Nonparametric Density Estimation From Covariate Information

Ryan T. ELMORE, Peter HALL, and Vladimir S. TROYNIKOV

An increasing number of statistical problems arise in connection with functional calibration. In each case, inexpensive indirect data in a particular context are combined with direct expensive-to-acquire data from different but related settings to estimate quantities in the former case. Sometimes (e.g., in chemometrics problems where spectroscopic calibration is used) the indirect data are functional. But more commonly, they are scalar or vector-valued, and the functional component is the quantity that we wish to estimate. The problem treated here is of the latter type. We observe data that give us access to the distribution of $\mathbf{U}$ given V, and from these and data on $\mathbf{U}$, we wish to estimate the density of V. The motivating real datasets are of age and covariate information in fish populations. We suggest two methodologies, each of which is based on transforming the problem to one involving inversion of a symmetric linear operator. Our techniques have connections to methods for functional data analysis and for a variety of mixture and deconvolution problems, as well as to calibration techniques.

KEY WORDS: Calibration; Deconvolution; Density estimation; Functional data analysis; Ill-posed problem; Kernel methods; Mixture model; Principal components; Regularization; Smoothing.

1. INTRODUCTION

The practical problem motivating this work is that of estimating the distribution of ages of fish of a specific species in a given population, say Π_1. Determining the age of a fish is difficult and expensive, and consequently, not all populations can have their age distributions estimated directly.

Indeed, determining the age of a fish typically involves killing it, cutting it open, manually removing the otolith bone from its head, using a low-speed, diamond-bladed saw to produce a thin section through the bone, and carefully counting the rings in the section. In contrast, such data as length or weight of fish are easy and inexpensive to gather.

These considerations motivate the following approach to estimating the distribution of fish age. Ages are determined for fish, usually of the same species as those in Π_1, in a reference population Π_0, using an authoritative method such as that described in the previous paragraph. The age distribution in Π_1 is then estimated by calibrating fish in Π_1 against those in Π_0, using covariate information such as fish length or weight. In fact, the density, rather than the distribution, of Π_1 is required, because it gives a better visual impression of such features as modality and skewness.

It should be stressed that even if Π_0 and Π_1 involve the same species of fish, the age distributions are generally different in the two cases, owing to the effects of different physical environments. Therefore, no direct data on the distribution of age in Π_1 are available.

The difficulty and expense of measuring fish age directly means that the size, n, of the sample of fish whose ages are known accurately is invariably less than the size, m, of the sample of fish from the population whose age distribution we wish to determine.

In a sense, the problem we are addressing is one of calibration. It involves comparing direct accurate data on age, which are expensive and difficult to obtain, against inexpensive indirect data which are more readily available, and using the indirect data to conduct inference about the population from which they came. (See, e.g., Osborne 1991 and Everett 1998, p. 50, for discussion of broad classes of problems of this type.) However, because the object of interest here is a density, not a parameter value, the problem has intrinsically different features. For example, in some important respects it resembles deconvolution.

The methods that we suggest for solving this problem show that it also has strong connections to principal components technology in functional data analysis, particularly through functional principal components analysis. In addition, the problem has points of contact with procedures used for inference for mixture distributions, because the sought-after density of Π_1 is a mixture, through a convolution-type operation, of the density of Π_0, which is directly estimable. To delineate the problem, we specify the mixture concisely, as follows.

Hypothetically, data $(\mathbf{U}, V)$, where $\mathbf{U}$ is a vector of covariates and V is a scalar representing age, could be gathered from Π_1. However, we are unable to observe V in this population, but we can obtain direct data on a related pair $(\mathbf{S}, T)$, where $\mathbf{S}$ and T are the versions of $\mathbf{U}$ and V in Π_0. If it may be assumed that the distribution of $\mathbf{U}$ given V is the same as that of $\mathbf{S}$ given T, then we may estimate the conditional density $f_{\mathbf{U}|V}(\mathbf{u}|v)$ using data from Π_0, estimate the marginal density $f = f_{\mathbf{U}}$ of $\mathbf{U}$ using data from Π_1, and thereby estimate the marginal density $g = f_V$ of V, by noting that these three functions are connected through the formula

$$f(\mathbf{u}) = \int g(v) f_{\mathbf{U}|V}(\mathbf{u}|v)\, dv. \tag{1}$$

We show how to invert empirical approximations to (1) so as to express g as a functional of f, where the functional is the inverse of a symmetric linear operator defined in terms of $f_{\mathbf{U}|V}$. We propose two methods of solution: a ridge-based approach, where the ridge parameter has a major influence on the amount of smoothing, and an empirical approximation to the spectrum of the linear operator, leading to an approximation to the operator itself.

Both of these approaches are versions of the singular value decomposition paradigm; the former is an example of Tikhonov, or quadratic, regularization (Tikhonov 1963). General results on the optimality, in certain cases, of these and related methodologies include those of Fan (1991, 1993), Donoho (1995),

Ryan T. Elmore is Assistant Professor, Department of Statistics, Colorado State University (E-mail: *elmore@stat.colostate.edu*). Peter Hall is Professor, Centre for Mathematics and Its Applications, Australian National University, Canberra, ACT 0200, Australia (E-mail: *halpstat@maths.anu.edu.au*). Vladimir S. Troynikov is Senior Research Scientist, Marine and Freshwater Systems, Research and Development Division, Department of Primary Industries, Queenscliff, VIC 3225, Australia (E-mail: *Vladimir.Troynikov@dpi.vic.gov.au*).

© 2006 American Statistical Association
Journal of the American Statistical Association
June 2006, Vol. 101, No. 474, Theory and Methods
DOI 10.1198/016214505000000916

Johnstone (1999), and Cavalier, Golubev, Picard, and Tsybakov (2002).

But although the approach taken in these contributions is particularly insightful and enlightening, it almost invariably pertains to an abstract, continuous, white-noise model. Neither the methodology nor the theoretical results apply directly to actual data, which would generally arise in the form of discrete samples of vectors. For example, the work of Efromovich and Koltchinskii (2001), like ours, treats cases where the operator to be inverted is unknown and estimated but addresses the white-noise model. In this article we bridge this gap, presenting specific methodology for real data vectors in the problem discussed earlier, and describing numerical and theoretical properties.

There is a massive related literature on deconvolution in density estimation, going back at least to work of Gaffey (1959). We mention here only a small number of recent contributions, including work of Barry and Diggle (1995) on smoothing parameter choice; of Neumann (1997) on estimating the error density; of Pensky and Vidakovic (1999), Walter (1999), Fan and Koo (2002), and Pensky (2002) on wavelet methods; of Youndje and Wells (2002) on cross-validation; and of Delaigle (2003) on general kernel methods. Deconvolution techniques play a major role in recent work in economics involving covariate measurement error (e.g., Horowitz and Markatou 1996; Li and Vuong 1998; Gilbert 2002; Li 2002; Linton and Whang 2002; Li and Hsiao 2004). There is also an extremely large literature on errors-in-variables problems (see, e.g., Carroll, Ruppert, and Stefanski 1995). In Troynikov (2004) a weak quasisolution approach was considered with discussion of the particular problem of fish age estimation (see also Troynikov 1998, 1999).

2. METHODOLOGY

2.1 Overview

The information that we have on the constituents of (1) is in the form of a dataset $\mathcal{S} = \{(\mathbf{S}_1, T_1), \dots, (\mathbf{S}_n, T_n)\}$ drawn from the distribution of $(\mathbf{S}, T)$, say, and a dataset $\mathcal{U} = \{\mathbf{U}_1, \dots, \mathbf{U}_m\}$ from the distribution of $\mathbf{U}$. In addition, we know that the distribution of $\mathbf{S}$ given T is the same as the distribution of $\mathbf{U}$ given V. It is assumed that $\mathcal{S}$ and $\mathcal{U}$ are independent sets of independent data.

In the practical problem that motivates this work, n is generally significantly less than m. This reflects the difficulty and expense of accurately determining the age of fish (i.e., the values of T_i in $\mathcal{S}$) compared with simply gathering covariate data (e.g., the values of $\mathbf{U}_i$ in $\mathcal{U}$). A detailed discussion of this point was given in Section 1.

