Testing and Estimating Shape-Constrained Nonparametric Density and Regression in the Presence of Measurement Error [1]

Raymond J. Carroll, Aurore Delaigle and Peter Hall

Abstract

In many applications we can expect that, or are interested to know if, a density function or a regression curve satisfies some specific shape constraints. 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 curtailed, at least for values of X that lie beyond the point at which monotonicity fails. In the case of a density, common shape constraints include log-concavity and unimodality. If we can correctly guess the shape of a curve, then nonparametric estimators can be improved by taking this information into account. Addressing such problems requires a method for testing the hypothesis that the curve of interest satisfies a shape constraint, and, if the conclusion of the test is positive, a technique for estimating the curve subject to the constraint. Nonparametric methodology for solving these problems already exists, but only in cases where the covariates are observed precisely. However in many problems, data can only be observed with measurement errors, and the methods employed in the error-free case typically do not carry over to this error context. In this paper we develop a novel approach to hypothesis testing and function estimation under shape constraints, which is valid in the context of measurement errors. Our method is based on tilting an estimator of the density or the regression mean until it satisfies the shape constraint, and we take as our test statistic the distance through which it is tilted. Bootstrap methods are used to calibrate the test. The constrained curve estimators that we develop are also based on tilting, and in that context our work has points of contact with methodology in the error-free case.

KEYWORDS. Bootstrap methods; Convexity; Errors in variables; Hypothesis testing; Kernel methods; Local polynomial estimators; Monotone function; Nonparametric regression; Shape constraint; Unimodality.

[1]Raymond J. Carroll is distinguished professor, Department of Statistics, Texas A & M University, College Station, TX 77843 (E-mail: carroll@stat.tamu.edu). Aurore Delaigle is a principal researcher and Queen Elizabeth II fellow, Department of Mathematics and Statistics, University of Melbourne, VIC, 3010, Australia (E-mail: A.Delaigle@ms.unimelb.edu.au). Peter Hall is Professor and federation fellow, Department of Mathematics and Statistics, University of Melbourne, VIC, 3010, Australia and Distinguished Professor, Department of Statistics, University of California at Davis, Davis, CA 95616, USA (E-mail: halpstat@ms.unimelb.edu.au). Carroll's research was supported by a grant from the National Cancer Institute (R37-CA057030) and by Award Number KUS-CI-016-04, made by King Abdullah University of Science and Technology (KAUST), and by National Science Foundation Instrumentation grant number 0922866. Delaigle's research was supported by a grant from the Australian Research Council, and Hall's research was supported by grants from the Australian Research Council and from the National Science Foundation. The authors thank the editor, the associate editor, and a referee for their valuable comments that helped improve a previous version of the manuscript. Address for correspondence: Aurore Delaigle, Department of Mathematics and Statistics, University of Melbourne, VIC, 3010, Australia

1 Introduction

In measurement errors problems, the interest is often in estimating a regression curve $g(x) = E(Y \,|\, X = x)$ 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. For instance, in dietary studies, the level X of saturated fat is often measured through a food frequency questionnaire (FFQ). Since FFQs that give information about X change upon repeated administration, the observed data W are regarded as noisy or contaminated values of the true response variable. Estimating g or f_X nonparametrically in this measurement error context is notoriously very hard. However, in many practical problems, we can anticipate that the regression curve or the density function f_X satisfies some shape constraints. For example, X might represent the level of a dietary component and Y a surrogate for a detrimental medical condition. If Y does not increase monotonically with X then we may need to reassess the usefulness of X as a pointer to the condition. In the density context, constraints that are commonly encountered include those of unimodality and log-concavity. In general, if our guess about a shape constraint of a curve is correct, then incorporating this information into the estimation procedure can improve the quality of estimators, which would be particularly useful in the difficult errors-in-variables context. If we wish to assess the validity of our guess, then we need to develop a procedure for testing hypotheses about the shape of the curve. Such methods exist in the literature, but only in the case where X is observed without measurement errors, and these techniques are usually not valid in the error case.

Relying only on observations of (W, Y) (in the regression case) or W (in the density case), in this paper we develop methods for testing the hypothesis that g or f_X satisfies a shape constraint, against the alternative that it does not, and for estimating g or f_X subject to the assumption that it satisfies the constraint. In both problems we adopt a tilting-based approach, a version of which was first suggested by Grenander (1956), for enforcing constraints. See Hall and Huang (2001, 2002), Müller *et al.* (2005) and Schick and Wefelmeyer (2009) for recent examples of this type of methodology. In the context of hypothesis testing we take our test statistic to be the distance through which the unconstrained estimator is tilted, and this represents a new approach to solving hypothesis testing problems. Further, we

use bootstrap methodology to calibrate our test. We study the properties of the suggested estimating and testing procedures both in practice and in theory. As part of our work, we describe what to the best of our knowledge is the first uniform convergence result for nonparametric regression with errors-in-variables.

Statistical methodology for estimating g or f_X, subject to shape constraints, or for testing the validity of the shape constraints, has been widely studied in the literature. See for example the relatively recent contributions by Groeneboom (2001), Hall and Huang (2001, 2002), Hall and Kang (2005), Dette *et al.* (2006), Antoniadis *et al.* (2007), Neumeyer (2007), Pal and Woodroofe (2007), Birke and Dette (2007), Birke (2009), Dümbgen and Rufibach (2009), Cule *et al.* (2010) and the references therein. Earlier work includes that of Friedman and Tibshirani (1984), Mukerjee (1988), Kelly and Rice (1990), Mammen (1991, 1995) and Sun and Woodroofe (1996).

The literature on estimating the density f_X and the regression mean g in errors-in-variables problems is particularly extensive, but can be accessed relatively easily through the monograph by Carroll *et al.* (2006). In an errors-in-variables setting the only existing work on inference subject to shape constraints appears to be that of Meister (2009), who suggests a test for local monotonicity of a density function, and Cordy and Thomas (1997), who estimate a distribution function under some unimodality constraint. Birke and Bissantz (2009) construct monotone estimators in a related problem involving convolution operators. However, these works do not provide adaptive, nonparametric approaches to estimation of, or hypothesis testing of general shape assumptions about, g and f_X.

2 Methodology

2.1 Model and constrained estimator

In the classical errors-in-variables regression problem, the interest is to estimate a regression curve g from data $(W_1, Y_1), \ldots, (W_n, Y_n)$ generated by the model

$$Y = g(X) + \epsilon\,, \quad W = X + U\,, \tag{2.1}$$

where $U \sim f_U$, $X \sim f_X$, U, X and ϵ are independent, $E(\epsilon^2) = \sigma^2$ and $E(\epsilon) = 0$. In this model the variable U is unobserved and represents a measurement error. In the density context, the standard problem is to estimate the density f_X from data $W_1, \ldots, W_n$ generated as at (2.1). Identifiability of f_X or g from data generated by the model (2.1) requires f_U to be known, and we shall also make this assumption. It is straightforward to extend our methodology to cases where f_U is unknown and estimated from replicated data, using the techniques of Delaigle *et al.* (2008).

Our goal in the constrained errors-in-variables problem is to estimate g and f_X nonparametrically 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 some a priori information that we have about a particular problem. Our approach to incorporating shape constraints is based on modifying standard non-restricted, nonparametric, errors-in-variables curve estimators.

We start by presenting the methodology in the density case. Let h be a bandwidth and K a kernel function, let $\phi_K(t) = \int e^{itx} K(x)\, dx$ denote the Fourier transform of K, and let $\phi_U(t) = \int e^{itx} f_U(x)\, dx$ be the characteristic function corresponding to f_U. The non-restricted errors-in-variables (or deconvolution) kernel estimator of f_X is defined by

$$\widehat{f}_X(x) = \frac{1}{nh} \sum_{j=1}^{n} K_U\{(x - W_j)/h\} \tag{2.2}$$

where

$$K_U(u) = \frac{1}{2\pi} \int e^{-itu}\, \phi_K(t)/\phi_U(t/h)\, dt\,. \tag{2.3}$$

See Carroll and Hall (1988) and Stefanski and Carroll (1990).

To incorporate a shape constraint in the estimation procedure we use a tilting approach which consists in first replacing the equal weights n^{-1}, applied to the sum in (2.2), by more general weights denoted respectively by p_j, where

$$p_j \geq 0 \quad \text{for} \quad 1 \leq j \leq n\,, \quad \text{and} \quad p_1 + \ldots + p_n = 1\,. \tag{2.4}$$

That is, we replace $\widehat{f}_X(x)$ in (2.2) by

$$\widehat{f}_X(x \,|\, p) = \frac{1}{h} \sum_{j=1}^{n} p_j\, K_U\{(x - W_j)/h\}\,. \tag{2.5}$$

Then, we choose the multinomial vector $p = (p_1, \ldots, p_n) = \widehat{p}$ so as to minimize the distance from p to the uniform vector $p^0 = (n^{-1}, \ldots, n^{-1})$, subject to our estimator of f_X satisfying the constraint. A variety of distance measures can be used in this context, including those suggested by Cressie and Read (1984), Read and Cressie (1988) and Hall and Presnell (1999). They are generally not metrics, and for example the two Kullback-Leibler divergences in (2.7) below are asymmetric in terms of the roles played by p and p^0. They are nevertheless readily interpretable from a statistical viewpoint. Suitable distance functions for $0 < \rho < 1$ are

$$D_\rho(p) = \frac{1}{\rho(1-\rho)} \left\{ n - \sum_{j=1}^n (np_j)^\rho \right\} \tag{2.6}$$

$$D_0(p) = -\sum_{j=1}^n \log(np_j)\,, \quad D_1(p) = n \sum_{j=1}^n p_j \log(np_j)\,. \tag{2.7}$$

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

Remark 1. Each of these quantities 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 avoids us having 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. The other required constraint, that $\sum_{1 \leq j \leq n} p_j = 1$, can be ensured more readily, for example by replacing p_1 by $1 - \sum_{2 \leq j \leq n} p_j$. Note too that, for $0 \leq \rho \leq 1$, $D_\rho(p^0) = 0$ and $D_\rho(p) > 0$ if $p \neq p^0$.

We use a similar approach in the regression context. Here, we modify the non-restricted errors-in-variables kernel estimator of Fan and Truong (1993), which is defined by

$$\widehat{g}(x) = \frac{\widehat{gf}_X(x)}{\widehat{f}_X(x)} = \sum_{j=1}^n S_j(x)\, Y_j\,, \tag{2.8}$$

where $\widehat{f}_X$ is defined in (2.2), and where

$$\widehat{gf}_X(x) = \frac{1}{nh} \sum_{j=1}^n Y_j\, K_U\{(x - W_j)/h\}\,, \tag{2.9}$$

$$S_j(x) = \frac{K_U\{(x - W_j)/h\}}{\sum_k K_U\{(x - W_k)/h\}}\,. \tag{2.10}$$

To incorporate a shape constraint we first replace the equal weights n^{-1}, applied to each Y_j in (2.9), by more general weights p_j satisfying (2.4). That is, we replace $\widehat{g}(x)$ in (2.8) by

$$\widehat{g}(x \,|\, p) = n \sum_{j=1}^{n} p_j \, S_j(x) \, Y_j \,. \tag{2.11}$$

Then, as in the density case, we choose the weights $\widehat{p}$ so as to minimize a distance from p to p^0 (for example $D_\rho(p)$ defined above) subject to $\hat{g}(\,|\,p)$ satisfying the constraint.

