Using SIMEX for Smoothing-Parameter Choice in Errors-in-Variables Problems

Aurore DELAIGLE and Peter HALL

SIMEX methods are attractive for solving curve estimation problems in errors-in-variables regression, using parametric or semiparametric techniques. However, nonparametric approaches are generally of quite a different type, being based on, for example, kernels, local-linear modeling, ridging, orthogonal series, or splines. All of these techniques involve the challenging (and not well studied) issue of empirical smoothing parameter choice. We show that SIMEX can be used effectively for selecting smoothing parameters when applying nonparametric methods to errors-in-variable regression. In particular, we suggest an approach based on multiple error-inflated (or remeasured) data sets and extrapolation.

KEY WORDS: Bandwidth; Bootstrap; Cross-validation; Ill-posed problem; Inverse problem; Kernel estimation; Monte Carlo simulation; Nonparametric curve estimation; Nonparametric regression; Parametric model; Statistical smoothing.

1. INTRODUCTION

The SIMEX method for deconvolution enjoys a range of attractive features, including a high degree of accuracy in parametric and semiparametric settings. However, in errors-in-variables problems in nonparametric regression, because the extrapolation function is unknown, it provides an approximation, rather than a consistent estimator, of the regression curve. There exist several consistent nonparametric regression estimators in the errors-in-variables setting, but, of course, these share the major feature of nonparametric techniques: their reliance on smoothing parameters. In this article we suggest a SIMEX-type method for choosing the smoothing parameter, rather than for constructing the estimator itself. This approach overcomes the notorious difficulty of smoothing parameter selection in problems such as errors-in-variables regression.

The methods that we propose have the potential for very broad application, for example, to deconvolution techniques based on kernels, local linear modeling, orthogonal series, or ridging and to both density deconvolution and regression in the presence of measurement error. Indeed, our methods could be used for every statistical problem to which SIMEX methods can be applied, in conjunction with quite different, statistically consistent nonparametric methods, to choose the smoothing parameter. We outline differences between our method and conventional SIMEX at the end of Section 2.3. Nevertheless, the method is a SIMEX approach, being based on distinct SIMulation (or remeasurement) and EXtrapolation steps, which we clearly identify.

For brevity and simplicity, we introduce the methodology in the case of kernel estimation in errors-in-variables regression, but in our numerical study we also consider local linear and ridging approaches. Several methods for tuning parameter choice already exist in the setting of density estimation (see, e.g., Hesse 1999; Delaigle and Gijbels 2004a,b); therefore, the need for smoothing parameter selectors is not as pressing in that problem as it is in the regression case, which is more important in practice and currently has few options for choosing the level of smoothing.

In the context of normally distributed measurement errors and regression errors, Berry, Carroll, and Ruppert (2002) developed a Bayesian approach that has very good performance in a variety of cases. Their simulation study, examining robustness of their method against those parametric assumptions, is promising. Their method can outperform the structural regression approach of Carroll, Maca, and Ruppert (1999), which itself can beat the deconvolution kernel estimator. Note, however, that the poor performance of the deconvolution kernel estimator reported by Carroll et al. (1999) may be related to the fact that a ridge parameter was not used (see Sec. 3.1), and perhaps also to subtle numerical issues (see Delaigle and Gijbels 2007). When these difficulties are removed, deconvolution kernel methods can outperform SIMEX, for example.

Despite the very strong competitors just mentioned, nonparametric methods (such as those based on kernels) remain popular, because of their simplicity, their wide range of application, and their guaranteed consistency. In this article we show how to choose smoothing parameters empirically to achieve good performance when using nonparametric methods. However, this contribution should not be construed as advocating nonparametric methods over alternative approaches in specific situations.

Early contributions to nonparametric methodology for deconvolution include those of Carroll and Hall (1988), Devroye (1989), Stefanski and Carroll (1990), Zhang (1990), Fan (1991), and Fan and Truong (1993). The SIMEX method was introduced by Cook and Stefanski (1994) in a parametric setting (see also Stefanski and Cook 1995; Carroll, Kuchenhoff, Lombard, and Stefanski 1996; Stefanski and Bay 1996; Kuchenhoff and Carroll 1997; Kim and Gleser 2000; Stefanski 2000; Devanarayan and Stefanski 2002). Recent applications of the SIMEX procedure have been reported by Li and Lin (2003), Staudenmayer and Ruppert (2004), Kuchenhoff, Mwalili, and Lesaffre (2006), and Luo, Stefanski, and Boos (2006).

2. METHODOLOGY

2.1 Model and Estimator

Data (W_j, Y_j) are generated by the model

$$Y_j = g(X_j) + V_j \qquad \text{and} \qquad W_j = X_j + U_j, \tag{1}$$

Aurore Delaigle is Lecturer, Department of Mathematics, University of Bristol, Bristol, BS81 TW, U.K. and Belz Research Fellow, Department of Mathematics and Statistics, University of Melbourne, Melbourne, VIC 3010, Australia (E-mail: *Aurore.Delaigle@bris.ac.uk*). Peter Hall is Professor, Department of Mathematics and Statistics, University of Melbourne, Melbourne, Australia and Professor, Department of Statistics, University of California, Davis, CA 95616 (E-mail: *halpstat@ms.unimelb.edu.au*). The work was supported by a Maurice Belz Fellowship and by grants from the Australian Research Council and National Science Foundation.

Journal of the American Statistical Association
March 2008, Vol. 103, No. 481, Theory and Methods
DOI 10.1198/016214507000001355

where the variates U_j, V_j, and X_j for $1 \le j < \infty$ are totally independent, and the sequences $U_1, U_2, \ldots$; $V_1, V_2, \ldots$; and $X_1, X_2, \ldots$ are identically distributed as U, V, and X. The distribution of U is known, the distribution of V is unknown but has mean 0, and we wish to estimate the smooth function g.

Let K denote a kernel function, and let $K^{\mathrm{Ft}}(t) = \int e^{itu} \times K(u)\,du$ be its Fourier transform. Let $h > 0$ be a bandwidth, write f_U for the density of U, let f_U^{Ft} be the characteristic function of the distribution of U, and define

$$K_U(u) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itu} K^{\mathrm{Ft}}(t) f_U^{\mathrm{Ft}}(t/h)^{-1}\,dt. \tag{2}$$

A kernel estimator of g, constructed from the n data pairs (W_j, Y_j), was proposed by Fan and Truong (1993) and is given by

$$\hat{g}(x) = \frac{\sum_j Y_j K_U\{(x - W_j)/h\}}{\sum_j K_U\{(x - W_j)/h\}}, \tag{3}$$

where, here and in (6), each summation is over $1 \le j \le n$.