To appreciate the sorts of methods that might be used to make inference from $\mathcal{S}$ and $\mathcal{U}$, it is helpful to study (1). Let $\mathcal{I}$ denote the support of the distribution of V, write $L_2(\mathcal{I})$ for the class of square-integrable functions from $\mathcal{I}$ to the real line, and define

$$A(v, w) = \int f_{\mathbf{U}|V}(\mathbf{u}|v) f_{\mathbf{U}|V}(\mathbf{u}|w)\, d\mathbf{u}. \tag{2}$$

Let us use the notation A also to denote the symmetric linear operator of which the kernel is the function A: $(Ag)(w) = \int A(v, w) g(v)\, dv$. Then, under regularity conditions on $f_{\mathbf{U}|V}(\mathbf{u}|v)$, (1) has a unique solution in g if and only if the operator A is invertible. In this case, defining

$$a(w) = \int f_{\mathbf{U}}(\mathbf{u}) f_{\mathbf{U}|V}(\mathbf{u}|w)\, d\mathbf{u}, \tag{3}$$

we see that (1) can be written equivalently as $a = Ag$, and so g can be written as $g = A^{-1}a$, assuming that the latter is well defined. That is, g is the solution of the equation

$$a(w) = \int g(v) A(v, w)\, dv. \tag{4}$$

This approach is, of course, familiar to statisticians. The basic equation (1) can be interpreted as $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$, where the role of the parameter vector $\boldsymbol{\beta}$ is played by the unknown function g, and $f_{\mathbf{U}|V}(\mathbf{u}|v)\, dv$ represents the $(\mathbf{u}, v)$th component of the "matrix" $\mathbf{X}$. The linear mapping determined by A corresponds to the matrix operator $\mathbf{X}^{\mathrm{T}}\mathbf{X}$, and the formula $g = A^{-1}a$ is the analog of the normal equation $\boldsymbol{\beta} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$.

If the operator A is invertible, then the function A is the kernel of a positive-definite operator on functions in $L_2(\mathcal{I})$ and may be represented in a spectral expansion,

$$A(v, w) = \sum_{j=1}^{\infty} \theta_j \phi_j(v) \phi_j(w), \tag{5}$$

where $\theta_1 \ge \theta_2 \ge \cdots > 0$ (see, e.g., Indritz 1963, chap. 4). Here (θ_j, ϕ_j) is the jth (eigenvalue, eigenvector) pair for the operator A. Necessarily, the functions ϕ_j form an orthonormal basis for $L_2(\mathcal{I})$.

In the foregoing notation, we may equivalently write (4) as $\alpha_j = \gamma_j \theta_j$ for $j \ge 1$, where $\alpha_j = \int a(w)\phi_j(w)\, dw$ and $\gamma_j = \int g(v)\phi_j(v)\, dv$. That is, $\gamma_j = \theta_j^{-1}\alpha_j$, and thus

$$g = \sum_{j=1}^{\infty} \theta_j^{-1} \alpha_j \phi_j. \tag{6}$$

We have data directly on the distribution of $\mathbf{U}$ given V and so we can estimate $f_{\mathbf{U}|V}$ and hence also A. Therefore, we can estimate θ_j and ϕ_j. We also have data directly on the distribution of $\mathbf{U}$, and so we can estimate the function a and thus also estimate α_j. It follows that we can estimate $\gamma_j = \theta_j^{-1}\alpha_j$, as well as g, by applying the expansion (6).

In more detail, using the dataset $\mathcal{S}$, we may easily construct an estimator $\hat{f}_{\mathbf{U}|V}$ of $f_{\mathbf{U}|V}$, and hence [e.g., by direct substitution into (2)], we may construct a symmetric estimator $\widehat{A}$ of A. Then we may develop an empirical expansion analogous to (5),

$$\widehat{A}(v, w) = \sum_{j=1}^{\infty} \hat{\theta}_j \hat{\phi}_j(v) \hat{\phi}_j(w), \tag{7}$$

where $\hat{\theta}_1 \ge \hat{\theta}_2 \ge \cdots \ge 0$. Here $(\hat{\theta}_j, \hat{\phi}_j)$ is the jth (eigenvalue, eigenvector) pair for the operator $\widehat{A}$, and the functions $\hat{\phi}_j$ form an orthonormal basis for $L_2(\mathcal{I})$. It is common to compute $\hat{\theta}_j$ and $\hat{\phi}_j$ by discretizing $\widehat{A}(\mathbf{u}, v)$ on a regular grid of values $(\mathbf{u}, v)$, and interpreting (7) as a conventional spectral decomposition of a matrix.

Using the datasets $\mathcal{S}$ and $\mathcal{U}$, we may construct estimators $\hat{f}_{\mathbf{S}|T}$ and $\hat{f}_{\mathbf{U}}$ of $f_{\mathbf{S}|T} = f_{\mathbf{U}|V}$ and $f_{\mathbf{U}}$, respectively. Then, by direct substitution of $(\hat{f}_{\mathbf{U}}, \hat{f}_{\mathbf{S}|T})$ for $(f_{\mathbf{U}}, f_{\mathbf{U}|V})$ in (3), we may construct an estimator $\hat{a}$ of a. Therefore, we may construct an estimator $\hat{\alpha}_j = \int \hat{a}(w)\hat{\phi}_j(w)\, dw$ of α_j. Our final estimator of g might then be inspired by (6),

$$\hat{g} = \sum_{j=1}^{\nu} \hat{\theta}_j^{-1} \hat{\alpha}_j \hat{\phi}_j, \tag{8}$$

where $\nu \geq 1$ denotes a smoothing parameter.

An alternative approach is to undertake a more direct empirical inversion of (4). This technique might involve using a ridge parameter, say λ, in place of the integer ν in (8) as the main smoothing parameter. (Subsidiary smoothing parameters would be used to construct $\widehat{A}$ and $\hat{a}$.) Specifically, let $\widehat{A}^-$ denote the inverse of the linear operator with kernel $\widehat{A} + \lambda I$, where $\lambda > 0$ is chosen to decrease to 0 as sample size increases and I is the identity operator. The resulting estimator, alternative to that at (8), would be

$$\tilde{g} = \widehat{A}^- \hat{a} = \sum_{j=1}^{\infty} (\hat{\theta}_j + \lambda)^{-1} \hat{\alpha}_j \hat{\phi}_j. \tag{9}$$

The infinite series on the right side of (9) converges in a mean squared sense.

In these discussions and those that follow, we have tacitly assumed that the sign of $\hat{\phi}_j$ is chosen so that $\int \phi_j \hat{\phi}_j \geq 0$. This enables us to consider $\hat{\phi}_j$ an approximation to ϕ_j rather than to $-\phi_j$. The convention does not affect the performance of estimators, of course, but it does make discussion a little simpler.

2.2 Estimators of $\hat{a}$ and $\widehat{A}$

We may equivalently write A and a as

$$A(v, w) = \frac{B(v, w)}{f_T(v) f_T(w)} \qquad \text{and} \qquad a(w) = \frac{b(w)}{f_T(w)}, \tag{10}$$

where

$$\begin{aligned} B(v, w) &= \int f_{\mathbf{S}T}(\mathbf{u}, v) f_{\mathbf{S}T}(\mathbf{u}, w)\, d\mathbf{u} \qquad \text{and} \\ b(w) &= \int f_{\mathbf{U}}(\mathbf{u}) f_{\mathbf{S}T}(\mathbf{u}, w)\, d\mathbf{u}, \end{aligned} \tag{11}$$

and $f_{\mathbf{U}}, f_T$, and $f_{\mathbf{S}T}$ denote the probability densities of $\mathbf{U}$, T, and $(\mathbf{S}, T)$. Let $\mathbf{S}$ and T (or, equivalently, $\mathbf{U}$ and V) be p-variate and univariate.

We use kernel methods to construct estimators of B, b, and f_T. Our definitions are conventional, and in particular do not take into account any edge effects that might exist. These may be accommodated by, for example, using boundary kernels toward the ends of $\mathcal{I}$ (see, e.g., Simonoff 1996, pp. 53–54), or by binning the data and fitting local polynomial smoothers to bin heights (e.g., Fan and Gijbels 1996).

Let K_1 and K_2 be p-variate univariate kernels (we take them to be compactly supported, spherically symmetric probability densities in the stated number of dimensions), and let h_1 and h_2 denote bandwidths suitable for p dimensions and one dimension. Estimators of $f_T, f_{\mathbf{S}T}$, B, and b are

$$\begin{aligned} \hat{f}_T(v) &= \frac{1}{nh_2} \sum_{i=1}^{n} K_2\left(\frac{v - T_i}{h_2}\right), \\ \hat{f}_{\mathbf{S}T}(\mathbf{u}, v) &= \frac{1}{nh_1^p h_2} \sum_{i=1}^{n} K_1\left(\frac{\mathbf{u} - \mathbf{S}_i}{h_1}\right) K_2\left(\frac{v - T_i}{h_2}\right), \\ \widehat{B}(v, w) &= \frac{1}{n(n-1)h_1^p h_2^2} \sum\sum_{i \neq j} K_1\left(\frac{\mathbf{S}_i - \mathbf{S}_j}{h_1}\right) \\ &\quad \times K_2\left(\frac{v - T_i}{h_2}\right) K_2\left(\frac{w - T_j}{h_2}\right), \end{aligned} \tag{12}$$

$$\hat{b}(w) = \frac{1}{m} \sum_{i=1}^{m} \hat{f}_{\mathbf{S}T}(\mathbf{U}_i, w), \tag{13}$$

where the bandwidths h_1 and h_2 are of course the same within each formula but are not necessarily identical between different formulas. Our estimators $\widehat{A}$ and $\hat{a}$ of A and a are obtained by substituting $\hat{f}_T$, $\widehat{B}$, and $\hat{b}$ for f_T, B, and b at (10),

$$\widehat{A}(v, w) = \frac{\widehat{B}(v, w)}{\hat{f}_T(v) \hat{f}_T(w)} \qquad \text{and} \qquad \hat{a}(w) = \frac{\hat{b}(w)}{\hat{f}_T(w)}. \tag{14}$$

From these functions, we may directly compute the sequences $\hat{\theta}_j$, $\hat{\phi}_j$, and $\hat{\alpha}_j$ and thence the estimators $\hat{g}$ and $\tilde{g}$ at (8) and (9). In principle, statistical or numerical difficulties could arise if the dimension, p, of the explanatory vectors $\mathbf{S}_i$ were large. This does not seem to occur in practice, however; in the datasets of which we are aware, p is usually equal to 1 and occasionally equal to 2.

