Semiparametric Binary Regression Models under Shape Constraints

Moulinath Banerjee [1], Debasri Mukherjee and Santosh Mishra

University of Michigan, Western Michigan University and Oregon State University

Abstract

We consider estimation of the regression function in a semiparametric binary regression model defined through an appropriate link function (with emphasis on the logistic link) using likelihood-ratio based inversion. The dichotomous response variable Δ is influenced by a set of covariates that can be partitioned as (X, Z) where Z (real valued) is the covariate of primary interest and X (vector valued) denotes a set of control variables. For any fixed X, the conditional probability of the event of interest ($\Delta = 1$) is assumed to be a non–decreasing function of Z. The effect of the control variables is captured by a regression parameter β. We show that the baseline conditional probability function (corresponding to $X = 0$) can be estimated by isotonic regression procedures and develop a likelihood ratio based method for constructing asymptotic confidence intervals for the conditional probability function (the regression function) that avoids the need to estimate nuisance parameters. Interestingly enough, the calibration of the likelihood ratio based confidence sets for the regression function no longer involves the usual χ^2 quantiles, but those of the distribution of a new random variable that can be characterized as a functional of convex minorants of Brownian motion with quadratic drift. Confidence sets for the regression parameter β can however be constructed using asymptotically χ^2 likelihood ratio statistics. The finite sample performance of the methods are assessed via a simulation study.

Keywords: convex minorants, likelihood ratio statistic, school attendance, semiparametric binary regression.

1 INTRODUCTION

Binary regression models are used frequently to model the effects of covariates on dichotomous outcome variables. A general formulation of parametric binary regression models runs as follows. If Δ is the indicator of the outcome and X is a set of (d–dimensional) covariates believed to influence the outcome, one can write $\tilde{g}(\mu(X)) = \beta^T X$, where the regression function $\mu(X) = P(\Delta = 1 \mid X)$ and $\tilde{g}$ is a smooth monotone increasing function from $(0, 1)$ to $(-\infty, \infty)$ and is called the

[1] Research supported in part by National Science Foundation grants DMS-0306235 and DMS-0705288

“link function”. Such models are very well–studied in the statistical literature (see, for example, McCullagh and Nelder (1989)) from both computational and theoretical angles. Some commonly used link functions are the logit (logistic regression), the probit and the complementary log-log link. In this paper, our interest is in situations where in addition to X, we have an additional (real–valued) covariate Z whose effect on the outcome variable is known qualitatively. More specifically, larger values of Z tend to make the outcome ($\Delta = 1$) more likely. The effect of Z in the model can be incorporated as follows. Write,

$$\tilde{g}(\mu(X, Z)) = \beta^T X + \psi(Z) \ (\psi \text{ increasing}). \tag{1.1}$$

Note that (a) $\mu(X, Z)$, the conditional probability of the outcome, is increasing in Z for fixed X and (b) the nonparametric component affects the conditional probability of an outcome additively on the scale of the link function.[1] Models of this kind are useful in a variety of settings and are therefore of considerable interest. See, for example, Dunson (2003) and Dunson and Neelon (2003) where nonparametric estimation of ψ as in (1.1) above is done in a Bayesian framework; for more general treatments, where the conditional mean of the outcome variable is monotone in one of the regressors, see Manski and Tamer (2002) and Magnac and Maurin (2007). See also, related work by Manski (1988) and Magnac and Maurin (2004), and some earlier work by the first author (Ghosh, Banerjee and Biswas (2004, 2008)) on monotone binary regression models considered in a fully nonparametric setting.

This paper treats a semiparametric binary regression model of the type described in (1.1), from the angle of likelihood inference, based on i.i.d. observations $\{\Delta_i, X_i, Z_i\}_{i=1}^n$ from the distribution of (Δ, X, Z). Our inference strategies are based on the maximization of the underlying likelihood function (to be described in Section 2). More specifically, we focus on testing the hypotheses: (a) $H_0 : \beta = \beta_0$ and (b) $\tilde{H}_0 : \psi(z_0) = \theta_0$ for some fixed point z_0, using the likelihood ratio statistic (henceforth LRS).

The key contributions of our approach are two–fold. The first is the break from the more conventional smoothness assumptions on the nonparametric component ψ; indeed, our smoothness assumptions are minimal as we only require the function to be continuously differentiable. Rather, we impose a shape constraint on ψ that is dictated by background knowledge about the effect of Z on the outcome. One major advantage of this approach stems from the fact that shape constraints automatically ”regularize” the estimation problem in the sense that the underlying likelihood function can be meaningfully maximized *without penalization or kernel smoothing.* Thus, this procedure avoids the well–known problems of choosing a penalization or smoothing parameter.

Secondly, the use of likelihood ratio statistics for making inferences on β and ψ, provides a simple but elegant way of constructing confidence sets, not only for these parameters but also for the conditional probability function/regression function ($\mu(x, z) = E\,(\Delta = 1 \mid X = x, Z = z)$) – a quantity of significant interest– circumventing the problem of estimating nuisance parameters. We

[1] The assumption that ψ is increasing is not restrictive; if the dependence on Z is decreasing, one can use the transformed covariate $\tilde{Z} \equiv -Z$.

elaborate on this in what follows.

We show that the LRS for testing H_0 has a limiting χ_d^2 distribution as in regular parametric problems, while that for testing $\tilde{H}_0 : \psi(z_0) = \theta_0$ converges in distribution to a "universal" random variable $\mathbb{D}$; "universal" in the sense that it does not depend on the underlying parameters of the model or the point of interest z_0. [2] This latter result is a new and powerful one as it can be used to obtain confidence sets for $\mu(x, z)$ by inverting likelihood ratio tests for testing a family of hypotheses of type (b). We emphasize that the computation of the LRS is completely objective and *does not involve smoothing/penalization* as discussed above. Furthemore, calibration of the likelihood ratio test only involves knowledge of the quantiles of (the asymptotic pivot) $\mathbb{D}$, which are well tabulated. Hence, nuisance parameters need not be estimated from the data. Of course, we reap a similar advantage while constructing confidence sets for the regression parameter β (or a sub–parameter as is discussed in Section 3) which involves inverting a family of likelihood ratio tests as in (a), calibrated via the usual χ^2 quantiles. In contrast, note that inferences for β and ψ (equivalently μ) could also be done using the limit distributions of the corresponding maximum likelihood estimates. However, we do not adopt this route as these distributions involve nuisance parameters that are difficult to estimate. One could argue that nuisance parameter estimation in this context could be obviated through the use of resampling techniques like subsampling (Politis, Romano and Wolf (1999)) or the m out of n bootstrap (the usual Efron–type bootstrap will not work in this situation (Sen, Banerjee and Woodroofe (2008)). But this once again introduces the problem of choosing m, the "block size", which can be regarded as a variant of a smoothing parameter.

Thus, the likelihood ratio method is much more *automated* and *objective* than its competitors.

As far as technicalities are concerned, the methodology and asymptotic theory developed in this paper is markedly different from those used in the smoothing literature. This arises from the fact that maximum likelihood estimation under monotonicity constraints can typically be reduced to an isotonic regression problem (for a comprehensive review see Robertson, Wright and Dykstra (1988) and more recently, Silvapulle and Sen (2004)). It is well–known in the isotonic regression literature that the asymptotic behavior of estimates (like the MLEs of ψ in our model) obtained through such regression cannot be analyzed using standard CLTs as they are highly non–linear functionals of the empirical distribution[3]; rather they are asymptotically distributed as the derivatives of convex minorants of Gaussian processes, which are non–normal. These technical details are discussed in later sections.

The rest of the paper is organized as follows. Maximum likelihood estimation and novel

[2] This limit distribution does not belong to the χ_1^2 family but can be thought of as an *analogue* of the χ_1^2distribution in nonregular statistical problems involving $n^{1/3}$ rate of convergence for maximum likelihood estimators and non–Gaussian limit distributions. Indeed, the maximum likelihood estimator $\hat{\psi}_n$ converges to the true ψ at rate $n^{1/3}$ in this problem, while the rate of convergence of $\hat{\beta}$ is $\sqrt{n}$.

[3] This is in contrast to estimates based on smoothing methods, which are essentially linear functionals of the empirical distribution, which enables the use of CLTs and leads to asymptotic normality.

likelihood ratio-based inferential procedures for a general link function are discussed in Section 2. In Section 3, for concreteness, we focus primarily on the logit link, which is the most widely used link function in statistical data analysis and discuss the asymptotic results and the associated methodology for construction of confidence sets in the setting of this model. We also indicate in Section 3 that similar results continue to hold for other commonly used link functions, like the probit link or the complementary log–log link. Proofs of some of the results in Section 3 are collected in the Appendix (Section 4).

2 COMPUTING MLES AND LIKELIHOOD RATIOS

The density function of the random vector (Δ, X, Z) is given by $p(\delta, x, z) = \mu(x,z)^{\delta}\,(1-\mu(x,z))^{1-\delta}\,f(z,x)$, where $f(z,x)$ is the joint density of (Z,X) with respect to Leb $\times\ \mu$ where Leb denotes Lebesgue measure on $[0,\infty)$ and μ is some measure defined on $\mathbb{R}^d$. We construct the likelihood function for the data, as:

$$L_n(\beta, \psi, \{\Delta_i, X_i, Z_i\}_{i=1}^n) = \Pi_{i=1}^n\, \mu(X_i, Z_i)^{\Delta_i}\,(1-\mu(X_i,Z_i))^{1-\Delta_i}\, f(Z_i, X_i)\,. \tag{2.2}$$

In what follows, we denote the true underlying values of the parameters (β, ψ) by (β_0, ψ_0). Using a link function $\tilde{g}$ (satisfying **Condition C:** For $\delta = 1$ or 0, the function $v(s) := \delta\,\log\,\tilde{h}(s) + (1-\delta)\,\log(1-\tilde{h}(s))$ is concave for $s \in (-\infty, \infty)$[4]), with inverse function $\tilde{h}$, the log–likelihood function for the sample, up to an additive factor that does not involve any of the parameters of interest, is given by $l_n(\beta,\psi) \equiv \log\,L_n(\beta, \psi, \{\Delta_i, X_i, Z_i\}_{i=1}^n) = \sum_{i=1}^n\, l(\beta, \psi, \Delta_i, X_i, Z_i)$where

$$l(\beta, \psi, \delta, x, z) = \delta\,\log\,\tilde{h}(\beta^T x + \psi(z)) + (1-\delta)\,\log(1-\tilde{h}(\beta^T x + \psi(z))). \tag{2.3}$$

We next introduce some notation and define some key quantities that will be crucial to the subsequent development. Let $Z_{(1)}, Z_{(2)}, \ldots, Z_{(n)}$ denote the ordered values of the Z_i's; let $\Delta_{(i)}$ and $X_{(i)}$ denote the indicator and covariate values associated with $Z_{(i)}$. Also, let $u_i \equiv \psi(Z_{(i)})$ and $R_i(\beta) = \beta^T X_{(i)}$. Denote the vector $(u_1, u_2, \ldots, u_n)$ by $\mathbf{u}$ and define the function g as: $g(\beta, \mathbf{u}) \equiv -l_n(\beta, \psi) = \sum_{i=1}^n\, \phi(\Delta_{(i)}, R_i(\beta), u_i)$ where $\phi(\delta, r, u) = -\delta\,\log\,\tilde{h}(r+u) - (1-\delta)\,\log(1-\tilde{h}(r+u))$. Note that the monotone function ψ is identifiable only up to its values at the $Z_{(i)}$'s; hence we identify ψ with the vector $\mathbf{u}$. Let $(\hat{\beta}_n, \hat{\mathbf{u}}_n) = \text{argmin}_{\beta\in\mathbb{R}^d, \{\mathbf{u}: u_1\leq u_2 \leq \ldots \leq u_n\}}\, g(\beta, \mathbf{u})$. The unconstrained MLE of (β, ψ) is given by $(\hat{\beta}_n, \hat{\psi}_n)$ where $\hat{\psi}_n$ is the (unique) right–continuous increasing step function that assumes the value $\hat{u}_{i,n}$ (the i'th component of $\hat{\mathbf{u}}_n$) at the point $Z_{(i)}$ and has no jump points outside of the set $\{Z_{(i)}\}_{i=1}^n$.

Next, for a fixed β, let $\hat{\mathbf{u}}_n^{(\beta)} = \text{argmin}_{\{\mathbf{u}: u_1\leq u_2\leq \ldots\leq u_n\}}\, g(\beta, \mathbf{u})$. Define $\hat{\psi}_n^{(\beta)}$ to be the (unique) right–continuous increasing step function that assumes the value $\hat{u}_{i,n}^{(\beta)}$ (the i'th component of

[4] It is easy to check that all three standard link functions used in binary regression: (a) the logit link for which $\tilde{h}(s) = e^s/(1+e^s)$, (b) the probit link for which $\tilde{h}(s) = \Phi(s)$, Φ denoting the normal cdf and (c) the complementary log-log link for which $\tilde{h}(s) = 1 - e^{-e^s}$, satisfy this property. The concavity of v implies the convexity of the function $g(\beta, \mathbf{u})$, to be introduced soon, in its arguments and guarantees a unique minimizer that may be obtained by using standard methods from convex optimization theory.

$\hat{\mathbf{u}}_n^{(\beta)}$) at the point $Z_{(i)}$ and has no jump points outside of the set $\{Z_{(i)}\}_{i=1}^n$. Then, note that: $\hat{\beta}_n = \text{argmin}_\beta \, g(\beta, \hat{\mathbf{u}}_n^{(\beta)})$ and $\hat{\psi}_n = \hat{\psi}_n^{(\hat{\beta}_n)}$.

2.1 The Likelihood Ratio Statistic for Testing the Value of β

The likelihood ratio statistic for testing $H_0 : \beta = \beta_0$ is given by:

$$\text{lrtbeta}_n = 2\,(l_n(\hat{\beta}_n, \hat{\psi}_n) - l_n(\beta_0, \hat{\psi}_n^{(\beta_0)}))\,. \tag{2.4}$$

Our computation of MLEs, $(\hat{\beta}_n, \hat{\psi}_n, \hat{\psi}_n^{(\beta_0)})$, in this semiparametric problem relies heavily on the convexity of $g(\beta, \mathbf{u})$ and is based on the following proposition.

Proposition 1: *Let $f(\gamma_1, \gamma_2)$ be a real–valued function defined on $\mathbb{R}^{k_1} \times \mathcal{C}$ where γ_1 varies in $\mathbb{R}^{k_1}$ and $\gamma_2 \in \mathcal{C}$, where $\mathcal{C}$ is a closed convex subset of $\mathbb{R}^{k_2}$. Assume that f is a continuously differentiable strictly convex function that is minimized (uniquely) at the point $(\gamma_1^\star, \gamma_2^\star)$. Consider the following updating algorithm. Start at an arbitrary point (γ_1^0, γ_2^0) in $\mathbb{R}^{k_1} \times \mathcal{C}$. Having defined (γ_1^m, γ_2^m) at stage m ($m \geq 0$), set $\gamma_2^{m+1} \equiv \text{argmin}_{\gamma_2 \in \mathcal{C}} f(\gamma_1^m, \gamma_2)$ and γ_1^{m+1} as the (unique) solution to $(\partial/\partial\,\gamma_1)\, f(\gamma_1, \gamma_2^{m+1}) = 0$. Then, irrespective of the starting value, the sequence of points $\{\gamma_1^m, \gamma_2^m\}_{m\geq 0}$ converges to $(\gamma_1^\star, \gamma_2^\star)$.*

Remark 1: We do not provide a proof of this proposition in this paper. The proposition follows as a direct consequence of Theorem 2.2 in Jongbloed (1998) on the convergence of an iteratively defined sequence using an algorithmic map which is adapted from a general convergence theorem (Theorem 7.2.3) from Bazaraa et. al. (1993). [5]

Consider first, the computation of the unconstrained MLEs $(\hat{\beta}_n, \hat{\psi}_n)$. The function g is defined on $\mathbb{R}^d \times \tilde{\mathcal{C}}$, where $\tilde{\mathcal{C}} \equiv \{\mathbf{u} = (u_1, u_2, \ldots, u_n) : u_1 \leq u_2 \leq \ldots \leq u_n\}$ is a closed convex cone in $\mathbb{R}^n$. Setting f in Proposition 1 to be g, $\gamma_1 = \beta$ and $\gamma_2 = \mathbf{u}$, it is easy to check that g is a continuously differentiable and strictly convex function that attains a unique minimum at $(\hat{\beta}_n, \hat{\mathbf{u}}_n)$. Thus, we are in the setting of Proposition 1 above. We next provide a step–by–step outline of the algorithm to evaluate $(\hat{\beta}_n, \hat{\psi}_n)$.

Computing the unconstrained MLEs:

Step 1. At Stage 0 of the algorithm, propose initial estimates $(\hat{\beta}_n^{(0)}, \hat{u}_n^{(0)})$. Also, set an initial tolerance level $\eta > 0$, small.

Step 2a. At Stage $p \geq 0$ of the algorithm, current estimates $(\hat{\beta}_n^{(p)}, \mathbf{u_n}^{(p)})$ are available. At Stage $p+1$, first update the second component to $\mathbf{u_n}^{(p+1)}$ by minimizing $g(\hat{\beta}_n^{(p)}, \mathbf{u})$ over $\mathbf{u} \in \tilde{\mathcal{C}}$, using *the*

[5] **Additional remark:** Note that the step of updating γ_1^m to γ_1^{m+1} which involves solving $(\partial/\partial\,\gamma_1)\, f(\gamma_1, \gamma_2^{m+1}) = 0$ is also a minimization step. Since $f(\gamma_1, \gamma_2^{m+1})$ is a continuously differentiable strictly convex function of γ_1, it is uniquely minimized at the point where its derivative is 0. Thus, we could alternatively have written: $\gamma_1^{m+1} \equiv \text{argmin}_{\gamma_1 \in \mathbb{R}^{k_1}} f(\gamma_1, \gamma_2^{m+1})$.

modified iterative convex minorant algorithm (MICM) due to Jongbloed (1998). Note that $\mathbf{u_n}^{(p+1)}$ is precisely the vector $\{\hat{\psi}_n^{(\beta)}(Z_{(i)})\}_{i=1}^n$, for $\beta = \hat{\beta}_n^{(p)}$.

