

MLE IN THE PROPORTIONAL ODDS MODEL

BY S.A. MURPHY¹, A.J. ROSSINI

AND A.W. VAN DER VAART

Pennsylvania State University, University of South Carolina

and Free University Amsterdam

June, 1996 revision

We consider maximum likelihood estimation of the parameters in the proportional odds model with right censored data. The estimator of the regression coefficient is shown to be asymptotically normal with efficient variance. The maximum likelihood estimator of the unknown monotonic transformation of the survival time converges uniformly at a parametric rate to the true transformation. Standard errors of the estimated regression coefficients are estimated by differentiation of the profile likelihood and are shown to be consistent. A likelihood ratio test for the regression coefficient is also considered.

1. Introduction

The proportional odds model represents an important alternative to the proportional hazards model in survival analysis. In the two sample problem, the proportional odds model constrains the ratio of odds of survival to be constant with time, whereas the proportional hazards model constrains the ratio of hazards to be constant with time. The proportionality constraint on the odds ratio implies that the ratio of the hazards converges to unity as time increases. A constant effect in terms of the odds ratio, yet a hazards ratio converging to one, can occur if the treatment effect on the hazard of failure dissipates with time. To generalize the proportional odds model to multiple samples and continuous covariates, Bennett (1983a, 1983b) parameterizes the survival function, S_Z , given the vector of covariates, Z , as

$$-\text{logit}(S_Z(t)) = G(t) + Z^T \beta,$$

where $\text{logit}(x) = \log(x/(1-x))$. The unknown parameters are G , where $G(t)$ is the baseline log odds of failure at time t , and β , is a p -vector of regression coefficients. When G is strictly increasing, this model is a linear transformation model; that is, for T the survival time,

$$G(T) = -Z^T \beta + \epsilon,$$

¹ Research partially supported by NSF grant DMS-9307255.

MSC 1991 subject classifications: 62G15, 62G20, 62F25.

Keywords and phrases: Survival analysis, profile likelihood, semiparametric model.

where ϵ is distributed according to a standard logistic distribution. Consideration of the linear transformation model lead Cheng et al. (1995) to estimate β by an estimating equation. Cheng et al. show that the resulting $\hat{\beta}$ is consistent and asymptotically normal. The proportional odds model has been studied by many other authors. Both Dabrowska & Doksum (1988) and Wu (1995) provide estimators in the two sample problem. Other authors include Cuzick (1988) who, in considering estimation in the linear transformation model, proposes a method based on ranks, and Pettitt (1984), who minimizes a marginal likelihood of the ranks in order to estimate the regression coefficient. Both Shen (1995) and Huang and Rossini (1995) use sieve maximum likelihood, the former for right censored failure times and the latter for interval censored failure times. Additionally, Bennett (1983) estimates both β and G via semiparametric maximum likelihood.

Bennett's estimator of β is the maximum profile likelihood estimator of β (G is profiled out). In this paper we demonstrate that the profile likelihood for β can be treated much as a parametric likelihood for β . In particular, the maximum profile likelihood estimator of β is consistent, asymptotically normal and efficient. Differentiation of the profile likelihood yields consistent estimators of the efficient information matrix. Additionally, a profile likelihood ratio statistic can be compared to percentiles of the chi-squared distribution to produce asymptotic hypothesis tests of the appropriate size. Section 2 contains a description of the asymptotic results and Section 3 gives a numerical comparison of the maximum profile likelihood estimator with the estimator of Cheng et al. (1995).

2. Results

As in Bennett (1983), it is convenient to use a slight reparametrization, $H(t) = e^{G(t)}$, so that H is the baseline odds of failure. The parameter, H , is a nondecreasing, right-continuous function with left-hand limits, from the positive real line to the positive real line, and $H(0) = 0$. In terms of the reparametrization, we have

$$S_Z(t) = \frac{e^{-Z^T \beta}}{H(t) + e^{-Z^T \beta}}.$$

We want to allow S_Z to be either a continuous or discrete or possibly a mixed distribution. Therefore we specify the conditional probability density with respect to the sum of the Lebesgue and the counting measure, by

$$\frac{e^{-Z^T \beta}}{(H(t) + e^{-Z^T \beta})(H(t-) + e^{-Z^T \beta})} h(t),$$

where $H(t-)$ is the left-hand limit of H at t . If H is absolutely continuous, then h is the derivative of H ; or if H is discrete, then $h(t) = \Delta H(t) = H(t) - H(t-)$. Bennett (1983)

assumes that S_Z is continuous, so that the denominator reduces to the square of $H(t) + e^{-z^T \beta}$.

