Corresponding Author: Dr. Jayanta Kumar Pal, Ph.D.

Corresponding Author's Institution: SAMSI

First Author: Jayanta Kumar Pal, Ph.D.

Order of Authors: Jayanta Kumar Pal, Ph.D.; Moulinath Banerjee, Ph.D.

Abstract: We consider the estimation of smooth regression functions in a class of conditionally parametric covariate-response models. Independent and identically distributed observations are available from the distribution of (Z,X), where Z is a real--valued covariate with some unknown distribution, and the response X conditional on Z is distributed according to the density p(.,psi(Z)), where p(.,theta) is a one-parameter exponential family. The function psi is a smooth monotone function. Under this formulation, the regression function E(X|Z) is monotone in the covariate Z (and can be expressed as a one--one function of psi); hence the term ``monotone response model''. Using a penalized least squares approach that incorporates both monotonicity and smoothness, we develop a scheme for producing smooth monotone estimates of the regression function and also the function psi across this entire class of models. Pointwise asymptotic normality of this estimator is established, with the rate of convergence depending on the smoothing parameter. This enables construction of Wald--type (pointwise) as well as
pivotal confidence sets for psi and also the regression function.

SAMSI,
19, T.W. Alexander Drive
E-mail : `jkp3@duke.edu`

To
The Editor,
Journal of Statistical Planning and Inference.

Respected Sir/Madam,

    This letter is to inform you about a manuscript submitted for possible publication in your esteemed journal. We submitted the article 'Estimation of smooth regression functions in monotone response models' by Jayanta Kumar Pal and Moulinath Banerjee, and request you to consider it for review. I look forward to your response and will be grateful to receive an acknowledgement and/or a reference number for future correspondence.
Thank you.
Yours truly,
Jayanta Kumar Pal.

Estimation of smooth regression functions in monotone response models

Jayanta Kumar Pal [a,*], Moulinath Banerjee [b]

[a]*SAMSI, 19 T.W. Alexander Drive, Research Triangle Park, NC 27606*

[b]*University of Michigan, 1085 South University, Ann Arbor, MI 48109*

Abstract

We consider the estimation of smooth regression functions in a class of conditionally parametric covariate-response models. Independent and identically distributed observations are available from the distribution of (Z, X), where Z is a real–valued covariate with some unknown distribution, and the response X conditional on Z is distributed according to the density $p(\cdot, \psi(Z))$, where $p(\cdot, \theta)$ is a one-parameter exponential family. The function ψ is a smooth monotone function. Under this formulation, the regression function $E(X \mid Z)$ is monotone in the covariate Z (and can be expressed as a one–one function of ψ); hence the term "monotone response model". Using a penalized least squares approach that incorporates both monotonicity and smoothness, we develop a scheme for producing smooth monotone estimates of the regression function and also the function ψ across this entire class of models. Pointwise asymptotic normality of this estimator is established, with the rate of convergence depending on the smoothing parameter. This enables construction of Wald–type (pointwise) as well as pivotal confidence sets for ψ and also the regression function.

Key words: monotone response models, smoothing spline, confidence interval

1 Introduction

Monotonicity is a shape-restriction that shows up very naturally in different areas of application like reliability, renewal theory, epidemiology and biomedical studies. For example, the monotone function of interest could be a distribution

* Corresponding author

Email addresses: `jpal@samsi.info` (Jayanta Kumar Pal), `moulib@umich.edu` (Moulinath Banerjee).

function or a cumulative hazard function (survival analysis), the mean function of a counting processes (demography, reliability, clinical trials), regression function (dose-response modeling, modeling disease incidence as a function of distance from a toxic source), density (inference in renewal theory and other applications) or hazard rate (reliability).

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 homoscedastic errors. A key feature of these monotone function problems is the slower rate of convergence (usually $n^{1/3}$) of the MLE, 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 was characterized by Groeneboom (1989).

Many of these monotone function models relate a covariate to the corresponding response through a monotone "dependence function". We call such models *monotone response models* and describe them as follows.

Consider independent and identically distributed observations $\{X_i, Z_i\}_{i=1}^n$, where each (X_i, Z_i) is distributed like (X, Z), and (X, Z) is distributed in the following way: The covariate Z is supported on a compact interval H and has a Lebesgue density p_Z (with distribution function F_Z). The conditional density of X given that $Z = z$ is given by $p(\cdot, \psi(z))$, where $\{p(\cdot, \theta) : \theta \in \Theta\}$ is a one–parameter exponential family of densities (with respect to some dominating measure μ) 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:

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

Now, it is easy to see that,

$$E[T(X)|Z = z] = B' \circ \psi(z) = m(z) \text{ (say)}.$$

Since B is infinitely differentiable on Θ (an open set) and ψ is continuous on H, $B^{(k)} \circ \psi$ is continuous and therefore bounded on H, for every $k > 0$. Also, $B''(\theta) = \text{Var}_\theta(T) > 0$ for every θ, implying that B' is a strictly increasing function. Hence, B' is invertible (with inverse function H, say) and estimating m is equivalent to estimating ψ. The function ψ is called the *dependence func-*

tion and is in 1-1 correspondence with the regression function m. Our target is to find a smooth estimate of ψ (or equivalently m) in the general scenario.

The monotone response model can be viewed as a generalization of simple monotone regression (see (a) below) in the same vein as the generalized linear model is viewed as an extension of the simple linear regression model. The usefulness of this general framework lies in the various dissimilar scenarios it comprises, as motivated by the following examples.

(a) **Monotone Regression Model:** Consider the model

$$X_i = \mu(Z_i) + \epsilon_i \,,$$

where $\{(\epsilon_i, Z_i)\}_{i=1}^n$ are i.i.d. random variables, ϵ_i is independent of Z_i, each ϵ_i is normal with mean 0 and variance σ^2, each Z_i has a Lebesgue density $p_Z(\cdot)$ and μ is a monotone function. The above model and its variants have been fairly well–studied in the literature on isotonic regression. Note the reference to Brunk (1970) above; more references follow in the discussion later. Here, $Z \sim p_Z(\cdot)$ and $X \mid Z = z \;\sim\; N(\mu(z), \sigma^2)$. This conditional density comes from the one–parameter exponential family $N(\eta, \sigma^2)$ (for fixed σ^2), η varying.

(b) **Binary Choice Model:** Here we have a dichotomous response variable $X = 1$ or 0 and a continuous covariate Z with a Lebesgue density $p_Z(\cdot)$ such that $P(X = 1 \mid Z) \equiv G(Z)$ is a smooth function of Z. Thus, conditional on Z, X has a Bernoulli distribution with parameter $G(Z)$. To cast this model in terms of the natural parameter, we take the link function as $\psi(z) = \log\left(G(z)/(1 - G(z))\right)$ and $p(x, \theta) = x\theta - \log(1 + e^\theta)$ for $\theta \in \mathbb{R}$, whence the natural sufficient statistic $T(x) = x$, $B(\theta) = \log(1 + e^\theta)$ and $m(z) = G(z)$. Models of this kind have been quite broadly studied in econometrics and statistics (see, for example, Dunson (2005), Newton et al (1996), Salanti & Ulm (2003)). In a biomedical context one could think of X as representing the indicator of a disease/infection and Z the level of exposure to a toxin, or the measured level of a biomarker 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 (that will appear in our data–analysis section) is the "Current Status Model" from survival analysis that is used extensively in epidemiology and has received much attention among biostatisticians and statisticians (see, for example, Sun & Kalbfleisch (1995), Huang (1996), Shiboski (1998), Sun (1999), Banerjee & Wellner (2001)). Consider n individuals who are checked for infection at independent random times $T_1, T_2, \ldots, T_n$; we set $X_i = 1$ if individual i is infected by time T_i and 0 other-

wise. We can think of X_i as $1\{W_i \leq T_i\}$ where W_i is the unobserved (random) time to infection, measured from some baseline period. The W_i's are assumed to be independent and also independent of the T_i's. We are interested in making inference on F, the common distribution function of the W_i's. We note that $\{X_i, T_i\}_{i=1}^{n}$ is an i.i.d. sample from the distribution of (X, T) where $T \sim h(\cdot)$ for some Lebesgue density h and $X \mid T \sim \text{Bernoulli}(F(T))$. This is precisely the (monotone) binary choice model considered above, with T and F playing the role of Z and G.

(c) **Poisson Regression Model:** Suppose that $Z \sim p_Z(\cdot)$ and $X \mid Z = z \sim \text{Poisson}(\psi(z))$ where ψ is a monotone function. Here one can think of Z as the distribution of a region from a hazardous point source (for example, a nuclear processing plant or a mine) and X the number of cases of disease incidence at distance Z (say, cancer occurrences due to radioactive exposure in the case of the nuclear processing plant, or Silicosis in the case of the mine). Given $Z = z$, the number of cases of disease incidence X at distance z from the source is assumed to follow a Poisson distribution with mean $\theta(z)$ where θ can be expected to be monotonically decreasing in z, since the harmful effect should decay as we move further out from the source. Variants of this model have received considerable attention in epidemiological contexts (Stone (1988), Diggle et al (1999), Morton-Jones et al (1999)).

In each of these examples, the regression function $E(X \mid Z)$ is a monotone function of Z, and one seeks to make inference on the regression function. Maximum likelihood estimation in each of the above problems (and for monotone function problems, in general) reduces to solving an equivalent isotonic regression problem, which can be done, for example, by using the standard Pool-Adjacent-Violator-Algorithm (PAVA). (See, Robertson et al (1989)). However, the MLE is piecewise constant and has abrupt discontinuities. This can often render it undesirable for scientific purposes; especially since the underlying function is thought to be smooth, jumps become difficult to interpret. A naive approach is to use kernel smoothing on the MLE of the regression function; however, this suffers from two main difficulties. Firstly, smoothing may destroy monotonicity and secondly, the number of support of points of the MLE is $O_p(n^{1/3})$, which is far smaller than the order of support points one has in standard density estimation and nonparametric regression contexts. This will typically make the kernel smoothed estimator unreliable.

