beta}^*) &= n\Big[W^{1/2}(\boldsymbol{\beta}^* - \boldsymbol{\beta}) - n^{-1/2}(W^-)^{1/2}U_n\Big]\Big[W^{1/2}(\boldsymbol{\beta}^* - \boldsymbol{\beta}) - n^{-1/2}(W^-)^{1/2}U_n\Big] \\ &+ R_3 + R_4(\boldsymbol{\beta}^*), \\ T(h) &= A_1h^{-1} + A_2nh^4 + R_5(h)\end{aligned}$$

where

$$U_n = \sum[X_i - E(X|X_i'\boldsymbol{\beta})]g'(X_i'\boldsymbol{\beta})\varepsilon_i,$$
$$\sup_{\|\boldsymbol{\beta}^* - \boldsymbol{\beta}\| = O(1/\sqrt{n})} \|R_4(\boldsymbol{\beta}^*)\| = o_p(1), \qquad \sup_{h = O(n^{-1/5})} \|R_5(h)\| = o_p(n^{1/5}).$$

g' is the derivative of g and R_3 is a constant independent of $\boldsymbol{\beta}^*$ and h. Note that our initial estimator $\hat{\boldsymbol{\theta}}$ is root-n consistent to $\boldsymbol{\theta}$. By the independence between $\tilde{Y}$ and Z and the root-n consistency of $\tilde{\boldsymbol{\beta}}$ to $\boldsymbol{\beta}$, we obtain easily that, and uniformly over $\boldsymbol{\beta}^*$ and h such that $\|\boldsymbol{\beta}^* - \boldsymbol{\beta}\| = O(1/\sqrt{n}), h = O(n^{-1/5})$,

$$\begin{aligned}&\|I_{nl}(\beta^*, h)\| = o_p(1), \quad l = 1, 2, 3 \\ &\|\frac{1}{\sqrt{n}}\boldsymbol{Z}'(I - \mathbf{S}_{X'\boldsymbol{\beta}^*})'(I - \mathbf{S}_{X'\boldsymbol{\beta}^*})\tilde{\boldsymbol{Y}}\| = o_p(1)\end{aligned}$$

Therefore, uniformly over $\boldsymbol{\beta}^*$ and h

$$\hat{D}(\boldsymbol{\beta}^*, h) = \tilde{D}(\boldsymbol{\beta}^*) + T(h) + o_p(n^{1/5}) + C_n,$$

where C_n is a constant independent of $\boldsymbol{\beta}^*$ and h. Hence the minimum of $\hat{D}(\boldsymbol{\beta}^*, h)$ within a radius $O(n^{-1/2})$ of $\boldsymbol{\beta}$ for the first variable and on a scale of $n^{-1/5}$ for the second variable satisfies for any unit vector $u \neq \boldsymbol{\beta}$

$$\begin{aligned}u'(\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}) &= n^{-1/2}u'(W^-)U_n + o_p(n^{-1/2}) \\ &= n^{-1/2}u'(W^-)\sum[X_i - E(X|\boldsymbol{\beta}'X_i)]g'(\boldsymbol{\beta}'X_i)\varepsilon_i + o_p(n^{-1/2})\end{aligned}$$

In other words,

$$n^{1/2}u'(\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}) \Longrightarrow N(0, u'\sigma^2(W^-)u).$$

The proof is completed. □

APPENDIX B: *Linear Smoother*

We first consider the simple case of one-dimensional smoothing to estimate the mean response, $m(x) = E\{Y|X = x\}$, based on data $\{(X_i, Y_i), i = 1, \cdots, n\}$. For a given scalar point x and bandwidth h, the local polynomial smoother (Fan and Gijbels (1996)) is based on a window, $(x - h, x + h)$, and a kernel weight function to fit locally a weighted polynomial regression, and then uses the fitted value at x as the estimate for $m(x)$. For instance, a locally linear smoother with a kernel K, using a linear polynomial to estimate the regression function via the least squares method, yields the following estimate:

$$\hat{m}(x) = \arg\min_a \min_b \sum_{i=1}^{n} [y_i - a - b(x_i - x)]^2 K_h(x_i - x),$$

where $K_h(u) = h^{-1}K(u/h)$. The solution to the minimization equation is

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} K_h(x_i - x)[\sum_{j=1}^{n} K_h(x_j - x)(x_j - x)^2 - (x_i - x)\sum_{j=1}^{n} K_h(x_j - x)(x_j - x)]y_i}{\sum_{i=1}^{n} K_h(x_i - x)[\sum_{j=1}^{n} K_h(x_j - x)(x_j - x)^2 - (x_i - x)\sum_{j=1}^{n} K_h(x_j - x)(x_j - x)]}.$$

This smoother belong to a class called the class of *linear smoothers*, which is a linear combination of the observed responses. For a linear smoother, we may construct a matrix $\mathbf{S_x}$ such that the estimated mean response is $\hat{\boldsymbol{y}} = \mathbf{S_x}\boldsymbol{y}$, where the subscript $\mathbf{x}$ denotes the covariate variables, $\{x_1, \cdots, x_n\}$, on which the smoothing is based. We will call $\mathbf{S_x}$ the *smoothing matrix*. It depends on the type of smoother and kernel function K used, the observed values of the covariates $\mathbf{x}$, and the smoothing parameter h.