We consider the use of this model for right censored data. That is, the observations are n i.i.d. copies of $X = (Y = T \wedge C, \delta, Z)$, where T denotes the failure time, and C is a censoring time. Conditional on Z , T is independent of C . The censoring indicator, δ , is one if $T \leq C$ and zero otherwise. Then the contribution of X is

$$p_{H,\beta}(x) = \left(\frac{e^{-z^T \beta} (1 - F_C(y - |z))}{(H(y) + e^{-z^T \beta})(H(y-) + e^{-z^T \beta})} h(y) \right)^\delta \left(\frac{e^{-z^T \beta}}{H(y) + e^{-z^T \beta}} f_C(y|z) \right)^{1-\delta} f_Z(z),$$

where F_Z is the marginal distribution of Z , F_C is the conditional distribution of C given Z , and lowercase letters denote the respective densities. If H is known to be absolutely continuous, then, as in the case of density estimation, there is no maximizer of the likelihood. However, if H is known to be discrete then a maximizer exists (as is proved below). To entertain a maximum likelihood estimator, we take the contribution to the likelihood for one observation to be,

$$\text{lik}(X; H, \beta) = \left(\frac{e^{-z^T \beta}}{(H(Y) + e^{-z^T \beta})(H(Y-) + e^{-z^T \beta})} \Delta H(Y) \right)^\delta \left(\frac{e^{-z^T \beta}}{H(Y) + e^{-z^T \beta}} \right)^{1-\delta},$$

where we have replaced h by ΔH . An alternative approach is to use a sieve; see Shen (1995) and Huang and Rossini (1995) for examples using the proportional odds model. The maximum likelihood estimator of H will be a nondecreasing step function with steps at the observed failure times.

The profile log likelihood for β is given by $\text{Prlik}_n(\beta) = \sum_{i=1}^n \log \text{lik}(X_i; \hat{H}_\beta, \beta)$, where \hat{H}_β maximizes the log likelihood for a fixed β . The maximum profile likelihood estimator, $\hat{\beta}$, maximizes $\text{Prlik}_n(\beta)$. This estimator is a function of the survival times only through their ranks. To see this, order the observed survival times from smallest to largest so that $t_{(1)} < t_{(2)} < \dots < t_{(k)}$, where k is the number of unique observed survival times. Assign the ranks 1 through k to these times. Censored times are assigned the rank corresponding to the largest $t_{(i)}$ less than or equal to the censored time. It turns out that the profile likelihood is the same, whether we use the Y_i 's, or replace them by their ranks. For a fixed β , the log likelihood evaluated at \hat{H}_β is maximal. The log likelihood with Y_i 's replaced by their ranks can attain the same maximum by setting $\Delta \hat{H}_\beta(i_j)$ equal to $\Delta \hat{H}_\beta(Y_j)$, where i_j is the rank of Y_j . The same argument can be made in the opposite direction hence the profile likelihood is invariant under any rank preserving transformation, and subsequently $\hat{\beta}$ is a function of the Y_i 's only through their ranks.

In the theorems below we make the following assumptions. For τ a finite time point, assume $P[C \geq \tau] = P[C = \tau] > 0$. That is, the study ends at time τ , and any remaining live individuals are considered censored at time τ . Also assume that, on the average, some individuals are at risk at time τ , that is $P[T > \tau] > 0$. Finally, for any possible covariate

pattern, the chance of observing a survival time should be positive, i.e. $P[T \leq C|Z] > 0$ almost surely under F_Z . Two technical conditions (that should not be necessary) are that β is known to lie in a compact set, say \mathcal{B} , and that the support of Z is bounded. In order for β to be identifiable, we need to assume that the covariance matrix of Z is positive definite. Since the maximum profile likelihood estimator of β is also the β -coordinate of the maximum likelihood estimator $(\hat{H}, \hat{\beta})$, the existence and consistency results can be expressed in terms of the joint maximum likelihood estimators.

THEOREM 2.1. (Existence). *There exists a pair (H, β) that maximizes $\prod_{i=1}^n \text{lik}(X_i; H, \beta)$ over the set of nondecreasing functions H with $H(0) = 0$ by \mathcal{B} .*

If we constrain $H(\tau)$ to be bounded by a constant, then the continuity of the likelihood in H and β can easily be seen to imply the existence of the maximum likelihood estimators. As a result, the proof of (2.1) needs to show only that the likelihood evaluated at any sequence H_m with $H_m(\tau)$ diverging to infinity as m increases will not approach the maximum value of the likelihood. This proof is omitted, as it is similar to the proof that the estimator of $H(\tau)$ does not diverge as n increases; this is done in Theorem 2.2. Denote the supremum norm on the interval $[0, \tau]$ by $\|\cdot\|_\infty$ and the Euclidean norm by $\|\cdot\|_2$.

