Competing statistics in exponential family regression models under certain shape constraints

Moulinath Banerjee[1]
University of Michigan

April 24, 2006

Abstract

We study a number of competing statistics for estimating the regression function in a family of conditionally parametric response models where, given a continuous covariate, the distribution of the response comes from a full rank exponential family with the parameter (which is in one to one correspondence with the conditional mean) being a shape constrained function of the covariate. Monotonicity, unimodality and U-shapes are considered. It is shown that the proposed competing statistics are approximately pivotal for large sample sizes and methods for constructing pointwise confidence sets for the regression function, away from a stationary point, are described. Results from a limited simulation study are presented.

Key words and phrases: competing statistics, conditionally parametric models, slope of greatest convex minorant, unimodal regression functions

1 Introduction and Background

Function estimation is a ubiquitous, and consequently well–studied problem in nonparametric statistics. In several scientific problems, qualitative background knowledge about the function is available, in which case it is sensible to incorporate such information in the statistical analysis. Shape–restrictions are typical examples of such qualitative knowledge, and appear in a large number of applications. In particular, monotonicity is a shape-restriction that shows up very naturally in different areas of application like reliability, renewal theory, epidemiology and biomedical studies. Closely related to monotonicity constraints are constraints like unimodality or U shapes/bath–tub shapes – functions satisfying such constraints are piecewise monotone. Some of the early work on monotone function estimation goes back to the 1950's. Grenander (1956) derived the MLE of a decreasing density as the slope of the least concave majorant of the empirical distribution function based on i.i.d. observations. The pointwise asymptotic distribution of Grenander's estimator was established by Prakasa Rao (1969). Brunk (1970) studied the

[1]Supported by NSF Grant DMS-0306235.

problem of estimating a monotone regression function in a signal plus noise model, with additive homoscedastic errors. A key feature of these monotone function problems is the slower pointwise rate of convergence (usually $n^{1/3}$) of the MLE (under the stipulation that the derivative of the monotone function at the point of interest does not vanish), as compared to the faster $\sqrt{n}$ rate in regular parametric models. Moreover, the pointwise limit distribution of the MLE turns out to be a non-Gaussian one, and seems to have first arisen in the work of Chernoff (1964).

In this paper, our goal is to study a number of competing statistics for estimating a regression function that is either monotone, or unimodal, or U–shaped, in a (general) setting where the conditional distribution of the response, given the covariate, comes from a full rank exponential family. We provide a generic description of such *conditionally parametric models* below, indicating why a study of such models is fraught with interest. In what follows, we initially assume that the regression function is monotone increasing. Having constructed the competing statistics and studied their limit behavior in this setting, we proceed to demonstrate how our methods extend to decreasing, unimodal and U–shaped regression functions.

1.1 Conditionally parametric response models: least squares and maximum likelihood estimates

Consider independent and identically distributed observations $\{Y_i, X_i\}_{i=1}^n$, where each (Y_i, X_i) is distributed like (Y, X), and (Y, X) is distributed in the following way: The covariate X is assumed to possess a Lebesgue density p_X (with distribution function F_X). The conditional density of Y given that $X = x$ is given by $p(\cdot, \psi(x))$, where $\{p(\cdot, \theta) : \theta \in \Theta\}$ is a one–parameter exponential family of densities (with respect to some dominating measure) parametrized in the natural or canonical form, and ψ is a smooth (continuously differentiable) monotone increasing function that takes values in Θ. Recall that the density $p(\cdot, \theta)$ can be expressed as:

$$p(y, \theta) = \exp[\theta T(y) - B(\theta)]h(y).$$

Now, it is easy to see that,

$$E[T(Y)|X = x] = B' \circ \psi(x) \equiv \mu(x) \text{ (say)}.$$

Since B is infinitely differentiable on Θ (an open set) and ψ is continuous, $B^{(k)} \circ \psi$ is continuous, for every $k > 0$. Moreover, for every θ we have: $B''(\theta) = I(\theta)$ where $I(\theta)$ is the information about θ and is equal to the variance of T in the parametric model $p(y, \theta)$. Therefore $B''(\theta) > 0$, which implies that B' is a strictly increasing function. It follows that B' is invertible (with inverse function, H, say), so that estimating the regression function μ is equivalent to estimating ψ. The function ψ is called the *link function* and as shown above, is in one–one correspondence with the monotone regression function μ. We give below several motivating examples of the above scenario that have been fairly well-studied in the literature.

Consider first, the *monotone regression model*. Here $Y_i = \mu(X_i) + \epsilon_i$ where $\{(\epsilon_i, X_i)\}_{i=1}^n$ are i.i.d. random variables, ϵ_i is independent of X_i, each ϵ_i has normal distribution with mean 0

and variance σ^2, each X_i has a Lebesgue density $p_X(\cdot)$ and μ is a monotone function. This model and its variants have been fairly well–studied in the literature on isotonic regression. Note, in particular, the reference to Brunk (1970) above. Here, $X \sim p_X(\cdot)$ and $Y \mid X = x \sim N(\mu(x), \sigma^2)$. This conditional density comes from the one–parameter exponential family $N(\eta, \sigma^2)$ (for fixed σ^2), η varying. To make the correspondence to the above framework, where we express the conditional density of Y in terms of the natural parameter, we take the natural sufficient statistic $T(Y) = Y$, the natural parameter $\theta = \eta/\sigma^2$, whence $B(\theta) = \theta^2\,\sigma^2/2$ and the link function $\psi(x) = \mu(x)/\sigma^2$.

Yet another example is the *binary choice model* under a monotonicity constraint. Here, we have a dichotomous response variable $Y = 1$ or 0 and a continuous covariate X with a Lebesgue density $p_X(\cdot)$ such that $P(Y = 1 \mid X) \equiv G(X)$ is a smooth increasing function of X. Thus, conditional on X, Y has a Bernoulli distribution with parameter $G(X)$. To cast this model in terms of the natural parameter, we take the link function as $\psi(x) = \log\left(G(x)/(1 - G(x))\right)$ and $p(\delta, \theta) = \delta\theta - \log(1 + e^\theta)$ for $\theta \in \mathbb{R}$, whence the natural sufficient statistic $T(\delta) = \delta$, $B(\theta) = \log(1 + e^\theta)$and $\mu(x) = G(x)$. Models of this kind have been quite broadly studied in econometrics and statistics (see, for example, Dunson (2004), Newton, Czado and Chappell (1996), Salanti and Ulm (2003)). In a biomedical context one could think of Y as representing the indicator of a disease/infection and X the level of exposure to a toxin, or the measured level of a bio-marker that is predictive of the disease/infection. In such cases it is often natural to impose a monotonicity assumption on G. An important special case of the above model is the "Current Status Model" from survival analysis that arises extensively in epidemiology and has received much attention among biostatisticians and statisticians (see, for example, Sun and Kalbfleisch (1993), Sun (1999), Shiboski (1998), Huang (1996), Banerjee and Wellner (2001)).

Finally consider the *Poisson regression model*: $X \sim p_X(\cdot)$ and $Y \mid X = x \sim \text{Poisson}(\lambda(x))$ where λ is a monotone function. We have n i.i.d. observations from this model. Here, $T(x) = x$, $\psi(x) = \log\,\lambda(x)$, and $p(y, \theta) = -e^\theta + y\,\theta - \log y!$, for $\theta \in \mathbb{R}$ and $\mu(x) = \lambda(x)$. Here one can think of X as the distribution of a region from a hazardous point source (for example, a nuclear processing plant or a mine) and Y the number of cases of disease incidence at distance X (say, cancer occurrences due to radioactive exposure in the case of the nuclear processing plant, or Silicosis in the case of the mine). Given $X = x$, the number of cases of disease incidence Y at distance x from the source is assumed to follow a Poisson distribution with mean $\lambda(x)$ where λ can be expected to be monotonically decreasing in x, since the harmful effect should decay as we move further out from the source. Variants of this model are quite well studied in epidemiological contexts (Stone (1988), Diggle, Morris and Morton–Jones (1999), Morton–Jones, Diggle and Elliott (1999)).

Let $\hat{\mu}_n$ denote the least squares estimate of μ and let $\hat{\mu}_n^0$ denote the constrained least squares estimate of μ, computed under the null hypothesis that $\mu(x_0) = \eta_0$ (equivalently $\psi(x_0) = \theta_0 \equiv H(\eta_0)$), for some interior point x_0 in the domain of p_X. Our goal is to characterize these estimates; as we will see shortly, this will lead directly to a characterization of the unconstrained and constrained MLEs of the function ψ. We first present some background on solving least squares problems under monotonicity constraints.

Cumulative sum diagram and greatest convex minorant: Consider a set of points in $\mathbb{R}^2$, $\{(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)\}$, where $x_0 = y_0 = 0$ and $x_0 < x_1 < \ldots < x_n$. Let $P(x)$ be the left–continuous function such that $P(x_i) = y_i$ and $P(x)$ is constant on (x_{i-1}, x_i). We will denote the vector of slopes (left–derivatives) of the greatest convex minorant (henceforth GCM) of $P(x)$ computed at the points $(x_1, x_2, \ldots, x_n)$ by $\text{slogcm}\,\{(x_i, y_i)\}_{i=0}^n$. The GCM of $P(x)$ is, of course, also the GCM of the function that one obtains by connecting the points $\{x_i, y_i\}_{i=0}^n$ successively, by means of straight lines. The slope of the convex minorant plays an important role in the characterization of solutions to least squares problems under monotonicity constraints.

Let $\mathcal{X} = \{x_1 < x_2 < \ldots < x_k\}$ be a linearly ordered set and let w be a positive (weight) function defined on this set. Let g be an arbitrary real-valued function defined on this set. Define $\mathcal{F}$ as the set of all functions f on $\mathcal{X}$ that are increasing with respect to the ordering on $\mathcal{X}$; i.e. $f(x_1) \leq f(x_2) \leq \ldots \leq f(x_k)$. Denote by $g^\star$, the least squares projection of g onto $\mathcal{F}$, with respect to the weight function w. Thus,

$$g^\star = \text{argmin}_{f \in \mathcal{F}} \sum_{i=1}^n (f(x_i) - g(x_i))^2 \, w(x_i) \quad (\star) \ .$$

It is well known (see, for example Robertson, Wright and Dykstra (1988)) that:

$$\{g^\star(x_i)\}_{i=1}^n = \text{slogcm}\,\{(W_i, G_i)\}_{i=0}^n$$

where $W_i = \sum_{j=1}^i w(x_j)$ and $G_i = \sum_{j=1}^i g(x_j)\, w(x_j)$.

Least squares estimates of μ: We are now in a position to characterize the least squares estimate of μ_n. The unconstrained least squares estimate $\hat{\mu}$ is given by:

$$\hat{\mu_n} = \text{argmin}_{\mu \text{ increasing}} \sum_{i=1}^n (T(Y_i) - \mu(X_i))^2 \, .$$

Let $\{X_{(i)}\}_{i=1}^n$ denote the ordered values of the X_i's and let $Y_{(i)}$ denote the response value corresponding to $X_{(i)}$. Since μ is increasing, the minimization problem is readily seen to reduce to one of minimizing $\sum_{i=1}^n (T(Y_{(i)}) - \mu_i)^2$ over all $\mu_1 \leq \mu_2 \leq \ldots \leq \mu_n$ (where $\mu_i = \mu(X_{(i)})$). The solution to this problem is known, by the characterization of the least squares projection stated above. It follows directly (take the g_i's to be the $T(Y_{(i)})$'s, the ordered set to be $\{1, 2, \ldots, n\}$ and the weight function to be identically equal to $1/n$) that $\{\hat{\mu}_{ni}\}_{i=1}^n$, the minimizer over all $\mu_1 \leq \ldots \leq \mu_n$ is given by:

$$\{\hat{\mu}_{ni}\}_{i=1}^n = \text{slogcm}\,\{G_n(X_{(i)}), V_n(X_{(i)})\}_{i=0}^n$$

where

$$G_n(x) = \frac{1}{n} \sum_{i=1}^n 1(X_i \leq x) \equiv \mathbb{P}_n \, 1(X \leq x) \ \text{ and } \ V_n(x) = \frac{1}{n} \sum_{i=1}^n T(Y_i)\, 1(X_i \leq x) \equiv \mathbb{P}_n \, (T(Y)\, 1(X \leq x)) \, ,$$

and $\mathbb{P}_n$ is the empirical measure of the data points that assigns mass $1/n$ to each point (X_i, Y_i). We interpret $X_{(0)}$ as $-\infty$, so that $G_n(0) = V_n(0) = 0$. The (unconstrained) least squares estimate of μ is formally taken to be the piecewise constant right continuous function, such that $\hat{\mu}(X_{(i)}) = \hat{\mu}_{ni}$ for $i = 1, 2, \ldots, n$.