From a computational viewpoint, our procedure for constructing a constrained density or regression estimator can be implemented by choosing p to minimize

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

where D_ρ is one of the distance measures at (2.6) and (2.7), $\mathrm{Pen}(p)$ is a positive penalty function of p which increases as the estimated curve departs further from the shape constraint, and λ is a parameter used to control the strength of the penalty. In practice we start with a small λ and repeat the procedure for successively larger values of λ 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

$$\pi(p, \lambda) = \lambda \sum_{k=1}^{m-1} \left| \widehat{g}(x_k \,|\, p) - \widehat{g}(x_{k+1} \,|\, p) \right|^r \mathbf{I}\{\widehat{g}(x_k \,|\, p) - \widehat{g}(x_{k+1} \,|\, p) > 0\} \,, \tag{2.12}$$

where $\mathbf{I}(\cdot)$ is the indicator function, $\widehat{g}(\,\cdot\,|\,p)$ is as at (2.11), and r denotes a positive integer. Similarly, the condition that f_X is log-concave (i.e., $f_X^{(1)}/f_X$ is decreasing) on $\mathcal{I} = [a, b]$ can be imposed by the penalty

$$\pi(p, \lambda) = \lambda \sum_{k=1}^{m-1} \left| \frac{\widehat{f}_X^{(1)}(x_k \,|\, p)}{\widehat{f}_X(x_k \,|\, p)} - \frac{\widehat{f}_X^{(1)}(x_{k+1} \,|\, p)}{\widehat{f}_X(x_{k+1} \,|\, p)} \right|^r \mathbf{I}\left\{ \frac{\widehat{f}_X^{(1)}(x_k \,|\, p)}{\widehat{f}_X(x_k \,|\, p)} - \frac{\widehat{f}_X^{(1)}(x_{k+1} \,|\, p)}{\widehat{f}_X(x_{k+1} \,|\, p)} < 0 \right\} , \tag{2.13}$$

where $\widehat{f}_X(\,\cdot\,|\,p)$ is as at (2.5). Note that ρ and r are not smoothing parameters and that their choice has a relatively minor impact on the success of the method. In our numerical implementation of the procedure we took $r = 2$ and $\rho = 1$.

Remark 2. Obviously, many standard methods, for example those based on splines and ridging, can be implemented in this way. Our choice of a kernel approach 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 methodology and theory extend easily to local polynomial estimators (Delaigle *et al.,* 2009).

Remark 3. Choosing the value of r to use in the constraint is quite similar to choosing a distance to minimize in general statistical estimation problems. Essentially, any $r \geq 0$ will lead to a consistent estimator that satisfies the shape constraint, but depending on the sample, one value of r will work better than the other. In finite samples the results will differ according to the value of r and, like the choice of a distance, the choice of r is up the preference of the user. For example, the constraint at (2.12) will provide a consistent monotone increasing estimator for any $r \geq 0$. Taking $r = 0$ penalizes equally regions that are almost monotone and regions that are strongly nonmonotone. Taking r larger puts a heavier penalty on regions that are severely nonmonotone, and the iterative minimization procedure first tries to correct severe violations, and then corrects smaller problems.

2.2 Hypothesis testing

The operation of choosing p, subject to each $p_j \geq 0$ and $\sum_j p_j = 1$, to ensure that $\widehat{f}_X(\,\cdot\,|\,p)$ or $\widehat{g}(\,\cdot\,|\,p)$ satisfies a shape constraint on $\mathcal{I}$, produces an empirical probability distribution $\widehat{p}$. We can interpret $D_\rho(\widehat{p})$ as the distance through which we have to tilt the data in order to ensure that the estimator satisfies the shape constraint. We expect that, as the shape of f_X or g moves further from the null hypothesis H_0 that a given shape constraint on $\mathcal{I}$ is satisfied, the value of $D_\rho(\widehat{p})$ will increase. Therefore we suggest testing H_0 by rejecting it if $D_\rho(\widehat{p})$ is too large.

In the density case, we calibrate the test using the bootstrap, as follows. (i) Compute a conventional deconvolution-based estimator $\widehat{f}_X$ of f_X, and a shape-constrained estimator $\widehat{f}_X(\,\cdot\,|\,\widehat{p})$ of f_X under H_0, from the data set $\mathcal{D} = \{W_1, \ldots, W_n\}$. (ii) Convert $\widehat{f}_X(\,\cdot\,|\,\widehat{p})$ to a proper density function $\widetilde{f}_X(\,\cdot\,|\,\widehat{p})$ (Hall and Murison, 1993), and sample data $X_1^*, \ldots, X_n^*$ from $\widetilde{f}_X(\,\cdot\,|\,\widehat{p})$, and $U_1^*, \ldots, U_n^*$ from f_U. Then construct $W_j^* = X_j^* + U_j^*$. (iii) Compute, from

the data set $\mathcal{D}^* = \{W_1^*, \ldots, W_n^*\}$, the bootstrap version $\widehat{f}_X^*(\cdot \,|\, p)$ of $\widehat{f}_X(\,\cdot \,|\, p)$. (iv) Calculate the version $\widehat{p}^*$ of $\widehat{p}$ by tilting to ensure that $\widehat{f}_X^*(\cdot \,|\, \widehat{p}^*)$ satisfies the shape constraint on $\mathcal{I}$, and compute $D_\rho(\widehat{p}^*)$. (v) Given a potential level, $\alpha \in (0,1)$, for a test of H_0, and using bootstrap simulation, compute the upper α-level critical point $\widehat{\xi}_\alpha$ of the conditional distribution of $D_\rho(\widehat{p}^*)$. (vi) Reject the null hypothesis if $D_\rho(\widehat{p}) > \widehat{\xi}_\alpha$.

In the regression case, we suggest calibrating the test using the bootstrap, as follows. (i) Compute a conventional deconvolution-based estimator $\widehat{f}_X$ of f_X, and a shape-constrained estimator $\widehat{g}(\,\cdot \,|\, \widehat{p})$ of g under H_0, from the data set $\mathcal{D} = \{(W_1, Y_1), \ldots, (W_n, Y_n)\}$. (ii) Compute an estimator $\widehat{\sigma}^2$ of the variance $\sigma^2 = \mathrm{var}(\epsilon)$, using methods of Delaigle and Hall (2010). (iii) Convert $\widehat{f}_X$ to a proper density function $\widetilde{f}_X$ and sample data $X_1^*, \ldots, X_n^*$ from $\widetilde{f}_X$, $U_1^*, \ldots, U_n^*$ from f_U, and $\epsilon_1^*, \ldots, \epsilon_p^*$ from a distribution with mean 0 and variance $\widehat{\sigma}^2$. Then set $W_j^* = X_j^* + U_j^*$ and $Y_j^* = \widehat{g}(X_j^* \,|\, \widehat{p}) + \epsilon_j^*$. (iv) Compute, from the data set $\mathcal{D}^* = \{(W_1^*, Y_1^*), \ldots, (W_n^*, Y_n^*)\}$, the bootstrap version $\widehat{g}^*(\cdot \,|\, p)$ of $\widehat{g}(\,\cdot \,|\, p)$. (v) Calculate the version $\widehat{p}^*$ of $\widehat{p}$ by tilting to ensure that $\widehat{g}^*(\cdot \,|\, \widehat{p}^*)$ satisfies the shape constraint on $\mathcal{I}$, and compute $D_\rho(\widehat{p}^*)$. (vi) and (vii): same as steps (v) and (vi) of the density case.

In step (iii) above we can for example assume the experimental errors to be uniformly distributed since the method proves to be quite robust against this assumption. For instance, first-order limit-theoretic properties depend on the error distribution only through its variance.

Remark 4. The method of Hall and Murison (1993) consists in replacing an estimator $\widehat{f}_X(x|p)$ by $\widetilde{f}_X(x|p) = \widehat{f}_X(x|p) \cdot 1_{[a,b]}(x) / \int_a^b \widehat{f}_X(x|p)\, dx$, where $[a,b]$ is the largest interval where $\hat{f}_X(x|p)$ is non negative. Note that $\widehat{f}_X(x|p)$ takes negative values only in areas where there are very few or no observations, which, for most samples, corresponds to areas where f_X is equal to, or very close to, zero. Since $\widehat{f}_X(x|p)$ takes negative values only in its tails, in most cases the Hall and Murison transformation will only affect the tails of the estimator, and hence the transformed estimator will either satisfy the constraint, or violate it only very mildly in the tails. To first order, this mild violation does not affect performance of the bootstrap testing procedures. At first sight, it may appear to the reader that instead of turning the constrained estimator into a density, we should rather directly modify the

unconstrained estimator to be a proper density and then change the weights to impose the shape constraint. However in general this would not guarantee that the constrained estimator itself is a proper density.

3 Numerical properties

Here we apply our estimation and testing procedures to several density and regression models.

3.1 Log-Concave Simulations

In the density context, we focused on the log-concavity assumption, which has received increasing attention in the recent error-free literature. We considered four densities f_X: one clearly log-concave, $f_X(x) = \phi(x)$, with ϕ the standard normal density; one just log-concave: $f_X(x) = 0.6\phi(x) + 0.4\phi(x-2)$; one just not log-concave: $f_X(x) = 0.6\phi(x) + 0.4\phi(x-3)$; and one clearly not log-concave: $f_X(x) = 0.6\phi(x) + 0.4\phi(x-4)$. Similar mixtures were considered by Cule, Samworth and Stewart (2010). In each case we generated 200 contaminated samples $W_1, \ldots, W_n$ of size $n = 250$, where $W_i = X_i + U_i$. We took U to be Laplace such that the noise to signal ratio $\mathrm{var}(U)/\mathrm{var}(X)$ equals 20%.

