Statistica Sinica **19** (2009), 869-883

DIMENSION REDUCTION FOR CONDITIONAL VARIANCE IN REGRESSIONS

Li-Ping Zhu and Li-Xing Zhu

East China Normal University and Hong Kong Baptist University

Abstract: Both the conditional mean and variance in regressions with high dimensional predictors are of importance in modeling. In this paper, we investigate estimation of the conditional variance. To attack the *curse of dimensionality*, we introduce a notion of central variance subspace (CVS) to capture the information contained in the conditional variance. To estimate the CVS, the impact from the conditional mean needs to be fully removed. To this end, a three-step procedure is proposed: Estimating exhaustively the CMS by an outer product gradient (OPG) method; estimating consistently the structural dimension of the CMS by a modified Bayesian information criterion (BIC); and estimating the conditional mean by a kernel smoother. After removing the conditional mean from the response, we suggest a squared residuals-based OPG method to identify the CVS. The asymptotic normality of candidate matrices, and hence of corresponding eigenvalues and eigenvectors, is obtained. Illustrative examples from simulation studies and a dataset are presented to assess the finite sample performance of the theoretical results.

Key words and phrases: Asymptotic normality, central variance subspace, dimension reduction, heteroscedasticity, outer product gradient.

1. Introduction

In full generality, the goal of regression analysis is to infer the conditional distribution of the univariate response Y given the $p \times 1$ vector of predictors X. Because in many statistical applications the dimension p is large, the statistical analysis is difficult. Therefore, it is important to reduce dimension without loss of information on the conditional distribution of $Y|X$. To address this issue, Cook (1994, 1998a)) proposed the idea of *sufficient dimension reduction* (SDR). Potential advantages accrue from working in the SDR context because no prespecified model for $Y|X$ is required, and the *curse of dimensionality* that may hinder other nonparametric methods is often avoided.

The SDR is based on a population meta-parameter, the *central subspace* (CS, Cook (1996)). The CS, usually denoted by $\mathcal{S}_{Y|X}$, is defined as the intersection of all subspaces $\mathcal{S} \subseteq \mathcal{R}^p$ satisfying $Y \perp\!\!\!\perp X|P_{\mathcal{S}}X$ where "$\perp\!\!\!\perp$" indicates independence and $P_{\mathcal{S}}$ is the orthogonal projection onto $\mathcal{S}$ in the usual inner product. Methods for estimating the CS include sliced inverse regression (SIR, Li (1991))

and its variations, such as the minimum discrepancy approach (MDA, Cook and Ni (2005)), sliced average variance estimation (SAVE, Cook and Weisberg (1991)), graphical regression (Cook (1994, 1998a)), parametric inverse regression (Bura and Cook (2001)), partial SIR (Chiaromonte, Cook and Li (2002)) when categorical predictors are involved, and the contour regression method (Li, Zha and Chiaromonte (2005)).

Note that the CS usually contains some important subspaces in regression models. One is the *central mean subspace* (CMS, Cook and Li (2002)), denoted by $\mathcal{S}_{E(Y|X)}$. The CMS captures the information in the conditional mean through some combinations of X, say, $\beta^T X$, and leaves the rest of $Y|X$ as a "nuisance parameter". In the literature there are many approaches to estimating the CMS, including ordinary least squares (OLS) and related methods based on convex objective functions, principal Hessian directions (pHd, Li (1992) and Cook (1998b)), iterative Hessian transformation (IHT, Cook and Li (2002)), average derivative estimation (ADE, Härdle and Stoker (1989)), outer product gradient (OPG) estimation and minimum average variance estimation (MAVE, Xia, Tong, Li and Zhu (2002)).

Another important subspace in the CS, referred to as the *central variance subspace* (CVS), is spanned by the directions in conditional variance. Heteroscedasticity is clearly of importance in modeling and understanding the variability of statistical data. However, estimation of the CVS has not received close attention. According to the definition of the conditional variance, to infer information on it, we need to remove the impact of the conditional mean. In a dimension-reduction framework, the conditional mean $E(Y|X)$ is usually of the form $E(Y|\beta^T X)$ where β is a $p \times d_M$ orthogonal matrix. That is, the column vectors of β span the CMS. To estimate $E(Y|X)$ efficiently, we can identify β first, and then estimate $E(Y|X)$ based on $\beta^T X$. If d_M is small, then we circumvent the problem of dimensionality successfully. When d_M is given, we only need to estimate $E(Y|\beta^T X)$ in a lower dimensional space based on $\beta^T X$. When d_M is unknown, however, we must estimate it in addition to identifying the CMS before approximating the regression function.

Clearly, to estimate the CVS, it is important to obtain the residuals that do not involve any information on the CMS. Otherwise, the CVS may not be identifiable. This requires us to exhaustively identify and estimate the CMS. Then, the ADE, the pHd, and the IHT are not appropriate because of the lack of exhaustibility. Although the MAVE can identify the CMS exhaustively, it is not employed in this paper because the asymptotic normality of the relevant estimators is still an open problem. As a result, the estimator of the CVS based on MAVE cannot be shown to have the desired theoretical properties. Thus we prefer the OPG method and, in Section 3.1, we provide the asymptotic normality of the OPG estimator for the CMS that is designed for recovery of the CMS.

Exhaustibility also requires a consistent estimator of the structural dimension of the CMS when d_M, the dimension of the CMS, is unknown. We propose, in Section 3.2, a modified Bayes Information Criterion (BIC) that improves the algorithm in Zhu, Miao and Peng (2006): Our proposed BIC-type criterion can be used even when the candidate matrix is not nonnegative definite. We investigate how to choose an optimal penalty factor. The major merit of this method is that the consistency of the estimator of the relevant matrix is enough to guarantee the consistency of the estimator of dimension. This is a general method which can be applied to determine the structural dimension of other dimension reduction subspaces, including the CVS.

