The Annals of Statistics
1997, Vol. 25, No. 1, 414–434

LOCAL ASYMPTOTICS FOR QUANTILE SMOOTHING SPLINES[1]

BY STEPHEN PORTNOY

University of Illinois

Quantile smoothing splines were introduced by Koenker, Ng and Portnoy as natural and appealing estimates of conditional quantiles of response variables. The natural setting for the problem considers minimization of a weighted combination of a "fit" penalty and a "roughness" penalty over the space of functions whose derivatives have bounded variation. Although this space is not traditional, Shen has shown recently that the quantile smoothing splines do indeed converge at the usual optimal rate ($n^{-2/5}$) in various norms. Here, local asymptotic results are obtained by establishing Bahadur representations for local parameters of the splines. These are used to obtain local rates of convergence, to establish uniform convergence rates, to provide local distribution theory for quantile B-splines and to expand the "fit" measure in order to analyze an information criterion for determining the smoothing parameter. Examples of using derivatives of the smoothing splines for estimating jump functions are also presented.

1. Introduction. Use of conditional quantiles plays an increasingly important role in modern statistical analysis. Following the pathbreaking work of Koenker and Bassett (1978), the theory and application of regression quantiles have undergone substantial development and now provide a reliable and efficient methodology for conditional quantile analysis in homoscedastic linear models. However, it has become clear that typical regression examples often exhibit important forms of nonstationarity. Responses that are especially high or low may often be expected to depend on the independent variables rather differently from median responses. Hendricks and Koenker (1992) present a good example of this in a study of consumer demand for electricity, where heavy users responded much more dramatically to weather and time variation (presumably because of air conditioner use). Thus, generalization of conditional quantile methods beyond the usual linear models is required.

Rather than try to model nonstationarity directly, we will take the natural approach of nonparametric estimation. Although there are a large number of methods available, Koenker, Ng and Portnoy (1994) make a strong case for a smoothing spline approach, which generalizes the L_2 theory developed for conditional means to the estimation of conditional quantiles. To be specific, consider optimization of a penalized form of the classical "check" function that defines quantiles: let $\tau \in [0, 1]$ and define

$$\rho_\tau(u) = u(\tau - I(u < 0)), \qquad \text{"fit"} = \sum_{i=1}^{n} \rho_\tau(Y_i - \hat{g}(x_i)). \tag{1.1}$$

Received June 1994; revised July 1996.

[1] Research partially supported by NSF Grant SES-91-21776.

AMS 1991 *subject classifications*. 62G05, 62E20.

Key words and phrases. Nonparametric regression, conditional quantiles, rates of convergence.

The classical L_2 roughness penalty is not especially natural for this measure of "fit," and it leads to rather formidable quadratic programming methods of computation. As noted in Koenker, Ng and Portnoy (1994), a roughness penalty based on the variation norm is quite natural, allows the use of efficient (parametric) linear programming methods and permits significant generalization of the space of fitting functions considered. Here we add the justification for this approach that the resulting conditional quantile estimates converge and possess natural local asymptotic and finite sample quantile-like properties. To define the roughness penalty formally, consider functions f of bounded variation and let $V(f)$ denote the total variation norm of f. The quantile smoothing spline may be defined as a function $\hat{g}_\tau(x)$ solving the following minimization problem:

$$\min_{V(g')<+\infty} \sum_{i=1}^{n} \rho_\tau(Y_i - g(x_i)) + \lambda V(g'), \tag{1.2}$$

where λ is the smoothing parameter. Note that $V(g') = \int |g''(x)|\,dx$ if g' is sufficiently smooth; so that the roughness penalty in (1.2) arises essentially by replacing the familiar L_2-norm of traditional spline theory with the L_1-norm here. Koenker, Ng and Portnoy (1994) show that the solutions $\hat{g}_n(x)$ of (1.2) are continuous, piecewise linear functions and that their computation is especially appealing. For λ sufficiently large, the solution is the corresponding globally linear regression quantile. By linear programming results, solutions remain constant on λ-intervals, and successive solutions can be found by single simplex pivots. Thus, the family of solutions for all λ values can be computed quite efficiently. Koenker, Ng and Portnoy (1994) present some examples and [following Schwarz (1978)] also introduce an information criterion (SIC, for "Schwarz information criterion") for choosing an appropriate value for the smoothing parameter λ. This criterion will be discussed further in Sections 4 and 5.

Although parametric linear programming provides efficient computation for moderate sample sizes, large n requires substantial computational resources. As a more easily computable alternative, consider functions $\tilde{g}_n(x)$ minimizing (1.2) subject to the condition that all breakpoints lie on a subgrid of larger mesh ε_n. The case where all of the points of the subgrid are breakpoints is just the case of B-splines, as described in He and Shi (1994). Thus, $\tilde{g}_n(x)$ may be called a penalized B-spline (with a specified subgrid). If n is large enough and ε_n is sufficiently small, $\tilde{g}_n(x)$ should be very similar to $\hat{g}_n(x)$.

The primary focus here is on asymptotics for these quantile smoothing splines. The results here will be developed under the simple nonparametric regression model:

$$Y_i = g_0(x_i) + U_i \quad \text{for } i = 1, \dots, n \tag{1.3}$$

where U_i form a random sample from a distribution F with density f. Generally $\{x_i\}$ will be taken to be equally spaced on the unit interval: $x_i \equiv i/n$, $i = 1, \dots, n$. In fact, since nonparametric methods are inherently local, general smooth heteroscedasticity is not difficult to incorporate. However, in an effort to simplify the already rather involved computations, the simpler model (1.3) will be assumed here, even though conditional quantiles are most important when they differ over τ. For (1.3), the τth conditional quantile is just $g_\tau(x) = g_0(x) + F^{-1}(\tau)$, and the subscript τ may be dropped if its value is clear from context.

In dealing with the asymptotic theory for solutions of (1.2), however, there is a fundamental difficulty with standard approaches to asymptotics: the appropriate space over which (1.2) is minimized is the space of functions whose derivatives have bounded (total) variation. Unfortunately, this space is too large in terms of metric entropy for standard methods to obtain the usual optimal rates ($n^{-2/5}$) [although results of Mammen and van de Geer (1997) may be applicable]. In a penetrating study of asymptotics for penalized methods, Shen (1994) obtained global convergence at the optimal rate using the following approach: take $\lambda_n = \mathscr{O}(n^{1/5})$ and consider a sequence of subspaces T_n whose metric entropy is bounded appropriately. Let g_n be any sequence of elements of T_n for which the penalized criterion in (1.2) is within δ_n of the minimizing value. If δ_n converges quickly enough, T_n has appropriately bounded metric entropy, and g_0 is sufficiently smooth (twice continuously differentiable), then g_n converges to g_τ at the optimal rate ($n^{-2/5}$) (in any reasonable norm, including L_1 or L_2). For the quantile smoothing spline penalties, Shen notes that one can choose T_n to be (norm bounded) subsets of the standard Sobolev space W_2 (whose metric entropy is well known to lead to $n^{-2/5}$ rates). In particular, let $\hat{g}_n$ be the minimizer of (1.2). For such a piecewise linear function (whose derivative has bounded variation), and *any* δ_n, there is a twice-differentiable function g_n^* (in W_2) such that both the fit and roughness penalties are within δ_n of the values for $\hat{g}_n$, as well as $\|g_n^* - \hat{g}_n\| \le \delta_n$. Let δ_n be of order $n^{-2/5}$. Then, as a consequence of Shen's result, g_n^* converges to the true g_τ at the optimal rate; by construction of the approximation, $\hat{g}_n$ must also converge in norm at the optimal rate.

Here the basic question is that of local asymptotics. Global convergence results giving rates of convergence in various norms have been well-developed for smoothing splines. However, local asymptotic results have been much less studied, even though such results are required for statistical inference. Some results are available, especially in some cases that are closely related to quantile smoothing splines. Mammen (1991) provides some local results for least squares estimates of piecewise concave or convex regression functions. These estimates are rather similar in form to the quantile smoothing splines. Some extensions are given by Wang (1993). Mammen and van de Geer (1997) also provide some local results in a problem like (1.2) but where the fit is measured by a sum of squares. The quadratic nature of the fitness criterion in these alternative approaches appears to simplify the theory, but the computational requirements seem substantially more complicated than the rather simple linear programming approach of quantile smoothing splines.

