UCLA
Department of Statistics Papers

Title
Regression Analysis with an Unknown Link Function: the Adjoint Projection Pursuit Regression

Permalink
https://escholarship.org/uc/item/0rw86170

Author
Duan, Naihua

Publication Date
1990-06-07

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

The Adjoint Projection Pursuit Regression
Author(s): Naihua Duan
Source: *Journal of the American Statistical Association*, Vol. 85, No. 412 (Dec., 1990), pp. 1029-1038
Published by: American Statistical Association
Stable URL: http://www.jstor.org/stable/2289599
Accessed: 16/05/2011 19:19

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at http://www.jstor.org/page/info/about/policies/terms.jsp. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/action/showPublisher?publisherCode=astata.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

American Statistical Association is collaborating with JSTOR to digitize, preserve and extend access to *Journal of the American Statistical Association*.

http://www.jstor.org

The Adjoint Projection Pursuit Regression

NAIHUA DUAN*

Consider a projection pursuit regression model $y = g(\alpha + \beta\mathbf{x}) + \varepsilon$, with an arbitrary monotonic link function g. We assume the link function is unknown; thus we can only estimate the direction of β, that is, the ratios among the components of β, say, β_1/β_2. The direction of β is a useful estimand that measures, say, how many units of x_2 is equivalent to one unit of x_1 in terms of potency in affecting the outcome y. Projection pursuit regression usually solves the equation $\text{cov}(\mathbf{x}, y - E(y \mid \beta\mathbf{x})) = \mathbf{0}$. As an alternative, we propose to solve the adjoint equation $\text{cov}(\mathbf{x} - E(\mathbf{x} \mid \beta\mathbf{x}), y) = \mathbf{0}$. The adjoint equation has the advantage that $E(\mathbf{x} \mid \beta\mathbf{x})$ might be easier to estimate than $E(y \mid \beta\mathbf{x})$. We establish two main results for the population case. First, β is the unique solution (up to a multiplicative scalar) to the adjoint equation. Second, β satisfies a fixed-point property based on a modified least squares regression derived from the adjoint equation. We apply the population results to two important empirical problems: diagnosis and estimation. For diagnosis, we check whether the linear least squares estimate for β is valid for the direction of β. In particular, we test whether the estimate agrees with either the adjoint equation or the fixed-point property. The tests are analogous to Tukey's 1 df test. For estimation, we propose the adjoint projection pursuit regression estimate, which solves the empirical adjoint equation. The estimate is $\sqrt{n}$-consistent for β up to a multiplicative scalar. The rate of convergence is insensitive to the choice of the smoothing parameter in estimating $E(\mathbf{x} \mid \beta\mathbf{x})$; any window size of order $O(1/\sqrt{n})$ or smaller achieves the optimal convergence rate. The estimate can be implemented numerically by iterating the modified least squares regression; a simulation study is given to illustrate this procedure.

KEY WORDS: Adjoint equation; Diagnostic test; Modified least squares regression; Nonparametric orthogonalization.

1. INTRODUCTION

Regression analysis is usually based on a working model that is at best an approximation. For example, we might assume a standard linear model

$$y = \alpha + \beta\mathbf{x} + \varepsilon, \qquad \varepsilon \mid \mathbf{x} \sim N(0, \sigma^2), \qquad (1.1)$$

where y denotes a scalar outcome variable and $\mathbf{x}$ denotes a p-dimensional column vector of regressor variables. Under this model, we might use the least squares linear regression of y on $\mathbf{x}$ to estimate the parameter vector (α, β). However, we might be concerned about possible violations of the model assumptions. For example, the standard linear model might be subject to *distribution violation:* the error distribution might not be normal. There is a rich literature on robust methods for estimating the linear model in the presence of distribution violation; see, for example, Huber (1981).

A more serious challenge to the working model is a violation of the functional form. The true model might be a power transformation model; the working model (1.1) might be based on a wrong transformation. For example, assume the true model is

$$y = (\alpha + \beta\mathbf{x} + \varepsilon)^3, \qquad \varepsilon \mid \mathbf{x} \sim N(0, \sigma^2). \qquad (1.1')$$

We should take the cubic root transformation for y before fitting the linear regression. If we use (1.1) as the working model, and fit the least squares linear regression of y on $\mathbf{x}$, our estimates for α and β might be substantially biased.

Table 1 (in Sec. 5) gives an illustration based on a Monte Carlo study. We assume $\alpha = 0$, $\beta = (3, 1)$, $\sigma^2 = 1$, and $\mathbf{x}$ is distributed uniformly over the square ($-1 \le x_1 \le 1$, $-1 \le x_2 \le 1$). For reasons to be discussed later, we focus on estimating $r = \beta_2/\beta_1$. For each replicate of the simulation, we sample 400 observations from the uniform distribution for $\mathbf{x}$ and model (1.1′). The simulation is replicated 1,000 times.

The least squares linear regression of y on $\mathbf{x}$ is summarized in "iteration" column 0. The Monte Carlo estimate for $E(\hat{r})$ is .44446, about 33% larger than the true value $r = \frac{1}{3}$. In the following paragraphs we introduce a new method for estimating r, called the *adjoint projection pursuit regression.* Iteration columns 1–4 summarize the results of applying this new method, using an iterative algorithm called the *modified least squares regression.* The rows (m) in the table refer to the amount of smoothing used in implementing the algorithm. After one iteration of the algorithm (column 1), the Monte Carlo estimates for $E(\hat{r})$ are very close to the true value; that is, the bias is essentially eliminated entirely. Furthermore, the behavior of the estimate is rather insensitive to the amount of smoothing: there is not much difference among rows in the table.

We now describe the general theory for the adjoint projection pursuit regression and the modified least squares regression; the Monte Carlo study is discussed further in Section 5. Instead of restricting the true model to be a power transformation model such as (1.1′), we assume the true model has the following form:

$$y = g(\alpha + \beta\mathbf{x}) + \varepsilon, \qquad E(\varepsilon \mid \mathbf{x}) = 0, \qquad \beta \neq \mathbf{0}, \quad (1.2)$$

where g is the *link function,* assumed to be arbitrary and unknown. We call a model of form (1.2) a *projection pursuit regression model* (PPRM) (with one ridge); see Friedman and Stuetzle (1981). Model (1.2) includes both transformation models (see Remark 1.1 at the end of the

* Naihua Duan is Senior Statistician, Economics and Statistics Department, RAND Corporation, Santa Monica, CA 90406. This research was supported in part by a cooperative agreement between RAND Corporation, SIMS, and the U.S. Environmental Protection Agency, and in part by RAND corporate funds. The author appreciates helpful discussions with Dennis Cook, Persi Diaconis, Brad Efron, Jerry Friedman, Hidehiko Ichimura, Jacob Klerman, Ker-Chau Li, Dan Relles, Tom Stoker, and Sandy Weisberg, and helpful comments from the associate editor and two referees.

© 1990 American Statistical Association

Introduction) and generalized linear models. To avoid trivialities, we assume the slope vector β is not null.

Stoker (1986) and Ichimura (1989) studied a broader class of models, the single-index models $y = g(h(\beta; \mathbf{x})) + \varepsilon$, $E(\varepsilon \mid \mathbf{x}) = 0$, where h is known up to the parameter β. Model (1.2) is a special case, with $h = \alpha + \beta\mathbf{x}$. We focus on model (1.2) in this paper.

For model (1.2), when the link function g is arbitrary and unknown, we cannot estimate the entire parameter vector (α, β). The most we can identify in (α, β) is the *direction* of β, which we define to be the collection of the ratios $\{\beta_j/\beta_k, j, k = 1, \ldots, p\}$. We cannot identify the intercept α and the length of β. We will therefore neglect the intercept α in the PPRM, and focus on estimating the direction of β.

The inability to identify the magnitudes of the slope coefficients might be disturbing. Is this too general a context for meaningful estimation and inference? No; the direction of β is a meaningful estimand; it measures the substitutability of different components of the regressor $\mathbf{x}$. If we can determine the ratio β_1/β_2, we know how many units of x_2 are equivalent to one unit of x_1 in terms of potency in affecting the outcome y; the ratio can be interpreted as the *relative potency* between x_1 and x_2.

For many empirical problems, we might be interested more in the relative potencies than in the absolute potencies. For example, consider a risk management decision problem: we must choose between two pesticides, A and B, both of which are carcinogenic. Assume that the extermination efficacies of both pesticides are linear and let c_1 (c_2) denote the efficacy for one unit of pesticide A (B). Assume that the carcinogenicity of the pesticides follows a PPRM, with β_1 (β_2) being the potency for one unit of pesticide A (B). Assume that the link function is increasing; that is, a larger value of $\beta\mathbf{x}$ corresponds to stronger carcinogenicity. Under those assumptions, we should choose pesticide A if $\beta_1/\beta_2 < c_1/c_2$, and B otherwise. It follows that we need only estimate the relative potency β_1/β_2 and not the absolute potencies β_1 and β_2.

When the prediction of y from $\mathbf{x}$ is of interest, we can first estimate the direction of β, then smooth y on $\hat{\beta}\mathbf{x}$ to predict y; $\hat{\beta}$ being an estimate that might be poor for estimating the length of β, but is good (e.g., consistent) for estimating the direction of β. Any bias in estimating the length of β is compensated by an opposite bias in estimating the link function g.

We assume throughout this article that the regressor $\mathbf{x}$ is sampled from a nondegenerate population:

Condition 0. The regressor $\mathbf{x}$ is sampled randomly from a probability distribution; the moments $\mu = E(\mathbf{x})$, $\Sigma(\mathbf{x}\mathbf{x}') = \text{cov}(\mathbf{x})$, and $\Sigma(\mathbf{x}\mathbf{x}')^{-1}$ exist.

We now review some relevant facts about the standard linear model (1.1), then extend them to the PPRM (1.2). For now we focus on the population case: $(y, \mathbf{x}')$ is a random vector. The population results are applied to the sampling case in Sections 3 and 4.

