# ISOTONIC NONPARAMETRIC REGRESSION IN THE PRESENCE OF MEASUREMENT ERROR

| | |
|---|---|
| Raymond J. Carroll | Texas A&M University |
| Aurore Delaigle | University of Melbourne |
| Peter Hall | U Melbourne & UC Davis |

# MOTIVATION AND MODEL

In measurement error problems we are often interested in estimating a regression curve $g(x) = E(Y \mid X = x)$ from data on $(W, Y)$, or in estimating the density $f_X$ of $X$ from data on $W$, where $W$ represents a contaminated version of $X$.

A typical model is

$$Y = g(X) + \epsilon\,, \quad W = X + U\,, \tag{1}$$

where $U \sim f_U$, $X \sim f_X$, $U$, $X$ and $\epsilon$ are independent, $E(\epsilon^2) = \sigma^2$ and $E(\epsilon) = 0$. In this model the variable $U$ is unobserved and represents measurement error.

Identifiability of $f_X$ or $g$ from data generated by the model at (1) requires $f_U$ to be known, and so we shall make this assumption. It is straightforward to extend our methodology to cases where $f_U$ is unknown and estimated from replicated data.

The presence of the noise variable, $U$, in (1) makes the model ill-posed. The solution to this inverse problem inevitably has reduced statistical performance, relatively to the case where $U \equiv 0$.

# INFERENCE UNDER CONSTRAINTS

It is often of interest to estimate $g$, and perhaps also $f_X$, nonparametrically but under shape restrictions. In principle, a shape constraint can be quite general, for example monotonicity, convexity, log-concavity or unimodality. In practice, it is usually motivated by prior information that we have about a particular problem.

In the context of errors-in-variables regression the most appropriate constraint is monotonicity. For example, when the explanatory variable, $X$, represents the value taken by a treatment or dosage, the conditional mean of the response, $Y$, is often anticipated to be a monotone function of $X$.

Indeed, if this regression mean is not monotone (in the appropriate direction) then the medical or commercial value of the treatment is likely to be significantly reduced, at least for values of $X$ that lie beyond the point at which monotonicity fails. In the case of a probability density, common shape constraints include log-concavity and unimodality.

# Brief literature survey

The literature on nonparametric inference in errors-in-variables problems is vast, but can be accessed relatively easily through the monograph by Carroll, Ruppert, Stefanski and Crainiceanu (2006). In earlier work on errors-in-variables problems, Meister (2009) suggested a test for local monotonicity of a density function, and Cordy and Thomas (1997) estimated a distribution function under a unimodality constraint.

More generally, recent contributions to nonparametric or semiparametric methodology, without shape constraints, include those of Schennach (2004), Huang *et al.* (2006), Delaigle and Meister (2007) and Delaigle, Fan and Carroll (2009).

We shall use a tilting-based approach to enforce constraints. This method was proposed in a special case by Grenander (1956), and substantially generalised by Hall and Presnell (1999). Recent examples of the use of this methodology can be found in work of Müller *et al.* (2005) and Schick and Wefelmeyer (2009).

# ESTIMATORS

First we describe methods for constructing unconstrained estimators of the density $f_X$ and regression mean $g$ in the model

$$Y = g(X) + \epsilon\,, \quad W = X + U\,.$$

If $K$ is a conventional kernel function and $h$ is a bandwidth then the estimators are

$$\hat{f}_X(x) = \frac{1}{h} \sum_{j=1}^{n} K_U\Big(\frac{x - W_j}{h}\Big)\,, \quad \hat{g}(x) = \frac{\widehat{gf}_X(x)}{\hat{f}_X(x)} = \sum_{j=1}^{n} S_j(x)\, Y_j\,,$$

where

$$\widehat{gf}_X(x) = \frac{1}{nh} \sum_{j=1}^{n} Y_j\, K_U\Big(\frac{x - W_j}{h}\Big)\,, \quad S_j(x) = \frac{K_U\{(x - W_j)/h\}}{\sum_k K_U\{(x - W_k)/h\}}\,,$$

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

and the subscript Ft denotes “Fourier transform.” (Thus, $f_U^{\mathrm{Ft}}$ is the characteristic function corresponding to the density $f_U$ of $U$.)

