Statistica Sinica **5**(1995), 667-688

BINARY REGRESSORS IN DIMENSION REDUCTION MODELS: A NEW LOOK AT TREATMENT COMPARISONS

R. J. Carroll and Ker-Chau Li

Texas A&M University and University of California at Los Angeles

Abstract. In this paper, new aspects of treatment comparison are brought out via the dimension reduction model of Li (1991) for general regression settings. Denoting the treatment indicator by Z and the covariate by X, the model $Y = g(v'X + \theta Z, \epsilon)$ is discussed in detail. Estimates of v and θ are obtained without assuming a functional form for g. Our method is based on the use of SIR (sliced inverse regression) for reducing the dimensionality of the covariate, followed by a partial-inverse mean matching method for estimating the treatment effect θ. Asymptotic theory and a simulation study are presented.

Key words and phrases: Conditioning, dimension reduction, linear design condition, nonparametric curve fitting, randomization, SIR, treatment effect.

1. Introduction

There is a growing interest in high dimensional data analysis. One approach in this area, taken by Li (1991), is to consider a dimension reduction model of the following form:

$$Y = g(\beta_1'\mathbf{x}, \ldots, \beta_q'\mathbf{x}, \epsilon), \tag{1.1}$$

where ϵ is independent of $\mathbf{x}$. The function g and the distribution of the random error ϵ are both unknown. The vectors $\beta_i, i = 1, \ldots, q$, span a q dimensional subspace in R^p, called the *effective dimension reduction* (e.d.r.) space. An e.d.r. direction refers to any vector in the e.d.r. space.

Sliced inverse regression (SIR) offers a promising methodology for estimating the e.d.r. directions. This procedure reverses the more natural forward regression methods which model Y as a function of $\mathbf{x}$. Instead of such direct data fitting, SIR exploits the conditional distribution of $\mathbf{x}$ given Y. The first moment method, based on $E(\mathbf{x}|Y)$, has been studied extensively in various contexts; see Carroll and Li (1992), Duan and Li (1991), Hsing and Carroll (1992), and Li (1991, 1992a). The second moment, $\text{cov}(\mathbf{x}|Y)$, is also useful; see Cook and Weisburg (1991), and Li (1991, 1992b).

Although model (1.1) is primarily proposed for handling continuous regressors, it is interesting to study how discrete variables can be incorporated in the model. In this paper, a single dichotomous variable Z will be considered. Denote the rest of the $p-1$ continuous regressors as X. While simple, this represents a natural situation for treatment comparison, as well as a first step towards more complex models. We call $Z=0$ the control group and $Z=1$ the treatment group.

We may rewrite model (1.1) as

$$y = g(B'X + Z\Theta, \epsilon), \tag{1.2}$$

where B is a $(p-1)$ by q matrix and Θ is a q-vector; both are unknown. Our model offers a variety of ways for comparing treatments, which have not been considered in the literature. To simplify the discussion, only the case $q=1$ will be studied in detail in this paper.

In Section 2, we compare our model with several others when X is only one-dimensional. This clarifies the connection of our approach to traditional parametric modeling as well as the more recent semiparametric methods for comparing nonparametric regression curves. Our model is flexible enough to deal with a variety of situations, ranging from clinical trials to industrial/engineering quality improvement settings.

Section 3 points out the theoretical difficulties in justifying direct application of the SIR technique. The discreteness of Z causes two related problems: (1) the linear design condition, namely Condition 3.1 of Li (1991), is violated; (2) the e.d.r space may not be identifiable without imposing proper constraints on the parameter values.

Section 4 presents our method for estimating the treatment effect Θ for one dimensional X. Since Θ is only a scalar now, it will be denoted by the lower case θ. Our estimate depends on a weight function. Of special interest is one which is related to double-slicing, an idea mentioned in the rejoinder of Li (1991). We refer to this method as *partial inverse mean matching.* It is based on the mean of Z conditional on Y and $X+\theta Z$. The estimate is obtained by finding a value of θ so that this conditional mean matches the mean of Z conditional on $X+\theta Z$ only. By contrast, other methods for comparing treatments are based on the conditional mean of Y given Z and X. We establish root n consistency and find the asymptotic distribution of our method. We also conduct a simulation study for illustration.

The multivariate covariate case is considered in Section 5. The general strategy is to reduce the dimension of X first. If the e.d.r. space has only one dimension, then we can apply the techniques in Section 4 to the reduced variable. We discuss how to apply the SIR methodology to find the correct projection direction for reduction. A simulation is conducted to illustrate our two-stage strategy. We do not discuss the case $q>1$, leaving this to future research.

A concluding note is given in Section 6. Technical proofs are given in Appendices.

2. Models for Trearment Comparison: Single Covariate

In this section, assume that X is only one-dimensional. The parameter Θ becomes a scalar now and for clarity, we denote it by the lower case θ. Since we have in total $p = 2$ regressors, one continuous X and one discrete Z, the e.d.r. space can have $q = 0, 1$, or 2 dimensions. The case $q = 0$ is trivial as Y will be independent of X and Z. The case $q = 2$ is also trivial because no dimension reduction is incurred; arbitrary joint distributions between X and Y are allowed for the treatment and the control groups and no relationship between them is drawn.

The most interesting case $q = 1$ is now discussed. It is worthwhile to further distinguish between the following two cases:

$$Y = g(Z, \epsilon), \tag{2.1}$$

and

$$Y = g(X + \theta Z, \epsilon), \tag{2.2}$$

where ϵ is independent of (X, Z). The first case, simplified from (1.2) with $B = 0$, represents the situation that the covariate X has no effect on Y in either the treatment or the control group.

We now focus on (2.2), which, because X is scalar, is equivalent to (1.2) with $B \neq 0$ by absorbing B into the unknown function g. The parameter θ in (2.2) can be interpreted as the treatment effect. The response Y is related to X through $Y = g(X, \epsilon)$ for the control group, and $Y = g(X + \theta, \epsilon)$ for the treatment group. Thus, a case from the treatment group behaves as if an additional amount θ had been added into its covariate X. Since we do not make any assumptions about (g, ϵ), there is no restriction on the joint distribution of Y and X.

Model (2.2) covers several special models for treatment comparison:

(1) Linear regression:

$$Y = a + bX + cZ + \epsilon. \tag{2.3}$$

For this case, parallel lines are anticipated in the plot of Y on X. The parameter c can be interpreted as the amount of vertical shift needed for the regression line with $Z = 0$ to match that with $Z = 1$. The same matching can also be achieved by shifting the treatment-group regression line horizontally by the amount of $\theta = c/b$.

(2) Nonlinear regression: There are two ways of generalizing (2.3) to nonlinear regression, namely vertical shift

$$Y = g(X) + cZ + \epsilon, \tag{2.4}$$

and horizontal shift

$$Y = g(X + \theta Z) + \epsilon. \tag{2.5}$$

The former appears in the literature concerning partly linear models or partial splines; see Chen (1988), Engle, Granger, Rice, and Weiss (1986), Heckman (1986), Rice (1986), Speckman (1988). The latter can be found in Härdle and Marron (1990), and Kneip and Gasser (1992).

Note that the vertical shift model (2.4) is not covered under (2.2). To fit it into our framework, we need to take $q = 2$ in (1.2).