Figure 1 compares, for the first two densities, the deconvolution kernel density estimator $\widehat{f}_X$ with its constrained version $\widehat{f}_X(\cdot \,|\, \widehat{p})$, where a log-concavity constraint is imposed on the intervals $\mathcal{I} = [-2, 2]$ and $\mathcal{I} = [-2, 4]$, respectively. The estimators were calculated using the plug-in bandwidth of Delaigle and Gijbels (2002, 2004). We show four curves, which we call "quantile curves", and which are the estimated curves constructed from the samples which gave the quantiles 0.2, 0.4, 0.6 and 0.8 of the values of the Integrated Squared Errors, $\mathrm{ISE}(\widehat{f}_X) = \int (\widehat{f}_X - f_X)^2$. We also show kernel estimators of the density of $\log[\mathrm{ISE}(\widehat{f}_X)/\,\mathrm{ISE}\{\widehat{f}_X(\cdot \,|\, \widehat{p})\}]$. The graphs in the third column of Figure 1 (and also in the third column of Figure 2, below) illustrate the significant improvement that one can get by incorporating a shape constraint; note that the densities graphed there are skewed to the right. To see this improvement numerically, we computed the Median Integrated Squared Errors (MISE), finding an improvement of 15% for the first standard normal den-

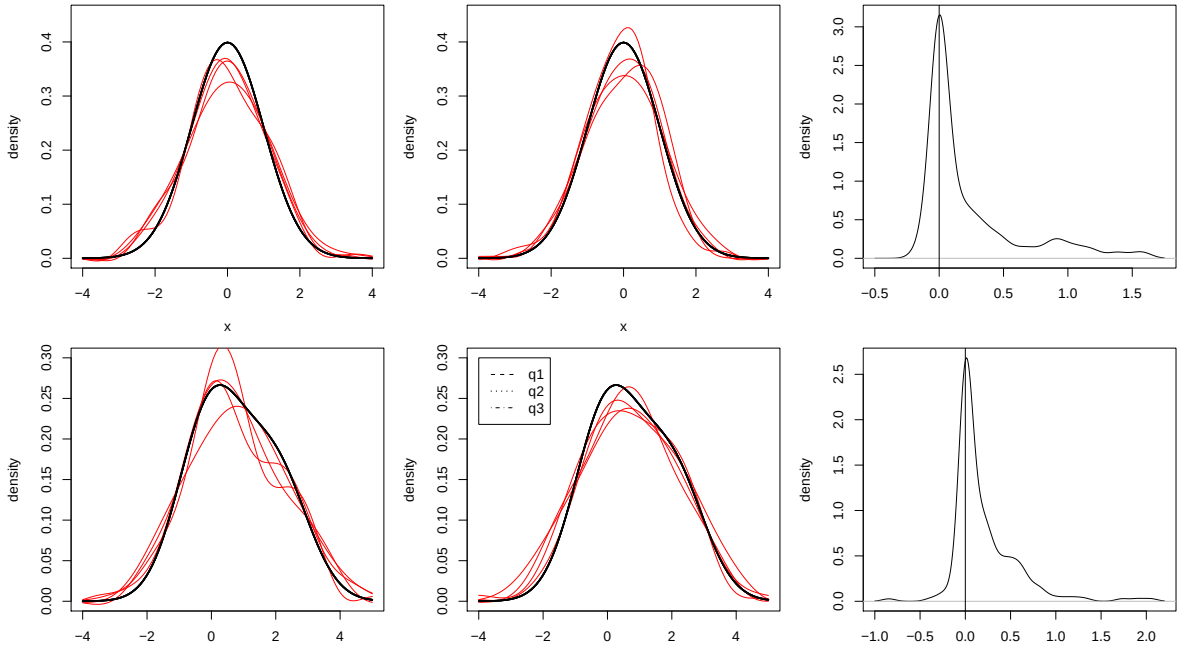

Figure 1: Quantile curves of the estimators: unrestricted estimator of f_X (left) or the log-concave estimators (middle), when $f_X = \phi(x)$ (top) or $f_X(x) = 0.6\phi(x) + 0.4\phi(x-2)$ (bottom). Right: kernel estimators of the density of $\log[\mathrm{ISE}(\widehat{f}_X)/\,\mathrm{ISE}\{\widehat{f}_X(\cdot\,|\,\widehat{p})\}]$; the vertical line indicates the value 0 for reference.

sity (respectively 22% for the mixture density). In addition, in these cases, the percentage of the times that the constrained estimator had smaller MISE was 70% (respectively 73%).

We applied our testing procedure to the four densities. For the first to the fourth densities, we tested for log-concavity on the intervals $\mathcal{I} = [-2, 2]$, $\mathcal{I} = [-2, 4]$, $\mathcal{I} = [-2, 5]$ and $\mathcal{I} = [-2, 5]$, respectively. Since, in the testing problem, our goal is not to estimate f_X, but to test a hypothesis about f_X, where this hypothesis is really a hypothesis on $f_X^{(1)}$, we used the plug-in bandwidth of Delaigle and Gijbels (2002, 2004) adapted to density derivative estimation. With this bandwidth, the proportion of times we rejected H_0 was 0.07 for the first density; 0.05 for the second density; 0.38 for the third density; 0.93 for the fourth density. In other words, the test approximately attained its desired level and had power to detect deviations from the null hypothesis.

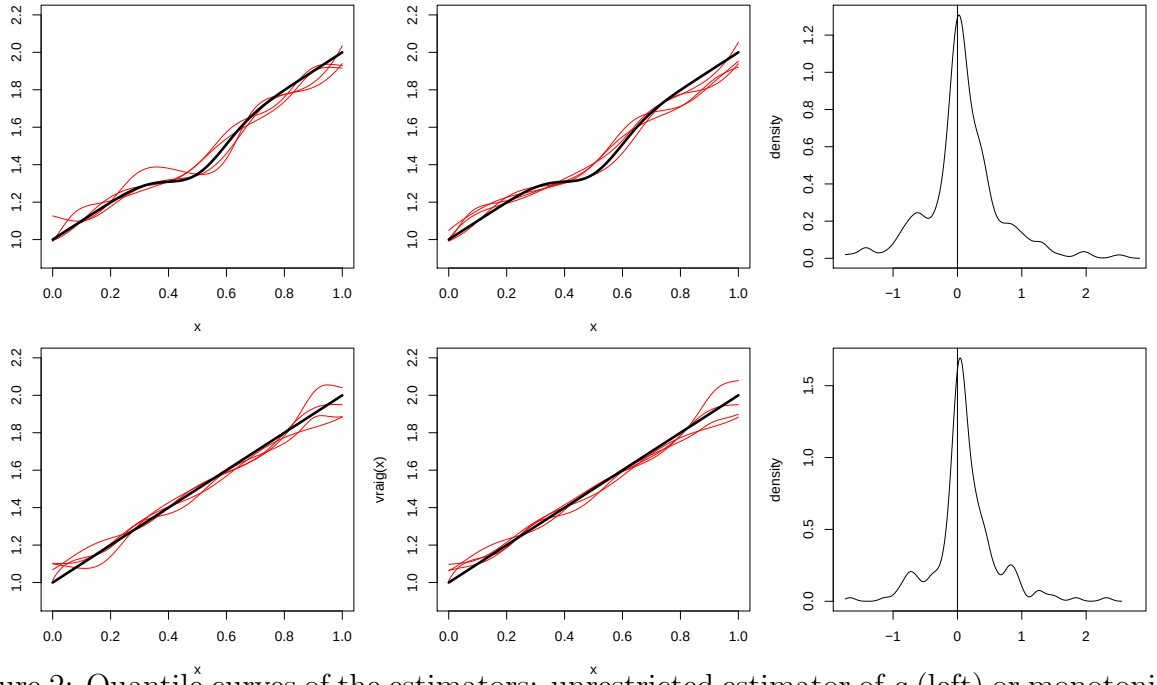

Figure 2: Quantile curves of the estimators: unrestricted estimator of g (left) or monotonized estimators (middle), when the regression curve corresponds to $a = 0.15$ (top) or $a = 0$ (bottom). Right: kernel estimators of the density of $\log[\mathrm{ISE}(\widehat{g})/\,\mathrm{ISE}\{\widehat{g}(\cdot\,|\,\widehat{p})\}]$; the vertical line indicates the value 0 for reference.

3.2 Monotonicity Simulations

Next we considered monotonicity constraints for a family of regression models, taken from Bowman *et al.* (1998), and defined by $g(x) = 1 + x - a\exp\{-50(x-0.5)^2\}$, $\epsilon = \mathrm{Normal}(0, 0.05^2)$, $X = \mathrm{Normal}(0.5, 0.1)$, 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 ($a = 0.45$). Note that the situation in Bowman *et al.* (1998) is much easier as, in their context, not only are the variables X_i observed without error, but they are also fixed and equispaced between 0 and 1. For the testing procedure we generated 200 samples $(X_1, Y_1), \ldots, (X_n, Y_n)$ of size $n = 250$ from each of those regression models, and added Laplace measurement errors U_i to the X_is, such that the noise to signal ratio $\mathrm{var}(U)/\mathrm{var}(X)$ was 20%.

Figure 2 compares the estimator $\widehat{g}$ with its monotonized version $\widehat{g}(\cdot\,|\,\widehat{p})$, when $a = 0$ or $a = 0.15$. For these two examples, the unconstrained estimator satisfied the constraint a number of times. Since in those cases the constrained estimator equalled the unconstrained

one, to illustrate better the amount of improvement that the constrained estimator can offer we generated random samples in the way described above, until we obtained 200 samples for which the unconstrained estimator did not satisfy the constraint. The results discussed in this paragraph correspond to those 200 samples. For both methods, 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 $\log[\mathrm{ISE}(\widehat{g})/\,\mathrm{ISE}\{\widehat{g}(\cdot\,|\,\widehat{p})\}]$ calculated from the 200 samples. We used the bandwidth of Delaigle and Hall (2008). To see numerically the improvement of our method, 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 $a = 0.15$ (respectively 67% when $a = 0$).

We tested monotonicity of these four curves on the interval $\mathcal{I} = [0, 1]$ and calibrated the test for a level $\alpha = 0.05$. We calculated the regression estimators using the bandwidth of Delaigle and Hall (2008), except for the estimator used to generate the bootstrap variables Y_j^*, where, as in Härdle and Marron (1995), we used a bandwidth of the order appropriate to estimate $g^{(2)}$ with a second order kernel. More precisely, we multiplied the bandwidth of Delaigle and Hall (2008) by n^b, where, in the Laplace case, $b = (1/9) - (1/13)$, see section 5. To reduce the occurrence of problems with the tails of the estimated regression curve, we discarded too extreme bootstrap data; in this case, those for which X^* was larger than 1 while Y^* was smaller than 1. For the same reason, to estimate the variance of ϵ, we used the minimum between three estimators: the naive difference-based estimator (see Delaigle and Hall, 2010), the corrected version of Delaigle and Hall (2010) and a SIMEX difference-based estimator constructed using the same principle as the SIMEX bandwidth of Delaigle and Hall (2008). To generate the bootstrap variables X_j^* we also needed a bandwidth for $\widehat{f}_X$, and we used the plug-in bandwidth of Delaigle and Gijbels (2002, 2004). With these choices, the proportion of times we rejected H_0 was 0.06 when $a = 0$; 0.11 when $a = 0.15$; 0.32 when $a = 0.25$ and 0.60 when $a = 0.45$. The power for $a = 0.45$ may seem a bit low, but the problem is very difficult because the variance of the measurement errors is high whereas the dip in the regression curve is narrow. Moreover, the variance of ϵ is very difficult to estimate

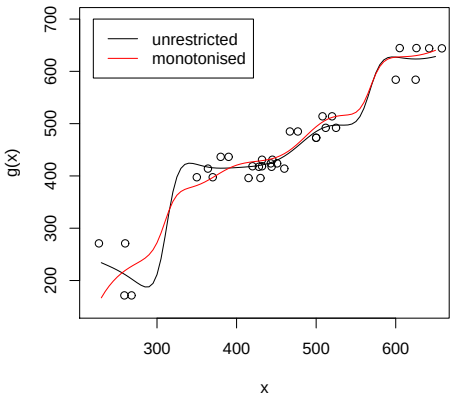

Figure 3: Unrestricted and monotonized estimated curves $g(x)$ for the Wright data.

in this case; when we used its real value instead of the estimated one, the power increased to 0.71.