# CONSTRAINED ESTIMATORS – (1)

First we perturb the conventional estimators,

$$\hat{f}_X(x) = \frac{1}{h} \sum_{j=1}^{n} K_U\Big(\frac{x - W_j}{h}\Big)\,, \quad \hat{g}(x) = \sum_{j=1}^{n} S_j(x)\, Y_j\,,$$

by tilting them to:

$$\hat{f}_X(x\,|\,p) = \frac{1}{h} \sum_{j=1}^{n} p_j\, K_U\Big(\frac{x - W_j}{h}\Big)\,, \quad \hat{g}(x\,|\,p) = n \sum_{j=1}^{n} p_j\, S_j(x)\, Y_j\,,$$

where the vector $p = (p_1, \ldots, p_n)$ is a multinomial probability distribution: each $p_j \geq 0$ and $\sum_j p_j = 1$.

Next we choose $p$ to minimise the distance of that vector from the uniform probability distribution, $p^0 = (1/n, \ldots, 1/n)$, subject to the constraint being satisfied. The constraint might be, say unimodality for $\hat{f}_X(\,\cdot\,|\,p)$; we shall focus on the constraint of monotonicity for $\hat{g}(\,\cdot\,|\,p)$.

# Constrained estimators – (2)

The concept of "tilting through the least distance subject to the constraint being satisfied" results in greatest fidelity to the data, subject to the constraint.

Distance measures that can be used include those suggested by Cressie and Read (1984) and Read and Cressie (1988):

$$D_\rho(p) = \frac{1}{\rho(1-\rho)} \left\{ n - \sum_{j=1}^{n} (np_j)^\rho \right\}, \quad D_0(p) = -\sum_{j=1}^{n} \log(np_j)\,, \quad D_1(p) = n \sum_{j=1}^{n} p_j \, \log(np_j)\,,$$

where $0 < \rho < 1$ in the definition of $D_\rho$.

These distance measures are generally not metrics, and for example the two Kullback-Leibler divergences, $D_0$ and $D_1$, are asymmetric in terms of the roles played by $p$ and $p^0$. They can nevertheless be readily interpreted from a statistical viewpoint.

The distance $D_0$ is arguably not as satisfactory as the others, since it takes the value infinity when one or more of the $p_j$s is zero and therefore strongly resists setting any of the $p_j$s to zero. This can result in other $p_j$s being altered unnecessarily.

# Computational considerations

A constrained density or regression estimator can be implemented by choosing $p$ to minimise

$$D_\rho(p) + \lambda \, \mathrm{Pen}(p) \,,$$

where $\mathrm{Pen}(p)$ is a positive penalty function of $p$ which increases as the estimated curve departs further from the shape constraint, and $\lambda$ is a parameter used to control the strength of the penalty. In practice we start with a small $\lambda$ and repeat the procedure for successively larger values of $\lambda$ until the constraint is satisfied.

For example, the condition that $g$ is monotone increasing on a specific interval $\mathcal{I} = [a, b]$, say, can be imposed computationally by dividing $\mathcal{I}$ up into a regular, discrete grid of points, $a = x_1 < \ldots < x_m = b$, and adding to the distance measure $D_\rho(p)$ the penalty

$$\lambda \, \mathrm{Pen}(p) = \lambda \sum_{k=1}^{m-1} \left| \hat{g}(x_k \,|\, p) - \hat{g}(x_{k+1} \,|\, p) \right|^r \mathbf{I}\left\{ \hat{g}(x_k \,|\, p) - \hat{g}(x_{k+1} \,|\, p) > 0 \right\} ,$$

where $\mathbf{I}(\cdot)$ is the indicator function, $\hat{g}(\,\cdot \,|\, p)$ is our tilted regression estimator, and $r$ denotes a positive integer.