We next consider the problem of determining the constrained least squares estimator, where the constraint is given by $H_0 : \mu(x_0) = \eta_0$. It is easy to see that this amounts to solving two separate optimization problems: (a) Minimize $\sum_{i=1}^{m} (T(Y_{(i)} - \mu_i)^2$ over all $\mu_1 \leq \mu_2 \leq \ldots \leq \mu_m \leq \eta_0$ and (b) Minimize $\sum_{i=m+1}^{n} (T(Y_{(i)}) - \mu_i)^2$ over all $\eta_0 \leq \mu_{m+1} \leq \mu_{m+2} \leq \ldots \leq \mu_n$; here m is that integer for which $X_{(m)} < x_0 < X_{(m+1)}$. The vector that solves (a) (say $\{\hat{\mu}^0_{ni}\}_{i=1}^m$) is given by:

$$\{\hat{\mu}^0_{ni}\}_{i=1}^m = \text{slogcm}\,\{G_n(X_{(i)}), V_n(X_{(i)})\}_{i=0}^m \ \wedge \eta_0\,,$$

where the minimum is interpreted as being componentwise. On the other hand, the vector that solves (b) (say $\{\hat{\mu}^0_{ni}\}_{i=m+1}^n$) is given by:

$$\{\hat{\mu}^0_{ni}\}_{i=m+1}^n = \text{slogcm}\,\{G_n(X_{(i)}) - G_n(X_{(m)}), V_n(X_{(i)}) - V_n(X_{(m)})\}_{i=m}^n \ \vee \eta_0\,,$$

where the maximum is also interpreted as being componentwise. The constrained MLE $\hat{\mu}^0_n$ is then taken to be the piecewise constant right–continuous function, such that $\hat{\mu}^0_n(X_{(j)}) = \hat{\mu}^0_{nj}$ for $j = 1, 2, \ldots, n$ and $\hat{\mu}^0_n(x_0) = \eta_0$, and such that $\hat{\mu}^0_n$ has no jumps outside the set $\{X_{(j)}\}_{j=1}^n \cup \{x_0\}$.

The above characterization of the constrained least squares estimator follows from the following proposition.

Proposition: Let $\mathcal{X} = \{x_1 < x_2 < \ldots < x_k\}$ be an ordered set and let w be a positive weight function defined on $\mathcal{X}$. Let g be a real-valued function defined on $\mathcal{X}$ and (i) $\mathcal{F}_{u,\eta}$ denote the set of increasing functions defined on $\mathcal{X}$ that are bounded above by η, (ii) $\mathcal{F}_{l,\eta}$ denote the set of increasing functions defined on $\mathcal{X}$ that are bounded below by η. Then, the least squares projection of g on $\mathcal{F}_{u,\eta}$ is given by:

$$g^\star_{u,\eta} \equiv \text{slogcm}\,\{(W_i, G_i)\}_{i=0}^n \wedge \eta\,,$$

and the least squares projection of g on $\mathcal{F}_{l,\eta}$ is given by:

$$g^\star_{l,\eta} \equiv \text{slogcm}\,\{(W_i, G_i)\}_{i=0}^n \vee \eta\,.$$

We next introduce the Kuhn–Tucker theorem which plays a crucial role in characterizing the unconstrained and constrained maximum likelihood estimates (MLEs) of the link function ψ.

Kuhn-Tucker theorem: Let ϕ be a strictly convex function defined on $\mathbb{R}^n$ and potentially assuming values in the extended real line. Define $R = \phi^{-1}(\mathbb{R})$ and consider the problem of minimizing ϕ on R, subject to a number of inequality and equality constraints that may be written

as $g_i(x) \leq 0$ for $i = 1, 2, \ldots, k$ and $g_i(x) = 0$ for $i = k+1, \ldots, m$. Here, the g_i's are convex functions. Then $\hat{x} \in R$ uniquely minimizes ϕ subject to the m constraints if and only if there exist non-negative (Lagrange multipliers) $\lambda_1, \lambda_2, \ldots, \lambda_m$ such that (a) $\sum_{i=1}^{m} \lambda_i \, g_i(\hat{x}) = 0$ and (b) $\bigtriangledown \phi(\hat{x}) + G_{n\times m}^T \, \lambda_{m\times 1} = 0$, where $G_{m\times n}$ is the total derivative of the function $(g_1, g_2, \ldots, g_m)^T$ at the point $\hat{x}$.

Maximum likelihood estimators of ψ: The likelihood function for ψ, up to a multiplicative factor not depending on ψ, is given by:

$$L_n(\psi, \{Y_i, X_i\}_{i=1}^n) = \prod_{i=1}^{n} \exp(\psi(X_i)\, T(Y_i) - B(\psi(X_i)))\, h(Y_i)\,,$$

whence the log-likelihood function for ψ, up to an additive factor that does not depend on ψ, is given by:

$$l_n(\psi, \{Y_i, X_i\}_{i=1}^n) = \sum_{i=1}^{n} [\psi(X_i)\, T(Y_i) - B(\psi(X_i))] \equiv \sum_{i=1}^{n} [\psi(X_{(i)})\, T(Y_{(i)}) - B(\psi(X_{(i)}))]\,.$$

Writing $\psi(X_{(i)}) = \psi_i$, it is seen that the problem of computing $\hat{\psi}_n$, the unconstrained MLE reduces to minimizing $\phi(\psi_1, \psi_2, \ldots, \psi_n) \equiv \sum_{i=1}^{n} [-\psi_i \, T(Y_{(i)}) + B(\psi_i)]$ over $\psi_1 \leq \psi_2 \leq \ldots \leq \psi_n$. The strict convexity of B implies the strict convexity of ϕ and the Kuhn-Tucker theorem may be invoked with $g_i(\tilde{\psi}) = \psi_i - \psi_{i+1}$ for $i = 1, 2, \ldots, n-1$ (here $\tilde{\psi}$ denotes the vector $(\psi_1, \psi_2, \ldots, \psi_n)$). Denoting the minimizer of ϕ by $\hat{\psi}_n = (\hat{\psi}_{n1}, \hat{\psi}_{n2}, \ldots, \hat{\psi}_{nn})$, the conditions (a) and (b) of that theorem translate to: $\lambda_i = \sum_{j=1}^{i} (T(Y_{(j)}) - B'(\hat{\psi}_{nj})) \geq 0$ for $i = 1, 2, \ldots, n-1$, and $\sum_{j=1}^{n} (T(Y_{(j)}) - B'(\hat{\psi}_{nj})) = 0$.

We now set $\hat{\psi}_{ni} = H(\hat{\mu}_{ni})$ (where H is the inverse function of B'), so that $\hat{\mu}_{ni} = B'(\hat{\psi}_{ni})$ and show that the above conditions are satisifed, whence it will follow immediately that the unconstrained MLE $\hat{\psi}_n = H(\hat{\mu}_n)$. Let $B_1, B_2, \ldots, B_r$ denote the (ordered) blocks of indices on which the solution $\hat{\mu}_n$ is constant, and let c_i denote the constant value of $\hat{\mu}_{nj}$ for $j \in B_i$; let $m_1 < m_2 < \ldots < m_r \equiv n$ denote the end-points of these blocks. Then, from the characterization of the isotonic regression, it follows that for each $1 \leq i \leq r$ (interpret m_0 as 0):

$$\frac{\sum_{j=m_{i-1}+1}^{m_i} T(Y_{(j)})}{m_i - m_{i-1}} = c_i \tag{1}$$

and for $m_{i-1} < s < m_i$

$$\frac{\sum_{j=m_{i-1}+1}^{s} T(Y_{(j)})}{s - m_{i-1}} \geq c_i\,. \tag{2}$$

This shows that

$$\lambda_i \equiv \sum_{j=1}^{i} (T(Y_{(j)}) - \hat{\mu}_{nj}) = 0 \;\text{ for }\; i = m_1, m_2, \ldots, m_{r-1}\,, \tag{3}$$

since

$$\lambda_{m_l} - \lambda_{m_{l-1}} = \sum_{j=m_{l-1}+1}^{m_l} (T(Y_{(j)}) - \hat{\mu}_{nj}) = (m_l - m_{l-1}) \left(\frac{\sum_{j=m_{l-1}+1}^{m_l} T(Y_{(j)})}{m_l - m_{l-1}} - c_i \right) = 0\,,$$

using (1), for $l = 1, 2, \ldots, r-1$ (interpret λ_0 as 0). Furthermore

$$\sum_{j=1}^{n} (T(Y_{(j)}) - B'(\hat{\psi}_{nj})) = \sum_{l=1}^{r} (m_l - m_{l-1}) \left(\frac{\sum_{j=m_{l-1}+1}^{m_l} T(Y_{(j)})}{m_l - m_{l-1}} - c_i \right) = 0\,.$$

Next consider $1 \leq s \leq n-1$ such that s is not the end-point of a block, and suppose that s lies strictly between m_{l-1} and m_l (where $1 \leq l \leq r$). Then,

$$\lambda_s = \lambda_{m_{l-1}} + (s - m_{l-1}) \left(\frac{\sum_{j=m_{i-1}+1}^{s} T(Y_{(j)})}{s - m_{l-1}} - c_l \right) \geq 0\,,$$

using (3) and (2). This shows that our choice of $\{\hat{\psi}_{nj}\}_{j=1}^{n}$ does indeed satisfy the Kuhn-Tucker conditions and completes the argument.

As in the case of the unconstrained least squares estimator, we can show, by splitting the likelihood maximization problem into two parts, and subsequently invoking the Kuhn-Tucker theorem, that the constrained MLE $\hat{\psi}_n^0$, computed under $H_0 : \psi(x_0) = \theta_0$ (where $\theta_0 = H(\eta_0)$), is given by $\hat{\psi}_n^0 = H(\hat{\mu}_n^0)$.

1.2 Competing statistics for estimating the regression function

We are now in a position to formulate a number of competing statistics to be used for estimating the value of μ (equivalently ψ) at a point. From the regression point of view, one natural statistic for testing the null hypothesis $H_0 : \mu(x_0) = \eta_0$ is the residual sum of squares given by:

$$RSS(\eta_0) = \sum_{i=1}^{n} (Y_i - \hat{\mu}_n^0(X_i))^2 - \sum_{i=1}^{n} (Y_i - \hat{\mu}_n(X_i))^2\,. \tag{4}$$

Large values of $RSS(\eta_0)$ provide evidence against the null hypothesis. To determine what is "large" will require investigation of the large sample distribution of $RSS(\eta_0)$, that we undertake in the next section. A competing statistic can be constructed by computing some sort of a global distance between the unconstrained and constrained least squares estimators. More specifically, consider:

$$L_2(\hat{\mu}_n, \hat{\mu}_n^0) = \sum_{i=1}^{n} (\hat{\mu}_n(X_i) - \hat{\mu}_n^0(X_i))^2 = n \int (\hat{\mu}_n - \hat{\mu}_n^0)^2 \, d\, G_n\,. \tag{5}$$

Once again, large values of this quantity provides evidence against the null hypothesis. In similar vein, we can consider:

$$L_2(\hat{\psi}_n, \hat{\psi}_n^0) = \sum_{i=1}^{n} (\hat{\psi}_n(X_i) - \hat{\psi}_n^0(X_i))^2 = n \int (\hat{\psi}_n - \hat{\psi}_n^0)^2 \, d\, G_n \, . \tag{6}$$

While the above provide valid measures of discrepancy, none of these use the likelihood function of the data to formulate the notion of discrepancy. The likelihood ratio statistic does precisely that by looking at the difference in the log–likelihood functions evaluated at the unconstrained and constrained MLEs. More precisely, the likelihood ratio statistic for testing H_0 is given by:

$$2 \, \log \, \lambda_n(\theta_0) = 2 \left\{ \sum_{i=1}^{n} [\hat{\psi}_n(X_{(i)}) \, T(Y_{(i)}) - B(\hat{\psi}_n(X_{(i)}))] - \sum_{i=1}^{n} [\hat{\psi}_n^0(X_{(i)}) \, T(Y_{(i)}) - B(\hat{\psi}_n^0(X_{(i)}))] \right\} . \tag{7}$$

The null hypothesis is rejected for large values of the likelihood ratio statistic. Of course, the maximum likelihood estimate $\hat{\psi}(x_0)$ can be used to make inference about $\psi(x_0)$. As will be shown later $n^{\gamma} \, (\hat{\psi}_n(x_0) - \psi(x_0))$ converges to a limit distribution, for some positive γ, which depends on the number of derivatives of ψ that vanish at the point x_0.