Step 2b. Having updated to $\mathbf{u_n}^{(p+1)}$, next update $\hat{\beta}_n^{(p)}$ to $\hat{\beta}_n^{(p+1)}$ by solving $(\partial/\partial\,\beta)\,g(\beta, \mathbf{u_n}^{(p+1)}) = 0$ using, for example, the Newton–Raphson procedure. In terms of the log–likelihood function, this amounts to solving $(\partial/\partial\,\beta) l_n(\beta, \psi) = 0$ for $\psi = \hat{\psi}_n^{(\hat{\beta}_n^{(p)})}$.

Step 3. (*Checking convergence*) If

$$\left| \frac{g(\hat{\beta}_n^{(p+1)}, \mathbf{u}_n{}^{(p+1)}) - g(\hat{\beta}_n^{(p)}, \mathbf{u}_n{}^{(p)})}{g(\hat{\beta}_n^{(p)}, \mathbf{u}_n{}^{(p)})} \right| \leq \eta$$

then stop and declare $(\hat{\beta}_n^{(p+1)}, \mathbf{u}_n{}^{(p+1)})$ as the MLEs. Otherwise, set $p = p + 1$ and return to Step 2a.

We now elaborate on Step 2a, the most involved segment of the above algorithm, that requires iterative quadratic optimization techniques under order constraints. This is precisely the problem of evaluating $\hat{\psi}_n^{(\beta)}$, the MLE of ψ for a fixed β. In particular, recall that $\hat{\psi}_n^{(\beta_0)}$ is the MLE of ψ under $H_0 : \beta = \beta_0$.

Characterizing and computing $\hat{\psi}_n^{(\beta)}$: This is characterized by the vector $\hat{\mathbf{u}}_n^{(\beta)} = (\hat{u}_{1,n}^{(\beta)} \leq \hat{u}_{2,n}^{(\beta)} \ldots \leq \hat{u}_{n,n}^{(\beta)})$ that minimizes $g(\beta, \mathbf{u})$ over all $u_1 \leq u_2 \leq \ldots \leq u_n$. [6] Before proceeding further, we 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 left-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 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 solution $\hat{\mathbf{u}}_n^{(\beta)}$ can be viewed as the slope of the greatest convex minorant (slogcm) of a random function defined in terms of $\hat{\mathbf{u}}_n^{(\beta)}$ itself. This *self–induced/self–consistent* characterization proves useful both for computational purposes and for the asymptotic theory. For the sake of notational convenience, we will denote $g(\beta, \mathbf{u})$ in the following discussion by $\xi(\mathbf{u})$ (suppressing the dependence on β) and $\hat{\mathbf{u}}_n^{(\beta)}$ by $\hat{\mathbf{u}} = (\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_n)$. For $1 \leq i \leq n$, set $d_i = \bigtriangledown_{ii}\,\xi(\hat{\mathbf{u}})$. Define the function η as follows:

$$\eta(\mathbf{u}) \quad = \quad \sum_{i=1}^n \left[u_i - \hat{u}_i + \bigtriangledown_i\,\xi(\hat{\mathbf{u}})\,d_i^{-1}\right]^2\,d_i = \sum_{i=1}^n \left[u_i - \left(\hat{u}_i - \bigtriangledown_i\,\xi(\hat{\mathbf{u}})\,d_i^{-1}\right)\right]^2\,d_i\,. \qquad (2.5)$$

[6] Without loss of generality one can assume that $\Delta_{(1)} = 1$ and $\Delta_{(n)} = 0$. If not, the effective sample size for the estimation of the parameters is $k_2 - k_1 + 1$ where k_1 is the first index i such that $\Delta_{(i)} = 1$ and k_2 is the last index such that $\Delta_{(i)} = 0$. It is not difficult to see that one can set $\hat{u}_{i,n}^{(\beta)} = -\infty$ for all $i < k_1$ and $\hat{u}_{i,n}^{(\beta)} = \infty$ for all $i > k_2$ without imposing any constraints on the other components of the minimizing vector.

It can be shown that $\hat{\mathbf{u}}$ minimizes η subject to the constraints that $u_1 \leq u_2 \leq \ldots \leq u_n$ (see Section 4.2 in the appendix for the details) and hence furnishes the isotonic regression of the function $h(i) = \hat{u}_i - \bigtriangledown_i \xi(\hat{\mathbf{u}})\, d_i^{-1}$ on the ordered set $\{1, 2, \ldots, n\}$ with weight function $d_i \equiv \bigtriangledown_{ii} \xi(\hat{\mathbf{u}})$. It is well known that the solution $\hat{\mathbf{u}} = (\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_n) = \text{slogcm} \left\{ \sum_{j=1}^{i} d_i \; , \; \sum_{j=1}^{i} h(i)\, d_i \right\}_{i=0}^{n}$. See, for example Theorem 1.2.1 of ROBERTSON ET.AL.(1988) and more generally, Chapter 1 of that book for an extensive discussion of isotonic regression.

Since $\hat{\mathbf{u}}$ is unknown, an iterative scheme is resorted to. For a fixed vector $\mathbf{v} \equiv (v_1, v_2, \ldots, v_n) \in \mathcal{C}$, set $d_{\mathbf{v},i} = \bigtriangledown_{ii} \xi(\mathbf{v})$ and define the function $\eta_{\mathbf{v}}(\mathbf{u}) \equiv \sum_{i=1}^{n} [u_i - (v_i - \bigtriangledown_i \xi(\mathbf{v})\, d_{\mathbf{v},i}^{-1}]^2\, d_{\mathbf{v}}, i$. Pick an initial guess for $\hat{\mathbf{u}}$, say $\mathbf{u}^{(0)} \in \mathcal{C}$, set $\mathbf{v} = \mathbf{u}^{(0)}$ and compute $\mathbf{u}^{(1)}$ by minimizing $\eta_{\mathbf{v}}(\mathbf{u})$ over $\mathcal{C}$; then, set $\mathbf{v} = \mathbf{u}^{0}$, obtain $\mathbf{u}^{(2)}$ by minimizing $\eta_{\mathbf{v}}(\mathbf{u})$ again, and proceed thus, until convergence. Generally, with $\mathbf{v} = \mathbf{u}^{(j)}$, we have: $\mathbf{u}^{(j+1)} = \text{slogcm} \left\{ \sum_{j=1}^{i} d_{\mathbf{v},i} \; , \; \sum_{j=1}^{i} (v_i - \bigtriangledown_i \xi(\mathbf{v})\, d_{\mathbf{v},i}^{-1})\, d_{\mathbf{v},i} \right\}_{i=0}^{n}$.

Remark 2: Certain convergence issues might arise with such a straightforward iterative scheme, since the algorithm could hit inadmissible regions in the search space. Jongbloed (1998) addresses this issue by using a modified iterated convex minorant (henceforth MICM) algorithm; see Section 2.4 of his paper for a discussion of the practical issues and a description of the relevant algorithm which incorporates a line search procedure to guarantee convergence to the minimizer. As this is a well–established algorithm in the isotonic regression literature, we do not discuss these subtleties any further, but refer the reader to Jongbloed's paper.

Remark 3: We have also not explicitly addressed the convergence issue. This is discussed in Jongbloed (1998). The iterations are stopped when the (necessary and sufficient) conditions that characterize the unique minimizer of ξ are satisified to a pre–specified degree of tolerance. For a discussion of these conditions, see Section 4.2.

2.2 The Likelihood Ratio Statistic For Testing the Value of ψ at a point

We next turn our attention to the likelihood ratio test for testing $\tilde{H}_0 : \psi(z_0) = \theta_0$ with $-\infty < \theta_0 < \infty$. This requires us to compute the constrained maximizers of β and ψ, say $(\hat{\beta}_{n,0}, \hat{\psi}_{n,0})$ under $\tilde{H}_0 : \psi(z_0) = \theta_0$. As in the unconstrained case, this maximization can be achieved in two steps. For each β, one can compute $\hat{\psi}_{n,0}^{(\beta)} = \text{argmax}_{\psi : \psi(z_0) = \theta_0}\, l_n(\beta, \psi)$. Then, $\hat{\beta}_{n,0} = \text{argmax}_{\beta}\, l_n(\beta, \hat{\psi}_{n,0}^{(\beta)})$ and $\hat{\psi}_{n,0} = \hat{\psi}_{n,0}^{(\hat{\beta}_{n,0})}$. The likelihood ratio statistic for testing $\tilde{H}_0 : \psi(z_0) = \theta_0$ is given by:

$$\text{lrtpsi}_n = 2\left(l_n(\hat{\beta}_n, \hat{\psi}_n) - l_n(\hat{\beta}_{n,0}, \hat{\psi}_{n,0})\right). \qquad (2.6)$$

Note that the monotone function $\hat{\psi}_{n,0}^{(\beta)}$ is identifiable only up to its values at the $Z_{(i)}$'s (and at the fixed point z_0 where it is required to equal θ_0) and we identify this function with the vector $\hat{\mathbf{u}}_{n,0}^{(\beta)} \equiv (\hat{u}_{1,n,0}^{(\beta)}, \hat{u}_{2,n,0}^{(\beta)}, \ldots, \hat{u}_{n,n,0}^{(\beta)})$ where $\hat{u}_{i,n,0}^{(\beta)} = \hat{\psi}_{n,0}^{(\beta)}(Z_{(i)})$. We will discuss the characterization of this vector shortly.

Before proceeding further, we introduce some notation. First, let m denote the number of Z values that are less than or equal to z_0. Then, we have $Z_{(m)} < z_0 < Z_{(m+1)}$ (with probability 1). Note that any monotone function ψ that satisfies $\tilde{H}_0$ will have: $\psi(Z_{(m)}) \leq \theta_0 \leq \psi(Z_{(m+1)})$. Define $\tilde{\mathcal{C}}_0$ to be the closed convex subset of $\mathbb{R}^n$ comprising all vectors $\mathbf{u}$ with $u_1 \leq u_2 \leq \ldots \leq u_m \leq \theta_0 \leq u_{m+1} \leq \ldots \leq u_n$. If $\hat{\mathbf{u}}_{n,0} \equiv (\hat{u}_{1,n,0}, \hat{u}_{2,n,0}, \ldots, \hat{u}_{n,n,0})$ denotes the vector $\{\hat{\psi}_{n,0}(Z_{(i)})\}_{i=1}^n$, then $(\beta_{n,0}, \hat{\mathbf{u}}_{n,0}) = \operatorname{argmin}_{\beta \in \mathbb{R}^d, \mathbf{u} \in \tilde{\mathcal{C}}_0} g(\beta, \mathbf{u})$. Since the function g is a continuously differentiable strictly convex function defined on the closed convex set $\mathbb{R}^d \times \tilde{\mathcal{C}}_0$ and assumes a unique minimum at $(\beta_{n,0}, \hat{\mathbf{u}}_{n,0})$, we can invoke Proposition 1 as before. The algorithm, which is similar to that in the preceding subsection, is formally presented below.

Computing the constrained MLEs under $\tilde{H}_0$:

Step 1. At Stage 0 of the algorithm, propose initial estimates $(\hat{\beta}_{n,0}^{(0)}, \hat{u}_{n,0}^{(0)})$. Also, set an initial tolerance level $\eta > 0$, small.

Step 2a. At Stage $p \geq 0$ of the algorithm, current estimates $(\hat{\beta}_{n,0}^{(p)}, \mathbf{u}_{n,0}^{(p)})$ are available. At Stage $p+1$, first update the second component to $\mathbf{u_{n,0}}^{(p+1)}$ by minimizing $g(\hat{\beta}_{n,0}^{(p)}, \mathbf{u})$ over $\mathbf{u} \in \tilde{\mathcal{C}}_0$. Note that $\mathbf{u}_{n,0}^{(p+1)}$ is precisely the vector $\{\hat{\psi}_{n,0}^{\beta}(Z_{(i)})\}_{i=1}^n$, for $\beta = \hat{\beta}_{n,0}^{(p)}$.

Step 2b. Next, update $\hat{\beta}_{n,0}^{(p)}$ to $\hat{\beta}_{n,0}^{(p+1)}$ by solving $(\partial/\partial\beta)\, g(\beta, \mathbf{u}_{n,0}^{(p+1)}) = 0$ using, say, the Newton–Raphson method.

Step 3. (*Checking convergence*) If $\left| \frac{g(\hat{\beta}_{n,0}^{(p+1)}, \mathbf{u}_{n,0}^{(p+1)}) - g(\hat{\beta}_{n,0}^{(p)}, \mathbf{u}_{n,0}^{(p)})}{g(\hat{\beta}_{n,0}^{(p)}, \mathbf{u}_{n,0}^{(p)})} \right| \leq \eta$, then stop and declare $(\hat{\beta}_{n,0}^{(p+1)}, \mathbf{u}_{n,0}^{(p+1)})$ as the MLEs. Otherwise, set $p = p+1$ and return to Step 2a.

It remains to elaborate on Step 2a, which involves computing $\hat{\psi}_{n,0}^{(\beta)}$ for some β.

Characterizing $\hat{\psi}_{n,0}^{(\beta)}$: Finding $\hat{\psi}_{n,0}^{(\beta)}$ amounts to minimizing $g(\beta, \mathbf{u}) = \sum_{i=1}^n \phi(\Delta_{(i)}, R_i(\beta), u_i)$ over all $u_1 \leq u_2 \ldots \leq u_m \leq \theta_0 \leq u_{m+1} \leq \ldots \leq u_n$. For the remainder of this discussion, we denote the minimizing vector $\hat{\mathbf{u}}_{n,0}^{(\beta)}$ by $\hat{\mathbf{u}}^{(0)}$. Finding $\hat{\mathbf{u}}^{(0)}$ can be reduced to solving two separate optimization problems. These are [1] Minimize $g_1(\beta, u_1, u_2, \ldots, u_m) \equiv \sum_{i=1}^m \phi(\Delta_{(i)}, R_i(\beta), u_i)$ over $u_1 \leq u_2 \leq \ldots \leq u_m \leq \theta_0$, and, [2] Minimize $g_2(\beta, u_{m+1}, u_{m+2}, \ldots, u_n) \equiv \sum_{i=m+1}^n \phi(\Delta_{(i)}, R_i(\beta), u_i)$ over $\theta_0 \leq u_{m+1} \leq u_{m+2} \leq \ldots \leq u_n$.

Consider [1] first. This is a problem that involves minimizing a smooth convex function over a convex set and one can easily write down the Kuhn–Tucker conditions characterizing the minimizer. It is easy to see that the solution $(\hat{u}_1^{(0)}, \hat{u}_2^{(0)}, \ldots, \hat{u}_m^{(0)})$ can be obtained as follows: Minimize $g_1(\beta, \mathbf{u_1})$ where $\mathbf{u_1} \equiv (u_1, u_2, \ldots, u_m)$ over $\mathcal{C}_1$, the closed convex cone in $\mathbb{R}^m$ defined as $\{\mathbf{u_1} : u_1 \leq u_2 \leq \ldots \leq u_m\}$, to obtain $\tilde{\mathbf{u}}_\mathbf{1} \equiv (\tilde{u}_1, \tilde{u}_2, \ldots, \tilde{u}_m)$. Then, $(\hat{u}_1^{(0)}, \hat{u}_2^{(0)}, \ldots, \hat{u}_m^{(0)}) = (\tilde{u}_1 \wedge \theta_0, \tilde{u}_2 \wedge \theta_0, \ldots, \tilde{u}_m \wedge \theta_0)$. The minimization of $g_1(\beta, \cdot)$ over $\mathcal{C}_1$ requires use of the MICM and follows the same technique as described in the preceding subsection in connection with estimating $\hat{\psi}_n^{(\beta)}$. On the other hand, the solution vector to [2], say $(\hat{u}_{m+1}^{(0)}, \hat{u}_{m+2}^{(0)}, \ldots, \hat{u}_n^{(0)})$, is given by $(\hat{u}_{m+1}^{(0)}, \hat{u}_{m+2}^{(0)}, \ldots, \hat{u}_n^{(0)}) = (\tilde{u}_{m+1} \vee \theta_0, \tilde{u}_{m+2} \vee \theta_0, \ldots, \tilde{u}_n \vee \theta_0)$ where $(\tilde{u}_{m+1}, \tilde{u}_{m+2}, \ldots, \tilde{u}_n) = \text{argmin}_{u_{m+1} \leq u_{m+2} \leq \ldots \leq u_n} g_2(\beta, u_{m+1}, u_{m+2}, \ldots, u_n)$ and uses the MICM, as in [1]. Finally $\hat{\mathbf{u}}^{(0)} = (\hat{u}_1^{(0)}, \hat{u}_2^{(0)}, \ldots, \hat{u}_n^{(0)})$.

3 ASYMPTOTIC RESULTS

In this section we present asymptotic results for the estimation of β and ψ. For the sake of concreteness and ease of exposition, we present results explicitly in the setting of logistic regression. The semiparametric logistic model is given by $\log \frac{\mu(X,Z)}{1-\mu(X,Z)} = \beta^T X + \psi(Z)$. The above display is equivalent to writing:

$$\mu(X,Z) = \frac{e^{\beta^T X} \Lambda(Z)}{1 + e^{\beta^T X} \Lambda(Z)} \quad , \text{ where } \Lambda(Z) = e^{\psi(Z)}. \tag{3.7}$$

The parameter space for β is taken to be a bounded subset of $\mathbb{R}^d$. We denote it by $\mathcal{B}$. The parameter space for $\Lambda = e^{\psi}$ is the space of all nondecreasing cadlag (i.e. right-continuous with left-hand limits) functions from $[0, \tau]$ to $[0, M]$ where M is some large positive constant. Let (β_0, Λ_0) denote the true model parameters (thus, $\Lambda_0 = e^{\psi_0}$). We make the following assumptions:

(A.1) The true regression parameter β_0 is an interior point of $\mathcal{B}$.

(A.2) The covariate X has bounded support. Hence, there exists x_0 such that $P(\|X\| \leq x_0) = 1$. Also $E(\text{Var}(X \mid Z))$ is positive definite with probability one.

(A.3) Let $\Lambda_0(0) = 0$. Let $\tau_{\Lambda_0} = \inf\{z : \Lambda_0(z) = \infty\}$. The support of Z is an interval $[\sigma, \tau]$ with $0 < \sigma < \tau < \tau_{\Lambda_0}$.

(A.4) We assume that $0 < \Lambda_0(\sigma-) < \Lambda_0(\tau) < M$. Also, Λ_0 is continuously differentiable on $[\sigma, \tau]$ with derivative λ_0 bounded away from 0 and from ∞.

(A.5) The marginal density of Z, which we denote by f_Z, is continuous and positive on $[\sigma, \tau]$.

(A.6) The function $h^{\star\star}$ defined below in (3.8) defined below has a version which is differentiable componentwise with each component possessing a bounded derivative on $[\sigma, \tau]$.