Suppose, for example, that $\mathbf{x} = (x_1, \ldots, x_n)$ is observed and that we are using a kernel K that has support $[-1, 1]$. The matrix $\mathbf{S}$ corresponding to the locally linear smoother above will have elements

$$S_{ij} = \frac{K_h(x_j - x_i)[\sum_{k=1}^{n} K_h(x_k - x_i)(x_k - x_i)^2 - (x_j - x_i)\sum_{k=1}^{n} K_h(x_k - x_i)(x_k - x_i)]}{\sum_{k=1}^{n} K_h(x_k - x_i)[\sum_{l=1}^{n} K_h(x_l - x_i)(x_l - x_i)^2 - (x_k - x_i)\sum_{l=1}^{n} K_h(x_l - x_i)(x_l - x_i)]}.$$

Automatic bandwidth choices based on Generalized cross validation

The bandwidth h which minimizes

$$GCV(h) = \frac{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{g}_h(X_i))^2}{(\frac{1}{n}\mathrm{tr}(\mathbf{I} - \mathbf{S}_h))^2}, \tag{41}$$

is the generalized cross-validated bandwidth, where $\mathbf{S}_h$ is the smoothing matrix corresponding to a bandwidth of h and $\hat{g}_h(X_i)$ is the estimated regression function corresponding to a bandwidth of h, evaluated at X_i.

Multivariate smoothing

For multivariate smoothing like in our setting of model (1), x will be a k-dimensional vector, the kernel will be replaced by a k-variate kernel function and the bandwidth by $(h_1, \cdots, h_k)$. The above simple linear regression based on weighted least squares fit will now be replaced by a multiple linear regression based on weighted least squares fits.

REFERENCES

BHATTACHARYA, P. K. & ZHAO, P.-L. (1997), Semiparametric inference in a partial linear model, *Ann. Statist.* **25**, 244-262.

BUJA, A., HASTIE, T., & TIBSHIRANI, R. (1989), Linear smoothers and additive models (with discussion), *Ann. Statist.* **17**, 453-555.

CARROLL, R. J., FAN, J., GIJBELS, I., & WAND, M. P. (1997), Generalized partially linear single-index models, *J. Amer. Statist. Assoc.* **92**, 477-489.

CHAUDHURI, P. & DOKSUM, K., & SAMAROV (1997), On average derivative quantile regression. *Ann. Statist.* **25**, 715-744.

CHEN, C.-H. & LI, K.-C. (1998), Can SIR be as popular as multiple linear regression?, *Statistica Sinica* **8**, 289-316.

CHEN, H. (1988), Convergence rates for parametric components in a partly linear model, *Ann. Statist.* **16** 136-146.

CHEN, H. & SHIAU, J.-J. H. (1994), Data-driven efficient estimators for a partially linear model, *Ann. Statist.* **22**, 211-237.

CHIOU, J.M. & MÜLLER, H.G. (1999), Nonparametric quasi-likelihood, *Ann. Statist.* **27**, 36-64.

CHONG, Y. S. (1999), *Dimension reduction mehtods for discrete and continuos covariates.* Unpublished Ph.D. Dissertaiton of the University of California.

COOK, R. D. (1998), Principal Hessian directions revisited (with discussion), *J. Amer. Statist. Assoc.* **93**, 84-100.

COOK, R. D. & WEISBERG, S. (1991), Discussion of "Sliced Inverse Regression," *J. Amer. Statist. Assoc.* **86**, 328-332.

CRAVEN, P. & WAHBA, G. (1979), Smoothing and noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation, *Numer. Math.* **31**, 377-403.

DABROWSKA, D. & DOKSUM (1988a), Partial likelihood in transformaiton models with censored data. *Scand. J. Statist.***15**, 1-23.

DABROWSKA, D. & DOKSUM (1988b), Estimation and testing in a two sample generalized odds rate model, *J. Amer. Statist. Assoc.* **83**, 744-749.

DENBY, L. (1986), Smooth regression functions, *Statistical Research Report 26*, AT&T Bell Laboratories, Murray Hill.

DOKSUM (1987), An extension of partial likelihood methods for proportional harzards model to general transformaiton models. *Ann. Statist.* **15**, 325-345.