For model (1.1), the usual likelihood theory determines the direction of β by solving the equation

$$\text{cov}(\mathbf{x}, y - L(y \mid \mathbf{b}\mathbf{x})) = \mathbf{0}, \tag{1.3}$$

where $\mathbf{b} \in R^p$ and $L(y \mid \mathbf{b}\mathbf{x})$ is the least squares linear regression of y on $\mathbf{b}\mathbf{x}$. Under (1.1), $\mathbf{b} = \beta$ is the unique solution (up to a multiplicative scalar) to (1.3). Equation (1.3) is equivalent to

$$\text{cov}(\mathbf{x} - L(\mathbf{x} \mid \mathbf{b}\mathbf{x}), y) = \mathbf{0}, \tag{1.4}$$

where

$$L(\mathbf{x} \mid \mathbf{b}\mathbf{x}) = \mu + \Sigma(\mathbf{x}\mathbf{x}')\mathbf{b}'\mathbf{b}(\mathbf{x} - \mu)/\mathbf{b}\Sigma(\mathbf{x}\mathbf{x}')\mathbf{b}' \tag{1.5}$$

is the linear regression of $\mathbf{x}$ on $\mathbf{b}\mathbf{x}$. In (1.3), we orthogonalize y against $\mathbf{b}\mathbf{x}$, then examine the association between $\mathbf{x}$ and the y residual. In (1.4), we orthogonalize $\mathbf{x}$ against $\mathbf{b}\mathbf{x}$, then examine the association between y and the $\mathbf{x}$ residual. The two equations are dual to each other; therefore we call (1.4) the (linear) *adjoint equation;* see, for example, Halmos (1958) for a general discussion of adjoint operators in linear algebra.

Both Equations (1.3) and (1.4) are equivalent to

$$\text{cov}(\mathbf{x} - L(\mathbf{x} \mid \mathbf{b}\mathbf{x}), y - L(y \mid \mathbf{b}\mathbf{x})) = \mathbf{0}, \tag{1.6}$$

which deals with the partial association between $\mathbf{x}$ and y after controlling for $\mathbf{b}\mathbf{x}$.

The analog of (1.3) for the PPRM is the equation

$$\text{cov}(\mathbf{x}, y - E(y \mid \mathbf{b}\mathbf{x})) = \mathbf{0}. \tag{1.7}$$

The analog of (1.4) is the (nonparametric) adjoint equation

$$\text{cov}(\mathbf{x} - E(\mathbf{x} \mid \mathbf{b}\mathbf{x}), y) = \mathbf{0}. \tag{1.8}$$

Both Equations (1.7) and (1.8) are equivalent to the partial association equation

$$\text{cov}(\mathbf{x} - E(\mathbf{x} \mid \mathbf{b}\mathbf{x}), y - E(y \mid \mathbf{b}\mathbf{x})) = \mathbf{0}. \tag{1.9}$$

The only difference between the two sets of equations is the way we orthogonalize y and $\mathbf{x}$ against $\mathbf{b}\mathbf{x}$. For the standard linear model, we use the linear regression to control for $\mathbf{b}\mathbf{x}$. For the PPRM, we use the nonparametric regression to control for $\mathbf{b}\mathbf{x}$.

Under the PPRM, $\mathbf{b} = \beta$ solves Equation (1.7). The (standard) projection pursuit regression is based on solving (1.7); see, for example, Friedman and Stuetzle (1981, 1985) and Huber (1985). Friedman and Stuetzle proposed an algorithm that alternates between smoothing y on $\mathbf{b}\mathbf{x}$ and optimizing $\mathbf{b}$ for a given link function.

Equations (1.7)–(1.9) are all equivalent to $E[\text{cov}(\mathbf{x}, y \mid \mathbf{b}\mathbf{x})] = \mathbf{0}$. We can therefore solve any of the three equations to determine β. We propose the adjoint projection pursuit regression, which is based on solving the adjoint equation (1.8).

The apparent advantage of solving (1.8) instead of (1.7) is that (1.8) does not depend explicitly on the unknown link function. For empirical applications, this might be an important advantage: $E(\mathbf{x} \mid \mathbf{b}\mathbf{x})$ might be easier to estimate than $E(y \mid \mathbf{b}\mathbf{x})$. For randomized experiments, we might know the distribution for $\mathbf{x}$, and therefore have complete

knowledge about $E(\mathbf{x} \mid \mathbf{bx})$. For sampling studies, the sample available for estimating $E(y \mid \mathbf{bx})$ might be much smaller than the sample available for estimating $E(\mathbf{x} \mid \mathbf{bx})$. For example, we might observe $\mathbf{x}$ for a large sample screened for possible enrollment in a study; we might observe y only for a smaller sample actually enrolled in the study.

When $\mathbf{x}$ and y are observed in the same sample, the estimation of $E(\mathbf{x} \mid \mathbf{bx})$ might require more computation than the estimation of $E(y \mid \mathbf{bx})$. However, $E(\mathbf{x} \mid \mathbf{bx})$ might be closer to being linear than $E(y \mid \mathbf{bx})$, and therefore easier to estimate. We can improve upon the linearity of $E(\mathbf{x} \mid \mathbf{bx})$ by deleting leverage points with extreme $\mathbf{x}$ values. Truncating the distribution of $\mathbf{x}$ does not affect the direction of β. It is more difficult to improve upon the linearity of $E(y \mid \mathbf{bx})$. For example, deleting or modifying outliers with extreme y values might affect the direction of β.

We introduce some notation: $\zeta(\mathbf{bx}, \mathbf{b}) = E(\mathbf{x} \mid \mathbf{bx})$ and $\delta(\mathbf{b}) = \operatorname{cov}(\mathbf{x} - \zeta(\mathbf{bx}, \mathbf{b}), y)$. For a given $\mathbf{b}$, ζ is the nonparametric regression of $\mathbf{x}$ on $\mathbf{bx}$. Note that its functional form also depends on $\mathbf{b}$. The covariance function $\delta(\mathbf{b})$ is analogous to the score function in parametric regression models; therefore, we call it the *adjoint score function.*

We establish two main results in Section 2. First, we establish in Theorem 2.1 that β is the unique solution (up to a multiplicative scalar) to the adjoint equation (1.8) under the PPRM and some mild conditions, the main one being

Condition 1. The link function g is strictly monotonic.

Condition 1 is plausible in many empirical applications. Although the scientist does not know the exact form of the link function, he might know the ranking of the expected outcome, and thus can verify a priori whether Condition 1 is valid. For example, an environmental epidemiologist might not know the exact form of a dose–response curve; nevertheless, he probably has a firm belief that a higher dose of PCB means a stronger carcinogenicity.

Our second main result relates the adjoint equation (1.8) to the least squares slope $\beta_{\mathrm{LS}} = \Sigma(y\mathbf{x}')\Sigma(\mathbf{x}\mathbf{x}')^{-1}$, where $\Sigma(y\mathbf{x}') = \operatorname{cov}(y, \mathbf{x}')$. Under the PPRM, β_{LS} solves (1.4) but might not solve (1.8); therefore, it might not be collinear with β. However, if

$$E(\mathbf{x} \mid \mathbf{bx}) = L(\mathbf{x} \mid \mathbf{bx}) \tag{1.10}$$

for all $\mathbf{b}$, then (1.8) coincides with (1.4), and β_{LS} solves both equations.

Under (1.10), β_{LS} is collinear with β, despite possible nonlinearity in the link function. This result was first established in Brillinger (1977, 1982). Related results for various parametric regressions are given in Chung and Goldberger (1984), Duan and Li (1987), Goldberger (1981), Greene (1981, 1983), Li and Duan (1989), Ruud (1983), and White (1981). Duan and Li (in press) studied a link-free regression method, the slicing regression, which is also based on (1.10).

It is unlikely for (1.10) to hold exactly in empirical applications. When (1.10) is not satisfied, we can modify $\mathbf{x}$ so that it will satisfy (1.10) for a given initial value $\mathbf{b}$. In particular, we consider the *modified regressor*

$$\tilde{\mathbf{x}} = L(\mathbf{x} \mid \mathbf{bx}) + (\mathbf{x} - E(\mathbf{x} \mid \mathbf{bx})). \tag{1.11}$$

It is easy to verify that $\tilde{\mathbf{x}}$ satisfies (1.10).

Our second main result is a fixed-point property for β. We take the least squares linear regression of y on $\tilde{\mathbf{x}}$, and call this regression the *modified least squares regression.* The slope vector from this regression is called the *modified least squares slope:*

$$\mathbf{b}^{(1)} = \Sigma(y\tilde{\mathbf{x}}')\Sigma(\tilde{\mathbf{x}}\tilde{\mathbf{x}}')^{-1}, \tag{1.12}$$

where $\Sigma(\tilde{\mathbf{x}}') = \operatorname{cov}(y, \tilde{\mathbf{x}}')$ and $\Sigma(\tilde{\mathbf{x}}\tilde{\mathbf{x}}') = \operatorname{cov}(\tilde{\mathbf{x}})$. We establish in Theorem 2.2 that $\mathbf{b}^{(1)}$ is collinear with $\mathbf{b}$ if and only if $\mathbf{b}$ is collinear with β. In other words, the direction of β is the only fixed point for the functional

$$\mathcal{F} : \mathbf{d}(\mathbf{b}) \to \mathbf{d}(\mathbf{b}^{(1)}), \tag{1.13}$$

where $\mathbf{d}(\mathbf{v})$ denotes the direction of the vector $\mathbf{v}$.

We apply the preceding population results to two important empirical problems: diagnosis and estimation. For the diagnosis problem, we use the standard linear model (1.1) as the working model, and take the least squares linear regression of y on $\mathbf{x}$. However, we suspect that the link function might be nonlinear. We want to diagnose whether the estimate $\hat{\beta}_{\mathrm{LS}}$ is valid for the direction of β.

The standard diagnostic techniques aim at detecting nonlinearity in the link function; see, for example, Tukey (1949) and Cook and Weisberg (1982). However, this might be inefficient if we are interested only in the direction of β. We might be willing to accept $\hat{\beta}_{\mathrm{LS}}$ if it estimates the direction of β properly, despite possible nonlinearity in the link function. We therefore propose to use diagnostic techniques that focus on the direction of β.

An obvious diagnostic technique is to test whether $\hat{\beta}_{\mathrm{LS}}$ agrees with Equations (1.7) or (1.8). There is an advantage in using (1.8) for this purpose: the null distribution for the test can be derived in a manner similar to Tukey's 1 df test (Theorem 3.1). We can also use the fixed-point result for such a diagnosis. We take $\hat{\beta}_{\mathrm{LS}}$ as the initial value, then modify the regressor accordingly to obtain the modified least squares slope. We then test whether the two estimated slopes are collinear (Theorem 3.2).