We finally define versions of the "score statistic" for this class of models. As will be seen later, these have natural connections to the likelihood ratio, least squares and the L_2 statistics introduced above. Consider, the log–likelihood function for the pair (Y, X):

$$l(Y, \psi(X)) \equiv \log \, p(Y, \psi(X)) = \psi(X) \, T(Y) - B(\psi(X)) \, .$$

The log–likelihood function for the data $\{Y_i, X_i\}_{i=1}^n$ is given by $l_n(\psi) = n \, \mathbb{P}_n \, l(Y, \psi(X))$, with $\mathbb{P}_n$ denoting the empirical measure of the data vector $\{Y_i, X_i\}_{i=1}^n$. Consider a perturbation of ψ in the direction of the monotone function η, defined by the parametric curve $\psi_{\eta,\epsilon} \equiv (1 - \epsilon) \, \psi(z) + \epsilon \, \eta(z)$. Set $l_{n,\epsilon} = n \, \mathbb{P}_n \, [\psi_{\eta,\epsilon}(X) \, T(Y) - B(\psi_{\eta,\epsilon}(X))]$. This can be viewed as the log–likelihood function from a one–dimensional model, parametrized by ϵ, and one can compute a score statistic at $\epsilon = 0$, by differentiating this parametric log–likelihood at $\epsilon = 0$. We get:

$$S_{n,\eta,\psi} = \frac{\partial}{\partial \epsilon} \, l_{n,\epsilon} \mid_{\epsilon=0} = n \, \mathbb{P}_n \, [(\eta(X) - \psi(X)) \, (T(Y) - B'(\psi(X))] \, .$$

Our proposed score statistics will be constructed by perturbing $\hat{\psi}_n$ in the direction of $\hat{\psi}_n^0$ and vice versa. Thus, we get:

$$S_{n,\hat{\psi}_n^0,\hat{\psi}_n} \equiv S_{n,1} = n \, \mathbb{P}_n \, \left[(\hat{\psi}_n^0(X) - \hat{\psi}_n(X)) \, (T(Y) - B'(\hat{\psi}_n(X))) \right]$$

and

$$S_{n,\hat{\psi}_n,\hat{\psi}_n^0} \equiv S_{n,2} = n \, \mathbb{P}_n \, \left[(\hat{\psi}_n(X) - \hat{\psi}_n^0(X)) \, (T(Y) - B'(\hat{\psi}_n^0(X))) \right] \, .$$

It is not difficult to see that $S_{n,1}$ is non-positive with probability one; this is a consequence of the fact that $\hat{\psi}_n$ is the MLE of ψ. The null hypothesis, $\psi(x_0) = \theta_0$ will be rejected for extreme values

(both large and small) of the score statistics. Note the contrast with the rejection region using the residual sum of squares, or likelihood ratio or L_2 statistics. With these statistics, it is sensible to reject only for large values, since each of these statistics will tend to increase as the data generating mechanism deviates more and more from the null hypothesis $\psi(x_0) = \theta_0$. In fact, small values are very compatible with the null hypothesis. However, with the score statistics the same cannot be inferred. The analytical forms of the statistics do not provide any insight regarding the nature of values of the score statistic (large of small) under deviation from the null hypothesis. All that may be inferred is that an atypical value of the score is less consistent with the null. Consequently, both small and large extremes need to be allowed.

In the next section, we study the limit distributions of these competing statistics under the null hypothesis, and show how the results can be used to construct various confidence intervals for ψ (equivalently μ) at a point of interest.

1.3 Relevant stochastic processes and derived functionals

To study the asymptotic distributions of these competing statistics, we introduce the relevant stochastic processes and certain derived functionals of these. For each $m \geq 1$ and for positive constants c and d, define $X_{c,d,m}(h) = cW(h) + d \mid h \mid^{m+1}$, for $h \in \mathbb{R}$. Here, $W(h)$ is standard two–sided Brownian motion starting from 0. For a real–valued function f defined on $\mathbb{R}$, let $\text{slogcm}(f, I)$ denote the left–hand slope of the GCM (greatest convex minorant) of the restriction of f to the interval I. We abbreviate $\text{slogcm}(f, \mathbb{R})$ to $\text{slogcm}(f)$. Also define:

$$\text{slogcm}^0(f) = (\text{slogcm}\,(f, (-\infty, 0]) \wedge 0)\, 1_{(-\infty,0]} + (\text{slogcm}\,(f, (0, \infty)) \vee 0)\, 1_{(0,\infty)}\,.$$

Set $g_{c,d,m} = \text{slogcm}(X_{c,d,m})$ and $g^0_{c,d,m} = \text{slogcm}^0\,(X_{c,d,m})$. The random function $g_{c,d,m}$ is increasing but piecewise constant with finitely many jumps in any compact interval. Also $g^0_{c,d,m}$, like $g_{c,d,m}$, is a piecewise constant increasing function, with finitely many jumps in any compact interval and differing, almost surely, from $g_{c,d,m}$ on a finite interval containing 0. In fact, with probability 1, $g^0_{c,d,m}$ is identically 0 in some random neighbourhood of 0, whereas $g_{c,d,m}$ is almost surely non-zero in some random neighbourhood of 0. Also, the length of the interval $D_{c,d,m}$ on which $g_{c,d,m}$ and $g^0_{c,d,m}$ differ is $O_p(1)$. The processes $g_{c,d,1}$ and $g^0_{c,d,1}$ in particular (slopes of convex minorants of Brownian motion with quadratic drift) are well-studied in the literature; see, for example, Banerjee and Wellner (2001) and Wellner (2003). The qualitative properties of the convex minorants, as described above, for $m > 1$ are similar to what one encounters in the case $m = 1$ (quadratic drift).

Brownian scaling allows us to relate the processes $(g_{c,d,m}, g^0_{c,d,m})$ to (g_m, g^0_m) where $g_m \equiv g_{1,1,m}$ and $g^0_m \equiv g^0_{1,1,m}$ are convex minorants of the canonical process $W(t) + \mid t \mid^{m+1}$. The following proposition holds.

Lemma 1 *The process* $\{(g_{c,d,m}(h), g^0_{c,d,m}(h) : h \in \mathbb{R}\}\}$ *has the same distribution as the process* $\{c\,(d/c)^{1/(2m+1)}\,(g_m((d/c)^{2/(2m+1)}\,h), g^0_m((d/c)^{2/(2m+1)}\,h)) : h \in \mathbb{R}\}$ *in the space* $\mathcal{L} \times \mathcal{L}$. *Here* $\mathcal{L}$ *denotes the space of monotone functions from* $\mathbb{R}$ *to* $\mathbb{R}$ *which are bounded on every compact set equipped with the topology of* L_2 *convergence (with respect to Lebesgue measure) on compact sets.*

Define random variables $\mathbb{D}_{c,d,m}, \mathbb{T}_{c,d,m}, \mathbb{M}_{1,c,d,m}, \mathbb{M}_{2,c,dm}$ in the following way:

$$\mathbb{D}_{c,d,m} = \int \{(g_{c,d,m}(h))^2 - (g^0_{c,d,m}(h))^2\}\, dh\,,$$

$$\mathbb{T}_{c,d,m} = \int (g_{c,d,m}(h) - g^0_{c,d,m}(h))^2\, dh\,,$$

$$\mathbb{M}_{1,c,d,m} = \int g^0_{c,d,m}(h)\,(g^0_{c,d,m}(h) - g_{c,d,m}(h))\, dh\,,$$

and

$$\mathbb{M}_{2,c,d,m} = \int g_{c,d,m}(h)\,(g_{c,d,m}(h) - g^0_{c,d,m}(h))\, dh\,,$$

and let $\mathbb{D}_m, \mathbb{T}_m, \mathbb{M}_{1,m}, \mathbb{M}_{2,m}$ denote the respective versions with $c = d = 1$. Using Lemma 1, we can show that the following holds.

Lemma 2 *We have:*

$$c^{-2}(\mathbb{D}_{c,d,m}, \mathbb{T}_{c,d,m}, \mathbb{M}_{1,c,d,m}, \mathbb{M}_{2,c,d,m}) \equiv_d (\mathbb{D}_m, \mathbb{T}_m, \mathbb{M}_{1,m}, \mathbb{M}_{2,m})\,.$$

See the appendix for a (partial) proof of this proposition, where we show the equality in distribution for the first component. The joint convergence can be proved by extending this argument but is skipped, since it is really the componentwise equality in distribution that gets used in studying the limit distributions of the competing statistics.

2 Limit distributions for the competing statistics and methodological implications

We will study the limit distribution of the competing statistics at a pre–fixed point x_0 under the assumption that the first $m-1$ derivatives of μ (and equivalently ψ) vanish at the point x_0 but the m'th does not, and is therefore strictly greater than 0 (under our assumption that μ is an increasing function). For $m = 1$, this reduces to the condition that the derivative at x_0 does not vanish. While the assumption of finitely many derivatives vanishing at x_0 is difficult to check from the methodological perspective (unless there happens to be compelling background knowledge that indicates that such is the case) and the case $m = 1$ is the one that can really be used effectively, formulating the results for a general m leads to a unified and aesthetically pleasing presentation of results at virtually no additional cost.

We first define localized versions of both the unconstrained and the constrained least squares estimates of μ, and the corresponding MLE's of ψ. Thus, we set:

$$X_n(h) = n^{m/(2m+1)}\,(\hat{\mu}_n(x_0 + h\,n^{-1/(2m+1)}) - \mu(x_0)) \;\;,\;\; Y_n(h) = n^{m/(2m+1)}\,(\hat{\mu}^0_n(x_0 + h\,n^{-1/(2m+1)}) - \mu(x_0))\,.$$

We also set:

$$\tilde{X}_n(h) = n^{m/(2m+1)}\,(\hat{\psi}_n(x_0 + h\,n^{-1/(2m+1)}) - \psi(x_0)) \;\;,\;\; \tilde{Y}_n(h) = n^{m/(2m+1)}\,(\hat{\psi}^0_n(x_0 + h\,n^{-1/(2m+1)}) - \psi(x_0))\,.$$

The following facts, to be (partially) established in the appendix, will be used.

Fact 1. The processes $(X_n(h), Y_n(h) : h \in \mathbb{R})$ converge in distribution to $(g_{a,b,m}(h), g^0_{a,b,m}(h) : h \in \mathbb{R})$ in the space $\mathcal{L} \times \mathcal{L}$, where $a = \sqrt{I(\psi(x_0))/p_X(x_0)}$ and $b = (1/(m+1)!) \mid \mu^{(m)}(x_0) \mid$.

Using the fact that $(\hat{\mu}_n(x), \hat{\mu}^0_n(x)) = (B'(\hat{\psi}_n(x)), B'(\hat{\psi}^0_n(x)))$ and the delta method, it is easily deduced from Fact 1 that $(\tilde{X}_n(h), \tilde{Y}_n(h) : h \in \mathbb{R})$ converge in distribution to $(g_{\tilde{a},\tilde{b},m}(h), g^0_{\tilde{a},\tilde{b},m}(h) : h \in \mathbb{R})$, where $\tilde{a} = \sqrt{1/(I(\psi(x_0))p_X(x_0))}$ and $b = (1/(m+1)!) \mid \psi^{(m)}(x_0) \mid$.

Fact 2. The estimators $\hat{\mu}_n$ and $\hat{\mu}^0_n$ differ on an interval D_n (around x_0) whose length is $O_p(n^{-1/(2m+1)})$.

Fact 3. Let J_n be the set of indices i such that $\hat{\mu}_n(X_{(i)}) \neq \hat{\mu}^0_n(X_{(i)})$. Then J_n can be broken up into ordered blocks of indices $\{B_j\}$ and $\{B^0_j\}$ such that the following holds: (a) For $i \in B_j$ $\hat{\mu}_n(X_{(i)})$ is constant and is given by $n_j^{-1} \sum_{i \in B_j} (T(Y_{(i)})$, n_j being the size of B_j. (b) A similar phenomenon holds for $\hat{\mu}^0_n$ on each B^0_j (which may be different from B_j) except on that single block where $\hat{\mu}^0_n$ assumes the constant value η_0.

We now state our main result.

Theorem 1 *Assume that the null hypothesis $\psi(x_0) = \theta_0$ (equivalently $\mu(x_0) = \eta_0$) holds. Then:*

(a) The statistic RSS(η_0) defined in (4) converges in distribution to $I(\psi(x_0))\, \mathbb{D}_m$, while the likelihood ratio statistic defined in (7) converges in distribution to $\mathbb{D}_m$.

(b) The statistic $L_2(\hat{\mu}_n, \hat{\mu}^0_n)$ converges in distribution to $I(\psi(x_0))\, \mathbb{T}_m$ while the statistic $L_2(\hat{\psi}_n, \hat{\psi}^0_n)$ converges in distribution to $(I(\psi(x_0))^{-1}\, \mathbb{T}_m$.

Define weighted versions of the statistics in (b) as follows. Set:

$$L_{2,w}(\hat{\mu}_n, \hat{\mu}^0_n) = \sum_{i=1}^{n} \frac{(\hat{\mu}_n(X_i) - \hat{\mu}^0_n(X_i))^2}{I(\hat{\psi}_n(X_i))}\,,$$

and

$$L_{2,w}(\hat{\psi}_n, \hat{\psi}^0_n) = \sum_{i=1}^{n} (\hat{\psi}_n(X_i) - \hat{\psi}^0_n(X_i))^2)\, I(\hat{\psi}_n(X_i))\,.$$

Both $L_{2,w}(\hat{\mu}_n, \hat{\mu}^0_n)$ and $L_2(\hat{\psi}_n, \hat{\psi}^0_n)$ converge to $\mathbb{T}_m$ in distribution.

(c) The statistic $S_{n,1}$ converges in distribution to $\mathbb{M}_{1,m}$ while the statistic $S_{n,2}$ converges in distribution to $\mathbb{M}_{2,m}$. Furthermore, $S_{n,1} + S_{n,2} = I(\psi(x_0))\, L_2(\hat{\psi}_n, \hat{\psi}^0_n) + o_p(1)$ while $S_{n,2} - S_{n,1} = 2 \log \lambda_n + o_p(1)$.

So, both the likelihood ratio and the L_2 statistics can be asymptotically linearly decomposed in terms of the score statistics.

