Biometrika (2007), **94**, 3, *pp*. 615–625
© 2007 Biometrika Trust
Printed in Great Britain

doi:10.1093/biomet/asm043
Advance Access publication 5 August 2007

Partial inverse regression

BY LEXIN LI

Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695, U.S.A.
li@stat.ncsu.edu

R. DENNIS COOK

School of Statistics, University of Minnesota, St Paul, Minnesota 55108, U.S.A.
dennis@stat.umn.edu

AND CHIH-LING TSAI

Graduate School of Management, University of California, Davis, California 95616, U.S.A.
cltsai@ucdavis.edu

SUMMARY

In regression with a vector of quantitative predictors, sufficient dimension reduction methods can effectively reduce the predictor dimension, while preserving full regression information and assuming no parametric model. However, all current reduction methods require the sample size n to be greater than the number of predictors p. It is well known that partial least squares can deal with problems with $n < p$. We first establish a link between partial least squares and sufficient dimension reduction. Motivated by this link, we then propose a new dimension reduction method, entitled partial inverse regression. We show that its sample estimator is consistent, and that its performance is similar to or superior to partial least squares when $n < p$, especially when the regression model is nonlinear or heteroscedastic. An example involving the spectroscopy analysis of biscuit dough is also given.

Some key words: Partial least squares; Single-index model; Sliced inverse regression.

1. INTRODUCTION

Regression analysis is widely used to study the relationship between a response variable Y and a $p \times 1$ predictor vector X. In practice, this relationship often depends solely on a few linear combinations $\beta^{\mathsf{T}} X$ of X, where β is a $p \times d$ matrix with $d < p$. For instance, a single-index model (Härdle et al., 1993) involves only one linear combination of the predictors, and has the form

$$Y = f(\beta^{\mathsf{T}} X) + \varepsilon, \tag{1}$$

where $d = 1$, $\|\beta\| = 1$, f is an unknown link function, and ε is an error term independent of X with its distribution unspecified. Other well-known models with $d = 1$ include the heteroscedastic model,

$$Y = f(\beta^{\mathsf{T}} X) + g(\beta^{\mathsf{T}} X)\varepsilon, \tag{2}$$

and the logistic single-index model,

$$Y \mid X \sim \text{Ber}\{\mu(X)\}, \text{ with } \mu(X) = \frac{1}{1 + \exp\{-h(\beta^{\mathsf{T}} X)\}}, \tag{3}$$

where g and h are unknown smooth link functions.

Sufficient dimension reduction methods estimate linear combinations $\beta^{\mathsf{T}} X$ that convey full regression information. The methodology stems from considering dimension reduction subspaces $\mathcal{S} \subset \mathbb{R}^p$, which are subspaces with the property $Y \perp\!\!\!\perp X \mid P_{\mathcal{S}} X$, where $P_{(\cdot)}$ denotes a projection operator with respect to the standard inner product. Under mild conditions (Cook, 1996) the intersection of all dimension reduction subspaces is itself a dimension reduction subspace, and is then called the central subspace $\mathcal{S}_{Y|X}$ (Cook, 1998, p.105). By definition, $\mathcal{S}_{Y|X}$ is a parsimonious population parameter that contains full information about the regression of Y on X, and thus is our main object of interest. We assume the existence of $\mathcal{S}_{Y|X}$ throughout this article, and let $d = \dim(\mathcal{S}_{Y|X})$. For models (1), (2) and (3), $\mathcal{S}_{Y|X} = \text{span}(\beta)$ with $d = 1$.

Commonly used pioneering model-free methods for estimating a basis of $\mathcal{S}_{Y|X}$ include ordinary least squares (Li & Duan, 1989) when $d = 1$, sliced inverse regression (Li, 1991) and sliced average variance estimation (Cook & Weisberg, 1991). Unfortunately, all of those methods require the number of observations n to be greater than p, and this limits applicability. By contrast, the method of partial least squares (Wold, 1975) does not have this limitation. Background about partial least squares estimator can be found in Helland (1988, 1990, 1992), Næs & Helland (1993), Garthwaite (1994) and Stone & Brooks (1990).

In this article we propose a natural extension of sliced inverse regression, entitled partial inverse regression, that can handle regressions with $n < p$, and that can also be effective for highly correlated predictors. We will concentrate only on regressions with $d = 1$, such as models (1), (2) and (3), because they often serve as the first projective approximation. We will use span(A) to denote a subspace of $\mathbb{R}^p$ that is spanned by the columns of the matrix A.

2. PARTIAL LEAST SQUARES AND PARTIAL INVERSE REGRESSION

2·1. *Sufficient dimension reduction methods*

Let $\Sigma_x = \text{var}(X) > 0$, let $\sigma_{xy} = \text{cov}(X, Y)$, and assume without loss of generality that $E(X) = 0$. Also, define $\Sigma_{x|y} = \text{var}\{E(X \mid Y)\}$, which is the covariance matrix of the inverse mean $E(X \mid Y)$. The usual linearity condition requires that

$$E(X \mid \beta^{\mathsf{T}} X = t) \text{ is a linear function of } t. \tag{4}$$

