Statistica Sinica **28** (2018), 2289-2307
doi:https://doi.org/10.5705/ss.202016.0393

WAVELET METHODS FOR ERRATIC REGRESSION MEANS IN THE PRESENCE OF MEASUREMENT ERROR

Peter Hall[1], Spiridon Penev[2] and Jason Tran[1]

[1] *University of Melbourne and* [2] *University of New South Wales, Sydney*

Abstract: In nonparametric regression with errors in the explanatory variable, the regression function is typically assumed to be smooth, and in particular not to have a rapidly changing derivative. Not all data applications have this property. When the property fails, conventional techniques, usually based on kernel methods, have unsatisfactory performance. We suggest an adaptive, wavelet-based approach, founded on the concept of explained sum of squares, and using matrix regularisation to reduce noise. This non-standard technique is used because conventional wavelet methods fail to estimate wavelet coefficients consistently in the presence of measurement error. We assume that the measurement error distribution is known. Our approach enjoys very good performance, especially when the regression function is erratic. Pronounced maxima and minima are recovered more accurately than when using conventional methods that tend to flatten peaks and troughs. We also show that wavelet techniques have advantages when estimating conventional, smooth functions since they require less sophisticated smoothing parameter choice. That problem is particularly challenging in the setting of measurement error. A data example is discussed and a simulation study is presented.

Key words and phrases: Chirp, cross-validation, deconvolution, discontinuity, errors in variables, error sum of squares, explained sum of squares, kernel methods.

1. Introduction

This paper was initiated in July 2013 when the second author was visiting Professor Peter Hall at The University of Melbourne as part of his special study leave. Most of the theoretical treatment was performed during this time and Peter's contribution was instrumental to the initiation of the paper. In particular, the idea to use the explained sum of squares approach which is fundamental for this paper, was Peter's. Since the numerical implementation, testing and simulation were challenging and time consuming, it took a while to get a manuscript that could be submitted for publication. Although theoretical and numerical results were presented at some seminars and conferences, the manuscript was never published until Peter's tragic and untimely death. Jason Tran was a PhD student of Peter's and contributed to proving the main statement of the paper by essentially simplifying some of the original assumptions. The current manuscript is a tribute to Peter's contributions to the methodology of measurement error models.

In some regression problems the explanatory variable, X, is not observed directly. Instead we know the value, $W = X + U$, of X corrupted by an error, U, incurred when measuring X. In this setting the observed data are pairs (W_i, Y_i) for $1 \leq i \leq n$, all distributed as (W, Y), where Y denotes the response when the explanatory variable is X :

$$W = X + U, \quad Y = g(X) + V. \tag{1.1}$$

Here g represents the regression mean $g(x) = E(Y|X = x), V$ represents experimental error, satisfying $E(V|X) = 0$, and we wish to estimate g. There is a large and extensive literature on estimating g in this setting when that function is relatively smooth. In particular, the existing methodology typically uses smoothing methods, for example modified kernel estimators, the performance of which deteriorates as g becomes more erratic. In this paper we suggest relatively adaptive methods using wavelet techniques and based on minimising an "explained sum of squares" (ESS) in errors-in-variables regression. This nonstandard technique is motivated by the difficulty of estimating wavelet coefficients in the presence of measurement error. In particular, those coefficients cannot be estimated using conventional arguments.

Our ESS approach can also be used in a variety of other settings and, in particular, it is straightforward to employ for non-wavelet techniques based on orthogonal series. In the wavelet context it enjoys very good performance, particularly in cases where g is more erratic than encompassed by the standard assumption that g has several bounded derivatives. For example, if g has a relatively pronounced peak or trough then our wavelet-based approach generally recovers that feature more accurately than does a conventional approach.

The existing methodology dates from work of Carroll and Hall (1988) and Stefanski and Carroll (1990). An excellent introduction to the literature is given in the monograph by Carrol et al. (2006). More recent contributions include those of Delaigle and Meister (2007), Maity and Apanasovich (2011), and Delaigle, Hall and Jamshidi (2015). The last paper deals for the first time with constructing confidence bands for the regression function.

An interesting approach to constructing an estimator of the regression function g has been suggested in Comte and Taupin (2007). To avoid the unrealistic situation in the usual Nadaraya-Watson construction, whereby it is assumed that g and the density f_X of the random variable X belong to the same smoothness class, the authors relax this assumption by applying a construction of an adaptive estimator as a ratio of penalized contrast function-based estimators of the product of gf_X and of the density f_X itself. However the behaviour of their estimator is driven by the slowest rate of the two estimators and it remains unclear how optimal such estimator can be. Hence they raise the question of constructing a good estimator that is different from the Nadaraya-Watson type estimator.

We will introduce our methodology in Section 2. There are, as yet, no competing approaches to constructing wavelet estimators in errors-in-variables regression. Arguably the closest existing technique has been developed in Chesneau (2010), who proposed methodology in the case where X is known to be uniformly distributed on a specified interval. In this case one can use "periodised" wavelet functions, and both practical implementation and theoretical justification are relatively straightforward. Chesneau (2010) focuses on cases where g is relatively smooth, and in fact much of the motivation for his approach derives from the fact that it can be readily used to estimate derivatives of g, as well as of g itself. The context of our work is quite different: the distribution of X is unknown, and is unlikely to be uniform, and the function g is relatively erratic and might not have several derivatives at all points in its support.

