Transactions of A. Razmadze
Mathematical Institute
Vol. **174** (2020), issue 2, 189–197

ALMOST PARAMETRIC SMOOTHING

ROGER KOENKER[1] AND JIAYING GU[2]

Dedicated to Estate Khmaladze on the occasion of his 75th birthday

Abstract. Shape constraints may be a powerful aid in nonparametric function estimation, often regularizing problems without any pesky choice of tuning parameters. In some special circumstances they also achieve a remarkable, adaptive, nearly parametric convergence rate. After reviewing some prominent examples of this phenomenon, we briefly consider a closely related problem arising in the context of monotone single index models for conditional quantile functions.

1. Introduction

In regular, finite-dimensional parametric models we expect that estimated parameters converge at a rate, proportional to $1/\sqrt{n}$ for sample size n. A nonparametric estimation of densities and regression functions is, generally, more challenging, and this is typically reflected at slower rates of convergence. Of course, a higher order kernel density estimation enables one to achieve nearly parametric rates at the price of producing embarrassing estimates that may violate the basic non-negativity requirement for estimated densities; consequently, they will not be considered further here. Instead, we will focus on settings where shape constraints enable nearly parametric convergence in various related smoothing problems.

2. Monotone Density Estimation

The leading example of the phenomenon that we wish to study is the celebrated monotone density estimator of [10]. Given independent observations, $X_1, \ldots, X_n$ from a distribution F_0 with a monotone decreasing density f_0, the classical prescription for the Grenander estimator is characterized as the left derivative of the least concave majorant of the empirical distribution function,

$$\mathbb{F}_n(x) = n^{-1} \sum_{i=1}^{n} I(X_i \le x).$$

This is illustrated in Figure 1, where the piecewise linear least concave majorant yields a piecewise constant density estimate. An especially appealing feature of this estimator is that it is fully automatic, not depending on any choice of tuning parameters. The location and mass associated with the resulting "histogram bins" are determined entirely from the data. This can be seen geometrically in the Figure: it is as if we have stretched a string over the empirical distribution function, and once this is done, the left derivative is determined. This may seem rather *ad hoc* on first encounter, so it is perhaps appropriate to find that the estimator can also be viewed as a nonparametric maximum likelihood estimator.

Consider the shape constrained density estimation problem,

$$\max_f \left\{ \int \log f(x) d\mathbb{F}_n(x) \mid f \text{ decreasing}, \int f(x)dx = 1 \right\}.$$

Lemma 2.2 of [13] establishes that the solution to this problem is the Grenander estimator, provided that we adopt the convention that $\hat{f}(x) = 0$ for $x < 0$. Jumps in $\hat{f}$ occur at the order statistics of a

2020 *Mathematics Subject Classification.* 62G07, 62G08.
Key words and phrases. Density estimation; Penalized likelihood, Quantile regression.

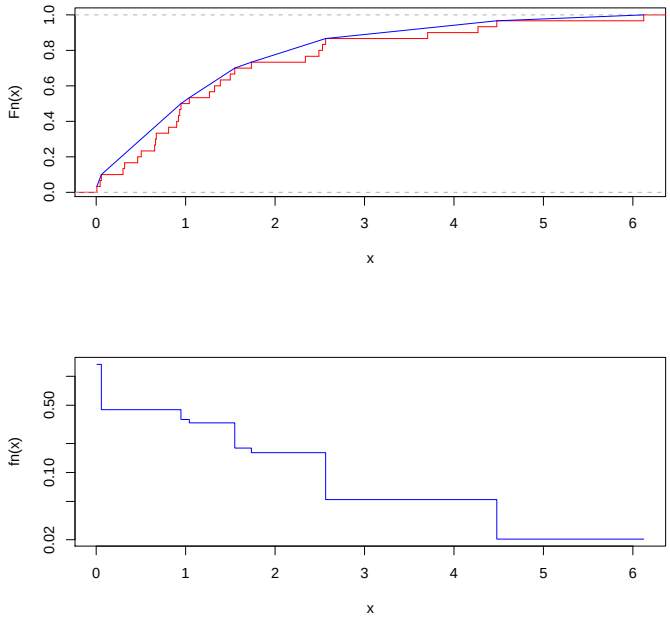

FIGURE 1. Grenander Estimator: The least concave majorant of the empirical distribution function in the upper panel, when differentiated yields the piecewise constant density estimate in the lower panel.

sample and at the origin. An alternative formulation, also grounded in maximum likelihood, involves the writing of our target density f, as a scale mixture of uniforms,

$$\max_{G\in\mathcal{G}}\left\{\int \log f(x)d\mathbb{F}_n(x) \mid f(x) = \int t^{-1}I(0\le x\le t)dG(t)\right\},$$

where $\mathcal{G}$ constitutes the set of proper distribution functions. In this case, solutions $\hat{G}$ assign a mass to a few discrete order statistics that then yield a mixture density $\hat{f}$, that is equal to the previous solutions.[1] The scale mixture formulation automatically imposes the constraint that the mixture density is supported on the positive half-line.