Under this condition, it is well-established that $\beta_{\text{OLS}} \equiv \Sigma_x^{-1} \sigma_{xy} \in \mathcal{S}_{Y|X}$, and that span$(\Sigma_x^{-1} \Sigma_{x|y}) \subseteq \mathcal{S}_{Y|X}$. These relationships are the population foundations for dimension reduction based on ordinary least squares and sliced inverse regression (Li, 1991). The linearity condition (4) is not a severe restriction, since it holds to a reasonable approximation as p increases (Hall & Li, 1993). Furthermore, when the predictors are elliptically symmetric, condition (4) holds for all p (Eaton, 1986). Given n observations $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, an estimator can then be obtained by substituting the sample versions of Σ_x^{-1}, σ_{xy} and $\Sigma_{x|y}$. However, when $n < p$, the sample version of Σ_x is singular, and consequently neither ordinary least squares nor sliced inverse regression is directly applicable. As discussed in § 2.2, partial least squares can be viewed as an adaptation of ordinary least squares for regressions with $n < p$. In order to adapt sliced inverse regression to estimation of $\mathcal{S}_{Y|X}$ in regressions with $d = 1$ and $n < p$, we first rewrite its population foundation as $\beta_{\text{SIR}} = \Sigma_x^{-1}\zeta_{x|y} \in \mathcal{S}_{Y|X}$, where $\zeta_{x|y}$ is the principal eigenvector of $\Sigma_{x|y}$. Given the linearity condition, the sample versions of β_{SIR} and β_{OLS} are both consistent estimators of the index β in (1), (2) and (3), without knowledge of the link functions. We next establish a connection between partial least squares and the central subspace $\mathcal{S}_{Y|X}$.

2·2. *Partial least squares*

Following Helland (1988, 1990), we can express the population version of the partial least squares coefficient vector in closed form as

$$\beta_{\text{PLS}} \equiv P_{R_q(\Sigma_x)}\beta_{\text{OLS}}, \tag{5}$$

where $P_{R_q(\Sigma_x)}$ denotes the projection on to the subspace spanned by the columns of the matrix $R_q \equiv (\sigma_{xy}, \Sigma_x\sigma_{xy}, \ldots, \Sigma_x^{q-1}\sigma_{xy})$ with respect to the Σ_x inner product, and q is a positive integer that is typically estimated in practice. When $R_q^{\mathrm{T}}\Sigma_x R_q$ is of full rank, $P_{R_q(\Sigma_x)} = R_q(R_q^{\mathrm{T}}\Sigma_x R_q)^{-1}R_q^{\mathrm{T}}\Sigma_x$ and $\beta_{\text{PLS}} = R_q(R_q^{\mathrm{T}}\Sigma_x R_q)^{-1}R_q^{\mathrm{T}}\sigma_{xy}$. Furthermore, σ_{xy} can always be written as a linear combination of $m \leqslant p$ eigenvectors of Σ_x: $\sigma_{xy} = \sum_{i=1}^m \theta_i\gamma_i$, where $\gamma_1, \ldots, \gamma_m$ are eigenvectors of Σ_x with corresponding eigenvalues $\lambda_1 \geqslant \ldots \geqslant \lambda_m > 0$, and $\theta_1, \ldots, \theta_m$ are nonzero real numbers. If the eigenvalue, λ_j say, of Σ_x is multiple, then the corresponding γ_j is the normed eigenvector obtained by rotation in the eigen-space (Helland, 1990). In particular, if Σ_x is a positive constant times the identity matrix, then we can take $m = 1$ and $\gamma_1 = \sigma_{xy}$.

To establish a relationship between β_{PLS} and the central subspace, first note that $\beta_{\text{OLS}} \in \text{span}(\gamma_1, \ldots, \gamma_m)$. Since $\sigma_{xy}, \Sigma_x\sigma_{xy}, \ldots, \Sigma_x^{m-1}\sigma_{xy}$ are linearly independent, and each belongs to $\text{span}(\gamma_1, \ldots, \gamma_m)$, we have $\text{span}(\gamma_1, \ldots, \gamma_m) = \text{span}(R_m) \subseteq \text{span}(R_q)$, for $q \geqslant m$. Consequently, $\beta_{\text{OLS}} \in \text{span}(R_q)$ and $\beta_{\text{PLS}} = P_{R_q(\Sigma_x)}\beta_{\text{OLS}} = \beta_{\text{OLS}}$. This result was also obtained by Helland & Almøy (1994, Lemma 1a), and a special case was discussed in Naik & Tsai (2000). When the linearity condition holds, $\beta_{\text{OLS}} \in \mathcal{S}_{Y|X}$, and thus $\beta_{\text{PLS}} \in \mathcal{S}_{Y|X}$. We summarize the above discussion in Proposition 1.

PROPOSITION 1. *If we assume the linearity condition* (4)*, and that* $q \geqslant m$*, then* $\beta_{\text{PLS}} \in \mathcal{S}_{Y|X}$.

Proposition 1 shows that, under the linearity condition, partial least squares may provide a useful estimator of the direction β without knowledge of the link function. Moreover, it motivates us to adapt sliced inverse regression to the $n < p$ problem by using basic ideas from partial least squares.