2.2 Cross-Validation Criteria

If we knew the values of $X_1, \ldots, X_n$ in addition to those of $(W_1, Y_1), \ldots, (W_n, Y_n)$, then we could compute a conventional cross-validation (CV) bandwidth $\hat{h}_0$: $\hat{h}_0 = \operatorname{argmin} \mathrm{CV}_0(h)$, where

$$\mathrm{CV}_0(h) = \frac{1}{n} \sum_{j=1}^{n} \{Y_j - \hat{g}_{-j}(X_j)\}^2 p(X_j). \tag{4}$$

Here $\hat{g}_{-j}$ denotes the version of $\hat{g}$, at (3), computed on omitting the jth data pair from the sample. The nonnegative function p in (4) represents a weight, used to prevent CV_0 from becoming too large through attempting to estimate $g(x)$ for values of x that lie in the tails of the distribution of X.

But the X_j's are not observable, and so CV_0 is not a practical criterion. We develop instead two versions of CV_0 for higher levels of observation error. Toward this end, let $U_1^*, U_2^*, \ldots$ and $U_1^{**}, U_2^{**}, \ldots$ denote independent and identically distributed random variables, also independent of the data pairs (W_j, Y_j) and having the distribution of U. Write $W_j^* = W_j + U_j^*$ and $W_j^{**} = W_j + U_j^* + U_j^{**}$ for $1 \le j \le n$, and consider the problem of estimating g_1 from the contaminated data (W_j^*, Y_j) or of estimating g_2 from values of (W_j^{**}, Y_j), where

$$\begin{aligned} g_1(x) &= E(Y \mid W = x) \qquad \text{and} \\ g_2(x) &= E(Y \mid W^* = x). \end{aligned} \tag{5}$$

Appropriate estimators can be based on (3),

$$\begin{aligned} \hat{g}_1^*(x) &= \frac{\sum_j Y_j K_U\{(x - W_j^*)/h\}}{\sum_j K_U\{(x - W_j^*)/h\}} \qquad \text{and} \\ \hat{g}_2^{**}(x) &= \frac{\sum_j Y_j K_U\{(x - W_j^{**})/h\}}{\sum_j K_U\{(x - W_j^{**})/h\}}. \end{aligned} \tag{6}$$

In this problem the variables W_j and W_j^* are known, and so we can use standard CV to determine the appropriate bandwidths for estimating g_1 and g_2. The respective criteria are

$$\mathrm{CV}^*(h) = \frac{1}{n} \sum_{j=1}^{n} \{Y_j - \hat{g}_{1,-j}^*(W_j)\}^2 p(W_j) \tag{7}$$

and

$$\mathrm{CV}^{**}(h) = \frac{1}{n} \sum_{j=1}^{n} \{Y_j - \hat{g}_{2,-j}^{**}(W_j^*)\}^2 p(W_j^*), \tag{8}$$

where the subscript "$-j$" indicates that we omit the jth data pair when constructing the estimator. The function p in (7) and (8) is identical to that in (4). The step leading from CV_0, at (4) to CV^* and CV^{**} at (7) and (8), is the SIMulation, or "remeasurement," step of SIMEX.

2.3 Using Cross-Validation to Choose Bandwidth

The cross-validation bandwidths derived from the criteria at (7) and (8) are $\hat{h}_1^* = \operatorname{argmin} \mathrm{CV}^*(h)$ and $\hat{h}_2^{**} = \operatorname{argmin} \mathrm{CV}^{**}(h)$. Of course, $\hat{h}_1^*$ depends on the simulated data W_j^*, and $\hat{h}_2^{**}$ also depends on the values of W_j^{**}. This relationship can be removed by averaging over a large number B of versions of CV^* and CV^{**}, at (7) and (8), for different simulated sequences (U_k^*, U_k^{**}), thus obtaining approximations to the quantities

$$\mathrm{CV}_1 = E(\mathrm{CV}^* \mid \mathcal{D}) \qquad \text{and} \qquad \mathrm{CV}_2 = E(\mathrm{CV}^{**} \mid \mathcal{D}), \tag{9}$$

where $\mathcal{D} = \{(W_1, Y_1), \ldots, (W_n, Y_n)\}$ denotes the data set. In practice, CV_1 and CV_2 would be computed as

$$\mathrm{CV}_1 = \frac{1}{B} \sum_{b=1}^{B} \mathrm{CV}_b^* \qquad \text{and} \qquad \mathrm{CV}_2 = \frac{1}{B} \sum_{b=1}^{B} \mathrm{CV}_b^{**}, \tag{10}$$

where CV_b^*, $1 \le b \le B$, and CV_b^{**}, $1 \le b \le B$, denote versions of CV^* and CV^{**} computed using independent values of U_j^* (in the definition W_j^*) and U_j^{**} (in W_j^{**}). We then define, for $j = 0, 1, 2$,

$$\hat{h}_j = \operatorname{argmin} \mathrm{CV}_j(h). \tag{11}$$

W^{**} measures W^* in the same way that W^* measures W, and W measures X. Thus we expect the relationship between $\hat{h}_0$ and $\hat{h}_1$ to be similar to that between $\hat{h}_1$ and $\hat{h}_2$, and so back-extrapolation, in the SIMEX fashion, can be used to produce an approximation to $\hat{h}_0$. For example, $\log(\hat{h}_0) - \log(\hat{h}_1) \approx \log(\hat{h}_1) - \log(\hat{h}_2)$, so that linear back-extrapolation from the pair $(\log \hat{h}_1, \log \hat{h}_2)$ might be used. This suggests taking the final bandwidth to be

$$\tilde{h}_0 = \hat{h}_1^2 / \hat{h}_2. \tag{12}$$

(Justification in the case of small error variance is given in Sec. 4.4.) This represents the EXtrapolation step of SIMEX. We show in Section 4.3 that under general constraints, the bandwidth $\tilde{h}_0$ is of the same size as the asymptotically optimal bandwidth, say h_0, that minimizes the asymptotic mean integrated squared error of $\hat{g}$. To be itself asymptotically optimal, this bandwidth requires adjustment by only a constant factor.

Although these methods are of SIMEX type, they differ from conventional, contemporary SIMEX in that they involve adding "whole errors" to the existing data. A more general approach would be to define W_j^* and W_j^{**} by adding "fractions" of the error variables U_j^* and U_j^{**}, for example, taking $W_j^* = W_j + \sqrt{\lambda} U_j^*$, where $0 < \lambda \le 1$, and interpolating either naively or by making reference to the way in which using $\lambda < 1$ rather than $\lambda = 1$ alters the effect of the error distribution on optimal choice of bandwidth. Also, more than a single extrapolant could be used.

