LSE Research Online

Article (refereed)

Oliver B. Linton

Efficient estimation of generalized additive nonparametric regression models

Originally published in Econometric theory, 16 (4). pp. 502-523 © 2000 Cambridge University Press.

You may cite this version as:
Linton, Oliver B. (2000). Efficient estimation of generalized additive nonparametric regression models [online]. London: LSE Research Online.
Available at: http://eprints.lse.ac.uk/archive/00000314
Available online: July 2005

LSE has developed LSE Research Online so that users may access research output of the School. Copyright © and Moral Rights for the papers on this site are retained by the individual authors and/or other copyright owners. Users may download and/or print one copy of any article(s) in LSE Research Online to facilitate their private study or for non-commercial research. You may not engage in further distribution of the material or use it for any profit-making activities or any commercial gain. You may freely distribute the URL (http://eprints.lse.ac.uk) of the LSE Research Online website.

http://eprints.lse.ac.uk
Contact LSE Research Online at: Library.Researchonline@lse.ac.uk

Econometric Theory, **16**, 2000, 502–523. Printed in the United States of America.

EFFICIENT ESTIMATION OF GENERALIZED ADDITIVE NONPARAMETRIC REGRESSION MODELS

Oliver B. Linton
London School of Economics
and
Yale University

We define new procedures for estimating generalized additive nonparametric regression models that are more efficient than the Linton and Härdle (1996, *Biometrika* 83, 529–540) integration-based method and achieve certain oracle bounds. We consider criterion functions based on the Linear exponential family, which includes many important special cases. We also consider the extension to multiple parameter models like the gamma distribution and to models for conditional heteroskedasticity.

1. INTRODUCTION

Additive models are widely used in both theoretical economics and in econometric data analysis. The standard text of Deaton and Muellbauer (1980) provides many microeconomics examples in which a separable structure is convenient for analysis and important for interpretability. There has been much recent theoretical and applied work in econometrics on semiparametric and nonparametric methods; see Härdle and Linton (1994) and Powell (1994) for bibliography and discussion. In such models additivity often has important implications for the rate at which the components can be estimated.

Let (X,Y) be a random variable with X of dimension d and Y a scalar. Consider the estimation of the regression function $m(x) = E(Y|X = x)$ based on a random sample $\{(X_i,Y_i)\}_{i=1}^n$ from this population. Stone (1980, 1982) and Ibragimov and Hasminskii (1980) show that the optimal rate for estimating m is $n^{-\ell/(2\ell+d)}$ with ℓ an index of smoothness of m. An additive structure for m is a regression function of the form

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

We thank two referees, Jean Opsomer, and Jens Perch Nielsen for helpful comments and the National Science Foundation and the North Atlantic Treaty Organization for financial support. Address correspondence to: Oliver B. Linton, Professor of Econometrics, London School of Economics, Houghton Street, London WC2A 2AE, UK.

© 2000 Cambridge University Press 0266-4666/00 $9.50

where $x = (x_1, \dots, x_d)'$ are the d-dimensional predictor variables and m_α are one-dimensional nonparametric functions operating on each element of the vector or predictor variables with $E\{m_\alpha(X_\alpha)\} = 0$. Stone (1986) shows that for such regression curves the optimal rate for estimating m is the one-dimensional rate of convergence with $n^{-\ell/(2\ell+1)}$ and does not increase with dimensions. In practice, the backfitting procedures proposed in Breiman and Friedman (1985) and Buja, Hastie, and Tibshirani (1989) are widely used to estimate the additive components. Buja et al. (1989, equation (18)) consider the problem of finding the projection of m onto the space of additive functions representing the right-hand side of (1). Replacing population by sample, this leads to a system of normal equations with $nd \times nd$ dimensions. To solve this in practice, the backfitting or Gauss–Seidel algorithm is usually used (see Hastie and Tibshirani, 1990, p. 91; Venables and Ripley, 1994, pp. 251–255). This technique is iterative and depends on the starting values and convergence criterion. These methods have been evaluated on numerous data sets and been refined quite considerably since their introduction.

Recently, Linton and Nielsen (1995), Tjøstheim and Auestad (1994), and Newey (1994) have independently proposed an alternative procedure for estimating m_α, which we call integration, that exploits the following idea. Suppose that $m(x_1, x_2)$ is any bivariate function and consider the quantities $\mu_1(x_1) = \int m(x_1, x_2)\, dP_2(x_2)$ and $\mu_2(x_2) = \int m(x_1, x_2)\, dP_1(x_1)$, where P_1 and P_2 are probability measures. If $m(x_1, x_2) = m_1(x_1) + m_2(x_2)$, then $\mu_1(\cdot)$ and $\mu_2(\cdot)$ are $m_1(\cdot)$ and $m_2(\cdot)$, respectively, up to a constant. In practice one replaces m by an estimate and integrates with respect to some known measure. The procedure is explicitly defined and its asymptotic distribution is easily derived: centered correctly, it converges to a normal distribution at the one-dimensional rate; the faster rate is because integration is averaging and hence reduces variance. The estimation procedure has been extended to a number of other contexts, such as the generalized additive model (Linton and Härdle, 1996), to dependent variable transformation models (Linton, Chen, Wang, and Härdle, 1997), to econometric time series models (Yang and Härdle, 1997), to panel data models (Porter, 1996), and to hazard models with time varying covariates and right censoring (Nielsen and Linton, 1997). Gozalo and Linton (1997) develop tests of additivity. In this wide variety of sampling schemes and models asymptotics for integration-based procedures have been derived because of the explicit form of the estimator. However, the integration method does not fully exploit the additive structure and is inefficient. Linton (1997) proposes a two-step procedure that took the integration estimate as a first step and then did one backfitting iteration from that. This procedure is argued to be oracle efficient, i.e., as efficient as the infeasible estimate that is based on knowing all components but the one of interest. The theoretical analysis of backfitting-like methods has only just begun and is thus far confined to regression. Opsomer and Ruppert (1997) provide conditional mean squared error expressions for bivariate independent and identically distributed (i.i.d.) data under strong conditions, whereas Linton, Mammen, and Nielsen (1997) establish a central limit theorem for a modified form of backfitting called empirical projection.

A generalized additive structure for m is of the form

$$G\{m(x)\} = c + \sum_{\alpha=1}^{d} m_\alpha(x_\alpha) \tag{2}$$

for some known, typically monotonic, link function G, where $x = (x_1, \ldots, x_d)^T$ are the d-dimensional predictor variables and m_α are one-dimensional nonparametric functions operating on each element of the vector of predictor variables. Here, $E\{m_\alpha(X_\alpha)\} = 0$ for identification. This class of models includes additive regression when G is the identity and multiplicative regression when G is the logarithm. For binary data it is appropriate to take G to be the inverse of a cumulative distribution function such as the normal or logit (this ensures that the regression function lies between 0 and 1 no matter what values $c + \sum_{\alpha=1}^{d} m_\alpha(x_\alpha)$ takes). Compare this specification with the semiparametric single index model considered in Ichimura (1993) in which the index on the right-hand side of (2) is linear but the link function $G(\cdot)$ is unrestricted (apart from the fact that it is the inverse of a cumulative distribution function [c.d.f.]). Both models considerably weaken the restrictions imposed by parametric binary choice models but are nonnested. One advantage of the additive model is that it allows for more general elasticity patterns: specifically, whereas in the single index model $\eta_{j:k} = (\partial \ln m/\partial x_j)/(\partial \ln m/\partial x_k)$ is restricted to be constant with respect to x, for the additive model $\eta_{j:k}$ can vary with x_j and x_k (although not with other x''s). Note that (2) is a partial model specification and we have not restricted in any way the variance or other aspects of the conditional distribution $\mathcal{L}(Y|X)$ of Y given X. A full model specification, widely used in this context, is to assume that $\mathcal{L}(Y|X)$ belongs to an exponential family with known link function G and mean m. This class of models is called *generalized additive* by Hastie and Tibshirani (1990). In some respects, econometricians would prefer the partial model specification in which we keep (2) but do not restrict ourselves to the exponential family. This flexibility is a relevant consideration for many data sets where there is overdispersion or heterogeneity.

Turning to estimation, Stone (1986) shows that for such models the optimal rate for estimating m (and m_α), based on a random sample $\{(Y_i, X_i)\}_{i=1}^{n}$ from this population, is the one-dimensional rate of convergence $n^{-\ell/2\ell+1)}$ to be compared with the best possible rate of $n^{-\ell/(2\ell+d)}$ when m is not so restricted. In practice, the backfitting procedures in conjunction with Fisher scoring are widely used to estimate generalized additive models (see Hastie and Tibshirani, 1990, p. 141). Linton and Härdle (1996) propose an alternative direct method for estimating the components by integrating a transformed pilot regression smoother. They provide sufficient conditions for a central limit theorem at the optimal one-dimensional rate. Nevertheless, this estimator is inefficient for the reasons given earlier.