Next we give motivation for $\hat{f}_T, \hat{f}_{\mathbf{S}T}$, $\widehat{B}$, and $\hat{b}$. The estimators $\hat{f}_T$ and $\hat{f}_{\mathbf{S}T}$ are, of course, conventional univariate and multivariate kernel density estimators. If in the definition of $\widehat{B}$ at (12), we were to replace K_1 by $K_1 * K_1$ (the convolution of K_1 with itself), then $\widehat{B}(v, w)$ would be identical to

$$\widetilde{B}(v, w) = \int \hat{f}_{\mathbf{S}T}(\mathbf{u}, v) \hat{f}_{\mathbf{S}T}(\mathbf{u}, w)\, d\mathbf{u} \tag{15}$$

[cf. the first part of (11)], except that the diagonal terms that arise from (15) would have been omitted. We use K_1 rather than $K_1 * K_1$ in the construction of $\widehat{B}$ only for simplicity. The estimator, $\hat{b}(w)$, of $b(w) = \int f_{\mathbf{S}T}(\mathbf{u}, w)\, dF_{\mathbf{U}}(\mathbf{u})$ is obtained in this formula by replacing $f_{\mathbf{S}T}$ by $\hat{f}_{\mathbf{S}T}$ and $F_{\mathbf{U}}$ by the empirical distribution function of $\mathbf{U}$.

2.3 Estimators of Principal Components

An alternative way of viewing the problem of estimating g is to consider that, rather than seeking an approximation to g, we wish to approximate the lower-dimensional principal components of g. In particular, we seek relatively concise estimators of $\hat{\theta}_j$, $\hat{\phi}_j$, and $\hat{\alpha}_j$. From these, we would form the estimator $\hat{g}$ at (8), using a relatively small value of ν rather than a value that would render $\hat{g}$ a particularly accurate approximation to g. We show in Section 5.2 that root-n–consistent estimation of θ_j and α_j is possible, even under nonparametric assumptions, and that $\hat{\phi}_j$ can be estimated at the standard rate commensurate with appropriate smoothness conditions imposed on the densities $f_{\mathbf{S}T}$ and $f_{\mathbf{U}V}$. In particular, the mean squared rate is $O(n^{-2r/(2r+1)})$ if these densities have r derivatives. Of course, these rates of convergence are not attainable uniformly in j.

3. EMPIRICAL SMOOTHING PARAMETER CHOICE

Because of the complex nature of this problem, it does not seem feasible to choose smoothing parameters in a theoretically optimal way, for example, by asymptotically minimizing a criterion such as mean integrated squared error (MISE). One obstacle is the sheer number of smoothing parameters; there are at least three (h_1, h_2, and λ), more if $p \geq 2$, and we use different bandwidths for different components. Second, no value of the variable V, the density of which we wish to estimate, is ever actually observed, and so we do not have access to a conventional cross-validation algorithm, where one would omit successive values V_i and use the others to estimate the density at V_i.

Third, although it is theoretically possible (under quite stringent conditions) to write a first-order formula for the mean squared error of $\hat{f}_V$ and to estimate the leading terms in that expression, this approach involves so many new smoothing parameters (all of which need to be estimated themselves) as to make it impractical. Therefore, developing a workable plug-in rule for smoothing parameter choice seems beyond reach.

For these reasons, the approach to smoothing parameter choice that we suggest is highly nonstandard. It is based on feasibility rather than optimality, and is as follows:

1. Apply the method in Section 2 to indirect estimation of f_T rather than of f_V.
2. Choose none, one, or two of the bandwidths h_1 and h_2 directly from data, as though our aim was to estimate $f_{\mathbf{S}T}$ and f_T.
3. Use a nonstandard criterion, which we introduce later, to select any remaining smoothing parameter, in particular the ν at (8) or the λ in (9), which was not chosen in step 2.

Step 1 may seem perverse, but, of course, the methodology described in Section 2 would remain valid if it were true that $(\mathbf{U}, V)$ and $(\mathbf{S}, V)$ had identical distributions. In step 1 we partially reproduce that setting, exploiting the fact that in practice m is larger than n, so that by constraining ourselves to working with a sample of size n, we are taking a conservative, or oversmoothed, approach to bandwidth choice.

For the sake of definiteness, we describe our method when $\tilde{g}$, defined at (9), is used to estimate g. The case of $\hat{g}$ is similar. To implement step 1, we replace the equation $a = Ag$, defining $g = f_V$, by $a_* = Ag_*$, defining $g_* = f_T$, where A is as at (10) and

$$a_*(w) = \int f_{\mathbf{S}}(\mathbf{u}) f_{\mathbf{S}|T}(\mathbf{u}|w)\, d\mathbf{u} = \frac{b_*(w)}{f_T(w)}$$

and

$$b_*(w) = \int f_{\mathbf{S}}(\mathbf{u}) f_{\mathbf{S}T}(\mathbf{u}, w)\, d\mathbf{u}.$$

We estimate A exactly as before, using the first formula in (14). However, because $\mathbf{S}_1, \dots, \mathbf{S}_n$ are correlated with $\hat{f}_{\mathbf{S}T}$, to estimate b, we use a leave-one-out approach, which gives the following version of $\hat{b}$ at (13):

$$\hat{b}_*(w) = \frac{1}{n} \sum_{i=1}^{n} \hat{f}_{\mathbf{S}T;-i}(\mathbf{S}_i, w),$$

where

$$\hat{f}_{\mathbf{S}T;-i}(\mathbf{u}, v) = \frac{1}{(n-1) h_1^p h_2} \sum_{j : j \neq i} K_1\left(\frac{\mathbf{u} - \mathbf{S}_j}{h_1}\right) K_2\left(\frac{v - T_j}{h_2}\right).$$

Our indirect estimator of g_*, mentioned in step 1, is the following version of $\tilde{g}$: $\tilde{g}_* = \hat{A}^{-1} \hat{a}_*$, where, analogously to (14), $\hat{a}_* = \hat{b}_* / \hat{f}_T$.

A reasonably simple, although unconventional, approach is to choose smoothing parameters for $\tilde{g}_*$ to minimize the criterion

$$\widehat{\mathrm{MISE}} = \int_{\mathcal{I}} (\hat{f}_T - \tilde{g}_*)^2. \tag{16}$$

The estimator $\hat{f}_T$ here would be constructed using conventional bandwidth choice methods for univariate density estimators. This approach exploits the fact that $\hat{f}_T$ converges at an inherently faster rate than $\hat{g}$, because the latter is impeded by the ill-posed nature of the inversion problem on which it is based. Therefore, errors in $\tilde{g}_*$ resulting from the choice of λ in the vicinity of the optimum, dominate errors in $\hat{f}_T$ as an approximation to f_T. This feature can be used to show that, provided that all smoothing parameters of $\tilde{g}_*$ are chosen in a reasonable range, the version of $\tilde{g}_*$ that is smoothed using parameters resulting from this argument is consistent for f_T.

4. NUMERICAL RESULTS

4.1 Simulation Study

We examined the numerical properties of the proposed estimator under various conditions in a Monte Carlo study. The results presented here are for situations where we generated T-data from a variety of distributions: symmetric (standard normal), right-skewed (gamma), left-skewed (square root of a chi-squared distribution), and bimodal (mixture of normals). All of the results are based on a "hybrid" estimator that uses both a frequency cutoff and a ridge parameter,

$$\check{g} = \sum_{j=1}^{\nu} (\hat{\theta}_j + \lambda)^{-1} \hat{\alpha}_j \hat{\phi}_j. \tag{17}$$

For each situation, we generated bivariate datasets $\mathcal{S} = \{(\mathbf{S}_1, T_1), \dots, (\mathbf{S}_n, T_n)\}$ by first constructing T_i distributed as F_T and $\mathbf{S}'_i$ distributed as $F_{\mathbf{S}'}$, independently for $i = 1, \dots, n$. Then we set $\mathbf{S}_i = \rho T_i + \sqrt{1 - \rho^2}\mathbf{S}'_i$ for some value of ρ, so that $\mathbf{S}_i$ and T_i were correlated. We generated $\mathcal{U} = \{\mathbf{U}_1, \dots, \mathbf{U}_m\}$ in the same way as $\mathbf{S}_i$. This ensured that the distribution of $\mathbf{S}$ given T was the same as the distribution of $\mathbf{U}$ given V which was our main assumption. We evaluated the performance of the density estimator for each dataset using the standard MISE criterion.

