J. R. Statist. Soc. B (2002)
64, *Part* 3, *pp.* 363–410

An adaptive estimation of dimension reduction space

Yingcun Xia,

University of Cambridge, UK, and Jinan University, People's Republic of China

Howell Tong,

University of Hong Kong, People's Republic of China, and London School of Economics and Political Science, UK

W. K. Li

University of Hong Kong, People's Republic of China

and Li-Xing Zhu

University of Hong Kong and Chinese Academy of Sciences, Beijing, People's Republic of China

[*Read before* The Royal Statistical Society *at a meeting organized by the* Research Section *on Wednesday, February 13th, 2002*, Professor D. Firth *in the Chair*]

Summary. Searching for an effective dimension reduction space is an important problem in regression, especially for high dimensional data. We propose an adaptive approach based on semiparametric models, which we call the (conditional) minimum average variance estimation (MAVE) method, within quite a general setting. The MAVE method has the following advantages. Most existing methods must *undersmooth* the nonparametric link function estimator to achieve a faster rate of consistency for the estimator of the parameters (than for that of the nonparametric function). In contrast, a faster consistency rate can be achieved by the MAVE method even without undersmoothing the nonparametric link function estimator. The MAVE method is applicable to a wide range of models, with fewer restrictions on the distribution of the covariates, to the extent that even time series can be included. Because of the faster rate of consistency for the parameter estimators, it is possible for us to estimate the dimension of the space consistently. The relationship of the MAVE method with other methods is also investigated. In particular, a simple outer product gradient estimator is proposed as an initial estimator. In addition to theoretical results, we demonstrate the efficacy of the MAVE method for high dimensional data sets through simulation. Two real data sets are analysed by using the MAVE approach.

Keywords: Average derivative estimation; Dimension reduction; Generalized linear models; Local linear smoother; Multiple time series; Non-linear time series analysis; Nonparametric regression; Principal Hessian direction; Projection pursuit; Semiparametrics; Sliced inverse regression estimation

1. Introduction

Let y and X be respectively $\mathbb{R}$-valued and $\mathbb{R}^p$-valued random variables. Without prior knowledge about the relationship between y and X, the regression function $g(x) = E(y|X = x)$ is often

Address for correspondence: Howell Tong, Department of Statistics, London School of Economics and Political Science, Houghton Street, London, WC2A 2AE, UK.
E-mail: h.tong@lse.ac.uk

 1369–7412/02/64363

modelled in a flexible nonparametric fashion. When the dimension of X is high, recent efforts have been expended in finding the relationship between y and X efficiently. The final goal is to approximate $g(x)$ by a function having simplifying structure which makes estimation and interpretation possible even for moderate sample sizes. There are essentially two approaches: the first is largely concerned with function approximation and the second with dimension reduction. Examples of the former are the additive model approach of Hastie and Tibshirani (1986) and the projection pursuit regression proposed by Friedman and Stuetzle (1981); both assume that the regression function is a sum of univariate smooth functions. Examples of the latter are the dimension reduction of Li (1991) and the regression graphics of Cook (1998).

A regression-type model for dimension reduction can be written as

$$y = g(B_0^{\mathrm{T}} X) + \varepsilon, \tag{1.1}$$

where g is an unknown smooth link function, $B_0 = (\beta_1, \ldots, \beta_D)$ is a $p \times D$ orthogonal matrix ($B_0^{\mathrm{T}} B_0 = I_{D\times D}$) with $D < p$ and $E(\varepsilon|X) = 0$ almost surely. The last condition allows ε to be dependent on X. When model (1.1) holds, the projection of the p-dimensional covariates X onto the D-dimensional subspace $B_0^{\mathrm{T}} X$ captures all the information that is provided by X on y. We call the D-dimensional subspace $B_0^{\mathrm{T}} X$ the effective dimension reduction (EDR) space. Li (1991) introduced the EDR space in a similar but more general context; the difference disappears for the case of additive noise as in model (1.1). See also Carroll and Li (1995), Chen and Li (1989) and Cook (1994). Note that the space spanned by the column vectors of B_0 is uniquely defined under some mild conditions (given in Section 3) and is our focus of interest. For convenience, we shall refer to these column vectors as EDR directions, which are unique up to orthogonal transformations. The estimation of the EDR space includes the estimation of the directions, namely B_0, and the corresponding dimension of the EDR space. For specific semiparametric models, methods have been introduced to estimate B_0. Next, we give a brief review of these methods.

One of the important approaches is the projection pursuit regression proposed by Friedman and Stuetzle (1981). Huber (1985) has given a comprehensive discussion. Chen (1991) has investigated a projection pursuit type of regression model. The primary focus of projection pursuit regression is more on the approximation of $g(x)$ by a sum of ridge functions $g_k(\cdot)$, namely

$$g(X) \approx \sum_{k=1}^{D} g_k(\beta_k^{\mathrm{T}} X),$$

than on looking for the EDR space.

A simple approach that is directly related to the estimation of EDR directions is the average derivative estimation (ADE) proposed by Härdle and Stoker (1989). For the single-index model $y = g_1(\beta_1^{\mathrm{T}} X) + \varepsilon$, the expectation of the gradient $\bigtriangledown g_1(X)$ is a scalar multiple of β_1. A nonparametric estimator of $\nabla g_1(X)$ leads to an estimator of β_1. There are several limitations of ADE.

(a) To estimate β_1, the condition $E\{g_1'(\beta_1^{\mathrm{T}} X)\} \neq 0$ is needed. This condition is violated when $g_1(\cdot)$ is an even function and X is symmetrically distributed.
(b) As far as we know, there is no successful extension to the case of more than one EDR direction.

The sliced inverse regression (SIR) method proposed by Li (1991) is perhaps up to now the most powerful method for searching for EDR directions and dimension reduction. However, the SIR method imposes some strong probabilistic structure on X. Specifically, the method requires

that, for any constant vector $b^{\mathrm{T}} = (b_1, \ldots, b_p)$, there are constants c_0 and $c^{\mathrm{T}} = (c_1, \ldots, c_D)$ depending on b such that, for the directions B_0 in model (1.1),

$$E(b^{\mathrm{T}} X | B_0^{\mathrm{T}} X) = c_0 + c^{\mathrm{T}} B_0^{\mathrm{T}} X. \tag{1.2}$$

As pointed out by Cook and Weisberg in their discussion of Li (1991), the most important family of distributions satisfying condition (1.2) is that of elliptically symmetric distributions. Now, in time series analysis we typically set $X = (y_{t-1}, \ldots, y_{t-p})^{\mathrm{T}}$, where $\{y_t\}$ is a time series. Then it is easy to prove that elliptical symmetry of X for all p with (second-order) stationarity of $\{y_t\}$ implies that $\{y_t\}$ is time reversible, a feature which is the exception rather than the rule in time series analysis. (For a discussion of time reversibility, see, for example, Tong (1990).)

Another aspect of searching for the EDR space is the determination of the corresponding dimension. The method proposed by Li (1991) can be applied to determine the dimension of the EDR space in some cases but for reasons mentioned above it is typically not relevant for time series data.

In this paper, we shall propose a new method to estimate the EDR directions. We call it the (conditional) minimum average variance estimation (MAVE) method. Our approach is inspired by the SIR method, the ADE method and the idea of local linear smoothers (see, for example, Fan and Gijbels (1996)). It is easy to implement and needs no strong assumptions on the probabilistic structure of X. Specifically, our methods apply to model (1.1) including its generalization within the additive noise set-up. The joint density function of covariate X is needed if we search for the EDR space globally. However, if we have some prior information about the EDR directions and we look for them locally, then existence of density of X in the directions around EDR directions will suffice. These cases include those in which some of the covariates are categorical or functionally related. The observations need not be independent, e.g. time series data. On the basis of the properties of the MAVE method, we shall propose a method to estimate the dimension of the EDR space, which again does not require strong assumptions on the design X and has wide applicability.

Let Z be an $\mathbb{R}^q$-valued random variable. A general semiparametric model can be written as

$$y = G\{\phi(B_0^{\mathrm{T}} X), Z, \theta\} + \varepsilon, \tag{1.3}$$

where G is a *known* smooth function up to a parameter vector $\theta \in \mathbb{R}^l$, $\phi(\cdot)$: $\mathbb{R}^D \mapsto \mathbb{R}^{D'}$ is an *unknown* smooth function and $E(\varepsilon | X, Z) = 0$ almost surely. Special cases are the generalized partially linear single-index model of Carroll *et al.* (1997) and the single-index functional coefficient model in Xia and Li (1999). Searching for the EDR space $B_0^{\mathrm{T}} X$ in model (1.3) is of theoretical as well as practical interest. However, the existing methods are not always appropriate for this model. An extension of our method to handle this model will be discussed.

The rest of this paper is organized as follows. Section 2 describes the MAVE procedure and gives some results. Section 3 discusses some comparisons with existing methods and proposes a simple average outer product of gradients (OPG) estimation method and an inverse MAVE method. To check the feasibility of our approach, we have conducted many simulations, typical ones of which are reported in Section 4. In Section 5 we study the circulatory and respiratory data of Hong Kong and the hitters' salary data of the USA using the MAVE methodology. In practice, we standardize our observations. Appendix A establishes the efficiency of the algorithm proposed. Some of our theoretical proofs are very lengthy and not included here. However, they are available on request from the authors. Finally, the programs are available at

`http://www.blackwellpublishers.co.uk/rss/`

2. Estimation of effective dimension reduction space

2.1. The estimation of effective dimension reduction directions

Let us denote the working dimension by d with $1 \leqslant d \leqslant p$. Therefore, we need to estimate only a set of orthogonal vectors. There are many related methods for this and similar purposes. Most of the existing methods adopt two separate cost functions. The first is used to estimate the link function and the second the directions based on the estimated link function. See, for example, Hall (1989), Härdle and Stoker (1989) and Carroll *et al.* (1997). It is therefore not surprising that the performance of the direction estimator suffers from the bias problem in nonparametric estimation. Härdle *et al.* (1993) noticed this and overcame the problem for a single-index model by minimizing a cross-validation-type sum of squares of the residuals *simultaneously* with respect to the bandwidth and the directions. However, the cross-validation-type sum of squares of residuals affects the performance of estimation. See Xia *et al.* (1999). Moreover, the minimization is not trivial. Härdle *et al.* (1993) used the grid search method in their simulations, which is quite inefficient when the dimension is high.

Consider the simple regression model (1.1). The direction B_0 is the solution of

$$\min_B [E\{y - E(y|B^{\mathrm{T}}X)\}^2]. \tag{2.1}$$

For any orthogonal matrix $B = (\beta_1, \ldots, \beta_d)$, the conditional variance given $B^{\mathrm{T}}X$ is

$$\sigma_B^2(B^{\mathrm{T}}X) = E[\{y - E(y|B^{\mathrm{T}}X)\}^2|B^{\mathrm{T}}X]. \tag{2.2}$$

It follows that

$$E\{y - E(y|B^{\mathrm{T}}X)\}^2 = E\{\sigma_B^2(B^{\mathrm{T}}X)\}.$$

Therefore, minimizing expression (2.1) is equivalent to minimizing, with respect to B,

$$E\{\sigma_B^2(B^{\mathrm{T}}X)\} \qquad \text{subject to } B^{\mathrm{T}}B = I. \tag{2.3}$$

We shall call this MAVE. Suppose that $\{(X_i, y_i)\ i = 1, 2, \ldots, n\}$ is a sample from (X, y). Let

$$g_B(v_1, \ldots, v_d) = E(y|\beta_1^{\mathrm{T}}X = v_1, \ldots, \beta_d^{\mathrm{T}}X = v_d).$$

For any given X_0, a local linear expansion of $E(y_i|B^{\mathrm{T}}X_i)$ at X_0 is

$$E(y_i|B^{\mathrm{T}}X_i) \approx a + b^{\mathrm{T}}B^{\mathrm{T}}(X_i - X_0), \tag{2.4}$$

where $a = g_B(B^{\mathrm{T}}X_0)$ and $b^{\mathrm{T}} = (b_{(1)}, \ldots, b_{(d)})$ with

$$b_{(k)} = \frac{\partial g_B(v_1, \ldots, v_d)}{\partial v_k}\bigg|_{v_1 = \beta_1^{\mathrm{T}}X_0, \ldots, v_d = \beta_d^{\mathrm{T}}X_0}, \qquad k = 1, \ldots, d.$$

Note that the right-hand side of approximation (2.4) is the tangent plane of g_B at $B^{\mathrm{T}}X_0$. The residuals are then

$$y_i - g_B(B^{\mathrm{T}}X_i) \approx y_i - \{a + b^{\mathrm{T}}B^{\mathrm{T}}(X_i - X_0)\}.$$

Following the idea of local linear smoothing estimation, we can estimate $\sigma_B^2(B^{\mathrm{T}}X_0)$ by exploiting the approximation

$$\sum_{i=1}^{n}\{y_i - E(y_i|B^{\mathrm{T}}X_i)\}^2 w_{i0} \approx \sum_{i=1}^{n}[y_i - \{a + b^{\mathrm{T}}B^{\mathrm{T}}(X_i - X_0)\}]^2 w_{i0}, \tag{2.5}$$

where $w_{i0} \geqslant 0$ are some weights with $\Sigma_{i=1}^{n} w_{i0} = 1$ and typically centred at $B^{\mathrm{T}}X_0$. The choice of the weights w_{i0} plays a key role in searching for the EDR directions. We shall discuss this issue in detail later. Usually,

$$w_{i0} = K_h\{B^{\mathrm{T}}(X_i - X_0)\} \Big/ \sum_{l=1}^{n} K_h\{B^{\mathrm{T}}(X_l - X_0)\},$$

where $K_h(\cdot) = h^d K(\cdot/h)$ and d is the dimension of $K(\cdot)$. For ease of exposition, $K(\cdot)$ denotes different kernel functions at different places. The estimators of a and b are just the minimum point of approximation (2.5). Therefore, the estimator of σ_B^2 at $B^{\mathrm{T}}X_0$ is just the minimum value of expression (2.5), namely

$$\hat{\sigma}_B^2(B^{\mathrm{T}}X_0) = \min_{a,b}\left(\sum_{i=1}^{n}[y_i - \{a + b^{\mathrm{T}}B^{\mathrm{T}}(X_i - X_0)\}]^2 w_{i0}\right). \tag{2.6}$$

Under some mild conditions, we have $\hat{\sigma}_B^2(B^{\mathrm{T}}X_0) - \sigma_B^2(B^{\mathrm{T}}X_0) = o_P(1)$. On the basis of expressions (2.1), (2.3) and (2.6), we can estimate the EDR directions by solving the minimization problem

$$\min_{B:B^{\mathrm{T}}B=I}\left\{\sum_{j=1}^{n}\hat{\sigma}_B^2(B^{\mathrm{T}}X_j)\right\} = \min_{\substack{B:B^{\mathrm{T}}B=I \\ a_j,b_j,j=1,\ldots,n}}\left(\sum_{j=1}^{n}\sum_{i=1}^{n}[y_i - \{a_j + b_j^{\mathrm{T}}B^{\mathrm{T}}(X_i - X_j)\}]^2 w_{ij}\right), \tag{2.7}$$

where $b_j^{\mathrm{T}} = (b_{j1}, \ldots, b_{jd})$. The MAVE method or the minimization in problem (2.7) can be seen as a combination of nonparametric function estimation and direction estimation, which is executed simultaneously with respect to the directions and the nonparametric link function. As we shall see, we benefit from this simultaneous minimization.

If the weights depend on B, the implementation of the minimization in problem (2.7) is non-trivial. The weight w_{i0} in approximation (2.5) should be chosen such that *the value of* w_{i0} *is a function of the distance between* X_i *and* X_0. Next, we give two choices of w_{i0}.

2.1.1. Multidimensional kernel weight

To simplify problem (2.7), a natural choice is

$$w_{i0} = K_h(X_i - X_0) \Big/ \sum_{l=1}^{n} K_h(X_l - X_0).$$

This kind of weight can be used as an initial step of estimation. Given d, we obtain a set of directions $\hat{B}$ via the minimization in problem (2.7). Let $\mathcal{S}(\hat{B})$ denote the subspace spanned by the column vectors of $\hat{B}$. The distance between the space $\mathcal{S}(B_0)$, the space spanned by the column vectors of B_0, and the space $\mathcal{S}(\hat{B})$ can be measured by $\|(I - B_0 B_0^{\mathrm{T}})\hat{B}\|$ if $d < D$ and $\|(I - \hat{B}\hat{B}^{\mathrm{T}})B_0\|$ if $d \geqslant D$. Here and later, obvious augmentations by zero vectors are understood and the distance is denoted by $m(\hat{B}, B_0)$.

Theorem 1. Suppose that conditions 1–6 (in Appendix A) hold, model (1.1) is true and as $n \to \infty$ both $nh^p/\log(n) \to \infty$ and $h \to 0$. If $d < D$, then

$$m(\hat{B}, B_0) = O_P(h^2 + h^{-1}\delta_n^2),$$

where $\delta_n = \{\log(n)/nh^p\}^{1/2}$. If $d \geqslant D$, then

$$m(\hat{B}, B_0) = O_P(h^3 + h^{-1}\delta_n^2).$$

Provided that the dimension is chosen correctly, the rate of consistency for $\hat{B}$ is $O_P\{h_{\mathrm{opt}}^3 \log(n)\}$ if we use the optimal bandwidth h_{opt} of the regression function estimation in the sense of minimizing the mean integrated squared errors. This is faster than the rate that is achieved by the other methods, which is $O_P(h_{\mathrm{opt}}^2)$. Note that the consistency rate for the local linear estimator of the link function is also $O_P(h_{\mathrm{opt}}^2)$. The faster rate is due to minimizing the average (conditional) variance with respect to both directions and the local linearization of the link function. Moreover, if we extend the idea to higher order local polynomial smoothers, root n consistency for the estimator of B_0 can be achieved; see the discussion in Section 6.

2.1.2. *Refined kernel weight*

If we know the dimension of the EDR space, which is usually less than p, we can then search for the EDR directions in a lower dimensional space, thereby reducing the effect of high dimension and improving the accuracy of the estimation. Suppose that we have an initial estimator of B_0, say $\hat{B}$. Let

$$\tilde{w}_{ij} = K_h\{\hat{B}^{\mathrm{T}}(X_i - X_j)\} \Big/ \sum_{l=1}^{n} K_h\{\hat{B}^{\mathrm{T}}(X_l - X_j)\}. \tag{2.8}$$

Re-estimate B_0 by the minimization in problem (2.7) with weights $\tilde{w}_{ij}$ replacing w_{ij}. By an abuse of notation, we denote the new estimator of B_0 by $\hat{B}$ also. Replace $\hat{B}$ in equation (2.8) by the latest $\hat{B}$ and estimate B_0. Repeat this procedure until $\hat{B}$ converges; we call the limit the refined MAVE (RMAVE) estimator. Results similar to those of theorem 1 can be obtained. We here use a lower dimensional kernel and the bandwidth now is smaller than that used in the multidimensional w_{ij}, leading to a faster rate of consistency.

One of the referees has drawn our attention to an unpublished paper by W. H. Wong and X. Shen, who have been working on a similar problem. They have proposed the nearest neighbour method and used the weights

$$w_{ij} = \frac{1}{n}\mathbf{1}_{\{X_i \text{ is one of the } \mathcal{N} \text{ nearest observations to } X_j\}},$$

where $\mathcal{N} < n$ is a suitable integer and $\mathbf{1}_A$ denotes the indicator function of the set A.

2.2. *Dimension of effective dimension reduction space*

Methods have been proposed for the determination of the number of the EDR directions. See, for example, Li (1992), Schott (1994) and Cook (1998). Their approaches tend to be based on similar probabilistic assumptions on the covariates X imposed by SIR. We now propose an alternative approach within our set-up. It is well known that a cross-validation approach penalizes the complexity of the model. See, for example, Stone (1974). We now extend the cross-validation method of Cheng and Tong (1992) and Yao and Tong (1994) to solve the above problem. A similar extension may be effected by using the approach of Auestad and Tjøstheim (1990), which is asymptotically equivalent to the cross-validation method.

Suppose that $\beta_1, \ldots, \beta_D$ are the EDR directions, i.e. $y = g(\beta_1^{\mathrm{T}}X, \ldots, \beta_D^{\mathrm{T}}X) + \varepsilon$ with $E(\varepsilon|X) = 0$ almost surely. If $D < p$, we can nominally extend the number of directions to p, say $\{\beta_1, \ldots, \beta_D, \ldots, \beta_p\}$, such that they are perpendicular to one another. Now, the problem becomes the selection of the covariates among $\{\beta_1^{\mathrm{T}}X, \ldots, \beta_p^{\mathrm{T}}X\}$. However, because $\beta_1, \ldots, \beta_p$ are unknown, we must replace β_ks by their estimators $\hat{\beta}_k$s. As we have proved that the rate of consistency of

the $\hat{\beta}_k$s is faster than that of the nonparametric link function estimators, the replacement is justified. Let

$$\hat{a}_{d0,j} = \sum_{i=1, i\neq j}^{n} K_{h_d}^{(i,j)} y_i \Big/ \sum_{i=1, i\neq j}^{n} K_{h_d}^{(i,j)},$$

where $K_{h_d}^{(i,j)} = K_{h_d}\{\hat{\beta}_1^{\mathrm{T}}(X_i - X_j), \ldots, \hat{\beta}_d^{\mathrm{T}}(X_i - X_j)\}$. Here, we use the suffix d to highlight the fact that the bandwidth depends on the working dimension d. Let

$$\mathrm{CV}(d) = n^{-1} \sum_{j=1}^{n} (y_j - \hat{a}_{d0,j})^2, \qquad d = 1, \ldots, p.$$

Suppose that model (1.1) holds and $B_d^{\mathrm{T}} X$ has a density $f_d(v_1, \ldots, v_d)$ with compact support, where $B_d = (\beta_1, \ldots, \beta_d)$. For ease of exposition, we temporarily abbreviate $g(v_1, \ldots, v_D)$ to $g(v)$ and $f_d(v_1, \ldots, v_d)$ to $f_d(v)$. When $d \geqslant D$, we have

$$\mathrm{CV}(d) = \sigma^2 + h_d^4 \int \left[\frac{1}{2}\mathrm{tr}\{\nabla^2 g(v)\} + f_d^{-1}(v)\, \nabla^{\mathrm{T}} g(v)\, \nabla\, f_d(v)\right]^2 f_d(v)\, \mathrm{d}v_1 \ldots \mathrm{d}v_d$$
$$+ \frac{\alpha_d}{n h_d^d}\{1 + o_P(1)\} + O_P(n^{-1/2} + h_d^5),$$

where $\sigma^2 = \mathrm{var}(\varepsilon)$,

$$\alpha_d = E\{E(\varepsilon^2 | B^{\mathrm{T}} X)/ f_d(B^{\mathrm{T}} X)\} \int K^2(v_1, \ldots, v_d)\, \mathrm{d}v_1 \ldots \mathrm{d}v_d$$

and $\nabla^2 g(v)$ is a $d \times d$ matrix whose (i, j)th element is $\partial^2 g(v)/\partial v_i \partial v_j$. If h_d is monotonic increasing such that $h_{d+1}^{d+1} = o(h_d^d)$, then $\mathrm{CV}(d)$ increases with d. Note that the optimal bandwidth $h_d \sim n^{-1/(d+4)}$ satisfies this requirement. When $d < D$, it is not difficult to see that $\mathrm{CV}(d) > \mathrm{CV}(D)$ because of the lack of fit. To include the case that y and X are independent, we define

$$\mathrm{CV}(0) = n^{-1} \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

It is easy to see that $\mathrm{CV}(0) = \sigma^2 + O_P(n^{-1/2})$. Thus, we estimate the dimension of EDR space as

$$\hat{d} = \arg \min_{0 \leqslant d \leqslant p} \{\mathrm{CV}(d)\}.$$

Theorem 2. Suppose that the assumptions 1–6 (in Appendix A) hold. Under model (1.1) with X having a density with compact support, we have

$$\hat{d} \to D \qquad \text{in probability}$$

If X is not bounded, we may consider only a compact domain over which the density is positive. Then we have a small probability of overestimating the dimension (Cheng and Tong, 1992; Yao and Tong, 1994). Note that $a_{d0,j}$ is the Nadaraya–Watson estimator of a. We can use alternatively the local linear estimator for $a_{d0,j}$, which also leads to a consistent $\hat{d}$. However, the local linear estimator involves more complicated computation. Moreover, as far as cross-validatory determination of the dimension is concerned, our experience shows that using the local linear estimator tends to lead to a poorer performance in comparison with using the Nadaraya–Watson estimator. Empirical evidence suggests that using the latter tends to incur a smaller bandwidth and to lead to a heavier penalty for overfitting.