(3) Multiplicative error: A simple multiplicative model is

$$Y = \mu + g(X + \theta Z)\epsilon. \tag{2.6}$$

Here the mean of Y is a constant independent of the covariate, but the standard deviation depends on (X, Z). This model is useful in dealing with problems associated with Taguchi's method for quality improvement. For example, increasing the stability of certain quality aspects of a product, measured by Y, may be the goal. Suppose the variance of Y depends on the level of an undesirable environmental factor X, an identifiable key source of noise which cannot be adjusted. Then model (2.6) can be used to assess how effective a treatment is in stabilizing the quality of the product. If g is an increasing nonnegative function, then a negative θ value shows that the treatment under study has an effect amounting to reducing the level of the uncontrollable environmental factor by the amount $|\theta|$.

Model (2.6) is a case where all the methods referred to in the discussion following (2.5) are not applicable, because in (2.6) treatment has no effect on the mean function.

(4) Generalized linear models:

$$Y \sim P_\tau, \ \tau = a + bX + cZ,$$

where the distribution P_τ belongs to an exponential family indexed by τ. Such models are often assumed when the response variable Y is discrete, e.g., the binomial or Poisson distributions. In such cases, both the mean and the variance of Y contain information about the treatment effect. Discrete output is quite popular in many applications, including clinical trials and quality engineering settings.

Remark 2.1. The popular single index model from the econometric literature is closely related to the dimension reduction framework (1.1) (with $q = 1$). Han (1987) and Sherman (1993) describe one such approach using a rank correlation procedure. However, the constructed index is required to have a *monotone* functional relationship with the response variable. This precludes cases like the multiplicative error in (2.6). In addition, the implementation of the procedure may not be easy to carry out although one Referee has informed us of an unpublished preprint by Chris Cavanagh and Robert Sherman of BELLCORE that discusses computation. Another related approach is projection pursuit regression, the theory of which as developed in Hall (1989), Chen (1991) and Härdle, Hall, and Ichimura (1993) does not directly apply to discrete regressors, nor does it apply to (2.6). A third approach is derivative-based methods; see Härdle and Stoker (1989) and Samarov (1993). Whether these methods are appropriate for treatment comparison or not is unclear. One difficulty is that we cannot take a partial derivative with respect to a discrete regressor. Another dimension reduction method not discussed here is the method of principal Hessian directions (Li (1992b)), which is related to the second derivative based method in Samarov (1993).

3. Difficulties

Since our model is a special case of SIR, why can't we apply SIR directly? There are two major reasons.

The first difficulty concerns the conditional linearity assumption, Condition 3.1 of Li (1991), needed for establishing the unbiasedness of SIR.

Linear Design Condition 3.1. For any b in R^p, the conditional expectation $E(b'\mathbf{x}|\beta_1'\mathbf{x},\ldots,\beta_q'\mathbf{x})$ is linear in $\beta_1'\mathbf{x},\ldots,\beta_q'\mathbf{x}$.

Let

$$T_\theta = X + \theta Z. \tag{3.1}$$

In the framework of (2.2), the Linear Design Condition 3.1 is reduced to the following:

$$E(Z|T_\theta = t) = a + bt, \text{ for some } a, b. \tag{3.2}$$

Because of the binary nature of Z, this identity rarely holds. To see this, note that

$$\begin{aligned} E(Z|T_\theta = t) &= P\{Z = 1|X + \theta Z = t\} \\ &= \frac{f_{X|Z=1}(t-\theta)P\{Z=1\}}{f_{X|Z=0}(t)P\{Z=0\} + f_{X|Z=1}(t-\theta)P\{Z=1\}}, \end{aligned}$$

where all f with subscripts denote conditional densities. If (3.2) holds, then the density ratio, $f_{X|Z=1}(t-\theta)/f_{X|Z=0}(t)$, must take the form of a special rational function, more precisely, the ratio of two linear functions.

The second issue concerns the identifiability of the e.d.r. space. For the general model (1.1), there is always more than one way of setting up the function g to represent a given joint distribution of $(Y, \mathbf{x})$. A less ambiguous statement, equivalent to (1.1), is that

$$\text{conditional on } \beta_i'\mathbf{x} \ (i = 1, \ldots, q), \ Y \text{ and } \mathbf{x} \text{ are independent.} \tag{3.3}$$

Similar to the definition of sufficient statistics, the notion of the e.d.r. space is to remove as many redundant dimensions as possible. Thus, although we can insist that any space containing an e.d.r. space could also be an e.d.r. space, the most useful e.d.r. space is surely the one with the smallest dimension. But an identifiability question arises immediately: is such a space unique? Cook (1994) shows that this is the case if the support of the distribution of $\mathbf{x}$ is equal to the entire R^p. In fact, as long as the support is connected and is contained in the closure of its interior, the minimum e.d.r space is unique. On the other hand, discreteness in a regressor is most likely to create an identifiability problem. For example, if the support of $\mathbf{x}$ has only a finite number of points, then any one-dimensional space spanned by any direction β can be an e.d.r. space if the projection $\beta'\mathbf{x}$ takes distinct values for distinct points in the support of $\mathbf{x}$.

Once again returning to our case with one continuous and one binary regressor, the support of the distribution of $\mathbf{x} = (X, Z)'$ is contained in two horizontal lines. Thus in the 2-D plane topology, the support of $\mathbf{x}$ has an empty interior. To illustrate what may go wrong, consider the case that X is restricted to the interval $[0, 1]$. Then for any $\theta > 1$, the support of $X + \theta Z$ will consist of two non-overlapping intervals, $[0, 1] \cup [\theta, \theta + 1]$. We are free to define the function $g(\cdot)$ in (2.2) for each segment separately, and the model would not bring any connection between the joint distribution of (Y, X) for the treatment group and that for the control group. Therefore, although model (2.2) appears as if the e.d.r. space has only one dimension, it already is the most general case and no dimension reduction is incurred.

The key for resolving this issue is to observe that if $g(\cdot)$ in (2.2) is periodic in the first argument with (smallest) period c, then we cannot distinguish θ from $\theta + c, \theta + 2c$, etc. To avoid any ambiguity, we may take θ to be the one with the smallest absolute value: $\theta \in [-c/2, c/2]$. For the case that the support of X is the interval $[0, 1]$, θ will then be restricted to the interval $[-1, 1]$, because we can always extend the g function periodically with period 2.

4. Estimation of the Treatment Effect

As made clear in the previous section, the current literature for dealing with model (2.5) cannot handle generalizations such as the multiplicative error model

(2.6). A different method, based on (3.3), will be presented here which applies to both (2.5) and (2.6).

For any real number θ, denote $T_\theta = X + \theta Z$, see (3.1), and $f_{YZT}(y,z,t;\theta) =$ joint density of Y, Z, T_θ; $f_{YT}(y,t;\theta) =$ joint density of Y, T_θ; $f_{ZT}(z,t;\theta) =$ joint density of Z, T_θ; $f_T(t;\theta) =$ density of T_θ. Also denote by θ_o the true value of θ. In our context, (3.3) is equivalent to

$$f_{YZT}(y,z,t;\theta_o)f_T(t;\theta_o) = f_{YT}(y,t;\theta_o)f_{ZT}(z,t;\theta_o). \tag{4.1}$$

Let $w(y,z,t)$ be any positive function. A broad class of estimates can be obtained based on implementing the following minimization problem:

$$\min_\theta \sum_z \int\!\!\int w(y,z,t)A^2(y,z,t;\theta)dydt; \text{ where} \tag{4.2}$$

$$A(y,z,t;\theta) = f_{YZT}(y,z,t;\theta)f_T(t;\theta) - f_{YT}(y,t;\theta)f_{ZT}(z,t;\theta). \tag{4.3}$$

The integration domain for (4.2) and elsewhere in this section is without restriction. A solution for (4.2) is of course $\theta = \theta_o$, and the minimum value is 0. Subject to constraints, if necessary, for dealing with bounded support of X and periodicity as discussed in Section 3, we now assume that the solution is unique. Taking the derivative with respect to θ, note that $\theta = \theta_o$ solves the equation

$$\sum_z \int\!\!\int w(y,z,t)A(y,z,t;\theta)B(y,z,t;\theta)dydt = 0;$$

$$B(y,z,t;\theta) = \frac{\partial}{\partial\theta}A(y,z,t;\theta).$$

To implement (4.2), we need to estimate the density functions. For simplicity, kernel density estimates will be used.

