*The Annals of Statistics*
1991, Vol. 19, No 2, 505–530

# SLICING REGRESSION: A LINK-FREE REGRESSION METHOD

BY NAIHUA DUAN[1] AND KER-CHAU LI[2]

*RAND Corporation and University of California at Los Angeles*

Consider a general regression model of the form $y = g(\alpha + \mathbf{x}'\beta, \varepsilon)$, with an arbitrary and unknown link function $g$. We study a link-free method, the slicing regression, for estimating the direction of $\beta$. The method is easy to implement and does not require iterative computation. First, we estimate the inverse regression function $E(\mathbf{x}|y)$ using a step function. We then estimate $\Gamma = \mathrm{Cov}[E(\mathbf{x}|y)]$, using the estimated inverse regression function. Finally, we take the spectral decomposition of the estimate $\hat{\Gamma}$ with respect to the sample covariance matrix for $\mathbf{x}$. The principal eigenvector is the slicing regression estimate for the direction of $\beta$. We establish $\sqrt{n}$-consistency and asymptotic normality, derive the asymptotic covariance matrix and provide Wald's test and a confidence region procedure. Efficiency is discussed for an important special case.

Most of our results require $\mathbf{x}$ to have an elliptically symmetric distribution. When the elliptical symmetry is violated, a bias bound is provided; the asymptotic bias is small when the elliptical symmetry is nearly satisfied. The bound suggests a projection index which can be used to measure the deviation from elliptical symmetry.

The theory is illustrated with a simulation study.

**1. Introduction.** Regression analysis is usually based on a working model. For example, we might assume the standard linear model

$$y = \alpha + \mathbf{x}'\beta + \varepsilon, \qquad \varepsilon|\mathbf{x} \sim N(0, \sigma^2),$$

where $y$ denotes a scalar outcome variable, $\mathbf{x}$ denotes a $d$-dimensional column vector of regressor variables and $\beta$ denotes a $d$-dimensional column vector of slope coefficients. Under this model, we might use the least squares regression to estimate the parameter vector $(\alpha, \beta')$.

In most empirical applications of regression analysis, the working model is at best an approximation. We probably do not believe in the working model, therefore we should be concerned about robustness against violations of the assumptions in the model. For example, we might consider *distribution viola-*

Received January 1989; revised August 1989.

[1]Research supported in part by a cooperative agreement between RAND Corporation, SIMS, U.S. EPA and in part by RAND corporate funds.

[2]Research supported by NSF Grant DMS-86-02018.

*AMS* 1980 *subject classifications.* 62J99.

*Key words and phrases.* Elliptical symmetry, general regression model, inverse regression, projection pursuit, spectral decomposition.

*tion*: the error distribution might not be normal. There is a rich literature on distribution violation for the linear model and robust methods which protect against distribution violation; see, e.g., Huber (1981).

On the other hand, the functional form of the model might also be violated. For example, in applying the least squares regression, one might be concerned whether $E(y|\mathbf{x})$ is indeed linear in $\mathbf{x}'\beta$. Instead, the true model might have the following form:

$$y = g(\alpha + \mathbf{x}'\beta, \varepsilon), \qquad \varepsilon|\mathbf{x} \sim F(\varepsilon), \tag{1.1}$$

where the bivariate function $g$ is the *link function*, $F$ is the error distribution; both $g$ and $F$ are assumed to be arbitrary and unknown. We call a model of form (1.1) a *general regression model*.

When the link function is arbitrary and unknown, we cannot estimate the entire parameter vector $(\alpha, \beta')$. The most that can be identified for $(\alpha, \beta')$ is the *direction* of the slope vector $\beta$, that is, the collection of the ratios $\{\beta_j/\beta_k, j, k = 1, \ldots, d\}$. In other words, we can only determine the line generated by $\beta$, but not the length or the orientation of $\beta$. Whether we can actually identify the direction of $\beta$ is examined in Appendix A.

Why should we be concerned about estimating the direction of $\beta$? In some situations, the direction of $\beta$ might be the estimand of interest; an interesting example from radiobiology is given in Vegesna, Withers and Taylor (1988). When the prediction of $y$ from $\mathbf{x}$ is of interest, we can first estimate the direction of $\beta$, then use nonparametric regression of $y$ on $\mathbf{x}'\hat{\beta}$ to estimate the link function $g$. For inference purposes, being able to identify the direction of $\beta$ means we can distinguish between $H_0$: $\beta_j = 0$ and $H_A$: $\beta_j \neq 0$, i.e., we can determine whether a specific regressor variable, say, $x_j$, has an effect on the outcome; see the inference results in Section 4.3.

Given sufficient prior information, we might specify a link function and an error distribution and proceed with the parametric regression method based on the working model. For example, we might specify the standard linear model and proceed with the least squares regression. If the true model does not have the specified link function, we have link violation. Brillinger (1977, 1983), Goldberger (1981), Greene (1981, 1983), White (1981), Chung and Goldberger (1984), Ruud (1983, 1986), Duan and Li (1985, 1987, 1991) and Li and Duan (1989) studied the behavior of various parametric regression methods under link violation.

There are many situations in which we do not have precise knowledge about the link function. It is therefore desirable to use estimation methods which do not require the specification of a link function. We will call such estimation methods *link-free* regression methods. Even when we do have some prior information on the link function, it might still be desirable to use link-free methods because the prior information might be highly imprecise.

We study a link-free regression method, the *slicing regression*, for estimating the direction of $\beta$. The slicing regression is very easy to implement and does not require iterative computation. The method is based on a crucial

relationship between the inverse regression $E(\mathbf{x}|y)$ and the forward regression slope $\beta$; see Theorem 2.1. The empirical algorithm is given in Section 3. First, we estimate the inverse regression curve $E(\mathbf{x}|y)$ using a step function: we partition the range of $y$ into slices and estimate $E(\mathbf{x}|y)$ in each slice of $y$ by the sample average of the corresponding $\mathbf{x}$'s. We then estimate the covariance matrix $\Gamma = \mathrm{Cov}[E(\mathbf{x}|y)]$, using the estimated inverse regression curve. Finally, we take the spectral decomposition of the estimate $\hat{\Gamma}$ with respect to the sample covariance matrix for $\mathbf{x}$. The principal eigenvector is the slicing regression estimate for the direction of $\beta$.

We establish the basic asymptotic theory for the slicing regression in Section 4: consistency, asymptotic normality, the asymptotic covariance matrix, Wald's test and a confidence region procedure. We also discuss efficiency for an important special case: there exists an unknown transformation of $y$ to the standard linear model and $\mathbf{x}$ is normally distributed. The slicing regression usually has good efficiency and is insensitive to how the range of $y$ is partitioned into slices.

Most of the results in this paper require the following design condition:

(DC.1) The regressor variable $\mathbf{x}$ is sampled randomly from a nondegenerate elliptically symmetric distribution.

We will refer to the distribution of $\mathbf{x}$ as the *design distribution*. We study the behavior of the slicing regression when the design distribution deviates from elliptical symmetry and establish a bias bound (Theorem 6.1); the asymptotic bias is small when the design distribution is nearly elliptically symmetric. The bias bound can be estimated empirically and suggests a projection index which can be used to measure the deviation from elliptical symmetry.

We can obtain different versions of the slicing regression by using different weights for the slices when estimating $\Gamma$. We derive the optimal weights in Section 5. If the distribution of $\mathbf{x}$ is normal, the optimal weights are proportional to the number of observations in each slice. If the distribution of $\mathbf{x}$ is elliptically symmetric but not normal, the optimal weights in essence impose a heterogeneity correction.

We give results from a simulation study in Section 7 to demonstrate the behavior of the slicing regression for moderate sample sizes.

REMARK 1.1. The slicing regression reduces to the usual discriminant analysis if we partition the range of $y$ into two slices. We usually assume normality for $\mathbf{x}$ (conditioned on $y$) for discriminant analysis [Fisher (1936), Haggstrom (1983)]. This condition is analogous to (DC.1); both conditions follow from the weaker condition (DC.1′) in Remark 2.2.

**2. Inverse regression.** An obvious way to estimate the direction of $\beta$ without specifying a link function is to use a suitable nonparametric regression method to estimate the *forward regression function*

$$\eta(\mathbf{x}) = E(y|\mathbf{x}).$$

Since $\eta$ depends on $\mathbf{x}$ only through $\mathbf{x}'\beta$ when $y$ follows a general regression model of form (1.1), it is possible to determine $\beta$ from $\eta$, e.g., using the gradient of $\eta$, which is proportional to $\beta$. This approach might be unsatisfactory due to the curse of dimensionality: For realistic sample sizes, it is difficult to implement standard nonparametric regression methods such as kernel methods, nearest neighbor methods, or smoothing splines, when the dimensionality $d$ is larger than two, because the design points are very sparse; see, e.g., Huber (1985).

In order to avoid the curse of dimensionality, we consider the *inverse regression function*

$$\xi(y) = E(\mathbf{x}|y). \tag{2.1}$$

The inverse regression function is easy to estimate because $y$ is a scalar; each component of the function can be estimated as a one-on-one nonparametric regression, thus we are free from the curse of dimensionality.

The inverse regression function might be of interest in its own right for studying the relationship between $y$ and $\mathbf{x}$. For example, inverse regression received a fair amount of attention in calibration problems; see, e.g., Krutchkoff (1967, 1969) and Hunter and Lamboy (1981). Conway and Roberts (1983) used a variant of the inverse regression, the reverse regression, to study job discrimination.

When we are mainly interested in the forward regression, it might still be useful to consider the inverse regression if the inverse regression provides useful information about the forward regression. This is established in the following theorem.