After an exhaustive estimate of the CMS is obtained, a nonparametric smoother is adopted to estimate $E(Y|\beta^T X)$, and then to obtain the residual $e =: Y - E(Y|\beta^T X)$. We propose, in Section 4, the e^2-based OPG method which can fully recover the CVS. Moreover, the e^2-based OPG method does not rely on the marginal distribution of predictors. These improvements significantly relax the restriction on the distribution of X and Y. Under mild conditions, asymptotic normality is established in this section for our proposed e^2-based OPG method.

In Section 5, we report on simulations that assess the performance of our methods. Horse mussel data is analyzed for illustration. The regularity conditions are listed in the Appendix. For detailed proof of all the results presented here, readers can refer to the online supplementary material attached to this paper.

2. Central Variance Subspace

For ease of illustration, we assume $\phi(X) =: E(Y|X) = E(Y|\beta^T X)$ where β is a $p \times d_M$ matrix.

Definition 2.1. Let $e = Y - \phi(X)$. If

$$e \perp\!\!\!\perp Var(Y|X)|\alpha^T X, \tag{2.1}$$

then the subspace spanned by the column vectors of α, denoted by $\mathcal{S}(\alpha)$, is a variance dimension reduction subspace for the regression of Y on X. The central variance subspace (CVS), denoted by $\mathcal{S}_{Var(Y|X)}$, is defined as the intersection of all the variance dimension reduction subspaces satisfying (2.1) provided it itself is a variance dimension reduction subspace.

Remark 1. As with the CS and the CMS, the CVS does not always exist. However, under some mild conditions, the existence and the uniqueness of the CVS can be guaranteed much as was the existence of the CS, see Cook (1998a)). We assume the existence of the CVS throughout the present context.

Parallel to Cook and Li (2002), the following theorem gives equivalent conditions for the conditional independence central to Definition 2.1.

Theorem 1. *The following statements are equivalent:*

(a) $e \perp\!\!\!\perp \mathrm{Var}(Y|X)|\alpha^T X$;

(b) $\mathrm{Cov}[e^2, \mathrm{Var}(Y|X)|\alpha^T X] = 0$;

(c) $\mathrm{Var}(Y|X)$ *is a measurable function of* $\alpha^T X$;

(d) *For any measurable function* $l(X)$ *such that* $\mathrm{Cov}[e^2, l(X)|\alpha^T X]$ *exists,* $\mathrm{Cov}[e^2, l(X)|\alpha^T X] = 0$.

Assertion (*c*) implies that $\sigma^2(X) =: \mathrm{Var}(Y|X) = E(e^2|X)$ is a function of $\alpha^T X$. It is not difficult to show that if $Z = A^T X + b$ for some invertible matrix A and some vector b, then $\mathcal{S}_{Var(Y|Z)} = A^{-1}\,\mathcal{S}_{Var(Y|X)}$ is the CVS for the regression of Y on Z. Consequently, there is no loss of generality in standardizing X to have mean 0 and identity covariance matrix. Hereafter we work with the standardized predictor X.

3. An Exhaustive Estimate of the CMS

Definition 2.1 shows that the conditional mean $E(Y|X)$ should be subtracted if the true error term is to be obtained. In order to obtain an exhaustive estimate of the CMS, and eventually to remove the effect of the conditional mean for estimating the CVS, a three-step procedure is suggested: identify the vectors of the CMS, estimate its dimension, approximate the link $\phi(.)$ in a nonparametric way.

3.1. Outer product gradient estimation and its asymptotic normality

Xia, Tong, Li and Zhu (2002) introduced an outer product gradient (OPG) method to estimate the CMS. The basic idea is to use the average of the square of the first derivative of the link function $\phi(x) = E(Y|X = x) = E(Y|\beta^T X = \beta^T x)$. To be more specific, let $\phi^{(1)}(x)$ denote the first derivative of $\phi(x)$ with respect to x. The population OPG matrix is defined as $\Delta = E[\phi^{(1)}(X)\phi^{(1)}(X)^T]$. By choosing β to be the eigenvectors corresponding to the largest d_M eigenvalues Δ, the OPG provides an exhaustive estimator of the CMS. When the *i.i.d.* observations $\{(x_i, y_i),\ i = 1, \ldots, n\}$ are available, we consider local $r-$th order polynomial fitting in the form of the minimization problem

$$\min_{a_j, b_j, c_{ji_1\cdots i_p}} \sum_{i=1}^{n} \Big[y_i - a_j - b_j^T (x_i - x_j) - \sum_{1<k\le r} \sum_{i_1+\cdots+i_p=k} c_{ji_1\cdots i_p} \{x_i - x_j\}_1^{i_1} \cdots \{x_i - x_j\}_p^{i_p} \Big]^2 K_w\Big\{\frac{(x_i - x_j)}{h_w}\Big\}, \quad (3.1)$$

where $\{x_i - x_j\}_k$ denotes the $k-$th element of vector $x_i - x_j$, and $K_w\{(x_i - x_j)/h_w\}$ is a $p-$variate kernel.