THEOREM 2.2. (Consistency and Asymptotic Normality). *The maximum likelihood estimator is consistent; $\|\hat{H} - H_0\|_\infty$ and $\|\hat{\beta} - \beta_0\|_2$ converge almost surely to zero as $n \rightarrow \infty$. Additionally if β_0 is in the interior of \mathcal{B} , then $\sqrt{n}\|\hat{H} - H_0\|_\infty$ is bounded in probability and $\sqrt{n}(\hat{\beta} - \beta_0)$ converges in distribution to a p -variate normal distribution with mean zero and efficient variance, Σ^{-1} .*

The efficient variance, Σ^{-1} (see (4.9), (4.10) and (4.11)), is not expressible in an explicit form. Estimation of this matrix is considered below. In the proof of the preceding theorem, we give an asymptotic normality result for $\sqrt{n}(\hat{H} - H_0)$ considered as a functional on the space of uniformly bounded functions of uniformly bounded variation. This implies, for example, that the asymptotic distribution of $\sqrt{n}(\hat{H}(t) - H_0(t))$ is normal.

Since \hat{H}_β is not an explicit function of β , we are unable to differentiate the profile log likelihood explicitly in β to form an estimator of Σ . Instead, we numerically differentiate this function, using forward finite differences for the off-diagonal elements, and a combination of a forward and backward finite differences for the diagonal elements. Let h_1, \dots, h_p be random variables, all converging in probability to zero as n increases. Denote the p -vector with a one in the i th location and zeros elsewhere by e_i .

THEOREM 2.3. (Estimation of the Standard Errors). *For $i \neq j$ let*

$$-\hat{\Sigma}_{ij} = \frac{1}{nh_i h_j} \left(\text{Prlik}_n(\hat{\beta} + h_i e_i + h_j e_j) - \text{Prlik}_n(\hat{\beta} + h_i e_i) - \text{Prlik}_n(\hat{\beta} + h_j e_j) + \text{Prlik}_n(\hat{\beta}) \right)$$

and put

$$-\hat{\Sigma}_{ii} = \frac{1}{nh_i^2} \left(\text{Prlik}_n(\hat{\beta} + h_i e_i) - 2\text{Prlik}_n(\hat{\beta}) + \text{Prlik}_n(\hat{\beta} - h_i e_i) \right).$$

If, for each i, j , $h_i \xrightarrow{P} 0$ and both h_i/h_j and $(\sqrt{n}h_i)^{-1}$ are bounded in probability as $n \rightarrow \infty$, then $\hat{\Sigma}_{ij}$ converges in probability to the (i, j) th component of Σ .

In the simulations, we use $h_i = \max(|\hat{\beta}_i|, 1) * \text{sign}(\hat{\beta}_i)/\sqrt{n}$ or $h_i = \text{sign}(\hat{\beta}_i)/\sqrt{n}$. The proof of this theorem is given in Murphy and van der Vaart (1996). Other, more complex, numerical methods could be considered. One possibility is to fit a higher order spline to the profile log likelihood, and to differentiate the spline. However the simulations below demonstrate that this simple approach works well.

In the following theorem we verify the use of the profile likelihood ratio statistic for hypothesis testing and for obtaining, via inversion, confidence intervals. For example, consider a test of $H_0: \beta_j = \beta_{j0}$ versus $H_1: \beta_j \neq \beta_{j0}$. Denote the maximum likelihood estimator of (H, β) under the null hypothesis by $(\hat{H}_0, \hat{\beta}_0)$, ie., $\hat{H}_{\hat{\beta}_0}$ is \hat{H}_0 .

THEOREM 2.4. (Likelihood Ratio Inference). Under $H_0: \beta_j = \beta_{j0}$ the profile likelihood ratio statistic,

$$\text{lrt}_n(\beta_{j0}) = 2 \left(\text{Prlik}_n(\hat{\beta}) - \text{Prlik}_n(\hat{\beta}_0) \right)$$

has an asymptotic chi-squared distribution with one degree of freedom.