THEOREM 2.1. *Assume the general regression model* (1.1) *and the design condition* (DC.1). *The inverse regression function* (2.1) *falls along a line*:

$$\xi(y) = \mu + \Sigma\beta\kappa(y), \tag{2.2}$$

*where* $\mu = E(\mathbf{x})$, $\Sigma = \operatorname{cov}(\mathbf{x})$ *and* $\kappa(y)$ *is a scalar function of* $y$:

$$\kappa(y) = \frac{E[(\mathbf{x} - \mu)'\beta|y]}{\beta'\Sigma\beta}. \tag{2.3}$$

PROOF. Design condition (DC.1) implies

$$E(\mathbf{x}|\mathbf{x}'\beta) = \mu + \frac{\Sigma\beta\beta'(\mathbf{x} - \mu)}{\beta'\Sigma\beta}. \tag{2.4}$$

The theorem follows from the fact that $\xi(y) = E[E(\mathbf{x}|\mathbf{x}'\beta)|y]$. □

According to the theorem,

$$\beta \propto \Sigma^{-1}(\xi(y) - \mu), \tag{2.5}$$

with the proportionality constant being $1/\kappa(y)$. For any $y$ with $\kappa(y) \neq 0$, we can determine the direction of $\beta$ using the right-hand side of (2.5). The following corollary allows us to combine the information from all $y$'s.

COROLLARY 2.2. *Assume the same conditions in Theorem* 2.1. *Let* $\Gamma = \operatorname{cov}(\xi(y))$. *The slope vector* $\beta$ *solves the following maximization problem*:

$$\max_{\mathbf{b}\in R^d} L(\mathbf{b}), \quad \textit{where} \quad L(\mathbf{b}) = \frac{\mathbf{b}'\Gamma\mathbf{b}}{\mathbf{b}'\Sigma\mathbf{b}}. \tag{2.6}$$

*The solution is unique* (*up to a multiplicative scalar*) *if and only if* $\kappa(y) \not\equiv 0$.

PROOF. According to (2.2),

$$\Gamma = \operatorname{Var}(\kappa(y))\Sigma\beta\beta'\Sigma$$

has rank one. The result follows from Cauchy's inequality. □

According to the corollary, $\beta$ is the principal eigenvector for $\Gamma$ with respect to the inner product

$$(\mathbf{b}, \mathbf{v}) = \mathbf{b}'\Sigma\mathbf{v}. \tag{2.7}$$

The maximum $L(\beta)$ is the principal eigenvalue. Since the rank of $\Gamma$ is one, the spectral decomposition for $\Gamma$ is trivial: all eigenvalues except the first are zero. The corollary simply restates the fact that $\xi(y)$ falls along the line (2.2). On the other hand, the method suggested in Corollary 2.2 is more useful than that in Theorem 2.1 when the design distribution deviates from elliptical symmetry. Although the inverse regression might no longer fall along a line, it is still possible that $\beta$ would (nearly) solve the maximization problem (2.6); see Remark 2.1.

The maximand $L(\mathbf{b})$ is the $R^2$ for the nonparametric regression of $\mathbf{x}'\mathbf{b}$ on $y$:

$$L(\mathbf{b}) = \frac{\operatorname{Var}[E(\mathbf{x}'\mathbf{b}|y)]}{\operatorname{Var}(\mathbf{x}'\mathbf{b})}.$$

It measures how well we can predict $\mathbf{x}'\mathbf{b}$ from $y$. The corollary indicates that among all linear combinations $\mathbf{x}'\mathbf{b}$, $y$ predicts $\mathbf{x}'\beta$ the best.

REMARK 2.1. If the stochastic term $\varepsilon$ is degenerate and the link function is invertible, Corollary 2.2 would hold for any design distribution, elliptically symmetric or not. To see this, note that conditioning on $y$ is equivalent to conditioning on $\mathbf{x}'\beta$, therefore $\Gamma = \operatorname{Cov}[E(\mathbf{x}|\mathbf{x}'\beta)]$. The maximand $L(\mathbf{b})$ is the $R^2$ for the regression of $\mathbf{x}'\mathbf{b}$ on $\mathbf{x}'\beta$, which is maximized for $\mathbf{b} \propto \beta$. Further discussions on Corollary 2.2 are given in Section 6.

REMARK 2.2. Design condition (DC.1) can be replaced by the following weaker condition in Theorem 2.1 and Corollary 2.2:

(DC.1′) The regressor $\mathbf{x}$ is sampled randomly from a nondegenerate probability distribution; the conditional expectation $E(\mathbf{x}'\mathbf{b}|\mathbf{x}'\beta)$ is linear in $\mathbf{x}'\beta$ for all $\mathbf{b} \in R^d$.

**3. Slicing regression.** We now apply the results in Section 2 to the sampling case. Given a random sample $\{(y_i, \mathbf{x}_i'),\ i = 1, \ldots, n\}$ from a general

regression model (1.1), we want to estimate the direction of the slope vector $\beta$. In order to apply Theorem 2.1 or Corollary 2.2, we need to estimate the inverse regression function $\xi(y)$. For simplicity, we use a step function estimate (cf. Remark 3.1). We partition the range of $y$ into, say, $H$ slices, $\{s_1, \ldots, s_H\}$. For each slice of $y$, we estimate $\xi(y) = E(\mathbf{x}|y)$ by the sample average of the corresponding $\mathbf{x}$'s. More specifically, our estimated inverse regression function is

$$\hat{\xi}(y) = \hat{\xi}_h = \frac{\sum_{i=1}^n \mathbf{x}_i 1_{ih}}{\sum_{i=1}^n 1_{ih}} \quad \text{if } y \in s_h, \tag{3.1}$$

where $1_{ih}$ is the indicator for the event $y_i \in s_h$.

The estimated inverse regression function converges to the true inverse regression function if we choose a suitable sequence of partitions whose meshes decrease to zero as $n \to \infty$. However, since $\xi(y)$ falls along a line, a crude estimate for $\xi(y)$ is adequate for estimating its direction. We assume for simplicity that the partition is fixed a priori and does not depend on $n$.

Under the same assumptions in Theorem 2.1, we have

$$\xi_h = E(\hat{\xi}_h) = \mu + \Sigma\beta k_h, \tag{3.2}$$

where

$$k_h = E[\kappa(y)|y \in s_h] = \frac{E[(\mathbf{x} - \mu)'\beta | y \in s_h]}{\beta'\Sigma\beta}, \tag{3.3}$$

thus the expectation of the estimated inverse regression function also falls along the line (2.2).

If the scalar $k_j$ for the $j$th slice is nonzero, we can estimate the direction of $\beta$ using the direction of $\tilde{\beta}^{(j)} = \hat{\Sigma}^{-1}(\hat{\xi}_j - \bar{\mathbf{x}})$, where $\bar{\mathbf{x}}$ is the sample average and $\hat{\Sigma}$ is the sample covariance matrix for the observed $\mathbf{x}$'s. By the central limit theorem, $\tilde{\beta}^{(j)}$ converges to $k_j\beta$ at rate $\sqrt{n}$, therefore the direction of $\tilde{\beta}^{(j)}$ is $\sqrt{n}$-consistent for the direction of $\beta$ if $k_j$ is nonzero.

Usually there is more than one slice for which $k_h$ is nonzero. We should combine the information from all the slices to estimate the direction of $\beta$. We will use a modification of the maximization problem (2.6) to do this. First we introduce some notations:

$$\begin{gathered} p_h = P(y \in s_h), \quad \mathbf{p} = (p_1, \ldots, p_H)', \quad \mathbf{k} = (k_1, \ldots, k_H)', \\ \xi = [\xi_1, \ldots, \xi_H], \quad \hat{\xi} = [\hat{\xi}_1, \ldots, \hat{\xi}_H]. \end{gathered} \tag{3.4}$$

We estimate $\Gamma$ by

$$\hat{\Gamma} = \hat{\xi} W \hat{\xi}', \tag{3.5}$$

where $W$ is an arbitrary symmetric nonnegative definite $H$ by $H$ matrix, chosen a priori, which satisfies

$$W\mathbf{1} = 0. \tag{3.6}$$

We can interpret $\hat{\Gamma}$ as a weighted covariance matrix for the data vector $\{\hat{\xi}_1, \ldots, \hat{\xi}_H\}$, using $W$ as the weight matrix. Condition (3.6) is required for $\hat{\Gamma}$ to

be location invariant. By the strong law of large numbers, $\hat{\Gamma}$ converges almost surely to

$$\xi W \xi' = \mathbf{k}' W \mathbf{k} \Sigma \beta \beta' \Sigma,$$

which is proportional to $\Gamma$.

For a given weight matrix $W$, we consider a maximization problem similar to (2.6):

$$\max_{\mathbf{b} \in R^d} \hat{L}(\mathbf{b}), \quad \text{where } \hat{L}(\mathbf{b}) = \frac{\mathbf{b}'\hat{\Gamma}\mathbf{b}}{\mathbf{b}'\hat{\Sigma}\mathbf{b}}. \tag{3.7}$$

We will refer to any solution to (3.7), $\hat{\beta}$, as a *slicing regression estimate* for the direction of $\beta$. This is usually defined uniquely up to a multiplicative scalar. The slicing regression estimate $\hat{\beta}$ is the principal eigenvector for $\hat{\Gamma}$ with respect to the inner product

$$[\mathbf{b}, \mathbf{v}] = \mathbf{b}'\hat{\Sigma}\mathbf{v}. \tag{2.7'}$$

The maximum

$$\hat{\lambda}_1 = \hat{L}(\hat{\beta}) \tag{3.8}$$

is the principal eigenvalue.

The estimate $\tilde{\beta}^{(j)}$ discussed earlier is a special case of the slicing regression estimate, when the weight matrix is taken to be $W = \mathbf{u}\mathbf{u}'$, where $u_j = 1 - \hat{p}_j$, $u_h = -\hat{p}_h$ for $h \neq j$ and $\hat{p}_h$ is the sample proportion of $y_i$'s in the $h$th slice.

There are many other choices for the weight matrix $W$. For example, we can take

$$W = W^{(\mathbf{r})} = D(\mathbf{r}) - \mathbf{r}\mathbf{r}'; \qquad \mathbf{r}'\mathbf{1} = 1; \qquad r_h \geq 0, \qquad h = 1, \ldots, H; \tag{3.9}$$

where $D(r)$ denotes the diagonal matrix with elements from the $H$-dimensional column vector $\mathbf{r}$. With this weight matrix, $\hat{\Gamma}$ is the covariance matrix for the data vectors $\{\hat{\xi}_1, \ldots, \hat{\xi}_H\}$, with the $h$th slice weighted by $r_h$. Note that $\mathbf{r}$ is a probability measure on the index set $\{1, \ldots, H\}$.