For the estimation problem, we want to estimate β without relying on a working model like (1.1). Friedman and Stuetzle's (1981) projection pursuit regression is an obvious choice. Hall (1989) established that the projection pursuit regression estimates the direction of β at the $\sqrt{n}$ rate. Ichimura (1989) established similar results for the single-index model, using the semiparametric least squares estimate, based on minimizing $n^{-1} \sum_i [y_i - \hat{E}(y \mid h(\beta, \mathbf{x}_i))]^2$.

Han (1987) proposed the maximum rank correlation estimator, which maximizes the rank correlation between y and $\mathbf{bx}$. Han established strong consistency for this estimator, and conjectured that it converges at the $\sqrt{n}$ rate.

Manski (1975, 1985) studied a closely related method, the maximum score estimator. Both Han's and Manski's estimators are fairly difficult to compute numerically.

Stoker (1986, 1989), Powell, Stock, and Stoker (1989), and Härdle and Stoker (1989) studied the average derivative estimate (ADE). Let $m(\mathbf{x}) = E(y \mid \mathbf{x})$, and $\nabla m(\mathbf{x})$ be the derivative of m at $\mathbf{x}$. Under model (1.2), we have $E[\nabla m(\mathbf{x})] = \gamma\beta$, where $\gamma = E[g'(\beta\mathbf{x})]$. We can estimate the direction of β by estimating $\nabla m(\mathbf{x})$, then taking the average. Alternatively, since $E[\nabla m(\mathbf{x})] = E[l(\mathbf{x})y]$, where $l(\mathbf{x}) = -\nabla \log f(\mathbf{x})$ and $f(\mathbf{x})$ is the density of $\mathbf{x}$, we can estimate the direction of β by estimating $l(\mathbf{x})y$, then taking the average. The ADE requires p-dimensional smoothing to estimate $\nabla m(\mathbf{x})$ or $f(\mathbf{x})$ and might be difficult to implement when p is not small. However, it is $\sqrt{n}$-consistent, does not require iterative computation, and does not require the monotonicity Condition 1. It does require that $\gamma \neq 0$.

We propose the adjoint projection pursuit regression estimate, which solves the adjoint equation (1.8) empirically. The estimate is shown in Theorem 4.1 to be $\sqrt{n}$-consistent for β up to a multiplicative scalar. The rate of convergence is insensitive to the choice of the smoothing parameter used in estimating $E(\mathbf{x} \mid \mathbf{bx})$; any window size of order $O(1/\sqrt{n})$ or smaller achieves the optimal convergence rate.

The adjoint projection pursuit regression can be implemented numerically by iterating the modified least squares regression. We do not have general results on the numerical properties of this algorithm. A heuristic argument is given in Section 4, which suggests that when the initial value is good, the algorithm converges very fast: one iteration might be nearly adequate.

The behavior of the adjoint projection pursuit regression estimate is illustrated with a simulation study in Section 5. For the example used in this study, one iteration eliminates the bias almost completely; the behavior of the estimate is also found to be insensitive to the choice of the smoothing parameter.

Remark 1.1. Transformation models such as (1.1′) are submodels of the PPRM (1.2) as long as the transformation is invertible and the error terms are identically distributed after the transformation. For example, consider a general transformation model

$$h(y) = \alpha + \beta\mathbf{x} + \varepsilon, \qquad \varepsilon \mid \mathbf{x} \sim F(\varepsilon). \tag{1.1″}$$

The regression function is given by

$$E(y \mid \mathbf{x}) = \int h^{-1}(\alpha + \beta\mathbf{x} + \varepsilon)\, dF(\varepsilon), \tag{1.14}$$

and depends on $\mathbf{x}$ only through $\alpha + \beta\mathbf{x}$. Therefore the transformation model (1.1″) can be written in the form of the PPRM (1.2), with the link function being the regression function (1.14) expressed as a function of $\alpha + \beta\mathbf{x}$. It should be noted that the PPRM (1.2) does not require the errors to be identically distributed.

Remark 1.2. Under model (1.1), (1.3) is the likelihood equation for the direction of β. Equation (1.7) can also be interpreted as the likelihood equation if the true model is a generalized linear model with the canonical link [see, e.g., McCullagh and Nelder (1983)], and we replace the nonparametric regression $E(y \mid \mathbf{bx})$ in (1.7) by the parametric mean function for the given generalized linear model.

Remark 1.3. For PPRM's with more than one ridge [$E(y \mid \mathbf{x}) = \sum_{j=1}^{K} g_j(\beta^{(j)}\mathbf{x})$], the adjoint projection pursuit regression can be applied to the residuals for each successive ridge component in steps similar to the standard projection pursuit regression.

2. MAIN RESULTS

We observed in Section 1 that β solves the adjoint equation (1.8). We now give the converse, which guarantees that the solution is unique up to a multiplicative scalar.

Theorem 2.1. Assume the random vector $(y, \mathbf{x}')$ follows a PPRM (1.2) and satisfies Conditions 0 and 1 and

Condition 2. If $\mathbf{b} \in R^p$ is not collinear with β, then $E[\mathrm{var}(\beta\mathbf{x} \mid \mathbf{bx})] > 0$.

Let $\mathbf{b} \in R^p$ be any initial value for β. The adjoint equation (1.8) is solved by $\mathbf{b}$ if and only if $\mathbf{b}$ is collinear with β.

(The proofs of all theorems are given in the Appendix.)

We observed in Section 1 that Condition 1 is plausible in many empirical problems. It cannot be ignored: without Condition 1, the adjoint equation might have solutions that are not collinear with β. An example is given in Remark 2.1.

Condition 2 indicates that $\beta\mathbf{x}$ is not a deterministic function of $\mathbf{bx}$. This is necessary in order for the direction of β to be identifiable; otherwise we can express the PPRM using either a link function for $\beta\mathbf{x}$ or a link function for $\mathbf{bx}$.

Condition 2 is satisfied if the support of $\mathbf{x}$ is p dimensional; that is, has an interior point in R^p. This might be too restrictive, though. For many empirical applications, some of the regressors might be discrete; we might also have functional constraints on $\mathbf{x}$; for example, some of the regressors might be higher-order terms or interaction terms. In order to identify the direction of β in those cases, we usually need an "anchoring" variable known to have an effect. We can then calibrate the potency of the other regressors relative to the anchoring variable.

Lemma 2.1. Condition 2 follows from Condition 0 and the following:

Condition 2′. The support for $\mathbf{x}$ has the form $R \times \mathcal{S}$, where $\mathcal{S} \subseteq R^{p-1}$ is the support for $(x_2, \ldots, x_p)'$.

Condition 2″. $\beta_1 \neq 0$.

Condition 2′ requires that the anchoring variable x_1 have infinite support. It might be possible to weaken this requirement if the direction of β is known to be bounded away from the hyperplane $\beta_1 = 0$; see Remark 2.2. We

also give examples in Remarks 2.2 and 2.3 to illustrate that Conditions 2′ and 2″ cannot be ignored.

We now give the fixed-point property for β: we can check the collinearity between $\mathbf{b}$ and $\mathbf{b}^{(1)}$ to determine the collinearity between $\mathbf{b}$ and β.

Theorem 2.2. Assume the random vector $(y, \mathbf{x}')$ follows a PPRM (1.2) and satisfies Conditions 0 and 1, and

Condition 2‴. If $\mathbf{b}, \mathbf{v} \in R^p$ are not collinear, then $E[\mathrm{var}(\mathbf{vx} \mid \mathbf{bx})] > 0$.

The initial value $\mathbf{b}$ and the modified least squares slope $\mathbf{b}^{(1)}$ in (1.12) are collinear if and only if $\mathbf{b}$ and β are collinear.

Condition 2‴ is stronger than Condition 2: it requires that all two-dimensional projections of $\mathbf{x}$ be nondegenerate. However, this stronger condition is only used to establish that $\Sigma(\tilde{\mathbf{x}}\tilde{\mathbf{x}}')$ is nonsingular; that is, $\mathbf{b}^{(1)}$ is well defined. The theorem remains true with Condition 2‴, if we redefine the modified least squares regression as

$$\mathbf{b}^{(1)} = \Sigma(y\tilde{\mathbf{x}}')\Sigma(\mathbf{xx}')^{-1}. \tag{1.12′}$$

With definition (1.12′), we do not need to verify that $\Sigma(\tilde{\mathbf{x}}\tilde{\mathbf{x}}')$ is nonsingular; thus only Condition 2 is needed.

There does not appear to be any intrinsic reason for preferring either (1.12) or (1.12′). We have focused on (1.12) instead of (1.12′) for two reasons, neither of which is very compelling. First, (1.12) is easier to implement on standard statistical packages. Second, (1.12) leads to a simpler null distribution for the diagnostic test to be given in Section 3.

Remark 2.1. We now give a counterexample for Theorem 2.1 when Condition 1 fails. Let $p = 2$ and let $\mathbf{x}$ follow a joint distribution such that $E(x_2 \mid x_1) \equiv 0$, but $E(x_1 \mid x_2)$ is not a constant. For example, take the uniform distribution over the triangle with vertices $(1, 1)$, $(1, -1)$, and $(0, 0)$. Let $\beta = (0, 1)$, $g(x_2) = E(x_1 \mid x_2)$, and $\mathbf{b} = (1, 0) \not\propto \beta$. Note that $\mathrm{cov}(x_1, x_2) = \mathrm{cov}(x_1, E(x_2 \mid x_1)) = 0$. Since $\mathrm{cov}(x_1, x_2) = \mathrm{cov}(x_2, g(x_2))$, the link function does not satisfy Condition 1. It is easy to verify that $\delta(\mathbf{b}) = \mathbf{0}$; thus the converse part of Theorem 2.1 fails.