4 Empirical example

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. As in Delaigle *et al.* (2008), to reduce the variance of ϵ, we take Y_i to be the average of the two Wright readings, and use the two replicated inaccurate measurements to form the sample of W_is (thus each Y_i is used twice). The variance of U is estimated from the replicated mini Wright readings, and for simplicity of calculation we assume a Laplace error. The aim is to determine whether the mini Wright readings are in agreement with the Wright readings. The data are plotted in Figure 3, as well as the two regression estimators (unrestricted and monotonized). Although the unrestricted estimator fluctuates somewhat, an application of our testing procedure does not permit us to reject the hypothesis that the readings by the mini Wright meter are a monotonic function of the readings of the Wright meter, and it is reasonable to infer that the fluctuations are probably artifacts caused by the low sample size, rather than a true characteristic of the curve. In particular the monotonized estimator

seems to improve the unrestricted estimator.

5 Theoretical properties

5.1 Introduction

Since the regression context is considerably more difficult to treat than the density case, in this section we restrict our theoretical study to the regression case. We state three theorems. They describe the uniform convergence of the standard, non-tilted estimator $\widehat{g}(\,\cdot\,|\,p^0)$ and its derivatives (Theorem 1), the rich variety of functions to which the standard estimator can be tilted (Theorem 2), and a lower bound to $D_\rho(\widehat{p})$ (Theorem 3). The lower bound complements an upper bound to $D_\rho(\widehat{p})$, implied by Theorem 2; see (5.8). Following each theorem the implications of that result are outlined and discussed. To our knowledge, no uniform convergence results of the type given in Theorem 1 exist in the setting of errors-in-variables problems. Therefore that result may be of independent interest.

5.2 Properties of the non-pivoted estimator $\widehat{g}(\,\cdot\,|\,p^0)$

If $\mathcal{J}$ denotes an interval where f_X vanishes, then it is easily seen that g is not identifiable at any point in the interior $\mathcal{J}^0$ of $\mathcal{J}$. Therefore, we shall confine attention to properties of $\widehat{g}(x\,|\,p)$ for x lying in intervals $\mathcal{I}$ where $f_X(x) > 0$. Let $\mathcal{I}$ be an interval on which f_X is bounded away from zero, let W, X, U, ϵ and g be as in (2.1), and let $\ell \geq 0$ denote an integer. When discussing convergence of the ℓth derivative, $\widehat{g}^{(\ell)}$, of $\widehat{g}$ to $g^{(\ell)}$ we shall assume that:

> (a) the error ϵ in (2.1) satisfies $E|\epsilon|^r < \infty$, where $r \geq 2$, and has zero mean; (b) f_X has $\ell + 2$ bounded derivatives; (c) ϕ_K is compactly supported and satisfies $\phi_K(0) = 1$; (d) there exist constants $B_1, \alpha > 0$ such that $|\phi_U(t)| \geq B_1\,(1 + |t|)^{-\alpha}$ for all t; and (e) $h = h(n) \to 0$ and, for some $\eta > 0$, $nh^{2(\alpha+\ell+1+\eta)} \geq 1$ for all sufficiently large n. (5.1)

The value of r that we need in (5.1)(a) depends only on α in (5.1)(d) and on η in (5.1)(e), and increases as α increases and η decreases. The resulting moment condition in

(5.1)(a), somewhat stronger than usual, is a consequence of the fact that, in Theorem 1 below, we establish uniform convergence of $\widehat{g}(\cdot \,|\, p^0)$, not just convergence at a single point. The condition on h in (5.1)(e) encompasses the optimal choice of bandwidth for $\widehat{g}^{(\ell)}(\cdot \,|\, p^0)$, which, under the assumption of $\ell + 2$ bounded derivatives of f_X and gf_X, is of size $n^{-1/(2\alpha+2\ell+5)}$. More generally, assumptions (b)–(e) in (5.1) are standard in nonparametric deconvolution; see Fan and Truong (1993).

Given r in (5.1)(a), f_X in (5.1)(b), an integer $\ell \geq 0$ and a constant $B_2 > 0$, we define a class $\mathcal{G} = \mathcal{G}(B_2, \ell, r, f_X)$ of functions g as follows:

$$\begin{array}{l} \mathcal{G} \text{ is the class of } (\ell+2)\text{-times differentiable functions } g \text{ such that, with} \\ \chi_1 = \max_{0 \leq m \leq \ell+2} |(gf_X)^{(m)}| \text{ and } \chi_2(w) = E\{|g(X)|^r \,|\, W = w\}\, f_X(w), \\ \text{(a) } |\chi_1(w)| \leq B_2 \text{ for all } w, \text{ (b) } \chi_2(w) \leq B_2 \text{ for all } w, \text{ and (c) } \int \chi_2(w)\, dw \leq B_2. \end{array} \tag{5.2}$$

Condition (5.2)(a) is a smoothness assumption on g, and, in combination with the smoothness condition on f_X in (5.1)(b), is conventional; and (5.2)(b) and (5.2)(c) are moment conditions on g, which, in conjunction with the moment condition on ϵ in (5.1), are also fairly standard.

Theorem 1. *Assume the model at* (2.1)*, and that* (5.1) *holds for $r \geq 2$ sufficiently large. Assume also that $\mathcal{G}$ satisfies (5.2). Then there exists $B_3 > 0$ such that*

$$\sup_{g \in \mathcal{G}} P\left[\sup_{x \in \mathcal{I}} \left| \widehat{g}^{(\ell)}(x) - g^{(\ell)}(x) \right| \geq B_3 \left\{ \left(nh^{2\alpha+2\ell+1} \right)^{-1/2} (\log n)^{1/2} + h^2 \right\} \right] \to 0\,. \tag{5.3}$$

The convergence rate given in Theorem 1 is best possible, in the following sense. If $\epsilon_1 > 0$ is given; if $g \in \mathcal{G}$, possibly depending on n; if $m = m(n)$ diverges to infinity at rate n^{ϵ_2}, for $\epsilon_2 > 0$; if $x_1, \ldots, x_m$ are equally spaced points in the interval $\mathcal{I}$; and if the $(\ell+2)$th derivatives of f_X and g are continuous, as well as bounded; then there exist small constants $B_8 > 0$ and $B_9 > 0$ such that, for all sufficiently large n,

$$P\left\{ \max_{j=1,\ldots,m} \left| \widehat{g}^{(\ell)}(x_j) - g^{(\ell)}(x_j) \right| \geq B_8 \left(nh^{2\alpha+2\ell+1} \right)^{-1/2} (\log n)^{1/2} + B_9\, h^2 \right\} \geq 1 - \epsilon_1\,. \tag{5.4}$$

Result (5.4) is proved by considering approximations to the joint distributions of $\widehat{g}^{(\ell)}(x_j)$ for $1 \leq j \leq m$.

Remark 5. The constant B_9 can be taken to be zero only if the order h^2 term in the standard formula for the bias of $\widehat{g}^{(\ell)}$, as an estimator of $g^{(\ell)}$, vanishes identically in $\mathcal{I}$, so that

bias equals $o(h^2)$ throughout $\mathcal{I}$. (For example, this would be the case if f_X and g were both constant on $\mathcal{I}$.) However, in this case the quantity $B_9 h^2$ would generally have to be replaced by a term of higher order in h.

To connect Theorem 1 with the problem of estimation under shape restrictions, consider for example the case where the shape constraint is that g is monotone on $\mathcal{I}$. Theorem 1, and the converse at (5.4), imply that

$$\zeta_n \equiv \left(nh^{2\alpha+3}\right)^{-1/2} (\log n)^{1/2} + h^2$$

is a lower bound to the gradient of a linear function g for which the standard estimator $\widehat{g}$ gives, with probability converging to 1, a monotone estimator. That is, if $g = g_n$ is linear over a nondegenerate subinterval of $\mathcal{I}$, and if the gradient of the line exceeds $B\{(nh^{2\alpha+3})^{-1/2} (\log n)^{1/2} + h^2\}$ for a sufficiently large constant B, then the probability that $\widehat{g}^{(1)}(x) > 0$ for all $x \in \mathcal{I}$ converges to 1; and, conversely, if bias is nonzero on a nondegenerate subinterval of $\mathcal{I}$, and if the probability that $\widehat{g}^{(1)}(x) > 0$ for all $x \in \mathcal{I}$ converges to 1, then the gradient of the line exceeds $B\{(nh^{2\alpha+3})^{-1/2} (\log n)^{1/2} + h^2\}$ for some $B > 0$. For example, if the bandwidth $h \asymp n^{-1/(2\alpha+5)}$ is chosen to ensure that $\widehat{g}$ has an optimal pointwise convergence rate, i.e. if the standard deviation and bias terms $(nh^{2\alpha+1})^{-1/2}$ and h^2 are of the same size, then $(nh^{2\alpha+3})^{-1/2} (\log n)^{1/2} + h^2 \asymp n^{-1/\{(2\alpha+5)\}} (\log n)^{1/2}$, which, up to a constant multiplier, represents the shallowest gradient (in asymptotic terms) that the straight line, in a graph of g, can have without it being necessary to tilt the estimator to ensure monotonicity of $\widehat{g}$. Analogous results can be stated for more general constraints of the form

$$\psi(g, g^{(1)}, \ldots, g^{(k)}) > 0 \text{ on a finite interval } \mathcal{I}\,, \tag{5.5}$$

where k is a positive integer and ψ is a smooth function, for example the condition corresponding to the constraint that g is log-concave.

