Spiking problem in monotone regression : penalized residual sum of squares

Jayanta Kumar Pal[1,2]

SAMSI, NC 27606, U.S.A.

Abstract

We consider the estimation of a monotone regression at its end-point, where the least square solution is inconsistent. We penalize the least square criterion to achieve consistency. We also derive the limit distribution for the residual sum of squares, that can be used to construct confidence intervals.

keywords & phrases : monotone regression, penalization, residual sum of squares, spiking problem.

1 Introduction

Monotone regression gained statistical importance in the last decades, in several scientific problems; e.g. dose-response experiments in biology, modeling disease incidences as function of toxicity levels, industrial experiments like the effect of temperature on the strength of steel etc. The least squares estimator (LSE henceforth), for this problem is described by Robertson et al (1987). We refer to Wright (1981) for its derivation and Groeneboom et al (2001) for its distribution.

In many of the problems, including the ones mentioned above, the value of the regression function at one of the end-points of the domain of the covariate is of special interest. For example,

[1]19, T.W. Alexander Drive, Research Triangle Park, NC 27709, U.S.A., jpal@samsi.info

[2]Research supported by NSF-SAMSI cooperative agreement No. DMS-0112069

the response level at the absence of any dose of medicine or the disease incidence at zero toxicity level (known as base level or placebo) is a quantity that needs specific attention. However, the LSE fails to be consistent at the end-points and has a systematic non-negligible asymptotic bias. This is known as the *spiking problem.* Similar problem also arises in the estimation of decreasing densities, where the maximum likelihood estimator is inconsistent at its left end-point. This has been addressed by Woodroofe and Sun (1993), using a penalization on the likelihood. Recently, Pal (2006, Chapter 5) investigates the behavior of the likelihood ratio in presence of the penalization in the density estimation. Analogously, here we apply the penalization on the least squares criterion and investigate the estimator at the end-point and the corresponding residual sum of squares.

In Section 2 we formulate the penalized least squares problem and characterize the estimators. We state the main result regarding the convergence of the penalized residual sum of squares to a limit distribution as well. In Section 3 we present the proofs of key propositions using technical arguments.

2 Main results

2.1 Preliminaries

We consider the problem of estimating the value of an increasing regression function at its left-end point. The problem of decreasing functions or estimation at the right end point follows similarly, by reflecting the regression on the vertical or horizontal axes. For simplicity, we restrict ourselves to the unit-interval as the support of the co-variate, and take the design points as equally spaced, viz. $x_i = i/n$. Suppose we have observations $Y_1, Y_2, \ldots, Y_n$ at those design points such that

$$Y_i = \mu(x_i) + \epsilon_i$$

for $i = 1, \ldots, n$, and ϵ_i are independent, identically distributed errors with zero means and common variance σ^2 . Assuming that the regression function μ is increasing in $[0, 1]$, the least square estimate (LSE) (MLE in case ϵ_i are Gaussian) of μ which minimizes

$$\sum_{i=1}^{n} (Y_i - \mu(x_i))^2$$

is the Pool-Adjacent-Violator-Algorithm (PAVA henceforth) estimator

$$\vec{\mu}(x_k) = \min_{s \geq k} \max_{r \leq k} \frac{Y_r + \ldots + Y_s}{s - r + 1}$$

Define $\mu_i = \mu(x_i)$ for $i \geq 1$, and $\vec{\mu}_i = \vec{\mu}(x_i)$. Then $\vec{\mu}(0+) = \vec{\mu}(x_1) = \min_{r \geq 1}(Y_1 + \ldots + Y_r)/r$ is inconsistent for $\mu(0+)$. In particular $P(\vec{\mu}(0+) \leq \mu(0+)) \to 1$, always yielding a negative non-negligible bias.

2.2 Penalization on the least square

One way to counter the problem is to penalize for too low estimate of $\mu(0+)$. In particular, we look at the penalized least square

$$\sum_{i=1}^{n} (Y_i - \mu_i)^2 - 2n\alpha\mu_1 \tag{1}$$

where $\alpha = \alpha_n$ is a penalty parameter that depends on the sample size n. For simplicity, we make that dependence implicit.