The fundamental asymptotic approximations required are local Bahadur representations for the parameters of a single linear segment of $\hat{g}$. It is easiest to establish the non-i.i.d. representation of Portnoy (1991) on the entire linear segment. However, this leads to two serious problems. First, the endpoints of the segment are random; so the representation is a random sum, whose distribution cannot be easily approximated. Second, for the representation to be useful, there must be sufficiently many points in the segment. Optimal convergence rates would require the number of points in the segment to be of exact order $n^{4/5}$. If a linear segment is not part of a concave or convex section of $\hat{g}$, then introducing an additional breakpoint along the segment will strictly increase the roughness penalty. This makes it possible to show that the length of such a segment must exceed $\varepsilon n^{-1/5}(\log n)^{1/3}$ (in probability). However, within a concave (or convex) section, introducing a new breakpoint that does *not* lose concavity (or convexity) does *not* change the roughness penalty at all. This seems to make it impossible to bound the length of such a segment from below: it seems that there very well might be exceedingly short linear segments along concave or convex sections. However, it is possible to bound the number of breakpoints by cn^a for some $a < 1$. This will yield uniform convergence, but at a somewhat slower rate. These problems may be circumvented by using the penalized B-splines $\tilde{g}_n(x)$ with an initial subgrid large enough to ensure that all linear segments are sufficiently long. Optimal asymptotic rates of convergence in the "sup" norm can be obtained by simply taking $\varepsilon_n = \mathscr{O}((n \log n)^{-1/5})$.

The basic Bahadur representations are introduced in Section 2. Section 3 shows that sufficiently many of the local linear segments are sufficiently large to provide the local and uniform convergence results. Section 4 presents an expansion of the "fit" criterion in (1.2) and applies it to defining a new information criterion for choosing λ. Section 5 notes that the piecewise linear form of the quantile smoothing splines suggests that derivatives of the splines may be used to estimate jump functions. Some examples are presented.

2. Asymptotic representations.

We posit the model (1.3) and the basic definitions given there, including the assumption that $x_i = i/n$. In addition, the following conditions are introduced:

F. The error distribution has a uniformly bounded, strictly positive density, with a uniformly bounded derivative.

G. The true regression function $g_0(x)$ has a uniformly bounded, continuous second derivative. This implies that for any partition of the unit interval by subintervals J_k centered at x_k and of length δ (or smaller), we have (uniformly)

$$g_0(x) = g_0(x_k) + g_0'(x_k)(x - x_k) + \tfrac{1}{2} g_0''(x_k)(x - x_k)^2 + o(\delta^2). \tag{2.1}$$

The following result now provides a representation for an entire local segment on which $\hat{g}(x)$ is linear. In particular, let J be the subscripts for an interval on which $\hat{g}(x)$ is linear, and let $\bar{x}_J$ be the midpoint of the interval. That is, the local parameter estimates may be defined by

$$\hat{g}(x) = \hat{\alpha} + \hat{\beta}(x - \bar{x}_J), \qquad \min\{x_i\colon i \in J\} \leq x \leq \max\{x_i\colon i \in J\}. \tag{2.2}$$

Let α and β be the (corresponding) value and derivative of $g_\tau(x) \equiv g_0(x) + F^{-1}(\tau)$ at $x = \bar{x}_J$. The representations depend critically on whether or not $\hat{g}'$ is monotonic over the three linear segments from the one preceding J to the one following J, that is, on whether or not $\hat{g}$ is concave or convex near J. To provide appropriate notation, let $I^*[\sim \mathrm{mon}(J)]$ denote the indicator function of the corresponding event; that is, with $(\hat{\beta}^-, \hat{\beta}, \hat{\beta}^+)$ denoting the three successive slopes (at J), define

$$I^*[\sim \mathrm{mon}(J)] \equiv \begin{cases} 1, & \hat{\beta}^- < \hat{\beta} < \hat{\beta}^+ \text{ or } \hat{\beta}^- > \hat{\beta} > \hat{\beta}^+, \\ 0, & \text{otherwise.} \end{cases} \tag{2.3}$$

THEOREM 2.1. *Assume conditions* F *and* G, *and let J be the subscripts for an interval on which $\hat{g}(x)$ is linear. Let $\{\lambda_n\}$ be any sequence of smoothing penalty coefficients in* (1.2):

(i) *With I^* defined by* (2.3) *and $\psi_\tau(u) \equiv \rho'_\tau(u)$, the following gradient conditions hold*:

$$\left| \sum_{i \in J} \psi_\tau(Y_i - \hat{g}(x_i)) \right| = \mathscr{O}(1); \tag{2.4}$$

$$\left| \sum_{i \in J} (x_i - \bar{x}_J)\psi_\tau(Y_i - \hat{g}(x_i)) \right| = \mathscr{O}(1 + I^*[\sim \mathrm{mon}(J)]\lambda_n). \tag{2.5}$$

(ii) *Thus, if the number of observations in J, $m \equiv m_n \equiv \sharp\{J\}$, satisfies*

$$\frac{m}{(\log n)^{1/2}} \to \infty, \tag{2.6}$$

then the following representation holds:

$$(\hat{\alpha} - \alpha) = \frac{1}{mf(F^{-1}(\tau))} \sum_{i \in J} \psi_\tau(U_i + F^{-1}(\tau)) + b_J(\tau) + \mathscr{O}_p(m^{-3/4}(\log n)); \tag{2.7}$$

$$\begin{aligned}(\hat{\beta} - \beta) = {} & \left(\frac{m^3}{12n^2}\right)^{-1} \frac{1}{f(F^{-1}(\tau))} \sum_{i \in J} (x_i - \bar{x}_J)\psi_\tau(U_i + F^{-1}(\tau)) \\ & + \mathscr{O}_p(m^{-3/4}(\log n)) + \mathscr{O}(I^*[\sim \mathrm{mon}(J)]n^2 \lambda_n m^{-3}).\end{aligned} \tag{2.8}$$

Here $b_J(\tau)$ is the bias term,

$$b_J(\tau) \equiv \left(\frac{m^2}{24n^2}\right) g''(\bar{x}_J). \tag{2.9}$$

PROOF. Since the departures from linearity in the model are nonstationary, even locally, we will follow the development of a representation given in Portnoy (1991). Many of the details will be relegated to that paper.

(i) The first step is to develop the gradient conditions. In the classical linear case, this involves taking directional derivatives with respect to the parameters of a perturbation of the linear fit. Here we need a local perturbation, that is, a perturbation defined on J. To do this, let $\check{J}$ be the interval J with both endpoints deleted. For a and b small, define

$$\begin{aligned}\check{g}(x) &= (\hat{\alpha}+a)+(\hat{\beta}+b)(x-\bar{x}_J),\\ \min\{x_i\colon i\in\check{J}\} &\le x \le \max\{x_i\colon i\in\check{J}\}.\end{aligned} \tag{2.10}$$

For a and b sufficiently small, we may extend this linear function and (if needed) the preceding and succeeding linear segments of $\hat{g}$ so that these segments meet within a length $1/n$ from the endpoints of J. Thus, a continuous, piecewise linear function $\check{g}$ is formed which is a linear perturbation of $\hat{g}$ on $\check{J}$ and agrees with $\hat{g}$ off J (and either equals $\hat{g}$ or is part of the linear segment in $\check{J}$ at the two endpoints of J).

Now, the values $a = 0$ and $b = 0$ minimize the difference Δ between the objective function (1.2) at $\check{g}$ and its value at $\hat{g}$. The contribution to the roughness penalty depends on whether or not $\hat{g}'$ is monotonic [as described above in (2.3)]. If the slopes are monotonic, there is no contribution to the difference in roughness penalty; but otherwise each slope change contributes a value b. Thus, the difference Δ is

$$\Delta = \sum_{i\in J}[\rho_\tau(Y_i-\check{g}(x_i))-\rho_\tau(Y_i-\hat{g}(x_i))]+2|b|I^*[\sim \mathrm{mon}(J)]\lambda. \tag{2.11}$$

Following the usual approach originally given in Koenker and Bassett (1978), the partial directional derivatives of the terms in (2.11) evaluated at $a = 0$ and $b = 0$ must sum to zero except for contributions at points where $Y_i = \hat{g}(x_i)$. Since these contributions are bounded, the a-partials and the b-partials yield (2.4) and (2.5).