Remark 2.2. Consider an ANOCOVA problem with $x_2 = 0$ or 1, and the support of x_1 given x_2 is the interval $(0, 1)$. Let $\mathbf{b} = (1, 1)$. Conditioned on $\mathbf{bx}$, $\mathbf{x}$ is degenerate; therefore, Condition 2 fails even if Condition 2″ holds. More generally, Condition 2 fails in this example if $|b_2/b_1| \geq 1$. For $|b_2/b_1| < 1$, the conditional distribution of $\mathbf{x}$ given $\mathbf{bx}$ is nondegenerate for some values of $\mathbf{bx}$; therefore, Condition 2 holds.

If we have prior information that $|\beta_2/\beta_1| < 1$, that is, x_2 is less potent than x_1, then we can restrict the initial value $\mathbf{b}$ to satisfy $|b_2/b_1| < 1$ also; therefore, Condition 2 and Theorem 2.1 hold. More generally, if we believe $|\beta_2/\beta_1| < c$, we can obtain Condition 2 by requiring the support of x_1 to be an interval with width at least $1/c$.

Remark 2.3. Consider a quadratic regression problem with $x_3 = x_2^2$ and $\beta_3 \neq 0$. If $\beta_1 = 0$, Condition 2 fails with $\mathbf{bx} = x_2$, even if Condition 2′ holds. Note that Lemma 2.1 identifies the direction of β by calibrating all potencies relative to x_1. If $\beta_1 = 0$, we can no longer compare the potencies of x_2 and x_3 by calibrating against x_1.

3. DIAGNOSIS

We now apply the results derived in Section 2 to the sampling case: we observe data $\{(y_i, \mathbf{x}_i'), i = 1, \ldots, n\}$ sampled randomly from a PPRM (1.2) that satisfies Condition 0. We need to estimate $\zeta(\mathbf{bx}, \mathbf{b}) = E(\mathbf{x} \mid \mathbf{bx})$. Since there is only one "predictor" $\mathbf{bx}$, we are free from the curse of dimensionality. There are many nonparametric regression methods available, such as the kernel method, the nearest-neighbor method, and smoothing splines. We assume that we have chosen a generic smoothing method, and denote the estimate as $\hat{\zeta}(\mathbf{bx}, \mathbf{b})$. We introduce some notation: $\zeta_i = \zeta(\mathbf{bx}_i, \mathbf{b})$ and $\hat{\zeta}_i = \hat{\zeta}(\mathbf{bx}_i, \mathbf{b})$. Let $\hat{\Sigma}(\mathbf{vw}')$ denote the sample covariance matrix between $\mathbf{v}$ and $\mathbf{w}$.

Since $\mathbf{b}\zeta_i = \mathbf{bx}_i$, we do not need to use smoothing to estimate $\mathbf{b}\zeta_i$. In order to take advantage of this fact, we reorient $\hat{\zeta}$ by

$$\hat{\zeta}_i = \hat{\zeta}_i + \hat{L}(\mathbf{x} - \hat{\zeta} \mid \mathbf{bx}_i), \tag{3.1}$$

where $\hat{L}$ is the estimated linear regression of $\mathbf{x} - \hat{\zeta}$ on $\mathbf{bx}$: $\hat{L}(\mathbf{x} - \hat{\zeta} \mid \mathbf{bx}_i) = (\bar{\mathbf{x}} - \bar{\hat{\zeta}}) + \hat{\Sigma}(\mathbf{xx}')\mathbf{b}'\mathbf{b}((\mathbf{x}_i - \bar{\mathbf{x}}) - (\hat{\zeta}_i - \bar{\hat{\zeta}}))/\mathbf{b}\hat{\Sigma}(\mathbf{xx}')\mathbf{b}'$, where $\bar{\mathbf{x}}$ ($\bar{\hat{\zeta}}$) is the sample average for the $\mathbf{x}_i$'s ($\hat{\zeta}_i$'s). It is easy to verify that $\mathbf{b}\hat{\zeta}_i = \mathbf{bx}_i$.

We estimate the adjoint score function $\delta(\mathbf{b}) = \mathrm{cov}(\mathbf{x} - \zeta(\mathbf{bx}, \mathbf{b}), y)$ by the sample covariance between $\mathbf{x} - \hat{\zeta}$ and y:

$$\begin{aligned}\hat{\delta}(\mathbf{b}) &= \hat{\Sigma}(\mathbf{x} - \hat{\zeta}, y) \\ &= ((\mathbf{X} - \bar{\mathbf{x}}) - (\hat{\mathbf{Z}} - \bar{z}))\mathbf{Y}'/(n - 1),\end{aligned} \tag{3.2}$$

where $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$, $\mathbf{Y} = (y_1, \ldots, y_n)$, $\hat{\mathbf{Z}} = [\hat{\zeta}_1, \ldots, \hat{\zeta}_n]$, and $\bar{z}$ is the sample average for the $\hat{\zeta}_i$'s.

For a given initial value $\mathbf{b}$, we estimate the modified regressor by

$$\hat{\mathbf{x}}_i = \hat{L}(\mathbf{x} \mid \mathbf{bx}_i) + (\mathbf{x}_i - \hat{\zeta}_i), \tag{3.3}$$

where $\hat{L}$ is the estimated linear regression of $\mathbf{x}$ on $\mathbf{bx}$. Note that $\mathbf{b}\hat{\mathbf{x}}_i = \mathbf{bx}_i$. We then estimate the modified least squares slope $\mathbf{b}^{(1)}$ by

$$\hat{\beta}^{(1)} = \hat{\Sigma}(y\hat{\mathbf{x}}')\hat{\Sigma}(\hat{\mathbf{x}}\hat{\mathbf{x}}')^{-1}. \tag{3.4}$$

For the rest of this section, we discuss the diagnosis problem. The estimation problem is studied in the next section.

We consider the estimated least squares slope $\hat{\beta}_{\mathrm{LS}} = \hat{\Sigma}(y\mathbf{x}')\hat{\Sigma}(\mathbf{xx}')^{-1}$. We test whether $\hat{\beta}_{\mathrm{LS}}$ agrees with either the adjoint equation (1.8) or the fixed-point property in Theorem 2.2. For simplicity of notation, we drop the subscript LS from $\hat{\beta}_{\mathrm{LS}}$ in this section.

First we test whether $\hat{\beta}$ agrees with the adjoint equation (1.8): we check whether $\hat{\delta}(\hat{\beta})$ is significantly different from the null vector. In order to implement this test, we need to derive the null distribution for $\hat{\delta}(\hat{\beta})$. We use a derivation

similar to Tukey's (1949) 1 df test, and consider the distribution for $\hat{\delta}(\hat{\beta})$ conditioned on $\hat{\beta}$. Under (1.1), we have the following conditional distribution: $\mathbf{Y} \mid (\hat{\beta}, \mathbf{X}) \sim N(\hat{\beta}\mathbf{X}, \sigma^2(I_n - P_n))$, where I_n is the $n \times n$ identity matrix and P_n is the projection matrix ("hat matrix") for $[\mathbf{1}_n, \mathbf{X}']$. The conditional distribution for $\hat{\delta}(\hat{\beta})$ is therefore $\hat{\delta}(\hat{\beta}) \mid (\hat{\beta}, \mathbf{X}) \sim N(v, \Gamma)$, where $v = (\hat{\Sigma}(\mathbf{xx}') - \hat{\Sigma}(\hat{\zeta}\hat{\zeta}'))\hat{\beta}'$, $\hat{\Sigma}(\hat{\zeta}\hat{\zeta}')$ is the sample covariance matrix for the $\hat{\zeta}_i$'s, and $\Gamma = \sigma^2\hat{\mathbf{Z}}(I_n - P_n)\hat{\mathbf{Z}}'/(n-1)^2$. This gives us a chi-squared test:

Theorem 3.1. Assume the observed data $\{(y_i, \mathbf{x}_i'), i = 1, \ldots, n\}$ are sampled randomly from the standard linear model (1.1) and satisfy Condition 0. We then have

$$(\hat{\delta}(\hat{\beta}) - v)'\Gamma^-(\hat{\delta}(\hat{\beta}) - v) \sim \chi_k^2, \tag{3.5}$$

where $k = \text{rank}(\Gamma)$ and Γ^- is any generalized inverse of Γ, such as the Penrose inverse.

An alternative diagnostic test is to check whether $\hat{\beta}$ agrees with the fixed-point property in Theorem 2.2. With $\mathbf{b} = \hat{\beta}$ as the initial value, we test whether $\hat{\beta}$ is collinear with the modified least squares slope $\hat{\beta}^{(1)}$ given in (3.4). By (3.4), the conditional distribution for $\hat{\beta}^{(1)}$ is given by $\hat{\beta}^{(1)} \mid (\hat{\beta}, \mathbf{X}) \sim N(\hat{\beta}, T)$, where $T = \hat{\Sigma}(\tilde{\mathbf{x}}\tilde{\mathbf{x}}')^{-1}\Gamma\hat{\Sigma}(\tilde{\mathbf{x}}\tilde{\mathbf{x}}')^{-1}$. We orthogonalize $\hat{\beta}^{(1)}$ against $\hat{\beta}$, and consider the residual $\hat{\beta}^{\perp} = \hat{\beta}\hat{D}$, where $\hat{D} = I_p - \hat{\Sigma}(\hat{\mathbf{x}}\hat{\mathbf{x}}')\Gamma^-\hat{\Sigma}(\hat{\mathbf{x}}\hat{\mathbf{x}}')\hat{\beta}'\hat{\beta}/\hat{\beta}\hat{\Sigma}(\hat{\mathbf{x}}\hat{\mathbf{x}}')\Gamma^-\hat{\Sigma}(\hat{\mathbf{x}}\hat{\mathbf{x}}')\hat{\beta}'$. The conditional distribution for $\hat{\beta}^{\perp}$ is given by $\hat{\beta}^{\perp} \mid (\hat{\beta}, \mathbf{X}) \sim N(\mathbf{0}, \Psi)$, where $\Psi = \hat{\Sigma}(\tilde{\mathbf{x}}\tilde{\mathbf{x}}')^{-1}\Gamma\hat{\Sigma}(\tilde{\mathbf{x}}\tilde{\mathbf{x}}')^{-1} - \hat{\beta}'\hat{\beta}/\hat{\beta}\hat{\Sigma}(\tilde{\mathbf{x}}\tilde{\mathbf{x}}')\Gamma^-\hat{\Sigma}(\tilde{\mathbf{x}}\tilde{\mathbf{x}}')\hat{\beta}'$. This gives us another chi-squared test:

Theorem 3.2. Assume the same conditions as in Theorem 3.1. We then have

$$\hat{\beta}^{\perp}\Psi^-\hat{\beta}^{\perp\prime} \sim \chi_m^2, \tag{3.6}$$

where Ψ^- is a generalized inverse of Ψ, and $m = \text{rank}(\Psi)$.

