

# NONPARAMETRIC LEAST SQUARES ESTIMATION IN DERIVATIVE FAMILIES

Peter Hall<sup>1</sup> Adonis Yatchew<sup>2,3</sup>

**ABSTRACT.** Cost function estimation often involves data on a function and a family of its derivatives. It is known that by using such data the convergence rates of nonparametric estimators can be substantially improved. In this paper we propose series-type estimators which incorporate various derivative data into a single, weighted, nonparametric, least-squares procedure. Convergence rates are obtained, with particular attention being paid to cases in which root- $n$  consistency can be achieved. For low-dimensional cases, it is shown that much of the beneficial impact on convergence rates is realized even if only data on ordinary first-order partials are available. For example, if one incorporates data on factor demands, a two or three-input cost function can be estimated at rates only slightly slower than if it were a function of a single nonparametric variable. In instances where the root- $n$  rate is attained (possibly up to a logarithmic factor), the smoothing parameter can often be chosen very easily, without resort to empirical methods. Moreover, the fact that the optimal rate is root- $n$  can be determined from knowledge of which derivatives are observed, and does not require complex empirical arguments. In other cases, standard cross-validation can be employed. The paper includes simulations and an illustration of cost function estimation.

**KEYWORDS.** Nonparametric regression, cost and factor demand estimation, partial derivative data, curse of dimensionality, dimension reduction, rates of convergence, orthogonal series methods, cross-validation, smoothing parameter selection, statistical smoothing.

**SHORT TITLE.** Derivative families.

March 31, 2008

---

<sup>1</sup> Department of Mathematics and Statistics, University of Melbourne, Parkville, VIC 3010, Australia. p.hall@ms.unimelb.edu.au

<sup>2</sup> Department of Economics, University of Toronto, 150 St George St, Toronto, Ontario M5S 3G7, Canada. yatchew@chass.utoronto.ca

<sup>3</sup> This research was supported by a grant from the Social Sciences and Humanities Research Council of Canada. Portions of it were conducted while the second author was visiting at the Centre for Mathematics and its Applications, Australian National University, Canberra, ACT 0200, Australia.

## 1. INTRODUCTION

Theoretical and empirical analysis of cost functions has a long tradition in economics (see e.g., Shephard 1953, 1970; McFadden 1978; Jorgenson 1986). Since, by Shephard's Lemma, factor demand functions are partial derivatives of the cost function, and therefore share its parameters, joint estimation of costs and factor demands can be implemented to increase the efficiency of parametric modeling.

In nonparametric settings the presence of derivative data can have an even more powerful impact, not only increasing efficiency but improving rates of convergence, in some cases permitting nonparametric estimation at parametric rates; see Hall and Yatchew (2007). In the present paper we suggest methods that optimally combine data on the function of interest and its various observed derivatives. To this end, we propose series-type estimators which incorporate all data within a single weighted nonparametric least squares procedure. The weights assigned to data on each derivative are chosen to reflect the information that they possess in relation to the function of interest. Interestingly, related problems arise also in the theory of stochastic control; for example, some motion sensors record derivatives of functions as well as the functions themselves (see e.g. Murray-Smith and Sbarbaro, 2002; Seeger, 2003, 2004, 2007; Solak *et al.*, 2004). See also Florens *et al.* (1996) who consider estimation of a function and its derivatives using finite-dimensional approximations.

To describe the setting in more detail, suppose noisy data are available on a family of derivatives, say  $\mathcal{G} = \{g^\beta : \beta \in B\}$ , where the argument of these functions may be a vector and  $B$  indexes the set of observed derivatives. It is assumed that  $g \in \mathcal{G}$ . Interest focuses on estimation of one or more members of  $\mathcal{G}$ , for instance  $g$ .

To appreciate the possibilities, consider the case where the argument  $x$  of  $g$  is a scalar, uniformly distributed on the unit interval, and  $g$  is a twice differentiable function which, for simplicity, satisfies  $g(0) = 0$ . Under standard second-derivative assumptions, and given a sample of size  $n$ , local averaging of data on  $y = g(x) + \varepsilon$  will produce an estimator with integrated mean squared error converging at the optimal nonparametric rate,  $O(n^{-4/5})$ . However, if data on  $y_1 = g'(x) + \varepsilon_1$  are available then numerical integration of this information yields an estimator which converges at the parametric mean-square rate,  $O(n^{-1})$ . Local averaging, the culprit underlying the slower rate of convergence (and, more generally, the curse of dimensionality), has

been effectively avoided.

On the other hand, the derivative-based estimator of  $g$ , which uses data only on  $y_1$ , ignores information contained in the level data which may be quite informative, particularly if the derivative data are much noisier than the level data. Moreover, if data are available on additional derivatives (or on integrals) of  $g$  then, though these cannot further improve the rate of convergence, they have the potential to enhance estimator performance and so should be incorporated.

Consider now the more general setting where  $x$  is of dimension  $p$ . Given a family of observed derivatives  $\mathcal{G}$ , and suitable assumptions thereon, there is an optimal rate that a class of estimators can achieve. Though that rate may be attained using data on one or other proper subset of  $\mathcal{G}$ , data on all elements of  $\mathcal{G}$  can be expected to improve the precision of estimation.

For cosine-series estimators, which we examine in detail, derivative data can also improve boundary behavior. Furthermore, in low-dimensional cases we find that much of the beneficial impact on convergence rates is realized even if only data on first-order ordinary partials are available. To illustrate, consider the bivariate setting where data on  $g$  are observed and the cosine-series estimator involves  $r$  terms for each component. In this case the variance component of mean integrated squared error is  $O(r^2/n)$ . However, if in addition data are available on  $\partial g/\partial x_i$ , for  $i = 1, 2$ , then the variance component becomes  $O(n^{-1} \log r)$ . For suitably smooth classes of functions this permits convergence at close to parametric rates.

These results have important implications for cost function estimation. In particular, if data are available on a two or three-input cost function and corresponding factor inputs, these series estimators can yield convergence rates only slightly slower than if costs were a function of a single nonparametric variable, (the quantity of output produced). Even if the production process involves four or more inputs, data on factor demands will produce very substantial improvements in the nonparametric convergence rates that can be achieved.

Our approach yields diverse additional benefits. For example, if root- $n$  consistency is possible, or if it is achievable up to a logarithmic factor, then the smoothing parameter may be easily chosen, without recourse to empirical methods. In other cases, standard cross-validation can be used. The method also readily generalizes to settings where one has data on functions of derivatives.

The paper is organized as follows. Section 2 sets out the basic model. Section 3 defines series-based estimators constructed by minimising weighted sums of squares and, in Theorem 3.1, provides a general formula for weighted mean integrated squared error in  $p$ -variate settings. Section 4 focuses on cosine-series regression and regularly-spaced design, and shows that substantial simplifications are available in this setting. Section 5 discusses mean squared error properties of our estimators for various configurations of observed derivatives. Section 6 provides results of simulations and illustrates the application of our approach to cost function estimation.

Given square matrices  $A = (a_{ij})$  and  $B = (b_{ij})$  of identical dimension, we employ the conventional matrix inner product,  $\langle A, B \rangle = \sum_{i,j} a_{ij} b_{ij} = \text{tr}(A^T B)$ , the latter being the trace of  $A^T B$ . We use  $\otimes$  to denote Kronecker or tensor products. Given a vector  $a$ ,  $\text{diag}(a)$  equals the diagonal matrix with entries given by the components of  $a$ ; and given a matrix  $A$ ,  $\text{diag}(A)$  is the vector consisting of the diagonal entries of  $A$ .

## 2. MODEL

Let  $B$ , denoting the class of derivative types on which we have data, be a set of  $p$ -vectors with nonnegative integer components. Given  $\beta = (\beta_1, \dots, \beta_p) \in B$  we define

$$g^\beta(x) = \frac{\partial^{\beta_1 + \dots + \beta_p}}{\partial x_1^{\beta_1} \dots \partial x_p^{\beta_p}} g(x), \quad (2.1)$$

where  $x = (x_1, \dots, x_p)$ . For each  $\beta \in B$  we observe data pairs  $(X_i, Y_{\beta i})$  generated by

$$Y_{\beta i} = g^\beta(X_i) + \varepsilon_{\beta i}, \quad i = 1, \dots, n, \quad (2.2)$$

where  $g$  is a scalar function of  $p$  variables,  $X_i$  is a  $p$ -vector that we take to lie in the unit cube  $\mathcal{R} = [0, 1]^p$ ,  $Y_{\beta i}$  is a scalar, and the errors  $\varepsilon_{\beta i}$  in (2.2) are independent for distinct pairs  $(\beta, i)$ , having a distribution that depends only on  $\beta$  and for which  $E(\varepsilon_{\beta i}) = 0$  and  $\text{var}(\varepsilon_{\beta i}) = \sigma_\beta^2 < \infty$ .

We assume that  $B$  includes the zero vector, so that data are observed on  $g$  itself as well as potentially on its derivatives. It is of course possible to take  $n$ , in (2.2), to depend on  $\beta$ . Indeed, if we replace  $n$  by  $n_\beta$ , say, then as long as we can choose an  $n$  with the property that  $\rho_\beta \equiv n_\beta/n$  is bounded away from zero and infinity for each  $\beta \in B$ , the results that we shall give below remain unchanged provided we replace

the variance  $\sigma_\beta^2$ , for example in the definition of  $U$  at (3.11), by  $\sigma_\beta^2/\rho_\beta$ . Therefore, in theoretical terms, the assertion (2.2) that for each derivative we have the same  $n$  is made without essential loss of generality. Likewise, our method for estimating  $g$  is readily generalised to cases where the value of  $n$  varies for different values of  $\beta$ .

It was shown by Hall and Yatchew (2007) that a set of conditions sufficient for estimating  $g$  root- $n$  consistently from the data at (2.2), is: (i)  $B$  includes the class of all  $p$ -vectors  $\beta$  of zeros and ones, (ii)  $g^\beta$  exists and is bounded on  $\mathcal{R}$  whenever  $\beta \in B$ , and (iii) the distributions of the design data  $X_i$  and the errors  $\varepsilon_{\beta_i}$  are sufficiently regular. In fact, (i) can be replaced by the following weaker condition: for each  $\beta' \in B$ ,  $B$  includes a vector  $\beta = \beta(\beta')$  which has a strictly positive integer in each position where  $\beta'$  has a 1, and a zero in each other position.

### 3. ESTIMATORS AND THEIR BASIC PROPERTIES

*3.1. Generalised Fourier expansion of  $g$ .* Write  $\mathcal{J}$  for the set of vectors  $j = (j_1, \dots, j_p)$  whose components are positive integers. Let  $r = (r_1, \dots, r_p)$  be a vector of positive integers and let  $\mathcal{J}^1$  be the subset of  $\mathcal{J}$  for which  $1 \leq j_s \leq r_s$ , for  $s = 1, \dots, p$ . Put  $\tilde{r} = \#\mathcal{J}^1 = \prod_{1 \leq s \leq p} r_s$  and define  $\mathcal{J}^2 \equiv \mathcal{J} \setminus \mathcal{J}^1$ . The set  $\mathcal{J}$  will index our generalised Fourier basis, and  $\mathcal{J}^1$  will index a subset of basis functions used to construct our estimator.