A number of authors have come up with different solutions to this problem of combining smoothness with shape restrictions in regression settings. Wright & Wegman (1980) for example, worked with spline based estimation of monotone and convex/concave functions. Mukerjee (1988) devised an estimate based on

4

kernel–type smoothing and showed that it is asymptotically normal. Mammen (1991)'s approach deals with successive kernel smoothing and isotonization, giving the estimator m_{SI}, or the interchanged estimator m_{IS}.

However, the assumption of a bivariate joint Lebesgue density for the regressor and the covariate limits the scope of Mukerjee's approach, since discrete response variables cannot be incorporated. Mammen's method, on the other hand, does not incorporate the case of random regressors, or heteroscedasticity of errors (mean–variance relationships), and is therefore not directly applicable to monotone response models either. In another approach, Kelly & Rice (2001) use monotone smoothing through B-splines, with application to estimating dose-response curves. However, they do not present a theoretical justification of their method, so effective comparison to other methods is difficult in this case. Recently, Ramsay (1998) worked on a similar spline-based method, but on a different transformed variable. Dette et al (2005) deal with a smooth binary regression model under monotonicity constraints but do not discuss extensions to the general exponential family case.

Recently, Pal & Woodroofe (2007) have developed a method of estimating a smooth monotone regression function in a model with fixed design points and additive mean 0 i.i.d. errors. This involves minimizing a penalized least squares criterion incorporating monotonicity constraints and weighing the penalties optimally. They show that the pointwise limit distribution of their estimate is normal , with rates faster than $n^{1/3}$, but slower than $\sqrt{n}$. Motivated by this approach, we adopt a similar strategy in the general monotone response model that incorporates monotonicity and smoothness constraints *simultaneously*. Pointwise asymptotic normality of the estimator is established, with a rate of convergence that depends on the smoothing parameter. This yields a method of constructing confidence intervals for the regression function at different points of interest. In particular, this gives us a novel way of constructing a smooth estimator of the survival time distribution in the current status model, along with pointwise confidence bands.

Furthermore, we compare this new method of constructing pointwise confidence intervals with the likelihood ratio based method due to Banerjee (2007). This is based on inversion of the likelihood ratio statistic for testing the null hypothesis that the regression function in a monotone response model assumes a fixed value. Numerical comparisons between the smooth estimator and the LRT method show that the former tends to produce sharper pointwise confidence sets on the average, though the coverage is marginally compromised. In view of this discussion, the proposed new method makes a fairly useful case, not least because of the fact that it can be readily applied to different examples

5

with very little methodological adaptation. In Section 2, we characterize the smooth estimator and propose a reasonably simple and fast iterative algorithm to compute the estimator. Section 3 is devoted to the pointwise asymptotic behavior of the estimator. We establish an equivalence between the spline-type estimator and a kernel estimator, which is instrumental in deriving the asymptotic normality. Section 4 presents simulation comparisons for these two methods in different scenarios (binary and Poisson regression) along with an application to a real-life data set involving binary responses, and is followed by references. Some of the relevant technical details are collected in an appendix.

2 The Smooth Estimator

2.1 Characterization and properties

In the monotone response model defined in the introduction, the log-likelihood for the data $(Z_1, X_1), \ldots, (Z_n, X_n)$ with realizations $(z_1, x_1), \ldots, (z_n, x_n)$ becomes

$$\sum_{i=1}^{n} \log p_Z(z_i) + \sum_{i=1}^{n} \log h(x_i) + \sum_{i=1}^{n} \{T(x_i)\psi(z_i) - B \circ \psi(z_i)\}. \qquad (1)$$

Our objective now centers on maximizing (1) subject to the constraint $\psi \nearrow$. Let $z_{(1)} < z_{(2)} < \ldots < z_{(n)}$ denote the ordered values of the z_i's and let $x_{(i)}$ denote the response value corresponding to $z_{(i)}$. Set $T_i = T(x_{(i)})$ and $\psi_i = \psi(z_{(i)})$. From the Kuhn-Tucker conditions, it is easy to show that,

$$\arg \max_{\psi_1 \leq \ldots \leq \psi_n} \sum_{i=1}^{n} [T_i \psi_i - B(\psi_i)] = \arg \min_{\psi_1 \leq \ldots \leq \psi_n} \sum_{i=1}^{n} [T_i - B'(\psi_i)]^2$$

Hence, finding the NPMLE is equivalent to solving the least squares problem: $\sum_{i=1}^{n} [T(x_i) - m(z_i)]^2$ on the set $\mathbb{B} = \{m \nearrow\}$, and computing $\psi = (B')^{-1} \circ m$, transforming back to get the link function.

The Maximum Likelihood Estimator or equivalently the Least Square Estimator is devoid of the desired smoothness and has abrupt jumps with non-removable discontinuities, and hence is often unsuitable for interpretation purposes. Tantiyaswasdikul & Woodroofe (1994) (TW94 from now on) suggested penalizing the least squares criterion using a smoothness penalty. Here, since maximum likelihood estimation is equivalent to least squares regression, we adopt a similar penalization criterion. To this end, we consider the λ-penalized

least squares problem

$$\frac{1}{n}\sum_{i=1}^{n}[T(x_i) - m(z_i)]^2 + \lambda \int_{-\infty}^{\infty} m'(z)^2 dz$$

in $\mathbb{B}' = \{m : m \text{ is absolutely continuous, } m' \text{ exists a.e. and } \geq 0\}$.

This formulation puts equal smoothness penalty to the entire range, even to regions where F_Z barely puts any mass. To alleviate this, we replace Z by its cumulative distribution function F_Z, as this will penalize more for roughness in a densely sampled region. Since F is an unknown function, we work with its estimate $\bar{F}_n$, defined as,

$$\bar{F}_n(z) = \frac{i}{n}\frac{z_{(i+1)} - z}{z_{(i+1)} - z_{(i)}} + \frac{i+1}{n}\frac{z - z_{(i)}}{z_{(i+1)} - z_{(i)}}, \qquad \text{for } z_{(i)} \leq z < z_{(i+1)}.$$

Outside $[z_{(1)}, z_{(n)}]$, it is only a question of definition since this interval will eventually cover the entire support of Z. For convenience, we define it as 1 for all z larger than $z_{(n)}$ and linearly in the stretch $[z_{(1)} - 1/n, z_{(1)}]$, with terminal values 0 and $1/n$ respectively and 0 to the left of $z_{(1)}$. This function is quite close to the actual empirical distribution function F_n, which being discontinuous is unsuitable for our purpose. Thus, $\bar{F}_n$ is continuous, piecewise linear, strictly increasing in $[z_{(1)}, z_{(n)}]$ and non-decreasing throughout, bounded between 0 and 1 and converges to the true CDF of Z. This transformation and the subsequent quantile transformation to get back z_i's is motivated by the large amount of work by E. Parzen on the quantile transformation, described in Parzen (2004).

Since the unpenalized problem requires finding the least squares solution of $\sum_{i=1}^{n}[T(x_i) - m(z_i)]^2 = \sum_{i=1}^{n}[T(x_{(i)}) - m \circ \bar{F}_n^{-1}(i/n)]^2$, setting $\phi = m \circ \overline{F}_n^{-1}$, we seek to solve the minimization problem

$$(1/n)\sum_{i=1}^{n}\left[T(x_{(i)}) - \phi\,(i/n)\right]^2 + \lambda \int_0^1 \phi'(z)^2 dz \qquad (2)$$

in $\{\phi : \phi \text{ is absolutely continuous from } [0,1] \rightarrow \mathbb{R}, \phi' \text{ exists a.e. and } \geq 0\}$ and transform back to get $\hat{m} = \hat{\phi} \circ \bar{F}_n$.

Characterization of the estimator: The estimator which minimizes (2) can be characterized as a solution to a set of recurrence relations. Let $T_i = T(X_{(i)})$ for $i = 1, \ldots, n$ and $G_k = (T_1 + \ldots + T_k)/n$ for $k = 1, \ldots, n$.

Define $\hat{\phi}$ to be a piecewise linear function with $\hat{\phi}(k/n) = \hat{\phi}_k$ (say) and $\hat{\phi}(t) = \hat{\phi}_1$

for $t < 1/n$. For such a function, let $\check{\Phi}(k/n) = \check{\Phi}_k$ for $k = 1, \ldots, n$, where $\check{\Phi}_k = (\hat{\phi}_1 + \ldots + \hat{\phi}_k)/n$. Then, the following holds.

Proposition 1. *A function $\hat{\phi}$ in $\{\phi : [0, 1] \rightarrow \mathbb{R}, \phi' \text{ exists a.e. and } \geq 0\}$ minimizes (2) among all such functions if it satisfies the recurrence equation*

$$\hat{\phi}_k = \hat{\phi}_{k-1} + (\check{\Phi}_{k-1} - G_{k-1})_+/(n\lambda) \tag{3}$$

and the boundary condition $\check{\Phi}_n = G_n$. Moreover, $\min_k nG_k/k \leq \hat{\phi}_1 \leq G_n$.

The result is analogous to Theorem 1 in TW94 and proved similarly. $\square$

Let $\tilde{\phi}$ denote the solution to the minimization problem in (2) with all restrictions on ϕ as before apart from monotonicity, i.e. ϕ' is no longer stipulated to be non–negative. We call $\tilde{\phi}$ the unconstrained minimizer of (2). The estimator $\hat{\phi}$ can be characterized as the projection of $\tilde{\phi}$ onto the set of monotone functions (using the norm $1/n \sum_{i=1}^{n} \phi(i/n)^2 + \lambda \int_0^1 \phi'(t)^2 dt$). This is explained in Mammen & Thomas-Agnan (1999) and Mammen et al (2001). Here, we devise an algorithm to compute the estimator directly, rather than following the two-step procedure.

