An Analysis of Transformations for Additive Nonparametric Regression

Oliver B. LINTON, Rong CHEN, Naiysin WANG, and Wolfgang HÄRDLE

We consider a nonparametric regression model with a parametric family of dependent variable transformations, one of which induces additive covariate effects. We estimate the additive regression effects using the integration method and estimate the transformation parameter from a profiled instrumental variable and pseudolikelihood criterion. The asymptotic distributions of the parameter and regression estimates are given. The practical performance is investigated via an application.

KEY WORDS: Box–Cox transformation; Dimensionality reduction; Kernels.

1. INTRODUCTION

Taking transformations of the data has been an integral part of statistical practice for many years. Transformations have been used to aid interpretability as well as to improve statistical performance. An important contribution to this methodology was made by Box and Cox (1964), who proposed a parametric power family of transformations that nested the logarithm and the level. They suggested that the power transformation, when applied to the dependent variable in a linear regression setting, might induce normality, error variance homogeneity, and additivity of effects. They proposed estimation methods for the regression and transformation parameters. Carroll and Ruppert (1984) suggested applying this and other transformations to both dependent and independent variables. A number of other dependent variable transformations have been suggested—for example, the Zellner–Revankar transform (see Zellner and Revankar 1969). The transformation methodology has been quite successful, and a large literature now exists on this subject for parametric models (see Carroll and Ruppert 1988). There are also a number of applications to economics data (see Ehrlich 1977; Heckman and Polachek 1974; Hulten and Wykoff 1981; Zarembka 1968; Zellner and Revankar 1969).

We work with transformations inside a regression setting. For many data, the linearity of covariate effect after transformation may be too strong. For example, a respected study of the effects of schooling and experience on earnings (Heckman and Polachek 1974, p. 350) found that although their data supported the logarithmic transformation of their dependent variable earnings, it was "somewhat less clear on the functional form for the independent variables." We thus consider a more general specification, allowing for nonparametric covariate effects. Let (X, Y) be a random variable with $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}$, and let $\{(X_i, Y_i)\}_{i=1}^n$ be an iid sample from this population. Consider the estimation of the regression function $m(x) = E\ (Y|X = x)$. Ibragimov and Hasminskii (1980) and Stone (1980, 1982) showed that the optimal rate for estimating m is $n^{-l/(2l+d)}$, with l a measure of the smoothness of m. This rate of convergence can be very slow for large dimensions d. One way of achieving better rates of convergence is through additive modeling. An additive structure for m is a regression function of the form $m(x) = c + \sum_{\alpha=1}^d m_\alpha(x_\alpha)$, where $x = (x_1, \ldots, x_d)'$ are the d-dimensional predictor variables and m_α are one-dimensional nonparametric functions with $E\{m_\alpha(X_\alpha)\} = 0$. Stone (1986) showed that for such regression curves, the optimal rate for estimating m is the one-dimensional rate of convergence $n^{-l/(2l+1)}$. Thus one speaks of dimensionality reduction through additive modeling.

We examine a semiparametric model that combines a parametric transformation with the flexibility of an additive nonparametric regression function. For a parametric family of transforms $\theta_\lambda(\cdot), \lambda \in \Lambda \subset \mathbb{R}$, define the regression functions $m_\lambda(x) = E\ \{\theta_\lambda(Y)|X_1 = x_1, \ldots, X_d = x_d\}$, and suppose that for some unique $\lambda_0 \in \Lambda$,

$$m_{\lambda_0}(x) = c + \sum_{\alpha=1}^d g_\alpha(x_\alpha), \tag{1}$$

where $g_\alpha(\cdot)$ are of unknown form with $E\{g_\alpha(X_\alpha)\} = 0, \alpha = 1, \ldots, d$. This model was previously addressed by Hastie and Tibshirani (1990, p. 187). They suggested estimation procedures based on the iterative backfitting method; however, they did not provide many results about the statistical properties of their procedures. Breiman and Friedman (1985) suggested a generalization of (1), called alternating conditional expectation (ACE), in which the transformation is not restricted to be parametric. Again, plausible estimation procedures are available, but little is known about their statistical properties. Finally, Tibshirani (1988) suggested a modification of the ACE estimation algorithm, called additivity and variance stabilization (AVAS). This has several advantages over the ACE algorithm: in particular, it reproduces model transformations and is equivariant under monotone transformations.

Oliver Linton is Associate Professor of Economics, Cowles Foundation for Research in Economics, Yale University, New Haven, CT 06520. Rong Chen is Associate Professor of Statistics and Naiysin Wang is Assistant Professor of Statistics, Department of Statistics, Texas A&M University, College Station, TX 77843. Wolfgang Härdle is Professor at the Institut für Statistik und Ökonometrie, Humboldt-Universität zu Berlin, Germany. The authors would like to thank Jens Perch Nielsen and Stefan Sperlich for helpful comments, and Nick Hengartner for sharpening their view on Section 2. They thank the National Science Foundation, the North Atlantic Treaty Organization, and the Sonderforschungsbereich 373 for financial support. GAUSS software is available from the first author on request.

Journal of the American Statistical Association
December 1997, Vol. 92, No. 440, Theory and Methods

Independently, Linton and Nielsen (1995), Newey (1994), and Tjøstheim and Auestad (1991, 1994), introduced a new method for estimating additive nonparametric models (when the transformation is known) that is based on direct integration of an initial pilot smoother of the regression function. It has been possible to prove a central limit theorem for this estimator because of its mathematical tractability. We extend their theory to the semiparametric model (1). We use the semiparametric profile method described by Bickel, Klaassen, Ritov, and Wellner (1993) to estimate the parameters $\lambda_0 \in \Lambda \subset \mathbb{R}$ in (1). This uses a preliminary integration estimate of the additive components of m_λ. We also show how to estimate the constant c, the univariate component functions $g_\alpha(\cdot)$, and the nonparametric regression function $m_\lambda(\cdot)$. We derive the asymptotic distributions of our estimators under standard regularity conditions. The estimator of λ_0 is root-n consistent, whereas the estimates of $g_\alpha(x_\alpha)$ and $m_\lambda(x)$ are consistent at the one-dimensional rate of $n^{2/5}$.

The article is organized as follows. In Section 2 we give an outline of the integration method and explain how it can circumvent the curse of dimensionality. In Section 3 we define the estimation procedures for both parametric and nonparametric parts. In Section 4 we give the asymptotic properties of the procedures. In Section 5 we illustrate our methods in an application and on simulated data. The Appendixes contain the proofs of all results.

2. THE INTEGRATION METHOD

Suppose that we have some pilot estimator $\hat{f}(x_1, x_2)$ of a function $f(x_1, x_2)$ of the scalar x_1 and the $d-1$ variables x_2. Define the integration estimator

$$\tilde{\varphi}_1(x_1) = \int \hat{f}(x_1, x_2)\, dQ(x_2) \tag{2}$$

for any $d-1$-dimensional signed measure Q. Usually, Q will be a probability measure. If the estimated function were additive—that is, $f(x_1, x_2) = f_1(x_1) + f_2(x_2)$—then $\tilde{\varphi}_1(\cdot)$ would consistently estimate $f_1(\cdot)$ plus a constant that is the same for all x_1. Likewise, if the function were multiplicative—that is, $f(x_1, x_2) = f_1(x_1) f_2(x_2)$—then $\tilde{\varphi}_1(\cdot)$ would consistently estimate $f_1(\cdot)$ times a constant that is the same for all x_1. This is the basic idea behind the integration method. Of course, it is also important to know whether one can also obtain the one-dimensional rate of convergence for $\tilde{\varphi}_1(\cdot)$. Intuitively, this will hold because integration is averaging and so reduces variance; a proof of this was first given by Newey (1994). However, the dependency of $\tilde{\varphi}_1(x_1)$ on the high-dimensional pilot smoother $\hat{f}(x_1, x_2)$ suggests that the $\tilde{\varphi}_1(x_1)$ may suffer somewhat from the curse of dimensionality. This viewpoint is supported by the fact that in the currently available central limit theorem proofs (see Chen, Härdle, Linton, and Severance-Lossin 1996; Linton and Härdle 1996; Newey 1994), it has been necessary to use bias reduction when the dimensions were greater than four. As usual, the conditions given there were sufficient but not necessary; in our opinion, they can be weakened considerably. To support this viewpoint, we give a different interpretation of the integration estimator.

To focus ideas, we look at density estimation. Let

$$\hat{f}(x_1, x_2) = \frac{1}{n h_1 h_2^{d-1}} \sum_{i=1}^{n} K_1\left(\frac{x_1 - X_{1i}}{h_1}\right) K_2\left(\frac{x_2 - X_{2i}}{h_2}\right) \tag{3}$$

be a product kernel density estimate based on an iid sample $\{X_i\}_{i=1}^n$ partitioned as earlier. When the dimensions d are large, $\hat{f}(x_1, x_2)$ has very poor performance. Now suppose that we integrate $\hat{f}(x_1, x_2)$ with respect to Lebesgue measure, that is, take $dQ(x_2) = dx_2$ in (2), and let

$$\begin{aligned}\tilde{\varphi}_1(x_1) &= \int \hat{f}(x_1, x_2)\, dx_2 \\ &= \frac{1}{n h_1 h_2^{d-1}} \sum_{i=1}^{n} K_1\left(\frac{x_1 - X_{1i}}{h_1}\right) \\ &\quad \times \int K_2\left(\frac{x_2 - X_{2i}}{h_2}\right) dx_2. \end{aligned} \tag{4}$$

By the usual change of variables, we obtain exactly

$$\tilde{\varphi}_1(x_1) = \frac{1}{n h_1} \sum_{i=1}^{n} K_1\left(\frac{x_1 - X_{1i}}{h_1}\right), \tag{5}$$

which is the one-dimensional kernel density estimate of $f_1(x_1) = \int f(x_1, x_2)\, dx_2$. Thus integration has taken a d-dimensional smoother into a one-dimensional smooth and completely eliminated the second bandwidth h_2.