An especially important weight matrix of form (3.9) is $W^{(\mathbf{p})}$, for which each slice is weighted by the probability for $y$ to fall inside the slice. We will refer to this weight matrix as the *proportional to size* (pps) weight matrix. We will show in Section 5 that the pps weight matrix is optimal when the design distribution is normal.

REMARK 3.1. The step function estimate (3.1) might not be very efficient for estimating $\xi(y)$ and can be improved upon, e.g., using kernel estimates. We might also choose the amount of smoothing adaptively, say, choose $H$ adaptively. In order to present the idea of slicing regression with a minimum of obfuscation, we have focused on the step function estimate with a fixed partition. The consideration of other smoothing methods and the adaptive choice of the smoothing parameter remains to be examined. However, the efficiency result in Section 4.5 indicates that the slicing regression estimate based on a fixed partition step function is nearly fully efficient for an impor-

tant special case, therefore the method might be insensitive to the smoothing method or the smoothing parameter.

REMARK 3.2. If the covariance matrix $\Sigma$ is known, we can use $\Sigma$ instead of $\hat{\Sigma}$ in our maximization problem:

$$\max_{\mathbf{b}\in R^d} \tilde{L}(\mathbf{b}), \quad \text{where } \tilde{L}(\mathbf{b}) = \frac{\mathbf{b}'\hat{\Gamma}\mathbf{b}}{\mathbf{b}'\Sigma\mathbf{b}}. \tag{3.7'}$$

When the distinction is necessary, we will refer to the slicing regression based on (3.7) as the *ignorant* slicing regression and refer to the slicing regression based on (3.7') as the *nonignorant* slicing regression. The nonignorant slicing regression estimate $\hat{\beta}$ is the principal eigenvector for $\hat{\Gamma}$ with respect to the inner product (2.7). The maximum

$$\hat{\lambda}_1 = \tilde{L}(\hat{\beta}) \tag{3.8'}$$

is the principal eigenvalue.

It might appear that the nonignorant estimate would perform better than the ignorant estimate. Contrary to this intuition, the ignorant estimate usually performs better; see Section 4.4.

**4. Asymptotic theory.** We now establish the basic asymptotic behavior for the slicing regression: consistency, asymptotic normality and the asymptotic covariance matrix. The results are then applied to two standard inference problems: testing a null hypothesis and constructing a confidence region. We also discuss efficiency for an important special case. Throughout this section we assume the weight matrix $W$ is given a priori and satisfies (3.6).

4.1. *Consistency.* It was noted in Section 3 that we can estimate the direction of $\beta$ consistently using $\tilde{\beta}^{(j)}$ if $k_j$ is nonzero. We now consider the consistency property for the slicing regression estimate in general. An estimate $\hat{\beta}$ is consistent for the direction of $\beta$ if the angle between $\hat{\beta}$ and $\beta$ converges to zero, i.e.,

$$\cos^2(\hat{\beta}, \beta) = \frac{(\hat{\beta}'\Sigma\beta)^2}{(\hat{\beta}'\Sigma\hat{\beta})(\beta'\Sigma\beta)} \to 1 \quad \text{(a.s.)}, \tag{4.1}$$

where the cosine function is taken with respect to the inner product (2.7).

THEOREM 4.1. *Assume the general regression model* (1.1), *the design condition* (DC.1) *and the following conditions*:

(DC.2) *The weight matrix $W$ is symmetric and nonnegative definite and satisfies* (3.6).

(DC.3) $$\mathbf{k}'W\mathbf{k} > 0.$$

The slicing regression estimate $\hat{\beta}$, which solves the maximization problem (3.7) or (3.7'), is consistent for the direction of $\beta$. Furthermore, the estimated

principal eigenvalue $\hat{\lambda}_1$ in (3.8) or (3.8′) is a consistent estimate for the population principal eigenvalue

$$\lambda_1 = \mathbf{k}'W\mathbf{k}. \tag{4.2}$$

The proof is sketched in Appendix C.

It might be difficult to verify (DC.3) because very little is known a priori about $\mathbf{k}$. If $\text{rank}(W) = H - 1$, then (DC.3) is satisfied if

$$\mathbf{k} \neq 0, \tag{DC.3′}$$

which is much easier to verify.

A sufficient condition for (DC.3′) is that $\kappa(y)$ be monotonic. This condition might be verifiable a priori in many empirical applications. A scientist might not have enough prior information to specify a link function; he might, however, have enough prior information on the ranking of the effects, so he can affirm the monotonicity of $\kappa$. Further discussions on the monotonicity condition are given in Appendix A.

4.2. *Asymptotic distribution.* We now discuss the asymptotic distribution for the slicing regression. We assume for convenience that $\beta$ has been normalized to have length one:

$$\beta'\Sigma\beta = 1. \tag{4.3}$$

In order to study the asymptotic covariance matrix, we also normalize the slicing regression estimate:

$$\hat{\beta}'\hat{\Sigma}\hat{\beta} = 1, \qquad \hat{\beta}'\hat{\Sigma}\beta > 0, \tag{4.4}$$

for the ignorant slicing regression and

$$\hat{\beta}'\Sigma\hat{\beta} = 1, \qquad \hat{\beta}'\Sigma\beta > 0, \tag{4.4′}$$

for the nonignorant slicing regression. Since we do not know $\beta$, we cannot determine empirically whether the second part of (4.4) or (4.4′) is satisfied or not: we cannot choose between $\hat{\beta}$ and $-\hat{\beta}$. Nevertheless, the distinction between the two solutions is irrelevant for inference about the *direction* of $\beta$; see Section 4.3.

Design condition (DC.1) implies that the conditional covariance matrix for $\mathbf{x}$ given $\mathbf{x}'\beta$ has the form

$$\text{Cov}(\mathbf{x}|\mathbf{x}'\beta) = a(\mathbf{x}'\beta)(\Sigma - \Sigma\beta\beta'\Sigma), \tag{4.5}$$

where $a$ is a scalar function and $E[a(\mathbf{x}'\beta)] = 1$. If the design distribution is normal, $a$ is identically one.

We introduce some more notations:

$$\mathbf{u} = W\mathbf{k} = (u_1, \ldots, u_H)',$$

$$a_h = E\big[a(\mathbf{x}'\beta)\big|y \in s_h\big], \qquad \mathbf{a} = (a_1, \ldots, a_H)', \tag{4.6}$$

$$c_h = E\big[a(\mathbf{x}'\beta)(\mathbf{x} - \mu)'\beta\big|y \in s_h\big], \qquad \mathbf{c} = (c_1, \ldots, c_H)'.$$

The asymptotic distribution for the nonignorant slicing regression is given in Theorem 4.2; the result for the ignorant slicing regression is given in Theorem 4.2′. The proofs are sketched in Appendix C.

THEOREM 4.2. *Assume the general regression model* (1.1), *the design condition* (DC.1), *the normalization* (4.3) *and conditions* (DC.2) *and* (DC.3). *The nonignorant slicing regression estimate, which solves the maximization problem* (3.7′) *and is normalized by* (4.4′), *has the following normal approximation*:

$$\sqrt{n}\,(\hat{\beta} - \beta) \to_{\mathscr{L}} N(\mathbf{0}, A(\Sigma^{-1} - \beta\beta')),$$

*where the scalar A is given by*

$$A = \frac{\sum_{h=1}^{H} a_h u_h^2 / p_h}{(\mathbf{u}'\mathbf{k})^2}. \tag{4.7}$$

THEOREM 4.2′. *Assume the same conditions in Theorem* 4.2. *The ignorant slicing regression estimate, which solves the maximization problem* (3.7) *and is normalized by* (4.4), *has the following normal approximation*:

$$\begin{aligned} &\sqrt{n}\,(\hat{\beta} - \beta) \to_{\mathscr{L}} N(\mathbf{0}, V), \\ &V = S(\Sigma^{-1} - \beta\beta') + T\beta\beta', \end{aligned} \tag{4.8}$$

*where S and T are nonnegative scalars,*

$$S = A + B - 2C, \tag{4.9}$$

*A is the same as in* (4.7), $B = E[a(\mathbf{x}'\beta)((\mathbf{x} - \mu)'\beta)^2]$, $C = \mathbf{u}'\mathbf{c}/\mathbf{u}'\mathbf{k}$ *and*

$$T = \tfrac{1}{4}\operatorname{Var}\left[((\mathbf{x} - \mu)'\beta)^2\right].$$

For inference purposes, we need to estimate the asymptotic covariance matrix. For some situations to be discussed later, we can use the estimated principal eigenvalue $\hat{\lambda}_1$ in (3.8) or (3.8′) to estimate the scalar $A$ in (4.7) or $S$ in (4.9). [The second term on the right-hand side of (4.8) does not affect the inference about the direction of $\beta$, therefore it is not necessary to estimate $T$.] Otherwise, we might need to estimate $\mathbf{p}$, $\mathbf{a}$, $\mathbf{k}$, $B$ and $\mathbf{c}$, in order to estimate $A$ or $S$. We can use the following method of moment estimates:

$$\hat{p}_h = \frac{1}{n}\sum_{i=1}^{n} 1_{ih}, \qquad \hat{k}_h = (\hat{\xi}_h - \bar{\mathbf{x}})'\hat{\beta},$$

$$\hat{a}_h = \frac{1}{d-1}\sum_{i=1}^{n} \frac{1_{ih}(\mathbf{x}_i - \bar{\mathbf{x}})'(\hat{\Sigma}^{-1} - \hat{\beta}\hat{\beta}')(\mathbf{x}_i - \bar{\mathbf{x}})}{n\hat{p}_h}, \tag{4.10}$$

$$\hat{B} = \frac{1}{d-1}\sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})'(\hat{\Sigma}^{-1} - \hat{\beta}\hat{\beta}')(\mathbf{x}_i - \bar{\mathbf{x}})((\mathbf{x}_i - \bar{\mathbf{x}})'\hat{\beta})^2, \tag{4.11}$$

$$\hat{c}_h = \frac{1}{d-1}\sum_{i=1}^{n} \frac{1_{ih}(\mathbf{x}_i - \bar{\mathbf{x}})'(\hat{\Sigma}^{-1} - \hat{\beta}\hat{\beta}')(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'\hat{\beta}}{n\hat{p}_h}. \tag{4.12}$$

The derivations of (4.10)–(4.12) are sketched in Appendix C.