2.3. Bandwidth and algorithm

An important feature of the MAVE method is that we do not need to undersmooth the link function estimator for the EDR direction estimator to achieve a higher rate of consistency than the former. Therefore, the optimal bandwidth in the sense of mean integrated squared error can be used and, in practice, a variable bandwidth is normally recommended, e.g. (in obvious notation)

$$K_{\mathbf{h}}(u_1, \ldots, u_d) = \frac{K(u_1/h_{(1)}, \ldots, u_d/h_{(d)})}{\prod_{k=1}^{d} h_{(k)}},$$

$\mathbf{h} = (h_{(1)}, \ldots, h_{(d)})$ and d is the dimension of $K(\cdot)$. There are many ways to obtain such a bandwidth $\mathbf{h}$. See, for example, Fan and Gijbels (1996) and Yang and Tschernig (1999).

Our search procedure is as follows.

Step 1 (directions): for each d, $1 \leqslant d \leqslant p$, we search for the d directions as follows.

(a) Initial value: use the multidimensional kernel weight to obtain an initial estimate of possible EDR directions $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_d$ by minimizing problem (2.7).
(b) Refined estimation: let $\hat{B} = (\hat{\beta}_1, \ldots, \hat{\beta}_d)$ constitute the latest estimator of B. Therefore we obtain refined kernel weights by using equation (2.8). We refine the estimator via expression (2.7) using the refined kernel weights. Continue this procedure until convergence. The CV(d) values can be obtained by using the final estimators of the directions.

Step 2 (dimension and output results): compare the CV(d), $0 \leqslant d \leqslant p$. The d with the smallest CV(d) value is the estimated dimension. The corresponding estimator of B in step 1(b) gives the estimated EDR directions.

Let $\hat{B}_a$ and $\hat{B}_b$ be the estimators of B in two adjacent iterations in step 1(b). A suggested stopping rule for step 1(b) is when the distances $m(\hat{B}_a, \hat{B}_b)$ in several adjacent iterations are each less than a pre-set tolerance. Next, we describe one method to implement the minimization in problem (2.7). For any d, let $B = (\beta_1, \ldots, \beta_d)$ be the initial value (set $\beta_1 = \beta_2 = \ldots = \beta_d = 0$ in step 1(a)). $B_{l,k} = (\beta_1, \ldots, \beta_{k-1})$ and $B_{r,k} = (\beta_{k+1}, \ldots, \beta_d)$, $k = 1, 2, \ldots, d$. Minimize

$$S_{n,k} = \sum_{j=1}^{n} \sum_{i=1}^{n} \left\{ y_i - a_j - (X_i - X_j)^{\mathrm{T}} (B_{l,k}, \beta, B_{r,k}) \begin{pmatrix} c_j \\ d_j \\ e_j \end{pmatrix} \right\}^2 w_{ij}$$

$$\text{subject to } B_{l,k}^{\mathrm{T}} \beta = 0 \text{ and } B_{r,k}^{\mathrm{T}} \beta = 0,$$

where c_j is a $(k-1) \times 1$ vector, d_j a scalar and e_j a $(d-k) \times 1$ vector. This is a typical constrained quadratic programming problem. See, for example, Rao (1973), page 232. Let

$$C_j = \sum_{i=1}^{n} w_{ij}(X_i - X_j),$$

$$D_j = \sum_{i=1}^{n} w_{ij}(X_i - X_j)(X_i - X_j)^{\mathrm{T}},$$

$$E_j = \sum_{i=1}^{n} w_{ij} y_i,$$

$$F_j = \sum_{i=1}^{n} w_{ij}(X_i - X_j) y_i.$$

With β given, the (a_j, c_j, d_j, e_j) which minimizes $S_{n,k}$ is given by

$$\begin{pmatrix} a_j \\ c_j \\ d_j \\ e_j \end{pmatrix} = \begin{pmatrix} 1 & C_j^{\mathrm{T}}(B_{l,k}, \beta, B_{r,k}) \\ (B_{l,k}, \beta, B_{r,k})^{\mathrm{T}} C_j & (B_{l,k}, \beta, B_{r,k})^{\mathrm{T}} D_j (B_{l,k}, \beta, B_{r,k}) \end{pmatrix}^{-1} \times \begin{pmatrix} E_j \\ (B_{l,k}, \beta, B_{r,k})^{\mathrm{T}} F_j \end{pmatrix},$$

$j = 1, \ldots, n$. If a_j, c_j, d_j and e_j are given, then the β which minimizes $S_{n,k}$ is given by

$$\begin{pmatrix} \beta \\ \lambda \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^{n} d_j^2 D_j & \tilde{B}_k \\ \tilde{B}_k^{\mathrm{T}} & 0 \end{pmatrix}^{+} \begin{pmatrix} \sum_{j=1}^{n} d_j \left\{ F_j - a_j C_j - D_j \tilde{B}_k \begin{pmatrix} c_j \\ e_j \end{pmatrix} \right\} \\ 0 \end{pmatrix},$$

where $\tilde{B}_k = (B_{l,k}, B_{r,k})$ and A^+ denotes the Moore–Penrose inverse of a matrix A. Here λ is the usual Lagrangian multiplier for the constraint minimization. Finally, we normalize β.

3. Links with other methods and generalization

3.1. Outer product of gradients estimation

Suppose that $y = \tilde{g}(X) + \varepsilon$ with $E(\varepsilon|X) = 0$ almost surely. Consider the minimization in problem (2.6). Under assumptions 1–6 (in Appendix A) and

$$w_{i0} = K_h(X_i - X_0) \Big/ \sum_{l=1}^{n} K_h(X_l - X_0),$$

we have

$$\min_{a,b} \left(\sum_{i=1}^{n} [y_i - \{a + b^{\mathrm{T}} B^{\mathrm{T}} (X_i - X_0)\}]^2 w_{i0} \right) = \hat{\sigma}^2(X_0) + h^2 \bigtriangledown^{\mathrm{T}} \tilde{g}(X_0) \times (I_{p\times p} - BB^{\mathrm{T}}) \bigtriangledown \tilde{g}(X_0) + o_P(h^2),$$

where $\hat{\sigma}^2(X_0) = \Sigma_{i=1}^{n} \varepsilon_i^2 w_{i0}$ does not depend on B. Thus, the minimization problem (2.7) depends mainly on

$$\begin{aligned} E\{\bigtriangledown^{\mathrm{T}} \tilde{g}(X)(I_{p\times p} - BB^{\mathrm{T}}) \bigtriangledown \tilde{g}(X)\} &= \mathrm{tr}[(I_{p\times p} - BB^{\mathrm{T}}) E\{\bigtriangledown \tilde{g}(X) \bigtriangledown^{\mathrm{T}} \tilde{g}(X)\}] \\ &= \mathrm{tr}[E\{\bigtriangledown \tilde{g}(X) \bigtriangledown^{\mathrm{T}} \tilde{g}(X)\}] - \mathrm{tr}[B^{\mathrm{T}} E\{\bigtriangledown \tilde{g}(X) \bigtriangledown^{\mathrm{T}} \tilde{g}(X)\} B]. \end{aligned}$$

Therefore, the B which minimizes this equation is the first d eigenvectors corresponding to the d largest eigenvalues of $E\{\bigtriangledown \tilde{g}(X) \bigtriangledown^{\mathrm{T}} \tilde{g}(X)\}$, which is the average OPG of $\tilde{g}(\cdot)$.

Lemma 1. Suppose that $\tilde{g}(\cdot)$ is differentiable. If model (1.1) is true, then B_0 is in the space spanned by the first D eigenvectors of $E[\bigtriangledown \tilde{g}(X) \bigtriangledown^{\mathrm{T}} \tilde{g}(X)]$ corresponding to the largest D eigenvalues.

This relationship was also noticed in Li (1991). By lemma 1, it is easy to see that the EDR space is unique up to orthogonal transformations if the density function of X has a compact support. We may use lemma 1 and propose the following estimation procedure. First, estimate

the gradients by local polynomial smoothing. Specifically, we consider the local linear fitting in the form of the minimization problem

$$\min_{a_j,b_j}\left(\sum_{i=1}^{n}[\{y_i - a_j - b_j^{\mathrm{T}}(X_i - X_j)\}^2 w_{ij}]\right). \tag{3.1}$$

We then estimate $E\{\nabla\tilde{g}(X)\ \nabla^{\mathrm{T}}\ \tilde{g}(X)\}$ by

$$\hat{\Sigma} = \frac{1}{n}\sum_{j=1}^{n}\hat{b}_j\hat{b}_j^{\mathrm{T}},$$

where $\hat{b}_j$ is the minimizer from expression (3.1). Finally, we estimate the EDR directions by the first d eigenvectors of $\hat{\Sigma}$. We call this method the method of OPG estimation.

Theorem 3. Let $\hat{\beta}_1, \ldots, \hat{\beta}_d$ be the first d eigenvectors of $\hat{\Sigma}$ corresponding to the largest d eigenvalues, and $\hat{B} = (\hat{\beta}_1, \ldots, \hat{\beta}_d)$. Suppose that conditions 1–6 (in Appendix A) hold and model (1.1) is true. If $nh^p/\log(n) \to \infty$ and $h \to 0$, then

$$m(\hat{B}, B_0) = O_P(h^2 + \delta_n^2 h^{-1}).$$

Unlike the ADE method, the OPG method still works even if $E\{\nabla\tilde{g}(X)\} = 0$. Moreover, the OPG method can handle multiple EDR directions simultaneously whereas the ADE method can only handle the first EDR direction (i.e. the single-index model). We can further refine the OPG estimator using refined weights as in the RMAVE method. Compared with the MAVE method, the OPG method still suffers from the effect of the bias term in nonparametric function estimation. Therefore, the rate of consistency is slower than that of the MAVE method when the dimension is chosen correctly. However, the OPG method is easy to implement and can be used as an initial value of other estimation methods. Li (1992) proposed the principal Hessian directions (PHD) method by estimating the Hessian matrix of $g(\cdot)$. Similarly to the OPG method, the directions are the eigenvectors of the Hessian matrix. For a normally distributed design X, the Hessian matrix can be properly estimated simply by Stein's lemma. However, the PHD method assumes some probabilistic structure on design X which is frequently violated in time series analysis. More fundamentally, the PHD method involves estimators of second derivatives whereas the OPG method involves only the first derivatives, which are considerably simpler and easier to estimate.

3.2. Inverse regression minimum average (conditional) variance estimation

We start with

$$\bar{w}_{ij} = K_h(y_i - y_j)\Big/\sum_{l=1}^{n}K_h(y_l - y_j). \tag{3.2}$$

Now, with this weight function, the minimization in equation (2.6) becomes the minimization of

$$\sum_{i=1}^{n}[y_i - \{a + b\beta^{\mathrm{T}}(X_i - X_0)\}]^2\bar{w}_{i0},$$

and the MAVE method involves the minimization of

$$\sum_{i=1}^{n}\sum_{j=1}^{n}[y_i - \{a_j + b_j\beta^{\mathrm{T}}(X_i - X_j)\}]^2\bar{w}_{ij}.$$

A 'dual' of this is the minimization of

$$\sum_{j=1}^{n}\sum_{i=1}^{n}\{\beta^{\mathrm{T}}X_i - c_j - d_j(y_i - y_j)\}^2\bar{w}_{ij}. \tag{3.3}$$

This may be considered an alternative derivation of the SIR method. The extension of expression (3.3) to more than one direction can be stated as follows. Suppose that the first k directions have been calculated and are denoted by $\hat{\beta}_1, \ldots, \hat{\beta}_k$ respectively. To obtain the $(k+1)$th direction, we need to perform

$$\min_{\alpha_{1j},\ldots,\alpha_{kj},c_j,d_j,\beta}\left[\sum_{j=1}^{n}\sum_{i=1}^{n}\{\beta^{\mathrm{T}}X_i + \alpha_{1j}\hat{\beta}_1^{\mathrm{T}}X_i + \ldots + \alpha_{kj}\hat{\beta}_k^{\mathrm{T}}X_i - c_j - d_j(y_i - y_j)\}^2\bar{w}_{ij}\right]$$
$$\text{subject to } \beta^{\mathrm{T}}(\hat{\beta}_1, \ldots, \hat{\beta}_k) = 0 \text{ and } \|\beta\| = 1. \tag{3.4}$$

We call the estimation method based on minimizing expression (3.3) with $\bar{w}_{ij}$ as defined in equation (3.2) the inverse MAVE (IMAVE) method. The IMAVE method is in line with the most predictable variate (Hotelling, 1935). The minimizations in expressions (3.3) and (3.4) can be seen as looking for linear combinations of X that are most predictable from y. Under a similar assumption on X as in SIR, we have the following result.

Theorem 4. Suppose that equation (1.2) and assumptions 1, 2(b), 3(b), 4, 5(b) and 6 (in Appendix A) hold. Let $\hat{b} = (\hat{\beta}_1, \ldots, \hat{\beta}_d)$. If $h \to 0$ and $nh/\log(n) \to \infty$, then

$$m(\hat{B}, B_0) = O_P\{h^2 + \log(n)/nh + n^{-1/2}\}.$$

This result is similar to that of Zhu and Fang (1996). As noted previously, the assumption on the design X can be a handicap as far as applications of the IMAVE method are concerned. Interestingly, simulations show that the SIR method and the IMAVE method can sometimes produce useful results in the case of independent data even when this assumption is mildly violated. However, for time series data, we find that this is often not so.

3.3. Semiparametric multi-index models

Consider the general model (1.3). Suppose that $G(v, Z, \theta)$ is differentiable. For ease of exposition we set $D' = 1$. Let $G'(v, Z, \theta) = \partial G(v, Z, \theta)/\partial v$. For $B^{\mathrm{T}}X_i$ close to $B^{\mathrm{T}}X_0$ we have

$$G\{\phi(B^{\mathrm{T}}X_i), Z_i, \theta\} \approx G\{\phi(B^{\mathrm{T}}X_0), Z_i, \theta\} + G'\{\phi(B^{\mathrm{T}}X_0), Z_i, \theta\} \bigtriangledown^{\mathrm{T}} \phi(B^{\mathrm{T}}X_0)B^{\mathrm{T}}(X_i - X_0).$$

To estimate B, we minimize

$$\sum_{j=1}^{n}\sum_{i=1}^{n}\{y_i - G(a_j, Z_i, \theta) - G'(a_j, Z_i, \theta)b_j^{\mathrm{T}}B^{\mathrm{T}}(X_i - X_j)\}^2 w_{ij}$$

with respect to a_j, b_j, $j = 1, \ldots, n$, θ and B. Similarly, we may first use the multidimensional kernel weight to obtain an initial estimate and then repeatedly use the refined kernel weight.

Model (1.3) includes many models with a fixed dimension of EDR space. Examples are the single-index model of Ichimura and Lee (1991), the generalized partially linear single-index model of Carroll *et al.* (1997) and Xia *et al.* (1999) and the single-index coefficient regression model of Xia and Li (1999). Here the estimation of the unknown function is also important. An obvious question is whether we can estimate both the function and the directions (multi-indices) with their optimal rates of consistency simultaneously. This problem has attracted much attention. See, for example, Härdle *et al.* (1993), Severini and Wong (1992) and Carroll *et al.* (1997).

For most methods, the estimator of the direction suffers from the effect of the bias in the estimator of the unknown link function. Therefore, undersmoothing the estimator of the link function is necessary for the estimator of the direction to achieve its optimal rate of consistency. We are not aware of any recommended method to select the undersmooth bandwidth. By minimizing a cross-validation-type sum of squares of residuals simultaneously with respect to both the bandwidth and the direction, Härdle *et al.* (1993) have given a positive answer to the question raised in the previous paragraph. However, we have discussed the problems with this approach in Section 2. In contrast, the MAVE-type methods can handle all the models mentioned above effectively. Specifically, when $D' = 1$, the root n rate of consistency for the direction estimator can be obtained and at the same time the optimal rate of consistency for the nonparametric function estimator can be achieved.

3.4. Discrete or functionally related covariates

Generally, dimension reduction methods cannot be applied to models with discrete or functionally related covariates because they are not estimable, in the sense that there can be more than one dimension reduction space up to orthogonal transformations.

We believe that, provided that the link function can be approximated locally by 'tangent' planes, the MAVE method can still be practically useful for discrete or functionally related covariates. The limiting accuracy will, of course, depend on the accuracy of the tangent plane approximation. We must keep in mind two points:

(a) the bandwidth cannot be selected to be smaller than a critical value because we must use adjacent points to estimate the 'tangent' plane and
(b) if none of the X design points has repeated measurements then bandwidth selection methods based on cross-validation may be considered. If the latter methods are ruled out, a feasible alternative may be one based on the idea of the nearest neighbours as follows. For any point x_k, we choose a nearest neighbour of x_k which includes observations $(\tilde{x}_1, \tilde{y}_1), \ldots, (\tilde{x}_{\tilde{p}}, \tilde{y}_{\tilde{p}})$, such that the plane $y = a + b^{\mathrm{T}}X$ is estimable, i.e. there is a unique solution of (a, b) to $\min_{a,b}\{\Sigma_{i=1}^{\tilde{p}}(\tilde{y}_i - a - b^{\mathrm{T}}\tilde{X}_i)^2\}$; cf. the nearest neighbour method due to Wong and Shen (unpublished) mentioned in Section 2.

If X includes continuous covariates as well as categorical or functionally related covariates, then the RMAVE method still applies with appropriate initial values. If we carry out a global search for the EDR directions, the procedure may be trapped by directions with positive probability due to the categorical data. If we have some prior information about the EDR directions such that we only need to search for the directions locally, then the density requirement can be relaxed, namely the density function of $B^{\mathrm{T}}X$ exists for all $B \in \mathcal{B} = \{B\colon B^{\mathrm{T}}B = I_D$ and $\|B - B_0\| < c\}$ for some $c > 0$. Suppose further that $E(XX^{\mathrm{T}}|B^{\mathrm{T}}X = v)$ and $E(X|B^{\mathrm{T}}X = v)$ exist and have continuous second-order derivatives. Then the RMAVE method in our paper applies with appropriate initial values in $\mathcal{B}$ and the search for the directions conducted within the same region.

4. Simulations

In this section, we carry out simulations to check the performance of the proposed OPG method and the MAVE-type methods. We shall use the square-distance function m^2, where m was defined in Section 2, to measure the error of estimation when we compare our method with others.

4.1. Example 1

We first adopt the examples used in Li (1991). Let $p = 10$ and $\varepsilon, x_1, x_2, \ldots, x_{10}$ be independent random variables each with a standard normal distribution. Consider two regression models:

$$y = x_1(x_1 + x_2 + 1) + 0.5\varepsilon, \tag{4.1}$$

$$y = x_1/\{0.5 + (x_2 + 1.5)^2\} + 0.5\varepsilon. \tag{4.2}$$

The sample size is set at $n = 200$ or $n = 400$ and 100 replications are drawn in each case. Let $\beta_1 = (1, 0, \ldots, 0)^{\mathrm{T}}$, $\beta_2 = (0, 1, \ldots, 0)^{\mathrm{T}}$ and $B_0 = (\beta_1, \beta_2)$. Fig. 1 shows the means of the estimation errors $m^2(\hat{\beta}_1, B_0)$ and $m^2(\hat{\beta}_2, B_0)$; they are labelled '1' and '2' for β_1 and β_2 respectively. In our simulations, the IMAVE method outperforms the SIR method but is outperformed by the MAVE method. The RMAVE method performs best of all the methods. Zhu and Fang (1996) proposed a kernel smooth version of the SIR method. However, their method does not show a significant improvement over that of the original SIR method.

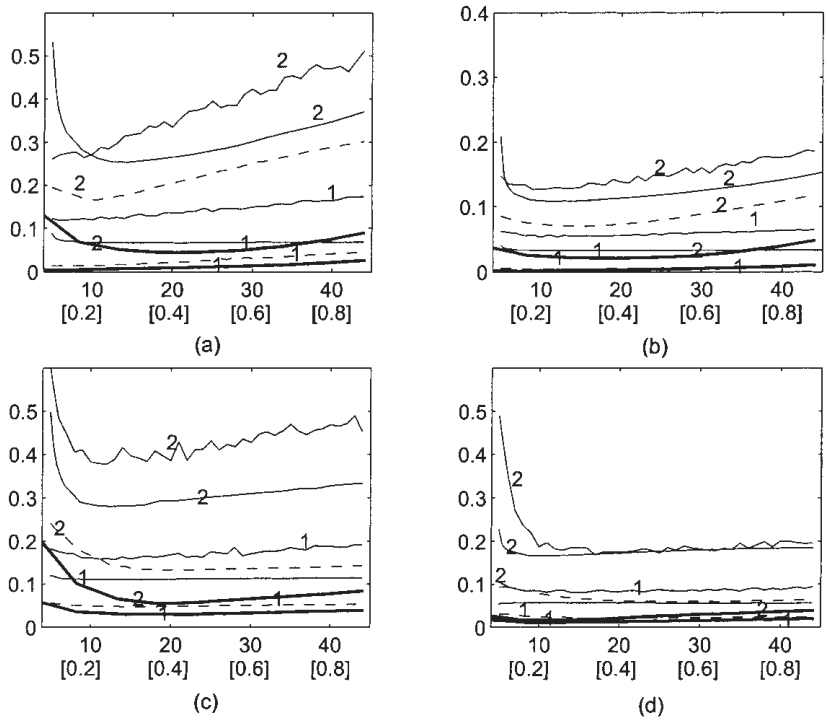


Fig. 1. Means of $m^2(\hat{\beta}_1, B_0)$ (labelled 1) and $m^2(\hat{\beta}_2, B_0)$ (labelled 2) (broken curves are based on the MAVE method; full curves are based on the IMAVE method; wavy curves are based on the SIR method; bold curves are based on the RMAVE method; the horizontal axes give the numbers of slices or the bandwidth (in square brackets) for the SIR method or IMAVE method respectively): (a) model (4.1), sample size 200, bandwidths 1–3 (MAVE method) and 0.1–1 (RMAVE method); (b) model (4.1), sample size 400, bandwidths 1–2 (MAVE method) and 0.1–1 (RMAVE method); (c) model (4.2), sample size 200, bandwidths 1–3 (MAVE method) and 0.1–1 (RMAVE method); (d) model (4.2), sample size 400, bandwidths 1–2 (MAVE method) and 0.1–1 (RMAVE method)

4.2. Example 2

Consider the model

$$y = X^{\mathrm{T}}\beta_1(X^{\mathrm{T}}\beta_2)^2 + (X^{\mathrm{T}}\beta_3)(X^{\mathrm{T}}\beta_4) + 0.5\varepsilon, \tag{4.3}$$

where $X \sim N(0, I_{10})$ and $\varepsilon \sim N(0, 1)$ and they are independent. In model (4.3), the coefficients $\beta_1 = (1, 2, 3, 4, 0, 0, 0, 0, 0, 0)^{\mathrm{T}}/\sqrt{30}$, $\beta_2 = (-2, 1, -4, 3, 1, 2, 0, 0, 0, 0)^{\mathrm{T}}/\sqrt{35}$, $\beta_3 = (0, 0, 0, 0, 2, -1, 2, 1, 2, 1)^{\mathrm{T}}/\sqrt{15}$, $\beta_4 = (0, 0, 0, 0, 0, 0, -1, -1, 1, 1)^{\mathrm{T}}/2$ and there are four EDR directions. Let $B_0 = (\beta_1, \beta_2, \beta_3, \beta_4)$. In our simulations, the SIR method and the IMAVE method perform quite poorly for this model. Next, we use this model to check the OPG method and the MAVE method.

With sample size $n = 100$, 200, 400, 200 independent samples are drawn in each case. The average distance from the estimated EDR directions to $\mathcal{S}(B_0)$ is calculated for the PHD method (Li, 1992), the OPG method, the MAVE method and the RMAVE method. The results are listed in Table 1. The results suggest that the MAVE method performs better than the OPG method, which performs better than the PHD method, whereas the RMAVE method shows a significant improvement over the MAVE method. Our method for the estimation of the number of EDR directions also gives satisfactory results.