(ii) The second part of the proof is to obtain the following uniform approximation: let $D \subset \mathbf{R}^2$ be the set defined by

$$D \equiv \left\{(\delta_1, \delta_2)\colon |\delta_k| \le c_1\left(\frac{m}{n}\right)^{1-k} m^{-1/2}(\log n)^{1/2},\ k=1,2\right\} \tag{2.12}$$

where c_1 is a constant. Let $z_i \equiv (1, (x_i-\bar{x}_J))'$. For $(\delta_1, \delta_2) \in D$, define

$$\begin{aligned}T(\delta_1,\delta_2) \equiv{}& \sum_{i\in J} z_i\psi_\tau(U_i+F^{-1}(\tau)-r_{Ji}+\delta_1+\delta_2(x_i-\bar{x}_J))\\ &-\sum_{i\in J} z_i\psi_\tau(U_i+F^{-1}(\tau)),\end{aligned} \tag{2.13}$$

where r_{Ji} is the bias of $g(x)$ from its linear approximation: for $i \in J$,

$$\begin{aligned}r_{Ji} &\equiv g(x_i)-g(\bar{x}_J)-g'(\bar{x}_J)(x_i-\bar{x}_J)\\ &= \frac{1}{2}g''(\bar{x}_J)(x_i-\bar{x}_J)^2+o\left(\left(\frac{m}{n}\right)^2\right).\end{aligned} \tag{2.14}$$

Then, with E^* denoting the expectation assuming J is fixed (not random),

$$\sup_{(\delta_1,\delta_2)\in D} \left|T(\delta_1, \delta_2) - E^* T(\delta_1, \delta_2)\right| = \mathscr{O}_p(m^{1/4}(\log n)). \tag{2.15}$$

This result is proved using a Bernstein exponential inequality for fixed (δ_1, δ_2) and then using the chaining argument. The proof follows those of Lemmas 3.3–3.5 in Portnoy (1991). The calculations of $E^* T(\delta_1, \delta_2)$ and $\mathrm{Var}^*\, T(\delta_1, \delta_2)$ are straightforward generalizations of the calculations in Portnoy (1991). The factor of $(\log n)$ in the stochastic error term permits the exponential inequality to bound the probability that the error exceeds this order by a value tending to zero faster than any fixed power of n. Since the endpoints of J are observations, there are at most $n(n-1)/2$ such intervals. Therefore, the stochastic error term in (2.15) is uniform over all such intervals, J.

(iii) Let $\hat{\delta}_1 = \hat{\alpha} - \alpha$ and $\hat{\delta}_2 = \hat{\beta} - \beta$. Since $E^* T(\delta_1, \delta_2)$ is linear in δ_1 and δ_2, (2.15) establishes a relationship among $(\hat{\alpha}, \hat{\beta})$, the representation on the right-hand side of (2.7) and (2.8), and the sums on the left-hand side of the gradient conditions (2.4) and (2.5); at least for $(\hat{\delta}_1, \hat{\delta}_2) \in D$. If one can show that the gradient conditions must fail on the boundary of D, then the monotonicity argument of Jurečková (1977) [exactly as used in the proof of Theorem 3.1 in Portnoy (1991)] shows that the gradient conditions must also fail outside D. This shows that $(\hat{\delta}_1, \hat{\delta}_2)$ must lie in D; hence the relationship noted above yields the resulting representations immediately. The remaining complication here is to show that the representations (2.7) and (2.8) are actually smaller than the bounds defining the set D (2.12) (so that the gradient conditions fail on the boundary of D). But this follows by applying an exponential inequality to sums of the form $\sum_{i\in J} z_i \psi_\tau(U_i + F^{-1}(\tau))$ for any fixed interval J. That is, for a fixed (nonrandom) J (with size m),

$$\sum_{i\in J} z_i \psi(U_i + F^{-1}(\tau)) = \mathscr{O}\left(\left(\overset{m}{\sum_{i\in J}(x_i - \bar{x}_J)^2}\right)(\log n)^{1/2}\right), \tag{2.16}$$

except with probability bounded by n^{-a} for any constant a. Uniformity over the entire unit interval follows since there are at most $n(n+1)/2$ intervals J. Lastly, the bias is especially easy to compute: denote the deviation of g_0 from linearity by

$$r_i = \frac{1}{2} g_0''(\bar{x}_J)(x_i - \bar{x}_J)^2 + o\left(\frac{m^2}{n^2}\right). \tag{2.17}$$

To compute the bias term in (2.7), note that

$$\begin{aligned} &E^*\left[\psi_\tau(U_i + F^{-1}(\tau) + r_i) - \psi_\tau(U_i + F^{-1}(\tau))\right] \\ &\quad = -r_i f(F^{-1}(\tau) + o\left(\frac{m^2}{n^2}\right). \end{aligned} \tag{2.18}$$

Using (2.14), the bias contribution for $\hat{\beta}$ involves $\sum_{i\in J}(x_i - \bar{x}_J)^3$, which vanishes, while the contribution for $\hat{\alpha}$ uses $\sum_{i\in J}(x_i - \bar{x}_J)^2 = m^3/(12n^2)$. □

As noted above, the fact the endpoints of the interval J are random makes the sum in the representations a random one, and this reduces the usefulness of the representation. However, one immediate consequence of (2.7) is the analogue of the fundamental Koenker–Bassett result [Koenker and Bassett (1978)] that the τth regression quantile lies above a fraction $\tau \pm p/n$ of the observations. The proof follows that of Koenker and Bassett (1978).

COROLLARY 2.1. *Suppose the gradient condition* (2.4) *holds, and let J be a linear segment with $\sharp\{J\} = m$. Then the number of observations in J that lie below the spline, $\hat{g}(x) = \alpha + \beta(x - \bar{x}_J)$, is within a constant of the value τm.*

The proof is as in Koenker and Bassett (1978). The constant arises from the $\mathscr{O}(1)$ term in (2.4), and the fact that there is a possible contribution to the error from each of the endpoints of J and from at most two points fit exactly by the local linear segment of the spline (and each of these contributions is less than 1 in absolute value).

It is often important to consider the case of penalized B-splines, where the knots are specified to lie along a fixed grid. A representation similar to that of Theorem 2.1 can be developed, but the proof requires one important modification: the perturbation, $\breve{g}$ (2.10), is not itself a B-spline in that two of its breakpoints need not lie on the grid. Thus, we require an alternative perturbation, whose knots lie on the grid, and this requires some extra computation. The basic steps are quite similar, so the proof will only be sketched.

The asymptotic results again depend on local representations, which themselves depend critically on the number of observations along an interval. For B-splines, these intervals may be fixed to have exactly m observations each. Uniform error bounds will be of order $((\log m)/m)^{1/2}$, the $\log m$ factor arising from an exponential inequality that provides the uniformity over all intervals. On the other hand, the bias will have order m^2/n^2. These orders will be equal for $m = n^{4/5}(\log n)^{1/5}$, which would lead to the optimal rate for sup norm convergence: $\mathscr{O}(n^{-2/5}(\log n)^{2/5})$. For a fixed interval, $m = n^{4/5}$ would give a better result. Thus the formulation below includes both of these rates. In fact, as long as $m = n^d$ for some $d > 0$, the basic expansions below would hold.