Remarks: The boundedness of $\mathcal{B}$ along with assumptions (A.1)–(A.3) are needed to deduce the consistency and rates of convergence of the maximum likelihood estimators. In particular, the boundedness of the covariate X does not cause a problem with applications. The utility of the assumption that the conditional dispersion of X given Z is positive definite is explained below. (A.4) and (A.5) are fairly weak regularity conditions on Λ_0 and the distribution of Z. The assumption (A.6) is a technical assumption and is required to ensure that one can define appropriate approximately *least favorable submodels*; these are finite–dimensional submodels of the given semiparametric model, with the property that the efficient score function for the semiparametric model at the true parameter values can be approximated by the usual score functions from these submodels. They turn out to be crucial for deriving the limit distribution of the likelihood ratio statistic for testing the regression parameter.

We now introduce the efficient score function for β in this model. The log density function for the vector (Δ, Z, X) is given by:

$$l_{\beta,\Lambda}(\delta, z, x) = \delta\,(\log\,\Lambda(z) + \beta^T\,x) - \log\,(1 + \Lambda(z)\,\exp(\beta^T\,x)) + \log\,f(z,x)\,.$$

The ordinary score function for β in this model is:

$$\dot{l}_\beta(\beta,\Lambda)(\delta,z,x) = (\partial/\partial\,\beta)\,l_{\beta,\Lambda}(\delta,x,z) = x\,\Lambda(z)\,Q((\delta,z,x);\beta,\Lambda)\,,$$

where

$$Q((\delta,z,x);\beta,\Lambda) = \frac{\delta}{\Lambda(z)} - \frac{e^{\beta^T\,x}}{1+\Lambda(z)\,e^{\beta^T\,x}}\,.$$

The score function for Λ is a linear operator acting on the space of functions of bounded variation on $[\sigma,\tau]$ and has the form:

$$\dot{l}_\Lambda(\beta,\Lambda)(h(\cdot))(\delta,z,x) = h(z)\,Q((\delta,z,x);\beta,\Lambda)\,.$$

Here h is a function of bounded variation on $[\sigma,\tau]$. To compute the form of this score function, we consider curves of the form $\Lambda + t\,h$ for $t \geq 0$ where h is a non–decreasing non–negative function on $[\sigma,\tau]$. Computing

$$B_\Lambda(h) = \frac{\partial}{\partial\,t}\,l_{\beta,\Lambda+t\,h}(\delta,x,z)\ \ |_{t=0}\,,$$

we get $B_\Lambda(h) = h(z)\,Q((\delta,z,x);\beta,\Lambda)$. The linear operator B_Λ now extends naturally to the closed linear span of all non–decreasing h's, which is precisely the space of all functions of bounded variation on $[\sigma,\tau]$.

The efficient score function for β at the true parameter values (β_0,Λ_0), which we will denote by $\tilde{l}$ for brevity, is given by

$$\tilde{l} = \dot{l}_\beta(\beta_0,\Lambda_0) - \dot{l}_\Lambda(\beta_0,\Lambda_0)h^\star$$

for functions $h^\star = (h_1^\star, h_2^\star, \ldots, h_d^\star)$ of bounded variation, such that $h_i^\star$ minimizes the distance

$$E_{\beta_0,\Lambda_0}(\dot{l}_{\beta,i}(\beta_0,\Lambda_0) - \dot{l}_\Lambda(\beta_0,\Lambda_0)\,h(\cdot))^2\,,$$

for h varying in the space of functions of bounded variation on $[\sigma,\tau]$. Here

$$\dot{l}_{\beta,i}(\beta_0,\Lambda_0) = x^{(i)}\,\Lambda(z)\,Q((\delta,z,x);\beta_0,\Lambda_0)$$

is the i'th component of the ordinary score function for β (and $x^{(i)}$ is the i'th component of x). It is not difficult to see that $h_i^\star$ must satisfy

$$E\,[B_{\Lambda_0}(h)\,(\dot{l}_{\beta,i}(\beta_0,\Lambda_0) - B_{\Lambda_0}(h_i^\star))] = 0\,,$$

for all h. This simplifies to

$$E\,[Q^2((\Delta,Z,X);\beta_0,\Lambda_0)h(Z)\,[X^{(i)}\,\Lambda_0(Z)-h_i^\star(Z)]]=0\,.$$

For the above to be satisfied it suffices to have:

$$E\,[Q^2((\Delta,Z,X);\beta_0,\Lambda_0)\,[X^{(i)}\,\Lambda_0(Z)-h_i^\star(Z)]\ \mid Z]=0\,,$$

whence

$$h_i^\star(Z)=\Lambda_0(Z)\,\frac{E(X^{(i)}\,Q^2((\Delta,Z,X);\beta_0,\Lambda_0)\ \mid Z)}{E(Q^2((\Delta,Z,X);\beta_0,\Lambda_0)\ \mid Z)}\,.$$

In vector notation we can therefore write

$$h^\star(Z)=\Lambda_0(Z)\,h^{\star\star}(Z)\equiv\Lambda_0(Z)\,\frac{E_{\beta_0,\Lambda_0}(X\,Q^2((\Delta,Z,X);\beta_0,\Lambda_0)\ \mid Z)}{E_{\beta_0,\Lambda_0}(Q^2((\Delta,Z,X);\beta_0,\Lambda_0)\ \mid Z)}\,. \tag{3.8}$$

The assumption (A.2) that $E(\mathrm{Var}(X\mid Z))$ is positive definite ensures that $\tilde{l}$, the efficient score function for β, is not identically zero, whence the efficient information $\tilde{I}_0=\mathrm{Disp}(\tilde{l})\equiv E_{\beta_0,\Lambda_0}(\tilde{l}\,\tilde{l}^T)$ is positive definite (Note that $E_{\beta_0,\Lambda_0}(\tilde{l})=0$). This entails that the MLE of β will converge at $\sqrt{n}$ rate to $beta_0$ and have an asymptotically normal distribution with a finite dispersion matrix.

Let $\tilde{\theta}_0=e^{\theta_0}$. Now, consider the problem of testing $H_0:\beta=\beta_0$ based on our data, but under the (true) constraint that $\Lambda(z_0)=\tilde{\theta}_0$. Thus, we define:

$$\mathrm{lrtbeta}_n^0=2\,\log\,\frac{\mathrm{argmax}_{\Lambda(z_0)=\tilde{\theta}_0}\,l_n(\beta,\Lambda)}{\mathrm{argmax}_{\beta=\beta_0,\Lambda(z_0)=\tilde{\theta}_0}\,l_n(\beta,\Lambda)}\,. \tag{3.9}$$

Thus,

$$\mathrm{lrtbeta}_n^0=2\,l_n(\hat{\beta}_{n,0},\hat{\Lambda}_{n,0})-2\,l_n(\beta_0,\hat{\Lambda}_{n,0}^{(\beta_0)})\,,$$

where $\hat{\Lambda}_{n,0}=\exp(\hat{\psi}_{n,0})$ and $\hat{\Lambda}_{n,0}^{(\beta_0)}=\exp(\hat{\psi}_{n,0}^{(\beta_0)})$. We now state a theorem describing the asymptotic behavior of $\hat{\beta}_n$ and $\hat{\beta}_{n,0}$ (which we subsequently denote by $\tilde{\beta}_n$) and the likelihood ratio statistics $\mathrm{lrtbeta}_n$ as defined in (2.4) and $\mathrm{lrtbeta}_n^0$ above.

Theorem 3.1 *Under Conditions (A.1) – (A.7), both $\hat{\beta}_n$ and $\tilde{\beta}_n$ are asymptotically linear in the efficient score function and have the following representation:*

$$\sqrt{n}\,(\hat{\beta}_n-\beta_0)=\frac{1}{\sqrt{n}}\,\tilde{I}_0^{-1}\,\sum_{i=1}^n\tilde{l}(\Delta_i,Z_i,X_i)+r_n \tag{3.10}$$

and

$$\sqrt{n}\,(\tilde{\beta}_n-\beta_0)=\frac{1}{\sqrt{n}}\,\tilde{I}_0^{-1}\,\sum_{i=1}^n\tilde{l}(\Delta_i,Z_i,X_i)+s_n \tag{3.11}$$

where r_n and s_n are $o_p(1)$. Hence both $\sqrt{n}\,(\hat{\beta}_n-\beta_0)$ and $\sqrt{n}\,(\tilde{\beta}_n-\beta_0)$ converge in distribution to $N(0,\tilde{I}_0^{-1})$.

Furthermore,

$$lrtbeta_n = n(\hat{\beta}_n - \beta_0)^T \, \tilde{I}_0 \, (\hat{\beta}_n - \beta_0) + o_p(1)\,, \tag{3.12}$$

while

$$lrtbeta_n^0 = n(\tilde{\beta}_n - \beta_0)^T \, \tilde{I}_0 \, (\tilde{\beta}_n - \beta_0) + o_p(1)\,. \tag{3.13}$$

It follows that both $lrtbeta_n$ and $lrtbeta_n^0$ are asympotically distributed like χ^2_d.

We next state asymptotic results concerning the nonparametric component of the model. In order to do so, we introduce the following processes. For positive constants c and d define the process $X_{c,d}(z) := c\,W(z) + d\,z^2$, where $W(z)$ is standard two-sided Brownian motion starting from 0. Let $G_{c,d}(z)$ denote the GCM of $X_{c,d}(z)$. Let $g_{c,d}(z)$ be the right derivative of $G_{c,d}$. This is a non–decreasing function that can be shown to be a piecewise constant, with finitely many jumps in any compact interval. Next, let $G_{c,d,L}(h)$ denote the GCM of $X_{c,d}(h)$ restricted to the set $h \leq 0$ and $g_{c,d,L}(h)$ denote its right–derivative process. For $h > 0$, let $G_{c,d,R}(h)$ denote the GCM of $X_{c,d}(h)$ restricted to the set $h > 0$ and $g_{c,d,R}(h)$ denote its right–derivative process. Define $g^0_{c,d}(h) = (g_{c,d,L}(h) \wedge 0)\, 1(h \leq 0) + (g_{c,d,R}(h) \vee 0)\, 1(h > 0)$. Then $g^0_{c,d}(h)$, like $g_{c,d}(h)$, is a non–decreasing function that is piecewise constant, with finitely many jumps in any compact interval and differs (almost surely) from $g_{c,d}(h)$ on a finite interval containing 0. In fact, with probability 1, $g^0_{c,d}(h)$ is identically 0 in some (random) neighborhood of 0, whereas $g_{c,d}(h)$ is almost surely non-zero in some (random) neighborhood of 0. Also, the interval $D_{c,d}$ on which $g_{c,d}$ and $g^0_{c,d}$ differ is $O_p(1)$. For more detailed descriptions of the processes $g_{c,d}$ and $g^0_{c,d}$, see Banerjee (2000), Banerjee and Wellner (2001) and Wellner (2003). Thus, $g_{1,1}$ and $g^0_{1,1}$ are the unconstrained and constrained versions of the slope processes associated with the "canonical" process $X_{1,1}(z)$. By Brownian scaling, the slope processes $g_{c,d}$ and $g^0_{c,d}$ can be related in distribution to the canonical slope processes $g_{1,1}$ and $g^0_{1,1}$. This is the content of the following proposition.

Lemma 3.1 *For any $M > 0$, the following distributional equality holds in the space $L_2[-M, M] \times L_2[-M, M]$:*

$$\left(g_{c,d}(h), g^0_{c,d}(h)\right) \;\overset{\mathcal{D}}{=}\; \left(c\,(d/c)^{1/3} g_{1,1}\left((d/c)^{2/3} h\right), c\,(d/c)^{1/3} g^0_{1,1}\left((d/c)^{2/3} h\right)\right)\,.$$

Here $L_2[-M, M]$ denotes the space of real–valued functions on $[-M, M]$ with finite L_2 norm (with respect to Lebesgue measure).

This is proved in Banerjee (2000), Chapter 3.

Let z_0 be an interior point of the support of Z. Define the (localized) slope processes U_n and V_n as follows:

$$U_n(h) = n^{1/3}\,(\hat{\psi}_n^{(\beta_0)}(z_0 + h\,n^{-1/3}) - \psi_0(z_0)) \;\text{ and }\; V_n(h) = n^{1/3}\,(\hat{\psi}_{n,0}^{(\beta_0)}(z_0 + h\,n^{-1/3}) - \psi_0(z_0))\,.$$

The following theorem describes the limiting distribution of the slope processes above.

Theorem 3.2 *Define,*

$$C(z_0) = \int \frac{e^{\beta_0^T x + \psi_0(z_0)}}{(1 + e^{\beta_0^T x + \psi_0(z_0)})^2} \, f(z_0, x) \, d\,\mu(x) \, .$$

Assume that $0 < C(z_0) < \infty$*. Let*

$$a = \sqrt{\frac{1}{C(z_0)}} \quad \text{and} \quad b = \frac{1}{2}\, \psi_0^{'}(z_0) \, ,$$

where $\psi_0^{'}$ *is the derivative of* ψ_0*. The processes* $(U_n(h), V_n(h))$ *converge finite dimensionally to the processes* $(g_{a,b}(h), g_{a,b}^0(h))$*. Furthermore, using the monotonicity of the processes* U_n *and* V_n*, it follows that the convergence holds in the space* $L_2[-K, K] \times L_2[-K, K]$ *for any* $K > 0$*.*

Setting $h = 0$ in the above theorem, we find that

$$n^{1/3} \, (\hat{\psi}_n^{(\beta_0)}(z_0) - \psi_0(z_0)) \to_d g_{a,b}(0) \equiv_d a \, (b/a)^{1/3} \, g_{1,1}(0) \equiv_d (8 \, a^2 \, b)^{1/3} \, \mathbb{Z} \, ,$$

where $\mathbb{Z} \equiv \text{argmin}_{h \in \mathbb{R}} \, (W(h) + h^2)$ and its distribution is referred to in the statistical literature as Chernoff's distribution. See, for example, Groeneboom and Wellner (2001) for a detailed description. The above display utilizes the result that $g_{1,1}(0) \equiv_d 2 \, \mathbb{Z}$ (since this result is not used in our proposed methodology for constructing confidence sets, discussed below, we do not establish this result in our paper). The random variable $\mathbb{Z}$ arises extensively in nonparametric problems involving *cube-root asymptotics* – problems where estimates of parameters converge at rate $n^{1/3}$ and in particular, is typically found to characterize the pointwise limit distribution of maximum likelihood estimators of monotone functions in nonparametric/semiparametric models. The distribution of $\mathbb{Z}$ is non-Gaussian and symmetric about 0. It can, in fact, be shown that

$$n^{1/3} \, (\hat{\psi}_n(z_0) - \psi_0(z_0)) \to_d (8 \, a^2 \, b)^{1/3} \, \mathbb{Z}$$

where $\hat{\psi}_n \equiv \hat{\psi}_n^{\hat{\beta}_n}$ is the unconstrained MLE of ψ. This is not surprising in view of the fact that $\sqrt{n}(\hat{\beta}_n - \beta_0) = O_p(1)$, so that $\hat{\beta}_n$ converges to β_0 at a faster rate than $n^{1/3}$, the convergence rate for $\hat{\psi}_n^{\hat{\beta}_n}$. Since the quantiles of $\mathbb{Z}$ are well–tabulated, this result can be used to construct asymptotic confidence sets of any pre-assigned level for $\psi_0(z_0)$ (equivalently $\Lambda_0(z_0)$), but the procedure requires estimating the constants a and b which turns out to be a tricky affair (one needs to estimate the joint density of the covariates that appears in the defining integral for $C(z_0)$ in addition to the derivative of ψ_0 at the point z_0, which is quite difficult, especially at modest sample sizes). Resampling techniques, like subsampling (m out of n bootstrap without replacement) as discussed in Politis, Romano and Wolf (1999), can circumvent the estimation of the nuisance parameters a and b, but are computationally quite intensive. To avoid these difficulties, we do not construct MLE based confidence sets for $\psi_0(z_0)$ in this paper; rather, we resort to inversion of the likelihood ratio statistic for testing the value of ψ_0 at a pre-fixed point of interest. Our next theorem is crucial for this purpose.

Theorem 3.3 *The likelihood ratio statistic for testing $\tilde{H}_0 : \psi(z_0) = \theta_0$, as defined in (2.6), converges in distribution to $\mathbb{D}$ where*

$$\mathbb{D} = \int \left((g_{1,1}(z))^2 - (g_{1,1}^0(z))^2 \right) \, dz \, .$$

The random variable $\mathbb{D}$ can be considered to be a non-regular analogue of the usual χ_1^2 random variable, in the sense that just as the χ_1^2 distribution describes the limiting likelihood ratio statistic for testing a real– valued parameter in a regular parametric model, similarly, the distribution of $\mathbb{D}$ describes the limiting likelihood ratio statistic for testing the value of a monotone function at a point in *conditionally parametric models* (see Banerjee (2007)) and more generally in pointwise estimation of monotone functions.

Construction of confidence sets for parameters of interest via likelihood ratio based inversion: Denote the likelihood ratio statistic for testing the null hypothesis $\psi(z_0) = \theta$ by $\text{lrtpsi}_n(\theta)$. The computation of the likelihood ratio statistic is dicussed, in detail, in Section 2. By Theorem 3.3, an approximate level $1 - \alpha$ confidence set for $\psi_0(z_0)$ is given by $S_{\psi_0(z_0)} \equiv \{\theta : \text{lrtpsi}_n(\theta) \leq q(\mathbb{D}, 1-\alpha)\}$, where $q(\mathbb{D}, 1-\alpha)$ is the $(1-\alpha)$'th quantile of the distribution of $\mathbb{D}$ (for $\alpha = 0.05$, this is approximately 2.28). Noting that $\Lambda_0(z_0) = \exp(\psi_0(z_0))$, the corresponding confidence set for $\Lambda_0(z_0)$ is simply $\exp(S_{\psi_0(z_0)})$. Furthermore, the corresponding confidence set for the baseline conditional probability function, $E(\Delta \mid X = 0, Z = z_0)$ is simply $e^{S_{\psi_0(z_0)}}/(1 + e^{S_{\psi_0(z_0)}})$.