REMARK 4.1. Design condition (DC.1) can be replaced in Theorems 4.2 and 4.2′ by the weaker conditions (DC.1′) and (4.5). Note that (4.5) is equivalent to

$$\frac{\mathrm{Var}(\mathbf{x}'\mathbf{b}|\mathbf{x}'\beta)}{E[\mathrm{Var}(\mathbf{x}'\mathbf{b}|\mathbf{x}'\beta)]} \quad \text{does not depend on } \mathbf{b}. \tag{4.5′}$$

The numerator in (4.5′) is the residual variance when we regress $\mathbf{x}'\mathbf{b}$ on $\mathbf{x}'\beta$. This regression is usually heteroscedastic; condition (4.5′) indicates that the pattern of heteroscedasticity does not depend on $\mathbf{b}$.

REMARK 4.2. The results in Theorems 4.2 and 4.2′ would still hold if we replace $W$ by a consistent estimate $\hat{W}$; see the proof for the theorems in Appendix C.

REMARK 4.3. If we normalize the ignorant slicing regression estimate by (4.4′) instead of (4.4), the asymptotic covariance matrix is given by the first term on the right-hand side of (4.8). For empirical applications, we usually have to normalize by (4.4) instead of (4.4′) because $\Sigma$ is unknown. For inference about the direction of $\beta$, the two asymptotic covariance matrices are equivalent.

4.3. *Inference.* We now apply the results in Section 4.2 to two basic inference problems: testing hypotheses and constructing confidence regions for $\beta$. Since we can only identify the direction of $\beta$, we can only test scale-invariant hypotheses of the form

$$H_0\colon L'\beta = 0, \tag{4.13}$$

where $L$ is a given $d \times q$ matrix of full rank $q \leq d$. Under the null hypothesis (4.13), the terms proportional to $\beta\beta'$ in the asymptotic covariance matrix are annihilated by $L$, therefore Wald's test is given by

$$\frac{n\hat{\beta}'L(L'\hat{\Sigma}^{-1}L)^{-1}L'\hat{\beta}}{\hat{S}} \to_{\mathscr{L}} \chi^2_q \tag{4.14}$$

for the ignorant slicing regression and

$$\frac{n\hat{\beta}'L(L'\Sigma^{-1}L)^{-1}L'\hat{\beta}}{\hat{A}} \to_{\mathscr{L}} \chi^2_q \tag{4.14′}$$

for the nonignorant slicing regression. Wald's test does not depend on the sign of $\hat{\beta}$: we can choose either $\hat{\beta}$ or $-\hat{\beta}$, disregarding the second part of (4.4) or (4.4′).

Wald's test can be inverted to obtain confidence regions. They have to be cone-shaped: If $\beta$ is in the confidence region, any scalar multiple of $\beta$ has to be in the region also. For example, we can test hypotheses of the form

$$H_0\colon \beta \propto \beta_0$$

and take the confidence region to be those $\beta_0$'s for which the previous null

hypothesis is accepted. This leads to the confidence region

$$\left\{\beta: \sin^2(\beta, \hat{\beta}) \leq \frac{\hat{S}\chi^2_{d-1,1-\alpha}}{n}\right\}, \tag{4.15}$$

which has asymptotic confidence level $1 - \alpha$; the sine function is taken with respect to the inner product (2.7′) and $\hat{\beta}$ is the ignorant slicing regression estimate. We can also construct confidence regions for subvectors or linear combinations of $\beta$.

4.4. *Normal design distribution.* The asymptotic theory for the slicing regression is greatly simplified if the design distribution is normal. Under normality, we have

$$a(\mathbf{x}'\beta) \equiv 1, \qquad a_h \equiv 1, \qquad B = 1, \qquad c_h \equiv k_h.$$

It follows that the scalars in Theorems 4.2 and 4.2′ are given by

$$A = \frac{\sum_{h=1}^{H} u_h^2/p_h}{(\mathbf{u}'\mathbf{k})^2}, \qquad S = \frac{\sum_{h=1}^{H} u_h^2/p_h}{(\mathbf{u}'\mathbf{k})^2} - 1.$$

The scalar $S$ for the ignorant slicing regression is smaller than the scalar $A$ for the nonignorant slicing regression. In other words, even when we know the true $\Sigma$, we are better off ignoring this information and using $\hat{\Sigma}$ in the maximization problem (3.7). Note that the maximization problem (3.7) is different from the usual two matrices spectral decomposition: $\hat{\Gamma}$ and $\hat{\Sigma}$ are dependent.

The benefit of ignorance depends on the design distribution and is not universally true. We give a somewhat artificial example in Appendix B for which the knowledge about $\Sigma$ does help. We can expect, though, that the benefit of ignorance will hold for design distributions reasonably close to being normal.

If we use the pps weight matrix,

$$W^{(\mathbf{p})} = D(\mathbf{p}) - \mathbf{p}\mathbf{p}', \tag{4.16}$$

the scalars in Theorems 4.2 and 4.2′ are given by

$$A = \frac{1}{\mathbf{u}'\mathbf{k}} = \frac{1}{\lambda_1}, \qquad S = \frac{1}{\lambda_1} - 1, \tag{4.17}$$

which can be estimated consistently by substituting $\hat{\lambda}_1$ for $\lambda_1$. In other words, when we use the pps weight matrix and the design distribution is normal, we do not need to estimate $\mathbf{a}$, $\mathbf{k}$ and $\mathbf{c}$. We will establish in Section 5 that the pps weight matrix (4.16) is optimal for the normal design distribution.

4.5. *Efficiency.* We consider the efficiency of the slicing regression estimate for an important special case. We assume there exists an unknown

transformation to the standard linear model:

$$t(y) = \alpha + \mathbf{x}'\beta + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2). \tag{4.18}$$

We also assume that the design distribution is normal. We use the ignorant slicing regression based on the pps weight matrix. As our benchmark for comparison, we use the least squares regression of $t(y)$ on $\mathbf{x}$, which gives an efficient estimate for $(\alpha, \beta')$. In order to implement this procedure, we need to have perfect prior information on the transformation $t$. This information is not available in most empirical applications. The slicing regression does not require any knowledge about the transformation $t$.

We now compare the performance for the two methods for estimating the direction of $\beta$. In order to make the comparison, we normalize the least squares slope to satisfy the first part of (4.4). We also assume (4.3). The normalized least squares slope is approximately

$$\beta + (\Sigma^{-1} - \beta\beta')S_{\varepsilon\mathbf{x}} - \tfrac{1}{2}\beta\beta'(\hat{\Sigma} - \Sigma)\beta, \tag{4.19}$$

where $S_{\varepsilon\mathbf{x}}$ is the sample covariance between $\varepsilon$'s and $\mathbf{x}$'s. This is asymptotically normal with mean $\beta$ and asymptotic covariance matrix

$$\frac{1}{n}\left[\sigma^2(\Sigma^{-1} - \beta\beta') + T\beta\beta'\right]. \tag{4.20}$$

We want to compare the scalar $\sigma^2$ in (4.20) with the scalar $S$ in (4.17), which in this special case is given by

$$S = \frac{1 - Q^2 + \sigma^2}{Q^2}, \tag{4.21}$$

where

$$Q^2 = \sum_{h=1}^{H} \frac{p_h\{t_h - E[t(y)]\}^2}{\mathrm{Var}(t(y))}, \qquad t_h = E\left[\, t(y) \middle| y \in s_h\right].$$

$Q^2$ is the proportion of the variance of $t(y)$ explained by the discretized $t(y)$'s, $(t_1, \ldots, t_H)$.

For estimating the direction of $\beta$, the efficiency for the ignorant slicing regression is

$$\text{efficiency} = \frac{\sigma^2 Q^2}{1 - Q^2 + \sigma^2} = \frac{(1 - R^2)Q^2}{1 - Q^2R^2}, \tag{4.22}$$

where $R^2$ is the usual $R^2$ for the linear model (4.18).

Unless $R^2$ is very close to one, the efficiency in (4.22) is usually fairly high, even when the number of slices is fairly small. For example, we assume $R^2 = 0.30$, a value typical of social science research. Assume for now that we use equal size slices: the partition is chosen so that the probability for $y$ to fall inside each slice is equal. For three slices, we have $Q^2 = 0.79$ and efficiency = 0.72. For ten slices, we have $Q^2 = 0.96$ and efficiency = 0.94. Ten slices are probably good enough for most purposes.

REMARK 4.4. It is possible to improve upon the equal size slices. For three slices, the optimal partition yields $Q^2 = 0.81$ and efficiency $= 0.75$ when $R^2 = 0.30$.

**5. Optimal weight matrix.** So far we have left open the choice of the weight matrix $W$. We now derive the weight matrix $W$ which minimizes the scalar $S$ or $A$ in the asymptotic covariance matrix. The pps weight matrix (4.16) appears to be reasonable and also guarantees that condition (DC.3) can be replaced by the weaker condition (DC.3′). We will establish that it is indeed optimal for the normal design distribution. For other elliptically symmetric design distributions, (4.16) might not be optimal; the optimal weight matrix in essence imposes a heterogeneity correction which is not necessary for the normal design distribution. The optimal weight matrix might depend on some unknown quantities, which can be estimated from the data in empirical applications.

The scalars $A$ and $S$ depend on $W$ only through

$$\mathbf{u} = W\mathbf{k}. \tag{5.1}$$

We will refer to $\mathbf{u}$ as the *scoring rule*. Two weight matrices with the same scoring rule give asymptotically equivalent slicing regression estimates. We shall find the optimal scoring rule under the constraints

$$\mathbf{u}'\mathbf{1} = 0, \qquad \mathbf{u}'\mathbf{k} \neq 0, \tag{5.2}$$

then find a corresponding weight matrix $W$ which satisfies (5.1).