5.3 Properties of the constrained estimator when g does not necessarily satisfy the shape constraint

Theorem 4.1 of Hall and Huang (2001) describes general circumstances where the probability distribution p, in any estimator $\widehat{g}(\,\cdot\,|\,p)$ of the type at (2.11), can be chosen so as to ensure

that the estimator is monotone increasing. We now discuss the very wide variety of other possible shapes that a constrained graph of $\widehat{g}(\,\cdot\,|\,p)$ can enjoy, even in the asymptotic limit. As a prelude to stating the theorem we introduce a general smooth function, $\gamma(\cdot)$, to which we shall show that $\widehat{g}(\,\cdot\,|\,p)$ can converge uniformly. With ℓ denoting a nonnegative integer we assume that:

$$\begin{aligned}&g \text{ and } \gamma, \text{ and their first } \ell+2 \text{ derivatives, are continuous, } g \text{ and } \gamma \text{ are positive}\\ &\text{and bounded away from zero and infinity, and } g \in \mathcal{G}(B_2, \ell, r, f_X) \text{ for some}\\ &B_2, r > 0.\end{aligned} \tag{5.6}$$

Cases where g and γ are not necessarily positive and bounded can be treated too, but at the expense of significantly more complex assumptions and theoretical arguments. Essentially, Theorem 2 below states that the ℓth derivative $\gamma^{(\ell)}(x)$ of any curve $\gamma(x)$ that satisfies (5.6) can be consistently estimated by the tilted estimator $\widehat{g}^{(\ell)}(x\,|\,p)$, defined to be the ℓth derivative of $\widehat{g}(x\,|\,p)$ at (2.11). As before, let $\mathcal{I}$ be an interval on which f_X is bounded away from zero.

Theorem 2. *Assume that* (5.1) *and* (5.6) *hold for $r \geq 2$ sufficiently large. Then there exist probability weights p_j satisfying* (2.4) *and such that: (i) the quantities np_j are uniformly bounded away from zero and infinity, and (ii) for a constant $B_3 > 0$,*

$$P\biggl[\sup_{x \in \mathcal{I}} \bigl|\widehat{g}^{(\ell)}(x\,|\,p) - \gamma^{(\ell)}(x)\bigr| \geq B_3 \Bigl\{\bigl(nh^{2\alpha+2\ell+1}\bigr)^{-1/2} (\log n)^{1/2} + h^2\Bigr\}\biggr] \to 0\,. \tag{5.7}$$

Under the conditions of the theorem we can construct data-dependent weights p_j such that np_j is bounded away from zero and infinity, and $\widehat{g}^{(\ell)}\,(\,\cdot\,|\,p) \to \gamma^{(\ell)}$ uniformly on $\mathcal{I}$. Since γ is virtually arbitrary then this property demonstrates the great flexibility of tilting methods for ensuring that, in the presence of errors in variables, an estimator of a regression mean accurately reflects the shape of a given function γ. In our context of shape restriction, γ is determined implicitly by the function in a given space, $\mathcal{S}$ say (e.g. the space of monotone increasing functions), that is the closest to g. We choose the weights $\widehat{p}_j$ so that $\widehat{g}(\,|\,\widehat{p})$ is the tilted estimator in $\mathcal{S}$ that is closest to $\widehat{g}$. More specifically, the constrained estimation problem we are solving takes the form: find p to minimize $D_\rho(p)$ subject to $\widehat{g}(x\,|\,p) \in \mathcal{S}$ for $x \in \mathcal{I}$, where $\mathcal{I}$ is an interval on which f_X is bounded away from zero.

Remark 6. We are not assuming that the true g lies in $\mathcal{S}$ on $\mathcal{I}$; we are constraining $\widehat{g}(\,\cdot\,|\,p)$ to be in $\mathcal{S}$ without requiring the true regression mean to have that property. In cases where g lies in $\mathcal{S}$ we have $\gamma \equiv g$, but the tilted estimator $\widehat{g}(\,\cdot\,|\,\widehat{p})$ gives better performance than the estimator $\widehat{g}$ at (2.8) because it incorporates the knowledge that we have about g. Whether g lies in $\mathcal{S}$ or not, we claim that, if $p = \widehat{p}$ minimizes $D_\rho(p)$ subject to $\widehat{g}(\,\cdot\,|\,p) \in \mathcal{S}$ on $\mathcal{I}$, then, as shown in Appendix A.2, as $n \to \infty$, there exists $B_4 > 0$ such that

$$P\{D_\rho(\widehat{p}) \le B_4\, n\} \to 1. \tag{5.8}$$

Remark 7. When the constraint is defined in terms of a strict inequality about a smooth function of derivatives of g, for example as at (5.5), then if the true g satisfies these shape constraint, the value of $D_\rho(\widehat{p})$ is an order of magnitude smaller than the $O_p(n)$ bound given at (5.8). More precisely, we show in appendix A.2 that, as $n \to \infty$,

$$P\{D_\rho(\widehat{p}) \le \epsilon n\} \to 1 \quad \text{for all } \epsilon > 0\,. \tag{5.9}$$

Remark 8. The case of a constraint defined in terms of a non-strict inequality, for example (5.5) with $>$ replaced by $\ge$, is statistically the most difficult, in the sense of requiring, generally speaking, the greatest amount of tilting, i.e. the largest value of $D_\rho(\widehat{p})$.

5.4 Consistent hypothesis testing

It remains to prove that the test suggested in section 2.2 is consistent. In Theorem 3, below, we give a formal result about the asymptotic behavior of the test statistic $D_\rho(\widehat{p})$ under the alternative that $g(\cdot)$ is not monotone nondecreasing on $\mathcal{I}_0$. As we shall show, it implies that the bootstrap-based hypothesis test suggested in section 2.2 gives statistically consistent results.

Theorem 3. *Assume the conditions of Theorem 2 pertaining to g and to the case $\ell = 1$, and that g is not monotone nondecreasing on $\mathcal{I}$. Then,*

$$\lim_{\epsilon \to 0} \liminf_{n \to \infty} P\{D_\rho(\widehat{p}) \ge n\epsilon\} = 1\,. \tag{5.10}$$

Remark 9. Although Theorem 3 treats only the case where the constraint is one of monotonicity, similar arguments can be used to prove that (5.10) also holds under $H_0 : g \in \mathcal{S}$ on $\mathcal{I}$ and $H_1 : g \notin \mathcal{S}$ on $\mathcal{I}$, where $\mathcal{S}$ is a set of functions satisfying a constraint defined in terms of an inequality such as that at (5.5). The more general version of Theorem 3, together with the bootstrap version of (5.9), imply that if the true $g \notin \mathcal{S}$ on $\mathcal{I}$ then the probability that the bootstrap test in section 2.2, which is based on an empirical critical point $\widehat{\xi}_\alpha$, results in rejection of H_0, converges to 1 as $n \to \infty$. That is, as shown in Appendix A.3,

$$P\big\{D_\rho(\widehat{p}) > \widehat{\xi}_\alpha\big\} \to 1\,. \tag{5.11}$$

Remark 10. If the true g is strictly monotone increasing on $\mathcal{I}$, and the conditions of Theorem 3 hold, then as shown in Appendix A.3, the probability that the hypothesis that g is monotone nondecreasing is rejected converges to zero:

$$P\{D_\rho(\widehat{p}) > \widehat{\xi}_\alpha\} \to 0\,. \tag{5.12}$$

This result can be generalized to constraints that have the form at (5.5).

Remark 11. The case where the true g is monotone nondecreasing, but not strictly so, is excluded above. It is problematic to treat because of the example where g is constant on at least part of $\mathcal{I}$. In the case of a constant g the probability that our test rejects the null hypothesis converges to neither 0 nor 1.

6 Discussion

We have developed a general methodology for doing measurement error analysis subject to a shape constraint, as well as a method for testing whether the constraint is actually satisfied. Our simulations as well as our example show that in cases where a standard measurement error estimator violates the constraint when it really holds, imposing the constraint provides improvement in estimation accuracy.

While we have concentrated on log-concavity of a density function and monotonicity of a regression function, many other shape constraints can be handled by our methodology. The key is to find a positive penalty function, such as at (2.12) for log-concavity and (2.13)

for monotonicity, which increases as the estimated curve departs further from the shape constraint.

A Appendix

A.1 Proof of Theorem 1

For $k = 0, 1$, define

$$T_{k\ell}(x) = \frac{1}{nh^{\ell+1}} \sum_{j=1}^{n} Y_j^k \, K_U^{(\ell)}\left(\frac{x - W_j}{h}\right).$$

Since, by (5.1)(b) and (5.2)(a), f_X and gf_X each have $\ell + 2$ uniformly bounded derivatives, then

$$E\{T_{k\ell}(x)\} = g_k^{(\ell)}(x) + O\big(h^2\big)\,, \tag{A.1}$$

uniformly in $k = 0, 1$ and $x \in \mathcal{I}$, where $g_0 = f_X$ and $g_1 = gf_X$. Also,

$$|T_{k\ell}(x_1) - T_{k\ell}(x_2)| \le \frac{|x_1 - x_2|}{nh^{\ell+2}} \left(\sup \left|K_U^{(\ell+1)}\right|\right) \sum_{j=1}^{n} |Y_j|^k\,. \tag{A.2}$$

For a constant $C_1 > 0$ and all x,

$$|K_U^{(\ell+1)}(x)| = \frac{1}{2\pi} \left| \int t^{\ell+1} \, e^{-itx} \, \phi_K(t) \, \phi_U(t/h)^{-1} \, dt \right| \le C_1 \, h^{-\alpha}\,;$$

here we used (5.1)(d) and the fact that, by (5.1)(c), ϕ_K is compactly supported. Therefore (A.2) implies that

$$|T_{k\ell}(x_1) - T_{k\ell}(x_2)| \le \frac{C_1 \, |x_1 - x_2|}{nh^{\ell+2+\alpha}} \sum_{j=1}^{n} |Y_j|^k\,, \tag{A.3}$$

uniformly in $k = 0, 1$ and $x_1, x_2 \in \mathcal{I}$.