For the linear model (1.1), consider the population case, with the initial value $\mathbf{b} = \beta_{\text{LS}}$: the adjoint score function $\delta(\beta_{\text{LS}})$ is null, and the modified least squares slope $\mathbf{b}^{(1)}$ coincides with β_{LS}. For the sampling case, with the initial value $\hat{\beta}_{\text{LS}}$, the conditional expectation of $\hat{\delta}(\hat{\beta}_{\text{LS}})$ is v, which might not be null; therefore, the estimated score function might be conditionally biased. On the other hand, the conditional expectation of $\hat{\beta}^{(1)}$ coincides with $\hat{\beta}_{\text{LS}}$; therefore, the estimated modified least squares slope is conditionally unbiased. This is an advantage for the diagnostic test in Theorem 3.2 over the test in Theorem 3.1.

4. ESTIMATION

For the estimation problem, we assume the observed data follow a PPRM (1.2) with an unknown link function and want to estimate β up to a multiplicative scalar. Either Theorem 2.1 or Theorem 2.2 can be used for this purpose. First we consider the empirical adjoint equation

$$\hat{\delta}(\mathbf{b}) = \mathbf{0}, \tag{4.1}$$

where $\hat{\delta}$ is the estimated adjoint score function given in (3.2). We estimate the direction of β by any solution to (4.1). Alternatively, we can estimate $\mathfrak{F}$ in (1.13) by

$$\hat{\mathfrak{F}} : \mathbf{d}(\mathbf{b}) \to \mathbf{d}(\hat{\beta}^{(1)}), \tag{4.2}$$

where $\hat{\beta}^{(1)}$ is given by (3.4). We then estimate the direction of β by any fixed point for $\hat{\mathfrak{F}}$.

We do not have general results on the existence or the uniqueness of solutions to (4.1) or of fixed points for (4.2). If the number of regressors, p, is odd, and $\hat{\mathfrak{F}}$ is continuous, a variant of Brouwer's fixed-point theorem holds (Brown 1971, p. 31), and guarantees the existence of a fixed point for (4.2). The continuity for $\hat{\mathfrak{F}}$ probably follows from suitable smoothness conditions on the PPRM and the distribution of $\mathbf{x}$; we have not been able to formulate this formally.

We will refer to either estimate defined above as the adjoint projection pursuit regression estimate $\hat{\beta}$ for the direction of β. The two definitions are approximately equivalent. By (3.2) and (3.3), we have $\hat{\Sigma}(y\hat{\mathbf{x}}') = c\mathbf{b}\hat{\Sigma}(\mathbf{xx}') + \hat{\delta}(\mathbf{b})'$, where c is a scalar. Since $\mathbf{b}\hat{\Sigma}(\mathbf{xx}') \cong \mathbf{b}\hat{\Sigma}(\hat{\mathbf{x}}\hat{\mathbf{x}}')$, $\hat{\beta}^{(1)} \cong c\mathbf{b} + \hat{\delta}(\mathbf{b})'\hat{\Sigma}(\hat{\mathbf{x}}\hat{\mathbf{x}}')^{-1}$. If the approximation can be replaced by an equality, then $\mathbf{b}$ solves (4.1) if and only if $\hat{\beta}^{(1)} \propto \mathbf{b}$; that is, the two definitions are equivalent. If we define $\mathbf{b}^{(1)}$ by (1.12′) instead of (1.12), and estimate it by $\hat{\beta}^{(1)} = \hat{\Sigma}(y\hat{\mathbf{x}})\hat{\Sigma}(\mathbf{xx}')^{-1}$ instead of (3.4), then the two definitions are exactly equivalent.

Under reasonable conditions, the smooth $\hat{\zeta}$ is consistent for ζ; thus the adjoint projection pursuit regression estimate is consistent for β up to a multiplicative scalar. This is established formally in Theorem 4.1 for the estimate based on (4.1).

For simplicity, we use a step function estimate for $\zeta(\mathbf{bx}, \mathbf{b})$. This is a rather crude smoothing method; however, it does achieve the optimal $\sqrt{n}$ convergence rate for $\hat{\beta}$. For empirical applications, it might be possible to improve the efficiency by using better nonparametric regression methods.

Without loss of generality, we normalize the initial value $\mathbf{b}$ to have length 1:

$$\|\mathbf{b}\| = (\mathbf{bb}')^{1/2} = 1. \tag{4.3}$$

We partition the range of $\mathbf{bx}$ into equal-width slices $\{\ldots, s_{-1}, s_0, s_1, \ldots, s_h, \ldots\}$, where the hth slice s_h is the interval $hd_n \le \mathbf{bx} < (h+1)d_n$. The window size d_n is chosen a priori and converges to 0 as $n \to \infty$. For each slice of $\mathbf{bx}$, we estimate ζ by the sample average for the corresponding $\mathbf{x}$'s. More specifically, we estimate $\zeta_i = \zeta(\mathbf{bx}_i, \mathbf{b})$ by

$$\tilde{\zeta}_i = \tilde{\zeta}(\mathbf{bx}_i, \mathbf{b}) = \sum_{j=1}^{n} \mathbf{x}_j 1_{jh}/n\hat{p}_h \quad \text{if} \quad \mathbf{bx}_i \in s_h, \tag{4.4}$$

where 1_{jh} is the indicator for the event "$\mathbf{bx}_j \in s_h$," and $\hat{p}_h$ is the proportion of $\mathbf{bx}_j$'s in the hth slice. We then reorient $\tilde{\zeta}$ into $\hat{\zeta}$ using (3.1).

We can now substitute the $\hat{\zeta}_i$'s derived from (4.4) and (3.1) into the empirical adjoint equation (4.1) to estimate the direction of β. In order to establish the $\sqrt{n}$-consistency of this estimate, we need the following regu-

larity conditions:

Condition 3. The moments $E(y^4)$ and $E(\|\mathbf{x}\|^4)$ exist.

Condition 4. $\zeta(\mathbf{bx}, \mathbf{b})$ is continuous in $\mathbf{b}$.

Condition 5. There exists $B < \infty$ such that $\|\dot{\zeta}(\mathbf{bx}, \mathbf{b})\| \leq B$ for all $\mathbf{b}$ and $\mathbf{x}$, where $\dot{\zeta}$ is the derivative of ζ with respect to $\mathbf{bx}$.

Theorem 4.1. Assume the observed data $\{(y_i, \mathbf{x}_i'), i = 1, \ldots, n\}$ are sampled randomly from a PPRM (1.2) and satisfy Conditions 0–5. For the step function estimate of $\zeta(\mathbf{bx}, \mathbf{b})$ given by (4.4) and (3.1), the adjoint projection pursuit regression estimate $\hat{\beta}$ that solves (4.1) is $\sqrt{n}$-consistent for β up to a multiplicative scalar if the window size d_n is of order $O(1/\sqrt{n})$ or smaller.

The $\sqrt{n}$-consistency for the adjoint projection pursuit regression estimate is insensitive to the choice of the smoothing parameter: Theorem 4.1 guarantees the optimal convergence rate for any reasonably "small" window sizes. It can be seen from the proof that the asymptotic mean squared error is dominated by variance (bias is negligible) if d_n is of order $o(1/\sqrt{n})$. We do not have the tradeoff between bias and variance for the usual nonparametric regression methods. This is because $\hat{\beta}$ does not depend on the entire estimated curve $\hat{\zeta}(\mathbf{bx}, \mathbf{b})$; instead it depends on the estimated curve only through an integrated summary measure, $\text{cov}(\zeta, y)$.

It is a nontrivial numerical task to solve (4.1) or to find a fixed point for (4.2). If the dimensionality of $\mathbf{x}$ is very low, it might be possible to find $\hat{\beta}$ using a grid search. Otherwise we need an efficient algorithm to find $\hat{\beta}$. A reasonable algorithm is to iterate $\hat{\mathfrak{F}}$ to find its fixed point. More specifically, we define the algorithm as follows:

Step 0. Choose any initial estimate $\mathbf{b}$, say, the least squares estimate $\hat{\beta}_{\text{LS}}$.

Step 1. Compute the estimated modified regressor $\hat{\mathbf{x}}$ in (3.3).

Step 2. Regress y on $\hat{\mathbf{x}}$ to obtain $\hat{\beta}^{(1)}$.

Step 3. Take $\hat{\beta}^{(1)}$ as the new initial estimate and return to Step 1 until convergence, say, if the angle between $\mathbf{b}$ and $\hat{\beta}^{(1)}$ is smaller than a given threshold.

If the algorithm converges, the limit is a fixed point for $\hat{\mathfrak{F}}$. We do not have general results on the numerical properties of this algorithm. A heuristic argument is given below which suggests that when the initial value is good, the algorithm converges very fast: one iteration might be nearly adequate.