In this paper, we suggest two-step procedures for estimating $m_\alpha(\cdot)$ in (2) that are more efficient than the integration method, thus extending the recent work of Linton (1997) in regression. We also provide more rigorous proofs of the claims made in that work. We base our procedures on a localized version of the likelihood function of linear exponential families (see Gourieroux, Monfort, and Trognon, 1984a, 1984b). This family includes what we are calling the partial model specification as a special case that corresponds to the homoskedastic normal likelihood function. Our estimators are nonlinear, and their asymptotics do not follow immediately from standard arguments for kernel estimators. Our proofs are based on a modification of some recent results of Gozalo and Linton (1995). For expositional purposes we shall work with the special case where we expect the "one-dimensional" rate of convergence $n^{2/5}$ for the additive estimates. The paper is organized as follows. In Section 2 we discuss infeasible oracle procedures for estimating one component that use knowledge of the other components. In particular, we introduce a criterion function based on linear exponential family density. We discuss feasible procedures and standard error construction. In Section 4 we discuss the extension to a model in which additive components enter into the local parameters of a general moment condition. We estimate the unknown functions using a local generalized method of moments (GMM) and local partial GMM criterion function. Our examples include the binomial and Poisson models and also models for conditional heteroskedasticity, known in time series as ARCH.

The symbol $\to_p$ denotes convergence in probability, whereas $\Rightarrow$ denotes convergence in distribution. For a random sequence X_n and deterministic decreasing sequences a_n, b_n we write $X_n \overset{AD}{=} N(a_n, b_n^2)$ whenever

$$\frac{X_n - a_n}{b_n} \Rightarrow N(0,1).$$

2. SINGLE PARAMETER LINEAR EXPONENTIAL FAMILY

2.1. Infeasible Procedures

We partition $X = (X_1, X_2)$ and $x = (x_1, x_2)$ where x_1 and X_1 are scalar, whereas x_2 and X_2 are in general of dimensions $d-1$. Let p_1 be the marginal density of X_1 and let p_2 and p be the densities of X_2 and X, respectively. Throughout, $m_2(\cdot)$ is an abbreviation for all the other components, i.e., $m_2(x_2) = \sum_{\alpha=2}^{d} m_\alpha(x_\alpha)$, and can be of any dimension. Let $\sigma^2(x) = \text{var}(Y|X = x)$.

Our purpose here is to define a standard by which to measure estimators of the components. The notion of efficiency in nonparametric models is not as clear and well understood as it is in parametric models. In particular, pointwise mean squared error comparisons do not provide a simple ranking between estimators such as kernel, splines, and nearest neighbors. Although minimax efficiencies can in principle serve this purpose, they are hard to work with and even harder to justify. Our approach is to measure our procedures against given infeasible (oracle) procedures for estimating $m_1(x_1)$ based on knowledge of c and $m_2(\cdot)$. Linton (1997) has already defined the oracle estimator when $G(\cdot)$ is the identity function, i.e., when we are in the additive regression setting (1). In this case, one smooths the partial errors $Y_i - c - m_2(X_{2i})$ on the direction of interest X_{1i}. He shows that indeed the oracle estimate has mean squared error smaller than the comparable integration-type estimator. In the general case though, one cannot find simple transformations of Y_i and $c + m_2(X_{2i})$ to which one can apply one-dimensional smoothing and that result in a more efficient procedure than the integration-type estimators. In summary, it was not immediately clear to us how to even define oracle efficiency in these nonlinear models. We suggest the following solution—impose our knowledge about $c + m_2(X_{2i})$ inside of a suitable criterion function.

We shall work with a criterion function motivated by the likelihood function of a complete specification of the conditional distribution of $Y|X$ along with the additivity restriction (2). In particular, we consider one-parameter linear exponential families, described in Gourieroux et al. (1984a), applied to the conditional distribution of Y given $X = x$. Every member of the family has a density with respect to some fixed measure μ, and this density function can be written as

$$\ell(y,m) = \exp\{A(m) + B(y) + C(m)y\}, \tag{3}$$

where $A(\cdot)$, $B(\cdot)$, and $C(\cdot)$ are known functions, with m being the mean of the distribution whose density is $\ell(y,m)$. The scalar $m \in \mathcal{M}$, a suitable parameter space. See Gourieroux et al. (1984a, 1984b) for parametric theory and applications in economics. The preceding likelihood function leads us to suggest the following class of criterion functions:

$$Q_n(\theta) = \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x_1 - X_{1i}}{h_n}\right)\{Y_i C_i(\theta) + A_i(\theta)\}, \tag{4}$$

where $C_i(\theta) = C(F(c + m_2(X_{2i}) + \theta_0 + \theta_1(X_{1i} - x_1)))$ and $A_i(\theta) = A(F(c + m_2(X_{2i}) + \theta_0 + \theta_1(X_{1i} - x_1)))$ with $F = G^{-1}$, whereas $\theta = (\theta_0, \theta_1)$. Here, h_n is a scalar bandwidth sequence and K is a kernel function. Let $\hat{\theta}$ maximize $Q_n(\theta)$ and let $\hat{m}_1(x_1) = \hat{\theta}_0(x_1)$ be our infeasible estimate of $m_1(x_1)$. We have the following result.

THEOREM 1. *Suppose that* (2) *holds. Then, under the regularity conditions* A *given in the Appendix, we have*

$$\hat{m}_1(x_1) - m_1(x_1) \overset{AD}{=} N\left[\frac{h_n^2}{2}\mu_2(K)m_1''(x_1), \frac{1}{nh_n}\|K\|_2^2 \frac{i_1(x_1)}{j_1^2(x_1)}\right],$$

where $\|K\|_2^2 = \int K^2(s)\,ds$ *and* $\mu_2(K) = \int K(s)s^2\,ds$, *whereas*

$$i_1(x_1) = \int C'(m(x))^2 F'(G(m(x)))^2 \sigma^2(x) p(x)\,dx_2,$$

$$j_1(x_1) = \int C'(m(x)) F'(G(m(x)))^2 p(x)\,dx_2.$$

We call $\hat{m}_1(x_1)$ an oracle estimator because its definition uses knowledge that only an oracle could have. A variety of smoothing paradigms could have been chosen here, and each will result in an "oracle" estimate. We have chosen the local linear with constant bandwidth kernel weighting because the local constant version, which does not include the slope parameter θ_1 and is slightly computationally easier, will result in "bad bias" behavior (for a discussion of the merits of local linear estimation see Fan, 1992). Higher order polynomials than linear can be used and will result in faster rates of convergence under appropriate smoothness conditions.

Remark 1. When (3) is true, we have $C'(m(x)) = 1/\sigma^2(x)$ by Property 3 of Gourieroux et al. (1984a). In this case, $j_1(x_1)$ is proportional to $i_1(x_1)$, and one obtains the simpler asymptotic variance proportional to

$$V_E = \frac{1}{\displaystyle\int \sigma^{-2}(x) F'(G(m(x)))^2 p(x)\, dx_2}.$$

The integration procedure of Linton and Härdle (1996) has asymptotic variance proportional to

$$V_H = \int G'\{m(x)\}^2 \sigma^2(x)\, \frac{p_2^2(x_2)}{p(x)}\, dx_2.$$

Because $G' = 1/F'$, we have, applying the Cauchy–Schwartz inequality, that $V_E \le V_{LH}$, and the oracle estimator has lower variance than the integration estimator. When (3) is not completely true, i.e., when the variance is misspecified, it is not possible to (uniformly) rank the two estimators unless the form of heteroskedasticity is restricted in some way (see the next section).

Remark 2. The bias of $\hat{m}_1(x_1)$ is what you would expect if $c + m_2(\cdot)$ were known to be exactly zero, and it is design adaptive. In the Linton and Härdle procedure there is an additional multiplicative factor to the bias,

$$\int \frac{p_2(x_2)}{F'(G(m(x)))}\, dx_2,$$

which can be either greater or less than one.

Remark 3. Note that $\hat{m}_1(x_1)$ is not guaranteed to satisfy $\int \hat{m}_1(x_1) p_1(x_1)\, dx_1 = 0$, but the recentered estimate

$$\hat{m}_{c1}(x_1) = \hat{m}_1(x_1) - \int \hat{m}_1(x_1) p_1(x_1)\, dx_1$$

does have this property. In fact, the variance of $\hat{m}_{c1}(x_1)$ and $\hat{m}_1(x_1)$ are the same to first order, whereas the bias of $\hat{m}_{c1}(x_1)$ has $m_1''(x_1)$ replaced by

$m_1''(x_1) - \int m_1''(x_1)p_1(x_1)\,dx_1$. According to integrated mean squared error, then, we are better off recentering because

$$\int \left\{ m_1''(x_1) - \int m_1''(x_1)p_1(x_1)\,dx_1 \right\}^2 p_1(x_1)\,dx_1 \leq \int \{m_1''(x_1)\}^2 p_1(x_1)\,dx_1.$$