4.3. Example 3

We next consider the non-linear time series model

$$y_t = -1 + 0.4\beta_1^{\mathrm{T}}X_{t-1} - \cos\left(\frac{\pi}{2}\beta_2^{\mathrm{T}}X_{t-1}\right) + \exp\{-(\beta_3^{\mathrm{T}}X_{t-1})^2\} + 0.2\varepsilon_t, \tag{4.4}$$

where $X_{t-1} = (y_{t-1}, \ldots, y_{t-6})^{\mathrm{T}}$, the ε are independent and identically distributed $N(0, 1)$, $\beta_1 = (1, 0, 0, 2, 0, 0)^{\mathrm{T}}/\sqrt{5}$, $\beta_2 = (0, 0, 1, 0, 0, 2)^{\mathrm{T}}/\sqrt{5}$ and $\beta_3 = (-2, 2, -2, 1, -1, 1)^{\mathrm{T}}/\sqrt{15}$. Fairly large simulations suggest that there is no discernible symmetry for the covariates; the SIR method does not appear appropriate or to perform well.

Now, the simulation results summarized in Table 2 show that both the OPG method and the MAVE method have quite small estimation errors. As expected, the RMAVE method works

Table 1. Average $m^2(\hat{\beta}_k, B_0)$ for model (4.3) by using different methods

n	*Method*	$m^2(\hat{\beta}_k, B_0)$				*Frequencies of estimated numbers of EDR directions*
		$k = 1$	$k = 2$	$k = 3$	$k = 4$	
100	PHD	0.2769	0.2992	0.4544	0.5818	$f_1 = 0,\ f_2 = 10,\ f_3 = 23,$
	OPG	0.1524	0.2438	0.3444	0.4886	$f_4 = 78,\ f_5 = 44,\ f_6 = 32,$
	MAVE	0.1364	0.1870	0.2165	0.3395	$f_7 = 11,\ f_8 = 1,\ f_9 = 1,$
	RMAVE	0.1137	0.1397	0.1848	0.3356	$f_{10} = 0$
200	PHD	0.1684	0.1892	0.3917	0.6006	$f_1 = 0,\ f_2 = 0,\ f_3 = 5,$
	OPG	0.0713	0.1013	0.1349	0.2604	$f_4 = 121,\ f_5 = 50,\ f_6 = 16,$
	MAVE	0.0710	0.0810	0.0752	0.1093	$f_7 = 8,\ f_8 = 0,\ f_9 = 0,$
	RMAVE	0.0469	0.0464	0.0437	0.0609	$f_{10} = 0$
400	PHD	0.0961	0.1151	0.3559	0.6020	$f_1 = 0,\ f_2 = 0,\ f_3 = 0,$
	OPG	0.0286	0.0388	0.0448	0.0565	$f_4 = 188,\ f_5 = 16,\ f_6 = 6,$
	MAVE	0.0300	0.0344	0.0292	0.0303	$f_7 = 0,\ f_8 = 0,\ f_9 = 0,$
	RMAVE	0.0170	0.0119	0.0116	0.0115	$f_{10} = 0$

Table 2. Average $m^2(\hat{\beta}_k, B_0)$ for model (4.4) by using different methods

n	*Method*	$m^2(\hat{\beta}_k, B_0)$			*Frequency of estimated number of EDR directions*
		$k = 1$	$k = 2$	$k = 3$	
100	PHD	0.1582	0.2742	0.3817	$f_1 = 3,\ f_2 = 73,$
	OPG	0.0427	0.1202	0.2803	$f_3 = 94,\ f_4 = 25,$
	MAVE	0.0295	0.1201	0.2924	$f_5 = 4,\ f_6 = 1$
	RMAVE	0.0096	0.1712	0.2003	
200	PHD	0.1565	0.2656	0.3690	$f_1 = 0,\ f_2 = 34,$
	OPG	0.0117	0.0613	0.1170	$f_3 = 160,\ f_4 = 5,$
	MAVE	0.0059	0.0399	0.1209	$f_5 = 1,\ f_6 = 0$
	RMAVE	0.0030	0.0224	0.0632	
300	PHD	0.1619	0.2681	0.3710	$f_1 = 0,\ f_2 = 11,$
	OPG	0.0076	0.0364	0.0809	$f_3 = 185,\ f_4 = 4,$
	MAVE	0.0040	0.0274	0.0666	$f_5 = 0,\ f_6 = 0$
	RMAVE	0.0017	0.0106	0.0262	

better than the MAVE method, which outperforms the OPG method. The PHD method does not fare very well. The number of the EDR directions is also estimated correctly most of the time.

5. Examples

5.1. Circulatory and respiratory problems in Hong Kong

Consider the effect of the levels of pollutants and weather on the total number y_t of daily hospital admissions of patients suffering from circulatory and respiratory problems. The pollutant and weather data are the daily average levels of sulphur dioxide (x_{1t} (μg m^{-3})), nitrogen dioxide (x_{2t} (μg m^{-3})), respirable suspended particulates (x_{3t} (μg m^{-3})), ozone (x_{4t} (μg m^{-3})), temperature (x_{5t} (°C)) and relative humidity (x_{6t} (%)). The data were collected daily in Hong Kong from January 1st, 1994, to December 31st, 1995, and are shown in Fig. 2. The basic question is this: are the prevailing levels of the pollutants a cause for concern?

A naïve approach may be to start with a simple linear regression model such as

$$\begin{matrix} y_t = & 255.45 & - & 0.55x_{1t} & + & 0.58x_{2t} & + & 0.18x_{3t} & - & 0.33x_{4t} & - & 0.12x_{5t} & - & 0.16x_{6t}. \\ & (20.64) & & (0.18) & & (0.17) & & (0.13) & & (0.11) & & (0.46) & & (0.23) \end{matrix} \tag{5.1}$$

Note that the coefficients of x_{3t}, x_{5t} and x_{6t} are not significantly different from 0 (at the 5% level of significance) by reference to their standard errors shown inside the parentheses and the negative and significant coefficients of x_{1t} and x_{4t} are difficult to interpret. Refinements of this model are, of course, possible within the linear framework but are unlikely to throw much light with respect to the opening question because, as we shall see, the situation is quite complex. Previous analyses, such as Fan and Zhang (1999) and Cai *et al.* (2000), have not included the weather effect. However, it turns out that the weather has an important role to play.

The daily admissions shown in Fig. 2(a) suggest non-stationarity in the form of almost a level shift taking place in early 1995 although none of the covariates seems to show a similar level shift. Now, a trend was also observed by Smith *et al.* (1999) in their study of the effect of particulates on human health. They conjectured that the trend was due to the epidemic effect. In our case, we understand from our data provider that additional hospital beds were released to accommodate circulatory and respiratory patients in the course of his joint project. As a

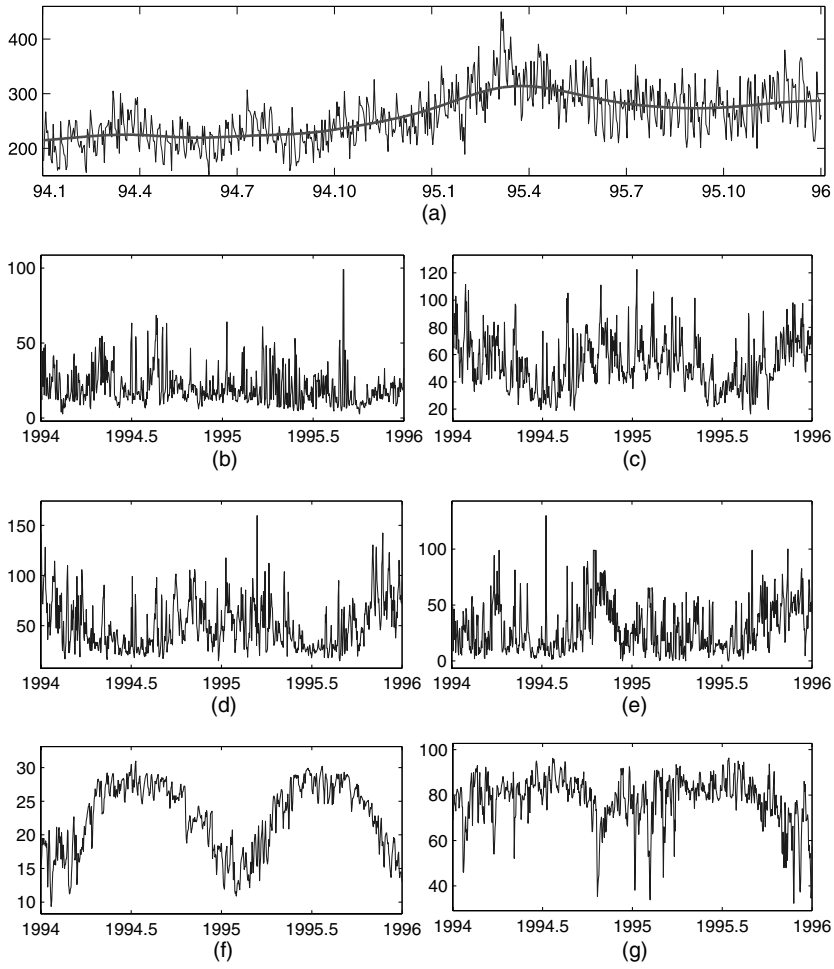


Fig. 2. (a) Total number of daily hospital admissions of circulatory and respiratory patients (———, time trend) and average levels of (b) sulphur dioxide, (c) nitrogen dioxide, (d) respirable suspended particulates, (e) ozone, (f) temperature and (g) humidity

result, we estimate the time dependence by a simple kernel method and the result is shown in Fig. 2(a). Another factor is the day of the week effect, presumably due to the hospital booking system. The day of the week effect can be estimated by a simple regression method using dummy variables. To assess the effect of pollutants better, we remove these two factors first. By an abuse of notation, we shall continue to use y_t to denote the 'filtered' data, now shown in Fig. 3.

As the pollutant-based and weather-based covariates may affect the circulatory and respiratory system with a time delay, we consider the six covariates in the last 7 days (1 week). Altogether, we have 42 covariates:

$$X_t = (x_{1,t-1}, x_{1,t-2}, \ldots, x_{1,t-7}, x_{2,t-1}, x_{2,t-2}, \ldots, x_{2,t-7}, \ldots, x_{6,t-1}, x_{6,t-2}, \ldots, x_{6,t-7})^{\mathrm{T}}.$$

Now, using the RMAVE method and with a cross-validation bandwidth, we have the results in Table 3. The cross-validation choice of the dimension is 3. The corresponding direction estimates are listed in Table 4.

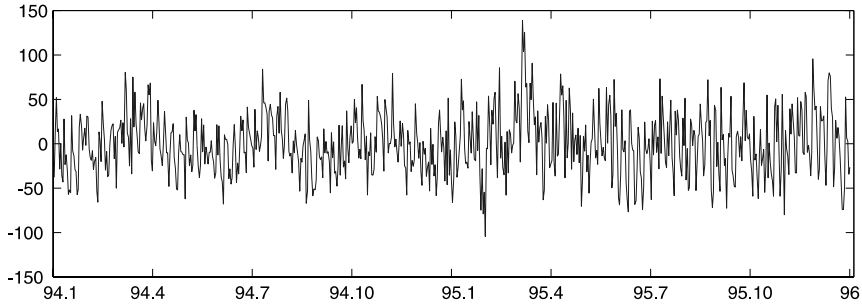


Fig. 3. 'Filtered' number of daily hospital admissions of circulatory and respiratory patients by removing the time trend and the day of the week effect

Figs 4(a)–4(c) show y_t plotted against the respective EDR directions. These plots and Table 4 suggest the following features.

(a) Rapid temperature changes play an important role. (Note the dominant coefficients for temperature in the two recent past days in $\hat{\beta}_{\mathrm{I}}^{\mathrm{T}} X$.)
(b) Of the pollutants, the most influential seems to be the particulates (note the large coefficient for particulates at lag 5 in $\hat{\beta}_{\mathrm{II}}^{\mathrm{T}} X$) and the least influential seems to be sulphur dioxide.
(c) The weather covariates are influential. (Note the many large coefficients for the weather covariates in all the three $\hat{\beta}$s.)

Comparing the levels of the individual pollutants in Hong Kong against the national ambient quality standard of the USA lends further support to feature (b).

Bearing these features in mind, we may explore further by focusing on the suspended particulates (x_3), the ozone level (x_4), the temperature (x_5) and its variation, and the relative humidity (x_6). First, we define the variation of temperature as

$$v_t = \mathrm{std}(x_{5,t-k}, k = 1, 2, 3, 4, 5).$$

Further simplification is obtained by selecting only one lag for each covariate. For this, we use the method of Yao and Tong (1994). The lagged covariates selected are $x_{3,t-2}$, $x_{4,t-6}$, $x_{5,t-4}$ and $x_{6,t-2}$. Let $Z_t = (x_{3,t-2}, x_{4,t-6}, v_t, x_{5,t-4}, x_{6,t-2})^{\mathrm{T}}$. We then consider a model of the form

$$y_t = g(Z_t) + \varepsilon_t.$$

Table 3. Results of the CV method

Dimension	*Bandwidth*	*CV(d) value*
1	0.10	0.33
2	0.13	0.28
3	0.16	0.27
4	0.20	0.29
5	0.21	0.29
6	0.24	0.31
7	0.24	0.34
8	0.29	0.31
9	0.31	0.34
10	0.31	0.37

Table 4. Estimated EDR directions $\hat{\beta}_{\mathrm{I}}$, $\hat{\beta}_{\mathrm{II}}$ and $\hat{\beta}_{\mathrm{III}}$†

Parameter	*Estimates for the following lags:*						
	1	*2*	*3*	*4*	*5*	*6*	*7*
x_1	0.0586	−0.0854	0.0472	−0.0152	**0.1083**	−0.0942	0.0734
x_2	0.0876	0.0313	**−0.1964**	0.0893	−0.0867	0.0951	**−0.1068**
x_3	**−0.2038**	**0.1103**	0.0153	0.0740	−0.0756	**0.1283**	−0.0520
x_4	0.0155	0.0692	**0.1622**	**−0.2624**	**0.1312**	**0.1342**	0.0976
x_5	**0.5065**	**−0.4079**	0.0743	0.0859	**−0.3024**	−0.1734	−0.0302
x_6	−0.0294	−0.0610	0.0129	−0.0392	−0.0075	**0.2850**	0.0513
x_1	**−0.1525**	0.0962	**−0.1112**	**0.1170**	−0.0388	−0.0605	−0.0326
x_2	−0.0029	**0.1614**	−0.0955	**−0.1160**	**−0.2185**	0.0826	**0.1696**
x_3	−0.0096	**−0.1874**	**0.2422**	−0.0047	**0.3272**	**−0.2646**	−0.0041
x_4	−0.0013	**−0.1162**	0.0673	**0.2113**	**−0.2193**	**0.1235**	**−0.1282**
x_5	**0.1410**	**0.1193**	**−0.1425**	**0.1819**	**−0.2793**	−0.0880	−0.0325
x_6	−0.0345	**−0.1479**	−0.0400	**0.4033**	0.0474	0.0899	**0.1336**
x_1	0.0701	0.0065	−0.0535	**−0.1570**	−0.0553	−0.0091	−0.0363
x_2	−0.0529	**0.1360**	0.0723	**0.1045**	−0.0045	−0.0200	0.0221
x_3	−0.0121	**−0.1189**	0.0715	−0.0814	0.0112	0.0155	**0.1214**
x_4	**0.2215**	0.0103	**−0.3304**	**0.1028**	0.0160	**−0.1805**	**0.1341**
x_5	**0.2909**	**−0.2372**	0.0621	−0.0211	0.0950	−0.0954	**0.2507**
x_6	**0.2797**	**−0.1094**	**−0.3038**	0.0452	**0.1754**	**−0.3937**	**0.2597**

†Entries in bold have relatively large absolute values.

The above proposed procedure yields the results in Table 5.

On the basis of Table 5 the dimension of EDR space is chosen to be 3 with the following estimated basis vectors for the space:

$$\begin{aligned}
\hat{\beta}_1 &= (-0.1317 \quad -0.0772 \quad 0.5256 \quad -0.8366 \quad -0.0235)^{\mathrm{T}},\\
\hat{\beta}_2 &= (0.4809 \quad 0.3154 \quad -0.6414 \quad -0.5078 \quad 0.0018)^{\mathrm{T}},\\
\hat{\beta}_3 &= (0.0101 \quad 0.3815 \quad 0.1345 \quad 0.0734 \quad -0.9115)^{\mathrm{T}}.
\end{aligned}$$

Figs 4(d)–4(f) show y_t plotted against the three directions. The 'price' of using the reduced set with five covariates instead of the original set with 42 covariates is, loosely speaking, an increase in the percentage of unexplained variation from about 27% to about 34%. (As we use standardized observations, we may interpret the CV(d) value as a percentage of unexplained variation.) In return, we can gain further insight.

(a) The first EDR direction is $-0.1317x_{3,t-2} - 0.0772x_{4,t-6} + 0.5256v_t - 0.8366x_{5,t-4} - 0.0235x_{6,t-2}$, with temperature and temperature variation being the two dominant

Table 5. Results of the cross-validation method

Dimension	*Bandwidth*	*CV(d) value*
1	0.325	0.3593
2	0.325	0.3516
3	0.325	0.3435
4	0.325	0.3523
5	0.475	0.3450

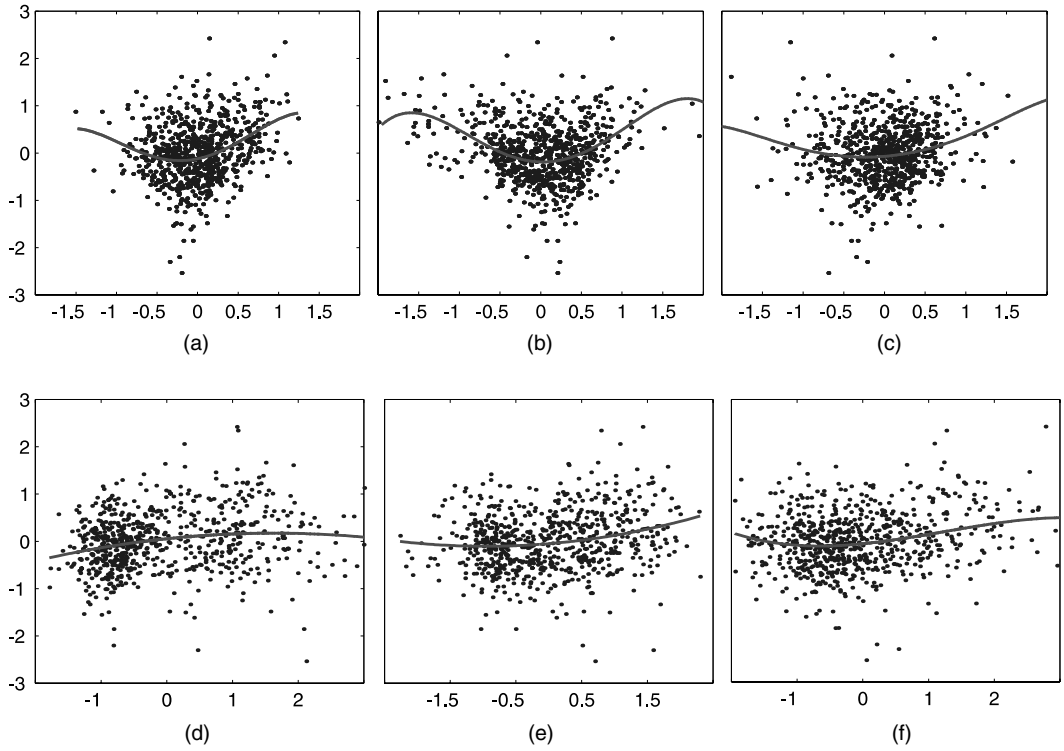


Fig. 4. y_t plotted against (a) $\hat{\beta}_{\mathrm{I}}^{\mathrm{T}} X_t$, (b) $\hat{\beta}_{\mathrm{II}}^{\mathrm{T}} X_t$, (c) $\hat{\beta}_{\mathrm{III}}^{\mathrm{T}} X_t$, (d) $\hat{\beta}_1^{\mathrm{T}} Z_t$, (e) $\hat{\beta}_2^{\mathrm{T}} Z_t$ and (f) $\hat{\beta}_3^{\mathrm{T}} Z_t$: ———, polynomial regression to make trends more visualizable

components. Fig. 4(d) suggests that this direction sees practically only the mean level of the hospital admissions.

(b) The second EDR direction is $0.4809x_{3,t-2} + 0.3154x_{4,t-6} - 0.6414v_t - 0.5078x_{5,t-4} + 0.0018x_{6,t-2}$, which, together with Fig. 4(e), suggests that high levels of suspended particulates and/or high levels of ozone during cold spells tend to cause high admissions.

(c) The third EDR direction is $0.0101x_{3,t-2} + 0.3815x_{4,t-6} + 0.1345v_t + 0.0734x_{5,t-4} - 0.9115x_{6,t-2}$, which, together with Fig. 4(f), suggests that high ozone levels on extremely dry days tend to cause high admissions.

This analysis suggests that pollutants have reached such a level in Hong Kong that it only takes the weather to enter the right regime to exacerbate the circulatory and respiratory problems there.

5.2. Hitters' salary data

The hitters' salary data set has attracted much attention among statisticians. The data consist of times at bat (x_1), hits (x_2), home runs (x_3), runs (x_4), runs batted in (x_5) and walks (x_6) in 1986, years in major leagues (x_7), times at bat (x_8), hits (x_9), home runs (x_{10}), runs (x_{11}), runs batted in (x_{12}) and walks (x_{13}) during their entire career up to 1986, annual salary (y) in 1987, put-outs (x_{14}), assistances (x_{15}) and errors (x_{16}). For ease of exposition, we abuse the notation and set y as the logarithm of annual salary in 1987, x_j the standardized x_j ($j = 1, \ldots, 16$) and

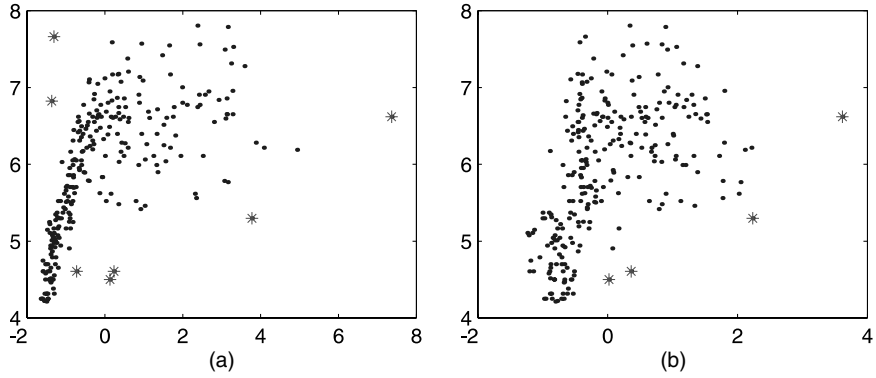


Fig. 5. y plotted against (a) $\hat{\beta}_1^{\mathrm{T}} X$ and (b) $\hat{\beta}_2^{\mathrm{T}} X$ for the hitters' salary data: $*$, outlier

X the vector $(x_1, \ldots, x_{16})^{\mathrm{T}}$. The main interest is 'why they make what they make', which was the main topic of a conference organized for the data by the American Statistical Association in 1988. More recent studies on this data include Chaudhuri *et al.* (1994) and Li *et al.* (2000). The latter suggested the existence of an 'aging effect' on salary.

Now, applying the RMAVE method to the data set and using model (1.1), we estimate the dimension of the EDR space as 2. We plot y against the two EDR directions as shown in Fig. 5. It suggests that there are seven outliers, in general agreement with an observation made by Li *et al.* (2000). Next, applying the RMAVE method to the data with the outliers removed, we have the following results. Table 6 shows that the dimension estimate remains at 2 and Fig. 6 shows the plots of y against the estimated EDR directions. The similarity between the results before and after the removal of outliers suggests a high degree of robustness enjoyed by the RMAVE method.

The EDR directions are given in the first pair of columns of Table 7. Note that, in the second direction, the negative coefficient (-0.23) of x_7 lends some support to the aging effect on salary suggested by Li *et al.* (2000).