4.2 Results

We first discuss the results of simulations involving normal data. In this situation, we took both F_T and $F_{\mathbf{S}'}$ to be the standard normal cumulative distribution function. Hence $(\mathbf{S}_i, T_i)$ has a bivariate normal distribution with zero means, unit variances, and correlation coefficient ρ. The results presented here are for the cases where $\rho = .5$ and .8. We considered samples of size $m, n = 300$, 500, and 1,000, with $m \geq n$. (In practice, m is always taken to be at least as large as n.) Thus six sample-size pairs were examined. We estimated the function g on an equally spaced mesh of $k + 1 = 31$ points defined on the interval $\mathcal{I}_v = [-3, 3]$.

The density estimator performed well in estimating this symmetric unimodal density, aside from a few instances where spurious modes were produced in the tails. The simulated MISEs (on a log scale) for the various combinations of sample sizes are presented in Figures 1(a) and 1(b) for the cases where $\rho = .5$ and .8. As expected, reducing ρ to .1 produces a very flat density estimator, reflecting the fact that if the correlation between the explanatory variable, $\mathbf{S}$ or $\mathbf{U}$, and the variable of major interest, T or V, is low, then data on the former contain little information about the distribution of the latter.

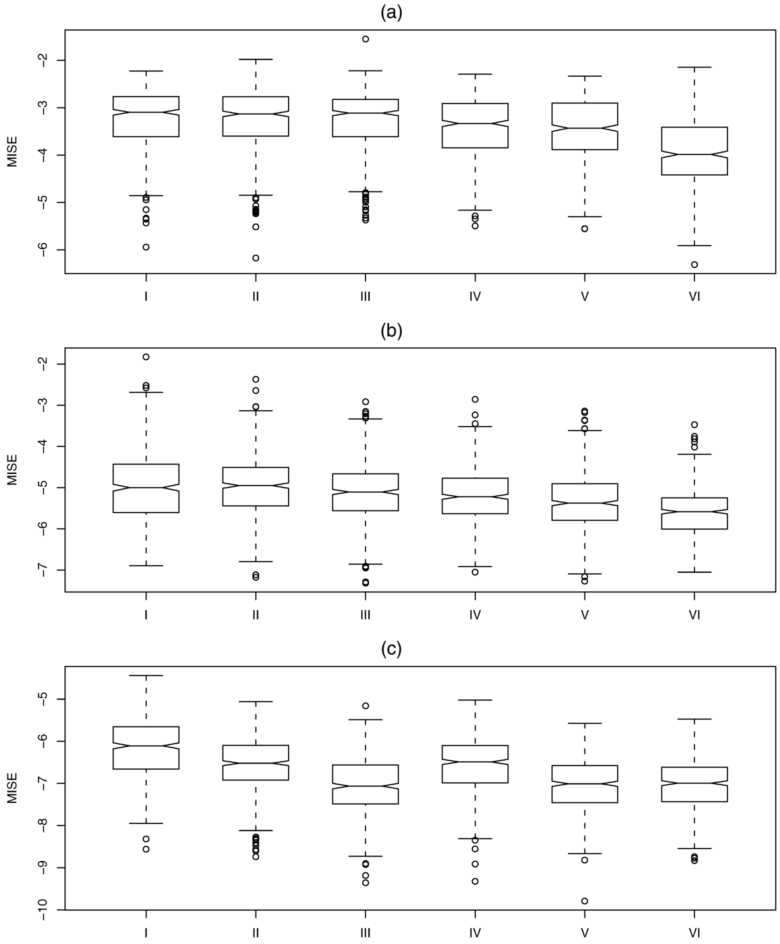

Figure 1. Boxplots of the Estimated MISE Under the Normal-Based Simulations for $\rho = .5$ (a), $\rho = .8$ (b), and Observing V Directly (c). From left to right, the numerals I–VI correspond to $n = 300$, $m = 300$, 500, and 1,000, $n = 500$, $m = 500$ and 1,000, and $n = m = 1{,}000$.

Figure 1(c) presents the simulated MISEs (on a log scale) in the case of observing the V-data directly and estimating the density using a standard density estimator. The boxplots there are based on 501 simulated datasets for each situation.

To assess the performance of our estimator in the case of nonnormal data, we generated T-data from the following three distributions: $f_1(t) = \frac{1}{2}t^2e^{-t}$ for $t > 0$, $f_2(t) = t^{1/4}\exp(-t^2/2)/2^{1/4}\Gamma(5/4)$ for $t > 0$, and $f_3(t) = \frac{1}{2}\phi(t;0) + \frac{1}{2}\phi(t;4)$ for $-\infty < t < \infty$, where $\phi(\cdot;\mu)$ represents the normal density with mean μ and unit variance. Thus f_1 is the gamma distribution with location parameter equal to three (right-skewed), f_2 is the square root of a chi-squared distribution with 2.5 degrees of freedom (left-skewed), and f_3 is a mixture of normal distributions whose means are separated by four standard deviations (bimodal). We set $\rho = .7, .85$, and .55 when using f_1, f_2, and f_3, so that the true correlation between $\mathbf{S}$ and T was approximately .85, .77, and .82. Regardless of the situation, we generated the covariate $\mathbf{S}_i'$ from the standard normal distribution and generated the $\mathbf{U}$-data in an analogous manner.

Figures 2(a.1), 2(b.1), and 2(c.1) display the simulated MISEs for 501 datasets under the different sample size combinations. Situations I, II, III, and IV in this figure refer to sample sizes $n = m = 1{,}000$; $n = 1{,}000$, $m = 1{,}500$; $n = 1{,}000$, $m = 2{,}000$; and $n = m = 2{,}000$. These sample sizes are small compared with those typically encountered in practice, because it is recognized that this deconvolution problem is especially challenging and cannot be adequately solved for relatively small datasets; see Section 4.3 for the sizes of typical real samples.

Figures 1 and 2 indicate that the sample size n of the complete sample plays a more important role than the sample size m of the incomplete sample in the estimation of g. Specifically, the estimated MISE decreases noticeably with increases in n. However, the MISE decreases only slightly as m increases for fixed n. The boxplots in Figures 2(a.2), 2(b.2), and 2(c.2) show the estimated MISE in directly estimating the density of V, if the data were available for samples of size $m = 1{,}000$, 1,500, and 2,000 (I, II, and III) under the different distributions.

In most cases, the shape of the unknown density is accurately represented by that of the estimator. This is illustrated in Figure 3, which presents the estimated density (dashed line) of the fit with median MISE, along with the true density (solid line),

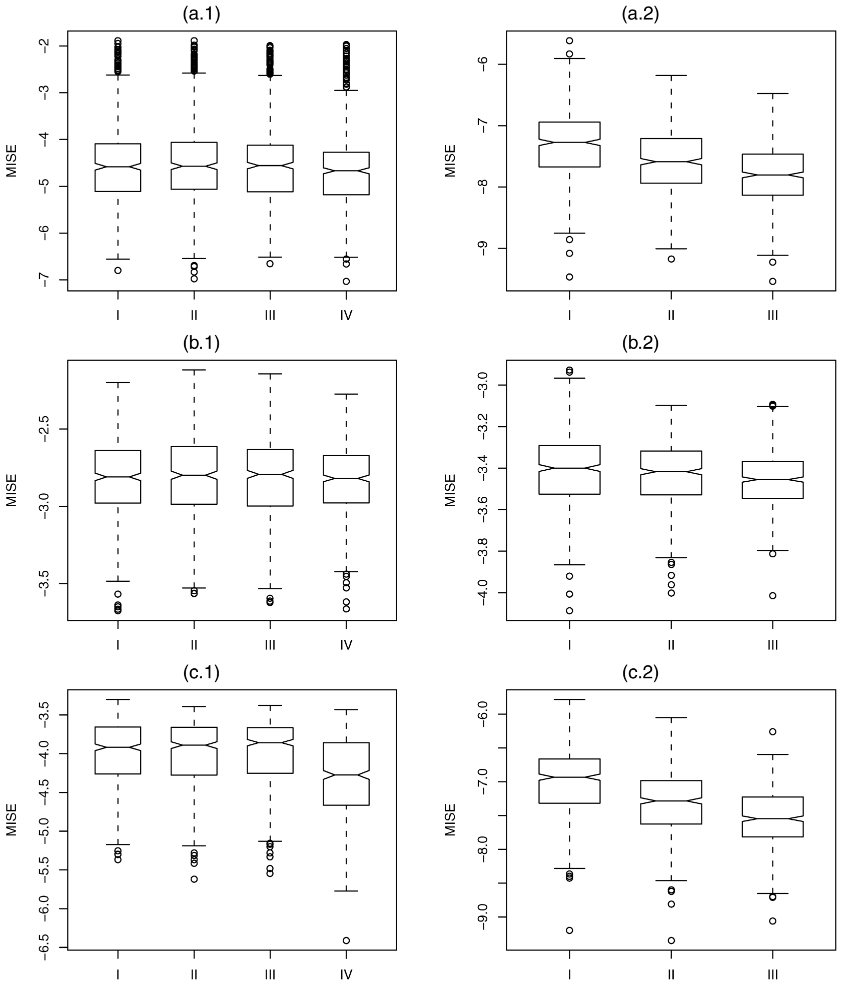

Figure 2. Boxplots of the Estimated MISE Under the Nonnormal Simulations for Gamma Data (a), Square Root of a Chi-Squared (b), and a Mixture of Normals (c). The samples sizes for I, II, III, and IV are $n = m = 1{,}000$; $n = 1{,}000$, $m = 1{,}500$; $n = 1{,}000$, $m = 2{,}000$; and $n = m = 2{,}000$. The boxplots in the right column are the estimated MISEs if we observe V-data directly in each situation: $m = 1{,}000$, 1,500, and 2,000.

