NONPARAMETRIC KERNEL AND REGRESSION SPLINE ESTIMATION IN THE PRESENCE OF MEASUREMENT ERROR

J. D. Maca, R. J. Carroll and David Ruppert *

January 15, 1997

Abstract

In many regression applications both the independent and dependent variables are measured with error. When this happens, conventional parametric and nonparametric regression techniques are no longer valid. We consider two different nonparametric techniques: regression splines and kernel estimation, of which both can be used in the presence of measurement error. Within the kernel regression context, we derive the limit distribution of the SIMEX estimate. With the regression spline technique, two different methods of estimations are used. The first method is the SIMEX algorithm which attempts to estimate the bias, and remove it. The second method is a structural approach, where one hypothesizes a distribution for the independent variable which depends on estimable parameters. A series of examples and simulations illustrate the methods.

Key Words and Phrases: Asymptotic theory; Bandwidth Selection; Bootstrap; Estimating Equations; Local Polynomial Regression, Measurement Error; Nonlinear Regression; Regression Splines; Sandwich Estimation; SIMEX.

Short title: Nonparametric Regression with Measurement Error

*J. D. Maca is a graduate student and R. J. Carroll is Professor of Statistics, Nutrition and Toxicology, Department of Statistics, Texas A&M University, College Station, TX 77843–3143. D. Ruppert is Professor, School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY 14853–3801. Maca's and Carroll's research was supported by a grant from the National Cancer Institute (CA–57030). Carroll's research was partially completed while visiting the Institut für Statistik und Ökonometrie, Sonderforschungsbereich 373, Humboldt Universität zu Berlin, with partial support from a senior Alexander von Humboldt Foundation research award. Ruppert's research was supported by grants from the National Science Foundation and the National Security Agency. We wish to acknowledge with thanks conversations with Simon Sheather.

1 INTRODUCTION

We consider the problem of nonparametric regression function estimation in the presence of measurement error in the predictor. Suppose that the regression of a response Y on a predictor X is given by $E(Y|X) = m(X)$. Instead of observing X, we can only observe W, an error-prone measurement related to X by an additive error model $W = X + U$, where U is a mean-zero normal random variable with variance σ_u^2. The question is: how can we estimate $m(\cdot)$ when observations on Y and W are all that are available?

This problem has been addressed previously, most notably by Fan & Truong (1993), who found the following discouraging result. Suppose that we allow $m(\cdot)$ to have up to k derivatives. They showed that if the measurement error was normally distributed, even with known error variance, then based on a sample of size n, no consistent nonparametric estimator of $m(\cdot)$ converges faster than the rate $\{\log(n)\}^k$. Since, for example, $\log(10,000,000) \approx 16$, effectively this result might be interpreted to say that consistent nonparametric regression function estimation in the presence of measurement error is impractical.

The Fan & Truong result can be interpreted in two ways. The first is pessimistic: nonparametric regression in the presence of measurement error is insolvable in practice. The second, and positive, interpretation focuses on the phrase "in practice". As reviewed by Carroll, Ruppert & Stefanski (1995), much of the enormous progress made in the field of measurement error for nonlinear models has been through the use of *approximately* consistent estimators, i.e., estimators which correct for most of measurement error induced bias, but not all. Practically, these classes of estimators do an effective job of removing such bias. Theoretically, for small errors ($\sigma_u^2 \to 0$), the bias of naive estimators is of the order $\mathcal{O}(\sigma_u^2)$, while the approximate error correctors have a bias of order $\mathcal{O}(\sigma_u^6)$ or less.

A second positive interpretation is to remember that the Fan & Truong result pertains to *globally* consistent estimation, i.e., estimators of $E(Y|X)$ which are consistent without anything but smoothness assumptions. Such results say nothing about estimators which are consistent for a flexible yet parametric subclass of the nonparametric family. For example, regression splines are a well-known parametric family with the capability of estimating wide classes of regression functions (although not all functions). It stands to reason that if one is willing to estimate $E(Y|X)$ by a regression spline, then effective semiparametric estimation of $E(Y|X)$ is possible even in the presence of measurement error.

This paper develops the two ideas of approximately consistent and regression spline estimation in the presence of measurement error. In Section 2 we show how to implement the SIMEX method (Cook & Stefanski, 1994; Carroll, Ruppert, and Stefanski 1995; Stefanski & Cook, 1995; Carroll, Küchenhoff, Lombard & Stefanski, 1996) in ordinary nonparametric regression, while Section 3 develops this idea for regression splines. The SIMEX method is a functional method, i.e., one that can be applied without estimation of the distribution of the unobservable X. In Section 4, we take up the structural approach in the context of regression splines, showing that the observed data follow a type of regression spline depending on the conditional distribution of X given W. If W given X is normally distributed, X given W depends on the marginal distribution of X, which we fit flexibly by a mixture of normal distributions with an unknown number of mixtures. This flexible distribution is fit by modifying the Gibbs sampling algorithm of Wasserman & Roeder (1997). Section 5 gives a number of numerical examples and simulations. Section 6 has concluding remarks.

While the discussion to follow is easiest in the case that the measurement error variance σ_u^2 is known, in practice this is not the case. In some instances, σ_u^2 is estimated by an external data set. Otherwise, internal replicates are used, so that we observe $W_{ij} = X_i + U_{ij}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, K_i \geq 1$, where the measurement errors (U_{ij}) are independent, mean zero, normally distributed random variables with variance σ_u^2; a components of variance estimate is given as equation (3.2) in Carroll, Ruppert, and Stefanski (1995). In theory, for either external or internal data, σ_u^2 is estimated at ordinary parametric rates $\mathcal{O}_P(n^{-1/2})$, and so the asymptotic effect of such estimation on nonparametric regression functions is often nil.

2 THE SIMEX ESTIMATOR

The SIMEX estimator was developed by Cook & Stefanski (1994), see Carroll, Küchenhoff, Lombard & Stefanski (1996) and Stefanski & Cook (1995) for related theory, and Carroll, Ruppert, and Ruppert (1995) for detailed discussion of implementation. We confine our discussion here to local linear kernel regression, although the methods are easily extended to higher order polynomial regression.

First consider the case that number of replicates $m_i = 1$ and that σ_u is known. Fix $B > 0$ to be a large but finite integer (50–200 in practice), and consider estimation of $E(Y|X)$ at x_0. For $b = 1, \ldots, B$ and any $\lambda > 0$, let $(\varepsilon_{ib})_1^n$ be a set of independent standard normal random variables which are then transformed to have sample mean zero, variance one and to be uncorrelated with the Y's and the W's. Define $W_{ib}(\lambda) = W_i + \sigma_u \lambda^{1/2} \varepsilon_{ib}$. For a kernel density function $K(\cdot)$ and bandwidth h, define $K_h(u) = h^{-1}K(u/h)$. Local linear kernel estimates solve the weighted least squares equation in $\mathcal{B} = (\beta_0, \beta_1)^t$,

$$0 = \sum_{i=1}^{n} \Big[Y_i - G_2^t \left\{ W_{ib}(\lambda) - x_0 \right\} \mathcal{B} \Big] G_2 \left\{ W_{ib}(\lambda) - x_0 \right\} K_h \left\{ W_{ib}(\lambda) - x_0 \right\}, \tag{1}$$

where $G_2(v) = (1, v)^t$. The kernel estimate is $\widehat{m}_{b,\lambda}(x_0, h) = \widehat{\beta}_0$. In general, one must estimate h as well, and we do this using EBBS (Empirical Bias Bandwidth Selection), see Ruppert (1997). The resulting implemented estimate is $\widehat{m}_{b,\lambda}(x_0)$. The average of these estimates over $b = 1, \ldots, B$ is $\widehat{m}_\lambda(x_0)$.