For ease of illustration, let $\{(x_i - x_j)^T_{(k)},\ i = 1, \ldots, n\}$ denote all distinct columns $\{x_i - x_j\}_1^{i_1} \cdots \{x_i - x_j\}_p^{i_p}$ satisfying $i_1 + \cdots + i_p = k$, $Y_n = (y_1, \ldots, y_n)^T$ is a vector, and W_{nj} is a diagonal matrix of weights, with entries $K_w\{(x_i - x_j)/h_w\}$. Denote by X_{ni} the predictor matrix whose (l, j)-block is $(x_l - x_i)^T_{(j)}$ for $l = 1, \ldots, n$, and $j = 0, \ldots, r$. When $j = 0$, $(x_l - x_i)_{(0)} = 1$ for all t and l. We re-organize the minimization problem (3.1) as

$$\min_{\beta_{0j}, \ldots, \beta_{rj}} \sum_{i=1}^{n} \Big[y_i - \beta_{0j} - \beta_{1j}^T (x_i - x_j)_{(1)} - \cdots - \beta_{rj}^T (x_i - x_j)_{(r)} \Big]^2 K_w\Big(\frac{x_i - x_j}{h_w}\Big). \quad (3.2)$$

Under the weighted least squares measure (3.2), we have $\widehat{\beta}_j =: (\widehat{\beta}_{0j}, \ldots, \widehat{\beta}_{rj})^T = (X_{nj}^T W_{nj} X_{nj})^{-1}\ (X_{nj}^T W_{nj} Y_n)$. Therefore, Δ_n, the estimator of Δ, is $1/n$ $\sum_{j=1}^n \widehat{\beta}_{1j} \widehat{\beta}_{1j}^T$.

To facilitate the development of our proposed methods, we first present the asymptotic normality of the OPG method. For notational clarity, we use the following notation. Let $\mu_j = \int u^j K_w(u) du$, $\nu_j = \int u^j K_w^2(u) du$, and write

$$S_r = (\mu_{i+j-2})_{1 \le i,j \le r+1},\ \tilde{S}_r = (\mu_{i+j-1})_{1 \le i,j \le r+1},$$
$$S_r^* = (\nu_{i+j-2})_{1 \le i,j \le r+1},\ \tilde{S}_r^* = (\nu_{i+j-1})_{1 \le i,j \le r+1}.$$

We introduce a block matrix $v_1 = (\mathbf{0}, I_p, \mathbf{0}, \ldots, \mathbf{0})$ with the $p \times p$ identity matrix I_p corresponding to the column indices of $\{(x_i - t)_{(1)}\}$ in X_{nt}, that is, $v_1 \beta_j = \beta_{1j}$ and $v_1 \hat{\beta}_j = \hat{\beta}_{1j}$. Let $Vech(C) = (c_{11}, \ldots, c_{p1}; c_{22}, \ldots, c_{p2}; c_{33}, \ldots, c_{pp})^T$ be a $p(p+1)/2$ dimensional column vector for a symmetric $p \times p$ matrix $C = (c_{kl})_{p \times p}$. Define $H_0 = v_1 S_r^{-1} (\mu_1, \ldots, \mu_r)^T [Y - \phi(X)] \beta^T v_1^T + v_1 \beta [Y - \phi(X)] (\mu_1, \ldots, \mu_r) (S_r^{-1})^T v_1^T$ and $V = \lambda^T \mathrm{Cov}[Vech(H_0)] \lambda$ for any $\lambda \in R^{p(p+1)/2}$.

Theorem 2. *Assume that conditions* C1−C8 *in the Appendix hold. Then*

$$\sqrt{n} h_w (\Delta_n - \Delta) \xrightarrow{d} H_0, \quad \text{as } n \to \infty, \quad (3.3)$$

where $\lambda^T Vech(H_0)$ has the normal distribution $N(0, V)$ for any $\lambda \neq 0$.

Following Zhu and Ng (2003) and Zhu and Fang (1996), we can easily derive the asymptotic normality of nonzero eigenvalues and their corresponding eigenvectors by using perturbation theory. Let $\lambda_1(A) \ge \cdots \ge \lambda_p(A) \ge 0$ and $b_i(A) = (b_{1i}(A), \ldots, b_{pi}(A))^T$, $i = 1, \ldots, p$, denote, respectively, the eigenvalues and their corresponding eigenvectors of a $p \times p$ matrix A.

Theorem 3. *In addition to the conditions of Theorem* 2, *assume that the nonzero* $\lambda_l(\Delta)'s$ *are distinct. Then for each nonzero eigenvalue* $\lambda_i(\Delta)$ *and the corresponding eigenvector* $b_i(\Delta)$, *we have, as* $n \to \infty$,

$$\begin{aligned}&\sqrt{n}h_w(\lambda_i(\Delta_n) - \lambda_i(\Delta))\\ &\quad = \sqrt{n}h_w b_i(\Delta)^T(\Delta_n - \Delta)b_i(\Delta) + o_p(\sqrt{n}h_w\|\Delta_n - \Delta\|) \xrightarrow{d} b_i(\Delta)^T H_0 b_i(\Delta),\end{aligned}$$

where H_0 *is given in Theorem* 2, *and*

$$\begin{aligned}&\sqrt{n}h_w(b_i(\Delta_n) - b_i(\Delta))\\ &\quad = \sqrt{n}h_w \sum_{l=1,l\neq i}^{p} \frac{b_i(\Delta)b_i(\Delta)^T(\Delta_n - \Delta)b_i(\Delta)}{\lambda_j(\Delta) - \lambda_l(\Delta)} + o_p(\sqrt{n}h_w\|\Delta_n - \Delta\|)\\ &\quad \xrightarrow{d} \sum_{l=1,l\neq i}^{p} \frac{b_i(\Delta)b_i(\Delta)^T H_0 b_i(\Delta)}{\lambda_j(\Delta) - \lambda_l(\Delta)},\end{aligned}$$

where $\|\Delta_n - \Delta\| = \sum_{1\le i,j\le p} |a_{ij}|$.