2·3. *Partial inverse regression*

Observing the similarity between β_{SIR}, β_{OLS} and β_{PLS}, we propose a partial inverse regression estimator based on the population quantity

$$\beta_{\text{PIRE}} \equiv P_{R_q^*(\Sigma_x)}\beta_{\text{SIR}}, \tag{6}$$

where $P_{R_q^*(\Sigma_x)}$ denotes the projection operator on to $\text{span}(R_q^*)$ with respect to the Σ_x inner product, $R_q^* \equiv (\zeta_{x|y}, \Sigma_x\zeta_{x|y}, \ldots, \Sigma_x^{q-1}\zeta_{x|y})$, and $\zeta_{x|y}$ is the principal eigenvector of $\Sigma_{x|y}$ as defined before. As with σ_{xy}, $\zeta_{x|y}$ can be represented uniquely as a linear combination of the eigenvectors γ_j of Σ_x: $\zeta_{x|y} = \sum_{i=1}^m \theta_i^*\gamma_i$, where $\theta_1^*, \ldots, \theta_m^*$ are nonzero real numbers. In effect, β_{PIRE} is a projection of β_{SIR} on to the subspace spanned by R_q^* with respect to Σ_x. When $q = m$, $R_q^{*\mathsf{T}}\Sigma_x R_q^*$ is of full rank, $P_{R_q^*(\Sigma_x)} = R_q^*(R_q^{*\mathsf{T}}\Sigma_x R_q^*)^{-1}R_q^{*\mathsf{T}}\Sigma_x$, and (6) becomes $\beta_{\text{PIRE}} = R_q^*(R_q^{*\mathsf{T}}\Sigma_x R_q^*)^{-1}R_q^{*\mathsf{T}}\zeta_{x|y}$. When $q > m$ and $R_q^{*\mathsf{T}}\Sigma_x R_q^*$ becomes singular, we can employ the generalized inverse to construct the projection in the population, so that

$$\beta_{\text{PIRE}} = R_q^*(R_q^{*\mathsf{T}}\Sigma_x R_q^*)^{-}R_q^{*\mathsf{T}}\zeta_{x|y}.$$

We can proceed analogously in the sample by using a generalized inverse, if the sample version of $R_q^{*\mathsf{T}}\Sigma_x R_q^*$ is ill-conditioned or singular; that is, we obtain the usual sample estimators $\hat{\Sigma}_x$, $\hat{\Sigma}_{x|y}$, and the first eigenvector $\hat{\zeta}_{x|y}$ of $\hat{\Sigma}_{x|y}$, and subsequently form the sample version $\hat{R}_q^*$ of R_q^* for a given q. The estimator $\hat{\beta}_{\text{PIRE}}$ of β_{PIRE} is then computed as $\hat{\beta}_{\text{PIRE}} = \hat{R}_q^*(\hat{R}_q^{*\mathsf{T}}\hat{\Sigma}_x\hat{R}_q^*)^{-}\hat{R}_q^{*\mathsf{T}}\hat{\zeta}_{x|y}$. Next, we describe the property of β_{PIRE} and $\hat{\beta}_{\text{PIRE}}$.

PROPOSITION 2. *If, for a regression with* $d = \dim(\mathcal{S}_{Y|X}) = 1$, *we assume that the linearity condition* (4) *is satisfied, the first eigenvalue of* $\Sigma_{x|y}$ *is greater than* 0, *and* $q \geqslant m$, *then* $\text{span}(\beta_{\text{PIRE}}) = \mathcal{S}_{Y|X}$. *In addition, the sample estimator* $\hat{\beta}_{\text{PIRE}}$ *is* $\surd n$*-consistent.*

Proof. Following the argument of Proposition 1, we can show that $\beta_{\text{SIR}} \in \text{span}(R_q^*)$. Consequently, $\beta_{\text{PIRE}} = P_{R_q^*(\Sigma_x)}\beta_{\text{SIR}} = \beta_{\text{SIR}} \in \mathcal{S}_{Y|X}$. Since $d = 1$ and the first eigenvalue of $\Sigma_{x|y}$ is positive, β_{PIRE} spans $\mathcal{S}_{Y|X}$. Moreover, the sample version $\hat{\zeta}_{x|y}$ is $\surd n$-consistent for $\zeta_{x|y}$ (Li, 1991), and the sample version $\hat{\Sigma}_x$ is $\surd n$-consistent for Σ_x. Therefore, $\hat{\beta}_{\text{PIRE}}$ is $\surd n$-consistent, which completes the proof. □

From the preceding discussion, we see that $\hat{\beta}_{\text{PIRE}}$ can be used to estimate the central subspace when $n < p$, provided that it is based on an estimator of q that is ideally larger than but reasonably close to m. As q increases beyond m, we are in effect adding unnecessary information to estimate parameters, which is reflected by having an ill-conditioned sample version of $R_q^{*\mathsf{T}}\Sigma_x R_q^*$. However there is not a discrete breakdown per se, as Fig. 1 in § 3 shows. We will discuss selection of q further in § 3·3. In addition, $\hat{\beta}_{\text{PIRE}}$ can retrieve information on both the conditional mean $E(Y \mid X)$ and the conditional variance $\text{var}(Y \mid X)$, whereas β_{PLS} can recover information only on the conditional mean $E(Y \mid X)$. For instance, consider the version of model (2) in which $f \equiv 0$ and $\text{var}\{g(\beta^{\mathsf{T}}X)\} > 0$. In this case the central subspace $\mathcal{S}_{Y|X} = \text{span}(\beta)$ may be consistently estimated by $\hat{\beta}_{\text{PIRE}}$, but not by $\hat{\beta}_{\text{PLS}}$. Proposition 2 is applicable to models (1), (2) and (3). We next use these models to study the performance of the partial inverse regression estimator.