Let  $\phi_1, \phi_2 \dots$  be a complete sequence in the class  $L_2([0, 1])$  of square-integrable functions on  $[0, 1]$ . Given  $j = (j_1, \dots, j_p) \in \mathcal{J}$  and  $x = (x_1, \dots, x_p) \in [0, 1]^p$ , define

$$\phi_j(x) = \phi_{j_1}(x_1) \dots \phi_{j_p}(x_p). \quad (3.1)$$

Then,

$$g = \sum_{j \in \mathcal{J}} \alpha_j^0 \phi_j, \quad (3.2)$$

where the series converges in  $L_2$  and  $\alpha_j^0 = \int_{\mathcal{R}} g \phi_j$  denotes the true value of a generalised Fourier coefficient.

We shall approximate the expansion (3.2) using a finite series based on the frequency cut-off vector  $r = (r_1, \dots, r_p)$ . Specifically, for each  $s = 1, \dots, p$ , assemble the  $r_s$ -vector  $(\phi_1(x_s), \dots, \phi_{r_s}(x_s))$  of scalar functions. From these, construct the  $\tilde{r}$ -dimensional tensor product of functions on  $\mathcal{R}$ ,

$$\begin{aligned} \Phi(x_1, \dots, x_p) = & (\phi_1(x_1), \dots, \phi_{r_1}(x_1)) \otimes (\phi_1(x_2), \dots, \phi_{r_2}(x_2)) \\ & \otimes \dots \otimes (\phi_1(x_p), \dots, \phi_{r_p}(x_p)). \end{aligned} \quad (3.3)$$

Our approximation to  $g$  will be the finite series,

$$g_1(x) = g_1(x | r) = \Phi(x) \alpha, \quad (3.4)$$

where  $\alpha$  is an  $\tilde{r}$ -dimensional column vector of real scalars. Defining derivatives as in (2.1), and in particular taking  $\phi_j^\beta(x) = \phi_j^\beta(x_1, \dots, x_p) = \phi_{j_1}^{\beta_1}(x_1) \dots \phi_{j_p}^{\beta_p}(x_p)$ , we have:  $g_1^\beta(x) = g_1^\beta(x | r) = \Phi^\beta(x) \alpha$ , where  $\Phi^\beta(x)$  is obtained by applying (2.1) to each component of  $\Phi(x)$  in (3.3).

**3.2. Definition of  $\hat{g}$ .** Write  $X$  for the  $n \times p$  design matrix with  $i$ th row  $X_i$ . Evaluate  $\Phi$  at each of  $X_1, \dots, X_n$ , and let  $\Phi_X$  denote the  $n \times \tilde{r}$  matrix with  $i$ th row  $\Phi(X_i)$ . Similarly, let  $\Phi_X^\beta$  be the  $n \times \tilde{r}$  matrix with  $i$ th row  $\Phi^\beta(X_i)$ . For any  $j \in \mathcal{J}$ , define the  $n$ -dimensional column vector

$$\phi_{jX}^\beta = (\phi_j^\beta(X_1), \dots, \phi_j^\beta(X_n))^T. \quad (3.5)$$

We may now rewrite (2.2) as

$$Y_\beta = g^\beta(X) + \varepsilon_\beta = \Phi_X^\beta \alpha^0 + \sum_{j \in \mathcal{J}^2} \alpha_j^0 \phi_{jX}^\beta + \varepsilon_\beta,$$

where  $Y_\beta = (Y_{\beta 1}, \dots, Y_{\beta n})^T$ ,  $g^\beta(X) = (g^\beta(X_1), \dots, g^\beta(X_n))^T$ ,  $\varepsilon_\beta = (\varepsilon_{\beta 1}, \dots, \varepsilon_{\beta n})^T$  and  $\alpha^0 = (\alpha_j^0)$  is the true form of the  $\tilde{r}$ -vector  $\alpha$  at (3.4).

We shall choose estimators  $\hat{\alpha}$  of  $\alpha$  by minimising the objective function

$$S(\alpha) = \frac{1}{n} \sum_{\beta \in B} \theta_\beta (Y_\beta - \Phi_X^\beta \alpha)^T (Y_\beta - \Phi_X^\beta \alpha), \quad (3.6)$$

where the  $\theta_\beta$  are weights whose choice will be discussed in section 3.4. Define the  $\tilde{r} \times \tilde{r}$  matrices

$$\Psi_X^\beta = n^{-1} (\Phi_X^\beta)^T \Phi_X^\beta, \quad M = \sum_{\beta \in B} \theta_\beta \Psi_X^\beta = n^{-1} \sum_{\beta \in B} \theta_\beta (\Phi_X^\beta)^T \Phi_X^\beta. \quad (3.7)$$

Differentiating the sum of squares at (3.6) with respect to  $\alpha$ , and equating to zero, yields

$$M\alpha = \left\{ \frac{1}{n} \sum_{\beta \in B} \theta_\beta (\Phi_X^\beta)^T \Phi_X^\beta \right\} \alpha = \frac{1}{n} \sum_{\beta \in B} \theta_\beta (\Phi_X^\beta)^T Y_\beta,$$

in which case

$$\begin{aligned} \hat{\alpha} &= M^{-1} \left\{ \frac{1}{n} \sum_{\beta \in B} \theta_\beta (\Phi_X^\beta)^T \left( \Phi_X^\beta \alpha^0 + \sum_{j \in \mathcal{J}^2} \alpha_j^0 \phi_{jX}^\beta + \varepsilon_\beta \right) \right\} \\ &= \alpha^0 + M^{-1} \left\{ \frac{1}{n} \sum_{\beta \in B} \theta_\beta (\Phi_X^\beta)^T \left( \sum_{j \in \mathcal{J}^2} \alpha_j^0 \phi_{jX}^\beta + \varepsilon_\beta \right) \right\}. \end{aligned} \quad (3.8)$$

Our estimator of  $g(x)$  is given by

$$\hat{g}(x) = \Phi(x) \hat{\alpha},$$

and the error of the estimator can be written as

$$\hat{g}(x) - g(x) = \Phi(x) (\hat{\alpha} - \alpha^0) - \sum_{j \in \mathcal{J}^2} \alpha_j^0 \phi_j(x). \quad (3.9)$$

Since both  $g$  and  $g_1$ , the latter defined at (3.4), are linear forms in the linearly independent functions  $\phi_j$ ; and since we may construct orthonormal functions (with respect to a weight,  $w$ ) from the  $\phi_j$  using the linear Gram-Schmidt procedure; then no loss of generality is incurred by supposing that the functions  $\phi_j$  are already orthonormal with respect to a weight function,  $\omega$  say. Therefore we shall impose this condition below. In this case the following expansion of weighted integrated squared error of  $\hat{g}$  follows from (3.9):

$$\int_{\mathcal{R}} (\hat{g} - g)^2 \omega = (\hat{\alpha} - \alpha^0)^{\text{T}} (\hat{\alpha} - \alpha^0) + \sum_{j \in \mathcal{J}^2} (\alpha_j^0)^2. \quad (3.10)$$

**3.3. Weighted mean integrated square error of  $\hat{g}$ .** In this section we expand the expected value of the left-hand side of (3.10), i.e. of weighted mean integrated squared error (MISE). Write  $E$  for expectation conditional on the design data  $X$ , and define

$$U = n^{-1} \sum_{\beta \in B} \theta_{\beta}^2 \sigma_{\beta}^2 \Psi_X^{\beta}, \quad u = n^{-1} \sum_{\beta \in B} \theta_{\beta} (\Phi_X^{\beta})^{\text{T}} \sum_{j \in \mathcal{J}^2} \alpha_j^0 \phi_{jX}^{\beta}, \quad (3.11)$$

denoting an  $\tilde{r} \times \tilde{r}$  matrix and an  $\tilde{r} \times 1$  vector, respectively.

**Theorem 3.1.** *Assume that the  $\phi_j$ , for  $j \in \mathcal{J}$ , form a complete orthonormal basis with respect to the weight function  $\omega$ , that  $\sigma_{\beta}^2 < \infty$  for each  $\beta \in B$ , that  $\int_{\mathcal{R}} g^2 < \infty$  and that  $M$ , defined in (3.7), is nonsingular. Then,*

$$\int_{\mathcal{R}} E(\hat{g} - g)^2 \omega = \langle M^{-2}, U \rangle + \langle M^{-2}, uu^{\text{T}} \rangle + \sum_{j \in \mathcal{J}^2} (\alpha_j^0)^2. \quad (3.12)$$

This theorem, and its analogues in related settings, can be used to derive rates of convergence and, more specifically, to point to settings where root- $n$  consistent

estimation of  $g$  is possible. Sections 4 and 5 will explore these issues. Theorems 3.1 and 4.1 will be derived in Appendices A.1 and A.2, respectively.

*3.4. Choice of  $\theta_\beta$ .* Only the first term,  $\langle M^{-2}, U \rangle$ , on the right-hand side of (3.12) represents a contribution from variance; the other two terms come from bias. It can be deduced from formulae (3.7) and (3.11) for  $M$  and  $U$ , respectively, that  $\langle M^{-2}, U \rangle$  is asymptotically minimised by choosing the weight,  $\theta_\beta$ , in the objective function  $S(\alpha)$  at (3.6), to be inversely proportional to  $\sigma_\beta^2$ . This approach to determining  $\theta_\beta$  is particularly relevant in problems where  $g$  can be estimated root- $n$  consistently, since in those instances the bias terms are asymptotically negligible. Details will be given in section 4.

Using this choice of the weights, the objective function becomes

$$S(\alpha) = \frac{1}{n} \sum_{\beta \in B} \frac{1}{\sigma_\beta^2} (Y_\beta - \Phi_X^\beta \alpha)^\top (Y_\beta - \Phi_X^\beta \alpha).$$

Provided  $g$  is continuous, the value of  $\sigma_\beta^2$  here can be estimated consistently using difference-based methods:

$$\hat{\sigma}_\beta^2 = \frac{1}{2n} \sum_{i=1}^n (Y_{\beta i} - Y_{\beta i'})^2,$$

where  $i'$  denotes an index which, among indices not equal to  $i$ , has the property that  $X_{i'}$  is nearest to  $X_i$ . (In the case of a tie for the latter property we should average over all the corresponding values of  $(Y_{\beta i} - Y_{\beta i'})^2$ .) This approach has been found to give good performance in related problems; see, for example, Buckley et al. (1988), Yatchew (1988), Hall et al. (1990), Dette et al. (1998), Fan and Yao (1998), Spokoiny (2002), Yatchew (2003) and Munk et al. (2005).

## 4. ORTHOGONAL BASIS

*4.1. Trigonometric-series example.* Among the cases covered by Theorem 3.1, the most transparent is that where the  $\tilde{r} \times \tilde{r}$  matrix  $M$  is diagonal. In this context we shall consider a cosine-series example. To give definitions, suppose temporarily that  $p = 1$  and consider the standard cosine basis on  $[0, 1]$ :

$$\phi_1(x) \equiv 1, \quad \phi_{j+1}(x) = 2^{1/2} \cos(j\pi x) \quad \text{for } j \geq 1. \quad (4.1)$$

The functions at (4.1) are exactly orthonormal on  $[0, 1]$ . Moreover, if we gather our data at

$$X_i = \frac{i}{n} - \frac{1}{2n}, \quad i = 1, \dots, n, \quad (4.2)$$

then,

$$\frac{1}{n} \sum_{i=0}^n \phi_j^\beta(X_i) \phi_k^\beta(X_i) = \delta_{jk} \{\pi^2(j-1)(k-1)\}^\beta, \quad \text{for } 1 \leq j, k \leq n, \quad (4.3)$$

where  $\delta_{jk}$  denotes the Kronecker delta. Conditions (4.1) and (4.2) define one of eight different discrete cosine transforms; see Martucci (1994).