In regression the pilot estimators typically used are more complicated than (3), and we cannot go from step (4) to (5) exactly; some approximation argument is necessary. However, we would like to suggest that even in this case the integration estimator can be interpreted as a one-dimensional smooth. Recent work by Hengartner (1996) confirms this view.

3. ESTIMATION

To estimate the quantities of interest, we use a multistage procedure. In the first part we estimate the transformation using the semiparametric profile method, discussed by Bickel et al. (1993). This requires an estimate of the regression functions for each parameter value. In the second part we estimate the additive regression function using the estimated transformation.

1. For each λ, m_λ is estimated by a multidimensional kernel smoother $\hat{m}_\lambda$.

2. For each direction α, the pilot estimate is then integrated to obtain an estimate of the individual effect of X_α on Y in the λ scale. The individual effect estimates are combined to form an "additive reconstruction" $\tilde{m}_\lambda$. A detailed description of the integration method is given in Section 3.1.

3. We define a criterion function for λ depending on the profiled estimate $\tilde{m}_\lambda$. We choose $\hat{\lambda}$ to optimize this criterion.

4. Finally, we construct a new kernel smoother $\hat{m}_{\hat{\lambda}}$ using the estimated $\hat{\lambda}$, and obtain the final additive estimate $\tilde{m}_{\hat{\lambda}}$ by integration.

Note that in Steps 1–3, we are concerned only with the properties of $\hat{\lambda}$.

3.1 Nonparametric Estimation

Step 1 involves estimating a regression function for the sample $\{X_i, \theta_\lambda(Y_i)\}_{i=1}^n$ for any λ. We estimate $m_\lambda(x)$ using the multidimensional Nadaraya–Watson estimator,

$$\hat{m}_\lambda(x) = \frac{\sum_{j=1}^n \prod_{\alpha=1}^d K_\alpha\left(\frac{x_\alpha - X_{\alpha j}}{h_\alpha}\right)\theta_\lambda(Y_j)}{\sum_{l=1}^n \prod_{\alpha=1}^d K_\alpha\left(\frac{x_\alpha - X_{\alpha l}}{h_\alpha}\right)}, \tag{6}$$

where $K_1, \ldots, K_d$ are univariate kernels and $h_1(n), \ldots, h_d(n)$ are bandwidth sequences, one for each direction. When our ultimate goal is to estimate λ, we take common bandwidth and kernel for notational simplicity; otherwise, we allow both kernel and bandwidth to vary with direction. For each $\alpha = 1, \ldots, d$, partition $x = (x_\alpha, x_{\underline{\alpha}})$, where x_α is a one-dimensional direction of interest and $x_{\underline{\alpha}}$ is a $d-1$-dimensional nuisance direction; do likewise with $X = (X_\alpha, X_{\underline{\alpha}})$ and $X_i = (X_{\alpha i}, X_{\underline{\alpha} i})$. Now define

$$\hat{\gamma}_\alpha(x_\alpha; \lambda) = n^{-1}\sum_{i=1}^n \hat{m}_\lambda(x_\alpha, X_{\underline{\alpha} i}). \tag{7}$$

This is the empirical integration estimator of Chen et al. (1996). Let p be the joint density of $X_1, \ldots, X_d$, let p_α be the marginal density of X_α, and let $p_{\underline{\alpha}}$ be the joint density of $X_{\underline{\alpha}}$. Then $\hat{\gamma}_\alpha(x_\alpha; \lambda)$ consistently estimates the population quantity

$$\gamma_\alpha(x_\alpha; \lambda) = \int m_\lambda(x_\alpha, x_{\underline{\alpha}}) p_{\underline{\alpha}}(x_{\underline{\alpha}})\, dx_{\underline{\alpha}}.$$

In fact, under the additive model (1), $\gamma_\alpha(x_\alpha; \lambda_0) = c + g_\alpha(x_\alpha)$. More generally, one could interpret $\gamma_\alpha(x_\alpha; \lambda)$ as one aspect of the univariate effect of the covariate on the transformed dependent variable. This is the point of view expressed by Newey (1994) in connection with econometric applications. Now, let

$$\tilde{m}_\lambda(x) = \sum_{\alpha=1}^d \hat{\gamma}_\alpha(x_\alpha; \lambda) - (d-1)\hat{c}_\lambda, \tag{8}$$

where $\hat{c}_\lambda = n^{-1}\sum_{i=1}^n \theta_\lambda(Y_i)$. Then $\tilde{\mathbf{m}}_\lambda(x)$ consistently estimates

$$\bar{m}_\lambda(x) = \sum_{\alpha=1}^d \gamma_\alpha(x_\alpha; \lambda) - (d-1)c_\lambda,$$

where $c_\lambda = E\{\theta_\lambda(Y)\}$. Note that $\bar{m}_{\lambda_0}(x) = m_{\lambda_0}(x)$; that is, combining the covariate effects in the λ_0 scale gives us the regression function. However, $\bar{m}_\lambda(x) \neq m_\lambda(x)$ for $\lambda \neq \lambda_0$. This information is used to identify λ_0. We implement (6), and hence (7) and (8), in several different ways according to our purpose.

3.2 Estimation of λ

To estimate λ, we use a nonlinear instrumental variable procedure. The assumed properties of the mean are used to generate first-order conditions that identify the parameter λ. This method has been widely used in estimating simultaneous equation systems and in certain dynamic models, and was used by Amemiya and Powell (1981) in the parametric Box–Cox model. [See Newey and McFadden (1994 p. 2116) for background references and discussion. See also Angrist, Imbens, and Rubin (1996) for a connection between instrumental variables and the Rubin causal model used in causal inference.]

Let $\mathbf{Z}_i, i = 1, \ldots, n$, be a J by 1 vector of iid instruments with the property that

$$E[\{\theta_{\lambda_0}(Y_i) - \bar{m}_{\lambda_0}(X_i)\}|\mathbf{Z}_i] = 0 \quad \text{a.s.}$$

for a unique λ_0. The instruments must satisfy an additional identification condition, which we discuss in Section 4. Examples of valid instruments include functions of the X's themselves. Let $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_n)', \boldsymbol{\Theta}_\lambda = (\theta_\lambda(Y_1), \ldots, \theta_\lambda(Y_n))'$, and $\tilde{\mathbf{M}}_\lambda = (\tilde{m}_\lambda(X_1), \ldots, \tilde{m}_\lambda(X_n))'$, and define $\hat{\lambda}$ to be any minimizer of

$$Q_n(\lambda) = \{(\boldsymbol{\Theta}_\lambda - \tilde{\mathbf{M}}_\lambda)'\mathbf{Z}/n\}\mathbf{W}_n\{\mathbf{Z}'(\boldsymbol{\Theta}_\lambda - \tilde{\mathbf{M}}_\lambda)/n\}, \tag{9}$$

where $\mathbf{W}_n$ is a sequence of J by J weighting matrices satisfying $\mathbf{W}_n \xrightarrow{p} \mathbf{W}$ a positive definite matrix. A simple feasible choice for $\mathbf{W}_n$ is the identity matrix. We discuss how more efficiency can be obtained by wiser choice of $\mathbf{W}_n$. In our application we computed $\hat{\lambda}$ by grid search, but iterative techniques such as Newton–Raphson also work well here.

An alternative method for estimating λ is the full Gaussian pseudolikelihood (see Box and Cox 1964). After profiling out a variance parameter, this amounts to choosing $\tilde{\lambda}$ that minimizes

$$L_n(\lambda) = n^{-1}\sum_{i=1}^n \ln J_\lambda(Y_i) - \ln\{n^{-1}(\boldsymbol{\Theta}_\lambda - \tilde{\mathbf{M}}_\lambda)'(\boldsymbol{\Theta}_\lambda - \tilde{\mathbf{M}}_\lambda)\},$$

where $J_\lambda(\cdot)$ is the Jacobian of the transformation $y \to \theta_\lambda(y)$. (See Carroll and Ruppert 1988, pp. 124–127, for more discussion of this method in the parametric regression problem. A robust version of this procedure was given in Carroll 1980.)

An alternative way to proceed is to profile both c and λ out by using $\tilde{m}_{\lambda,c}(x) = \sum_{\alpha=1}^d \hat{\gamma}_\alpha(x_\alpha; \lambda) - (d-1)c$ in place of $\tilde{m}_{\lambda,c}(x)$, and to construct profiled criteria $Q_n(\lambda, c)$ or $L_n(\lambda, c)$, which are then optimized with respect to both c and λ. We restrict attention to the simpler procedure based on profiling only λ. In this case, given estimates of λ, we define estimators of $c, g_\alpha(\cdot)$, and $m_{\lambda_0}(\cdot)$ as in Section 3.1.