2.2. Feasible Procedures

The previous section established the standard by which we choose to measure our estimators. We now show that one can achieve the oracle efficiency bounds given in Theorem 1 by substituting a suitable pilot estimator of $c + m_2(X_{2i})$ in the criterion function (4). Suppose that $\tilde{c} + \tilde{m}_2(X_{2i})$ is some initial consistent estimate and let

$$\tilde{Q}_n(\theta) = \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x_1 - X_{1i}}{h_n}\right)\{Y_i\,\tilde{C}_i(\theta) + \tilde{A}_i(\theta)\}, \tag{5}$$

where $\tilde{A}_i(\theta) = A(F(\tilde{c} + \tilde{m}_2(X_{2i}) + \theta_0 + \theta_1(X_{1i} - x_1)))$ and $\tilde{C}_i(\theta) = C(F(\tilde{c} + \tilde{m}_2(X_{2i}) + \theta_0 + \theta_1(X_{1i} - x_1)))$. Now let $\hat{\theta}^*(x_1) = (\hat{\theta}_0^*(x_1), \hat{\theta}_1^*(x_1))$ minimize $\tilde{Q}_n(\theta)$ and let $\hat{m}_1^*(x_1) = \hat{\theta}_0^*(x_1)$ be our feasible estimate of $m_1(x_1)$. Suitable initial estimates are provided by the Linton and Härdle (1996) procedure, which is explicitly defined. Finding $\hat{\theta}^*$ still involves solving a nonlinear optimization problem in general; an alternative approach here is to use the linearized two-step estimator

$$\begin{pmatrix} \hat{m}_1^{**}(x_1) \\ \hat{m}_1^{**\prime}(x_1) \end{pmatrix} \equiv \hat{\theta}^{**} = \tilde{\theta} - \left[\frac{\partial^2 \bar{Q}_n(\tilde{\theta})}{\partial\theta\partial\theta^T}\right]^{-1} \frac{\partial \tilde{Q}_n(\tilde{\theta})}{\partial\theta},$$

where $\tilde{\theta}$ is the full vector of preliminary estimates.

To provide asymptotic results we shall suppose that the initial estimator satisfies a linear expansion. Specifically, we suppose that

$$\tilde{c} - c + \tilde{m}_2(X_{2i}) - m_2(X_{2i}) = \sum_{\alpha=2}^{d} \frac{1}{ng_n}\sum_{j=1}^{n} \bar{K}\left(\frac{X_{\alpha i} - X_{\alpha j}}{g_n}\right) w_\alpha(X_i, X_j)\varepsilon_j + \delta_{ni}, \tag{6}$$

where $\varepsilon_j = Y_j - E(Y_j|X_j)$, where $\bar{K}$ is a kernel function, g_n is a bandwidth sequence, and w_α is some fixed function. The expansion (6) is assumed to obey the regularity conditions B given in the Appendix, which include the requirement that the remainder term $\delta_{ni} = o_p(n^{-2/5})$ uniformly in i. A number of procedures have recently been proposed for estimating components in additive models under a variety of sampling schemes (see, e.g., among others Linton and Nielsen, 1995; Linton and Härdle, 1996; Yang and Härdle, 1997; Kim, Linton, and Hengartner, 1997). The expansion (6) can be achieved by all of these methods by undersmoothing under various conditions.[1] One might need to assume stronger smoothness conditions than made in Assumption A to achieve this, although recent work by Hengartner (1996) suggests this may not be necessary.

We now have the following result.

THEOREM 2. *Suppose that Assumptions* A *and* B *given in the Appendix hold*. *Then*, *under* (2), *we have*

$$n^{2/5}\{\hat{m}_1^*(x_1) - \hat{m}_1(x_1)\}, n^{2/5}\{\hat{m}_1^{**}(x_1) - \hat{m}_1(x_1)\} \to_p 0.$$

This says that efficient estimates can be constructed by the two-step procedure and by the linearized two-step estimator; estimation of the nuisance parameter $c + m_2(\cdot)$ has no effect on the limiting distribution. This is not generally the case in parametric estimation problems, unless there is some orthogonality between the estimating equations. In our case, there is an intrinsic local orthogonality that affects smoothing operations.

Standard error and bandwidth choice issues can now be addressed via the mean squared error expressions given in Theorem 1, using modifications of standard methods. Thus, under the conditions of Theorem 2 and provided $nh_n^5 \to 0$, the following interval,

$$\hat{m}_1^*(x_1) \pm z_{\alpha/2}\sqrt{\frac{1}{nh_n}\,\|K\|_2^2\,\frac{\hat{i}_1(x_1)}{\hat{j}_1^2(x_1)}},$$

provides $1 - \alpha$ coverage of the true function $m_1(x_1)$, where z_α is the α critical value from the standard normal distribution, whereas

$$\hat{i}_1(x_1) = \frac{1}{n}\sum_{i=1}^{n} C'(\tilde{m}(x_1, X_{2i}))^2 F'(G(\tilde{m}(x_1, X_{2i})))^2 \tilde{\sigma}^2(x_1, X_{2i}),$$

$$\hat{j}_1(x_1) = \frac{1}{n}\sum_{i=1}^{n} C'(\tilde{m}(x_1, X_{2i})) F'(G(\tilde{m}(x_1, X_{2i})))^2,$$

in which $\tilde{m}(\cdot)$ and $\tilde{\sigma}^2(\cdot)$ are any uniformly consistent estimates of $m(\cdot)$ and $\sigma^2(\cdot)$ (see Härdle and Linton, 1994).

3. MULTIPARAMETER EXTENSIONS

The models we have examined thus far were one-parameter families as has been the case in most of the literature on additive models; we now consider extensions to multiple parameter families. The quadratic exponential family of Gourieroux et al. (1984a) can be analyzed similarly to the process described previously. This would amount to having an additional set of equations that impose additivity on some transformation of the variance. We shall consider a slightly more general situation based on the generalized method of moments, which allows the additivity to be imposed on any set of moments. We suppose that there exists a known function $\varphi: \mathbb{R}^{m+d+p} \to \mathbb{R}^q$ such that there exists a vector of additive functions $g^0(x) = (g_1^0(x), \ldots, g_p^0(x))$ with

$$g_l^0(x) = c_l + \sum_{\alpha=1}^{d} g_{l\alpha}(x_\alpha), \qquad l = 1, \ldots, p,$$

where $g_{l\alpha}(X_\alpha)$ are mean zero for identification, for which

$$E[\varphi(U, g^0(X))|X = x] = 0, \tag{7}$$

where $U = (Y, X)$. We assume that $q > 1$ and that there is a unique solution to (7). This sort of information could arise from an economic model or through partial specification of moments, as happens in the ARCH models (see the discussion that follows). It also includes a full likelihood specification as a special case. For example, suppose that $\ell(U, g^0(X))$ is the logarithm of the density function of $Y|X$ in which $g^0(X)$ is a vector of parameters. Then, $g^0(x)$ is the unique quantity that satisfies

$$\frac{\partial}{\partial g} E[\ell(U, g^0(X))|X = x] = 0.$$

This system of equations is of the form (7).

This leads naturally to the following estimation scheme. First, estimate $g^0(x)$ by any unrestricted smoothing method—we propose a sort of local GMM. Second, integrate out the directions not of interest to get a preliminary estimate of the univariate effects. Finally, reestimate the local GMM criterion function replacing the parameters of the components not of interest by preliminary estimates.

Let $\tilde{\theta}(x) = (\tilde{\theta}_1(x), \ldots, \tilde{\theta}_p(x))$ minimize the following criterion:

$$\left\| \frac{1}{nh_n^d} \sum_{i=1}^{n} \mathcal{K}\left(\frac{x - X_i}{h_n} \right) \varphi(U_i, \theta) \right\|_{A_n}^2 \tag{8}$$

with respect to $\theta = (\theta_1, \ldots, \theta_p)$, where $U_i = (Y_i, X_i)$, $\mathcal{K}$ is a multivariate kernel, whereas $\|x\|_{A_n} = (x^T A_n x)^{1/2}$ for some sequence of positive definite matrices $A_n \to_p A$, and let $\tilde{g}(x) = \tilde{\theta}(x)$. We are using a local constant approach here for simplicity. The asymptotic properties of this procedure can be derived using an extension of Gozalo and Linton (1995); we expect that $\tilde{g}(x)$ is asymptotically normal with pointwise mean squared error rate of $n^{-4/(4+d)}$ and indeed has an expansion like (6). To obtain estimates of the component functions, we simply integrate this pilot procedure as follows, letting, for example,

$$\tilde{g}_{l1}(x_1) = \int \tilde{g}_l(x) p_2(x_2)\, dx_2, \qquad l = 1, \ldots, p \tag{9}$$

and the other components similarly.[2] To estimate c_l we can use $\tilde{c}_l = \int \tilde{g}_{l1}(x_1) p_1(x)\, dx_1$. Thus, $\tilde{g}_{lj}(\cdot)$ are feasible preliminary estimates of $g_{lj}(\cdot)$. To achieve efficiency, we must modify this procedure to impose additivity. We first describe the oracle estimate. Let $\hat\theta = (\hat\theta_0, \hat\theta_1) = (\hat\theta_{01}, \dots, \hat\theta_{0p}, \hat\theta_{11}, \dots, \hat\theta_{1p})$ minimize the partial GMM criterion

$$G_n(\theta) = \left\| \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x_1 - X_{1i}}{h_n} \right) \varphi[U_i, c + \theta_0 + \theta_1 \cdot (X_{1i} - x_1) + g_{\cdot 2}(X_{2i})] \right\|_{A_n}^2$$