When  $p \geq 1$  we assume that the design points  $X_i$  are regularly spaced in the unit cube  $[0, 1]^p$ , in the sense that:

$$\{X_1, \dots, X_n\} \text{ can be written as the set of all vectors } ((i_1 - \frac{1}{2})/m, \dots, (i_p - \frac{1}{2})/m), \text{ where } 1 \leq i_1, \dots, i_p \leq m \text{ and } n = m^p. \quad (4.4)$$

This requirement for a regular grid can be relaxed, but doing so involves greater complexity. When (4.4) holds, the basis functions can be constructed as at (3.1) and (3.3), starting with the functions  $\phi_j$  defined at (4.1). In this case, equation (4.3) becomes

$$\frac{1}{n} \sum_{i=0}^n \phi_j^\beta(X_i) \phi_k^\beta(X_i) = \delta_{jk} \{\pi^2(j_1-1)(k_1-1)\}^{\beta_1} \dots \{\pi^2(j_p-1)(k_p-1)\}^{\beta_p} \quad \text{for } 1 \leq j_1, \dots, j_p, k_1, \dots, k_p \leq m, \quad (4.5)$$

where  $\delta_{jk} = 1$  if the vectors  $j$  and  $k$  are identical and zero otherwise.

**4.2. Properties of mean integrated squared error.** Let  $m$  be as in (4.4), and given  $s = 1, \dots, p$ , let  $r_s \in [1, m]$  be an integer. Define the following vectors of length  $r_s$ :  $\kappa_s = (0, 1, 2, \dots, (r_s - 1))^T$ ,  $\kappa_s^0 = (1, 1, \dots, 1)^T$  and  $\kappa_s^b = (0, 1^b, \dots, (r_s - 1)^b)^T$ , the latter for any integer  $b \neq 0$ . Here the superscripts denote indices, not powers. As in section 3.1, write  $\mathcal{J}$  for the set of all  $j = (j_1, \dots, j_p)$  for which each  $j_s$  is a positive integer, let  $\mathcal{J}^1$  be the set of  $j$  such that  $1 \leq j_s \leq r_s$  for  $1 \leq s \leq p$ , and define  $\mathcal{J}^2 = \mathcal{J} \setminus \mathcal{J}^1$ . Recall that user-chosen positive weights  $\theta_\beta$  appear in the objective function at (3.6), and put  $c_\beta = \pi^{2(\beta_1 + \dots + \beta_p)}$ ,

$$\Lambda = \sum_{\beta \in B} c_\beta \theta_\beta \text{diag}(\otimes_{s=1}^p \kappa_s^{2\beta_s}), \quad \Upsilon = \sum_{\beta \in B} c_\beta \theta_\beta^2 \sigma_\beta^2 \text{diag}(\otimes_{s=1}^p \kappa_s^{2\beta_s}), \quad (4.6)$$

$$\text{var} = n^{-1} \langle \Lambda^{-2}, \Upsilon \rangle = n^{-1} \text{tr}(\Lambda^{-2} \Upsilon), \quad \text{bias}^2 = \sum_{j \in \mathcal{J}^2} (\alpha_j^0)^2. \quad (4.7)$$

Consider the following conditions:

- (a) When  $p = 1$  the functions  $\phi_j$  are as given by (4.1) and for general  $p$  they are constructed as suggested by that formula and (3.1); (b) the design points  $X_i$  are regularly spaced, as indicated at (4.4); (c)  $0 < \sigma_\beta^2 < \infty$  for each  $\beta \in B$ ; (d)  $B$  contains zero, i.e. the  $p$ -vector  $(0, \dots, 0)$ ; (e) for each  $\beta \in B$ ,  $|\alpha_j^0| = O(j_1^{-(\beta_1 + \delta)} \dots j_p^{-(\beta_p + \delta)})$  uniformly in  $j$ , where  $\delta > \max\{1, p/2\}$ ; and (f)  $\omega \equiv 1$ . (4.8)

Parts (a)–(c) of (4.8) merely assert the context of a cosine-series basis, regularly-spaced design and error distributions with finite, nonvanishing variance; part (d) asks that we have data on  $g$  itself, and part (e) requires that the function  $g$  be sufficiently smooth, in the sense of a periodic function. If  $g$  is supported on an interval  $[a, b]$  that lies strictly inside  $[0, 1]$ , i.e. for which  $0 < a < b < 1$ , and if  $g(x)$  decreases smoothly to zero as  $x$  decreases to  $a$  or increases to  $b$ , then (4.8)(e) holds.

**Theorem 4.1.** *If (4.8) holds then*

$$\int_{\mathcal{R}} E(\hat{g} - g)^2 = \text{var} + \text{bias}^2 + o(\text{var}). \quad (4.9)$$

**4.3. Asymptotic distribution of mean integrated squared error.** It is conventional, in function-estimation problems where the convergence rate is slower than  $n^{-1/2}$ , for integrated squared error (ISE) to be asymptotic to MISE. That is, while ISE and MISE both converge to zero (the former converging in probability), their ratio converges to 1 in probability. Therefore ISE is, to first order, non-random. See, for example, Hall (1984), Ioannides (1992), Tenreiro (1998), Kim and Cox (2001) and Ghorai (2003).

However, in the root- $n$  consistent cases discussed in this paper, the ratio of ISE and MISE for  $\hat{g}$  has a nondegenerate limiting distribution. Equivalently, ISE multiplied by  $n$  converges in distribution, although not in probability, to a random variable which is not almost surely constant.

To construct the asymptotic distribution mentioned above, assume for simplicity that  $p = 1$ . Suppose too that the errors  $\epsilon_{\beta i}$ , for  $\beta \in B$  and  $1 \leq i \leq n$ , introduced at (2.2), are totally independent. Let  $Z_{\beta j}$ , for  $\beta \in B$  and  $j \geq 1$ , denote a sequence of independent, standard normal random variables, and put  $\lambda_j = \{\sum_{\beta \in B} c_{\beta} \theta_{\beta} (j-1)^{2\beta}\}^2$  and

$$Z = \sum_{\beta \in B} c_{\beta} \theta_{\beta}^2 \sigma_{\beta}^2 \sum_{j=1}^{\infty} \frac{(j-1)^{2\beta}}{\lambda_j} Z_{\beta j}^2. \quad (4.10)$$

Note that, provided  $B$  contains a nonzero element, the infinite series in  $j$  at (4.10) converges, with probability 1, for each  $\beta \in B$ . A proof of the following theorem is similar to that of Theorem 4.1.

**Theorem 4.2.** *Assume that (4.8) holds, that  $B$  contains a nonzero element, and*

that  $(\text{bias})^2 = o(n^{-1})$ . Then,

$$n \int_{\mathcal{R}} (\hat{g} - g)^2 \rightarrow Z \quad (4.11)$$

in distribution as  $n \rightarrow \infty$ .

The condition  $(\text{bias})^2 = o(n^{-1})$ , imposed in Theorem 4.2, is ensured by simply taking  $r = n$ . Comparing (4.9) and (4.11) it becomes apparent that  $n \text{var} \sim E(Z)$ , where  $\text{var}$  is as defined at (4.7). This property can be proved to hold under the conditions of Theorem 4.2. Versions of these results can be derived in  $p$ -variate problems.

*4.4. Choice of the smoothing parameter,  $r$ .* Cases where root- $n$  consistency, i.e.  $\text{MISE} = O(n^{-1})$ , is attainable can be identified directly from the set  $B$ . They do not require empirical arguments. Section 5.2 will give details. In those instances, choosing each  $r_s = m$ , the largest possible value, is often appropriate. The variance component remains of order  $n^{-1}$ , even for such large values of  $r_s$ .

Also of practical interest are cases where the optimal order of MISE equals  $n^{-1} \log n$ . This occurs, for example, when  $p = 2$  and  $B = \{(0, 0), (1, 0), (0, 1)\}$ . (See section 5.2 for details.) Here too taking each  $r_s = m$  gives the optimal convergence rate.

In other cases, standard methods can be used to select  $r$ . Indeed, let  $\hat{g}_{-i}(\cdot | r)$  denote the version of the estimator  $\hat{g}$  computed when the pair  $(X_i, Y_{0i})$  is omitted from the dataset, and write

$$T(r) = \frac{1}{n} \sum_{i=1}^n \{Y_{0i} - \hat{g}_{-i}(X_i | r)\}^2$$

for the standard cross-validation form of the sum of squares for error. An empirical procedure for choosing  $r$  is to select it to minimise  $T(r)$ .

## 5. PARTICULAR CASES OF THE MEAN INTEGRATED SQUARED ERROR EXPANSION

*5.1. Calculating the term  $\text{bias}^2$  in (4.9).* The value of  $\text{bias}^2$  depends on the size of the  $\alpha_j^0$ , and can be assessed by analytical methods, as follows. Let  $[j]^\beta = j_1^{\beta_1} \dots j_p^{\beta_p}$  and define  $\vec{2} = (2, \dots, 2)$  to be the  $p$ -vector of twos. If the second derivative of  $g$  with respect to each component, i.e. if the function  $g^{\vec{2}}$ , is bounded on  $\mathcal{R}$ , then an

integration by parts argument shows that  $[j]^{\bar{2}} |\alpha_j^0|$  is bounded uniformly in  $j$ , in which case  $\text{bias}^2 = O(r_1^{-3} + \dots + r_p^{-3})$ .

This in turn can lead to circumstances where  $\text{bias}^2$  dominates  $\text{var}$ . For example, suppose data are available on all first-order own- and cross-partials of  $g$ . Then  $\text{var} = O(n^{-1})$ , its fastest rate. Even if  $r_1, \dots, r_p$  are set equal to  $m$ ,  $\text{bias}^2$  may be as large as  $O(m^{-3}) = O(n^{-3/p})$ , in which case it would dominate  $\text{var}$  so long as  $p > 3$ .

*5.2. Calculating the term  $\text{var}$  appearing in (4.7) and (4.9).* In this section we elucidate properties of  $n \text{var} = \langle \Lambda^{-2}, \Upsilon \rangle$ . We shall put special emphasis on cases where root- $n$  consistency is possible, which requires in particular that

$$\limsup_{r_1, \dots, r_p \rightarrow \infty} \langle \Lambda^{-2}, \Upsilon \rangle < \infty. \quad (5.1)$$

Assume first that  $p = 1$ . Then if  $B = \{0\}$ , meaning that only data on  $g$  itself are observed, we have  $\theta_0 = 1$ ,  $\Lambda = I$  (the  $r \times r$  identity matrix),  $\Upsilon = \sigma_0^2 I$  and  $\langle \Lambda^{-2}, \Upsilon \rangle = \sigma_0^2 r$ , whence it follows that  $\text{var} = n^{-1} \sigma_0^2 r$ . The latter is the conventional asymptotic formula for MISE of a nonparametric estimator based on a generalised Fourier series; see, for example, Kronmal and Tarter (1968). In this setting the size of MISE involves a first-order compromise between “var” and “bias<sup>2</sup>”.