An algorithm to compute the estimator: To compute $\hat{\phi}$ we will set up a binary search algorithm to satisfy both (3) and the boundary condition. The two bounds on $\hat{\phi}_1$ in Proposition 1 are required to initiate the iteration. The iteration converges due to the contraction of the binary search algorithm and is developed as follows.

Step 1 Let $A = \min_k nG_k/k$, $B = G_n$, and ϵ some pre-assigned tolerance limit.
Step 2 $\hat{\phi}_1 = (A + B)/2$, $\hat{\phi}_k = \hat{\phi}_{k-1} + (\check{\Phi}_{k-1} - G_{k-1})_+/(n\lambda)$, for $k = 2, \ldots, n$,
  $\check{\Phi}_k = (\hat{\phi}_1 + \ldots + \hat{\phi}_k)/n$, for $k = 1, \ldots, n$,
Step 3 If $|\check{\Phi}_n - G_n| \leq \epsilon$, stop.
  If $\check{\Phi}_n > G_n + \epsilon$, set $B = \hat{\phi}_1$. Repeat from Step 2.
  If $\check{\Phi}_n < G_n - \epsilon$, set $A = \hat{\phi}_1$. Repeat from Step 2.

This algorithm leads to an estimator which satisfies the criteria in Proposition 1, hence solving the transformed minimization problem (2).

2.2 Discussion of the procedure

The computational simplicity of the algorithm is highlighted by the fact that it takes only an ordered set of response values and estimates the link function using an iterative algorithm. The covariate values are not required in every

iteration. Rather, they are only used to order the responses, and calibrate the final estimate. The run time is much lower compared to the PAVA algorithm and can be implemented easily using any standard software package. The pre-assigned tolerance limit should be independent of the covariates, but chosen on the basis of the magnitude and precision of the response.

Here we have used an L_2 penalty on the first derivative of the estimator, which leads us to a piecewise linear function, analogous to the usual linear spline. If we use an L_1 penalty instead, it will make the integral penalty a scaled version of the difference between the maximum and the minimum of the estimator, and the minimizer will be attained at an estimator which is closer to a constant. It will fail to penalize for the roughness of the estimator at intermediate points. On the other hand, if we base our penalization criterion on the second derivative, it will lead us to a cubic spline, and the monotonicity constraint will be difficult to impose. For example, a cubic spline need not be monotone on its range to ensure that it preserves the ordering when restricted to the design points. The problem of constraining the penalized least squares criterion is discussed in Mammen & Thomas-Agnan (1999) and Mammen et al (2001), with emphasis on different smoothness penalties and constraints.

Finally, the quantile transformation makes the estimator invariant under any monotone covariate transformation, and therefore robust to the covariate density. This is a property shared by the PAVA estimator. In fact, the covariate distribution does not need to be estimated to construct confidence intervals. Finally, we emphasize the fact that the method tries to balance between monotonicity and smoothing, instead of prioritizing one over the other.

3 Large Sample Properties

In this Section, we discuss the pointwise asymptotic behavior of the estimator. Henceforth, we make the following assumptions.

(a) The covariate Z has a compact support.
(b) On the support of Z, the function ψ (and consequently m) is twice continuously differentiable with bounded derivatives and so is the distribution function F.
(c) The covariate density f does not vanish anywhere on the support of Z.

Let z_0 be any interior point in the support of Z. The target is to estimate $\psi(z_0)$ or $m(z_0)$, the influence of the covariate $Z = z_0$ on the distribution of X.

Using $\hat{m}(z_0) = \hat{\phi} \circ \bar{F}_n(z_0)$, the estimate of $\psi(z_0)$ is,

$$\hat{\psi}(z_0) = (B')^{-1} \circ \hat{\phi} \circ \bar{F}_n(z_0).$$

We start with a central proposition that establishes an asymptotic equivalence between $\hat{m}(z_0) - m(z_0)$ and a kernel estimator. The proof is fairly technical and we defer it to the appendix for the convenience of the reader.

Proposition 2. *Define $U_i = T(X_i) - m(Z_i)$ for $i = 1, \ldots, n$, and $U_{(i)} = T(X_{(i)}) - m(Z_{(i)})$, for $i = 1, \ldots, n$. Set $\beta = \lambda^{-1/2}$. Then, under our assumptions and for $\lambda \to 0$ and $n\lambda \to \infty$,*

$$\hat{m}(z_0) = m(z_0) + \beta/(2n) \sum_{i=1}^{n} e^{-\beta|F(z_0) - \frac{i}{n}|} U_{(i)} + \beta O_P(n^{-\frac{2}{3}} \log n)$$
$$+ \beta O_P(e^{-\beta F(z_0)(1 - F(z_0))}) + O_P(\lambda) + O_P(1/\sqrt{n}).$$

The main result regarding the asymptotic distribution can be stated as follows.

Theorem 1. *If $\lambda n^{2/5} \to 0$, but $\lambda n^{2/3} \to \infty$, then,*

$$\sqrt{(n/\beta)} (\hat{m}(z_0) - m(z_0)) \Rightarrow N\left(0, (B'' \circ \psi(z_0))/4\right).$$

The following corollary is direct from the theorem.

Corollary 1. *Under the same conditions on λ as in Theorem 1,*

$$\sqrt{(n/\beta)} \left(\hat{\psi}(z_0)) - \psi(z_0)\right) \Rightarrow N\left(0, \kappa(z_0)\right).$$

where, $\kappa(z_0) = 1/(4B'' \circ \psi(z_0))$.

Proof of Theorem 1: Conditional on the sequence $(Z_1, Z_2, \ldots)$, the random variables $U_1, U_2, \ldots$, are independent random variables, with common mean 0 and variances $B'' \circ \psi(Z_i)$ respectively. Now, using Proposition 2, we have (under the assumptions $\lambda n^{2/5} \to 0$, but $\lambda n^{2/3} \to \infty$) that

$$\sqrt{(n/\beta)} (\hat{m}(z_0) - m(z_0)) = \sqrt{(\beta/4n)} \sum_{i=1}^{n} e^{-\beta|F(z_0) - i/n|} U_{(i)} + o_P(1)$$
$$= \frac{1}{2} \sum_{i=1}^{n} a_{nk_i} U_i \text{ (say)}$$

where $k_i = \text{rank}(Z_i)$ (among $Z_1, Z_2, \ldots, Z_n$) and $a_{nk} = \sqrt{(\beta/n)} e^{-\beta|F(z_0) - k/n|}$. We will study the limit distribution of $\sum_{i=1}^{n} a_{nk_i} U_i$ conditionally on the se-

10

quence $(Z_1, Z_2, Z_3, \ldots)$. Let, $\tilde{a}_{nk} = e^{-\beta|F(z_0)-k/n|}$, $m_n^2 = \sum_{i=1}^{n} \tilde{a}_{nk_i}^2 B''(\psi(Z_i))$ and

$$\xi_n = \frac{\sum_{i=1}^{n} a_{nk_i} U_i}{\sqrt{\mathrm{Var}(\sum_{i=1}^{n} a_{nk_i} U_i)}} = \frac{\sum_{i=1}^{n} \tilde{a}_{nk_i} U_i}{\sqrt{\sum_{i=1}^{n} \tilde{a}_{nk_i}^2 B''(\psi(Z_i))}} = \frac{\sum_{i=1}^{n} \tilde{a}_{nk_i} U_i}{m_n}.$$

Now, conditionally on the sequence $(Z_1, Z_2, Z_3, \ldots)$, ξ_n converges in distribution to $N(0,1)$, provided we can verify the Lindeberg condition, which in this case is $Q_{n,\epsilon} \to 0$, where

$$Q_{n,\epsilon} = \frac{1}{m_n^2} \sum_{i=1}^{n} E_{Z_i} \left[\tilde{a}_{nk_i}^2 U_i^2 \, 1\{| \, \tilde{a}_{nk_i} U_i \, | \geq \epsilon \, m_n\} \right].$$

In the above display E_{Z_i} denotes the conditional expectation given Z_i. Using the fact that the Z_i's all belong to a compact interval H on which the common distribution of the Z_i's is supported, and the fact that $0 < \inf_{z \in H} B''(\psi(z)) \leq \sup_{z \in H} B''(\psi(z)) < \infty$ we readily conclude that $Q_{n,\epsilon} \to 0$ if

$$\tilde{Q}_{n,\epsilon} = \frac{1}{\sum_{i=1}^{n} \tilde{a}_{nk_i}^2} \sum_{i=1}^{n} \tilde{a}_{nk_i}^2 E_{Z_i} \left[U_i^2 \, 1 \left\{ | \, U_i \, | \geq \frac{\epsilon}{\max_{1 \leq i \leq n} \tilde{a}_{nk_i}/m_n} \right\} \right] \to 0.$$

Let $\eta > 0$ be preassigned. Now, conditionally on the sequence $\{Z_i\}_{i=1}^{\infty}$, the sequence $\{U_i\}_{i=1}^{\infty}$ is uniformly integrable (since the sequence of conditional variances is uniformly bounded). Also, $\max_{1 \leq i \leq n} \tilde{a}_{nk_i}/m_n$ converges to 0 as n increases; hence for all sufficiently large n, we can ensure that $\epsilon/(\max_{1 \leq i \leq n} \tilde{a}_{nk_i}/m_n)$ is greater than u, where u is such that $\sup_{1 \leq i < \infty} E[U_i^2 \, 1(| \, U_i \, | \geq u) \mid Z_i] < \eta$. It follows, that for all sufficiently large n, $\tilde{Q}_{n,\epsilon} < \eta$. The fact that $\max_{1 \leq i \leq n} \tilde{a}_{nk_i}/m_n$ converges to 0 is a consequence of the facts that

$$\max_{1 \leq i \leq n} e^{-\beta|F(z_0)-k_i/n|} \equiv \max_{1 \leq i \leq n} e^{-\beta|F(z_0)-i/n|} \to 1$$

and that

$$\sum_{i=1}^{n} e^{-2\beta|F(z_0)-k_i/n|} \equiv \sum_{i=1}^{n} e^{-2\beta|F(z_0)-i/n|} \to \infty.$$

Hence, the Lindeberg condition is satisfied.