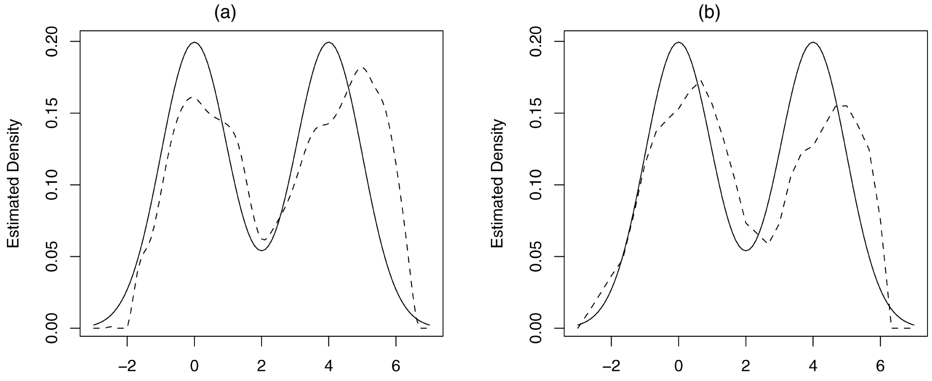

Figure 3. The Plot of $\breve{g}$ Having Median MISE (- - - -) and the True Underlying Density (——) for the Mixture of Normal Data for Samples of Size (a) $n = m = 1{,}000$ and (b) $n = m = 2{,}000$.

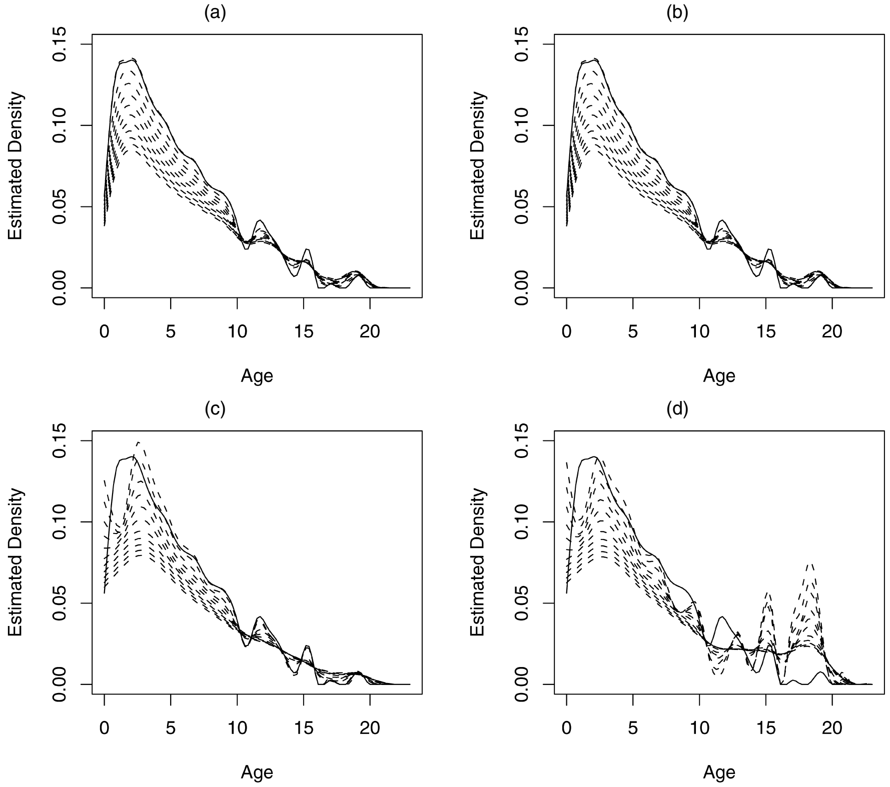

Figure 4. The Estimated Density for the Sand Flathead Fish Data for Smoothing Parameters λ = .0035 and ν = 3. Panels (a)–(b) show the estimated density curves (——) for ν = 2, 3, 4, and 5 and the estimated density curves from varying λ between 0 and .3 (- - - -).

of V for the case where $T \sim f_3$ with $n = m = 1{,}000$ in (a) and $n = m = 2{,}000$ in (b).

Our method of estimation is not guaranteed to produce nonnegative functions $\tilde{g}$. To overcome this problem, $\tilde{g}\mathbb{1}_{[\tilde{g}\geq 0]}$ can be used to estimate g. This does not affect the convergence results given in Section 5.1, because $\int(\tilde{g}\mathbb{1}_{[\tilde{g}\geq 0]} - g)^2 \leq \int(\tilde{g} - g)^2$. Figures 3 and 4 were obtained using this approach.

4.3 Example

We applied the methods to a sample of sand flathead fish (*Platycephalus arenarius*). The complete sample ($n = 3{,}081$) has measurements on both age and length of the fish, whereas the incomplete sample ($m = 9{,}046$) has only covariate information on length of the fish. Thus the goal is to estimate the age density of the second population. Results similar to those reported here were obtained in a real-data resampling experiment using only an ($\mathbf{S}$, T) sample of size $n = 6{,}017$ (see Elmore et al. 2004). There the species was school whiting (*Sillago bassensis*), and we repeatedly drew datasets of size $m = n = 500$ by sampling without replacement, to elucidate the performance of our method using datasets of substantially smaller sizes than are typically encountered in practice.

Returning to the sand flathead data, by minimizing the criterion in (16), we obtain the smoothing parameters $\lambda = .0035$ and $\nu = 3$. The estimated density based on (16), using the aforementioned values of the smoothing parameters, is given as the solid line in Figures 4(a)–4(d). To gauge the sensitivity of the choice of smoothing parameters on the overall estimate, we varied λ from 0 to .3 and let $\nu = 2, 3, 4$, and 5. The estimated densities for the various values of λ and $\nu = 2, 3, 4$, and 5 are given by the dashed curves in Figures 4(a)–4(d). Note that as λ increases from 0 to .3, the estimated densities tend to move from an undersmoothed estimate to an oversmoothed estimate. Further, as we increase the number of terms, ν, included in the estimate, the magnitude of the spurious modes tends to increase as well.

There are drawbacks to these methods, however. In particular, it is not unlikely that the procedure will produce spurious modes in the final density estimate. This problem apparently occurs for two reasons: We are estimating orthonormal functions in L_2 with orthonormal vectors in $\mathbb{R}^2$, and the data-driven choices of ν and λ are stochastically variable. The former issue is known to be a problem; in fact, Silverman (1996) stated (albeit in a different context) that it may yield an "excessively variable or rough" approximation to the functions. The latter problem can usually be remedied by post hoc tuning of the parameters by the researcher.

5. THEORETICAL PROPERTIES

5.1 Rate of Convergence of $\tilde{g}$ to g

Define $\|\tilde{g} - g\|^2 = \int(\tilde{g} - g)^2$. Our aim in this section is to provide a rate of convergence of $E\|\tilde{g} - g\|^2$ to 0, under elementary conditions on the densities involved. It is possible to obtain slightly improved rates under more specific assumptions, for example, if θ_j and γ_j decrease in asymptotic proportion to $j^{-\alpha}$ and $j^{-\beta}$, say, as j increases. However, we wish to describe convergence in quite general circumstances. Indeed, an attractive feature of Theorem 1 is that there is no specific mention of θ_j or γ_j among the assumptions.

As in Section 2, we let $\mathcal{I}$ denote the compact interval on which the distributions of V and T are supported. Because the construction of $\tilde{g}$ involves dividing by estimators of f_T [see (14)], it is convenient to assume that f_T is bounded away from 0 on $\mathcal{I}$; without that condition, the convergence rate of $E\|\tilde{g}-g\|^2$ to 0 is influenced by the rates at which f_T decreases at points in $\mathcal{I}$ where it vanishes.