2. Methodology

2.1. Overview of wavelet expansions and regression estimators

The classical monograph Daubechies (1992) can be consulted for a rigorous introduction of the wavelet theory as an important tool in approximation theory. For statistical and inferential aspects of wavelets, including software implementations in the popular **R** software, see Nason (2008). The wavelet expansion of a function g is given by

$$g(u) = \sum_j \alpha_j^0 \phi_j(u) + \sum_{k=0}^{\infty} \sum_j \alpha_{jk}^0 \psi_{jk}(u), \tag{2.1}$$

where $\phi_j(u) = \rho^{1/2}\phi(\rho u - j)$ and $\psi_{jk}(u) = \rho_k^{1/2}\psi(\rho_k u - j)$; ϕ and ψ denote compactly supported "father" and "mother" wavelet functions, respectively; $\rho \geq 1$, a positive number potentially depending on n, is sometimes called the primary resolution level (see, e.g., Hall and Penev (2001)); and $\rho_k = 2^k\rho$ for positive integers k. In (2.1), and in similar formulae below, $\sum_j$ denotes summation over all positive and negative integers j. However, since φ and ψ in our methodology are always assumed to be compactly supported, all but a finite number of terms in the infinite series in (2.1) vanish. Sometimes in the literature, j is used to denote the scale and k is used as a location index.

Our goal is to estimate $g(u)$ for values of u in the compact interval $[a, b]$ supporting f_X and g.

Using a conventional hard thresholding rule to ensure adaptivity, one would estimate g by the function $\hat{g}(u) = \hat{g}_{t,\rho,m}(u)$:

$$\hat{g}(u) = \sum_j \hat{\alpha}_j \phi_j(u) + \sum_{k=0}^{m} \sum_j \hat{\alpha}_{jk} I(|\hat{\alpha}_{jk}| > t\hat{\sigma}_{jk}) \psi_{jk}(u), \tag{2.2}$$

where $\hat{\alpha}_j$ and $\hat{\alpha}_{jk}$ estimate α_j and α_{jk}, respectively; $I(|\hat{\alpha}_{jk}| > t\hat{\sigma}_{jk}) = 1$ if $|\hat{\alpha}_{jk}| > t\hat{\sigma}_{jk}$, and equals 0 otherwise; $\hat{\sigma}_{jk}$ is an estimator of the variance of $\hat{\alpha}_{jk}$; and the threshold, $t > 0$, is a tuning parameter. Wavelet estimators are generally robust against choice of m, often taken to be a constant multiple of $\log n$ and, in particular, usually is not chosen specifically from the data. The value of $\hat{\sigma}_{jk}$ can be computed using bootstrap methods, although it is often adequate to employ an upper bound. Theoretical arguments (Donoho and Johnstone (1994)) suggest taking $t = (C \log n)^{1/2}$ where C, for example $C = 2$, is a positive constant.

We take an unconventional approach, motivated by the presence of error in measurements of the explanatory variables. Specifically, we first estimate α_j^0 and α_{jk}^0 directly, using an argument based on explained (or error) sum of squares; see Section 2.3. Next we fit the model at (2.2), but with the thresholding term $I(|\hat{\alpha}_{jk}| > t\hat{\sigma}_{jk})$ replaced by 1. Then we use matrix regularisation to reduce noise; see Section 2.4 for details. The resulting estimator, $\hat{g}$, is

$$\hat{g}(u) = \sum_j \hat{\alpha}_j \phi_j(u) + \sum_{k=0}^{m} \sum_j \hat{\alpha}_{jk} \psi_{jk}(u), \tag{2.3}$$

rather than (2.2). We have conducted numerical experiments using explicit regularisation, as at (2.2), and found that its performance is generally similar to, but on average slightly inferior to, matrix regularisation.

2.2. Regression estimation in the measurement error model

In the case of regression with measurement error, the observed data pairs (W_i, Y_i), for $1 \leq i \leq n$, are generated by the model at (1.1), where U denotes a measurement error, V is an experimental error and satisfies $E(V) = 0$, and the variables U, V and X are completely independent of one another. The distribution of U is assumed known. Since this assumption seems strong, different attempts have been made in the past to alleviate it. The literature on such methods is vast. Rather than listing some of it, we refer to Section 1 of the recent paper Delaigle and Hall (2016), and to the reference list therein. This paper demonstrates, via a completely new approach, that identification in the deconvolution problem can in principle be achieved even when the distribution of U is unknown (but symmetric) and without extra data of any type. The paper Delaigle and Hall (2016) deals with density deconvolution but, as the authors point out, the methodology can be extended to regression with meeasurement error. However, we will continue using the assumption of a known distribution of U in the rest of this paper.

The methodology in Section 2.3 will rely on estimators $\hat{f}_W$ and $\hat{f}_X$ of the densities f_W and f_X of W and X, respectively. Construction of those quantities is straightforward. Indeed, since we observe data $W_1, \ldots, W_n$ then $\hat{f}_W$ can be computed very conventionally:

$$\hat{f}_W(w) = \frac{1}{nh_1} \sum_{i=1}^{n} L\left(\frac{w - W_i}{h_1}\right),$$

where L is a standard univariate kernel, which we take to be a bounded, symmetric, univariate probability density.

The deconvolution kernel estimator $\hat{f}_X(x)$ has been introduced first by Carroll and Hall (1988) and by Stefanski and Carroll (1990). Let K be a bounded, integrable function on the real line, satisfying $\int_{-\infty}^{\infty} K(x)dx = 1$, and let K^{Ft} denote the Fourier transform of $K : K^{Ft}(t) = \int_{-\infty}^{\infty} \exp(itu)K(u)du$. Let f_U^{Ft} be the characteristic function corresponding to the density f_U of U, let $h_2 > 0$ be a bandwidth, and define

$$K_U(u) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp(-itu) \frac{K^{Ft}(t)}{f_U^{Ft}(t/h_2)} dt,$$

where the notation i, only when appearing in an exponent, is $\sqrt{-1}$ rather than an index. The deconvolution kernel estimator of f_X is given by

$$\hat{f}_X(x) = \frac{1}{nh_2} \sum_{i=1}^{n} K_U\left(\frac{x - W_i}{h_2}\right)$$

for $x \in [a, b]$.