THEOREM 2.2. *Let $\tilde{g}_n(x)$ be the penalized B-spline defined in Section* 1, *with λ_n arbitrary and with m, the number of observations between breakpoints, satisfying $m = n^{4/5}(\log n)^a$ for some real number a. Note that the interval length is $L_n \equiv m/n = n^{-1/5}(\log n)^a$. Let $\{x_j^*\colon j = 0, 1, \dots, n/m\}$ denote the breakpoints, and parameterize the B-spline by its values at x_j^*: $\theta_j \equiv g_0(x_j^*)$. Thus the penalized B-spline is the solution $\tilde{g}$ to* (1.2) *among piecewise linear functions with breakpoints only at $\{x_j^*\}$, and we may define $\tilde{g}$ by the parameter estimates $\tilde{\theta}_j \equiv \tilde{g}(x_j^*)$. Then the following representation holds uniformly in n and $\log n \le j \le n/m - \log n$:*

$$(\tilde{\theta}_j - \theta_j) = \sqrt{3}A_j^* + \sqrt{3}b_j^* + \mathscr{O}_p(m^{-3/4}(\log n)^{1/2}) + \mathscr{O}(n^{-3/5}(\log n)^{4a}) + \mathscr{O}(n\lambda_n/m^2). \tag{2.19}$$

Here

$$A_j^* \equiv \sum_{k=j-\log n}^{j+\log n} (-\gamma)^{|k|} A_{j-k} \tag{2.20}$$

with $\gamma = 2 - \sqrt{3}$ *and*

$$A_j \equiv \frac{1}{mf(F^{-1}(\tau))} \sum_{i=j-m}^{j+m} \left(\frac{m-|i|}{m}\right) \psi_\tau(U_i + F^{-1}(\tau)). \tag{2.21}$$

Also, $\{b_j^*\}$ *is the bias*:

$$b_j^* \equiv \frac{m^2}{12n^2} \sum_{k=j-\log n}^{j+\log n} (-\gamma)^{|k|} g_0''(x_{j-k}^*). \tag{2.22}$$

Consequently, *since* $\max_j |A_j^*| = \mathscr{O}_p((\log n/m)^{1/2})$ *and* $\max_j |b_j^*| = \mathscr{O}(m^2/n^2)$, *we may take* $a = 1/5$, *to obtain*

$$\max_j |\tilde{\theta}_j - \theta_j| = \mathscr{O}_p(n^{-2/5}(\log n)^{2/5}),$$
$$\max_j |\tilde{\beta}_j - \beta_j| = \mathscr{O}_p(n^{-1/5}(\log n)^{1/5}),$$

where $\tilde{\beta}_j = (\tilde{\theta}_{j+1} - \tilde{\theta}_j)/L_n$ *is the slope on the* j*th interval*. *Furthermore*,

$$\sup_{n^{-1/5}\log n \le x \le 1 - n^{-1/5}\log n} |\tilde{g}(x) - g_0(x)| = \mathscr{O}_p\left(n^{-2/5}(\log n)^{2/5}\right). \tag{2.23}$$

PROOF. The basic ideas are in the proof of Theorem 2.1, so only the important differences are presented here. Most important is the need for an alternative perturbation: fix j and consider $\theta_j \mapsto \theta_j + a$. This changes the penalized B-spline only on the intervals adjacent to x_j^*; by piecewise linearity, the perturbed B-spline is

$$g_a(x_{j+i}) = g(x_{j+i}) + \left(\frac{m-|i|}{m}\right) a, \qquad i = 0, 1, \dots, m. \tag{2.24}$$

It follows that the gradient condition becomes

$$\left| \sum_{i=-m}^{m} \left(\frac{m-|i|}{m}\right) \psi(Y_i - \tilde{g}(x_i)) \right| = \mathscr{O}(1) + \mathscr{O}\left(\frac{n\lambda_n}{m}\right), \tag{2.25}$$

where the second error term is the contribution of the smoothness penalty and follows from the fact that if the value changes by a, then the slope changes by a divided by the length of the interval ($L_n = m/n$). To compute the representation requires taking the expectation of the difference between the summand in the gradient condition and its value at the true mean, $g_0(x_i)$. With some work this expectation is

$$\begin{aligned} f(F^{-1}(\tau))\biggl\{ & mb_j + (\tilde{\theta}_{j-1} - \theta_{j-1}) \sum_{i=0}^{m} \frac{m-i}{m}\frac{i}{m} \\ & + (\tilde{\theta}_j - \theta_j) \sum_{i=-m}^{m} \frac{(m-|i|)^2}{m^2} + (\tilde{\theta}_{j+1} - \theta_{j+1}) \sum_{i=0}^{m} \frac{m-i}{m}\frac{i}{m}\biggr\} \\ & = mf(F^{-1}(\tau)\biggl\{ b_j + \frac{1}{6}(\tilde{\theta}_{j-1} - \theta_{j-1}) + \frac{2}{3}(\tilde{\theta}_j - \theta_j) \\ & \qquad + \frac{1}{6}(\tilde{\theta}_{j+1} - \theta_{j+1}) + \mathscr{O}(1)\biggr\} \\ & \equiv mf(F^{-1}(\tau)\biggl\{ b_j + \eta_j + \mathscr{O}(1)\biggr\}, \end{aligned} \tag{2.26}$$

where this last equation defines η_j, and where

$$\begin{aligned} mb_j &= \frac{1}{2}\sum_{i=-m}^{m} \frac{m-|i|}{m}\left(\frac{i}{n}\right)^2 g_0{}''(x_j^*) + \mathscr{O}\left(\frac{m^4}{n^3}\right) \\ &= \frac{m^3}{12n^2} g_0{}''(x_j^*) + \mathscr{O}\left(\frac{m^4}{n^3}\right). \end{aligned} \tag{2.27}$$

Now, developing an appropriate analogue of (2.15) and combining the expectation calculation and gradient condition above yields the following representation:

$$\begin{aligned} \eta_j = A_j + b_j &+ \mathscr{O}_p(m^{-3/4}(\log n)^{1/2}) + \mathscr{O}(n^{-3/5}(\log n)^{4a}) \\ &+ \mathscr{O}(n\lambda_n/m^2), \end{aligned} \tag{2.28}$$

where the middle big-$\mathscr{O}$ term combines the various errors above, and where A_j and b_j are given by (2.21) and (2.27), respectively.

To obtain a representation for $\tilde{\theta}_j$, a specific exponential smoothing can be applied to collapse η_j. Define $\gamma = 2 - \sqrt{3}$ as for (2.20) so that $\gamma^2 - 4\gamma + 1 = 0$. Then the exponential smoothing of $\{\eta_j\}$ telescopes, to give

$$\sum_{|k| \le \log n} (-\gamma)^{|k|}\eta_{j-k} = \left(\frac{2-\gamma}{3}\right)\left(\tilde{\theta}_j - \theta_j\right) + \mathscr{O}\left(\gamma^{\log n}\right). \tag{2.29}$$

Note that $\gamma^{\log n} \le 1/n$. Therefore, the desired representations (2.19) follow from (2.28) and the definitions of A_j^* and b_j^*. The remainder of the theorem follows from applying exponential bounds on the A_j's, which gives the $((\log m)/m)^{1/2}$ term in the errors that provides the uniformity of the bounds. The bound on the supremum uses the fact that $|\tilde{g}(x) - g_0(x)|$ is bounded by the difference at the nearest breakpoint plus $(n \log n)^{-1/5}$ times the slope at x. □

Finally, we formulate a result for the case where the breakpoints are completely specified as being equally spaced. Here the representation theorem is adequate for distributional results since the $\mathscr{O}(n\lambda_n/m^2)$ term does not appear. So consider functions $g(x)$ which are piecewise linear with fixed breakpoints at the indices jm, for $j = 1, \dots, p = [n/m]$. Again, take $m = n^{4/5}(\log n)^a$. Then the number of breakpoints is exactly $n/m = n^{1/5}(\log n)^{-a}$. This is just the usual case of B-splines, whose global asymptotics were given in He and Shi (1994). Define $\tilde{g}^*$ to be the function of this form minimizing the objective function (1.1) *without the penalty*. For each linear segment $J_j = \{i\colon jm \le i \le (j+1)m\}$, define the local parameter estimates $\tilde{\alpha}^*$ and $\tilde{\beta}^*$ so that $\tilde{g}^* = \tilde{\alpha}^*_j + \tilde{\beta}^*_j(x_i - \bar{x}_j)$ (for $i \in J_j$), where $\bar{x}_j$ corresponds to the midpoint of J_j. Then $\tilde{\alpha}^*_j$ and $\tilde{\beta}^*_j$ are simple linear functions of η_j; hence, from Theorem 2.2, these parameters satisfy representations from which asymptotic normality follows immediately. It remains to compute variances and the covariance (which is straightforward though tedious) to obtain the following result.