To start with, we define the Cusum Diagram F of the data as follows. Let $F_r = (Y_1 + \ldots + Y_r)/n$ for $r = 1, \ldots, n$. F is the piecewise linear interpolant of $F_1, \ldots, F_n$ in the interval $[0, 1]$ such that $F(i/n) = F_i$. The following characterization of the penalized least square estimator, denoted $\hat{\mu}$, is similar to the estimator of Woodroofe and Sun (1993). The proof is presented in Section 3.

Proposition 1 *$\hat{\mu}$, the minimizer of 1 can be characterized as,*

$$\hat{\mu}(x_k) = \begin{cases} \min_{0 \leq x \leq 1} \frac{F(x)+\alpha}{x}, & \text{if } k \leq m_0 \\ \vec{\mu}(x_k), & \text{if } k > m_0 \end{cases}$$

where $m_0 = n \, argmin_{0 \leq x \leq 1}(F(x) + \alpha)/x$. In particular $\hat{\mu}(0+) = \min_r \frac{F_r + \alpha}{r/n}$.

The goal is to construct confidence sets for $\mu(0+)$, and for that we need sampling distributions for the estimate $\hat{\mu}(0+)$. The estimator achieves consistency and its limit distribution is derived subsequently. Alternately, we can use the residual sum of squares. However, the regular RSS is not relevant here because of the penalization. Under the constraint $\mu(0+) = c$, a specific value, the penalized sum of square will increase, and reflect the effect of $\mu(0+)$. Hence, we focus on the difference between the constrained and unconstrained penalized sum of squares, and conclude that a large difference indicates a lack of evidence in favor of the hypothesis $\mu(0+) = c$. To be more precise, we measure $Q = \min_{\mu\uparrow, \mu_1 = c} \left[\sum_{i=1}^n (Y_i - \mu_i)^2 - 2n\alpha c\right] - \min_{\mu\uparrow} \left[\sum_{i=1}^n (Y_i - \mu_i)^2 - 2n\alpha\mu_1\right]$ as a test statistic for that hypothesis.

To find the constrained (penalized) least square solution, one would try to minimize the quantity $\sum_{i=1}^n (Y_i - \mu_i)^2$ under the constraint $c = \mu_1 \leq \mu_2 \ldots \leq \mu_n$. Taking β as a Lagrange multiplier, we minimize

$$\sum_{i=1}^{n} (Y_i - \mu_i)^2 - 2n\beta\mu_1 \tag{2}$$

subject to $\mu_1 = c$. The following characterization, similar to the penalized LSE, can be proved using similar arguments. See Section 3, for a proof.

Lemma 1 *$\hat{\mu}^c$, the minimizer of (2), can be characterized as,*

$$\hat{\mu}^c(x_k) = \begin{cases} c, & \text{if } k \leq s_0 \\ \vec{\mu}(x_k), & \text{if } k > s_0 \end{cases}$$

where $s_0 = n\ argmin_{0 \leq x \leq 1}(F(x) - cx)$.

For convenience, we define, $x_{s_0} = s_0/n$ and $x_{m_0} = m_0/n$.

2.3 Limit distribution for the estimators

We state the main result related to the asymptotic behavior of the LSE and the RSS (both penalized) in the following theorem. First, we need a few definitions.

Define, $\mathbb{G}(t) = \mathbb{B}(t) + t^2$, and $\mathbb{G}_1(t) = (\mathbb{G}(t) + 1)/t$, where $\mathbb{B}$ is a standard Brownian motion and let $\mathbb{Z}(t)$ be the right derivative process of the Greatest Convex Minorant (GCM) of the process $\mathbb{G}(t)$. Further, we define $\mathbb{X} = \text{argmin}\mathbb{G}$, $\mathbb{X}_1 = \text{argmin}\mathbb{G}_1$ and $\mathbb{U}_1 = \inf \mathbb{G}_1$. Let,

$$\mathbb{J} = \begin{cases} \mathbb{X}_1\mathbb{U}_1^2 + \int_{\mathbb{X}_1}^{\mathbb{X}} \mathbb{Z}^2(t)dt, & \text{if } \mathbb{X}_1 \leq \mathbb{X} \\ \mathbb{X}_1\mathbb{U}_1^2 - \int_{\mathbb{X}}^{\mathbb{X}_1} \mathbb{Z}^2(t)dt, & \text{if } \mathbb{X} < \mathbb{X}_1 \end{cases}$$

We state the main result as follows. The proof is presented in Section 3.