It is important to note that the asymptotic normality holds for $\lambda_i(\Delta) > 0$; otherwise, $\lambda_i(\Delta_n)$ converges to 0 faster than root$-(nh_w^2)$ by direct application of Theorem 3.1 in Eaton and Tyler (1991).

3.2. Determination of the structural dimension

If the structural dimension of the CMS, d_M, is unknown, its estimation is necessary. Although the sequential test has been popularly used in practice, see Li (1991) and later developments in this area, it is not consistent (Ferré (1998)). Therefore, to remove the impact of the CMS exhaustively, we must develop a consistent estimator of the structural dimension of the CMS so as to achieve a consistent estimator of the CMS.

Zhu, Miao and Peng (2006) recommended a modified BIC type algorithm, together with several choices of the penalty factor. To avoid selecting the penalty factor, Zhu and Zhu (2007) suggested choosing the penalty term simply to be of order $\log n$. Moreover, the BIC type algorithms of Zhu, Miao and Peng (2006) cannot be directly used when the candidate matrix, such as the $Y-$based pHd (Li (1992)), is not nonnegative-definite. Thus, a more general method is desired.

Because we must estimate the dimensions of both the CMS and the CVS in this context, without notational confusion we write Λ, with a sample version Λ_n, as a candidate matrix that targets Ψ, which can be the CMS or the CVS or some other subspaces. Let K_1 be the true dimension of space Ψ, which can be either d_M or d_V, and $\widehat{K_1}$ be the estimate of K_1. To tackle the problem that Λ is

not always non-negative definite, we introduce $\Omega = \Lambda^2 + I_p$ where I_p is the $p \times p$ identity matrix. Denote by Ω_n the estimate of Ω.

Recall the definition of $\lambda_i(A)$. Clearly, $\lambda_i(\Omega) = \lambda_i^2(\Lambda) + 1$. Determining the dimension of Ψ now turns to estimating K_1, the number of the eigenvalues of Ω greater than 1. Following Zhu, Miao and Peng (2006), we define the quasi log likelihood function as

$$\log L(\lambda(\Omega)) = -\frac{n}{2}\log|\Omega| - \frac{n}{2} tr\Omega^{-1}\Omega_n, \tag{3.4}$$

where $\lambda(\Omega) = (\lambda_1(\Omega), \ldots, \lambda_p(\Omega))^T$. Let Θ_k be the set consisting of all values such that $\lambda_1(\Omega) \geq \cdots \geq \lambda_k(\Omega) > 1$ and $\lambda_{k+1}(\Omega) = \cdots = \lambda_p(\Omega) = 1$. In addition, let τ denote the number of $\lambda_i(\Omega_n)$'s which are greater than 1. According to Zhu, Miao and Peng (2006), we have

$$\sup_{\lambda(\Omega)\in\Theta_k} \log L(\lambda(\Omega)) = -\frac{n}{2}\sum_{i=1}^{p}\log\lambda_i(\Omega_n) - \frac{np}{2} + \frac{n}{2}\sum_{i=1+\min(\tau,k)}^{p}(\log\lambda_i(\Omega_n) + 1 - \lambda_i(\Omega_n)).$$

Note that this supremum does not involve the unknowns relating to Ω and its eigenvalues. For defining the estimator $\widehat{K}_1$ of the true dimension K_1, we suggest using the equivalent form

$$\log L_k =: \frac{n}{2}\sum_{i=1+\min(\tau,k)}^{p}(\log(\lambda_i(\Omega_n)) + 1 - \lambda_i(\Omega_n)).$$

Therefore, a modified BIC type criterion can be defined as

$$G(k) = \log L_k - \frac{C_n k(2p-k+1)}{2}, \tag{3.5}$$

where the second term is the penalty term, C_n is a penalty constant, and $k(2p-k+1)/2$ is the number of free parameters of (3.4) needed to be estimated when $\lambda(\Omega) \in \Theta$. The estimator of K_1 is taken as the maximizer $\widehat{K_1}$ of $G(k)$ over $k \in \{0, \ldots, p-1\}$, that is,

$$G(\widehat{K_1}) = \max_{0\leq k\leq p-1} G(k). \tag{3.6}$$

Theorem 4. *Suppose* $\lambda_i(\Lambda_n) - \lambda_i(\Lambda) = O_P(n^{-1/2}h_w^{-1})$ *for* $i = 1, \ldots, K_1$, *and* $\lambda_i(\Lambda_n) = o_P(n^{-1/2}h_w^{-1})$, $i = K_1+1, \ldots, p$. *Then, for* C_n *satisfying* $C_n/n \to 0$ *and* $C_n h_w^2 \to \infty$, $\widehat{K_1} \to K_1$ *in probability.*

After we have identified the vectors in CMS and estimated consistently the structural dimension of CMS, we can simply use kernel smoothing to approximate

the link function $\phi(\cdot)$. In doing so, the effect of the conditional mean can be removed exhaustively. In the following section we turn to the recovery of the CVS through the residual $e = Y - \phi(\beta^T X)$.

4. Identification and Estimation of the CVS

In this section, we discuss how to recover the CVS. Since the OPG does not rely on the distribution of the predictor vector, we investigate how to extend the idea of the OPG to target the CVS. For notational clarity, we write $\sigma^{2(k)}(x)$ as the $k-$th derivative of $\sigma^2(x) =: E(e^2|X = x) = E(e^2|\alpha^T X = \alpha^T x)$ with respect to x, and d_V as the structural dimension of CVS.