# Remarks on constraining by tilting – (1)

**Remark 1.** Many standard methods, for example those based on splines and ridging, can be constrained using the above tilting-based approach. Our choice of kernel methods enables us to develop relatively detailed theoretical properties, which can be expected to reflect those in other cases where such a concise account is out of reach. In the kernel case our work extends easily to local polynomial estimators (Delaigle, Fan and Carroll, 2009).

**Remark 2.** Note that, for $0 \leq \rho \leq 1$, we have $D_\rho(p^0) = 0$ and $D_\rho(p) > 0$ if $p \neq p^0$. In our numerical work we found that $D_1$ generally gave very good performance, although results for other $D_\rho$, for $\rho > 0$, were often similar.

# REMARKS ON CONSTRAINING BY TILTING – (2)

**Remark 3.** Each of the distance measures has the advantage that it is not well-defined unless each $p_j \geq 0$, or in fact $p_j > 0$ in the case of $D_0$. This means that we do not need to impose nonnegativity as an additional constraint. In contrast, the standard quadratic measure of distance between $p$ and $p^0$ does not automatically ensure that the components of $p$ are nonnegative.

**Remark 4.** The other required constraint, $\sum_{1\leq j\leq n} p_j = 1$, can be ensured quite readily, for example by replacing $p_1$ by $1 - \sum_{2\leq j\leq n} p_j$.

# Summary of theoretical properties

Under conventional regularity conditions, and assuming that the true regression function $g$ satisfies the constraints, the constrained regression estimator $\hat{g}(\,\cdot\,|\,p)$ attains optimal convergence rates, both pointwise and uniformly.

Additionally, optimal convergence rates to derivatives of $g$ are enjoyed by the respective derivatives of $\hat{g}(\,\cdot\,|\,p)$.

Moreover, the probability distribution $p = (p_1, \ldots, p_n)$ can be chosen so that $\hat{g}(\,\cdot\,|\,p)$, and its derivatives, converge to a general smooth function $\gamma$, and its respective derivatives.

This shows that the tilting approach is particularly flexible: $p$ can be chosen so that the $\ell$th derivative $\gamma^{(\ell)}$ of any function $\gamma$ that satisfies mild regularity conditions can be consistently estimated by the tilted estimator $\hat{g}^{(\ell)}(\,\cdot\,|\,p)$, defined to be the $\ell$th derivative of $\hat{g}(\,\cdot\,|\,p)$.

# Applications to hypothesis testing

The operation of choosing $p$, subject to each $p_j \geq 0$ and $\sum_j p_j = 1$, to ensure that $\hat{f}_X(\,\cdot\,|\,p)$ or $\hat{g}(\,\cdot\,|\,p)$ satisfies a shape constraint on $\mathcal{I}$, produces an empirical probability distribution $\hat{p}$. We can interpret $D_\rho(\hat{p})$ as the distance through which we have to tilt the data in order to ensure that the estimator $\hat{g}(\,\cdot\,|\,\hat{p})$ satisfies the shape constraint.

We expect that, as the shape of $g$ moves further from that prescribed by the null hypothesis $H_0$, the value of $D_\rho(\hat{p})$ will increase. Therefore we suggest testing $H_0$ by rejecting it if $D_\rho(\hat{p})$ is too large.

# Implementing the test – (1)

We use bootstrap methods to calibrate the test, as follows.

(i) Compute a conventional deconvolution-based estimator $\hat{f}_X$ of $f_X$, and a shape constrained estimator $\hat{g}(\,\cdot\,|\,\hat{p})$ of $g$, under the null hypothesis $H_0$ of monotonicity, from the dataset $\mathcal{D} = \{(W_1, Y_1), \ldots, (W_n, Y_n)\}$.

(ii) Compute an estimator $\hat{\sigma}^2$ of the variance $\sigma^2 = \mathrm{var}(\epsilon)$. (Methods are given by Delaigle and Hall, 2010).