(d) The statistic $n^{m/(2m+1)}(\hat{\mu}_m(x_0) - \eta_0)$ *converges in distribution to* $(a^{2m}\, b)^{1/(2m+1)}\, g_m(0)$.

Methodological consequences: The above theorem has significant methodological consequences for the estimation of the regression function μ (equivalently, the function ψ). The results in (a), (b) and (c) of the above theorem give a number of competing pivots through the inversion of which confidence sets for $\mu(x_0)$ can be obtained. Let $d_{m,\beta}, t_{m,\beta}, M_{1,m,\beta}$ and $M_{2,m,\beta}$ denote the β'th quantiles of the distributions of $\mathbb{D}_m, \mathbb{T}_m, \mathbb{M}_{1,m}$ and $\mathbb{M}_{2,m}$ respectively. Consider the null hypothesis $H_\eta : \mu(x_0) = \eta$ (equivalently $\psi(x_0) = H(\eta)$), and denote the constrained least squares estimate of μ under this hypothesis by $\hat{\mu}_n^\eta$ and the corresponding MLE of ψ by $\hat{\psi}_n^\eta$. Let $RSS(\eta)$ denote the residual sum of squares statistic for testing this hypothesis and 2 log $\lambda_n(H(\eta))$ denote the corresponding likelihood ratio statistic. From (a), we obtain two asymptotic level $1-\alpha$ confidence sets for $\mu(x_0)$ as:

$$\{\eta : I(H(\eta))^{-1}\, \mathrm{RSS}(\eta) \leq d_{m,1-\alpha}\} \text{ and } \{\eta : 2 \log \lambda_n(H(\eta)) \leq d_{m,1-\alpha}\}\,.$$

Confidence sets for $\mu(x_0)$ based on the first two statistics in Part (b) are given by:

$$\{\eta : I(H(\eta))^{-1}\, L_2(\hat{\mu}_n, \hat{\mu}_n^\eta) \leq t_{m,1-\alpha}\} \text{ and } \{\eta : I(H(\eta))\, L_2(\hat{\psi}_n, \hat{\psi}_n^\eta) \leq t_{m,1-\alpha}\}\,,$$

while, using the weighted versions we get confidence sets:

$$\{\eta : L_{2,w}(\hat{\mu}_n, \hat{\mu}_n^\eta) \leq t_{m,1-\alpha}\} \text{ and } \{\eta : L_2(\hat{\psi}_n, \hat{\psi}_n^\eta) \leq t_{m,1-\alpha}\}\,.$$

Using the results of Part (c), we get the following confidence sets:

$$\{\eta : M_{1,m,\alpha/2} \leq S_{n,\hat{\psi}_n^\eta,\hat{\psi}_n} \leq M_{1,m,1-\alpha/2}\} \text{ and } \{\eta : M_{2,m,\alpha/2} \leq S_{n,\hat{\psi}_n,\hat{\psi}_n^\eta} \leq M_{2,m,1-\alpha/2}\}\,.$$

The result in (d) can be used to construct a confidence set of the form $[\hat{\mu}(x_0) - n^{-m/(2m+1)}\, (\hat{a}^{2m}\, \hat{b})^{1/(2m+1)}\, q_{m,\alpha/2}, \hat{\mu}(x_0) + n^{-m/(2m+1)}\, (\hat{a}^{2m}\, \hat{b})^{1/(2m+1)}\, q_{m,\alpha/2}]$, where $q_{m,\alpha/2}$ is the $(1-\alpha/2)$'th quantile of the (symmetric) distribution of $g_{1,1,m}(0)$. When $m = 1$, the slope of the greatest convex minorant at 0 is distributed like twice the minimizer of $\{W(h)+h^2 : h \in \mathbb{R}\}$, whose distribution is very well–studied (see, for example, Groeneboom and Wellner (2001)) and is referred to in the literature as Chernoff's distribution. The prime issue with this method lies in the fact that the m'th derivative at the point x_0 needs to be estimated, and this is a difficult affair. However, it is possible to bypass parameter estimation in this case by resorting to resampling techniques. Efron's bootstrap does not work in this situation, but subsampling (see Politis, Romano and Wolf (1999)) does.

2.1 Incorporating further shape constraints

We now discuss how the above methodology can be extended to incorporate further shape constraints. Our discussion, thus far, has focused on estimating a monotone increasing regression function μ, but works equally well for decreasing functions, and also for unimodal/U–shaped regression functions, provided one stays away from the mode/minimizer of the regression function. We first investigate the case of a decreasing regression function.

Decreasing regression function: The results of Theorem 1 continue to hold for a decreasing regression function (under the assumption that the first $m-1$ derivatives of the decreasing regression function μ vanish at the point x_0 and the m'th does not). In this case, the unconstrained and constrained least squares estimates of μ are no longer characterized as the slopes of greatest convex minorants, but as slopes of least concave majorants. We present the characterizations of the least squares estimates in the decreasing case below. We first introduce some notation. For points, $\{(x_0, y_0), (x_1, y_1), \ldots, (x_k, y_k)\}$ where $x_0 = y_0 = 0$ and $x_0 < x_1 < \ldots < x_k$, consider the right–continuous function $P(x)$ such that $P(x_i) = y_i$ and such that $P(x)$ is constant on (x_{i-1}, x_i). We will denote the vector of slopes (left–derivatives) of the LCM (least concave majorant) of $P(x)$ computed at the points $(x_1, x_2, \ldots, x_k)$ by slolcm $\{(x_i, y_i)\}_{i=0}^{k}$. With G_n and V_n as before, it is not difficult to see that

$$\{\hat{\mu}_n(X_{(i)})\}_{i=1}^{n} = \text{slolcm } \left\{G_n(X_{(i)}), V_n(X_{(i)})\right\}_{i=0}^{n} . \tag{8}$$

Also, the MLE under $H_0 : \mu(x_0) = \eta_0$ is given by:

$$\{\hat{\mu}_n^0(X_{(i)})\}_{i=1}^{m} = \eta_0 \vee \text{slolcm } \left\{G_n(X_{(i)}), V_n(X_{(i)})\right\}_{i=0}^{m} , \tag{9}$$

where the maximum is interpreted as being taken componentwise, while

$$\{\hat{\mu}_n^0(X_{(i)})\}_{i=m+1}^{n} = \eta_0 \wedge \text{slolcm } \left\{G_n(X_{(i)}) - G_n(X_{(m)}), V_n(X_{(i)}) - V_n(X_{(m)})\right\}_{i=m}^{n} , \tag{10}$$

where the minimum is once again interpreted as being taken componentwise.

As with an increasing regression function, the unconstrained and constrained MLE's of ψ are given by $\hat{\psi}_n = H(\hat{\mu}_n)$ and $\hat{\psi}_n^0 = H(\hat{\mu}_n^0)$ respectively.

Unimodal regression functions: Suppose now that the regression function is unimodal. Thus, there exists $M > 0$ such that the regression function is increasing on $[0, M]$ and decreasing to the right of M, with the derivative at M being equal to 0. The goal is to construct a confidence set for the regression function at a point $x_0 \neq M$ under the assumption that the first $m-1$ derivatives of μ vanish at x_0 and the m'th does not. We consider the more realistic case for which M is unknown.

First compute a consistent estimator, $\hat{M}_n$, of the mode M. With probability tending to 1, $x_0 < \hat{M}_n$ if x_0 is to the left of M and $x_0 > \hat{M}_n$ if x_0 is to the right of M.

Assume first that $x_0 < M \wedge \hat{M}_n$. Let m_n be such that $X_{(m_n)} \leq \hat{M}_n < X_{(m_n+1)}$. Let $\hat{\mu}_n$ denote the unconstrained LSE of μ, using $\hat{M}_n$ as the mode. Then, $\hat{\mu}_n$ is obtained by minimizing $\sum_{i=1}^{n} (T(Y_{(i)}) - \mu_i)^2$ over all $\mu_1, \mu_2, \ldots, \mu_n$ with $\mu_1 \leq \mu_2 \ldots \leq \mu_{m_n}$ and $\mu_{m_n+1} \geq \mu_{m_n+2} \geq \ldots \geq \mu_n$. It is not difficult to verify that

$$\{\hat{\mu}_n(X_{(i)})\}_{i=1}^{m_n} = \text{slogcm } \left\{G_n(X_{(i)}), V_n(X_{(i)})\right\}_{i=0}^{m_n} .$$

while

$$\{\hat{\mu}_n(X_{(i)})\}_{i=m_n+1}^{n} = \text{slolcm } \left\{G_n(X_{(i)}) - G_n(X_{(m_n)}), V_n(X_{(i)}) - V_n(X_{(m_n)})\right\}_{i=m_n}^{n} .$$

Now, consider testing the (true) null hypothesis that $\mu(x_0) = \eta_0$. Let $m < m_n$ be the number of $X_{(i)}$'s that do not exceed x_0. Denoting, as before, the constrained MLE by $\hat{\mu}_n^0$, it can be checked that $\hat{\mu}_n^0(X_{(j)}) = \hat{\mu}_n(X_{(j)})$ for $j > m_n$, whereas

$$\{\hat{\mu}_n^0(X_{(i)})\}_{i=1}^{m} = \eta_0 \wedge \text{slogcm } \left\{G_n(X_{(i)}), V_n(X_{(i)})\right\}_{i=0}^{m} ,$$

and

$$\{\hat{\mu}_n^0(X_{(i)})\}_{i=m+1}^{m_n} = \eta_0 \vee \text{slogcm } \left\{G_n(X_{(i)}) - G_n(X_{(m)}), V_n(X_{(i)}) - V_n(X_{(m)})\right\}_{i=m}^{m_n} .$$

The corresponding unconstrained and constrained MLE's of ψ are obtained by transforming $\hat{\mu}_n$ and $\hat{\mu}_n^0$ by H. The competing statistics from Section 1.2 have the same form as in the monotone function case, with the effective contribution coming from response–covariate pairs with the covariate in a shrinking ($O_p(n^{-1/3})$) neighborhood of the point x_0. The asymptotic distributions of the competing statistics are identical to those in the monotone function case. For a similar result for the maximum likelihood estimator, in the setting of unimodal density estimation away from the mode, we refer the reader to Theorem 1 of Bickel and Fan (1996). A rigorous derivation of the asymptotics for the unimodal case involves some embellishments of the arguments in the monotone function scenario and are omitted. Intuitively, it is not difficult to see why the asymptotic behavior remains unaltered. The characterization of the MLE/LSE of ψ/μ on the interval $[0, M_n]$, with M_n converging to M is in terms of unconstrained/constrained slopes of convex minorants exactly as in the monotone function case. Furthermore, the behavior at the point x_0, which is bounded away from M_n with probability increasing to 1, is only influenced by the behavior of localized versions of the processes V_n and G_n in a shrinking $n^{-1/3}$ neighborhood of the point x_0 (where the unconstrained and the constrained MLE's differ), and these behave asymptotically in exactly the same fashion as for the monotone function case. Consequently, the behavior of the MLE's/LSE's (and consequently, the various competing statistics) stay unaffected. An asymptotic confidence interval of level $1 - \alpha$ for $\mu(x_0)$ can therefore be constructed in the exact same way, as for the monotone function case.

The other situation is when $M \vee \hat{M}_n < x_0$. In this case $\hat{\mu}_n$ has the same form as above. Now, consider testing the (true) null hypothesis that $\mu(x_0) = \eta_0$. Let m be the number of $X_{(i)}$'s such that $\hat{M}_n < X_{(i)} \leq x_0$. Now, $\hat{\mu}_n^0(X_{(j)}) = \hat{\mu}_n(X_{(j)})$ for $1 \leq j \leq m_n$, while

$$\{\hat{\mu}_n^0(X_{(i)})\}_{i=m_n+1}^{m_n+m} = \eta_0 \vee \text{slolcm } \left\{G_n(X_{(i)}) - G_n(X_{(m_n)}), V_n(X_{(i)}) - V_n(X_{(m_n)})\right\}_{i=m_n}^{m_n+m} ,$$

and

$$\{\hat{\mu}_n^0(X_{(i)})\}_{i=m_n+m+1}^{n} = \eta_0 \wedge \text{slolcm } \left\{G_n(X_{(i)}) - G_n(X_{(m_n+m)}), V_n(X_{(i)}) - V_n(X_{(m_n+m)})\right\}_{i=m_n+m}^{n} .$$

The competing statistics continue to have the same limit distributions and confidence sets may be constructed in the usual fashion.

U–shaped regression functions: Our methodology extends also to U-shaped regression

functions. A U-shaped function is a unimodal function turned upside down (we assume a unique minimum for the function). As in the unimodal case, once a consistent estimator of the point at which the regression function attains its minimum has been obtained, the competing statistics for testing the null hypothesis $\mu(x_0) = \eta_0$ can be conducted in a manner similar to the unimodal case. The alterations of the above formulas that need to be made are quite obvious, given that the regression function is now intially decreasing and then increasing. For the sake of conciseness, we have omitted these formulas. The limit distribution of the competing statistics are identical to those in the unimodal case.