3. SIMULATIONS AND EXAMPLE

3·1. *Simulation settings*

We have done extensive simulation studies, and for the sake of brevity we report the results for the following four single-index models:

$$Y = x_1 + x_2 + \ldots + x_{10} + \sigma_0\varepsilon, \tag{7}$$

$$Y = \exp\{-(x_1 + x_2 + \ldots + x_{10})\} + \sigma_0\varepsilon, \tag{8}$$

$$Y = \log(|x_1 + x_2 + \ldots + x_{10} - 4|) + \sigma_0\varepsilon, \tag{9}$$

$$Y = \exp\{0{\cdot}75(x_1 + x_2 + \ldots + x_{10})\}\varepsilon. \tag{10}$$

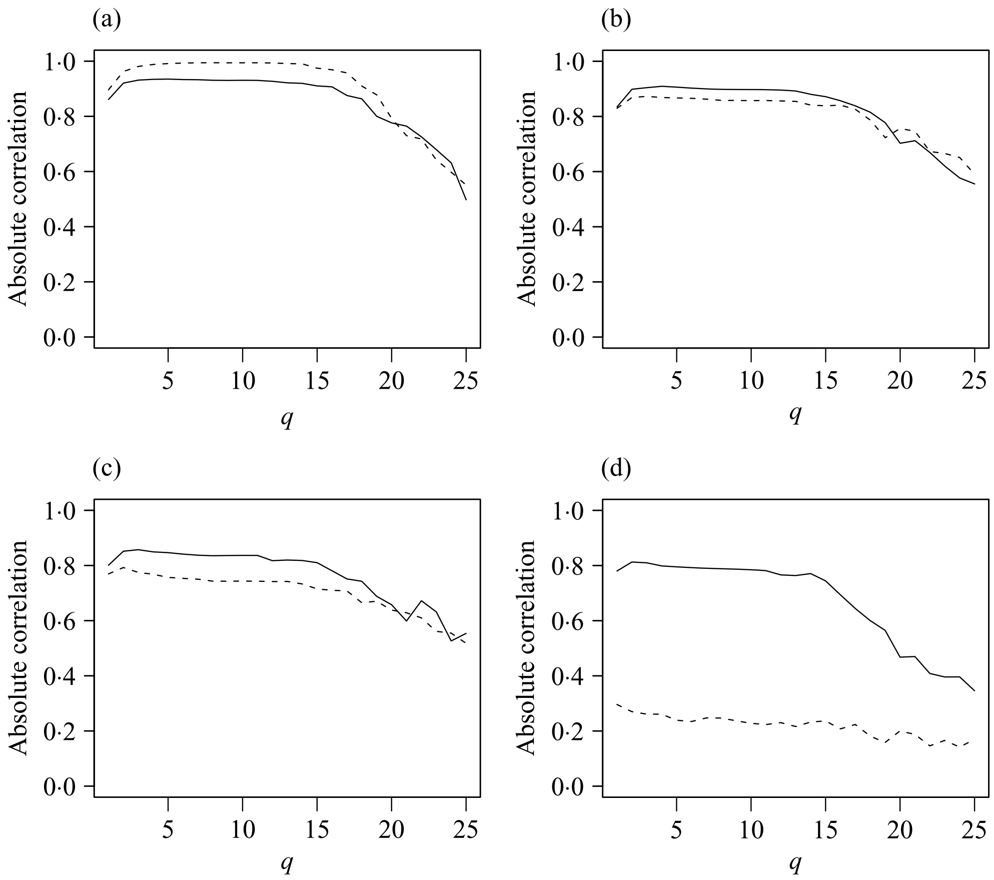

Fig. 1. Simulation study. Median of absolute correlation as a function of q for (a) Model (7), (b) Model (8), (c) Model (9), (d) Model (10): solid line, partial inverse regression; dashed line, partial least squares.

The link function of model (7) is linear, whereas the link functions of models (8) and (9) are nonlinear. Model (10) represents a heteroscedastic error structure and is in the same category as the single-index model (2). First, for each model, $n = 100$ observations were generated from $p = 110$ independent predictors, which follow the Un(0, 1) distribution in models (7) to (9), and the standard normal distribution for model (10). The linearity condition does not hold for the uniform predictors. However, this violation of the linearity condition is negligible because most low-dimensional projections of a high-dimensional data cloud are close to being normal (Hall & Li, 1993). The central subspace $\mathcal{S}_{Y|X}$ is spanned by the vector $\beta = (1, \ldots, 1, 0, \ldots, 0)^{\mathsf{T}}$, where the first 10 elements are ones and the remainder are zeros. The error term ε follows the standard normal distribution, and σ_0 in models (7) to (9) controls the magnitude of noise. Initially, σ_0 was chosen so that the signal-to-noise ratio, defined as the ratio of the range of the conditional mean of the response to the error standard deviation, equals 50. Since our focus is on the accuracy of the estimated direction, we follow Li (1991) and Carroll & Li (1992) and use the absolute correlation $|\text{corr}(\beta^{\mathsf{T}}X, \hat{\beta}^{\mathsf{T}}X)|$ as the performance measure, where $\hat{\beta}$ is the estimated direction, and X denotes the random variable with the specified distribution. Since $\beta^{\mathsf{T}}X$ is the true index, this measure is informative even when $n < p$.