3.3 Discussion

Among the many examples of interest, the following ones are used most commonly:

Box–Cox: $\theta_\lambda(y) = (y^\lambda - 1)/\lambda; J_\lambda(y) = y^{\lambda-1}$
Zellner–Revankar: $\theta_\lambda(y) = \ln y + \lambda y^2; J_\lambda(y) = y^{-1} + 2\lambda y$
arcsinh: $\theta_\lambda(y) = \sinh^{-1}(\lambda y)/\lambda; J_\lambda(y) = (1 + \lambda y^2)^{-1/2}$.

The arcsinh transform was discussed by Johnson (1949) and more recently by Robinson (1991). The main advantage of the arcsinh transform is that it works for y taking any value, whereas the Box–Cox and the Zellner–Revankar transforms are defined only if y is positive. For these transformations, the error term cannot be normally distributed except for a few isolated parameters, and so the Gaussian likelihood is misspecified. In fact, as Amemiya and Powell (1981) pointed out, the resulting estimators (in the parametric case) are inconsistent only when $n \to \infty$. This is the main advantage of the instrumental variable criterion; it is consistent even for these transformations under general sampling conditions. However, we note that Bickel and Doksum (1981) established consistency of the Gaussian pseudolikelihood procedure when both $n \to \infty$ and $\sigma \to 0$, where σ is the scale of the error term.

We subsequently focus on the instrumental variable procedure in our treatment of the asymptotics.

4. ASYMPTOTIC PROPERTIES

We develop asymptotic approximations for the instrumental variable estimator $\hat{\lambda}$, as $n \to \infty$. Under similar conditions, $\tilde{\lambda}$ is also consistent for some transformations (although not the Box–Cox as discussed earlier). Finally, we establish the asymptotic distribution of the covariate effects. We give three theorems. As far as the properties of $\hat{\lambda}$ are concerned, we use a common kernel K and bandwidth h in (6). For consistency of $\hat{\lambda}$, second-order kernels suffice, whereas to achieve root-n consistency we must use bias reduction in all directions when the dimensions exceed 3. This type of condition is common in other semiparametric situations, see for example Robinson (1988). In the last theorem, which is about the properties of $\hat{\gamma}_\alpha(\cdot)$, we allow for bias reduction in directions other than α when the dimensions exceed four (as in Linton and Härdle 1996).

4.1 Consistency of $\hat{\lambda}$

Our proof uses primarily that $\tilde{m}_\lambda(x)$ is uniformly consistent as an estimator of $\bar{m}_\lambda(x)$; no rates are needed. For this reason, we take the standard kernel estimator (6) with common bandwidth h and kernel K for each direction. We work with iid instruments $\mathbf{Z}_i$ that are mean 0 (this is without loss of generality and can be achieved by subtracting off sample means) and make some additional technical assumptions, stated in Appendix A.

Theorem 1. Suppose that conditions A1–A6 given in Appendix A hold. Suppose further that the following condition holds:

A7: For all $\lambda \neq \lambda_0$ in a neighborhood $\mathcal{N}(\lambda_0)$ of λ_0, $E[\{\bar{m}_\lambda(X_i) - m_\lambda(X_i)\}|\mathbf{Z}_i] \neq 0$. Then $\hat{\lambda}$ is locally consistent.

Remark. The identification condition A7 is difficult to verify from primitive conditions, as is true in related situations (see Newey and McFadden 1994, p. 2127). Nevertheless, we expect that condition A7 is obtained through additivity. For $\lambda \neq \lambda_0$, the regression function m_λ is not additive, and so the difference between the additive reconstruction $\tilde{m}_\lambda$ and m_λ should, in the limit, be correlated with the instruments.

4.2 Asymptotic Normality of $\hat{\lambda}$

We use a common kernel K and bandwidth $h(n)$ for each direction α. We need some additional conditions given in Appendix A. Specifically, to obtain $n^{1/2}$ consistency for $\hat{\lambda}$, it is necessary to use bias-reducing kernels of order $q \geq 2$ to estimate the regression function when the dimensions $d > 4$. Define the mean 0 iid random variables $\varepsilon_\lambda(i) = \theta_\lambda(Y_i) - m_\lambda(X_i), i = 1, \ldots, n$, and also let $\tilde{\varepsilon}_\lambda(i) = \theta_\lambda(Y_i) - \tilde{m}_\lambda(X_i)$ for any λ. Also, let $r_j(\lambda) = E[\partial^j/\partial\lambda^j\{m_\lambda(X_i) - \bar{m}_\lambda(X_i)\}\mathbf{Z}_i], j = 0, 1, 2$, and set $r_{j0} = r_j(\lambda_0)$. Finally, let

$$\Omega = E[\varepsilon^2_{\lambda_0}(i)\{\mathbf{Z}_i - \mathbf{Z}_i^*\}\{\mathbf{Z}_i - \mathbf{Z}_i^*\}'],$$

where

$$\mathbf{Z}_i^* = \sum_{\alpha=1}^{d} E(\mathbf{Z}_i|X_{\alpha i})\,\frac{p_\alpha(X_{\alpha i})p_{\underline{\alpha}}(X_{\underline{\alpha} i})}{p(X_i)}.$$

Theorem 2. Suppose that conditions A1–A7 and B1–B6 given in the Appendix hold. Then

$$n^{1/2}(\hat{\lambda} - \lambda_0) \Rightarrow \mathrm{N}\{0, (r_{10}'\mathbf{W}r_{10})^{-2}r_{10}'\mathbf{W}\Omega\mathbf{W}r_{10}\}. \quad (10)$$

We can consistently estimate r_{10} by $\hat{r}_1 = n^{-1}\sum_{i=1}^n \mathbf{Z}_i(\partial/\partial\lambda)\{\hat{m}_{\hat{\lambda}}(X_i) - \tilde{m}_{\hat{\lambda}}(X_i)\}$ and Ω by $\hat{\Omega} = n^{-1}\sum_{i=1}^n \mathbf{Z}_i\mathbf{Z}_i'\tilde{\varepsilon}^2_{\hat{\lambda}}(i)$, and thereby consistently estimate the asymptotic variance V by

$$\hat{V} = (\hat{r}_1'\mathbf{W}_n\hat{r}_1)^{-2}\hat{r}_1'\mathbf{W}_n\hat{\Omega}^{-1}\mathbf{W}_n\hat{r}_1.$$

Carroll and Wand (1991) and Wang and Ruppert (1996) obtained similar results for their semiparametric estimators.

By straightforward calculations one can show that the optimal weighting matrix is $\mathbf{W}_n = \Omega^{-1}$ or any consistent estimate thereof (see Newey and McFadden 1994), in which case the asymptotic variance in (10) becomes $V_{\mathrm{opt}} = (r_{10}'\Omega^{-1}r_{10})^{-1}$. If we take $\mathbf{W}_n = \hat{\Omega}^{-1}$, then the efficient variance is achieved.

Consider the infeasible procedure $\bar{\lambda}$ that replaces $\tilde{m}_\lambda$ by $\bar{m}_\lambda$ in (6). This has asymptotic variance the same as (10) with Ω replaced by $\Omega_0 = E[\varepsilon^2_{\lambda_0}(i)\mathbf{Z}_i\mathbf{Z}_i']$. Now work with the special case that $\varepsilon^2_{\lambda_0}(i)$ is independent of $\mathbf{Z}_i, E(\mathbf{Z}_i|X_i)$ is additive and the components of X are mutually independent. In this case $\mathbf{Z}_i^* = E(\mathbf{Z}_i|X_i)$, so that $\Omega \leq \Omega_0$; that is, the asymptotic variance of $\hat{\lambda}$ can be smaller than that for $\bar{\lambda}$. Although this appears anomalous, it has been found by

other authors (see Gutierrez and Carroll 1995; Robins, Rotnizky, and Zhao 1994). It is mainly due to the fact that the instrumental variable method may not reach the semiparametric efficiency bound as defined by Newey (1990). Even though $\hat{\lambda}$ is not efficient, it does not impose either normality of the errors or homoscedasticity in the conditional variance, and it is easy to implement. In this sense it is robust and practical. One can impose additional restrictions, such as constant variance, within this framework by adding an additional moment condition to (9). This results in more efficient estimates of λ when the restrictions are true.

4.3 Asymptotic Normality of Regression Function Estimates

We now consider the final stage in which the root-n consistent estimate $\hat{\lambda}$ is used to estimate the one-dimensional functions. We use a procedure that has bandwidth $h = \beta n^{-1/5}$ and second-order kernel K for the direction of interest and $(d-1)$-dimensional kernel L of order q with common bandwidth g for the remaining directions (as in Chen et al. 1996 for additive nonparametric regression). As pointed out there, if $d \leq 4$, then the theory is consistent with using second-order kernels. When $d > 4$, one must use higher-order kernels to satisfy the conditions of the theorem.

Let $\|K\|_2^2 = \int K^2(u)\,du$ and $\mu_2(K) = \int u^2 K(u)\,du$.