Consistent estimation of the mode: It remains to prescribe a consistent estimate of the mode in the unimodal case. Let $\hat{\mu}^{(k)}$ be the LSE of μ based on $\{Y_{(j)}, X_{(j)} \neq k\}$, assuming that the mode of the regression function is at $X_{(k)}$ (so the least squares criterion is minimized subject to μ increasing on $[0, X_{(k)}]$ and decreasing to the right of $X_{(k)}$) and let $s_{n,k}$ be the corresponding minimum value of the least squares criterion. Then, a consistent estimate of the mode is given by $X_{(k^\star)}$, where $k^\star = \text{argmin}_{1 \leq k \leq n} s_{n,k}$. Our estimate here is similar to that proposed in Shoung and Zhang (2001). An alternative consistent estimate of the mode could be obtained in the same fashion as above, but replacing minimization of the least squares criterion by the maximization of the likelihood function. An estimator of this type, in the setting of a unimodal density is given in Bickel and Fan (1996). An analogous prescription applies to a U–shaped regression function.

3 Some simulation results

In this section we illustrate some of the distributional convergences stated in Theorem 1 and also report results from a limited simulation study.

Figure 3 shows the empirical distributions of the likelihood ratio statistic, the scaled residual sum of squares statistic ($RSS(\eta_0)/I(\psi(x_0))$) and the scaled L_2 statistic $L_2(\hat{\mu}_n, \hat{\mu}_n^0)/I(\psi(x_0))$ for the simulation setting described below, along with the empirical distributions of (fine discrete approximations to) the corresponding limit distributions.

5000 replicates from the distributions of these three statistics were generated from a simulation setting where $Z \sim \text{Unif}(-1, 1)$ and $X \mid Z = z$ is distributed as a Poisson with mean $\mu(z) \equiv 1 + z^2$ and $n = 2000$. Thus, the regression function is U–shaped and attains its minimum at 0. The (true) null hypothesis under which the three competing statistics are computed is $H_0 : \mu(.5) = 1.25$. Note that $\mu'(1.5) \neq 0$, so that $m = 1$. Thus, the common limit distribution of the likelihood ratio statistic and the scaled residual sum of squares statistic is given by $\mathbb{D}_1$ and the limit distribution of the scaled L_2 statistic is given by $\mathbb{T}_1$. We plotted the empirical distribution of the likelihood ratio statistic in red and that of the scaled residual sum of squares statistic in green. The empirical distribution of $\mathbb{D}_1$ was plotted in black. The green curve and the red curve coincide almost completely; since the green curve was plotted later, the red curve is virtually hidden except for a few minor streaks of red visible towards the tail. The empirical distribution of $\mathbb{D}_1$ was plotted in black and shows very good agreement with the red and green curves. The empirical distribution of the scaled L_2

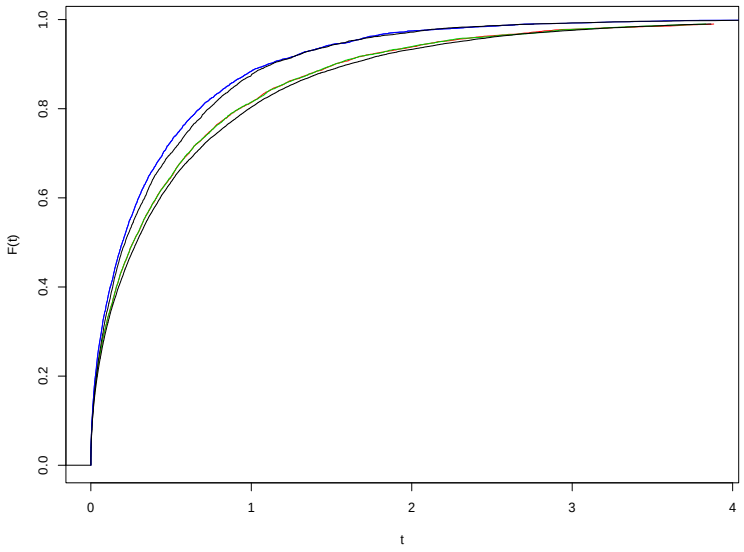

Figure 1: The empirical distributions of some asymptotic pivots and their limit distributions

statistic was plotted in blue and that of $\mathbb{T}_1$ in black. Once again, these curves are in good agreement.

The properties of confidence sets using the three competing pivots above were also studied. For this simulation study, we chose Z to follow Unif$(0,1)$ and $X \mid Z = z$ to follow Poisson$(\mu(z) \equiv 1 + z)$. Asymptotically 95% confidence intervals for $\mu(0.5)$ were obtained by using these different pivots (via inversion) for sample sizes $n = 50, 100, 200, 500, 1000, 1500$ and 1000 replicates for each sample size. Table 1 summarizes the behavior of these confidence intervals: for each sample size, the average length of these intervals and the observed coverage are reported for the likelihood ratio method, the RSS based method and the L_2 based method respectively. The observed coverage matches the nominal coverage extremely well, even for modest sample sizes. For each method, the average length of the confidence intervals (expectedly) diminishes with increasing sample size with the L_2 based confidence sets being the shortest on average. Table 2 shows selected quantiles of the distributions of $\mathbb{D}_1$ and $\mathbb{T}_1$ used in the construction of the confidence intervals referred to above.

4 Concluding discussion

In this paper, we have demonstrated how inference for a shape constrained regression function may be done in a broad class of regression models. The only structural constraint on the regression models lies in the assumption that the conditional distribution of the response given the covariate belongs to a regular exponential family, with the regression function being monotonic/unimodal/U–shaped in the covariate. The formulation is however shown to be fairly general in the sense that

Table 1: *Poisson Regression Model: average length (AL) and empirical coverage (C) of asymptotic 95% confidence intervals using likelihood ratio (PL), residual sum of squares (RSS) and L_2 based (L2) methods.*

	LRT		RSS		L2	
n	AL	C	AL	C	AL	C
50	0.852	0.952	0.846	0.955	0.802	0.930
100	0.714	0.953	0.715	0.955	0.672	0.953
200	0.569	0.943	0.569	0.939	0.535	0.950
500	0.424	0.956	0.425	0.958	0.396	0.947
1000	0.341	0.954	0.341	0.955	0.317	0.954
1500	0.294	0.952	0.294	0.954	0.275	0.947

Table 2: *Selected quantiles of* $\mathbb{D}_1$ *and* $\mathbb{T}_1$

p	$F^{-1}_{\mathbb{D}_1}(p)$	$F^{-1}_{\mathbb{T}_1}(p)$
.50	0.285	0.210
.75	0.803	0.588
.80	0.987	0.716
.85	1.230	0.887
.90	1.611	1.139
.95	2.287	1.591
.99	3.865	2.759

many well known regression models of interest, involving both discrete and continuous responses, can be captured in the framework. For unimodal (or U–shaped) functions, so long as inference is restricted to points away from the mode (or the minimizer), the techniques for estimating a monotone regression function can be conveniently adapted.

It should be noted that the monotone regression model $Y = \mu(X) + \epsilon$ with an additive error ϵ that is independent of X does not exactly fall in this set-up unless the assumption of Gaussianity (on the error) is made, and even under this assumption the error variance must be estimated for inference on μ to be carried out under the allowed shape constraints. However, the results of this paper continue to hold under fairly general error distributions. Consistent estimation of the error variance is readily accomplished and therefore poses no major challenges.

There are several natural directions in which the results of this paper can be extended. Firstly, the proposed estimation strategies do not work at a stationary point of the regression function (like the maximizer of a unimodal function, or the minimizer of a U–shaped function). The isotonic estimates computed in the first section are, in fact, not even consistent at these points. Construction of a confidence set for a unimodal regression function, say, at its modal value will require a penalization based likelihood criterion as in Woodroofe and Sun (1993) or in the more recent work of Pal (2006). Yet another problem is the construction of a confidence set for the mode itself, and to our knowledge, a nonparametric solution to this remains unavailable as yet. Secondly, this paper deals with a unidimensional covariate but from the perspective of applications one would like to incorporate auxiliary covariates into the model whose effect could be modelled parametrically. One way to address this issue is to postulate that the conditional mean of the response Y given covariates (X, W), where X is the primary one–dimensional covariate of interest and W is a vector of supplementary covariates is given by $B'(\beta^T W + \psi(X))$, β being a regression parameter and ψ being constrained by the usual shape restrictions. Models like the semilinear regression model or the logistic regression model can be readily seen to belong to this category. There are reasons to believe that confidence intervals for the regression function at a fixed covariate profile (w_0, x_0) should be constructible using asymptotic pivots of the type that have been described above, even though there are several technical (semiparametric) issues that would need to be resolved before such a general statement can be accurately made. Furthermore, it is not very difficult to see that in a semiparametric model of this type, the maximum likelihood estimates of ψ will no longer have an explicit representation as in the current scenario. Rather, self–induced characterizations will be needed, making the asymptotics significantly more difficult to handle. A full discussion of these issues is well beyond the scope of this paper and is left as a topic for future research.

Finally, an extensive and carefully designed simulation study of the proposed methods will need to be done before any judgment can be passed as to the relative optimality of any of these pivots in comparison to the others. Our limited simulations indicate that the $L2$ based method is marginally superior; however, there is not enough evidence, computational or theoretical, to come to a definitive conclusion as to whether this is indeed, in general, the case. On the computational

front, yet another goal would be to disseminate the ensuing methodology (from this paper) through freely available software packages like R–modules, which could be expected to promote their use substantially. At present, a beta version of the software is available as a Fortran program (on request) from the author.

5 Proof of Theorem 1

For the proof, we refer to $\hat{\mu}_n$ and $\hat{\mu}_n^0$ as $\hat{\mu}$ and $\hat{\mu}^0$ respectively, and to $\hat{\psi}_n$ and $\hat{\psi}_n^0$ as $\hat{\psi}$ and $\hat{\psi}^0$ respectively. We first establish (a). It is easy to see that:

$$\begin{aligned}
\mathrm{RSS}(\eta_0) &= \sum_{i\in J_n} [(T(Y_{(i)}) - \mu(x_0)) - (\hat{\mu}^0(X_{(i)}) - \mu(x_0))]^2 - \sum_{i\in J_n} [(T(Y_{(i)}) - \mu(x_0)) - (\hat{\mu}(X_{(i)}) - \mu(x_0))]^2 \\
&= \sum_{i\in J_n} (\hat{\mu}^0(X_{(i)}) - \mu(x_0))^2 - \sum_{i\in J_n} (\hat{\mu}(X_{(i)}) - \mu(x_0))^2 \\
&\quad -2 \sum_{i\in J_n} (T(Y_{(i)}) - \mu(x_0))(\hat{\mu}^0(X_{(i)}) - \mu(x_0)) + 2 \sum_{i\in J_n} (T(Y_{(i)}) - \mu(x_0))(\hat{\mu}(X_{(i)}) - \mu(x_0)) \\
&\equiv I_n - II_n - III_n + IV_n \,.
\end{aligned}$$

Consider the term III_n. Let $B_1^0, \ldots, B_L^0$ denote the ordered blocks of indices into which J_n decomposes, such that on each block $\hat{\mu}^0$ assumes the constant value c_j^0 and let B_l^0 denote that single block on which $\hat{\mu}^0$ assumes the constant value η_0. Then, for $j \neq l$, $c_j^0 = m_j^{-1} \sum_{i\in B_j^0} T(Y_{(i)})$, where m_j is the cardinality of B_j^0, by Fact 3. We have:

$$\begin{aligned}
III_n &= 2 \sum_{j=1}^{L} \sum_{i\in B_j^0} (T(Y_{(i)}) - \eta_0)(c_j^0 - \eta_0) \\
&= 2 \sum_{j=1}^{L} (c_j^0 - \eta_0) \sum_{i\in B_j^0} (T(Y_{(i)}) - \eta_0) \\
&= 2 \sum_{j=1}^{L} m_j \, (c_j^0 - \eta_0)^2 \\
&= 2 \sum_{j=1}^{L} \sum_{i\in B_j^0} (\hat{\mu}^0(X_{(i)}) - \eta_0)^2 \\
&= 2 \sum_{i\in J_n} (\hat{\mu}^0(X_{(i)}) - \mu(x_0))^2 \,.
\end{aligned}$$

Similarly, it follows that $IV_n = 2 \sum_{i\in J_n} (\hat{\mu}(X_{(i)}) - \mu(x_0))^2$. Consequently:

$$\mathrm{RSS}_n(\eta_0) = \sum_{i\in J_n} (\hat{\mu}(X_{(i)}) - \mu(x_0))^2 - \sum_{i\in J_n} (\hat{\mu}^0(X_{(i)}) - \mu(x_0))^2$$

$$
\begin{aligned}
&= n\,\mathbb{P}_n\left[(\hat{\mu}(x)-\mu(x_0))^2\,1(x\in D_n)\right] - n\,\mathbb{P}_n\left[(\hat{\mu}^0(x)-\mu(x_0))^2\,1(x\in D_n)\right] \\
&= n^{1/(2m+1)}\,(\mathbb{P}_n - P)\left[(X_n^2(h)-Y_n^2(h))\,1(h\in\tilde{D}_n)\right] \\
&\qquad + n^{1/(2m+1)}\,P[(X_n^2(h)-Y_n^2(h))\,1(h\in\tilde{D}_n)]\,, \qquad (11)
\end{aligned}
$$

where $\tilde{D}_n = n^{1/(2m+1)}(x-x_0)$, $h = n^{1/(2m+1)}(x-x_0)$ and X_n and Y_n are as defined at the beginning of Section 2. By Fact 2, $\tilde{D}_n$ is eventually contained in a compact set with arbirtarily high (pre-assigned) probability. Also, the monotone (in h) processes X_n and Y_n are eventually bounded on any compact set with arbitrarily high probability, by Fact 1. Using preservation properties for Donsker classes of functions, it is readily concluded that the class of functions $\{h \mapsto (X_n^2(h)-Y_n^2(h))\,1(h\in\tilde{D}_n)\}$ is eventually contained in a P–Donsker class of functions with arbitrarily high (pre-assigned) probability. It follows that $n^{1/(2m+1)}\,(\mathbb{P}_n - P)\left[(X_n^2(h)-Y_n^2(h))\,1(h\in\tilde{D}_n)\right]$ is $o_p(1)$ and we can write:

$$
\begin{aligned}
\mathrm{RSS}(\eta_0) &= n^{1/(2m+1)}\,P\left[(X_n^2(h)-Y_n^2(h))\,1(h\in\tilde{D}_n)\right] + o_p(1) \\
&= \int_{\tilde{D}_n} (X_n^2(h)-Y_n^2(h))\,p_X(x_0 + h\,n^{-1/(2m+1)})\,dh \\
&= p_X(x_0)\int_{\tilde{D}_n} (X_n^2(h)-Y_n^2(h))\,dh + o_p(1) \\
&\to_d\ p_X(x_0)\int \left[(g_{a,b,m}(h))^2 - (g^0_{a,b,m}(h))^2\right]\,dh\,,
\end{aligned}
$$

where the last step is a consequence of Fact 1. By Lemma 2, it follows that:

$$
\mathrm{RSS}(\eta_0) \to_d a^2\,p_X(x_0)\,\mathbb{D}_m\,;
$$

but $a^2\,p_X(x_0) = I(\psi(x_0))$ and this finishes the proof.