Thus, conditionally on almost every sequence $(Z_1, Z_2, \ldots)$

$$\frac{\sum_{i=1}^{n} \sqrt{\frac{\beta}{n}} e^{-\beta|F(z_0)-\frac{k_i}{n}|} U_i}{\sqrt{\frac{\beta}{n} \sum_{i=1}^{n} e^{-2\beta|F(z_0)-\frac{k_i}{n}|} B'' \circ \psi(Z_i)}} = \frac{\sum_{i=1}^{n} \sqrt{\frac{\beta}{n}} e^{-\beta|F(z_0)-\frac{i}{n}|} U_{(i)}}{\sqrt{\frac{\beta}{n} \sum_{i=1}^{n} e^{-2\beta|F(z_0)-\frac{i}{n}|} B'' \circ \psi(Z_{(i)})}} \Rightarrow N(0,1).$$

Since the limit is independent of the sequence $(Z_1, Z_2, \ldots)$ being conditioned on, it follows that unconditionally as well, the above display holds. Next, it is

not difficult to show that, since $\beta^2/n \to 0$, we have,

$$\frac{\beta}{n} \sum_{i=1}^{n} e^{-2\beta|F(z_0)-\frac{i}{n}|} B'' \circ \psi(Z_{(i)}) \to_p B'' \circ \psi(z_0).$$

Therefore, the denominator in the above display converges in probability, and using this in conjunction with Slutsky's theorem yields that

$$\sqrt{\frac{\beta}{n}} \frac{\sum_{i=1}^{n} e^{-\beta|F(z_0)-\frac{i}{n}|} U_{(i)}}{\sqrt{B'' \circ \psi(z_0)}} \Rightarrow N(0,1).$$

Hence,

$$\frac{1}{2} \sqrt{\frac{\beta}{n}} \sum_{i=1}^{n} e^{-\beta|F(z_0)-\frac{i}{n}|} U_{(i)} \Rightarrow N\left(0, \frac{B'' \circ \psi(z_0)}{4}\right)$$

The theorem follows. $\square$

Remark 1: As is evident from Proposition 2, the quantity $\hat{m}(z_0) - m(z_0)$ is asymptotically equivalent to a kernel estimator with a Laplacian kernel and bandwidth $1/\beta$. In fact, the estimate $\tilde{m} \equiv \tilde{\phi} \circ \overline{F}_n$ based on the unconstrained minimizer of (2) will exhibit the same behavior. This is reminiscent of the results in Silverman (1984) and Nychka (1995) where we observe a similar equivalence between smoothing splines and kernel estimators. In fact, under the assumptions of Theorem 1, with probability tending to one, $\tilde{m}$ is monotone and hence equals the constrained estimator $\hat{m}$. As a result, the asymptotic performances of the constrained and the unconstrained estimators coincide and are driven entirely by the smoothness assumptions. Of coure, imposition of the monotonicity constraint (when it is true) will lead to improved finite sample performance.

Remark 2: The rate of convergence, $\sqrt{\beta/n}$, depends directly on $\beta = \lambda^{-1/2}$ which lies within $n^{1/3}$ and $n^{1/5}$, by the conditions of Theorem 1. As a result, the best achievable rate of convergence is $n^{-2/5}$, as expected under the assumption that the true function has bounded second derivative. To achieve the best rate of $n^{-2/5}$, however, λ needs to be of order $n^{-2/5}$. This leads to a non-negligible bias term in the asymptotics. The asymptotic normality will still be valid, but with a non-central mean.

Remark 3: As an artifact of our method, we obtain an asymptotic distribution for the estimate that does not rely on the underlying covariate density. This stems from the use of the quantile transformation. Since the limit distribution is pivotal this buys us significant mileage over other methods as far as

construction of confidence sets is concerned since nuisance parameters do not need to be estimated.

3.1 Construction of confidence intervals:

Theorem 1 immediately provides a way to construct an asymptotic level $1 - \alpha$ confidence interval for $m(z_0)$. Clearly,

$$P\left[-z_{\alpha/2} \leq \sqrt{\frac{n}{\beta}}\, \frac{2(\hat{m}(z_0) - m(z_0))}{\sqrt{B''(\psi(z_0))}} \leq z_{\alpha/2} \right] \to 1 - \alpha\,,$$

with increasing n. Here $z_{\alpha/2}$ is the upper $\alpha/2$'th quantile of the standard normal distribution. Now, $B''(\psi(z_0))$ is estimated consistently by $B''(\hat{\psi}(z_0))$. It then follows that

$$\left[\hat{m}(z_0) - \sqrt{\frac{\beta}{n}}\, \frac{\sqrt{B''(\hat{\psi}(z_0))}\, z_{\alpha/2}}{2},\, \hat{m}(z_0) + \sqrt{\frac{\beta}{n}}\, \frac{\sqrt{B''(\hat{\psi}(z_0))}\, z_{\alpha/2}}{2} \right]$$

is an approximate asymptotic level $1 - \alpha$ confidence set for $m(z_0)$. We refer to this, subsequently, as the two–sided confidence interval. Transforming the above confidence interval by $(B')^{-1}$ gives a corresponding confidence interval for $\psi(z_0)$. We can also use Corollary 1 to construct a symmetric confidence interval, though we prefer the former for accuracy.

Denote the inverse function of B' by H. An alternative way of constructing a confidence set for $m(z_0)$ is to note that, under the null hypothesis $H_0 : m(z_0) = \eta_0$, the quantity

$$\frac{2\sqrt{\frac{n}{\beta}}\, (\hat{m}(z_0) - \eta_0)}{\sqrt{B''(H(\eta_0))}} \to_d N(0, 1)\,,$$

and is therefore an asymptotic pivot. An asymptotic level $1 - \alpha$ confidence set for $m(z_0)$ is then given by:

$$\left\{ \eta : \left| \frac{2\sqrt{\frac{n}{\beta}}\, (\hat{m}(z_0) - \eta)}{\sqrt{B''(H(\eta))}} \right| \leq z_{\alpha/2} \right\}\,,$$

where $z_{\alpha/2}$ is the upper $\alpha/2$'th quantile of the standard normal distribution. Thus, a confidence set for the regression function (and equivalently, the link function, via back–transformation) can be obtained without the need to estimate nuisance parameters. This is a virtue that the proposed method shares

with the likelihood ratio method to be discussed in the next section. This kind of confidence sets will be referred as pivot-based confidence sets. For Binomial and Poisson regression, these reduce to proper intervals with the estimator being an interior point.

4 Simulation comparisons and real data examples

4.1 Binomial and Poisson simulations

We first study the relative behavior of confidence sets for the regression function in a monotone response model obtained using the three different methods explained above: (i) two-sided intervals using the limit distribution of the smooth estimator, (ii) pivot based confidence sets using the limit distribution of the smooth estimator, and (iii) likelihood ratio based confidence sets obtained via inversion of the likelihood ratio statistic (the reader is encouraged to refer to Banerjee (2007) for a detailed discussion). The Binomial and Poisson regression models (Examples (b) and (c) respectively in Section 1) are considered. It can be shown, that in these two cases, confidence sets of type (ii) reduce to proper intervals with the estimator being an interior point. Confidence sets of type (iii) are necessarily intervals (for the monotone response models considered in this paper).

(i) **Bernoulli simulations:** We chose sample sizes $n = 100, 200, \ldots, 1000$. For each n, we generated 1000 replicates from the distribution of $\{X_i, Z_i\}_{i=1}^n$, where the (X_i, Z_i)'s are i.i.d random variables, with Z_i following a standard double exponential distribution truncated in the interval $[-100, 100]$ and X_i given Z_i following a Bernoulli distribution, with link function $\psi(z) = e^z$. For each replicate, we computed 95% confidence intervals for $m(z_0) = P(X_i = 1 \mid Z_i = z_0)$ (for $z_0 = 0.4$) using methods (i),(ii) and (iii) (with $\lambda = n^{-.65}$). The lengths of all three types of C.I's were then averaged over the 1000 replicates and the empirical coverage of the three different types of intervals were computed.

(ii) **Poisson simulations:** The same selection of sample sizes was chosen, as for the Bernoulli case. 1000 replicates from the distribution of $\{X_i, Z_i\}_{i=1}^n$ were generated for each n. As in the Bernoulli case, the (X_i, Z_i)'s are i.i.d., Z_i has the same distribution as in (i), and given Z_i, X_i is distributed as a Poisson random variable with link function $2 \arctan(z)/\pi$. 95% confidence intervals for $m(z_0) = E(X_i \mid Z_i = z_0)$ (for $z_0 = -0.4$) using the three different methods were computed for each of the 1000 replicates (as in the Bernoulli case, $\lambda = n^{-.65}$ for the smooth estimators), and as before the

14

average length of the intervals and the empirical coverage were computed.

TABLE 1 IS ABOUT HERE

TABLE 2 IS ABOUT HERE

Tables 1 and 2 report the coverage probabilities and the interval lengths from all the simulations respectively. It transpires that the smooth estimator based intervals (both the two-sided and the pivot-based confidence intervals) behave less conservatively than those produced by the likelihood ratio based method. They produce systematically narrower intervals, but the small sample coverage probability is marginally less (though fairly close to the nominal). Moreover, the pivot-based method is seen to produce shorter intervals with lower coverage than the direct two-sided C.I. This comparison is also shown graphically in Figure 1.