3. NUMERICAL PROPERTIES

3.1 Generalization of the Method

Although our detailed exposition focuses on the deconvolution kernel estimator, the procedure can be applied to select the smoothing parameter of other consistent deconvolution procedures, such as the ridge-based approach of Hall and Meister (2007) and the local linear (LL) procedure of Carroll, Delaigle, and Fan (2008). In all cases, we let h denote the smoothing parameter. For these three procedures, the estimator of the regression curve g is of the type

$$\hat{g}(x) = N_W(x)/D_W(x),$$

where the subscript "W" indicates the dependence on the sample $W_1, \ldots, W_n$. For example, in the kernel case, $N_W(x) = \sum_j Y_j K_U\{(x - W_j)/h\}$ and $D_W(x) = \sum_j K_U\{(x - W_j)/h\}$. In practice, as in the error-free case, such estimators can perform rather poorly if the denominator is too close to 0. To circumvent this difficulty, we modify the estimators $\hat{g}$ and their remeasured counterparts $\hat{g}_1^*$ and $\hat{g}_2^{**}$, incorporating a ridge parameter in the denominator; that is, at each stage of the procedure, we use the following version of the estimators:

$$\bar{g}(x) = N_T(x)/\max\{D_T(x), \rho\}, \tag{13}$$

where the pair $(\bar{g}, T)$ denotes any one of $(\hat{g}, W)$, $(\hat{g}_1^*, W^*)$, or $(\hat{g}_2^{**}, W^{**})$. Carroll, Delaigle, and Hall (2007) discussed properties of a related estimator.

Instead of choosing ρ and h simultaneously through our CV procedure, which would be technically involved and not always adequate, we select ρ by a preliminary CV procedure for the estimation of $\hat{g}_1^*$. More precisely, we find the value (h^*, ρ^*) that minimizes $\mathrm{CV}^*(h, \rho) = B^{-1} \sum_b CV_b^*(h, \rho)$. Taking $\rho = \rho^*$, we apply our CV procedure to select h.

3.2 Simulation Settings

We considered four different regression models:

(a) $g(x) = 5\sin(2x)\exp(-16x^2/50)$, $X \sim \mathrm{N}(0, 1.5)$, $V \sim \phi_{0,1.19}$;

(b) $g(x) = x^2$, $X \sim .7X_1 + .3X_2$, where $X_1 \sim f_{X_1}(x) = 1.5x^2 1_{[-1,1]}(x)$, $X_2 \sim \mathrm{U}[-1, 1]$, $\mathrm{V} \sim \phi_{0,.26}$;

(c) $Y|X = x \sim \mathrm{Be}\{g(x)\}$, $g(x) = .45\sin(2\pi x) + .5$, $X \sim \mathrm{U}[0, 1]$; and

(d) $g(x) = \phi_{0,1.5}(4x) + \phi_{1,2}(4x) + \phi_{2,5}(4x)$, $X \sim \mathrm{N}(0, 1.5)$, $V \sim \phi_{0,.0028}$,

where Be represents Bernoulli and $\phi_{\mu,\sigma}(x)$ denotes the density of a $\mathrm{N}(\mu, \sigma^2)$ variable. We took $U \sim$ Laplace or centered normal with $\mathrm{var}(U)/\mathrm{var}(X) = 10\%$ or 20%. We considered normal errors U to illustrate the fact that the results given in Section 4 can be extended to errors that have a characteristic function tending to 0 exponentially fast. We took $B = 15$, $p(W_j) = \hat{f}_W(W_j)$, and $p(W_j^*) = \hat{f}_{W^*}(W_j^*)$, where, for $T = W$ or W^*, $\hat{f}_T$ is the usual kernel density estimator of f_T with normal reference bandwidth. For the kernel and the LL methods, we used the kernel for which $K^{\mathrm{Ft}}(t) = (1 - t^2)^3 1_{[-1,1]}(t)$. The ridge-based approach of Hall and Meister (2007) does not need a kernel.

In each case, we generated 200 samples of size $n = 100$, 250, or 500 from (W, Y), and constructed the corresponding 200 estimators of g. For each we calculated the integrated squared error, $\mathrm{ISE}_{\hat{g}} = \int_a^b (\hat{g} - g)^2$, on the interval $[a, b]$ where the curves are presented. In the figures we show the three estimates of g corresponding to the first (q1), second (q2), and third (q3) quartiles of these 200 ISEs. We show only part of the simulation results, but our conclusions can be extended to cases not presented here. In all graphs, the target curve is represented by the solid curve.

To illustrate the importance of taking the error U into account, we calculated the naive LL smoother that ignored the error and used a CV method to select (h, ρ), where the ridge ρ was introduced to avoid problems in the denominator. Finally, we calculated a nonparametric SIMEX LL regression procedure. We generated the SIMEX samples $W_{ib}(\lambda) = W_i + \sqrt{\lambda}\epsilon_{ib}$, where, in the notation of Carroll et al. (1999), $\epsilon_{ib} \sim f_U$, $\lambda = (0, .5, 1, 1.5, 2)$ and $B = 20$. To the best of our knowledge, there does not exist a sophisticated algorithm for SIMEX that includes a ridge; thus we selected the smoothing parameters (ridge and bandwidths) by a CV procedure. We used linear back-extrapolation. Note that we include SIMEX only for illustration purposes, and that our implementation of SIMEX is not necessarily optimal; for example, we could use larger values of B and choose the smoothing parameters in a more elaborate way.

3.3 Results of Simulations

The main goal of this section is not to advocate a particular regression technique, but rather to use a few examples to show that our method for selecting the smoothing parameters works well in practice. Overall, we found that, except for the naive estimator, all methods (SIMEX, ridge, local constant kernel, and LL kernel) often gave similar results.

Figure 1 shows the results obtained for estimating curve (a) from samples of size $n = 250$ contaminated by normal errors with $\mathrm{var}(U) = 20\%\,\mathrm{var}(X)$. We give the quartile curves obtained for the naive LL estimator (NAIVE) and the ridge approach of Hall and Meister (2007), using the data-driven method (R DD) or the theoretical optimal values of ρ and h (R OPT). We calculated the latter for each sample by minimizing the ISE. We also provide the boxplots of ISEs calculated for these estimators. The first group of three boxplots is for $\mathrm{var}(U) = 10\%\,\mathrm{var}(X)$, and the second is for $\mathrm{var}(U) = 20\%\,\mathrm{var}(X)$. We can see that in both cases, the data-driven method performs quite well and the naive estimator performs rather poorly. Our implementation of the SIMEX method did not give good results in this example.