Theorem 5. *If $\sigma^2(x)$ is differentiable, then α is in the space spanned by the first d_V eigenvectors of $\Delta_e = E[\sigma^{2(1)}(X)\sigma^{2(1)}(X)^T]$ corresponding to the largest d_V eigenvalues.*

The above theorem implies that the OPG method, with the response Y replaced by e^2, the square of the residuals, can be used to infer the CVS. This is referred as the e^2-based OPG method hereafter.

To obtain the residuals, we first need to estimate the mean function $\phi(x) = E(Y|X^T\beta = x^T\beta)$. For ease of exposition, we use kernel estimation. Specifically, for any $p \times d_M$ orthogonal matrix β, denote by $f(x^T\beta)$ the density function of $X^T\beta$. When the independent and identically distributed sample points $\{(x_i, y_i), i = 1, \ldots, n\}$ are available, the kernel estimator $\hat{f}(x^T\beta)$ takes the form

$$\hat{f}(x^T\beta) = \frac{1}{n}\sum_{i=1}^{n} K_{lh_l}(x_i^T\beta - x^T\beta) = \frac{1}{nh_l^{d_M}}\sum_{i=1}^{n} K_l\Big(\frac{x_i^T\beta - x^T\beta}{h_l}\Big).$$

Write $\hat{g}(x^T\beta) = 1/(nh_l^{d_M})\sum_{i=1}^{n} y_i K_l[(x_i^T\beta - x^T\beta)/h_l]$, and then $\hat{\phi}(x) = \hat{g}(x^T\beta)/\hat{f}(x^T\beta)$. Recall that we use second order kernel function $K_w(\cdot)$ as a weight function when the OPG method is suggested. However, here we can use a different kernel function $K_l(\cdot)$, accompanied by a separate bandwidth h_l, to approximate the link function $\phi(\cdot)$.

When β and d_M are unknown, and the estimators $\widehat{d_M}$ and $\hat{\beta}_{\widehat{d_M}}$ are applicable, and the resulting estimator of $\phi(x)$ is

$$\begin{aligned}\hat{\phi}(x) &= \frac{\hat{g}(x^T\hat{\beta}_{\widehat{d}_M})}{\hat{f}(x^T\hat{\beta}_{\widehat{d}_M})} \\ &= \frac{\frac{1}{nh_l^{\widehat{d}_M}}\sum_{i=1}^{n} y_i K_l\Big(\frac{x_i^T\hat{\beta}_{\widehat{d}_M} - x^T\hat{\beta}_{\widehat{d}_M}}{h_l}\Big)}{\frac{1}{nh_l^{\widehat{d}_M}}\sum_{i=1}^{n} K_l\Big(\frac{x_i^T\hat{\beta}_{\widehat{d}_M} - x^T\hat{\beta}_{\widehat{d}_M}}{h_l}\Big)}. \end{aligned} \tag{4.1}$$

Therefore, $\hat{e}_i = y_i - \hat{\phi}(x_i^T \hat{\beta}_{\widehat{d_M}})$ are the estimators of corresponding e_i's.

When $\hat{e}_i$'s are available, consider the minimization problem

$$\min_{\beta_{0j}^*,\ldots,\beta_{rj}^*} \sum_{i=1}^{n} \Big[\hat{e}_i^2 - \beta_{0j}^* - \beta_{1j}^{*,T}(x_i - x_j)_{(1)} - \cdots - \beta_{rj}^{*,T}(x_i - x_j)_{(r)}\Big]^2 K_w\Big(\frac{x_i - x_j}{h_w}\Big). \quad (4.2)$$

Under this least square criterion, we obtain the minimizer $\widehat{\beta}_j^* =: (\widehat{\beta}_{0j}^*, \ldots, \widehat{\beta}_{rj})^{*,T} = (X_{nj}^T W_{nj} X_{nj})^{-1}(X_{nj}^T W_{nj}(\hat{e}_1^2, \ldots, \hat{e}_n^2)^T$. Similar to the original OPG method, we can estimate $\Delta_e = E[\sigma^{2(1)}(X)\sigma^{2(1)}(X)^T]$ by $\Delta_{en} = (1/n)\sum_{j=1}^n \beta_{1j}^* \beta_{1j}^{*,T}$. Applying the spectral decomposition of Δ_{en}, we can obtain the estimate of the CVS.

For notational clarity, we define $H_{OPG} = v_1 S_r^{-1}(\mu_1, \ldots, \mu_r)^T[e^2 - \sigma^2(X)]\beta^{*,T} v_1^T + v_1 \beta^* [e^2 - \sigma^2(X)](\mu_1, \ldots, \mu_r)(S_r^{-1})^T v_1^T$, and $V_{OPG} = \lambda^T \mathrm{Cov}[Vech(H_{OPG})]\lambda$ for any $\lambda \in R^{p(p+1)/2}$. The asymptotic normality of the e^2-based OPG method is at hand.

Theorem 6. *Assume that conditions* C1−C10 *in the Appendix hold. As* $n \to \infty$, *we have*

$$\sqrt{n} h_w(\Delta_{en} - \Delta_e) \xrightarrow{d} H_{OPG},$$

where $\lambda^T Vech(H_{OPG})$ *is* $N(0, V_{OPG})$ *for any* $\lambda \neq 0$.

5. Illustrative Examples

5.1. Simulation study

The following six models were simulated to evaluate the performance of the proposed e^2-based OPG method when recovering the CVS.

Model 1: $y = x^T\beta_1 + (x^T\alpha_1)\varepsilon$;
Model 2: $y = x^T\beta_1 + (x^T\alpha_2)\varepsilon$;
Model 3: $y = (x^T\beta_2)^2 + (x^T\alpha_1)\varepsilon$;
Model 4: $y = (x^T\beta_1)^3 + (x^T\alpha_1)\varepsilon$;
Model 5: $y = (x^T\beta_1) + [(x^T\alpha_1) + e^{(x^T\alpha_2)}]\varepsilon$;
Model 6: $y = [(x^T\beta_1) + e^{(x^T\beta_2)}] + [(x^T\alpha_1) + e^{(x^T\alpha_2)}]\varepsilon$.