Confidence sets for the regression function at values $X = x_0, Z = z_0$, i.e. $\mu(x_0, z_0) = E(\Delta \mid X = x_0, Z = z_0)$ can also be constructed in a similar fashion. This requires redefining the covariate X, so as to convert $\mu(x_0, z_0)$ to a baseline conditional probability. Set $\tilde{X} = X - x_0$. Then $\mu(x_0, z_0) = P(\Delta = 1 \mid \tilde{X} = 0, Z = z)$. Define $\tilde{\mu}(\tilde{x}, z) = E(\Delta \mid \tilde{X} = \tilde{x}, Z = z)$. We have,

$$\tilde{\mu}(\tilde{x}, z) = \mu(\tilde{x} + x_0, z) = \frac{e^{\beta_0^T\,(\tilde{x}+x_0)}\,\Lambda_0(z)}{1 + e^{\beta_0^T\,(\tilde{x}+x_0)}\,\Lambda_0(z)} = \frac{e^{\beta_0^T\,\tilde{x}}\,\tilde{\Lambda}_0(z)}{1 + e^{\beta_0^T\,\tilde{x}}\,\tilde{\Lambda}_0(z)}$$

where $\tilde{\Lambda}_0(z) = e^{\beta_0^T\,x_0}\,\Lambda_0(z)$, with $\tilde{\psi}_0(z) \equiv \log\,\tilde{\Lambda}_0(z) = \beta_0^T\,x_0 + \psi_0(z)$. This is exactly the model considered at the beginning of Section 3 in terms of new covariates $(\tilde{X}, Z)$ and satisfies the regularity conditions A.1 – A.6 (with X replaced by $\tilde{X}$). Now, $\mu(x_0, z_0) = \tilde{\mu}(0, z_0) = e^{\tilde{\psi}_0(z_0)}/(1 + e^{\tilde{\psi}_0(z_0)})$. An approximate level $1 - \alpha$ confidence set for $\tilde{\psi}_0(z_0)$, say $\tilde{S}_{\tilde{\psi}_0(z_0)}$ can be found in exactly the same fashion as before; i.e. $\tilde{S}_{\tilde{\psi}_0(z_0)} = \{\theta : \widetilde{\text{lrtpsi}}_n(\theta) \leq q(\mathbb{D}, 1-\alpha)\}$, where $\widetilde{\text{lrtpsi}}_n(\theta)$ is the likelihood ratio statistic for testing $\tilde{\psi}(z_0) = \theta$ and is computed in exactly the same way as the statistic in (2.6), but using the covariates $\tilde{X}$ and Z, instead of X and Z. Correspondingly, the confidence set for $\tilde{\mu}(0, z_0)$ is $e^{\tilde{S}_{\tilde{\psi}_0(z_0)}}/(1 + e^{\tilde{S}_{\tilde{\psi}_0(z_0)}})$. This principle is applied extensively to construct confidence sets for the conditional probabilites in the data analysis example in Section 4.

The construction of joint confidence sets is also of importance in certain applications. Thus, one may be interested in a joint confidence set for $(\mu(x_0, z_0), \mu(x_0, z_1))$ for $z_0 < z_1$. To this end consider the hypothesis $\tilde{H}_{0,1} : \psi(z_0) = \theta_0, \psi(z_1) = \theta_1$ where $z_0 < z_1$ and $\theta_0 < \theta_1$. A natural statistic to test this hypothesis is $M_n \equiv \max(\mathrm{lrtpsi}_n^{(z_0)}(\theta_0), \mathrm{lrtpsi}_n^{(z_1)}(\theta_1))$ where $\mathrm{lrtpsi}_n^{(z_0)}(\theta_0)$ is the likelihood ratio statistic for testing $\psi(z_0) = \theta_0$ and $\mathrm{lrtpsi}_n^{(z_1)}(\theta_1)$, the likelihood ratio statistic for testing $\psi(z_1) = \theta_1$. It can be shown that when the null hypothesis is true, $\mathrm{lrtpsi}_n^{(z_0)}(\theta_0)$ and $\mathrm{lrtpsi}_n^{(z_1)}(\theta_1)$ are asymptotically independent and M_n converges in distribution to $\mathbb{D}^{(2)} \equiv \max(\mathbb{D}_1, \mathbb{D}_2)$ where $\mathbb{D}_1$ and $\mathbb{D}_2$ are identical copies of $\mathbb{D}$. The quantiles of this distribution are well–tabulated and joint confidence sets for $(\psi(z_0), \psi(z_1))$ are therefore readily constructed by inversion. This leads to joint confidence sets for $(\mu(x_0, z_0), \mu(x_0, z_1))$ by centering X around x_0, as in the previous paragraph. Consider the pair (θ, θ') (with $\theta \le \theta'$) and let $\widetilde{\mathrm{lrtpsi}}_n^{(z_0)}(\theta)$ and $\widetilde{\mathrm{lrtpsi}}_n^{(z_1)}(\theta')$ denote, respectively, the likelihood ratio statistics for testing $\tilde{\psi}(z_0) = \theta$ and $\tilde{\psi}(z_1) = \theta'$, these being computed in exactly the same way as the statistic in (2.6) but using covariates $(\tilde{X}, Z)$ instead of (X, Z). Let:

$$\tilde{S}_{\tilde{\psi}_0(z_0),\tilde{\psi}_0(z_1)} = \{(\theta, \theta') : \theta \le \theta', \max(\widetilde{\mathrm{lrtpsi}}_n^{(z_0)}(\theta), \widetilde{\mathrm{lrtpsi}}_n^{(z_1)}(\theta')) \le q(\mathbb{D}^{(2)}, 1-\alpha)\}\,.$$

This set has a simple characterization as a polygon in $\mathbb{R}^2$ (it is either a triangle or a trapezium or a pentagon). Let $\overline{S}_{\tilde{\psi}_0(z_0)} = \{\theta : \widetilde{\mathrm{lrtpsi}}_n^{(z_0)}(\theta) \le q(\mathbb{D}^{(2)}, 1-\alpha)\}$ and $\overline{S}_{\tilde{\psi}_0(z_1)} = \{\theta' : \widetilde{\mathrm{lrtpsi}}_n^{(z_1)}(\theta') \le q(\mathbb{D}^{(2)}, 1-\alpha)\}$. Then:

$$\tilde{S}_{\tilde{\psi}_0(z_0),\tilde{\psi}_0(z_1)} = (\overline{S}_{\tilde{\psi}_0(z_0)} \times \overline{S}_{\tilde{\psi}_0(z_1)}) \cap \mathcal{C}_2$$

where $\mathcal{C}$ is the cone given by $\{(\theta, \theta') \in \mathbb{R}^2 : \theta \le \theta'\}$. A two–dimensional joint confidence set of level $1-\alpha$ for $(\mu(x_0, z_0), \mu(x_0, z_1))$ is given by:

$$\{(e^{\theta}/(1+e^{\theta}), e^{\theta'}/(1+e^{\theta'})) : (\theta, \theta') \in \tilde{S}_{\tilde{\psi}_0(z_0),\tilde{\psi}_0(z_1)}\}\,.$$

This method can be extended to provide confidence sets at more than 2 points. However, for a fixed sample size, the performance of this procedure will deteriorate as the number of z_i's increases, owing to the finite sample dependence among the pointwise likelihood ratio statistics.

Confidence sets for the finite dimensional regression parameter β_0 can be constructed in the usual fashion as: $\{\beta : \mathrm{lrtbeta}_n(\beta) \le q_{\chi^2_d, 1-\alpha}\}$, where $\mathrm{lrtbeta}_n(\beta)$ is the likelihood ratio statistic for testing the null hypothesis that the true regression parameter is β (see (2.7)), and $q_{\chi^2_d, 1-\alpha}$ is the $(1-\alpha)$'th quantile of the χ^2_d distribution. This method can be adapted to construct confidence sets for a sub–vector of the regression parameters as well. So, consider a situation where the regression parameter vector can be partitioned as $\beta = (\eta_1, \eta_2)$. Let $\beta_0 = (\eta_{10}, \eta_{20})$ denote the true parameter value and suppose that we are interested in a confidence set for η_{10}. Let d_1 and d_2 denote the dimensions of η_1 and η_2 respectively. To test $\overline{H}_0 : \eta_1 = \eta_{10}$, the log–likelihood function $l_n(\beta, \psi)$ is maximized over all β of the form (η_{10}, η_2) (where η_2 varies freely in $\mathbb{R}^{d_2}$) and ψ monotone increasing. If we identify ψ, as before, with the vector $\mathbf{u} = \{\psi(Z_{(i)})\}_{i=1}^n$, then, $\overline{g}(\eta_2, \mathbf{u}) \equiv -l_n((\eta_{10}, \eta_2), \psi)$ is a continuously differentiable strictly convex function defined on $\mathbb{R}^{d_2} \times \mathcal{C}$ and its minimizer can be obtained using Proposition 1. If $(\hat{\eta}_2, \hat{\mathbf{u}}_n^{(\eta_{10})})$ denotes the minimizer of $\overline{g}$, then the constrained MLEs of (β, ψ) under $\overline{H}_0$ are: $((\eta_{10}, \hat{\eta}_2), \hat{\psi}_n^{(\eta_{10})})$ where $\hat{\psi}_n^{(\eta_{10})})$

is the (unique) right–continuous increasing step function that assumes the value $\hat{u}_{i,n}^{(\eta_{10})}$ (the i'th component of $\hat{\mathbf{u}}_n^{(\eta_{10})}$) at the point $Z_{(i)}$ and has no jump points outside of the set $\{Z_{(i)}\}_{i=1}^n$. The likelihood ratio statistic for testing $\overline{H}_0$ is then given by:

$$2\,[l_n(\hat{\beta}_n, \hat{\psi}_n) - l_n((\eta_{10}, \hat{\eta}_2), \hat{\psi}_n^{(\eta_{10})})]$$

and converges to the $\chi^2_{d_1}$ distribution. Therefore, a level $1-\alpha$ confidence set for the sub–vector η_{10} can be readily computed via inversion and calibration using $\chi^2_{d_1}$ quantiles. We skip the details.

General Link Functions: Under regularity conditions analogous to those described at the beginning of this section, similar results are obtained for more general link functions, so long as the inverse link $\tilde{h}$ satisfies Condition (C) described in Section 2. Thus, for any $\tilde{h}$ satisfying the concavity constraints, (a) the likelihood ratio statistic for testing $\beta = \beta_0$ as described in (2.4) converges to a χ^2_d distribution *and* (b) the likelihood ratio statistic for testing $\tilde{H}_0 : \psi(z_0) = \theta_0$ as described in (2.6) converges in distribution to $\mathbb{D}$ (when $\tilde{H}_0$ is true). Confidence sets for $\beta, \psi(z_0)$ and $\mu(x_0, z_0)$ as well as a sub–vector of β may be obtained by methods analogous to those used in the logistic regression framework. Once again, owing to space constraints, we skip the details.

4 APPENDIX

Proof of Theorem 3.1: For simplicity, we assume that X is 1–dimensional, so that β is also 1–dimensional. This will make the proof easier to understand, without any essential loss of generality. Also, in what follows, $P_{\beta,\Lambda}$ will denote the distribution of (Δ, Z, X) under parameter value (β, Λ) and $p_{\beta,\Lambda}$ the corresponding density. Also p_0 denotes the density under the true values (β_0, Λ_0) and $P_0 \equiv P_{\beta_0,\Lambda_0}$.

The consistency of $(\hat{\beta}_n, \hat{\Lambda}_n)$ for (β_0, Λ_0) and of $\hat{\Lambda}_{n,0}$ for Λ_0 can be established via a standard consistency proof. Here we only provide a sketch. We can use the method of Wald (see, for example, Theorem 5.14 of Van der Vaart (1998)) with criterion function $m_{\beta,\Lambda} = \log{(p_{\beta,\Lambda} + p_{\beta_0,\Lambda_0})}/2$. It is not difficult to see that $\mathbb{P}_n(m_{\hat{\beta}_n,\hat{\Lambda}_n}) \geq \mathbb{P}_n(m_{\beta,\Lambda})$ for any (β, Λ) and $P_0(m_{\beta_0,\Lambda_0}) \geq P_0(m_{\beta,\Lambda})$ for any (β, Λ) where P_0 is the distribution under (β_0, Λ_0). Equipping the parameter space with the Euclidean topology on $\mathbb{R}^d$ times the weak topology on the space of bounded cadlag functions defined on $[0, \tau]$ renders it compact. Furthermore, the criterion functions are uniformly bounded. Conditions (5.12) and (5.13) on page 48 of Van der Vaart (1998) are then easily verified and Theorem 5.14 can be invoked to conclude that $\hat{\beta}_n$ converges in probability to β_0 under the usual Euclidean topology and that $\hat{\Lambda}_n$ converges to Λ_0 in the weak topology on the interval $[\sigma, \tau]$ (on which Λ_0 is identifiable). Pointwise convergence of $\hat{\Lambda}_n$ to Λ_0 (in probability) for any $\sigma < x < \tau$ is easily deduced. Invoking the uniform continuity of Λ_0 on any compact interval strictly contained in $[\sigma, \tau]$ along with the monotonicity of the functions $\hat{\Lambda}_n$ and Λ_0 allows us to strengthen this to convergence in probability under the topology of uniform convergence on compact subsets of $[\sigma, \tau]$.

In what follows, we will freely use the facts that (a) $(\hat{\beta}_n, \hat{\Lambda}_n)$ converges to (β_0, Λ_0) in the product topology $\tau_1 \times \tau_2$ where τ_1 is the Euclidean topology on $\mathbb{R}^d$ and τ_2 is the topology of uniform convergence on compact sets. Also, $\hat{\Lambda}_{n,0}$ converges to Λ_0 under τ_2. Another result that we will need is a rate of convergence of $\hat{\Lambda}_n$ and $\hat{\Lambda}_{n,0}$ to Λ_0 in an L_2 metric. We have:

$$\int_\sigma^\tau (\hat{\Lambda}_n(u) - \Lambda_0(u))^2\, du = O_p(n^{-2/3}) \text{ and } \int_\sigma^\tau (\hat{\Lambda}_{n,0}(u) - \Lambda_0(u))^2\, du = O_p(n^{-2/3})\,.$$

This can be deduced by using arguments similar to those in Section A.3 of Murphy and Van der Vaart (1997) for the Cox model with interval censored data, where a rate of convergence for the MLE of Λ is deduced from the rate of convergence of of density estimators with respect to (an appropriate modification of) the Hellinger distance. See also Theorem 3.3 of Huang (1996).

We will use Theorem 3.1 of Murphy and Van der Vaart (1997) to establish that $\mathrm{lrtbeta}_n$, the likelihood ratio statistic for testing $\beta = \beta_0$ is asymptotically χ_1^2 and has the representation (3.12); in the process, we will establish the asymptotically linear representation of the MLE $\hat{\beta}$ in the efficient score function (display (3.10)). We start by constructing the "approximately least favorable" one–dimensional submodels that satisfy conditions (3.6) and (3.7) on Page 1482 of Murphy and Van der Vaart (1997). For parameter values (β, Λ) we define the corresponding submodel

$$s_t(\beta, \Lambda) = (t, \Lambda_t(\beta, \Lambda)) \equiv (t, \Lambda + (\theta - t)\, \phi(\Lambda)\, h^{\star\star} \circ \Lambda_0^{-1} \circ \Lambda)\,.$$

Here ϕ is a function mapping $[0, M]$ into $[0, \infty)$, such that (i) $\phi(y) = y$ on $[\Lambda_0(\sigma), \Lambda_0(\tau)]$, (ii) $y \mapsto \phi(y)/y$ is Lipschitz and (iii) $\phi(y) \leq c\,(y \wedge M - y)$ for a sufficiently large constant c that depends on (θ_0, Λ_0) only. By the assumption that $[\Lambda_0(\sigma-), \Lambda_0(\tau)] \subset (0, M)$ such a function ϕ exists. It is not difficult to show that for t sufficiently close to θ, $(t, \Lambda_t(\theta, \Lambda))$ is a valid parameter.

We next compute the scores from these approximately least favorable submodels. Let $p(\delta, z, x; t, \Lambda, \beta)$ denote the density function under parameter value $(t, \Lambda_t(\beta, \Lambda))$ and $l(\delta, z, x; t, \Lambda, \beta)$ denote the log–density. We have

$$l(\delta, x, z; t, \Lambda, \beta) = \delta\, \log\, \Lambda_t(\beta, \Lambda)(z) + \delta\, t\, x - \log\,(1 + \Lambda_t(\beta, \Lambda)(z)\, e^{t\,x})\,.$$

Straightforward computations yield that the score function is:

$$\dot{l}(\delta, x, z; t, \Lambda, \beta) = \frac{\delta}{\Lambda_t(\beta, \Lambda)(z)}\, [-\phi(\Lambda(z))\, h^{\star\star} \circ \Lambda_0^{-1} \circ \Lambda(z)] + \delta\, x$$
$$-\frac{1}{1 + \Lambda_t(\beta, \Lambda)(z)\, e^{t\,x}}\, \left[-\phi(\Lambda(z))\, h^{\star\star} \circ \Lambda_0^{-1} \circ \Lambda(z)\, e^{tx} + \Lambda_t(\beta, \Lambda)(z)\, x\, e^{tx}\right]\,.$$

As (t, β, Λ) converges to $(\beta_0, \beta_0, \Lambda_0)$ (the convergence is taken to be with respect to the product of the Euclidean topology on $\mathbb{R}^2$ and the topology of uniform convergence on compact subsets of $[\sigma, \tau]$), it is easy to verify that

$$\dot{l}(\delta, x, z; t, \Lambda, \beta) \to \frac{\delta}{\Lambda_0(z)}\, [-\Lambda_0(z)\, h^{\star\star}(z)] + \delta\, x - \frac{1}{1 + \Lambda_0(z)\, e^{\beta_0\, x}}\, [-\Lambda_0(z)\, h^{\star\star}(z) + \Lambda_0(z)\, x\, e^{\beta_0\, x}]\,,$$

almost everywhere with respect to P_0, the true measure and this is precisely $\tilde{l}(\delta, x, z)$, the efficient score function (as can be checked by arranging terms).

