Chapter 1

Inference In Exponential Family Regression Models Under Certain Shape Constraints Using Inversion Based Techniques

Moulinath Banerjee*

451 West Hall, 1085 South University Ave.
Ann Arbor, MI, 48109–1107
USA
moulib@umich.edu

We address the problem of pointwise estimation of a regression function under certain shape constraints, using a number of different statistics that can be viewed as measures of discrepancy from a postulated null hypothesis. Pointwise confidence sets are obtained via the usual inversion technique that exploits the duality between construction of confidence sets for a parameter of interest and testing pointwise hypotheses about that parameter. Monotonicity, unimodality and U–shapes are considered. A major advantage of these proposed methods lies in the fact that the statistics of interest are approximately pivotal for large sample sizes and therefore enable inference to be carried out without the need to estimate difficult nuisance parameters. Multivariate generalizations are briefly discussed.

1.1. 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 body of applications. In particular, monotonicity is a shape-restriction that shows up very naturally in different areas 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 problem of estimating a monotone regression function in a signal plus noise model, with additive errors. A key feature of these monotone function problems is the slower pointwise rate of convergence of the MLE ($n^{1/3}$ 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 introduce and study a variety of statistics for estimating a regression function (at a point) that is either monotone, or unimodal, or U–shaped, in a 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*

*The author is an Associate Professor of Statistics and Biostatistics at University of Michigan, Ann Arbor.

below. In what follows, we initially assume that the regression function is monotone increasing. Having constructed statistics of interest and studied their limit behavior in this setting, we proceed to demonstrate how our methods extend to decreasing, unimodal and U–shaped regression functions.

1.1.1. *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: $p(y, \theta) = \exp[\theta T(y) - B(\theta)]h(y)$. Also, recall that $E[T(Y)|X = x] = B' \circ \psi(x) \equiv \mu(x)$. 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, as shown above, is in one–one correspondence with the monotone regression function μ.

Special cases of this generic formulation abound in the literature.

(a) For example, consider 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. 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 and η varying) and can be readily represented in the canonical form.

(b) 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 $G(X)$. 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.

(c) Finally consider the *Poisson regression model* which is useful for modelling count data: $X \sim p_X(\cdot)$ and $Y \mid X = x \sim \text{Poisson}(\lambda(x))$ where λ is a monotone function. Here, one can think of X as the distance 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. Variants of this model have been explored in epidemiological contexts. (Stone (1988), Diggle, Morris and Morton–Jones (1999), Morton–Jones, Diggle and Elliott (1999)).

Let $\hat{\mu}_n$ ($\hat{\psi}_n$) denote the least squares estimate (MLE) of μ and let $\hat{\mu}_n^0$ ($\hat{\psi}_n^0$) denote the constrained least squares estimate (MLE) 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. Each of our proposed statistics is, simply, a measure of discrepancy between the unconstrained and constrained least squares estimates (or MLEs) and can be used to test the null hypothesis above. Confidence sets of a given level for $\mu(x_0)$ will be obtained by inverting these tests, with critical values determined by the quantiles of the corresponding limit distributions under the null hypothesis. Before introducing the statistics of interest, we characterize these estimates below.

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.

The following proposition characterizes least squares estimates under monotonicity constraints and is adapted from Robertson et. al. (1988).

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 let $\mathcal{F}$ denote the set of all real–valued increasing functions defined on $\mathcal{X}$. For $\mathcal{G} \subset \mathcal{F}$, denote the least squares projection of g onto $\mathcal{G}$ with respect to the weight function w by $g_{\mathcal{G}}^\star$; i.e: $g_{\mathcal{G}}^\star \equiv \text{argmin}_{f \in \mathcal{G}} \sum_{i=1}^k (f(x_i) - g(x_i))^2\, w(x_i)$. Now, let (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. For $0 \le i \le k$, let $W_i = \sum_{j=1}^i w(x_j)$ and $G_i = \sum_{j=1}^i w(x_j) g(x_j)$. Then:

$$g_{\mathcal{F}_{u,\eta}}^\star \equiv \text{slogcm}\,\{(W_i, G_i)\}_{i=0}^k \wedge \eta \quad \text{and} \quad g_{\mathcal{F}_{l,\eta}}^\star \equiv \text{slogcm}\,\{(W_i, G_i)\}_{i=0}^k \vee \eta\,.$$

In the above display, the maximum is interpreted as being taken componentwise and so is the minimum.