(iii) Convert $\hat{f}_X$ to a proper density function $\widetilde{f}_X$ and sample data $X_1^*, \ldots, X_n^*$ from $\widetilde{f}_X$, sample $U_1^*, \ldots, U_n^*$ from $f_U$, and sample $\epsilon_1^*, \ldots, \epsilon_p^*$ from a distribution with mean $0$ and variance $\hat{\sigma}^2$ (e.g. a normal distribution). Then set $W_j^* = X_j^* + U_j^*$ and $Y_j^* = \hat{g}(X_j^*\,|\,\hat{p}) + \epsilon_j^*$.

# Implementing the test – (2)

(iv) Compute, from the dataset $\mathcal{D}^* = \{(W_1^*, Y_1^*), \ldots, (W_n^*, Y_n^*)\}$, the bootstrap version $\hat{g}^*(\cdot \,|\, p)$ of $\hat{g}(\,\cdot\,|\,p)$.

(v) Calculate the version $\hat{p}^*$ of $\hat{p}$ by tilting to ensure that $\hat{g}^*(\cdot \,|\, \hat{p}^*)$ satisfies the shape constraint on $\mathcal{I}$, and compute $D_\rho(\hat{p}^*)$.

(vi) Given a potential level, $\alpha \in (0,1)$, for a test of $H_0$, and using bootstrap simulation, compute the upper $\alpha$-level critical point $\hat{\xi}_\alpha$ of the conditional distribution of $D_\rho(\hat{p}^*)$.

(vii) Reject the null hypothesis if $D_\rho(\hat{p}) > \hat{\xi}_\alpha$.

# SIMULATIONS – (1)

We considered a family of regression models introduced by Bowman, Jones and Gijbels (1998), and defined by

$$g(x) = 1 + x - a\exp\left\{ -50\,(x-0.5)^2 \right\},$$

$\epsilon \sim \mathrm{Normal}\,(0, 0.05^2)$, $X \sim \mathrm{Normal}\,(0.5, 0.1)$, where $a$ is chosen so that $g$ is clearly monotone increasing ($a = 0$), only just monotone increasing ($a = 0.15$), slightly nonmonotone increasing ($a = 0.25$) or more clearly nonmonotone ($a = 0.45$).

The measurement errors were Laplace, chosen so that the noise to signal ratio, $\mathrm{var}(U)/\mathrm{var}(X)$, was $20\%$. Sample size was $n = 250$.

The figure compares the estimator $\hat g$ with its monotonised version $\hat g(\cdot\,|\,\hat p)$, when $a = 0$ or $a = 0.15$. We used the bandwidth suggested by Delaigle and Hall (2008).