We may combine the MAVE methodology with ideas such as thresholds (e.g. Tong (1990)) and regression trees to fit different regression models to different parts of the data set. For regression trees, we may mention the classification and regression trees method of Breiman *et al.* (1984), the SUPPORT algorithm of Chaudhuri *et al.* (1994) and the PHDRT algorithm of Li *et al.* (2000) and others. As an illustration, the left-hand 'regime' in Fig. 6(a) can be fitted

Table 6. Results of the cross-validation method (with the outliers removed)

Dimension	*Bandwidth*	*CV(d) value*
1	0.148	0.265
2	0.395	0.118
3	0.473	0.134
4	0.609	0.158
5	0.544	0.139
6	0.596	0.135
7	0.662	0.146
8	0.572	0.133
9	0.655	0.132
10	0.927	0.178

Table 7. Estimated EDR directions in model (1.1) and (5.2)†

$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}$	$\hat{\theta}$
−0.25 (x_1)	0.08 (x_1)	−0.05 (x_1)	0.14 (x_1)	−0.01 (x_1)	**−0.27** (x_1)
0.24 (x_2)	0.04 (x_2)	0.04 (x_2)	**−0.20** (x_2)	0.15 (x_2)	0.17 (x_2)
0.09 (x_3)	−0.01 (x_3)	−0.03 (x_3)	−0.09 (x_3)	0.02 (x_3)	0.09 (x_3)
0.00 (x_4)	0.07 (x_4)	0.03 (x_4)	**0.40** (x_4)	0.01 (x_4)	−0.04 (x_4)
−0.01 (x_5)	−0.04 (x_5)	−0.03 (x_5)	−0.03 (x_5)	−0.06 (x_5)	0.01 (x_5)
0.05 (x_6)	0.04 (x_6)	−0.09 (x_6)	**−0.29** (x_6)	0.06 (x_6)	0.04 (x_6)
0.52 (x_7)	**−0.23** (x_7)	0.18 (x_7)	0.02 (x_7)	0.01 (x_7)	**0.51** (x_7)
0.55 (x_8)	**−0.49** (x_8)	**0.51** (x_8)	**−0.57** (x_8)	0.03 (x_8)	**0.74** (x_8)
0.37 (x_9)	**0.75** (x_9)	**−0.81** (x_9)	**−0.26** (x_9)	**0.90** (x_9)	−0.11 (x_9)
0.10 (x_{10})	0.15 (x_{10})	0.00 (x_{10})	−0.08 (x_{10})	**0.24** (x_{10})	0.03 (x_{10})
0.23 (x_{11})	0.12 (x_{11})	0.08 (x_{11})	**0.27** (x_{11})	0.14 (x_{11})	0.16 (x_{11})
0.08 (x_{12})	0.17 (x_{12})	−0.10 (x_{12})	0.08 (x_{12})	0.04 (x_{12})	−0.13 (x_{12})
0.30 (x_{13})	**0.22** (x_{13})	0.07 (x_{13})	**0.43** (x_{13})	**0.27** (x_{13})	0.08 (x_{13})
−0.01 (x_{14})	0.09 (x_{14})	−0.06 (x_{14})	0.09 (x_{14})	0.08 (x_{14})	−0.04 (x_{14})
0.04 (x_{15})	−0.03 (x_{15})	0.03 (x_{15})	0.12 (x_{15})	−0.03 (x_{15})	0.06 (x_{15})
0.00 (x_{16})	0.04 (x_{16})	−0.05 (x_{16})	−0.08 (x_{16})	0.05 (x_{16})	−0.02 (x_{16})

†Entries in bold have relatively large absolute values.

by a simple straight line, say

$$y = \underset{(0.07)}{7.24} + \underset{(0.02)}{0.09}x_3 + \underset{(0.07)}{0.38}x_7 + \underset{(0.15)}{1.49}x_9 + \underset{(0.15)}{0.83}x_{13}.$$

The standard deviation of the fitted residuals, $\hat{\sigma}$, is 0.26 and $R^2 = 0.865$. The threshold is set at -0.47. The right-hand regime is much more volatile and we may return to the RMAVE method. The estimated dimension is still 2 and the estimated directions are given in the second pair of columns in Table 7. Let $z_1 = \hat{\beta}_1^{\mathrm{T}} X$ and $z_2 = \hat{\beta}_2^{\mathrm{T}} X$. We may fit to the right-hand regime a polynomial regression such as

$$y = \underset{(0.03)}{6.61} - \underset{(0.11)}{1.86}z_1 + \underset{(0.09)}{0.21}z_2 - \underset{(0.19)}{1.19}z_2^2.$$

For this model, $\hat{\sigma} = 0.28$ and $R^2 = 0.714$. The overall $\hat{\sigma}$ is 0.27. A simple calculation shows

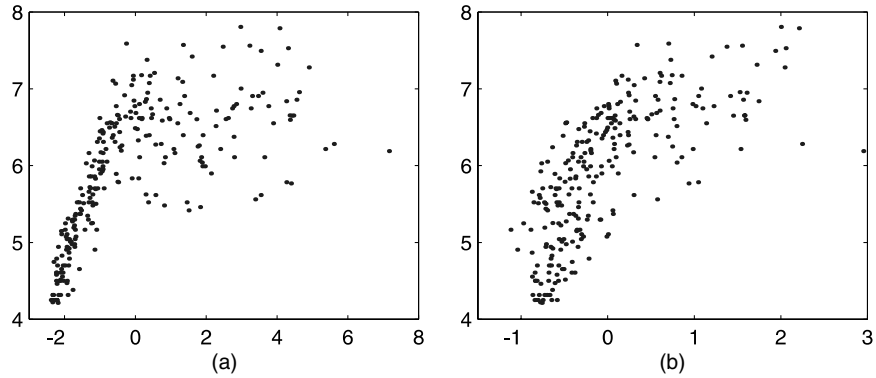


Fig. 6. *y* plotted against (a) $\hat{\beta}_1^{\mathrm{T}} X$ and (b) $\hat{\beta}_2^{\mathrm{T}} X$ for the hitters' salary data with the outliers removed

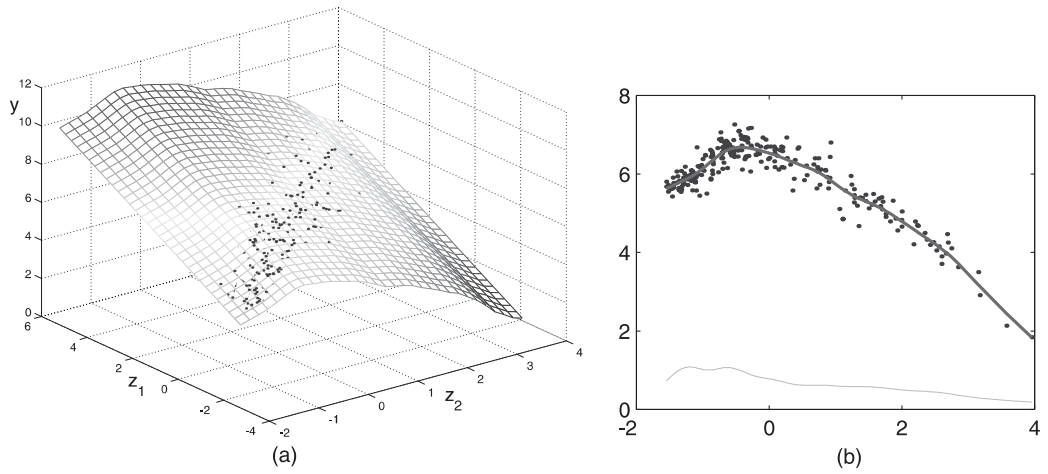


Fig. 7. (a) Estimated regression surface of model (5.2) (•, observations) and (b) estimated regression function of g (———) and estimate of the density function along the direction (· · · · · · ·) (•, residuals after removing the linear part in model (5.2))

that the coefficient of x_7 in the model for the right-hand regime is again negative, with the implication mentioned previously. As a comparison with the regression tree results obtained by Li *et al.* (2000), we quote $\hat{\sigma} = 0.422$ for the classification and regression trees method with five bases, 0.33 for the multivariate adaptive regression splines method with 13 bases, 0.44 for the SUPPORT algorithm with two bases and 0.35 for the PHDRT algorithm. For our simple-minded hybrid, the overall $\hat{\sigma} = 0.27$ with two bases.

Finally, we may consider the model

$$y = a\beta^{\mathrm{T}}X + g(\theta^{\mathrm{T}}X) + \varepsilon, \tag{5.2}$$

where $\beta \perp \theta$ with $\|\theta\| = \|\beta\| = 1$. This is a special case of model (1.3). See Xia *et al.* (1999) for details. Using the method described in Section 3, we obtain estimates of β and θ as listed in the third pair of columns in Table 7, $\hat{a} = 0.75$, $\hat{\sigma} = 0.26$ and the estimate of the function g as shown in Fig. 7(b). (Because the density of $\hat{\theta}^{\mathrm{T}}X$ is not so uniform, a variable bandwidth is used. See Fan and Gijbels (1996), page 152.) The dominant covariates in $z_1 = \hat{\beta}^{\mathrm{T}}X$ are x_2, x_9, x_{10}, x_{11} and x_{13}, all with positive coefficients. Four out of these five covariates measure past performance and so we may interpret z_1 as principally a measure of past performance. Fig. 7(a) shows that, along the z_1-axis, players with better past performance are paid better. Note also that the number of years in the major league (x_7) only features in z_2, i.e. $\hat{\theta}^{\mathrm{T}}X$, and quite prominently so. The estimated $g(z_2)$ lends support to the existence of an aging effect, now with the salary peaking at around $z_2 = -0.5$.

6. Conclusions

Our theoretical analysis, simulations and real applications have led us to believe that the MAVE methodology has many attractive attributes. Different from most existing methods for the estimation of the directions, the MAVE estimators of the directions have a faster rate of consistency than the corresponding estimators of the link function. On the basis of the faster rate of consistency, a consistent method for the determination of the number of EDR directions has been proposed. The MAVE method can easily be extended to more complicated models. It does not

require strong assumptions on the design of X and the regression functions and can be applied to both independent data and dependent data.

As a by-product, we have extended the ADE method of Härdle and Stoker (1989) to the case of more than one EDR direction, resulting in the OPG method. This method has wider applicability with respect to designs for X and regression functions. Our basic idea has also led to the IMAVE method, which is closely related to the SIR method and the most predictable problem of Hotelling (1935), but in our simulations IMAVE seems to enjoy a better performance than SIR. The refined kernel based on the determination of the number of the directions can further improve the accuracy of estimation of the directions. Our simulations show that substantial improvements can be achieved.

Theoretical improvements on the MAVE method and the OPG method can be made by using higher order local polynomial smoothing. For example, we may replace expressions (2.7) and (3.1) by

$$\min_{\substack{B:B^{\mathrm{T}}B=I \\ a_j,b_j,c_j}} \left[\sum_{j=1}^{n}\sum_{i=1}^{n}\{y_i - a_j - b_j B^{\mathrm{T}}(X_i - X_j) \right.$$

$$\left. - \sum_{1<k\leqslant r}\ \sum_{i_1+\ldots+i_p=k} (c_{j,i_1,i_2,\ldots,i_p}\{X_i - X_j\}_1^{i_1}\{X_i - X_j\}_2^{i_2}\ldots\{X_i - X_j\}_p^{i_p})\}^2 w_{ij}\right],$$

where $c_j = \{c_{j,i_1,i_2,\ldots,i_p}, i_1 + \ldots + i_p = k, 1 < k \leqslant r\}$, and

$$\min_{a_j,b_j,c_j} \left[\sum_{i=1}^{n}\{y_i - a_j - b_j^{\mathrm{T}}(X_i - X_j) \right.$$

$$\left. - \sum_{1<k\leqslant r}\sum_{i_1+\ldots+i_p=k} (c_{j,i_1,i_2,\ldots,i_p}\{X_i - X_j\}_1^{i_1}\{X_i - X_j\}_2^{i_2}\ldots\{X_i - X_j\}_p^{i_p})\}^2 K_h(X_i - X_j)\right]$$

respectively. Higher rates of consistency can then be obtained.

Unlike the SIR method, the MAVE method is well adapted to time series; our experience suggests that the MAVE method is also robust against outliers. Furthermore, all our simulations show that the MAVE method has a much better performance than the SIR method (and OPG method). Although theorem 2 furnishes a partial explanation, we are still intrigued because SIR uses the one-dimensional kernel (for the kernel version) whereas the MAVE method uses a multidimensional kernel. However, because the SIR method uses y to produce the kernel weight, its efficiency will suffer from fluctuations in the link function. The gain by using the y-based one-dimensional kernel does not seem to be sufficient to compensate for the loss in efficiency caused by these fluctuations, but further research is needed here.

Acknowledgements

We thank the Biotechnology and Biological Science Research Council and Engineering and Physical Sciences Research Council of the UK, the Research Grants Council of Hong Kong, the Committee on Research and Conference Grants of the University of Hong Kong, the Friends of London School of Economics (Hong Kong) and the Wellcome Trust for partial support. We are most grateful to two referees for constructive comments. We thank Professor Wing Hung Wong and Professor X. Shen for making available to us their unpublished work and Professor T. S. Lau for providing the Hong Kong data and some background information.

Appendix A

A.1. Assumptions and remarks

The observations of X should be standardized before the analysis. Define the generalized conditional density

$$p_{\xi|\zeta}(u|v) = \lim_{\mathrm{d}u\to 0,\,\mathrm{d}v\to 0}\left\{\frac{P(\xi\in[u,u+\mathrm{d}u),\,\zeta\in[v,v+\mathrm{d}v))}{\mathrm{d}u\,P(\zeta\in[v,v+\mathrm{d}v))}\right\},$$

and we define $0/0=0$. In our proofs, we need the following conditions. (In all our theorems, weaker conditions can be adopted at the expense of much lengthier proofs.)

Condition 1: $\{(X_i, y_i)\}$ is a stationary (with the same distribution as (X, y)) and absolutely regular sequence, i.e.

$$\beta(k)=\sup_{i\geqslant 1}\left[E\left\{\sup_{A\in\mathcal{F}_{i+k}^{\infty}}|P(A|\mathcal{F}_1^i)-P(A)|\right\}\right]\to 0 \qquad \text{as } k\to\infty,$$

where $\mathcal{F}_i^k$ denotes the σ-field generated by $\{(X_l, y_l) : i\leqslant l\leqslant k\}$. Further, $\beta(k)$ decreases at a geometric rate.

Condition 2:

(a) $E|y|^k<\infty$ for all $k>0$;
(b) $E\|X\|^k<\infty$ for all $k>0$.

Condition 3:

(a) the density function f of X has bounded fourth derivative and is bounded away from 0 in a neighbourhood $\mathcal{D}$ around 0;
(b) the density function f_y of y has bounded derivative and is bounded away from 0 on a compact support.

Condition 4: the generalized conditional densities $p_{X|y}(x|y)$ of X given y and $p_{(X_0,\,X_l)|(y_0,\,y_l)}$ of (X_0, X_l) given (y_0, y_l) are bounded for all $l\geqslant 1$.

Condition 5:

(a) g has bounded, continuous third derivatives;
(b) $E(X|y)$ and $E(XX^{\mathrm{T}}|y)$ have bounded, continuous third derivatives.

Condition 6: $K(\cdot)$ is a spherical symmetric density function with a bounded derivative. All the moments of $K(\cdot)$ exist.

Condition 1 is made only for the purpose of simplicity of proof. It can be weakened to $\beta(k)=O(k^{-\iota})$ for some $\iota>0$. Many time series models, including the autoregressive single-index model (Xia and An, 1999), satisfy assumption 1. Condition 2(a) is also made for simplicity of proof. See, for example, Härdle *et al.* (1993). The existence of finite moments is sufficient. Condition 3(a) is needed for the uniform rate of consistency of the kernel smoothing methods. Condition 4 is needed for kernel estimation of dependent data. Condition 5(a) is made to meet the continuous requirement for kernel smoothing. The kernel assumption 6 is satisfied by most of the commonly used kernel functions. For ease of exposition, we further assume that

$$\int UU^{\mathrm{T}}K(U)\,\mathrm{d}U=I.$$

A.2. The efficiency of the algorithm

To explain the mechanism of the MAVE method, we here consider only the single-index model, i.e.

$$y = g(\beta_0^{\mathrm{T}} X) + \varepsilon.$$

We estimate β_0 by minimizing

$$\sum_{j=1}^{n}\sum_{i=1}^{n}\{y_i - a_j - b_j\beta^{\mathrm{T}}(X_i - X_j)\}^2 w_{ij} \tag{A.1}$$

iteratively with respect to (a_j, b_j) and β. Let

$$s_{n,0}(x) = n^{-1}\sum_{i=1}^{n} K_{h,i}(x),$$

$$s_{n,1}(x) = n^{-1}\sum_{i=1}^{n} K_{h,i}(x)((X_i - x)/h),$$

$$s_{n,2}(x) = n^{-1}\sum_{i=1}^{n} K_{h,i}(x)((X_i - x)/h)((X_i - x)/h)^{\mathrm{T}}.$$

Then $w_{ij} = n^{-1}K_{h,i}(X_j)/s_{n,0}(X_j)$. According to our estimation procedure, if we begin with any unit norm vector β, we have by minimizing expression (A.1)

$$\hat{a}_j = \frac{n^{-1}\sum_{i=1}^{n}\{\beta^{\mathrm{T}}s_{n,2}\beta K_{h,i}(X_j) - \beta^{\mathrm{T}}s_{n,1}\beta^{\mathrm{T}}K_{h,i}(X_j)((X_i - X_j)/h)\}y_i}{\beta^{\mathrm{T}}(s_{n,0}(X_j)s_{n,2}(X_j) - s_{n,1}(X_j)s_{n,1}^{\mathrm{T}}(X_j))\beta},$$

$$\hat{b}_j h = \frac{n^{-1}\sum_{i=1}^{n}\{s_{n,0}(X_j)\beta^{\mathrm{T}}K_{h,i}(X_j)((X_i - x)/h) - \beta^{\mathrm{T}}s_{n,1}K_{h,i}(X_j)\}y_i}{\beta^{\mathrm{T}}(s_{n,0}(X_j)s_{n,2}(X_j) - s_{n,1}(X_j)s_{n,1}^{\mathrm{T}}(X_j))\beta}.$$

After one step of iteration, we obtain the estimate of β_0 as

$$\tilde{\beta} = \left\{\sum_{j=1}^{n} h^2\hat{b}_j^2\sum_{i=1}^{n}\frac{K_{h,i}(X_j)((X_i - X_j)/h)((X_i - X_j)/h)^{\mathrm{T}}}{s_{n,0}(X_j)}\right\}^{-1} \times \sum_{j=1}^{n} h\hat{b}_j\sum_{i=1}^{n}\frac{K_{h,i}(X_j)((X_i - X_j)/h)(y_i - \hat{a}_j)}{s_{n,0}(X_j)}.$$

If β is not perpendicular to β_0, we have

$$\tilde{\beta} = \{1 + (1 - \beta^{\mathrm{T}}\beta_0) + o_P(1)\}\beta_0 + O_P[(h^2 + \delta_n)\{h + m(\beta, \beta_0)\} + h^{-1}\delta_n^2]\beta_0^{\perp}, \tag{A.2}$$

where $\beta_0^{\perp}$ is a vector perpendicular to β_0. Equation (A.2) means that the effect of the initial value is quite small. Note that $\delta_n \sim h^2\log(n)^{1/2}$ if we use the optimal bandwidth of the estimation of the regression function, i.e. $h \sim n^{-1/(p+4)}$. Suppose that we start with an initial estimator of β_0 which has a consistency rate of $O_P\{h^2\log(n)\}$. Then $m(\beta, \beta_0) = O_P\{h^2\log(n)\}$ and we have

$$\tilde{\beta} = \{1 + (1 - \beta^{\mathrm{T}}\beta_0) + o_P(1)\}\beta_0 + O_P(h^3 + h\delta_n + h^{-1}\delta_n^2)\beta_0^{\perp}.$$

Therefore,

$$m(\tilde{\beta}, \beta_0) = O_P(h^3 + h\delta_n + h^{-1}\delta_n^2).$$

This estimation procedure is very efficient in that, in theory, after two steps the estimate from our procedure can achieve the final consistency rate.

A similar result was discovered in a different context by Hannan (1969). Specifically, he developed an estimation procedure for the parameters of autoregressive moving average processes. Starting with arbitrary consistent estimators of the parameters, a modification by one step of the Newton–Raphson-type iteration can make the estimators asymptotically efficient. In the MAVE method, the first step is to find a consistent 'initial' estimator. The second step is to modify the 'initial' estimator, which can also make the estimate asymptotically efficient. In spite of the asymptotic efficiency, the iterative application of the procedure beyond the two steps was suggested by Hannan (1969) as a way of further improving the estimator. For the MAVE method, our simulation also suggests that further iterations are beneficial.

References

Auestad, B. and Tjøstheim, D. (1990) Identification of nonlinear time series: first order characterization and order determination. *Biometrika*, **77**, 669–688.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984) *Classification and Regression Trees.* Belmont: Wadsworth.

Cai, Z., Fan, J. and Yao, Q. (2000) Functional-coefficient regression models for nonlinear time series. *J. Am. Statist. Ass.*, **95**, 941–956.

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

Carroll, R. J. and Li, K. C. (1995) Binary regressors in dimension reduction models: a new look at treatment comparisons. *Statist. Sin.*, **5**, 667–688.

Chaudhuri, P., Huang, M. C., Loh, W. Y. and Yao, R. (1994) Piecewise-polynomial regression trees. *Statist. Sin.*, **4**, 143–167.

Chen, C.-H. and Li, K. C. (1989) Can SIR be as popular as multiple linear regression? *Statist. Sin.*, **8**, 289–316.

Chen, H. (1991) Estimation of a projection-pursuit type regression model. *Ann. Statist.*, **19**, 142–157.

Cheng, B. and Tong, H. (1992) On consistent nonparametric order determination and chaos (with discussion). *J. R. Statist. Soc.* B, **54**, 427–449.

Cook, R. D. (1994) On the interpretation of regression plots. *J. Am. Statist. Ass.*, **89**, 177–189.

——— (1998) Principal Hessian directions revisited (with discussions). *J. Am. Statist. Ass.*, **93**, 85–100.

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

Fan, J. and Zhang, W. Y. (1999) Statistical estimation in varying coefficient models. *Ann. Statist.*, **27**, 1491–1518.

Friedman, J. H. and Stuetzle, W. (1981) Projection pursuit regression. *J. Am. Statist. Ass.*, **76**, 817–823.

Hall, P. (1989) On projection pursuit regression. *Ann. Statist.*, **17**, 573–588.

Hannan, E. J. (1969) The estimation of mixed moving average autoregressive system. *Biometrika*, **56**, 579–593.

Härdle, W., Hall, P. and Ichimura, H. (1993) Optimal smoothing in single-index models. *Ann. Statist.*, **21**, 157–178.

Härdle, W. and Stoker, T. M. (1989) Investigating smooth multiple regression by method of average derivatives. *J. Am. Statist. Ass.*, **84**, 986–995.

Hastie, T. J. and Tibshirani, R. (1986) Generalized additive models (with discussion). *Statist. Sci.*, **1**, 297–318.

Hotelling, H. (1935) The most predictable criterion. *J. Educ. Psychol.*, **26**, 139–142.

Huber, P. J. (1985) Projection pursuit (with discussion). *Ann. Statist.*, **13**, 435–525.

Ichimura, H. and Lee, L. (1991) Semiparametric least squares estimation of multiple index models: single equation estimation. In *Nonparametric and Semiparametric Methods in Econometrics and Statistics* (eds W. Barnett, J. Powell and G. Tauchen). Cambridge: Cambridge University Press.

Li, K. C. (1991) Sliced inverse regression for dimension reduction (with discussion). *J. Am. Statist. Ass.*, **86**, 316–342.

——— (1992) On principal Hessian directions for data visualisation and dimension reduction: another application of Stein's Lemma. *Ann. Statist.*, **87**, 1025–1039.

Li, K. C., Lue, H. H. and Chen, C. H. (2000) Interactive tree-structured regression via principal Hessian directions. *J. Am. Statist. Ass.*, **95**, 547–560.

Rao, C. R. (1973) *Linear Statistical Inference and Its Applications*. New York: Wiley.

Schott, J. R. (1994) Determining the dimensionality in sliced inverse regression. *J. Am. Statist. Ass.*, **89**, 141–148.

Severini, T. A. and Wong, W. H. (1992) Profile likelihood and conditionally parametric models. *Ann. Statist.*, **20**, 1768–1802.