Remark: Taking $\eta = \infty$, $\mathcal{F}_{u,\eta}$ becomes $\mathcal{F}$ and it follows that $g_{\mathcal{F}}^\star \equiv \text{slogcm}\,\{(W_i, G_i)\}_{i=0}^k$.

Least squares estimates of μ: We are now in a position to characterize the least squares estimate of μ. 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 $\mu_1 \le \mu_2 \le \ldots \le \mu_n$ (where $\mu_i = \mu(X_{(i)})$). Using the above proposition, we find that $\{\hat{\mu}_{ni}\}_{i=1}^n$, the minimizer over all $\mu_1 \le \ldots \le \mu_n$, is given by $\{\hat{\mu}_{ni}\}_{i=1}^n = \text{slogcm}\,\{G_n(X_{(i)}), V_n(X_{(i)})\}_{i=0}^n$, where $G_n(x) = \frac{1}{n}\sum_{i=1}^n 1(X_i \le x)$ and $V_n(x) = \frac{1}{n}\sum_{i=1}^n T(Y_i)\, 1(X_i \le x)$. Using operator notation (for a measure Q on the underlying sample space and a real valued function f defined on the sample space, denote $\int f dQ$ as Qf),

$G_n(x) = \mathbb{P}_n\, 1(-\infty, x]$ and $V_n(x) = \mathbb{P}_n(T(y)\, 1(-\infty, x])$, where $\mathbb{P}_n$ 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: $\{\hat{\mu}^0_{ni}\}_{i=1}^m = \text{slogcm}\, \{G_n(X_{(i)}), V_n(X_{(i)})\}_{i=0}^m \wedge \eta_0$. On the other hand, the vector that solves (b) (say $\{\hat{\mu}^0_{ni}\}_{i=m+1}^n$) is given by:

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

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 characterization of the unconstrained and constrained maximum likelihood estimates (MLEs) of ψ depends heavily on the Kuhn–Tucker theorem. Though this is a very standard theorem in the literature on convex function estimation, we state it briefly, for convenience.

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) $\bigtriangledown \phi(\hat{x}) + G^T_{n\times m}\, \lambda_{m\times 1} = 0$, 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:

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

whence the log-likelihood function for ψ is:

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

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 $\sum_{j=1}^n (T(Y_{(j)}) - B'(\hat{\psi}_{nj})) = 0$.

Setting $\hat{\psi}_{ni} = H(\hat{\mu}_{ni})$ (where H is the inverse function of B'), so that $\hat{\mu}_{ni} = B'(\hat{\psi}_{ni})$ it can be shown that the above conditions are satisifed, whence it follows that the unconstrained MLE $\hat{\psi}_n = H(\hat{\mu}_n)$. For the details of this argument, see Banerjee (2007A).

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. Discrepancy statistics for testing the null hypothesis

We are now in a position to formulate the discrepancy statistics to be used for testing H_0. Inversion of these discrepancy statistics will provide confidence sets for μ (equivalently ψ) at the point x_0. From the regression point of view, one natural statistic for testing the null hypothesis $H_0 : \mu(x_0) = \eta_0$ is the difference in sum of squares given by:

$$DSS(\eta_0) = \sum_{i=1}^{n} (Y_i - \hat{\mu}_n^0(X_i))^2 - \sum_{i=1}^{n} (Y_i - \hat{\mu}_n(X_i))^2 \,. \tag{1.1}$$