THEOREM 2.3. *Assume the conditions for Theorem* 2.2. *Then*, *under the above specification of a fixed grid for the breakpoints*,

$$m^{1/2}(\tilde{\alpha}^*_j - \alpha_j) - b_j \to_D \mathscr{N}\left(0, \frac{(3-\sqrt{3})}{4}\sigma^2_\tau\right), \tag{2.30}$$

$$(m^{1/2}L_n)(\tilde{\beta}^*_j - \beta_j) \to_D \mathscr{N}(0, 6(\sqrt{3}-1)\sigma^2_\tau), \tag{2.31}$$

where $\sigma^2_\tau \equiv \tau(1-\tau)/f^2(F^{-1}(\tau))$ *and* $(m^{1/2}L_n) = m^{3/2}/n$.

Again (*taking* $a = 1/5$), *uniform convergence of the* B-*splines holds*:

$$\sup_{0\le x\le 1} |\tilde{g}^*(x) - g_0(x)| = \mathscr{O}_p(n^{-2/5}(\log n)^{2/5}). \tag{2.32}$$

REMARKS. (i) Note that $\tilde{\alpha}^*_j$ and $\tilde{\beta}^*_j$ are asymptotically independent.

(ii) This result is a bit stronger than Theorem 2.2 in that it continues the result to the penultimate endpoints of the x-interval. This requires some additional work, or it can be obtained by recognizing that the B-spline problem without any penalty is just a $(p+1)$-parameter regression quantile problem with $p^3 \log n/n \to 0$. Hence, the results of Welsh (1989) apply. Again, somewhat complicated computations are needed to compute elements of the appropriate $(X'X)^{-1}$ matrix for the variances and covariances, but the same distributional results can be obtained.

3. Length of linear segments. The theorems of the preceding section provide immediately the appropriate rate of convergence for the parameters of a sufficiently long local linear segment. Here, Proposition 3.1 shows that linear segments that are *not* part of a convex or concave section of $\hat{g}(x)$ must be sufficiently long that convergence occurs at nearly the optimal rate ($\mathscr{O}_p(n^{-2/5}\log n)$). However, as noted in the Introduction, there appears to be no way to control the length of a segment along a concave or convex section of the spline. Nonetheless, it is relatively straightforward to show that the number of segments cannot be too large. As a consequence, it is possible to prove uniform convergence of $\hat{g}$, although at a rather slow rate. For penalized B-splines, of course, Theorem 2.2 provides a rate of uniform convergence that is optimal.

Now, consider linear segments that are not part of a convex or concave section of the spline, that is, linear segments for which the slope of the preceding segment, its own slope and the slope of the succeeding segment are not monotonic.

PROPOSITION 3.1. *Under the model assumptions of Section* 1, *assume that* $\lambda = cn^{1/5}$. *Let* $J \equiv \{i\colon i_1 \le i \le i_2\}$ *denote the indices corresponding to a "nonconcave" or "nonconvex" segment* (*as described above*), *and let the number of observations in* J *be* $m \equiv i_2 - i_1 + 1$ *be the number of observations in* J. *Then there is a constant* c *such that*, *with probability tending to* 1,

$$m \ge cn^{4/5}(\log n)^{-1/3}. \tag{3.1}$$

PROOF. Crude bounds will be used first to show that m is moderately large, and then more careful expansions will yield the result in (3.1). Let $\hat{g}(x) = \hat{\alpha} + \hat{\beta}(x - x_{i_1})$ be the linear segment of the spline along J (i.e., for $x = x_i = i/n$ with $i \in J$). Let $\breve{g}(x) = \hat{\alpha} + (\hat{\beta} + b)(x - x_{i_1})$ be a perturbation of $\hat{g}$ (and note that no perturbation of $\hat{\alpha}$ is needed here). As in (2.11), it is easy to see that

$$\begin{aligned}\Delta F &\equiv \sum_{i=1}^{n} \rho_\tau(Y_i - \breve{g}(x_i)) - \sum_{i=1}^{n} \rho_\tau(Y_i - \hat{g}(x_i))\\ &\le \sum_{i=0}^{m} |\breve{g}(x_i) - \hat{g}(x_i)| = b\sum_{i=0}^{m} \frac{i}{n} = \frac{bm(m+1)}{2n}.\end{aligned} \tag{3.2}$$

Let $\hat{\beta}_-$ and $\hat{\beta}_+$ denote the slopes of the preceding and succeeding linear segments. Consider the case where $\hat{\beta}_- > \hat{\beta}$ and $\hat{\beta}_+ > \hat{\beta}$ (the other case will follow analogously). Then, again, as in (2.11), the difference in the roughness penalty is

$$\begin{aligned}\Delta S &\equiv V(\breve{g}) - V(\hat{g})\\ &= \hat{\beta}_- - (\hat{\beta} + b) + \hat{\beta}_+ - (\hat{\beta} + b) - [\hat{\beta}_- - \hat{\beta} + \hat{\beta}_+ - \hat{\beta}] = -2b.\end{aligned} \tag{3.3}$$

As a consequence, since $\Delta F + \lambda\,\Delta S \ge 0$,

$$bm(m+1)/(2n) - 2b\lambda \ge 0 \quad \Rightarrow \quad m \ge 2cn^{3/5} - 1 \tag{3.4}$$

since $\lambda = cn^{1/5}$.

It is now possible to use (3.4) to obtain a more precise expansion of ΔF. In particular, (3.4) permits one to follow the proof of Lemma 3.1 in Gutenbrunner, Jurečková, Koenker and Portnoy (1993) and (with some effort) to obtain the following expansion: there are constants c_1 and c_2 such that, with probability tending to 1,

$$\left|\Delta F - b\sum_{i=0}^{m}\frac{i}{n}\psi_\tau(U_i)\right| \le c_1 b\left(\log n\sum_{i=0}^{m}\frac{i^2}{n^2}\right)^{1/2} + c_2 m^{-k} \tag{3.5}$$

for b sufficiently small and for fixed k (here k may be taken to be 3, and b may be taken to be m^{-3} also). The first (c_1) bound comes from the exponential inequality and is essentially $(\log n)^{1/2}$ times the standard deviation of ΔF (with $\hat{\alpha}$ and $\hat{\beta}$ taken as fixed); and the second (c_2) bound comes from the contribution of the discontinuity of ψ_τ and uses the argument at the end of the proof of Lemma 3.1 in Gutenbrunner, Jurečková, Koenker and Portnoy (1993). Combining the inequality $\Delta F + \lambda\,\Delta S \ge 0$ with (3.3) and (3.5), one immediately obtains

$$m^{3/2} \ge c_1 n^{6/5}(\log n)^{-1/2} + c_2 m^{-k}/b, \tag{3.6}$$

with probability tending to 1; the result (3.1) follows by taking $b \le m^{-k}$. □

A precise result on the number of breakpoints and, consequently, on the size of the locally linear seqments will now be presented. Let $\hat{g}(x)$ be a quantile smoothing spline satisfying (1.2).

LEMMA 3.1. *Assume conditions* F *and* G *of Section* 2, *and suppose* $\lambda_n = \mathscr{O}(n^{1/5})$. *Let* p *be the number of interpolated points, that is,* $\hat{g}(x_{i_j}) = Y_{i_j}$ *for* $j = 1, \dots, p$. *Then, with probability tending to* 1, $p = \mathscr{O}(n^{11/15})$.