First we treat the case that Y is discrete. Then (4.2) becomes

$$\min_\theta \sum_{y,z} \int w(y,z,t)A(y,z,t;\theta)^2 dt. \tag{4.4}$$

Suppose $(y_i, x_i, z_i), i = 1, \ldots, n$, are the sample values. Denote $t_i(\theta) = x_i + \theta z_i$. Then consider the kernel estimates:

$$\widehat{f}_{YZT}(y,z,t;\theta) = (nh)^{-1}\sum_{i=1}^n I(y_i = y)I(z_i = z)K\left\{\frac{t_i(\theta) - t}{h}\right\}; \tag{4.5}$$

$$\widehat{f}_{YT}(y,t;\theta) = (nh)^{-1}\sum_{i=1}^n I(y_i = y)K\left\{\frac{t_i(\theta) - t}{h}\right\}; \tag{4.6}$$

$$\widehat{f}_{ZT}(z,t;\theta) = (nh)^{-1}\sum_{i=1}^n I(z_i = z)K\left\{\frac{t_i(\theta) - t}{h}\right\};$$

$$\widehat{f}_T(t;\theta) = (nh)^{-1}\sum_{i=1}^n K\left\{\frac{t_i(\theta) - t}{h}\right\},$$

where $I(\cdot)$ is the indicator function, $K(\cdot)$ is a suitable kernel function with $\int uK(u)du = 0$, $\int K(u)du = 1$, and h is the bandwidth.

Our estimate $\widehat{\theta}$ is given by the solution of

$$\min_{\theta} \sum_{y,z} \int w(y,z,t)\widehat{A}^2(y,z,t;\theta)dt; \text{ where} \tag{4.7}$$
$$\widehat{A}(y,z,t;\theta) = \widehat{f}_{YZT}(y,z,t;\theta)\widehat{f}_T(t;\theta) - \widehat{f}_{YT}(y,t;\theta)\widehat{f}_{ZT}(z,t;\theta).$$

Consistency of $\widehat{\theta}$ is simple to establish, and we omit the details. We proceed to discuss the rate of convergence and find the resulting asymptotic distribution. Differentiating (4.7) with respect to θ, note that $\widehat{\theta}$ solves the equation

$$\sum_{y,z} \int w(y,z,t)\widehat{A}(y,z,t;\theta)\widehat{B}(y,z,t;\theta)dt = 0; \tag{4.8}$$
$$\widehat{B}(y,z,t;\theta) = \frac{\partial}{\partial\theta}\widehat{A}(y,z,t;\theta).$$

With subscripts omitted, we shall need the standard rate of convergence result from nonparametric density estimation for the four kernel density estimates $\widehat{f}$, namely that for each of these estimates, $\widehat{f} = f + O_p\left\{h^2 + (nh)^{-1/2}\right\}$.

The rest of the derivation can then be carried out by Taylor's expansion; see Appendix A. There we show that

$$\widehat{\theta} - \theta_o = (nG)^{-1} \sum_i \Delta\left\{y_i, z_i, t_i(\theta_o); \theta_o\right\} + o_p(n^{-1/2}), \tag{4.9}$$

where

$$G = \sum_{y,z} \int w(y,z,t)B^2(y,z,t;\theta_o)dt;$$

$$\begin{aligned}&\Delta(y,z,t;\theta_o)\\ &= f_T(t;\theta_o)\Big\{w(y,z,t)B(y,z,t;\theta_o) - E[w(y,Z,t)B(y,Z,t;\theta_o)|Y=y, T_{\theta_o}=t]\\ &\quad - E[w(Y,z,t)B(Y,z,t;\theta_o)|Z=z, T_{\theta_o}=t] + E[w(Y,Z,t)B(Y,Z,t;\theta_o)|T_{\theta_o}=t]\Big\}.\end{aligned}$$

The term $B(y,z,t:\theta_o)$ can be expressed more explicitly; see equations (B.1) and (B.2) in Appendix B.

From (4.9), the rate of convergence for estimating θ_o is seen to be root-n. By the central limit theorem, $\widehat{\theta}$ is asymptotically normal, with asymptotic variance equal to $G^{-2}E\Delta(Y,Z,T_{\theta_o};\theta_o)^2$. In order for (4.9) to hold, the bandwidth has to satisfy the condition

$$n^{-\frac{1}{4}} >> h >> n^{-\frac{1}{3}}. \tag{4.10}$$

The lower bound $h >> n^{-\frac{1}{3}}$ is needed for the consistency of $\widehat{B}(y,z,t;\theta)$. The upper bound of h arises from a bias calculation; see Appendix A.

The case that Y is continuous can be handled similarly. The simplest device is to discretize Y by slicing the range of Y into a small number of groups. Models such as (2.2) still hold in general for this sliced Y, so that we can apply the method discussed above for the discrete case. Alternatively, and with more complications numerically, we may apply bivariate kernel density estimates to replace (4.5) and (4.6):

$$\begin{aligned}\widehat{f}_{YZT}(y,z,t;\theta) &= (nhh')^{-1}\sum_{i=1}^{n} I(z_i=z)K\left(\frac{t_i(\theta)-t}{h}\right)K\left(\frac{y_i-y}{h'}\right);\\ \widehat{f}_{YT}(y,t;\theta) &= (nhh')^{-1}\sum_{i=1}^{n} K\left(\frac{t_i(\theta)-t}{h}\right)K\left(\frac{y_i-y}{h'}\right).\end{aligned}$$

We replace (4.7) with

$$\min_{\theta}\sum_{z}\int\int w(y,z,t)\widehat{A}^2(y,z,t;\theta)dydt.$$

The asymptotic expansion of $\widehat{\theta}$ takes the same form as (4.9) with G replaced by

$$G=\sum_{z}\int\int w(y,z,t)B^2(y,z,t;\theta_o)dydt.$$

There is some flexibility in setting up the weight function $w(y,z,t)$ in (4.2) or in (4.4). We shall allow it to depend on θ as well. Two interesting special cases are discussed below.

4.1.1 Partial-inverse mean matching

Instead of using (4.1), we can use (3.3) to write $E(Z|Y,T_{\theta_o}) = E(Z|T_{\theta_o})$. From this we see that $\theta=\theta_o$ is a solution of the minimization problem:

$$\min_{\theta} E\left[\mathrm{Var}\left\{E(Z|Y,T_\theta)|T_\theta\right\}\right]. \tag{4.11}$$

This becomes a special case of the doubly sliced inverse regression method mentioned in the Rejoinder of Li (1991). Denote

$$\begin{aligned}m_Z(t;y,\theta) &= E(Z|Y=y,T_\theta=t) = f_{YZT}(y,1,t;\theta)/f_{YT}(y,t;\theta);\\ m_Z(t;\theta) &= E(Z|T_\theta=t) = f_{ZT}(1,t;\theta)/f_T(t;\theta).\end{aligned}$$

Then we can rewrite (4.11) as

$$\min_{\theta} \sum_{y} \int \{m_Z(t; y, \theta) - m_Z(t; \theta)\}^2 f_{YT}(y, t; \theta) dt. \tag{4.12}$$

The sample version of (4.12) is then to minimize the function

$$\text{PIMM}(\theta) = \sum_{y} \int \{\widehat{m}_Z(t; y, \theta) - \widehat{m}_Z(t; \theta)\}^2 \hat{f}_{YT}(y, t; \theta) dt, \tag{4.13}$$

where

$$\widehat{m}_Z(t; y, \theta) = \widehat{f}_{YZT}(y, 1, t; \theta) / \hat{f}_{YT}(y, t; \theta);$$
$$\widehat{m}_Z(t; \theta) = \widehat{f}_{ZT}(1, t; \theta) / \hat{f}_T(t; \theta).$$

It is easy to see that (4.12) is the same as (4.4) with the weight function

$$w(y, z, t) = \frac{1}{2} f_{YT}(y, t; \theta)^{-1} f_T(t; \theta)^{-2}. \tag{4.14}$$

The dependence of the weight on θ does not change the main asymptotic result. This is similar to the use of weighted least squares in generalized linear models. However, as in the standard nonparametric smoothing context, we do need some standard minor modifications and/or regularity conditions on the denominator densities so as to bound them away from zero; details are routine and therefore omitted here.