2.3. Explained sum of squares

As

$$E\{g(X)|W=w\} = \frac{1}{f_W(w)} \int_a^b g(x) f_U(w-x) f_X(x) dx, \tag{2.4}$$

we deduce that if g is given by (2.1) then

$$E\{g(X)|W=w\} = \frac{1}{f_W(w)} \left\{ \sum_j \alpha_j^0 \int_a^b \phi_j(x) f_U(w-x) f_X(x) dx \right. \\ \left. + \sum_{k=0}^{\infty} \sum_j \alpha_{jk}^0 \int_a^b \psi_{jk}(x) f_U(w-x) f_X(x) dx \right\}. \tag{2.5}$$

Formula (2.5) motivates computing $\hat{\alpha}_j$ and $\hat{\alpha}_{jk}$ by minimising the explained sum of squares,

$$S(\alpha) = \sum_{i=1}^n \left[Y_i - \frac{1}{\hat{f}_W(W_i)} \left\{ \sum_j \alpha_j \int_a^b \phi_j(x) f_U(W_i - x) \hat{f}_X(x) dx \right. \right. \\ \left. \left. + \sum_{k=0}^{m} \sum_j \alpha_{jk} \int_a^b \psi_{jk}(x) f_U(W_i - x) \hat{f}_X(x) dx \right\} \right]^2 w_i, \tag{2.6}$$

where $\hat{f}_W$ and $\hat{f}_X$ are estimators of f_W and f_X, respectively. Here w_i denotes a nonnegative weight and α, a vector of finite length p, say, denotes the concatenation of values of α_j for all j, and α_{jk}, the latter only for $1 \le k \le m$ but for all j. The right-hand side of (2.6) becomes a little simpler, and numerically more robust, if we take $w_i = \hat{f}_W(W_i)^2 v_i$, where v_i is a bounded, nonnegative weight, which gives:

$$S(\alpha) = \sum_{i=1}^n \left\{ Y_i \hat{f}_W(W_i) - \sum_j \alpha_j \int \phi_j(x) f_U(W_i - x) \hat{f}_X(x) dx \right. \\ \left. - \sum_{k=0}^{m} \sum_j \alpha_{jk} \int \psi_{jk}(x) f_U(W_i - x) \hat{f}_X(x) dx \right\}^2 v_i. \tag{2.7}$$

Since only a finite number of values α_j, α_{jk} are nonzero, and those values are identified from the support of the wavelet, then, prior to regularisation, computing $\hat{\alpha} = \text{argmin}_{\alpha} S(\alpha)$ is a matter of matrix inversion.

2.4. Regularisation

Taking each $v_i = 1$ for simplicity, we write (2.7) as

$$S(\alpha) = \|Z - A\alpha\|^2,$$

where Z is an n-vector with components $Y_i, \hat{f}_W(W_i)$ and A, a function of the data, is an $n \times p$ matrix. Thus, $A = Q_1 \Lambda Q_2^T$ where Q_1 and Q_2 are, respectively, $n \times n$ and $p \times p$ orthogonal matrices, and Λ is a diagonal matrix with components $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{\min(n,p)} \geq 0$ down the main diagonal. The minimum-norm value of $\hat{\alpha} = \mathrm{argmin}_\alpha S(\alpha)$ is

$$\hat{\alpha} = A^- Z = Q_2 \Lambda^- Q_1^T Z, \tag{2.8}$$

where A^- and Λ^- denote Moore-Penrose inverses and, in particular, $\Lambda^- = D(p_1)$, with p_1 denoting the largest $j \leq p$ such that $\lambda_j > 0$, and where, for $q \leq p_1$, we define

$$D(q) = \mathrm{diag}(\lambda_1^{-1}, \ldots, \lambda_q^{-1}, 0, \ldots, 0).$$

We regularise by choosing $q \leq p_1$ to minimise a cross-validation estimator of mean squared prediction error, and then we take

$$\hat{\alpha} = \hat{\alpha}(q) = Q_2 D(q) Q_1^T Z, \tag{2.9}$$

in place of the value at (2.8). Our estimator of g is then defined as at (2.3), although with $\hat{\alpha}$ given by (2.9), and depends on the q chosen largest eigenvalues of Λ.

To construct the cross-validation criterion, let $\hat{f}_{W,-i}, \hat{f}_{X,-i}$ and $\hat{g}_{-i}$ denote the versions of $\hat{f}_W, \hat{f}_X$ and $\hat{g}$ computed from the dataset $\{(W_1, Y_1), \ldots, (W_n, Y_n)\}$ excluding the ith pair. Let $\mathcal{I}$ be a subset of the set $\{1, \ldots, n\}$ for which $\hat{f}_W(W_i)$, or perhaps $\hat{f}_{W,-i}(W_i)$, is bounded above a given positive constant. Reflecting on (2.4), define

$$T(q) = \sum_{i \in \mathcal{I}} \left\{ Y_i - \frac{1}{\hat{f}_{W,-i}} \int \hat{g}_{-i}(x) f_U(W_i - x) \hat{f}_{X,-i}(x) dx \right\}^2. \tag{2.10}$$

Alternatively we could multiply the ith summand on the right-hand side of (2.10) by $\hat{f}_{W,-i}(W_i)^2$, and take the sum over $i = 1, \ldots, n$. The quantity $\hat{g}_{-i}$, on the right-hand side of (2.10), depends on q, although this is suppressed in the notation. That dependence is the source of the dependence on q of the left-hand side of (2.10). We choose $q = \hat{q}$ to minimise $T(q)$. In numerical practice we have found that replacing $\hat{f}_{W,-i}$ and $\hat{f}_{X,-i}$ by their non leave-one-out variants, $\hat{f}_W$ and $\hat{f}_X$, usually has negligible effect on results. Of course, it produces significant savings in computation time.