Smith, R. L., Davis, J. M. and Speckman, P. (1999) Assessing the human health risk of atmospheric particles. In *Environmental Statistics: Analysing Data for Environmental Policy*. New York: Wiley.

Stone, M. (1974) Cross-validatory choice and assessment of statistical predictions (with discussion). *J. R. Statist. Soc.* B, **36**, 111–147.

Tong, H. (1990) *Nonlinear Time Series Analysis: a Dynamical System Approach*. Oxford: Oxford University Press.

Xia, Y. and An, H.-Z. (1999) Projection pursuit autoregression in time series. *J. Time Ser. Anal.*, **20**, 693–714.

Xia, Y. and Li, W. K. (1999) On single-index coefficient regression models. *J. Am. Statist. Ass.*, **94**, 1275–1285.

Xia, Y., Tong, H. and Li, W. K. (1999) On extended partially linear single-index models. *Biometrika*, **86**, 831–842.

Yang, L. and Tschernig, R. (1999) Multivariate bandwidth selection for local linear regression. *J. R. Statist. Soc.* B, **61**, 793–815.

Yao, Q. and Tong, H. (1994) On subset selection in nonparametric stochastic regression. *Statist. Sin.*, **4**, 51–70.

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

Discussion on the paper by Xia, Tong, Li and Zhu

J. T. Kent (*University of Leeds*)

The paper is an ambitious attempt to tackle high dimensional regression problems. There are connections to

several areas of statistics, including multivariate analysis, nonparametric regression and linear regression. I would like to direct some comments to each area in turn.

Multivariate analysis

A standard model in multivariate analysis of variance involves k groups of p-dimensional observations X with different means. The group membership can be represented in terms of a random variable y taking integer values $j = 1, \ldots, k$, with probabilities π_j. Conditional on $y = j$, the distribution of X is modelled by $N_p(\mu_j, \Sigma)$, $j = 1, \ldots, k$. Let $\bar{\mu}$ denote the average of these mean values. Canonical variate analysis is a tool for improving the interpretability in this setting via dimension reduction. It is assumed that these means lie on a lower dimensional plane of dimension D, say, where $D < \min(k - 1, p)$, i.e. we assume that the $\{\mu_j - \bar{\mu}\}$ span a subspace of dimension D. Let B $(p \times D)$ be a matrix whose columns span this subspace and let C $(p \times (p - D))$ be a complementary matrix so that (B, C) is non-singular. Reversing the conditioning yields the logistic-type regression model

$$P(y = j|X) \propto \pi_j \exp\{(\mu_j - \bar{\mu})^{\mathrm{T}}\Sigma^{-1}(X - \bar{\mu}) - \tfrac{1}{2}(\mu_j - \bar{\mu})^{\mathrm{T}}\Sigma^{-1}(\mu_j - \bar{\mu})\}$$

in which the exponent is a linear function of X with different coefficients for each j.

It can be checked that this conditional probability in fact depends only on $B^{\mathrm{T}}X$, not on all of X, and so yields the conditional independence statement

$$(y \perp C^{\mathrm{T}}X)|B^{\mathrm{T}}X.$$

Thus this model can be regarded as a discrete and parametric version of the authors' model (1.1). In passing, note that similar conditional independence statements form the building-blocks of graphical models, except that in our setting B is unknown.

In the k-groups model, the marginal distribution of X is a mixture of p-variate normals. However, when attention is focused on the conditional distribution of $y|X$ in the logistic-type regression model, it is usual to allow more general possibilities for the marginal distribution of X. The k-groups model can be viewed as a motivating example for the sliced inverse regression approach to nonparametric multiple regression, whereas the logistic-type regression model better matches the tone of the current paper.

Nonparametric regression

A generalized additive model takes the form $y = \Sigma_{j=1}^{D} g_j(\beta_j^{\mathrm{T}}X) + \varepsilon$. The ridge terms $g_j(\beta_j^{\mathrm{T}}X)$ can be viewed as 'main effects' in the directions β_j. In contrast, the more general model (1.1), $y = g(B_0^{\mathrm{T}}X) + \varepsilon$, which forms the foundation of the paper, also allows 'interaction terms'. However, I am concerned that there is a tendency in practice to interpret the columns of B_0 as main effects and to ignore possible interactions. For example, consider the plots of y *versus* $\hat{\beta}_1^{\mathrm{T}}X$ and y *versus* $\hat{\beta}_2^{\mathrm{T}}X$ in Fig. 5. There are two related problems with these plots. First any possible interactions are ignored; it might be better to represent the whole response surface. The second problem is that these two directions $\hat{\beta}_1$ and $\hat{\beta}_2$ have no preferred status. It is possible to take any other basis of their column space without affecting the validity of the model.

Linear regression

Reduced rank models are also of interest in linear regression analysis. Of course the ordinary least squares regression model is a special case of model (1.1) with $D = 1$ and g linear. However, when p is large, it is well known that the least squares estimator can be unstable, so attempts are often made to reduce the dimensionality of X. One class of methods involves variable selection. However, a class of methods that is more in keeping with the current paper involves the construction of new linear composite variables from X. One of the simplest such methods is principal components regression in which X is replaced by its first few dominant principal components. Unfortunately, this method is rather unsatisfactory since the dominant principal components depend just on the X-variability and not on the relationship to y. A hybrid approach between ordinary least squares and principal components regression is partial least squares; see Stone and Brooks (1990) for a unified treatment. Of course these methods of dimension reduction (including variable selection methods as well) depend heavily on the covariance structure of X.

Are there any lessons from this methodology for this paper? In particular, what happens when there is very high correlation between the X-variables or, more generally, when the X-variables become nearly collinear? My concern is that the estimate of the column space of B will become unstable and that problem (2.7) might have multiple solutions.

I have found the paper tremendously stimulating, and it gives me great pleasure to propose the vote of thanks.

Adrian Bowman (*University of Glasgow*)

It is a great pleasure to add my thanks for this paper. I enjoyed both its reading and its presentation. Over the past few years there has been a considerable amount of work in the dimension reduction area. Regression used to be a topic which we thought we understood. Now we are not so sure. It is one of the merits of this paper that it brings together a variety of approaches in this area and synthesizes them into a simple but potentially powerful idea. Direct and simultaneous estimation of both the nonparametric and the directional components of the model brings some significant benefits. These include an avoidance of some of the usual difficulties with bias incurred by smoothing, a weakening of assumptions, the ability to handle the special but important case of time series, some impressively strong supporting asymptotics and evidence of good behaviour in numerical work. However, it is difficult to believe that these properties are not bought at some price and I would like to explore one or two aspects of where the costs may lie.

The first relevant feature is that, although the central idea is attractively simple, the implementation is necessarily more sophisticated. It involves a variety of steps. The first is smoothing in, possibly high dimensional, covariate space. Most people feel comfortable when applying smoothing in one, two or occasionally three dimensions. The authors have been courageous in going rather beyond that. In the hospital admissions data courage gives way to heroism by smoothing in 42 dimensions. Of course, the refinements introduced by the authors quickly reduce attention to the much smaller dimensional space defined by the current effective dimension reduction (EDR) directions where smoothing can be applied without difficulty. At the same time, there is a high dimensional minimization in operation to identify the EDR directions. Beyond this lies a cross-validation step to compare the EDR dimensions. Finally, there is some mention in the paper of the possibility of using a data-dependent bandwidth choice, although the authors wisely do not routinely incorporate this. The end result is a set of EDR directions which have been produced by a set of complex operations on the data. However, there is no difficulty in principle with that. Complex data may require complex methods of analysis and if the end result brings insight then it has been worthwhile.

On the question of insight, I would like to use the hospital admissions data as a means of raising some practical issues. The first concerns the *robustness* and *sensitivity* of the procedure. A scatterplot matrix reveals a variety of features in the covariates. One is the presence of substantial skewness. The sulphur dioxide variable is a good example of this and it includes in particular two very large observations. Since the sulphur dioxide, nitrogen dioxide and particulates covariates are all concentrations, it would be natural to take a log-transformation of each. Ozone, although also skewed, contains observations at or close to zero and so it may be best left unaltered, along with temperature. Humidity is a percentage, with many observations at high values and so the logistic transformation would be natural here. The question is whether the broad qualitative conclusions of the analysis will remain unchanged when repeated using the variables on these, arguably more natural, scales. The assumptions of the model are weak but one can only feel that there will be greater stability if the variables exhibit approximately normal variation. A second issue arises from the scatterplot of log(nitrogen dioxide) and log(particulates) which shows a strong linear relationship between these two variables. This is exactly the situation assumed by the model. However, it then seems surprising that particulates feature strongly in the conclusions whereas nitrogen dioxide does not. This raises the question of whether the decisions being made by the procedure on the weights to assign to variables are ones which we shall always feel comfortable with.

An issue of the *appropriateness* of the model is raised by the scatterplot of nitrogen dioxide against temperature. This shows a clear non-linear pattern which will be obscured by the linear combinations around which the model is built. Of course, a second dimension will, in this case, allow the full relationship between the covariates to be expressed. However, it would seem more appropriate to incorporate specific non-linear relationships into the model in a more direct way, where these are appropriate.

Finally, some important issues arise under the heading of *interpretation*. The first derives from the fact that EDR delivers a subspace, not a co-ordinate system. The same subspace can be represented by EDR directions which are rotated in different ways. This makes the interpretation of specific elements of the EDR direction vectors rather difficult. The nonparametric surface g has an unspecified shape, built from all EDR directions simultaneously. The marginal space may change radically as the EDR co-ordinate system is rotated. An interpretation can therefore only be made from the entire collection of EDRs and this is not an easy task. In addition, if we simulate data where y is unrelated to x we are still likely to identify EDR directions of apparent meaning. This highlights the need for some statistical methods of model comparison, beyond CV(d), to ensure that the results of EDR can safely be attributed to meaningful structure rather than to noise.

When the authors have come so far, it may seem churlish to ask them to go yet further. However, I raise

these issues in the hope that the authors will be able to devote their considerable powers to addressing them. To return to the original remarks, this is clearly a simple but potentially powerful idea which deserves to be considered carefully. I have great pleasure again in congratulating the authors on their paper and in warmly seconding the vote of thanks.

The vote of thanks was passed by acclamation.

Santiago Velilla (*Universidad Carlos III de Madrid, Getafe*)
In developing minimum average variance estimation (MAVE) the authors seem to have in mind a first-order regression problem in which all the information that X carries on the response y is captured by the conditional expectation $E(y|X)$. In this sense, the populational objective function (2.1) and its sample version (2.7) seem to be appropriate when the error ε in model (1.1) not only satisfies the condition $E(\varepsilon|X) = 0$ but also $\text{var}(\varepsilon|X) = \sigma^2$. If the conditional variance is not constant, expressions (2.1) and (2.7) should perhaps be modified accordingly.

In comparing the four new methods proposed in this interesting paper, I find that both the outer product of gradients method, in Section 3.1, and inverse MAVE, in Section 3.2, have a natural *nested* character. Once a decision has been taken on the value of the dimension of the effective dimension reduction space, directions are determined sequentially. In contrast, both MAVE and refined MAVE seem to require specific computation in each step $d = 1, 2, \ldots$. Moreover, as indicated in the algorithm of Section 2.3, computation is required for all $1 \leqslant d \leqslant p$. In view of the pattern of Tables 3, 5 and 6 in the examples in Sections 5.1 and 5.2, where the change in the CV(d) value is 'small' when spurious directions are considered, for 'large' values of d the algorithm could be initialized using the results for $d - 1$ making it 'nested', i.e. looking only for $\hat{\beta}_d$, once $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_{d-1}$ have been determined. Of course, this is just a suggestion based on the pattern of the tables in the examples, but this simplified scheme for *spurious* values of d might save some computational time.

Finally, in connection with condition (1.2), in Velilla (1998), section 4.1, I proposed a method for generating regressors X satisfying condition (1.2) that are not necessarily elliptical. This method has been applied, for example, in Bura and Cook (2001a, b) for assessing by simulation the performance of some methods for testing for dimension.

Wenyang Zhang (*University of Kent at Canterbury*)
I have two comments to make on this interesting paper.

Shannon's entropy
A measure of uncertainty, Shannon's entropy, was introduced by Shannon (1948), which is extremely useful in communication theory. It also can be used to reduce dimension in regression to avoid the 'curse of dimensionality'.

Let ξ and η be two random variables with joint density function $f(x, y)$. $p(x)$ is the density of ξ, the entropy of ξ is defined as

$$H(\xi) = -\int p(x)\, \log\{p(x)\}\, \mathrm{d}x$$

and the conditional entropy of ξ given η is

$$H_\eta(\xi) = -\int\int f(x, y)[\log\{f(x, y)\} - \log\{q(y)\}]\, \mathrm{d}x\, \mathrm{d}y$$

where $q(y)$ is the density of η. The information contained in η about ξ is

$$I(\xi, \eta) = H(\xi) - H_\eta(\xi).$$

Let Y be the response, X be the covariate with high dimension p and $(X_i, Y_i), i = 1, \ldots, n$, be a sample from (X, Y). For any fixed β, the estimate $\hat{I}(Y, \beta^\tau X)$ of $I(Y, \beta^\tau X)$ can be obtained by standard density estimation; see Fan and Gijbels (1996). An alternative dimension reduction procedure is maximize $\hat{I}(Y, \beta^\tau X)$ subject to $||\beta|| = 1$, to find the maximizer β_1 and maximum I_1, then maximize $\hat{I}(Y, \beta^\tau X)$, subject to $\beta^\tau\beta_1 = 0$ and $||\beta|| = 1$, to find the maximizer β_2 and maximum I_2, and continue this exercise until I_q is less than a selected critical value c which may be obtained by cross-validation. $(\beta_1, \ldots, \beta_q)$ forms the efficient directions to reduce the dimension. It would be very interesting to compare this approach with that in the paper.

Curse of dimensionality
In Section 2.1.1, the initial B is obtained based on

$$w_{i0} = K_h(X_i - X_0) \Big/ \sum_{l=1}^{n} K_h(X_l - X_0).$$

If the dimension p of X is very large, it would be impossible to obtain an initial B with small bias owing to the 'curse of dimensionality'. My question is does this bias matter in your procedure? If not, why could we not take the whole range of X_i as the initial bandwidth?

Frank Critchley (*The Open University, Milton Keynes*)
In welcoming the faster rate of consistency and time series extensions afforded by the paper, I would like to make the following points in which $Y_x := (Y|X = x)$ and $\varepsilon_x := (\varepsilon|X = x)$.

(a) I was somewhat surprised not to find fuller reference to the important body of work by Cook and co-workers, surveyed to that date in Cook (1998). Among other attractive features, such as its graphical emphasis, this approach examines how the whole distribution of Y_x—not just, as here, its mean $g(x)$—varies with x. Again, it exploits a conditional independence formulation throughout, that is both logically cogent and statistically intuitive. I would also like to draw attention to two forthcoming papers, available on the *Annals of Statistics* Web site and directly relevant to this paper: Cook and Li (2002), which addresses dimension reduction for $g(x)$, and Chiaromonte *et al*. (2002), which overlaps with Section 3.4.
(b) There are two apparent significant errors of omission.
   (i) In the sentence two after equation (1.1), a simple counter-example is

$$X \sim N_2(0, I),$$

$$Y = g(X_1) + \varepsilon$$

and

$$\varepsilon_x \sim N(0, \sigma^2 x_2^2).$$

   The omission appears to be that model (1.1) should be augmented by the *location regression* requirement $Y \perp\!\!\!\perp X | E(Y|X)$ (Cook (1998), page 111); a similar remark applies to model (1.3).
   (ii) In the sentence including expression (2.1), additional conditions—such as constancy of $\mathrm{var}(\varepsilon_x)$ over x—apparently are required.
(c) The benefits of this paper—including relaxation of condition (1.2) on X—come at the price of other non-trivial restrictions to its applicability: in particular, to additive error models that are special cases of location regression and in which certain additional conditions hold.
(d) In unpublished preliminary discussions with Cook, it was noted that the conditional independence approach seems natural in a variety of time series contexts, autoregressive processes being obvious examples. This would seem a promising line of enquiry.
(e) In view of the quadratic nature of the criterion minimized, I was somewhat surprised by the robustness to outliers claim (Section 6) and would value further details.
(f) Concerning Section 2.1.2, under what conditions is convergence (to a unique solution) guaranteed?

Anthony Atkinson (*London School of Economics and Political Science*)
I congratulate the authors on an interesting paper which stimulated an excellent discussion. I have five points.

(a) John Kent placed the authors' proposal in the context of other dimension reduction methods, including partial least squares. This method is often used with p close to n. Is this likely to cause any problems? Partial least squares is also often used with $p \gg n$, e.g. in the spectroscopic data set analysed again by Brown *et al*. (2001). Can the authors' method be extended to this important class of problems?
(b) The interpretation of results like those of Table 4 seems beset with difficulties, since the directions

can be rotated in the D-dimensional subspace. Basilevsky (1994), section 6.10, discussed the similar problem of rotation and interpretation in factor analysis.

(c) On pages 378–379 the data have the effects of two factors removed, so that the y_t are indeed notationally abused, being residuals. The method of added variables (e.g. Atkinson and Riani (2000), section 2.2) indicates that the same regression should be performed on the explanatory variables as on the response, so that the analysis becomes one of residuals on residuals. Incidentally, this use of only one set of residuals is a frequent occurrence in time series analysis, where a series is 'pre-whitened', but the regressors left untouched.

(d) Some dicussants have mentioned robustness. It has been the experience of Marco Riani and myself that use of the forward search (Atkinson and Riani, 2000) reveals masked outliers and their effects in a way that is impossible by looking at a fit to all the data. The data are fitted to subsets of increasing size and parameter estimates, residuals and other quantities monitored. The starting-point for the searches is a robustly chosen subset of p, or a few more, observations. Could relatively small subsets of the data be used here to start such a process?

(e) Many statistical methods, including, I suspect, that described here, tend to work better if the data are approximately normal. In applications of inverse regression for dimension reduction, the data are sometimes transformed to approximate multivariate normality by using a multivariate Box–Cox transformation. An example is the analysis of data on New Zealand mussels in chapters 10 and 11 of Cook and Weisberg (1994). A robust version of this transformation using the forward search is illustrated in Riani and Atkinson (2001). What is the effect here of such transformations both on computation time and on the conclusions drawn from Tables 4 and 7?

Qiwei Yao (*London School of Economics and Political Science*)

The authors should be congratulated for making a further contribution along their impressive list of publications on nonparametric multivariate regression—a very important and immensely difficult topic.

Theorem 1 may be presented in a slightly stronger form by defining the weights w_{ij} in terms of $\{B^{\mathrm{T}}X_i\}$ instead of $\{X_i\}$. This effectively changes a p-dimensional smoothing problem into a d-dimensional one. The gain in convergence rate would now be $h_{\mathrm{opt}} \log(n) = O\{n^{-1/(d+4)} \log(n)\}$ at the price of the added computational complication in the minimization of problem (2.7).

As B_0 is only defined up to any orthogonal transforms, will the alternating iteration between refined kernel weights and estimating β_j in step 1(b) lead to stable $\hat{\beta}_j$? The use of refined kernel weights only makes sense if such a stable solution is guaranteed.

An alternative version for the distance measure would be

$$m(\hat{B}, B_0) = ||(I - B_0 B_0^{\mathrm{T}})\hat{B}|| + ||(I - \hat{B}\hat{B}^{\mathrm{T}})B_0||.$$

Then $m(\hat{B}, B_0) \to 0$ in probability if and only if $\hat{B}$ estimates B_0 'correctly'.

Finally the method proposed is most useful when D is small such as 2 or 3, as we still need to estimate the link function even if we have the right effective dimension reduction. If model (1.1) does not hold, will the procedure lead to a 'good' approximation for the conditional expectation of y given X?

A. H. Welsh (*University of Southampton*)

Comparisons of minimum average variance estimation (MAVE) with sliced average variance estimation (SAVE) proposed by Cook and Weisberg (1991) (see Cook and Yin (2001) for recent references) in addition to sliced inverse regression may be interesting and more insightful. Robustness issues in sliced inverse regression and SAVE were raised at the 2000 Australian conference in a presentation by Ursula Gather and the discussion to Cook and Yin (2001). The issues are subtle so the claim that MAVE has good robustness properties needs a proper investigation.

In the single-index model, the asymptotic distribution of $\tilde{\beta}$ is essentially determined by

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \hat{b}_j \, w_{ij}(X_i - X_j)\{\varepsilon_i + g(X_i^{\mathrm{T}}\beta_0) - \hat{a}_j\},$$

the 'numerator' in $\tilde{\beta}$. The approach in which we estimate g and g' by smoothing (as in the present paper) but estimate β_0 by standard maximum likelihood (Brillinger, 1992; Weisberg and Welsh, 1994) seems rather different. However, it is important to centre X_i about an estimate of $E(X|X^{\mathrm{T}}\beta_0 = X_i^{\mathrm{T}}\beta_0)$ and, under the simplifying conditions of the present paper and using local linear smoothing (Ruckstuhl and Welsh, 1999), the equivalent expression for this estimator is

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\hat{b}_i\tilde{w}_{ji}(X_i - X_j)\{\varepsilon_i + g(X_i^{\mathrm{T}}\beta_0) - \hat{a}_i\}.$$

Whereas we usually use undersmoothing, higher order kernels or higher order polynomials in local polynomial smoothing to increase the rate of convergence of $\hat{a}_i - g(X_i^{\mathrm{T}}\beta_0)$ so that it is asymptotically negligible, MAVE estimates integrals of g rather than g so we can use optimal bandwidths for g while estimating β_0. If the above expressions are correct, MAVE should have the same asymptotic distribution (possibly up to centring of the covariates) as the maximum likelihood estimator but this needs to be checked carefully. Finally, MAVE should also be extended to other distributions, presumably by maximizing the average local log-likelihood.

Hengjian Cui (*Beijing Normal University*) **and Guoying Li** (*Academy of Mathematics and System Sciences, Beijing*)

This paper is very interesting and very provocative! The authors give us new ideas to search the effective dimension reduction (EDR) space in nonparametric regression settings.

The minimum average variance estimation (MAVE) is effective provided that model (1.1) is correct. It is different from projection pursuit (PP) (Huber, 1985; Li and Cheng, 1993), which assumes that the link function is a sum of several ridge functions; we call it the PP regression (PPR) model here. If model (1.1) is true, the first PP approximation is $E(y|\beta_1^{\mathrm{T}}X)$. However, β_1 is not necessarily in the space spanned by B_0 although $E(y|\beta_1^{\mathrm{T}}X)$ is the first-order optimal PP approximation of $g(B_0^{\mathrm{T}}X)$. If the PPR model is true and the number of ridge functions is less than p, model (1.1) holds obviously. However, MAVE concentrates on finding the EDR directions whereas the PP approach provides estimators for both the directions and the link function. Another point is that MAVE uses a high dimensional kernel whereas PP needs only a one-dimensional kernel. To simplify computation in MAVE, we may use the following iterative algorithm to search the EDR directions one by one:

$$\min_{\substack{\beta_d \perp \hat{\beta}_1,\ldots,\hat{\beta}_{d-1}\\ ||\beta_d||=1}} \left\{ \sum_{j=1}^{n} \hat{\sigma}^2_{\hat{B}_{d-1},\,\beta_d}(\hat{\beta}_1^{\mathrm{T}}X_j, \ldots, \hat{\beta}_{d-1}^{\mathrm{T}}X_j, \beta_d^{\mathrm{T}}X_j) \right\}$$

where $\hat{B}_{d-1} = (\hat{\beta}_1, \ldots, \hat{\beta}_{d-1})$. Then, the associated p-dimensional kernel can be taken as a product of p one-dimensional kernels. This intuitively makes sense by theorem 1 and lemma 1. Also, we may refine the kernel weights and determine the number D by the procedures described in Sections 2.1.2 and 2.2 respectively.

The example in Section 5.2 shows that the (refined) MAVE method is robust. It seems to us that it is robust against outliers in X-space because the local smoother puts lower weights on further X_js. If the outliers occur in Y-space the story may be different.

There are at least two obvious questions. One is the inference of the EDR directions, which involves the asymptotic normality of the $\hat{B}$. This is true for single-index models (Härdle *et al.*, 1993; Xia and Li, 1999). We believe that the $\hat{B}$ obtained by (refined) MAVE has $\sqrt{n}$-consistency and asymptotic normality under some regular conditions. The expression of the asymptotic covariance matrix of $\hat{B}$ could be complicated, and its consistent estimator is needed. This may be given by, say, a bootstrap method. Moreover, the estimation of the link function is also important. In particular, we may first ask whether the link function is additive (Cui *et al.*, 2001). Also, it is expected that the MAVE method may be extended to the case that X includes continuous as well as categorical (or, generally, discrete) or functionally related covariates, as mentioned in Section 3.4. Further work is definitely needed in this area.