There is an extensive literature on the asymptotic behavior of the Grenander estimator beginning with [26], who established that pointwise,

$$n^{1/3}(\hat{f}_n(x_0) - f(x_0))/[4f(x_0)f'(x_0)]^{1/3} \rightsquigarrow Z$$

where Z is the maximizer of two-sided Brownian motion minus a parabola,

$$Z = \text{argmax}_t\{W(t) - t^2\}$$

The $\mathcal{O}(n^{-1/3})$ rate may be somewhat disappointing, however, it should be kept in mind that this result applies to the entire class of decreasing densities without the (second-order) differentiability conditions routinely assumed by kernel estimators to achieve their familiar $\mathcal{O}(n^{-2/5})$ rate.

Global convergence of the Grenander estimator was studied by [11], who established that for any bounded decreasing density f, with compact support on $[a,\infty)$ and continuous first derivative,

$$\lim_{n\to\infty} n^{1/3}R_n(f,\hat{f}_n) = K\int_a^\infty |f(x)f'(x)/2|^{1/3}dx$$

[1]For further computational details on these alternative formulations see `demo(Grenander)` in the R package `REBayes`.

where $R_n(f, \hat{f}_n) = \mathbb{E}_f \int |f(x) - \hat{f}_n(x)|dx$. Here, the constant K is more explicitly expressed as $2\mathbb{E}|V(0)| \approx 0.82$, where

$$V(a) = \sup\{t \in \mathbb{R} \mid W(t) = (t-a)^2 = \max!\}$$

and $W(t)$ is a two-sided Brownian motion on $\mathbb{R}$. [3] provide a more refined analysis of the local behavior at zero including the possibility of an unbounded target density. As is noted by [4], the uniformity is still problematic, so it is of considerable interest to have non-asymptotic risk bounds for $R_n(f, \hat{f}_n)$. To this end, L. Birgé shows that the piecewise constant, histogram-like nature of the Grenander estimator is adaptive in the sense that it tends to select an optimal partition for the binning strategy of the histogram. For smooth target densities this still yields a $\mathcal{O}(n^{-1/3})$ convergence rate, however, in the very special case that the target density is piecewise constant with a finite number of jumps, the results imply that $\hat{f}_n$ achieves the parametric rate $\mathcal{O}(n^{-1/2})$. The piecewise constant, histogram-like nature of the Grenander estimator is adaptive in the stronger sense that it selects a binning strategy suited to histogram nature of the true density as if it were a parametric object, which of course in a sense it is. Birgé is very careful to stress the special character of this result, so it may be easy to lose sight of this truly remarkable feature. In contrast to an adaptive kernel density estimation that requires a pilot estimate to guide the choice of the local bandwidth selection, the Grenander estimator constitutes its own pilot estimator, automatically selecting bins without the benefit of any preliminary bandwidth selection.[2]

3. Unimodal Density Estimation

This parametric rate performance of the Grenander estimator, however special its circumstances may be, turns out to have interesting extensions and counterparts in a wide variety of other shape constrained smoothing problems. For unimodal densities with the known mode results for the Grenander estimator can be immediately extended, and with some further effort an estimated mode can be accommodated. Closely related is the problem of estimating *strongly* unimodal, i.e., log-concave, densities. This is also a shape constrained problem susceptible to a maximum likelihood treatment,

$$\max_f \left\{ \sum_{i=1}^n \log f(x_i) \mid \log f \text{ concave}, \int f(x)dx = 1 \right\},$$

and can be reformulated as the convex optimization problem,

$$\min_g \left\{ \sum_{i=1}^n g(x_i) \mid g \in \mathcal{K}, \int e^{g(x)}dx = 1 \right\},$$

where $\mathcal{K}$ denotes the closed convex cone of convex functions. Solutions $\hat{g}_n$ are now piecewise linear with knots at the data points, so $\hat{f}_n = e^{\hat{g}_n}$ is piecewise exponential, and vanishes off the empirical support of the observations. Recently, [19] have proved that $\hat{f}_n$ achieves the minimax rate of convergence,