The criterion function (4.12) is based on the conditional mean of Z given $(Y, X + \theta Z)$. By comparison, methods for handling (2.4) or (2.5) are based on the conditional mean of Y given (Z, X). Our method thus partially inverts the roles of the involved variables in taking conditional expectations. To emphasize the difference, our method can be referred to as the *partial inverse mean matching* (PIMM) as opposed to the forward mean matching used by the other techniques.

4.1.2. Chi-squared statistics

Another interesting choice of the weight function is

$$w(y, z, t) = \frac{1}{f_T(t; \theta) f_{YT}(y, t; \theta) f_{ZT}(z, t; \theta)}. \tag{4.15}$$

This is related to the chi-squared statistic for testing the conditional independence (3.3). More explicitly, conditional on $T_\theta = t$, suppose we want to test if Y (assumed to be discrete for clarity) is independent of Z. Then the population version of the Pearson chi-squared statistic amounts to

$$\sum_{y,z} \frac{\{f_{Y,Z|T=t}(y, z) - f_{Y|T=t}(y) f_{Z|T=t}(t)\}^2}{f_{Y|T=t}(y) f_{Z|T=t}(t)},$$

where $f_{Y,Z|T=t}, f_{Y|T=t}$, and $f_{Z|T=t}$ denote conditional densities. Combining chi-squares from each t, we obtain

$$\sum_{y,z} \int \frac{\left\{f_{Y,Z|T=t}(y,z) - f_{Y|T=t}(y) f_{Z|T=t}(t)\right\}^2}{f_{Y|T=t}(y) f_{Z|T=t}(t)} f_T(t;\theta)dt. \tag{4.16}$$

Using the relationships $f_{Y,Z|T=t}(y,z) = f_{YZT}(y,z,t;\theta)/f_T(t;\theta)$, $f_{Y|T=t}(y) = f_{YT}(y,t;\theta)/f_T(t;\theta)$, and $f_{Z|T=t}(z) = f_{ZT}(z,t;\theta)/f_T(t;\theta)$, we can express (4.16) in the form (4.4) with the weight function given by (4.15).

4.2. Simulation

A small-scale simulation is reported here to illustrate the partial-inverse mean matching method. To generate the data, we use the following model:

$$\begin{aligned} &X, \epsilon \sim N(0,1);\ Z \sim \text{ binomial}(1,.5);\ X, \epsilon, Z, \text{ are independent} \\ &\qquad Y^* = X + .5Z + \sigma\epsilon; \\ &\qquad Y = 0,1,2, \text{ for } Y^* \le 0, Y^* \ge 0.5, 0 \le Y^* < 0.5, \text{ respectively.} \end{aligned}$$

The treatment effect $\theta = .5$ is equivalent to shifting X up by a half of its standard deviation. The variable Y^* is not observable. We only observe Y which has three classes.

We took $n = 200$ and used $\sigma = .1$ to study the low noise level case first. Figure 4.1 shows a typical set of data generated. The three classes of Y are formed by the partition made with the two horizontal lines as shown there.

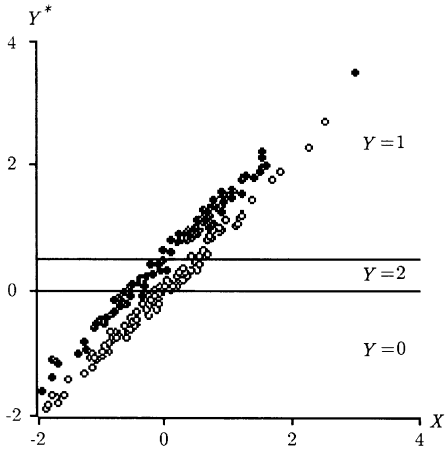

Figure 4.1. Data for low noise level.

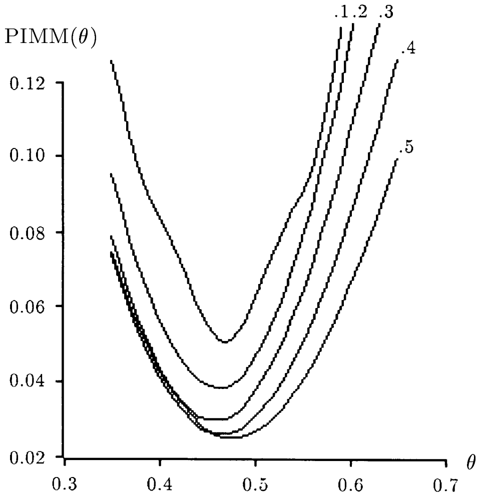

Figure 4.2. The PIMM curves for $h = .1, .2, .3, .4, .5$.

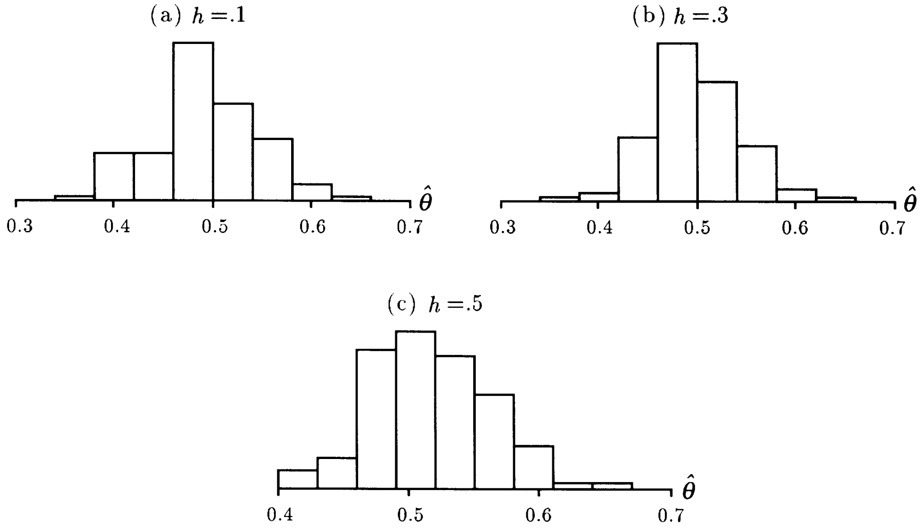

Figure 4.3(a)-(c). Histograms of the estimate $\widehat{\theta}$ in 100 runs.