Figure 2 shows the results obtained for samples of size $n = 500$ generated from curve (b) when $U \sim$ Laplace and $\mathrm{var}(U) = 10\%\,\mathrm{var}(X)$, using our data-driven method applied to the LL deconvolution estimator of Carrol et al. (2008), the SIMEX LL estimator (SIMEX), and the naive local linear estimator (NAIVE). Here the design density has large discontinuities at the endpoints of its support; thus the LL estimator is particularly appropriate. We also show boxplots of the ISEs for these three estimators and the deconvolution LL method using the theoretically optimal values of ρ and h calculated for each sample by minimizing the ISE (LL OPT). In this case, both the SIMEX and the data-driven consistent method gave good results, both strongly outperforming the naive estimator.

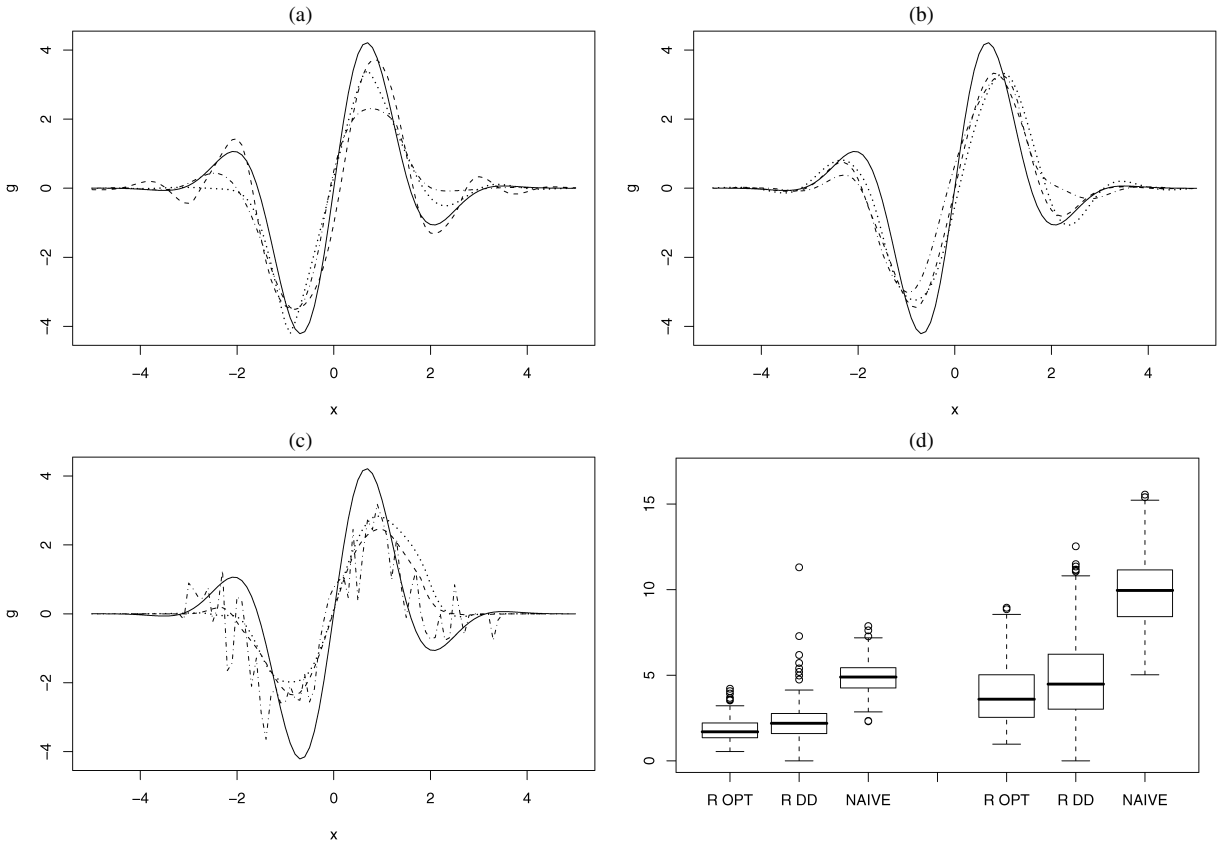

Figure 1. Quartile curves of 200 estimates of curve (a) when U is normal with $\text{var}(U) = 20\%\,\text{var}(X)$ and $n = 250$, using the optimal (a) or data-driven (b) ridge-based method or the naive method (c) (- - -, q1; ·····, q2; - · -, q3). (d) Boxplots of the ISEs for $\text{var}(U) = 10\%\,\text{var}(X)$ (first group of three) or $\text{var}(U) = 20\%\,\text{var}(X)$ (last group of three).

3.4 A Note on the Unknown Error Case

Our procedure can be adapted to the case in which the symmetric error density f_U is unknown and the observations are of the form (W_{i1}, W_{i2}, Y_i), $i = 1, \ldots, n$, where, for $j = 1, 2$, $W_{ij} = X_i + U_{ij}$ and the errors U_{ij} are iid. The unknown f_U^{Ft}'s appearing in K_U can be replaced by an estimator constructed from the replicated data, and our bandwidth selection procedure may be applied in exactly the same way to this regression estimator; the only change is in the how we generate the errors U^* and U^{**}, because the error density is unknown. A simple method involves applying the estimator of Delaigle, Hall, and Meister (2008) to the averaged data $(\bar{W}_i, Y_i)$, where $\bar{W}_i = (W_{i1} + W_{i2})/2 = X_i + Z_i$ with $Z_i \equiv (U_{i1} + U_{i2})/2$. Then the errors Z^* and Z^{**} can be generated by drawing with replacement from the sample $(W_{i1} - W_{i2})/2$, $i = 1, \ldots, n$. A more elaborate procedure would be to generate the errors from a standard kernel density estimator of f_Z, but our simulations indicated that this did not seem necessary.

To illustrate the method, we generated 200 samples of size $n = 500$, where g was curve (c), the error U was $\text{N}(0, \sigma_U^2)$ with $\sigma_U^2 = 20\%\,\text{var}(X)$, and each W_i was replicated once. The results are shown in Figure 3, with quartile curves of the data-driven deconvolution kernel method (KERNEL), SIMEX LL method (SIMEX), and naive LL estimator (NAIVE). Also shown are boxplots of the ISEs of these estimators and the data-driven ridge estimator (RIDGE). In this example, all methods performed quite similarly, and the naive estimator gave reasonable results (although inferior to the other methods).