Large values of $DSS(\eta_0)$ provide evidence against the null hypothesis. To determine what is "large" will require investigation of the large sample distribution of $DSS(\eta_0)$ that we undertake in the next section. Writing the regression model as $T(Y_i) = \mu(X_i) + \tilde{\epsilon}_i$ (where $\tilde{\epsilon}_i$ has mean 0), note that $DSS(\eta_0)$ can be interpreted as a pseudo likelihood ratio statistic that is obtained by *pretending* that the $\tilde{\epsilon}_i$ are (conditionally homoscedastic) normal errors. Another discrepancy statistic can be constructed by computing a global distance between the unconstrained and constrained least squares estimators. More specifically, consider:

$$L_2(\hat{\mu}_n, \hat{\mu}_n^0) = \sum_{i=1}^{n} (\hat{\mu}_n(X_i) - \hat{\mu}_n^0(X_i))^2 = n \int (\hat{\mu}_n - \hat{\mu}_n^0)^2 \, d\, G_n \,. \tag{1.2}$$

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

$$L_2(\hat{\psi}_n, \hat{\psi}_n^0) = \sum_{i=1}^{n} (\hat{\psi}_n(X_i) - \hat{\psi}_n^0(X_i))^2 = n \int (\hat{\psi}_n - \hat{\psi}_n^0)^2 \, d\, G_n \,. \tag{1.3}$$

While the above provide valid measures of discrepancy, none of these use the actual 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:

$$2 \log \lambda_n(\theta_0) = 2 \left\{ \sum_{i=1}^{n} [\hat{\psi}_n(X_{(i)})\, T(Y_{(i)}) - B(\hat{\psi}_n(X_{(i)}))] - \sum_{i=1}^{n} [\hat{\psi}_n^0(X_{(i)})\, T(Y_{(i)}) - B(\hat{\psi}_n^0(X_{(i)}))] \right\} . \tag{1.4}$$

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 or small) under deviation from the null hypothesis. All that may be inferred is that an atypical value (i.e. either very small or very large) 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 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.

1.2.1. *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 $W(t) + \mid t \mid^{m+1}$), as follows.

Lemma 1.1. *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 $\mathbb{D}_{c,d,m}, \mathbb{T}_{c,d,m}, \mathbb{M}_{1,c,d,m}, \mathbb{M}_{2,c,dm}$ as follows:

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

$$\mathbb{M}_{1,c,d,m} = \int g^0_{c,d,m}(h)\,(g^0_{c,d,m}(h) - g_{c,d,m}(h))\,dh \;\;,\;\; \mathbb{M}_{2,c,d,m} = \int g_{c,d,m}(h)\,(g_{c,d,m}(h) - g^0_{c,d,m}(h))\,dh\,,$$

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 1.1, we can show that the following holds.

Lemma 1.2. *We have:* $c^{-2}(\mathbb{D}_{c,d,m}, \mathbb{T}_{c,d,m}, \mathbb{M}_{1,c,d,m}, \mathbb{M}_{2,c,d,m}) \equiv_d (\mathbb{D}_m, \mathbb{T}_m, \mathbb{M}_{1,m}, \mathbb{M}_{2,m})$.

For proofs of these lemmas (which rely on Brownian scaling arguments) see Section 6 of Banerjee (2007A).

1.3. Limit distributions for the discrepancy statistics and methodological implications

We will study the limit distribution of the discrepancy statistics at the 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 presentation of results at 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 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 each j, for $i \in B_j$, $\hat{\mu}_n(X_{(i)})$ is constant with the constant value depending on j and given by $n_j^{-1} \sum_{i \in B_j} T(Y_{(i)})$, where n_j is the size of B_j. (b) For each B^0_j, for i in B^0_j, $\hat{\mu}^0_n(X_{(i)})$ is constant with the constant value depending on j; moreover, if the constant value on B^0_j is different from η_0, then it is given by $(n^0_j)^{-1} \sum_{i \in B^0_j} T(Y_{(i)})$, where n^0_j is the size of B^0_j.