Theorem 3. Suppose that the condition C1 given in the Appendix holds. Then

$$n^{2/5}\{\hat{\gamma}_\alpha(x_\alpha;\hat{\lambda}) - \gamma_\alpha(x_\alpha;\lambda_0)\} \Rightarrow \mathrm{N}\{b_\alpha(x_\alpha), v_\alpha(x_\alpha)\},$$

where

$$b_\alpha(x_\alpha) = \beta^2\mu_2(K)\left\{\frac{1}{2}\,g''_\alpha(x_\alpha) + g'_\alpha(x_\alpha) \times \int \frac{\partial \ln p}{\partial x_\alpha}\,(x)p_{\underline{\alpha}}(x_{\underline{\alpha}})\,dx_{\underline{\alpha}}\right\}$$

and

$$v_\alpha(x_\alpha) = \beta^{-1}\|K\|_2^2 \int \sigma^2(x)\,\frac{p^2_{\underline{\alpha}}(x_{\underline{\alpha}})}{p(x)}\,dx_{\underline{\alpha}},$$

where $\sigma^2(x) = \mathrm{var}\{\theta_{\lambda_0}(Y)|X = x\}$. Finally, $v_\alpha(x_\alpha)$ can be consistently estimated by $\sum_{j=1}^n w_j^2\tilde{\varepsilon}^2_{\hat{\lambda}}(j)$, where $\{w_j\}_{j=1}^n$ are the weights in $\hat{\gamma}_\alpha(x_\alpha;\hat{\lambda}) = \sum_{j=1}^n w_j\theta_{\hat{\lambda}}(y_j)$. Similar results can be obtained for local linear regression smoothers (Fan 1992), with the bias function given by the simpler form $b_\alpha(x_\alpha) = \beta^2\mu_2(K)g''_\alpha(x_\alpha)/2$.

The estimated regression function has the one-dimensional convergence rate and is unaffected by the estimation of the transformation parameter. Bandwidth selection and order selection can be addressed exactly as for the regression problem. We do not advocate using higher-order kernels in practice, even when the dimensions are high, because of their well-known poor small-sample performance and because we think that with better proof technology, one can prove Theorem 3 without bias reduction. (See Fan, Härdle,

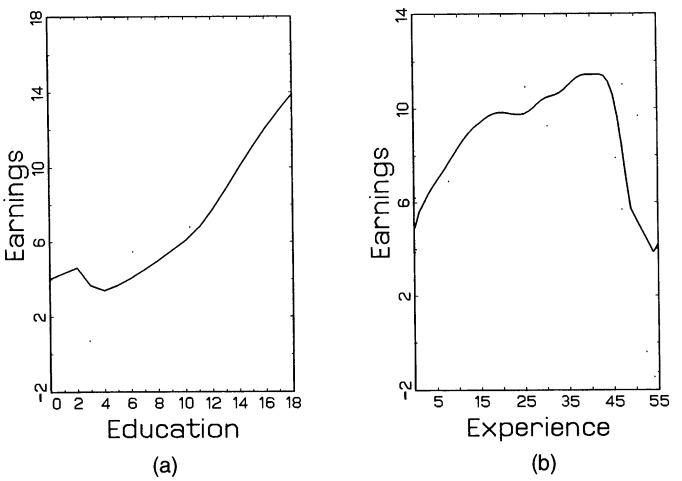

Figure 2. The Estimated Covariate Effects (a) Education and (b) Earnings on Earnings (λ = 1) Shown With Pointwise Symmetric 95% Confidence Intervals.

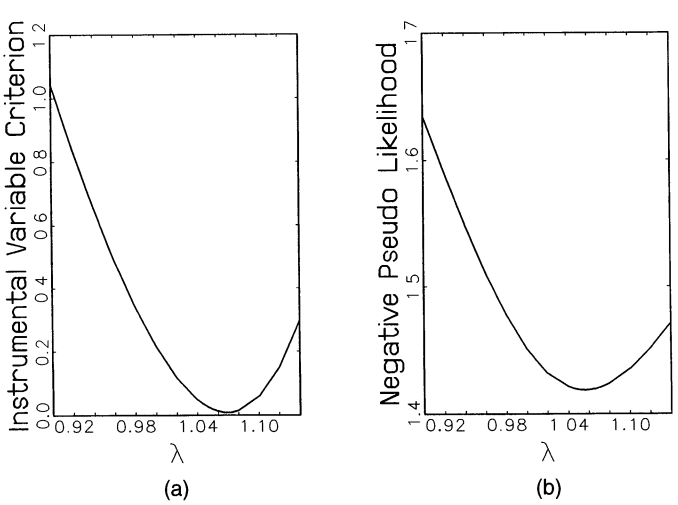

Figure 1. The Criterion Functions (a) Instrumental Variables ($\times 10^{-4}$) and (b) Negative Pseudolikelihood Function ($\times 10^{-3}$) Against λ for a Grid of Values Between [.90, 1.14] for the Earnings/Education Data.

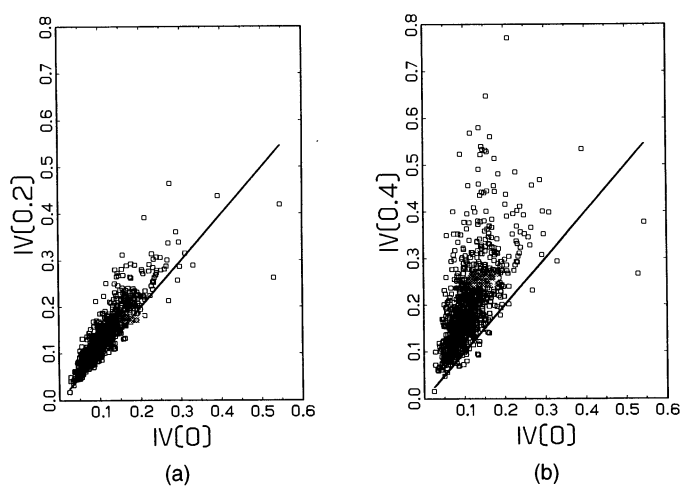

Figure 3. The Relative Magnitudes of the Criterion Functions for the Simulated Data at Different Parameter Values (a) λ = .2 Versus λ = 0 and (b) λ = .4 Versus λ = 0. The 45-degree line is shown for comparison. Sample size n = 100.

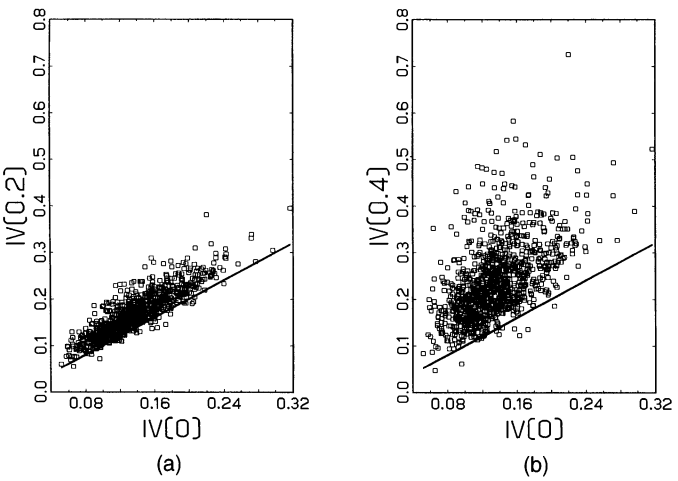

Figure 4. The Relative Magnitudes of the Criterion Functions for the Simulated Data at Different Parameter Values (a) $\lambda = .2$ Versus $\lambda = 0$ and (b) $\lambda = .4$ Versus $\lambda = 0$. The 45-degree line is shown for comparison. Sample size $n = 250$.

and Mammen 1996 and Hengartner 1996 for further discussion.)

5. NUMERICAL RESULTS

5.1 Application

We examine again the dataset used by Linton and Nielsen (1995) obtained from a random sample of 534 individuals from the 1985 Current Population Survey conducted by the U.S. Department of Commerce. (Details of this dataset can be found in Berndt 1991, chap. 5.) We examine the relationship between wages (y) and the covariates education in years (x_1) and experience in years (x_2). Linton and Nielsen (1995) used a logarithmic transformation of the dependent variable, as suggested by much other work. We use the Box–Cox transformation on the dependent variable that nests both logarithm and level. We implemented the additive regression procedure using a Gaussian kernel and rule of thumb bandwidth selection, following Linton and Neilson (1995). Figure 1 shows the instrumental variable (IV) criterion and the pseudolikelihood (LIKE) criterion computed at a grid of λ values. For instruments we took 1, x_1, x_1^2, x_2, and x_2^2, and we used $\mathbf{W}_n = I$ as weighting. The IV criterion was maximized at $\lambda = 1.069$, and LIKE was minimized at $\lambda = 1.056$. (The optima were found by grid search to a precision of $\pm.0005$.) The standard error was .0749 for the IV estimator. Thus the optimal transformation here is not far from linear; that is, the effects of education and experience on earnings itself appear to be additive.

Figure 2 plots the fitted additive regression for the untransformed y; that is, $\lambda = 1$. The effect of education on earnings is mildly convex, whereas that of experience is somewhat concave, with rapidly increasing returns to the first 10 years of experience followed by a slow but steady increase through 40 years, followed by a decline in later years. This is consistent with other studies, although some have found a similar dip in the returns to education, (see Mukarjee and Stern 1994).