Theorem 1 *Suppose $\mu'(0+) > 0$ and define $a = \mu'(0+)/2$. Under the penalization $\alpha = (\sigma^2/n^2 a)^{1/3}$, we have,*

$$\begin{aligned} n^{1/3}(\hat{\mu}(0+) - c) &\Rightarrow (a\sigma^2)^{1/3}\mathbb{U}_1 \\ Q &\Rightarrow \sigma^2 \mathbb{J} \end{aligned}$$

where $\Rightarrow$ denotes convergence in distribution.

In the following, we state some preparatory results, with their proofs deferred until Section 3. They will be essential in showing Theorem 1.

To investigate the large sample properties of Q, we need to derive limits of $\hat{\mu}$ and $\hat{\mu}^c$ respectively. Equivalently, we look at the joint limit distribution of $(x_{m_0}, \hat{\mu}(0+), x_{s_0}, \vec{\mu})$. Our next proposition provides the result.

Proposition 2 *Suppose a and α are defined as in Theorem 1. We define a scaled version of the LSE, viz. $\mathbb{Z}_n(t) = n^{1/3}(\vec{\mu}((\sigma^2/na^2)^{1/3}t) - c)$. Then, as $n \to \infty$,*

$$\begin{pmatrix} n^{1/3}(\hat{\mu}(0+) - c) \\ n^{1/3} x_{m_0} \\ n^{1/3} x_{s_0} \\ \{\mathbb{Z}_n(t)\}_{t \geq 0} \end{pmatrix} \Rightarrow \begin{pmatrix} (a\sigma^2)^{1/3}\mathbb{U}_1 \\ (\sigma^2/a^2)^{1/3}\mathbb{X}_1 \\ (\sigma^2/a^2)^{1/3}\mathbb{X} \\ \{\mathbb{Z}(t)\}_{t \geq 0} \end{pmatrix}.$$

The proof is deferred until the next Section. $\square$

The last step before proving Theorem 1 is to express Q in terms of $(x_{m_0}, \hat{\mu}(0+), x_{s_0}, \vec{\mu})$, or equivalently $(m_0, \hat{\mu}(0+), s_0, \vec{\mu})$. The proof of the following Lemma requires algebraic manipulation and several properties of the estimator $\vec{\mu}$.

Lemma 2 *Q can be expressed as follows :*

$$Q = \begin{cases} m_0(\hat{\mu}(0+) - c)^2 + \sum_{i=m_0+1}^{s_0} (\vec{\mu}_i - c)^2, & \text{if } m_0 < s_0 \\ m_0(\hat{\mu}(0+) - c)^2 - \sum_{i=s_0+1}^{m_0} (\vec{\mu}_i - c)^2, & \text{if } s_0 \leq m_0 \end{cases}$$

Using Proposition 2 and Lemma 2, we can deduce Theorem 1 using continuous mapping theorem etc. The details are sketched in the proof in the next Section.

2.4 Discussion

Theorem 1 shows consistency for the penalized LSE and gives the limit distribution for both LSE and the RSS. This is analogous to the asymptotic normality of the usual LSE and the χ^2 convergence of the corresponding RSS in the regular parametric models. Moreover, this can be compared to the corresponding results in case of monotone regression at an interior point. There, the LSE converges to a non-normal distribution at a rate $n^{1/3}$, and the RSS or the likelihood ratio converges to a universal distribution different from χ^2. See, e.g. Banerjee and Wellner (2001). However, at the end-point, we get a completely different distribution, and the effect of $\mu'(0+)$ can not be neglected.

However, it should be noted that we need a consistent estimator of σ^2 as well as $\mu'(0+)$ to achieve a confidence interval. Both of them are relatively tricky. Meyer and Woodroofe (2000) discusses a method that yields consistent estimator for σ^2 using the LSE $\vec{\mu}$. The estimation of $\mu'(0+)$ can be done using similar techniques as in density estimation, discussed by Pal (2006).