with respect to $\theta_0 = (\theta_{01}, \dots, \theta_{0p})$ and $\theta_1 = (\theta_{11}, \dots, \theta_{1p})$, where the vectors $g_{\cdot 2}(\cdot) = (g_{12}(\cdot), \dots, g_{p2}(\cdot))$ and $c = (c_1, \dots, c_p)$ are assumed known, and let $\hat g_{\cdot 1}(x_1) = (\hat g_{11}(x_1), \dots, \hat g_{p1}(x_1)) = \hat\theta_0(x_1)$. Finally, the feasible version of this replaces $g_{\cdot 2}(\cdot)$ and c by a vector of preliminary estimates provided by the integration principle, i.e., we let $\hat\theta^* = (\hat\theta_0^*, \hat\theta_1^*) = (\hat\theta_{01}^*, \dots, \hat\theta_{0p}^*, \hat\theta_{11}^*, \dots, \hat\theta_{1p}^*)$ minimize

$$\tilde G_n(\theta) = \left\| \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x_1 - X_{1i}}{h_n} \right) \varphi[U_i, \tilde c + \theta_0 + \theta_1 \cdot (X_{1i} - x_1) + \tilde g_{\cdot 2}(X_{2i})] \right\|_{A_n}^2$$

with respect to $\theta = (\theta_0, \theta_1)$, where $\tilde c$ and $\tilde g_{\cdot 2}(X_{2i})$ are obtained from (8) and (9), and let $\hat g_{\cdot 1}^*(x_1) = (\hat g_{11}^*(x_1), \dots, \hat g_{p1}^*(x_1)) = \hat\theta_0^*(x_1)$.

3.1. Asymptotics

Define the following $q \times p$ and $q \times q$ matrices:

$$\Psi(x,t) = E\left[\frac{\partial \varphi(U,t)}{\partial t} \middle| X = x \right]; \qquad R(x,t) = E[\varphi(U,t)\varphi^T(U,t) | X = x],$$

and let $\Psi_1 = \Psi_1(x_1) = \int \Psi(x, g^0(x)) p(x)\, dx_2$ and $R_1 = R_1(x_1) = \int R(x, g^0(x)) p(x)\, dx_2$. Furthermore, suppose that each of the preliminary estimators described in (8) and (9) satisfies a linear expansion such as (6). We have the following result.

THEOREM 3. *Under the regularity conditions* A′ *and* B′ *given in the Appendix, we have under the specification* (3) *that* $n^{2/5}[\hat g_{\cdot 1}^*(x_1) - \hat g_{\cdot 1}(x_1)] = o_p(1)$ *and that*

$$\hat g_{\cdot 1}(x_1) - g_{\cdot 1}(x_1) \overset{AD}{=} N\left[\frac{h_n^2}{2} \mu_2(K) g_1''(x_1), \frac{1}{nh_n} \|K\|_2^2 (\Psi_1^T A \Psi_1)^{-1} (\Psi_1^T A R_1 A \Psi_1)(\Psi_1^T A \Psi_1)^{-1} \right]. \tag{10}$$

Furthermore, if we take $A_n = \hat R_1^{-1}(x_1)$, *where* $\hat R_1(x_1) \to_p R_1(x_1)$, *then* $n^{2/5}[\hat g_{\cdot 1}^*(x_1) - \hat g_{\cdot 1}(x_1)] = o_p(1)$ *and*

$$\hat g_{\cdot 1}(x_1) - g_{\cdot 1}(x_1) \overset{AD}{=} N\left[\frac{h_n^2}{2} \mu_2(K) g_1''(x_1), \frac{1}{nh_n} \|K\|_2^2 (\Psi_1^T R_1^{-1} \Psi_1)^{-1} \right].$$

The choice of $A_n = \hat{R}_1^{-1}(x_1)$ as weighting gives minimum variance among the class of all such procedures. Note that the efficiency standard we erect here is not as high as in the one-parameter models. This is because, generically, we can expect correlation between $\hat{g}_{j1}(x_1)$ and $\hat{g}_{k1}(x_1)$ for $j,k = 1,\dots,p$. In other words, it is not possible to estimate $g_{11}(x_1)$, say, as well as if one knew every other component function in the model, although it is possible to estimate the vector $g_{.1}(\cdot)$ as well as if $g_{.2}(\cdot)$ were known.

As before, the preceding result can be used for bandwidth choice and standard error construction by replacing unknown quantities in (10) by estimates. Thus, under the conditions of Theorem 3 and provided $nh_n^5 \to 0$, for any vector $a = (a_1,\dots,a_p)^T$, the following interval,

$$a^T\hat{g}^*_{.1}(x_1) \pm z_{\alpha/2}\sqrt{\frac{1}{nh_n}\|K\|_2^2 a^T(\hat{\Psi}_1^T A_n \hat{\Psi}_1)^{-1}(\hat{\Psi}_1^T A_n \hat{R}_1 A_n \hat{\Psi}_1)(\hat{\Psi}_1^T A_n \hat{\Psi}_1)^{-1}a},$$

provides $1-\alpha$ coverage of the true function $a^T g_{.1}(x_1)$, where

$$\hat{\Psi}_1 = \frac{1}{nh_n}\sum_{i=1}^n K\left(\frac{x_1 - X_{1i}}{h_n}\right)\frac{\partial\varphi}{\partial t}[U_i,\tilde{c} + \tilde{g}_{.1}(x_1) + \tilde{g}_{.2}(X_{2i})],$$

$$\hat{R}_1 = \frac{1}{nh_n}\sum_{i=1}^n K\left(\frac{x_1 - X_{1i}}{h_n}\right)(\varphi\cdot\varphi^T)[U_i,\tilde{c} + \tilde{g}_{.1}(x_1) + \tilde{g}_{.2}(X_{2i})].$$

3.2. Examples

Example 1 (gamma and beta)

Suppose that there exist functions $\alpha(x)$ and $\beta(x)$, both themselves additively separable functions of x, that satisfy the equations

$$E(Y|X=x) = \alpha(x)\beta(x); \qquad E(Y^2|X=x) = \beta^2(x)\alpha(x)[1+\alpha(x)].$$

This partial model specification is implied by $Y|X=x$ being gamma distributed but is somewhat weaker. In this case, (7) is satisfied with $\varphi_1(Y,X|\alpha,\beta) = Y - \alpha\beta$ and $\varphi_2(Y,X|\alpha,\beta) = Y^2 - \beta^2\alpha(1+\alpha)$. A full model specification can be based on the gamma (log) density function of (Y,X), from which we obtain

$$E[\ell(U,\alpha,\beta|X=x)]$$
$$= (\alpha(x)-1)m_\ell(x) - \beta(x)^{-1}m(x) - \ln\Gamma(\alpha(x)) - \alpha(x)\ln\beta(x), \qquad \mathbf{(11)}$$

where $\Gamma(\cdot)$ is the gamma function, whereas $m(x) = E[Y|X=x]$ and $m_\ell(x) = E[\ln Y|X=x]$. This generates the following moment conditions:

$$\varphi_1(U|\alpha,\beta) = \ln Y - \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} - \ln\beta; \qquad \varphi_2(U|\alpha,\beta) = \frac{Y-\alpha\beta}{\beta^2}.$$

The asymptotic variance of these procedures can be found by direct calculation.[3] The beta distribution, which is frequently used in the study of rate or proportion data, can also be put in this framework. See Heckman and Willis (1977) for an econometric application of the beta distribution.

Example 2 (variance models [ARCH])

Suppose that with probability one

$$E(Y|X=x) = m(x) = F_m[\alpha(x)], \qquad \alpha(x) = c_m + m_1(x_1) + m_2(x_2), \tag{12}$$

$$\operatorname{var}(Y|X=x) = \sigma^2(x) = F_\sigma[\beta(x)], \qquad \beta(x) = c_\sigma + \sigma_1(x_1) + \sigma_2(x_2) \tag{13}$$

for some known functions F_m and F_σ. Estimates of $m_j(\cdot)$ and $\sigma_j(\cdot)$ can be obtained by integrating (transformed) nonparametric estimates of the mean and variance, as in Yang and Härdle (1997). Note that their procedure ignores the cross-equation information, which can be imposed in our framework. Using only the mean and variance specification gives the following moment functions: $\varphi_1(Y,X|\alpha,\beta) = Y - F_m(\alpha)$ and $\varphi_2(Y,X|\alpha,\beta) = Y^2 - F_m^2(\alpha) - F_\sigma(\beta)$; the asymptotic variance of the GMM procedure is as in (10) with

$$R(x, g^0(x)) = \begin{bmatrix} \sigma^2(x) & \kappa_3(x) \\ \kappa_3(x) & \kappa_4(x) + 2 \end{bmatrix};$$

$$\Psi(x, g^0(x)) = \begin{bmatrix} F_m'(\alpha(x)) & 0 \\ 2F_m(\alpha(x))F_m'(\alpha(x)) & F_\sigma'(\alpha(x)) \end{bmatrix},$$

where $\kappa_3(x) = E[\{Y - E(Y|X=x)\}^3|X=x]$. The optimal estimator has lower asymptotic variance than the procedure of Yang and Härdle (1997, Theorem 2.4) because it uses cross-equation information.[4]

A convenient complete model specification here is that $Y|X=x$ is $N(m(x), \sigma^2(x))$, which leads to the following moments:

$$\varphi_1(Y,X|\alpha,\beta) = \frac{Y - F_m(\alpha(x))}{F_\sigma(\beta(x))} F_m'(\alpha(x));$$

$$\varphi_2(Y,X|\alpha,\beta) = \frac{1}{2}\left\{\left(\frac{Y - F_m(\alpha(x))}{F_\sigma(\beta(x))}\right)^2 - 1\right\} \frac{F_\sigma'}{F_\sigma}(\beta(x)).$$

The corresponding procedure has asymptotic variance as in (10) with

$$R(x, g^0(x)) = \Psi(x, g^0(x)) = \begin{bmatrix} \dfrac{F_m'^2(\alpha(x))}{F_\sigma(\beta(x))} & 0 \\ 0 & \dfrac{1}{2}\left[\dfrac{F_\sigma'}{F_\sigma}(\beta(x))\right]^2 \end{bmatrix}.$$

4. CONCLUDING REMARKS

We have provided a general principle for obtaining efficient estimates that works in almost any model with separable nonparametric components, whether fully specified or only partially specified. We did not consider models with parametric components or discrete explanatory variables, because such models can be viewed as special cases of ours. The only new issue that arises in such models is how to impose the restriction of parametric effects efficiently.