5.2 Simulations

We first investigate by simulation how well the IV criterion function does at discriminating between rival models for small samples. This bears on how well $\hat{\lambda}$ behaves in small samples.

We generated 1,000 samples of sizes 100 and 250 from

$$\ln Y = X_1 + \left(X_2^2 - \frac{1}{12}\right) + \varepsilon,$$

where $\varepsilon \sim$ N(0, .1), while X_1 and X_2 were mutually independent uniforms on $[-.5, .5]$. We work again with the Box–Cox model for which the foregoing data-generation process is equivalent to $\lambda_0 = 0$. Figure 3a plots $Q_n(.2)$ against $Q_n(0)$, and Figure 3b shows $Q_n(.4)$ against $Q_n(0)$, both for the smaller sample size of 100. We used the same instruments as in the application. The median values were .101, .129, and .182. In 927 cases the criterion preferred $\lambda = 0$ to $\lambda = .2$, and in 973 cases it preferred $\lambda = 0$ to $\lambda = .4$. Figure 4 shows the same plots as in Figure 3, for the larger sample size of 250. In this case the criterion preferred $\lambda = 0$ to $\lambda = .2$ in 961 cases, and preferred $\lambda = 0$ to $\lambda = .4$ in 989 cases. Thus as sample size increases, the evidence in favor of the true value mounts.

Secondly, we check whether the integration method breaks down when applied to high-dimensional data. We generated data from the following model:

$$Y = \sum_{\alpha=1}^{d} X_\alpha^3 + \varepsilon,$$

where X_α are independent uniforms on $[-.5, .5]$ and $\varepsilon \sim$ N(0, .1). We examine our integration estimate $\tilde{m}_1(0)$ of $m_1(0)$, assuming that $\lambda = 1$ is known, for the cases $d = 1, 2, \ldots, 10$. Note that the estimate of $m_1(0)$ for the case $d = 1$ is the one-dimensional smooth of Y on X_1; its asymptotic variance cannot be bettered when $d \geq 2$, even by the backfitting method. We used a uniform kernel and a single bandwidth of size precisely $n^{-1/5}$. Our results are given in Table 1.

There is a deterioration in performance as dimension increases but less so at the larger sample size. Even in the case where $n = 100$, there is only about a 50% increase in the root mean squared error for $d = 10$.

6. FINAL REMARKS AND CONCLUSIONS

There was some controversy about how to conduct infer-

Table 1. Root Mean Squared Error of $\tilde{m}_1(0)$ Based on 500 Replications

	$d = 1$	$d = 2$	$d = 3$	$d = 4$	$d = 5$	$d = 6$	$d = 7$	$d = 8$	$d = 9$	$d = 10$
$n = 100$	.0121	.0122	.0130	.0138	.0150	.0158	.0167	.0172	.0181	.0188
$n = 200$	.0098	.0101	.0104	.0103	.0107	.0115	.0120	.0126	.0129	.0133

ence in fully parametric transformation models; in particular, whether one should take into account the uncertainty due to not knowing the transformation when evaluating estimates of the regression coefficients (see Bickel and Doksum 1981; Hinkley and Runger 1984). In our case such questions are moot, because the uncertainty in measuring the transformation is of smaller order than that in estimating the regression function. Thus the asymptotic distributions that we have given are appropriate from both points of view. However, controversy would reemerge if one were interested in the constant c or some other root-n consistently estimable functional of the covariate effects.

APPENDIX A: PROOFS OF THEOREMS

The proofs of our theorems are based on several lemmas established in Appendix B. We make use of the uniform weak laws of large numbers (ULLN) established by Andrews (1995) and Newey (1991). We use the following notations:

$$M_\lambda = \begin{pmatrix} m_\lambda(X_1) \\ \vdots \\ m_\lambda(X_n) \end{pmatrix}; \quad \bar{M}_\lambda = \begin{pmatrix} \bar{m}_\lambda(X_1) \\ \vdots \\ \bar{m}_\lambda(X_n) \end{pmatrix}; \quad \Xi_\lambda = \begin{pmatrix} \varepsilon_\lambda(1) \\ \vdots \\ \varepsilon_\lambda(n) \end{pmatrix}.$$

We also use the following regularity conditions:

Assumption A

1. The kernel function $K(\cdot)$ is bounded, nonnegative, compactly supported, and Lipschitz continuous, and $\int K(u)\,du = 1$.
2. The sequence of bandwidths satisfies $h \to 0$ and $nh^{d+1} \to \infty$.
3. The densities $p_\alpha, p_{\underline{\alpha}}$, and p are bounded away from 0 and infinity and are Lipschitz continuous on their compact supports $\mathcal{X}_\alpha, \mathcal{X}_{\underline{\alpha}}$, and $\mathcal{X}$. Furthermore, $m_\lambda(x)$ is Lipschitz continuous on $\mathcal{X} \times \mathcal{N}(\lambda_0)$.
4. The random variables Z_i have a positive definite second moment matrix.
5. The stochastic process $\{Z_i\varepsilon_\lambda(i)\}_{\lambda\in\mathcal{N}(\lambda_0)}$ has a uniformly bounded first absolute moment and is also Lipschitz continuous in the sense that

$$|Z_i\varepsilon_{\lambda_1}(i) - Z_i\varepsilon_{\lambda_2}(i)| \le |\lambda_1 - \lambda_2|A$$

for all $\lambda_1, \lambda_2 \in \mathcal{N}(\lambda_0)$, where $E(A) < \infty$. Furthermore, $E[\varepsilon_\lambda(i)|Z_i] = 0$ for $\lambda \in \mathcal{N}(\lambda_0)$.
6. The stochastic process $\{[\bar{m}_\lambda(X_i) - m_\lambda(X_i)]Z_i\}_{\lambda\in\mathcal{N}(\lambda_0)}$ has a uniformly bounded first absolute moment.

These conditions are standard. For asymptotic normality, we require these additional conditions:

Assumption B

1. The density p (and hence p_α and $p_{\underline{\alpha}}$) and the regression function m_{λ_0} are q times (Lipschitz) continuously differentiable on $\mathcal{X}$.
2. $\int u^j K(u)\,du = 0, j = 1, \ldots, q$, and $K(\cdot)$ has a Lipschitz-continuous derivative.
3. The sequence of bandwidths satisfies $nh^{d+1} \to \infty, nh^{2q} \to 0$.
4. The stochastic processes $(Z_i\{[\partial\varepsilon_\lambda(i)]/\partial\lambda\})_{\lambda\in\mathcal{N}(\lambda_0)}$ and $(Z_i\{[\partial^2\varepsilon_\lambda(i)]/\partial\lambda^2\})_{\lambda\in\mathcal{N}(\lambda_0)}$ have uniformly bounded first absolute moments and are Lipschitz continuous with respect to λ on $\mathcal{N}(\lambda_0)$.
5. The stochastic processes $\{Z_i(\partial^j/\partial\lambda^j)[m_\lambda(X_i) - \bar{m}_\lambda(X_i)]\}_{\lambda\in\mathcal{N}(\lambda_0)}, j = 0, 1, 2$, have uniformly bounded first absolute moments and are also Lipschitz continuous with respect to λ on $\mathcal{N}(\lambda_0)$. Let $r_j(\lambda) = E[Z_i(\partial^j/\partial\lambda^j)\{m_\lambda(X_i) - \bar{m}_\lambda(X_i)\}], j = 0, 1, 2$, and assume that $r_1(\lambda_0) \ne 0$.
6. $E\{Z_iZ_i'\varepsilon^2_{\lambda_0}(i)\} < \infty$.

Note that conditions A5 and B4 imply that $E(\{[\partial\varepsilon_\lambda(i)]/\partial\lambda\}|Z_i)$ and $E(\{[\partial^2\varepsilon_\lambda(i)]/\partial\lambda^2\}|Z_i)$ are both 0.

Finally, for Theorem 3 we use the following conditions:

Assumption C

1. The conditions of Chen et al. (1996) hold. Suppose also that the family of random variables $\partial\theta_\lambda(Y)/\partial\lambda$ is tight in λ. Finally, suppose that $n^{1/2}(\hat{\lambda} - \lambda) = O_p(1)$.

Proof of Theorem 1

Write

$$\boldsymbol{\Theta}_\lambda - \tilde{\mathbf{M}}_\lambda = (\boldsymbol{\Theta}_\lambda - M_\lambda) + (M_\lambda + \bar{M}_\lambda) + (\bar{M}_\lambda - \tilde{\mathbf{M}}_\lambda).$$

Condition A5 is sufficient to guarantee that the ULLN holds for the first term; that is,

$$\sup_{\lambda\in\mathcal{N}(\lambda_0)} |n^{-1}\mathbf{Z}'(\boldsymbol{\Theta}_\lambda - M_\lambda)| = o_p(1).$$

Furthermore,

$$\sup_{\lambda\in\mathcal{N}(\lambda_0)} |n^{-1}\mathbf{Z}'(\bar{M}_\lambda - \tilde{\mathbf{M}}_\lambda)| = o_p(1)$$

by Cauchy–Schwarz, a weak law of large numbers applied to $\mathbf{Z}'\mathbf{Z}/n$, and Lemma 1. Hence, using Cauchy–Schwarz once again, we have

$$\begin{aligned} Q_n(\lambda) &= n^{-2}(M_\lambda - \bar{M}_\lambda)'\mathbf{Z}\mathbf{W}_n\mathbf{Z}'(M_\lambda - \bar{M}_\lambda) + o_p(1) \\ &= Q(\lambda) + o_p(1), \end{aligned}$$

where $Q(\lambda) = r_0'(\lambda)\mathbf{W}r_0(\lambda)$. The last equality is due to ULLN applied to $n^{-1}\mathbf{Z}'(M_\lambda - \bar{M}_\lambda)$, which holds under conditions A3 and A6. Because $m_{\lambda_0} - \bar{m}_{\lambda_0} = 0$, we have $Q(\lambda_0) = 0$.