For simplicity, we consider the population version of the algorithm. Without loss of generality, we assume that the regressor $\mathbf{x}$ and the initial value $\mathbf{b}$ are orthonormalized as follows: $E(\mathbf{x}) = \mathbf{0}$, $\text{cov}(\mathbf{x}) = I_p$, and $\mathbf{b} = (1, 0, \ldots, 0)$. Assume that $\mathbf{b}$ is nearly collinear with β, so we can take the first-order Taylor approximation for the link function g:

$$E(y \mid \mathbf{x}) = g(\beta_1 x_1 + \beta_2 \mathbf{x}_2) \cong g(\beta_1 x_1) + \beta_2 \mathbf{x}_2 g'(\beta_1 x_1);$$

therefore, $\mathbf{b}_2^{(1)} \cong \beta_2 E[g'(\beta_1 x_1)\text{cov}(\mathbf{x}_2 \mid x_1)]\{E[\text{cov}(\mathbf{x}_2 \mid x_1)]\}^{-1}$. If $g'(\beta_1 x_1)$ is nearly uncorrelated with $\text{cov}(\mathbf{x}_2 \mid x_1)$, we have

$$\mathbf{b}_2^{(1)} \cong \beta_2 E[g'(\beta_1 x_1)]; \tag{4.5}$$

thus $\mathbf{b}_2^{(1)}$ is nearly collinear with β_2. The condition preceding (4.5) is not crucial. We can avoid it if we redefine the modified least squares regression: instead of using the unconditional least squares regression of y on $\tilde{\mathbf{x}}$, we use the conditional least squares regressions

$$\mathbf{b}_2^{(1)}(x_1) = \text{cov}(y, \tilde{\mathbf{x}}_2' \mid x_1)[\text{cov}(\tilde{\mathbf{x}}_2 \mid x_1)]^{-1} \cong \beta_2 g'(\beta_1 x_1),$$

then average them over x_1. We then have

$$\mathbf{b}_2^{(1)} = E[\mathbf{b}_2^{(1)}(x_1)] \cong \beta_2 E[g'(\beta_1 x_1)], \tag{4.5'}$$

without requiring that $g'(\beta_1 x_1)$ be nearly uncorrelated with $\text{cov}(x_2 \mid x_1)$.

It follows from either (4.5) or (4.5') that $\mathbf{b}^{(1)}$ is nearly a linear combination of $\mathbf{b}$ and β. Therefore, we need only search over the two-dimensional space spanned by $\mathbf{b}$ and $\mathbf{b}^{(1)}$ to find β. Since we only need the direction for β, this is actually a one-dimensional search over the unit circle in this two-dimensional space.

We might not even need to perform this one-dimensional search. The first component of $\mathbf{b}^{(1)}$, $b_1^{(1)}$, is given approximately by $b_1^{(1)} \cong E[x_1 g(\beta_1 x_1)] + \beta_2 E[g'(\beta_1 x_1) x_1 \mathbf{x}_2']'$. If $g'(\beta_1 x_1)$ is nearly uncorrelated with $x_1 \mathbf{x}_2'$, the second term is approximately 0. We then have

$$b_1^{(1)} \cong E[x_1 g(\beta_1 x_1)] = \beta_1 E[\beta_1 x_1 g(\beta_1 x_1)]/\text{var}(\beta_1 x_1). \tag{4.6}$$

If x_1 is nearly normal, the scalars $E[g'(\beta_1 x_1)]$ in (4.5) and $E[\beta_1 x_1 g(\beta_1 x_1)]/\text{var}(\beta_1 x_1)$ in (4.6) are nearly identical, and thus $\mathbf{b}^{(1)}$ is already nearly collinear with β. Otherwise, the former scalar equals $E[-\log(f(\beta_1 x_1))g(\beta_1 x_1)]$, where $f(\beta_1 x_1)$ is the density function for $\beta_1 x_1$. It is therefore possible to estimate the ratio between the two scalars, then adjust $b_1^{(1)}$ and $\mathbf{b}_2^{(1)}$ by these scalars to estimate β.

Remark 4.1. Condition 5 indicates that the curves $\zeta(\mathbf{bx}, \mathbf{b})$, parameterized by $\mathbf{bx}$, have a common speed limit. Both Conditions 4 and 5 probably follow from suitable smoothness conditions on the distribution function for $\mathbf{x}$; we have not been able to formulate this formally.

Remark 4.2. Since bias is dominated by variance for the adjoint projection pursuit regression estimate $\hat{\beta}$, we can derive inference procedures for $\hat{\beta}$ from the asymptotic variance. Alternatively, we can make inferences for $\hat{\beta}$ using resampling methods such as Efron's (1982) bootstrap method. Efron and Tibshirani (1986) and Efron (1988) applied the bootstrap method to the (standard) projection pursuit regression.

5. A SIMULATION STUDY

We now discuss in further detail the simulation study mentioned in Section 1. The study demonstrates the behavior of the adjoint projection pursuit regression estimate, using the modified least squares regression algorithm Steps 0–3. We assume the regressor $\mathbf{x}$ is distrib-

uted uniformly over the square ($-1 \le x_1 \le 1$, $-1 \le x_2 \le 1$). We consider two PPRM's:

$$y = (\beta\mathbf{x} + \varepsilon)^3, \qquad \varepsilon \mid \mathbf{x} \sim N(0, 1), \tag{5.1}$$

$$y = \beta\mathbf{x} + \varepsilon, \qquad \varepsilon \mid \mathbf{x} \sim N(0, 1), \tag{5.2}$$

where $\beta = (3, 1)$. We assume that we do not know the true link function, and therefore can estimate only the direction of β, that is, the ratio $r = \beta_2/\beta_1$. The true value is $r = \frac{1}{3}$.

For the standard linear model (5.2), the least squares linear regression is unbiased and efficient. The value of R^2 for the linear model is $10/13 \approx .77$. For the cubic model (5.1), we should take the cubic root transformation for y. If we were unaware of this, and took the least squares linear regression of y on $\mathbf{x}$, the estimate would be substantially biased. It is easy to verify that the population value is $\beta_{LS,2}/\beta_{LS,1} = 21/47 \approx .4468$, which is quite different from the true value $\frac{1}{3}$. Our simulation results indicate that the modified least squares regression algorithm corrects this bias essentially in just one iteration.

For each replicate of the simulation, we sample 400 observations from the uniform distribution for $\mathbf{x}$ and model (5.2). The regressors are redrawn for each replicate. For model (5.1), we take the cube of the y's generated for model (5.2). We then apply the modified least squares regression algorithm for four iterations, using the least squares slope as the initial value in Step 0. The simulation is replicated 1,000 times.

We determine the window size d by choosing the nominal number of slices, m. For an initial value $\mathbf{b}$, we divide the range of $\mathbf{bx}$, $(-|b_1| - |b_2|, |b_1| + |b_2|)$, into m equal-width slices, with $d = 2(|b_1| + |b_2|)/m$. Since the realized values of $\mathbf{bx}$ might not cover the entire range, the actual number of nonempty slices might be smaller than m. We take a wide range of nominal slice numbers: $m = 10$, 20, 40, 80, and 200. For $m = 10$, we smooth a lot: on the average, each slice has 40 observations. For $m = 200$, we smooth very little: on the average, each slice has only two observations.

The results for the simulation study are given in Tables 1–6. For each table, the column "m" designates the nominal slice number. The columns below "Iteration" designate the iteration number: 0 refers to the linear regression, 1 refers to the first iteration of the modified least squares regression, and 2, 3, and 4 refer to the second, third, and fourth iterations, respectively.

Tables 1 and 4 give the estimated mean values for $\hat{r} = \hat{\beta}_2/\hat{\beta}_1$. Tables 2 and 5 give the estimated mean squared error $\hat{E}[(\hat{r} - \frac{1}{3})^2]$. For the cubic model (5.1), the linear regression is biased; $E(\hat{r})$ is indistinguishable from its population value $r \approx .4468$. The first iteration eliminates most of the bias; the remaining bias is small (about 5%), but significant relative to the Monte Carlo precision available. The mean squared error is reduced from the linear regression by a factor of 4–6. Further iterations reduce the bias and the mean squared error more, on a smaller scale.

For the standard linear model (5.2), the linear regression is unbiased for β; $E(\hat{r})$ is indistinguishable from the true value $r = \frac{1}{3}$. The adjoint projection pursuit regression estimates also exhibit little or no bias. The only bias (about 1%) detectable under the available Monte Carlo precision occurs when we smooth a lot ($m = 10$). The mean squared errors for the adjoint projection pursuit regression estimates are larger than that for the linear regression estimate; this is to be expected since the latter is efficient. Unless we smooth very little ($m = 200$), the loss of precision is minor: the mean squared error increases by at most 25%. For empirical applications in which we lack

Table 1. Estimated Direction for β: Cubic Model

	Iteration				
m	0	1	2	3	4
10	.44446	.34845	.34414	.34319	.34316
	(.00162)	(.00153)	(.00142)	(.00142)	(.00141)
20	.44446	.34435	.33734	.33663	.33646
	(.00162)	(.00153)	(.00141)	(.00141)	(.00140)
40	.44446	.34361	.33665	.33542	.33534
	(.00162)	(.00156)	(.00143)	(.00144)	(.00143)
80	.44446	.34455	.33717	.33688	.33614
	(.00162)	(.00162)	(.00151)	(.00149)	(.00148)
200	.44446	.34795	.33542	.33686	.33494
	(.00162)	(.00178)	(.00172)	(.00170)	(.00167)

NOTE: In each entry pair the top row is $E(\hat{r}) = E(\hat{\beta}_2/\hat{\beta}_1)$ and the bottom row is $\sigma(\hat{E}(\hat{r}))$.

Table 2. Mean Squared Error for $\hat{r}$: Cubic Model

	Iteration				
m	0	1	2	3	4
10	.014975	.002578	.002126	.002126	.002095
	(.000367)	(.000111)	(.000098)	(.000097)	(.000097)
20	.014975	.002446	.002011	.002009	.001993
	(.000367)	(.000107)	(.000092)	(.000091)	(.000093)
40	.014975	.002544	.002041	.002085	.002061
	(.000367)	(.000110)	(.000092)	(.000096)	(.000095)
80	.014975	.002757	.002297	.002235	.002192
	(.000367)	(.000120)	(.000109)	(.000099)	(.000102)
200	.014975	.003383	.002975	.002897	.002781
	(.000367)	(.000143)	(.000140)	(.000137)	(.000127)

NOTE: $\sigma(\widehat{\text{MSE}})$ in parenthesis.

Table 3. Average Relative Change in $\hat{r}$: Cubic Model

	Iteration			
m	0–1	1–2	2–3	3–4
10	.21679	.04342	.01784	.01178
	(.00161)	(.00106)	(.00056)	(.00045)
20	.22615	.04088	.01617	.01094
	(.00159)	(.00099)	(.00055)	(.00046)
40	.22802	.05084	.02501	.02090
	(.00169)	(.00122)	(.00079)	(.00079)
80	.22592	.06944	.04748	.04529
	(.00197)	(.00171)	(.00145)	(.00144)
200	.21881	.11156	.10811	.10612
	(.00256)	(.00284)	(.00351)	(.00309)

NOTE: σ(ARC) in parenthesis.