If the additive structure (2) does not hold, then $\hat{m}_1(x_1)$ is estimating some other functional of the joint distribution (depending of course on what $c + m_2(\cdot)$ is) (see, e.g., Newey, 1994). Specifically, $\hat{m}_1(x_1)$ consistently estimates the minimizer of a Kullback–Liebler distance with respect to θ. Centered correctly, the asymptotic distributions take a similar form, with some relabeling, and are efficient for estimating these particular functionals.

NOTES

[1] Note that the expansion (6) contains no bias terms, which can be achieved by undersmoothing or additional bias reduction.

[2] A computationally efficient estimate of $g_{l1}(x_1)$ can be constructed by generalizing Kim et al. (1997) as follows. Let

$$\tilde{g}_{l1}(x_1) = \frac{1}{n}\sum_{i=1}^{n} K_h(x_1 - X_{1i})\tilde{g}_l(X_i)\,\frac{\tilde{p}_2(X_{2i})}{\tilde{p}(X_i)},$$

where $\tilde{p}_2$ and $\tilde{p}$ are kernel estimates of p_2 and p, respectively.

[3] With regard to preliminary estimation in the full model specification, there are two estimation strategies. First, simply substitute estimates of $m(x)$ and $m_\ell(x)$ in (11) and maximize to obtain $\tilde{\alpha}(x)$ and $\tilde{\beta}(x)$. Second, one can estimate the local parameters $\alpha(x)$ and $\beta(x)$ by local likelihood; i.e., let $\tilde{\alpha}(x)$ and $\tilde{\beta}(x)$ maximize

$$\frac{1}{nh_n^d}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right)[(\alpha - 1)\ln Y_i - \beta^{-1}Y_i - \ln\Gamma(\alpha) - \alpha\ln\beta]$$

with respect to α, β. In both cases, we then integrate $\tilde{\alpha}(x)$ and $\tilde{\beta}(x)$ with respect to $p_2(x_2)\,dx_2$.

[4] Strictly speaking our results only apply to the i.i.d. case, but recent work of Kim (1998) has extended this to a time series setting.

REFERENCES

Brieman, L. & J.H. Friedman (1985) Estimating optimal transformations for multiple regression and correlation, (with discussion). *Journal of the American Statistical Association* 80, 580–619.

Buja, A., T. Hastie, & R. Tibshirani (1989) Linear smoothers and additive models (with discussion). *Annals of Statistics* 17, 453–555.

Deaton, A. & J. Muellbauer (1980) *Economics and Consumer Behavior.* Cambridge: Cambridge University Press.

Fan, J. (1992) Design-adaptive nonparametric regression. *Journal of the American Statistical Association* 87, 998–1004.

Gourieroux, C., A. Monfort, & A. Trognon (1984a) Pseudo maximum likelihood methods: Theory. *Econometrica* 52, 681–700.

Gourieroux, C., A. Monfort, & A. Trognon (1984b) Pseudo maximum likelihood methods: Theory. *Econometrica* 52, 701–720.

Gozalo, P. & O.B. Linton (1995) Using Parametric Information in Nonparametric Regression. Working paper 95-40, Brown University.

Gozalo, P. & O.B. Linton (1997) A nonparametric test of additivity in generalized nonparametric regression. Available at http://www.econ.yale.edu/~linton.

Härdle, W. & O.B. Linton (1994) Applied nonparametric methods. In D.F. McFadden & R.F. Engle III (eds.), *The Handbook of Econometrics*, vol. IV. Amsterdam: North Holland.

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

Heckman, J.J. & R.J. Willis (1977) A beta-logistic model for the analysis of sequential laborforce participation by married women. *Journal of Political Economy* 85, 27–58.

Hengartner, N.W. (1996) Rate Optimal Estimation of Additive Regression via the Integration Method in the Presence of Many Covariates. Preprint, Department of Statistics, Yale University. Http://www.stats.yale.edu/Preprints.

Ibragimov, I.A. & Hasminskii, R.Z. (1980) On nonparametric estimation of regression. *Soviet Math. Dokl.* 21, 810–814.

Ichimura, H. (1993) Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics* 58, 71–120.

Kim, W. (1998) Quick and Efficient Estimation of Additive Time Series Models of Volatility. Manuscript, Yale University.

Kim, W., O.B. Linton, & N. Hengartner (1997) A computationally efficient oracle estimator for additive nonparametric regression with bootstrap confidence intervals. Available at http://econ.lse.ac.uk/~olinton.

Linton, O.B. (1997) Efficient estimation of additive nonparametric regression models. *Biometrika* 84, 469–474.

Linton, O.B., R. Cheng, N. Wang, & W. Härdle (1997) An analysis of transformations for additive nonparametric regression. *Journal of the American Statistical Association* 92, 1512–1521.

Linton, O.B. & W. Härdle (1996) Estimating additive regression models with known links. *Biometrika* 83, 529–540.

Linton, O.B., E. Mammen, & J.P. Nielsen (1997) The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. Available at http://econ.lse.ac.uk/~olinton.

Linton, O.B. & J.P. Nielsen (1995) A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika* 82, 93–100.

Newey, W.K. (1994) Kernel estimation of partial means. *Econometric Theory* 10, 233–253.

Nielsen, J.P. & O.B. Linton (1997) Multiplicative and Additive Marker Dependent Hazard Estimation Based on Marginal Integration. Unpublished manuscript.

Opsomer, J.D. & D. Ruppert (1997) Fitting a bivariate additive model by local polynomial regression. *Annals of Statistics* 25, 186–211.

Porter, J. (1996) Thesis, Department of Economics, MIT.

Powell, J.L. (1994) Estimation of semiparametric models. In D.F. McFadden & R.F. Engle III (eds.), *The Handbook of Econometrics*, vol. IV. Amsterdam: North Holland.

Stone, C.J. (1980) Optimal rates of convergence for nonparametric estimators. *Annals of Statistics* 8, 1348–1360.

Stone, C.J. (1982) Optimal global rates of convergence for nonparametric regression. *Annals of Statistics* 8, 1040–1053.

Stone, C.J. (1986) The dimensionality reduction principle for generalized additive models. *Annals of Statistics* 14, 592–606.

Tjøstheim, D. & B. Auestad (1994) Nonparametric identification of nonlinear time series: Projections. *Journal of the American Statistical Association* 89, 1398–1409.

Venables, W.N. & B.D. Ripley (1994) Modern applied statistics with S-plus. Berlin: Springer Verlag.

Yang, L.J. & W. Härdle (1997) Nonparametric autoregression with multiplicative volatility and additive mean. *Journal of Time Series Analysis*, forthcoming.

APPENDIX

Let $L(z) = C(F(z))$, $P(z) = A(F(z))$, and

$$D(x,z) = m(x)L(z) + P(z).$$

We shall let $D^{(j)}(x,z)$, $j = 1,2,\dots$ denote partial derivatives of D with respect to z. We let $|A| = (\mathrm{tr}(A^T A))^{1/2}$ for any matrix A.

We use the following assumptions.

Assumption A.

1. The random sample $\{(Y_i, X_i)\}_{i=1}^n$, $Y_i \in \mathbb{R}$, $X_i \in \mathcal{X}$ a compact subset of $\mathbb{R}^d$ is i.i.d. with $E(Y^4) < \infty$.
2. Let $p(x)$ be the marginal density of X with respect to Lebesgue measure and let $m(x) \equiv E(Y\|X = x)$. We suppose that $p(x)$ and $m_1(x_1)$ are twice continuously differentiable with respect to x_1 at all x and that $\inf_{x\in\mathcal{X}} p(x) > 0$.
3. The variance function $\sigma^2(x) = \mathrm{var}(Y|X = x)$ is Lipschitz continuous at all $x \in \mathcal{X}$; i.e., there exists a constant c such that for all x, x', we have $|\sigma^2(x) - \sigma^2(x')| \le c|x - x'|$.
4. The functions $A(\cdot)$, $C(\cdot)$, $G(\cdot)$, and $F(\cdot)$ have bounded continuous second derivatives over any compact interval. The function G is strictly monotonic.
5. The kernel weighting function K is continuous, symmetric about zero, of bounded support, and satisfies $\int K(v)\,dv = 1$.
6. $\{h_n : n \ge 1\}$ is a sequence of nonrandom bounded positive constants satisfying $h_n \to 0$ and $nh_n/\log n \to \infty$.
7. The true parameters $\theta_0^0(x_1) = m_1(x_1)$ and $\theta_1^0(x_1) = m_1'(x_1)$ lie in the interior of the compact parameter space $\Theta = \Theta_0 \times \Theta_1$.