3 Proofs

3.1 Proof of Proposition 1

We observe, $\sum_{i=1}^n (Y_i - \mu_i)^2 - 2n\alpha\mu_1 = \sum_{i=1}^n (\tilde{Y}_i - \mu_i)^2 - 2n\alpha Y_1 - n^2\alpha^2$ where $\tilde{Y}_1 = Y_1 + n\alpha$ and $\tilde{Y}_i = Y_i$ for $i \geq 2$. Therefore the minimizer $\hat{\mu}$ is the PAVA estimator with new data $\tilde{Y}_1, Y_2, \ldots, Y_n$ i.e. $\hat{\mu}_k = \min_{s \geq k} \max_{r \leq k} (\tilde{Y}_r + \ldots + \tilde{Y}_s)/(s - r + 1)$. Therefore,

$$\begin{aligned}
\hat{\mu}(0+) &= \min_{r \geq 1} \frac{\tilde{Y}_1 + \ldots + \tilde{Y}_r}{r} \\
&= \min_{r \geq 1} \frac{Y_1 + \ldots + Y_r + n\alpha}{r} \\
&= \min_{r \geq 1} \frac{n(F_r + \alpha)}{r} \\
&= \min_{0 \leq x \leq 1} \frac{F(x) + \alpha}{x}
\end{aligned}$$

Moreover, except for the first block (i.e. until the point where $\hat{\mu}_i = \hat{\mu}_1$), the contribution of $\tilde{Y}_1$ does not influence the estimator $\hat{\mu}$, and hence $\hat{\mu}_k = \vec{\mu}_k$. Let m_0 denote the first break-off point of the estimator $\hat{\mu}$. It is characterized as

$$m_0 = \text{argmin}_{r\geq 1}\frac{Y_1+\ldots+Y_r+n\alpha}{r} = \text{argmin}\frac{n(F_r+\alpha)}{r} = n\,\text{argmin}_{0\leq x\leq 1}\frac{F(x)+\alpha}{x}$$

The proposition follows. $\square$

3.2 Proof of Lemma 1

As in Proposition 2, $\hat{\mu}^c$ is the PAVA estimator of a new data set $Y_1+n\hat{\beta}, Y_2, \ldots, Y_n$ where $\hat{\beta}$ solves the equation

$$\hat{\mu}_1^c = \min_{s\geq 1}\frac{Y_1+\ldots+Y_s+n\beta}{s} = c$$

Therefore, $\hat{\beta} = \max_{s\geq 1}(cs/n - F_s) = \max_{0\leq x\leq 1}(cx - F(x))$. To see this, observe that

$$\begin{aligned}
&\min_{s\geq 1}\frac{Y_1+\ldots+Y_s+n\hat{\beta}}{s} = c \\
\Leftrightarrow\quad &\frac{Y_1+\ldots+Y_s+n\hat{\beta}}{s} \geq c \qquad \text{for all s with equality at least once,} \\
\Leftrightarrow\quad &nF_s + n\hat{\beta} \geq cs \qquad \text{for all s with equality at least once,} \\
\Leftrightarrow\quad &\hat{\beta} \geq \frac{cs}{n} - F_s \qquad \text{for all s with equality at least once,}
\end{aligned}$$

and the definition of $\hat{\beta}$ follows. Moreover, $\hat{\mu}^c = c$ until its first break-off point s_0, and equal to $\vec{\mu}$ beyond that. Clearly, $s_0 = \text{argmin}(F_s - cs/n) = n\,\text{argmin}_{0\leq x\leq 1}(F(x) - cx)$. $\square$

3.3 Proof of Proposition 2

Define the step function $\Lambda_n(x) = \lfloor nx \rfloor/n$. To prove the result, we invoke the Hungarian embedding for the limit of partial sums of i.i.d. random variables. See, e.g. Karatzas & Shreve (1991) for a discussion.