3. Computation and Numerical Examples

3.1. Computing wavelet and kernel estimators, and tuning parameters

Our numerical experiments were performed in FORTRAN 95. We employed compactly supported wavelets from the Daubechies family, with extremal phase and highest number of vanishing moments compatible with the width of the wavelets support (see Daubechies (1992, p.195)). A parameter N determines the length of the support for both father and mother wavelets as $2N-1$. The smoothness (regularity) of these wavelets increases with N. We chose $N=5$ throughout. This, in our experience, represents a compromise between smoothness and length of the compact support for this family. Although these wavelets do not have an explicit analytic formula, they can be approximated arbitrarily accurately via some refinement schemes.

Smoothing parameters involved in the definition of $S(\alpha)$ at (2.6), include the initial resolution ρ and the bandwidths h_1 and h_2 used to construct the density approximations $\hat{f}_W$ and $\hat{f}_X$, respectively. We chose h_1 using the iterative bandwidth selector proposed by Engel, Herrmann and Gasser (1994), after experimenting with other methods. Although h_2 ideally should be chosen larger than h_1, we found it to be convenient, and to give good results, if we took $h_1 = h_2$.

Some theoretical work by Hall and Penev (2001) has been done for the choice of ρ in the setting of wavelet regression without measurement error. In principle, a practical choice could be to include ρ itself as a part of the cross-validation functional of Section 2.4, and to also optimize it with respect to ρ. A much less computationally demanding alternative is to choose ρ from a certain range of values and then to compensate for a slight inaccuracy via a choice of the value of q through the cross-validation procedure. This approach works well because the choice of q turns out to be by far the most important choice of a tuning parameter in our model and, in that sense, the precise choice of ρ is only of secondary importance. Using experimentation we found that $m=5$, and ρ in the range from 3 to 10, gave excellent results for n between 100 and 1,000, and we recommend these choices in applications. We took $\rho=3$ throughout the simulations reported in Section 3.2, and employed cross-validation to select q; see (2.10).

In Sections 3.2 and 3.3 the results obtained using our approach are compared with those found by conventional errors-in-variables regression estimator of Nadaraya-Watson type, first proposed in Fan and Truong (1993), with bandwidth chosen using the SIMEX method of Delaigle and Hall (2008). Intuitively, if the noise-to-signal ratio, or nsr, for the measurement error is small then it could simply be ignored without having much impact on the resulting estimators. Taking the same approach when the nsr is larger, however, would be expected to cause serious damage to the fit. Our method, which takes the influence of the measurement error in its stride, should perform better in such cases, especially for more complex signals, as long as the sample size is large enough for model parameters to be estimated sufficiently accurately. To assess this hypothesis we also constructed standard kernel and wavelet estimators without acknowledging that noise is present. We found that the standard conventional wavelet estimator that simply disregards the measurement error in the input variable, performed poorly in our simulation. The empirical wavelet coefficients in this case need to be evaluated using observations on an irregular grid, and that is known to cause difficulties for the wavelet estimator (see, e.g., Hall and Turlach (1997); Cai and Brown (1998)).

By now, there are some remedies for such cases, based on the second-generation wavelets (or lifting) that have performed well in empirical work. The original lifting approach of W. Sweldens was exploited in a series of papers (see, e.g., Claypoole et al. (2003); Delouille, Simoens and von Sachs (2004); Jansen, Nason and Silverman (2009); Nunes, Knight and Nason (2006) and the references therein). However, the difficulties are further exacerbated in our case due to the fact that the input variables are not uniformly distributed, and that their values are observed inexactly due to the presence of measurement error. As a result the wavelet coefficients were evaluated inaccurately, which degraded the quality of the estimated curve. For this reason we do not include results obtained using the conventional wavelet estimator.

The standard kernel regression estimator of Nadaraya-Watson type that does not acknowledge the noise U performed noticeably better, and we briefly discuss the performance of this method in our simulation comparison in Section 3.2. To make the comparison as fair as possible, we employed a particularly reliable method for kernel regression, with a sophisticated tuning parameter choice in this case. In particular, we borrowed the procedure described theoretically in Chapter 7 of the monograph by Müller (1988). We used a second-order minimum variance kernel, discussed in Chapter 5 of Müller (1988), for the estimation step, and applied cross-validation with a choice of 20 different widths, utilising routines in Chapter 12 of Müller (1988). In this way a fair comparison, described in the next section, was implemented using three methods: Our new wavelet technique in the presence of measurement error (WAVERR), the kernel method in the presence of measurement error (KERERR), and conventional kernel regression ignoring measurement error.

3.2. Simulation study

We took the variance of U to be $0.05, 0.10$ or 0.15 times that of X; these values represent nsr. The kernel K, in Section 2.2, was the standard normal density. Sample sizes used predominantly were $n = 300, 600$ or 900, which are relatively small compared with those in conventional statistical applications of wavelets, where typically $n = 512, 1{,}024, 2{,}048, 4{,}096$ or more. We shall discuss results for smaller and larger sample sizes in text. We first report results for model 1, where $g(x) = 5\sin(2x)\exp(-16x^2/50)$, and model 2, where $g(x) = -3\cos^2\{-x-(\pi/15)\}$ if $x < 0$, equals 0 if $x = 0$, and equals $3\cos^2\{x-(\pi/15)\}$ if $x > 0$. In particular, model 1 is particularly smooth but has increasingly small oscillations, of constant wavelength, in the tails, whereas model 2 has a marked discontinuity at the origin. In each case the distributions of X and V were both normal $N(0, 1.5)$, and the distribution of the measurement error, U, was Laplace.

The discontinuity in model 2 could in principle be handled with kernel methods. For example, in the error-free case, if we knew that there was a discontinuity, we could estimate the discontinuity and make adjustments to the estimator taking the discontinuity into account (see e.g., Gijbels, Hall and Kneip (1999)). The problem seems not to have been studied in the measurement error literature for the regression case until now. In any case, the essential advantage of the wavelet-based procedure we propose for estimating erratic regression means, is that the adjustment to the discontinuity happens in an automatic way.