In these models, the covariates x and the error ε are independent, and follow respectively normal distribution $N(0, I_3)$ and $N(0, 1)$ where I_3 is a 3×3 identity matrix. We chose $\alpha_1 = (0, 1, 0)^T$, $\alpha_2 = \beta_1 = (1, 0, 0)^T$ and $\beta_2 = (1, 1, 0)^T/\sqrt{2}$. The basic experiment was replicated to obtain 200 data sets, each of size $n = 400$.

We chose these models based on the following considerations. In all six models, $\mathcal{S}_{Y|X} = \mathcal{S}_{E(Y^2|X)} = \mathcal{S}_{E(Y|X)} + \mathcal{S}_{Var(Y|X)}$. The CMS and the CVS are

orthogonal in Model 1 but identical in Model 2, with each model having a linear link. The CMS and the CVS overlap in Model 3 but are orthogonal in Model 4. The link functions in the latter two models are nonlinear. In the previous four models, both the CMS and the CVS are single-index. In contrast, in Models 5 and 6, the CVS is multi-index. The CMS is a proper subspace of the CVS in Model 5, while the CMS is identical to the CVS in Model 6.

When the OPG is used to target the CMS, the product kernel K_w was the product of p kernel functions each of the form $15/16(1-u^2)^2 I_{(|u|\leq 1)}$ in (3.2), see Härdle and Mammen (1993). To select a bandwidth h_w for undersmoothing the estimator under the constraint of condition C8, which is needed for asymptotic normality, we needed a smaller bandwidth than the one that is optimal in a nonparametric regression scheme. Following the idea in Carroll, Fan, Gijbels and Wand (1997) and Stute and Zhu (2005), we employed an algorithm which could be easily implemented. We chose $h_w = n^{-2/15} h_{opt,w}$, where $h_{opt,w} = O(n^{-1/5})$ is the optimal bandwidth in terms of the generalized cross-validation (GCV) criterion. A similar idea was also used in Zhu (2003), Zhu and Ng (2003), and Zhu and Zhu (2007).

Subsequently, the spectral decomposition was applied to the estimation of the kernel matrix of the OPG method. Then the BIC type criterion (3.6) with $C_n = \sqrt{n}$ was used to estimate the structural dimension of the CMS .

After the OPG estimator was obtained, kernel smoothing was used to estimate the link functions for all models. We chose a kernel function $K_l(\cdot)$ with order higher than 2 to recover the CVS consistently, because higher-order kernel functions can estimate the link functions with smaller bias. Specifically, we used the fourth-order kernel $K_l(u)$ which is the product of $\widehat{d_M}$ kernel functions each of the form $k(u)I_{(|u|\leq 1)}/\int_{-1}^{1} k(u)du$ with $k(u) = [(3-u^2)/(2\sqrt{2\pi})]e^{-u^2/2}$. Then we chose the bandwidth h_l by GCV criterion. We used these two kernel functions (K_w, K_l), and two bandwidths (h_w, h_l), throughout the investigation of the illustrative examples.

To assess the performance of our proposed methods, we used the trace correlation coefficient $R^2 = trace(P_\alpha P_{\hat{\alpha}})/d_V$, proposed by Ferré (1998), where d_V is the dimension of the CVS, P_A is the orthogonal projection onto A in the usual inner product.

The mean μ and the standard deviation s of the trace correlation coefficients over 200 repetitions were used to evaluate the efficiency of our proposed methods when the OPG was used to recover the CMS:

$$\mu = \frac{1}{200}\sum_{i=1}^{200} R_i^2 \quad \text{and} \quad s = \left[\frac{1}{200}\sum_{i=1}^{200}(R_i^2-\mu)^2\right]^{\frac{1}{2}}. \tag{5.1}$$

Table 1. The frequency of decisions of dimension with $n = 400$.

	$dim = 0$	$dim = 1$	$dim = 2$
Model 1	0.00	0.99	0.01
Model 2	0.00	1.00	0.00
Model 3	0.01	0.95	0.04
Model 4	0.00	0.99	0.01
Model 5	0.05	0.10	0.85
Model 6	0.08	0.12	0.80

In Model 1, the e^2- based OPG was employed to recover the CVS. We found $\mu = 0.9725$ and $s = 0.0321$. That is, under the linear model, the e^2-based OPG performed well when the CVS and the CMS were orthogonal.

In Model 2, $\mathcal{S}_{Var(Y|X)} = \mathcal{S}_{Y|X} = \mathcal{S}_{E(Y^2|X)} = \mathcal{S}_{E(Y|X)}$, the e^2-based OPG after the three-step procedure gave $\mu = 0.9712$ and $s = 0.0231$. By comparing the performance of the e^2-based OPG in Model 1 with that in Model 2, we can clearly see that it recovered the CVS very well whether the CVS and the CMS were orthogonal or not.

Model 3 and Model 4 were selected to show the performance of the e^2-based OPG in nonlinear models. In Model 3, $\mu = 0.9230$ and $s = 0.0543$ means that e^2-based OPG performed well in nonlinear models.

Model 4 is used for comparison with Model 1. The results $\mu = 0.9239$ and $s = 0.1222$ suggest that the performance of the e^2-based OPG method in nonlinear models is not as stable as in linear model.