FIGURE 1 IS ABOUT HERE

4.2   Analysis of Austrian Rubella data

To illustrate the method in a real example, we select a data set concerning 230 Austrian males with their exact age as covariate. The goal is to investigate the age-specific immunization intensity of Rubella sero-prevalence. For each person, the current immunization status was tested at the Institute of Virology, Vienna. The goal is to estimate the distribution of the time to infection by Rubella in that population. Though the exact time of immunization is unobserved, we assume that the effect of immunization is lifelong. Hence, we can treat this as a current status study, assuming that the age of immunization and the current age are independent. For details, see Keiding et al (1996).

Let T_i denote the current time/age of the i-th individual, and W_i be his unobserved immunization age. The status we observe at the current age is $X_i = 1_{W_i \leq T_i}$. The regression function is given by $P(X_i = 1|T_i) = F(T_i)$, where F is the cumulative distribution function of X. This is our function of interest. We compare the performance of the three methods discussed in the last Section for ages 1-28. For a detailed discussion of the LRT-based method in the context of the Rubella data set, we refer the reader to Section 4 of Banerjee and Wellner (2005). The intervals are shown in Table 3, for selected ages. It is clear from the table that the smooth estimator produces shorter confidence intervals in general.

TABLE 3 IS ABOUT HERE

15

As the simulations show that only a mild compromise is made in terms of the coverage, the effect of smoothing is rather encouraging. We plot the smooth estimator and the pointwise confidence intervals, for both direct and pivot-based methods in Figures 2. Interestingly enough, for this particular data set, the pivot based confidence intervals tend to be larger than the two-sided ones (though they remain substantially shorter than the likelihood ratio based ones).

FIGURE 2 IS ABOUT HERE

Conclusion

The paper describes an approach to obtain a smooth estimate of a monotone regression function in a generalized exponential family regression setting. Since both discrete and continuous response variables as well as general mean–variance relationships can be incorporated in this set–up, the proposed method has wide applicability. A key feature of the proposed method is the use of the quantile transformation that reduces the optimization problem to the covariate space $[0, 1]$ and yields an estimator of the regression function whose asymptotic distribution is free from nuisance parameters. This facilitates the construction of confidence sets for the regression function significantly. Unlike the maximum likelihood estimator of the regression which is piecewise constant and whose distribution is non–standard (and is given by Chernoff's distribution, see Groeneboom & Wellner (2001) for a reference), the smooth estimator is asymptotically normal and has a (faster) rate of convergence that depends on the smoothing parameter.

There are several natural directions in which the work in this paper can be extended. In certain applications one may have more information on the shape of the function. For example, production functions in economics are known to be increasing and concave (on an appropriate scale). It would be interesting to extend the approach of this paper to construct smooth estimates satisfying both monotonicity and concavity/convexity constraints. This would entail solving the minimization problem in (2) with an added constraint on the second derivative of ϕ. Yet another direction is a semiparametric extension where in addition to Z, there is another set of covariates W that influences the response and whose effect is captured by a finite–dimensional regression parameter. Thus, for example, one can have $E(X \mid (Z, W)) = B'(\psi(Z) + \beta^T W)$ and inference needs to be made on both β and ψ. This will be an interesting direction from the point of view of applications and is left as a topic for future research.

16

Acknowledgments

A. Appendix

The proof of Proposition 2 is fairly involved, and we present it step-by-step for the sake of clarity. The following proposition forms the main basis of the proof.

Proposition 3. *Let $\phi_{(n)} = m \circ \bar{F}_n^{-1}$. For each n, this is a random function that depends only on the Z-s (and not on the X-s). For $0 < t < 1$ and any λ, $\hat{\phi}(t)$ can be approximated as,*

$$\hat{\phi}(t) = \phi_{(n)}(t) + O_P(\lambda) + \frac{\beta}{2n} \sum_{i=1}^{n} e^{-\beta|t - \frac{i}{n}|} U_{(i)} + \beta O_P(n^{-\frac{2}{3}} \log n) + \beta O_P(e^{-\beta t(1-t)}).$$

We need some introductory results to establish Proposition 3.

A.1. Preliminaries

As the estimator depends on the smoothing parameter λ, we will write it as a function of λ for the time being. Moreover, let $c_\lambda = \hat{\phi}_1(\lambda)$. Let $\vec{\phi}$ be the isotonic regression of $(T_1, \ldots, T_n)$ with weight vector $(1, 1, \ldots, 1)$. Then,

$$\vec{\phi}_k = \max_{i \leq k} \min_{j \geq k} \frac{T_i + \ldots + T_j}{j - i + 1}.$$

Also, define $G(t) = \int_0^t g(s)\,\omega(ds)$, where ω is the measure on $[0, 1]$ that puts mass $1/n$ at each of the knots i/n, for $1 \leq i \leq n$, and g is a function with $g(k/n) = T_k$. Let $\tilde{G}$ be the greatest convex minorant (GCM) of the points $\{(k/n, G_k)\}_{k=0}^n$, with knot values $\tilde{G}_k$. Note that $\vec{\phi}_k$ is the left–derivative of $\tilde{G}$, computed at the point k/n. For such a function, define the functions $\check{\Phi}(t)$ and $\hat{\Phi}(t)$ as $\int_0^t \hat{\phi}(s)\omega(ds)$ and $\int_0^t \hat{\phi}(s)ds$ respectively.

Then,

A: c_λ is non-decreasing in λ.

B: $\check{\Phi}_k(\lambda)$ is non-decreasing in λ.

C: For all $k = 1, \ldots, n$, and for all λ, $\check{\Phi}_k(\lambda) \geq \tilde{G}_k$.

D: For all λ and t between 0 and 1, $\vec{\phi}_1 \leq \hat{\phi}(\lambda, t) \leq \vec{\phi}_n$.

E: Both $\vec{\phi}$ and $\hat{\phi}$ are bounded uniformly with probability 1.
Proof:

A: For $c \in \mathbb{R}$, define $\phi_0(\lambda, c) = c$ for all λ, and $\phi_k(\lambda, c) = \phi_{k-1}(\lambda, c) + (\Phi_{k-1}(\lambda, c) - G_{k-1})_+/(n\lambda)$, $\Phi_k(\lambda, c) = (\phi_1(\lambda, c) + \ldots + \phi_k(\lambda, c))/n$. Then, both ϕ_k and Φ_k are non-increasing in λ, by induction. Let $\lambda < \lambda'$. Clearly, $\Phi_n(\lambda, c_\lambda) = G_n \geq \Phi_n(\lambda', c_\lambda)$, implies $c_\lambda \leq c_{\lambda'}$. $\square$

B: Let $\lambda < \lambda'$. Clearly, $\check{\Phi}_1(\lambda) = c_\lambda/n \leq c_{\lambda'}/n = \check{\Phi}_1(\lambda')$ and $\check{\Phi}_n(\lambda) = \check{\Phi}_n(\lambda') = G_n$. Suppose, $\check{\Phi}_j(\lambda) > \check{\Phi}_j(\lambda')$ for some j between 2 and $n-1$, and let k be the smallest such j. Then

$$\hat{\phi}_k(\lambda) > \hat{\phi}_k(\lambda').$$

Therefore,

$$\hat{\phi}_{k+1}(\lambda) = \hat{\phi}_k(\lambda) + \frac{1}{\lambda}(\check{\Phi}_k(\lambda) - G_k)_+(t_{k+1} - t_k)$$
$$> \hat{\phi}_k(\lambda') + \frac{1}{\lambda'}(\check{\Phi}_k(\lambda') - G_k)_+(t_{k+1} - t_k)$$
$$= \hat{\phi}_{k+1}(\lambda')$$

and $\check{\Phi}_{k+1}(\lambda) > \check{\Phi}_{k+1}(\lambda')$. Proceeding like this, we get, $\check{\Phi}_n(\lambda) > \check{\Phi}_n(\lambda')$, a contradiction to our assertion. $\square$

C: Clearly, $\check{\Phi}_1(\lambda) = c_\lambda/n \geq \min_{1 \leq k \leq n} G_k/k = \tilde{G}_1$ and $\check{\Phi}_n(\lambda) = G_n = \tilde{G}_n$. As above, suppose, $\check{\Phi}_j(\lambda) < \tilde{G}_j$ for some j between 2 and $n-1$, and let k be the smallest such j. Then, as $\tilde{G}_k \leq G_k$ for all k,

$$\hat{\phi}_{k+1}(\lambda) = \hat{\phi}_k(\lambda) + \frac{1}{\lambda}(\check{\Phi}_k(\lambda) - G_k)_+(t_{k+1} - t_k)$$
$$= \hat{\phi}_k(\lambda)$$
$$= n[\check{\Phi}_k(\lambda) - \check{\Phi}_{k-1}(\lambda)]$$
$$< n[\tilde{G}_k - \tilde{G}_{k-1}]$$
$$\leq n[\tilde{G}_{k+1} - \tilde{G}_k] \qquad \text{(as } \tilde{G} \text{ is convex).}$$

Therefore,

$$\check{\Phi}_{k+1}(\lambda) = \check{\Phi}_k(\lambda) + \frac{\hat{\phi}_{k+1}(\lambda)}{n}$$
$$< \tilde{G}_{k+1}.$$

Consequently, $\hat{\phi}_{k+2}(\lambda) = \hat{\phi}_{k+1}(\lambda) = \hat{\phi}_k(\lambda) \leq n[\tilde{G}_{k+1} - \tilde{G}_k] \leq n[\tilde{G}_{k+2} - \tilde{G}_{k+1}]$ (as above) and $\check{\Phi}_{k+2}(\lambda) < \tilde{G}_{k+2}$ as well. Proceeding in the same line, we conclude, $\check{\Phi}_n(\lambda) < \tilde{G}_n$, a contradiction. $\square$

18

D: Clearly, for all t and λ, $\hat{\phi}(\lambda, t) \geq \hat{\phi}_1(\lambda) = c_\lambda \geq \min_k nG_k/k = \vec{\phi}(t_1)$, establishing the first inequality. The other inequality follows by symmetry. $\square$

E:

$$\vec{\phi}_1 = \min_{i \geq 1} \frac{T_1 + \ldots + T_i}{i}$$

$$\geq \inf B' + \min_{i \geq 1} \frac{U_{(1)} + \ldots + U_{(i)}}{i}$$

However, $(U_{(1)} + \ldots + U_{(n)})/n = (U_1 + \ldots + U_n)/n \to 0$ w.p.1, using Strong Law of Large Numbers. Hence, $\min_{i \geq 1}(U_{(1)} + \ldots + U_{(i)})/i > -\infty$ w.p. 1. Consequently, $\vec{\phi}_1 > -\infty$ w.p. 1. Using symmetry we get, $\vec{\phi}_n < \infty$ w.p. 1 as well. The lemma follows. The boundedness of $\hat{\phi}$ follows from its being sandwiched by the two extreme values of $\vec{\phi}$ (by Property D of the estimator). $\square$

The following two results from analysis will be useful for our purpose.