3·2. *Finite-sample performance*

We first examined the performance of partial inverse regression and partial least squares as q, the number of columns of R_q^* and R_q, increases. Figure 1 depicts the median of the absolute correlations out of 100 realizations for a series of q values ranging from 1 to 25. For models (7) to (9), the two estimators perform similarly, with partial least squares performing slightly better in the linear model (7), and partial inverse regression performing a little better in models (8) and (9) with nonlinear link functions. In model (10), where the predictor effects are present in the conditional variance, the partial inverse regression estimator clearly outperforms partial least squares, because of the former method's capability of extracting information from both the conditional mean and variance.

We next studied the performance as the number of predictors p increases with n fixed. The simulation settings are the same as those in § 3·1. Figure 2 shows the median of $|\text{corr}(\beta^{\mathsf{T}}X, \hat{\beta}^{\mathsf{T}}X)|$ when p takes values in $\{10, 20, \ldots, 300\}$ with $n = 100$ and $q = 5$. We employed the usual sliced inverse regression and ordinary least squares estimators when $p < n$, but switched to their partial versions when $p \geqslant n$. It can be seen that the partial estimators perform as natural extensions of their large sample counterparts. It is also noted that the performance of partial inverse regression is satisfactory when p is as large as three times the sample size.

In models (7) to (10), the predictors were simulated as independent identically distributed random variables. As a result, var(X) has a single eigenvalue with multiplicity p, which

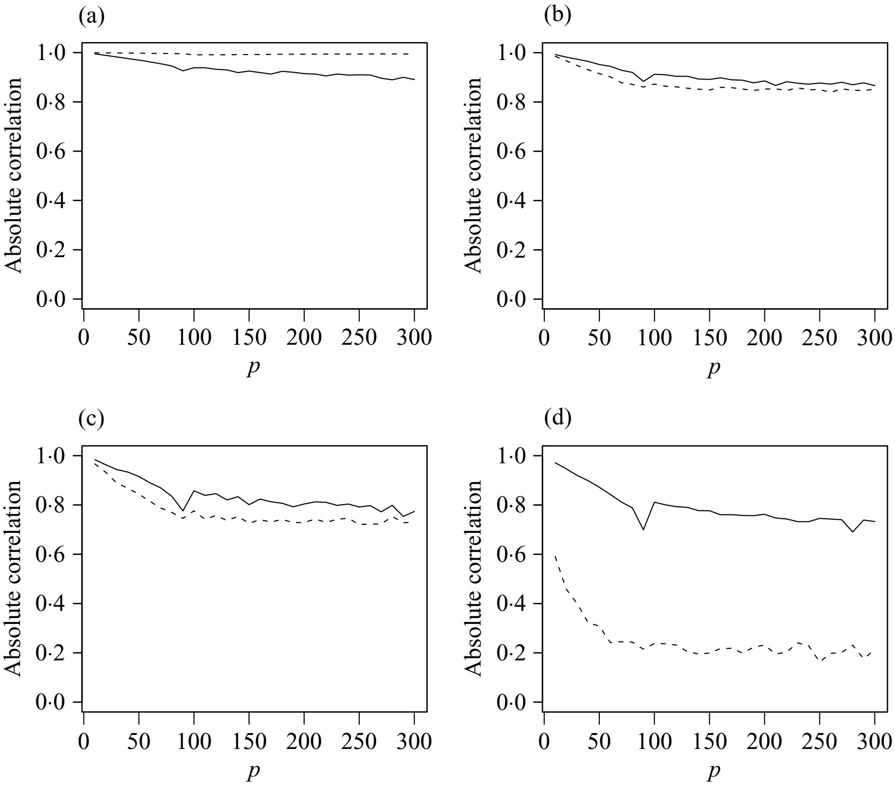

Fig. 2. Simulation study. Median of absolute correlation as a function of p for (a) Model (7), (b) Model (8), (c) Model (9), (d) Model (10): solid line, sliced and partial inverse regression; dashed line, ordinary and partial least squares.

implies that $m = 1$. From Fig. 1 and other simulations, we observed that the performance of the partial estimators is insensitive to the choice of q over a fairly wide range between 1 and 12, whereas their performance degrades when q exceeds a certain value. This is because the estimator of the inverse of the matrix ${R_q^*}^{\mathsf{T}} \Sigma_x R_q^*$, or $R_q^{\mathsf{T}} \Sigma_x R_q$, becomes much more variable for $q >> m$.

3·3. *Selection of q*

Here we present two approaches for estimating q of the matrices R_q and R_q^*. The first is the threshold approach. Given an estimator $\hat{R}$ of $R(p) = (\upsilon, \Sigma_x \upsilon, \ldots, \Sigma_x^{p-1}\upsilon)$, let

$$q = \sum_{j=1}^{p-1} I(r_j > \alpha), \text{ with } r_j = \lambda_j/\lambda_{j+1},\ j = 1, \ldots, p-1,$$

where $\upsilon = \zeta_{x|y}$ for partial inverse regression, $\upsilon = \sigma_{xy}$ for partial least squares, $I(\cdot)$ is the indicator function, α is a prespecified threshold value, and $\lambda_1, \ldots, \lambda_p$ are eigenvalues of the matrix $\hat{R}\hat{R}^{\mathsf{T}}$ in descending order. With $q \geqslant m$, the last $p - m$ of these eigenvalues equal zero in the population. However, their sample versions will be small but not exactly zero. Therefore, we choose the ratio of two adjacent eigenvalues to be slightly larger than one for the purpose of determining q.