3.5 Real Data Example

We applied the bandwidth selection procedure to the Framingham data used by Carroll, Ruppert, Stefanski, and Crainiceanu (2006), where the goal was to predict the risk of coronary heart disease (CHD) from systolic blood pressure (SBP). For 1,615 male patients, SBP was measured twice at each of two exams. We took W_{i1} and W_{i2}, equal to the logarithm of (the average -50) of the two replicated measurements obtained at exam 1, (resp. exam 2); Y_i indicated the presence (1) or absence (0) of CHD over an 8-year follow-up period.

We applied the method of Section 3.4 to the data $(\bar{W}_i, Y_i)$, using the deconvolution kernel estimator (KERNEL). We also calculated the naive estimator (NAIVE) and the SIMEX estimator (SIMEX), but for both of them we used a local constant version, because the LL method appeared to be too wiggly to be realistic. Finally, we calculated the parametric logistic (LOGIST) fit using regression calibration, as done by Carroll et al. (2006). The estimated curves, shown in Figure 4, are very similar between 4 and 4.6, but then the three nonparametric curves deviate from the logistic model, with the naive estimator being a bit smoother than the other two. Of course, it is impossible to say whether or not the true curve fits a logistic model, but it would be interesting for future research to develop techniques for testing such hypotheses.

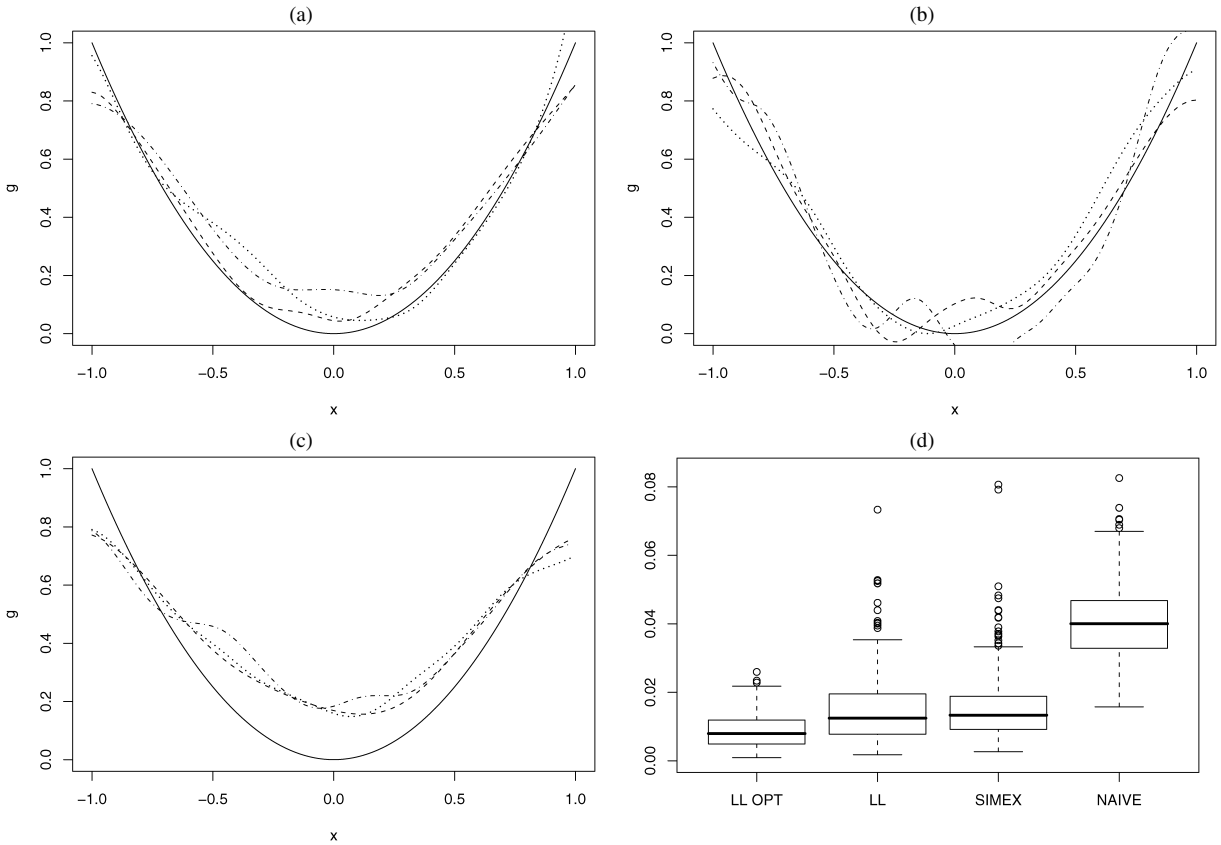

Figure 2. Quartile curves of 200 estimates of curve (b) when U is Laplace with $\text{var}(U) = 10\% \text{var}(X)$ and $n = 500$, using the data-driven local linear (a) or SIMEX (b) method, or the naive estimator (c) (- - -, q1; ·····, q2; -·-·-, q3). (d) Boxplots of the ISEs for several estimators.