More generally, but still in the case  $p = 1$ ,

$$\langle \Lambda^{-2}, \Upsilon \rangle = \sum_{j=0}^{r-1} \left( \sum_{\beta \in B} c_\beta \theta_\beta j^{2\beta} \right)^{-2} \left( \sum_{\beta \in B} c_\beta \theta_\beta^2 \sigma_\beta^2 j^{2\beta} \right), \quad (5.2)$$

and the series on the right-hand side is convergent provided that  $B$  contains an element in addition to 0. In this case we can write  $\text{var} \sim C n^{-1}$ , where  $C > 0$  equals the limit as  $r \rightarrow \infty$  of the right-hand side of (5.2).

For general  $p \geq 1$ ,

$$\langle \Lambda^{-2}, \Upsilon \rangle \asymp \sum_{j_1=0}^{r_1} \dots \sum_{j_p=0}^{r_p} \left( \sum_{\beta \in B} j_1^{\beta_1} \dots j_p^{\beta_p} \right)^{-2}, \quad (5.3)$$

where we interpret  $0^0$  as 1, and  $a \asymp b$  means that  $a/b$  is bounded away from zero and infinity as the parameter values (here, the values of  $r_1, \dots, r_p$ ) alter. At times it is convenient to approximate the multiple summation at (5.3) using multiple integrals. To give a representation of this type, let  $\mathcal{V}$  denote the set of all  $2^p$  distinct

$p$ -vectors of zeros and ones. Given  $v = (v_1, \dots, v_p) \in \mathcal{V}$ , write  $a(v)$  for the number of nonzero components of  $v$ , and let  $\mathcal{S}(v) = (v_{k_{v_1}}, \dots, v_{k_{v_{a(v)}}})$  be the sequence of these components. Write  $\mathcal{J}^1(v)$  for the set of vectors  $j(v) = (j_{k_{v_1}}, \dots, j_{k_{v_{a(v)}}})$ , of length  $a(v)$ , for which  $j_{k_{v_1}} \dots j_{k_{v_{a(v)}}} \neq 0$  and  $1 \leq j_s \leq r_s$  for  $s = k_{v_1}, \dots, k_{v_{a(v)}}$ . In particular,  $j(v) \in \mathcal{J}^1(v)$  contains just  $a(v)$  specific nonzero components of  $j \in \mathcal{J}^1$ . Put

$$B(v) = \left\{ \beta \in B : \beta \neq 0, \quad \text{and} \quad \beta_s \neq 0 \quad \text{if and only if} \quad s \in \mathcal{S}(v) \right\}.$$

Let  $\mathcal{V}(B)$  denote the set of  $v$  for which  $B(v)$  is not empty. Then,

$$\begin{aligned} \langle \Lambda^{-2}, \Upsilon \rangle &\asymp \sum_{v \in \mathcal{V}(B)} \sum_{j(v) \in \mathcal{J}^1(v)} \int_0^{r_{k_{v_1}}} \dots \int_0^{r_{k_{v_{a(v)}}}} \left( 1 + \sum_{\beta \in B(v): \beta \neq 0} x_{k_{v_1}}^{\beta_{k_{v_1}}} \dots x_{k_{v_{a(v)}}}^{\beta_{k_{v_{a(v)}}}} \right)^{-2} \\ &\quad \times dx_{k_{v_1}} \dots dx_{k_{v_{a(v)}}} \\ &\asymp \sum_{v \in \mathcal{V}(B)} \sum_{j(v) \in \mathcal{J}^1(v)} \int_0^{r_{k_{v_1}}} \dots \int_0^{r_{k_{v_{a(v)}}}} \left( 1 + \max_{\beta \in B(v): \beta \neq 0} x_{k_{v_1}}^{\beta_{k_{v_1}}} \dots x_{k_{v_{a(v)}}}^{\beta_{k_{v_{a(v)}}}} \right)^{-2} \\ &\quad \times dx_{k_{v_1}} \dots dx_{k_{v_{a(v)}}}; \end{aligned} \quad (5.4)$$

see Appendix A.5. Condition (5.1) holds if and only if the right-hand side of (5.3), or equivalently the right-hand side of (5.4), evaluated at  $r = (\infty, \dots, \infty)$ , is finite.

In many cases of interest the expansion at (5.4) can be greatly simplified, leading to simple conditions for root- $n$  consistency and related properties. For example, if  $B$  contains just the zero vector and the  $p$  nonzero vectors

$$(\gamma_1, 0, \dots, 0), \quad (0, \gamma_2, 0, \dots, 0), \quad \dots, \quad (0, \dots, 0, \gamma_p), \quad (5.5)$$

where each  $\gamma_i$  is a positive integer, then var converges at rate  $O(n^{-1})$  if and only if

$$\sum_{i=1}^p \gamma_i^{-1} < 2. \quad (5.6)$$

See Appendix A.6. In particular, when each  $\gamma_i$  takes the same value,  $\gamma$  say, a necessary condition for root- $n$  consistency of  $\hat{g}$  for  $g$  is  $\gamma \geq \frac{1}{2}(p+1)$ , if  $\gamma$  is odd, or  $\gamma \geq \frac{1}{2}p+1$ , if  $p$  is even. Moreover, it can be readily shown that a necessary condition for root- $n$  consistency is that a sequence of  $p$  distinct vectors of the form (5.5), where each  $\gamma_i$  is an integer, is contained in  $B$ . See Appendix A.7.

If  $p = 2$  then there are just five different types of set  $B$  that do not consist solely of the zero vector: (a)  $B = \{(0, 0), (k, 0)\}$ , (b)  $B = \{(0, 0), (k, \ell)\}$ , (c)  $B =$

$\{(0, 0), (k, 0), (0, \ell)\}$ , (d)  $B = \{(0, 0), (k, 0), (u, v)\}$  and (e)  $B = \{(0, 0), (k, 0), (0, \ell), (u, v)\}$ , where  $k, \ell, u, v$  are positive integers. (We have excluded  $B = \{(0, 0), (0, k)\}$  and  $B = \{(0, 0), (0, k), (u, v)\}$  since these are symmetric to cases (a) and (d), respectively.) Root- $n$  consistency is impossible in cases (a), (b) and (d), since the set  $B$  does not contain the set of vectors at (5.5). We know from (5.6) that in case (c), root- $n$  consistency is possible if  $k^{-1} + \ell^{-1} < 2$ , which is equivalent to asking that at least one of  $k$  and  $\ell$  be at least as large as 2. Finally, root- $n$  consistency is always possible in case (e); see Appendix A.8.

Asymptotic formulae for “var” in cases (a)–(e) above can be derived in those settings where root- $n$  consistency cannot occur. For example,  $\text{var} \asymp n^{-1} r_1 r_2$  if only level data are observed, i.e. if  $B = \{0\}$ . Moreover, in cases (a) and (d) from the previous paragraph,  $\text{var} \asymp n^{-1} r_2$ ; in case (b),  $\text{var} \asymp n^{-1} (r_1 + r_2)$ ; and in case (c), if  $k = \ell = 1$ ,  $\text{var} \asymp n^{-1} \log \min(r_1, r_2)$ . See Appendix A.9.

To generalise the result in case (c) we note that, if  $p \geq 2$  and  $B (= B_1, \text{ say})$  consists of the zero vector and the  $p$  nonzero vectors at (5.5), with  $\gamma_i = 1$  in each, then  $\text{var} \asymp n^{-1} J_p(r)$ , where

$$J_p(r) = \int_0^{r_1} \cdots \int_0^{r_{p-2}} \log \left\{ \frac{(1 + x_1 + \cdots + x_{p-2} + r_{p-1})(1 + x_1 + \cdots + x_{p-2} + r_p)}{(1 + x_1 + \cdots + x_{p-2})(1 + x_1 + \cdots + x_{p-2} + r_{p-1} + r_p)} \right\} \times dx_1 \cdots dx_{p-2}. \quad (5.7)$$

If  $p \geq 3$  and each  $r_i$  equals a constant multiple of  $r_0$ , say, then  $J_p(r) \asymp r_0^{p-2}$ . See Appendix A.9. (Taking each  $r_i$  to be a multiple of  $r_0$  is reasonable here, and in the example discussed immediately below, since for each  $i$  where we make the assumption, we have the same level of derivative information.)

A related example is that where  $p \geq 2$  and just one of the vectors at (5.5) is omitted from  $B_1$ , in particular where

$$B = \left\{ (0, 0, \dots, 0), (0, 1, 0, \dots, 0), \dots, (0, \dots, 0, 1) \right\}. \quad (5.8)$$

Here, if  $C_1, \dots, C_p$  denote positive constants,

$$n \text{ var} \asymp r_1 J_{p-1}(r_2, \dots, r_p) \asymp \begin{cases} r_1 & \text{if } p = 2 \\ r_1 \log \min(r_2, r_3) & \text{if } p = 3 \\ r_1 r_0^{p-3} & \text{if } p \geq 4 \text{ and } r_i = C_i r_0 \text{ for } i \geq 2. \end{cases} \quad (5.9)$$

In particular, if  $p = 2$  or  $3$  then “var” differs by at most a logarithmic factor from its form in the conventional case where  $p = 1$  and  $B = \{0\}$ . Even for  $p \geq 4$  there is a substantial reduction in variance. This indicates that significant improvements in the nonparametric,  $p$ -variate convergence rate can be achieved by observing data on all but one first partial derivative.

The latter result is of particular relevance to cost-function estimation, where data are available on the cost function and all but one of its first-order partials. More precisely, let costs be given by  $g(x_1, \dots, x_p)$ , where  $x_1$  is the level of output produced and  $(x_2, \dots, x_p)$  are the prices of the  $p - 1$  factor inputs. The ordinary partials of  $g$ , with respect to input prices, equal the quantities of factor inputs, which are often observed along with the costs themselves. In this setting,  $B$  is given by (5.8). As before, let  $r = (r_1, \dots, r_p)$  denote the vector of smoothing parameters. Noting that homogeneity of degree one in prices reduces the number of price arguments by one, we conclude that cost functions with two or three factor inputs can be estimated at rates only slightly slower, by a logarithmic factor, than those that would arise if costs were a function of a single variable.

## 6. NUMERICAL RESULTS

*6.1. Estimation of a univariate function.* We generate data  $(Y_{0i}, Y_{1i}, X_i)$ , for  $i = 1, \dots, n$ , on a function and its first derivative, where  $Y_{0i} = g(X_i) + \epsilon_{0i}$  and  $Y_{1i} = g'(X_i) + \epsilon_{1i}$ ; the  $X_i$  are equally spaced on the unit interval, as in (4.2);  $g(x) = \sum_1^\infty \alpha_j^0 \phi_j$ , where the  $\phi_j$  are defined at (4.1) and  $\alpha_j^0 = j^{-2.01}$ , so that the smoothness condition (4.8e) is satisfied; and  $\epsilon_{0i}, \epsilon_{1i}$  are totally independent, mean-zero normal random variables with standard deviations  $\sigma_0 = 0.15$  and  $\sigma_1 = 1.2$ , respectively. The approximate  $R^2$  values for the two equations are 0.57 and 0.55. We generate 100 datasets for sample sizes ranging from  $n = 25$  to 500. For each  $r \leq n$  we calculate the traditional estimator of  $g$  using solely the level data (i.e., for which  $B = \{0\}$ ), and the estimator using both level and derivative data (i.e., for which  $B = \{0, 1\}$ ), where in the latter we set  $\theta_0 = 1/\sigma_0^2$  and  $\theta_1 = 1/\sigma_1^2$ .