The SIMEX estimator is then defined by a three step process: (a) select a finite set of λ's such as $\lambda = 0$, 1/2, 1, 3/2, 2 and compute $\widehat{m}_\lambda(x_0)$; (b) fit a convenient function of λ, such as a quadratic, to the terms $\widehat{m}_\lambda(x_0)$; (c) extrapolate this fit back to $\lambda = -1$, resulting in $\widehat{m}(x_0)$. Some asymptotic distribution theory is derived in the appendix; in particular the limiting values of bias and variance are derived for a quadratic extrapolant.

When σ_u is unknown, it is replaced by an estimate. If the number of replicates is constant, $(\kappa_i \equiv \kappa)$ then $W_{ib}(\lambda) = \overline{W}_{i\cdot} + (\sigma_u \kappa_i^{-1/2} \lambda^{1/2} \varepsilon_{ib})$. The only remaining step is how to handle the case that the number of replicates is not constant. Carroll, et al. (1995) use the definition of $W_{ib}(\lambda)$ given immediately above, but there is a theoretical difficulty, namely $E\left(Y | \overline{W}_{i\cdot}, \kappa_i = 1\right)$ $\neq E(Y | \overline{W}_{i\cdot}, \kappa_i = 2)$. This causes some problems of theory and even more of notation because if we define $m_\lambda(x_0, \kappa_i) = E(Y | \overline{W}_{i\cdot}, \kappa_i)$, then the naive kernel regression which ignores measurement error converges to $n^{-1} \sum_{i=1}^{n} m_\lambda(x_0, \kappa_i)$, which is a mixture of regression functions depending on the design. Despite this technical complication, the results derived in the appendix extend immediately.

Finally, we note that the results in the appendix show an interesting feature, namely that the bias and variance of the SIMEX estimator depends only on the bias and variance of the naive estimator at $\lambda = 0$. The bias contribution to the SIMEX estimator from the naive estimator is proportional to h_0^2, and the variance contribution is a delta-method derivable linear function of the variance of the naive estimator. In principle, at least, one can select h_0 to obtain a good SIMEX estimator, and not merely a good naive estimator. However, the best method of bandwidth selection for the SIMEX method is an open problem.

3 REGRESSION SPLINES AND SIMEX

We write the regression spline of order p and with ℓ knots $(\xi_1, \ldots, \xi_\ell)$ as

$$m_{p\ell}(x; \boldsymbol{\beta}) = \sum_{j=0}^{p} \beta_j x^j + \sum_{j=1}^{\ell} \beta_{p+j} (x - \xi_j)_+^p, \tag{2}$$

where $v_+ = vI(v > 0)$, and $I(\cdot)$ is the indicator function. If the number of knots and the knots themselves are fixed, then fitting (2) to error-prone data is simply a parametric problem, to which the SIMEX idea applies. Extrapolation can take one of two forms: (a) direct application of the SIMEX algorithm requires that one extrapolate the coefficients back to $\lambda = -1$ and then announce the resulting function; and (b) for each fixed x, extrapolate the fitted function $m_{p\ell}\{x; \widehat{\boldsymbol{\beta}}(\lambda)\}$ back to $\lambda = -1$. Both methods have something to recommend to them. With either option, the fixed knot selection method can be implemented extremely quickly using an idea of Ruppert & Carroll (1996). Here one uses a large number of knots, and then obtains smoothness by a type of ridge regression and C_p, see the appendix for details.

Alternatively, one may allow either or both of the number of knots ℓ or their locations to vary with each of the $1 + B(C-1)$ data sets formed by C values of λ, including zero, and $b = 1, \ldots, B$ simulations in SIMEX. Here one would clearly use option (b), because the meaning of β would vary for each of the data sets formed by combinations of (b, λ). This variable knot size-location method would appear to have an advantage over the fixed knot method in that for each set of computer-generated "data", one is in some sense optimizing to the data at hand. One has to be aware of a weakness of this approach, besides the fact that it is not at all clear that each data set is much better fit this way than by our fixed knot method. The issue here is computation. If $C = 5$ and $B = 100$, with variable knot size and selection, the computing time required to implement knot size-location is at least 400 times that of a single fit, which may become prohibitive, especially if the bootstrap is used to form confidence intervals.

4 STRUCTURAL APPROACH TO REGRESSION SPLINES

Structural estimation in measurement error models means that one hypothesizes a distribution for X depending on a parameter Θ. Since W given X is normal with variance σ_u^2, (σ_u, Θ) together produce the conditional distribution of X given W. Thus, if Y given X has mean determined by the spline (2), Y given W has mean

$$E(Y|W) = \sum_{j=0}^{p} \beta_j E(X^j|W) + \sum_{j=1}^{\ell} \beta_{p+j} E\left\{(X - \xi_j)_+^p |W\right\}. \tag{3}$$

Under a parametric model for X given W, all the conditional expectations in (3) are easily calculated numerically, and the β's can be estimated by ordinary least squares, or ridge regression (see Section 7.1 in the Appendix)

The asymptotic distributions of the parameter estimates and fitted values are also easily obtained. Estimation of (σ_u, Θ) can be done based only on the W's themselves, by solving an equation of the form $0 = \sum_1^n \psi_1(W_i, \sigma_u, \Theta)$. If we define

$$\begin{aligned} G(W_i, \sigma_u, \Theta) &= \Big[1, E(X|W, \Theta, \sigma_u), \ldots, E(X^p|W, \Theta, \sigma_u), \\ &\qquad E\left\{(X-\xi_1)_+^p|W, \Theta, \sigma_u\right\}, \ldots, E\left\{(X-\xi_\ell)_+^p|W, \Theta, \sigma_u\right\}\Big]^t, \end{aligned}$$

then we estimate $\mathcal{B}$ by solving $0 = \sum_1^n \psi_2(Y_i, W_i, \Theta, \sigma_u, \mathcal{B})$, where

$$\psi_2(Y, W\Theta, \sigma_u, \mathcal{V}) = \left\{Y - G^t(W, \sigma_u, \Theta)\mathcal{B}\right\} G(W, \sigma_u, \Theta).$$