$$\inf_{f_n} \sup_{f_0 \in \mathcal{F}} \mathbb{E}_{f_0} d_H^2(f_n, f_0) \asymp n^{-4/5},$$

where $d_H^2(f, g) = \int (\sqrt{f(x)} - \sqrt{g(x)})^2 dx$ is the squared Hellinger distance, $\mathcal{F}$ denotes the set of all upper semi-continuous log concave densities, and f_n is any estimator of f_0. Again, it may be tempting to ask, "So what? Can't I achieve this same rate with conventional kernel methods?" When the target density f_0 is strictly log concave, the shape constraint is eventually rendered irrelevant, since any reasonable estimator would remain in the interior of the constraint set. What if, instead, f_0 lies in the boundary of the constraint set? In the log concave case this would mean that $g_0 = \log f_0$ was itself piecewise affine with k distinct pieces. In such k-affine cases, [17] establish that the non-parametric

[2]Indeed, this may lead one to wonder whether, in circumstances where the monotonicity assumption is plausible, it might be advantageous to use the Grenander $\hat{f}_n$ as a pilot estimator, simply convolving it with a smooth density if its piecewise constant appearance was deemed unattractive.

maximum likelihood estimator $\hat{f}_n$ achieves a nearly parametric rate of convergence, that is, there is a universal constant C such that for every $n \geq 2$ and every k-affine f_0,

$$\mathbb{E}_{f_0} d_H^2(f_n, f_0) \leq \frac{Ck}{n} \log^{5/4} n.$$

Thus, again without any prior knowledge about the number of affine pieces, the NPMLE achieves the almost parametric rate of $\mathcal{O}(1/\sqrt{n})$, without any required tuning parameter selection. In fact, something considerably more general is proved for f_0 that are nearly k-affine. It would also be possible to generalize to weaker forms of concavity as in [22], but we will resist going into the details. Instead, we will turn our attention to the estimation of a general class of mixture models.

4. Nonparametric Estimation of Mixture Densities

Many statistical problems can be formulated as parametric mixtures, leading examples are the Gaussian location mixture

$$f(x) = \int \varphi(x - \theta) dG(\theta)$$

and the Gaussian scale mixture

$$f(x) = \int \theta^{-1} \varphi(x/\theta) dG(\theta).$$

Given a sample of independent observations, $X_1, X_2, \ldots, X_n$, we can consider these as models with $X_i \sim \mathcal{N}(\theta_i, 1)$ and $X_i \sim \mathcal{N}(0, \theta_i^2)$, respectively. We would like to estimate the mixing distribution G when the observations are assumed to be exchangeable. [16] proposed estimating G by a nonparametric maximum likelihood,

$$\max_{G \in \mathcal{G}} \left\{ \sum_{i=1}^{n} \log f(X_i) \mid f(x) = \int \varphi(x, \theta) dG(\theta) \right\}, \tag{1}$$

and proved consistency of the resulting $\hat{G}$. Computation by the EM algorithm was suggested by [23], but remained quite challenging. Modern convex optimization methods provide a much more efficient and scalable approach to computation as is shown in [21]. However, many problems regarding the statistical performance of these methods remain open.

An important step forward in this respect is the recent work of [27] who consider the Gaussian location mixture model in $\mathbb{R}^d$. They evaluate performance relative to the oracle Bayes estimator that knows the empirical measure of the true θ's, $\mathbb{G}_n(t) = n^{-1} \sum_{i=1}^{n} I(t - \theta_i)$. Their Proposition 2.3 establishes that when $\mathbb{G}_n$ is discrete, supported on a set of cardinality k, there exists a constant C_d such that

$$\mathbb{E} d_H^2(\hat{f}_n, f_{\mathbb{G}_n}) \leq C_d \left(\frac{k}{n} (\sqrt{\log n})^{d + (4-d)_+} \right).$$

It follows easily that this is the minimax attainable rate. Again, we have an almost parametric convergence rate up to the logarithmic factor for the nonparametric MLE of the mixture density.