Figure 1 illustrates average mean squared errors (AMSE). The U-shaped bias-variance tradeoff is evident for the estimators using only level data, which is as expected since each additional  $\hat{\alpha}_j$  contributes  $\sigma_0^2 n^{-1}$  to var.

Once derivative data are introduced, AMSE declines initially, then remains

relatively flat essentially because each successive  $\alpha_j^0$  is estimated with rapidly increasing accuracy, so that the incremental contribution to variance is  $O(j^{-2}n^{-1})$ . Moreover, the estimator appears to be well-behaved even if the number of parameters estimated equals the number of observations, i.e., if  $r = n$ . See the discussion of this choice below Theorem 4.2.

Table 1 summarizes the minimum AMSEs for each of the estimators. Consistent with its faster rate of convergence, minimum AMSE for the estimator which uses derivative data declines more rapidly as sample size increases, relative to minimum AMSE for the traditional estimator.

*6.2. Estimation of a bivariate function.* Consider data on a bivariate function and all its first-order partial derivatives; that is,  $(Y_{00i}, Y_{10i}, Y_{01i}, Y_{11i}, X_{1i}, X_{2i})$ , for  $i = 1, \dots, n$ , where  $Y_{00i} = g(X_{1i}, X_{2i}) + \epsilon_{00i}$ ,  $Y_{10i} = \{\partial g(X_{1i}, X_{2i})/\partial x_1\} + \epsilon_{10i}$ ,  $Y_{01i} = \{\partial g(X_{1i}, X_{2i})/\partial x_2\} + \epsilon_{01i}$  and  $Y_{11i} = \{\partial^2 g(X_{1i}, X_{2i})/\partial x_1 \partial x_2\} + \epsilon_{11i}$ . The design points are assumed to be regularly spaced on the unit square as in (4.4);  $g(x_1, x_2) = \sum_{1 \leq j_1 < \infty} \sum_{1 \leq j_2 < \infty} \alpha_{j_1 j_2}^0 \phi_{j_1}(x_1) \phi_{j_2}(x_2)$ , where the  $\phi_j$  are as defined at (4.1) and  $\alpha_{j_1 j_2}^0 = j_1^{-2.01} j_2^{-2.01}$ , and  $\epsilon_{00i}, \epsilon_{10i}, \epsilon_{01i}, \epsilon_{11i}$  are everywhere independent, mean-zero normal random variables with standard deviations  $\sigma_{00} = 0.4$ ,  $\sigma_{10} = 1.0$ ,  $\sigma_{01} = 1.0$  and  $\sigma_{11} = 2.0$ . The approximate  $R^2$  values for the four equations are 0.51, 0.64, 0.64 and 0.60 respectively. We set  $\theta_{00} = \sigma_{00}^{-2}$ ,  $\theta_{10} = \sigma_{10}^{-2}$ ,  $\theta_{01} = \sigma_{01}^{-2}$  and  $\theta_{11} = \sigma_{11}^{-2}$ .

We generate 100 datasets for sample sizes  $n = m^2 = 25, 49, 100, 400$ . Given  $(r_1, r_2)$  satisfying  $1 \leq r_1 \leq m$ ,  $1 \leq r_2 \leq m$ , we calculate our estimator for each of the eight subsets  $B$  of  $\{(0, 0), (1, 0), (0, 1), (1, 1)\}$  which contain  $(0, 0)$ .

Table 2 contains minimum AMSEs for various sample sizes. As expected, lowest AMSE is achieved if data on all first-order partial derivatives are observed, followed by the case where ordinary partials are observed for each variable  $x_1$  and  $x_2$ , (i.e., where  $B = \{(0, 0), (1, 0), (0, 1)\}$ ). However, the inclusion of data on any derivative has a materially beneficial impact on AMSE.

Figures 2a and 2b display AMSE contour plots for  $n = 100$ . In the classical case  $B = \{(0, 0)\}$ ,  $\text{var} \asymp n^{-1} r_1 r_2$ . There is a clear tradeoff between bias and variance in both directions. The value of AMSE is minimized when  $r_1$  and  $r_2$  equal 3.

If in addition one observes data on  $\partial g(X_{1i}, X_{2i})/\partial x_1$ , that is if  $B = \{(0, 0),$

$(1, 0)\}$ , then  $\text{var} \asymp n^{-1}r_2$ . AMSE becomes relatively flat in the  $r_1$  direction for  $r_1 \geq 3$ . However, as  $r_2$  increases there is initially a decline in AMSE, then an increase, indicating a tradeoff between bias and variance in the  $r_2$  direction. The case  $B = \{(0, 0), (0, 1)\}$  is symmetric.

The contour plots for the case  $B = \{(0, 0), (1, 1)\}$ , where  $\text{var} \asymp n^{-1}(r_1 + r_2)$ , resemble those for the case  $B = \{(0, 0)\}$ . There is an observable bias-variance tradeoff in both the  $r_1$  and  $r_2$  directions because the data on  $\partial^2 g(X_{1i}, X_{2i})/\partial x_1 \partial x_2$  contain no information on two infinite subsets of parameters associated with  $x_1$  and  $x_2$ , respectively, in particular on  $\{\alpha_{j1}^0, j \geq 1\}$  and  $\{\alpha_{1j}^0, j \geq 1\}$ .

The contour plots for the case  $B = \{(0, 0), (1, 0), (1, 1)\}$  are similar those for  $B = \{(0, 0), (1, 0)\}$ . In both instances the derivative data contain no information on  $\{\alpha_{1j}^0, j \geq 1\}$ , leading to a bias-variance tradeoff in the  $r_2$  direction. The case  $B = \{(0, 0), (0, 1), (1, 1)\}$  is analogous.

In the case  $B = \{(0, 0), (1, 0), (0, 1)\}$ ,  $\text{var} \asymp n^{-1} \log \min(r_1, r_2)$ . The value of AMSE declines initially, then becomes relatively flat in both the  $r_1$  and  $r_2$  directions. The high efficiency of this combination of observed derivatives, relative to the root- $n$  consistent case where  $B = \{(0, 0), (1, 0), (0, 1), (1, 1)\}$ , derives from the fact that all parameters, save the constant term, can be estimated from derivative data.

*6.3. Cost function estimation.* Consider  $y = h(Q, w) + \epsilon$ , where  $y$  is the observed level of costs of producing the level of output  $Q$  given input prices  $w = (w_1, \dots, w_p)$ . By Shephard's Lemma, the ordinary partials of  $h$  with respect to  $w$  yield factor inputs. Since cost functions are homogeneous of degree one in  $w$ , we may rewrite  $h(Q, w) = w_p g(Q, w_1/w_p, \dots, w_{p-1}/w_p)$ , that is, as a function of nonparametric dimension  $p$  rather than  $p + 1$ . For expositional convenience we set  $w_p = 1$  or, equivalently, we choose the  $p$ th input to be the numeraire good.

We calibrate our simulations using a two-factor Cobb-Douglas production function  $cx_1^{c_1}x_2^{c_2}$ , where  $x_1$  and  $x_2$  are the levels of inputs. The corresponding cost function, and demand for the first input, are given respectively by

$$g(Q, w_1) = \tilde{c} Q^{1/(c_1+c_2)} w_1^{c_1/(c_1+c_2)},$$

$$\partial g(Q, w_1)/\partial w_1 = \frac{c_1 \tilde{c}}{c_1 + c_2} Q^{1/(c_1+c_2)} w_1^{-c_2/(c_1+c_2)},$$

where  $\tilde{c} = \{(c_1/c_2)^{c_2/(c_1+c_2)} + (c_2/c_1)^{c_1/(c_1+c_2)}\} c^{-1/(c_1+c_2)}$ .

Data are generated for  $y = g(Q, w_1) + \epsilon$  and  $y_{01} = \{\partial g(Q, w_1)/\partial w_1\} + \epsilon_{01}$ , where  $c = 0.1$ ,  $c_1 = 0.5$  and  $c_2 = 0.6$ . We further assume that  $\epsilon$  and  $\epsilon_{01}$  are independent normal residuals with zero means and standard deviations 1. The  $R^2$  is approximately 0.41 for the level equation and 0.07 for the derivative equation. Data on  $Q, w_1$  constitute a uniform grid on the unit square  $[1, 2]^2$ . We set  $n = 100, r_1 = 4$  and  $r_2 = 4$ .

Figure 3 illustrates the true cost function  $g(Q, w_1)$ , and its contours or isocost curves; an estimate based solely on the cost data; and, an estimate which incorporates derivative or factor price data.

## A. APPENDICES

A.1. *Proof of Theorem 3.1.* Applying (3.8) and taking expectations, we obtain,

$$\begin{aligned} & E\left\{(\hat{\alpha} - \alpha^0)^\top (\hat{\alpha} - \alpha^0)\right\} \\ &= E\left[\left\{\frac{1}{n} \sum_{\beta \in B} \theta_\beta (\Phi_X^\beta)^\top \varepsilon_\beta\right\}^\top M^{-2} \left\{\frac{1}{n} \sum_{\beta \in B} \theta_\beta (\Phi_X^\beta)^\top \varepsilon_\beta\right\}\right] \\ &\quad + \left\{\frac{1}{n} \sum_{\beta \in B} \theta_\beta (\Phi_X^\beta)^\top \sum_{j \in \mathcal{J}^2} \alpha_j^0 \phi_{jX}^\beta\right\}^\top M^{-2} \left\{\frac{1}{n} \sum_{\beta \in B} \theta_\beta (\Phi_X^\beta)^\top \sum_{j \in \mathcal{J}^2} \alpha_j^0 \phi_{jX}^\beta\right\} \\ &= T_1 + T_2, \end{aligned} \tag{A.1}$$

with  $T_1$  and  $T_2$  denoting the respective terms on the right-hand side. Now,

$$\begin{aligned} T_1 &= \frac{1}{n} \sum_{\beta \in B} \theta_\beta^2 \sigma_\beta^2 \operatorname{tr} \left\{ \frac{1}{n} (\Phi_X^\beta)^\top \Phi_X^\beta M^{-2} \right\} \\ &= \operatorname{tr} \left\{ \left( \frac{1}{n} \sum_{\beta \in B} \theta_\beta^2 \sigma_\beta^2 \Psi_X^\beta \right) M^{-2} \right\} = \langle M^{-2}, U \rangle, \end{aligned} \tag{A.2}$$

and, using the definition of  $u$  given at (3.11),  $T_2 = \langle M^{-2}, uu^\top \rangle$ . Theorem 3.1 follows from this property, (3.10), (A.1) and (A.2).

A.2. *Lemma 1.* Assume  $p = 1$ , the basis functions are as at (4.1) and data are gathered on the uniform design given at (4.2). Define

$$d_{jk}^\beta = \frac{1}{n} \sum_{i=1}^n \phi_j^\beta(X_i) \phi_k^\beta(X_i).$$

Let  $\pi_{jk}^\beta = \{\pi^2(j-1)(k-1)\}^\beta$  where  $0^0 \equiv 1$ . If  $2 \leq j \leq n, k \geq 1$ , then

$$|d_{jk}^\beta| = \begin{cases} \pi_{jk}^\beta & \text{if } k-j = 2sn \text{ or } k+j = 2sn+2; s = 0, 1, 2, \dots \\ 0 & \text{otherwise.} \end{cases}$$

If  $k \geq 1$ , then

$$|d_{1k}^\beta| = \begin{cases} 1 & \text{if } \beta = 0 \text{ and } k = 1 \\ \sqrt{2} & \text{if } \beta = 0 \text{ and } k = 2sn + 1; s = 1, 2, 3, \dots \\ 0 & \text{otherwise.} \end{cases}$$

*Proof of Lemma 1.* The proof is similar to that by Eubank (1999, pp.144, Lemma 3.4). See also Martucci (1994).