Since, by (5.1)(e), $h \to 0$ and $nh^{2(\alpha+\ell+1+\eta)} \ge 1$ for all sufficiently large n, then for large n, $(h^{\ell+2+\alpha})^{-1} \le n^{\tau}$, where

$$\tau = \frac{\ell + 2 + \alpha}{2\,(\alpha + \ell + 1)}\,.$$

Therefore (A.3) implies that

$$|T_{k\ell}(x_1) - T_{k\ell}(x_2)| \le C_1 \, n^{\tau-1} \, |x_1 - x_2| \sum_{j=1}^{n} |Y_j|^k\,.$$

Hence, if x_1 and x_2 are constrained by

$$|x_1 - x_2| \le n^{-\tau-1} \tag{A.4}$$

then

$$|T_{k\ell}(x_1) - T_{k\ell}(x_2)| \le \frac{C_1}{n^2} \sum_{j=1}^{n} |Y_j|^k \,. \tag{A.5}$$

Put

$$\sigma_{k\ell}(x)^2 = \mathrm{var}\{T_{k\ell}(x)\} = O\big\{(nh^{2\alpha+2\ell+1})^{-1}\big\}\,, \tag{A.6}$$

where the second identity holds uniformly in $k = 0, 1$ and $x \in \mathcal{I}$; see (A.12), below, for explicit calculation. Define $S_{k\ell}(x) = (1-E)\,T_{k\ell}(x)/\sigma_{k\ell}(x)$, so that $S_{k\ell}(x)$ has zero mean and unit variance, and assume for the present that, with $\lambda_n = (\log n)^{1/2}$,

$$\begin{aligned} &P\{S_{k\ell}(x) > c\,\lambda_n\} = (2\pi)^{-1/2}\,(c\lambda_n)^{-1}\,n^{-c^2/2}\,\{1 + o(1)\} \text{ uniformly in} \\ &k = 0, 1 \text{ and } x \in \mathcal{I}. \end{aligned} \tag{A.7}$$

Let $\mathcal{I}_n$ denote a set consisting of $O(n^{c^2/2})$ elements of $\mathcal{I}$. Then (A.1), (A.6) and (A.7) together imply that if $C_2 > c$ is sufficiently large then

$$P\bigg[\sup_{x\in\mathcal{I}_n} \big|T_{k\ell}(x) - g_k^{(\ell)}(x)\big| \ge C_2\,\Big\{\big(nh^{2\alpha+2\ell+1}\big)^{-1/2}\,(\log n)^{1/2} + h^2\Big\}\bigg] \to 0 \tag{A.8}$$

as $n \to \infty$. Combining (A.5) and (A.8), and noting that, by a law of large numbers and since $E(|Y|) < \infty$ (see (5.1)(a) and (5.2)(b)),

$$P\bigg(n^{-1}\sum_{j=1}^{n} |Y_j|^k > E|Y|^k + \eta\bigg) \to 0$$

for each $\eta > 0$, then if (A.4) holds and, in (A.7), we take $c > 2\,(\tau+1)^{1/2}$, we have, for any $C_3 > C_1\,E|Y| + C_2$,

$$P\bigg[\sup_{x\in\mathcal{I}} \big|T_{k\ell}(x) - g_k^{(\ell)}(x)\big| \ge C_3\,\Big\{\big(nh^{2\alpha+2\ell+1}\big)^{-1/2}\,(\log n)^{1/2} + h^2\Big\}\bigg] \to 0\,. \tag{A.9}$$

Using (A.9), and using Taylor expansion, we deduce that Theorem 1 holds if $nh^{2\alpha+2\ell+1} \to \infty$.

It remains to derive (A.7), which we do by following the argument used to prove Theorem 1 of Rubin and Sethuraman (1965). Note that $n^{1/2}S_{k\ell}(x)$ equals a sum of n independent random variables, each with the distribution of

$$R = \{(1-E)\,R_0\}/(\mathrm{var}R_0)^{1/2}\,, \tag{A.10}$$

where $R_0 = R_0(x) = h^{-\ell-1}\,R_1$ and

$$R_1 = Y^k\,K_U^{(\ell)}\{(x-W)/h)\}\,. \tag{A.11}$$

In our context, the random variables denoted by X_{ni} in Theorem 1 of Rubin and Sethuraman (1965) are all independent and identically distributed as $R(x)$. This ensures that Rubin and Sethuraman's (1965) assumption (8) holds, with their a and A both equal to 1, and their assumption (7) holds with (in their notation) $\sigma_{ni}^2 = 1$.

Recall the definition of g_k above (A.2), and observe that, for $k = 0, 1$,

$$E(R_0) = \int g_k^{(\ell)}(x - hu)\, K(u)\, du = g_k(x) + O\big(h^2\big)\,,$$

uniformly in $x \in \mathcal{I}$ as $n \to \infty$. Also, defining $\gamma_0(w) = E\{g(X)^2 \,|\, W = w\}$ and $\sigma^2 = \mathrm{var}(\epsilon)$, we have, in the case $k = 1$:

$$\begin{aligned}
\mathrm{var}(R_0) &= h^{-(2\ell+1)} \int \{\gamma_0(x - hu) + \sigma^2\}\, f_W(x - hu)\, K_U^{(\ell)}(u)^2\, du - (ER_0)^2 \\
&= h^{-(2\ell+1)}\, \{\gamma_0(x) + \sigma^2\}\, f_W(x) \int K_U^{(\ell)}(u)^2\, du + o\big(h^{-(2\ell+1-2\alpha)}\big) \\
&\sim \text{const.}\, h^{-(2\ell+2\alpha+1)}\, \{\gamma_0(x) + \sigma^2\}\, f_W(x)\,, \qquad \text{(A.12)}
\end{aligned}$$

uniformly in $x \in \mathcal{I}$. An identical formula, but with $\gamma_0(x) + \sigma^2$ replaced by 1, holds in the case $k = 0$. Hence, for $k = 0, 1$,

$$\begin{aligned}
&R(x) = h^{\alpha-(1/2)}\, \big\{R_1(x) - h^{\ell+1}\, a(x,h)\big\}\big/ b(x,h), \text{ where } |a(x,h)| \text{ and } b(x,h) \\
&\text{are bounded, and } b(x,h) \text{ is bounded above zero, uniformly in } x \in \mathcal{I} \text{ as} \\
&n \to \infty. \qquad \text{(A.13)}
\end{aligned}$$

Using (5.1) we deduce that

$$2\pi \left|K_U^{(\ell)}(x)\right| = \left| \int e^{-itx}\, (-it)^\ell\, \phi_K(t)\, f_U(t/h)^{-1}\, dt \right| \le \text{const.}\, h^{-\alpha}\,, \qquad \text{(A.14)}$$

uniformly in all x. Combining (A.13) and (A.14) we see that, defining $a_n = n^{1/2}/(\log n)^\beta$ for any fixed $\beta > 0$, if $k = 0$ in the definition of R_1 at (A.11) (and thence in the definition of R at (A.10)) then the event $|R(x)| \le a_n$ is implied by $h^{-1/2} \le \text{const.}\, n^{1/2}/(\log n)^\beta$, provided the constant is sufficiently small; and that this inequality holds for all sufficiently large n since, by (5.1)(e), $nh^{2(\alpha+\ell+1+\eta)} \ge 1$ for some $\eta > 0$, or equivalently, $h^{-(\alpha+\ell+1)} \le \text{const.}\, n^{(1/2)-\eta_1}$ for some $\eta_1 > 0$. Likewise, if $k = 1$ in (A.11) then $|R(x)| \le a_n$ is implied by $|Y| \le n^{\eta_2}$ for some $\eta_2 > 0$, and if $C_4 > 0$ is given, and we choose r, in (5.1)(a) and (5.2), sufficiently large, then $P(|Y| \le n^{\eta_2}) = 1 - O(n^{-C_4})$. Therefore $\inf_{x\in\mathcal{I}} P\{|R(x)| \le a_n\} = 1 - O(n^{-C_4})$, and similarly it can be proved that, again for sufficiently large r, $\sup_{x\in\mathcal{I}} E[|R(x)|^2\, I\{|R(x)| >$

$a_n\}] = O(n^{-C_4})$. Combining these results we deduce that:

> if $k = 0$ then there exists $n_0 \geq 1$ such that $P\{\sup_x |R(x)| \leq a_n\} = 1$ for all $n \geq n_0$; and if $k = 1$ and $C_4 > 0$ is given then we can choose r, in (5.1)(a) and (5.2), so large that $\inf_{x \in \mathcal{I}} P\{|R(x)| \leq a_n\} = 1 - O(n^{-C_4})$ and $\sup_{x \in \mathcal{I}} E[|R(x)|^2 \, I\{|R(x)| > a_n\}] = O(n^{-C_4})$. (A.15)

This property, when $\beta = 1/2$, implies results (30), (31) and (43) of Rubin and Sethuraman (1965). Therefore, although condition (11) of Rubin and Sethuraman (1965) does not hold in our context, since it is used only for their (30), (31) and (43) then we do not require it.

Result (A.15), in the case $\beta = 3/2$, also implies property (9) of Rubin and Sethuraman (1965). Since Rubin and Sethuraman's (1965) conditions (7), (8), (9) and (11) are all that is needed for their Theorem 1 then that result holds in our case, uniformly in $x \in \mathcal{I}$ and for $k = 0, 1$, where it implies (A.7). This completes the proof of our Theorem 1.

A.2 Proof of Theorem 2

For brevity we treat only the case $\ell = 0$. Take $p_j = \phi(W_j)/\{\sum_j \phi(W_j)\}$, where ϕ denotes a function satisfying

$$C_1 \leq \phi(x) \leq C_2 \quad \text{for all } x\,, \tag{A.16}$$

with $0 < C_1 < C_2 < \infty$ and $E\{\phi(W)\} = 1$. (The latter condition serves only to define concisely the scale of ϕ.) Then the constraints at (2.4) hold, and part (i) of Theorem 2 follows from (A.16).

Using the definitions of $S_j(x)$ and $\widehat{g}(x \,|\, p)$ at (2.10) and (2.11), respectively, and defining $\delta = (nh^{2\alpha+1})^{-1/2} \, (\log n)^{1/2}$ and $\psi_1(w) = E\{g(X) \,|\, W = w\}$, we see that:

$$\begin{aligned}
\widehat{g}(x \,|\, p) &= n \sum_{j=1}^n p_j \, S_j(x) \, Y_j = \big\{1 + O_p\big(n^{-1/2}\big)\big\} \sum_{j=1}^n \phi(W_j) \, S_j(x) \, Y_j \\
&= \big\{1 + O_p\big(n^{-1/2}\big)\big\} \, \frac{\sum_k \phi(W_k) \, Y_k \, K_U\{(x - W_k)/h\}}{\sum_k K_U\{(x - W_k)/h\}} \\
&= \frac{E[\phi(W) \, \psi_1(W) \, K_U\{(x - W)/h\}]}{E[K_U\{(x - W)/h\}]} + O_p(\delta)\,,
\end{aligned} \tag{A.17}$$

uniformly in $x \in \mathcal{I}$. (In (A.17) the first identity comes from the definition of $\widehat{g}(\,\cdot \,|\, p)$; the second and third from the definition of p_j, using the fact that $\sum_j \phi(W_j) = n \, E\{\phi(W)\} +$

$O_p(n^{1/2})$; and the third from property (A.16), using the arguments employed to prove Theorem 1.) As h tends to zero,

$$h^{-1}\,E[K_U\{(x-W)/h\}] \;=\; f_X(x) + O\big(h^2\big)\,, \tag{A.18}$$

$$h^{-1}\,E[\phi(W)\,\psi_1(W)\,K_U\{(x-W)/h\}] \;=\; \psi_2(x) + O\big(h^2\big)\,, \tag{A.19}$$

where

$$\psi_2(x) = \frac{1}{2\pi}\int e^{-itx}\,(\phi\psi_1 f_W)^{\rm Ft}(t)\,\phi_U(t)^{-1}\,dt\,. \tag{A.20}$$

Now,

$$\begin{aligned}(\phi\psi_1 f_W)^{\rm Ft}(t) &= \int e^{itw}\,E\{g(X)\,|\,W=w\}\,(\phi f_W)(w)\,dw \\ &= \int e^{itw}\,\phi(w)\,dw \int g(x)\,f_X(x)\,f_U(w-x)\,dx \\ &= \int e^{itw}\,\phi(w)\,\{(gf_X)*f_U\}(w)\,dw\,.\end{aligned}$$

If we define

$$\phi = \{(gf_X)*f_U\}^{-1}\,\{(\gamma f_X)*f_U\} \tag{A.21}$$

then $(\phi\psi_1 f_W)^{\rm Ft} = (\gamma f_X)^{\rm Ft}\,\phi_U$, and hence, by (A.20), $\psi_2 = \gamma f_X$. From this result and (A.17)–(A.19) we deduce part (ii) of Theorem 2.