We now state our main result.

Theorem 1.1. *Assume that the null hypothesis $\psi(x_0) = \theta_0$ (equivalently $\mu(x_0) = \eta_0$) holds. Then:*

(a) The statistic DSS(η_0) defined in (1.1) converges in distribution to $I(\psi(x_0))\, \mathbb{D}_m$, while the likelihood ratio statistic defined in (1.4) 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, $S_{n,1}+S_{n,2} = I(\psi(x_0))\, L_2(\hat{\psi}_n, \hat{\psi}_n^0)+o_p(1)$ 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 $n^{m/(2m+1)}(\hat{\mu}_n(x_0) - \eta_0)$ converges in distribution to $(a^{2m}\, b)^{1/(2m+1)}\, g_m(0)$.

Remarks: For a proof outline of Fact 1, see Section 6 of Banerjee (2007A) which uses the "switching relationship" as demonstrated in Examples 3.2.14 and 3.2.15 of Van der Vaart and Wellner (1996). For an alternative approach, that uses continuous mapping arguments for slope-of-greatest-convex-minorant estimators for the case $m = 1$ in a more general model, see the proof of Theorem 2.1 of Banerjee (2007B). Banerjee (2007B) also studies the limit behavior of the likelihood ratio statistic in this general model but does not deal with the other discrepancy measures considered in this paper; indeed, some of the discrepancy measures like the DSS cannot be tackled in the setup of Banerjee (2007B). The natural connection between least squares and maximum likelihood estimates in the current setting (as noted in Section 1.1) is an outcome of the nice structure of exponential family models but is absent in the general setting of Banerjee (2007B). A proof of Fact 2 is given in Banerjee (2007A). Fact 3 is a straightforward consequence of the characterization of isotonic regression estimators as blockwise averages.

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 Theorem 1 provide a number of 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 $DSS(\eta)$ 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:

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

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

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

while, using the weighted versions we get confidence sets:

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

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

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

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.

1.3.1. *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 $\text{slolcm}\,\{(x_i, y_i)\}_{i=0}^{k}$. With G_n and V_n as before, it is not difficult to see that $\{\hat{\mu}_n(X_{(i)})\}_{i=1}^{n} = \text{slolcm}\,\left\{G_n(X_{(i)}), V_n(X_{(i)})\right\}_{i=0}^{n}$. Also, the MLE under $H_0 : \mu(x_0) = \eta_0$ is given by: $\{\hat{\mu}_n^0(X_{(i)})\}_{i=1}^{m} = \eta_0 \vee \text{slolcm}\,\left\{G_n(X_{(i)}), V_n(X_{(i)})\right\}_{i=0}^{m}$ while $\{\hat{\mu}_n^0(X_{(i)})\}_{i=m+1}^{n} = \eta_0 \wedge \text{slolcm}\,\left\{G_n(X_{(i)}) - G_n(X_{(m)}), V_n(X_{(i)}) - V_n(X_{(m)})\right\}_{i=m}^{n}$. 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. 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 discrepancy statistics from Section 1.2 have the same form as in the monotone function case, with the effective contribution coming from response–covariate pairs for which the covariate lies in a shrinking ($O_p(n^{-1/(2m+1)})$) neighborhood of the point x_0. The asymptotic distributions of these 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. For example, the characterization of the LSE of 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 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 LSEs (and consequently, the discrepancy statistics based on these) stays 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 discrepancy 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 discrepancy statistics for testing the null hypothesis $\mu(x_0) = \eta_0$ can be constructed 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 $\{(Y_{(j)}, X_{(j)}) : j \neq k\}$, 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.

1.4. Concluding Discussion

In this paper, we have demonstrated how pointwise 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 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.