Given a scoring rule $\mathbf{u}$, we can always find a representation of the form

$$\begin{gathered} W \propto [D(\mathbf{r}) - \mathbf{r}\mathbf{r}'] + \mathbf{q}\mathbf{q}'; \qquad \mathbf{r}'\mathbf{1} = 1; \qquad \mathbf{q}'\mathbf{1} = 0; \\ r_h \geq 0, \qquad h = 1, \ldots, H, \end{gathered} \tag{5.3}$$

which satisfies (5.1). The corresponding $\hat{\Gamma}$ can be interpreted as a weighted covariance matrix with a location correction:

$$\hat{\Gamma} = \sum_{h=1}^{H} r_h(\hat{\xi}_h - \tilde{\xi})(\hat{\xi}_h - \tilde{\xi})', \tag{5.4}$$

where $\tilde{\xi}$ is the weighted sample average $\tilde{\xi} = \sum_{h=1}^{H}(r_h + q_h)\hat{\xi}_h$.

We give the optimal scoring rule $\mathbf{u}$ in Theorems 5.1 and 5.1′, respectively, for the nonignorant and ignorant slicing regression estimates. The proofs are straightforward applications of the Lagrange multipliers method, with the constraints $\mathbf{u}'\mathbf{1} = 0$ and $\mathbf{u}'\mathbf{k} = 1$.

We introduce some new notations:

$$\tilde{p}_h = \frac{p_h/a_h}{\sum_{j=1}^{H} p_j/a_j}, \qquad \tilde{\mathbf{p}} = (\tilde{p}_1, \ldots, \tilde{p}_H). \tag{5.5}$$

Since $\tilde{\mathbf{p}}$ is a probability measure on the index set $\{1, \ldots, H\}$, we have moment

operators such as

$$E_{\tilde{\mathbf{p}}}(\mathbf{k}) = \sum_{h=1}^{H} \tilde{p}_h k_h, \qquad \sigma^2_{\tilde{\mathbf{p}}}(\mathbf{k}) = \sum_{h=1}^{H} \tilde{p}_h \big(k_h - E_{\tilde{\mathbf{p}}}(\mathbf{k})\big)^2.$$

THEOREM 5.1. *Under the same assumptions in Theorem* 4.2, *the optimal nonignorant slicing regression estimate, which minimizes the scalar* $A$ *in* (4.7), *is based on the scoring rule*

$$u_h = \frac{p_h\big(k_h - E_{\tilde{\mathbf{p}}}(\mathbf{k})\big)}{a_h}. \tag{5.6}$$

*The minimized scalar is*

$$A = \frac{1}{\mathbf{u}'\mathbf{k}} = \frac{E_{\tilde{\mathbf{p}}}(\mathbf{a})}{\sigma^2_{\tilde{\mathbf{p}}}(\mathbf{k})} = \frac{1}{\lambda_1}, \tag{5.7}$$

*which can be estimated consistently by the reciprocal of* $\hat{\lambda}_1$ *in* (3.8′).

THEOREM 5.1′. *Under the same assumptions in Theorem* 4.2′, *the optimal ignorant slicing regression estimate, which minimizes the scalar* $S$ *in* (4.9), *is based on the scoring rule*

$$u_h = \frac{p_h\big[\big(c_h - E_{\tilde{\mathbf{p}}}(\mathbf{c})\big) + \nu\big(k_h - E_{\tilde{\mathbf{p}}}(\mathbf{k})\big)\big]}{a_h}, \tag{5.6′}$$

*where* $\nu = [E_{\tilde{\mathbf{p}}}(\mathbf{a}) - \sigma_{\tilde{\mathbf{p}}}(\mathbf{c}, \mathbf{k})]/\sigma^2_{\tilde{\mathbf{p}}}(\mathbf{k})$. *The minimized scalar is*

$$S = B + \frac{\nu^2\sigma^2_{\tilde{\mathbf{p}}}(\mathbf{k}) - \sigma^2_{\tilde{\mathbf{p}}}(\mathbf{c})}{E_{\tilde{\mathbf{p}}}(\mathbf{a})}. \tag{5.7′}$$

For normal design distributions, the optimal scoring rules (5.6) and (5.6′) are both given by $u_h = p_h k_h$, which can be represented by the pps weight matrix $W^{(\mathbf{p})}$. In other words, the pps weight matrix is optimal for the slicing regression when the design distribution is normal. The minimized scalars $A$ and $S$ are given by (4.17).

When the design distribution is nonnormal, we need to find suitable representations for the optimal weight matrices corresponding to the optimal scoring rules. The optimal scoring rule (5.6) can be represented by $W^{(\tilde{\mathbf{p}})}$, the weight matrix of form (3.9) with $\mathbf{r} = \tilde{\mathbf{p}}$. This optimal weight matrix can be interpreted as the pps weight matrix with a heterogeneity correction: each slice is weighted by $p_h/a_h$ instead of $p_h$, thus slices with less dispersion (smaller $a_h$) are weighted heavier. $\hat{\Gamma}$ can be interpreted as a weighted covariance matrix for the data vectors $\{\hat{\xi}_1, \ldots, \hat{\xi}_H\}$, with each slice weighted by $\tilde{p}_h$.

For the optimal scoring rule (5.6′), there does not appear to be a closed form representation of form (3.9), therefore we will use a representation of form

(5.3). Taking $\mathbf{r} = \tilde{\mathbf{p}}$, we have

$$q_h \propto \frac{p_h(c_h - E_{\tilde{\mathbf{p}}}(\mathbf{c}))}{a_h}. \tag{5.8}$$

The optimal weight matrix can be interpreted as $W^{(\tilde{\mathbf{p}})}$ with a location correction. Instead of centering $\hat{\xi}_h$'s by $\bar{\xi} = \sum_{h=1}^{H} \tilde{p}_h \hat{\xi}_h$ to derive the estimate $\hat{\Gamma}$, we center them by $\tilde{\xi} = \bar{\xi} + \sum_{h=1}^{H} q_h \hat{\xi}_h$ as in (5.4). The location correction can be interpreted as follows. We approximate the parameter $c_h$ roughly by

$$\begin{aligned} c_h &= E\left[a(\mathbf{x}'\beta)(\mathbf{x} - \mu)'\beta \,|\, y \in s_h\right] \\ &\approx E\left[a(\mathbf{x}'\beta) \,|\, y \in s_h\right] E\left[(\mathbf{x} - \mu)'\beta \,|\, y \in s_h\right] = a_h k_h, \end{aligned}$$

thus the optimal scoring rule is approximated by

$$u_h \approx p_h k_h + \frac{\nu p_h\left[k_h - E_{\tilde{\mathbf{p}}}(\mathbf{k})\right]}{a_h}, \tag{5.9}$$

$$\nu \approx \frac{E_{\tilde{\mathbf{p}}}(\mathbf{a})\left[1 - \sigma^2_{\tilde{\mathbf{p}}}(\mathbf{k})\right]}{\sigma^2_{\tilde{\mathbf{p}}}(\mathbf{k})} \geq 0.$$

The first term on the right-hand side of (5.9) is the scoring rule for the pps weight matrix, $W^{(\mathbf{p})}$. The second term is the scoring rule for $W^{(\tilde{\mathbf{p}})}$, the pps weight matrix corrected for heterogeneity. Therefore the optimal scoring rule (5.6′) is roughly a convex compromise between the two.

REMARK 5.1. Design condition (DC.1) can be replaced by the weaker conditions (DC.1′) and (4.5) in Theorems 5.1 and 5.1′.

REMARK 5.2. In order to implement the optimal slicing regression when the design distribution is nonnormal, we might have to estimate nuisance parameters $(\mathbf{a}, \mathbf{k}, \mathbf{c})$, which depend on $\beta$. We can use the pps weight matrix to obtain an initial estimate for $\beta$, then estimate the nuisance parameters from this initial estimate. We can then estimate the optimal scoring rule and the optimal weight matrix and reestimate the direction of $\beta$ using the estimated optimal weights (cf. Remark 4.2).

Prior to carrying out the reestimation, we can estimate the optimal scalar (5.7) or (5.7′) and compare it with the estimated scalar for the original weight matrix. The ratio between the two scalars can be used to estimate the potential improvement in efficiency. If the potential improvement is small, the reestimation might not be worthwhile. This is likely to be true if the heterogeneity is moderate, i.e., $a_h$'s are close to each other. We give an extreme example in Appendix B to demonstrate that the weight adjustment can result in a substantial improvement if the heterogeneity is severe.

**6. Violation of elliptical symmetry.** We have assumed until now that the design distribution is elliptically symmetric. It is natural to ask whether

the slicing regression still provides a good estimate for the direction of $\beta$ when the elliptical symmetry is violated. We now establish a bias bound for the population case.

Let $\tilde{\beta}$ be a solution to the maximization problem (2.6) and $\lambda = L(\tilde{\beta})$ be the maximum. They are, respectively, the principal eigenvector and eigenvalue for the spectral decomposition of $\Gamma$ with respect to the inner product (2.7). $\tilde{\beta}$ is the population version of the slicing regression estimate for the direction of $\beta$. When the design distribution is not elliptically symmetric, $\tilde{\beta}$ might not be collinear with $\beta$. We measure the noncollinearity between $\beta$ and $\tilde{\beta}$ by $\sin^2(\beta, \tilde{\beta})$, where the sine function is taken with respect to the inner product (2.7).

We now consider another spectral decomposition. Let $\Lambda = \mathrm{Cov}[E(\mathbf{x}|\mathbf{x}'\beta)]$. We take the spectral decomposition of $\Lambda$ with respect to the inner product (2.7). The principal eigenvector is $\beta$; the principal eigenvalue is one. Let $\tau$ be the second eigenvalue. We can interpret $\tau$ as a measure of the deviation from elliptical symmetry. Under elliptical symmetry, $E(\mathbf{x}|\mathbf{x}'\beta)$ falls along the line (2.4), therefore $\tau = 0$. When elliptical symmetry is violated, $E(\mathbf{x}|\mathbf{x}'\beta)$ is a curve which meanders around (2.4) and $\tau$ measures the largest mean squared deviation from (2.4).

The comparison between $\lambda$ and $\tau$ gives the following theorem.

THEOREM 6.1. *Assume the general regression model* (1.1). *Assume the regressor* $\mathbf{x}$ *is sampled randomly from a probability distribution which might not be elliptically symmetric. Let* $\tilde{\beta}$ *be a solution to the maximization problem* (2.6). *The noncollinearity between* $\beta$ *and* $\tilde{\beta}$ *satisfies the following bound*:

$$\sin^2(\beta, \tilde{\beta}) \leq \frac{\tau/(1-\tau)}{\lambda/(1-\lambda)}. \tag{6.1}$$

The proof is sketched in Appendix C.