Using the condition A7 and the fact that $\mathbf{W}$ is positive definite, we have $Q(\lambda) > 0$ for $\lambda \ne \lambda_0$ and $\lambda \in \mathcal{N}(\lambda_0)$. Hence λ_0 uniquely minimizes $Q(\lambda)$.

Proof of Theorem 2

By a Taylor expansion,

$$\begin{aligned} 0 = n^{1/2}s_n(\hat{\lambda}) &= n^{1/2}s_n(\lambda_0) + H_n(\lambda_0)n^{1/2}(\hat{\lambda} - \lambda_0) \\ &\quad + \{H_n(\lambda^*) - H_n(\lambda_0)\}n^{1/2}(\hat{\lambda} - \lambda_0), \quad \text{(A.1)} \end{aligned}$$

where $s_n(\lambda) = \partial Q_n/\partial\lambda$ and $H_n(\lambda) = \partial^2 Q_n/\partial\lambda^2$, whereas λ^* is between $\hat{\lambda}$ and λ_0. We have

$$s_n(\lambda) = 2n^{-2}(\boldsymbol{\Theta}_\lambda - \tilde{\mathbf{M}}_\lambda)'\mathbf{Z}\mathbf{W}_n\mathbf{Z}'\left(\frac{\partial\boldsymbol{\Theta}_\lambda}{\partial\lambda} - \frac{\partial\tilde{\mathbf{M}}_\lambda}{\partial\lambda}\right)$$

and

$$\begin{aligned} n^2 H_n(\lambda) &= 2(\boldsymbol{\Theta}_\lambda - \tilde{\mathbf{M}}_\lambda)\mathbf{Z}\mathbf{W}_n\mathbf{Z}'\left(\frac{\partial^2\boldsymbol{\Theta}_\lambda}{\partial\lambda^2} - \frac{\partial^2\tilde{\mathbf{M}}_\lambda}{\partial\lambda^2}\right) \\ &\quad + 2\left(\frac{\partial\boldsymbol{\Theta}_\lambda}{\partial\lambda} - \frac{\partial\tilde{\mathbf{M}}_\lambda}{\partial\lambda}\right)'\mathbf{Z}\mathbf{W}_n\mathbf{Z}'\left(\frac{\partial\boldsymbol{\Theta}_\lambda}{\partial\lambda} - \frac{\partial\tilde{\mathbf{M}}_\lambda}{\partial\lambda}\right). \end{aligned}$$

We show that

(a) $n^{1/2}s_n(\lambda_0) \Rightarrow \mathrm{N}(0, 4r_{10}'\mathbf{W}\Omega\mathbf{W}r_{10})$,
(b) $H_n(\lambda_0) = 2r_{10}'\mathbf{W}r_{10} + o_p(1)$, and
(c) $\{H_n(\lambda^*) - H_n(\lambda_0)\}n^{1/2}(\hat{\lambda} - \lambda_0)$ is of smaller order than the other terms.

We first establish (a). We have

$$n^{1/2}s_n(\lambda_0) = 2\{n^{-1/2}(\boldsymbol{\Theta}_{\lambda_0} - \tilde{\mathbf{M}}_{\lambda_0})'\mathbf{Z}\}\mathbf{W}_n \times \left\{n^{-1}\mathbf{Z}'\left(\frac{\partial\boldsymbol{\Theta}_{\lambda_0}}{\partial\lambda} - \frac{\partial\tilde{\mathbf{M}}_{\lambda_0}}{\partial\lambda}\right)\right\} \equiv 2T_{1n}T_{2n}T_{3n},$$

where $T_{2n} = \mathbf{W}_n = \mathbf{W} + o_p(1)$. An application of Lemma 2 yields

$$T_{3n} = r_1(\lambda_0) + o_p(1).$$

Finally, by Lemma 4, $T_{1n} \Rightarrow \mathrm{N}(0, \Omega)$. Therefore,

$$n^{1/2}s_n(\lambda_0) \Rightarrow \mathrm{N}(0, 4r_{10}'\mathbf{W}\Omega\mathbf{W}r_{10}).$$

We now turn to (b) and (c). These are proved by establishing that

$$\sup_{\lambda\in\mathcal{N}(\lambda_0)} |H_n(\lambda) - H(\lambda)| = o_p(1), \tag{A.2}$$

where

$$H(\lambda) = 2r_1(\lambda)'\mathbf{W}r_1(\lambda) + 2r_0(\lambda)'\mathbf{W}r_2(\lambda),$$

which implies (c). Then note that

$$H(\lambda_0) = 2r_1(\lambda_0)'\mathbf{W}r_1(\lambda_0),$$

because $r_0(\lambda_0) = 0$ by condition A7. This gives (b). Using the results in Theorem 1 and Lemmas 1, 2, and 3, (A.2) follows by the Cauchy–Schwarz inequality.

Using (a), (b), and (c), (10) follows. The consistency of the standard errors follows from the uniform consistency results used in the foregoing argument.

Proof of Theorem 3

By a Taylor expansion,

$$\begin{aligned} &n^{2/5}\{\hat{\gamma}_\alpha(x_\alpha;\hat{\lambda}) - \gamma_\alpha(x_\alpha;\lambda_0)\} \\ &\quad = n^{2/5}\{\hat{\gamma}_\alpha(x_\alpha;\lambda_0) - \gamma_\alpha(x_\alpha;\lambda_0)\} \\ &\qquad + n^{-1/10}\left\{\frac{\partial\hat{\gamma}_\alpha}{\partial\lambda}(x_\alpha;\lambda^*)n^{1/2}(\hat{\lambda} - \lambda_0)\right\}, \end{aligned} \tag{A.3}$$

where λ^* is an intermediate point between $\hat{\lambda}$ and λ_0. Then

$$\begin{aligned} &\frac{\partial\hat{\gamma}_\alpha}{\partial\lambda}(x_\alpha;\lambda) \\ &\quad = n^{-1}\sum_{i=1}^n \frac{\sum_{j=1}^n K\left(\frac{x_\alpha - X_{\underline{\alpha}j}}{h}\right) L\left(\frac{X_{\underline{\alpha}i} - X_{\underline{\alpha}j}}{g}\right)\frac{\partial\theta_\lambda}{\partial\lambda}(Y_j)}{\sum_{l=1}^n K\left(\frac{x_\alpha - X_{\underline{\alpha}l}}{h}\right) L\left(\frac{X_{\underline{\alpha}i} - X_{\underline{\alpha}l}}{g}\right)} \\ &\quad = O_p(1), \end{aligned}$$

by C1. Therefore, the second term in (A.3) is $o_p(1)$. The result then follows from theorem 1 of Chen et al. (1996).

APPENDIX B: LEMMAS

Lemma 1. Suppose that conditions A1–A6 hold. Then

$$\sup_{x\in\mathcal{X}}\sup_{\lambda\in\mathcal{N}(\lambda_0)} |\tilde{m}_\lambda(x) - \bar{m}_\lambda(x)| = o_p(1).$$

Proof. We first show that

$$\sup_{x\in\mathcal{X}}\sup_{\lambda\in\mathcal{N}(\lambda_0)} |\hat{\gamma}_\alpha(x_\alpha;\lambda) - \gamma_\alpha(x_\alpha;\lambda)| = o_p(1).$$

We have

$$\begin{aligned} &|\hat{\gamma}_\alpha(x_\alpha;\lambda) - \gamma_\alpha(x_\alpha;\lambda)| \\ &\quad \le \left|n^{-1}\sum_{i=1}^n \{\hat{m}_\lambda(x_\alpha, X_{\underline{\alpha}i}) - m_\lambda(x_\alpha, X_{\underline{\alpha}i})\}\right| \\ &\qquad + \left|n^{-1}\sum_{i=1}^n \{m_\lambda(x_\alpha, X_{\underline{\alpha}i}) - \gamma_\alpha(x_\alpha;\lambda)\}\right| \end{aligned}$$

by the triangle inequality. We can bound the first term on the right side by $\sup_{x_{\underline{\alpha}}} |\hat{m}_\lambda(x) - m_\lambda(x)|$, where

$$\sup_{x\in\mathcal{X}}\sup_{\lambda\in\mathcal{N}(\lambda_0)} |\hat{m}_\lambda(x) - m_\lambda(x)| = o_p(1)$$

(by Andrews 1995, thm. 3). Finally,

$$\sup_{x\in\mathcal{X}}\sup_{\lambda\in\mathcal{N}(\lambda_0)} \left|n^{-1}\sum_{i=1}^n m_\lambda(x_\alpha, X_{\underline{\alpha}i}) - \gamma_\alpha(x_\alpha;\lambda)\right| = o_p(1)$$

(by Newey 1991, cor. 3.1). Therefore, $\hat{\gamma}_\alpha(x_\alpha;\lambda)$ is uniformly consistent.