A natural question here is the extent to which the monotonicity assumption is relevant in making pointwise inference on the regression function, since the monotonicity assumption describes the shape of the function globally. We point out that the monotonicity assumption is *crucial* to the inference strategies that we employ: the discrepancy statistics which are functionals of least squares/ maximum likelihood estimates under the monotonicity constraint turn out to be asymptotically pivotal, thereby enabling us to construct pointwise confidence sets *without the need to estimate nuisance parameters or bandwidth parameters*, a difficulty that smoothing techniques have to contend with. This, in our view, is a major statistical advantage. It must be noted, however, that our methods apply only to globally monotone or piecewise monotone functions (unlike smoothing techniques which apply more generally) and should therefore not be employed if such information is unavailable.

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 likelihood based 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 deal with multivariate generalizations. At the very least, it will be important to incorporate auxiliary covariates into the model whose effect can 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. In semiparametric models of this type, the maximum likelihood estimates of ψ do not necessarily admit explicit representations as in the current scenario. In such cases, self–induced characterizations are needed, making the asymptotics more difficult to handle. For some work in this direction, see Banerjee et. al. (2006, 2008). A full discussion of these models 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 be needed before conclusions can be drawn about the relative optimality of any of these pivots with respect to the others.

1.5. Proof of Theorem 1.1

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:

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

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: $III_n = 2 \sum_{j=1}^L \sum_{i\in B_j^0} (T(Y_{(i)}) - \eta_0)(c_j^0 - \eta_0) = 2 \sum_{j=1}^L (c_j^0 - \eta_0) \sum_{i\in B_j^0} (T(Y_{(i)}) - \eta_0) = 2 \sum_{j=1}^L m_j (c_j^0 - \eta_0)^2 = 2 \sum_{j=1}^L \sum_{i\in B_j^0} (\hat{\mu}^0(X_{(i)}) - \eta_0)^2 = 2 \sum_{i\in J_n} (\hat{\mu}^0(X_{(i)}) - \mu(x_0))^2$. Similarly, it follows that $IV_n = 2 \sum_{i\in J_n} (\hat{\mu}(X_{(i)}) - \mu(x_0))^2$.

Consequently:

$$\begin{aligned}\mathrm{DSS}(\eta_0) &= \sum_{i\in J_n}(\hat{\mu}(X_{(i)})-\mu(x_0))^2 - \sum_{i\in J_n}(\hat{\mu}^0(X_{(i)})-\mu(x_0))^2 \\ &= n\,\mathbb{P}_n\left[(\hat{\mu}(x)-\mu(x_0))^2\,1(x\in D_n)\right] - n\,\mathbb{P}_n\left[(\hat{\mu}^0(x)-\mu(x_0))^2\,1(x\in D_n)\right] \\ &= n^{1/(2m+1)}\,(\mathbb{P}_n - P)\left[(X_n^2(h)-Y_n^2(h))\,1(h\in\tilde{D}_n)\right] \\ &\qquad + n^{1/(2m+1)}\,P[(X_n^2(h)-Y_n^2(h))\,1(h\in\tilde{D}_n)]\,, \end{aligned} \tag{1.5}$$

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 $n^{1/(2m+1)}\,(\mathbb{P}_n-P)\,[(X_n^2(h)-Y_n^2(h))\,1(h\in\tilde{D}_n)]$ is $o_p(1)$ and we can write: $\mathrm{DSS}(\eta_0) = n^{1/(2m+1)}\,P\,[(X_n^2(h)-Y_n^2(h))\,1(h\in\tilde{D}_n)] + o_p(1) = \int_{\tilde{D}_n}(X_n^2(h)-Y_n^2(h))\,p_X(x_0+h\,n^{-1/(2m+1)})\,dh = p_X(x_0)\int_{\tilde{D}_n}(X_n^2(h)-Y_n^2(h))\,dh + o_p(1) \to_d p_X(x_0)\int[(g_{a,b,m}(h))^2-(g^0_{a,b,m}(h))^2]\,dh$, where the last step is a consequence of Fact 1. By Lemma 2, it follows that $\mathrm{DSS}(\eta_0)\to_d a^2\,p_X(x_0)\,\mathbb{D}_m$; 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 see that up to an $o_p(1)$ term II_n equals:

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

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

$$2\log\lambda_n(\theta_0) = 2\left[\sum_{i\in J_n}((\hat{\psi}(X_{(i)})-\theta_0)-(\hat{\psi}^0(X_{(i)})-\theta_0))\,(T(Y_{(i)})-B'(\theta_0))\right] - Q_n + o_p(1)\,.$$

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

$$2\sum_{i\in J_n}(\hat{\psi}(X_{(i)})-\theta_0)\,(B'(\hat{\psi}(X_{(i)}))-B'(\theta_0)) - 2\sum_{i\in J_n}(\hat{\psi}^0(X_{(i)})-\theta_0)\,(B'(\hat{\psi}^0(X_{(i)}))-B'(\theta_0))\,. \tag{1.6}$$

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), (1.6) 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:

$$\begin{aligned}2\,\log\,\lambda_n &= I(\psi(x_0))\,n\,\mathbb{P}_n\left[\left\{(\hat{\psi}(x)-\theta_0)^2-(\hat{\psi}^0(x)-\theta_0)^2\right\}1(x\in D_n)\right]+o_p(1)\\ &= I(\psi(x_0))\,n^{1/(2m+1)}\,(\mathbb{P}_n-P)\,[(\tilde{X}_n^2(h)-\tilde{Y}_n^2(h))\,1(h\in\tilde{D}_n)]\\ &\quad +I(\psi(x_0))\,n^{1/(2m+1)}\,P[(\tilde{X}_n^2(h)-\tilde{Y}_n^2(h))\,1(h\in\tilde{D}_n)]+o_p(1)\,. \qquad (1.7)\end{aligned}$$

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 (1.5), 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, $L_{2,w}(\hat{\mu},\hat{\mu}^0)=\sum_{i=1}^n(\hat{\mu}(X_i)-\hat{\mu}^0(X_i))^2\,I(\hat{\psi}(X_i))^{-1}=A_n+B_n$, where $A_n=\sum_{i=1}^n I(\psi(x_0))^{-1}\,(\hat{\mu}(X_i)-\hat{\mu}^0(X_i))^2$ and $B_n=\sum_{i=1}^n[(\hat{\mu}(X_i)-\hat{\mu}^0(X_i))^2\,(I(\hat{\psi}(X_i))-I(\psi(x_0)))]\,[I(\hat{\psi}(X_i))\,I(\psi(x_0))]^{-1}$. Next,

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

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:

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

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:

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

The expression on the right side of the above display can be eventually bounded, with probability larger than any pre–specified amount, by a constant times

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

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:

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

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

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.

The proof of part (c) uses very similar arguments and is skipped. The details are available in Banerjee (2007A). 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 1.1 to arrive at the result.

Acknowledgements: I would like to thank Jayanta Pal for some very helpful discussion. The research for this paper was partly supported by NSF grant DMS-0306235.

1.6. References

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

Banerjee, M., Biswas, P. and Ghosh, D. (2006) A Semiparametric Binary Regression Model Involving Monotonicity Constraints. *Scandinavian Journal of Statistics* **33,** 4, 673–697.

Banerjee, M. (2007A). Competing statistics in exponential family regression models under certain shape constraints. *Technical Report*, **442**, University of Michian, Department of Statistics. Available at http://www.stat.lsa.umich.edu/∼moulib/unimodal-expfamily.pdf

Banerjee, M. (2007B). Likelihood based inference for monotone response models. *Ann. Statist.*, **35**, 3, 931–956.

Banerjee, M., Mukherjee, D. and Mishra, S. (2008) Semiparametric binary regression models under shape constraints with an application to Indian schooling data. Tentatively accepted by *Journal of Econometrics*.

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.

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

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

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) End-point estimation of decreasing densities – asymptotic behavior of penalized likelihood ratio. Available at http://www.samsi.info/jpal/publication.html

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.

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.

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.