Remark 1. Our choice of the kernel K as the standard normal density is possible, and justified in view of the choice of the measurement error model to be Laplace distributed, (the ordinary smooth family of distributions (Fan and Truong (1993))). It would not be possible to choose K as above if, for example, the measurement error was also normally distributed (super smooth in Fan and Truong's sense), because then the integral in the definition of the deconvolution kernel K_U would not exist if n was large enough.

For each combination of model and sample size, we repeated the data generating process and the curve fitting steps 101 times for the two methods WAVERR and KERERR, and calculated the corresponding integrated squared error,

$$ISE = \int_{-2}^{2} \{\hat{g}(x) - g(x)\}^2 \, dx,$$

Table 1. Values of 10^3 times the median of 101 simulated values of ISE (integrated squared error) for the WAVERR and KERERR estimators, for noise-to-signal ratios 0.05, 0.10 and 0.15 and sample sizes $n = 150, 300, 600, 900$ and 1,200 in cases where g is the smooth function in model 1, or the function with a pronounced discontinuity in model 2.

nsr		0.05		0.10		0.15	
model	n	WAVERR	KERERR	WAVERR	KERERR	WAVERR	KERERR
1	150	4.97	3.98	8.63	7.64	15.60	11.58
1	300	3.33	4.24	6.27	6.32	13.04	8.10
1	600	2.20	2.93	3.71	5.07	7.53	6.70
1	900	1.84	2.94	3.39	4.80	6.14	5.90
1	1,200	1.41	2.39	2.67	3.56	5.24	5.25
2	150	6.48	7.07	10.18	10.60	18.40	10.81
2	300	5.47	6.88	8.52	8.75	12.98	9.58
2	600	4.87	5.44	7.07	7.32	9.91	8.27
2	900	4.25	5.35	5.59	6.79	7.57	8.17
2	1,200	3.99	5.24	5.52	6.77	7.46	7.80

of the fit for each method. Quantities proportional to medians of the 101 ISE values are reported in Table 1. From Table 1 in the cases nsr = 0.05 and nsr = 0.10, and for each model and each sample size, the median value of ISE is less for the wavelet based estimator than for its kernel counterpart. In the setting of model 1, the wavelet estimator operates for smooth functions as though it were a kernel estimator with a kernel of slightly higher order than the actual kernel method that we employed. For smaller sample sizes the relationship is reversed; for example, when $n = 150$ and nsr = 0.05, the median values of ISE for WAVERR and KERERR are, respectively, 4.97 and 3.98, reflecting the relatively poor performance of WAVERR for smaller sample sizes, discussed below.

The trend is also observed for model 2, where the wavelet estimator outperforms its kernel counterpart for all sample sizes and for nsr = 0.05 and nsr = 0.10. The regression mean in model 2 has a jump discontinuity, and the pattern of performance in this setting reflects the known advantages of using wavelet methods in such instances. It can be shown that, in this case, even when there is no measurement error, the expected value of ISE for the kernel estimator and an optimal choice of bandwidth is of size no larger than a constant multiple of $n^{-1/2}$, and cannot be reduced by using higher order kernels.

To get a visual idea about the reasons behind the better overall performance of the wavelet-based procedure as illustrated in the table, we present graphs of "typical" estimated curves when using the three estimation methods for each

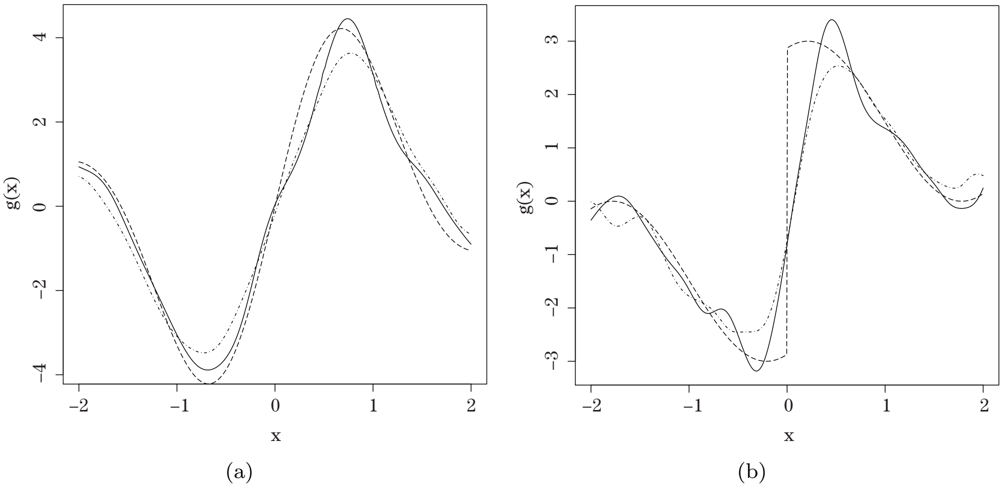

Figure 1. Left panel: Estimators of a smooth curve (model 1). Right panel: estimators of a discontinuous curve (model 2). Typical (median ISE) curves, $n = 300$, noise-to-signal ratio 0.05. True model: longdashed line, WAVEERR estimator: solid line, KERERR estimator: dot-dashed line.

of the chosen sample sizes. We define as typical the fits that produced the median ISE values for the WAVEERR method out of the 101 simulated functions. The main reason for the better performance of the WAVEERR method is that, as a wavelet-based procedure, WAVEERR allows one to zoom in better into discontinuities, peaks and troughs. Figure 1 illustrates an overlay of the fits when the noise-to-signal ratio was 0.05.