It follows from estimating equation methods (Carroll, Ruppert, and Stefanski, 1995) that $n^{1/2}\left(\widehat{\sigma}_u - \sigma_u, \widehat{\Theta} - \Theta, \widehat{\mathcal{B}} - \mathcal{B}\right)$ is asymptotically normally distributed with mean zero and variance $A^{-1}B(A^{-1})^t$, where

$$A = E\begin{bmatrix} \dfrac{\partial}{\partial\sigma_u}\psi_1 & \dfrac{\partial}{\partial\Theta^t}\psi_1 & 0 \\ \dfrac{\partial}{\partial\sigma_u}\psi_2 & \dfrac{\partial}{\partial\Theta^t}\psi_2 & -GG^t \end{bmatrix}; \qquad B = \begin{bmatrix} E\psi_1\psi_1^t & 0 \\ 0 & E\psi_2\psi_2^t \end{bmatrix},$$

since $E\psi_1\psi_2^t = E\left\{E\left(\psi_1\psi_2^t|W\right)\right\} = E\left\{\psi_1 E\left(\psi_2^t|W\right)\right\} = 0$ under the model. Of course, A and B can be estimated consistently by sample average of the terms within its expectation. In specific applications, better estimates may be obtainable. For example, if ψ_1 is the likelihood score for estimating (σ_u, Θ) from the W's, then $E\psi_1\psi_1^t = -E\left\{(\partial/\partial\sigma_u)\psi_1 \ \ (\partial/\partial\Theta^t)\psi_1\right\}$ is the Fisher information.

The remaining issue is to specify a distribution for X. The obvious one is the normal distribution in which case $W = X + U$ would be marginally normally distributed, so that the assumption of normal X can be checked empirically from the observed data. To build some model robustness, one could use instead a flexible parametric family which includes the normal distribution, e.g., the seminonparametric family of Davidian & Gallant (1993) or the mixture of normals family.

A mixture of k-normals has the means $\widetilde{\underline{\boldsymbol{\mu}}}_k = (\mu_{1k}, \ldots, \mu_{kk})$, standard deviations $\widetilde{\underline{\boldsymbol{\sigma}}}_k = (\sigma_{1k}, \ldots, \sigma_{kk})$ and proportions $\widetilde{\underline{\mathbf{p}}}_k = (p_{1k}, \ldots, p_{kk})$, where $\sum_{j=1}^k p_{jk} = 1$. When X is observable, Wasserman & Roeder (1997) propose a Bayesian method for estimating $(k, \widetilde{\underline{\boldsymbol{\mu}}}_k, \widetilde{\underline{\boldsymbol{\sigma}}}_k, \widetilde{\underline{\mathbf{p}}}_k)$ when k is constrained to lie in the set $1 \le k \le L$ for some fixed L. Here we modify their method to account for the measurement error. Suppose that we observe $W_{ij} = X_i + U_{ij}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, m_i$. Let σ_u have the inverse-chi prior density (A_u, r_u) where r_u is known,

$$[\sigma_u] \sim \frac{A_u^{r_u/2}}{2^{(r_u/2)-1}\Gamma(r_u/2)}\sigma_u^{-r_u-1}\exp\left\{-\frac{A_u}{2\sigma_u^2}\right\}.$$

Fix k. Let $\widetilde{\underline{\mathbf{W}}}$ consist of all the observed W's, $\widetilde{\underline{\mathbf{X}}}$ the latent X's, $\widetilde{\underline{\mathbf{G}}}_k$ the latent group assignment indicators $(G_{k1}, \ldots, G_{kn})$ telling from which of the k normal subpopulations $\widetilde{\underline{\mathbf{X}}}$ is drawn, $[A_k]$ be proportional to a scaling constant, and $[\widetilde{\underline{\boldsymbol{\mu}}}_k, \widetilde{\underline{\boldsymbol{\sigma}}}_k, \widetilde{\underline{\mathbf{p}}}_k]$ be the prior defined by Wasserman & Roeder.

The joint density for given k is

$$\begin{aligned}[\underline{\widetilde{\mathbf{W}}}, \underline{\widetilde{\mathbf{X}}}, \underline{\widetilde{\mathbf{G}}}_k, \sigma_u, A_k, \underline{\widetilde{\boldsymbol{\mu}}}_k, \underline{\widetilde{\boldsymbol{\sigma}}}_k, \underline{\widetilde{\mathbf{p}}}_k] \quad \sim \quad & [\underline{\widetilde{\mathbf{W}}} | \underline{\widetilde{\mathbf{X}}}, \sigma_u][\sigma_u][\underline{\widetilde{\mathbf{X}}} | \underline{\widetilde{\mathbf{G}}}_k, A_k, \underline{\widetilde{\boldsymbol{\mu}}}_k, \underline{\widetilde{\boldsymbol{\sigma}}}_k, \underline{\widetilde{\mathbf{p}}}_k] \\ & \times [\underline{\widetilde{\mathbf{G}}}_k | A_k, \underline{\widetilde{\boldsymbol{\mu}}}_k, \underline{\widetilde{\boldsymbol{\sigma}}}_k, \underline{\widetilde{\mathbf{p}}}_k][A_k, \underline{\widetilde{\boldsymbol{\mu}}}_k, \underline{\widetilde{\boldsymbol{\sigma}}}_k, \underline{\widetilde{\mathbf{p}}}_k]. \end{aligned} \tag{4}$$

Inspection of (4) reveals that the Gibbs sampler has an especially convenient form. Once one has generated the latent variables $\underline{\widetilde{\mathbf{X}}}$ and σ_u in a Gibbs step, the generation of $(\underline{\widetilde{\mathbf{G}}}_k, A_k, \underline{\widetilde{\boldsymbol{\mu}}}_k, \underline{\widetilde{\boldsymbol{\sigma}}}_k, \underline{\widetilde{\mathbf{p}}}_k)$ is exactly the same as if $\underline{\widetilde{\mathbf{X}}}$ were known and there were no measurement error, we can adapt without change the Gibbs steps derived by Wasserman & Roeder. The Gibbs steps for σ_u and $\underline{\widetilde{\mathbf{X}}}$ are also easy. One sees that σ_u given all the rest is inverse-chi with parameters $A_r + \sum_{i=1}^{n} \sum_{j=1}^{m_i} (W_{ij} - X_i)^2, r_u + \sum_{i=1}^{n} m_i$, while any X_i given all the rest and that $G_{ki} = j$ is normal with

$$\mu = \frac{W_{i\cdot}\sigma_j^2 + \mu_j \sigma_u^2}{m\sigma_j^2 + \sigma_u^2} \quad \text{and} \quad \sigma^2 = \frac{\sigma_u^2 \sigma_j^2}{m\sigma_j^2 + \sigma_u^2}, \quad \text{where} \quad W_{i\cdot} = \sum_{j=1}^{m} W_{ij}.$$