The second approach to selecting q is based on the average squared error. The procedures are as follows: first, for each given q, obtain parameter estimates $\hat{\beta}(q)$ by the two partial methods; secondly, use cubic smoothing splines (Green & Silverman, 1994), or local linear regression (Fan & Gijbels, 1996; Fan & Yao, 1998) to estimate $\hat{f}$ for model (1), and estimate $\hat{f}$ and $\hat{g}$ for model (2), with the bandwidth selected by Craven & Wahba's (1979) generalized crossvalidation criterion; thirdly, find q that minimizes $\sum_{i=1}^{n}[Y_i - \hat{f}\{\hat{\beta}(q)^{\mathsf{T}} X_i\}]^2/n$ for model (1), and $\sum_{i=1}^{n}[Y_i - \hat{f}\{\hat{\beta}(q)^{\mathsf{T}} X_i\}]^2/[n\hat{g}\{\hat{\beta}(q)^{\mathsf{T}} X_i\}]$ for model (2).

We have studied the sensitivity of the threshold value α in the estimation of q. Our extensive simulations have found that $\alpha = 1{\cdot}5$ often yields a good selection of q. We also compared the performance of the threshold approach with $\alpha = 1{\cdot}5$ to that of the squared-error approach. Unreported simulations suggest that the two selection methods are comparable, whereas the threshold method enjoys computational simplicity since it does not require the estimation of the link functions.

3·4. *Correlated predictors*

Independent predictors were employed in § 3·1, resulting in an identity matrix multiplied by a constant as the predictor covariance matrix. We next examined the performance of the partial estimators under four correlation structures: (i) no correlation, (ii) correlation ρ between the first two predictors, (iii) pairwise correlation ρ among the first m predictors, and (iv) pairwise correlation ρ among all predictors. Table 1 summarizes the median and the median absolute deviation, in parentheses, based on 100 replications when $\rho = 0{\cdot}999$. The performance of the partial methods is similar for all four correlation structures. We also experimented with other values of ρ between 0·3 and 0·9, and observed the same patterns.

Following the simulations of large sample regressions reported by Naik & Tsai (2000), we also considered a linear model with extremely highly correlated predictors,

$$Y = \beta^{\mathsf{T}} X + \varepsilon, \qquad (11)$$

Table 1. *Median and median absolute deviation, in parentheses, of the absolute correlation for a variety of correlation structures, with* $\rho = 0{\cdot}999$. *Results are based on* 100 *data replications*

		Structure 1	Structure 2	Structure 3	Structure 4
Model (7)	PIRE	0·933 (0·024)	0·935 (0·015)	0·954 (0·014)	0·970 (0·007)
	PLS	0·994 (0·001)	0·994 (0·001)	0·996 (0·001)	0·998 (0·001)
Model (8)	PIRE	0·900 (0·037)	0·903 (0·031)	0·556 (0·198)	0·634 (0·174)
	PLS	0·849 (0·041)	0·841 (0·038)	0·483 (0·106)	0·593 (0·109)
Model (9)	PIRE	0·832 (0·057)	0·905 (0·028)	0·950 (0·016)	0·968 (0·008)
	PLS	0·747 (0·067)	0·830 (0·047)	0·981 (0·009)	0·997 (0·001)
Model (10)	PIRE	0·799 (0·046)	0·816 (0·045)	0·907 (0·019)	1·000 (0·000)
	PLS	0·191 (0·160)	0·225 (0·146)	0·320 (0·073)	0·902 (0·145)

PIRE, partial inverse regression; PLS, partial least squares.

where $p = 5$ and $\beta = (1, 1, 1, 1, 10)^{\mathsf{T}}$. The first four predictors are distributed independently as standard normal random variables. The fifth predictor x_5 was generated to be highly correlated with the fourth predictor x_4: $x_5 = x_4 + \delta e$, with the scalar parameter δ controlling the predictor correlation. Both ε and e are independent standard normal errors. Naik & Tsai (2000) considered only the case where $n > p$, with $n = 1000$ and $p = 5$. They compared sliced inverse regression and partial least squares when the correlation between x_4 and x_5 ranges from moderate to extremely large. They employed $|\hat{\beta}_5/\hat{\beta}_1|$ as their performance criterion, with the true value equal to 10. We employed the same set-up as theirs, except that the partial inverse regression estimator has been added. Figure 3 was constructed in the same manner as Fig. 1(a) of Naik & Tsai (2000), which shows the median of $|\hat{\beta}_5/\hat{\beta}_1|$ over 1000 replications for three estimators. The performance of sliced inverse regression and partial least squares that we observed agrees with that reported by Naik & Tsai (2000). In particular, sliced inverse regression was observed to break down for very small δ, since then the sample predictor covariance matrix is close to singular and its inverse becomes highly variable. By contrast, both partial methods are seen to be capable of dealing effectively with highly correlated predictors. The correlation between x_4 and x_5 at $\delta = 0{\cdot}005$ is about 0·99999. Naik & Tsai (2000) also studied a nonlinear link function in a set-up similar to model (11). Our results, not shown here, again verify their findings and demonstrate the effectiveness of partial inverse regression. Naik & Tsai (2000) warned about the poor performance of partial least squares for moderate sample size when the relationship between Y and X is nonlinear, which agrees well with our observations on the partial estimators when $n < p$.