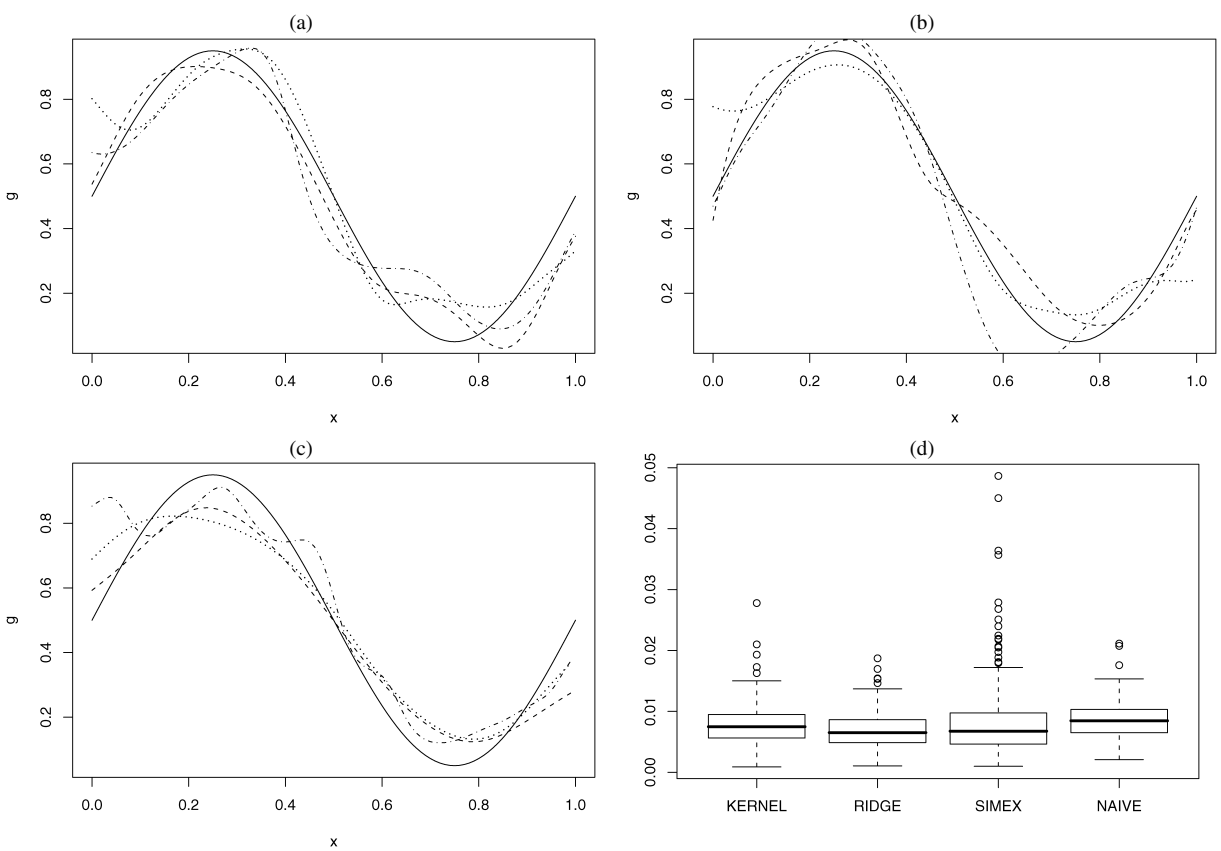

Figure 3. Quartile curves of 200 estimates of the regression function (d) in the unknown error case when $n = 500$ and U is normal, using the data-driven kernel (a), SIMEX (b), or naive (c) method (- - -, q1; ·····, q2; -·-·-, q3). (d) Boxplots of the ISEs for several estimators.

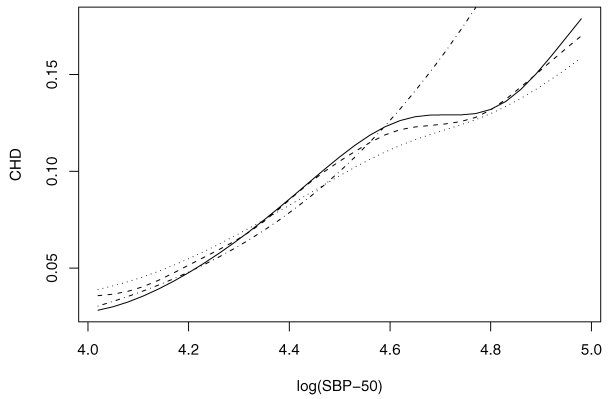

Figure 4. Estimation of the risk of CHD based on the Framingham data for the deconvolution kernel (——), SIMEX (- - -), naive (·····), and logistic (-·-) estimators.

4. THEORETICAL PROPERTIES

4.1 Relationship Between Cross-Validation and Integrated Squared Error

We show that the cross-validation criteria CV_0, CV_1, and CV_2 [see (4) and (9)] can be viewed as having been constructed with the aim of finding an empirical approximation to the bandwidths that minimize certain weighted forms of asymptotic mean integrated squared error (AMISE). Define $g_0 = g$, $\hat{g}_0 = \hat{g}$, $\hat{g}_1 = \hat{g}_1^*$, and $\hat{g}_2 = \hat{g}_2^{**}$, with $\hat{g}$, $\hat{g}_1^*$, and $\hat{g}_2^{**}$ as in (3) and (6) and let $W^{(j)}$ denote X, W, and W^* in the cases where $j = 0, 1, 2$. Then, for $j = 0, 1, 2$, each of the bandwidths $\hat{h}_j$ defined at (11) can be viewed as an approximation to the bandwidth h_j that minimizes

$$\mathrm{AMISE}(\hat{g}_j - g_j) = \int \mathrm{AMSE}\{\hat{g}_j(x) - g_j(x)\} \times f_{W^{(j)}}(x) p(x)\, dx, \quad (14)$$

where the functions g, g_1, and g_2 are as in (1) and (5) and AMSE denotes asymptotic mean squared error. In Section 4.2 we argue that $\hat{h}_j / h_j \to 1$ in probability as $n \to \infty$; see (16).

To simplify the presentation, we give the main results in the next two sections and then gather the technical conditions in two separate Sections 4.4 and 4.5. Proofs of the theorems are available in a longer version of the article, available from the authors request.

4.2 Properties of Cross-Validation Bandwidths

We start by studying the asymptotic behavior of the bandwidths. In particular, (16) states bandwidth consistency.

Theorem 1. Assume that (10), (24)–(27), (29) and (31) hold. If (30) is true for $Z = X$ (resp. for $Z = W$ or W^*), then in the case where $j = 0$ (resp., where $j = 1$ or $j = 2$), we have

$$\mathrm{CV}_j(h) = \{1 + o_p(1)\}\,\mathrm{AMISE}(\hat{g}_j - g_j) + T \quad (15)$$

uniformly in $h \in \mathcal{H}$, where $T = T(n)$ denotes a quantity that does not depend on h. Moreover,

$$\hat{h}_j / h_j \to 1 \quad (16)$$

in probability, where h_j denotes the bandwidths that minimize (14).

To appreciate the implications of (16), assume for simplicity that the characteristic function f_U^{Ft} satisfies the polynomial decay condition

$$|f_U^{\mathrm{Ft}}(t)| \sim \text{const.}|t|^{-\alpha} \quad (17)$$

as $|t|$ increases, where $\alpha > 0$ is a measure of the smoothness of the distribution of U. (Other, more general constraints also lead to the conclusions that we draw later.) Then asymptotic formulas for the bias and variance of deconvolution kernel regression estimators derived by Fan and Truong (1993) can be used to show that

$$\mathrm{AMISE}(\hat{g}_j - g_j) \sim (nh^{2\alpha+1})^{-1} C + h^4 B_j, \quad (18)$$

where C and D_j denote fixed positive constants and where

$$B_j = \frac{1}{4}\kappa_2^2 \int f_{W^{(j)}} p q_j^2, \quad (19)$$

with $\kappa_2 = \int x^2 K(x)\, dx$ and $q_j = g_j'' + 2 f_{W^{(j)}}' g_j' f_{W^{(j)}}^{-1}$. Results (16) and (18) imply that $\hat{h}_j$ and the bandwidth h_j satisfy, for $j = 0, 1, 2$,

$$\hat{h}_j \sim_p h_j \sim D_j n^{-1/(2\alpha+5)}, \quad (20)$$

which in turn implies that, regardless of whether $j = 0$, 1, or 2, the bandwidth $\hat{h}_j$ that minimizes CV_j is of size $n^{-1/(2\alpha+5)}$, and that the bandwidth $\tilde{h}$, defined at (12), satisfies

$$\tilde{h} \sim_p D n^{-1/(2\alpha+5)},$$

where $D = D_1^2 D_2^{-1}$. Therefore, $\tilde{h}$ is of the same order as the asymptotically optimal bandwidth, h_0, that minimizes $\mathrm{AMISE}(\hat{g}_0 - g_0)$.