Define the partial sums $S_r = \epsilon_1 + \ldots + \epsilon_r$ for $r = 1, \ldots, n$. Assume that ϵ_i's have finite moment generating function and common variance σ^2. Then, we can ascertain the existence of a standard

Brownian Motion $\mathbb{B}_n$ for each n, and a sequence of random variables T_n having the same distribution as S_n such that,

$$\sup_t |T_{\lfloor nt \rfloor} - \sigma\sqrt{n}\mathbb{B}_n(t)| = O(\log n)$$

almost surely. This result is also known as the strong approximation. Since we are looking at a distribution convergence, we treat T_n and S_n likewise, and replace one of them by other. We observe that,

$$\begin{aligned}
F(x) &= \frac{1}{n}\sum_{i=1}^{\lfloor nx \rfloor} Y_i + O_p(\frac{1}{n}) \\
&= \frac{1}{n}\sum_{i=1}^{\lfloor nx \rfloor} [\mu(\frac{i}{n}) + \epsilon_i] + O_p(\frac{1}{n}) \\
&= \int_0^x \mu(t)d\Lambda_n(t) + \frac{1}{n}S_{n\Lambda_n(x)} + O_p(\frac{1}{n}) \\
&= \int_0^x \mu(t)dt + \frac{\sigma}{\sqrt{n}}\mathbb{B}_n(x) + O_p(\frac{1}{n}) + O(\frac{\log n}{n}) \\
&= \int_0^x (\mu(0) + t\mu'(0) + \frac{t^2}{2}\mu''(t^*))dt + \frac{\sigma}{\sqrt{n}}\mathbb{B}_n(x) + O_p(\frac{1}{n}) + O(\frac{\log n}{n}) \\
&= cx + ax^2 + o(x^2) + \frac{\sigma}{\sqrt{n}}\mathbb{B}_n(x) + O_p(\frac{1}{n}) + O(\frac{\log n}{n}) \\
&= cx + ax^2 + \frac{\sigma}{\sqrt{n}}\mathbb{B}_n(x) + R_n(x)
\end{aligned}$$

where the remainder $R_n(x)$ is negligible for calculations. Define $b = (\sigma^2/na^2)^{1/3}$ and $\mathbb{B}'_n(t) = \mathbb{B}_n(t)/\sqrt{b}$. Then,

$$\begin{aligned}
x_{m_0} &= \text{argmin}\frac{F(x)+\alpha}{x} \\
&= \text{argmin}_{0\le x\le 1}\left[\frac{cx + ax^2 + \frac{\sigma}{\sqrt{n}}\mathbb{B}_n(x) + \alpha + R_n(x)}{x}\right] \\
&= b\,\text{argmin}_{0\le t\le \frac{1}{b}}\left[c + \frac{ab^2t^2 + \frac{\sigma}{\sqrt{n}}\mathbb{B}_n(bt) + \alpha}{bt}\right] + o_p(b) \qquad \text{taking } x = bt \\
&= b\,\text{argmin}_{0\le t\le \frac{1}{b}}\left[\frac{ab^2t^2 + \frac{\sigma\sqrt{b}}{\sqrt{n}}\mathbb{B}'_n(t) + \alpha}{t}\right] + o_p(n^{-1/3})
\end{aligned}$$

Using $\alpha = (\sigma^4/n^2a)^{1/3}$, so that $ab^2 = \sigma\sqrt{b/n} = \alpha$, we conclude,

$$x_{m_0} = (\frac{\sigma^2}{na^2})^{\frac{1}{3}}\text{argmin}_{0\le t\le (\frac{na^2}{\sigma^2})^{\frac{1}{3}}}\left[\frac{t^2 + \mathbb{B}'_n(t) + 1}{t}\right] + o_p(n^{-1/3})$$

and therefore,

$$n^{\frac{1}{3}} x_{m_0} \Rightarrow (\frac{\sigma^2}{a^2})^{\frac{1}{3}} \mathbb{X}_1$$

Moreover,

$$\begin{aligned}
\hat{\mu}(0+) - c &= \min_x \frac{F(x) + \alpha}{x} - c \\
&= \min_x \frac{ax^2 + \frac{\sigma}{\sqrt{n}} \mathbb{B}_n(x) + \alpha + R_n(x)}{x} \\
&= \min_{0 \le t \le \frac{1}{b}} ab \frac{t^2 + \mathbb{B}'_n(t) + 1}{t} + o_p(n^{-1/3})
\end{aligned}$$

i.e.

$$n^{\frac{1}{3}} (\hat{\mu}(0+) - c) \Rightarrow (a\sigma^2)^{\frac{1}{3}} \mathbb{U}_1$$