Table 4. Estimated Direction for β: Linear Model

	Iteration				
m	0	1	2	3	4
10	.33217 (.00094)	.32910 (.00100)	.32919 (.00099)	.32923 (.00100)	.32913 (.00100)
20	.33217 (.00094)	.33213 (.00100)	.33195 (.00100)	.33207 (.00100)	.33197 (.00100)
40	.33217 (.00094)	.33272 (.00101)	.33251 (.00100)	.33269 (.00101)	.33262 (.00100)
80	.33217 (.00094)	.33326 (.00105)	.33294 (.00103)	.33349 (.00106)	.33279 (.00104)
200	.33217 (.00094)	.33260 (.00124)	.33239 (.00118)	.33361 (.00122)	.33244 (.00122)

NOTE: In each entry pair the top row is $E(\hat{r}) = E(\hat{\beta}_2/\hat{\beta}_1)$ and the bottom row is $\sigma(\hat{E}(\hat{r}))$.

Table 6. Average Relative Change in $\hat{r}$: Linear Model

	Iteration			
m	0–1	1–2	2–3	3–4
10	.02778 (.00068)	.00979 (.00029)	.00615 (.00019)	.00475 (.00016)
20	.02578 (.00063)	.00834 (.00025)	.00488 (.00016)	.00388 (.00014)
40	.02959 (.00073)	.01309 (.00041)	.00983 (.00035)	.00890 (.00035)
80	.03808 (.00094)	.02787 (.00082)	.02653 (.00082)	.02603 (.00083)
200	.06222 (.00152)	.07189 (.00200)	.07328 (.00201)	.07537 (.00204)

NOTE: σ(ARC) in parenthesis.

precise prior information about the form of the link function, it might be worthwhile to pay this premium for the protection against possible nonlinearity in the link function.

For both models, the performance of the adjoint projection pursuit regression estimate is rather insensitive to the choice of m. Both $E(\hat{r})$ and mean squared error are indistinguishable for $m = 20, 40$, and 80. When we smooth a lot ($m = 10$), $E(\hat{r})$ is slightly different. When we smooth very little ($m = 200$), the mean squared error is substantially higher.

Tables 3 and 6 give the average relative change in $\hat{r}$, ARC $= E(|(\hat{r}^{(i+1)} - \hat{r}^{(i)})/\hat{r}^{(i)}|)$, from one iteration to the next. For the cubic model (5.1), a large movement occurs in the first iteration. The average changes in the later iterations are much smaller than $\sigma(\hat{r})$, except when we smooth very little ($m = 200$), in which case the algorithm still makes a fairly large movement in the fourth iteration. For the standard linear model (5.2), there is little change, especially after the first iteration. The average changes in the later iterations are an order of magnitude smaller than $\sigma(\hat{r})$, except when we smooth very little.

The modified least squares regression algorithm is very effective in reducing the bias in a few iterations, and does not increase the variability substantially. The algorithm converges fairly fast, and the performance appears to be insensitive to the amount of smoothing, except when we smooth very little ($m = 200$).

Table 5. Mean Squared Error for $\hat{r}$: Linear Model

	Iteration				
m	0	1	2	3	4
10	.000888 (.000040)	.001025 (.000046)	.001004 (.000045)	.001021 (.000047)	.001013 (.000046)
20	.000888 (.000040)	.001003 (.000045)	.000992 (.000045)	.001008 (.000046)	.000999 (.000046)
40	.000888 (.000040)	.001022 (.000045)	.000995 (.000045)	.001015 (.000046)	.001004 (.000045)
80	.000888 (.000040)	.001104 (.000048)	.001061 (.000048)	.001115 (.000051)	.001075 (.000048)
200	.000888 (.000040)	.001540 (.000071)	.001396 (.000062)	.001496 (.000071)	.001500 (.000071)

NOTE: $\sigma(\widehat{\text{MSE}})$ in parenthesis.

APPENDIX: TECHNICAL PROOFS

Proof of Theorem 2.1. The "if" part of the theorem follows from (1.2): conditioned on $\beta\mathbf{x}$, y and $\mathbf{x}$ are uncorrelated. To prove the converse, we assume without loss of generality that the link function is increasing. If $\mathbf{b}$ is not collinear with β, then

$$\beta \operatorname{cov}(\mathbf{x}, y \mid \mathbf{bx}) = \operatorname{cov}(\beta\mathbf{x}, g(\beta\mathbf{x}) \mid \mathbf{bx}) > 0$$

under Conditions 1 and 2. Therefore $\mathbf{b}$ does not solve (1.8).

Proof of Lemma 2.1. We partition $\mathbf{x}$, β, and $\mathbf{b}$ into $(x_1, \mathbf{x}_2')'$, et cetera. We need to prove that the conditional distribution of $\beta\mathbf{x}$ given $\mathbf{bx}$ is nondegenerate.

If $b_1 = 0$, it is sufficient to prove that $\beta\mathbf{x}$ given $\mathbf{x}_2$ is nondegenerate; that is, that $\beta_1 x_1$ given $\mathbf{x}_2$ is nondegenerate. By Condition 2″, we need only prove that x_1 given $\mathbf{x}_2$ is nondegenerate. This follows from Condition 2′: the support of x_1 given $\mathbf{x}_2$ is R.

We now assume $b_1 \neq 0$. Assume Condition 2 fails; thus $\beta\mathbf{x}$ given $\mathbf{bx}$ is degenerate for some $\mathbf{b} \not\propto \beta$. Note that $\beta\mathbf{x} = \mathbf{bx}\beta_1/b_1 + (\beta_2 - \beta_1\mathbf{b}_2/b_1)\mathbf{x}_2$, where $\beta_2 - \beta_1\mathbf{b}_2/b_1 \neq \mathbf{0}$. Therefore, the support of $\mathbf{x}_2$ given $\mathbf{bx}$ is contained in a $p - 2$-dimensional subspace. By Condition 2′, the support of $\mathbf{x}_2$ given $\mathbf{bx}$ is the same as the support of $\mathbf{x}_2$, namely, $\mathcal{S}$. Therefore, $\mathbf{x}_2$ is not of full rank, contradicting Condition 0: cov($\mathbf{x}$) is nonsingular. The contradiction proves Condition 2.

Proof of Theorem 2.2. First we verify that $\Sigma(\tilde{\mathbf{x}}\tilde{\mathbf{x}}') = \operatorname{cov}(\tilde{\mathbf{x}})$ is nonsingular, and hence $\mathbf{b}^{(1)}$ is well defined. We need to verify that $\operatorname{var}(\mathbf{v}\tilde{\mathbf{x}}) > 0$ for all $\mathbf{v} \neq \mathbf{0}$. Since the two terms on the right-hand side of (1.11) are uncorrelated, we have

$$\operatorname{var}(\mathbf{v}\tilde{\mathbf{x}}) = (\mathbf{v}\Sigma(\mathbf{xx}')\mathbf{b}')^2/\mathbf{b}\Sigma(\mathbf{xx}')\mathbf{b}' + E[\operatorname{var}(\mathbf{vx} \mid \mathbf{bx})].$$

If $\mathbf{v}$ is collinear with $\mathbf{b}$, the first term on the right-hand side is positive. For other $\mathbf{v}$'s, Condition 2‴ implies the second term is positive. Thus $\Sigma(\tilde{\mathbf{x}}\tilde{\mathbf{x}}')$ is nonsingular.

Note that $\mathbf{b}\tilde{\mathbf{x}} = \mathbf{bx}$, $\mathbf{b}\Sigma(\tilde{\mathbf{x}}\tilde{\mathbf{x}}') = \operatorname{cov}(\mathbf{b}\tilde{\mathbf{x}}, \tilde{\mathbf{x}}') = \operatorname{cov}(\mathbf{bx}, \tilde{\mathbf{x}}') = \mathbf{b}\Sigma(\mathbf{xx}')$, and

$$\mathbf{b}^{(1)} = c\mathbf{b} + \delta(\mathbf{b})'\Sigma(\tilde{\mathbf{x}}\tilde{\mathbf{x}}')^{-1}, \tag{A.1}$$

where c is a scalar. The "if" part of the theorem follows from (A.1) and Theorem 2.1. To prove the converse, we assume $\mathbf{b} \not\propto \beta$. We decompose β as follows: $\beta = a\mathbf{b} + \beta^{\perp}$, where $\beta^{\perp}\Sigma(\mathbf{xx}')\mathbf{b}' = 0$ and a is a scalar. By Conditions 1 and 2, we have $\beta^{\perp}\delta(\mathbf{b}) = E[\operatorname{cov}(\beta^{\perp}\mathbf{x}, g(a\mathbf{bx} + \beta^{\perp}\mathbf{x}) \mid \mathbf{bx})] \neq 0$. Since $\beta^{\perp}\Sigma(\mathbf{xx}')\mathbf{b}' = 0$,

this implies $\delta(\mathbf{b}) \not\propto \Sigma(\mathbf{xx}')\mathbf{b}'$. By (A.1), this implies $\mathbf{b}^{(1)} \not\propto \mathbf{b}$, which completes the proof.

Proof of Theorem 4.1. Without loss of generality, we restrict the proof to $\mathbf{b}$'s that satisfy (4.3) and $\mathbf{b}\beta' \geq 0$. The domain $\mathcal{B}$ of those $\mathbf{b}$'s is compact. We claim that for any $\varepsilon' > 0$, we can find $\varepsilon > 0$ such that $\|\delta(\mathbf{b})\| < \varepsilon$ implies $\|\mathbf{b} - \beta\| < \varepsilon'$. Assume the claim is false; we can find $\varepsilon' > 0$ such that for all $\varepsilon > 0$, there is a $\mathbf{b}$ satisfying $\|\delta(\mathbf{b})\| < \varepsilon$ and $\|\mathbf{b} - \beta\| > \varepsilon'$. By compactness and Condition 4, those $\mathbf{b}$'s have an accumulation point β^* such that $\|\delta(\beta^*)\| = 0$ and $\|\beta^* - \beta\| \geq \varepsilon'$. This contradicts the converse part of Theorem 2.1, and proves the claim.