We take the kernels K_1 and K_2 in Section 2 to be bounded, supported on $[-1,1]^p$ and on $[-1,1]$, and of rth order, and when v is within h_2 of the left (resp. right) side of $\mathcal{I}$, we replace K_2, in the formulas $K_2\{(v-T_i)/h_2\}$ and $K_2\{(w-T_j)/h_2\}$ in Section 2, by a one-sided compactly supported rth-order kernel that vanishes on $[-1,0]$ (resp. $[0,1]$). This eliminates the main edge-effect problems suffered by our kernel-based estimators.

Because we wish to bound the expected value of the squared error, it is also necessary to modify the estimator so as to guard against instances where denominators take values too close to 0. This could be done by incorporating a conventional ridge parameter, but for technical simplicity we simply replace $\hat{f}_T$ in the denominators in (14) by a constant $c>0$ if $\hat{f}_T$ takes a value $<c$, where c is any number satisfying $0<c<\min_{\mathcal{I}} f_T$. We refer to the conventions in this and the previous paragraph as the "edge effect" and "ridging" rules.

Before stating our main result, some discussion of the relative impacts of m and n is in order. For simplicity, we consider m to be a function of n, and allow m and n to diverge together; that is, $m=m(n)\to\infty$ as n increases. If n is sufficiently greater than m, then, as is intuitively clear and may be proved theoretically, errors in estimating f_T and $f_{\mathbf{S}T}$ are negligible relative to those arising when estimating the function b, defined at (11). Consequently, for n/m sufficiently large, errors that arise when estimating the eigenvalues θ_j and eigenfunctions ϕ_j are negligible relative to those that occur when estimating the Fourier coefficient α_j, and if the bandwidths h_1 and h_2 are chosen appropriately,

$$E\|\tilde{g}-g\|^2 = O\left[\lambda^{-2}m^{-1}\sum_{j=1}^{\infty}\min(1,\theta_j/\lambda) + \sum_{j=1}^{\infty}\min\{1,(\lambda/\theta_j)^2\}\gamma_j^2\right]. \quad (18)$$

This result is closely analogous to (19).

However, as discussed in Section 1, in practice it is invariably the case that m is much larger than n, not the other way around. Therefore, although results such as (17) might be of theoretical interest, they have little practical bearing on the problem that motivates our work.

These considerations motivate us to make the realistic assumption that n/m is bounded as $n\to\infty$. Because the rate of convergence of $\tilde{g}$ to g is determined by the smaller of the two sample sizes, the fact that n/m is bounded implies that the regularity conditions will be governed by n, not by m. In particular, it is natural that m does not appear in condition (18), because if m were as small as a constant multiple of n, or as large as infinity, then we would still require that (18) hold. Note that a sufficient condition for n/m to be bounded is that $n\sim Cm^s$, where $C>0$ and $0<s\le 1$.

Theorem 1. Assume that the distributions of $(\mathbf{S},T)$ and $(\mathbf{U},V)$ have proper joint densities, with compact supports and with the marginal densities f_T and f_V bounded away from 0 on their common support, $\mathcal{I}$; that $f_{\mathbf{S}|T}$ and $f_{\mathbf{U}|V}$ are identical; that $f_{\mathbf{U}V}$ and $f_{\mathbf{S}T}$ both have $r\ge 1$ bounded derivatives on $\mathbb{R}^p\times\mathcal{I}$; that the kernels K_1 and K_2 used to construct $\tilde{g}$ are bounded, compactly supported, and of rth order; that the edge effect and ridging rules apply; that n/m is bounded as $n\to\infty$; and that for some $\epsilon>0$, the smoothing parameters h_1, h_2, and λ depend on m and n, and satisfy $n^{1-\epsilon}\min(h_1^p,h_2)\to\infty$, $n^{\epsilon}\max(h_1,h_2)\to 0$ and

$$0<\lambda\to 0, \qquad n^{\epsilon}\lambda^{-2}\{(nh_2)^{-1}+h_1^{2r}+h_2^{2r}\}\to 0. \quad (19)$$

Then

$$E\|\tilde{g}-g\|^2 = O\left[\lambda^{-2}(n^{-1}+h_1^{2r}+h_2^{2r})\sum_{j=1}^{\infty}\min(1,\theta_j/\lambda) + \lambda^{-2}(nh_2)^{-1} + \sum_{j=1}^{\infty}\min\{1,(\lambda/\theta_j)^2\}\gamma_j^2\right]. \quad (20)$$

It can be seen from (19) that condition (18) is virtually superfluous. As long as the smoothing parameters are chosen so that the right side of (19) converges to 0 at a polynomial rate, (18) will follow from (19).

To appreciate the implications of Theorem 1, assume for simplicity that θ_j and γ_j decrease in asymptotic proportion to $j^{-\alpha}$ and $j^{-\beta}$, as $j\to\infty$. (Necessarily, $\alpha>r+\frac{1}{2}$ and $\beta>r$.) It is intuitively clear that (a) the problem of estimating g is particularly difficult when α is large, and that (b) this difficulty is alleviated if β is also large. Indeed, the problem of estimating g becomes nonidentifiable as α diverges. To appreciate why this is so, note that if $\alpha=\infty$, then the expansion of A at (5) is only finite-dimensional, and so knowing the distribution of $\mathbf{S}|T$ gives us access to only a finite number of the components of g, whereas in general, g depends on an infinite number of components. However, having both β and α large provides amelioration, because it is then relatively less important to know high-order Fourier coefficients of g.

These properties are reflected in the convergence rate in Theorem 1. If we optimize the right side of (18) over the smoothing parameters h_1, h_2, and λ, then the right side of (19) equals $O(n^{-c(\alpha)})$, where, as $\alpha\to\infty$ for fixed β and r, $c(\alpha)\sim c_0(\alpha)\equiv r(2\beta-1)/\alpha(2r+1)$. The fact that $c_0(\alpha)$ decreases with increasing α reflects property (a), and the fact that it increases with increasing β reflects property (b). If α and β are permitted to diverge to infinity together, then it may be shown that $c(\alpha)\to r/(2r+1)$.

5.2 Root-n–Consistent Estimation of θ_j and α_j, and Estimation of ϕ_j

To appreciate properties in this case, let us assume that r, the number of derivatives, and p, the dimension of the explanatory variable S, satisfy $r>\frac{1}{2}p$. Choose h_2 to be of size between

$n^{-1/(2r)}$ and $n^{-1/(2r+1)}$ and h_1 to be of size between $n^{(\epsilon-1)/p}$ and h_2, for some $0 < \epsilon < (2r-p)/(2r)$. Then the bandwidth conditions imposed in Theorem 1, with the exception of (18), hold. For these bandwidths, the condition

$$h_1 = O(h_2) \tag{21}$$

is always valid, and if h_2 is at the lower end of its range, then the condition

$$h_1^{2r} + h_2^{2r} = O(n^{-1}) \tag{22}$$

applies. In Theorem 2, result (22) establishes root-n consistency of $\hat{\theta}_j$ and $\hat{\alpha}_j$ under (21). In addition, (23) shows that under (20), the mean squared difference between $\hat{\phi}_j$ and ϕ_j has the same variance and squared bias expansion that we commonly associate with much simpler problems of this type, when the target function has r bounded derivatives. In particular, the variance and squared bias are of orders $(nh_2)^{-1}$ and h_2^{2r}. As a result, the L_2 rate at which $\hat{\phi}_j$ converges to ϕ_j can be rendered equal to the familiar rate, $n^{-2r/(2r+1)}$, associated with r times differentiable functions, by choosing h_2 to be of size $n^{-1/(2r+1)}$ and h_1 to be of order between $n^{(\epsilon-1)/p}$ and $n^{-1/(2r+1)}$, for some $\epsilon > 0$. The fact that ϕ_j has r derivatives follows from the assumption of this level of smoothness of $f_{\mathbf{S}T}$.

Theorem 2. Assume all of the conditions in Theorem 1, except (18), and suppose also that $\theta_1, \ldots, \theta_{j+1}$ are distinct. If in addition (21) holds, then

$$|\hat{\theta}_j - \theta_j| + |\hat{\alpha}_j - \alpha_j| = O_p\bigl(n^{-1/2}\bigr). \tag{23}$$

If instead of (21), we assume (20), then

$$\|\hat{\phi}_j - \phi_j\| = O_p\bigl\{(nh_2)^{-1/2} + h_2^r\bigr\}. \tag{24}$$

APPENDIX: PROOF OF THEOREM 1

Given a function q and a functional Q, define $\|q\| = (\int q^2)^{1/2}$ and $\|Q\| = \sup \|Qq\|$, where the supremum is taken over q with $\|q\| = 1$. Note also that $\|Q\| \le |||Q|||$, where $|||Q|||^2 = \int Q^2$ and, in the latter integral, we view Q as a function rather than an operator. The inequality in terms of $|||Q|||$ allows us to use simple moment methods to bound quantities such as $E\|\widehat{A} - A\|^k$, for even integers k, for example, in (A.2) and (A.3) herein.

When the ridging rule is used to ward off problems caused by denominators, the effect is to add terms, of smaller order than n^{-C} for any $C > 0$, to the right sides of the upper bounds. The terms are derived using large-deviation bounds. But we do not explicitly include the terms in the inequalities that follow, because they may be assumed to be incorporated through increasing the values of unspecified constants on right sides, including those constants that are implicit in "big oh" bounds.