PROOF. The result of Shen(1994) giving convergence of the quantile splines in the L_2-norm at the optimal rate will be applied. To do this, it is necessary to bound the average squared error along the grid of all observations below by the L_2-norm. To obtain this bound, consider an interval (x_i, x_{i+1}). On this interval, $\hat{g}(x)$ is linear, and $g(x)$ is within c/n^2 of a linear function $g_L(x)$ [since $g''(x)$ is uniformly bounded]. Using linearity, direct computation gives

$$\begin{aligned}\int_{x_i}^{x_{i+1}}(\hat{g}(x) - g_L(x))^2 &= \frac{1}{3n}\{\Delta g(x_{i+1})^2 + \Delta g(x_{i+1})\,\Delta g(x_i) + \Delta g(x_i)^2\}\\ &\ge \frac{1}{6n}\{\Delta g(x_{i+1})^2 + \Delta g(x_i)^2\},\end{aligned} \tag{3.7}$$

where $\Delta g(x) \equiv \hat{g}(x) - g_L(x)$. It follows that

$$\|\hat{g} - g\|_{L_2}^2 \ge \frac{1}{3n}\sum_{i=1}^{n}(\hat{g}(x_i) - g(x_i))^2 - \frac{c}{n^2}. \tag{3.8}$$

Now, for interpolated points, $\hat{g}(x_{i_j}) - g(x_{i_j}) = U_{i_j}$. Thus, letting $|U|_{(k)}$ denote the ordered absolute values of the errors and using (3.8),

$$\begin{aligned}\|\hat{g} - g\|_{L_2}^2 + \frac{c}{n^2} &\geq \frac{1}{3n}\sum_{i=1}^{n}(\hat{g}(x_i) - g(x_i))^2 \\ &\geq \frac{1}{3n}\sum_{j=1}^{p}(\hat{g}(x_{i_j}) - g(x_{i_j}))^2 \geq \frac{1}{3n}\sum_{j=1}^{p}(|U|_{(j)})^2,\end{aligned} \tag{3.9}$$

with probability tending to 1. Now let $V = |U|$, let F_V denote the c.d.f. of V and note that the density f_V is bounded strictly above zero on any interval $[0, a]$. Thus, letting $Z_{(i)}$ denote uniform order statistics, and expanding F_V, the last sum in (3.9) is

$$\begin{aligned}\frac{1}{n}\sum_{j=1}^{p}(F_V^{-1}(Z_{(j)}))^2 &\geq \frac{1}{n}\sum_{j=1}^{p\wedge(n/2)}\left(\frac{Z_{(j)}}{f_V(c)}\right)^2 \\ &\geq \frac{c'}{n}\sum_{j=1}^{p\wedge(n/2)}\left(\frac{j}{n}\right)^2 \geq c_1\frac{p^3}{n^3},\end{aligned} \tag{3.10}$$

with probability tending to 1, where the last step uses well-known properties of the empirical (uniform) distribution function [see, e.g., Shorack and Wellner (1986)].

Now, the basic convergence result of Shen [(1994), Example 4, page 12] shows that with probability tending to 1,

$$\|\hat{g} - g\|_{L_2}^2 \leq c_2 n^{-4/5} \tag{3.11}$$

for some constant c_2. It follows from (3.10) and (3.11) that

$$p \leq c^* n\left(n^{-4/5} + \frac{c}{n^2}\right)^{1/3} = \mathscr{O}(n^{11/15}),$$

with probability tending to 1. □

Lemma 3.1 will now be applied to obtain the following uniform convergence result.

THEOREM 3.1. *Assume the conditions for Lemma* 3.1. *Then*

$$\sup_{x\in[0,1]} |\hat{g}_n(x) - g(x)| = \mathscr{O}_p(n^{-4/45}(\log n)^{1/2}). \tag{3.12}$$

REMARK. It is possible to get a slightly better rate of convergence using this approach by replacing the L_2-norms of (3.9) by L_r-norms. Lemma 3.1 would then provide a bound $p = \mathscr{O}(n^{-(3/5)(r/(r+1))})$. This is arbitrarily close to $\mathscr{O}(n^{3/5})$, which would lead to a bound of $\mathscr{O}(n^{-2/15}(\log n)^{1/2})$ in (3.12).

PROOF OF THOEREM 3.1. Consider linear segments of length less than n^{-a}, where $a = 37/45$, and let T denote the union of all such segments. By Lemma 3.1, the total length of T (and thus of any segment of T) is bounded by $pn^{-a} \le c^* n^{11/15-a} = c^* n^{-4/45}$. So points in such small segments must lie within $c^* n^{-4/45}$ of points in larger segments. Now use Theorem 2.1 with the number of points in the local linear segment exceeding n^{-a}. It follows that, uniformly for $x \notin T$,

$$|\hat{g}_n(x) - g(x)| = \mathscr{O}_p(n^{-(1-a)/2}(\log n)^{1/2}) = \mathscr{O}_p(n^{-4/45}(\log n)^{1/2}). \tag{3.13}$$

Now $\hat{g}'(x)$ is uniformly bounded in probability for the following reasons: the first and last linear segments have in fact nonmonotonic derivatives. Hence, on the first and last linear segments, $(\hat{\beta} - \beta) = \mathscr{O}_p(n^{-1/5}(\log n)^{1/2})$, by Proposition 3.1 and (2.8); and β is bounded by hypothesis G. Therefore, we have

$$\begin{aligned} \sup_{x\in[0,1]} |\hat{g}_n(x) - g(x)| &\le \sup_{x\notin T} |\hat{g}_n(x) - g(x)| + c_0 \times \text{length}(T) \\ &= \mathscr{O}_p(n^{-(1-a)/2}(\log n)^{1/2} + n^{11/15-a}) \\ &= \mathscr{O}_p(n^{-4/45}(\log n)^{1/2}). \qquad \square \end{aligned} \tag{3.14}$$

4. Expansion of ρ: the SIC criterion. Koenker, Ng and Portnoy (1994) introduced the following "Schwarz information criterion" as a formal way of choosing the smoothing parameter λ:

$$SIC(\lambda) = \log\left(n^{-1}\sum_{i=1}^{n} \rho_\tau(Y_i - \hat{g}(x_i))\right) + \frac{\log n}{2n} p(\lambda), \tag{4.1}$$

where $p(\lambda)$ is the number of points interpolated exactly by $\hat{g}$. The coefficient of $p(\lambda)$ [i.e., $\log n/(2n)$] was chosen in exact analogy with the corresponding coefficient that Schwarz (1978) introduced to prevent overfitting. However, it is possible to expand the first term in (4.1) and show that the coefficient $(1/2)\log n$ should be replaced by a constant depending on τ, the error density and the true regression function g_0. This is rather reminiscent of the "bias corrected" AIC criterion of Hurvich and Tsai (1989), but where the normal assumption is avoided. The expansion here uses the local representations in the special case of Theorem 2.3 to obtain an expansion of the first term in (4.1)

First note that it is equivalent to use the following form for the *SIC* criterion, which will be denoted as ADIC (asymptotically defined information criterion):

$$\text{ADIC}(\lambda) \equiv \log\left(\frac{(1/n)\sum_{i=1}^{n} \rho_\tau(Y_i - \hat{g}(x_i))}{(1/n)\sum_{i=1}^{n} \rho_\tau(Y_i - g_\tau(x_i))}\right) + c_n \frac{p(\lambda)}{n}, \tag{4.2}$$

where c_n is a quantity to be analyzed. This form immediately permits application of the following expansion.

THEOREM 4.1. *Assume the conditions for Theorem* 2.3 *with* $p = [n^{1/5}]$ *fixed breakpoints lying on a lattice. Then*

$$\sum_{i=1}^{n} \rho_\tau(Y_i - \hat{g}(x_i)) - \sum_{i=1}^{n} \rho_\tau(Y_i - g_\tau(x_i)) = -c_0 p + o(p), \tag{4.3}$$

where

$$c_0 \equiv \frac{\tau(1-\tau)}{f(F^{-1}(\tau))} + \frac{\tau(1-\tau) f(F^{-1}(\tau))}{1152} \int_0^1 (g_0''(x))^2\,dx. \tag{4.4}$$

PROOF. Let J_j denote the jth linear segment for $j = 1, \dots, p$, and let $\bar{x}_j$ denote the midpoint of the corresponding segment. Then the difference of the ρ-functions in (4.3) can be written as $\sum_{j=1}^{p} \Delta_j$, where

$$\Delta_j \equiv \sum_{i \in J_j} \{\rho_\tau(Y_i - \hat{g}(x_i)) - \rho_\tau(U_i + F^{-1}(\tau) + r_{ij})\}, \tag{4.5}$$

where r_{ij} is the deviation of g_0 from its linear part:

$$r_{ij} \equiv \tfrac{1}{2} g_0''(\bar{x}_j)(x_i - \bar{x}_j)^2 + o(d^2). \tag{4.6}$$