5. Shape Constrained Regression

It should not come as a big surprise that the shape constraints can also play an important role in regression, as well as in density estimation. Most of the literature has focused on the least squares fidelity criterion. The simplest setting is the isotonic regression model,

$$Y_i = \theta_i + u_i \quad i = 1, 2, \ldots, n,$$

where the θ_i are assumed to satisfy $\theta_1 \leq \theta_2 \leq \cdots \leq \theta_n$. Implicitly, we can think of this model as one in which we observe Y_i's at a sequence of increasing design points. The apparently more general formulation of the model with $Y_i = g(x_i) + u_i$ reduces to the former model under general convex loss; if the observations are not ordered in the covariate x_i, we can simply reorder the Y_i's according to the order of the x_i's and proceed as before. Under the monotonicity constraint, the solutions are piecewise constant with jumps at the design points and loss depends only on the estimated function values at these design points.

For i.i.d. Gaussian u_i with variance $\sigma^2 < \infty$ the nonparametric MLE can again be formulated as a convex optimization problem

$$\min_{\theta} \left\{ \sum_{i=1}^{n} (Y_i - \theta_i)^2 \mid \theta \in \mathcal{K}_n \right\},$$

where $\mathcal{K}_n$ is the convex polyhedral cone of nondecreasing sequences. This problem has a long history going back to [5] and perhaps even before. Computation of solutions are typically carried out with the pool-adjacent-violaters algorithm (PAVA), although various modern variants of quadratic programming could also be used.

[30] showed that the empirical risk of the nonparametric MLE, $\hat{\theta}_n$

$$R_n = n^{-1} \sum_{i=1}^{n} (\theta_i - \hat{\theta}_i)^2 \leq C \left[\left(\frac{\sigma^2 V_n}{n} \right)^{2/3} + \frac{\sigma^2 \log n}{n} \right],$$

where $V_n = \theta_n - \theta_1$ and C is a fixed constant. However, more recent refinements establish that improvement over this $\mathcal{O}(n^{-1/3})$ rate can be achieved under the special circumstances that the θ_i are piecewise constant with a small number k of pieces. In that case it is proved in [7] that

$$R_n \leq \inf_k \left(\frac{4\sigma^2(1+k)}{n} \log \frac{en}{1+k} \right).$$

Thus, up to the log factor we again have almost parametric convergence determined by the number of distinct piecewise constant elements in the target function. And again, it is worth stressing that adaptation is achieved over the number and locations of these pieces without any intervention of tuning parameters. When the monotonicity is misspecified, there is, obviously, a bias effect and this is also characterized in the general formulation of this result.

When the monotonicity constraint is replaced by a convexity (or concavity) constraint, the nonparametric MLE under Gaussian error is piecewise linear with knots at the observed design points. In the simplest setting with equally spaced design points this imposes the constraint that the second differences of θ_i's are nonnegative. [14] and [7] prove that when the target regression function is k-affine, that is, piecewise linear with k distinct pieces, the NPMLE again achieves an adaptive parametric rate of convergence up to a log factor.

Although the prior literature has focused exclusively on the least squares, i.i.d. Gaussian noise setting, as has most of the PAVA literature, there is nothing that prohibits us from entertaining other fidelity criteria. A natural alternative is the family of quantile loss functions that yield estimates of the conditional quantile functions of the response. Again, we have a convex optimization problem

$$\min \left\{ \sum_{i=1}^{n} \rho_\tau(y_i - g(x_i)) \mid g \in \mathcal{K} \right\}$$

, where $\rho_\tau(u) = u(\tau - I(u < 0))$, and $\mathcal{K}$ is the closed convex cone representing either monotone, convex or concave functions. An implementation of such estimators is available in the R package `quantreg` with the function `rqss`. In contrast to the least squares version of PAVA, the algorithmic complexity of the quantile implementation via interior point methods is not carefully analyzed, but sparsity of the underlying constraint matrix assures efficient practical performance. This implementation expands the formulation in several respects: (i) there is an option to impose further smoothness on the shape constrained estimate; (ii) general, unequally spaced design points are permitted; and (iii) additive models with several shape constrained components are permitted. To elaborate briefly on the first point, the general form of the `rqss` function permits the user to impose a total variation penalty on the first derivative of the fitted function

$$TV(g') = \int |g''(x)| dx$$

, controlled by a tuning parameter λ. When λ is sufficiently large, the $\hat{g}_n$ is constrained to be linear, while when λ is sufficiently close to zero, the TV penalty has no effect, and only the shape constraint determines the fit.