The relatively poor performance of the wavelet approach for relatively high nsr and small sample sizes, as seen in Table 1 when nsr = 0.15, can be seen too for sufficiently small sample sizes, smaller than those addressed by Table 1. The consistent pattern is that for each value of nsr, and each model, there is a sample size, n_{nsr} say, above which the kernel method is outperformed by the wavelet approach, and below which the kernel method gives best performance and the ratio of median ISE for the wavelet method, to median ISE for the kernel method, decreases as n increases. The value of n_{nsr} increases with nsr. Any improvements that the wavelet approach has to offer arise partly through the qualitatively different appearance of wavelet estimators, rather than just through reduced values of integrated squared error. This is seen in the context of model 2, where the wavelet estimator tracks the jump discontinuity relatively closely. It is more subtle, however, for relatively smooth models such as model 1. In those settings, particularly when turning points in the function g are relatively narrow, even though they have many bounded derivatives, the wavelet estimator reaches more deeply into the peak or trough.

Table 1 compares WAVERR and KERERR. In an extended version of this paper, we made a thorough comparison with the Nadaraya-Watson conventional regression method that completely ignores the measurement error (we call this method REGNOERR). The results for the same sample sizes and the same nsr, as expected, were worse than those in Table 1, and we do not include them. Some numbers for the median ISE for model 2 in the case of REGNOERR are listed for comparison: when $nsr = 0.05 : 5.71$ and 5.45 for $n = 600$ and $n = 900$; when $nsr = 0.10 : 8.08$ and 7.91 for $n = 600$ and $n = 900$; when $nsr = 0.15 : 10.06$ and 9.78 for $n = 600$ and $n = 900$. These results are worse than the related results in Table 1 and the median ISEs virtually do not change when the sample size is increased from 600 to 900, which indicates a non-consistency of the REGNOERR method. This is well-known in the measurement error literature.

We have also experimented, in an extended version of this paper, with the regression function $g(x) = 3\sin\left\{2\pi(x + 0.1x^2)\right\}$ that has the features of a chirp. Our simulations show clearly that the wavelet estimator does a much better job of reaching into peaks and troughs of the true regression mean. Here we also assessed the performance of the standard kernel estimator when no allowance was made for measurement error. Even when nsr = 0.05, there was always an increase in median ISE if we did not allow for measurement error. Of course, improvements in performance were greater for larger values of nsr.

3.3. Application to GDP data

Gross Domestic Product (GDP) is recorded quarterly and is prone to significant error (Fixler (2009)). It represents a sum of several (seven in the case of the USA) components, each of which is measured with error. Sources of error are discussed by Young (1994), and on web pages of the US Bureau of Economic Analysis. The average absolute values of the revisions are reported, typically with an error of about 2%. Economic theory suggests that there is a long-term association between economic activity and stock prices. Additionally, the Dow Jones Industrial Average (DJIA) is an average of the price of 30 of the largest and most widely traded US stocks, and so it should be interpretable as a noisy function of stock prices.

With this in mind we examined a scatterplot of the Dow Jones Industrial Average (DJIA) index, scaled by the factor 1,000, and the seasonally adjusted US Real GDP, in trillions of 2009 US dollars. The data are available from the

US Federal Reserve at http://research.stlouisfed.org/fred2/series/. We extracted quarterly data from January 1 1973 to January 1 2015, a sample size of 169. The year 1973 is motivated as the starting point since in that year the US decoupled completely the value of the US dollar and the gold standard. The relationship between GDP and DJIA was much less volatile prior to 1973, and so wavelet methods are not as well motivated in that time period.

Work of Young (1994) suggests taking the standard deviation, σ, of the Laplace distribution to equal 0.2, which we did. The fit is stable with respect to values of σ in the range 0.15 to 0.25. The graph in Figure 2 compares two fits, obtained using conventional deconvolution kernel and wavelet methods, respectively. Bandwidths for the conventional method were chosen using SIMEX. The parameters involved in the wavelet fit were $\rho = 5, m = 4$ and $N = 5$, and q was determined by cross-validation. Historical records, for example during financial crises in 2002, 2003 and 2009, demonstrate sharp falls in DJIA despite the relative stability of GDP figures. These falls influenced significantly the regression means at certain GDP values. However, the conventional estimator is not able to estimate these erratic values accurately. As Figure 2 shows, the wavelet estimator adapts better.

4. Theoretical Properties

In this section we show that our estimator $\hat{g}$, for an empirically chosen set of wavelet coefficients, is consistent for g. Let $[a, b]$ be the support of f_X, and (c, d) be a bounded interval containing $[a, b]$. *Assumption* 1(c) below, asserts that $\int_{[a,b]} |dg(x)| < \infty$ where the integral denotes the total variation of g. Let $B_1 > 0$ denote the finite value of the latter integral. We put $B_2 = \sup_{x\in[a,b]} |g(x)|$ and, given $B_3 > \max(B_1, B_2)$ and an integer $m \geq 1$, define $\mathcal{A}_m$ to be the set of vectors α of potential wavelet coefficients such that (4.1) and (4.2) hold:

$$g(x|\alpha, m) = 0 \text{ for all } x \notin [c, d], \tag{4.1}$$

$$\sup_{x\in[c,d]} |g(x|\alpha, m)| \leq B_3, \quad \int_{[c,d]} |dg(x|\alpha, m)| \leq B_3. \tag{4.2}$$

We have simplified notation by not indicating the dependence of $\mathcal{A}_m$ on B_3.

From Section 2, ϕ and ψ are compactly supported. We assume too that either they are Haar wavelets, or they are differentiable and hence of bounded variation. Then, if α^0 is taken to be the true sequence of wavelet coefficients α_j^0 for $j \in \mathbb{N}$ and α_{jk}^0 for $k \in \{0, \ldots, m\}$, $j \in \mathbb{N}$ in (2.1), conditions (4.1) and (4.2) both hold for all sufficiently large m. Therefore $\mathcal{A}_m$ is nonempty for such m.