Assumption B.

1. For each $\alpha = 2,\dots,d$, the functions w_α and $\bar{K}$ are continuous on their bounded supports. Furthermore, $\bar{K}$ is Lipschitz continuous; i.e., there exists a finite constant c such that $|\bar{K}(t) - \bar{K}(s)| \le c|t - s|$ for all t, s.
2. The bandwidths satisfy $g_n/h_n \to 0$, $nh_n g_n \to \infty$, and $n^3 g_n^5/\log n \to \infty$.
3. The remainder term in (6) satisfies

$$\max_{1\le i\le n} |\delta_{ni}| = o_p(n^{-2/5}).$$

4. The functions $A(\cdot)$, $C(\cdot)$, and $F(\cdot)$ have bounded continuous third derivatives over any compact interval.

Assumptions A′ and B′ are like A and B except that we replace m, σ^2, A, C, and F by the corresponding quantities derived from φ.

Proof of Theorem 1. Let $\theta_0^0(x_1)$ and $\theta_1^0(x_1)$ be the true local parameters, i.e., $\theta_0^0(x_1) = m_1(x_1)$ and $\theta_1^0(x_1) = m_1'(x_1)$. We first show that $\hat{\theta}(x_1) = (\hat{\theta}_0(x_1), \tilde{\theta}_1(x_1))^T$ consistently estimates $\theta(x_1) = (\theta_0(x_1), \theta_1(x_1))$. By the uniform law of large numbers in Gozalo and Linton (1995), we have

$$\sup_{\theta \in \Theta} |Q_n(\theta) - \bar{Q}_n(\theta)| \to_p 0,$$

where $\bar{Q}_n(\theta) = E\{Q_n(\theta)\}$. This applies because of the smoothness and boundedness conditions on A, C, and F. Furthermore,

$$\begin{aligned}\bar{Q}_n(\theta) &= \int D(X, c + m_2(X_2) + \theta_0 + \theta_1(X_{1i} - x_1)) \frac{1}{h_n} K\left(\frac{x_1 - X_1}{h_n}\right) p(X)\, dX \\ &= \int D(x_1 - uh_n, x_2, c + m_2(x_2) + \theta_0 + \theta_1 h_n u) K(u) p(x_1 - uh_n, x_2)\, du\, dx_2 \\ &\to \int D(x, c + m_2(x_2) + \theta_0) p(x)\, dx_2 := Q_0(\theta_0) \qquad \textbf{(A.1)}\end{aligned}$$

uniformly in $\theta \in \Theta$. The second equality follows by the change of variables $X_1 \to u = (x_1 - X_1)/h_n$, and convergence follows by dominated convergence and continuity. We now apply property 4 of Gourieroux et al. (1984a), which says that, provided F is monotonic,

$$Q_0(\theta_0) \le Q_0(\theta_0^0)$$

with equality if and only if $\theta_0 = \theta_0^0$. This establishes consistency of $\hat{\theta}_0(x_1)$. The derivative parameter $\theta_1(x_1)$ is determined by the next order term (in h_n) through a Taylor expansion of (A.1). When evaluated at (θ_0^0, θ_1), this quantity is, apart from terms that do not depend on θ_1 or are of smaller order, h_n^2 times a constant times

$$Q_1(\theta_1) = \int \left\{ a(x)\theta_1 + \frac{1}{2} b(x)\theta_1^2 \right\} p(x)\, dx_2, \qquad \textbf{(A.2)}$$

where

$$a(x) = \frac{\partial m}{\partial x_1}(x) C'(m(x)) F'(G(m(x))); \qquad b(x) = D''(x, G(m(x))).$$

Note that by properties 1 and 2 of Gourieroux et al. we have

$$D''(x, G(m(x)) = -C'(m(x)) F'(G(m(x)))^2, \qquad \textbf{(A.3)}$$

and we can see that the unique minimum of $Q_1(\theta_1)$ is $\theta_1(x_1) = m_1'(x_1)$ ($C'(m) > 0$ by property 3 of Gourieroux et al.). See Gozalo and Linton (1995) for further discussion. This establishes the consistency of $\hat{\theta}(x_1)$.

We now turn to asymptotic normality. By an asymptotic expansion we have

$$H_n[\hat{\theta}(x_1) - \theta^0(x_1)] = -\left[H_n^{-1}\frac{\partial^2 Q_n(\theta^*(x_1))}{\partial\theta\partial\theta^T}H_n^{-1}\right]^{-1} H_n^{-1}\frac{\partial Q_n(\theta^0(x_1))}{\partial\theta}, \tag{A.4}$$

where $H_n = \text{diag}(1, h_n)$ and $\theta^*(x_1)$ is a vector intermediate between $\hat{\theta}(x_1)$ and $\theta^0(x_1)$. The presentation of (A.4) assumes that the matrix in square brackets is invertible, which we shall show is true with probability tending to one. The score function is

$$\frac{\partial Q_n(\theta^0(x_1))}{\partial\theta} = \frac{-2}{nh_n}\sum_{i=1}^n K\left(\frac{x_1 - X_{1i}}{h_n}\right)\begin{pmatrix} 1 \\ (X_{1i} - x_1)\end{pmatrix}\{Y_i L'(\bar{Z}_i) + P'(\bar{Z}_i)\},$$

whereas the Hessian matrix is

$$\frac{\partial^2 Q_n(\theta)}{\partial\theta\partial\theta^T} = \frac{2}{nh_n}\sum_{i=1}^n K\left(\frac{x_1 - X_{1i}}{h_n}\right)\begin{pmatrix} 1 & (X_{1i} - x_1) \\ (X_{1i} - x_1) & (X_{1i} - x_1)^2\end{pmatrix}\{Y_i L''(\bar{Z}_i(\theta)) + P''(\bar{Z}_i(\theta))\},$$

where

$$\bar{Z}_i(\theta) = c + m_2(X_{2i}) + \theta_0 + \theta_1(X_{1i} - x_1),$$

$$Z_i = c + m_2(X_{2i}) + m_1(X_{1i}),$$

and $\bar{Z}_i = \bar{Z}_i(\theta^0(x_1))$.

We next show that the vector $H_n^{-1}\partial Q_n(\theta^0(x_1))/\partial\theta$ satisfies a central limit theorem, whereas $H_n^{-1}\{\partial^2 Q_n(\theta^*(x_1))/\partial\theta\partial\theta^T\}H_n^{-1}$ is, asymptotically, a positive definite diagonal matrix. Write the score function as

$$\begin{aligned} H_n^{-1}\frac{\partial Q_n(\theta^0(x_1))}{\partial\theta} &= \frac{-2}{nh_n}\sum_{i=1}^n K\left(\frac{x_1 - X_{1i}}{h_n}\right)\varepsilon_i L'(\bar{Z}_i)\begin{pmatrix} 1 \\ \dfrac{X_{1i} - x_1}{h_n}\end{pmatrix} \\ &\quad - \frac{2}{nh_n}\sum_{i=1}^n K\left(\frac{x_1 - X_{1i}}{h_n}\right)D'(X_i, \bar{Z}_i)\begin{pmatrix} 1 \\ \dfrac{X_{1i} - x_1}{h_n}\end{pmatrix} \\ &\equiv T_{n1} + T_{n2}, \end{aligned}$$

where $\varepsilon_i = Y_i - m(X_i) = Y_i - E(Y_i|X_i = x)$. The first random vector is mean zero and has variance matrix

$$\begin{aligned}\mathrm{var}(T_{n1}) &= \frac{4}{nh_n}\,\frac{1}{nh_n}\sum_{i=1}^{n} E\left[K^2\left(\frac{x_1-X_{1i}}{h_n}\right)\sigma^2(X_i)L'(\bar{Z}_i)^2\begin{pmatrix}1 & \dfrac{X_{1i}-x_1}{h_n}\\ \dfrac{X_{1i}-x_1}{h_n} & \left(\dfrac{X_{1i}-x_1}{h_n}\right)^2\end{pmatrix}\right]\\ &= \frac{4}{nh_n}\int K^2(u)\sigma^2(x_1-uh_n,x_2)L'(c+m_2(x_2)+m_1(x_1)+h_n m_1'(x_1)u)^2\\ &\quad\cdot\begin{pmatrix}1 & u\\ u & u^2\end{pmatrix}p(x_1-uh_n,x_2)\,dx_2\,du\\ &= \frac{4}{nh_n}\begin{pmatrix}\|K\|_2^2 & 0\\ 0 & \mu_2(K^2)\end{pmatrix}i_1(x_1)\{1+o(1)\}\end{aligned}$$

by the law of iterated expectation, Fubini's theorem, and dominated convergence, which can be applied using the boundedness and continuity conditions. Finally,

$$\frac{e_1^T T_{n1}}{\sqrt{\dfrac{4}{nh_n}\|K\|_2^2 i_1(x_1)}} \Rightarrow N(0,1),$$

where $e_1^T = (1,0)$, by the Lindeberg central limit theorem (see Gozalo and Linton, 1995, Lemma CLT).