From a computational viewpoint the polyhedral cone and total variation constraints are especially appealing in the quantile regression setting because they maintain the linear programming structure of the estimation problem. Due to the relative sparsity of the design matrices in such problems, modern interior point algorithms are quite efficient even for large scale problems. It should be noted that the form of the solutions, that is whether they are piecewise constant, piecewise linear, etc., is entirely determined by the form of the constraints and, in particular, by the order of the differential operator appearing there. Thus, if $g \in \mathcal{K}$ requires that $Dg \geq 0$ to impose monotonicity, then solutions will be piecewise constant. If instead $D^2 g \geq 0$ is imposed to achieve convexity, then solutions will be piecewise linear. Likewise, total variation penalties on g, itself, yield piecewise constant solutions, while total variation penalties on Dg, thereby controlling the L_1 norm of $D^2 g$ yield piecewise linear solutions. Although such methods have a long history in imaging and actuarial science, they only have become widely appreciated in statistics through the relatively recent works of [18] and [28].

We conjecture that these shape constrained conditional quantile function estimators enjoy the same almost parametric convergence as their least squares counterparts and hope to report on this at a later time.

6. Shape Constrained Transformation Models

This brings us to our final category of shape constrained estimators: transformation models take a variety of forms, but typically they have a single index structure like

$$\mathbb{E}Y_i|X_i = \Psi(X_i^\top \beta). \tag{2}$$

The covariates and the parameter $\beta \in \mathbb{R}^p$ are wrapped in a function $\Psi : \mathbb{R} \to \mathbb{R}$ that may be parametric or nonparametric. Motivated by the revival of interest in the Grenander estimator, there has been an increased interest in transformation models with *monotonic* Ψ. Clearly, when $p = 1$, the β parameter is irrelevant and with Ψ monotonic, we are back to the methods described in the previous section. When $p > 1$, we can regard such models as an heroic attempt to circumvent the curse of dimensionality by assuming a simple form for the way the covariates enter the model while preserving some semblance of nonlinear structure. There are a variety of closely related models, some of which replace Ψ on the right-hand side by some transformation of the response variable itself. The monograph [6] provides a systematic treatment of many of these models, both parametric and nonparametric.

Recently, [2] and [12] have very thoroughly explored various approaches to estimating the model (2). They argue that it is preferable to avoid the direct profiling approach and focus on methods that find approximate zeros of the score (gradient) equations. It is clear that the vector β is identified only up to scale, so it is natural to impose the constraint that its Euclidean norm is one, $\|\beta\| = 1$. This can be accomplished in a variety of ways, either by transformation to spherical coordinates, or by adding a Lagrangian term. They employ the former scheme for their asymptotics, but prefer the latter from a practical, computational standpoint. Another option is to simply set one of coordinates of β equal to 1 or -1.

A drawback of the conditional mean formulation of the model, one that also afflicts a much broader class of nonlinear transformation models for conditional means, is the necessity of assuming that in the additive error formulation of the model

$$Y_i = \Psi(X_i^\top \beta) + U_i, \tag{3}$$

there is a full independence between the observed covariates X_i and U_i. One way to circumvent this requirement is to replace the mean formulation by a conditional quantile formulation

$$\mathbb{Q}_{Y|X}(\tau|X) = \Psi_\tau(X^\top \beta_\tau). \tag{4}$$

The quantile formulation also renders superfluous the unsightly moment conditions that appear inevitably in the analysis of the mean formulation. Such models were first considered by [8] who provide a very thorough motivation and contextualization for this class of models. Drawing on earlier work of Chaudhuri, they propose an average derivative estimator for β based on the nonparametric kernel weighted quantile regression.

When the τth conditional quantile function of Y given X is postulated to be a monotone function of a linear predictor in X_i, as seems plausible in many applications, we can try to exploit shape constrained methods to estimate both Ψ and β. Since quantiles are equivariant to monotone transformation, interpretation of the family of such models is also much more straightforward, than their mean counterparts. Our initial computational strategy arose immediately from the equivariance property of the quantiles, (4) implies

$$\mathbb{Q}_{\Psi_\tau^{-1}(Y)|X}(\tau|X) = X^\top \beta_\tau. \tag{5}$$

Since this linear quantile regression formulation can be efficiently estimated even for a high-dimensional β, a simple iterative strategy in which alternate back and forth from estimation of Ψ to estimation of β seems attractive. At each iteration we can modify the resulting β so that it has norm one. Given a β, an estimate of Ψ can be obtained by solving the monotone quantile regression problem described above. Both steps are linear programs. The biconvex structure of the problem is common to many mathematical contexts (see [1] and [9] for further details). Unfortunately, there is no general assurance that such an iterative procedure converges to a global optimum. Indeed, contrary to our initial, naive expectations, it performed abysmally.