*A.3. Lemma 2.* Invoke the assumptions underlying Lemma 1. Let  $\beta$  be a nonnegative integer,  $\kappa = (0, 1, 2, \dots, (r-1))^T$ , and suppose  $c_j = O(j^{-(\beta+\delta)})$  uniformly in  $j = 1, \dots, \infty$  for some  $\delta > 1$ . Then

$$\frac{1}{n} \sum_{j=1}^r (\Phi_X^\beta)^T \phi_{jX}^\beta c_j = O(\kappa^{\beta-\delta}) \quad \text{and} \quad \frac{1}{n} \sum_{j=r+1}^{\infty} (\Phi_X^\beta)^T \phi_{jX}^\beta c_j = O(\kappa^\beta/n^\delta).$$

*Proof of Lemma 2.* To prove the first equality, note that  $c \equiv (c_1, \dots, c_r)^T = O(\kappa^{-(\beta+\delta)})$ . Applying (4.3) we have,

$$\frac{1}{n} \sum_{j=1}^r (\Phi_X^\beta)^T \phi_{jX}^\beta c_j = \frac{1}{n} (\Phi_X^\beta)^T \Phi_X^\beta c = O\{\text{diag}(\kappa^{2\beta}) \kappa^{-(\beta+\delta)}\} = O(\kappa^{\beta-\delta}).$$

To establish the second condition, note that

$$\begin{aligned} \frac{1}{n} \sum_{j=r+1}^{\infty} (\Phi_X^\beta)^T \phi_{jX}^\beta c_j &= \frac{1}{n} \sum_{j=n+1}^{\infty} (\Phi_X^\beta)^T \phi_{jX}^\beta c_j = O\left\{ \frac{1}{n} \sum_{j=n+1}^{\infty} (\Phi_X^\beta)^T \phi_{jX}^\beta \frac{1}{j^{\beta+\delta}} \right\} \\ &= O\left\{ \frac{n^\beta}{n^{\beta+\delta}} + \frac{(2n)^\beta}{(2n)^{\beta+\delta}} + \frac{(3n)^\beta}{(3n)^{\beta+\delta}} + \dots \right\} \kappa^\beta \\ &= O\left\{ \frac{1}{n^\delta} \left( 1 + \frac{1}{2^\delta} + \frac{1}{3^\delta} + \dots \right) \right\} \kappa^\beta = O(\kappa^\beta/n^\delta). \end{aligned}$$

The first equality follows from (4.3), in particular,  $n^{-1}(\phi_{jX}^\beta)^T \phi_{j'X}^\beta = 0$  for  $1 \leq j \neq j' \leq n$ . The third equality follows from Lemma 1, which calculates the values of  $n^{-1}(\phi_{jX}^\beta)^T \phi_{kX}^\beta$  for  $1 \leq j \leq n, k > n$ .

*A.4. Proof of Theorem 4.1.* Take  $\bar{X}$  to be the vector of  $m$  equally spaced design points on the unit interval given at (4.2). Let the  $m$ -vector  $\phi_{j_s \bar{X}}^{\beta_s}$  be the  $\beta_s$  derivative of  $\phi_{j_s}$ , defined at (4.1) and evaluated at  $\bar{X}$ . Let  $\Phi_{\bar{X}}^{\beta_s}$  be the  $m \times r_s$  matrix with columns  $\phi_{j_s \bar{X}}^{\beta_s}$ , for  $j_s = 1, \dots, r_s$ . Define  $\phi_{jX}^\beta = \otimes_{s=1}^p \phi_{j_s \bar{X}}^{\beta_s}$  and  $\Phi_X^\beta = \otimes_{s=1}^p \Phi_{\bar{X}}^{\beta_s}$ .

Using (3.7), (4.5) and (4.6) we conclude that  $M = \Lambda$ . Similarly, using (3.11), (4.5) and (4.6) we conclude that  $U = n^{-1}\Upsilon$ . Thus,  $\langle M^{-2}, U \rangle = n^{-1} \langle \Lambda^{-2}, \Upsilon \rangle = \text{var}$  where the latter is defined at (4.7).

Now write  $u$ , defined at (3.11), as

$$\begin{aligned}
u &= \sum_{\beta \in B} \theta_\beta \left[ m^{-p} \left( \otimes_{s=1}^p \Phi_{\bar{X}}^{\beta_s} \right)^\top \left\{ \sum_{j \in \mathcal{J}^2} \left( \otimes_{s=1}^p \phi_{j_s \bar{X}}^{\beta_s} \right) \alpha_j^0 \right\} \right] \\
&= \sum_{\beta \in B} \theta_\beta \left[ m^{-p} \sum_{j \in \mathcal{J}^2} \left( \otimes_{s=1}^p \Phi_{\bar{X}}^{\beta_s} \right)^\top \left\{ \otimes_{s=1}^p \phi_{j_s \bar{X}}^{\beta_s} O(j_s^{-(\beta_s + \delta)}) \right\} \right] \\
&= \sum_{\beta \in B} \theta_\beta \left[ \sum_{j \in \mathcal{J}^2} \otimes_{s=1}^p \left\{ m^{-1} \left( \Phi_{\bar{X}}^{\beta_s} \right)^\top \phi_{j_s \bar{X}}^{\beta_s} O(j_s^{-(\beta_s + \delta)}) \right\} \right].
\end{aligned} \tag{A.3}$$

The summation over  $\mathcal{J}^2$  can be decomposed into  $2^p - 1$  summations, of the form

$$\begin{aligned}
&\sum_{j_1} \cdots \sum_{j_p} \otimes_{s=1}^p \left\{ m^{-1} \left( \Phi_{\bar{X}}^{\beta_s} \right)^\top \phi_{j_s \bar{X}}^{\beta_s} O(j_s^{-(\beta_s + \delta)}) \right\} \\
&= \otimes_{s=1}^p \left\{ m^{-1} \sum_{j_s} \left( \Phi_{\bar{X}}^{\beta_s} \right)^\top \phi_{j_s \bar{X}}^{\beta_s} O(j_s^{-(\beta_s + \delta)}) \right\},
\end{aligned} \tag{A.4}$$

where, for each  $s = 1, \dots, p$ ,  $j_s$  ranges over  $\{1, \dots, r_s\}$  or  $\{r_{s+1}, r_{s+2}, \dots\}$  and at least one of the indexes  $j_s$  ranges over the latter infinite set.

Each expression in braces at (A.4) corresponds to one of the two types in Lemma 2, with  $n$  replaced by  $m$ , and at least one is of the second type. Thus, the multiple summation in (A.4) is of order no larger than  $O(m^{-\delta} \otimes_{s=1}^p \kappa_s^{\beta_s})$ . Inserting this result into (A.3) we have,  $u = O(m^{-\delta} \sum_{\beta \in B} \otimes_{s=1}^p \kappa_s^{\beta_s})$ .

Finally,  $\Lambda^{1/2} \asymp \sum_{\beta \in B} \text{diag}(\otimes_{s=1}^p \kappa_s^{\beta_s})$ , so that  $\Lambda^{-1/2} u = O(m^{-\delta} \otimes_{s=1}^p \kappa_s^0)$ , where  $\kappa_s^0$  is an  $r_s$ -vector of ones. Thus,  $\langle M^{-2}, uu^\top \rangle = (\Lambda^{-1/2} u)^\top \Lambda^{-1} (\Lambda^{-1/2} u) = O(m^{-2\delta} \text{tr}(\Lambda^{-1})) = O(n^{-2\delta/p} \text{tr}(\Lambda^{-1}))$ . Since, by assumption,  $2\delta/p > 1$ ; and, from (A.6) below,  $\text{var} = O\{n^{-1} \text{tr}(\Lambda^{-1})\}$ ; then we may conclude that  $\langle M^{-2}, uu^\top \rangle = o(\text{var})$ .

*A.5. Properties of  $\text{var} = n^{-1} \langle \Lambda^{-2}, \Upsilon \rangle$ .* For general  $p \geq 1$ ,

$$\begin{aligned}
\langle \Lambda^{-2}, \Upsilon \rangle &= \sum_{j_1=0}^{r_1-1} \cdots \sum_{j_p=0}^{r_p-1} \left( \sum_{\beta \in B} c_\beta \theta_\beta j_1^{2\beta_1} \cdots j_p^{2\beta_p} \right)^{-2} \left( \sum_{\beta \in B} c_\beta \theta_\beta^2 \sigma_\beta^2 j_1^{2\beta_1} \cdots j_p^{2\beta_p} \right) \\
&\asymp \sum_{j_1=0}^{r_1} \cdots \sum_{j_p=0}^{r_p} \left( \sum_{\beta \in B} j_1^{2\beta_1} \cdots j_p^{2\beta_p} \right)^{-1} \\
&\asymp \sum_{j_1=0}^{r_1} \cdots \sum_{j_p=0}^{r_p} \left( \sum_{\beta \in B} j_1^{\beta_1} \cdots j_p^{\beta_p} \right)^{-2},
\end{aligned} \tag{A.5}$$

where the asymptotic relation holds as  $r_1, \dots, r_p \rightarrow \infty$ . Thus,

$$\text{var} = n^{-1} \langle \Lambda^{-2}, \Upsilon \rangle \asymp \frac{1}{n} \sum_{j_1=0}^{r_1} \cdots \sum_{j_p=0}^{r_p} \left( \sum_{\beta \in B} j_1^{2\beta_1} \cdots j_p^{2\beta_p} \right)^{-1} \asymp n^{-1} \text{tr}(\Lambda^{-1}). \tag{A.6}$$

To derive (5.4), note that by (5.3),

$$\begin{aligned}
\langle \Lambda^{-2}, \Upsilon \rangle &\asymp \sum_{j_1=0}^{r_1} \dots \sum_{j_p=0}^{r_p} \left( 1 + \sum_{\beta \in B: \beta \neq 0} j_1^{\beta_1} \dots j_p^{\beta_p} \right)^{-2} \\
&\asymp \sum_{v \in \mathcal{V}(B)} \sum_{j(v) \in \mathcal{J}^1(v)} \left( 1 + \sum_{\beta \in B(v): \beta \neq 0} j_1^{\beta_1} \dots j_p^{\beta_p} \right)^{-2} \\
&= \sum_{v \in \mathcal{V}(B)} \sum_{j(v) \in \mathcal{J}^1(v)} \left( 1 + \sum_{\beta \in B(v): \beta \neq 0} j_{k_{v1}}^{\beta_{k_{v1}}} \dots j_{k_{va(v)}}^{\beta_{k_{va(v)}}} \right)^{-2} \\
&\asymp \sum_{v \in \mathcal{V}(B)} \sum_{j(v) \in \mathcal{J}^1(v)} \int_0^{r_{k_{v1}}} \dots \int_0^{r_{k_{va(v)}}} \left( 1 + \sum_{\beta \in B(v): \beta \neq 0} x_{k_{v1}}^{\beta_{k_{v1}}} \dots x_{k_{va(v)}}^{\beta_{k_{va(v)}}} \right)^{-2} \\
&\quad \times dx_{k_{v1}} \dots dx_{k_{va(v)}}. \tag{A.7}
\end{aligned}$$

*A.6. Proof that (5.6) is necessary and sufficient for  $\text{var} = O(n^{-1})$  when the nonzero vectors in  $B$  are those at (5.5).* Note that in this case, (5.4) evaluated at  $r = (\infty, \dots, \infty)$  is finite if and only if

$$\int_1^\infty \dots \int_1^\infty \left( \max_{1 \leq i \leq p} x_i^{\gamma_i} \right)^{-2} dx_1 \dots dx_p < \infty. \tag{A.8}$$

Changing variable in (A.8) from  $x_i$  to  $y_i = x_i^{2\gamma_i}$  for  $1 \leq i \leq p$ , we see that finiteness of the integral at (A.8) is equivalent to

$$\int_1^\infty \dots \int_1^\infty \left( \prod_{i=1}^p y_i^{(1/2\gamma_i)-1} \right) \left( \max_{1 \leq i \leq p} y_i \right)^{-1} dy_1 \dots dy_p < \infty,$$

and hence to finiteness of

$$I(\gamma_1, \dots, \gamma_p) \equiv \int_{1 \leq y_1 \leq \dots \leq y_p < \infty} \left( \prod_{i=1}^p y_i^{(1/2\gamma_i)-1} \right) y_p^{-1} dy_1 \dots dy_p$$

for each ordering of the positive integers  $\gamma_1, \dots, \gamma_p$ . The value of  $I(\gamma_1, \dots, \gamma_p)$  can be worked out exactly, and shown to equal a finite constant multiple of

$$\int_1^\infty y_1^{(1/2\gamma_1)+\dots+(1/2\gamma_p)-2} dy_1,$$

the finiteness of which is equivalent to (5.6).