Following Wasserman & Roeder, having generated estimates of $\Theta_k = (\sigma_u, \underline{\widetilde{\boldsymbol{\mu}}}_k, \underline{\widetilde{\boldsymbol{\sigma}}}_k, \underline{\widetilde{\mathbf{p}}}_k)$ for given k, namely the median of the value $(\sigma_u, \underline{\widetilde{\boldsymbol{\mu}}}_k, \underline{\widetilde{\boldsymbol{\sigma}}}_k)$ and the mean of the values $\underline{\widetilde{\mathbf{p}}}_k$ in the Gibbs steps, we estimate the posterior probability that there are k mixtures as $n^{-3k/2} \ell \left(\widehat{\Theta}_k \right)$, where $\ell(\Theta_k)$ is the likelihood of $\underline{\widetilde{\mathbf{W}}}_k$ evaluated at the parameters Θ_k. This likelihood is

$$\ell\left(\Theta_k\right) = \prod_{i=1}^{n} \prod_{j=1}^{m_i} \sum_{k=1}^{l} p_k (\sigma_k^2 + \sigma_u^2)^{-1/2} (2\pi)^{-1/2} \exp \left\{ \frac{-(W_{ij} - \mu_k)^2}{2(\sigma_k^2 + \sigma_u^2)} \right\}.$$

We now return to (3). To implement this, we need the conditional distribution of X_i given $(W_{i1}, \ldots, W_{im_i})$ for $i = 1, \ldots, n$. When X is a mixture of k-normals, this conditional distribution is easily seen to be a mixture of k-normals with

j^{th} mean $= (\mu_j \sigma_u^2 + \overline{W} m \sigma_j^2)(\sigma_u^2 + m\sigma_j^2)^{-1}$;

j^{th} variance $= \sigma_j^2 \sigma_u^2 (\sigma_u^2 + m\sigma_j^2)^{-1}$;

j^{th} proportion $= p_j \left(\widetilde{\sigma}_j \sum_{i=1}^{k} \left[p_j \widetilde{\sigma}_i^{-1} \exp \left\{ \frac{-(\overline{W} - \mu_i)^2}{2\widetilde{\sigma}_i^2} \right\} \right] \right)^{-1} \exp \left\{ \frac{-(\overline{W} - \mu_j)^2}{2\widetilde{\sigma}_j^2} \right\}$;

where $\widetilde{\sigma}_j = (\sigma_j^2 + m^{-1}\sigma_u^2)^{1/2}$. If $(\widehat{\eta}_1, \ldots, \widehat{\eta}_L)$ are the estimated posterior probabilities formed from Gibbs sampling, we take X_i given $(W_{i1}, \ldots, W_{im_i})$ to be a mixture of the previously defined mixture normals, with mixing proportions $(\widehat{\eta}_1, \ldots \widehat{\eta}_L)$.

5 EXAMPLES AND SIMULATIONS

5.1 Simulations

In this section we present a few examples showing the improved estimating abilities of the Gibbs regression spline. For all simulations the average squared error (ASE) and average absolute error

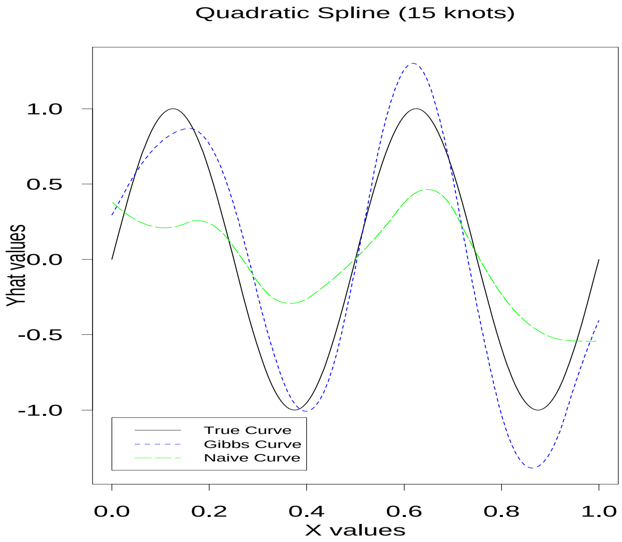

Figure 1: *Comparison of Gibbs spline to Naive spline*

(AAE) were computed for each type of spline to facilitate comparisons. The spline coefficients were found using the ridge regression method outlined in Section 7.1, using a quadratic order spline with 15 equally spaced knot points.

First, we compare the Gibbs spline to the naive spline. For this example, there were 500 data points, (X), which were generated from an Uniform(0,1) random variable. The true curve was then generated as $Y = \sin(4\pi X) + \epsilon$, where $\epsilon \sim N(0, .05^2)$. The measurement error was generated from the $N(0, .2^2)$ distribution, and there were two replications at each X point. The comparison of the naive spline to the Gibbs spline can be found in Figure 1. The Gibbs spline had an ASE and AAE of 0.064 and 0.220 respectively, which compares to the ASE and AAE values for the naive spline of 0.213 and 0.413.

The SIMEX estimator was found by the following algorithm. At each point in a grid of equally spaced points over the X range, $m_{p\ell}\{x; \widehat{\beta}(\lambda)\}$ was found. This was done for $B = 1000$ generated datasets, and the mean was recorded for each level of $\lambda = (0, .2, .4, .6, .8, 1)$. The SIMEX estimate for that point was then found by linearly and quadratically extrapolating back to $\lambda = -1$. These estimators will be referred to as SIMEX(L) and SIMEX(Q). For the SIMEX(Q) estimator, the ASE and AAE were found as 0.333 and 0.519 respectively, whereas SIMEX(L) estimator had ASE and AAE values of 0.239 and 0.441.

The second example compares the Gibbs spline to the SIMEX estimator. For this example, 500 X values were generated from a Beta$(5, 5)$ distribution. The true curve was then generated

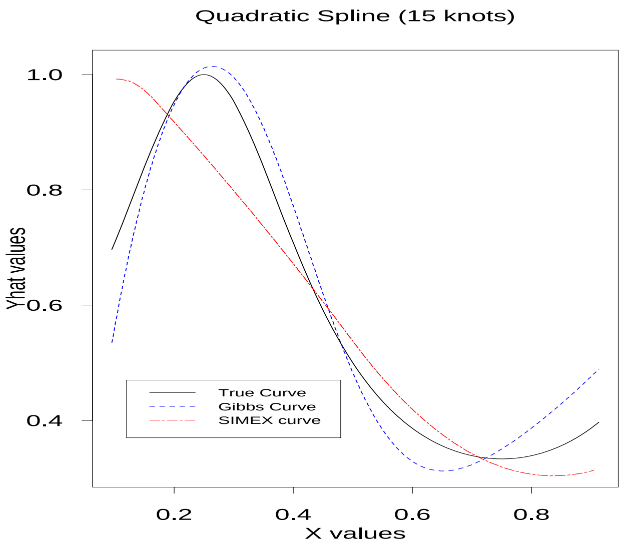

Figure 2: *Comparison of Gibbs spline to SIMEX(L) spline*

by $Y = \{2 - \sin(2\pi X)\}^{-1} + \epsilon$, where $\epsilon \sim N(0, .02^2)$. The measurement error was generated from the $N(0, .2^2)$ distribution, with two replications at each point. The Gibbs spline, and the SIMEX spline(L) produced with this data can be found in Figure 2. The Gibbs curve had an ASE and AAE of 0.00274 and 0.04391 respectively, whereas the SIMEX spline(L) had ASE and AAE of 0.00494 and 0.05113 respectively. The SIMEX(Q) spline was also found, and had ASE and AAE values of 0.01182 and 0.08903 respectively. All of these can be compared to the naive spline which had ASE and AAE values of 0.00964 and 0.08016.