We use the standard biweight kernel, available in XLISP.STAT (Tierney (1990)), for smoothing. To simplify the computation, we minimize a discretized version of (4.13),

$$\mathrm{PIMM}(\theta) \approx \sum_{y} \sum_{i=1}^{40} \{\widehat{m}_Z(t_i; y, \theta) - \widehat{m}_Z(t_i; \theta)\}^2 \, \widehat{f}_{YT}(t_i, y; \theta); \; t_i = -1.75 + .1i. \tag{4.17}$$

The integration domain $[-1.75, 2.25]$ is chosen because it already covers about 95% of T_θ at the true value $\theta_o = .5$. For the data displayed in Figure 4.1, five curves of PIMM(θ) (4.17),with h set at $.1, .2, .3, .4, .5$, respectively, are shown in Figure 4.2. The minimum $\widehat{\theta}$ is obtained at $.47, .46, .46, .47, .48$ respectively. This indicates that in a wide range of h, the estimate $\widehat{\theta}$ does not change much. To confirm this, we repeated the simulation 100 times. The means and the standard deviations (in parentheses) for these 100 runs are

$$0.498\ (0.053),\ 0.501\ (0.046),\ 0.505\ (0.046),\ 0.511\ (0.046),\ 0.518\ (0.047),$$

for $h = .1, .2, .3, .4, .5$ respectively. Three histograms, $h = .1, .3, .5$, are given in Figures 4.3(a-c).

To simulate a high noise level case, we used $\sigma = .5$. Figure 4.4 shows a typical data set. Note that even with Y^*, it is not easy to tell one group from the other. The criterion curves for minimization are shown in Figure 4.5, with the minimum being $.41, .39, .42, .44, .46$ respectively. Except for $h = .1$, the curves are fairly smooth. The means and standard deviations (in parentheses) of $\widehat{\theta}$ in 100 simulation runs (Figure 4.6) were

$$0.56\ (0.17),\ 0.54\ (0.17),\ 0.54\ (0.14),\ 0.54\ (0.13),\ 0.54\ (0.13).$$

The bias is still small compared to the variance of the estimate. Since the variance of the noise is now 25 times as big as the low noise level case, the mean squared error in estimating θ is also bigger than before, but only by about six or seven times.

Another simple benchmark for comparison is to compute the standard deviation of $\overline{Y}_1^* - \overline{Y}_2^*$, the difference in the average of Y^* for the two groups, which is about the same as the best linear estimate of θ if Y is observable and the linear model is assumed. This gives a value of $.05\sqrt{2}$, or about .071. Thus, the standard error for our estimate $\widehat{\theta}$ is only about twice as large, despite of the loss of information due to classification of Y^* and of course despite the fact that we have made no model assumptions.

Remark 4.1. It would be desirable to have an automatic bandwidth selection rule for selecting h, but this requires further study. What we have demonstrated here is the existence of a range of h for which the estimate $\widehat{\theta}$ is stable. In practice, we suggest plotting the minimization curves for a reasonable range of h as demonstrated in our simulation study. This allows us to assess the sensitivity of the estimate to the bandwidth selection.

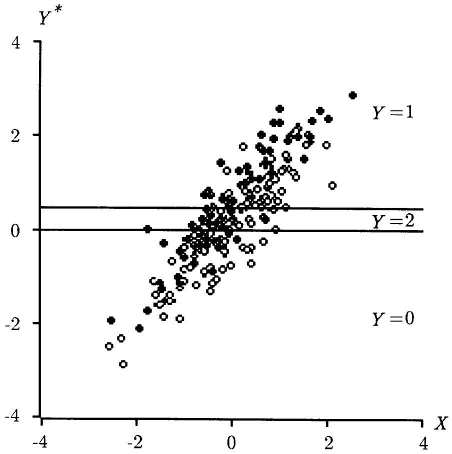

Figure 4.4. Data for high noise level.

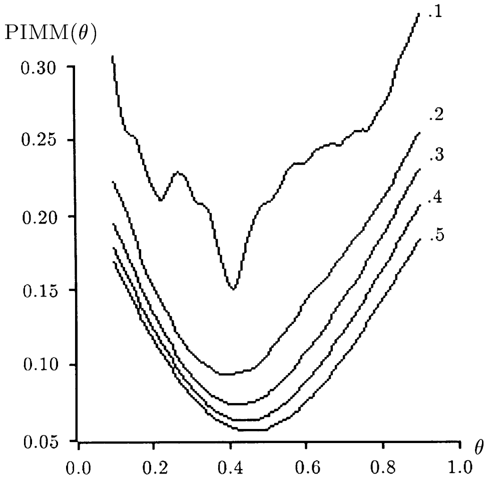

Figure 4.5. The PIMM curves for high noise kevel data, $h = .1, .2, .3, .4, .5$.

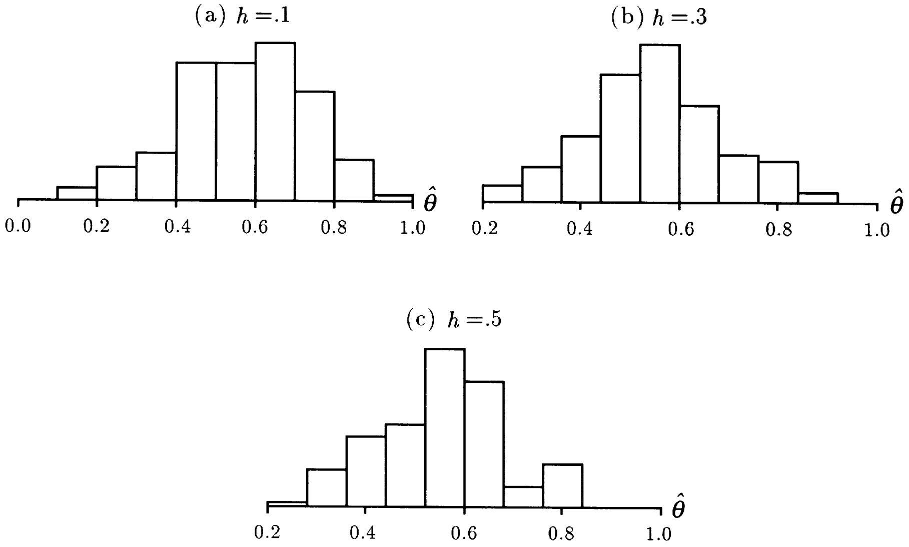

Figure 4.6(a)–(c). Histograms of the estimate $\widehat{\theta}$ in 100 runs for high noise level model.

5. Treatment Comparison with Multivariate Covariates

Treatment comparison in the presence of a high dimensional covariate X is a much broader issue. In this section, we shall discuss an extension of (2.2) to the multivariate case:

$$Y = g(v'X + \theta Z, \epsilon). \tag{5.1}$$

Our strategy is first to estimate the direction v so that the dimensionality of X is reduced to just one. After obtaining a good estimate $\widehat{v}$, we may compute the reduced variable $\widehat{v}'X$, which we then treat as the univariate covariate in the preceding section and apply the one-dimensional technique discussed there to estimate θ.

One must be careful in interpreting the value of θ, which is only defined relative to the size and the sign of v. Note that we can multiply v and θ simultaneously by a constant and absorb the constant into g. This slight ambiguity, which is unavoidable in this context, should not cause any practical difficulties.

The question is how to estimate the direction of v. We use the SIR technique as discussed in the following two subsections.

5.1. Randomized treatment assignment

Randomization is frequently applied in assigning patients to treatment or control groups in medical trials. This practice leads to the condition:

$$X\text{and } Z \text{ are independent.} \qquad (5.2)$$

In order to apply the SIR technique, we assume the linear design condition on the distribution of X:

Linear Design Condition 5.1. For any b in R^{p-1}, $E(b'X|v'X)$ is linear in $v'X$.