If the design distribution is elliptically symmetric, we have $\tau = 0$, thus the right-hand side of (6.1) is zero, i.e., the slicing regression is Fisher consistent (cf. Corollary 2.2). If $\lambda = 1$, the bound is again zero, thus we have Fisher consistency even though the design distribution might not be elliptically symmetric. In order for $\lambda = 1$, we must have $y = g(\mathbf{x}'\beta)$, where $g$ is invertible (cf. Remark 2.1).

If the right-and side of (6.1) is close to zero, the slicing regression would be nearly Fisher consistent for the direction of $\beta$. This is true if the design distribution is nearly elliptically symmetric ($\tau \approx 0$) or if $\lambda$ is close to one.

For the sampling case, the bias bound (6.1) can be estimated from observed data. Note that both $\Lambda$ and $\tau$ depend on $\beta$. If we have a consistent initial estimate for the direction of $\beta$, we can then estimate $T$ using this initial estimate and carry out the spectral decomposition to estimate $\tau$. If such an initial estimate is not available, we can replace $\tau$ in (6.1) by

$$\tau^{\sup} = \sup_{\beta \in R^d} \tau(\beta).$$

In order to estimate $\tau^{\sup}$, we need to maximize the estimate $\hat{\tau}(\beta)$ over $\beta$. This is a projection-pursuit problem, with $\hat{\tau}(\beta)$ as the projection index. Huber [(1985) and discussions] gave a comprehensive review of the projection pursuit problem. Cox (1985) suggested a projection index, the maximum curvature in Cox and Small (1978), which is analogous to $\tau(\beta)$.

**7. A simulation study.** We have conducted a simulation study to demonstrate the performance for the slicing regression estimate. We consider two general regression models:

$$y = \mathbf{x}'\beta + \varepsilon, \tag{7.1}$$

$$y = 0.1(\mathbf{x}'\beta + \varepsilon)^3, \tag{7.2}$$

where $\beta = (1, 1, 1, 0, 0, 0)'$, $\mathbf{x} \sim N(\mathbf{0}, I_6)$ and $\varepsilon \sim N(0, 1)$. We generate samples of size $n = 100$ each for each model, then estimate the direction of $\beta$ by the ignorant slicing regression estimate $\hat{\beta}$, normalized by (4.4'); cf. Remark 4.3. In order to study the sensitivity of $\hat{\beta}$ to changes in the number of slices, $H$, we take $H = 6, 10, 20$. For each $H$, the grid points are equally spaced between $-3$ and 3. In other words, the first slice is $y \leq -3$, the last slice is $y > 3$ and there are $H - 2$ slices in between. We use a thousand replicates to estimate the expectation and the standard deviation for each component of $\hat{\beta}$. We also estimate the total variance for $\hat{\beta}$, i.e., the trace of the covariance matrix for $\hat{\beta}$.

The results of the simulation study are given in Tables 1 and 2. For both models, $E(\hat{\beta})$ is very close to $(0.577, 0.577, 0.577, 0, 0, 0)'$, the true slope vector $\beta$ normalized to have length one. The estimate is insensitive to changes in the number of slices. For each model, the total variance varies by less than twenty percent when $H$ changes from 6 to 20. This suggests that the choice of $H$ for the slicing regression problem might not be as crucial as the choice of the smoothing parameter for the typical nonparametric regression or density estimation problems.

We also report the performance of the least squares estimate (after normalization) for comparison. For model (7.1), the least squares estimate is the

TABLE 7.1

*Expectation and standard deviation (in parentheses) of* $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_6)'$ *for* (7.1), $n = 100$. *The last row is the least squares estimate after normalization*

| $H$ | $\hat{\beta}_1$ | $\hat{\beta}_2$ | $\hat{\beta}_3$ | $\hat{\beta}_4$ | $\hat{\beta}_5$ | $\hat{\beta}_6$ | Total variance |
|---|---|---|---|---|---|---|---|
| 6 | 0.571 | 0.572 | 0.569 | 0.000 | 0.000 | −0.004 | 0.0232 |
| | (0.056) | (0.055) | (0.057) | (0.069) | (0.069) | (0.066) | |
| 10 | 0.570 | 0.572 | 0.570 | 0.001 | −0.001 | −0.004 | 0.0223 |
| | (0.054) | (0.054) | (0.056) | (0.068) | (0.067) | (0.065) | |
| 20 | 0.569 | 0.571 | 0.569 | 0.001 | 0.001 | −0.004 | 0.0263 |
| | (0.059) | (0.060) | (0.062) | (0.072) | (0.071) | (0.072) | |
| Least squares | 0.572 | 0.573 | 0.571 | 0.002 | −0.001 | −0.003 | 0.0179 |
| | (0.048) | (0.048) | (0.050) | (0.060) | (0.060) | (0.060) | |

TABLE 7.2
*Expectation and standard deviation (in parentheses) of* $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_6)'$ *for* (7.2), $n = 100$. *The last row is the least squares estimate after normalization*

| $H$ | $\hat{\beta}_1$ | $\hat{\beta}_2$ | $\hat{\beta}_3$ | $\hat{\beta}_4$ | $\hat{\beta}_5$ | $\hat{\beta}_6$ | Total variance |
|---|---|---|---|---|---|---|---|
| 6 | 0.568 | 0.570 | 0.570 | 0.000 | 0.003 | −0.004 | 0.0284 |
| | (0.062) | (0.062) | (0.064) | (0.076) | (0.075) | (0.072) | |
| 10 | 0.570 | 0.570 | 0.570 | −0.001 | 0.002 | −0.004 | 0.0268 |
| | (0.060) | (0.060) | (0.062) | (0.073) | (0.073) | (0.071) | |
| 20 | 0.569 | 0.569 | 0.570 | −0.001 | 0.002 | −0.004 | 0.0273 |
| | (0.060) | (0.062) | (0.063) | (0.074) | (0.074) | (0.071) | |
| Least squares | 0.561 | 0.560 | 0.562 | 0.003 | 0.001 | −0.004 | 0.0554 |
| | (0.082) | (0.085) | (0.086) | (0.107) | (0.104) | (0.108) | |

maximum likelihood estimate, therefore it outperforms the slicing regression estimate. However, the slicing regression estimate has a reasonably good relative efficiency, about eighty percent (0.0179/0.0223) for $H = 10$. For model (7.2), the slicing regression estimate is about twice as efficient as the least squares estimate. It is also interesting to observe that the performance of the slicing regression estimate is roughly the same under the two models. This further confirms that the slicing regression estimate is insensitive to the choice of the slices; for the slicing regression, (7.1) differs from (7.2) only in transforming the grid points which determine the slices.

It is perhaps unfair to compare the slicing regression estimate with the least squares estimate for model (7.2). After examining the residuals, we probably will use a Box–Cox transformation model, which might do a better job than the least squares regression without a transformation. However, the improvement in efficiency is at most thirty-three percent ($1 - 0.0179/0.0266$) for $H = 10$, because the transformation model estimate cannot perform better than the least squares regression based on the correct transformation.

The example chosen here is fairly unfavorable for the slicing regression because $R^2$ is rather high: $R^2 = 0.75$ for model (7.1). If $R^2$ is lower, e.g., if $R^2 = 0.30$, the relative efficiency for the slicing regression would be higher than what is shown here; see Section 4.5. Finally, the simplicity of the slicing regression suggests itself as a good initial estimate if one wishes to pursue adaptive estimation.

## APPENDIX A

**Identifiability.** We noted in Section 1 that the most we can identify in the parameter vector $(\alpha, \beta')$ is the direction of $\beta$. We now discuss whether the direction itself can be identified.

A.1. *Identification using inverse regression.* If the scalar function $\kappa(y)$ in (2.3) is not identically zero, we can use (2.5) or Corollary 2.2 to identify the

direction of $\beta$. If we know further that $\kappa$ is monotonic, we can use Theorem 4.1 and (DC.3') to verify the consistency of the slicing regression estimate. The following theorem establishes these properties for two rich classes of general regression models.

THEOREM A.1. *If the conditional distribution of y given* $\mathbf{x}'\beta$ *is stochastically monotonic in* $\mathbf{x}'\beta$, *i.e., if* $P(Y \leq y|\mathbf{x}'\beta)$ *is monotonic in* $\mathbf{x}'\beta$, *then* $\kappa(y) \not\equiv 0$. *If the conditional distribution of y given* $\mathbf{x}'\beta$ *has monotone likelihood ratio, then* $\kappa(y)$ *is monotonic in y.*

If we impose the stochastic monotonicity condition on the true model, we can then identify the direction of $\beta$ using the inverse regression function. The class of general regression models which satisfy this condition is very rich and includes location families, monotonic transformation families and natural exponential families. Similarly, if we impose the monotone likelihood ratio condition, we can use the slicing regression to identify the direction of $\beta$.

We now give an example to demonstrate that the monotonicity conditions are neither necessary nor redundant:

EXAMPLE A.1. We consider a class of heterogeneous models of the form

$$y = \varepsilon g(\mathbf{x}'\beta), \quad \text{where} \quad E(\varepsilon|\mathbf{x}) = 0.$$

The nonparametric forward regression E(y|**x**) is identically zero and provides no information about $\beta$; y depends on $\beta$ through heteroscedasticity instead of through the mean. The inverse regression might or might not be informative about $\beta$. If g is nonnegative and increasing in $\mathbf{x}'\beta$, then $\kappa$(y) is increasing in |y| and therefore cannot be identically zero: larger |y|'s are more likely to come from larger $\mathbf{x}'\beta$'s. If g is symmetric about zero and the distributions for **x** and $\varepsilon$ are both symmetric about zero, then $\kappa$(y) is identically zero.

Theorem A.1 follows from the following lemma about Bayesian estimation, which might be of interest in itself.