5.2 Framingham Heart Study Data

For this example, we use non-simulated data to again see the flattening effects of measurement error. First let X be a person's true systolic blood pressure and Y be the person's true diastolic blood pressure. Then clearly when blood pressure measurements are taken, both X and Y are measured with error. The data for this example comes from the well known Framingham Heart Study. This dataset contains 1642 individuals who had repeated blood pressures measurements taken. The spline estimates for this dataset are found in Figure 3. As in the simulated datasets, the naive splines estimates tended to be pulled toward the null line.

5.3 Fan & Truong Simulation Comparison

In the final example, we compare the Gibbs and SIMEX spline to the deconvoluting kernel estimators found in Fan and Truong (1993). The notation from their paper has been changed to be

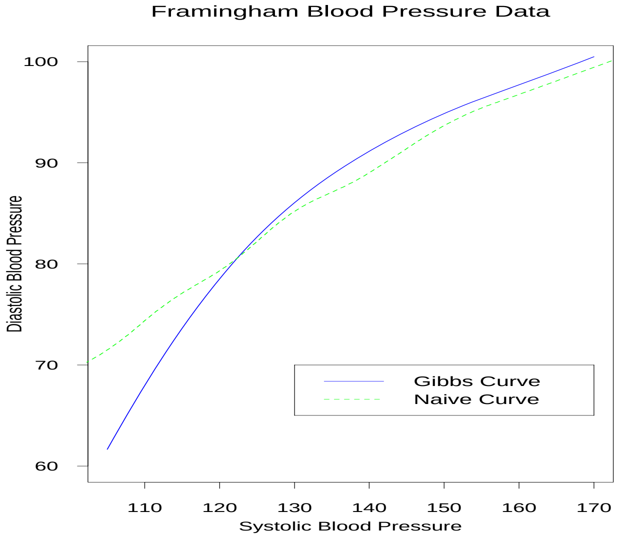

Figure 3: *Comparison of Gibbs spline to Naive spline for Blood pressure data*

consistent with the notation used in this paper. In their simulation they had $X \sim \text{Normal}(0.5, .25^2)$ and defined the measurement error variance to be $\sigma_0^2 = \dfrac{3}{7} \cdot \text{var}(X)$, and $W = X + \varepsilon$, where the measurement error was either $\varepsilon \sim N(0, \sigma_0^2)$, or, $\varepsilon \sim \text{LogNormal}(0, \sigma_0^2)$. Random samples $(W_1, Y_1), \ldots, (W_n, Y_n)$ were generated, where $Y = X_+^3(1-X)_+^3 + \epsilon, \quad \epsilon \sim N(0, .0015^2)$ or $Y = 1 + 4X + \epsilon, \quad \epsilon \sim N(0, .25^2)$. The Average Squared Error (ASE) was then computed over a grid of 101 equally spaced points from .1 to .9. This was done for sample sizes of $n = 200, n = 400, n = 800$ and for 3 different kernels. We will compare the SIMEX splines and Gibbs splines only to the deconvoluting kernel with non–data dependent bandwidth chosen to have the lowest ASE. The results from the first model can be found in Table 1; the results for the second model are similar. The quadratically interpolated SIMEX spline again performed poorly, but the linearly interpolated estimate, SIMEX(L), had a significant improvement over the estimators found by Fan and Truong. The Gibbs spline had an even further improvement to the SIMEX(L) estimate for the normal measurement error case. However, in the double exponential error case, the Gibbs spline tended to be undersmoothed, causing the poorer performance compared to the SIMEX(L) estimator although both still outperformed the estimators found by Fan and Truong.

6 DISCUSSION AND GENERALIZATIONS

This paper focuses on ordinary nonparametric regression estimation. However, it can be extended to the class of generalized linear models with mean $\mu\{m(x)\}$ and variance $\sigma^2 V\{m(x)\}$ with known functions $\mu(\cdot)$ and $V(\cdot)$. Here we provide a brief discussion: further details will be described in a future publication.

Nonparametric kernel regression in GLIM's was described by Fan, Heckman & Wand (1995). The bias and variance formulae are of a similar order of magnitude as in the ordinary regression case, and hence application of SIMEX should follow the same general outline as in Section 2. Ruppert & Carroll (1996) discuss regression spline estimation in GLIM's, and their methods can be combined directly with SIMEX as in Section 3.

The structural approach of Section 4 is more complicated in the GLIM context. Writing the spline as before as $m(x;\beta)$, we have

$$\begin{aligned} E(Y|W) &= E[\mu\left\{m(X;\beta)\right\}|W]; \\ \mathrm{var}(Y|W) &= \sigma^2 E[V\{m(X;\beta)\}|W] + \mathrm{var}[\mu\{m(X;\beta)\}|W]. \end{aligned}$$

Both the mean and variance functions are easily calculated numerically given a model for $[X|W]$. The parameter β can be estimated using quasilikelihood ideas; SIMEX provides starting values. Selection of knots is more complex, and would follow the traditional AIC criterion, or an extension of the ridge regression method.

We have assumed without comment that $W = X + U$, with U normally distributed and having mean zero. In fact, for purposes of (nearly) nonparametric estimation, it suffices merely that some monotone transformation of originally observed W's follow this additive error model, i.e., $g(W) = g(X) + U$, because if $g(\cdot)$ is any strictly monotone function, $E(Y|X = x_0) = E\left\{Y|g(X) = g(x_0)\right\}$. See Nusser, Carriquiry, Dodd & Fuller (1996) and Eckert & Carroll (1997) for such methods of transformation; the former differs from the latter in requiring that $g(X)$ also be normally distributed.

REFERENCES

Carroll, R. J., Küchenhoff, H., Lombard, F., & Stefanski, L. A. (1996). Asymptotics for the SIMEX estimator in structural measurement error models. *Journal of the American Statistical Association*, 91, 242–250.

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

Carroll, R. J., Ruppert, D. & Welsh, A. H. (1997) Nonparametric estimation via local estimating equations, with applications to nutrition calibration. Preprint.

Cook, J. R. & Stefanski L. A. (1994) Simulation–extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, 89, 1314-1328.

Davidian, M. & Gallant, R. A. (1993) The nonlinear mixed effects model with a smooth random effects density. *Biometrika*, 80, 475- 488.

Eckert, R. S. & Carroll, R. J. (1997). Transformations to additivity in measurement error models. *Biometrics*, to appear.