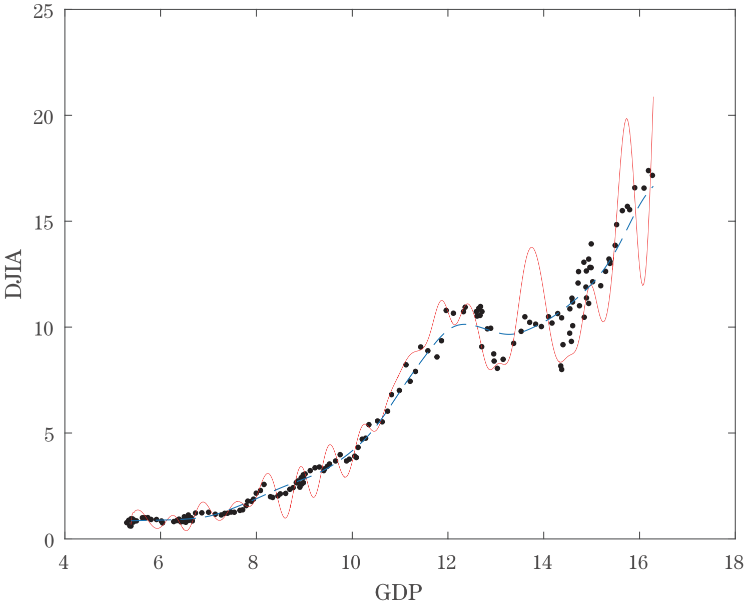

Figure 2. The black data dots are values of (GDP, DJIA) pairs, the dashed line depicts the conventional estimator KERERR of the regression mean, and the solid line shows the wavelet-based estimator WAVERR.

Assumption A1. (a) f_X is continuous, with support equal to the compact interval $[a, b]$, and of bounded variation; (b) f_W is continuous and $f_W(w) > 0$ for all $w \in \mathbb{R}$; (c) g is of bounded variation on $[a, b]$; (d) for each m, the coefficient vector α used to construct $g(x|\alpha, m)$ is constrained to lie in $\mathcal{A}_m$; (e) f_U^{Ft} has at most a countable number of zeros on the real line; and (f) the distribution of V is compactly supported.

Remark 2. Cohen, Daubechies and Vial (1993) construct wavelets that can serve as an orthonormal basis on a compact interval rather than on the whole real line. If one were to use such wavelets to estimate g, $[c, d]$ in the construction of $\mathcal{A}_m$ can be replaced precisely by $[a, b]$.

Remark 3. Examination of the form of the deconvolution kernel K_U suggests that the Fourier transform f_U^{Ft} should not vanish at one or more points in the real line, otherwise there are poles in the integral defining K_U and a problem arises since the integral does not exist. In the numerical examples we present in this paper with the Laplace density model for U, the Fourier transform does not vanish and there is no issue. However, a more advanced construction, suggested first in

Hall and Meister (2007), that modifies the calculation of K_U via introduction of ridging, helps to avoid the issue even in cases where f_U^{Ft} has at most a countable number of zeros on the real line, hence our Assumption A1(e).

In Fan and Truong (1993) and in other papers in the regression with measurement error literature, there is no assumptions of compact support for X whereas this assumption is utilised in the consistency proof for our procedure. While this seems a limitation of our method, we make much less restrictive assumptions on the regression function by allowing it to be discontinuous whereas other approaches require smoothness of the regression function (existence of derivatives of certain order). Also, we believe that the assumption of compact support of X is not too restrictive because this compact support can be made very large to satisfy all practically interesting situations. Not least, our numerical examples and experimentations demonstrate that the method performs very well even when the support of X is not necessarily compact.

Assumption A2. (a) the density estimators $\hat{f}_X$ and $\hat{f}_W$ converge at rate n^{-r} in L_2, in the sense that $\int_a^b(\hat{f}_X - f_X)^2 = O_p(n^{-2r})$ and $\int_{-\infty}^{\infty}(\hat{f}_W - f_W)^2 = O_p(n^{-2r})$, and moreover $\sup_{w\in\mathbb{R}}|\hat{f}_W(w) - f_W(w)| = O_p(n^{-r})$ where $0 < r < 1/2$; and (b) $m = m(n)$ diverges, subject to $m \le m_0(n)$ where $m_0(n) \to \infty$ and $2^{m_0(n)} = o(n^r)$.

This assumption is somewhat generous toward the estimator $\hat{f}_W$; we would expect $\hat{f}_W$ to converge to f_W an order of magnitude faster than does $\hat{f}_X$ to f_X, although we only require the same rate for both estimators. In addition, we mention that the required polynomial convergence rate for $\hat{f}_X$ in Assumption 2 is only possible to achieve if the distribution of U is ordinary smooth (otherwise the attainable rate is much slower (Fan and Truong (1993)).

Define $S(\alpha)$ as at (2.7), with $v_i \equiv 1$ there.

Theorem 1. *If Assumptions* A1–A2 *hold, and the estimator $\hat{g}$ at* (2.3) *is computed using wavelet coefficients $\hat{\alpha}_j$ and $\hat{\alpha}_{jk}$ that minimise $S(\alpha)$, at* (2.7), *with $\hat{\alpha}$ constrained to lie in $\mathcal{A}_m$, then, for all $q \in (0,\infty)$,*

$$\int_{[a,b]} |\hat{g} - g|^q \to 0 \tag{4.3}$$

in probability as $n \to \infty$.

A similar argument can be used to justify using the cross-validatory criterion $T(q)$, at (2.10), weighted by the square of $\hat{f}_{W,-i}(W_i)$:

$$T(q) = \sum_{i \in \mathcal{I}} \left\{ Y_i \hat{f}_{W,-i}(W_i) - \int \hat{g}_{-i}(x) f_U(W_i - x) \hat{f}_{X_,-i}(x) dx \right\}^2 .$$

At present we are able only to establish consistency, not convergence rates. With the weak modelling assumptions and the specificity of our estimation method, establishing convergence rates seems to be a challenging problem.