Thus, following the lead of [12], but not without some trepidation, we turned to the global methods of optimization, in particular, the patterned search method of [15]. Convergence of such pattern search algorithms to a stationary point was established in [29]. An R implementation is available from the `optimx` package of [24], and an Rcpp implementation is available from the github site of Piet Groeneboom. Provisionally, we have experimented with the former implementation which has been performed quite well. In Figure 2, we illustrate three realizations from a sample in which the true Ψ is piecewise constant with only one jump; there are 5 covariates drawn as independent standard Gaussians. It is apparent from this Figure that the location and magnitude of the jump in Ψ is quite accurately estimated, this is hardly surprising in view of the fact that all the information about the linear predictor is contained in the neighborhood around this jump. Estimation of the level of Ψ before and after the jump is more problematic, which is again not surprising given that in the absence of a jump we would not be able to consistently estimate the linear index at all.

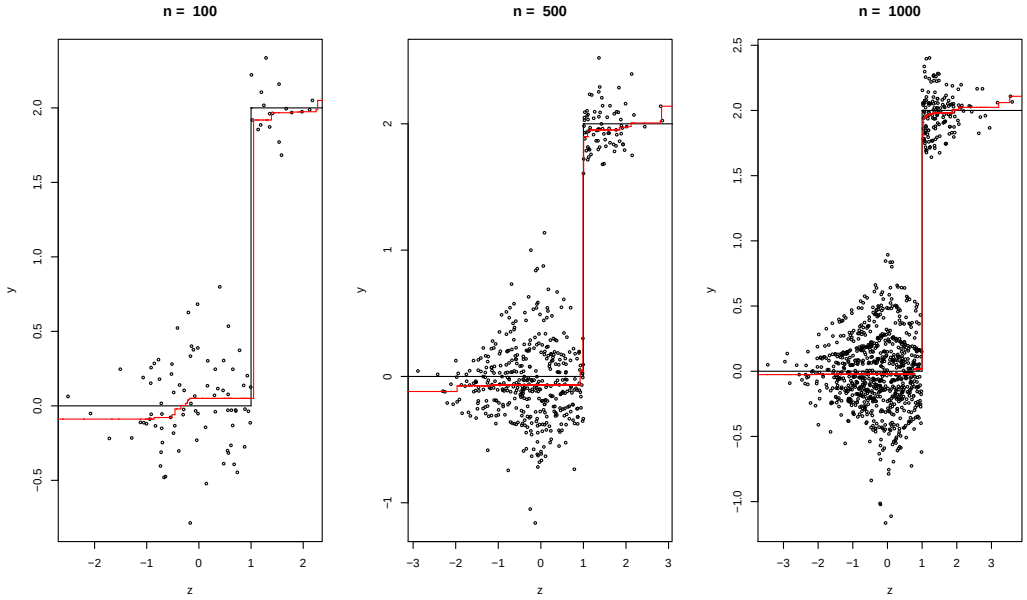

FIGURE 2. Single Index Median Regression Monotone Transformation Model: The plotted points depict the pairs $(\Psi_\tau(x_i^\top \beta_0), y_i)$. Three realizations for difference sample sizes of the final estimate of the monotone $\hat{\Psi}_n$ based on five Gaussian covariates in the single index. The true Ψ is depicted in black, while the estimate is in red.

To begin to explore the asymptotic behavior of our proposed estimator it is useful to reconsider the case for a known transformation, Ψ. As is described in [20], Section 4.4, if we adopt the model

$$\mathbb{Q}_{Y_i|X_i=x_i}(\tau|x_i) = g(x_i, \beta_0),$$

it is natural to try to estimate β_0 by

$$\hat{\beta}_n = \text{argmin}_{b\in\mathcal{B}} \sum \rho_\tau(y_i - g(x_i, b)).$$

We emphasize the verb "try" since optimization need no longer be an assured attack on a convex problem with a unique solution. In keeping with the vast literature on nonlinear least squares, we will assume that the domain $\mathcal{B}$ is compact. In addition, we will assume that the conditional distribution functions F_i of $Y_i|X_i$ are absolutely continuous with continuous derivatives $f_i(\xi_i)$ at the points $\xi_i = g(x_i, \beta_0)$, and the following conditions on design.