Fan, J., Heckman, N. E., and Wand, M. P. (1995). Local Polynomial Kernel Regression for Generalized Linear Models and Quasi-Likelihood Functions. *Journal of the American Statistical Association*, 90, 141–150.

Fan, J. & Truong, Y. K. (1993). Nonparametric regression with errors in variables. *Annals of Statistics*, 21, 1900–1925.

Nusser, S. M., Carriquiry, A. L., Dodd, K. W. & Fuller, W. A. (1995). A semiparametric transformation approach to estimating usual intake distributions. *Journal of the American Statistical Association*, to appear.

Ruppert, D. (1997). Empirical-Bias Bandwidths for Local Polynomial Nonparametric Regression and Density Estimation. *Journal of the American Statistical Association*, to appear.

Ruppert, D. & Carroll, R. J. (1996). A simple roughness penalty approach to regression spline estimation. Preprint.

Ruppert, D. & Wand, M. P. (1994). Multivariate Locally Weighted Least Squares Regression. *Annals of Statistics*, 22, 1346–1370.

Stefanski, L. A., & Cook, J. R. (1995). Simulation–Extrapolation: The measurement error jackknife. *Journal of the American Statistical Association*, 90, 1247–1256.

Wasserman, L. & Roeder, K. (1997). Bayesian Density Estimation Using Mixtures of Normals. *Journal of the American Statistical Association*, to appear.

7 APPENDIX

7.1 Roughness Penalty Approach to Regression Spline Estimation

Here we briefly review the work of Ruppert & Carroll (1996). Suppose that we have data (X_i, Y_i) where X_i is univariate, $Y_i = m(X_i) + \epsilon_i$, and m is a smooth function. To estimate m we let $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p, \beta_{p+1}, \ldots, \beta_{p+k})$ and use a regression spline model

$$m(x; \boldsymbol{\beta}) = \beta_0 + \beta_1 x + \cdots + \beta_p x^p + \sum_{j=1}^{\ell} \beta_{p+j}(x - \xi_j)_+^p.$$

where $p \geq 1$ is an integer and $\xi_1 < \ldots < \xi_\ell$ are fixed knots. The traditional method of "smoothing" the estimate is through knot selection. Ruppert & Carroll (1996) use a different approach by allowing ℓ to be large and using a roughness penalty on $\{\beta_{p+j}\}_{j=1}^{\ell}$ which is the set of jumps in the pth derivative of $m(x; \boldsymbol{\beta})$. They use this as a penalty on the $(p+1)$th derivative of $m(x; \boldsymbol{\beta})$ where that derivative is a generalized function. They recommend ℓ between 10 and 20 and letting k_j be the j/ℓ th sample quantile of the X_i's.

Define $\hat{\boldsymbol{\beta}}(\alpha)$ to be the minimizer of

$$\sum_{i=1}^{n}\left(Y_i - [\beta_0 + \beta_1 x + \cdots + \beta_p x^p + \sum_{j=1}^{\ell}\beta_{p+j}(x-k_j)_+^p]\right)^2 + \alpha\sum_{j=1}^{\ell}\beta_{p+j}^2.$$

Let X be the "design matrix" for the regression spline and let D be a diagonal matrix whose first $(1+p)$ diagonal elements are 0 and whose remaining diagonal elements are 1. Then simple calculations show that $\hat{\boldsymbol{\beta}}(\alpha)$ is given by

$$\hat{\boldsymbol{\beta}}(\alpha) = \left(X^T X + \alpha D\right)^{-1} X^T Y \tag{5}$$

This is a ridge regression estimator that shrinks the regression spline towards the least-squares fit to a pth degree polynomial model, with the amount of shrinkage determined by the smoothing parameter α.

Computing (5) is extremely quick, even for a relatively large number, say 30, values of α. This allows rapid selection of α by C_p, or perhaps by GCV. Here we look at C_p. Let

$$\mathrm{ASR}(\alpha) = n^{-1}\sum_{i=1}^{n}\{Y_i - m(X_i;\boldsymbol{\beta}(\alpha))\}^2$$

be the average squared residuals using α. Let $S(\alpha) = X\left(X^T X + \alpha D\right)^{-1} X^T$ be the "smoother" or "hat" matrix. Then, let α^* be a small value of α implying little smoothing. Then

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}\{Y_i - m(X_i;\boldsymbol{\beta}(\alpha^*))\}^2}{n - \mathrm{tr}[2S(\alpha^*) - S(\alpha^*)^2]}.$$

is a nearly unbiased estimator of the variance of the ϵ_i's. Finally,

$$C_p(\alpha) = \mathrm{ASR}(\alpha) + 2\mathrm{tr}(S(\alpha))\hat{\sigma}^2/n$$

is the C_p statistic. We choose α by computing $C_p(\alpha)$ for a grid of α values and choosing the minimizer of C_p.

7.2 SIMEX Estimate in Kernel Regression

Let $f_\lambda(\cdot)$ be the density function of $W + \lambda^{1/2}\sigma_u\varepsilon$, so that if $f_W(\cdot)$ is the density of W,

$$f_\lambda(x_0) = \int \left(\lambda^{1/2}\sigma_u^{1/2}\right)^{-1} f_W(z)\varphi\left\{(x_0 - z)/(\lambda^{1/2}\sigma_u)\right\}dz.$$

Let $m_\lambda(x_0) = E\left\{Y|W + \lambda^{1/2}\sigma_u\varepsilon = x_0\right\}$. Implicit in the work of Fan (1993) and Ruppert & Wand (1994) and as explicitly derived by Carroll, Ruppert & Welsh (1997) is the expansion for any fixed b that as $h \to 0$ and $nh \to \infty$ (assuming that K is scaled so that $\int z^2 K(z)dz = 1$),

$$\widehat{m}_{b,\lambda}(x_0,h) - m_\lambda(x_0) - (h^2/2)m_\lambda^{(2)}(x_0) \approx \{nf_\lambda(x_0)\}^{-1}\sum_{i=1}^{n}\left[Y_i - m_\lambda\left\{W_{ib}(\lambda)\right\}\right]K_h\left\{W_{ib}(\lambda) - x_0\right\},$$

where the error is of order $o_p\left\{h^2 + (nh)^{-1/2}\right\}$.

In what follows, it is convenient notationally to use the same bandwidth h for every $b = 1, \ldots, B$, but to allow this bandwidth to depend on λ, hence h_λ. Of course, in practice one might estimate h for each λ and b, but as $n \to \infty$ the error in estimating this bandwidth becomes negligible, and hence asymptotically the same bandwidths are being used. As stated in the text, the best method of bandwidth selection for the SIMEX method remains an open problem.