DOKSUM, K & GASKO, M. (1990), On a correspondence between modles in binary regression analysis and

in survival analysis. *Internat. Statist. Rev.* **58**, 243-252.
DOKSUM, K. & SAMAROV, A. (1995), Nonparametric estimation of global functionals and a measure of the explanatory power of covariates in regression. *Ann. of Statist.* **23**, 1443-1473.
ENGLE, R., GRANGER, C., RICE, J., & WEISS, A. (1986), Semiparametric estimates of the relation between weather and electricity sales, *J. Amer. Statist. Assoc.* **81**, 310-320.
FAN, J. (1993), Local linear regression smoothers and their minimax efficiency. *Ann. Statist.* **21**, 196-216.
FAN, J. & GIJBELS, I. (1996), *Local Polynomial Modelling and Its Applications*, Chapman and Hall, London.
FRIEDMAN, J. H. & STUETZLE, W. (1981), Projection pursuit regression, *J. Amer. Statist. Assoc.* **76**, 817-823.
smoothing and cross-validation, Biometrika **72**, 527-537. HALL, P. (1989), On projection pursuit regression, *Ann. Statist.* **17**, 573-588.
HAMILTON, S. A. & TRUONG, Y. K. (1997), Local linear estimation in partly linear models, *J. Multivariate Analysis* **60**, 1-19.
HARDLE, W. & HALL, P. & ICHIMURA, H. (1993), Optimal smoothing in single-index models. *Ann. Statist.* **21**, 157-178.
HÄRDLE, W. & STOKER, T. M. (1989), Investigating smooth multiple regression by the method of average derivatives, *J. Amer. Statist. Assoc.* **84**, 986-995.
HECKMAN, N. E. (1986), Spline smoothing in a partly linear model, *J. Royal Statist. Soc. B* **48**, 244-248.
HRISTACHE, M. & JUDITSKY, A. & SPOKOINY, V. (2001), Direct estimation of the index coefficient in a single-index model. *Ann. Statist.* **29**, 595-623.
LI, K.-C. (1991), Sliced inverse regression for dimension reduction (with discussion), *J. Amer. Statist. Assoc.* **86**, 316-342.
LI, K.-C. (1992), On principal Hessian directions for data visualization and dimension reduction: another application of Stein's lemma, *J. Amer. Statist. Assoc.* **87**, 1025-1039.
POLLARD, D. (1984), *Convergence of Stochastic Processes*, Springer-Verlag, New York.
RICE, J. (1986), Convergence rates for partially splined models, *Statist. Prob. Lett.* **4**, 203-208.
SAMAROV, A. (1993), Exploring structure using nonparametric functional estimation, *J. Amer. Statist. Assoc.* **88**, 836-847.
SEVERINI, T. A., & STANISWALIS, J. G. (1994), Quasi-likelihood estimation in semiparametric models, *J. Amer. Statist. Assoc.* **89**, 501-511.
SPECKMAN, P. (1988), Kernel smoothing in partial linear models, *J. Royal Statist. Soc. B* **50**, 413-436.
STOKER, T. M. (1986), Consistent estimation of scaled coefficient, *Economoetrics.* **54**, 1461-1481.
STUTE, W. & ZHU, L.-X. (2005), Nonparametric checks for single-index models, *Annals of Statistics*, in press.
WAHBA, G. (1984), Partial spline models for the semiparametric estimation of functions of several variables, *Statistical Analyses for Time Series*, pp. 319-329, Institute of Statistical Mathematics, Tokyo.
XIA, Y.,TONG, H., LI, W. K. and ZHU, L.-X. (2002), An adaptive estimation of optimal regression subspace, *Journal of Royal Statistical Society, Series B*, **64**, 363-410.
YU, Y. & RUPPERT, D. (2002). Penalized spline estomation for partial linear single-index models. *J. Amer. Statist. Assoc.* **97**, 1042-1054.
ZHU, L.-X. & NG, K. W. (1995), Asymptotics of sliced inverse regression, *Statistica Sinica* **5**, 727-736.
ZHU, L.-X. & FANG, K. T. (1996), Asymptotics for the kernel estimates of sliced inverse regression, *Ann. Statist.* **24**, 1053-1067.

Yun Sam Chong
William E. Wecker Associates, Inc.
505 San Marin Dr.
Novato, CA 94945, USA
E-mail:chong@mail.wecker.com

Jane-Ling Wang
Department of Statistics,
University of California,
Davis, CA 95616, USA
E-mail: wang@wald.ucdavis.edu

Lixing Zhu
Depaerment of Statistics and Acruarial Science
The University of Hong Kong
Hong Kong, China
E-mail:lzhu@hku.hk