Vladimir Spokoiny (*Weierstrass Institute and Humboldt University, Berlin*)

The authors discuss an excellent idea for solving the dimension reduction problem by minimizing the sum

$$\sum_{j=1}^{n}\sum_{i=1}^{n}[y_i - \{a_j + b_j^{\mathrm{T}}B^{\mathrm{T}}(X_i - X_j)\}]^2 w_{ij}$$

over all $p \times D$ matrices B fulfilling $B^{\mathrm{T}}B = 1$. Here w_{ij} are non-negative weights. The approach has genuine benefits compared with the existing methods like sliced inverse regression or average derivative estimation. The choice of the weights w_{ij} plays the central role in this method. The authors discuss two possibilities. The first is to apply the usual multidimensional kernel weights

$$w_{ij} = K_h(X_i - X_j) \Big/ \sum_l K_h(X_l - X_j).$$

This approach, similarly to the average derivative estimation or outer product of gradients methods, suffers from the *curse of dimensionality* problem. Indeed, even for the optimal choice of the bandwidth h, the accuracy of estimation of the effective dimension reduction space is very low if the dimensionality p is large. The refined weights

$$\tilde{w}_{ij} = K_h\{\hat{B}^{\mathrm{T}}(X_i - X_j)\} \Big/ \sum_l K_h\{\hat{B}^{\mathrm{T}}(X_l - X_j)\}$$

are based on the knowledge of the structure of the model and they allow us to obtain better accuracy of estimation corresponding to the problem of the reduced dimension. However, the refined weights proposal utilizes the estimator $\hat{B}$ which comes from the first-step estimation with the multidimensional weights. If this first-step estimator is not sufficiently precise then the advantage of using the refined weights disappears and the whole procedure may fail in estimating the true effective dimension reduction. Hristache, Juditski and Spokoiny (2001) and Hristache, Juditski, Polzehl and Spokoiny (2001) proposed another way of selecting the refined weights w_{ij} based on the idea of structural adaptation. The idea is to pass progressively from multidimensional weights w_{ij} to the low dimensional weights of type $\tilde{w}_{ij}$. In this context, an interesting question is the possibility of joining the proposal of this paper (to estimate the index space by minimizing the mean average squared error) with the structural adaptation method.

The following contributions were received in writing after the meeting.

K. S. Chan (*University of Iowa, Iowa City*) **and Ming-Chung Li** (*EMMES Corporation, Rockville*)

We congratulate the authors for their masterly piece of work that will certainly stimulate much research on semiparametric modelling and non-linear time series.

The authors considered the case of univariate responses. Interestingly, we have independently done some related work with multivariate responses. Li and Chan (2001) (and also Li (2000)) proposed the semiparametric reduced rank regression model

$$Y_t = C\, f(BX_t) + \varepsilon_t,$$

where Y_t and X_t are m- and n-dimensional componentwise standardized random vectors, ε_t is of zero mean and identical variance given the current X and past Xs and Ys, C and B are $m \times r_1$ and $r_2 \times n$ coefficient matrices and r_1 and r_2 are the ranks of the model. The unknown (link) function f maps from R^{r_2} to R^{r_1}. The model is unaltered on replacing C, $f(\cdot)$ and B by CP, $P^{-1}\, f(Q^{-1}\cdot)$ and QB for any two invertible matrices P and Q. So, identification requires constraining, for example, the leading subsquare matrices of C and B as identity matrices, after suitable permutations of the variables. We may interpret the r_1 components of $f(BX_t) = (f_1(U_{1,t}, \ldots, U_{r_2,t}), \ldots, f_{r_1}(U_{1,t}, \ldots, U_{r_2,t}))^{\mathrm{T}}$ as non-linear principal components which depend on the indices $BX_t = (U_{1,t}, \ldots, U_{r_2,t})^{\mathrm{T}}$. Li and Chan (2001) proposed an estimation procedure that resembles the minimum average variance estimation method for $m = 1$.

We now use the respiratory problem data to illustrate the semiparametric reduced rank regression model with some preliminary analysis of the dynamic structure of air pollution in Honk Kong. Let Y consist of (log-transformed) sulphur dioxide (S), nitrogen dioxide (N), (log-transformed) respirable suspended particulates (P) and (square-root-transformed) ozone (O); X consists of lags 1, 2 and 7 of the Y-variable and lags 0 and 1 of temperature (T) and humidity (H). From cross-validation, $r_1 = r_2 = 2$. B is estimated to equal (standard errors are given in parentheses; NA denotes 'not applicable')

$$\begin{pmatrix}
S_{t-1} & N_{t-1} & P_{t-1} & O_{t-1} & S_{t-2} & N_{t-2} & P_{t-2} & O_{t-2} \\
-0.617 & 1 & -0.011 & 0.523 & 0.038 & -0.033 & 0.046 & -0.146 \\
(0.085) & (\mathrm{NA}) & (0.104) & (0.122) & (0.087) & (0.117) & (0.099) & (0.083) \\
0.510 & 0 & 0.159 & -0.121 & -0.110 & -0.057 & 0.036 & 0.034 \\
(0.064) & (\mathrm{NA}) & (0.076) & (0.079) & (0.070) & (0.087) & (0.074) & (0.061)
\end{pmatrix}$$

$$\begin{pmatrix}
S_{t-7} & N_{t-7} & P_{t-7} & O_{t-7} & T_t & T_{t-1} & H_t & H_{t-1} \\
-0.136 & 0.036 & 0.104 & 0.084 & 0 & 0.210 & -1.145 & 0.071 \\
(0.064) & (0.085) & (0.087) & (0.060) & (\mathrm{NA}) & (0.075) & (0.177) & (0.112) \\
0.120 & 0.018 & -0.038 & -0.047 & 1 & -1.167 & 0.349 & -0.179 \\
(0.049) & (0.067) & (0.060) & (0.048) & (\mathrm{NA}) & (0.056) & (0.093) & (0.074)
\end{pmatrix}.$$

Here, the subsquare matrix corresponding to N_{t-1} and T_t is normalized as the identity matrix. Fig. 8 displays the smoothed graphs of the non-linear principal components $\hat{f}_i$ *versus* the indices u_1 and u_2. Whereas $\hat{f}_1$ seems linear, $\hat{f}_2$ appears to be piecewise linear. Below is the estimate of C and that after transformation that renders the two non-linear principal components uncorrelated and of unit variance:

$$\hat{C} = \begin{pmatrix} 0.000 & 1.000 \\ (\text{NA}) & (\text{NA}) \\ 1.000 & 0.000 \\ (\text{NA}) & (\text{NA}) \\ 1.065 & -0.189 \\ (0.038) & (0.063) \\ 0.977 & -0.890 \\ (0.050) & (0.097) \end{pmatrix}; \qquad \hat{C}_{\text{rotated}} = \begin{pmatrix} 0.124 & 0.526 \\ (\text{NA}) & (\text{NA}) \\ 0.729 & 0.124 \\ (\text{NA}) & (\text{NA}) \\ 0.753 & 0.033 \\ (0.029) & (0.034) \\ 0.601 & -0.347 \\ (0.036) & (0.050) \end{pmatrix}$$

The Euclidean distance between any two rows of the rotated C measures the dissimilarity in the dynamics of the corresponding variables. The rotated $\hat{C}$ suggests that the sulphur dioxide variable enjoyed different dynamics from other variables whereas the suspended particulates and nitrogen dioxide variables shared similar dynamics, over the study period; see also Fig. 8.

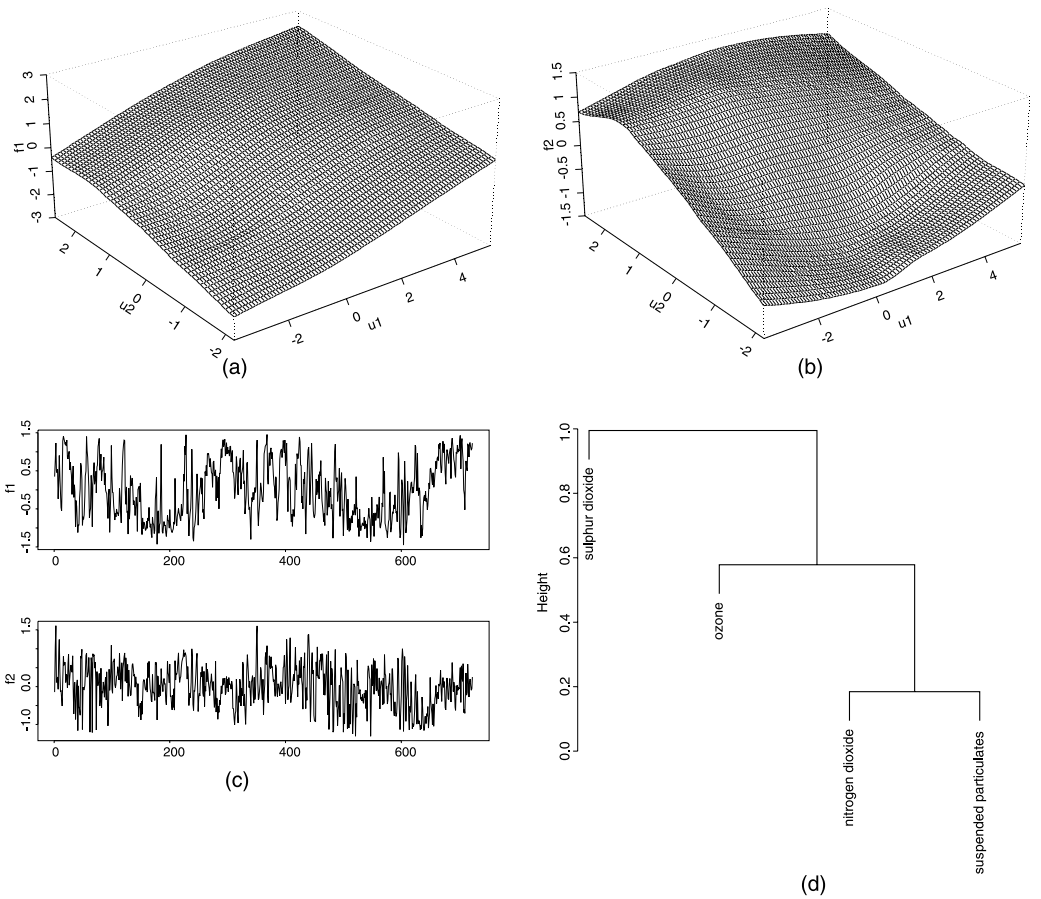


Fig. 8. (a) Smoothed graph of $\hat{f}_1$, (b) smoothed graph of $\hat{f}_2$, (c) time series plots of the two non-linear principal components and (d) dendrogram from a cluster analysis of the dynamics of the four pollution variables, based on $\hat{C}_{\text{rotated}}$

Pavel Čížek and Wolfgang Härdle (*Humboldt University, Berlin*) **and Lijian Yang** (*Michigan State University, East Lansing*)

This paper addresses the challenging problem of dimension reduction and we congratulate the authors for this new insight into modelling high dimensional data. They provide the new minimum average variance estimation (MAVE) approach that creates a variety of semiparametric modelling strategies. The technical treatment is excellent and the algorithms derived are directly implementable. From a practitioner's point of view, there are probably questions about the performance of the method in non-standard situations.

For an assumed number of directions, the MAVE method is based on the local linear approximation of a regression function. The main idea is to use this approximation (conditionally on yet unknown indices) directly in the local linear smoothing procedure by using a multidimensional kernel. This is just a simultaneous minimization with respect to function and direction estimates, which is broader than the usual methods that estimate only function values or only directions. According to theorem 1, this makes undersmoothing of the bandwidth selection unnecessary. Additionally, MAVE together with a cross-validation procedure can be used to estimate the effective dimension reduction (EDR) dimension.

On the basis of MAVE, the authors design generalizations of several existing methods (e.g. the outer product of gradients (OPG) method is a generalization of additive derivatives estimation by Härdle and Stoker (1989)). Additionally, these extensions even outperform the original methods. However, we must keep in mind that these generalizations are valid only under assumptions on the smoothness of all the variables and cannot therefore replace the corresponding single- and multi-index methods that can also handle discrete variables (e.g. semiparametric least squares by Ichimura (1993)).

Finally, the MAVE method is claimed to be robust against outliers, supposedly in the space of explanatory variables. We examined the robustness of the choice of the EDR dimension and the OPG and MAVE methods to outliers and random noise in more detail. In the first case, our simulations regarding the cross-validation procedure in the presence of a single outlier show two main effects: the outlier results generally in an upwardly biased estimate of the EDR dimension, and additionally, in most cases, model estimates under contamination do not reduce the variance of the dependent variable conditionally on the regression function. In the second case, we studied the behaviour of MAVE and OPG under contamination. The most interesting result is that OPG, which for clean data is always worse than MAVE, can keep up with or even outperform MAVE when applied to contaminated data. We achieved similar results also under no contamination and a high variance of the error term.

R. D. Cook (*University of Minnesota, St Paul*)

The authors refer to span(B_0) from model (1.1) as the effective dimension reduction (EDR) subspace, but I find this characterization to be incorrect. Li (1991) defined the EDR subspace as the span(B) in the representation $y = g(B^{\mathrm{T}}X, \delta)$, where the error $\delta \perp\!\!\!\perp X$ and $B = (b_1, \ldots, b_k)$. Because ε may depend on X, equation (1.1) permits a model with $\varepsilon = \sigma(C_0^{\mathrm{T}}X)\delta$, where $\sigma(C_0^{\mathrm{T}}X) \geqslant 0$. For this version of model (1.1), the EDR subspace is span(B_0) + span(C_0), not span(B_0) as the paper implies. This confusion is unfortunate but perhaps understandable because published descriptions of the EDR subspace are not explicitly constructive.

A mean subspace is any subspace span(B) of $\mathbb{R}^p$ such that $y \perp\!\!\!\perp E(y|X)|B^{\mathrm{T}}X$. If the intersection of all mean subspaces is itself a mean subspace it is called the central mean subspace (CMS) and may be taken as the subject of a regression inquiry. Recently introduced by Cook and Li (2002), the CMS seems to be the subspace pursued in this paper.

A dimension reduction subspace (DRS) is any subspace span(B) such that $y \perp\!\!\!\perp X|B^{\mathrm{T}}X$. When the intersection of all DRSs is itself a DRS it is called the central subspace (CS; Cook (1996a, b, 1998)), which is a metaparameter for dimension reduction. The CS may not exist when the EDR subspace does exist. And the CS may exist straightforwardly when the construction of the EDR subspace is problematic (e.g. binary responses). I find the CS to be *much* easier to handle in theory and widely applicable in practice. The CMS is contained in the CS. The CS is invariant under strictly monotonic transformations of Y, whereas the CMS and span(B_0) are not. Compactness of the support of X is not required for the CMS or the CS (see the discussion following lemma 1).

I do not regard sliced inverse regression (SIR) and refined minimum average variance estimation (RMAVE) to be direct competitors. SIR estimates directions in the CS, whereas RMAVE apparently estimates the CMS. The authors demonstrate that RMAVE does better than SIR in some situations that RMAVE was designed to handle. I wonder how RMAVE would perform across the many situations where SIR, sliced average variance estimation and related methods have apparently uncovered key regression structures.

The fact that SIR will not perform well in models like model (4.3) is known (Cook and Weisberg, 1991). Does the performance of RMAVE degrade when there are strong non-linear relationships among the predictors, the kind that would render SIR ineffective?

I found this paper interesting because of the suggestion that local methods might mitigate the need for restrictions on the predictors.

Jianqing Fan (*University of North Carolina at Chapel Hill*)

Model parameters and identifiability

The basic assumption of the paper is that model (1.1) holds. In practice, it is at best an approximation. In general, following Fan *et al*. (2001), the parameters $\mathbf{B}_0$ and the function g can be defined as the minimizer of

$$\min_{g\mathbf{B}:\mathbf{B}^{\mathrm{T}}\mathbf{B}=I_D} [E\{Y - g(\mathbf{B}^{\mathrm{T}}\mathbf{X})\}^2] = \min_{\mathbf{B}:\mathbf{B}^{\mathrm{T}}\mathbf{B}=I_D} [E\{Y - E(Y|\mathbf{X}^{\mathrm{T}}\mathbf{B})\}^2].$$

This is the same as expression (2.1). Hence, the model assumption (2.1) is not needed as far as the procedure for estimating $\mathbf{B}_0$ and g is concerned. Under what conditions does the optimization problem (2.1) have a unique solution, namely when is the parameter $\mathbf{B}_0$ identifiable? (Indeed, only the space spanned by the columns of $\mathbf{B}_0$ is possibly identifiable.)

The identifiability condition is necessary for asymptotic results to hold. To elaborate the identifiability issue, consider the model studied by Fan *et al*. (2001):

$$Y = \sum_{j=0}^{p} g_i(\mathbf{B}^{\mathrm{T}}\mathbf{X})X_j + \varepsilon$$

with $X_0 = 1$. Consider the specific case where $D = 1$ and write $\mathbf{B} = \boldsymbol{\beta}$. When $g_j(x) = \alpha_j x$ with $\alpha_0 = 0$, this model becomes

$$Y = (\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{X})(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}) + \varepsilon,$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^{\mathrm{T}}$. When they are not parallel, the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are not identifiable for $D = 1$. This is the only case where the parameters are not identifiable for $D = 1$, following theorem 1 of Fan *et al*. (2001). This case does not appear in model (1.1), since the authors implicitly assume that $g(B_0^{\mathrm{T}}\mathbf{X}) = E(Y|B_0^{\mathrm{T}}\mathbf{X})$.

Minimum average variance estimation and profile likelihood

The profile likelihood is commonly used to estimate parameters and nonparametric functions in semiparametric models. The basic idea, in the current context, is to estimate the function g for a given $\mathbf{B}$ by using a nonparametric approach, resulting in an estimator $\hat{g}_{\mathbf{B}}(\cdot)$. Now, find the parameter $\mathbf{B}$ to minimize

$$\sum_{i=1}^{n} \{Y_i - \hat{g}_{\mathbf{B}}(\mathbf{B}^{\mathrm{T}}\mathbf{X}_i)\}^2.$$

The fully iterated procedure in Carroll *et al.* (1997) used this idea. Minimum average variance estimation is a nice variation of the profile likelihood method. It is motivated from estimating the conditional variance by a kernel estimator rather than minimizing directly the mean-square errors. As a result, it has the nice expression (2.7) which facilitates theoretical studies but involves an extra loop of summation in computation. The merits of both approaches are worth exploring further. However, it is worthwhile to mention that the profile likelihood method generally gives semiparametric efficient estimators (see, for example, Carroll *et al.* (1997) and Murphy and van der Vaart (2000)). Whether minimum average variance estimation has this kind of optimality remains to be seen. Two procedures share at least one merit in common: no undersmoothing is needed for estimating parametric components (Carroll *et al.* (1997) and theorem 1 of the present paper). In fact, the criteria that the two procedures optimize are approximately the same.

Expression (2.7) is somewhat informal, since its minimization with respect to $\mathbf{B}$ is not unique though its effective dimension reduction is. Could the authors therefore explain how problem (2.7) is minimized and clarify the convergence criterion in Section 2.3?

L. Ferré (*University of Toulouse le Mirail*)
The paper is interesting since it substitutes local linear smoothing for inverse regression for estimating the effective dimension space. The main advantage of the method over inverse regression is that condition (1.2) is relaxed, allowing applications to time series. Even if my own experience of the application of sliced inverse regression in times series is quite positive, time reversibility is indeed an awkward condition derived from equation (1.2). However, an argument in favour of inverse regression is simplicity: estimates of the effective dimension reduction space are deduced from a simple eigenvalue decomposition of a matrix independently from g. This feature allows in particular extensions to functional data (see for example Dauxois *et al.* (2001)). This necessary reduction of the dimension (recall the goal: overcome the 'curse of dimensionality') comes before (and independently of) the nonparametric estimation of g. For deriving this dimension, tests have been proposed, relying, in the original papers, on distributional assumptions. These assumptions can be removed since recent unpublished work has shown that the existence of the first four moments is sufficient. An alternative is to use a model selection approach based on the distance between $S(B_0)$ and $S(\hat{B}_d)$ by letting d vary (Ferré, 1998). The main idea is that a working dimension that is lower than the 'true' dimension D can be preferable and the distance between $S(\hat{B}_d)$ and a d-subspace of the unknown $S(B_0)$ is finally used. Simple estimates of this criterion have been proposed for elliptically distributed explanatory variates but also for the general case by using the bootstrap or jackknife (see Ferré (1997, 1998)). Local linear smoothing intends to estimate at the same time the regression function and the effective dimension reduction space. The price to pay is that more local linear smoothing is needed than covariates are included in the model. For the dimensionality a global model selection approach is considered, but cross-validation, in addition to the high computational cost, does not avoid the curse of dimensionality. Indeed, $\hat{a}_{d0,j}$ is the Nadaraya–Watson estimator which may perform poorly for large values of d and my feeling is that overparameterization is to be feared.

Ker-chau Li (*University of California at Los Angeles*)
The dramatic improvement of the methods proposed over sliced inverse regression (SIR) and the principal Hessian directions method for the three examples deserves some non-asymptotic explanations. For $n = 200$ and $p = 10$, it is difficult to tell why the nice asymptotic theorems are relevant. For the first two examples, a simple explanation goes like this. First, least squares regression is known to be consistent in finding an effective dimension reduction direction (Brillinger, 1983; Li and Duan, 1989) under condition (1.2). It is straightforward to extend this result to weighted least squares regression provided that the weight function depends on (y, x) only through $(y, B_0^{\mathrm{T}}X)$. Now because equation (2.6) is basically a weighted least squares regression, one can prove that, for the population version of equation (2.6), $b^{\mathrm{T}}B^{\mathrm{T}}$ should be in the effective dimension reduction space. If condition (1.2) does not hold, then the result may be biased and an upper bound of bias can be evaluated (Duan and Li, 1991; Li, 1997). Problem (2.7) amounts to averaging over a number of weight functions. Averaging may help the cancellation of bias in the time series context.

For fairness, I would like to point out that weighted versions of SIR and similar procedures have been proposed before to temper the bias problem; see the discussion and rejoinder in Li (1991). It is worth pointing out the difference between condition (1.2) and elliptical symmetry (Hall and Li, 1993). Also SIR and principal Hessian directions can be applied to residuals after deterministic components have been taken out. Iteration does improve the results. However, the issue of non-linear confounding (Li, 1997) sets a limitation that is difficult to bypass by any procedure. It is not clear to me whether the new approach can do anything about it.

For brevity, I shall not go over the long list of clever ideas that I found interesting in this path breaking work by the authors. Let me close by noting that they did not compare their procedure with projection pursuit regression. A dozen years ago when I submitted my SIR paper to the *Journal of the American Statistical Association*, the Associate Editor recommended rejection because he or she thought that SIR was not as good as projection pursuit regression. Luckily my paper was salvaged by the Editor, who allowed me to explain the difference between the two approaches. Apparently the authors have done more than enough to convince the reviewers just as they have convinced me!

Lexin Li (*University of Minnesota, St Paul*)
Adopting the notation in model (1.1) and following the definitions of the central mean subspace (CMS) (Cook and Li, 2002), the minimum average variance estimation (MAVE) methods seem to pursue the CMS only. To confirm this, simulations were done on models of the form $y = g(B_1^{\mathrm{T}}X) + h(B_2^{\mathrm{T}}X)\varepsilon$, where g and h are both unknown functions, ε is independent of X and $E(\varepsilon) = 0$. My results indicate that MAVE

methods can successfully estimate B_1 in the mean structure $E(y|X)$, whereas they always miss B_2 in the error structure.

Refined MAVE (RMAVE) does not require sliced inverse regression's (Li, 1991) linearity condition. Simulations were done to examine the performance of RMAVE when there are strong non-linear relationships among the predictors X. I considered one-dimensional models only, where $B \in \Re^p$. The results show that RMAVE has good performance for one-dimensional models when the non-linearity in X is strong.