Result 1: *The differential equation $\lambda\Phi''(t) = \Phi(t) - H(t)$ for $0 \leq t \leq 1$, with boundary conditions $\Phi(0) = 0$ and $\Phi(1) = A$ has the unique solution*

$$\Phi(t) = c_0(\beta)e^{-\beta t} + c_1(\beta)e^{-\beta(1-t)} + \frac{1}{2}\int_0^1 \beta e^{-\beta|t-s|}H(s)ds$$

where, $\beta = \lambda^{-1/2}$, and c_0, c_1 are two bounded functions, provided H itself is bounded.

To simplify notation, we define the kernel K_λ on $[0,1]^2$ as

$$K_\lambda(s,t) = \frac{1}{2}\beta e^{-\beta|t-s|}$$

and the operator $\mathcal{K}_\lambda$ on the set of L^1 functions on $[0,1]$ as,

$$\mathcal{K}_\lambda l(t) = \int_0^1 K_\lambda(s,t)l(s)ds.$$

Result 2: *If h is twice continuously differentiable at $t \in (0,1)$, then $\mathcal{K}_\lambda h(t) = h(t) + \lambda h''(t) + o(\lambda)$ as $\lambda \to 0$.*
The following Proposition helps us to approximate the estimator with a solution to a differential equation.

Proposition 4.

$$\|\lambda\hat{\Phi}'' - (\hat{\Phi} - \tilde{G})\| = O_P(n^{-\frac{2}{3}}(\log n)^{\frac{2}{3}}).$$

19

Before embarking on the proof, we start with a lemma.

Lemma 1.
$$(\check{\Phi}(\lambda, t) - \tilde{G}(t))_- = O_p(1/n), \tag{4}$$
where x_- denotes the negative part of x.

Proof: For $k/n \le t < (k+1)/n$, $\check{\Phi}(\lambda, t) = \check{\Phi}_k(\lambda)$, and we have

$$\check{\Phi}(\lambda, t) - \tilde{G}(t) = \check{\Phi}_k(\lambda) - \tilde{G}_k + \tilde{G}_k - \tilde{G}(t).$$

Since, $\check{\Phi}_k(\lambda) - \tilde{G}_k \ge 0$, by Property C, it follows that

$$(\check{\Phi}(\lambda, t) - \tilde{G}(t))_- \le | \tilde{G}_k - \tilde{G}(t) | \le |\tilde{G}_k - \tilde{G}_{k+1}|,$$

since $\tilde{G}(t)$ lies between $\tilde{G}_k$ and $\tilde{G}_{k+1}$. Clearly, (4) holds, since $|\tilde{G}_k - \tilde{G}_{k+1}| = \vec{\phi}_k/n$, and $\vec{\phi}$ is bounded uniformly by Property E. $\Box$

We also need to state a result establishing the proximity between the cusum diagram G and its GCM $\tilde{G}$. The following result will be referred as the Kiefer-Wolfowitz phenomenon, as a similar result was first established by Kiefer & Wolfowitz (1976) for monotone density estimation.

Result 3: *We have ,*

$$\|G - \tilde{G}\| = O_P(n^{-2/3}(\log n)^{2/3}).$$

This can be proved by appropriately modifying the arguments in Pal & Woodroofe (2004). The proof is detailed in Pal (2006), Chapter 4.5. This is where we require assumption (c) to hold. $\Box$

Proof of Proposition 4: From (3), we get:

$$n(\hat{\phi}_k - \hat{\phi}_{k-1})\lambda = \left(\check{\Phi}_{k-1} - G_{k-1}\right)_+.$$

We note that on the interval $[(k-1)/n, k/n)$, the function $\check{\Phi} - G$ is constant, and assumes the value $\check{\Phi}_{k-1} - G_{k-1}$. Furthermore, using the piecewise linearity of $\hat{\phi}$, we get that $n(\hat{\phi}_k - \hat{\phi}_{k-1})$ is the constant value of the right–derivative of $\hat{\phi}$ on $[(k-1)/n, k/n)$. Interpreting $\hat{\Phi}''$ as a right derivative, we therefore conclude that:

$$\lambda \hat{\Phi}'' = \left(\check{\Phi} - G\right)_+. \tag{5}$$

Next, we observe that, $\|\hat{\Phi} - \check{\Phi}\| \le \|\hat{\phi}\|/2n = O_P(1/n)$. Using this and equation (5), we get

20

$$\|\lambda\hat{\Phi}'' - (\hat{\Phi} - \tilde{G})\| = \|(\check{\Phi} - G)_+ - (\hat{\Phi} - \tilde{G})\|$$
$$= \|(\check{\Phi} - G)_+ - (\hat{\Phi} - \tilde{G}) - (\hat{\Phi} - \check{\Phi})\|$$
$$\leq \|(\check{\Phi} - G)_+ - (\check{\Phi} - \tilde{G})_+\| + \|(\hat{\Phi} - \check{\Phi})\| + O_p(1/n)$$
$$\leq \|\tilde{G} - G\| + O_p(1/n)$$
$$= O_P(n^{-\frac{2}{3}}(\log n)^{\frac{2}{3}}).$$

In the above sequence of equations, we use Lemma 1 in passing from Step 2 to Step 3, and the Kiefer-Wolfowitz phenomenon (Result 3) in the last step.
□

A.2 Proof of Proposition 3

We define,

$$\bar{\Phi}_{(n)}(x) = \int_0^x \phi_{(n)}(t)\omega(dt),$$
$$\Phi_{(n)}(x) = \int_0^x \phi_{(n)}(t)dt$$

Let, $R = (\hat{\Phi} - \tilde{G}) - \lambda\hat{\Phi}''$. Then, $\|R\|_\infty = O_P(n^{-\frac{2}{3}}(\log n)^{\frac{2}{3}})$ and $\hat{\Phi}$ satisfies,

$$\lambda\hat{\Phi}'' - \hat{\Phi} = -\tilde{G} - R = -H \text{ (say)}.$$

Hence, using the uniqueness of the solution of Result 1,

$$\hat{\Phi}(t) = \int_0^1 K_\lambda(t,s)H(s)ds + C_0(\beta)e^{-\beta t} + C_1(\beta)e^{\beta(t-1)}$$
$$= \int_0^1 K_\lambda(t,s)\tilde{G}(s)ds + \int_0^1 K_\lambda(t,s)R(s)ds + C_0(\beta)e^{-\beta t} + C_1(\beta)e^{\beta(t-1)},$$

Consequently, the solution $\hat{\Phi}$ can be written as,

$$\hat{\Phi}(t) = \mathcal{K}_\lambda\tilde{G}(t) + \mathcal{K}_\lambda R(t) + C_0(\beta)e^{-\beta t} + C_1(\beta)e^{-\beta(1-t)} \qquad (6)$$

where C_0 and C_1 are stochastically bounded variables, and $\|R\| = O_p(\log n/n)^{2/3}$. We observe that $\frac{\partial}{\partial t}K_\lambda(s,t) = -\frac{\partial}{\partial s}K_\lambda(s,t)$. Now, differentiating both sides of Equation (6),

$$\hat{\phi}(t) = \hat{\Phi}'(t)$$
$$= \int_0^1 \frac{\partial}{\partial t} K_\lambda(s,t)[\tilde{G}(s) + R(s)]ds + \beta\{C_1(\beta)e^{-\beta(1-t)} - C_0(\beta)e^{-\beta t}\}$$
$$= \int_0^1 \frac{\partial}{\partial t} K_\lambda(s,t)\Phi_{(n)}(s)ds + \int_0^1 \frac{\partial}{\partial t} K_\lambda(s,t)[G(s) - \bar{\Phi}_{(n)}(s)]ds$$
$$+ \beta\{C_1(\beta)e^{-\beta(1-t)} - C_0(\beta)e^{-\beta t}\} + \int_0^1 \frac{\partial}{\partial t} K_\lambda(s,t)[\bar{\Phi}_{(n)} - G + \tilde{G} + R - \Phi_{(n)}](s)ds$$
$$= A_n(t) + B_n(t) + C_n(t) + D_n(t) \text{ (say)}$$