Consider the constrained estimator now. First,

$$\begin{aligned}
x_{s_0} &= \operatorname{argmin}(F(x) - cx) \\
&= \operatorname{argmin}_x \Big[ax^2 + \frac{\sigma}{\sqrt{n}} \mathbb{B}_n(x) + R_n(x) \Big] \\
&= (\frac{\sigma^2}{na^2})^{\frac{1}{3}} \operatorname{argmin}_t (t^2 + \mathbb{B}'_n(t)) + o_p(n^{-\frac{1}{3}})
\end{aligned}$$

so that

$$n^{\frac{1}{3}} x_{s_0} \Rightarrow (\frac{\sigma^2}{a^2})^{\frac{1}{3}} \mathbb{X}.$$

We look at the behavior of $\vec{\mu}$, properly scaled, in the region (x_{s_0}, x_{m_0}). We remember that it is the right-derivative of the Greatest Convex Minorant of the Cusum diagram $F(x)$ at the point $x = s$. Approximating $F(x)$ by $ax^2 + cx + \sigma/\sqrt{n}\mathbb{B}_n(x)$ in the interval, we write,

$$\vec{\mu}(s) = \frac{d}{dx} (GCM(ax^2 + cx + (\frac{\sigma}{\sqrt{n}}) \mathbb{B}_n(x)))|_{x=s}$$

Using the scaling $x = bt$ as before, and the fact that the GCM of a process added with a linear term is same as the linear term added with the GCM of the process itself, we get,

$$\begin{aligned}
\vec{\mu}(bt) - c &= \frac{1}{b} \frac{d}{dt} (GCM(ab^2 t^2 + \frac{\sigma\sqrt{b}}{\sqrt{n}} \mathbb{B}'_n(t))) \\
&= ab \frac{d}{dt} (GCM(t^2 + \mathbb{B}'_n(t))) \qquad \text{since } ab^2 = \frac{\sigma\sqrt{b}}{\sqrt{n}} \\
&= (\frac{a\sigma^2}{n})^{\frac{1}{3}} \frac{d}{dt} (GCM(t^2 + \mathbb{B}'_n(t)))
\end{aligned}$$

which yields,

$$n^{\frac{1}{3}}(\vec{\mu}(t(\frac{\sigma^2}{na^2})^{\frac{1}{3}}) - c) \Rightarrow (a\sigma^2)^{\frac{1}{3}}\mathbb{Z}(t)$$

The remainders in all these calculations are negligible as we look at points those are at a distance $O_p(n^{-1/3})$ from zero. Moreover, since the individual distributions are derived as limit of the same Brownian motion $\mathbb{B}'_n$, the joint convergence follows. $\square$

3.4 Proof of Lemma 2

Proof : The following identities are useful to prove the Lemma.

$$\sum_{i=1}^{m_0} Y_i = \sum_{i=1}^{m_0} \vec{\mu}_i = m_0\hat{\mu}_1 - n\alpha \tag{A}$$

$$\sum_{i=1}^{s_0} Y_i = \sum_{i=1}^{s_0} \vec{\mu}_i = cs_0 - n\beta \tag{B}$$

If $m_0 < s_0$

$$\sum_{i=m_0+1}^{s_0} Y_i = \sum_{i=m_0+1}^{s_0} \vec{\mu}_i \tag{C1}$$

$$\sum_{i=m_0+1}^{s_0} \vec{\mu}_i Y_i = \sum_{i=m_0+1}^{s_0} \vec{\mu}_i^2 \tag{C2}$$

The first part of (A) follows from the properties of the PAVA estimator, which is constant over each block, and equals the average of the Y-values in that block. The second part follows similarly, since m_0 is the first break-off point for $\tilde{Y_1}, Y_2$ etc, and therefore $\hat{\mu}_1 = (\tilde{Y_1} + \ldots + Y_{m_0})/m_0 = (\sum_{i=1}^{m_0} Y_i + n\alpha)/m_0$. (B) and (C1) follows similarly. (C2) follows by splitting the summation over blocks first, and then taking the sum of the Y-s in each block. In case $s_0 \leq m_0$, the equalities (C1) and (C2) remain still valid, with the role of m_0 and s_0 being interchanged. Now,