Put $A^- = (A + \lambda I)^{-1}$ and $q = A^- a = A^- A g$. Now

$$\begin{aligned}\tilde{g} - g = (\widehat{A}^- - A^-)(\hat{a} - a) + A^-(\hat{a} - a) \\ + (\widehat{A}^- - A^-)a + (A^- - A^{-1})a,\end{aligned}$$

from which it follows that

$$\begin{aligned}&\Bigl|(E\|\tilde{g} - g\|^2)^{1/2} \\ &\quad - \bigl\{E\|A^-(\hat{a} - a) - A^-(\widehat{A} - A)q + (A^- - A^{-1})a\|^2\bigr\}^{1/2}\Bigr| \\ &\le \bigl\{E\|(\widehat{A}^- - A^-)(\hat{a} - a)\|^2\bigr\}^{1/2} \\ &\quad + \bigl\{E\|(\widehat{A}^- - A^-)a + A^-(\widehat{A}^- - A^-)q\|^2\bigr\}^{1/2}.\end{aligned} \tag{A.1}$$

Note also that for each $j_0 \ge 2$, and for constants depending on j_0 but not on n,

$$\begin{aligned}&\bigl\{E\|(\widehat{A}^- - A^-)(\hat{a} - a)\|^2\bigr\}^{1/2} \\ &\quad \le \text{const} \cdot \Biggl(\sum_{j=1}^{j_0-1} \bigl[E\bigl\|\{A^-(\widehat{A} - A)\}^j A^-(\hat{a} - a)\bigr\|^2\bigr]^{1/2} \\ &\qquad + \lambda^{-(j_0+1)}(E\|\widehat{A} - A\|^{2j_0})^{1/4}(E\|\hat{a} - a\|^4)^{1/4}\Biggr),\end{aligned} \tag{A.2}$$

$$\begin{aligned}&\bigl\{E\|(\widehat{A}^- - A^-)a + A^-(\widehat{A} - A)q\|^2\bigr\}^{1/2} \\ &\quad \le \text{const} \cdot \Biggl(\sum_{j=2}^{j_0-1} \bigl[E\bigl\|\{A^-(\widehat{A} - A)\}^j A^- q\bigr\|^2\bigr]^{1/2} \\ &\qquad + \lambda^{-j_0}(E\|\widehat{A} - A\|^{2j_0})^{1/2}\Biggr).\end{aligned} \tag{A.3}$$

Moment calculations show that for some $\epsilon > 0$ and each $k \ge 1$,

$$\begin{aligned}E|||\widehat{A} - A|||^{2k} &= O\bigl[\{(nh_2)^{-1} + (n^2 h_1^p h_2)^{-1} + h_1^{2r} + h_2^{2r}\}^k\bigr] \\ &= O\{(n^{-\epsilon}\lambda^2)^k\}.\end{aligned} \tag{A.4}$$

Therefore, if $C > 0$ is given, then, by choosing j_0 sufficiently large, we may ensure that the terms involving $E\|\widehat{A} - A\|^{2j_0}$, on the right sides of (A.2) and (A.3), both equal $O(n^{-C})$. Simpler arguments can be used to show that the other terms on the right sides of (A.2) and (A.3) are all of smaller order than the square root of the term on the right side of (19). Therefore, the correctness of Theorem 1 rests in showing that the bound on the right side of (19) applies to the square of the subtracted term on the left side of (A.1), that is, in proving that

$$\begin{aligned}\xi &\equiv E\|A^-(\hat{a} - a) - A^-(\widehat{A} - A)q + (A^- - A^{-1})a\|^2 \\ &= O\Biggl\{\lambda^{-2}(n^{-1} + h_1^{2r} + h_2^{2r}) \sum_{j=1}^{\infty} \min(1, \theta_j/\lambda) \\ &\qquad + \lambda^{-2}(nh_2)^{-1} + \sum_{j=1}^{\infty} \min\{1, (\lambda/\theta_j)^2\}\gamma_j^2\Biggr\}.\end{aligned} \tag{A.5}$$

Define

$$\begin{aligned}D_1(v, w) &= E\bigl[\{\hat{a}(v) - a(v)\}\{\hat{a}(w) - a(w)\}\bigr], \\ D_2(v_1, v_2) &= \int q(w_1)q(w_2)E\bigl[\{\widehat{A}(v_1, w_1) - A(v_1, w_1)\} \\ &\qquad \times \{\widehat{A}(v_2, w_2) - A(v_2, w_2)\}\bigr]\, dw_1\, dw_2,\end{aligned}$$

$$\begin{aligned}J_1 &= \int E\{A^-(\hat{a} - a)\}^2 \\ &= \int (A^-)^2(v, w) D_1(v, w)\, dv\, dw,\end{aligned} \tag{A.6}$$

$$\begin{aligned}J_2 &= \int E\{A^-(\widehat{A} - A)q\}^2 \\ &= \int (A^-)^2(v, w) D_2(v, w)\, dv\, dw.\end{aligned} \tag{A.7}$$

In either case, we may write $J_k = \int (A^-)^2(v, w) D_k(v, w)\, dv\, dw$. We can expand $D_k(v, w)$, giving a sequence of explicit terms followed by a remainder that equals $\delta D_{k1}(v, w)$, say, where $\delta = \delta(n) \to 0$ and $\sup_{v,w} |D_{k1}(v, w)| \le C$, with $C > 0$ not depending on n. A bound for the contribution of the remainder to J_k thus may be obtained using the following result: If the function D, defined on $\mathcal{I}^2$, satisfies $\sup|D| \le C$, and if $0 < \lambda \le \frac{1}{2}$, then

$$\left|\int (A^-)^2(v,w)D(v,w)\,dv\,dw\right| \le C_1 C\lambda^{-2}\sum_{j=1}^{\infty}\min(1,\theta_j/\lambda), \qquad \text{(A.8)}$$

where $C_1 > 0$ depends only on θ_1 and $|\mathcal{I}|$, the latter denoting the length of the compact interval $\mathcal{I}$ on which the distributions of V and T are supported. The following lemma is related. It, and (A.8), were proved by Elmore, Hall, and Troynikov (2004).

Lemma A.1. Let $0 < h \le \frac{1}{2}$ and $C > 0$, and assume that $\mathcal{I} = [0, 1]$. Suppose also that the function $D(v, w)$ is defined on $[0, 1]^2$ and satisfies

$$D(v,w) = \int G(s,t,v,w)\,ds\,dt,$$

where G is defined on $[0, 1]^4$ and

$$|G(s,t,v,w)| \le CI(|u_1 - u_2| \le h), \qquad \text{(A.9)}$$

with u_1 and u_2 being any two distinct members of s, t, v, and w. Define $\xi = 1$ if $\{u_1, u_2\} = \{v, w\}$ and $\xi = 0$ otherwise. Then

$$\left|\int (A^-)^2(v,w)D(v,w)\,dv\,dw - \lambda^{-2}\xi\int D(v,v)\,dv\right| \le CC_1 h\lambda^{-2}\sum_{j=1}^{\infty}\min(1,\theta_j/\lambda), \qquad \text{(A.10)}$$

where $C_1 > 0$ depends only on θ_1. If, instead of (A.9),

$$|G(s,t,v,w)| \le CI(|s-t| \le h, |v-w| \le h), \qquad \text{(A.11)}$$

then (A.10) holds with h^2 rather than h on the right side.

By Taylor expansion, we obtain, for each $j_0 \ge 2$,

$$\begin{aligned}
&\Bigg|E\big[\{\hat a(v) - a(v)\}\{\hat a(w) - a(w)\}\big] \\
&\quad - E\Bigg[\Bigg[a(v)\sum_{j=1}^{j_0-1}\left\{-\frac{\hat f_T(v) - f_T(v)}{f_T(v)}\right\}^j + \frac{\hat b(v) - b(v)}{f_T(v)}\sum_{j=0}^{j_0-1}\left\{-\frac{\hat f_T(v) - f_T(v)}{f_T(v)}\right\}^j\Bigg] \\
&\quad \times \Bigg[a(w)\sum_{j=1}^{j_0-1}\left\{-\frac{\hat f_T(w) - f_T(w)}{f_T(w)}\right\}^j + \frac{\hat b(w) - b(w)}{f_T(w)}\sum_{j=0}^{j_0-1}\left\{-\frac{\hat f_T(w) - f_T(w)}{f_T(w)}\right\}^j\Bigg]\Bigg]\Bigg| \\
&= O\big\{(nh_2)^{-j_0/2} + h_2^{j_0 r}\big\}, \qquad \text{(A.12)}
\end{aligned}$$

where the "big oh" bound applies uniformly in v and w. The subtracted expected value within modulus signs on the left side of (A.12) can be expanded term by term. We deal directly only with the dominant term there,

$$\begin{aligned}
&\frac{a(v)a(w)}{f_T(v)f_T(w)}E\big[\{\hat f_T(v) - f_T(v)\}\{\hat f_T(w) - f_T(w)\}\big] \\
&\quad + \frac{1}{f_T(v)f_T(w)}E\big[\{\hat b(v) - b(v)\}\{\hat b(w) - b(w)\}\big] \\
&\quad - \frac{a(v)}{f_T(v)f_T(w)}E\big[\{\hat f_T(v) - f_T(v)\}\{\hat b(w) - b(w)\}\big] \\
&\quad - \frac{a(w)}{f_T(v)f_T(w)}E\big[\{\hat f_T(w) - f_T(w)\}\{\hat b(v) - b(v)\}\big].
\end{aligned}$$