G1: There exist constants k_0, k_1 and n_0 such that for $\beta_1, \beta_2 \in \mathcal{B}$ and $n > n_0$,

$$k_0 \parallel \beta_1 - \beta_2 \parallel \leq \left(n^{-1} \sum_{i=1}^{n} (g(x_i, \beta_1) - g(x_i, \beta_2))^2 \right)^{1/2} \leq k_1 \parallel \beta_1 - \beta_2 \parallel .$$

G2: There exist positive definite matrices D_0 and $D_1(\tau)$ such that with $\dot{g}_i = \partial g(x_i, \beta)/\partial\beta|_{\beta=\beta_0}$,

$$\begin{aligned} &\text{(i)} \quad \mathbb{E}\dot{g}_i\dot{g}_i^\top = D_0/ \\ &\text{(ii)} \quad \mathbb{E}f_i(\xi_i)\dot{g}_i\dot{g}_i^\top = D_1(\tau), \\ &\text{(iii)} \quad \max_{i=1,\ldots,n} \parallel \dot{g}_i \parallel /\sqrt{n} \to 0. \end{aligned}$$

Under these conditions, it can be shown that we have the Bahadur representation

$$\sqrt{n}(\hat{\beta}_n - \beta_0) = D_1^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{g}_i \psi_\tau(u_i) + o_p(1),$$

where $\psi_\tau = \rho'_\tau$ and $u_i = y_i - g(x_i, \beta_0)$. Consequently,

$$\sqrt{n}(\hat{\beta}_n - \beta_0) \rightsquigarrow \mathcal{N}(0, \tau(1-\tau)D_1^{-1}D_0D_1^{-1}),$$

(for further details see [25]).

In the special case of the single index model, $g(x, \beta) = \Psi(x^\top\beta)$ and $\beta \in \mathcal{B}$ is replaced by $\beta \in \mathcal{S}^{p-1} \equiv \{b \in \mathbb{R}^p \mid \|b\| = 1\}$. Thus, $\dot{g} = \partial g/\partial\beta$ becomes $J\dot{\Psi}X$, where J denotes the Jacobian of the transformation that maps β into its $(p-1)$-dimensional counterpart. When Ψ is strictly increasing as is commonly assumed in the literature, this returns the expressions for D_0 and D_1 to something closely resembling their linear quantile regression equivalents, except for the weighting factors from the $\dot{\Psi}_i$ terms and the dimension reduction effect of the Jacobian terms. Inverses in the sandwich formulae now of course need to be interpreted as generalized inverses due to the dimension reduction.

At this point the obvious question is: How does all this change when Ψ is *estimated*? Surprisingly, the answer would seem to be: very little. Following the arguments of [2] and several prior authors cited there in the mean regression setting, this would entail replacing $XX^\top$ in the modified expressions for D_0 and D_1 by the conditional covariance $\text{Cov}(X|X^\top\beta = x^\top\beta)$. This change reflects a reduction in the precision of the estimator $\hat{\beta}_n$. For smoothly increasing Ψ, as in the least squares theory, it is inevitable that we would obtain cube root convergence for $\hat{\Psi}_n$. A much more intriguing question, but a considerably more difficult one, is this: Can $\sqrt{n}$ convergence of $\hat{\Psi}_n$ be salvaged if we are willing to assume that the true Ψ is piecewise constant? The highly accurate estimates of the jump component of Ψ in Figure 2 offers a hint that this may indeed be plausible. Unfortunately, we must leave this intriguing problem for a future research.

References

1. R. J. Aumann, S. Hart, Bi-convexity and bi-martingales. *Israel J. Math.* **54** (1986), no. 2, 159–180.
2. F. Balabdaoui, P. Groeneboom, K. Hendrickx, Score estimation in the monotone single-index model. *Scand. J. Stat.* **46** (2019), no. 2, 517–544.