Here d is the segment length, $d = 1/p \approx n^{-1/5}$. Now, applying equation (3.36) of Gutenbrunner, Jurečková, Koenker and Portnoy (1993), it is not difficult to express Δ_j as follows:

$$\Delta_j = -\frac{1}{2}\left(\frac{1}{f(F^{-1}(\tau))} \left\| Q_n^{-1/2} \sum_{i \in J_j} z_i \psi_\tau(U_i + F^{-1}(\tau) + r_{ij}) \right\|^2\right) + o_p(1), \tag{4.7}$$

where

$$z_i = \begin{pmatrix} 1 \\ x_i - \bar{x}_j \end{pmatrix}, \qquad Q_n = \begin{pmatrix} nd & 0 \\ 0 & nd^3/12 \end{pmatrix}. \tag{4.8}$$

It follows that there is a random variable $W \sim \chi_2^2$ such that

$$\Delta_j = -\frac{\tau(1-\tau)}{2 f(F^{-1}(\tau))} \left(W_j + \left\| \sum_{i \in J_j} Q_n^{-1/2} z_i (\text{bias}_{ij}) \right\|^2 \right), \tag{4.9}$$

where $(\text{bias}_{ij}) = r_{ij} f(F^{-1}(\tau)) + o(d^2)$. Note that the error terms are uniform in j. Therefore,

$$\begin{aligned} \sum_{i \in J_j} z_i (\text{bias}_{ij}) &= \tfrac{1}{2} f(F^{-1}(\tau)) \begin{pmatrix} \sum_{i \in J_j} g_0''(\bar{x}_j)(x_i - \bar{x}_j)^2 \\ \sum_{i \in J_j} g_0''(\bar{x}_j)(x_i - \bar{x}_j)^3 \end{pmatrix} \\ &= \begin{pmatrix} f(F^{-1}(\tau)) g_0''(\bar{x}_j) n d^3/24 \\ 0 \end{pmatrix}. \end{aligned} \tag{4.10}$$

It follows that

$$(4.11)\qquad \left\| \sum_{i\in J_j} Q_n^{-1/2} z_i(\mathrm{bias}_{ij}) \right\|^2 = f^2(F^{-1}(\tau))(g_0''(\bar{x}_j))^2 nd^5/576 + o(d^2).$$

Using the fact that $\sum_{j=1}^{p} \chi_2^2 = 2p + o_p(p)$ and inserting (4.11) (with $nd^5 = 1$) into (4.9),

$$(4.12)\qquad \sum_{j=1}^{p} \Delta_j = -p\left(\frac{\tau(1-\tau)}{f(F^{-1}(\tau))} + \frac{\tau(1-\tau)f(F^{-1}(\tau))}{1152} \int_0^1 (g_0''(x))^2\,dx \right) + o(p),$$

which is the desired result. □

Using Theorem 4.1, the following expansion for ADIC is immediate:

$$(4.13)\qquad \mathrm{ADIC} = -\frac{p}{n}\frac{c_0}{n^{-1}\sum_{i=1}^{n}\rho_\tau(Y_i - g_\tau(x_i))} + c_n\frac{p}{n} + o\left(\frac{p}{n}\right),$$

where c_0 is given by (4.4). It now follows that if ADIC is to be minimized, the linear coefficients of p in (4.13) must cancel. That is, we must have

$$(4.14)\qquad c_n = \frac{c_0}{n^{-1}\sum_{i=1}^{n}\rho_\tau(Y_i - g_\tau(x_i))}.$$

REMARKS. (i) The explicit appearance of the factor $g_0''(\bar{x}_j)$ emphasizes the necessity of assuming that g_0 is twice continuously differentiable. That is, g_0 is in fact smoother than the "bounded variation" functions over which the quantile smoothing spline is defined [see Koenker, Ng and Portnoy (1994)]. If g_0 is not sufficiently smooth, it is possible that the rate of convergence is slower than that given by the results here or by the results of Shen (1994).

(ii) As noted in the Introduction, parametric linear programming provides efficient computation of the ADIC function of (4.2) for all p. In fact, starting at the global linear fit ($p = 2$), one needs only to pivot until n is somewhat larger than $n^{1/5}$; although in the examples I have tried with n less than 300, it was little harder to compute the quantile smoothing splines for all the (finitely many) values of λ. The value p^* minimizing (4.2) [with c_n given by (4.14)] can be found by simple finite minimization. I conjecture that, under appropriate conditions, the corresponding $\hat{\lambda}_n(p^*)$ will be of order $n^{1/5}$. If this conjecture is not true, it would be possible to truncate $\hat{\lambda}_n$ to lie in an interval of the form $(an^{1/5}, bn^{1/5})$ with a small and b large. If this is done, it should be rather straightforward to show that the asymptotic results of Sections 2 and 3 will hold for the truncated $\hat{\lambda}_n$. The basic problem would be to consider a fixed linear segment J (as in Section 2) and to show that the sum of the "check" function in (4.2) may be replaced by a sum over $\{i \notin J\}$ without changing $\hat{\lambda}_n$ appreciably. Since the local asymptotic results depend only on observations in J, and $\hat{\lambda}_n$ is of order $n^{1/5}$ by construction, the results of Sections 2 and 3 would follow.

(iii) Application of the ADIC criterion with c_0 given by (4.4) clearly requires estimates of $f(F^{-1}(\tau))$, of the denominator in c_n and of $\int(g''(x))^2dx$.

This last ("curvature") term can be estimated in several ways. Perhaps the best approach is to use (4.10) directly. The quantity $g_0''(\bar{x}_j)$ could be estimated locally either using the slopes $\hat{\beta}_{j-1}$, $\hat{\beta}_j$ and $\hat{\beta}_{j+1}$ from the quantile smoothing spline or (perhaps more accurately) by refitting a local quadratic approximation near $\bar{x}_j$. Given density and c.d.f. estimates, the second term in (4.4) could be estimated by

$$\frac{1}{2}\tau(1-\tau)\hat{f}(\hat{F}^{-1}(\tau))\frac{1}{p}\sum_{j=1}^{p}\left(\frac{1}{2}\left(\frac{p}{n}\right)^{1/2}\sum_{i\in J_j}\hat{g}''(\bar{x}_j)(x_i-\bar{x}_j)^2\right)^2.$$

The density term, however, may require somewhat more work, although the experience of estimating $f(F^{-1}(\tau))$ reported in Portnoy and Koenker (1989) offers hope that this can be done in an effective manner. Whether or not this version of the information criterion is really appropriate will await extensive experience with examples and simulations. In some examples ADIC is close to the originally suggested SIC, and in most cases it seems to work remarkably well. Future work should clarify the utility of this potentially valuable approach.

5. Derivatives of QSS for estimating jump functions. For penalized B-splines, the derivative $\hat{g}'(x)$ converges to $g_\tau'(x)$ at the nearly optimal rate $n^{-1/5}(\log n)^{1/2}$, which is essentially the best one can expect without introducing more stringent smoothness assumptions. However, since $\hat{g}$ is piecewise linear, its derivative is a piecewise constant jump function. Thus, it may be possible to use derivatives of quantile smoothing splines to estimate (discontinuous) jump functions. The basic idea would be to integrate the data [i.e., take partial sums of Y_i times $\Delta x_i \equiv (x_i - x_{i-1})$], fit a quantile smoothing spline $\hat{h}_\tau(x)$ to the integrated data and then differentiate: $\hat{g}_\tau(x) = \hat{h}_\tau'(x)$. Here, two examples of this idea will be presented. The first is a simulated jump function, and the second is a simulated density, with which most density estimators appear to have great difficulty.

The jump function example is as follows: for each interval $[j, j+1]$, for $j = 1, 2, 3, 4$, take 20 observations with x_i on a regular grid of mesh 0.05 and with $Y_i \sim \mathcal{N}(j, 1)$. To describe the method explicitly, let $\tilde{Y}_i = 0.05\sum_{k=1}^{i} Y_k$, fix $\tau = 0.5$ and let $\hat{h}_\lambda(x_i) = \hat{\alpha}_j(\lambda) + \hat{\beta}_j(\lambda)(x_i - \bar{x}_j)$ be the L_1 ($\tau = 0.5$ quantile smoothing spline) fit to $\tilde{Y}_i$ (where the subscript j indexes the jth linear segment J_j of the spline). Define the estimator of the original jump function (i.e., the fit to the original Y_i) by $\hat{g}_\lambda(x_i) = \hat{\beta}_j(\lambda)$ for $i \in J_j$. Although the error structure for the partial sums is no longer i.i.d. globally, it may not be too far from i.i.d. locally. That is, conditional on the value of the partial sums $\tilde{Y}_i$ at some point, the successive further errors for a relatively small range of x_i would still look somewhat stationary.