$$
\begin{aligned}
Q &= \min_{\mu\uparrow,\mu_1=c}\Big[\sum_{i=1}^{n}(Y_i-\mu_i)^2-2n\alpha\mu_1\Big]-\min_{\mu\uparrow}\Big[\sum_{i=1}^{n}(Y_i-\mu_i)^2-2n\alpha\mu_1\Big] \\
&= \sum_{i=1}^{n}(Y_i-\hat{\mu}_i^c)^2-2n\alpha c-\sum_{i=1}^{n}(Y_i-\hat{\mu}_i)^2+2n\alpha\hat{\mu}_1 \\
&= \sum_{i=1}^{s_0}(Y_i-c)^2+\sum_{i=s_0+1}^{n}(Y_i-\vec{\mu}_i)^2-\sum_{i=1}^{m_0}(Y_i-\hat{\mu}_1)^2-\sum_{i=m_0+1}^{n}(Y_i-\vec{\mu}_i)^2+2n\alpha(\hat{\mu}(0+)-c)
\end{aligned}
$$

We use the identities (A) etc in the following two cases.

Case 1: $m_0 < s_0$

$$
\begin{aligned}
Q &= 2n\alpha(\hat{\mu}(0+)-c)+\sum_{i=1}^{m_0}[(Y_i-c)^2-(Y_i-\hat{\mu}_1)^2]+\sum_{i=m_0+1}^{s_0}[(Y_i-c)^2-(Y_i-\vec{\mu}_i)^2] \\
&= 2n\alpha(\hat{\mu}(0+)-c)+2(\hat{\mu}_1-c)\sum_{i=1}^{m_0}Y_i+m_0[c^2-\hat{\mu}_1^2]+\sum_{i=m_0+1}^{s_0}[c^2-2cY_i-\vec{\mu}_i^2+2\vec{\mu}_iY_i] \\
&= 2m_0(\hat{\mu}_1-c)\hat{\mu}_1+m_0[c^2-\hat{\mu}_1^2]+\sum_{i=m_0+1}^{s_0}[c^2-2c\vec{\mu}_i-\vec{\mu}_i^2+2\vec{\mu}_i^2] \\
&= m_0(\hat{\mu}(0+)-c)^2+\sum_{i=m_0+1}^{s_0}(\vec{\mu}_i-c)^2
\end{aligned}
$$

Case 2: $s_0 \le m_0$

$$
\begin{aligned}
Q &= 2n\alpha(\hat{\mu}_1-c)+\sum_{i=1}^{s_0}[(Y_i-c)^2-(Y_i-\hat{\mu}_1)^2]+\sum_{i=m_0+1}^{s_0}[(Y_i-\vec{\mu}_i)^2-(Y_i-\hat{\mu}_1)^2] \\
&= 2n\alpha(\hat{\mu}_1-c)+2(\hat{\mu}_1-c)\sum_{i=1}^{s_0}Y_i+s_0[c^2-\hat{\mu}_1^2]+\sum_{i=s_0+1}^{m_0}[\vec{\mu}_i^2-\hat{\mu}_1^2-2\vec{\mu}_iY_i+2\hat{\mu}_1Y_i] \\
&= 2n\alpha(\hat{\mu}_1-c)+2(\hat{\mu}_1-c)(cs_0-n\beta)+s_0[c^2-\hat{\mu}_1^2]-(m_0-s_0)\hat{\mu}_1^2 \\
&\qquad +\sum_{i=s_0+1}^{m_0}[\vec{\mu}_i^2-2\vec{\mu}_i^2]+2\hat{\mu}_1\sum_{i=s_0+1}^{m_0}\vec{\mu}_i \\
&= 2(\hat{\mu}_1-c)(n\alpha+cs_0-n\beta+\sum_{i=s_0+1}^{m_0}\vec{\mu}_i)+s_0c^2-m_0\hat{\mu}_1^2-\sum_{i=s_0+1}^{m_0}(\vec{\mu}_i-c)^2+c^2(m_0-s_0) \\
&= m_0(\hat{\mu}(0+)-c)^2-\sum_{i=s_0+1}^{m_0}(\vec{\mu}_i-c)^2, \qquad (\text{ since } \sum_{i=s_0+1}^{m_0}\vec{\mu}_i=m_0\hat{\mu}_1-n\alpha-cs_0+n\beta)
\end{aligned}
$$