Note: In going from the first to the second line of the display preceding the last, the factor of $n^{1/(2m+1)}$ *does disappear*. This is due to the fact that if X has density $p_X(x)$, the random variable $R \equiv n^{1/(2m+1)}(X-x_0)$ has density $p_R(r) = p_X(x_0 + r\,n^{-1/(2m+1)})\,n^{-1/(2m+1)}$.

We now establish the limit distribution of the likelihood ratio statistic $2\log\lambda_n(\theta_0)$. This will be done by establishing a simple representation for the likelihood ratio statistic in terms of the slope processes $\tilde{X}_n$ and $\tilde{Y}_n$. We write $2\log\lambda_n(\theta_0) = I_n - II_n$ where

$$
I_n = 2\left(\sum_{i\in J_n} (\hat{\psi}(X_{(i)})T(Y_{(i)}) - \hat{\psi}^0(X_{(i)})T(Y_{(i)})\right)
$$

and $II_n = 2\sum_{i\in J_n}(B(\hat{\psi}(X_{(i)})) - B(\hat{\psi}^0(X_{(i)})))$. On Taylor expansion of $B(\hat{\psi}(X_{(i)}))$ and $B(\hat{\psi}^0(X_{(i)}))$ around $\psi(x_0)$, up to the third order, we get:

$$
II_n = 2\sum_{i\in J_n} B'(\theta_0)\,\left\{(\hat{\psi}(X_{(i)})-\theta_0) - (\hat{\psi}^0(X_{(i)})-\theta_0)\right\}
$$

$$+\sum_{i\in J_n} B''(\theta_0)\,\left\{(\hat{\psi}(X_{(i)})-\theta_0)^2-(\hat{\psi}^0(X_{(i)})-\theta_0)^2\right\}+o_p(1)\,.$$

Denote the second term on the right side of the above display by Q_n. Combining I_n and II_n we obtain:

$$2\,\log\,\lambda_n(\theta_0)=2\,\left[\sum_{i\in J_n}((\hat{\psi}(X_{(i)})-\theta_0)-(\hat{\psi}^0(X_{(i)})-\theta_0))\,(T(Y_{(i)})-B'(\theta_0))\right]-Q_n\,.$$

The first term on the right side of the above display is equal to:

$$2\sum_{i\in J_n}(\hat{\psi}(X_{(i)})-\theta_0)\,(B'(\hat{\psi}(X_{(i)}))-B'(\theta_0))-2\sum_{i\in J_n}(\hat{\psi}^0(X_{(i)})-\theta_0)\,(B'(\hat{\psi}^0(X_{(i)}))-B'(\theta_0))\,. \quad (12)$$

The above is a direct consequence of Fact 3, and can be established by looking at the blocks of indices on which the unconstrained and constrained MLE's $\hat{\psi}$ and $\hat{\psi}^0$ are constant. Note that these are precisely the same blocks on which the unconstrained and constrained least squares estimates are constant. For example, if B_j is a block of indices contained in J_n on which $\hat{\psi}$ attains the constant value a_0 (and $\hat{\mu}$ attains the constant value $B'(a_0)$), then it is readily checked that

$$\sum_{i\in B_j}(\hat{\psi}(X_{(i)})-\theta_0)\,(T(Y_{(i)})-B'(\theta_0))=\sum_{i\in J_n}(\hat{\psi}(X_{(i)})-\theta_0)\,(B'(\hat{\psi}(X_{(i)})-B'(\theta_0))$$

with the common value being given by $m_j\,(a_0-\theta_0)(B'(a_0)-B'(\theta_0))$, m_j being the cardinality of B_j. A similar argument applies for the constrained MLE $\hat{\psi}^0$. Next, by Taylor expansion of $B'(\hat{\psi}(X_{(i)}))$ and $B'(\hat{\psi}^0(X_{(i)}))$ around $\psi(x_0)$ (up to the second order), (12) can be simplified to:

$$2\,\left[\sum_{i\in J_n}B''(\theta_0)\,((\hat{\psi}(X_{(i)})-\theta_0)^2-(\hat{\psi}^0(X_{(i)})-\theta_0)^2)\right]+o_p(1)\equiv 2\,Q_n+o_p(1)\,.$$

It follows that $2\,\log\,\lambda_n=Q_n+o_p(1)$. Recalling that $B''(\theta_0)=I(\psi(x_0))$, we have:

$$\begin{aligned}2\,\log\,\lambda_n &= I(\psi(x_0))\,n\,\mathbb{P}_n\,\left[\left\{(\hat{\psi}(x)-\theta_0)^2-(\hat{\psi}^0(x)-\theta_0)^2\right\}\,1(x\in D_n)\right]+o_p(1)\\ &= I(\psi(x_0))\,n^{1/(2m+1)}\,(\mathbb{P}_n-P)\,[(\tilde{X}_n^2(h)-\tilde{Y}_n^2(h))\,1(h\in\tilde{D}_n)]\\ &\qquad +I(\psi(x_0))\,n^{1/(2m+1)}\,P[(\tilde{X}_n^2(h)-\tilde{Y}_n^2(h))\,1(h\in\tilde{D}_n)]+o_p(1)\,. \quad (13)\end{aligned}$$

Apart from the constant $I(\psi(x_0))$ and the $o_p(1)$ term, the above display has the exact same form as the representation for $\mathrm{RSS}(\eta_0)$ in (11), but with X_n and Y_n replaced by $\tilde{X}_n$ and $\tilde{Y}_n$ respectively. Following (almost) identical steps to those for $\mathrm{RSS}(\eta_0)$ we conclude that:

$$2\,\log\,\lambda_n(\theta_0)\to_d I(\psi(x_0))\,p_X(x_0)\,\tilde{a}^2\,\mathbb{D}_m=\mathbb{D}_m\,.$$

This finishes the proof of Part (a).

We next establish Part (b). We only establish the asymptotics for $L_{2,w}(\hat{\mu}, \hat{\mu}^0)$. The remaining three statistics can be handled by similar arguments. We have:

$$\begin{aligned} L_{2,w}(\hat{\mu}, \hat{\mu}^0) &= \sum_{i=1}^{n} \frac{(\hat{\mu}(X_i) - \hat{\mu}^0(X_i))^2}{I(\hat{\psi}(X_i))} \\ &= \sum_{i=1}^{n} \frac{(\hat{\mu}(X_i) - \hat{\mu}^0(X_i))^2}{I(\psi(x_0))} - \sum_{i=1}^{n} \frac{(\hat{\mu}(X_i) - \hat{\mu}^0(X_i))^2 \, (I(\hat{\psi}(X_i)) - I(\psi(x_0)))}{I(\hat{\psi}(X_i)) \, I(\psi(x_0))} \\ &\equiv A_n - B_n \, . \end{aligned}$$

Next,

$$\begin{aligned} A_n &= \frac{1}{I(\psi(x_0))} \, n \, \mathbb{P}_n \left[((\hat{\mu}(x) - \mu(x_0)) - (\hat{\mu}^0(x) - \mu(x_0)))^2 \, 1(x \in D_n) \right] \\ &= \frac{1}{I(\psi(x_0))} \, n^{1/(2m+1)} \, (\mathbb{P}_n - P) \left[(X_n(h) - Y_n(h))^2 \, 1(h \in \tilde{D}_n) \right] \\ &\qquad + \frac{1}{I(\psi(x_0))} \, n^{1/(2m+1)} \, P[(X_n(h) - Y_n(h))^2 \, 1(h \in \tilde{D}_n)] \\ &\equiv A_{1,n} + A_{2,n} \, . \end{aligned}$$

As in Part (a), $A_{1,n}$ is $o_p(1)$ and the contribution in the limit comes from $A_{2,n}$ alone. As in Part (a), we get:

$$\begin{aligned} A_{2,n} &= \frac{1}{I(\psi(x_0))} \int_{\tilde{D}_n} (X_n(h) - Y_n(h))^2 \, p_X(x_0 + h \, n^{-1/(2m+1)}) \, dh \\ &= \frac{1}{I(\psi(x_0))} \, p_X(x_0) \int_{\tilde{D}_n} (X_n(h) - Y_n(h))^2 \, dh + o_p(1) \\ &\rightarrow_d \frac{p_X(x_0)}{I(\psi(x_0))} \int \left[(g_{a,b,m}(h) - g^0_{a,b,m}(h))^2 \right] \, dh \\ &\equiv_d \frac{p_X(x_0)}{I(\psi(x_0))} \, a^2 \, \mathbb{T}_m \\ &= \mathbb{T}_m \end{aligned}$$

where the convergence in distribution is deduced as in Part (a). The equivalence in distribution is a direct consequence of Lemma 2. It only remains to show that B_n converges in probability to 0. We write:

$$B_n \leq \mathbb{P}_n \left[\frac{n^{1/(2m+1)} \, (X_n(h) - Y_n(h))^2 \mid I(\hat{\psi}(x_0 + hn^{-1/(2m+1)})) - I(\psi(x_0)) \mid 1(h \in \tilde{D}_n)}{I(\hat{\psi}(x_0 + hn^{-1/(2m+1)})) \, I(\psi(x_0))} \right] .$$

It is not very difficult to see that we can eventually bound the expression on the right side of the above display, with probability larger than any pre–specified amount by a constant times

$$\mathbb{P}_n \left[n^{-(m-1)/(2m+1)} \, (X_n(h) - Y_n(h))^2 \mid \tilde{X}_n(h) \mid 1(h \in [-K, K]) \right] ,$$

for some large $K > 0$. For $m > 1$ this is $o_p(1)$ by virtue of the facts that the random functions X_n, Y_n and $\tilde{X}_n$ are $O_p(1)$ for $h \in [-K, K]$ and that $n^{-(m-1)/(2m+1)}$ converges to 0. For $m = 1$, this last term is identically 1 and the argument proceeds as follows. We decompose the above display as:

$$(\mathbb{P}_n - P)\left((X_n(h) - Y_n(h))^2 \mid \tilde{X}_n(h) \mid 1(h \in [-K, K])\right) \\ + P\left((X_n(h) - Y_n(h))^2 \mid \tilde{X}_n(h) \mid 1(h \in [-K, K])\right).$$

The first term in the above display goes to 0 in probability yet again using standard arguments from empirical process theory. Also,by direct computation, one finds that the second term in the display is $O_p(n^{-1/(2m+1)})$, and hence $o_p(1)$. Hence B_n does not contribute asymptotically to the limit.