*A.7. Proof that root- $n$  consistency implies that  $B$  contains the sequence of  $p$ -vectors  $(\gamma_1, 0, \dots, 0), (0, \gamma_2, 0, \dots, 0), \dots, (0, \dots, 0, \gamma_p)$ , where each  $\gamma_i$  is a positive integer.*

Without loss of generality, suppose that  $B$  does not contain a vector of the form  $(\gamma_1, 0, \dots, 0)$ . Root- $n$  consistency requires that the last multiple summation in (A.5) be finite. Recall that the zero vector is always an element of  $B$  and evaluate that summation at  $j_2 = \dots = j_p = 0$ , to obtain  $\sum_{j_1=0}^{r_1} 1$ , which diverges.

A.8. *Proof that, when (i)  $p = 2$ , (ii)  $B = \{(0, 0), (k, 0), (0, \ell), (u, v)\}$  and (iii)  $k, \ell, u, v \geq 1$ , root- $n$  consistency is always possible if all second derivatives are bounded.* Note that in this case,  $\text{bias}^2 = o(n^{-1})$ , (see section 5.1). Applying (A.5), we have:

$$\begin{aligned} n \text{ var} &\asymp \sum_{j_1=0}^{r_1} \sum_{j_2=0}^{r_2} (1 + j_1^k + j_2^\ell + j_1^u j_2^v)^{-2} \\ &= 1 + \sum_{j_1=1}^{r_1} (1 + j_1^k)^{-2} + \sum_{j_2=1}^{r_2} (1 + j_2^\ell)^{-2} + \sum_{j_1=1}^{r_1} \sum_{j_2=1}^{r_2} (1 + j_1^k + j_2^\ell + j_1^u j_2^v)^{-2} \\ &\leq 1 + \sum_{j_1=1}^{r_1} j_1^{-2k} + \sum_{j_2=1}^{r_2} j_2^{-2\ell} + \sum_{j_1=1}^{r_1} \sum_{j_2=1}^{r_2} j_1^{-2u} j_2^{-2v}, \end{aligned}$$

and each summation in the last line converges as  $r_1$  and  $r_2$  increase.

A.9. *Discussion of other particular cases.* If  $B = \{(0, 0), (k, 0)\}$  then, applying (A.5) we have

$$\begin{aligned} n \text{ var} &\asymp \sum_{j_1=0}^{r_1} \sum_{j_2=0}^{r_2} (1 + j_1^k)^{-2} \\ &= 1 + \sum_{j_1=1}^{r_1} (1 + j_1^k)^{-2} + \sum_{j_2=1}^{r_2} 1 + \sum_{j_1=1}^{r_1} \sum_{j_2=1}^{r_2} (1 + j_1^k)^{-2} \asymp r_2. \end{aligned}$$

The case  $B = \{(0, 0), (k, 0), (u, v)\}$  is similar. If  $B = \{(0, 0), (k, \ell)\}$  then

$$\begin{aligned} n \text{ var} &\asymp \sum_{j_1=0}^{r_1} \sum_{j_2=0}^{r_2} (1 + j_1^k j_2^\ell)^{-2} \\ &= 1 + \sum_{j_1=1}^{r_1} 1 + \sum_{j_2=1}^{r_2} 1 + \sum_{j_1=1}^{r_1} \sum_{j_2=1}^{r_2} (1 + j_1^k j_2^\ell)^{-2} \asymp r_1 + r_2. \end{aligned}$$

If  $B$  contains the zero vector and the nonzero vectors at (5.5), with each  $\gamma_i = 1$  in the latter, then

$$\begin{aligned} n \text{ var} = \langle \Lambda^{-2}, \Upsilon \rangle &\asymp \sum_{j_1=0}^{r_1} \dots \sum_{j_p=0}^{r_p} (1 + j_1 + \dots + j_p)^{-2} \\ &\asymp \sum_{j_1=1}^{r_1} \dots \sum_{j_p=1}^{r_p} (1 + j_1 + \dots + j_p)^{-2} \asymp I_1(r), \end{aligned}$$

where

$$I_1(r) \equiv \int_0^{r_1} \dots \int_0^{r_p} (1 + x_1 + \dots + x_p)^{-2} dx_1 \dots dx_p.$$

Now,  $J_p(r) \asymp \log \min(r_1, r_2)$  when  $p = 2$ , and if  $p \geq 2$ ,

$$I_1(r) = \int_0^{r_1} \dots \int_0^{r_{p-1}} \left\{ (1 + x_1 + \dots + x_{p-1})^{-1} - (1 + x_1 + \dots + x_{p-1} + r_p)^{-1} \right\} \\ \times dx_1 \dots dx_{p-1} = J_p(r),$$

the latter defined at (5.7). If each  $r_i = r_0$ , say, then with  $x = (x_1, \dots, x_p)^\top$ ,

$$I_1(r_0) \asymp \int_0^{r_0} \dots \int_0^{r_0} (1 + \|x\|)^{-2} dx_1 \dots dx_p \asymp \int_{\|x\| \leq r_0} (1 + \|x\|)^{-2} dx \\ \asymp \int_0^{r_0} t^{p-3} dt \asymp \begin{cases} \log r_0 & \text{if } p = 2 \\ r_0^{p-2} & \text{if } p \geq 3. \end{cases}$$

## REFERENCES

- BUCKLEY, M.J., EAGLESON, G.K. AND SILVERMAN, B.W. (1988). The estimation of residual variance in nonparametric regression. *Biometrika* **75**, 189–199.
- DETTE, H., MUNK, A. AND WAGNER, T. (1998). Estimating the variance in nonparametric regression — what is a reasonable choice? *J. Roy. Statist. Soc. Ser. B* **60**, 751–764.
- EUBANK, R. (1999). *Nonparametric Regression and Spline Smoothing*. Marcel Dekker, New York.
- FAN, J. AND YAO, Q. (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika* **85**, 645–660.
- Florens, J.-P., M. Ivaldi and S. Larribeau (1996). Sobolev estimation of approximate regressions. *Econometric Theory* **12**, 753–772.
- GHORAI, J.K. (2003). A central limit theorem for the  $L_2$  error of positive wavelet density estimator. *Ann. Inst. Statist. Math.* **55**, 619–637.
- HALL, P. (1984). Central limit theorem for integrated squared error of nonparametric density estimators. *J. Multivar. Anal.* **14**, 1–16.
- HALL, P., KAY, J. AND TITTERINGTON, D.M. (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika* **77**, 521–528.

- HALL, P. AND YATCHEW, A. (2007). Nonparametric estimation when data on derivatives are available. *Ann. Statist.* **35**, 300–323.
- IOANNIDES, D.A. (1992). Integrated square error of nonparametric estimators of regression function — the fixed design case. *Statist. Probab. Lett.* **15**, 85–94.
- JORGENSON, D. (1986). Econometric methods for modeling producer behavior. In: *Handbook of Econometrics III*, Eds. Z. Griliches and M.D. Intriligator, pp. 1841–1915. Amsterdam: Elsevier.
- KIM, T.Y. AND COX, D.D. (2001). Convergence rates for average square errors for kernel smoothing estimators. *J. Nonparam. Statist.* **13**, 209–228.
- KRONMAL, R. AND TARTER, M. (1968). The estimation of probability densities and cumulatives by Fourier series methods. *J. Amer. Statist. Assoc.* **63**, 925–952.
- MARTUCCI, S.A. (1994). Symmetric convolution and the discrete sine and cosine transformations. *IEEE Trans. Signal Process.* **42**, 1038–1051.
- MCFADDEN, D. (1978). Cost revenue and profit functions. In: *Production Economics: A Dual Approach to Theory and Applications I*, Eds. M. Fuss and D. McFadden, pp. 1-109. Amsterdam: North-Holland.
- MUNK, A., BISSANTZ, N., WAGNER, T. AND FREITAG, G. (2005). On difference-based variance estimation in nonparametric regression when the covariate is high dimensional. *J. Roy. Statist. Soc. Ser. B* **67**, 19–41.
- MURRAY-SMITH, R. AND SBARBARO, D. (2002). Nonlinear adaptive control using non-parametric Gaussian process prior models. In: *15th IFAC World Congress on Automatic Control* (electronic – CD-ROM), Barcelona.
- SEEGER, M. (2003). *Bayesian Gaussian Process Models: PAC-Bayesian Generalisation Error Bounds and Sparse Approximations*. PhD Thesis, University of Edinburgh.
- SEEGER, M. (2004). Gaussian processes for machine learning. *J. Neural Systems* **14**, 1–38.
- SEEGER, M. (2007). Cross-validation optimization for large scale structured classification kernel methods. <http://www.kyb.tuebingen.mpg.de/bs/people/seeger/papers/klr-jmlr.pdf>
- SHEPHARD, R. (1953). *Cost and Production Functions*. Princeton, NJ: Princeton University Press.
- SHEPHARD, R. (1970). *Theory of Cost and Production Functions*. Princeton, NJ: Princeton University Press.
- SOLAK, E., MURRAY-SMITH, R., LEITHEAD, W.E., LEITH, D.J. AND RASMUSSEN, C.E. (2004). Derivative observations in Gaussian Process models of dynamic systems. In: *Advances in Neural Information Processing Systems 16* (electronic – CD-ROM), Vancouver, Canada.

- SPOKOINY, V. (2002). Variance estimation for high-dimensional regression models. *J. Multivar. Anal.* **82**, 111–133.
- TENREIRO, C. (1998). Asymptotic distribution for a discrete version of integrated square error of multivariate density kernel estimators. *J. Statist. Plann. Inf.* **69**, 133–151.
- YATCHEW, A. (1988) Some tests of nonparametric regression models. In: *Dynamic Econometric Modelling, Proceedings of the Third International Symposium on Economic Theory*, Eds. W. Barnett, E. Berndt and H. White, pp. 121–135. Cambridge: Cambridge University Press.
- YATCHEW, A. (2003) *Semiparametric Regression for the Applied Econometrician*. Cambridge: Cambridge University Press.