Under the assumption $D = 1$, however, there is still room for improvement, compared with RMAVE, to estimate the underlying true direction without the requirement of the linearity condition. Cook and Nachtsheim (1994) suggested a co-ordinatewise reweighting approach to remove the non-linearity in X and to make X elliptically contoured. I have been investigating the possibility of extending the idea of removing the non-linearity in X by clustering on X-space as the first step. An ordinary least squares (OLS) estimate is obtained from each cluster, and all those estimates are combined to estimate the true direction. Intuitively, the clusterwise OLS method works because non-linearity in X is broken and within each cluster the linearity condition should hold approximately. Then the Li–Duan proposition (Li and Duan (1989), theorem 2.1, and Cook (1998), proposition 8.1) is applicable within each cluster. I also consider an iterative version of the algorithm, which obtains the estimate by iteratively clustering on $\hat{B}_i^{\mathrm{T}} X$, where $\hat{B}_i$ is the estimate from the ith iteration. Simulations show that the OLS estimate with clustering achieves a better performance than RMAVE. As an example, consider the model $x_1 \sim \text{uniform}(0,1)$ and $x_2 = \log(x_1) + e$, where $e \sim \text{uniform}(-0.3, 0.3)$, and $y = \log(x_1) + \varepsilon$, where $\varepsilon \sim N(0, 0.01)$. The actual direction is $B = (1, 0)^{\mathrm{T}}$. With 100 observations, RMAVE gives an estimate of $\hat{B} = (0.991, 0.133)^{\mathrm{T}}$ with the angle to B equal to 7.626°, whereas OLS with five clusters produces $\hat{B} = (0.999, 0.038)^{\mathrm{T}}$ with the angle to B equal to 2.196°. Here the number of clusters, 5, is chosen before we see the computational results, to make the comparison fair. Details of this work will be reported elsewhere.

Oliver Linton (*London School of Economics and Political Science*)

This is a comprehensive paper. I shall just focus on the new implementation of Ichimura's semiparametric least squares method for estimating index models. In expression (A.1) the authors sequentially minimize

$$\sum_{j=1}^{n} \sum_{i=1}^{n} \left\{ y_i - a_j - b_j \beta^{\mathrm{T}} (X_i - X_j) \right\}^2 w_{ij}$$

with respect to (a, b, β) holding w_{ij} constant and starting from some initial consistent estimator $\tilde{\beta}_0$. The Ichimura (1993) procedure involves sequential minimization with the difference that he uses only local constant but also includes the dependence of w_{ij} on β; this leads to a nasty non-linear optimization problem, whereas the authors' procedure is just bilinear least squares, and so is conditionally linear. They apparently prove that after two iterations their $\tilde{\beta}$ behaves as if (a, b) were known in expression (A.1). I think that this is an important idea that will make estimation of these models much easier. The authors develop many useful tools and apply them impressively. I have some comments and questions.

The initial consistent estimator that lurks in Appendix A.2 is either the average derivative estimator (in which case the criticisms in (a) and (b) of the second page apply) or some non-linear least squares estimator, which itself will be heavily computational.

I suppose that the authors' estimator achieves the semiparametric efficiency bound in for example the special case of Appendix A.2 with independent and identically distributed ε, but it is not so clear to me.

In time series, we come across special sorts of indices like $\Sigma_{k=0}^{\infty} \beta^k X_{t-k}$, where β is unknown; this would generalize the linear model $y_t = \beta y_{t-1} + \gamma X_t + \varepsilon_t$ that is widely used. Have the authors thought about this case?

I do not think that the optimal amount of smoothing for the function will always be the same as the optimal amount of smoothing for the parameter. Generally speaking it seems that in 'adaptive' cases the optimal bandwidth for the parameter and the function have the same magnitude, although not the same constant. See for example Carroll and Härdle (1989). In non-adaptive cases this is not usually so. In the partially linear model $y = \beta_x + g(z) + e$, Linton (1995) showed that the Robinson (1998) estimator $\hat{\beta}$ for β has expansion $\hat{\beta} - \beta = O_p(n^{-1/2}) + O_p(h^4) + O_p(n^{-1}h^{-1/2})$ under twice continuous differentiability of g, which suggests an optimal bandwidth rate of $h \propto n^{-1/9}$, i.e. it is optimal to undersmooth. Although maybe the authors can find an estimator of β that has the optimal bandwidth rate of $h \propto n^{-1/5}$.

Liqiang Ni (*University of Minnesota, St Paul*)

I applaud the authors for the promising refined minimum average variance estimation (RMAVE) algorithm and the intriguing idea of determining the dimension in a cross-validation approach. Many methods have been proposed to estimate directions in the effective dimension reduction space (Li, 1991), or the central subspace (Cook, 1996). Sliced inverse regression (SIR) can discover directions of linear terms in mean functions but fails in symmetric situations like $y = (\beta^{\mathrm{T}}X)^2 + \varepsilon$ with X normal, $E(X) = 0$ and $\varepsilon \perp\!\!\!\perp X$, where the direction can be detected by sliced average variance estimation (Cook and Weisberg, 1991). In my experience, RMAVE can estimate both linear and quadratic terms well.

Suppose that we have a continuous predictor $X \in R^p$ and a categorical predictor $C \in R$ representing different subpopulations. If the mean function of Y does have a form as

$$g\left\{\beta^{\mathrm{T}}\begin{pmatrix} X \\ C \end{pmatrix}\right\},$$

which may indicate shifts between subpopulations, RMAVE can be practically useful under the circumstances described by the authors. However, when $y = G_C(\beta_C^{\mathrm{T}}X) + \varepsilon$, so each subpopulation may have its own unique directions and functions, mixing continuous and categorical predictors may be inappropriate. Partial SIR (Chiaromonte *et al.*, 2002) directly addresses this issue. In the same spirit, we may consider 'partial RMAVE'. One way to do this may be simply to let the weight w_{ij} in expression (3.8) multiply an indicator function $I(C_i = C_j)$ and modify the cross-validation (CV) function as well. Details of this approach, which seems to work quite well, will be reported elsewhere.

The selection of the bandwidth seems tricky. The estimation of dimension is much more stable when CV adopts the Nadaraya–Watson estimator than when using a local linear estimator. Neverthless, it is still sensitive to the bandwidth. I applied RMAVE to the AIS data (Chiaromonte *et al.*, 2002) which consist of a mixture of two linear regressions determined by the only categorical predictor—gender. Considering only continuous predictors, the Nadaraya–Watson CV values suggested two dimensions with larger bandwidth and only one dimension with smaller bandwidth. The partial RMAVE method described as above, however, suggested one dimension consistently, which confirmed that both linear regressions associate with the same direction, $y = G_C(\beta^{\mathrm{T}}X) + \varepsilon$.

I have a question about inverse MAVE. The essence of SIR is that, under the linearity condition (1.2), the space spanned by $E(Z|Y)$ where $E(Z) = 0$ and $\mathrm{cov}(X) = I$ is a subset of the EDR space. To estimate this space, Li (1991) proposed slicing on Y, and Zhu and Fang (1996) proposed kernel methods. I am not sure whether inverse MAVE is intended to estimate $\mathrm{span}\{E(Z|Y)\}$ also.

Megu Ohtaki and Yasunori Fujikoshi (*Hiroshima University*)

We praise the authors of this paper, which has a highly original and fascinating content. The paper is sure to be one of the monumental works in the field of multivariate analysis.

In the paper it is clearly shown that the minimum average variance estimation (MAVE) method and its algorithm have many advantages over existing methods for searching an effective dimension reduction (EDR) space. Just like the sliced inverse regression method, however, no description for the reduction in the number of the original covariables was given. It is also important to consider selection of the original variables as well as the covariables $\beta_1^{\mathrm{T}}X, \ldots, \beta_p^{\mathrm{T}}X$. In practical situations of data analysis, a model with a small number of original covariables is preferable while the bias is negligible. This problem may be formulated mathematically as below.

Suppose, for example, in model (1.1)

$$y = g(B_0^{\mathrm{T}}X) + \varepsilon,$$

where B_0 and X are decomposed as

$$B_0 = \begin{pmatrix} B_{01} \\ B_{02} \end{pmatrix}_{p\times D}, \qquad X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}_{p\times 1},$$

and hence $B_0^{\mathrm{T}}X = B_{01}^{\mathrm{T}}X_1 + B_{02}^{\mathrm{T}}X_2$. If $B_{02} = O$, then it is expected by analogy (Akaike, 1973; Mallows, 1973) for cases of linear regression that we shall be able to have a more efficient EDR.

For not only such a mathematical background but also economical reasons, those covariables which have no effect on the response should not be used in regression analysis. Therefore, we propose the regression model

$$y = g(B^{\mathrm{T}}X) + \varepsilon,$$

where $Q \subset \{1, \ldots, p\}$ and $D_Q = \mathrm{diag}(q_1, \ldots, q_p)_{p \times p}$, with $q_i = 1$ if $i \in Q$ and $q_i = 0$ otherwise, for selecting the optimal model, and to choose a model attaining $\min_{d,Q}\{\mathrm{CV}(d, Q)\}$ that will be constructed by modifying the cross-validation criterion, $\mathrm{CV}(d)$, which is given in the paper. Thus the MAVE method may be extended easily to reduce the number of the original covariables as well as the dimension of an EDR space simultaneously. Furthermore, the MAVE method has the advantage that it may be generalized to multivariate regression.

In linear statistical inference, it has been reported that the model selection method using Akaike's information criteria AIC is not consistent for estimating the true model (see, for example, Shibata (1976) and Fujikoshi (1985)). Stone (1974) showed that the cross-validation criterion and AIC are asymptotically equivalent for model selection. Given these results, we wonder whether theorem 2 is consistent with the classical results.

James R. Schott (*University of Central Florida, Orlando*)

Over the past decade, there has been a considerable amount of work on dimensionality reduction techniques in the regression setting. This paper represents a substantial contribution to that area. I have just a couple of minor comments relating to the sliced inverse regression (SIR) procedure of Li (1991) and subsequent similar types of procedure such as the sliced average variance estimate of Cook and Weisberg (1991).

The linear condition given in equation (1.2) is a fundamental requirement for most of these procedures. Additional assumptions may be needed; for instance, sliced average variance estimation requires a constant variance assumption, and inferential methods, associated with these procedures, for determining the correct dimension often require stronger conditions. These additional assumptions are certainly restrictive, but it is important to note that equation (1.2) is a fairly mild condition. It is weaker than elliptical symmetry because it only has to hold for the directions B_0. Thus, we may not have elliptical symmetry but be sufficiently lucky still to have condition (1.2) hold. In fact, Hall and Li (1993) have shown that, loosely speaking, if the dimension of X is high, then it is likely that condition (1.2) holds at least approximately.

A further point to note is that procedures like SIR estimate a space that may be a proper subspace of the space spanned by the columns of B_0. Have we missed any important directions? If so, how do we recover them? These are questions that may need to be answered when using SIR. However, they are not relevant questions for the adaptive procedures proposed here since they directly estimate the space spanned by the columns of B_0.

C. M. Setodji (*University of Minnesota, St Paul*)

We have been presented with a constructive and useful paper and the authors are to be congratulated. Minimum average variance estimation (MAVE) seems to be an interesting and intriguing method for dimension reduction estimation. Equation (1.1) is applicable to any regression problem since, for any Y and X, we can always define $\varepsilon = Y - E(Y|X)$ which depends on X and satisfies the conditions in the paper. I have applied MAVE to three well-known sets of data that have been studied in the dimension reduction literature, and the optimal bandwidth was used throughout. Background on the examples was given by Cook and Critchley (2000). In all three examples, MAVE fails to produce the directions obtained by other methods.

First the methods proposed were applied to the bank-note data. With a binary response (the bank-note's authenticity) and six predictors, all the information in the regression is contained in the mean function. The refined MAVE method gave $\hat{d} = 21$, which is the same as the result produced by sliced average variance estimation (SAVE) (Cook and Critchley, 2000; Chiaromonte *et al*., 2002) and projection pursuit analysis (Posse, 1995). Whereas the first MAVE and SAVE directions are essentially the same, the second directions are quite different. The second SAVE direction shows two kinds of forged notes, but the role of the second MAVE direction is unclear. It misses the clustering in the counterfeit notes.

We also applied MAVE to the Hawkins data, designed to challenge traditional and robust regression methods with outliers. Although the data with four covariates and a continuous response have two directions in the mean function, refined MAVE and inverse MAVE suggest independence whereas the outer products of gradients method suggests only one direction. SAVE correctly identifies the regression structure. Lastly, the method was applied to the AIS data, a data set with mixtures. MAVE gave $\hat{d} = 1$, suggesting one direction, whereas sliced inverse regression infers $\hat{d} = 2$. MAVE evidently missed the 'joining information' for males and females.

Many regression problems are filled with 'mixtures' which is the one thing that all these data sets have in common. Mixtures increase the dimension of the mean function. My experience suggests that the MAVE

methods fail to detect mixture regressions. Is it possible to enhance the proposed method to face such an issue?

Finally, for me, one of the weaknesses of the method proposed is the fact that it is not invariant under linear transformations. Using (x_1, x_2) or $(x_1 + x_2, x_2)$ as predictors may yield different first directions when $d = 1$. More developments need to be pursued for these methods.

Nils Chr. Stenseth and Ole Chr. Lingjærde (*University of Oslo*)

Lynx populations undergo regular density cycles all across the boreal forest of Canada (see, for example, Stenseth *et al*. (1998)). In a previous analysis of the lynx dynamics (Stenseth *et al*., 1999) two competing hypotheses were put forward regarding the spatial structure of the dynamics. One predicts that the dynamical structure clusters into groups defined according to ecological-based features, whereas the other predicts that it clusters into groups according to climatic-based features. On the basis of an analysis of 21 time series from 1821 onwards, Stenseth *et al*. (1999) found evidence in support of the latter hypothesis, assuming a piecewise linear autoregressive model for each population. However, their model did not explicitly include any climatic effects.

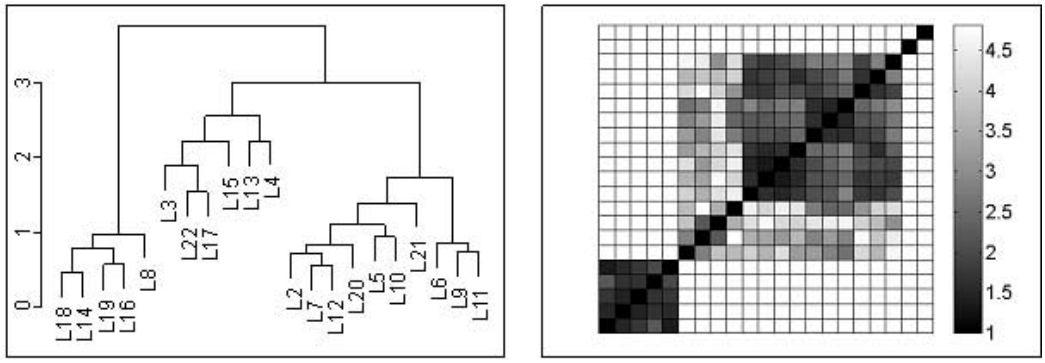
Here, we propose to use the authors' minimum average variance estimation (MAVE) methodology to study the spatial structure of the Canadian lynx populations, on the basis of a more general nonparametric model of the dynamics that includes as a covariate the potentially important climatic variable known as the North Atlantic oscillation winter index. Specifically, let L_t^s denote the natural logarithm of the abundance of lynx in region s in year t, and let NAO_t denote the North Atlantic oscillation winter index in year t. For each s and t define the response $y_t^s = L_t^s$ and the vector of covariates

$$\mathbf{X}_t^s = (L_{t-1}^s, L_{t-2}^s, L_{t-3}^s, L_{t-4}^s, \mathrm{NAO}_t, \mathrm{NAO}_{t-1}, \mathrm{NAO}_{t-2})^{\mathrm{T}}.$$

For each region s we assume the model

$$y_t^s = g_s(\mathbf{B}_{s,0}^{\mathrm{T}}\mathbf{X}_t^s) + \varepsilon_{s.t} = g_s(\boldsymbol{\beta}_{s,1}^{\mathrm{T}}\mathbf{X}_t^s, \ldots, \boldsymbol{\beta}_{s,d}^{\mathrm{T}}\mathbf{X}_t^s) + \varepsilon_{s.t}$$

where g_s is an unknown smooth link function, $\mathbf{B}_{s,0} = (\boldsymbol{\beta}_{s,1}, \boldsymbol{\beta}_{s,2}, \ldots, \boldsymbol{\beta}_{s,d}) \in \mathbb{R}^{7,d(s)}$ is an orthogonal matrix and $E(\varepsilon_{s,t}|\mathbf{X}_s^t) = 0$ almost surely. Using refined MAVE and cross-validation, we estimated $d(s)$ and $\mathbf{B}_{s,0}$ for each s. To compare the dynamics in two regions s and s' we considered the largest principal angle $\varphi(s, s')$ between the subspaces spanned by the columns of $\mathbf{B}_{s,0}$ and $\mathbf{B}_{s',0}$ respectively. This angle can be determined from the relationships $0 \leqslant \varphi(s, s') \leqslant \pi/2$ and $\sin\{\varphi(s, s')\} = ||\mathbf{B}_{s,0}\mathbf{B}_{s,0}^{\mathrm{T}} - \mathbf{B}_{s',0}\mathbf{B}_{s',0}^{\mathrm{T}}||_2$. See Fig. 9 for results when $d(s)$ is estimated by cross-validation. Note that rows and columns are permuted to obtain coherent blocks of similar dynamics.



(a) (b)

Fig. 9. Comparison of dynamic structures across Canada, using cross-validation estimates for the orders $d(s)$ (the comparison is based on the largest principal angles between the estimated reduction subspaces for each region): (a) average linkage hierarchical clustering of the 21 time series; (b) pseudocolour checker-board plot of distances (the plotted values are non-linearly scaled as $\exp\{\varphi(s, s')\}$ to accentuate the regions of similar dynamics; order of regions (from left to right) with two major clusters emphasized, **L18, L19, L16, L8, L14**, L3, L22, L17, L15, **L2, L7, L12, L20, L6, L9, L11, L5, L10, L21**, L13, L4)

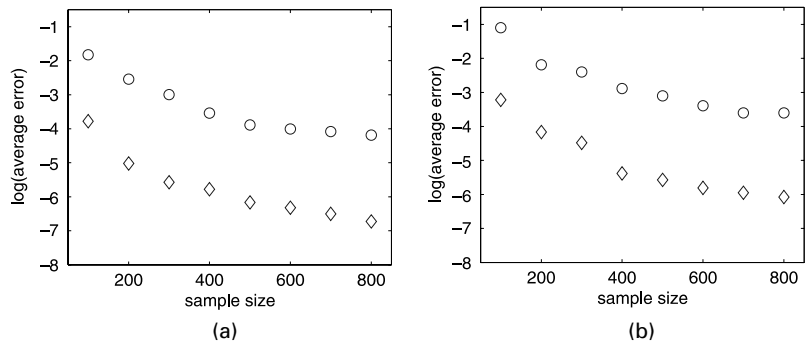


Fig. 10. (a) Parametric estimation and (b) nonparametric estimation (◇, results with uncorrelated design; ○, results for designs with functional relationships)

The results are strikingly similar to what we proposed as the ecological region structuring, and there is no strong support for the climatic region structuring, the latter of which was concluded to be the most appropriate region by Stenseth *et al.* (1999). To understand the underlying reasons for these differences certainly requires further work, both on the ecological and on the statistical side—work that we would like to pursue.

The **authors** replied later, in writing, as follows.

The extraordinarily kind words from so many distinguished discussants have overwhelmed us. We thank all the discussants for their constructive remarks and stimulating questions. Limitations of time and space prevent us from answering every question raised. Moreover, some of the suggestions will keep us busy for a while!

We thank Professor Kent for pointing out possible connections with other areas. His point regarding reduced rank models is clearly related to Chan and M. Li's important contribution. Turning to partial least squares, one of us has studied a nonparametric partial least squares regression after transformation. For data (y, X), a spline transformation $G(\cdot)$ of the response y is carried out so that the partial least squares regression can be modelled without knowing the exact form of $G(\cdot)$. Readers can refer to Zhu (2002) for more details. The basic idea is to 'linearize' a smooth function $G(\cdot)$ of the response y by $\pi(\cdot)^{\mathrm{T}}\theta$, where $\pi(\cdot)$ is a vector of B-spline basis functions of y and θ is an unknown projection parameter.

Concerning the issue of possible confounding between the covariates sulphur dioxide, nitrogen dioxide and the particulates (Bowman), the contribution by Professor Chan and Dr M. Li is relevant.

Concerning the challenging non-linear confounding problem mentioned by Professor K. C. Li, let us study the model used in Li (1997). Let $u_1 \sim \text{uniform}(0, 1)$, $u_2 = \log(u_1) + e$ with $e \sim \text{uniform}(-0.5, 0.5)$; $u_3, u_4, u_5 \sim^{\text{IID}} N(0, 1)$ and $x_1 = u_1 + u_3$, $x_2 = u_2 + u_4 + u_5$, $x_3 = u_3 - u_4$, $x_4 = u_4$ and $x_5 = u_5$. A relationship of y with $X = (x_1, \ldots, x_5)^{\mathrm{T}}$ via u_1 is

$$y = \log(u_1) + 0.1\varepsilon, \tag{1}$$

where $\varepsilon \sim^{\text{IID}} N(0, 1)$. The sample size $n = 100$. We estimate the directions by refined minimum average variance estimation (RMAVE) with $h = 0.05$. From 200 independent replications, the mean and the standard deviation of the estimated directions (we constrain the first component to be positive) are

$$\begin{array}{ccccc} (0.5662 & 0.0311 & -0.5660 & -0.5972 & -0.0316)^{\mathrm{T}}. \\ (0.0046) & (0.0107) & (0.0043) & (0.0067) & (0.0119) \end{array}$$

Because $u_1 = (1, 0, -1, -1, 0)^{\mathrm{T}} X$, the true direction is $(0.5774, 0, -0.5774, -0.5774, 0)^{\mathrm{T}}$. Our estimation results are quite encouraging especially since the structure of model (1) can hardly be detected by any of the other procedures. See for example Li (1997).

We agree with Professor Kent and Professor Bowman that the issue of collinearity is important. With a large set of near collinear covariates, some prescreening is recommended using such devices as principal components and others. Our limited simulations suggest that the MAVE method can still give some useful information when there is strong collinearity of functional relationships between covariates. Here we report the simulations for the model

Table 8. Means (and standard deviations) of the estimated EDR directions for model (5) with a sample size 200 and 100 replications†

β_1	(0.4538	0.4387	0.4467	0.4443	−0.0041	0.0013	0.0242	0.0067	0.0193	0.0128)$^{\mathrm{T}}$
Standard deviation	(0.0985	0.0988	0.1089	0.0969	0.0765	0.1025	0.1973	0.1900	0.1815	0.1961)$^{\mathrm{T}}$
β_2	(0.0106	−0.0033	0.0223	0.0127	0.0016	0.0071	0.3983	0.3833	0.3765	0.3428)$^{\mathrm{T}}$
Standard deviation	(0.2290	0.2300	0.2435	0.2520	0.1339	0.1511	0.2063	0.1963	0.1869	0.2223)$^{\mathrm{T}}$

†$h = 0.6$ was used.

$$y = (\beta_1^{\mathrm{T}} X)^2 (a + \beta_2^{\mathrm{T}} X) + \varepsilon, \tag{2}$$

where $a = 1$, $\beta_1 = (-\frac{2}{3}, \frac{1}{3}, \frac{2}{3}, 0)^{\mathrm{T}}$, $\beta_2 = (1, 0, 1, 0)^{\mathrm{T}}/\sqrt{2}$ and $X = (x_1, x_2, x_3, x_4)^{\mathrm{T}}$. Two cases are considered:

(a) an uncorrelated design, $x_1, x_2, x_3, x_4, \varepsilon \sim^{\mathrm{IID}} N(0, 1)$, and
(b) a design with functional relationships, $x_3 = (2x_1 + 2x_2 + \varepsilon_1)/3$, $x_4 = \{\mathrm{sgn}(x_1)|x_1|^2 + \varepsilon_2\}/2$ and $x_1, x_2, \varepsilon_1, \varepsilon_2, \varepsilon \sim^{\mathrm{IID}} N(0, 1)$.