Now, because $\bar{m}_\lambda(x) = \sum_{\alpha=1}^d \gamma_\alpha(x_\alpha;\lambda) - (d-1)c_\lambda$ and $\tilde{m}_\lambda(x) = \sum_{\alpha=1}^d \hat{\gamma}_\alpha(x_\alpha;\lambda) - (d-1)\hat{c}_\lambda$, the conclusion of Lemma 2 follows by the triangle inequality and the fact that $\sup_{\lambda\in\mathcal{N}(\lambda_0)} |\hat{c}_\lambda - c_\lambda| = o_p(1)$.

Lemma 2. Suppose that conditions A1–A6 and B4 and B5 hold. Then

$$\sup_{\lambda\in\mathcal{N}(\lambda_0)} \left|n^{-1}\mathbf{Z}'\left(\frac{\partial\boldsymbol{\Theta}_\lambda}{\partial\lambda} - \frac{\partial\tilde{\mathbf{M}}_\lambda}{\partial\lambda}\right) - r_1(\lambda)\right| = o_p(1).$$

Proof. We have

$$\begin{aligned} &n^{-1}\mathbf{Z}'\frac{\partial}{\partial\lambda}(\boldsymbol{\Theta}_\lambda - \tilde{\mathbf{M}}_\lambda) \\ &\quad = n^{-1}\mathbf{Z}'\frac{\partial}{\partial\lambda}\{(\boldsymbol{\Theta}_\lambda - M_\lambda) + (M_\lambda - \bar{M}_\lambda) + (\bar{M}_\lambda + \tilde{\mathbf{M}}_\lambda)\}. \end{aligned}$$

The first term is uniformly $o_p(1)$, by the ULLN guaranteed by condition B4. The second term converges in probability to its mean constant vector $r_1(\lambda)$ by the same reasoning with condition B5. The last term can be proven to be $o_p(1)$ by noting that

$$\sup_{x\in\mathcal{X}}\sup_{\lambda\in\mathcal{N}(\lambda_0)} \left|\frac{\partial\tilde{m}_\lambda}{\partial\lambda}(x) - \frac{\partial\bar{m}_\lambda}{\partial\lambda}(x)\right| = o_p(1),$$

by the same arguments as in Lemma 1.

Lemma 3. Suppose that conditions A1–A6 and B4 and B5 hold. Then

$$\sup_{\lambda\in\mathcal{N}(\lambda_0)} \left|n^{-1}\mathbf{Z}'\left(\frac{\partial^2\boldsymbol{\Theta}_\lambda}{\partial\lambda^2} - \frac{\partial^2\tilde{\mathbf{M}}_\lambda}{\partial\lambda^2}\right) - r_2(\lambda)\right| = o_p(1).$$

Proof. Similar to Lemma 2.

Lemma 4. Suppose that conditions A1, A3–A5, A7, B1–B3, and B6 hold. Then

$$n^{-1/2}(\boldsymbol{\Theta}_{\lambda_0} - \tilde{\mathbf{M}}_{\lambda_0})'\mathbf{Z} \Rightarrow \mathrm{N}(0, \Omega).$$

Proof. Write

$$\begin{aligned} T_{1n} &= n^{-1/2}\{(\boldsymbol{\Theta}_{\lambda_0} - \dot{M}_{\lambda_0}) + (M_{\lambda_0} - \bar{M}_{\lambda_0}) \\ &\quad + (\bar{M}_{\lambda_0} - \tilde{\mathbf{M}}_{\lambda_0})\}'\mathbf{Z} \\ &= n^{-1/2}\Xi'_{\lambda_0}\mathbf{Z} + n^{-1/2}(\bar{M}_{\lambda_0} - \tilde{\mathbf{M}}_{\lambda_0})'\mathbf{Z}. \end{aligned}$$

because $M_{\lambda_0} - \bar{M}_{\lambda_0} = 0$. We have

$$n^{-1/2}(\tilde{\mathbf{M}}_{\lambda_0} - \bar{M}_{\lambda_0})'\mathbf{Z} = \mathrm{I} + \mathrm{II} + \mathrm{III},$$

where

$$\mathrm{I} = n^{-3/2}\sum_{\alpha=1}^{d}\sum_{i=1}^{n}\sum_{j=1}^{n} Z_i\{\hat{m}_{\lambda_0}(X_{\alpha i}, X_{\underline{\alpha} j}) - m_{\lambda_0}(X_{\alpha i}, X_{\underline{\alpha} j})\},$$

$$\mathrm{II} = -(d-1)n^{-1}\sum_{i=1}^{n} Z_i \times n^{-1/2}\sum_{j=1}^{n}\varepsilon_{\lambda_0}(j),$$

and

$$\mathrm{III} = n^{-3/2}\sum_{\alpha=1}^{d}\sum_{i=1}^{n}\sum_{j=1}^{n} Z_i\left\{m_{\lambda_0}(X_{\alpha i}, X_{\underline{\alpha} j}) - \int m_{\lambda_0}(X_{\alpha i}, x_{\underline{\alpha}})p_{\underline{\alpha}}(x_{\underline{\alpha}})\,dx_{\underline{\alpha}}\right\}.$$

because Z_i is mean 0, $\mathrm{II} = o_p(1)$. Second,

$$\mathrm{III} = n^{-1}\sum_{i=1}^{n} Z_i \times (d-1)n^{-1/2}\sum_{\alpha=1}^{d}\sum_{j=1}^{n} g_\alpha(X_{\alpha j}) = o_p(1),$$

because $E\{g_\alpha(X_{\alpha j})\} = 0$. Finally, we show that

$$\mathrm{I} = n^{-1/2}\sum_{j=1}^{n}\sum_{\alpha=1}^{d} w_{\alpha j}\varepsilon_{\lambda_0}(j) + o_p(1), \tag{B.1}$$

where $\omega_{\alpha j} = E\,(Z_i|X_\alpha = X_{\alpha j})\{[p_\alpha(x_{\alpha j})p_{\underline{\alpha}}(X_{\underline{\alpha} j})]/[p(X_j)]\}$.

To prove (B.1), let $L_h(t_{\underline{\alpha}}) = \prod_{\beta\neq\alpha} K_h(t_\beta)$ for any vector $\mathbf{t} = (t_\alpha, t_{\underline{\alpha}})$, and write

$$\mathrm{I} = n^{-1/2}\sum_{i=1}^{n}\sum_{\alpha=1}^{d}\left\{\frac{1}{n}\sum_{j=1}^{n}\frac{\hat{a}(X_{\alpha i}, X_{\underline{\alpha} j})}{\hat{p}(X_{\alpha i}, X_{\underline{\alpha} j})}\right\} Z_i, \tag{B.2}$$

where

$$\hat{a}(x) = \hat{r}(x) - \hat{p}(x)m_{\lambda_0}(x),$$

$$\hat{r}(x) = \frac{1}{n}\sum_{l=1}^{n} L_h(x_{\underline{\alpha}} - X_{\underline{\alpha} l})K_h(x_\alpha - X_{\alpha l})\theta_{\lambda_0}(Y_l),$$

and

$$\hat{p}(x) = \frac{1}{n}\sum_{l=1}^{n} L_h(x_{\underline{\alpha}} - X_{\underline{\alpha} l})K_h(x_\alpha - X_{\alpha l}).$$

Using the arguments of Chen et al. (1996), $\mathrm{I} = (T_{4n} + T_{5n})\{1 + o_p(1)\}$, where

$$T_{4n} = n^{-1/2}\sum_{i=1}^{n}\sum_{\alpha=1}^{d}\frac{1}{n}\sum_{j=1}^{n}\frac{\hat{a}(X_{\alpha i}, X_{\underline{\alpha} j}) - E_*\{\hat{a}(X_{\alpha i}, X_{\underline{\alpha} j})\}}{p(X_{\alpha i}, X_{\underline{\alpha} j})} Z_i$$

and

$$T_{5n} = n^{-1/2}\sum_{i=1}^{n}\sum_{\alpha=1}^{d}\frac{1}{n}\sum_{j=1}^{n}\frac{E_*\{\hat{a}(X_{\alpha i}, X_{\underline{\alpha} j})\}}{p(X_{\alpha i}, X_{\underline{\alpha} j})} Z_i,$$

where E_* denotes expectation conditional on $X_1, \ldots, X_n$. Note that

$$\begin{aligned} &\hat{a}(X_{\alpha i}, X_{\underline{\alpha} j}) - E_*\{\hat{a}(X_{\alpha i}, X_{\underline{\alpha} j})\} \\ &\qquad = \frac{1}{n}\sum_{l=1}^{n} L_h(X_{\underline{\alpha} j} - X_{\underline{\alpha} l})K_h(X_{\alpha i} - X_{\alpha l})\varepsilon_{\lambda_0}(l), \end{aligned}$$

so

$$T_{4n} = n^{-1/2}\sum_{\alpha=1}^{d}\sum_{l=1}^{n}\varepsilon_{\lambda_0}(l)\left\{\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\omega_{\alpha l i j}\right\},$$

where $\omega_{\alpha l i j} = L_h(X_{\underline{\alpha} j} - X_{\underline{\alpha} l})K_h(X_{\alpha i} - X_{\alpha l})Z_i/p(X_{\alpha i}, X_{\underline{\alpha} j})$. By the same integration arguments of Chen et al. (1996, thm. 1), we can replace $1/n^2\sum_i\sum_j \omega_{\alpha l i j}$ by $\omega_{\alpha l}$. Thus

$$T_{4n} = n^{-1/2}\sum_{l=1}^{n}\sum_{\alpha=1}^{d}\omega_{\alpha l}\varepsilon_{\lambda_0}(l) + o_p(1),$$

as required. By direct calculation, the bias term T_{5n} is $O_p(n^{1/2}h^q)$ by conditions B1 and B2 and is $o_p(1)$ by B3.