3·5. *Application*: *spectroscopy of biscuit dough*

Finally, we applied the partial inverse regression estimator to a quantitative near-infrared spectroscopy study on the composition of biscuit dough. A full description of the experiment can be found in Osborne et al. (1984). The data were analyzed by Brown et al. (2001). They consist of 39 samples of dough pieces in a training set and 31 samples in an independent test set. A reflectance spectrum is available for each dough piece, measured from 1100 to 2498 nanometers in steps of 2. Following Brown et al. (2001), we reduced the number of spectral points in our analysis. The first 140 and last 49 wavelengths, which are believed to contain little information, were removed, with the resulting wavelength ranging from 1380 to 2400 nm. We then took one spectrum for every 8 points, thus increasing the spectral step

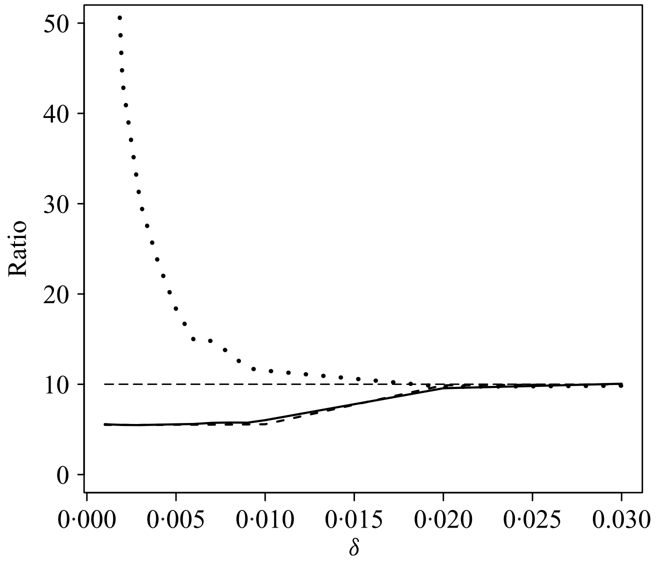

Fig. 3. Simulation study. Median of $|\hat\beta_5/\hat\beta_1|$ as a function of correlation between x_4 and x_5 in Model (11): solid line, partial inverse regression; dashed line, partial least squares; dotted line, sliced inverse regression; horizontal line, the true value of the ratio.

size to 16 nanometers and yielding 64 predictors. The response is the percentage of sucrose in the dough.

We applied partial inverse regression to the training data to estimate the linear combination of the predictors. The estimated q equals 6. We then employed a cubic smoothing spline to estimate the conditional mean $E(Y|\hat\beta^{\mathrm{T}}X)$. The correlation between the fitted and the actual responses is 0·97, which indicates a good fit to the training data. Based on the estimated predictor combination and the smoothing function obtained from the training data, we further predicted the response in the test data. The correlation between the predicted and the actual responses in the test set is 0·91, showing a reasonably accurate prediction. Figure 4 depicts the analysis, where $\hat\beta^{\mathrm{T}}X$ is shown on the horizontal axis and Y on the vertical axis, with the solid line indicating the nonparametrically estimated mean function $E(Y|\hat\beta^{\mathrm{T}}X)$. Figure 4(a), (b) and (c) show the estimates based on $q = 1$, the estimated value of 6, and 25, respectively, for the training data. When q is either too small or too big, the estimate is not satisfactory. Figure 4(d) shows the test data with the estimated $q = 6$.

4. Discussion

We have also conducted simulations to contrast the new approach with a method which uses the standard sliced inverse regression estimator, except that $\hat\Sigma_x^{-1}$ is replaced with a generalized inverse when $n < p$. Our results indicate that this method performs much worse than partial inverse regression. The reason for this can be traced back to our use of span$(\hat R_q^*)$, which serves as an estimated upper bound on the