It remains to ensure that ϕ satisfies (A.16). If $0 < C_4 \le \min(g,\gamma) \le \max(g,\gamma) \le C_5$ then, using (A.21) and the fact that f_U and f_X are probability densities,

$$\phi \le C_4^{-1}\,C_5\,(f_X * f_U)^{-1}\,(f_X * f_U) = C_4^{-1}\,C_5\,,$$

and similarly $\phi \ge C_5^{-1}\,C_4$, where we interpret 0/0 as 1 and in particular define ϕ to be an arbitrary positive constant at points where $f_X * f_U = 0$. This establishes (A.16).

A.2.1 Argument leading to (5.8)

To appreciate why (5.8) is correct, suppose initially that the values of np_j are constrained to satisfy $B_6 \le np_j \le B_7$ for all j, where $0 < B_6 < B_7 < \infty$. Then it can be shown directly from (2.6) and (2.7) that $0 \le D_\rho(p) \le B_4\,n$, where B_4 depends only on B_6, B_7 and ρ. Therefore Theorem 2 asserts the existence of B_4 such that, with probability converging to 1 as $n \to \infty$, there exists at least one p for which $D_\rho(p) \le B_4\,n$ and $\widehat{g}(x\,|\,p) \in \mathcal{S}$ for all $x \in \mathcal{I}$. Hence, since $\widehat{p}$ minimizes D_ρ subject to $\widehat{g}(x\,|\,p) \in \mathcal{S}$, then $D_\rho(\widehat{p}) \le D_\rho(p) \le B_4\,n$ with probability converging to 1. Therefore (5.8) holds.

A.2.2 Argument leading to (5.9)

To appreciate why, consider the case where the true g is strictly monotone nondecreasing; more general constraints, such as that at (5.5), can be treated similarly. Then (5.9) is readily proved using a slight modification of the argument leading to Theorem 2, noting that a strictly increasing function γ for which the ratio g/γ is no further than a small, fixed positive number δ from 1 can be achieved in the limit by tilting the standard estimator using weights p_j for which np_j differs from 1 by no more than $\delta\,\{1 + O_p(n^{-1/2})\}$, uniformly in j. (Here it is necessary only to appeal to formula (A.21).) Therefore $P\{D_\rho(\widehat{p}) \le \epsilon(\delta)\,n\} \to 1$, where $\epsilon(\delta)$ depends only on g, ρ and δ and converges to zero as δ decreases. Property (5.9) follows immediately.

A.3 Proof of Theorem 3

First we state a lemma. Let $\alpha_0 \ge \alpha + 1$ be an integer, and let ψ_1 denote a nonnegative function with support equal to $[-1, 1]$ and with at least α_0 bounded derivatives on the real line. Define $\psi(x \,|\, x_1, x_2) = \psi_1[\{x - \frac{1}{2}\,(x_1 + x_2)\}/(x_2 - x_1)]$, where x_1 and x_2 are in $\mathcal{I}$ and $x_1 \ne x_2$.

Lemma 1. *Assume that $E(\epsilon^2) < \infty$ and $E(\epsilon) = 0$, that $|\phi_U(t)| \ge B_1\,(1 + |t|)^{-\alpha}$ for all t, that the functions $\chi_3(w) = E\{g(X)^2 \,|\, W = w\}$ and f_W satisfy $\sup\{(\chi_3 + 1)\,f_W\} < \infty$, and that f_X is continuous and bounded away from zero on $[x_1, x_2]$. Then,*

$$\left| \int \{\widehat{g}(x \,|\, p) - \widehat{g}(x \,|\, p^0)\}\,\psi(x \,|\, x_1, x_2)\,\widehat{f}_X(x)\,dx \right|^2 = O_p\left\{ n^{-1} \sum_{j=1}^n (np_j - 1)^2 \right\}, \qquad \text{(A.22)}$$

uniformly in n-variate probability distributions p.

Proof of Lemma 1. To derive Lemma 1, observe that the function

$$\kappa(w \,|\, h) = \frac{1}{h} \int K_U\left(\frac{x - w}{h}\right) \psi(x \,|\, x_1, x_2)\,dx$$

has Fourier transform $\kappa^{\mathrm{Ft}}(t \,|\, h) = \phi_K(ht)\,\psi^{\mathrm{Ft}}(t \,|\, x_1, x_2)\,\phi_U(t)^{-1}$. Since $\psi(\cdot \,|\, x_1, x_2)$ has $\alpha_0 \ge \alpha + 1$ bounded derivatives then, for a constant $C_6 > 0$,

$$\left|\psi^{\mathrm{Ft}}(t \,|\, x_1, x_2)\,\phi_U(t)^{-1}\right| \le C_6\,(1 + |t|)^{-\alpha_0}\,(1 + |t|)^{\alpha} \le C_6\,(1 + |t|)^{-1}\,.$$

Therefore, defining $C_7 = \sup\{(\chi_3 + \sigma^2)\, f_W\}$, we have:

$$\begin{aligned} E\big\{Y^2\,\kappa(W\,|\,h)^2\big\} &= \int \big\{\chi_3(w) + \sigma^2\big\}\, f_W(w)\,\kappa(w\,|\,h)^2\,dw \\ &\le C_7 \int \kappa(w\,|\,h)^2\,dw = (2\pi)^{-1}\,C_7 \int \big|\kappa^{\mathrm{Ft}}(t\,|\,h)\big|^2\,dt \\ &\le (2\pi)^{-1}\,C_6^2\,C_7 \int (1+|t|)^{-2}\,dt < \infty\,. \end{aligned} \tag{A.23}$$

Furthermore, the left-hand side of (A.22) equals the square of:

$$\begin{aligned} &\int \{\widehat{g}(x\,|\,p) - \widehat{g}(x\,|\,p^0)\}\,\psi(x\,|\,x_1,x_2)\,\widehat{f}_X(x)\,dx \\ &\quad = \sum_{j=1}^n (np_j - 1)\,Y_j \int S_j(x)\,\psi(x\,|\,x_1,x_2)\,\widehat{f}_X(x)\,dx \\ &\quad = \frac{1}{n}\sum_{j=1}^n (np_j - 1)\,Y_j\,\frac{1}{h}\int K_U\!\left(\frac{x - W_j}{h}\right)\psi(x\,|\,x_1,x_2)\,dx \\ &\quad = \frac{1}{n}\sum_{j=1}^n (np_j - 1)\,Y_j\,\kappa(W_j\,|\,h)\,. \end{aligned}$$

Therefore, by the Cauchy-Schwarz inequality, the left-hand side of (A.22) does not exceed the value of

$$\left\{\frac{1}{n}\sum_{j=1}^n (np_j - 1)^2\right\}\left\{\frac{1}{n}\sum_{j=1}^n Y_j^2\,\kappa(W_j\,|\,h)^2\right\} = O_p\left\{\frac{1}{n}\sum_{j=1}^n (np_j - 1)^2\right\},$$

where the identity follows from (A.23) and holds uniformly in p. This proves Lemma 1. □

To establish Theorem 3, if g is not monotone nondecreasing on $\mathcal{I} = [a,b]$ then there exists a sequence of subintervals of $\mathcal{I}$, say $[x_{k1}, x_{k2}]$ for $1 \le k \le m$ where $a \le x_{11} \le \ldots \le x_{m2} \le b$ and $x_{k1} \le x_{k2}$ for each k, where on each interval g is strictly decreasing; and there exists a constant $C_8 > 0$; such that

$$\inf_{\text{mon. nondecr. }\gamma}\ \max_{1\le k\le m}\left|\int \{\gamma(x) - g(x)\}\,\psi(x\,|\,x_{k1},x_{k2})\,f_X(x)\,dx\right| \ge C_8\,, \tag{A.24}$$

with the infimum taken over all monotone nondecreasing functions γ on $\mathcal{I}$. To appreciate why (A.24) must hold, for all functions g that are not monotone nondecreasing on $\mathcal{I}$, there exists $C_9 > 0$ such that

$$\inf_{\text{mon. nondecr. }\gamma}\int_{\mathcal{I}} |\gamma(x) - g(x)|\,f_X(x)\,dx \ge C_9\,.$$

Hence, using an argument by contradiction, we see that if $a = y_0 < y_1 < \ldots < y_m = b$ is a decomposition of $\mathcal{I}$ into a regular grid of edge width $\delta = (b-a)/m$ then we can choose δ so small that

$$\inf_{\text{mon. nondecr. } \gamma} \sum_{k=1}^m \left| \int \{\gamma(x) - g(x)\}\, \psi(x \,|\, y_{k-1}, y_k)\, f_X(x)\, dx \right| \geq C_{10}\,,$$

for a positive constant C_{10}. Therefore,

$$\inf_{\text{mon. nondecr. } \gamma} \max_{1 \leq k \leq m} \left| \int \{\gamma(x) - g(x)\}\, \psi(x \,|\, y_{k-1}, y_k)\, f_X(x)\, dx \right| \geq C_{10}/(m)\,,$$

which implies (A.24).