4.3 Optimality of Linear Back-Extrapolation in the Low-Noise Case

For the sake of simplicity, we again assume that the characteristic function f_U^{Ft} satisfies (17). In this setting, (18) holds. From (16) and (18), it can be seen that the bandwidth h_j that minimizes $\mathrm{AMISE}(\hat{g}_j - g_j)$ satisfies

$$\hat{h}_j \sim_p h_j \sim \left\{ \frac{(2\alpha+1)C}{4B_j n} \right\}^{1/(2\alpha+5)} \quad (21)$$

as $n \to \infty$. Suppose that we can write $U = \sigma_U T$, where the distribution of the random variable T has mean 0 and unit variance, and σ_U^2 is fixed but small. (This property characterizes the "low-noise" case referred to in the heading of this section.) Also let ψ (with or without subscripts) denote a function determined by f_T, f_X, and g and satisfying $\psi(u) = o(u)$ as $u \to 0$. The next theorem describes the behavior of B_j as $\sigma_U \to 0$.

Theorem 2. If (32) holds, then we have

$$B_j = B_0\{1 + j Q \sigma_U^2 + \psi_{1j}(\sigma_U^2)\}, \quad (22)$$

where the constant Q depends only on f_T, f_X, and g and in particular does not depend on σ_U^2.

Properties (21) and (22) imply that

$$\log \hat{h}_j = d \log(A/nB_0) - jQ\sigma_U^2 + \psi_{2j}(\sigma_U^2) + o_p(1)$$

for $j = 0, 1, 2$, where $A > 0$. It follows from this result, and from (21) for $j = 0$, that if we compute $\log \tilde{h}_0$ by linear back-extrapolation from $\log \hat{h}_1$ and $\log \hat{h}_2$, then $\tilde{h}_0$ satisfies

$$\tilde{h}_0 \sim_p \hat{h}_0\{1 + \psi(\sigma_U^2)\} \sim_p h_0\{1 + \psi(\sigma_U^2)\}. \tag{23}$$

This result justifies linear back-extrapolation when the variance of U is small. In particular, comparing (21) and (23), we see that if σ_U^2 is sufficiently small, although fixed, then the linearly back-extrapolated bandwidth estimator $\tilde{h}_0$ is, with probability converging to 1 as $n \to \infty$, closer to the desired bandwidth h_0 than either $\hat{h}_1$ or $\hat{h}_2$ is.

4.4 Technical Assumptions for Theorem 1

In addition to asking that the densities f_X, f_W, and f_{W^*} of X, W, and W^* are well defined, we assume that the following hold:

$$(1 + x^2)|K(x)| \text{ is integrable, } K^{\mathrm{Ft}} \text{ has a compact support, } K^{\mathrm{Ft}}(0) \neq 0; \tag{24}$$

$$f_U^{\mathrm{Ft}} \text{ does not vanish on the real line;} \tag{25}$$

$$p \text{ is nonnegative, bounded, and supported on a compact interval } \mathcal{S}_p; \tag{26}$$

$$g \text{ is bounded and has two continuous derivatives on the real line, } \sigma_V \neq 0 \text{ and } E(|V|^C) < \infty, \text{ for all } C > 0, \tag{27}$$

where $\sigma_V^2 = \operatorname{var} V$. Define

$$\lambda(h) = \int K_U(v)^2\, dv = \frac{1}{2\pi} \int |K^{\mathrm{Ft}}(t)|^2 |f_U^{\mathrm{Ft}}(t/h)|^{-2}\, dt, \tag{28}$$

and, given $\epsilon \in (0, 1)$, let $\mathcal{H} = \mathcal{H}(n)$ denote a set of values of $h > 0$ with the property that, as $n \to \infty$,

$$\text{(a) } \sup_{h \in \mathcal{H}} h = O(n^{-\epsilon}), \text{ (b) } \sup_{h \in \mathcal{H}} \lambda(h)/h = O(n^{1-\epsilon}), \text{ and (c) the bandwidths } h_j \text{ that minimize } \operatorname{amise}(\hat{g}_j - g_j), \text{ are, for all sufficiently large } n \text{ and for } j = 0, 1, 2, \text{ in } \mathcal{H}. \tag{29}$$

Taking Z to denote X, W, or W^*, we ask that

$$\text{(a) } f_Z \text{ has two bounded and continuous derivatives on the real line and is bounded away from 0 on } \mathcal{S}_p; \text{ (b) } \sup f_U < \infty \text{ and } E|U|^\epsilon < \infty \text{ for some } \epsilon > 0; \text{ (c) } E|X|^\epsilon < \infty \text{ for some } \epsilon > 0; \text{ and (d) the function } \lambda, \text{ defined at (28), satisfies } C_1 h^{-C_2} \le \lambda(h) \le C_3 h^{-C_4} \text{ for constants } 0 < C_1 < C_3 < \infty \text{ and } 0 < C_2 < C_4 < \infty, \text{ and all } h \in (0, 1]. \tag{30}$$

To relate (29) and (30) to earlier discussion, note that (30)(d) is implied by (20), which in turn follows from (17), and if f_U^{Ft} satisfies (17) for a constant $\alpha > 0$, then (20) holds and implies that a sufficient condition for part (b) of (29) is $\inf_{h \in \mathcal{H}} h \ge \text{const.} n^{-(1-\epsilon)/(2\alpha+1)}$. Moreover, if $\kappa_2 \neq 0$, then h_0, h_1, and h_2 are each asymptotic to constant multiples of $n^{-1/(2\alpha+5)}$. Therefore, if we define $\mathcal{H}$ to be the set of values h for which $n^{-(1-\epsilon_1)/(2\alpha+1)} \le h \le n^{-\epsilon_2}$, where $0 < \epsilon_1 < 4/(2\alpha + 5)$ and $0 < \epsilon_2 < 1/(2\alpha + 5)$, then each of parts (a)–(c) of (29) holds, with $\epsilon \in (0, \min(\epsilon_1, \epsilon_2))$.