Next, we note that the class of functions $\dot{l}(\delta, x, z; t, \Lambda, \beta)$ with β and t ranging in a neighborhood of β_0 and Λ ranging over the class of increasing cadlag functions on $[0, \tau]$ and taking values in $[0, M]$ is a uniformly bounded and Donsker class. We can write the score function as:

$$-\delta\left[\frac{\phi(\Lambda(z))/\Lambda(z)}{1+(\beta-t)\,(\phi(\Lambda(z))/\Lambda(z))\,h^{\star\star}\circ\Lambda_0^{-1}\circ\Lambda(z)}\right]+\delta\,x-\frac{\Lambda_t(\beta,\Lambda)(z)\,x\,e^{tx}-\phi(\Lambda(z))\,h^{\star\star}\circ\Lambda_0^{-1}\circ\Lambda(z)\,e^{tx}}{1+\Lambda_t(\beta,\Lambda)(z)\,e^{tx}}\,.$$

In the wake of standard preservation properties of Donsker classes of functions (see pages 192 – 193 of Van der Vaart and Wellner (1996)) it suffices to show that each of the above three components is uniformly bounded and Donsker. Conside the first component. Since the class of functions $\{\Lambda\}$ being considered here is uniformly bounded and is contained in the class of monotone functions on $[\sigma, \tau]$, Theorem 2.7.5. of Van der Vaart and Wellner (1996) can be invoked (with minor modification) to conclude that this is a universally Donsker class. Since $y \mapsto \phi(y)/y$ is bounded and Lipschitz, the composition of this function with the class $\{\Lambda\}$, i.e. the class $\{\phi(\Lambda)/\Lambda\}$ is bounded and Donsker, by Theorem 2.10.6 of Van der Vaart and Wellner (1996). Next, using (A.4) and (A.6), it is easily verfied that the function $h^{\star\star}\cdot\Lambda_0^{-1}$ has a uniformly bounded derivative and is therefore Lipschitz; it follows then that $h^{\star\star}\cdot\Lambda_0^{-1}\cdot\Lambda$ is bounded and Donsker. Since the product of a number of uniformly bounded Donsker classes is Donsker and addition of constants preserves the Donsker property, conclude that $\{1+(\beta-t)\,(\phi(\Lambda(z))/\Lambda(z))\,h^{\star\star}\cdot\Lambda_0^{-1}\cdot\Lambda(z)\}$ is a Donsker class. Furthermore if (β, t) vary in a sufficiently small neighborhood of β_0, then

$$1+(\beta-t)\,(\phi(\Lambda(z))/\Lambda(z))\,h^{\star\star}\cdot\Lambda_0^{-1}\cdot\Lambda(z)\geq 1-\mid\beta-t\mid\;G\,\|h^{\star\star}\cdot\Lambda_0^{-1}\|_\infty\geq 1-\epsilon>0\,,$$

if $\mid\beta-t\mid<\epsilon/(G\,\|h^{\star\star}\cdot\Lambda_0^{-1}\|_\infty)$ where $G=\sup\mid\phi(y)/y\mid$ and $1>\epsilon>0$ is preassigned. It follows that $\{(1+(\beta-t)\,(\phi(\Lambda(z))/\Lambda(z))\,h^{\star\star}\cdot\Lambda_0^{-1}\cdot\Lambda(z))^{-1}\}$ is also bounded and Donsker. Conclude, using preservation properties yet again, that the first term is a bounded Donsker class.

To show that the third term is Donsker, note that each of the following classes $\{\phi(\Lambda)\}$, $\{\Lambda_t(\beta,\Lambda)\}$, $\{h^{\star\star}\cdot\Lambda_0^{-1}\cdot\Lambda\}$, $\{e^{tx}\}$ are bounded Donsker classes as t and β range in a bounded neighborhood of β_0 and Λ ranges in the class of nondecreasing cadlag functions on $[0, M]$; also the class $\{(1+\Lambda_t(\beta,\Lambda)(z)\,e^{tx})^{-1}\}$ is bounded above by 1. Now, employ standard preservation properties to arrive at the desired conclusion. The second term is of course a fixed bounded function, hence Donsker. Thus, we conclude that the class of score functions considered above is indeed Donsker and uniformly bounded.

Furthermore, the class

$$\mathcal{F}=\{g(\delta,z,x;t,\Lambda,\beta)=(1/p(\delta,z,x;t,\Lambda,\beta))\,(\partial^2/\partial t^2)\,p(\delta,z,x;t,\Lambda,\beta)\}$$

as (β, t, Λ) vary in a sufficiently small neighborhood of $(\beta_0, \beta_0, \Lambda_0)$ is a Glivenko–Cantelli class of functions (this can be shown by employing techniques similar to those used to show that the score functions above are Donsker). Furthermore $g(\delta, z, x; t, \Lambda, \beta) \to g(\delta, z, x; \beta_0, \Lambda_0, \beta_0)$ almost surely under P_0 (the true underlying distribution generating the data), as $(t, \Lambda, \beta) \to (\beta_0, \Lambda_0, \beta_0)$. It now follows from the discussion on Pages 71-77 of Banerjee (2000) that Condition (3.8) on Page 1482 of Murphy and Van der Vaart (1997) is satisfied.

Next, we need to verify the "unbiasedness condition" – this is condition (3.9) of Murphy and Van der Vaart (1997) and in this case can be written as:

$$\sqrt{n}\,\mathbb{P}_n\,(\dot{l}(\delta, x, z; \beta_0, \hat{\Lambda}_n^{(\beta_0)}, \beta_0) - \tilde{l}) \to_p 0\,.$$

By the consistency of $\hat{\Lambda}_n^{(\beta_0)}$ for Λ_0 and the facts that (i) the class of score functions obtained from the least favorable submodels are Donsker and (ii) $P_0\,\tilde{l} = 0$, the above condition is equivalent to $\sqrt{n}\,P_0\,(\dot{l}(\delta, x, z; \beta_0, \hat{\Lambda}_n^{(\beta_0)}, \beta_0)) \to_p 0$. For notational convenience, abbreviate $\dot{l}(\delta, x, z; \beta_0, \Lambda, \beta_0)$ to $\dot{l}(\Lambda)$ and $\hat{\Lambda}_n^{(\beta_0)}$ to $\hat{\Lambda}_0$. We have,

$$P_0\left(\dot{l}(\hat{\Lambda}_0)\right) = (P_0 - P_{\beta_0,\hat{\Lambda}_0})(\dot{l}(\Lambda_0)) + (P_0 - P_{\beta_0,\hat{\Lambda}_0})(\dot{l}(\hat{\Lambda}_0) - \dot{l}(\Lambda_0))\,. \tag{4.14}$$

To write this decomposition we use the fact that $P_{\beta,\Lambda}(\dot{l}(.; \beta, \Lambda, \beta)) = 0$ for all (β, Λ). It now suffices to show that each of the two terms on the right side of the above display is $o_p(n^{-1/2})$. Consider the first term. This can be written as:

$$(P_0 - P_{\beta_0,\hat{\Lambda}_0})(\dot{l}(\Lambda_0)) = P_0\left\{(\dot{l}(\Lambda_0))\left[\frac{p_0 - p_{\beta_0,\hat{\Lambda}_0}}{p_0} - \dot{l}_\Lambda(\beta_0, \Lambda_0)(\Lambda_0 - \hat{\Lambda}_0)\right]\right\}\,. \tag{4.15}$$

Here we are using the fact that $\dot{l}(\Lambda_0)$ is the efficient score function and hence orthogonal to all functions in the span of $\dot{l}_\Lambda(\beta_0, \Lambda_0)$. We now simplify the expression

$$\frac{p_0 - p_{\beta_0,\hat{\Lambda}_0}}{p_0} - \dot{l}_\Lambda(\beta_0, \Lambda_0)(\Lambda_0 - \hat{\Lambda}_0)$$

using a simple Taylor expansion. We have:

$$\begin{aligned} p(\beta_0, \hat{\Lambda}_0) &= p(\beta_0, \Lambda_0 + (\hat{\Lambda}_0 - \Lambda_0)) \\ &= p(\beta_0, \Lambda_0) + \left\{\frac{d}{dt}p(\beta_0, \Lambda_0 + t(\hat{\Lambda}_0 - \Lambda_0))\right\}_{t=0} + \frac{1}{2}\frac{d^2}{dt^2}p(\beta_0, \Lambda_0 + t^*(\hat{\Lambda}_0 - \Lambda_0)) \end{aligned}$$

for some $0 < t^* < 1$. Here t^* can depend on the point of evaluation of the density (note that the arguments (δ, x, z) are being suppressed). But

$$(d/dt)\{p(\beta_0, \Lambda_0 + t(\hat{\Lambda}_0 - \Lambda_0))\}_{t=0} = \dot{l}_\Lambda(\beta_0, \Lambda_0)(\hat{\Lambda}_0 - \Lambda_0) \times p(\beta_0, \Lambda_0)$$

and using this we get:

$$\frac{p(\beta_0, \Lambda_0) - p(\beta_0, \hat{\Lambda}_0)}{p(\beta_0, \Lambda_0)} = \dot{l}_\Lambda(\beta_0, \Lambda_0)(\Lambda_0 - \hat{\Lambda}_0) - \frac{1}{p(\beta_0, \Lambda_0)}\frac{1}{2}\frac{d^2}{dt^2}p(\beta_0, \Lambda_0 + t^*(\hat{\Lambda}_0 - \Lambda_0))\,.$$

Now using (4.15) it follows that

$$(P_0 - P_{\beta_0,\hat{\Lambda}_0})\left[\dot{l}(\Lambda_0)\right] = -P_0\left[\dot{l}(\Lambda_0)\frac{1}{p_0}\frac{1}{2}\frac{d^2}{dt^2}p(\beta_0,\Lambda_0 + t^*(\hat{\Lambda}_0 - \Lambda_0))\right]. \tag{4.16}$$

We now compute $\left(d^2/dt^2\right) p(\beta_0,\Lambda_0 + t^* h)$ where h denotes $\hat{\Lambda}_0 - \Lambda_0$. Straightforward differentiation yields that

$$\begin{aligned}
\frac{d^2}{dt^2}p(\beta_0,\Lambda_0 + th) &= \frac{d}{dt}\left[\frac{d}{dt}\left[\delta + (1 - 2\,\delta)\,\frac{1}{1 + e^{\beta_0\,x}(\Lambda_0 + th)(z)}\right]\right] \\
&= (1 - 2\,\delta)\,e^{2\,\beta_0\,x}\,h^2(z)\,\frac{2}{(1 + e^{\beta_0\,x}(\Lambda_0 + th)(z))^3}\,.
\end{aligned}$$

In writing $p(\beta_0,\Lambda_0 + th)$ above, the joint density of (X,Z) has been absorbed into the dominating measure; call the resulting measure μ_{dom}. From the above display, easily we have:

$$\left|\frac{d^2}{dt^2}p(\beta_0,\Lambda_0 + th)\right| \;\leq\; h^2 K \tag{4.17}$$

for all t (where K is a constant not depending upon t), provided that $\Lambda_0 + t\,h \geq 0$. This is indeed the case for $h = \hat{\Lambda}_0 - \Lambda_0$. Thus we get,

$$\begin{aligned}
\mid (P_0 - P_{\beta_0,\hat{\Lambda}_0})(\dot{l}(\Lambda_0)) \mid &\leq \int \mid \dot{l}\,(\Lambda_0) \mid \frac{1}{p_0}\left|\frac{d^2}{dt^2}p(\beta_0,\Lambda_0 + t^*(\hat{\Lambda}_0 - \Lambda_0))\right| p_0\, d\mu_{dom}(\delta,x,z) \\
&\leq K' \int (\hat{\Lambda}_0 - \Lambda_0)^2(z) d\mu_{dom}(\delta,x,z) \\
&= K'' \int_\sigma^\tau (\hat{\Lambda}_0 - \Lambda_0)^2(z) f_Z(z) dz\,.
\end{aligned}$$

Here the K's are constants. Since f_Z is continuous on $[\sigma,\tau]$ it attains its maximum, showing that

$$\left|(P_0 - P_{\beta_0,\hat{\Lambda}_0})[\dot{l}(\Lambda_0)]\right| \leq K'' \int_\sigma^\tau (\hat{\Lambda}_0 - \Lambda_0)^2(z) f_Z(z) dy \leq K''' \int_\sigma^\tau (\hat{\Lambda}_0 - \Lambda_0)^2(z) dy\,.$$

Now, $\int_\sigma^\tau (\hat{\Lambda}_0 - \Lambda_0)^2(z)dz$ is $O_p(n^{-2/3})$, and hence certainly $o_p(n^{-1/2})$ showing that so is the left side of the above display. Consider the second term in (4.14). To tackle this term, note first of all, that

$$\mid \dot{l}(\hat{\Lambda}_0) - \dot{l}(\Lambda_0) \mid \;\leq\; C \mid \hat{\Lambda}_0 - \Lambda_0 \mid \tag{4.18}$$

for some constant C not depending on δ, x, z. Now write the second term as:

$$\begin{aligned}
\int (\dot{l}(\hat{\Lambda}_0) - \dot{l}(\Lambda_0))(p_0 - p_{\beta_0,\hat{\Lambda}_0}) d\mu_{dom}(\delta,z,x) &= \int (\dot{l}(\hat{\Lambda}_0) - \dot{l}(\Lambda_0))[l_\Lambda(\beta_0,\Lambda_0)(\Lambda_0 - \hat{\Lambda}_0)\,p(\beta_0,\Lambda_0) \\
&\qquad - \frac{1}{2}\frac{d^2}{dt^2}p(\beta_0,\Lambda_0 + t^*(\hat{\Lambda}_0 - \Lambda_0))] d\mu_{dom} \\
&= \int (\dot{l}(\hat{\Lambda}_0) - \dot{l}(\Lambda_0)) l_\Lambda(\beta_0,\Lambda_0)(\Lambda_0 - \hat{\Lambda}_0)\,p(\beta_0,\Lambda_0)\,d\mu_{dom} \\
&\qquad - \int (\dot{l}(\hat{\Lambda}_0) - \dot{l}(\Lambda_0))\frac{1}{2}\frac{d^2}{dt^2}p(\beta_0,\Lambda_0 + t^*(\hat{\Lambda}_0 - \Lambda_0)) d\mu_{dom}\,.
\end{aligned}$$

Now $l_\Lambda(\beta_0,\Lambda_0)(\Lambda_0 - \hat{\Lambda}_0) = (\Lambda_0 - \hat{\Lambda}_0)Q(x;\beta_0,\Lambda_0)$. Using this along with the fact that $Q(x;\beta_0,\Lambda_0)$ is bounded, the bound on $\frac{1}{2}\frac{d^2}{dt^2}p(\beta_0,\Lambda_0 + t^*(\hat{\Lambda}_0 - \Lambda_0))$ obtained through (4.17), the inequality (4.18) and the uniform boundedness of $\mid \dot{l}(\hat{\Lambda}_0) - \dot{l}(\Lambda_0) \mid$ and the boundedness of the density of Y, we find that the sum of the absolute values of the integrals in the last expression of the above display is bounded by:

$$C_1 \int_\sigma^\tau (\hat{\Lambda}_0 - \Lambda_0)^2 dy + C_2 \int_\sigma^\tau (\hat{\Lambda}_0 - \Lambda_0)^2 dy$$

showing that

$$\left| \int (\dot{l}(\hat{\Lambda}_0) - \dot{l}(\Lambda_0))(p_0 - p_{\beta_0,\hat{\Lambda}_0}) d\mu_{dom}(\delta,z,x) \right| \le C_3 \int_\sigma^\tau (\hat{\Lambda}_0 - \Lambda_0)^2 dy$$

and as before we conclude that $\int (\dot{l}(\hat{\Lambda}_0) - \dot{l}(\Lambda_0))(p_0 - p_{\beta_0,\hat{\Lambda}_0}) d\mu_{dom}(\delta,z,x)$ is $o_p(n^{-1/2})$. This completes the proof that the unbiasedness condition holds.

It only remains to show the asymptotic linearity of the MLE in the efficient score function – in other words, establishing the representation (3.17). (3.10). Since $\hat{\beta}$ maximizes the function

$$t \mapsto \mathbb{P}_n \log p(t, \Lambda_t(\hat{\beta},\hat{\Lambda}))$$

over t, it follows that:

$$\mathbb{P}_n(\dot{l}(\cdot;\hat{\beta},\hat{\Lambda},\hat{\beta})) = 0 \,.$$

Let $\mathbb{G}_n$ denote the empirical process $\sqrt{n}\,(\mathbb{P}_n - P_0)$. By the Donsker property of the class of functions $\dot{l}(\cdot;t,\Lambda,\beta)$ and the consistency of $(\hat{\beta},\hat{\Lambda})$ we have that:

$$\mathbb{G}_n(\dot{l}(\cdot;\hat{\beta},\hat{\Lambda},\hat{\beta}) - \dot{l}(\cdot;\beta_0,\Lambda_0,\beta_0)) \to_P 0 \,.$$

The preceding two displays then jointly imply that:

$$-\sqrt{n}P_0(\dot{l}(\cdot;\hat{\beta},\hat{\Lambda},\hat{\beta})) = \mathbb{G}_n(\dot{l}(\cdot;\beta_0,\Lambda_0,\beta_0)) + o_p(1) \,.$$

By exactly the same arguments as used to show that $\sqrt{n}\,P_0\,\dot{l}(\cdot;\beta_0,\hat{\Lambda}_0,\beta_0) \to_P 0$ we can show that $\sqrt{n}\,P_0\,\dot{l}(\cdot;\beta_0,\hat{\Lambda},\beta_0) \to_P 0$; we need to replace $\hat{\Lambda}_0$ everywhere in that chain of arguments by $\hat{\Lambda}$ and use that $\int_\sigma^\tau (\hat{\Lambda} - \Lambda_0)^2(y)dy$ is $O_p(n^{-2/3})$. Adding $\sqrt{n}\,P_0\,\dot{l}(\cdot;\beta_0,\hat{\Lambda},\beta_0)$ to the left side of the preceding display and using the fact that it converges in probability to 0 gives

$$-\sqrt{n}P_0(\dot{l}(\cdot;\hat{\beta},\hat{\Lambda},\hat{\beta})) + \sqrt{n}P_0(\dot{l}(\cdot;\beta_0,\hat{\Lambda},\beta_0)) = \mathbb{G}_n(\dot{l}(\cdot;\beta_0,\Lambda_0,\beta_0)) + o_p(1) \,. \tag{4.19}$$

Now define $\ddot{\kappa}(\cdot;t,\Lambda) = \frac{\partial}{\partial t}\dot{l}(\cdot;t,\Lambda,t)$. Then,

$$\begin{aligned} P_0\,(\ddot{\kappa}(\cdot;t,\Lambda)) &= P_0\left[(\partial/\partial t)\dot{l}(\cdot;t,\Lambda,t)\right] \\ &= (\partial/\partial t)P_0\left[\dot{l}(\cdot;t,\Lambda,t)\right] \,. \end{aligned}$$