The CVS is multi-index in Models 5 and 6. We found $\mu = 0.8823$ and $s = 0.1523$ in Model 5, and $\mu = 0.8551$ and $s = 0.2041$ in Model 6. Thus, the e^2-based OPG method still worked well.

We also used the BIC type criterion with $C_n = \sqrt{n}$ to estimate the structural dimension of the CVS. The results are reported in Table 1. We can see that the BIC type method is worthy of recommendation.

5.2. Horse mussel data

A sample of 201 horse mussels was collected at 5 sites in the Malborough Sounds at the Northeast of New Zealand's South Island (Camden (1999)). The response variable is muscle mass Y, the edible portion of the mussel, in grams. The quantitative predictors are all related characteristics of mussel shells: shell width W and shell length L, each in mm, and shell mass S, in grams. To ease the interpretation we assume that the data are independent and identically distributed from the total combined population over the five sites. There are a few additional predictors that are not used here.

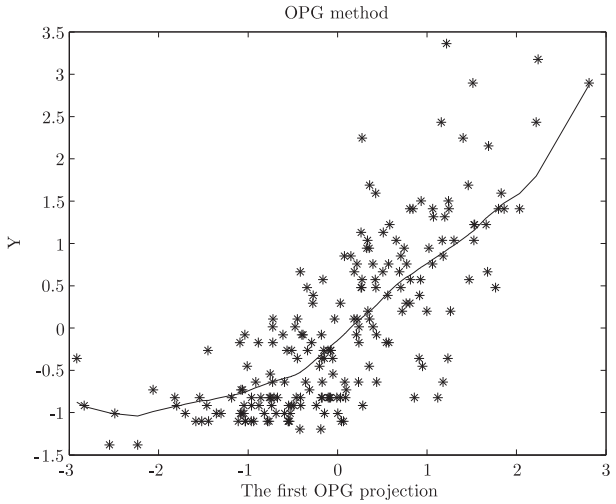

Figure 1. Horse mussel data. The vertical line maps the response Y, the horizontal line, the first OPG predictors. The line is obtained through kernel smoothing.

Consider the regression problem with the response Y and the predictors (L, W, S) transformed to $X = (L, W^{0.36}, S^{0.11})^T$ to comply with the linearity condition. Cook (1998a)) analyzed this dataset using SIR, and suggested that the central subspace $\mathcal{S}_{Y|X}$ is one-dimensional.

We re-visit this dataset to explore the model structure. We first used the OPG method to find the CMS, and used the BIC type algorithm to estimate its dimension. We adopted the same kernel K_w and a similar bandwidth selector to that are used in the previous simulations. The resulting bandwidth was $h_w = 0.4$. Our proposed BIC suggested a one-dimensional subspace. Therefore, we chose the first projection

$$\beta_1 = (0.8585, 0.5090, -0.0623)^T.$$

This estimation is similar to Cook (1998a). The scatter-plot of Y versus $\beta_1^T X$ in Figure 1 shows that this direction is contained in the CMS.

Figure 1 also shows that the dispersion of Y becomes larger with larger value of $\beta_1^T X$, which indicates that there may exist a heteroscedasticity structure in the data. To verify this argument, we further studied these data using the e^2-based OPG method to explore heteroscedasticity. We chose the same kernel function $K_l(\cdot)$ and the bandwidth $h_l = 0.7$. Our BIC still suggested a one-dimensional CVS when the e^2-based OPG method was used. Therefore, we also chose the first vector $\alpha_1 = (0.9785, 0.1275, -0.1622)^T$.

To assess the similarity between β_1 and α_1, we used the trace correlation coefficient and found $R^2 = (\beta_1^T \alpha_1)^2 = 0.9151^2$. Together with Cook's (1998a) result

that $\mathcal{S}_{Y|X}$ is one dimensional, we suggest that $\mathcal{S}_{Y|X} = \mathcal{S}_{E(Y|X)} = \mathcal{S}_{Var(Y|X)} = Span\{\alpha_1\} = Span\{\beta_1\}$. Therefore, it is reasonable to propose that, for the horse mussel data, $y = \phi(x^T\alpha_1) + \sigma(x^T\alpha_1)\varepsilon$. This analysis verifies that the e^2-based OPG can identify the heteroscedasticity when the CMS and the CVS are not orthogonal.

Acknowledgement

The first author was supported by an NSF grant from National Natural Science Foundation of China (No. 10701035), ChenGuang project of Shanghai Education Development Foundation (No. 2007CG33), a special fund for young teachers in Shanghai universities (No. 79001320), and Science and Technology Commission of Shanghai Municipality Pujiang Project (No. 07PJ14037). The second author was supported by a grant (HKU 7058/05P) from the Research Grants Council of Hong Kong, and a FRG grant from Hong Kong Baptist University, Hong Kong. The authors are grateful to the Editor, an associate editor, and the two referees for their generous help, constructive comments, and suggestions, which led to a great improvement of our earlier draft.

Appendix: Some conditions

The regularity conditions we used are listed here.

(C1) The kernel function $K_w(\cdot)$ is a continuous density function having bounded support.

(C2) The density function of X satisfies: $0 < \inf_t f_X(t) \leq \sup_t f_X(t) < \infty$, and its second derivative $f_X^{(2)}(t)$ satisfies a Local Liptschitz condition over the support $\mathcal{T}$ of X, namely, there exist a constant c such that $|f_X^{(2)}(t+v) - f_X^{(2)}(t)| \leq c|t|$ for any t in a neighborhood of zero.

(C3) If $\phi(x) =: E(Y|X = x)$, the $(r+3)-$th derivative $\phi^{(r+3)}(\cdot)$ exists and is continuous over $\mathcal{T}$.