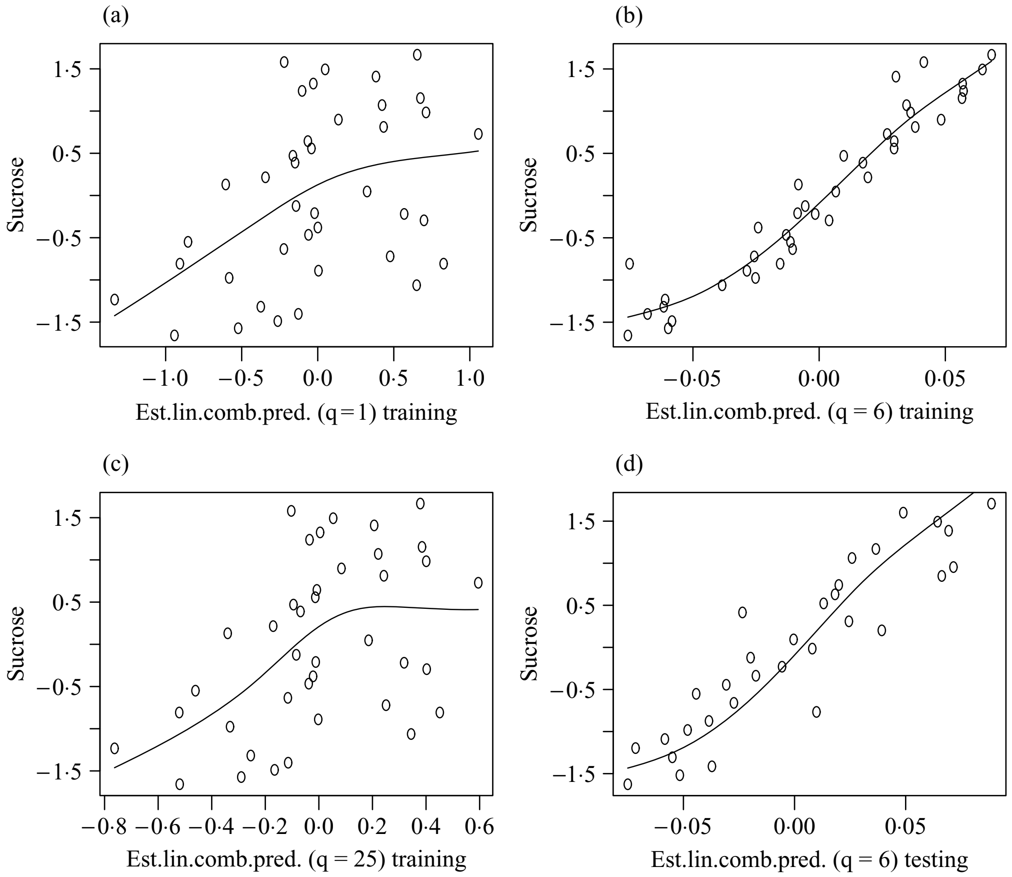

Fig. 4. Spectroscopy data. Scatterplots of sucrose as a function of extracted linear combination of predictors based on partial inverse regression. The solid line represents the cubic smoothing spline estimate.

central subspace, but is lost among the eigenvectors of $\hat{\Sigma}_x$ when using a generalized inverse.

Acknowledgement

We are grateful to the two referees and the editor for their constructive comments. We also thank Professors T. Fearn and M. Vannucci for providing us with the biscuit dough data. This research was supported in part by a grant from the U.S. National Science Foundation awarded to R.D.C.

References

Brown, P. J., Fearn, T. & Vannucci, M. (2001). Bayesian wavelet regression on curves with application to a spectroscopic calibration problem. *J. Am. Statist. Assoc.* **96**, 398–408.

Carroll, R. & Li, K. C. (1992). Measurement error regression with unknown link: Dimension reduction and visualization. *J. Am. Statist. Assoc.* **87**, 1040–50.

Cook, R. D. (1996). Graphics for regressions with a binary response. *J. Am. Statist. Assoc.* **91**, 983–92.

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

Cook, R. D. & Weisberg, S. (1991). Discussion of a paper by K.-C. Li. *J. Am. Statist. Assoc.* **86**, 328–32.

Craven, P. & Wahba, G. (1979). Smoothing noisy data with spline functions. *Numer. Math.* **31**, 377–403.

Eaton, M. (1986). A characterization of spherical distributions. *J. Mult. Anal.* **20**, 272–76.

Fan, J. & Gijbels, I. (1996). *Local Polynomial Modelling and its Application*. London: Chapman and Hall.

Fan, J. & Yao, Q. (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika* **85**, 645–60.

Garthwaite, P. H. (1994). An interpretation of partial least squares. *J. Am. Statist. Assoc.* **89**, 122–7.

Green, P. J. & Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. London: Chapman and Hall.

Hall, P. & Li, K. C. (1993). On almost linearity of low dimensional projections from high dimensional data. *Ann. Statist.* **21**, 867–89.

Härdle, W., Hall, P. & Ichimura, H. (1993). Optimal smoothing in single-index models. *Ann. Statist.* **21**, 157–78.

Helland, I. S. (1988). On the structure of partial least squares regression. *Commun. Statist.* B **17**, 581–607.

Helland, I. S. (1990). Partial least squares regression and statistical models. *Scand. J. Statist.* **17**, 97–114.

Helland, I. S. (1992). Maximum likelihood regression on relevant components. *J. R. Statist. Soc.* B **54**, 637–47.

Helland, I. S. & Almøy, T. (1994). Comparison of prediction methods when only a few components are relevant. *J. Am. Statist. Assoc.* **89**, 583–91.

Li, K. C. (1991). Sliced inverse regression for dimension reduction (with Discussion). *J. Am. Statist. Assoc.* **86**, 316–42.

Li, K. C. & Duan, N. (1989). Regression analysis under link violation. *Ann. Statist.* **17**, 1009–52.

Naik, P. & Tsai, C.-L. (2000). Partial least squares estimator for single-index models. *J. R. Statist. Soc.* B **62**, 763–71.

Naes, T. & Helland, I. S. (1993). Relevant components in regression. *Scand. J. Statist.* **20**, 239–50.

Osborne, B. G., Fearn, T., Miller, A. R. & Douglas, S. (1984). Application of near infrared reflectance spectroscopy to compositional analysis of biscuits and biscuit doughs. *J. Sci. Food Agric.* **35**, 99–105.

Stone, M. & Brooks, R. J. (1990). Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression (with Discussion). *J. R. Statist. Soc.* B **52**, 237–69.

Wold, H. (1975). Soft modelling by latent variables: The Nonlinear Partial Least Squares (NIPALS) approach. In *Perspectives in Probability and Statistics, Papers in Honour of M.S. Barlett*, Ed. J. Gani, pp. 117–42. London: Academic Press.

[*Received March* 2005. *Revised December* 2006]