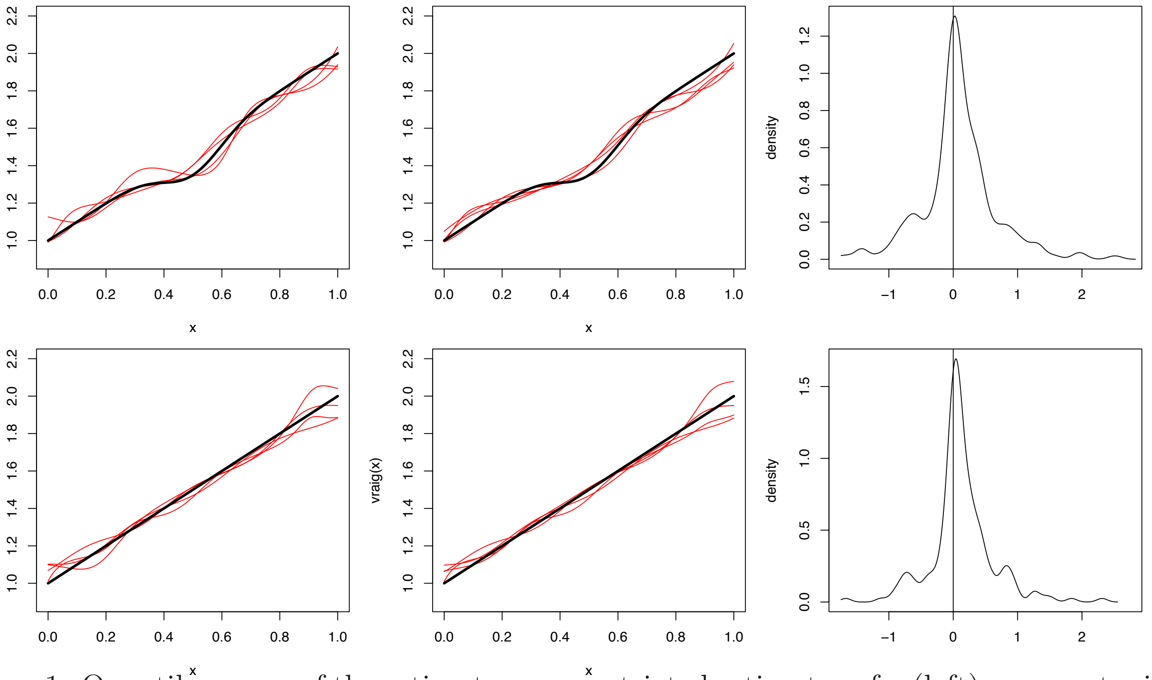

Figure 1: Quantile curves of the estimators: unrestricted estimator of $g$ (left) or monotonized estimators (middle), when the regression curve corresponds to $a = 0.15$ (top) or $a = 0$ (bottom). Right: kernel estimators of the density of $\log[\mathrm{ISE}(\widehat{g})/\,\mathrm{ISE}\{\widehat{g}(\cdot\,|\,\widehat{p})\}]$; the vertical line indicates the value 0 for reference.

# Simulations – (2)

For both constrained and unconstrained estimators we show four curves corresponding to the samples which gave the quantiles $0.2$, $0.4$, $0.6$ and $0.8$ of the values of the Integrated Squared Error, $\mathrm{ISE}(\widehat{g}) = \int (\widehat{g} - g)^2$.

We also show kernel estimators of the density of $\log[\mathrm{ISE}(\widehat{g})/\mathrm{ISE}\{\widehat{g}(\cdot \mid \hat{p})\}]$, calculated from the 200 samples. It can be seen that there is a slight skewness to the right, indicating a slight reduction in ISE by constraining.

To explore this property we computed the Median Integrated Squared Errors (MISE), finding an improvement of 12% when $a = 0.15$ (respectively 7% when $a = 0$). In addition, in these cases the percentage of the times that the constrained estimator had smaller MISE was 61% when $a = 0.15$ (and 67% when $a = 0$).

# Real-data example – (1)

We applied our procedure to the peak expiratory flow rate (PEFR) data of Bland and Altman (1986). The data concern measurements of the PEFR on 17 individuals, using two procedures: two replicated accurate measurements obtained by a Wright peak flow meter, and two replicated inaccurate measurements obtained by a mini Wright meter.

The aim was to determine whether the mini Wright readings are in agreement with the Wright readings.

The variance of $U$ was estimated from the replicated mini Wright readings, and for simplicity of calculation we assumed a Laplace error.

The data and the two regression estimators (unrestricted and monotonised, respectively) are plotted in the figure.

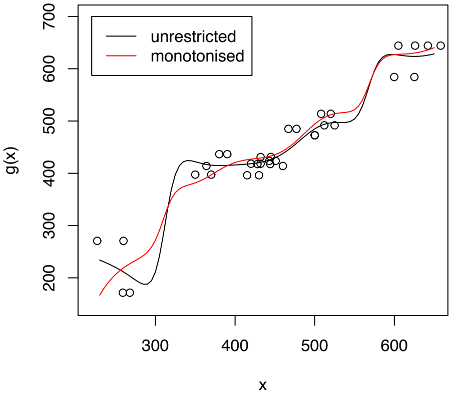

Figure 2: Unrestricted and monotonized estimated curves $g(x)$ for the Wright data.

# Real-data example – (2)

Although the unrestricted estimator fluctuates somewhat, an application of our testing procedure does not permit us to reject the hypothesis that the readings on the mini Wright meter are a monotone function of the readings on the Wright meter. It is perhaps reasonable to infer that the fluctuations are artifacts caused by the sample sample size, rather than a true characteristic of the curve.