We next turn to the proof of Part (c). Consider the score statistic $S_{n,1}$. We have:

$$\begin{aligned} S_{n,1} &= n\,\mathbb{P}_n\left[(\hat{\psi}^0(x) - \hat{\psi}(x))\,(T(y) - B'(\hat{\psi}(x))\right] \\ &= n\,\mathbb{P}_n\left[(\hat{\psi}^0(x) - \psi(x_0))\{T(y) - B'(\psi(x_0)) - (B'(\hat{\psi}(x)) - B'(\psi(x_0)))\}1(x \in D_n)\right] \\ &\qquad - n\,\mathbb{P}_n\left[(\hat{\psi}(x) - \psi(x_0))\{T(y) - B'(\psi(x_0)) - (B'(\hat{\psi}(x)) - B'(\psi(x_0)))\}1(x \in D_n)\right] \\ &= n\,\mathbb{P}_n\left[(\hat{\psi}^0(x) - \psi(x_0))\{T(y) - B'(\psi(x_0)) - (B'(\hat{\psi}(x) - B'(\psi(x_0)))\}1(x \in D_n)\right], \end{aligned}$$

the second term vanishing as a consequence of Fact 3 – the argument is similar to that involved in establishing the representation (12) and is not reiterated. Hence,

$$\begin{aligned} S_{n,1} &= n\,\mathbb{P}_n[(\hat{\psi}^0(x) - \psi(x_0))\,\{(\hat{\mu}^0(x) - \mu(x_0)) - (\hat{\mu}(x) - \mu(x_0))\}\,1(x \in D_n)] \\ &= n^{1/(2m+1)}\,P[n^{m/(2m+1)}(\hat{\psi}^0(x) - \psi(x_0))\{n^{m/(2m+1)}(\hat{\mu}^0(x) - \mu(x_0)) \\ &\qquad - n^{m/(2m+1)}(\hat{\mu}(x) - \mu(x_0))\}\,1(x \in D_n)] + o_p(1) \\ &= o_p(1) + \int_{\tilde{D}_n} \tilde{Y}_n(h)\,(Y_n(h) - X_n(h))\,p_X(x_0 + h\,n^{-1/(2m+1)})\,dh \end{aligned}$$

by (at this point) standard arguments. Since the length of $\tilde{D}_n$ is $O_p(1)$ and for $h \in [-K, K]$, $(X_n(h), Y_n(h)) = B''(\psi(x_0))\,(\tilde{X}_n(h), \tilde{Y}_n(h)) + o_p(1)$ (by the Delta method), we can write:

$$S_{n,1} = o_p(1) + I(\psi(x_0))p_X(x_0) \int_{\tilde{D}_n} (\tilde{Y}_n^2(h) - \tilde{Y}_n(h)\,\tilde{X}_n(h))\,dh\,. \tag{14}$$

Using Fact 1, and the fact that $\tilde{D}_n$ is eventually in a compact set with arbitrarily high (preassigned) probability, we deduce that:

$$\begin{aligned} S_{n,1} &\to_d I(\psi(x_0))p_X(x_0) \int g^0_{\tilde{a},\tilde{b},m}(h)\,(g^0_{\tilde{a},\tilde{b},m}(h) - g_{\tilde{a},\tilde{b},m}(h)) \\ &\equiv_d \tilde{a}^2\,I(\psi(x_0))p_X(x_0)\,\mathbb{M}_{1,m}\,, \end{aligned}$$

by Lemma 2. But $\tilde{a}^2 = (I(\psi(x_0))p_X(x_0))^{-1}$ which finishes the proof.

We can similarly show that $S_{n,2}$ admits the representation:

$$S_{n,2} = o_p(1) + I(\psi(x_0))p_X(x_0) \int_{\tilde{D}_n} (\tilde{X}_n^2(h) - \tilde{Y}_n(h)\,\tilde{X}_n(h))\,dh\,, \tag{15}$$

and the fact that $S_{n,2}$ converges in distribution to $\mathbb{M}_{2,m}$ follows from this fairly directly. Next, using the representations (14) and (15)), we see that

$$S_{n,2} - S_{n,1} = I(\psi(x_0))p_X(x_0) \int_{\tilde{D}_n} (\tilde{X}_n^2(h) - Y_n^2(h))\,dh + o_p(1)\,.$$

Now consider the representation (13) for $2\log\lambda_n$. The first term there is $o_p(1)$ and writing the second term there as an integral, it is not difficult to see that the first term on the right side of the above display is asymptotically equivalent to that second term. It follows that $S_{n,2} - S_{n,1} = 2\log\lambda_n + o_p(1)$. The proof of the fact that $S_{n,1} + S_{n,2} = I(\psi(x_0))\,L_2(\hat{\psi}, \hat{\psi}^0) + o_p(1)$ follows along similar lines and is omitted. This finishes the proof of Part (c).

The proof of Part (d) is a direct consequence of Fact 1; look at the limit distribution of $X_n(0)$ and use the Brownian scaling relations from Lemma 1 to arrive at the result.

6 Appendix

Proof of Lemma 1: We first relate $X_{c,d,m}(h) = cW(h) + d|h|^{m+1}$ to the canonical process $X_m(h) = W(h) + |h|^{m+1}$ using Brownian scaling. The goal is to find constants k_1, k_2 such that

$$k_1 X_{c,d,m}(k_2 h) \equiv_D X_m(h)$$

as processes in the space of locally bounded functions on the line equipped with the topology of uniform convergence on compacta. For any $\alpha > 0$, $W(h) \equiv_d \alpha^{-1/2}\,W(\alpha t)$, so

$$k_1 X_{c,d,m}(k_2 h) \equiv_d k_1 \mid k_2 \mid^{m+1} d \mid h \mid^{m+1} + k_1 c\alpha^{-1/2} W(k_2\alpha h)\,.$$

To satisfy the first display, it suffices to choose the constants k_1 and k_2 such that

$$k_1 k_2^{m+1} d = 1\,, k_1 c\alpha^{-1/2} = 1 \text{ and } k_2\alpha = 1\,.$$

Eliminating α from the second and third equations we obtain that $k_2 = k_1^{-2}c^{-2}$. Using this in conjunction with the first equation, we readily solve for k_1 and k_2 to get:

$$k_1 = d^{1/(2m+1)} c^{-2(m+1)/(2m+1)} \text{ and } k_2 = c^{2/(2m+1)} d^{-2/(2m+1)}\,.$$

Conclude that:

$$X_{c,d,m}(h) \equiv_d c^{2(m+1)/(2m+1)} d^{-1/(2m+1)}\, X_m(c^{-2/(2m+1)} d^{2/(2m+1)} h)\,.$$

Hence,

$$(\text{slogcm}X_{c,d,m}(h), \text{slogcm}^0 X_{c,d,m}(h)) \equiv_d c^{2(m+1)/(2m+1)} d^{-1/(2m+1)}\, c^{-2/(2m+1)} d^{2/(2m+1)}$$
$$\times \left\{ \text{slogcm}X_m(c^{-2/(2m+1)} d^{2/(2m+1)} h), \text{slogcm}^0 X_m(c^{-2/(2m+1)} d^{2/(2m+1)} h) \right\} ,$$

since we are computing slopes of convex minorants with respect to h. On simplifying, we obtain:

$$(g_{c,d,m}(h), g^0_{c,d,m}(h)) \equiv_d (c^{2m} d)^{1/(2m+1)} \, (g_m((d/c)^{2/(2m+1)}), g^0_m((d/c)^{2/(2m+1)})) \, ,$$

the equality holding in the space $\mathcal{L} \times \mathcal{L}$. □

Proof of Lemma 2: We only show that $U \equiv c^{-2}\, \mathbb{D}_{c,m} \equiv_d \mathbb{D}_m = V$. The remaining marginal distributional identities can be shown similarly and the method demonstrated below can be extended to establish the joint distributional identity claimed in the lemma. Since the joint identity is of little statistical consequence we refrain from proving it. We will need the following lemma due to Prakasa Rao (1969).

Lemma 3 *Suppose that* $\{\mathcal{X}_{n\epsilon}\}, \{\mathcal{Y}_n\}$ *and* $\{W_\epsilon\}$ *are three sets of random variables such that*

(*i*) $\lim_{\epsilon \to 0} \limsup_{n \to \infty} P[\mathcal{X}_{n\epsilon} \neq \mathcal{Y}_n] = 0 \, ,$

(*ii*) $\lim_{\epsilon \to 0} P[W_\epsilon \neq \mathcal{Y}] = 0 \, ,$

(iii) For every $\epsilon > 0$, $\mathcal{X}_{n\epsilon} \to_d W_\epsilon$ *as* $n \to \infty$.

Then $\mathcal{Y}_n \to_d \mathcal{Y}$, *as* $n \to \infty$.

Denote by $D_{c,d}$ the set where the processes $g_{c,d,m}$ and $g^0_{c,d,m}$ differ, and by $D_{1,1}$ the set where the processes g_m and g^0_m differ. Now, for each $\epsilon > 0$, we can find $K_\epsilon > 0$ such that

$$P \left[\, D_{c,d} \subset [-K_\epsilon, K_\epsilon] \, \right] > 1 - \epsilon$$

and

$$P \left[\, D_{1,1} \subset [-(d/c)^{2/(2m+1)} \, K_\epsilon, (d/c)^{2/(2m+1)} \, K_\epsilon] \right] > 1 - \epsilon \, .$$

Thus

$$\begin{aligned} U_\epsilon &= \int_{[-K_\epsilon, K_\epsilon]} c^{-2} \, \left((g_{c,d,m}(z))^2 - (g^0_{c,d,m}(z))^2 \right) dz \\ &\equiv_d \int_{[-K_\epsilon, K_\epsilon]} c^{-2} \, (c^2 \, (d/c)^{2/(2m+1)}) \, \left((g_m((d/c)^{2/(2m+1)} h))^2 - (g^0_m((d/c)^{2/(2m+1)} h))^2 \right) dh \\ &= \int_{[-(d/c)^{2/(2m+1)} K_\epsilon, (d/c)^{2/(2m+1)} K_\epsilon]} \left((g_m(h))^2 - (g^0_m(h))^2 \right) dh \\ &= V_\epsilon \, , \end{aligned}$$

where in passing from the first line of the above display to the second, we have invoked Lemma 1. Now, note that

$$P\left[U_\epsilon \neq U\right] < \epsilon \text{ and } P\left[V_\epsilon \neq V\right] < \epsilon\,.$$

Therefore by Lemma 3 (set $\mathcal{X}_{n\epsilon} = U_\epsilon$ for all n, set $\mathcal{Y}_n = U$ for all n, set $W_\epsilon = V_\epsilon$, and set $\mathcal{Y} = V$), it follows that $U \equiv_d V$. $\square$

Proof of Fact 1: The proof of Fact 1 relies on extensive use of "switching relationships" which allow us to translate the behavior of the slope of the convex minorant of a random cumulative sum diagram (this is how the estimators $\hat{\mu}_n$ and $\hat{\mu}_n^0$ are characterized) in terms of the minimizer of a stochastic process. The limiting behavior of the slope process can then be studied in terms of the limiting behavior of the minimizer of this stochastic process by applying argmin continuous mapping theorems. Switching relationships on the limit process allow interpretation of the behavior of the minimizer of the limit process in terms of the slope of the convex minorant of the limiting versions of the cumulative sum diagrams (appropriately normalized).

The first step is to establish finite–dimensional convergence of the processes $(X_n(h), Y_n(h))$ to $(g_{a,b,m}(h), g^0_{a,b,m}(h))$. Thus, it is shown that for any $(h_1, h_2, \ldots, h_k)$, the random vector

$$\left(\{X_n(h_i)\}_{i=1}^k, \{Y_n(h_i)\}_{i=1}^k\right) \rightarrow_d \left(\{g_{a,b,m}(h_i)\}_{i=1}^k, \{g^0_{a,b,m}(h_i)\}_{i=1}^k\right)\,,$$

in the space $\mathbb{R}^{2k}$. Next, to deduce the convergence in $\mathcal{L} \times \mathcal{L}$ note firstly that $X_n(h)$ and $Y_n(h)$ are monotone functions. Now, given a sequence (ψ_n, ϕ_n) in $\mathcal{L} \times \mathcal{L}$ such that ψ_n and ϕ_n are monotone functions and (ψ_n, ϕ_n) converges pointwise to (ψ, ϕ) (where (ψ, ϕ) is in $\mathcal{L} \times \mathcal{L}$), we can conclude that $(\psi_n, \phi_n) \rightarrow (\psi, \phi)$ in $\mathcal{L} \times \mathcal{L}$. It follows, in the wake of distributional convergence of all the finite - dimensional marginals of (X_n, Y_n) to those of $(g_{a,b,m}(h), g^0_{a,b,m}(h))$, that

$$(X_n(h), Y_n(h)) \rightarrow_d (g_{a,b,m}(h), g^0_{a,b,m}(h))$$

in $\mathcal{L} \times \mathcal{L}$ (see the result of Corollary 2 following Theorem 3 of Huang and Zhang (1994)).

In the remainder of this proof we will sketch the proof of convergence of $X_n(h)$ to $g_{a,b,m}(h)$ for any h. The general proof of finite–dimensional convergence is cumbersome to write out and contains minor extensions of the ideas expounded here. We now use "the switching relationship" for the unconstrained MLE $\hat{\mu}_n$ to get:

$$\hat{\mu}_n(x) \leq a \Leftrightarrow \text{argmin}_x \left[V_n(x) - a\, G_n(x)\right] \geq X_x \tag{16}$$

where X_x is the largest X value not exceeding x. By argmin we denote the largest element in the set of minimizers. This can be chosen to be one of the X_i's. The above equivalence is a direct characterization of the fact that the vector $\{\hat{\mu}_n(X_{(i)})\}_{i=1}^n$ is the vector of slopes (left–derivatives) of the cumulative sum diagram formed by the points $\{G_n(X_{(i)}), V_n(X_{(i)})\}_{i=0}^n$, computed at the points $\{G_n(X_{(i)})\}_{i=1}^n$. The easiest way to verify this is by drawing a picture.