Denote the left side of (4.19) by L. Then

$$\begin{aligned}
L &= -\sqrt{n}\left[P_0\left(\dot{l}(\cdot;\hat{\beta},\hat{\Lambda},\hat{\beta})\right) - P_0\left(\dot{l}(\cdot;\beta_0,\hat{\Lambda},\beta_0)\right)\right] \\
&= -\sqrt{n}\left[\frac{\partial}{\partial t}P_0\left(\dot{l}(\cdot;\tilde{\beta},\hat{\Lambda},\tilde{\beta})\right)\right]\left(\hat{\beta}_n - \beta_0\right) \\
&= -\sqrt{n}\left[P_0\left(\frac{\partial}{\partial t}\dot{l}(\cdot;t,\hat{\Lambda},t)\mid_{t=\tilde{\beta}}\right)\right]\left(\hat{\beta}_n - \beta_0\right) \\
&= -\sqrt{n}\left[P_0\left(\ddot{\kappa}(\cdot;\tilde{\beta},\hat{\Lambda})\right)\right]\left(\hat{\beta}_n - \beta_0\right)
\end{aligned}$$

where $\tilde{\beta}$ lies between $\hat{\beta}_n$ and β_0 and does not depend on (δ, z, x). We now claim that:

$$\left[P_0\left(\ddot{\kappa}(\cdot;\tilde{\beta},\hat{\Lambda})\right)\right] \quad \to_P \quad -I_0\,,$$

as $(\hat{\beta}_n, \hat{\Lambda}_n) \to_P (\beta_0, \Lambda_0)$. This is proved in the following way. Denote the quantity on the left hand side of the above display by M_n. It is easy to check that $\ddot{\kappa}(\cdot;t,\Lambda)$ is uniformly bounded as (t,Λ) range in a finite neighborhood of (β_0,Λ_0). Also $\ddot{\kappa}(\cdot;t,\Lambda) \to \ddot{\kappa}(\cdot;\beta_0,\Lambda_0)$ for P_0 almost every x as $(t,\Lambda) \to (\beta_0,\Lambda_0)$. Hence by the DCT $P_0\left(\ddot{\kappa}_i(\cdot;t,\Lambda)\right) \to P_0\left(\ddot{\kappa}(\cdot;\beta_0,\Lambda_0)\right)$ and consequently $P_0\left(\ddot{\kappa}(\cdot;\tilde{\beta},\hat{\Lambda}_n)\right) \to_P P_0\left(\ddot{\kappa}(\cdot;\beta_0,\Lambda_0)\right)$. Thus $M_n \to_P P_0\left(\ddot{\kappa}(\cdot;\beta_0,\Lambda_0)\right)$. We now need to show that this equals $-I_0$. Now,

$$\int \dot{l}(\cdot;\beta,\Lambda,\beta)\,p(\cdot;\beta,\Lambda)\,d\mu_{dom} = 0$$

for all (β,Λ). Differentiating this relation with respect to β gives:

$$\begin{aligned}
0 &= \int \frac{\partial}{\partial\beta}\left(\dot{l}(\cdot;\beta,\Lambda,\beta)\,p(\cdot;\beta,\Lambda)\right)d\mu_{dom} \\
&= \int \ddot{\kappa}(\cdot;\beta,\Lambda)\,p(\cdot,\beta,\Lambda)d\mu_{dom} + \int \dot{l}(\cdot;\beta,\Lambda,\beta)\,\dot{l}_{\beta}(\cdot;\beta,\Lambda)p(\cdot;\beta,\Lambda)d\mu_{dom}\,,
\end{aligned}$$

where $\dot{l}_{\beta}(\cdot;\beta,\Lambda)$ is the ordinary score function for β. For $\beta = \beta_0$ and $\Lambda = \Lambda_0$ we then have:

$$\begin{aligned}
\int \ddot{\kappa}(\cdot;\beta_0,\Lambda_0)\,p(\cdot;\beta_0,\Lambda_0)d\mu_{dom} &= -\int \dot{l}(\cdot;\beta_0,\Lambda_0,\beta_0)\,\dot{l}_{\beta}(\cdot;\beta_0,\Lambda_0)^T\,p(\cdot;\beta_0,\Lambda_0)d\mu_{dom} \\
&= -\int \tilde{l}_0\,\left(\tilde{l}_0 + \dot{l}_{\beta}(\cdot;\beta_0,\Lambda_0) - \tilde{l}_0\right)^T p_0\,d\mu_{dom} \\
&= -I_0
\end{aligned}$$

since $\tilde{l}_0 \perp \dot{l}_{\beta}(\cdot;\beta_0,\Lambda_0) - \tilde{l}_0$. This completes the proof of the claim.

Now from (4.19) and the display following it, we have,

$$\begin{aligned}
-\sqrt{n}\,M_n(\hat{\beta}_n - \beta_0) &= \sqrt{n}\,(\mathbb{P}_n - P_0)\,\tilde{l}_0 + o_p(1) \\
&= \sqrt{n}\,\mathbb{P}_n\,\tilde{l}_0 + o_p(1)
\end{aligned}$$

and this can be rewritten as

$$\begin{aligned}\sqrt{n}\left(\hat{\beta}_n - \beta_0\right) &= -M_n^{-1}\sqrt{n}\,\mathbb{P}_n\,\tilde{l}_0 - M_n^{-1}o_p(1)\\ &= I_0^{-1}\sqrt{n}\,\mathbb{P}_n\,\tilde{l}_0 + o_p(1)\end{aligned}$$

by Slutsky's theorem, completing the proof of asymptotic efficiency.

Hence, facts (3.10) and (3.12) are established. The asymptotically linear representation for $\tilde{\beta}_n$ and the limiting χ^2 distribution for $\mathrm{lrtbeta}_n^0$ may be established by similar steps. Some additional care needs to be exercised, since the parameter space for Λ is now restricted by fixing the value at the point z_0. Roughly the intuition is the following: The unconstrained MLE of β, is $\sqrt{n}$–consistent and asymptotically efficient for the given model. The unconstrained likelihood ratio statisitic for testing $\beta = \beta_0$, which we denote by $\mathrm{lrtbeta}_n$ is asymptotically χ^2. These properties will be preserved even when we compute the above statistics under the *single (true) constraint* that $\Lambda(z_0) = \theta_0$. In fact, the same asymptotic representations for the above statistics will continue to hold when we constrain Λ at finitely many points. Note however, that the limit distribution of the MLE will generally be affected under infinitely many constraints on Λ. This is easily seen when we constrain Λ on the support of Z. In this case Λ is completely known and the asymptotic variance of β is the inverse of the ordinary information for θ as opposed to the efficient information. □

Proof–sketch of Theorem 3.2: The proof of this theorem relies on extensive use of "switching relationships" which allow us to translate the behavior of the slope of the convex minorant of a random cumulative sum diagram (this is how the estimators $\hat{\psi}_n^{(\beta_0)}$ and $\hat{\psi}_{n,0}^{(\beta_0)}$ are characterized) in terms of the minimizer of a stochastic process. The limiting behavior of the slope process can then be studied in terms of the limiting behavior of the minimizer of this stochastic process by applying argmin continuous mapping theorems. Switching relationships on the limit process allow interpretation of the behavior of the minimizer of the limit process in terms of the slope of the convex minorant of the limiting versions of the cumulative sum diagrams (appropriately normalized).

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

$$\left(\{U_n(h_i)\}_{i=1}^k, \{V_n(h_i)\}_{i=1}^k\right) \rightarrow_d \left(\{g_{a,b}(h_i)\}_{i=1}^k, \{g_{a,b}^0(h_i)\}_{i=1}^k\right),$$

in the space $\mathbb{R}^{2k}$. Next, to deduce the convergence in $L_2[-K,K] \times L_2[-K,K]$ note firstly that $U_n(h)$ and $V_n(h)$ are monotone functions. Now, given a sequence (ψ_n, ϕ_n) in $L_2[-K,K] \times L_2[-K,K]$ such that ψ_n and ϕ_n are monotone functions and (ϕ_n, ψ_n) converges pointwise to (ϕ, ψ) (where (ϕ, ψ) is in $L_2[-K,K] \times L_2[-K,K]$), we can conclude that $(\psi_n, \phi_n) \rightarrow (\psi, \phi)$ in $L_2[-K,K] \times L_2[-K,K]$. It follows, in the wake of distributional convergence of all the finite - dimensional marginals of (U_n, V_n) to those of $(g_{a,b}(h), g_{a,b}^0(h))$, that

$$(U_n(h), V_n(h)) \rightarrow_d (g_{a,b}(h), g_{a,b}^0(h))$$

in $L_2[-K,K] \times L_2[-K,K]$ (this parallels the result of Corollary 2 following Theorem 3 of Huang and Zhang (1994)).

In the remainder of this proof we will sketch the proof of convergence of $U_n(h)$ to $g_{a,b}(h)$ for any h; the general proof of finite–dimensional convergence is cumbersome to write out and contains minor extensions of the ideas expounded here. In what follows, we denote $\hat{\psi}_n^{(\beta_0)}$ by $\tilde{\psi}$. For a fixed ψ we define the following processes:

$$W_{n,\psi}(r) = \mathbb{P}_n \left[\left(\Delta - \frac{\exp(\psi(Z) + \beta_0^T X)}{1 + \exp(\psi(Z) + \beta_0^T X)} \right) 1(Z \leq r) \right] ,$$

$$G_{n,\psi}(r) = \mathbb{P}_n \left[\left(\frac{\exp(\psi(Z) + \beta_0^T X)}{(1 + \exp(\psi(Z) + \beta_0^T X))^2} \right) 1(Z \leq r) \right] ,$$

and

$$B_{n,\psi}(r) = W_{n,\psi}(r) + \int_0^r \psi(z)\, d\, G_{n,\psi}(z) \,.$$

We will denote by W_n, G_n, B_n the above processes when $\psi = \tilde{\psi}$.

We can now use "the switching relationship" for the unconstrained MLE $\tilde{\psi}(z)$ to get:

$$\tilde{\psi}(z) \leq a \Leftrightarrow \text{argmin}_{r \geq 0} \left[B_n(r) - a\, G_n(r) \right] \geq Z_z \tag{4.20}$$

where Z_z is the largest Z value not exceeding z. By argmin we denote the largest element in the set of minimizers. This can be chosen to be one of the Z_i's. The above equivalence is a direct characterization of the fact that the vector $\{\tilde{\psi}(Z_{(i)})\}_{i=1}^n$ is the vector of slopes (left–derivatives) of the cumulative sum diagram formed by the points $\{G_n(Z_{(i)}), B_n(Z_{(i)})\}_{i=0}^n$, computed at the points $\{G_n(Z_{(i)})\}_{i=1}^n$. The easiest way to verify this is by drawing a picture.

Now, $U_n(h_0) = n^{1/3}\, (\tilde{\psi}(z_0 + h_0\, n^{-1/3}) - \psi_0(z_0))$. We want to find

$$\lim_{n\to\infty} P\, (n^{1/3}\, (\tilde{\psi}(z_0 + h_0\, n^{-1/3}) - \psi_0(z_0)) \leq x) \,.$$

Now, define

$$A_n = \{ n^{1/3}\, (\tilde{\psi}(z_0 + h_0\, n^{-1/3}) - \psi_0(z_0)) \leq x \} \,.$$

Consider the event A_n. We have

$$\begin{aligned} n^{1/3}\, (\tilde{\psi}(z_0 + h_0\, n^{-1/3}) - \psi_0(z_0)) \leq x \quad & \Leftrightarrow \quad \tilde{\psi}(z_0 + h_0\, n^{-1/3}) \leq \psi_0(z_0) + x\, n^{-1/3} \\ & \Leftrightarrow \quad \text{argmin}_r \left[B_n(r) - (\psi_0(z_0) + x\, n^{-1/3})\, G_n(r) \right] \geq Z_{(z_0 + h_0\, n^{-1/3})} \\ & \Leftrightarrow \quad \text{argmin}_r \left[V_n(r) - x\, n^{-1/3}\, G_n(r) \right] \geq Z_{(z_0 + h_0\, n^{-1/3})} \,, \end{aligned}$$

where the second bidirectional implication in the above display follows from the first on using (4.20), and $V_n(r) = B_n(r) - \psi_0(z_0)\, G_n(r)$. Thus,

$$\begin{aligned} A_n &= \left\{ n^{1/3} \left(\text{argmin}_r \left[V_n(r) - x\, n^{-1/3}\, G_n(r) \right] - z_0 \right) \geq n^{1/3}\, (Z_{(z_0 + h_0\, n^{-1/3})} - z_0) \right\} \\ &= \left\{ \text{argmin}_h \left[V_n(z_0 + h\, n^{-1/3}) - x\, n^{-1/3}\, G_n(z_0 + h\, n^{-1/3}) \right] \geq h_0 + o_p(1) \right\} \\ &= \left\{ \text{argmin}_h\, \mathbb{M}_n(h) - x\, \mathbb{G}_n(h) \geq h_0 + o_p(1) \right\}\ , \end{aligned}$$

where

$$\mathbb{M}_n(h) = n^{2/3} \left[V_n(z_0 + h\, n^{-1/3}) - V_n(z_0) \right]$$

and

$$\mathbb{G}_n(h) = n^{1/3} \left[G_n(z_0 + h\, n^{-1/3}) - G_n(z_0) \right]\ .$$

The process $\mathbb{M}_n(h) - x\, \mathbb{G}_n(h)$ converges in the space $B_{loc}(\mathbb{R})$ (here $B_{loc}(\mathbb{R})$ is the space of real–valued functions on the real line that are bounded on every compact set and equipped with the topology of uniform convergence on compact sets) to the process $L(h) \equiv \tilde{a}\, W(h) + \tilde{b}\, h^2 - x\, C(z_0)\, h$. Here $\tilde{a} = \sqrt{C(z_0)}$, $\tilde{b} = \psi_0'(z_0)\, C(z_0)/2$ and $W(h)$ is a fixed two-sided Brownian motion process starting from 0. This result is obtained by using the fact that the process $\mathbb{M}_n(h)$ converges to the limiting process $\tilde{a}\, W(h) + \tilde{b}\, h^2$ under the topology of uniform convergence on compact sets. The convergence of $\mathbb{M}_n(h)$ can be deduced from the convergence of the process

$$\tilde{P}_{n,\psi_0}(h) = n^{2/3} \left[(B_{n,\psi_0}(z_0 + h\, n^{-1/3}) - B_{n,\psi_0}(z_0)) - \psi_0(z_0)\, (G_{n,\psi_0}(z_0 + h\, n^{-1/3}) - G_{n,\psi_0}(z_0)) \right]$$

to $\tilde{a}\, W(h) + \tilde{b}\, h^2$ along with the fact that $\sup_{h \in [-M,M]} \mid \tilde{\psi}(z_0 + h\, n^{-1/3}) - \psi_0(z_0) \mid = O_p(n^{-1/3})$ which entails that $\sup_{h \in [-K,K]} \mid \tilde{P}_{n,\psi_0}(h) - \mathbb{M}_n(h) \mid \to_p 0$, for every $K > 0$. Furthermore, the process $\mathbb{G}_n(h)$ converges uniformly in probability on every $[-K, K]$ to the deterministic process $C(z_0)\, h$.

The convergence in distribution of $\text{argmin}_h\, \mathbb{M}_n(h) - x\, \mathbb{G}_n(h)$ to $\text{argmin}_h\, L(h)$ is accomplished by appealing to an appropriate argmin continuous mapping theorem. The key facts that guarantee the convergence of the minimizers are (i) the fact that the limiting process possesses a unique minimizer almost surely and (ii) the minimizers of the finite sample processes are tight. This involves application of an appropriate "rate theorem" for minimizers of stochastic processes (for example Theorem 3.2.5 or Theorem 3.4.1 of Van der Vaart and Wellner (1996)). The computations are tedious but straightforward and skipped here. For a flavor of the key steps involved in establishing tightness, we refer the reader to Section 3.2.3 of Van der Vaart and Wellner (1996) and in particular Example 3.2.15 (current status data) which is naturally related to binary regression.

It follows that

$$\lim_{n \to \infty} P\, (n^{1/3}\, (\tilde{\psi}(z_0 + h_0\, n^{-1/3}) - \psi_0(z_0)) \leq x) = P\, (\text{argmin}_{\mathbb{R}}\, \tilde{a}\, W(h) + \tilde{b}\, h^2 - x\, C(z_0)\, h \geq h_0)\,. \tag{4.21}$$

We now use the switching relationships on the limit process. From the work of Groeneboom (1989) it follows that

$$\mathrm{argmin}_{\mathbb{R}}\, \tilde{a}\, W(h) + \tilde{b}\, h^2 - x\, C(z_0)\, h > h_0 \Leftrightarrow g_{\tilde{a},\tilde{b}}(h_0) < x\, C(z_0)\,,$$

with probability one. Therefore,

$$\lim_{n\to\infty} P\,(n^{1/3}\,(\tilde{\psi}(z_0 + h_0\, n^{-1/3}) - \psi_0(z_0)) \leq x) = P\,(g_{\tilde{a},\tilde{b}}(h_0) < x\, C(z_0))\,.$$

On noting that:

$$\frac{1}{C(z_0)}\, \left(g_{\tilde{a},\tilde{b}}(\cdot), g^0_{\tilde{a},\tilde{b}}(\cdot)\right) \equiv_d (g_{a,b}(\cdot), g^0_{a,b}(\cdot))\,,$$

with a and b as defined in the statement of the theorem (this follows readily from Lemma 3.1). □

4.1 Further Details About The Unconstrained And Constrained MLEs From Section 2

Details of the "self–consistent" characterization of $\hat{\mathbf{u}}_n^{(\beta)}$ in Section 2: As in Section 2, we denote the function $g(\beta, \mathbf{u})$ in the discussion that follows by $\xi(\mathbf{u})$. This is strictly convex in $\mathbf{u} \equiv (u_1, u_2, \ldots, u_n)$ and for simplicity we denote $\hat{\mathbf{u}}_n^{(\beta)}$, its minimizer over the region $\mathcal{C} := \{\mathbf{u} : u_1 \leq u_2 \leq \ldots \leq u_n\}$, by $\hat{\mathbf{u}}$ (suppressing the dependence on n and β). Let $\bigtriangledown_j\, \xi(\mathbf{u})$ denote the j'th partial derivative of ξ with respect to $\mathbf{u}$. Using the Kuhn–Tucker theorem for optimizing a convex function over a closed convex set, we find that $\hat{\mathbf{u}} = (\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_n)$ is uniquely characterized by the conditions:

$$\sum_{j=i+1}^{n} \bigtriangledown_j\, \xi(\hat{\mathbf{u}}) \geq 0 \quad \text{, for} \quad i = 1, 2, \ldots, (n-1) \tag{4.22}$$

and

$$\sum_{j=1}^{n} \bigtriangledown_j\, \xi(\hat{\mathbf{u}}) = 0\,. \tag{4.23}$$

Consider now, the following (quadratic) function $\tilde{\xi}(\mathbf{u}) = \frac{1}{2}\left[\mathbf{u} - \hat{\mathbf{u}} + \mathcal{K}^{-1} \bigtriangledown \xi(\hat{\mathbf{u}})\right]^T \mathcal{K} \left[\mathbf{u} - \hat{\mathbf{u}} + \mathcal{K}^{-1} \bigtriangledown \xi(\hat{\mathbf{u}})\right]$ where $\mathcal{K}$ is some positive definite matrix. Note that $\mathrm{Hess}(\tilde{\xi}) = \mathcal{K}$ which is positive definite; thus $\tilde{\xi}$ is a strictly convex function. It is also finite and continuously differentiable over $\mathbb{R}^n$. Also, $\bigtriangledown \tilde{\xi}(\mathbf{u}) = \mathcal{K}\left(\mathbf{u} - \hat{\mathbf{u}} + \mathcal{K}^{-1} \bigtriangledown \xi(\hat{\mathbf{u}})\right)$. Now, consider the problem of minimizing $\tilde{\xi}$ over $\mathcal{C}$. If $\mathbf{u}^\star$ is the (unique) global minimizer, then necessary and sufficient conditions are given by conditions (4.22) (for $i = 1, 2, \ldots, n-1$) and (4.23), with ξ replaced by $\tilde{\xi}$ and $\hat{\mathbf{u}}$ replaced by $\mathbf{u}^\star$. Now, $\bigtriangledown \tilde{\xi}(\hat{\mathbf{u}}) = \bigtriangledown \xi(\hat{\mathbf{u}})$, so that the vector $\hat{\mathbf{u}} \in \mathcal{C}$ does indeed satisfy the conditions (4.22) (for $i = 1, 2, \ldots, n-1$) and (4.23), with ξ replaced by $\tilde{\xi}$. It follows that $\hat{\mathbf{u}}$ is the unique minimizer of $\tilde{\xi}$ over $\mathcal{C}$, i.e. $\mathbf{u}^\star = \hat{\mathbf{u}}$.