Using the decomposition of Carroll, et al. (1996), since B is fixed, and since $\widehat{m}_\lambda(x_0, h) = B^{-1}\sum_{b=1}^{B} \widehat{m}_{b,\lambda}(x_0, h_\lambda)$,

$$\begin{aligned} &\widehat{m}_\lambda(x_0, h_\lambda) - m_\lambda(x_0) - \left(h_\lambda^2/2\right) m_\lambda^{(2)}(x_0) \\ &\approx \{n f_\lambda(x_0)\}^{-1} \sum_{i=1}^{n} \left(B^{-1} \sum_{b=1}^{B} \left[Y_i - m_\lambda\{W_{ib}(\lambda)\}\right] K_{h_\lambda}\{W_{ib}(\lambda) - x_0\}\right). \end{aligned} \tag{6}$$

In what follows, we will use the following slight abuse of notation. We will write expressions for moments of $\widehat{m}_\lambda(x_0, h_\lambda)$, but these will actually apply to the asymptotically equivalent version on the right side of (6). The terms inside the parentheses on the right hand side of (6) are independent mean zero random variables. Letting $\widetilde{\underline{\mathbf{Y}}} = (Y_1, \ldots, Y_n)$ and $\widetilde{\underline{\mathbf{W}}} = (W_1, \ldots, W_n)$, and using right hand side of (6) as equivalent to the left side, consider

$$\begin{aligned} \operatorname{var}\{\widehat{m}_\lambda(x_0, h_\lambda)\} \quad \approx \quad & E\left[\operatorname{var}\left\{\widehat{m}_\lambda(x_0, h_\lambda) | \widetilde{\underline{\mathbf{Y}}}, \widetilde{\underline{\mathbf{W}}}\right\}\right] \\ & + \operatorname{var}\left[E\left\{\widehat{m}_\lambda(x_0, h_\lambda) - m_\lambda(x_0) - (h_\lambda^2/2) m_\lambda^{(2)}(x_0) | \widetilde{\underline{\mathbf{Y}}}, \widetilde{\underline{\mathbf{W}}}\right\}\right]. \end{aligned} \tag{7}$$

If $\lambda = 0$ or if $\sigma_u^2 = 0$, then $m_0(x_0) = E(Y|W = x_0)$, $f_0(x_0) = f_W(x_0)$, and (6) becomes

$$\widehat{m}_0(x_0, h_0) - m_0(x_0) - (h_0^2/2) m_0^{(2)}(x_0) \approx \{n f_0(x_0)\}^{-1} \sum_{1=i}^{n} \{Y_i - m_0(W_i)\} K_{h_0}(W_i - x_0),$$

which has mean zero and asymptotic variance

$$\{n h_0 f_0(x_0)\}^{-1} \ \operatorname{Var}(Y|W = x_0) \int K^2(v) dv. \tag{8}$$

If $\lambda > 0$ and $\sigma_u^2 > 0$, we study the terms of (7) in turn. For the first, note that given $\widetilde{\underline{\mathbf{Y}}}$ and $\widetilde{\underline{\mathbf{W}}}$, the only remaining random variables are the (ε_{ib}), which are all mutually independent. Hence

$$\begin{aligned} &\operatorname{var}\left\{\widehat{m}_\lambda(x_0, h_\lambda) | \widetilde{\underline{\mathbf{Y}}}, \widetilde{\underline{\mathbf{W}}}\right\} \\ &\approx \left\{nB f_\lambda^2(x_0)\right\}^{-1} n^{-1} \sum_{i=1}^{n} \operatorname{var}\left[\left\{Y_i - m_\lambda(W_i + \sigma_u \lambda^{1/2}\varepsilon)\right\} K_{h_\lambda}(W_i + \sigma_u \lambda^{1/2}\varepsilon - x_0) | Y_i, W_i\right] \\ &= \left\{nB f_\lambda^2(x_0)\right\}^{-1} n^{-1} \sum_{i=1}^{n} \int \left\{Y_i - m_\lambda\left(W_i + \sigma_u \lambda^{1/2}\varepsilon\right)\right\}^2 K_{h_\lambda}^2\left(W_i + \sigma_u \lambda^{1/2}\varepsilon - x_0\right) \varphi(\varepsilon) d\varepsilon \\ &- \left\{nB f_\lambda^2(x_0)\right\}^{-1} n^{-1} \sum_{i=1}^{n} \left[\int \left\{Y_i - m_\lambda(W_i + \sigma_u \lambda^{1/2}\varepsilon)\right\} K_{h_\lambda}\left(W_i + \sigma_u \lambda^{1/2}\varepsilon - x_0\right) \varphi(\varepsilon) d\varepsilon\right]^2. \end{aligned}$$

Setting $z = (W_i + \sigma_u \lambda^{1/2}\varepsilon - x_0)/h_\lambda$ so that $W_i + \sigma_u \lambda^{1/2}\varepsilon = x_0 + z h_\lambda$ and $\varepsilon = (x_0 + z h_\lambda - W_i)/\sigma_u \lambda^{1/2}$, we compute the two terms as

$$\begin{aligned} &= \left\{n h_\lambda B f_\lambda^2(x_0) \sigma_u \lambda^{1/2}\right\}^{-1} n^{-1} \sum_{i=1}^{n} \int \{Y_i - m_\lambda(x_0 + z h_\lambda)\}^2 K^2(z) \varphi\left(\frac{z h_\lambda + x_0 - W_i}{\sigma_u \lambda^{1/2}}\right) dz \\ &- \left\{nB f_\lambda^2(x_0) \sigma_u^2 \lambda\right\}^{-1} n^{-1} \sum_{i=1}^{n} \left[\int \{Y_i - m_\lambda(x_0 + z h_\lambda)\} K(z) \varphi\left(\frac{z h_\lambda + x_0 - W_i}{\sigma_u \lambda^{1/2}}\right) dz\right]^2. \end{aligned}$$

The second term is $\mathcal{O}(n^{-1})$, so we are left with

$$\begin{aligned}&\operatorname{var}\left\{\widehat{m}_\lambda(x_0,h)|\underline{\widetilde{\mathbf{Y}}},\underline{\widetilde{\mathbf{W}}}\right\} \approx \\ &\left\{nh_\lambda B f_\lambda^2(x_0)\sigma_u\lambda^{1/2}\right\}^{-1}\left\{\int K^2(z)dz\right\} n^{-1}\sum_{i=1}^n \{Y_i - m_\lambda(x_0)\}^2 \varphi\left(\frac{W_i - x_0}{\sigma_u\lambda^{1/2}}\right). \end{aligned} \tag{9}$$

Note the curious fact that there is a B in the denominator. This means that if B is large, (9) is small in comparison to what happens when $\lambda = 0$, see (8). In fact, as $B \to \infty$, (6) converges to

$$\{nf_\lambda(x_0)\}^{-1}\sum_{i=1}^n E\left[\left\{Y_i - m_\lambda\left(W_i + \sigma_u\lambda^{1/2}\varepsilon\right)\right\} K_{h_\lambda}\left(W_i + \sigma_u\lambda^{1/2}\varepsilon\right)\middle| Y_i, W_i\right], \tag{10}$$

and this random variable has zero variance given $(\underline{\widetilde{\mathbf{Y}}},\underline{\widetilde{\mathbf{W}}})$, just as predicted by (9).