$n$	25	50	100	200	500
$B = \{0\}$	.0135	.0079	.0045	.0029	.0014
$B = \{0, 1\}$	.0063	.0032	.0016	.0008	.0003

**TABLE 1. Minimum AMSE – Univariate Function**

$n$	25	49	100	400
$B = \{(0, 0)\}$	0.0501	0.0388	0.0266	0.0138
$B = \{(0, 0), (1, 0)\}$	0.0337	0.0208	0.0136	0.0054
$B = \{(0, 0), (0, 1)\}$	0.0339	0.0210	0.0138	0.0053
$B = \{(0, 0), (1, 1)\}$	0.0389	0.0255	0.0164	0.0070
$B = \{(0, 0), (1, 0), (0, 1)\}$	0.0215	0.0120	0.0074	0.0019
$B = \{(0, 0), (1, 0), (1, 1)\}$	0.0284	0.0175	0.0114	0.0040
$B = \{(0, 0), (0, 1), (1, 1)\}$	0.0286	0.0176	0.0112	0.0041
$B = \{(0, 0), (1, 0), (0, 1), (1, 1)\}$	0.0184	0.0097	0.0062	0.0012

**TABLE 2. Minimum AMSE – Bivariate Function**

Figure 1. AMSE for estimators of a univariate function.

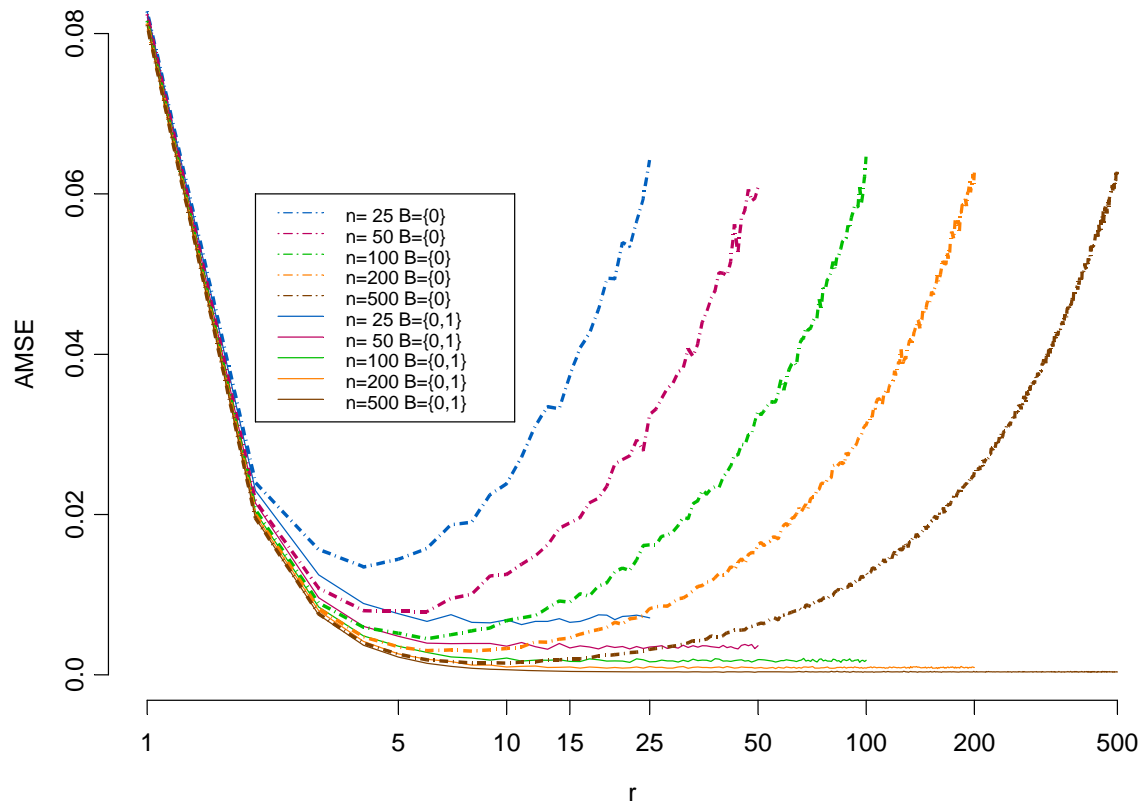
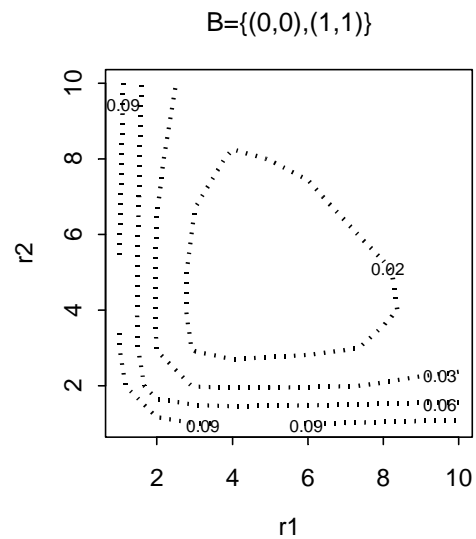
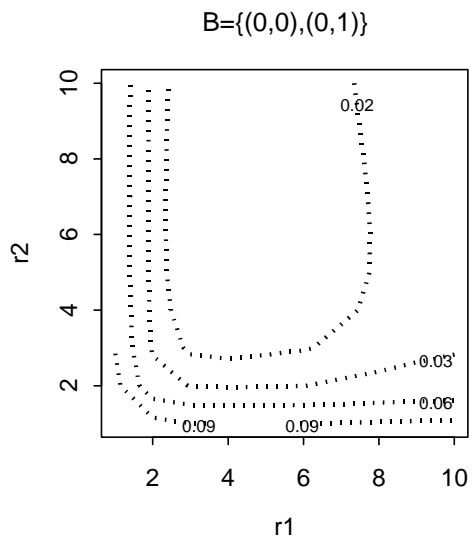
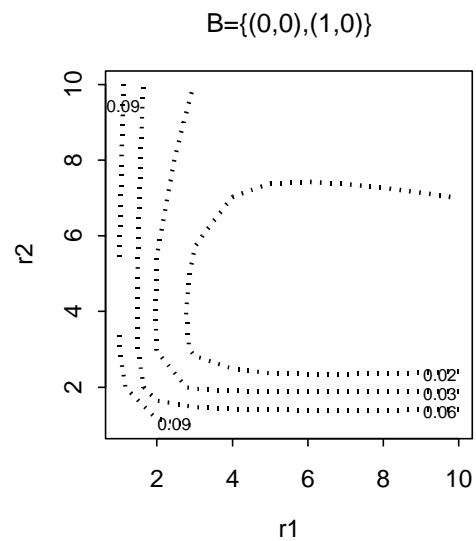
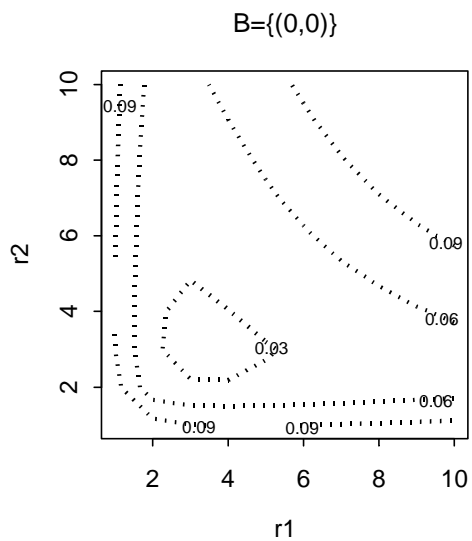
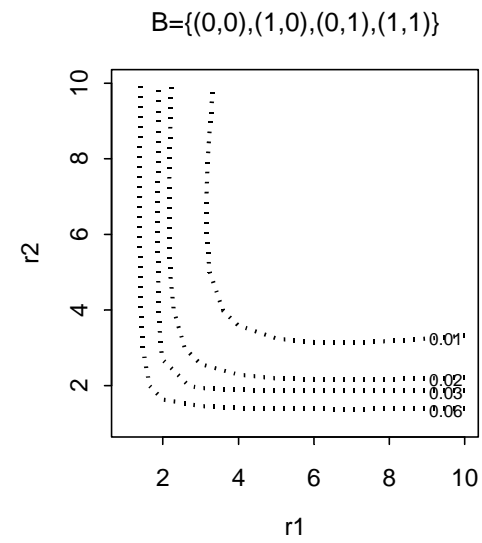
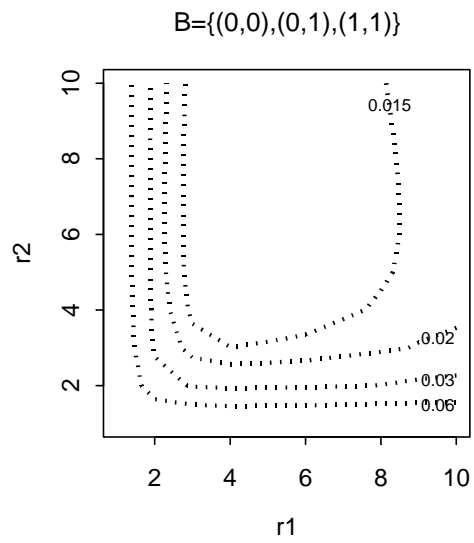
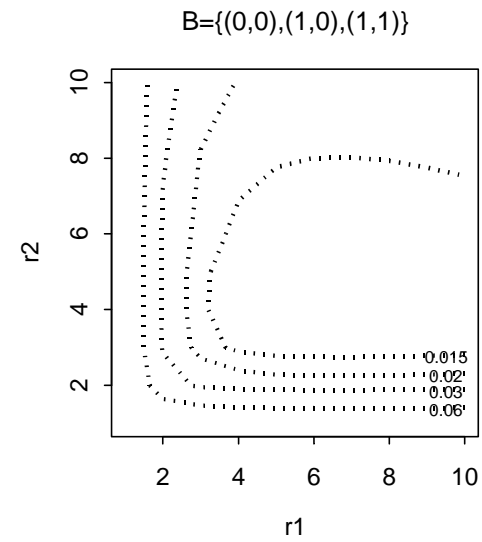
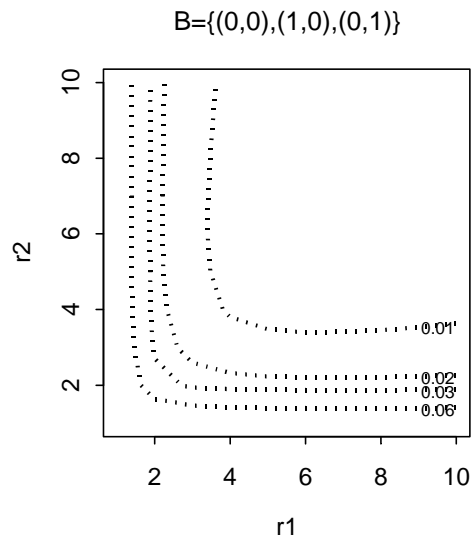


Figure 2a. AMSE contour plots for estimators of a bivariate function



**Figure 2b. AMSE contour plots for estimators of a bivariate function**



**Figure 3. Cost function example.**