LEMMA A.2. *Let* $\mathscr{K} = \{K_\theta(y)\}$ *be a one parameter family of sampling distributions, parametrized by* $\theta$. *Assume that* $\mathscr{K}$ *has densities* $\{k_\theta(y)\}$. *Assume that* $\theta$ *follows a prior distribution* $\Pi(\theta)$. *Let* $\hat{\theta}(y) = E(\theta|y)$ *be the posterior expectation and* $\Pi(\theta|y)$ *be the posterior distribution.*

(i) *If* $\mathscr{K}$ *is stochastically monotonic in* $\theta$, *then* $\hat{\theta}$ *cannot be a constant.*

(ii) *The following three conditions are equivalent*: (a) *The sampling distributions in* $\mathscr{K}$ *have monotone likelihood ratio.* (b) *The posterior distributions* $\Pi(\theta|y)$, *as a one parameter family parametrized by y, have monotone likelihood ratio.* (c) *The posterior expectation* $\hat{\theta}(y)$ *is monotonic in y for all* $\Pi(\theta)$.

Theorem A.1 follows immediately from Lemma A.2: we treat $\mathbf{x}'\beta$ as the parameter $\theta$ and treat the conditional distribution of $y$ given $\mathbf{x}'\beta$ as the sampling distributions $\mathscr{K}$. The proof of Lemma A.2 is sketched in Appendix C.

A.2. *Identification using forward regression.* Li and Duan (1989) established results similar to Theorem 2.1 for a rich class of parametric forward regressions. Let $L(\theta, y)$ be a criterion function such as the negative of a log-likelihood function. Let the regression estimate $(\hat{\alpha}, \hat{\beta}')$ be a solution to the minimization problem

$$\min_{(a, \mathbf{b}')} \sum_{i=1}^{n} L(a + \mathbf{x}_i'\mathbf{b}, y_i).$$

The population version of this estimate is the solution $(\alpha^*, \beta^{*\prime})$ for the minimization problem

$$\min_{(a, \mathbf{b}')} E[L(a + \mathbf{x}'\mathbf{b}, y)], \tag{A.1}$$

where the expectation is taken over the joint distribution of $y$ and $\mathbf{x}$. Li and Duan [(1989), Theorem 2.1] have shown that

$$\beta^* \propto \beta \tag{A.2}$$

under design condition (DC.1) and a convexity condition on $L$. If the proportionality constant in (A.2) is nonzero, $\beta^*$ identifies the direction of $\beta$. The following theorem compares identification using forward and inverse regressions.

THEOREM A.3. *Assume the general regression model* (1.1) *and the design condition* (DC.1). *Let* $L(\theta, y)$ *be a criterion function which is convex in* $\theta$ *for all* $y$. *Assume that the minimization problem* (A.1) *has a unique solution* $(\alpha^*, \beta^{*\prime})$. *If* $\kappa(y) \equiv 0$, *then* $\beta^*$ *is null.*

PROOF. Using the result in Li and Duan [(1989), Theorem 2.1], we need only minimize the following expectation over $(a, c)$:

$$E[L(a + c\mathbf{x}'\beta, y)] = E[E\{L(a + c\mathbf{x}'\beta, y)|y\}]; \tag{A.3}$$

the regression slope is then given by $\beta^* = \beta c^*$, where $(a^*, c^*)$ minimizes (A.3). By the convexity of $L$ and Jensen's inequality, the right-hand side of (A.3) is bounded from below by $E[L(a + cE(\mathbf{x}'\beta|y), y)]$. If $\kappa(y) \equiv 0$, the minimization problem reduces to minimizing $E[L(a, y)]$ over $a$, i.e., the optimal value for $c$ is zero. □

According to the theorem, if any parametric forward regression based on (A.1) identifies the direction of $\beta$, then $\kappa(y) \not\equiv 0$; it follows that the inverse regression would also identify the direction of $\beta$. In other words, the inverse regression is at least as effective as any parametric forward regression based on (A.1) for identifying the direction of $\beta$.

Design condition (DC.1) is used in Theorem A.3 only to reduce (A.1) to the minimization of (A.3), not in the subsequent derivations. For the special case

of a simple regression, we have the following result which might be interesting in its own right.

OBSERVATION A.4. Assume that $(y, x)$ follows an arbitrary probability distribution, where $y$ and $x$ are both scalars. Let $L(\theta, y)$ be a criterion function which is convex in $\theta$ for all $y$. If $E(x|y)$ is a constant, then the parametric forward regression of $y$ on $x$ based on

$$\min_{(a,b)} E[L(a + xb, y)]$$

has zero slope, i.e., the optimal value for $b$ is zero.

If inverse regression fails to reveal any relationship between $y$ and $x$, parametric forward regressions based on convex criterion functions cannot either. When the joint distribution for $(y, x)$ is normal, this is a well-known fact: If $E(x|y)$ is a constant, then $E(y|x)$ is also a constant; actually, $y$ and $x$ are independent in this case.

## APPENDIX B

**Weight adjustment.** For nonnormal design distributions, the pps matrix might not be optimal. If the design distribution is nearly homogeneous, that is, $a_h$'s are close to each other, we would expect the pps weight matrix to have good efficiency relative to the optimal weight matrix. However, the pps weight matrix might be very inefficient if the design distribution is highly heterogeneous.

We now construct a rather extreme example with severe heterogeneity. We take $\mathbf{x}$ to be two-dimensional, with the design distribution being uniform on the circle centered at $\mathbf{0}$ with radius $\sqrt{2}$, thus $\mathrm{Cov}(\mathbf{x}) = I$. Let $\theta$ be the angle on the circle. We have

$$\theta \sim U(0, 2\pi),$$

$$x_1 = \sqrt{2}\cos(\theta), \qquad x_2 = \sqrt{2}\sin(\theta).$$

Let the slope vector be $\beta = (1, 0)'$ and the model be $y = \mathbf{x}'\beta = x_1$. We divide the range of $y$ into four slices, using the partition $(-\sqrt{2}\cos(\delta), 0, \sqrt{2}\cos(\delta))$, where $\delta$ is a small positive constant. The dispersion of $\mathbf{x}$ in the two extreme slices are therefore very small.

We compare five slicing regression estimates for this example.

1. $\hat{\beta}^{\mathrm{opt}}$ is the optimal ignorant slicing regression based on the scoring rule (5.6)′).
2. $\hat{\beta}^{\tilde{\mathbf{p}}}$ is the ignorant slicing regression based on the scoring rule (5.6).
3. $\hat{\beta}^{\mathbf{p}}$ is the ignorant slicing regression based on the pps weight matrix.
4. $\tilde{\beta}^{\mathrm{opt}}$ is the optimal nonignorant slicing regression based on the scoring rule (5.6).

5. $\tilde{\beta}^{\mathbf{p}}$ is the nonignorant slicing regression based on the pps weight matrix. For $\delta$ near zero, the scalars $S$ or $A$ for these estimates are approximately

$$S^{\text{opt}} \doteq \frac{1}{2} - \frac{32}{9\pi^2} \approx 0.140, \qquad S^{\tilde{\mathbf{p}}} \doteq \frac{1}{2}, \qquad S^{\mathbf{p}} \doteq \frac{\pi^2}{8} - \frac{5}{6} \approx 0.401,$$

$$A^{\text{opt}} \doteq \frac{\pi}{6}\delta \approx 0.524\delta, \qquad A^{\mathbf{p}} \doteq \frac{\pi^2}{8} \approx 1.234.$$

For $\delta$ sufficiently small, $A^{\text{opt}}$ is smaller than $S^{\text{opt}}$, thus the benefit of ignorance does not hold for this example. When $\Sigma$ is unknown, $\hat{\beta}^{\text{opt}}$ outperforms the other two ignorant slicing regressions substantially: the relative efficiency of $\hat{\beta}^{\tilde{\mathbf{p}}}$ is about 0.28, while the relative efficiency of $\hat{\beta}^{\mathbf{p}}$ is about 0.35.

## APPENDIX C

### Technical proofs.

PROOF OF THEOREM 4.1. Since $\hat{\Gamma}$ converges almost surely to a matrix proportional to $\Gamma$, both $\hat{L}(\mathbf{b})$ and $\tilde{L}(\mathbf{b})$ converge to a criterion function proportional to $L(\mathbf{b})$ in (2.6). The convergence is uniform in $\mathbf{b}$. The rest of the theorem follows from Corollary 2.2. □

PROOF OF THEOREM 4.2 AND THEOREM 4.2′. Without loss of generality, assume that $\mathbf{u}'\mathbf{k} = 1$. We approximate $\hat{\Gamma}$ by

$$\hat{\Gamma} \doteq \mathbf{u}'\mathbf{k}\Sigma\beta\beta'\Sigma + (\hat{\xi} - \xi)\mathbf{u}\beta'\Sigma + \Sigma\beta\mathbf{u}'(\hat{\xi} - \xi)' \doteq (\Sigma\beta + \Delta)(\Sigma\beta + \Delta)',$$

where $\Delta = (\hat{\xi} - \xi)\mathbf{u}$. The nonignorant slicing regression maximizes

$$\tilde{L}(\mathbf{b}) \doteq \frac{[\mathbf{b}'(\Sigma\beta + \Delta)]^2}{\mathbf{b}'\Sigma\mathbf{b}}. \tag{C.1}$$

The nonignorant slicing regression estimate, normalized by (4.4′), is approximated by

$$\hat{\beta} \doteq \beta + (\Sigma^{-1} - \beta\beta')\Delta. \tag{C.2}$$

The right-hand side is asymptotically normal with mean $\beta$. The asymptotic covariance matrix is given by

$$\text{Cov}(\hat{\beta}) = (\Sigma^{-1} - \beta\beta')\text{Cov}(\Delta)(\Sigma^{-1} - \beta\beta').$$

Using the approximation $\hat{\xi}_h - \xi_h \doteq \sum_{i=1}^{n} 1_{ih}(\mathbf{x}_i - \xi_h)/np_h$, it is straightforward to derive the asymptotic covariance matrix for $\Delta$ and verify (4.7) and Theorem 4.2.