In conclusion, we have shown that

$$n^{-1/2}(\tilde{\mathbf{M}}_{\lambda_0} - \bar{M}_{\lambda_0})'\mathbf{Z} = n^{-1/2}\sum_{j=1}^{n} Z_j^*\varepsilon_{\lambda_0}(j) + o_p(1).$$

By B4 and B6, we can apply the central limit theorem to

$$n^{-1/2}\sum_{j=1}^{n}\{Z_j - Z_j^*\}\varepsilon_{\lambda_0}(j),$$

as required.

[Received September 1995. Revised December 1996.]

REFERENCES

Amemiya, T., and Powell, J. L. (1981), "A Comparison of the Box–Cox Maximum Likelihood Estimator and the Non-Linear Two-Stage Least Squares Estimator," *Journal of Econometrics*, 17, 351–381.

Andrews, D. W. K. (1995), "Nonparametric Kernel Estimation for Semiparametric Models," *Econometric Theory*, 11, 560–596.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996), "Identification of Causal Effects Using Instrumental Variables," (with discussion) *Journal of the American Statistical Association*, 91, 444–472.

Auestad, B., and Tjøstheim, D. (1991), "Functional Identification in Nonlinear Time Series," in *Nonparametric Functional Estimation and Related Topics*, ed. G. Roussas, Dordrecht: Kluwer, pp. 493–507.

Berndt, E. (1991), *The Practice of Econometrics*, Reading, MA: Addison-Wesley.

Bickel, P. J., and Doksum, K. A. (1981), "An Analysis of Transformations Revisited," *Journal of the American Statistical Association*, 76, 296–311.

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

Box, G. E. P., and Cox, D. R. (1964), "An Analysis of Transformations," *Journal of the Royal Statistical Society*, Ser. B, 26, 211–252.

Breiman, L., and Friedman, J. H. (1985), "Estimating Optimal Transformations for Multiple Regression and Correlation," (with discussion) *Journal of the American Statistical Association*, 80, 580–619.

Burbidge, J. B., Magee, L., and Robb, A. L. (1988), "Alternative Transformations to Handle Extreme Values of the Dependent Variable," *Journal of the American Statistical Association*, 83, 123–127.

Buja, A., Hastie, T., and Tibshirani, R. (1989), "Linear Smoothers and Additive Models," (with discussion), *The Annals of Statistics*, 17, 453–555.

Carroll, R. J. (1980), "A Robust Method of Testing Transformations to Achieve Approximate Normality," *Journal of the Royal Statistical Society*, Ser. B, 42, 71–78.

——— (1982), "Prediction and Power Transformations When the Choice of Power is Restricted to a Finite Set," *Journal of the American Statistical Association*, 77, 908–915.

Carroll, R. J., and Ruppert, D. (1984), "Power Transformation When Fitting Theoretical Models to Data," *Journal of the American Statistical Association*, 79, 321–328.

——— (1988), *Transformation and Weighting in Regression*, New York: Chapman and Hall.

Carroll, R. J., and Wand, M. P. (1991), "Semiparametric Estimation in Logistic Measurement Error Models," *Journal of the Royal Statistical Society*, Ser. B, 53, 573–585.

Chen, R., Härdle, W., Linton, O. B., and Severance-Lossin, E. (1996), "Estimation in Additive Nonparametric Regression," *Proceedings of the COMPSTAT Conference Semmering*, eds. W. Härdle and M. Schimek, Heidelberg: Physika Verlag.

Ehrlich, I. (1977), "Capital Punishment and Deterrence: Some Further Thoughts and Additional Evidence," *Journal of Political Economy*, 85, 741–788.

Engle, R. F., Granger, C. W. J., Rice, J., and Weiss, A. (1986), "Semiparametric Estimates of the Relation Between Weather and Electricity Sales," *Journal of the American Statistical Association*, 81, 310–320.

Fan, J. (1992), "Design-Adaptive Nonparametric Regression," *Journal of the American Statistical Association*, 87, 998–1004.

Fan, J., Härdle, W., and Mammen, E. (in press), "Direct Estimation of Low-Dimensional Components in Additive Models," submitted to *The Annals of Statistics*.

Gutierrez, R., and Carroll, R. J. (1995), "Plug-In Semiparametric Estimating Equations," unpublished manuscript, Texas A&M University.

Härdle, W. (1990), *Applied Nonparametric Regression*, Cambridge, U.K.: Cambridge University Press.

Hastie, T. J., and Tibshirani, R. J. (1990), *Generalized Additive Models*, London: Chapman and Hall.

Heckman, J. J., and Polachek, S. (1974), "Empirical Evidence on the Functional Form of the Earnings–Schooling Relationship," *Journal of the American Statistical Association*, 69, 350–354.

Hengartner, N. W. (1996), "Rate-Optimal Estimation of Additive Regression via the Integration Method in the Presence of Many Covariates," preprint, Yale University, Dept. of Statistics.

Hinkley, D. V., and Runger, G. (1984), "The Analysis of Transformed Data" (with discussion), *Journal of the American Statistical Association*, 79, 302–320.

Hulten, C. R., and Wykoff, F. C. (1981), "The Estimation of Economic Depreciation Using Vintage Asset Prices," *Journal of Econometrics*, 15, 367–396.

Ibragimov, I. A., and Hasminskii, R. Z. (1980), "On Nonparametric Estimation of Regression," *Soviet Mathematics. Doklady*, 21, 810–4.

Johnson, N. L. (1949), "Systems of Frequency Curves Generated by Methods of Translation," *Biometrika*, 36, 149–176.

Linton, O. B., and Härdle, W. (1996), "Estimating Additive Regression With Known Links," *Biometrika*, 83, 529–540.

Linton, O. B., and Nielsen, J. P. (1995), "Estimating Structured Nonparametric Regression of the Kernel Method," *Biometrika*, 82, 93–101.

Mukarjee, H., and Stern, S. (1994), "Feasible Nonparametric Estimation of Multiargument Monotone Functions," *Journal of the American Statistical Association*, 89, 77–80.

Newey, W. K. (1990), "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics*, 5, 99–135.

——— (1991), "Uniform Convergence in Probability and Stochastic Equicontinuity," *Econometrica*, 59, 1161–1168.

——— (1994), "Kernel Estimation of Partial Means," *Econometric Theory*, 10, 233–253.

Newey, W. K., and McFadden, D. (1994), "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, Vol. 4, eds. R. F. Engle and D. L. McFadden, Amsterdam: Elsevier.

Robins, J., Rotnizky, A., and Zhao, L. (1994), "Estimation of Regression Coefficients When Some Regressors are not Always Observed," *Journal of the American Statistical Association*, 89, 846–857.

Robinson, P. M. (1988), "Root-n-Consistent Semiparametric Regression," *Econometrica*, 56, 931–954.

——— (1991), "Best Nonlinear Three-Stage Least Squares Estimation of Certain Econometric Models," *Econometrica*, 59, 755–786.

Stone, C. J. (1980), "Optimal Rates of Convergence for Nonparametric Estimators," *The Annals of Statistics*, 8, 1348–1360.

——— (1982), "Optimal Global Rates of Convergence for Nonparametric Regression," *The Annals of Statistics*, 8, 1040–1053.

——— (1986), "The Dimensionality Reduction Principle for Generalized Additive Models," *The Annals of Statistics*, 14, 592–606.

Tibshirani, R. (1988), "Estimating Optimal Transformations for Regression via Additivity and Variance Stabilization," *Journal of the American Statistical Association*, 83, 394–405.

Tjøstheim, D., and Auestad, B. (1994), "Nonparametric Identification of Nonlinear Time Series: Projections," *Journal of the American Statistical Association*, 89, 1398–1409.

Wang, N., and Ruppert, D. (1996), "Estimation of Regression Parameters in a Semiparametric Transformation Model," *Journal of Statistical Planning and Inference*, 52, 331–351.

Zarembka, P. (1968), "Functional Form in the Demand for Money," *Journal of the American Statistical Association*, 63, 502–511.

——— (1974), "Transformations of Variables in Econometrics," in *Frontiers in Econometrics*, ed. P. Zarembka, Boston: Academic Press.

Zellner, A., and Revankar, N. (1969), "Generalized Production Functions," *Review of Economic Studies*, 37, 241–250.