The following lemma shows that a direct application of SIR for Y on X (without worrying about the presence of Z) gives a consistent estimate of v.

Lemma 5.1. *Under* (5.1), (5.2), *and the Linear Design Condition* 5.1, *we have* $E(X|Y) - EX \propto \Sigma_X v$, *where* Σ_X *denotes the covariance matrix of* X.

Proof. First assume $EX = 0$ without loss of generality. By conditioning,

$$E(X|Y) = E\left\{E(X|v'X, Z, \epsilon)|Y\right\} = E\left\{E(X|v'X)|Y\right\}.$$

Now apply the *Linear Design Condition* to get $E(X|v'x) \propto \Sigma_X v$. With this, we may return to the preceding expression and complete the proof.

From Lemma 5.1, it follows that the eigenvalue decomposition of $\mathrm{Cov}[E(X|Y)]$ with respect to $\mathrm{Cov}(X)$ has at most one nonzero eigenvalue, and the nonzero eigenvector has to be in the direction of v: $\mathrm{Cov}[E(X|Y)]v = \lambda\Sigma_X v$. This shows that we can estimate v by the first eigenvector of SIR.

Denote the SIR estimate by $\widehat{v}_{\mathrm{sir}}$, from which we derive the reduced variable $\widehat{v}'_{\mathrm{sir}}X$. With this, one can apply the technique from Section 4 to estimate the treatment effect θ.

We illustrate this two-stage procedure by a simulation.

Example 5.1. We used a five-dimensional regressor $X = (X_1, \ldots, X_5)'$ from the standard normal distribution. The model is

$$\begin{aligned}&Y = (\alpha + v'X + \theta Z + \sigma\epsilon)^2;\\ &Z \sim \text{binomial } (1, .5), \quad Z, X \quad \text{independent,}\end{aligned}$$

where ϵ is the standard normal random error. We took $v = (1, 1, -1, -1, 0)'$, $\theta = 1$, $\alpha = 5, \sigma = .5$. After generating $n = 200$ observations, we first applied SIR for Y on X with 10 slices. The eigenvalues are $.85, .092, .032, \ldots$. As expected, only the first one is significant. The corresponding eigenvector is $\widehat{v}_{\mathrm{sir}} = (.53, .50, -.52, -.56, .04)'$. We computed the reduced variable $\widehat{v}'X$ for each of the 200 observations and then used this reduced variable to estimate the treatment effect θ. We discretized Y values into 5 consecutive classes, each with

40 observations, and then applied the partial-inverse mean matching method. The PIMM curves with $h = .1, .2, .3, .4, .5$, are shown in Figure 5.1. The minimum points are about the same: (.46, .50, .48, .48, .49). Since the curve for $h = .1$ shows some instability, we decided to ignore it. Taking the average of the remaining four points, our estimate is $\widehat{\theta} = .48$. Figure 5.2 shows the plot of Y against the variable $T = \widehat{\theta} + \widehat{v}'_{\text{sir}}X$. This plot reveals a clear quadratic pattern.

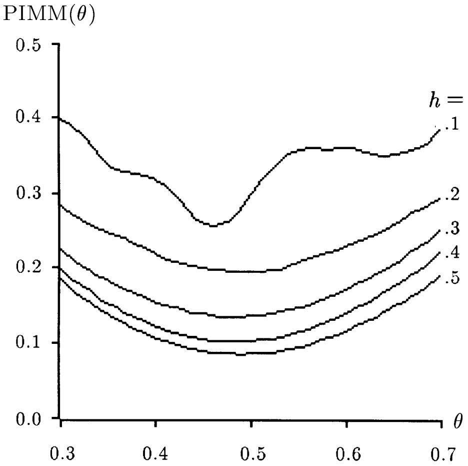

Figure 5.1. Curves of PIMM(θ) for Example 5.1.

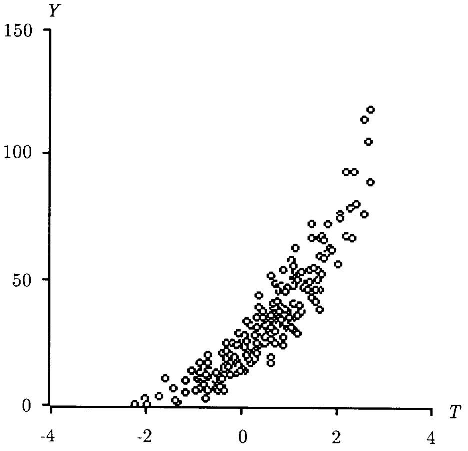

Figure 5.2. Plot of Y against T for Example 5.1.

Note that $\widehat{v}_{\text{sir}}$ is approximately equal to $.5v$. This is why $\widehat{\theta}$ is only about a half of $\theta = 1$. As mentioned earlier, we can absorb the multiplicative constant .5 into the g function in (5.1). On the other hand, the eigenvector which is the output of SIR is usually standardized so that the reduced variable has unit variance. With this convention, the corresponding treatment effect $\widehat{\theta}$ is no longer dependent on the unit of X, thus relieving the ambiguity in the scale interpretation problem.

5.2 Combining SIR

If the treatment assignment is not independent of X, the result of Lemma 5.1 is not true, and direct application of SIR on X may incur some bias. A simple strategy is to estimate v from each group. We need the *Linear Design Condition* for each group :

for any b in R^{p-1}, $E(b'X|v'X, Z = z)$ is linear in $v'X$ for each $z = 0, 1$.

The SIR estimate from each group is the largest eigenvector of the decomposition:

$$\widehat{\Sigma}_{\mathbf{m}z}\widehat{v}_z = \widehat{\lambda}_z\widehat{\Sigma}_{Xz},\ \ z = 0, 1, \tag{5.3}$$

where Σ_{Xz} is the sample covariance of X given $Z = z$ and $\widehat{\Sigma}_{mz}$ is the estimate of Cov $\{E(X|Y, Z = z)\}$, which is constructed by first partitioning the data with $Z = z$ into H_z slices as evenly as possible according to the order of the Y values, computing $\widehat{\mathbf{m}}_{hz}$, the mean of X for the hth slice, and then forming the covariance of these slice means:

$$\widehat{\Sigma}_{\mathbf{m}z} = \sum_h \widehat{p}_{hz}(\widehat{\mathbf{m}}_{hz} - \overline{X}_z)(\widehat{\mathbf{m}}_{hz} - \overline{X}_z)', \tag{5.4}$$

where $\overline{X}_z$ is the sample mean of X for group $Z = z$, and $\widehat{p}_{hz}$ denotes the proportion of cases in slice h.

Having computed $(\widehat{v}_0, \widehat{v}_1)$, an immediate question is how to combine them. There are several ways to proceed. The simplest is to take the average $\widehat{v}_c = (n_0\widehat{v}_0 + n_1\widehat{v}_1)/n$, where n_0 and n_1 denote the sample size for each group. A less simple alternative that incorporates the covariance matrix of each individual estimate is discussed in Appendix C.

The combination can take place before (5.3). There are numerical methods for performing "simultaneous" eigenvalue decomposition. For example, we may take $\widehat{v}_c$ to be the vector v that maximizes the following:

$$\max_v \sum_z \frac{n_z}{n} \cdot \frac{v'\widehat{\Sigma}_{\mathbf{m}z}v}{v'\widehat{\Sigma}_{Xz}v}.$$

When the covariance matrix of X for each group is assumed to be the same, we can pool the two covariance matrices in (5.4) together and conduct a single eigenvalue decomposition:

$$\widehat{\Sigma}_{\mathbf{m}p}\widehat{v}_c = \widehat{\lambda}_c\widehat{\Sigma}_{Xp}\widehat{v}_c,$$

where $\widehat{\Sigma}_{\mathbf{m}p} = n^{-1}\sum_z n_z\widehat{\Sigma}_{\mathbf{m}z}$; $\widehat{\Sigma}_{Xp} = n^{-1}\sum_z n_z\widehat{\Sigma}_{Xz}$.