Now, $X_n(h_0) = n^{m/(2m+1)}\,(\hat{\mu}_n(x_0 + h_0\,n^{-1/(2m+1)}) - \mu(x_0))$. We want to find

$$\lim_{n\to\infty} P\left(n^{m/(2m+1)}\,(\hat{\mu}_n(x_0 + h_0\,n^{-1/(2m+1)}) - \mu(x_0)) \le w\right).$$

Now, define

$$A_n = \left\{n^{m/(2m+1)}\,(\hat{\mu}_n(x_0 + h_0\,n^{1/(2m+1)}) - \mu(x_0)) \le w\right\}.$$

Consider the event A_n. We have

$$\begin{aligned}
n^{m/(2m+1)}\,(\hat{\mu}_n(x_0 + h_0\,n^{1/(2m+1)}) - \mu(x_0)) \le w \quad &\Leftrightarrow \quad \hat{\mu}_n(x_0 + h_0\,n^{1/(2m+1)}) \le \mu(x_0) + w\,n^{-m/(2m+1)} \\
&\Leftrightarrow \quad \text{argmin}_r \left[V_n(r) - (\mu(x_0) + w\,n^{-m/(2m+1)})\,G_n(r)\right] \\
&\qquad\qquad \ge X_{(x_0+h_0\,n^{1/(2m+1)})} \\
&\Leftrightarrow \quad \text{argmin}_r \left[\tilde{V}_n(r) - w\,\tilde{G}_n(r)\right] \ge X_{(x_0+h_0\,n^{1/(2m+1)})},
\end{aligned}$$

where

$$\tilde{V}_n(r) = n^{(m+1)/(2m+1)}\left\{(V_n(r) - V_n(x_0)) - \mu(x_0)(G_n(r) - G_n(x_0))\right\}$$

and $\tilde{G}_n(r) = n^{1/(2m+1)}\,(G_n(r) - G_n(x_0))$. The second bidirectional implication in the above display follows from the first on using (16), and the third follows from the second upon centering the relevant processes around their values at x_0 and multiplying by a factor of $n^{(m+1)/(2m+1)}$. These operations do not disturb the argmin. Now, introduce the local variable $h = n^{1/(2m+1)}(r - x_0)$. Using the fact that $n^{1/(2m+1)}(X_{(x_0+h_0\,n^{1/(2m+1)})} - x_0)$ is $h_0 + o_p(1)$ we get:

$$A_n = \text{argmin}_h \left[\tilde{V}_n(x_0 + h\,n^{-1/(2m+1)}) - w\,\tilde{G}_n(x_0 + h\,n^{-1/(2m+1)})\right] \ge h_0 + o_p(1).$$

Set $\mathbb{M}_n(h) = \tilde{V}_n(x_0 + h\,n^{-1/(2m+1)})$ and $\mathbb{G}_n(h) = \tilde{G}_n(x_0 + h\,n^{-1/(2m+1)})$. Now,

$$\mathbb{G}_n(h) = h\,\frac{G_n(x_0 + h\,n^{-1/(2m+1)}) - G_n(x_0)}{n^{-1/(2m+1)}h}$$

converges uniformly, in probability, on compact subsets of the real line to $p_X(x_0)h$. The process $\mathbb{M}_n(h)$ can be written as:

$$\mathbb{M}_n(h) = \sqrt{n}\,(\mathbb{P}_n - P)\,f_{n,h}(y,x) + n^{(m+1)/(2m+1)}P[(T(y) - \mu(x_0))(1(x \le x_0 + hn^{-1/(2m+1)}) - 1(x \le x_0))]$$

with

$$f_{n,h} = n^{1/(4m+2)}(T(y) - \mu(x_0))(1(x \le x_0 + hn^{-1/(2m+1)}) - 1(x \le x_0)).$$

By tedious but straightforward computations, the first term in the above decomposition of $\mathbb{M}_n(h)$ converges in distribution under the topology of uniform convergence on compact sets to $cW(h)$ where $W(h)$ is standard Brownian motion and $c^2 = B''(\psi(x_0))p_X(x_0) = I(\psi(x_0))p_X(x_0)$ (see, for example, Theorem 2.11.22 of van der Vaart and Wellner (1996) which provides sufficient conditions for establishing distributional convergences of the above type). The second term in the above

decomposition is nonstochastic and can be handled by writing it as an integral with respect to distribution of X and then invoking the Taylor expansion about the point x_0. Using the fact that the first $m-1$ derivatives of the function μ vanish at the point x_0, this term can be shown to converge uniformly for h in compacta to the function $d \mid h \mid^{m+1}$ where $d = \mid \mu^{(m)}(x_0) \mid p_X(x_0)/(m+1)!$. Thus, the process

$$S_n(h) \equiv \mathbb{M}_n(h) - w\mathbb{G}_n(h) \rightarrow_d L(h) \equiv c\,W(h) + d \mid h \mid^{m+1} - wp_X(x_0)h\,.$$

The limiting process $L(h)$ possesses a unique minimizer almost surely. Also the sequence $\operatorname{argmin}_h S_n(h)$ is tight. It follows by the argmin continuous mapping theorem (see Kim and Pollard (1990) or van der Vaart and Wellner (1996)) that $\operatorname{argmin}_h S_n(h) \rightarrow_d \operatorname{argmin}_h L(h)$. Consequently, $P(A_n) = P(\operatorname{argmin}_h S_n(h) \geq h_0 + o_p(1))$ converges to $P(\operatorname{argmin} L(h) \geq h_0)$. Next, we use "switching" on the limit process. On a set of probability 1, $\operatorname{argmin} L(h) \geq h_0$ is equivalent to $g_{c,d,m}(h_0) \leq wp_X(x_0)$, so that the limiting probability can be written as $P(p_X(x_0)^{-1}\, g_{c,d,m}(h_0) \leq w)$. But this is equal to $P(g_{c/p_X(x_0),d/p_X(x_0),m}(h_0) \leq w)$, which is what we sought to prove.

Here, we have glossed over the proof of the tightness of the sequence of minimizers $\operatorname{argmin}_h S_n(h)$. This may be achieved through a direct application of the rate theorem Theorem 3.2.5 of van der Vaart and Wellner (1996) with an appropriate choice of the distance function $(d(\theta, \theta_0) = \mid \theta - \theta_0 \mid^{(m+1)/2})$. We skip the details. □

Proof of Fact 2: First, note that D_n is either the null set, or it is an interval containing the point x_0. Second, note that if $\hat{\mu}_n(x) \neq \hat{\mu}_n^0(x)$, then either $\hat{\mu}_n(x) = \hat{\mu}_n(x_0)$ or $\hat{\mu}_n^0(x_0) = \eta_0$.

Let $\tilde{D}_n$ denote the set $n^{1/(2m+1)}\,(D_n - x_0)$. It suffices to show that given $\epsilon > 0$, there exists $M > 0$, such that $P(\tilde{D}_n \subset [-M, M]) > 1 - \epsilon$ eventually.

To prove this, proceed in the following way. Let $\tilde{D}_n = [\tilde{A}_n, \tilde{B}_n)$. Now the event $\{\tilde{D}_n \subset [-M, M]\}$ is the same as $\{-M < \tilde{A}_n \leq \tilde{B}_n < M\}$. It suffices to show that $P(\tilde{B}_n < M) > 1 - \epsilon/2$ and $P(-M < \tilde{A}_n) > 1 - \epsilon/2$ eventually for M sufficiently large. We shall prove the first assertion; the second follows similarly. To prove the first assertion it suffices to show that $P(\tilde{B}_n > M) < \epsilon/2$ for M sufficiently large. But $\tilde{B}_n > M$ means that $x_0 + M\, n^{-1/(2m+1)}$ is in the difference set D_n and this implies that either $\hat{\mu}_n^0(x_0 + M\, n^{-1/(2m+1)}) = \eta_0$ or $\hat{\mu}_n(x_0 + M\, n^{-1/(2m+1)}) = \hat{\mu}_n(x_0)$ by the second observation at the start of the proof. This can be written in terms of the processes X_n and Y_n in the following way:

$$\{\tilde{B}_n > M\} \quad \subset \quad \{Y_n(M) = 0\} \cup \{X_n(0) = X_n(M)\}\,. \tag{17}$$

Now

$$P(Y_n(M) = 0) \rightarrow P(g_{a,b,m,R}(M) \leq 0)$$

where $g_{a,b,m,R}(M)$ is the left derivative of the greatest convex minorant of the process $aW(h) + b \mid h \mid^2$ on $[0, \infty)$. By choosing M large enough we can ensure that the probability on the right of

the above display is strictly less than $\epsilon/4$. Consider now $P(X_n(0) = X_n(M))$. This is the same as $P(X_n(0) - X_n(M) = 0)$. Now by Fact 1 again, we have

$$(X_n(0), X_n(M)) \rightarrow_d (g_{a,b,m}(0), g_{a,b,m}(M))$$

so that

$$\text{limsup}\, P\left(X_n(0) - X_n(M) = 0\right) \leq P\left(g_{a,b,m}(0) - g_{a,b,m}(M) = 0\right)$$

and the right hand side is again less than $\epsilon/4$ for M sufficiently large. It now follows from (17) that

$$P(\tilde{B}_n > M) < \epsilon/2$$

eventually. □

Acknowledgements: I would like to thank Jayanta Pal for some very helpful discussion.

7 References

Banerjee, M. and Wellner, J. A (2001) Likelihood ratio tests for monotone functions. *Ann. Statist.* **29**, 1699–1731.

Bickel, P. and Fan, J. (1996) Some problems on the estimation of unimodal densities. *Statist. Sinica* **6**, 23 – 45.

Brunk, H.D. (1970). Estimation of isotonic regression. *Nonparametric Techniques in Statistical Inference.*,M.L. Puri, ed.

Chernoff, H. (1964). Estimation of the mode. *Ann. Statist. Math* **16**, 31–41.

Diggle, P., Morris, S. and Morton–Jones, T. (1999) Case–control isotonic regression for investigation of elevation in risk around a risk source. *Statistics in Medicine*, **18**, 1605–1613.

Dunson, D.B.(2004) Bayesian isotonic regression for discrete outcomes. *Working Paper*, available at *http://ftp.isds.duke.edu/ WorkingPapers/03-16.pdf*

Grenander, U. (1956) On the theory of mortality measurements II. Skand. Akt. 39 125-153.

Groeneboom, P. (1989) Brownian motion with a parabolic drift and Airy functions. *Probability Theory and Related Fields* **81**, 79 - 109.

Groeneboom, P. and Wellner J.A. (2001) Computing Chernoff's distribution. *Journal of Computational and Graphical Statistics.* **10**, 388-400.

Huang, Y. and Zhang, C. (1994) Estimating a monotone density from censored observations. *Ann. Statist.* **24**, 1256 – 1274.

Kim, J. and Pollard, D. (1990) Cube root asymptotics. *Ann. Statist.* **18**, 191-219.

Morton–Jones, T., Diggle, P. and Elliott, P. (1999) Investigation of excess environment risk around putative sources: Stone's test with covariate adjustment. *Statistics in Medicine*, **18**, 189 – 197.

Pal, J. (2006) A penalized likelihood ratio method to construct confidence intervals for f(0+) when f is a decreasing density. *Working paper*, University of Michigan, Department of Statistics.

Newton, M.A., Czado, C. and Chappell, R. (1996) Bayesian inference for semiparametric binary regression. *JASA*, **91**, 142-153.

Politis, D.M., Romano, J.P., and Wolf, M. (1999) *Subsampling*, Springer–Verlag, New York.

Prakasa Rao, B.L.S. (1969). Estimation of a unimodal density. *Sankhya. Ser. A*, **31** , 23 - 36.

Robertson,T., Wright, F.T. and Dykstra, R.L. (1988). *Order Restricted Statistical Inference*. Wiley, New York

Salanti, G. and Ulm, K. (2003) Tests for trend with binary response. *Biometrical Journal*, **45**, 277-291.

Shiboski, S.(1998) Generalized additive models with current status data. *Lifetime Data Analysis*, **4**, 29 – 50.

Shoung, J-M. and Zhang, C-H. (2001) Least squares estimators of the mode of a unimodal regression function. *Ann. Statist.*, **29**, 648–665.

Stone, R. A. (1988) Investigations of excess environmental risks around putative sources: Statistical problems and a proposed test. *Statistics In Medicine*, **7**, 649–660.

Sun, J. and Kalbfleisch, J.D. (1993) The analysis of current status data on point processes. *JASA*, **88**, 1449-1454.

Sun, J. (1999) A nonparametric test for current status data with unequal censoring. *Jour. of the Royal Stat. Soc. B*, **61**, 243–250.

Van der Vaart, A. and Wellner, J.A. (1996) *Weak Convergence and Empirical Processes*. Springer, New York.

Wellner, J. (2003) Gaussian white noise models: some results for monotone functions. *Crossing Boundaries: Statistical Essays in Honor of Jack Hall*, IMS Lecture Notes-Monograph Series, Vol **43** (2003), 87 – 104. J.E. Kolassa and D. Oakes, editors.

Woodroofe, M. and Sun, J. (1993) A penalized likelihood estimate of f 0+ when f is nonincreasing. *Statist. Sinica*, **3**, 501-515.