Since, $\int_0^1 \frac{\partial}{\partial t}|K_\lambda(s,t)| \leq \beta$ and $\|\bar{\Phi}_{(n)} - G + \tilde{G} + R - \Phi_{(n)}\| = O_P(n^{-\frac{2}{3}}\log n)$, we have $D_n(t) = \beta O_P(n^{-\frac{2}{3}}\log n)$. Also, that $C_n(t) = \beta O_P(e^{-\beta t(1-t)})$, is clear. Next,

$$B_n(t) = -\int_0^1 \frac{\partial}{\partial s} K_\lambda(s,t)[G(s) - \bar{\Phi}_{(n)}(s)]ds$$
$$= \int_0^1 K_\lambda(s,t)[dG(s) - d\bar{\Phi}_{(n)}(s)] - K_\lambda(1,t)[G(1) - \bar{\Phi}_{(n)}(1)]$$
$$= \frac{\beta}{2n}\sum_{i=1}^n e^{-\beta|t-\frac{i}{n}|}U_{(i)} + \beta O_P(e^{-\beta t(1-t)}).$$

Finally, using Result 2, we conclude that, $\mathcal{K}_\lambda \phi_{(n)}(t) = \phi_{(n)}(t) + O_P(\lambda)$. Hence,

$$A_n(t) = -\int_0^1 \frac{\partial}{\partial s} K_\lambda(s,t)\Phi_{(n)}(s)ds$$
$$= \int_0^1 K_\lambda(s,t)\phi_{(n)}(s)ds - K_\lambda(1,t)\Phi_{(n)}(1)$$
$$= \phi_{(n)}(t) + O_P(\lambda) + \beta O_P(e^{-\beta t(1-t)}).$$

Hence, combining them, we deduce, for $0 < t < 1$, and any λ,

$$\hat{\phi}(t) = \phi_{(n)}(t) + O_P(\lambda) + \sum_{i=1}^n \frac{\beta}{2n}e^{-\beta|t-\frac{i}{n}|}U_{(i)} + \beta O_P(n^{-\frac{2}{3}}\log n) + \beta O_P(e^{-\beta t(1-t)}).$$

A.3. Proof of Proposition 2

As $0 < F(z_0) < 1$, from Proposition 3, we have:

$$\hat{\phi}(F(z_0)) = \phi_{(n)}(F(z_0)) + O_P(\lambda) + \frac{\beta}{2n}\sum_{i=1}^n e^{-\beta|F(z_0)-\frac{i}{n}|}U_{(i)}$$
$$+ \beta O_P(n^{-\frac{2}{3}}\log n) + \beta O_P(e^{-\beta F(z_0)(1-F(z_0))}).$$

Next, we observe that, both $\hat{\phi}$ and $\phi_{(n)}$ have bounded slopes near $F(z_0)$ (Both of them are piecewise differentiable functions). To see this, consider a small neighborhood of $F(z_0)$ where F is strictly increasing, i.e. the covariate density f is strictly positive. In this neighborhood,

$$|\phi_{(n)}(\bar{F}_n(z_0)) - \phi_{(n)}(F(z_0))| \leq |\bar{F}_n(z_0) - F(z_0)| \sup \frac{|\phi_{(n)}(t) - \phi_{(n)}(s)|}{|t - s|}$$

$$\leq |\bar{F}_n(z_0) - F(z_0)| ||m'||_\infty \sup \frac{t - s}{\bar{F}_n(t) - \bar{F}_n(s)}.$$

The last supremum is taken in a neighborhood around z_0 (Since $\bar{F}_n(z_0) \to F(z_0)$, it will eventually be inside that interval around $F(z_0)$). Clearly, that supremum equals the reciprocal of $\min n(z_{i+1} - z_i)$ which converges to $\inf f$ in that neighborhood. (See, Theorem 3 in Shorack (1972) for a reference).

Therefore,

$$|\phi_{(n)}(\bar{F}_n(z_0)) - \phi_{(n)}(F(z_0))| = |\bar{F}_n(z_0) - F(z_0)|O_P(1). \tag{7}$$

Next, in a neighborhood of $F(z_0)$ which does not contain 0 or 1, we have,

$$\sup \frac{\hat{\phi}(s) - \hat{\phi}(t)}{s - t} = \frac{1}{\lambda} \sup(\check{\Phi} - G)_+ \leq \frac{1}{\lambda}||\check{\Phi} - \tilde{G}||$$

using (3) and Marshall's lemma respectively. Since $\lambda n^{2/3} \to 0$ and the term $||\check{\Phi} - \tilde{G}||$ is of order $1/n$ (using the same arguments as in Lemma 1), we conclude that

$$|\hat{\phi}(\bar{F}_n(z_0)) - \hat{\phi}(F(z_0))| = |\bar{F}_n(z_0) - F(z_0)|O_P(1). \tag{8}$$

in the neighborhood.
Now, since $\bar{F}_n(z_0) - F(z_0) = O_P(1/\sqrt{n})$ (by the central limit theorem), (7) and (8) yield:

$$\hat{m}(z_0) - m(z_0) = \hat{\phi}(\bar{F}_n(z_0)) - \phi_{(n)}(\bar{F}_n(z_0)) = \hat{\phi}(F(z_0)) - \phi_{(n)}(F(z_0)) + O_P(1/\sqrt{n})$$

The result follows. $\square$

Sample	Binomial			Poisson		
size	two-sided	pivot-based	LRT	two-sided	pivot-based	LRT
100	0.934	0.922	0.942	0.930	0.925	0.943
200	0.945	0.938	0.957	0.934	0.927	0.957
300	0.948	0.944	0.952	0.923	0.916	0.954
400	0.948	0.949	0.961	0.933	0.928	0.952
500	0.948	0.948	0.949	0.935	0.930	0.948
600	0.949	0.947	0.958	0.944	0.937	0.954
700	0.954	0.948	0.952	0.944	0.938	0.949
800	0.950	0.943	0.952	0.942	0.937	0.950
900	0.940	0.939	0.958	0.950	0.941	0.944
1000	0.936	0.936	0.953	0.954	0.942	0.946

Table 1

Relative comparison of the coverage for smoothing based C.I.s and the likelihood ratio based method obtained from simulations with sample sizes 100,...,1000 in case of binary and Poisson regression.

Sample	Binomial			Poisson		
size	two-sided	pivot-based	LRT	two-sided	pivot-based	LRT
100	0.1960	0.1896	0.2928	0.5557	0.4773	0.8755
200	0.1551	0.1529	0.2324	0.4514	0.3791	0.6862
300	0.1353	0.1342	0.1999	0.3997	0.3312	0.5970
400	0.1228	0.1223	0.1805	0.3666	0.3007	0.5430
500	0.1139	0.1137	0.1672	0.3429	0.2788	0.5020
600	0.1071	0.1071	0.1564	0.3246	0.2622	0.4715
700	0.1016	0.1018	0.1487	0.3100	0.2489	0.4470
800	0.0972	0.0974	0.1421	0.2978	0.2378	0.4280
900	0.0934	0.0937	0.1361	0.2875	0.2285	0.4090
1000	0.0901	0.0905	0.1309	0.2785	0.2204	0.3930

Table 2

Relative comparison of the interval lengths for smoothing based C.I.s and the likelihood ratio based method obtained from simulations with sample sizes 100,...,1000 in case of binary and Poisson regression.

age	Estimators		C.I.		
	MLE	smooth	lrt-based	smooth (two-sided)	smooth (pivot-based)
2.00	0.486	0.459	.360-.594 (.234)	.393-.597 (.204)	.338-.659 (.320)
4.00	0.486	0.459	.362-.594 (.232)	.393-.598 (.205)	.338-.659 (.320)
6.00	0.500	0.489	.376-.678 (.302)	.436-.640 (.204)	.367-.696 (.329)
8.00	0.500	0.513	.376-.714 (.338)	.467-.669 (.202)	.389-.725 (.336)
10.00	0.571	0.582	.403-.818 (.415)	.542-.738 (.196)	.455-.808 (.352)
12.00	0.571	0.631	.405-.898 (.493)	.568-.761 (.193)	.503-.867 (.363)
14.00	0.857	0.808	.593-.940 (.347)	.626-.810 (.184)	.690-1.000 (.310)
16.00	0.867	0.862	.693-.947 (.254)	.672-.847 (.175)	.753-1.000 (.247)
18.00	0.867	0.862	.731-.947 (.216)	.699-.868 (.169)	.753-1.000 (.247)
20.00	0.867	0.862	.731-.947 (.216)	.726-.887 (.161)	.753-1.000 (.247)
22.00	0.875	0.893	.757-.959(.202)	.760-.912 (.152)	.789-1.000 (.211)
24.00	0.929	0.924	.793-.965 (.172)	.794-.934 (.140)	.828-1.000 (.172)
26.00	0.940	0.939	.841-.967 (.126)	.848-.967 (.119)	.848-1.000 (.152)
28.00	0.940	0.939	.863-.967 (.104)	.861-.974 (.113)	.848-1.000 (.152)