It follows from Theorem 2 that, under the assumptions of Theorem 3, the event $\mathcal{E}$ that there exists a value $\widehat{p}$ of p such that $\widehat{g}(\,\cdot \,|\, \widehat{p})$ is monotone nondecreasing on $\mathcal{I}$, satisfies $P(\mathcal{E}) \to 1$ as $n \to \infty$. If $\mathcal{E}$ obtains, let $\widehat{p}$ denote a p that minimizes $D_\rho(p)$ subject to $\widehat{g}(\,\cdot \,|\, p)$ being monotone nondecreasing. Then, since $\widehat{g}(\,\cdot \,|\, p^0)$ and $\widehat{f}_X$ converge in probability to g and f_X, respectively, uniformly on compact intervals where f_X is bounded away from zero (in the case of $\widehat{f}_X$ this convergence can be proved as in the derivation of Theorem 1 in the case $\ell = 0$), then (A.24) implies that:

$$P\left\{ \max_{1 \leq k \leq m} \left| \int \{\widehat{g}(x \,|\, \widehat{p}) - \widehat{g}(x \,|\, p^0)\}\, \psi(x \,|\, x_{k1}, x_{k2})\, \widehat{f}_X(x)\, dx \right| \geq \tfrac{1}{2}\, C_8 \right\} \to 1\,.$$

Therefore, by Lemma 1,

$$\lim_{\epsilon \to 0} \liminf_{n \to \infty} P\left\{ \sum_{j=1}^n (n\widehat{p}_j - 1)^2 \geq n\epsilon \right\} = 1\,. \tag{A.25}$$

If D_ρ is as defined as at (2.6) or (2.7) then it can be shown by a Lagrange multiplier argument, differentiating $D_\rho(p) + \lambda\,(\sum_j p_j - 1)$, that, for any given $p = (p_1, \ldots, p_n)$, the value of $D_\rho(p)$ is decreased by decreasing components p_j for which $p_j > n^{-1}$ and increasing components for which $p_j < n^{-1}$. Hence, if there exists a distribution p such that $\widehat{g}(\,\cdot \,|\, p)$ is nondecreasing on $\mathcal{I}$ and $C_{10}^{-1} \leq np_j \leq C_{10}$, where $C_{10} > 0$ is given, then the distribution $\widehat{p} = (\widehat{p}_1, \ldots, \widehat{p}_n)$ that minimizes $D_\rho(p)$ subject to $\widehat{g}(\,\cdot \,|\, p)$ being nondecreasing also satisfies $C_{10}^{-1} \leq n\widehat{p}_j \leq C_{10}$. Call this property ($\mathrm{P}_1$), and let ($\mathrm{P}_2$) denote the property that the distance measure D_ρ, defined by (2.6) or (2.7), satisfies $C_{11} \sum_j (np_j - 1)^2 \leq D_\rho(p) \leq C_{12} \sum_j (np_j - 1)^2$ uniformly in n-variate probability distributions p such that $C_{10}^{-1} \leq np_j \leq C_{10}$, where C_{11} and C_{12} depend only on ρ and C_{10}. Property (P_2) can be derived by elementary calculus. Together, (P_1) and

(P_2) imply that, if

> there exists $C_{10} > 1$ such that, with probability converging to 1 as $n \to \infty$, there is a probability distribution p such that $C_{10}^{-1} \le np_j \le C_{10}$ for all j and $\widehat{g}(\,\cdot\,|\,p)$ is monotone nondecreasing on $\mathcal{I}$, (A.26)

then with probability converging to 1 the particular distribution $\widehat{p}$ that minimizes D_ρ subject to monotonicity satisfies

$$P\biggl\{C_{11}\sum_{j=1}^{n}(n\widehat{p}_j - 1)^2 \le D_\rho(\widehat{p})\biggr\} \to 1 \tag{A.27}$$

as $n \to \infty$. Property (A.26) is implied by Theorem 2 in the case $\ell = 1$, on taking γ in the theorem to be a strictly monotone increasing function. In this setting, (A.25) and (A.27) imply (5.10).

A.3.1 Argument leading to (5.11)

Step (i) of the bootstrap algorithm in section 2.2 involves constructing an estimator $\widehat{g}(\,\cdot\,|\,\widehat{p})$ which lies in $\mathcal{S}$ on $\mathcal{I}$. Step (iii) takes this function to be the true g in the bootstrap step, and then tilts the resulting bootstrap estimator, $\widehat{g}^*(\cdot\,|\,p)$, to $\widehat{g}^*(\cdot\,|\,\widehat{p}^*)$, so that the latter lies in $\mathcal{S}$ on $\mathcal{I}$. As argued at (5.9), this produces a tilting distance $D_\rho(\widehat{p}^*)$ that is of smaller order than n, in the following sense: As n diverges,

$$P\{D_\rho(\widehat{p}^*) \le \epsilon n\,|\,\mathcal{D}\} \to 1 \quad \text{in probability, for all } \epsilon > 0\,. \tag{A.28}$$

Recall that $\mathcal{D}$, introduced in section 2.2, denotes the full data set. A formal proof of (A.28) differs in only minor respects from that of (5.9), and so is not given here. Result (A.28) implies that the critical point $\widehat{\xi}_\alpha$ satisfies $P(\widehat{\xi}_\alpha \le n\epsilon) \to 1$ for each $\epsilon > 0$. This property and (5.10) imply (5.11).

A.3.2 Argument leading to (5.12)

To appreciate why, note first that $P(\widehat{p} = p^0) \to 1$ as $n \to \infty$; this follows from the fact that, by Theorem 1, $\sup_{x\in\mathcal{I}} |\widehat{g}^{(\ell)}(x\,|\,p^0) - g^{(\ell)}(x)| \to 0$ in probability. In consequence, $P(\widehat{p}^* = p^0\,|\,\mathcal{D}) \to 1$ in probability as $n \to \infty$. Therefore, $P\{D_\rho(\widehat{p}^*) = 0\,|\,\mathcal{D}\} \to 1$, and so the value of the critical point satisfies $P(\widehat{\xi}_\alpha = 0) \to 1$ as $n \to \infty$. Hence, with probability converging to 1 the statement $D_\rho(\widehat{p}) > \widehat{\xi}_\alpha$ is identical to $0 > 0$, and so fails to hold. Therefore (5.12) obtains. This argument, too, can be generalized to constraints that have the form at (5.5).

References

Antoniadis, A., Bigot, J. and Gijbels, I. (2007). Penalized wavelet monotone regression. *Statistics and Probability Letters*, **77**, 1608–1621.

Birke, M. (2009). Shape constrained kernel density estimation. *Journal of Statistical Planning & Inference*, **139**, 2851–2862.

Birke, M. and Bissantz, N. (2008). Shape constrained estimators in inverse regression models with convolution-type operator. Technical report SFB 475.

Birke, M. and Dette, H. (2007). Estimating a convex function in nonparametric regression. *Scandinavian Journal of Statistics*, **34**, 384–404.

Bland, J. M. and Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, **i**, 307–310.

Bowman, A. W., Jones, M. C. and Gijbels, I. (1998). Testing Monotonicity of Regression. *Journal of Computational and Graphical Statistics*, **7**, 489–500.

Carroll, R. J. and Hall, P. (1988). Optimal rates of convergence for deconvolving a density. *Journal of the American Statistical Association*, **83**, 1184–1186.

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

Cordy, C. B. and Thomas, D. R. (1997). Deconvolution of a distribution function. *Journal of the American Statistical Association*, **92**, 1459-1465.

Cressie, N. A. C. and Read, T. R. C. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society, Series B*, **46**, 440–464.

Cule, M. L., Samworth, R. J. and Stewart, M. I. (2010). Maximum likelihood estimation of a multidimensional log-concave density. *Journal of the Royal Statistical Society, Series B*, to appear.

Delaigle, A., Fan, J. and Carroll, R. J. (2009). A design-adaptive local polynomial estimator for the errors-in-variables problem. *Journal of the American Statistical Association*,**104**, 348–359.

Delaigle, A. and Gijbels, I. (2002). Estimation of integrated squared density derivatives from a contaminated sample. *Journal of the Royal Statistical Society, Series B*, **64**, 869–886.

Delaigle, A. and Gijbels, I. (2004). Comparison of data-driven bandwidth selection procedures in deconvolution kernel density estimation. *Computational Statistics & Data Analysis*, **45**, 249–267.

Delaigle, A., and Hall, P. (2008). Using SIMEX for smoothing-parameter choice in errors-in-variables problems. *Journal of the American Statistical Association*, **103**, 280–287.

Delaigle, A. and Hall, P. (2010). Estimation of observation-error variance in errors-in-variables regression. *Statistica Sinica*, to appear.

Delaigle, A., Hall, P. and Meister, A.. (2008). On Deconvolution with repeated measurements. *Journal of the American Statistical Association*, **36**, 665–685.

Dette, H., Neumeyer, N. and Pilz, K. (2009). A simple nonparametric estimator of a strictly monotone regression function. *Bernoulli* **12**, 469–490.

Dümbgen, L. and Rufibach, K. (2009). Maximum likelihood estimation of a log-concave density and its distribution function: Basic properties and uniform consistency. *Bernoulli* **15**, 40–68.

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

Friedman, J. H. and Tibshirani, R. J. (1984). The monotone smoothing of scatterplots. *Technometrics*, **26**, 243–250.

Grenander, U. (1956). On the theory of mortality measurement II. *Skandinavian Aktuarietidskr*, **39**, 125–153.

Groeneboom, P. (2001). Estimation of a convex function: Characterizations and asymptotic theory. *Annals of Statistics*, **29**, 1653–1698.

Hall, P. and Huang, L. S. (2001). Nonparametric kernel regression subject to monotonicity constraints. *Annals of Statistics*, **29**, 624–647.

Hall, P. and Huang, L. S. (2002). Unimodal density estimation using kernel methods. *Statistica Sinica*, **12**, 965–990.

Hall, P. and Kang, K. H. (2005). Unimodal kernel density estimation by data sharpening. *Statistica Sinica*, **15**, 73–98.

Hall, P. and Murison, R. D. (1993). Correcting the negativity of high-order kernel density estimators. *Journal of Multivariate Analysis*, **47**, 103–122.

Hall, P. and Presnell, B. (1999). Intentionally biased bootstrap methods. *Journal of the Royal Statistical Society, Series B*, **61**, 143–158.

Härdle, W. and Marron, J. S. (1991). Bootstrap simultaneous error bars for nonparametric regression. *Annals of Statistics*, **19**, 778–796.

Kelly, C. and Rice, J. (1990). Monotone smoothing with application to dose-response curves and assessment of synergism. *Biometrics*, **46**, 1071–1085.

Mammen, E. (1991). Nonparametric regression under qualitative smoothness assumptions. *Annals of Statistics*, **19**, 741–759 .

Mammen, E. (1995). On qualitative smoothness of kernel density estimates. *Statistics*, **26**, 253–267.

Meister, A. (2009). On testing for local monotonicity in deconvolution problems. *Statistics and Probability Letters*, **79**, 312–319.

Mukerjee, H. (1988). Monotone nonparametric regression. *Annals of Statistics*, **16**, 741–750.

Müller, U. U., Schick, A. and Wefelmeyer, W. G. (2005). Weighted residual-based density estimators for nonlinear autoregressive models. *Statistica Sinica*, **15**, 177–195.

Neumeyer, N. (2007). A note on uniform consistency of monotone function estimators. *Statistics and Probability Letters*, **77**, 693–703.

Pal, J. K. and Woodroofe, M. (2007). Large sample properties of shape restricted regression estimators with smoothness adjustments. *Statistica Sinica*, **17**, 1601–1616.

Read, T. R. C. and Cressie, N. A. C. (1988). *Goodness-of-Fit Statistics for Discrete Multivariate Data.* Springer, New York.

Rubin, H. and Sethuramen, J. (1965). Probabilities of moderate deviations. *Sankhyā, Series A*, **27**, 325–346.

Schick, A. and Wefelmeyer, W. G. (2009). Improved density estimators for invertible linear processes. *Communications in Statistics, Theory & Methods*, **38**, 3123–3147.

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

Sun, J. and Woodroofe, M. (1996). Adaptive smoothing for a penalized NPMLE of a non-increasing density. *Journal of Statistical Planning & Inference*, **52**, 143–159.