The four terms in this formula are dominated by constant multiples of $(nh_2)^{-1}I(|v-w| \le C_1h_2) + h_1^{2r} + h_2^{2r}$,

$$\{(nh_2)^{-1} + (mnh_1^p h_2)^{-1}\}I(|v-w| \le C_1h_2) + m^{-1} + h_1^{2r} + h_2^{2r},$$

$t \equiv (nh_2)^{-1}I(|v-w| \le C_1h_2)$, and t, where $C_1 > 0$. Higher-order expansions admit the same bounds; in view of (A.8), (A.12), and the assumptions imposed on h_1 and h_2 in the theorem, we may stop the series at a sufficiently large value of j_0 and incur a remainder that is no larger than the right side of (A.13). Therefore, by (A.8) and Lemma A.1,

$$\begin{aligned}
J_1 = O\Bigg[&\lambda^{-2}\{n^{-1} + (mnh_1^p)^{-1} + h_1^{2r} + h_2^{2r}\}\sum_{j=1}^{\infty}\min(1,\theta_j/\lambda) \\
&+ \lambda^{-2}\{(nh_2)^{-1} + (mnh_1^p h_2)^{-1}\}\Bigg].
\end{aligned}$$

A similar argument, detailed by Elmore et al. (2004), shows that if we replace m by n on the right side of (A.13) and J_1 by J_2 [defined in (A.7)] on the left side, then (A.13) continues to hold. More simply, $J_3 \equiv \int\{(A^- - A^{-1})a\}^2 = \lambda^2\int(A^-g)^2$ satisfies

$$J_3 = \lambda^2\sum_{j=1}^{\infty}(\theta_j + \lambda)^{-2}\gamma_j^2 \le \sum_{j=1}^{\infty}\min\{1, (\lambda/\theta_j)^2\}\gamma_j^2. \qquad \text{(A.13)}$$

Using (A.13), its analog for J_2, and (A.14) to bound ξ, on the left side of (A.5), and using (A.2), (A.3), and the argument outlined in the paragraph after (A.3), followed by application of (A.8) and Lemma A.1, we deduce that

$$\begin{aligned}
\xi = O\Bigg[&\lambda^{-2}\{n^{-1} + (n^2h_1^p)^{-1} + h_1^{2r} + h_2^{2r}\}\sum_{j=1}^{\infty}\min(1,\theta_j/\lambda) \\
&+ \lambda^{-2}\{(nh_2)^{-1} + (n^2h_1^p h_2)^{-1}\} + \sum_{j=1}^{\infty}\min\{1, (\lambda/\theta_j)^2\}\gamma_j^2\Bigg].
\end{aligned}$$

This result, and the assumptions made about h_1 and h_2, imply (A.5) and hence Theorem 1.

[Received July 2004. Revised June 2005.]

REFERENCES

Barry, J., and Diggle, P. (1995), "Choosing the Smoothing Parameter in a Fourier Approach to Nonparametric Deconvolution of a Density Function," *Journal of Nonparametric Statistics*, 4, 223–232.

Bhatia, R., Davis, C., and Mcintosh, A. (1983), "Perturbation of Spectral Subspaces and Solution of Linear Operator Equations," *Linear Algebra Applications*, 52/53, 45–67.

Carroll, R. J., Ruppert, D., and Stefanski, L. A. (1995), *Measurement Error in Nonlinear Models*, London: Chapman & Hall.

Cavalier, L., Golubev, G. K., Picard, D., and Tsybakov, A. B. (2002), "Oracle Inequalities for Inverse Problems," *The Annals of Statistics*, 30, 843–874.

Delaigle, A. (2003), "Kernel Estimation in Deconvolution Problems," unpublished doctoral thesis, Institut de Statistique, Université Catholique de Louvain.

Donoho, D. L. (1995), "Nonlinear Solution of Linear Inverse Problems by Wavelet–Vaguelette Decomposition," *Applied Computational Harmonic Analysis*, 2, 101–126.

Efromovich, S., and Koltchinskii, V. (2001), "On Inverse Problems With Unknown Operators," *IEEE Transactions on Information Theory*, 47, 2876–2894.

Elmore, R. T., Hall, P., and Troynikov, V. S. (2004), "Nonparametric Density Estimation From Covariate Information," available at *http://www.stat.colostate.edu/~elmore/papers/eht_pp_jasa.pdf.*

Everett, B. S. (1998), *Cambridge Dictionary of Statistics*, Cambridge, U.K.: Cambridge University Press.

Fan, J. (1991), "On the Optimal Rates of Convergence for Nonparametric Deconvolution Problems," *The Annals of Statistics*, 19, 1257–1272.

——— (1993), "Adaptively Local One-Dimensional Subproblems With Application to a Deconvolution Problem," *The Annals of Statistics*, 21, 600–610.

Fan, J., and Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications*, London: Chapman & Hall.

Fan, J., and Koo, J. Y. (2002), "Wavelet Deconvolution," *IEEE Transactions on Information Theory*, 48, 734–747.

Gaffey, W. R. (1959), "A Consistent Estimator of a Component of a Convolution," *The Annals of Mathematical Statistics*, 30, 198–205.

Gilbert, S. (2002), "Testing the Distribution of Error Components in Panel Data Models," *Economics Letters*, 77, 47–53.

Horowitz, J. L., and Markatou, M. (1996), "Semiparametric Estimation of Regression Models for Panel Data," *Review Economic Studies*, 63, 145–168.

Indritz, J. (1963), *Methods in Analysis*, New York: Macmillan.

Johnstone, I. M. (1999), "Wavelet Shrinkage for Correlated Data and Inverse Problems: Adaptivity Results," *Statistica Sinica*, 9, 51–83.

Li, T. (2002), "Robust and Consistent Estimation of Nonlinear Errors-in-Variables, Models," *Journal of Econometrics*, 110, 1–26.

Li, T., and Hsiao, C. (2004), "Robust Estimation of Generalised Linear Models With Measurement Errors," *Journal of Econometrics*, 118, 51–65.

Li, T., and Vuong, Q. (1998), "Nonparametric Estimation of the Measurement Error Model Using Multiple Indicators," *Journal of Multivariate Analysis*, 65, 139–165.

Linton, O., and Whang, Y. J. (2002), "Nonparametric Estimation With Aggregated Data," *Economic Theory*, 18, 420–468.

Neumann, M. H. (1997), "On the Effect of Estimating the Error Density in Nonparametric Deconvolution," *Journal of Nonparametric Statistics*, 7, 307–330.

Osborne, C. (1991), "Statistical Calibration: A Review," *International Statistical Review*, 59, 309–336.

Pensky, M. (2002), "Density Deconvolution Based on Wavelets With Bounded Supports," *Statistics & Probability Letters*, 56, 261–269.

Pensky, M., and Vidakovic, B. (1999), "Adaptive Wavelet Estimator for Nonparametric Density Deconvolution," *The Annals of Statistics*, 27, 2033–2053.

Silverman, B. W. (1996), "Smoothed Functional Principal Components Analysis by Choice of Norm," *The Annals of Statistics*, 24, 1–24.

Simonoff, J. S. (1996), *Smoothing Methods in Statistics*, New York: Springer-Verlag.

Tikhonov, A. N. (1963), "On the Solution of Incorrectly Put Problems and the Regularisation Method," in *Outlines Joint Symposium on Partial Differential Equations*, Moscow: Academy of Science USSR Siberian Branch, pp. 261–265.

Troynikov, V. S. (1998), "Probability Density Functions Useful for Parametrization of Heterogeneity in Growth and Allometry Data," *Bulletin of Mathematical Biology*, 60, 1099–1122.

——— (1999), "Use of Bayes Theorem to Correct Size-Specific Sampling Bias in Growth Data," *Bulletin of Mathematical Biology*, 60, 355–363.

——— (2004), "Use of Weak Quasi-Solutions of the Fredholm First-Kind Equation in Problems With Scarce Data," *Applied Mathematics and Computations*, 150, 855–863.

Youndje, T., and Wells, M. T. (2002), "Least Squares Cross-Validation for the Kernel Deconvolution Density Estimator," *Comples Rendus Mathématique Academic des Sciences Paris*, 334, 509–513.

Walter, G. G. (1999), "Density Estimation in the Presence of Noise," *Statistics & Probability Letters*, 41, 237–246.