Table 3

Comparison of intervals given by different techniques for the Rubella data set. The first column indicates the age; the second and the third column give the corresponding value of the MLE and the smooth estimator. The next three columns indicate the confidence intervals with their widths for the likelihood ratio based method, the smoothing based two-sided interval and the pivot-based confidence interval.
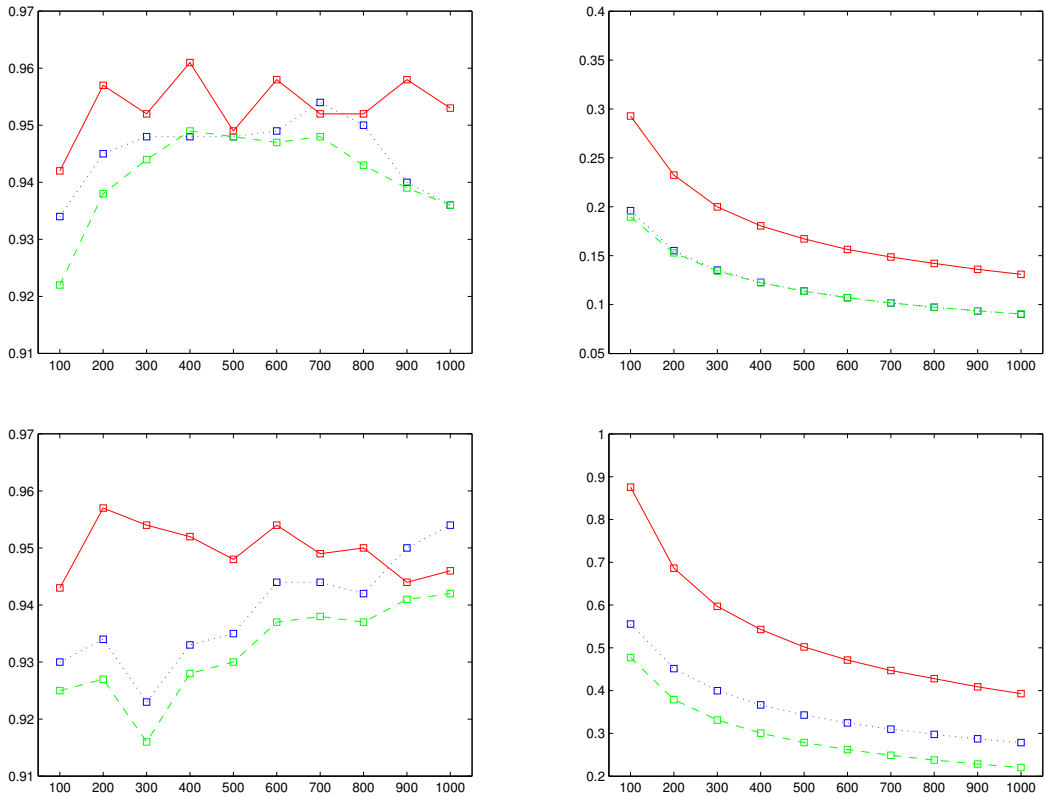
Fig. 1. The Coverage and the interval length for the smooth estimate based methods and the LRT based interval in the Binomial and Poisson families. The solid lines indicate the LRT method, the dotted lines show the two–sided C.I. and the dashed lines are for the pivot-based C.I. The first row of figures are for the coverage and expected length respectively for binary regression, whereas the second row corresponds to Poisson regression.
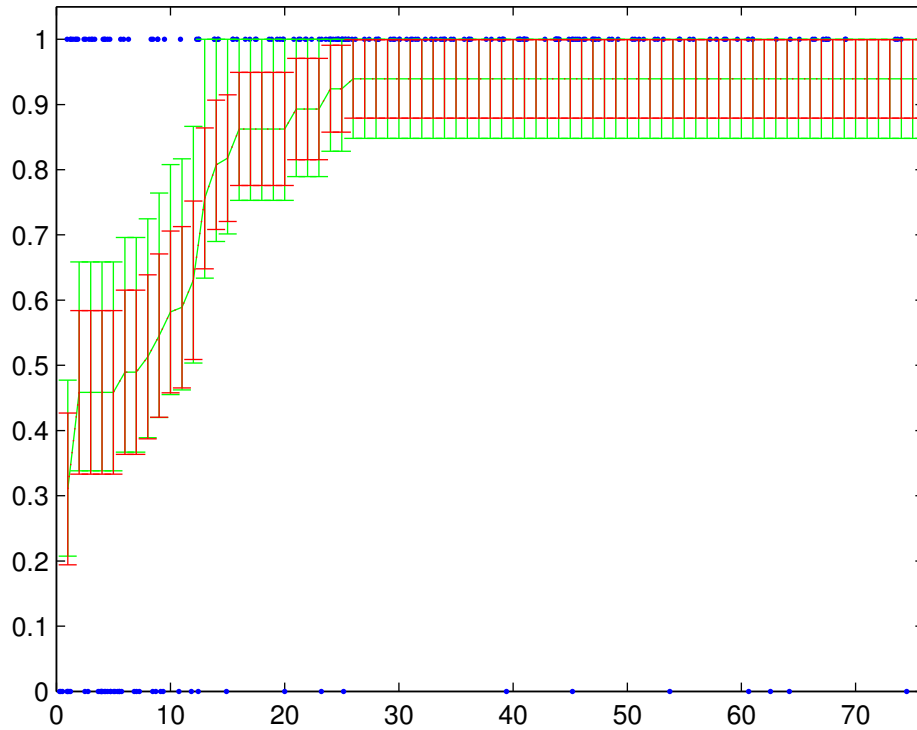
27

Fig. 2. *Confidence intervals based on the Rubella data set for ages 1-80 using the two smoothing methods are shown. The intervals are shown with vertical bars and the smooth estimator is plotted. The data points are also shown as tiny dots. The two-sided intervals are shown using the bold bands and the pivot-based in a lighter shade.*

References

Banerjee, M., 2007. Likelihood based inference for monotone response models. To appear in Ann. Statist. **35**. Available at www.stat.lsa.umich.edu/∼moulib/b1.pdf

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

Banerjee, M., Wellner, J. A., 2005. Confidence intervals for current status data. Scand. J. Statist. **32**, 405–424.

Brunk, H. D., 1970. Estimation of isotonic regression. Nonparametric Techniques in Statistical Inference. Edited by M.L. Puri.

Dette, H., Neumeyer, N., Pilz, K. F., 2005. A note on Nonparametric estimation of the effective dose in quantal bioassay. J. Amer. Statist. Assoc. **100**, 503-510.

Diggle, P., Morris, S., Morton-Jones, T., 1999. Case-control isotonic regression

for investigation of elevation in risk around a risk source. Stat. Med. **18**, 1605-1613.

Dunson, D.B., 2005. Bayesian semiparametric isotonic regression for count data. J. Amer. Statist. Assoc. **100**, 618-627.

Grenander, U., 1956. On the Theory of mortality Measurement, Part II, Skand. Akt. **39**, 125-153.

Groeneboom, P., 1989. Brownian motion with a parabolic drift and airy functions. Probab. Theory Related Fields, **81**, 79-109.

Groeneboom, P., Wellner, J. A., 2001. Computing Chernoff's distribution. J. Comput. Graph. Statist. **10**, 388-400.

Huang, J., 1996. Efficient estimation for the proportional hazards model with interval censoring. Ann. Statist., **24**, 540-568.

Keiding, N., Begtrup, K., Scheike, T., Hasibeder, G., 1996. Estimation from current status data in continuous time. Lifetime Data Anal. **2**, 119-129.

Kelly, C., Rice, J., 1990. Monotone smoothing with application to dose-response curves and the assessment of Synergism. Biometrics. **46**, 1071-1085.

Kiefer, J., Wolfowitz, J., 1976. Asymptotically Minimax Estimation of concave and convex Distribution Functions. Z. Wahrsch. verw. Gebiete, **34**, 73-85.

Mammen, E. 1991. Estimating a Smooth Monotone Regression Function. Ann. Statist. **19**, 724-740.

Mammen, E., Marron, J. S., Turlach, B., Wand, M. P. (2001). A general projection framework for constrained smoothing. Statist. Sci. **16** 232-248.

Mammen, E., Thomas-Agnan, C., 1999. Smoothing splines and shape restrictions. Scand. J. Statist. **26**, 239-252.

Morton-Jones, T., Diggle, P., Elliott, P., 1999. Investigation of excess environment risk around putative sources: Stone's test with covariate adjustment. Stat. Med., **18**, 189-197.

Mukerjee, H., 1988. Monotone nonparametric regression. Ann. Statist., **16**, 741-750.

Newton, M. A., Czado, C., Chappell, R., 1996. Bayesian inference for semi-parametric binary regression. J. Amer. Statist. Assoc., **91**, 142-153.

Nychka, D. (1995). Splines as local smoothers. Ann. Statist., **23**, 1175-1197.

Pal, J., Woodroofe, M. B., 2004. On the distance between the Cumulative Sum Diagram and its Greatest Convex Minorant under unequally spaced design points. Scand. J. Statist. **33**, 279-291.

Pal, J., Woodroofe, M. B., 2007. Large Sample Properties of Shape Restricted Regression estimators with Smoothness Adjustments To appear in Statist. Sinica. Available at www.samsi.info/jpal/ss05029-revise-2.pdf

Pal, J., 2006. Statistical analysis and inference in shape restricted problems with applications to astronomy. Ph.D. dissertation, University of Michigan.

Parzen, E., 2004. Quantile probability and statistical data modeling. Statist. Sci. **19**, 652-662.

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

Ramsay, J.O., 1998. Estimating Smooth Monotone Functions. J. R. Stat. Soc. Ser. B. **60**, 365-375.

Robertson, T., Wright, F. T., Dykstra, R. L., 1987. Order Restricted Statistical Inference. John Wiley and Sons.

Salanti, G., Ulm, K., 2003. Tests for trend with binary response. Biom. J., **45**, 277-291.

Shiboski, S., 1998. Generalized additive models with current status data. Lifetime Data Anal., **4**, 29-50.

Shorack, G., 1972. Convergence of quantiles and spacings processes with applications. Ann. Math. Statist., **43**, 1400-1411.

Silverman, B., 1984. Spline Smoothing: the equivalent variable Kernel Method. Ann. Statist. **12**, 898-916.

Stone, R. A., 1988. Investigations of excess environmental risks around putative sources: Statistical problems and a proposed test. Stat. Med., **7**, 649-660.

Sun, J. and Kalbfleisch, J.D., 1995. Estimation of the mean function of a point process based on panel count data. Statist. Sinica, **5**, 279-290.

Sun, J., 1999. A nonparametric test for current status data with unequal censoring. J. R. Stat. Soc. Ser. B, **61**, 243-250.

Tantiyaswasdikul, Chim and Woodroofe, M. B. (1994). Isotonic Smoothing Splines under Sequential Designs. J. Statist. Plan. Inference. **38**, 75-87.

Wright, I. W., Wegman, E. J., 1980. Isotonic, Convex and related Splines. Ann. Statist., **8**, 1023-1035.