We estimate model (2) under respectively the nonparametric setting and the non-linear parametric setting. With different sample sizes and bandwidths 0.6, 0.5, 0.45, 0.4, 0.35, 0.3, 0.28 and 0.25, results for the parametric estimators (obtained with the SAS software) and RMAVE estimators are shown in Fig. 10, where the error is defined as $m^2(\hat{\beta}_1, B_0) + m^2(\hat{\beta}_2, B_0)$ with $B_0 = (\beta_1, \beta_2)$. It is clear that both methods suffer from functional relationships between covariates. The relative degradation of efficiency for RMAVE due to collinearity and functional relationships between covariates is similar to that for the parametric case.

Our remark on the apparent robustness, based on our experience with MAVE, has somewhat to our surprise aroused substantial interest among the discussants (Critchley, Atkinson, Cui, G. Li, Yao, Čížek, Härdle, Yang and Welsh). The issue is important but we have as yet no theoretical results to offer.

We take Professor Cook's point about effective dimension reduction (EDR), a name which we adopted only after a suggestion from a referee. We also thank Professor Cook (and Professor Critchley) for clarifying the differences between the central subspace and the central mean subspace and their roles in the sliced inverse regression and the RMAVE methods. Professor Cook, Professor Critchley, Dr L. Li, Dr Schott and Dr Velilla raise concerns about heteroscedastic variance and wonder whether RMAVE can detect directions in the variance specification. If the conditional mean and the (not necessarily homogeneous) noise are additive, a two-step procedure may be adopted as follows. MAVE is first used to search the directions in the conditional mean and then applied to the squares of the residuals to look for the other directions. An alternative approach is as follows. Suppose that

$$y = g(B_0^{\mathrm{T}} X, \varepsilon). \tag{3}$$

Some of the EDR directions will be ignored if only the usual conditional mean is investigated. For any values δ and Δ, the data $(X_j + \delta, |y_j - \Delta|)$ are from the following model which has the same EDR space:

$$|y - \Delta| = E[|g^{\delta,\Delta}\{B_0^{\mathrm{T}}(X + \delta), \varepsilon\}|\,|B_0^{\mathrm{T}} X] + \eta^{\delta,\Delta} \tag{4}$$

with $g^{\delta,\Delta}$ denoting some measurable function, where

$$\eta^{\delta,\Delta} = |g^{\delta,\Delta}\{B_0^{\mathrm{T}}(X + \delta), \varepsilon\}| - E[|g^{\delta,\Delta}\{B_0^{\mathrm{T}}(X + \delta), \varepsilon\}|\,|B_0^{\mathrm{T}} X]$$

with $E(\eta^{\delta,\Delta}|X) = 0$. By choosing Δ appropriately, the conditional mean of model (4) can detect the other EDR directions. To avoid the difficulty of choosing Δ, we may use several of them together. For the following model, we consider three different pairs of (δ, Δ) and then we have four samples $\{(X_i, y_i)\}$, $\{(X_i + \delta_k, |y_i - \Delta_k|)\}$, $k = 1, 2, 3$. We re-denote them as $\{(X_{ki}, y_{ki})\}$, $k = 1, 2, 3, 4$. Using MAVE, for each sample we have from problem (2.7) a double summation to look for B:

$$S_k(B) = \sum_{j=1}^{n} \sum_{i=1}^{n} \{y_{ki} - b_{kj}^{\mathrm{T}} B^{\mathrm{T}}(X_{ki} - X_{kj})\}^2 w_{k,ij},$$

$k = 1, 2, 3, 4$. The common thing in these double summations is the direction matrix B. To find B, we can minimize $S_1(B) + S_2(B) + S_3(B) + S_4(B)$. We illustrate this approach with the model

Table 9. Means of estimation error $m^2\{(\beta_1, \beta_2), \hat{\beta}_1\}$, $m^2\{(\beta_1, \beta_2), \hat{\beta}_2\}$ for model (6) based on different algorithms†

Method	*Means of estimation errors for various bandwidths or spans*		
PPR (S-PLUS)	[0.4] (0.3459, 0.2876)	[0.5] (0.2997, 0.2613)	[0.6] (0.3707, 0.2776)
RMAVE (additive)	[0.2] (0.0415, 0.0355)	[0.3] (0.0088, 0.0170)	[0.4] (0.0214, 0.0212)
RMAVE (non-additive)	[0.3] (0.0305, 0.0516)	[0.4] (0.0481, 0.0731)	[0.5] (0.1104, 0.0586)

†Bandwidths or spans are given in square brackets.

$$y = \exp\{-2(\beta_1^{\mathrm{T}} X)^2\} + 0.5(\beta_2^{\mathrm{T}} X)\varepsilon, \tag{5}$$

where $X = (x_1, \ldots, x_{10})^{\mathrm{T}}$ with ε, x_j, $j = 1, \ldots, 10 \sim^{\mathrm{IID}} N(0, 1)$, $\beta_1 = (0.5, 0.5, 0.5, 0.5, 0, \ldots, 0)^{\mathrm{T}}$ and $\beta_2 = (0, \ldots, 0, 0.5, 0.5, 0.5, 0.5)^{\mathrm{T}}$. The simulation results are reported in Table 8. And Fig. 11 shows that by using the conditional expectation $E(|y - \Delta_k||X)$ we can capture all the EDR directions.

To answer questions concerning the minimization of problem (2.7) raised by the following discussants in this paragraph, we state some additional properties of RMAVE here. First, the estimation error for RMAVE is

$$m(\hat{B}, B_0) = O_p\left\{h_d^3 + \frac{\log(n)}{nh_d^{d+1}} + n^{-1/2}\right\},$$

provided that $d \geqslant D$. The estimation error depends only on d (and not on p). When d is small, root n consistency can be achieved (similar results were obtained by Hristache *et al.* (2002) from an approach that is analogous to the outer product of gradients method using refined weights). This answers the question of Professor K. C. Li and Professor Yao and gives an intuitive reason why our simulation works well. Secondly, the MAVE method can be applied easily to semiparametric models such as the model given in Professor Fan's comments. For all the single-index type of models that we have investigated (e.g. the single-index model and the generalized partially linear single-index model; see Xia *et al.* (2002)), the estimators are efficient in the semiparametric sense (Bickel *et al.*, 1993), and undersmoothing is unnecessary. This addresses Professor Linton's question.

We welcome the mention of projection pursuit regression (PPR) by Dr Cui, Dr G. Li, Professor K. C. Li and Dr Zhang, who have reiterated the differences between MAVE and PPR. Consider the PPR model

$$y = g_1(\beta_1^{\mathrm{T}} X) + \ldots + g_D(\beta_D^{\mathrm{T}} X) + \varepsilon, \tag{6}$$

where $E(\varepsilon|X) = 0$ and $(\beta_1^{\mathrm{T}}, \ldots, \beta_D^{\mathrm{T}})$ spans an EDR space. In the absence of extra conditions, we cannot ensure that the directions searched by PPR are in the EDR space. We compare the RMAVE algo-

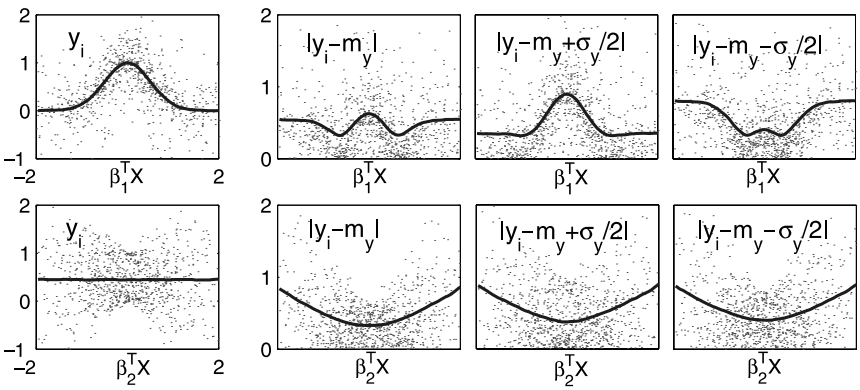


Fig. 11. 1000 observations from model (3) (·) and conditional expectations (———), based on kernel regression from 1 million observations

Table 10. Results of the CV methods

Data	*Method*†	*Results for the following dimensions:*						
		0	*1*	*2*	*3*	*4*	*5*	*6*
Bank-note	Bandwidth	—	0.1	0.3	0.5	0.6	0.7	0.7
	LL–CV value	0.2525	**0.0016**	0.0029	0.0045	0.0061	0.0093	0.0153
	NW–CV value	0.2525	**0.0036**	0.0049	0.0047	0.0079	0.0078	0.0085
AIS	Bandwidth	—	0.6	0.7	0.9			
	LL–CV value	150.5675	13.7718	**12.4450**	12.9045			
	NW–CV value	150.5675	20.2026	**19.8053**	27.1200			
Hawkins	Bandwidth	—	0.28	0.26	0.28	1		
	LL–CV value	9.2133	**7.8666**	8.8623	10.4843	18.5332		
	NW–CV value	9.2133	**7.6566**	9.0900	11.1208	11.6386		

†LL, local linear; NW, Nadaraya–Watson.

rithm with the PPR program in S-PLUS by reference to the distances from the EDR space based on the estimated directions. In our simulations, we take $D = 2$, $g_1(v) = \exp(-2v^2)$, $g_2(v) = -\cos(2v)$, $X = (x_1, x_2, \ldots, x_{15})^{\mathrm{T}}$, $\beta_1 = (1, 2, 3, 4, 5, 6, 7, 8, 7, 6, 5, 4, 3, 2, 1)^{\mathrm{T}}/\sqrt{344}$, $\beta_2 = (-7, -6, -5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5, 6, 7)^{\mathrm{T}}/\sqrt{280}$ and $x_1, \ldots, x_{15}, 2\varepsilon \sim^{\mathrm{IID}} N(0, 1)$. With a sample size of 200 and 200 independent replications, the estimated errors are listed in Table 9. The PPR algorithm in S-PLUS performs much worse than the MAVE algorithm; even without the benefit of the additive noise structure, the RMAVE method still outperforms the PPR algorithm in S-PLUS.

We refer Professor Ohtaki and Professor Fujikoshi to Cheng and Tong (1992), which establishes consistency of the cross-validation (CV) estimate, and to Professor Ferré's contribution.

We now consider Professor Setodji's examples. Because of the estimation of the remainder term, we have fewer problems to face than undersmoothing. It allows us to use the optimal bandwidth chosen by data-driven methods. For example, the CV method for the local linear smoothing of y_i on $X_i^{\mathrm{T}}\hat{B}$ can be applied to step 1(b) of our algorithm to choose the bandwidth that is used for the next iteration of estimation. Using this kind of bandwidth, we have re-examined the data sets cited by Professor Setodji. As usual we standardize each covariate before applying the RMAVE method. Table 10 shows our results with the smallest CV values highlighted in bold.

For the bank-note data, the dimension is estimated by CV to be 1 (instead of Setodji's 2). The corresponding direction is estimated as $\beta_1 = (-0.0521, 0.1438, -0.2036, 0.8103, 0.2242, -0.4779)^{\mathrm{T}}$. On the basis of this direction, we further have the following fit, which turns out to be practically deterministic:

$$y_i = f(\beta_1^{\mathrm{T}} X_i), \qquad \text{where } f(v) = 1 \text{ if } v \leqslant -0.2 \text{ or } f(v) = 0 \text{ otherwise.}$$

See also Fig. 12(a). With this simple deterministic single-index relationship, it seems difficult to believe that the efficient dimension is 2 as suggested by the sliced average variance estimation (SAVE) method in Cook and Critchley (2000). One possible explanation for suggesting a second dimension is that, if we classify $\{\beta_1^{\mathrm{T}} X_i, i = 1, \ldots, n\}$ into two groups then one of the notes might be in the wrong group on the basis of the SIR (or SAVE) direction as shown in Fig. 12(c). However, on the basis of the RMAVE direction above there is no such 'outlier'. See Fig. 12(b). For the AIS data, the CV estimated dimension is 2, which is the same as that suggested by SAVE. The results are shown in Figs 12(d) and 12(e). It seems to us that RMAVE has not missed any information. For the Hawkins data, the dimension is estimated to be 1. The model seems to give a reasonable fit to the data although the estimated dimension is lower than 2; see Fig. 12(f). Since the data set was generated from two regression models, we have also explored RMAVE with dimension 2 (and bandwidth 0.2). The directions are estimated as $\beta_1 = (0.0326, 0.7432, -0.2440, 0.6221)^{\mathrm{T}}$ and $\beta_2 = (0.7139, -0.1634, 0.5653, 0.3796)^{\mathrm{T}}$. The difference between these directions and the directions β_{01} and β_{02} that are estimated on the basis of the two regressions above is very small. See also Figs 12(g)–12(j). Fig. 12(g) can distinguish the observations by their models. The rotation in Figs 12(g) and 12(h) is useful for interpretation purposes and is related to questions about the effect of rotation raised by Professor Bowman, Professor Atkinson, Professor Chan and Dr M. Li, and Professor Yao.

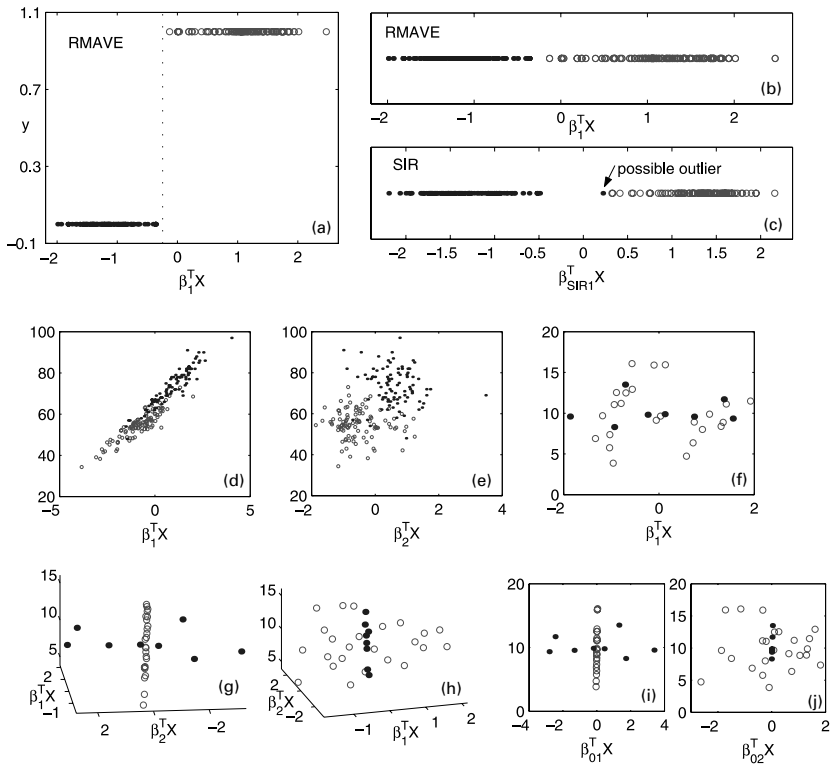


Fig. 12. Calculations for (a), (b), (c) the bank-note data (○, y = '1'; •, y = '0'), (d), (e) the AIS data (○, females; •, males) and (f), (g), (h), (i), (j) the Hawkins data (○, primary regression; •, second regression)

Professor Stenseth and Dr Lingjærde's application of the RMAVE method to the Canadian lynx populations is clearly very interesting. We also look forward to using the partial RMAVE method suggested by Professor Ni.

Concerning Professor Spokoiny's question, a further improvement on MAVE can be made. For example, we can improve the stability of the algorithm along the lines suggested by him.

References in the discussion

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In *Proc. 2nd Int. Symp. Information Theory* (eds B. N. Petrov and F. Csáki), pp. 267–281. Budapest: Akadémiai Kiadó.

Atkinson, A. C. and Riani, M. (2000) *Robust Diagnostic Regression Analysis*. New York: Springer.

Basilevsky, A. (1994) *Statistical Factor Analysis and Related Methods*. New York: Wiley.

Bickel, P. J., Klaassen, A. J., Ritov, Y. and Wellner, J. A. (1993) *Efficient and Adaptive Inference in Semiparametric Models*. Baltimore: Johns Hopkins University Press.

Brillinger, D. R. (1983) A generalized linear model with "Gaussian" regressor variables. In *A Festschrift for Erich L. Lehmann* (eds P. J. Bickel, K. A. Doksum and J. L. Hodges, Jr), pp. 97–114. Belmont: Wadsworth.

———(1992) Nerve cell spike train data analysis: a progression of technique. *J. Am. Statist. Ass.*, **87**, 260–271.

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

Bura, E. and Cook, R. D. (2001a) Extending sliced inverse regression: the weighted chi-squared test. *J. Am. Statist. Ass.*, **96**, 990–1003.

———(2001b) Estimating the structural dimension of regressions via parametric inverse regression. *J. R. Statist. Soc.* B, **63**, 393–410.

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

Carroll, R. J. and Härdle, W. (1989) Second order effects in semiparametric weighted least squares regression. *Statistics*, **2**, 179–186.
Cheng, B. and Tong, H. (1992) On consistent nonparametric order determination and chaos (with discussion). *J. R. Statist. Soc.* B, **54**, 427–449, 451–474.
Chiaromonte, F., Cook, R. D. and Li, B. (2002) Sufficient dimension reduction in regressions with categorical predictors. *Ann. Statist.*, **30**, in the press.
Cook, R. D. (1996a) Graphics for regressions with a binary response. *J. Am. Statist. Ass.*, **91**, 983–992.
———(1996b) *Regression Graphics: Ideas for Studying Regressions through Graphics*. New York: Wiley.
———(1998) *Regression Graphics*. New York: Wiley.
Cook, R. D. and Critchley, F. (2000) Identifying regression outliers and mixtures graphically. *J. Am. Statist. Ass.*, **95**, 781–794.
Cook, R. D. and Li, B. (2002) Dimension reduction for conditional mean in regression. *Ann. Statist.*, **30**, in the press.
Cook, R. D. and Nachtsheim, C. J. (1994) Re-weighting to achieve elliptically contoured covariates in regression. *J. Am. Statist. Ass.*, **89**, 592–600.
Cook, R. D. and Weisberg, S. (1991) Discussion on 'Sliced inverse regression' (by K. C. Li). *J. Am. Statist. Ass.*, **86**, 316–342.
———(1994) *An Introduction to Regression Graphics*. New York: Wiley.
Cook, R. D. and Yin, X. (2001) Dimension reduction and visualization in discriminant analysis (with discussion). *Aust. New Z. J. Statist.*, **43**, 147–200.
Cui, H., He, X. and Liu, L. (2001) Testing additivity with regression splines. Submitted to *Ann. Statist*.
Dauxois, J., Ferré, L. and Yao, A. F. (2001) Un modèle semi-paramétrique pour variables hilbertiennes. *C. R. Acad. Sci.*, **333**, 947–952.
Duan, N. and Li, K. C. (1991) A bias bound for least squares linear regression. *Statist. Sin.*, **1**, 127–136.
Fan, J. and Gijbels, I. (1996) *Local Polynomical Modelling and Its Applications*. London: Chapman and Hall.
Fan, J., Yao, Q. and Cai, Z. (2001) Adaptive varying-coefficient linear models. Submitted to *J. R. Statist. Soc.* B.
Ferré L. (1997) Dimension choice for Sliced Inverse Regression based on ranks. *Student*, **2**, 95–108.
———(1998) Determination of the dimension in SIR and related methods. *J. Am. Statist. Ass.*, **93**, 132–140.
Fujikoshi, Y. (1985) Selection of variables in two-group discriminant analysis by error rate and Akaike's information criteria. *J. Multiv. Anal.*, **17**, 27–37.
Hall, P. and Li, K. C. (1993) On almost linearity of low dimensional projections from high dimensional data. *Ann. Statist.*, **21**, 867–889.
Härdle, W., Hall, P. and Ichimura, H. (1993) Optimal smoothing in single-index models. *Ann. Statist.*, **21**, 157–178.
Härdle, W. and Stoker, T. M. (1989) Investigating smooth multiple regression by method of average derivatives. *J. Am. Statist. Ass.*, **84**, 986–995.
Hristache, M., Juditsky, A., Polzehl, J. and Spokoiny, V. (2001) Direct estimation of the index coefficients in a single-index model. *Ann. Statist.*, **29**, 1537–1566.
———(2002) Structure adaptive approach for dimension reduction. *Ann. Statist.*, to be published.
Hristache, M., Juditsky, A. and Spokoiny, V. (2001) Direct estimation of the index coefficients in a single-index model. *Ann. Statist.*, **29**, 595–623.
Huber, P. J. (1985) Projection pursuit (with discussion). *Ann. Statist.*, **13**, 435–525.
Ichimura, H. (1993) Semiparametric least squares (SLS) and weighted SLS estimation of single index models. *J. Econometr.*, **58**, 71–120.
Li, G. and Cheng, P. (1993) Some recent developments in projection pursuit in China. *Statist. Sin.*, **3**, 35–51.
Li, K. C. (1991) Sliced inverse regression for dimension reduction (with discussion). *J. Am. Statist. Ass.*, **86**, 316–342.
———(1997) Nonlinear confounding in high dimensional regression. *Ann. Statist.*, **57**, 577–612.
Li, M.-C. (2000) Multivariate nonlinear time series modeling. *PhD Thesis*. University of Iowa, Iowa City.
Li, M.-C. and Chan, K. S. (2001) Semiparametric reduced-rank regression. *Technical Report 310*. Department of Statistics and Actuarial Science, University of Iowa, Iowa City. (Available from `http://www.stat.uiowa.edu/techrep/.`)
Li, K. C. and Duan, N. (1989) Regression analysis under link violation. *Ann. Statist.*, **17**, 1009–1052.
Linton, O. B. (1995) Second order approximation in the partially linear regression model. *Econometrica*, **63**, 1079–1112.
Mallows, C. L. (1973) Some comments on C_p. *Technometrics*, **15**, 661–671.
Murphy, S. A. and van der Vaart, A. W. (2000) On profile likelihood (with discussion). *J. Am. Statist. Ass.*, **95**, 449–485.
Posse, C. (1995) Projection pursuit exploratory data analysis. *Comput. Statist. Data Anal.*, **20**, 669–687.
Riani, M. and Atkinson, A. C. (2001) A unifed approach to outliers, influence, and transformations in discriminant analysis. *J. Comput. Graph Statist.*, **10**, 513–544.
Robinson, P. M. (1988) Root-N-consistent semiparametric regression. *Econometrica*, **56**, 931–954.
Ruckstuhl, A. F. and Welsh, A. H. (1999) Reference band for nonparametrically estimated link function. *J. Comput. Graph. Statist.*, **8**, 699–714.

Shannon, C. E. (1948) A mathematical theory of communication. *Bell Syst. Tech. J.*, **27**, 379–423, 623–656.
Shibata, R. (1976) Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika*, **63**, 117–126.
Stenseth, N. C., Chan, K.-S., Tong, H., Boonstra, R., Boutin, S., Krebs, C. J., Post, E., O'Donoghue, M., Yoccoz, N. G., Forchhammer, M. C. and Hurrell, J. W. (1999) Common dynamic structure of Canada lynx populations within three climatic regions. *Science*, **285**, 1017–1073.
Stenseth, N. C., Falck, W., Chan, K.-S., Bjørnstad, O. N., O'Donoghue, M., Tong, H., Boonstra, R., Boutin, S., Krebs, C. J. and Yoccoz, N. G. (1998) From patterns to processes: phases and density dependencies in the Canadian lynx cycle. *Proc. Natn. Acad. Sci. USA*, **95**, 15430–15435.
Stone, M. (1974) Cross-validatory choice and assessment of statistical predictions (with discussion). *J. R. Statist. Soc.* B, **36**, 111–147.
Stone, M. and 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**, 237–269.
Velilla, S. (1998) Assessing the number of linear components in a general regression problem. *J. Am. Statist. Ass.*, **93**, 1008–1098.
Weisberg, S. and Welsh, A. H. (1994) Adapting for the missing link. *Ann. Statist.*, **22**, 1674–1700.
Xia, Y. and Li, W. K. (1999) On single-index coefficient regression models. *J. Am. Statist. Ass.*, **94**, 1275–1285.
Xia, Y., Tong, H. and Li, W. K. (2002) Single index volatility model and its estimation. *Statist. Sin.*, to be published.
Zhu, L. X. (2002) Transforming a response variable for partial least squares regression. *Technical Report*. Department of Statistics and Actuarial Science, Univesity of Hong Kong, Hong Kong.
Zhu, L. X. and Fang, K.-T. (1996) Asymptotics for kernel estimate of sliced inverse regression. *Ann. Statist.*, **24**, 1053–1068.