The second term in the score function determines the bias of $\hat{m}_1(x_1)$. By Taylor expansion

$$\begin{aligned}D(X_i,\bar{Z}_i) &= D(X_i,Z_i) + D'(X_i,Z_i)[m_1(X_{1i}) - m_1(x_1) - m_1'(x_1)(X_{1i}-x_1)]\\ &\quad + D''(X_i,Z_i^*)[m_1(X_{1i}) - m_1(x_1) - m_1'(x_1)(X_{1i}-x_1)]^2,\end{aligned}$$

where Z_i^* are intermediate between $\bar{Z}_i$ and Z_i. Note that $D(X_i,Z_i) = 0$ by property 1 of Gourieroux et al. (1984a). Therefore,

$$\begin{aligned}T_{n2} &= \frac{-2}{nh_n}\sum_{i=1}^{n} K\left(\frac{x_1-X_{1i}}{h_n}\right)\begin{pmatrix}1\\ \dfrac{X_{1i}-x_1}{h_n}\end{pmatrix}\\ &\quad\times D'(X_i,Z_i)[m_1(X_{1i}) - m_1(x_1) - m_1'(x_1)(X_{1i}-x_1)]\\ &\quad+\frac{-2}{nh_n}\sum_{i=1}^{n} K\left(\frac{x_1-X_{1i}}{h_n}\right)\begin{pmatrix}1\\ \dfrac{X_{1i}-x_1}{h_n}\end{pmatrix}\\ &\quad\times D''(X_i,Z_i^*)[m_1(X_{1i}) - m_1(x_1) - m_1'(x_1)(X_{1i}-x_1)]^2\\ &= \frac{-1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x_1-X_{1i}}{h_n}\right)\begin{pmatrix}1\\ \dfrac{X_{1i}-x_1}{h_n}\end{pmatrix}m_1''(x_1)(X_{1i}-x_1)^2 D'(X_i,Z_i) + o_p(h_n^2) \qquad \textbf{(A.5)}\\ &= -h_n^2\begin{pmatrix}\mu_2(K)\\ 0\end{pmatrix}m_1''(x_1)i_1(x_1) + o_p(h_n^2), \qquad \textbf{(A.6)}\end{aligned}$$

where (A.5) follows from the fact that for some $c < \infty$,

$$\sup_{|t_1 - x_1| < ch_n} |m_1(t_1) - m_1(x_1) - m_1'(x_1)(t_1 - x_1) - \tfrac{1}{2}m_1''(x_1)(t_1 - x_1)^2| = o_p(h_n^2)$$

and the fact that $\bar{Z}_i$ and Z_i, and hence Z_i^*, are bounded, whereas (A.6) follows by a standard law of large numbers, change of variables, and dominated convergence arguments.

By applying the same uniform law of large numbers and dominated convergence arguments we used in the consistency proof, we have that

$$\sup_{\theta \in \Theta_n} \left| H_n^{-1} \left(\frac{\partial^2 Q_n(\theta)}{\partial\theta\partial\theta^T} - \frac{\partial^2 Q_n(\theta^0(x_1))}{\partial\theta\partial\theta^T} \right) H_n^{-1} \right| = o_p(1),$$

where Θ_n is a shrinking neighborhood of θ^0. Note that this only requires two continuous derivatives, because if $\sup|\bar{Z}_i(\theta) - Z_i| = o_p(1)$, then $\sup|g(\bar{Z}_i(\theta)) - g(Z_i)| = o_p(1)$ for any uniformly continuous function g. Furthermore,

$$\begin{aligned}
\frac{\partial^2 Q_n(\theta^0(x_1))}{\partial\theta\partial\theta^T} &= E\left[\frac{\partial^2 Q_n(\theta^0(x_1))}{\partial\theta\partial\theta^T} \right] + o_p(1) \\
&= \frac{-2}{nh_n} \sum_{i=1}^{n} E\left[K\left(\frac{x_1 - X_{1i}}{h_n} \right) D''(X_i, \bar{Z}_i) \begin{pmatrix} 1 & \dfrac{X_{1i} - x_1}{h_n} \\ \dfrac{X_{1i} - x_1}{h_n} & \left(\dfrac{X_{1i} - x_1}{h_n} \right)^2 \end{pmatrix} \right] \\
&\quad + o_p(1) \\
&\to_p -2 \int D''(x, G(m(x))) p(x)\, dx_2 \begin{pmatrix} 1 & 0 \\ 0 & \mu_2(K) \end{pmatrix},
\end{aligned} \tag{A.7}$$

where the equalities follow by a law of large numbers, whereas the third line follows using dominated convergence and continuity arguments as previously. Applying (A.3), we find that the (1,1) element of (A.7) is $-2j_1(x_1)$ as required. ■

Proof of Theorem 2. Assumption B implies that

$$\tilde{c} - c + \max_{1 \le i \le n} |\tilde{m}_2(X_{2i}) - m_2(X_{2i})| = O_p\left(\sqrt{\frac{\log n}{ng_n}} \right) + o_p(n^{-2/5}). \tag{A.8}$$

For any θ, let

$$\tilde{Z}_i(\theta) = \tilde{c} + \tilde{m}_2(X_{2i}) + \theta_0 + \theta_1(X_{1i} - x_1)$$

and let $\tilde{Z}_i = \tilde{Z}_i(\theta^0)$ and $\bar{Z}_i = \bar{Z}_i(\theta^0)$ as before. Define also $\eta_{ni}(\theta) = \tilde{Z}_i(\theta) - \bar{Z}_i(\theta) = \tilde{Z}_i - \bar{Z}_i = \eta_{ni}$. Expanding $D(X_i, \tilde{Z}_i(\theta))$ and its derivatives about $D(X_i, \bar{Z}_i(\theta))$ in a Taylor series, we get (for $j = 0,1$ and $r = 1,2,\ldots$),

$$D^{(j)}(X_i, \tilde{Z}_i(\theta)) = \sum_{\ell=0}^{r-1} D^{(j+\ell)}(X_i, \bar{Z}_i(\theta))\eta_{ni}^{\ell} + D^{(j+r)}(X_i, \bar{Z}_i^{*j}(\theta))\eta_{ni}^{r}, \tag{A.9}$$

provided the relevant derivatives exist, where $\bar{Z}_i^{*j}(\theta)$ are intermediate between $\widetilde{Z}_i(\theta)$ and $\bar{Z}_i(\theta)$. Our conditions B imply that $\max_{1\le i\le n}|\eta_{ni}| = O_p\big(\sqrt{\log n/ng_n}\big) + o_p(n^{-2/5})$. Furthermore, although $\widetilde{Z}_i(\theta)$ can be unbounded, with a probability tending to one all $\widetilde{Z}_i(\theta)$ lie in the compact support of $\bar{Z}_i(\theta)$. Therefore, we have that

$$\sup_{\theta\in\Theta_n}\max_{1\le i\le n}|D^{(j)}(X_i,\widetilde{Z}_i(\theta)) - D^{(j)}(X_i,\bar{Z}_i(\theta))| = O_p\left(\sqrt{\frac{\log n}{ng_n}}\right) + o_p(n^{-2/5}),$$
$$j = 0,1,2, \tag{A.10}$$

by Assumption B4. A similar result evidently holds for L, P, and their derivatives. Therefore,

$$\begin{aligned}
&\sup_{\theta\in\Theta}|\widetilde{Q}_n(\theta) - Q_n(\theta)|\\
&\quad\le \frac{2}{nh_n}\sum_{i=1}^{n}\left|K\left(\frac{x_1 - X_{1i}}{h_n}\right)\varepsilon_i\right|\cdot\sup_{\theta\in\Theta}\max_{1\le i\le n}|L(\widetilde{Z}_i(\theta)) - L(\bar{Z}_i(\theta))|\\
&\qquad+\frac{1}{nh_n}\sum_{i=1}^{n}\left|K\left(\frac{x_1 - X_{1i}}{h_n}\right)\right|\cdot\sup_{\theta\in\Theta}\max_{1\le i\le n}|D(X_i,\widetilde{Z}_i(\theta)) - D(X_i\,\bar{Z}_i(\theta))|\\
&\quad= O_p\left(\sqrt{\frac{\log n}{ng_n}}\right) + o_p(n^{-2/5}) = o_p(1).
\end{aligned}$$

Therefore, $\hat{m}_1^*(x_1) \to_p m_1(x_1)$.