6. Conclusion

The purpose of this paper is two-fold: (1) to demonstrate how a discrete regressor can be incorporated into the SIR methodology; and (2) to broaden aspects of treatment comparison in the presence of covariates. The entire study takes place in the context of the dimension reduction model (1.1).

We have illustrated the diversity of applications that can be incorporated into our model. The treatment effect was defined and an inverse mean matching estimation method was introduced. For a high dimensional covariate, we suggested a two stage approach: first apply dimension reduction techniques like SIR to reduce the multivariate covariate into a single dimension, and then apply the inverse mean matching estimation method.

In reducing the covariate, we have required the *Linear Design Condition.* In general, SIR is not overly sensitive to a mild departure from this condition, particularly when the covariate is continuous; see the rejoinder of Li (1991), as well as Hall and Li (1993). On the other hand, we can apply subsampling or reweighting techniques to force elliptic symmetry on X; see Brillinger (1991) and Cook and Nachtsheim (1994).

Other dimension reduction methods such as those discussed in Remark 2.1 are feasible alternatives in reducing the covariate dimension. How they compare with SIR is an issue raised and discussed before; see the discussion and rejoinder of Li (1991). Due to the lack of extensive numerical studies in the literature about how well they deal with various design distributions and dimensionality, we cannot comment more about their relative merits at this point.

We have not discussed the case $q > 1$ in (1.2). The dimension reduction technique can be applied separately in each group to reduce the dimensionality of X. The details for estimating the parameter vector Θ will be explored in the future.

Acknowledgement

The research of the first author was supported by a grant from the National Cancer Institute. The research of the second author was supported by an NSF Grant and a Guggenheim Fellowship.

Appendices

Appendix A. Derivation of (4.9)

Consider the local Taylor expansions

$$\widehat{A}(y,z,t;\theta) = \widehat{A}(y,z,t;\theta_o) + \widehat{B}(y,z,t;\theta_o)(\theta-\theta_o) + o_p(\theta-\theta_o);$$

$$\widehat{B}(y,z,t;\theta) = \widehat{B}(y,z,t;\theta_o) + o_p(1).$$

From (4.8), we obtain

$$\begin{aligned}\widehat{\theta}-\theta_o &= -\left\{\sum_{y,z}\int w(y,z,t)\widehat{B}^2(y,z,t;\theta_o)dt\right\}^{-1}\left\{\sum_{y,z}\int w(y,z,t)\widehat{A}(y,z,t;\theta_o)\widehat{B}(y,z,t;\theta_o)dt\right\}\\ &\qquad + o_p(\widehat{\theta}-\theta)\\ &= -\left\{\sum_{y,z}\int w(y,z,t)B^2(y,z,t;\theta_o)dt\right\}^{-1}\left\{\sum_{y,z}\int w(y,z,t)\widehat{A}(y,z,t;\theta_o)\widehat{B}(y,z,t;\theta_o)dt\right\}\\ &\qquad + o_p(\widehat{\theta}-\theta).\end{aligned}$$

To proceed, define

$$\begin{aligned}\tilde{A}(y,z,t;\theta) = {} & \widehat{f}_{YZT}(y,z,t;\theta)f_T(t;\theta_o) + f_{YZT}(y,z,t;\theta_o)\widehat{f}_T(t;\theta) \\ & - \widehat{f}_{YT}(y,t;\theta)f_{ZT}(z,t;\theta_o) - f_{YT}(y,t;\theta_o)\widehat{f}_{ZT}(z,t;\theta). \qquad \text{(A.1)}\end{aligned}$$

Using (4.1), it follows that

$$\begin{aligned} & \widehat{A}(y,z,t;\theta) \\ = {} & \tilde{A}(y,z,t;\theta) + \Big\{\widehat{f}_{YZT}(y,z,t;\theta) - f_{YZT}(y,z,t;\theta)\Big\}\Big\{\widehat{f}_T(t;\theta) - f_T(t;\theta)\Big\} \\ & + \Big\{\widehat{f}_{YT}(y,t;\theta) - f_{YT}(y,t;\theta)\Big\}\Big\{\widehat{f}_{ZT}(z,t;\theta) - f_{ZT}(z,t;\theta)\Big\} \\ = {} & \tilde{A}(y,z,t;\theta) + o_p(n^{-\frac{1}{2}}),\end{aligned}$$

where the last expression is obtained under the assumption that the rate of convergence in density estimation is faster than $n^{-1/4}$, which is very mild and is satisfied under (4.10).

It remains to derive

$$\begin{aligned} & \sum_{y,z}\int w(y,z,t)\tilde{A}(y,z,t;\theta_o)B(y,z,t;\theta_o)dt \\ & = n^{-1}\sum_i \Delta(y_i,z_i,t_i(\theta_o);\theta_o) + o_p(n^{-\frac{1}{2}}). \qquad \text{(A.2)}\end{aligned}$$

This can be done by change of variables. More precisely, the left side of (A.2) can be expanded into four terms, each involving one of the four terms in (A.1). The right side of (A.2) can also be expanded into four terms, each associated with one of the four terms in the definition of $\Delta(y,z,t;\theta_o)$. For the first term on the left side, we have

$$\begin{aligned} & \sum_{y,z}\int w(y,z,t)\widehat{f}_{YZT}(y,z,t;\theta_o)f_T(t;\theta_o)B(y,z,t;\theta_o)dt \\ = {} & (nh)^{-1}\sum_{y,z}\sum_{i=1}^n I(y_i=y)I(z_i=z)\int K(\frac{t_i(\theta_o)-t}{h})w(y,z,t)f_T(t;\theta_o)B(y,z,t;\theta_o)dt \\ = {} & n^{-1}\sum_{i=1}^n\int w(y_i,z_i,t_i(\theta_o)+ht')B(y_i,z_i,t_i(\theta_o)+ht')f_T(t_i(\theta_o)+ht';\theta_o)K(t')dt' \\ = {} & n^{-1}\sum_{i=1}^n\int w(y_i,z_i,t_i(\theta_o))f_T(t_i(\theta_o);\theta_o)B(y_i,z_i,t_i(\theta_o))dt + O_p(h^2).\end{aligned}$$

This gives the first term on the right hand side. The bandwidth interval given by (4.10) implies $O_p(h^2) = o_p(n^{-\frac{1}{2}})$. All other three terms can be obtained similarly.