It now suffices to try to minimize $\tilde{\xi}$; of course the problem here is that $\hat{\mathbf{u}}$ is unknown and $\tilde{\xi}$ is defined in terms of $\hat{\mathbf{u}}$. However, an iterative scheme can be developed along the following lines. Choosing $\mathcal{K}$ to be a diagonal matrix with the i, i'th entry being $d_i \equiv \bigtriangledown_{ii}\, \xi(\hat{\mathbf{u}})$ ($\mathcal{K}$

thus defined is a p.d. matrix, since the diagonal entries of the Hessian of ξ at the minimizer $\hat{\mathbf{u}}$, which is a positive definite matrix, are positive), we see that the above quadratic form reduces to $\eta(\mathbf{u})/2$ where $\eta(\mathbf{u}) = \sum_{i=1}^n \left[u_i - \left(\hat{u}_i - \bigtriangledown_i \xi(\hat{\mathbf{u}}) d_i^{-1}\right)\right]^2 d_i$. Thus, $\hat{\mathbf{u}}$ minimizes $\eta(\mathbf{u})$ subject to the constraints that $u_1 \leq u_2 \leq \ldots \leq u_n$ and therefore furnishes the isotonic regression of the function $h(i) = \hat{u}_i - \bigtriangledown_i \xi(\hat{\mathbf{u}})\, d_i^{-1}$ on the ordered set $\{1, 2, \ldots, n\}$ with weight function d_i. From the theory of isotonic regression, it is well known that the solution $\hat{\mathbf{u}} \equiv (\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_n) = \text{slogcm } \left\{\sum_{j=1}^i d_i \, , \, \sum_{j=1}^i h(i)\, d_i\right\}_{i=0}^n$. This representation leads to the MICM as outlined in Section 2.

Implications of the self–consistent/self–induced characterization of $\hat{\mathbf{u}}_n^{(\beta)}$: Recall that $\hat{\mathbf{u}}_n^{(\beta)} = (\hat{u}_{1,n}^{(\beta)}, \hat{u}_{2,n}^{(\beta)}, \ldots, \hat{u}_{n,n}^{(\beta)})$. Let $B_1, B_2, \ldots, B_k$ be the unique partitioning of $1, 2, \ldots, n$ into ordered blocks of indices (say $B_1 = \{1, 2, \ldots, l_1\}, B_2 = \{l_1 + 1, l_1 + 2, \ldots, l_2\}$ and so on) such that, for each i, for all $j \in B_i$, $\hat{u}_{j,n}^{(\beta)}$ equals w_i, with the common block values, the w_i's, satisfying $w_1 < w_2 < \ldots < w_k$. Since the $\hat{u}_{j,n}^{(\beta)}$'s are increasing in j, this is possible. An important consequence of the self–consistent characterization is the fact that each w_i can be written as a weighted average of the $h(j)$'s for the j's in B_i, with the weights given by the d_j's. The B_i's are called the *level blocks* of $\hat{\mathbf{u}}_n^{(\beta)}$ and the w_i's are called the *level values*.

We now introduce some notation that will be useful in the proof of Theorem 3.3. Denote $\phi(\Delta_{(i)}, R_i(\beta), t)$ by $\phi_{i,\beta}(t)$ and its first and second derivatives with respect to t by $\phi'_{i,\beta}(t)$ and $\phi''_{i,\beta}(t)$. Identifying the function $\hat{\psi}_n^{(\beta)}$ with the vector $\hat{\mathbf{u}}_n^{(\beta)}$ in the usual fashion, we can write

$$\hat{\psi}_n^{(\beta)} \equiv \text{slogcm} \left\{\sum_{i=1}^k \phi''_{i,\beta}(\hat{\psi}_n^{(\beta)}(Z_{(i)}))\, , \, \sum_{i=1}^k \left[\hat{\psi}_n^{(\beta)}(Z_{(i)}) - \frac{\phi'_{i,\beta}(\hat{\psi}_n^{(\beta)}(Z_{(i)}))}{\phi''_{i,\beta}(\hat{\psi}_n^{(\beta)}(Z_{(i)}))}\right] \phi''_{i,\beta}(\hat{\psi}_n^{(\beta)}(Z_{(i)}))\right\}_{k=0}^n .$$

Hence, we can write w_i as

$$w_i = \hat{\psi}_n^{(\beta)}(Z_{(j)}) = \frac{\sum_{k \in B_i} \{\hat{\psi}_n^{(\beta)}(Z_{(k)})\, \phi''_{k,\beta}(\hat{\psi}_n^{(\beta)}(Z_{(k)})) - \phi'_{k,\beta}(\hat{\psi}_n^{(\beta)}(Z_{(k)}))\}}{\sum_{k \in B_i} \phi''_{k,\beta}(\hat{\psi}_n^{(\beta)}(Z_{(k)}))} \quad \text{for } \; j \in B_i\,. \qquad (4.24)$$

Further details about $\hat{\mathbf{u}}_{n,0}^{(\beta)}$: The vector $\hat{\mathbf{u}}_{n,0}^{(\beta)}$, which we identify with $\hat{\psi}_{n,0}^{(\beta)}$ (as explained in Section 2) also has a self–consistent/self–induced characterization in terms of the slope of the greatest convex minorant of a random function. This follows in the same way as in the case of $\hat{\mathbf{u}}_n^{(\beta)}$ by formulating a quadratic optimization problem based on the Kuhn–Tucker conditions for the corresponding minimization problem. We skip the details but give the self-consistent characterization. As before, we abbreviate $g(\beta, \mathbf{u})$ to $\xi(\mathbf{u})$, suppressing the dependence on β. We also abbreviate $\hat{\mathbf{u}}_{n,0}^{(\beta)}$ to $\hat{\mathbf{u}}^{(0)}$. For each i, set $d_i = \bigtriangledown_{ii}\, \xi(\hat{\mathbf{u}}^{(0)})$. Then, $\hat{\mathbf{u}}^{(0)}$ minimizes, $A(u_1, u_2, \ldots, u_n) = \sum_{i=1}^n \left[u_i - \left(\hat{u}_i^{(0)} - \bigtriangledown_i \xi(\hat{\mathbf{u}}^{(0)}) d_i^{-1}\right)\right]^2 d_i$ subject to the constraints that $u_1 \leq u_2 \leq \ldots \leq u_m \leq \theta_0 \leq u_{m+1} \leq \ldots \leq u_n$. Let $\tilde{B}_1, \tilde{B}_2, \ldots, \tilde{B}_l$ denote the level blocks of $\hat{\mathbf{u}}_{n,0}^{(\beta)}$ and let

$\{\tilde{w}_i\}_{i=1}^{l}$ denote the corresponding level values. Then, as long as $\tilde{w}_i \neq \theta_0$, it can be written as

$$\tilde{w}_i = \hat{\psi}_{n,0}^{(\beta)}(Z_{(j)}) = \frac{\sum_{k \in \tilde{B}_i} \{\hat{\psi}_{n,0}^{(\beta)}(Z_{(k)})\, \phi''_{k,\beta}(\hat{\psi}_{n,0}^{(\beta)}(Z_{(k)})) - \phi'_{k,\beta}(\hat{\psi}_{n,0}^{(\beta)}(Z_{(k)}))\}}{\sum_{k \in \tilde{B}_i} \phi''_{k,\beta}(\hat{\psi}_{n,0}^{(\beta)}(Z_{(k)}))} \text{ for } j \in \tilde{B}_i\,. \quad (4.25)$$

This representation is once again, a direct outcome of the self–induced characterization, and will prove useful in what follows.

4.2 Proof of Theorem 3.3

The likelihood ratio statistic of interest can be written as

$$\begin{aligned}
\mathrm{lrtpsi}_n &= 2\,(l_n(\hat{\beta}_n, \hat{\psi}_n) - l_n(\hat{\beta}_{n,0}, \hat{\psi}_{n,0})) \\
&= 2\,(l_n(\beta_0, \hat{\psi}_n^{(\beta_0)}) - l_n(\beta_0, \hat{\psi}_{n,0}^{(\beta_0)})) + 2\,(l_n(\hat{\beta}_n, \hat{\psi}_n) - l_n(\beta_0, \hat{\psi}_n^{(\beta_0)})) - 2(l_n(\hat{\beta}_{n,0}, \hat{\psi}_{n,0}) - l_n(\beta_0, \hat{\psi}_{n,0}^{(\beta_0)}))\,.
\end{aligned}$$

It will follow from Theorem 3.1 that

$$\tilde{R}_n \equiv 2\,(l_n(\hat{\beta}_n, \hat{\psi}_n) - l_n(\beta_0, \hat{\psi}_n^{(\beta_0)})) - 2(l_n(\hat{\beta}_{n,0}, \hat{\psi}_{n,0}) - l_n(\beta_0, \hat{\psi}_{n,0}^{(\beta_0)}))$$

is $o_p(1)$ whence it suffices to find the asymptotic distribution of

$$C_n = 2\,(l_n(\beta_0, \hat{\psi}_n^{(\beta_0)}) - l_n(\beta_0, \hat{\psi}_{n,0}^{(\beta_0)}))\,.$$

This is precisely the likelihood ratio statistic for testing $\psi(z_0) = \theta_0$ holding β fixed at its true value β_0. We can write C_n as,

$$C_n = 2\left[\sum_{i=1}^{n} \phi(\Delta_{(i)}, R_i(\beta_0), \hat{\psi}_{n,0}^{(\beta_0)}(Z_{(i)})) - \sum_{i=1}^{n} \phi(\Delta_{(i)}, R_i(\beta_0), \hat{\psi}_n^{(\beta_0)}(Z_{(i)}))\right]$$

where ϕ is as defined in (**??**). For the sake of notational compactness, in the remainder of the proof, we will write $\hat{\psi}_n^{(\beta_0)}(Z_{(i)})$ as $\tilde{\psi}(Z_{(i)})$, $\hat{\psi}_{n,0}^{(\beta_0)}(Z_{(i)})$ as $\tilde{\psi}_0(Z_{(i)})$, and $\phi(\Delta_{(i)}, R_i(\beta_0), t)$ as $\phi_i(t)$. Furthermore $\partial/\partial t\, \phi(\Delta_{(i)}, R_i(\beta_0), t)$ will be written as $\phi'_i(t)$ and so on. The set of indices i on which $\tilde{\psi}(Z_{(i)})$ and $\tilde{\psi}_0(Z_{(i)})$ differ is denoted by J_n. Now, $C_n = -2\,T_n$ where

$$\begin{aligned}
T_n &= \sum_{i=1}^{n} \phi_i(\tilde{\psi}(Z_{(i)})) - \sum_{i=1}^{n} \phi_i(\tilde{\psi}_0(Z_{(i)})) \\
&= \sum_{i \in J_n} \phi_i(\tilde{\psi}(Z_{(i)})) - \sum_{i \in J_n} \phi_i(\tilde{\psi}_0(Z_{(i)})) \\
&= \sum_{i \in J_n} \phi_i^{'}(\psi_0(z_0))\,[(\tilde{\psi}(Z_{(i)}) - \psi_0(z_0)) - (\tilde{\psi}_0(Z_{(i)}) - \psi_0(z_0))] \\
&\qquad\qquad + \sum_{i \in J_n} \frac{1}{2}\,\phi_i^{''}(\psi_0(z_0))\,\left[(\tilde{\psi}(Z_{(i)}) - \psi_0(z_0))^2 - (\tilde{\psi}_0(Z_{(i)}) - \psi_0(z_0))^2\right] + R_n \\
&\equiv T_{n,1} + T_{n,2} + R_n\,,
\end{aligned}$$

by Taylor–expanding $\phi_i(t)$ around $\psi_0(z_0)$. Here,

$$R_n = \sum_{i\in J_n} \frac{1}{6}\,\phi_i^{'''}\,(\tilde{\psi}(Z_{(i)})^\star)\,\Big(\tilde{\psi}(Z_{(i)}) - \psi_0(z_0)\Big)^3 - \sum_{i\in J_n} \frac{1}{6}\,\phi_i^{'''}\,(\tilde{\psi}_0(Z_{(i)})^\star)\,\Big(\tilde{\psi}_0(Z_{(i)}) - \psi_0(z_0)\Big)^3$$

(where $\tilde{\psi}(Z_{(i)})^\star$ is some point between $\tilde{\psi}(Z_{(i)})$ and $\psi_0(z_0)$ and $\tilde{\psi}_0(Z_{(i)})^\star$ is some point between $\tilde{\psi}_0(Z_{(i)})$ and $\psi_0(z_0)$) and can be shown to converge to 0 in probability by using the facts that (a) $\sup_{i\in J_n} \mid \phi_i^{'''}(\tilde{\psi}(Z_{(i)})^\star) \mid$ and $\sup_{i\in J_n} \mid \phi_i^{'''}(\tilde{\psi}_0(Z_{(i)})^\star) \mid$ are $O_p(1)$, (b) $\sup_{z\in D_n} \mid \tilde{\psi}(z) - \psi_0(z_0) \mid$ and $\sup_{z\in D_n} \mid \tilde{\psi}_0(z) - \psi_0(z_0) \mid$ are $O_p(n^{-1/3})$ where D_n is the set on which $\tilde{\psi}$ and $\tilde{\psi}_0$ differ, and (c) the length of D_n is $O_p(n^{-1/3})$. Now consider $T_{n,2}$. Once again, by Taylor expansion, we have

$$\begin{aligned}
T_{n,2} &\equiv \sum_{i\in J_n} \frac{1}{2}\,\phi_i^{''}(\psi_0(z_0))\,\Big[(\tilde{\psi}(Z_{(i)}) - \psi_0(z_0))^2 - (\tilde{\psi}_0(Z_{(i)}) - \psi_0(z_0))^2\Big] \\
&= \sum_{i\in J_n} \frac{1}{2}\phi_i^{''}(\tilde{\psi}(Z_{(i)}))[\tilde{\psi}(Z_{(i)}) - \psi_0(z_0)]^2 - \sum_{i\in J_n} \frac{1}{2}\phi_i^{''}(\tilde{\psi}_0(Z_{(i)}))[\tilde{\psi}_0(Z_{(i)}) - \psi_0(z_0)]^2 \\
&\qquad + o_p(1)\,. \qquad (4.26)
\end{aligned}$$

Now consider,

$$T_{n,1} \equiv \sum_{i\in J_n} \phi_i^{'}(\psi_0(z_0))(\tilde{\psi}(Z_{(i)}) - \psi_0(z_0)) - \sum_{i\in J_n} \phi_i^{'}(\psi_0(z_0))\,(\tilde{\psi}_0(Z_{(i)}) - \psi_0(z_0)) \equiv S_1 - S_2\,.$$

Consider the term S_2. Note that for each $i \in J_n$, we can write:

$$\phi_i^{'}(\psi_0(z_0)) = \phi_i^{'}(\tilde{\psi}_0(Z_{(i)})) + (\psi_0(z_0) - \tilde{\psi}_0(Z_{(i)}))\,\phi_i^{''}(\tilde{\psi}_0(Z_{(i)})) + \frac{1}{2}\,\phi_i^{'''}(\tilde{\psi}_0(Z_{(i)})^{\star\star})(\psi_0(z_0) - \tilde{\psi}_0(Z_{(i)}))^2$$

where $\tilde{\psi}_0(Z_{(i)})^{\star\star}$ is a point between $\tilde{\psi}_0(Z_{(i)})$ and $\psi_0(z_0)$. We then have,

$$\begin{aligned}
S_2 &= \sum_{i\in J_n} \Big[\phi_i^{'}(\tilde{\psi}_0(Z_{(i)})) + (\psi_0(z_0) - \tilde{\psi}_0(Z_{(i)}))\,\phi_i^{''}(\tilde{\psi}_0(Z_{(i)})) + \frac{1}{2}\,\phi_i^{'''}(\tilde{\psi}_0(Z_{(i)})^{\star\star})(\psi_0(z_0) - \tilde{\psi}_0(Z_{(i)}))^2\Big] \\
&= \qquad\qquad \times(\tilde{\psi}_0(Z_{(i)}) - \psi_0(z_0)) \\
&= \sum_{i\in J_n} \Big[\phi_i^{'}(\tilde{\psi}_0(Z_{(i)})) + (\psi_0(z_0) - \tilde{\psi}_0(Z_{(i)}))\,\phi_i^{''}(\tilde{\psi}_0(Z_{(i)}))\Big]\;(\tilde{\psi}_0(Z_{(i)}) - \psi_0(z_0)) + o_p(1) \\
&= -\sum_{i\in J_n} \phi_i^{''}(\tilde{\psi}_0(Z_{(i)}))\,\left[\tilde{\psi}_0(Z_{(i)}) - \frac{\phi_i^{'}(\tilde{\psi}_0(Z_{(i)}))}{\phi_i^{''}(\tilde{\psi}_0(Z_{(i)}))} - \psi_0(z_0)\right]\,(\tilde{\psi}_0(Z_{(i)}) - \psi_0(z_0)) + o_p(1)\,,
\end{aligned}$$

where the fact that the term involving $\phi_i^{'''}$ is $o_p(1)$ is deduced by arguments similar to those needed to show that R_n is $o_p(1)$. Now, let $B_1^0, B_2^0, \ldots, B_r^0$ denote the level blocks for $\tilde{\psi}_0(Z_{(i)})$ that constitute