We now turn to the asymptotic distribution. The argument is based on showing that the feasible score and Hessian matrix are sufficiently close to their infeasible counterparts. We show that

$$\left|H_n^{-1}\left(\frac{\partial\widetilde{Q}_n(\theta^0(x_1))}{\partial\theta_0} - \frac{\partial Q_n(\theta^0(x_1))}{\partial\theta_0}\right)\right| = o_p(n^{-2/5}) \tag{A.11}$$

$$\sup_{\theta\in\Theta_n}\left|H_n^{-1}\left(\frac{\partial^2\widetilde{Q}_n(\theta)}{\partial\theta\partial\theta^T} - \frac{\partial^2 Q_n(\theta)}{\partial\theta\partial\theta^T}\right)H_n^{-1}\right| = o_p(1). \tag{A.12}$$

We first show (A.11). We have

$$\begin{aligned}
&H_n^{-1}\left(\frac{\partial\widetilde{Q}_n(\theta^0(x_1))}{\partial\theta_0} - \frac{\partial Q_n(\theta^0(x_1))}{\partial\theta_0}\right)\\
&\quad= \frac{-2}{nh_n}\sum_{i=1}^{n}K\left(\frac{x_1 - X_{1i}}{h_n}\right)\begin{pmatrix}1\\ \left(\dfrac{X_{1i} - x_1}{h_n}\right)\end{pmatrix}\varepsilon_i\{L'(\widetilde{Z}_i) - L'(\bar{Z}_i)\}\\
&\qquad+\frac{2}{nh_n}\sum_{i=1}^{n}K\left(\frac{x_1 - X_{1i}}{h_n}\right)\begin{pmatrix}1\\ \left(\dfrac{X_{1i} - x_1}{h_n}\right)\end{pmatrix}\{D'(X_i,\widetilde{Z}_i) - D'(X_i,\bar{Z}_i)\}\\
&\quad= T_{n3} + T_{n4}.
\end{aligned}$$

In the sequel we shall restrict attention to the first component of these vectors. The second component behaves similarly—the functions $K(u)$ and $K(u)u$ have similar properties. We first examine the first element of T_{n4}, which satisfies

$$T_{n4_1} = \frac{2}{nh_n}\sum_{i=1}^{n} K\left(\frac{x_1 - X_{1i}}{h_n}\right) D''(X_i, \bar{Z}_i)\eta_{ni} + \left\{O_p\left(\sqrt{\frac{\log n}{ng_n}}\right) + o_p(n^{-2/5})\right\}^2$$

by (A.9). By Assumption B3, the remainder term is $o_p(n^{-2/5})$. Furthermore, the leading term of T_{n4_1} is

$$\begin{aligned}
&\frac{2}{nh_n}\sum_{i=1}^{n} K\left(\frac{x_1 - X_{1i}}{h_n}\right) D''(X_i, \bar{Z}_i)\\
&\qquad \cdot\left\{\frac{1}{ng_n}\sum_{\alpha=2}^{d}\sum_{j=1}^{n}\bar{K}\left(\frac{X_{\alpha i} - X_{\alpha j}}{g_n}\right) w_\alpha(X_i, X_j)\varepsilon_j + o_p(n^{-2/5})\right\}\\
&\quad = \frac{1}{n}\sum_{j=1}^{n}\varepsilon_j\left\{\frac{1}{ng_n h_n}\sum_{i=1}^{n} K\left(\frac{x_1 - X_{1i}}{h_n}\right)\sum_{\alpha=2}^{d}\bar{K}\left(\frac{X_{\alpha i} - X_{\alpha j}}{g_n}\right) w_\alpha(X_i, X_j) D''(X_i, \bar{Z}_i)\right\}\\
&\qquad + o_p(n^{-2/5})\\
&\quad \equiv \frac{1}{n}\sum_{j=1}^{n}\varepsilon_j\zeta_{nj} + o_p(n^{-2/5})\\
&\quad = O_p(n^{-1/2}) + o_p(n^{-2/5}).
\end{aligned} \tag{A.13}$$

The reason for (A.13) is as follows. We have $E[T_{n61}|X_1,\ldots,X_n] = 0$, whereas $\text{var}[T_{n61}|X_1,\ldots,X_n] = n^{-2}\sum_{j=1}^{n}\sigma^2(X_j)\zeta_{nj}^2$, where for any $c > 0$,

$$\Pr\left[\frac{1}{n}\sum_{j=1}^{n}\sigma^2(X_j)\zeta_{nj}^2 \geq c\right] \leq \frac{\sup_x \sigma^2(x)E(\zeta_{n1}^2)}{c} \tag{A.14}$$

by identity of distribution and the Markov inequality. Now, $E(\zeta_{n1}^2) = E^2(\zeta_{n1}) + \text{var}(\zeta_{n1})$, where $\sup_n E(\zeta_{n1}) < \infty$ and $\text{var}(\zeta_{n1}) = O(1/ng_nh_n) = o(1)$ by Assumption B2. Therefore, the numerator of the right-hand side of (A.14) is finite, so that $\text{var}[T_{n61}|X_1,\ldots,X_n] = O_p(n^{-1})$. From this, (A.13) follows by an application of Lemma 1, which follows. The same result applies to the second component of T_{n41}.

To handle T_{n3} substitute (A.9) with $j = 1$ to yield that

$$T_{n3_1} = \frac{2}{nh_n}\sum_{i=1}^{n} K\left(\frac{x_1 - X_{1i}}{h_n}\right)\varepsilon_i L''(\bar{Z}_i)\eta_{ni} + \left\{O_p\left(\sqrt{\frac{\log n}{ng_n}}\right) + o_p(n^{-2/5})\right\}^2$$

as before. Finally, the leading term of leading term of T_{n3_1} is

$$\begin{aligned}
&\frac{2}{nh_n}\sum_{i=1}^{n} K\left(\frac{x_1 - X_{1i}}{h_n}\right)\varepsilon_i L''(\bar{Z}_i)\\
&\qquad \cdot\left[\frac{1}{ng_n}\sum_{\alpha=2}^{d}\sum_{j=1}^{n}\bar{K}\left(\frac{X_{\alpha i} - X_{\alpha j}}{g_n}\right) w_\alpha(X_i, X_j)\varepsilon_j + o_p(n^{-2/5})\right]\\
&\quad = \sum_{i=1}^{n}\sum_{j=1}^{n} U_n(V_i, V_j) + o_p(n^{-2/5}),
\end{aligned} \tag{A.15}$$

where $V_i = (X_i, \varepsilon_i)$ and the "U-statistic" kernel is

$$U_n(V_i, V_j) = \frac{2}{n^2 g_n h_n} K\left(\frac{x_1 - X_{1i}}{h_n}\right) \sum_{\alpha=2}^{d} \bar{K}\left(\frac{X_{\alpha i} - X_{\alpha j}}{g_n}\right) \varepsilon_i L''(\bar{Z}_i) w_\alpha(X_i, X_j) \varepsilon_j.$$

The error in (A.15) is $o_p(n^{-2/5})$ because of the uniformity in the expansion (6). We have $E[U_n(V_i, V_i)] = O(1/n^2 g_n)$ and $\mathrm{var}[U_n(V_i, V_i)] = O(1/n^4 g_n^2 h_n)$, whereas $E[U_n(V_i, V_j)] = 0$ and $\mathrm{var}[U_n(V_i, V_j)] = O(1/n^4 g_n h_n)$ for $i \neq j$. Furthermore, $E[U_n(V_i, V_j) U_n(V_i, V_k)] = E[U_n(V_i, V_j) U_n(V_k, V_j)] = 0$. Therefore, by standard arguments, $\sum_{i=1}^{n} \sum_{j=1}^{n} U_n(V_i, V_j) = O_p(1/n\sqrt{g_n h_n})$.

The proof of (A.12) follows by another application of (A.9). The proof for $\hat{m}_1^{**}(x_1)$ is similar and is omitted. ■

Proof of Theorem 3. The proof is very similar to that of Theorems 1 and 2 and is omitted.

In the proof of Theorem 2 we made use of the following lemma (which may be well known, although we have not found any reference to such).

LEMMA 1. *Let (Y_n, X_n) be a sequence of random variables with Y_n scalar and $X_n \in \mathbb{R}^{\ell(n)}$ for some $\ell(n)$. Suppose that $E(Y_n | X_n) = \mu_n(X_n)$ and* $\mathrm{var}(Y_n | X_n) = \sigma_n^2(X_n)$ *almost surely, where $\mu_n(X_n), \sigma_n^2(X_n) \to_p 0$. Then, $Y_n \to_p 0$.*

Proof of Lemma 1. Define $\varepsilon_n = [Y_n - \mu_n(X_n)]/\sigma_n(X_n)$, which has $E(\varepsilon_n | X_n) = 0$ and $\mathrm{var}(\varepsilon_n | X_n) = 1$ (we can suppose without loss of generality that $\mu_n(X_n)$ and $\sigma_n^2(X_n)$ are real valued), and for any sequence c_n,

$$Y_n'(c_n) = \mu_n(X_n) \mathbf{1}[|\mu_n(X_n)| < c_n] + \varepsilon_n \sigma_n(X_n) \mathbf{1}[|\sigma_n^2(X_n)| < c_n].$$

Because both $\mu_n(X_n)$ and $\sigma_n^2(X_n)$ tend to zero in probability, there exists a sequence $c_n \to 0$ for which $Y_n'(c_n) - Y_n \to_p 0$. For this sequence, both $E[Y_n'(c_n)]$ and $\mathrm{var}[Y_n'(c_n)]$ exist and tend to zero, which implies that $Y_n'(c_n) \to_p 0$. ■