Supplementary Materials

A detailed proof of the theorem is given in the Supplementary Material.

Acknowledgment

Peter Hall was supported by a laureate fellowship and a discovery project of the Australian Research Council, and by the Australian Resarch Council Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS). The work of Spiridon Penev was supported by Australian Research Council's Discovery Project Funding Scheme (project DP160103489). We are truly indebted to an associate editor and to the two anonymous referees for the careful reading and the numerous suggestions. Their input helped us to significantly improve the style and the presentation of the paper.

References

Cai, T. and Brown, L. (1998). Wavelet shrinkage for nonequispaced samples. *Ann. Statist.* **26**, 1783–1799.

Carroll, R. and Hall, P. (1988). Optimal rates of convergence for deconvolving a density. *J. Amer. Statist. Assoc.* **83**, 1184–1186.

Carrol, R. J., Ruppert, D., Stefanski, L. A. and Crainiceau, C. M. (2006). *Measurement Error in Nonlinear Models.* 2nd Edition, Chapman & Hall, CRC Press, Boca Raton.

Comte, F. and Taupin, M.-L. (2007). Nonparametric estimation of the regression function in an errors-in-varaibles model. *Statist. Sinica* **17**, 1065–1090.

Chesneau, C. (2010). On adaptive wavelet estimation of the regression function and its derivatives in an errors-in-variables model. *Current Development in Theory and Applications of Wavelets* **4**, 131–151.

Claypoole, R. L., Davis, G. M., Sweldens, W. and Baraniuk, R. G. (2003). Nonlinear wavelet transforms for image coding via lifting. *IEEE Trans. Image Process.* **12**, 1449–1459.

Cohen, A., Daubechies, I. and Vial, P. (1993). Wavelets on the interval and fast wavelet transforms. *Appl. Comput. Harmon. Anal.* **1**, 54–81.

Comte, F. and Taupinm M-L. (2007). Adaptive estimation in a non-parametric regression model with error-in-variables. *Statist. Sinica* **17**, 1065–1090.

Daubechies, I. (1992). *Ten lectures on wavelets.* SIAM.

Delaigle, A. and Hall, P. (2008). Using SIMEX for smoothing-parameter choice in errors-in-variables problems. *J. Amer. Statist. Assoc.* **103**, 280–287.

Delaigle, A., Hall, P. and Jamshidi, F. (2015). Confidence bands in non-parametric error-in-variables regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77**, 149–169.

Delaigle, A. and Hall, P. (2016). Methodology for non-parametric deconvolution when the error distribution is unknonwn, *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78**, 231–252.

Delaigle, A. and Meister, A. (2007). Nonparametric regression estimation in the heteroscedastic errors-in-variables problem. *J. Amer. Statist. Assoc.* **102**, 1416–1426.

Delouille, V., Simoens, J. and von Sachs, R. (2004). Smooth design-adapted wavelets for nonparametric stochastic regression. *J. Am. Statist. Soc.* **99**, 643–658.

Donoho, D. and Johnstone, I. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425–455.

Engel, J., Herrmann, E. and Gasser, T. (1994). An iterative bandwidth selector for kernel estimation of densities and their derivatives. *J. Nonparametr. Stat.* **4**, 21–34.

Fan, J. and Truong, Y. (1993). Nonparametric regression with errors in variables. *Ann. Statist.* **21**, 1900–1925.

Fixler, D. (2009). Measurement error in the national accounts. In: *Measurement Error: Consequences, Applications and Solutions. Advances in Econometrics* **24**, (Edited by J. Birner, D. Edgerton and T. Elger.), 91–105.

Gijbels, I., Hall, P. and Kneip, A. (1999). On the estimation of jump points in smooth curves. *Ann. Inst. Statist. Math.* **51**, 231–251.

Hall, P. and Meister, A. (2007). A ridge-parameter approach to deconvolution. *Ann. Statist.* **35**, 1535–1558.

Hall, P. and Penev, S. (2001). Cross-validation for choosing resolution level for nonlinear wavelet curve estimators. *Bernoulli* **7**, 317–341.

Hall, P. and Turlach, B. (1997). Interpolation methods for nonlinear wavelet regression with irregularly spaced design. *Ann. Statist.* **25**, 1912–1925.

Jansen, M., Nason, G. P. and Silverman, B. W. (2009). Multiscale methods for data on graphs and irregular multidimensional situations. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71**, 97–125.

Maity, A. and Apanasovich, I. (2011). Estimation via corrected scores in general semiparametric regression models with error-prone covariates. *Electron. J. Stat.* **5**, 1424–1449.

Müller, H.-G. (1988). *Nonparametric Regression Analysis of Longitudinal Data.* Springer.

Nason, G. P. (2008). *Wavelet Methods in Statistics with R.* Springer.

Nunes, M., Knight, M. and Nason, G. P. (2006). Adaptive lifting for nonparametric regression. *Stat. Comput.* **16**, 143–159.

Stefanski, L. and Carroll, R. (1990). Deconvoluting kernel density estimators. *Statistics* **21**, 169–184.

Young, A. (1994). The statistics corner: Reliability and accuracy of quarterly GDP. *Business Economics* **29**, 63–67.

School of Mathematics and Statistics, University of Melbourne, Victoria 3010, Australia.

Department of Statistics, School of Mathematics and Statistics, UNSW, Sydney 2052 NSW, Australia.

E-mail: s.penev@unsw.edu.au

School of Mathematics and Statistics, University of Melbourne, Victoria 3010, Australia.

E-mail: j.tran8@student.unimelb.edu.au

(Received August 2016; accepted April 2017)