3. F. Balabdaoui, H. Jankowski, M. Pavlides, A. Seregin, J. Wellner, On the Grenander estimator at zero. *Statist. Sinica* **21** (2011), no. 2, 873–899.
4. L. Birgé, The Grenander estimator: a nonasymptotic approach. *Ann. Statist.* **17** (1989), no. 4, 1532–1549.
5. H. D. Brunk, Maximum likelihood estimates of monotone parameters. *Ann. Math. Statist.* **26** (1955), 607–616.
6. R. J. Carroll, D. Ruppert, *Transformation and Weighting in Regression*. Monographs on Statistics and Applied Probability. Chapman and Hall, New York, 1988.
7. S. Chatterjee, A. Guntuboyina, B. Sen, On risk bounds in isotonic and other shape restricted regression problems. *Ann. Statist.* **43** (2015), no. 4, 1774–1800.
8. P. Chaudhuri, K. Doksum, A. Samarov, On average derivative quantile regression. *Ann. Statist.* **25** (1997), no. 2, 715–744.
9. J. Gorski, F. Pfeuffer, K. Klamroth, Biconvex sets and optimization with biconvex functions: a survey and extensions. *Math. Methods Oper. Res.* **66** (2007), no. 3, 373–407.
10. U. Grenander, On the theory of mortality measurement. II. *Skand. Aktuarietidskr.* **39** (1956), 125–153 (1957).
11. P. Groeneboom, *Estimating a Monotone Density*. Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer, vol. II (Berkeley, Calif., 1983), 539–555, Wadsworth Statist./Probab. Ser., Wadsworth, Belmont, CA, 1985.
12. P. Groeneboom, K. Hendrickx, Estimation in monotone single-index models. *Stat. Neerl.* **73** (2019), no. 1, 78–99.
13. P. Groeneboom, G. Jongbloed, *Nonparametric Estimation Under Shape Constraints. Estimators, Algorithms and Asymptotics*. Cambridge Series in Statistical and Probabilistic Mathematics, 38. Cambridge University Press, New York, 2014.
14. A. Guntuboyina, B. Sen, Global risk bounds and adaptation in univariate convex regression. *Probab. Theory Related Fields* **163** (2015), no. 1-2, 379–411.
15. R. Hooke, T. A. Jeeves, Direct search solution of numerical and statistical problems. *Journal of the ACM* **8** (1961), 212–229.
16. J. Kiefer, J. Wolfowitz, Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.* **27** (1956), 887–906.
17. A. K. H. Kim, A. Guntuboyina, R. J. Samworth, Adaptation in log-concave density estimation. *Ann. Statist.* **46** (2018), no. 5, 2279–2306.
18. S.-J. Kim, K. Koh, S. Boyd, D. Gorinevsky, l_1 trend filtering. *SIAM Rev.* **51** (2009), no. 2, 339–360.
19. A. K. H. Kim, R. J. Samworth, Global rates of convergence in log-concave density estimation. *Ann. Statist.* **44** (2016), no. 6, 2756–2779.
20. R. Koenker, *Quantile Regression*. Econometric Society Monographs, 38. Cambridge University Press, Cambridge, 2005.
21. R. Koenker, I. Mizera, Convex optimization, shape constraints, compound decisions, and empirical Bayes rules. *J. Amer. Statist. Assoc.* **109** (2014), no. 506, 674–685.
22. R. Koenker, I. Mizera, Shape constrained density estimation via penalized Rnyi divergence. *Statist. Sci.* **33** (2018), no. 4, 510–526.
23. N. Laird, Nonparametric maximum likelihood estimation of a mixed distribution. *J. Amer. Statist. Assoc.* **73** (1978), no. 364, 805–811.
24. J. C. Nash, R. Varadhan, Unifying Optimization Algorithms to Aid Software System Users: optimx for R. *Journal of Statistical Software* **43** (2011), 1–14.
25. W. Oberhofer, H. Haupt, Asymptotic theory for nonlinear quantile regression under weak dependence. *Econometric Theory* **32** (2016), no. 3, 686–713.
26. B. L. S. Prakasa Rao, Estimation of a unimodal density. *Sankhyā Ser. A* **31** (1969), 23–36.
27. S. Saha, A. Guntuboyina, On the nonparametric maximum likelihood estimator for Gaussian location mixture densities with application to Gaussian denoising. *Ann. Statist.* **48** (2020), no. 2, 738–762.
28. R. J. Tibshirani, Adaptive piecewise polynomial estimation via trend filtering. *Ann. Statist.* **42** (2014), no. 1, 285–323.
29. V. J. Torczon, On the convergence of pattern search algorithms. *SIAM J. Optim.* **7** (1997), no. 1, 1–25.
30. C.-H. Zhang, Risk bounds in isotonic regression. *Ann. Statist.* **30** (2002), no. 2, 528–555.

(Received 19.11.2019)

[1]Roger Koenker, Department of Economics, UCL, London, WC1H OAX, UK
E-mail address: r.koenker@ucl.ac.uk

[2]Jiaying Gu, Department of Economics, University of Toronto, Ontario, M5S 3G7, Canada
E-mail address: jiaying.gu@utoronto.ca