Appendix B. Derivation of $B(y,z,t;\theta_o)$

The term $B(y,z,t;\theta_o)$ can be expressed more explicitly. This is based on the relationship $f_{YZT}(y,z,t;\theta) = f_{YZT}(y,z,t+(\theta_o-\theta)z;\theta_o)$. Taking partial derivatives with respect to θ, it follows that

$$\frac{\partial}{\partial\theta} f_{YZT}(y,z,t;\theta)|_{\theta_o} = -z\frac{\partial}{\partial t} f_{YZT}(y,z,t;\theta_o).$$

We also have

$$\frac{\partial}{\partial\theta} f_{YT}(y,t;\theta)|_{\theta_o} = \sum_z \frac{\partial}{\partial\theta} f_{YZT}(y,z,t;\theta)|_{\theta_o} = -\frac{\partial}{\partial t} f_{YZT}(y,1,t;\theta_o).$$

Similarly,

$$\begin{aligned}\frac{\partial}{\partial\theta} f_{ZT}(z,t;\theta)|_{\theta_o} &= -z\frac{\partial}{\partial t} f_{ZT}(z,t;\theta_o);\\ \frac{\partial}{\partial t} f_{T}(z,t;\theta)|_{\theta_o} &= -\frac{\partial}{\partial t} f_{ZT}(1,t;\theta_o).\end{aligned}$$

The case that $z=0$ is now obvious:

$$\begin{aligned}B(y,0,t;\theta_o) =& f_{ZT}(0,t;\theta_o)\frac{\partial}{\partial t} f_{YZT}(y,1,t;\theta_o)\\ &- f_{YZT}(y,0,t;\theta_o)\frac{\partial}{\partial t} f_{ZT}(1,t;\theta_o).\end{aligned} \tag{B.1}$$

For the case $z=1$, we need to use the identity obtained from taking the partial derivative with respective to t on both sides of (4.1). This leads to

$$\begin{aligned}B(y,1,t;\theta_o) =& f_{YZT}(y,1,t;\theta_o)\frac{\partial}{\partial t} f_{ZT}(0,t;\theta_o)\\ &- f_{ZT}(1,t;\theta_o)\frac{\partial}{\partial t} f_{YZT}(y,0,t;\theta_o).\end{aligned} \tag{B.2}$$

Appendix C. Combination via covariance weighting

This method is motivated from the estimation of a normal mean vector μ, given two independent observations $\mathbf{u}_0, \mathbf{u}_1$ with the common mean μ but different covariance matrices, Σ_0, Σ_1. The maximum likelihood estimate is equal to $(\Sigma_0^{-1} + \Sigma_1^{-1})^{-1}(\Sigma_0^{-1}\mathbf{u}_0 + \Sigma_1^{-1}\mathbf{u}_1)$, which can be interpreted as a matrix-weighted average of $\mathbf{u}_0$ and $\mathbf{u}_1$ with the covariance matrices as the weights. This combination method can be applied to $\widehat{v}_0, \widehat{v}_1$ which have been shown to be asymptotically normal (Duan and Li 1991). However, some modification is needed because the covariance matrix for $\widehat{v}_z, z=0,1$, denoted by Σ_z, is always degenerate due to the normalization constraint on the length of each eigenvector $\|\widehat{v}_z\| = 1$, forcing $\Sigma_z v = 0$.

Let $\widehat{\Sigma}_z$ be an estimate of Σ_z with $\widehat{\Sigma}_z \widehat{v}_z = 0$. The maximum likelihood estimate is obtained by $\min_v \sum_z (\widehat{v}_z - v)' \widehat{\Sigma}_z^{-1} (\widehat{v}_z - v)$. There are difficulties in inverting the matrix. To circumvent this problem, we may minimize $\sum_z (\widehat{v}_z' v)^{-2} (\widehat{v}_z - v)' (\widehat{\Sigma}_z)^- (\widehat{v}_z - v)$, or equivalently, minimize

$$\sum_z (\widehat{v}_z' v)^{-2} v' (\widehat{\Sigma}_z)^- v. \tag{C.1}$$

The insertion of the term $\widehat{v}_z' v = \|v\| \cos(\widehat{v}_z, v)$ acts like a regularization procedure in ill-posed problems. It assures that the solution v may not be too far away from each $\widehat{v}_z$. Another advantage is that (C.1) is invariant under the scale change of v, a desirable property for us because our interest is mainly in the direction (but not the length) of the combined estimate.

First decompose v into two orthogonal parts, $v = u + \tilde{u} = Pv + Qv$, where u is in the space spanned by $(\widehat{v}_0, \widehat{v}_1)$; $\tilde{u}$ is in the orthogonal complement; and P, Q are the projection matrices. Then one can write (C.1) as

$$\tilde{u}' A(u) u + 2 \tilde{u}' B(u) u + u' C(u) u, \tag{C.2}$$

where $A(u) = \sum_z (u' \widehat{v}_z)^{-2} Q \widehat{\Sigma}_z^- Q$; $B(u) = \sum_z (u' \widehat{v}_z)^{-2} Q \widehat{\Sigma}_z^- P$; $C(u) = \sum_z (u' \widehat{v}_z)^{-2}$ $P \widehat{\Sigma}_z^- P$. One need only take $\tilde{u} = A(u)^- B(u) u$ as the solution. This reduces to the minimization of $u'[C(u) - B(u)' A(u)^- B(u)] u$. This minimization involves only one parameter, but there is no closed form solution. An initial value of u can be found by minimizing $u' C(u) u$.

References

Brillinger, D. R. (1991). Discussion of "Sliced inverse regression". *J▷Amer▷Statist▷Assoc▷* **86**, 333-333.

Carroll, R. J. and Li, K. C. (1992). Measurement error regression with unknown link : Dimension reduction and data visualization. J. Amer. Statist. Assoc. **87**, 1040-1050.

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

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

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

Cook, R. D. and Nachtsheim, C. J. (1994). Reweighting to achieve elliptically contoured covariates in regression. J. Amer. Statist. Assoc. **89**, 592-599.

Cook, R. D. and Weisberg, S. (1991). Discussion of "Sliced inverse regression". J. Amer. Statist. Assoc. **86**, 328-332.

Duan, N. and Li, K. C. (1991). Slicing regression: A link-free regression method. Ann. Statist. **19**, 505-530.

Engle, R. F., Granger, C. W. J., Rice, J. and Weiss, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. J. Amer. Statist. Assoc. **81**, 310-320.

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

Hall, P. and Li, K. C. (1993). On almost linearity of low dimensional projections from high dimensional data. *Ann▷Statist▷* **21**, 867-889.

Han, A. K. (1987). Non-parametric analysis of a generalized regression model. J. Econometrics **35**, 303-316.

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

Härdle, W. and Marron, J. S. (1990). Semiparametric comparison of regression curves, Ann. Statist. **18**, 63-89.

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

Heckman, N. E. (1986). Spline smoothing in a partly linear model. J. Roy. Statist. Soc. Ser.B **48**, 244-248.

Hsing, T. and Carroll, R. J. (1992). An asymptotic theory for sliced inverse regression. Ann. Statist. **20**, 1040-1061.

Kneip, A. and Gasser, T. (1992). Statistical tools to analyze data representing a sample of curves. Ann. Statist. **20**, 1266-1305.

Li, K. C. (1991). Sliced inverse regression for dimension reduction, with discussions. J. Amer. Statist. Assoc. **86**, 316-342.

Li, K. C. (1992a). Uncertainty analysis for mathematical models with SIR. In Probability and Statistics (Edited by Ze-Pei Jiang, Shi-Jian Yan, Ping Cheng and Rong Wu), 138-162, World Scientific, Singapore.

Li, K. C. (1992b). On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. J. Amer. Statist. Assoc. **87**, 1025-1039.

Rice, J. (1986). Convergence rates for partially splined models. Statist. Probab. Lett. **4**, 203-208.

Samarov, A. M. (1993). Exploring regression structure using nonparametric functional estimation. J. Amer. Statist. Assoc. **88**, 836-847.

Sherman, R. P. (1993). The limiting distribution of the maximum rank correlation estimator. Econometrica **61**, 123-137.

Speckman, P. (1988). Kernel smoothing in partial linear models. J. Roy. Statist. Soc. Ser.B **50**, 413-436.

Tierney, L. (1990). LISP-STAT: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics. John Wiley, New York.

Department of Statistics, Texas A&M University, College Station, TX 77843-3143, U.S.A.

Department of Mathematics, University of California, Los Angeles, CA 90024, U.S.A.

(Received June 1993; accepted March 1995)