The lemma follows. $\square$

3.5 Proof of Theorem 1

To start with, let $s_0 \leq m_0$. Then, $\sum_{i=s_0+1}^{m_0}(\vec{\mu}_i - c)^2 = n\int_{x_{s_0}}^{x_{m_0}}(\vec{\mu}(s) - c)^2 d\Lambda_n(s) = n\int_{x_{s_0}}^{x_{m_0}}(\vec{\mu}(s) - c)^2 ds + o_p(1)$, since both x_{m_0} and x_{s_0} are of order $n^{-1/3}$. Similarly for $m_0 < s_0$, we have $\sum_{i=m_0+1}^{s_0}(\vec{\mu}_i - c)^2 = n\int_{x_{m_0}}^{x_{s_0}}(\vec{\mu}(s) - c)^2 ds + o_p(1)$.

From Proposition 2, $(na^2/\sigma^2)^{1/3}(x_{m_0}, x_{s_0})$ converges weakly to $(\mathbb{X}_1, \mathbb{X})$ where the limit processes are functions of the same standard Brownian Motion $\mathbb{B}(t)$. Therefore,

$$\begin{aligned} n\int_{x_{m_0}}^{x_{s_0}}(\vec{\mu}(s) - c)^2 ds &= n\int_{\frac{x_{m_0}}{b}}^{\frac{x_{s_0}}{b}}(\vec{\mu}(bt) - c)^2 b\, dt \\ &= \int_{\frac{x_{m_0}}{b}}^{\frac{x_{s_0}}{b}}\left[n^{\frac{1}{3}}(\vec{\mu}(t(\frac{\sigma^2}{na^2})^{\frac{1}{3}}) - c)\right]^2 (n^{\frac{1}{3}}b)dt \\ &\Rightarrow \sigma^2\int_{\mathbb{X}_1}^{\mathbb{X}}\mathbb{Z}^2(t)dt \end{aligned}$$

Similarly, the same result holds when on the set $x_{s_0} \leq x_{m_0}$, with their roles being interchanged.

Finally, we recall that,

$$\begin{aligned} m_0(\hat{\mu}(0+) - c)^2 &= n^{\frac{1}{3}x_{m_0}}\left[n^{\frac{1}{3}}(\hat{\mu}(0+) - c)\right]^2 \\ &\Rightarrow (\frac{\sigma^2}{a^2})^{\frac{1}{3}}\mathbb{X}_1(a\sigma^2)^{\frac{2}{3}}\mathbb{U}_1^2 \\ &= \sigma^2\mathbb{X}_1\mathbb{U}_1^2 \end{aligned}$$

Therefore, using Lemma 2,

$$Q \Rightarrow \sigma^2(\mathbb{X}_1\mathbb{U}_1^2 + \int_{\mathbb{X}_1}^{\mathbb{X}}\mathbb{Z}^2(t)dt) = \sigma^2\mathbb{J}$$

where the integral changes sign when $\mathbb{X}_1 > \mathbb{X}$. $\square$

Acknowledgments

The research is supported by NSF-SAMSI cooperative agreement No. DMS-0112069, during post-doctoral research in SAMSI. Discussions with Professor Moulinath Banerjee have been instrumental for the inception of the problem. The author also acknowledges the contribution of Professor Michael Woodroofe as the doctoral thesis supervisor.

References

[1] Banerjee, M. and Wellner, J. A. (2001), Likelihood Ratio Tests for Monotone Functions, *Ann. Statist.* **29**, 1699-1731.

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

[3] Karatzas, I. and Shreve, S. E. (1991), *Brownian motion and Stochastic calculus* (Springer-Verlag, New York)

[4] Meyer, M. and Woodroofe, M. B. (2000), On the degrees of freedom of shape restricted regression. *Ann. Statist.* **28**, 1083-1104.

[5] Pal, J. K. (2006), *Statistical analaysis and inference in shape-restricted problems with applications to astronomy* (Doctoral Dissertation, University of Michigan).

[6] Robertson, T., Wright, F. T. and Dykstra, R. L. (1987), *Order Restricted Statistical Inference* (John Wiley, New York).

[7] Woodroofe, M. B. and Sun, J. (1993), A Penalized Maximum Likelihood estimate of f(0+) when f is non-increasing, *Statist. Sinica* **3** 501-515.

[8] Wright, F. T. (1981), The asymptotic behavior of monotone regression estimates, *Ann. Statist.* **9**, 443-448.