J_n, with level values $w_1^0, w_2^0, \ldots, w_r^0$ and suppose that $w_l^0 = \psi_0(z_0) \equiv \theta_0$. Then,

$$\begin{aligned}
S_2 + o_p(1) &= -\sum_{j=1}^{r} \sum_{i \in B_j} \left[\phi_i^{''}(\tilde{\psi}_0(Z_{(i)}))\left(\tilde{\psi}_0(Z_{(i)})) - \frac{\phi_i^{'}(\tilde{\psi}_0(Z_{(i)}))}{\phi_i^{''}(\tilde{\psi}_0(Z_{(i)}))}\right) - \psi_0(z_0)\,\phi_i^{''}(\tilde{\psi}_0(Z_{(i)}))\right] \\
&\qquad\qquad \times (\tilde{\psi}_0(Z_{(i)}) - \psi_0(z_0)) \\
&= -\sum_{j=1}^{r} \sum_{i \in B_j} \left[\phi_i^{''}(w_j^0)\left(w_j^0 - \frac{\phi_i^{'}(w_j^0)}{\phi_i^{''}(w_j^0)}\right) - \psi_0(z_0)\,\phi^{''}(w_j^0)\right](w_j^0 - \psi_0(z_0)) \\
&= -\sum_{j \neq l} (w_j^0 - \psi_0(z_0)) \left[\sum_{i \in B_j} (\phi_i^{''}(w_j^0)\,w_j^0 - \phi_i^{'}(w_j^0)) - \psi_0(z_0) \sum_{i \in B_j} \phi_i^{''}(w_j^0)\right] \\
&= -\sum_{j \neq l} (w_j^0 - \psi_0(z_0)) \left[\left(\sum_{i \in B_j} \phi_i^{''}(w_j^0)\right)\left[\frac{\sum_{i \in B_j} (\phi_i^{''}(w_j^0)\,w_j^0 - \phi_i^{'}(w_j^0))}{\sum_{i \in B_j} \phi_i^{''}(w_j^0)} - \psi_0(z_0)\right]\right] \\
&= -\sum_{j \neq l} \sum_{i \in B_j} \phi_i^{''}(w_j^0)\,(w_j^0 - \psi_0(z_0))^2 \,,
\end{aligned}$$

where this last step follows from the following observation: If B' is a level block for $\tilde{\psi}_0$ contained in J_n with level value $w^{(0)}$, then

$$w^{(0)} = \frac{\sum_{k \in B'} (w^{(0)}\,\phi_k^{''}(w^{(0)}) - \phi_k^{'}(w^{(0)}))}{\sum_{k \in B'} \phi_k^{''}(w^{(0)})} \,.$$

provided $w^{(0)} \neq \theta_0$. This is a direct consequence of the representation (4.25). It follows that

$$\begin{aligned}
S_2 + o_p(1) &= -\sum_{j \neq l} \sum_{i \in B_j} \phi_i^{''}(w_j^0)\,(w_j^0 - \psi_0(z_0))^2 \\
&= -\sum_{j=1}^{r} \sum_{i \in B_j} \phi_i^{''}(w_j^0)\,(w_j^0 - \psi_0(z_0))^2 \\
&= -\sum_{j=1}^{r} \sum_{i \in B_j} \phi_i^{''}(\tilde{\psi}_0(Z_{(i)}))\,(\tilde{\psi}_0(Z_{(i)}) - \psi_0(z_0))^2 \\
&= -\sum_{i \in J_n} \phi_i^{''}(\tilde{\psi}_0(Z_{(i)}))\,(\tilde{\psi}_0(Z_{(i)}) - \psi_0(z_0))^2 \,.
\end{aligned}$$

It is similarly established (using (4.24)) that

$$S_1 + o_p(1) = -\sum_{i \in J_n} \phi_i^{''}(\tilde{\psi}(Z_{(i)}))\,(\tilde{\psi}(Z_{(i)}) - \psi_0(z_0))^2 \,.$$

It follows that

$$T_{n,1} = -\sum_{i \in J_n} \phi_i^{''}(\tilde{\psi}(Z_{(i)}))\,(\tilde{\psi}(Z_{(i)}) - \psi_0(z_0))^2 + \sum_{i \in J_n} \phi_i^{''}(\tilde{\psi}_0(Z_{(i)}))\,(\tilde{\psi}_0(Z_{(i)}) - \psi_0(z_0))^2 + o_p(1) \,.$$

Now, on using (4.26) and the fact that R_n is $o_p(1)$ we get

$$\begin{aligned} T_n &= T_{n,1} + T_{n,2} + o_p(1) \\ &= -\frac{1}{2}\sum_{i \in J_n} \phi_i''(\tilde{\psi}(Z_{(i)}))\,(\tilde{\psi}(Z_{(i)}) - \psi_0(z_0))^2 + \frac{1}{2}\sum_{i \in J_n} \phi_i''(\tilde{\psi}_0(Z_{(i)}))\,(\tilde{\psi}_0(Z_{(i)}) - \psi_0(z_0))^2 + o_p(1)\,, \end{aligned}$$

whence

$$\begin{aligned} C_n &= -2\,T_n = \sum_{i \in J_n} \phi_i''(\tilde{\psi}(Z_{(i)}))\,(\tilde{\psi}(Z_{(i)}) - \psi_0(z_0))^2 - \sum_{i \in J_n} \phi_i''(\tilde{\psi}_0(Z_{(i)}))\,(\tilde{\psi}_0(Z_{(i)}) - \psi_0(z_0))^2 + o_p(1) \\ &= \sum_{i \in J_n} \phi_i''(\psi_0(Z_{(i)}))\,(\tilde{\psi}(Z_{(i)}) - \psi_0(z_0))^2 - \sum_{i \in J_n} \phi_i''(\psi_0(Z_{(i)}))\,(\tilde{\psi}_0(Z_{(i)}) - \psi_0(z_0))^2 + o_p(1)\,. \end{aligned}$$

Now,

$$\phi_i''(\psi_0(Z_{(i)})) = \frac{\exp(\psi_0(Z_{(i)}) + \beta_0^T\,X_{(i)})}{(1 + \exp(\psi_0(Z_{(i)}) + \beta_0^T\,X_{(i)}))^2},$$

whence

$$\begin{aligned} C_n &= \sum_{i \in J_n} \frac{\exp(\psi_0(Z_{(i)}) + \beta_0^T\,X_{(i)})}{(1 + \exp(\psi_0(Z_{(i)}) + \beta_0^T\,X_{(i)}))^2} \left[(\tilde{\psi}(Z_{(i)}) - \psi_0(z_0))^2 - (\tilde{\psi}_0(Z_{(i)}) - \psi_0(z_0))^2\right] + o_p(1) \\ &= n^{1/3}\,(\mathbb{P}_n - P)\,\xi_n(\delta, z, x) + n^{1/3}\,P\,\xi_n(\delta, z, x) + o_p(1) \end{aligned}$$

where $\mathbb{P}_n$ is the empirical measure of the observations $\{\Delta_i, Z_i, X_i\}_{i=1}^n$, P denotes the true underlying distribution of (Δ, Z, X), ξ_n is the random function given by

$$\xi_n(\delta, z, x) = \frac{\exp(\psi_0(z) + \beta_0^T\,x)}{(1 + \exp(\psi_0(z) + \beta_0^T\,x))^2} \left[(n^{1/3}(\tilde{\psi}(z) - \psi_0(z_0)))^2 - (n^{1/3}\,(\tilde{\psi}_0(z) - \psi_0(z_0)))^2\right] 1(z \in D_n)\,.$$

We are using operator notation here for expectations; thus $\mathbb{P}_n\,g$ denotes the expectation of g under the measure $\mathbb{P}_n$ and $P\,g$ denotes the expectation of g under the measure P. The function g is allowed to be a random function. Now,

$$n^{1/3}\,(\mathbb{P}_n - P)\,\xi_n(\delta, z, x) = n^{-1/6}\,\sqrt{n}\,(\mathbb{P}_n - P)\,\xi_n(\delta, z, x)\,.$$

Using the facts that (i) D_n is eventually contained in a set of the form $[z_0 - M\,n^{-1/3}, z_0 + M\,n^{-1/3}]$ with arbirtrarily high preassigned probability (ii) the processes U_n and V_n are $O_p(1)$ on compacts and monotone increasing, along with standard preservation properties of Donsker classes of functions, it can be argued that with arbitrarily high preassigned probability, the function $\xi_n(\delta, z, x)$ lies in a Donsker class, whence it follows that $\sqrt{n}\,(\mathbb{P}_n - P)\,\xi_n(\delta, z, x)$ is $O_p(1)$; consequently $n^{1/3}\,(\mathbb{P}_n - P)\,\xi_n(\delta, z, x)$ is $O_p(n^{-1/6})$ and hence $o_p(1)$.

To find the asymptotic distribution of C_n we can therefore concentrate on the asymptotic distribution of

$$n^{1/3}\,P\,\xi_n(\delta, z, x) = n^{1/3}\,P\,\left[h(z, x)\,K_n(z)\right]$$

where

$$K_n(z) = \left[(n^{1/3}(\tilde{\psi}(z) - \psi_0(z_0)))^2 - (n^{1/3}(\tilde{\psi}_0(z) - \psi_0(z_0)))^2\right] \; 1(Z \in D_n)$$

and

$$h(z,x) = \frac{\exp(\psi_0(z) + \beta_0^T\, x)}{(1 + \exp(\psi_0(z) + \beta_0^T\, x))^2}\,.$$

Thus,

$$\begin{aligned} n^{1/3}\, P\, \xi_n(\delta, z, x) &= n^{1/3}\, P\, \left[K_n(z)\, h(z,x)\right] \\ &= n^{1/3} \int_{D_n} K_n(z)\, E(h(Z,X) \mid Z = z)\, f_Z(z)\, dz \\ &= n^{1/3} \int_{\tilde{D}_n} K_n(z_0 + h\, n^{-1/3})\, w(z_0 + h\, n^{-1/3})\, f_Z(z_0 + h\, n^{-1/3})\, dh \end{aligned}$$

where $h = n^{1/3}\,(z - z_0)$, $\tilde{D}_n = n^{1/3}\,(D_n - z_0)$ and $w(z) = E(h(Z,X) \mid Z = z)$. Now note that,

$$K_n(z_0 + h\, n^{-1/3}) = (U_n^2(h) - V_n^2(h))\, 1\,(h \in \tilde{D}_n)$$

where $\tilde{D}_n$ is the set on which U_n and V_n differ. Now, note that w is continuous in z and is given by:

$$w(z) = \int \frac{\exp(\psi_0(z) + \beta_0^T\, x)}{(1 + \exp(\psi_0(z) + \beta_0^T\, x))^2}\, \frac{f(z,x)}{f_Z(z)}\, d\,\mu(x)\,.$$

On using the facts that $\tilde{D}_n$ is eventually contained with arbitrarily high probability in a compact set and the boundedness in probability of the processes U_n and V_n on compacts along with the continuity of the functions w and f_Z, we get,

$$n^{1/3}\, P\, \xi_n(\delta, z, x) = \int w(z_0)\, f_Z(z_0)\, (U_n^2(h) - V_n^2(h))\, dh + o_p(1)\,.$$

But $C(z_0) = w(z_0)\, f_Z(z_0) = 1/a^2$ where a is as defined in Theorem 3.2. An application of Theorem 3.2 and Slutsky's theorem yields

$$n^{1/3}\, P\, \xi_n(\delta, z, x) \rightarrow_d \frac{1}{a^2} \int \left((g_{a,b}(h))^2 - (g_{a,b}^0(h))^2\right)\, dh\,,$$

and the fact that

$$\frac{1}{a^2} \int \left((g_{a,b}(h))^2 - (g_{a,b}^0(z))^2\right)\, dh \equiv_d \int \left((g_{1,1}(h))^2 - (g_{1,1}^0(z))^2\right)\, dh \equiv \mathbb{D}$$

follows as a direct application of Lemma 3.1 followed by the change of variable theorem from calculus.

It remains to show that

$$\begin{aligned} \tilde{R}_n &= 2\,(l_n(\hat{\beta}_n, \hat{\psi}_n) - l_n(\beta_0, \hat{\psi}_n^{\beta_0})) - 2(l_n(\hat{\beta}_{n,0}, \hat{\Lambda}_{n,0}) - l_n(\beta_0, \hat{\psi}_{n,0}^{\beta_0})) \\ &\equiv 2\,(l_n(\hat{\beta}_n, \hat{\Lambda}_n) - l_n(\beta_0, \hat{\Lambda}_n^{\beta_0})) - 2(l_n(\hat{\beta}_{n,0}, \hat{\Lambda}_{n,0}) - l_n(\beta_0, \hat{\Lambda}_{n,0}^{\beta_0})) \end{aligned}$$

is $o_p(1)$. This is precisely $\mathrm{lrtbeta}_n - \mathrm{lrtbeta}_n^0$. From Theorem 3.1 we get:

$$\begin{aligned}
\mathrm{lrtbeta}_n - \mathrm{lrtbeta}_n^0 &= n\,(\hat{\beta}_n - \beta_0)^T\,\tilde{I}_0\,(\hat{\beta}_n - \beta_0) - n\,(\tilde{\beta}_n - \beta_0)^T\,\tilde{I}_0\,(\tilde{\beta}_n - \beta_0) + o_p(1) \\
&= n\,(\hat{\beta}_n - \tilde{\beta}_n)^T\,\tilde{I}_0\,(\hat{\beta}_n - \tilde{\beta}_n) + 2\,n\,(\tilde{\beta}_n - \beta_0)^T\,\tilde{I}_0\,(\hat{\beta}_n - \tilde{\beta}_n) + o_p(1) \\
&= \sqrt{n}\,(\hat{\beta}_n - \tilde{\beta}_n)^T\,\tilde{I}_0\,\sqrt{n}\,(\hat{\beta}_n - \tilde{\beta}_n) + 2\,\sqrt{n}\,(\tilde{\beta}_n - \beta_0)^T\,\tilde{I}_0\,\sqrt{n}(\hat{\beta}_n - \tilde{\beta}_n) + o_p(1) \\
&\equiv I_n + II_n + o_p(1)\,.
\end{aligned}$$

The fact that I_n is $o_p(1)$ follows from the observation that $\sqrt{n}\,(\hat{\beta}_n - \tilde{\beta}_n) = r_n - s_n$, which is $o_p(1)$ (by Theorem 3.1). The fact that II_n is $o_p(1)$ follows on using the facts that $\sqrt{n}\,(\hat{\beta}_n - \tilde{\beta}_n)$ is $o_p(1)$ and that $\sqrt{n}\,(\tilde{\beta}_n - \beta_0)$ is $O_p(1)$. □

References

Banerjee, M. (2000). *Likelihood Ratio Inference in Regular and Nonregular Problems.* Ph.D. dissertation, University of Washington.

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

Banerjee, M. (2007). Likelihood based inference for monotone response models. *Annals of Statistics*, **35**, 931 – 956.

Bazaraa, M.S., Sherali, H.D. and Shetty, C.M. (1993). *Nonlinear Programming: Theory and Applications.* John Wiley & Sons (1993).

Dunson, D.B., (2003). Bayesian isotonic regression for discrete outcomes. *Working Paper*, available at *http://ftp.isds.duke.edu/ WorkingPapers/03-16.pdf*

Dunson, D.B. and Neelon, B. (2003). Bayesian inference on order–constrained parameters in generalized linear models. *Biometrics*, **59** (2), 286 – 295.

Ghosh, D., Banerjee, M. and Biswas, P. (2004). Binary isotonic regression procedures, with application to cancer biomarkers. *The University of Michigan Department of Biostatistics Working Paper Series*, Working Paper **38**. Available at *http://www.bepress.com/umichbiostat/paper38*

Ghosh, D., Banerjee, M. and Biswas, P. (2008). Inference for constrained estimation of tumor size distributions. *Biometrics*, to appear. Available at *http://www.stat.lsa.umich.edu/∼moulib/cancconstrainednpmle4.pdf.*

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

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

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

Huang, J. (1996). Efficient estimation for the Proportional Hazards Model with Interval Censoring. *Ann. Statist.* **24**, 540 – 568.

Jongbloed, G. (1998). The iterative convex minorant algorithm for nonparametric estimation. *J. Comput. Graph. Statist.* **7**, 310-321.

Magnac, T. and Maurin, E. (2004). Partial identification in monotone binary models: discrete regressors and interval data. *Working Paper*, available at *http://www.crest.fr/doctravail/document/2004-11.pdf*

Magnac, T. and Maurin, E. (2007). Identification and information in monotone binary models. *Journal of Econometrics* **139**, Issue 1, 76–104.

Manski, C.F. (1988). Identification of binary response models. *JASA* **83**, 729–738.

Manski, C.F. and Tamer, E. (2002). Inference in regressions with interval data on a regressor or outcome *Econometrica*, **70**, 519–546.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*, **37**, Monographs on Statistics and Applied Probability. Chapman and Hall, London.

Murphy, S.A. and Van der Vaart, A.W. (1997). Semiparametric Likelihood Ratio Inference. *Ann. Statist.* **25**, 1471 – 1509.

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

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

Sen, B., Banerjee, M. and Woodroofe, M.B. (2008). Inconsistency of bootstrap: the Grenander estimator. Available at *http://www.stat.lsa.umich.edu/∼moulib/Grenboots.pdf*

Silvapulle, M. J. and Sen, P.K. (2004). *Constrained Statistical Inference*, Wiley Series in Probability and Statistics.

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

Van der Vaart, A. (1998). *Asymptotic Statistics.* Cambridge University Press, Cambridge.

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.