We next turn to the second term in (7). Continuing to assume that $\lambda > 0$ and $\sigma_u^2 > 0$, the expectation in question is just

$$\begin{aligned} &\{nf_\lambda(x_0)\}^{-1}\sum_{i=1}^n \int \left\{Y_i - m_\lambda\left(W_i + \sigma_u\lambda^{1/2}\varepsilon\right)\right\} K_{h_\lambda}\left(W_i + \sigma_u\lambda^{1/2}\varepsilon - x_0\right)\varphi(\varepsilon)d\varepsilon \qquad (11)\\ &= \{nf_\lambda(x_0)\}^{-1}\sum_{i=1}^n \int \{Y_i - m_\lambda(x_0 + zh_\lambda)\} K(z)\varphi\left(\frac{zh_\lambda + x_0 - W_i}{\sigma_u\lambda^{1/2}}\right)\left(\sigma_u\lambda^{1/2}\right)^{-1}dz \\ &\approx \{nf_\lambda(x_0)\}^{-1}\sum_{i=1}^n \{Y_i - m_\lambda(x_0)\}\varphi\left(\frac{W_i - x_0}{\sigma_u\lambda^{1/2}}\right)\left(\sigma_u\lambda^{1/2}\right)^{-1}, \end{aligned}$$

which has variance of order $\mathcal{O}(n^{-1})$. We have thus shown that for $\lambda > 0$, $\sigma_u^2 > 0$,

$$\operatorname{var}\{\widehat{m}_\lambda(x_0,h)\} = \mathcal{O}\left\{(nhB)^{-1}\right\} + \mathcal{O}(n^{-1}). \tag{12}$$

It is important to note that the second term in (7) is $\mathcal{O}\left(n^{-1}\right)$ only when $\lambda > 0$ and $\sigma_u^2 > 0$. If either equal 0, the expectation calculated above is

$$\{nf_0(x_0)\}^{-1}\sum_{i=1}^n \{Y_i - m_0(W_i)\} K_{h_0}(W_i - x_0),$$

which has mean zero and variance $\mathcal{O}\left\{(nh)^{-1}\right\}$. The difference is that when $\lambda > 0$ and $\sigma_u^2 > 0$, (12) represents a "double-smooth", i.e., summation and integration, and it is well-known that double smoothing increases rates of convergence.

If we compare (8) with (12), we note that for n and B sufficiently large, the latter will be negligible with respect to the former, at least in practice. Hence, in what follows, we will ignore this variability by treating B as if it were equal to infinity. This makes the analysis of the SIMEX extrapolants easy. For example, if one is fitting a quadratic model, one minimizes in $(\alpha_0, \alpha_1, \alpha_2)$, $\sum_\lambda \{\widehat{m}_\lambda(x_0, h_\lambda) - \alpha_0 - \alpha_1\lambda - \alpha_2\lambda^2\}^2$, and thus solves

$$0 = \sum_\lambda \left\{\widehat{m}_\lambda(x_0,h_\lambda) - \alpha_0 - \alpha_1\lambda - \alpha_2\lambda^2\right\}\left(1,\lambda,\lambda^2\right)^t.$$

Using standard least-squares results, we get

$$\begin{pmatrix}\widehat{\alpha}_0 - \alpha_0\\ \widehat{\alpha}_1 - \alpha_1\\ \widehat{\alpha}_2 - \alpha_2\end{pmatrix} = \left\{\sum_\lambda \left(1,\lambda,\lambda^2\right)^t (1,\lambda,\lambda^2)\right\}^{-1}\sum_\lambda \left\{\widehat{m}_\lambda(x_0,h_\lambda) - \alpha_0 - \alpha_1\lambda - \alpha_2\lambda^2\right\}\left(1,\lambda,\lambda^2\right)^t. \tag{13}$$

Assuming the terms $m_\lambda(x_0)$ actually follow a quadratic, this means that the left-hand-side of (13) has approximate mean

$$\left\{\sum_\lambda \left(1,\lambda,\lambda^2\right)^t(1,\lambda,\lambda^2)\right\}^{-1}\sum_\lambda \left(h_\lambda^2/2\right)m_\lambda^{(2)}(x_0)(1,\lambda,\lambda^2)^t,$$

and, because B is large, its approximate variance is

$$\begin{aligned}&\{nh_\lambda f_0(x_0)\}^{-1}\,\mathrm{var}(Y|W=x_0)\int K^2(z)dz\\ &\times\left\{\sum_\lambda(1,\lambda,\lambda^2)^t(1,\lambda,\lambda^2)\right\}^{-1}(1,0,0)^t(1,0,0)\left\{\sum_\lambda(1,\lambda,\lambda^2)^t(1,\lambda,\lambda^2)\right\}^{-1}.\end{aligned}$$

The SIMEX estimate is just $\widehat{\alpha}_0-\widehat{\alpha}_1+\widehat{\alpha}_2=(1,-1,1)\left(\widehat{\alpha}_0,\widehat{\alpha}_1,\widehat{\alpha}_2\right)^t$ so that its asymptotic bias is

$$c(x_0)^t\sum_\lambda\left(h_\lambda^2/2\right)m_\lambda^{(2)}(x_0)(1,\lambda,\lambda^2)^t$$

and its asymptotic variance is

$$\{nh_0f_0(x_0)\}^{-1}\,\mathrm{var}(Y|W=x_0)\int K^2(z)dz\;c^t(x_0)(1,0,0)^t(1,0,0)c(x_0),$$

where

$$c^t(x_0)=(1,-1,1)\left\{\sum_\lambda(1,\lambda,\lambda^2)^t(1,\lambda,\lambda^2)\right\}^{-1}.$$

It is clearly easy to derive the bias and variance for any extrapolation function.

Table 1: ASE($\times 10^{-6}$) for estimating $m(x) = x_+^3(1-x)_+^3$

Size	Measure	Gibbs	Simex(L)	Simex(Q)	Fan&Truong
	Normal Measurement Error				
200	Bias^2	0.121	0.560	0.139	
	90%$\|\hat{m} - m\|$	2312.562	1934.057	3073.203	
	ASE	2.764	5.399	10.961	7.99
400	Bias^2	0.009	0.558	0.238	
	90%$\|\hat{m} - m\|$	1734.504	955.660	1641.664	
	ASE	1.511	4.153	10.358	6.72
800	Bias^2	0.037	0.492	0.181	
	90%$\|\hat{m} - m\|$	1232.777	499.026	827.847	
	ASE	0.777	4.084	10.062	5.86
	Double Exponential Measurement Error				
200	Bias^2	0.190	0.243	0.066	
	90%$\|\hat{m} - m\|$	2458.087	1061.776	2439.172	
	ASE	2.736	2.203	8.272	4.08
400	Bias^2	0.141	0.333	0.117	
	90%$\|\hat{m} - m\|$	2157.324	429.897	1313.959	
	ASE	2.257	1.653	7.177	2.92
800	Bias^2	0.289	0.218	0.052	
	90%$\|\hat{m} - m\|$	2366.121	205.467	660.400	
	ASE	2.720	1.165	6.621	2.30