For the ignorant slicing regression, the denominator in (C.1) is based on $\hat{\Sigma}$ instead of $\Sigma$. Using the approximation

$$\hat{\Sigma}^{-1} \doteq \Sigma^{-1} - \Sigma^{-1}(\hat{\Sigma} - \Sigma)\Sigma^{-1},$$

the ignorant slicing regression estimate, normalized by (4.4), is approximated by

$$\text{(C.2')} \qquad \hat{\beta} \doteq \beta + (\Sigma^{-1} - \beta\beta')\Delta - (\Sigma^{-1} - \beta\beta'/2)(\hat{\Sigma} - \Sigma)\beta.$$

The covariance matrix for $\hat{\Sigma}\beta$ gives the $B$ term and the $T$ term in (4.8) and (4.9). The covariance between $\Delta$ and $\hat{\Sigma}\beta$ gives the $C$ term.

For Remark 4.2, note that if we replace $W$ by $\hat{W}$, $\mathbf{k}$ by $\hat{\mathbf{k}}$ and $\mathbf{u}$ by $\hat{\mathbf{u}}$, the same approximation for $\hat{\Gamma}$ holds with $\Delta$ replaced by $\hat{\Delta} = (\hat{\xi} - \xi)\hat{\mathbf{u}}$. Since $\hat{\Delta}$ differs from $\Delta$ by a lower order term, the results in Theorems 4.2 and 4.2' remain the same. □

DERIVATION OF (4.10)–(4.12). All three estimates follow from

$$E\big[(\mathbf{x} - \mu)(\Sigma^{-1} - \beta\beta')(\mathbf{x} - \mu)\big|\mathbf{x}'\beta\big] = (d - 1)a(\mathbf{x}'\beta).$$

(4.10) follows from taking the expectation on both sides of this equation over $\mathbf{x}'\beta$ conditioned on $y \in s_h$. (4.11) follows from multiplying both sides by $(\beta(\mathbf{x} - \mu))^2$, then taking expectation; (4.12) follows from multiplying both sides by $\beta(\mathbf{x} - \mu)$, then taking the conditional expectation.

PROOF OF THEOREM 6.1. Without loss of generality, assume $E(\mathbf{x}) = \mathbf{0}$, $\mathrm{Cov}(\mathbf{x}) = \mathbf{I}$, $\beta'\beta = 1$, $\tilde{\beta}'\tilde{\beta} = 1$. Let $\theta$ be the angle between $\beta$ and $\tilde{\beta}$. We decompose $\tilde{\beta}$ as follows:

$$\tilde{\beta} = \cos(\theta)\beta + \sin(\theta)\,\delta, \quad \text{where } \delta'\beta = 0,\ \delta'\delta = 1.$$

For any $\mathbf{b} \in R^d$, we have the inequality

$$\text{(C.3)} \qquad \lambda\big(\mathbf{b}'\tilde{\beta}\big)^2 \le \mathbf{b}'\Gamma\mathbf{b} \le \mathbf{b}'\Lambda\mathbf{b};$$

the second inequality follows from

$$\mathrm{Var}(\mathbf{x}'\mathbf{b}|\mathbf{x}'\beta) = \mathrm{Var}(\mathbf{x}'\mathbf{b}|\mathbf{x}'\beta, \varepsilon) \le \mathrm{Var}(\mathbf{x}'\mathbf{b}|y).$$

Taking $\mathbf{b} = \cos(\theta)\beta + (\sin(\theta)/\pi)\delta$, we have $\mathbf{b}'\tilde{\beta} = \cos^2(\theta) + \sin^2(\theta)/\tau$ and

$$\begin{aligned}\mathbf{b}'\Lambda\mathbf{b} &= \mathrm{Var}\big[E(\mathbf{x}'\mathbf{b}|\mathbf{x}'\beta)\big] \\ &= \cos^2(\theta) + (\sin^2(\theta)/\tau^2)\mathrm{Var}\big[E(\mathbf{x}'\delta|\mathbf{x}'\beta)\big] \\ &\quad + (2\sin(\theta)\cos(\theta)/\tau)\mathrm{Cov}\big[\mathbf{x}'\beta, E(\mathbf{x}'\delta|\mathbf{x}'\beta)\big].\end{aligned}$$

Since $\mathrm{Var}[E(\mathbf{x}'\delta|\mathbf{x}'\beta)] \le \tau$ and $\mathrm{Cov}[\mathbf{x}'\beta, E(\mathbf{x}'\delta|\mathbf{x}'\beta)] = 0$, we have $\mathbf{b}'\Lambda\mathbf{b} \le \cos^2(\theta) + \sin^2(\theta)/\tau$. It follows from (C.3) that $\lambda(\cos^2(\theta) + \sin^2(\theta)/\tau) \le 1$, which proves the theorem. □

PROOF OF LEMMA A.2. Without loss of generality, assume $E(\theta) = 0$. Assume that $\hat{\theta} \equiv 0$, i.e., $\int \theta k_\theta(y)\, d\Pi(\theta) \equiv 0$. This is equivalent to $\int \theta K_\theta(y)\, d\Pi(\theta) \equiv 0$. Since $E(\theta) = 0$, the integral in the last identity is the covariance between $\theta$ and $K_\theta(y)$, where $y$ is treated as a constant. If $\mathscr{K}$ is stochastically mono-

tonic, $K$ is monotonic in $\theta$, therefore this covariance cannot be identically zero. This establishes a contradiction and proves part (i).

By definition, the monotone-likelihood ratio property for $\mathscr{K}$ is equivalent to

$$(y - y')(\theta - \theta')\left[k_\theta(y)k_{\theta'}(y') - k_\theta(y')k_{\theta'}(y)\right] > 0, \tag{C.4}$$

where $y \neq y'$ and $\theta \neq \theta'$ are arbitrary. The posterior density is given by

$$d\Pi(\theta|y)/d\Pi(\theta) = k_\theta(y)/k_\Pi(y),$$

where $k_\Pi(y) = \int k_\theta(y)\, d\Pi(\theta)$ is the marginal density for $y$. Therefore we can divide the term inside the square bracket in (C.4) by $k_\Pi(y)k_\Pi(y')$ to obtain the monotone-likelihood ratio property for the posterior distributions and vice versa. This establishes the equivalence between (a) and (b) in (ii).

It is a well-known fact that (b) implies (c). To verify that (c) implies (a), take the prior distribution to be uniform over two given points $\theta$ and $\theta'$. It follows from (c) that for any $y < y'$,

$$\frac{\theta k_\theta(y) + \theta' k_{\theta'}(y)}{k_\theta(y) + k_{\theta'}(y)} < \frac{\theta k_\theta(y') + \theta' k_{\theta'}(y')}{k_\theta(y') + k_{\theta'}(y')}.$$

This inequality is equivalent to (C.4). □

**Acknowledgments.** We appreciate useful discussions with Jerry Friedman and helpful comments from two anonymous referees.

## REFERENCES

BRILLINGER, D. R. (1977). The identification of a particular nonlinear time series system. *Biometrika* **64** 622–654.

BRILLINGER, D. R. (1983). A generalized linear model with "Gaussian" regressor variables. In *A Festschrift for Erich L. Lehmann in Honor of His Sixty-Fifth Birthday* (P. J. Bickel, K. A. Doksum and J. L. Hodges, eds.) 97–114. Wadsworth, Belmont, Calif.

CHUNG, C. F. and GOLDBERGER, A. S. (1984). Proportional projections in limited dependent variable models. *Econometrica* **52** 531–534.

CONWAY, D. A. and ROBERTS, H. V. (1983). Reverse regression, fairness, and employment discrimination. *J. Business Econom. Statist.* **1** 75–85.

COX, D. R. (1985). Discussion of "Projection pursuit" by P. J. Huber. *Ann. Statist.* **13** 493–494.

COX, D. R. and SMALL, N. J. H. (1978). Testing multivariate normality. *Biometrika* **65** 263–272.

DUAN, N. and LI, K.-C. (1985). The ordinary least squares estimation for the general-link linear models, with applications. Technical Report Ser. 2880, Math. Res. Center, Univ. Wisconsin, Madison.

DUAN, N. and LI, K.-C. (1987). Distribution-free and link-free estimation for the sample selection model. *J. Econometrics* **35** 25–35.

DUAN, N. and LI, K.-C. (1991). A bias bound for least squares linear regression. *Statistica Sinica* **1**.

FISHER, R. A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugenics* **7** 179–188.

GOLDBERGER, A. S. (1981). Linear regression after selection. *J. Econometrics* **15** 357–366.

GREENE, W. (1981). On the asymptotic bias of ordinary least squares estimates of the Tobit model. *Econometrica* **49** 505–514.

GREENE, W. (1983). Estimation of limited dependent variable models by ordinary least squares and the method of moments. *J. Econometrics* **21** 195–212.

HAGGSTROM, G. W. (1983). Logistic regression and discriminant analysis by ordinary least squares. *J. Business Econom. Statist.* **1** 229–238.

HUBER, P. J. (1981). *Robust Statistics*. Wiley, New York.

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

HUNTER, W. G. and LAMBOY, W. F. (1981). A Bayesian analysis of the linear calibration problem. *Technometrics* **23** 323–328.

KRUTCHKOFF, R. G. (1967). Classical and inverse regression methods of calibration. *Technometrics* **9** 425–439.

KRUTCHKOFF, R. G. (1969). Classical and inverse regression methods of calibration in extrapolation. *Technometrics* **11** 605–608.

LI, K.-C. and DUAN, N. (1989). Regression analysis under link violation. *Ann. Statist.* **17** 1009–1052.

RUUD, P. (1983). Sufficient conditions for the consistency of maximum likelihood estimation despite misspecification of distribution in multinomial discrete choice models. *Econometrica* **51** 225–228.

RUUD, P. (1986). Consistent estimation of limited dependent variable models despite misspecification of distribution. *J. Econometrics* **32** 157–187.

VEGESNA, V., WITHERS, H. R. and TAYLOR, J. M. G. (1988). Epilation in mice after single and multifractionated irradiation. *Radiotherapy and Oncology* **12** 233–239.

WHITE, H. (1981). Consequences and detection of misspecified nonlinear regression models. *J. Amer. Statist. Assoc.* **76** 419–433.

RAND CORPORATION
1700 MAIN STREET
SANTA MONICA, CALIFORNIA 90406–2138

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF CALIFORNIA
AT LOS ANGELES
405 HILGARD AVENUE
LOS ANGELES, CALIFORNIA 90024