A potential problem here, however, is that g_0 is no longer smooth enough for the asymptotic theorems to hold. Nonetheless, it seems reasonable to examine these estimates. Choosing an appropriate λ-value is especially problematic. Visual assessment and some very rough calculations based on Theorem 4.1

suggest the possibility of using an ADIC criterion:

$$\log\left(\frac{1}{n}\sum_{i=1}^{n}\rho(\tilde{Y}_i - \hat{h}_\lambda(x_i))\right) + c_n\frac{p}{n},$$

with c_n between 10 and 50 (or so). Two solutions corresponding to "optimal" λ-values for $c_n = 40$ ($\lambda^* = 5.227$) and for $c_n = 20$ ($\lambda^* = 3.244$) are plotted in Figure 1, and show remarkable agreement with the true jump function (as well as moderate robustness to the choice of c_n). Use of the original SIC criterion with $(p\log n)/n$ does *not* work at all in this case. It suggests a solution with far too few breakpoints (i.e., with λ far too large).

The second example is a simulated density used by Roger Koenker as a classroom example. A random sample of size $n = 200$ was taken from a trimodal density defined as a weighted combination of three lognormals and plotted in Figure 2. Students in the class tried a variety of nonparametric density estimators, but the only method that could resolve the second mode very well was a version of Stone's logspline methods [see Kooperberg and Stone (1991)]. Here, the empirical distribution function was computed from the data, median smoothing splines were calculated, and they were differentiated to give piecewise constant density estimators. An ADIC criterion with $c_n = 10$ was applied to the spline estimate of the c.d.f., giving $\lambda = 3.831$. The corresponding density estimate together with a visually appealing one with a nearby λ-value are plotted in Figure 2. Here, the true density is smooth, so one might expect the slower convergence rate of the derivative to be a serious handicap; but the results appear to be remarkably good, especially if the piecewise flat appearance is acceptable.

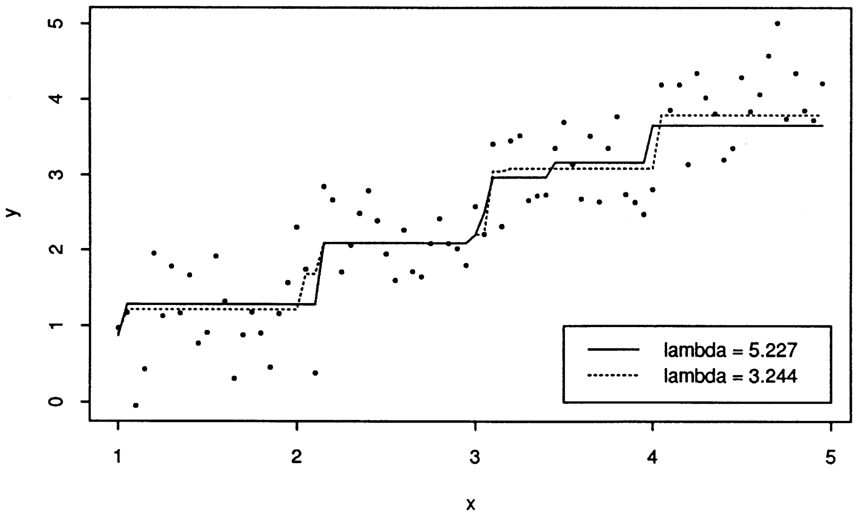

FIG. 1. *Derivatives of spline fits to integrated jump data.*

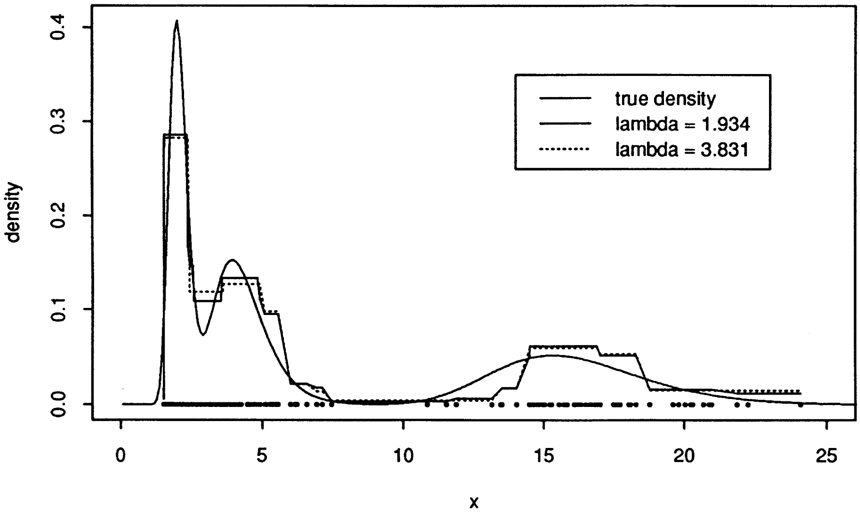

FIG. 2. *Derivatives of spline c.d.f. estimates.*

Clearly considerably more work is needed, but these initial attempts show real potential, especially considering the rather small amount of fine-tuning required. Note that since the computational method uses parametric programming, the computation of $\hat{g}_\lambda(x)$ for all λ is quite fast. Once this is done, viewing all solutions in turn or adjusting and applying various information criteria is computationally trivial.

Acknowledgments. The author expresses sincere gratitude to the referees and to Xuming He and Peide Shi for careful readings that led to significant improvements and corrections of errors in earlier versions.

REFERENCES

GUTENBRUNNER, C., JUREČKOVÁ, J., KOENKER, R. and PORTNOY, S. (1993). Tests of linear hypotheses based on regression rank scores. *J. Nonparametric Statist.* **2** 307–331.

HE, X. and SHI, P. (1994). Convergence rate of B-spline estimators of nonparametric conditional quantile functions. *J. Nonparametric Statist.* **3** 299–308.

HENDRICKS, W. and KOENKER, R. (1992). Hierarchical spline models for conditional quantiles and the demand for electricity. *J. Amer. Statist. Assoc.* **87** 58–68.

HURVICH, C. and TSAI, C.-L. (1989). Model selection in small samples. *Biometrika* **76** 297–307.

JUREČKOVÁ, J. (1977). Asymptotic relations of M-estimates and R-estimates in linear regression model. *Ann. Statist.* **5** 464–472.

KOENKER, R. and BASSETT G. (1978). Regression quantiles. *Econometrica* **46** 33–50.

KOENKER, R., NG, P. T. and PORTNOY, S. (1994). Quantile smoothing splines. *Biometrika* **81** 673–680.

KOOPERBERG, C. and STONE, C. (1991). A study of logspline density estimation. *Comput. Statist. Data Anal.* **12** 327–347.

MAMMEN, E. (1991). Nonparametric regression under qualitative smoothness assumptions. *Ann. Statist.* **19** 741–759.

MAMMEN, E. and VAN DE GEER, S. (1997). Locally adaptive regression splines. *Ann. Statist.* **25** 387–413.

PORTNOY, S. (1991). Asymptotic behavior of regression quantiles in non-stationary dependent cases. *J. Multivariate Anal.* **38** 100–113.

PORTNOY, S. and KOENKER, R. (1989). Adaptive L-estimation for linear models. *Ann. Statist.* **17** 362–381.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464.

SHEN, X. (1994). On the method of penalization. Technical report, Dept. Statistics, Ohio State Univ.

SHORACK, G. and WELLNER, J. (1986). *Empirical Processes with Applications to Statistics*. Wiley, New York.

WANG, Y. (1993). The limiting distribution in concave regression. Technical report, Univ. Missouri–Columbia.

WELSH, A. (1989). On M-processes and M-estimation. *Ann. Statist.* **17** 337–361.

DEPARTMENT OF STATISTICS
UNIVERSITY OF ILLINOIS
101 ILLINI HALL
725 S. WRIGHT STREET
CHAMPAIGN, ILLINOIS 61820
E-MAIL: portnoy@steve.stat.uiuc.edu