Finally, we take B in (10) to be no more than polynomially large as a function of n,

$$B = B(n) \to \infty \quad \text{and} \quad B = O(n^C) \quad \text{for some } C > 0. \tag{31}$$

This ensures that certain very pathological cases (the probabilities of which are exponentially small as functions of n) have a negligibly small chance of occurring among any of the summands in the definitions at (10), thereby guaranteeing that the denominators in the definitions of $\hat{g}^*_{1,-j}$ and $\hat{g}^{**}_{2,-j}$ in (6) and (7) are not too close to 0.

4.5 Technical Assumptions for Theorem 2

$$U = \sigma_U T, \text{ where the distribution of the random variable } T \text{ is held fixed, and } \sigma_U > 0 \text{ is permitted to decrease to zero; } E(T^2) < \infty \text{ and } E(T) = 0; \text{ the support, } \mathcal{S}_p, \text{ of the bounded, nonnegative function } p \text{ is a compact set; } f_X \text{ has three continuous derivatives, and } g \text{ has four continuous derivatives, on an open set containing } \mathcal{S}_p;\ f_X > 0 \text{ on } \mathcal{S}_p. \tag{32}$$

[Received April 2007. Revised September 2007.]

REFERENCES

Berry, S., Carroll, R. J., and Ruppert, D. (2002), "Bayesian Smoothing and Regression Splines for Measurement Error Problems," *Journal of the American Statistical Association* 97, 160–169.

Carroll, R. J., and Hall, P. (1988), "Optimal Rates of Convergence for Deconvolving a Density," *Journal of the American Statistical Association*, 83, 1184–1186.

Carroll, R. J., Delaigle, A., and Hall, P. (2007), "Nonparametric Regression Estimation From Data Contaminated by a Mixture of Berkson and Classical Errors," *Journal of the Royal Statistical Society*, Ser. B, 69, 859–878.

Carroll, R. J., Kuchenhoff, H., Lombard, F., and Stefanski, L. A. (1996), "Asymptotics for the SIMEX Estimator in Nonlinear Measurement Error Models," *Journal of the American Statistical Association*, 91, 242–250.

Carroll, R. J., Maca, J. D., and Ruppert, D. (1999), "Nonparametric Regression in the Presence of Measurement Error," *Biometrika*, 86, 541–554.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006), *Measurement Error in Nonlinear Models* (2nd ed.), Boca Raton, FL: Chapman & Hall CRC Press.

Cook, J. R., and Stefanski, L. A. (1994), "Simulation-Extrapolation Estimation in Parametric Measurement Error Models," *Journal of the American Statistical Association*, 89, 1314–1328.

Delaigle, A., and Gijbels, I. (2004a), "Bootstrap Bandwidth Selection in Kernel Density Estimation From a Contaminated Sample," *Annals of the Institute of Statistical Mathematics*, 56, 19–47.

——— (2004b), "Practical Bandwidth Selection in Deconvolution Kernel Density Estimation," *Computational Statistics—Data Analysis*, 45, 249–267.

——— (2007), "Frequent Problems in Calculating Integrals and Optimizing Objective Functions: A Case Study in Density Deconvolution," *Statistics and Computing*, 17, 349–355.

Delaigle, A., Fan, J., and Carroll, R. J. (2008), "Local Polynomial Estimator for the Errors-in-Variables Problem," working paper.

Delaigle, A., Hall, P., and Meister, A. (2008), "On Deconvolution With Repeated Measurements," *The Annals of Statistics*, 36, 665–685.

Devanarayan, V., and Stefanski, L. A. (2002), "Empirical Simulation Extrapolation for Measurement Error Models With Replicate Measurements," *Statistics and Probability Letters*, 59, 219–225.

Devroye, L. (1989), "Consistent Deconvolution in Density Estimation," *The Canadian Journal of Statistics*, 17, 235–239.

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

Fan, J., and Truong, Y. K. (1993), "Nonparametric Regression With Errors in Variables, *The Annals of Statistics*, 21, 1900–1925.

Hall, P. (1984), "Central Limit Theorem for Integrated Square Error of Multivariate Nonparametric Density Estimators," *Journal of Multivariate Analysis*, 14, 1–16.

Hall, P., and Meister, A. (2007), "A Ridge-Parameter Approach to Deconvolution," *The Annals of Statistics*, 35, 1535–1558.

Hesse, C. H. (1999), "Data-Driven Deconvolution," *Journal of Nonparametric Statistics*, 10, 343–373.

Kim, J., and Gleser, L. J. (2000), "SIMEX Approaches to Measurement Error in ROC Studies," *Computational Statistics—Theory and Methods*, 29, 2473–2491.

Kuchenhoff, H., and Carroll, R. J. (1997), "Segmented Regression With Errors in Predictors: Semi-Parametric and Parametric Methods," *Statistic in Medicine*, 16, 169–188.

Kuchenhoff, H., Mwalili, S. M., and Lesaffre, E. (2006), "A General Method for Dealing With Misclassification in Regression: The Misclassification SIMEX," *Biometrics*, 62, 85–96.

Li, Y., and Lin, X. H. (2003), "Functional Inference in Frailty Measurement Error Models for Clustered Survival Data Using the SIMEX Approach," *Journal of the American Statistical Association*, 98, 191–203.

Luo, X. H., Stefanski, L. A., and Boos, D. D. (2006), "Tuning Variable Selection Procedures by Adding Noise," *Technometrics*, 48, 165–175.

Staudenmayer, J., and Ruppert, D. (2004), "Local Polynomial Regression and Simulation-Extrapolation," *Journal of the Royal Statistical Society*, Ser. B, 66, 17–30.

Stefanski, L. A. (2000), "Measurement Error Models, *Journal of the American Statistical Association*, 95, 1353–1358.

Stefanski, L. A., and Bay, J. M. (1996), "Simulation Extrapolation Deconvolution of Finite Population Cumulative Distribution Function Estimators," *Biometrika*, 83, 407–417.

Stefanski, L. A., and Carroll, R. J. (1990), "Deconvoluting Kernel Density Estimators," *Statistics*, 21, 169–184.

Stefanski, L. A., and Cook, J. R. (1995), "Simulation-Extrapolation: The Measurement Error Jackknife," *Journal of the American Statistical Association* 90, 1247–1256.

Zhang, C. H. (1990), "Fourier Methods for Estimating Mixing Densities and Distributions," *The Annals of Statistics*, 18, 806–830.