According to the claim, we need only prove that $\hat{\delta}(\mathbf{b})$ converges to $\delta(\mathbf{b})$ uniformly over $\mathcal{B}$. By (3.2), $\hat{\delta}(\mathbf{b}) = (I - \hat{P})(\hat{\Sigma}(\mathbf{x}y) - \hat{\Sigma}(\tilde{\zeta}y))$, where $\hat{P} = \hat{\Sigma}(\mathbf{xx}')\mathbf{b}'\mathbf{b}/\mathbf{b}\hat{\Sigma}(\mathbf{xx}')\mathbf{b}'$. It is easy to verify that $\hat{P}$ converges to $P = \Sigma(\mathbf{xx}')\mathbf{b}'\mathbf{b}/\mathbf{b}\Sigma(\mathbf{xx}')\mathbf{b}'$ at $\sqrt{n}$ rate uniformly over $\mathbf{b} \in \mathcal{B}$, and $(I - P)\delta(\mathbf{b}) = \delta(\mathbf{b})$. Since $\hat{\Sigma}(\mathbf{x}y)$ also converges to $\Sigma(\mathbf{x}y)$ at $\sqrt{n}$ rate, we need only prove that $\hat{\Sigma}(\tilde{\zeta}y)$ converges to $\Sigma(\zeta y)$ at $\sqrt{n}$ rate uniformly over $\mathcal{B}$.

By (4.4), we have $\hat{\Sigma}(\tilde{\zeta}y) = (n-1)^{-1}\sum_{i=1}^n \tilde{\zeta}_i(y_i - \bar{y}) \doteq n^{-1}\sum_{ij} \mathbf{x}_j(y_i - \bar{y})w_{ij}$, where $w_{ij} = 1_{ih}1_{jh}/n\hat{p}_h$. It follows that $\hat{\Sigma}(\tilde{\zeta}y) \doteq \sum_h \hat{p}_h\bar{\mathbf{x}}_h\bar{y}_h - \bar{\mathbf{x}}\bar{y}$, where $\bar{\mathbf{x}}_h$ ($\bar{y}_h$) is the sample average of the $\mathbf{x}_i$'s (y_i's) for observations with $\mathbf{bx}_i \in s_h$. Since $\bar{\mathbf{x}}\bar{y}$ converges to $E(\zeta)E(y)$ at $\sqrt{n}$ rate, we need only prove that $\sum_h \hat{p}_h\bar{\mathbf{x}}_h\bar{y}_h$ converges to $E(\zeta y)$.

Let $\zeta_h = E[\zeta(\mathbf{bx}, \mathbf{b}) \mid \mathbf{bx} \in s_h]$, $\eta_h = E(y \mid \mathbf{bx} \in s_h)$, and $p_h = E(\hat{p}_h) = \Pr(\mathbf{bx} \in s_h)$. We have $\sum_h \hat{p}_h\bar{\mathbf{x}}_h\bar{y}_h - E(\zeta y) \doteq$ bias + error, where bias $= \sum_h p_h\zeta_h\eta_h - E(\zeta y)$ and

$$\text{error} = \sum_h [(\hat{p}_h - p_h)\zeta_h\eta_h + \hat{p}_h(\bar{\mathbf{x}}_h - \zeta_h)\eta_h + \hat{p}_h\zeta_h(\bar{y}_h - \eta_h)].$$

Since $E(\text{error}) = E[E(\text{error} \mid \hat{p})] = \mathbf{0}$, we have

$$\begin{aligned} E(\|\text{error}\|^2) &\leq \sum_h \{n^{-1}p_h\|\zeta_h\|^2\eta_h^2 + E[\hat{p}_h^2\eta_h^2 \operatorname{tr}(\operatorname{cov}(\bar{\mathbf{x}}_h \mid \hat{p}_h))] \\ &\qquad + E[\hat{p}_h^2\|\zeta_h\|^2 \operatorname{var}(\bar{y}_h \mid \hat{p}_h)]\} \\ &\doteq n^{-1}\sum_h \{p_h\|\zeta_h\|^2\eta_h^2 + p_h\eta_h^2 \operatorname{tr}(\operatorname{cov}(\mathbf{x} \mid \mathbf{bx} \in s_h)) \\ &\qquad + p_h\|\zeta_h\|^2 \operatorname{var}(y \mid \mathbf{bx} \in s_h)\} \\ &\leq 3[E(y^4)E(\|\mathbf{x}\|^4)]^{1/2}/n. \end{aligned}$$

The last inequality follows from the Cauchy–Schwarz inequality and Jensen's inequality. Since bias $= -E[\operatorname{cov}(\zeta, y \mid \mathbf{bx} \in s_h)]$, we have

$$\begin{aligned} \|\text{bias}\|^2 &\leq E\{\|\operatorname{cov}(\zeta, y \mid \mathbf{bx} \in s_h)\|^2\} \\ &\leq \sum_{j=1}^p E\{\operatorname{var}(\zeta(\mathbf{bx}, \mathbf{b})_j \mid \mathbf{bx} \in s_h)\operatorname{var}(y \mid \mathbf{bx} \in s_h)\}, \end{aligned}$$

where $\zeta(\cdot, \cdot)_j$ denotes the jth component of $\zeta(\cdot, \cdot)$. By Condition 5, we have

$$\operatorname{var}(\zeta(\mathbf{bx}, \mathbf{b})_j \mid \mathbf{bx} \in s_h) \leq B^2d_n^2/4,$$
$$\|\text{bias}\|^2 \leq pB^2d_n^2E(y^2)/4.$$

If $d_n = o(1)$, both bias and error converge to $\mathbf{0}$ uniformly over $\mathcal{B}$; therefore, $\hat{\beta}$ is consistent for β up to a multiplicative scalar. If $d_n = O(1/\sqrt{n})$, the convergence rate is $\sqrt{n}$; bias and error are of the same order. If $d_n = o(1/\sqrt{n})$, the convergence rate is $\sqrt{n}$, but bias is of a smaller order than error.

[*Received January 1989. Revised March 1990.*]

REFERENCES

Brillinger, D. R. (1977), "The Identification of a Particular Nonlinear Time Series System," *Biometrika,* 64, 622–654.

——— (1982), "A General Linear Model with 'Gaussian' Regressor Variables," in *A Festschrift for Erich L. Lehmann,* eds. P. J. Bickel, K. A. Doksum, and J. L. Hodges, Belmont, CA: Wadsworth.

Brown, R. F. (1971), *The Lefschetz Fixed Point Theorem,* Glenview, IL: Scott, Foresman.

Chung, C. F., and Goldberger, A. S. (1984), "Proportional Projections in Limited Dependent Variable Models," *Econometrica,* 52, 531–534.

Cook, R. D., and Weisberg, S. (1982), *Residuals and Influence in Regression,* New York: Chapman & Hall.

Duan, N., and Li, K.-C. (1987), "Distribution-Free and Link-Free Estimation for the Sample Selection Model," *Journal of Econometrics,* 35, 25–35.

——— (in press), "Slicing Regression: A Link-Free Regression Method," *The Annals of Statistics.*

Efron, B. (1982), "The Jackknife, the Bootstrap, and Other Resampling Plans," *CBMS–NSF Regional Conference Series in Applied Mathematics 38,* Philadelphia: Society for Industrial and Applied Mathematics.

——— (1988), "Computer-Intensive Methods in Statistical Regression," *SIAM Review,* 30, 421–449.

Efron, B., and Tibshirani, R. (1986), "Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy," *Statistical Science,* 1, 54–77.

Friedman, J. H., and Stuetzle, W. (1981), "Projection Pursuit Regression," *Journal of the American Statistical Association,* 76, 817–823.

——— (1985), Discussion of "Projection Pursuit," by P. J. Huber, *The Annals of Statistics,* 13, 475–481.

Goldberger, A. S. (1981), "Linear Regression After Selection," *Journal of Econometrics,* 15, 357–366.

Greene, W. (1981), "On the Asymptotic Bias of Ordinary Least Squares Estimates of the Tobit Model," *Econometrica,* 49, 505–514.

——— (1983), "Estimation of Limited Dependent Variable Models by Ordinary Least Squares and the Method of Moments," *Journal of Econometrics,* 21, 195–212.

Hall, P. (1989), "On Projection Pursuit Regression," *The Annals of Statistics,* 17, 573–588.

Halmos, P. R. (1958), *Finite-Dimensional Vector Spaces,* Princeton: Van Nostrand.

Han, A. K. (1987), "Nonparametric Analysis of a Generalized Regression Model: The Maximum Rank Correlation Estimator," *Journal of Econometrics,* 35, 303–316.

Härdle, W., and Stoker, T. M. (1989), "Investigating Smooth Multiple Regression by the Method of Average Derivatives," *Journal of the American Statistical Association,* 84, 986–995.

Huber, P. J. (1981), *Robust Statistics,* New York: John Wiley.

——— (1985), "Projection Pursuit" (with discussion), *The Annals of Statistics,* 13, 435–525.

Ichimura, H. (1989), "Semiparametric Least Squares Estimation of Single Index Models," mimeograph, University of Minnesota, Dept. of Economics.

Li, K.-C., and Duan, N. (1989), "Regression Analysis Under Link Violation," *The Annals of Statistics,* 17, 1009–1052.

McCullagh, P., and Nelder, J. A. (1983), *Generalized Linear Models,* New York: Chapman & Hall.

Manski, C. F. (1975), "Maximum Score Estimation of the Stochastic Utility Models of Choice," *Journal of Econometrics,* 3, 205–228.

——— (1985), "Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator," *Journal of Econometrics,* 27, 313–334.

Powell, J. L., Stock, J. H., and Stoker, T. M. (1989), "Semiparametric Estimation of Index Coefficients," *Econometrica,* 57, 1403.

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.

Stoker, T. M. (1986), "Consistent Estimation of Scaled Coefficients," *Econometrica,* 54, 1461–1481.

——— (1989), "Equivalence of Direct, Indirect and Slope Estimators of Average Derivatives," to appear in *Nonparametric and Semiparametric Methods in Econometrics and Statistics,* eds., W. A. Barnett, J. Powell, and G. Tauchen, New York: Cambridge University Press.

Tukey, J. W. (1949), "One Degree of Freedom for Nonadditivity," *Biometrics,* 5, 232–242.

White, H. (1981), "Consequences and Detection of Misspecified Nonlinear Regression Models," *Journal of the American Statistical Association,* 76, 419–433.