(C4) The variance function $\sigma^2(x) = E[(Y - \phi(X))^2|X = x]$ has a bounded second derivative over $\mathcal{T}$.

(C5) The kernel function $K_l(u)$ is bounded and symmetric, and is Lipschitz continuous on $\mathcal{T}$; moreover, it satisfies $\int_{\mathcal{T}} K_l(u) = 1$; $\int_{\mathcal{T}} u^i K_l(u) = 0$, $i = 1, \ldots, d-1$, $\int_{\mathcal{T}} u^d K_l(u) = M_K \neq 0$, $d \geq 2$.

(C6) The bandwidth h_l satisfies $nh_w^2 h_l^{2d_M+2} \to \infty$ as $n \to \infty$.

(C7) The link function $\phi(x)$ and the variance function $\sigma^2(x)$ are bounded on $\mathcal{T}$.

(C8) The bandwidth h_w satisfies $\sqrt{n}h_w^{p+1} \to \infty$ and $\sqrt{n}h_w^{r+1} \to 0$.

(C9) $E[(Y-\phi(X))^4|X=x]$ has a bounded second derivative over $\mathcal{T}$.

(C10) The density function $f(x^T\beta)$ of $X^T\beta$, the density function $f_1(x^T\alpha)$ of $X^T\alpha$, $\phi(x)=E(Y|X=x)$ and $\sigma^2(x)=E(e^2|X=x)$ are $d-$times differentiable on $\mathcal{T}$, and their derivatives satisfy the following condition. If $H(\cdot)$ denotes $f(x^T\beta)$, $f_1(x^T\alpha)$, $\phi(x)$ or $\sigma^2(x)$, there exists a neighborhood of the origin, say U, and a constant $c>0$ such that, for any $u\in U$,

$$H^{(d-1)}(t+u)-H^{(d-1)}(t)\le c|u|,$$

where $H^{(d-1)}(t)$ denotes the $(d-1)-$th derivatives of the function $H(\cdot)$.

References

Bura, E. and Cook, R. D. (2001). Estimating the structural dimension of regressions via parametric inverse regression. *J. Roy. Statist. Soc. Ser. B* **63**, 393-410.

Camden, M. (1999). *The Data Bundle*. New Zealand Statistical Association, WeHington, New Zealand.

Carroll, R. J., Fan, J., Gijbels, I. and Wand, M. P. (1997). Generalized partially linear single-index models. *J. Amer. Statist. Assoc.* **92**, 477-489.

Chiaromonte, F., Cook, R. D. and Li, B. (2002). Sufficient dimension reduction in regression with categorical predictors. *Ann. Statist.* **30**, 475-497.

Cook, R. D. (1994). On the interpretation of regression plots. *J. Amer. Statist. Assoc.* **89**, 177-189.

Cook, R. D. (1996). Graphics for regressions with a binary response. *J. Amer. Statist. Assoc.* **91**, 983-992.

Cook, R. D. (1998a). *Regression Graphics: Ideas for Studying Regressions through Graphics*. Wiley, New York.

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

Cook, R. D. and Li, B. (2002). Dimension reduction for conditional mean in regression. *Ann. Statist.* **30**, 455-474.

Cook, R. D. and Ni, L. (2005). Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. *J. Amer. Statist. Assoc.*, **100**, 410-428.

Cook, R. D. and Weisberg, S. (1991). Discussion to "Sliced inverse regression for dimension reduction". *J. Amer. Statist. Assoc.* **86**, 316-342.

Eaton, M. L. and Tyler, D. E. (1991). On Wielandt's Inequality and its applications to the asymptotic distribution of the eigenvalues of a random symmetric matrix. *Ann. Statist.* **19**, 260-271.

Ferré, L. (1998). Determining the dimension in sliced inverse regression and related methods. *J. Amer. Statist. Assoc.* **93**, 132-140.

Härdle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *Ann. Statist.* **21**, 1926-1947.

Härdle, W. and Stoker, T. M. (1989). Investigating smooth multiple regression by the method of average derivatives. *J. Amer. Statist. Assoc.* **84**, 986-995.

Li, B., Zha, H. and Chiaromonte, F. (2005). Contour regression: A general approach to dimension reduction. *Ann. Statist.* **33**, 1580-1616.

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.

Stute, W. and Zhu, L. X. (2005). Nonparametric checks for single-index models. *Ann. Statist.* **33**, 1048-1083.

Xia, Y., Tong, H., Li, W. K. and Zhu, L. X. (2002). An adaptive estimation of optimal regression subspace, *J. Roy. Statist. Soc. Ser. B* **64**, 363-410.

Zhu, L. X. (2003). Model checking of dimensional-reduction type for regression. *Statist. Sinica* **13**, 283-296.

Zhu, L. X. and Fang, K. T. (1996). Asymptotics for kernel estimate of sliced inverse regression. *Ann. Statist.* **24**, 1053-1068.

Zhu, L. X. and Ng, K. W. (2003). Checking the adequacy of a partial linear model. *Statist. Sinica* **13,** 763-781.

Zhu, L. X., Miao, B. Q. and Peng, H. (2006). Sliced inverse regression with large dimensional covariates. *J. Amer. Statist. Assoc.* **101**, 630-643.

Zhu, L. P. and Zhu, L. X. (2007). On kernel method for sliced average variance estimation. *J. Multivariate Anal.* **98**, 970-991.

School of Finance and Statistics, East China Normal University, No. 500 Dong Chuang Road, Shanghai, China.

E-mail: lpzhu@stat.ecnu.edu.cn

Department of Mathematics, Hong Kong Baptist University, Hong Kong, China.

E-mail: lzhu@hkbu.edu.hk

(Received March 2007; accepted January 2008)