3. Computation and Simulation Results

The likelihood can be written as

$$\prod_{i=1}^n \left(\frac{e^{-Z_i^T \beta}}{H(Y_i) + e^{-Z_i^T \beta}} \right) \left(\frac{\Delta H(Y_i)}{H(Y_i-) + e^{-Z_i^T \beta}} \right)^{\delta_i}.$$

The presence of the terms $\Delta H(Y_i)$ forces the estimator of H to have positive jumps at the observed failure times. Further considerations (see e.g. the representation for \hat{H} below) show that \hat{H} does not have jumps at any other points. Thus, the number of parameters is p plus the number of observed failure times. If the largest Y , say $Y_{(n)}$, is an observed failure time and there are no censorings at this time, then $\Delta H(Y_{(n)})$ appears in the likelihood only through the term $\Delta H(Y_{(n)}) / (H(Y_{(n)}) + e^{-Z_{(n)}^T \beta})$. This term is increasing in $\Delta H(Y_{(n)})$ (when keeping $H(Y_{(n)}-)$ fixed), and maximally one for $\Delta \hat{H}(Y_{(n)}) = \infty$. By our stipulation that $P(T > \tau) > 0$, the largest Y will be censored at τ with probability tending to one. Hence in our proofs we need not worry about this case. In practice the largest Y may be

an observed failure time, and then the likelihood is given by

$$\prod_{\substack{i=1 \\ Y_i < Y_{(n)}}}^n \left(\frac{e^{-Z_i^T \beta}}{H(Y_i) + e^{-Z_i^T \beta}} \right) \left(\frac{\Delta H(Y_i)}{H(Y_{i-}) + e^{-Z_i^T \beta}} \right)^{\delta_i} \prod_{\substack{i=1 \\ Y_i = Y_{(n)}}}^n \left(\frac{e^{-Z_i^T \beta}}{(H(Y_{(n)-}) + e^{-Z_i^T \beta})^{\delta_i}} \right).$$

We see that all observed failure times equal to the largest Y enter the likelihood as if they were censored failure times. So when the largest Y is an observed failure time, we set $\Delta \hat{H}(Y_{(n)})$ equal to positive infinity, and for estimation of the remaining jumps of H we act in the computations as if all observed failure times equal to the largest Y are censored. In the computations the number of parameters is $p + k$, where k is the number of observed failure times not equal to the largest Y .

To maximize the log likelihood, we use the IMSL routine, UMIAH, which employs a modification of the Newton-Raphson algorithm. This routine requires both the specification of the gradient and the Hessian. The gradient is a $p + k$ dimensional vector found by differentiating the log likelihood with respect to the p -dimensional vector β and with respect to the k jumps, ΔH . Similarly, the Hessian is a $p + k \times p + k$ matrix. In a problem with two covariates, estimation of the parameters requires one maximization, estimation of the standard errors requires five maximizations, and a likelihood ratio test requires an additional maximization. However, the routine works quite fast as can be seen from the CPU times in Table 1. These are the average times to simulate a sample, calculate the estimators and their standard errors, and to perform a likelihood ratio test for the first regression coefficient.

In the simulations we compared the profile likelihood approach with the estimating equation approach of Cheng et al. Cheng et al. allow for a variety of weight functions. It was our experience that the method using the suggested optimal weight function occasionally lead to computational difficulties and when there were no computation difficulties, the results were virtually identical to using a weight function of 1. As a result we used Cheng et al.'s estimator based on a weight function of 1 in the following comparisons. We considered two sample sizes, 50 and 100, along with two levels of censoring, 10% and 20% censoring. There are two covariates: the first covariate is a Bernoulli(.5) variable and the second covariate is either a uniform or an exponential variable. When the censoring is independent, the failure times are censored at a fixed time point. Furthermore, we considered two types of dependent censoring. In the first type of dependent censoring, the censoring varied by the covariate pattern, but was at a fixed time point corresponding to 10% or 20% censoring, conditionally on the covariate pattern. In the second type of dependent censoring, the failures with the first covariate equal to zero were not censored, and the failures with the first covariate equal to one were censored at a fixed time which corresponded to 20% censoring. Two possibilities were considered for the function H : $H(t) = t$ and $H(t) = t^2$. The two regression coefficients were set to either 1 or 0.

To compare the two methods, we calculated the mean square errors of the estimators (Cheng et al.'s estimator and the profile estimator), and constructed Wald 95% confidence intervals for the regression coefficients. In general, the mean square errors as the estimators of β_1 are comparable, with the Cheng et. al estimator apparently slightly more accurate. For the estimators of β_2 the results depend on the distribution of the covariate. For a uniformly distributed covariate the estimators were comparable, whereas for an exponential distribution the profile likelihood estimators appears to be between 10 % and 13 % more efficient. Both methods produce confidence intervals with slightly lower confidence than 95%. However, the confidence level of the Wald interval based on the profile estimator is rarely significantly different from 95%, whereas the confidence of the Wald interval based on Cheng et al.'s estimator is often significantly different from 95%. The reason for this is that, although the Cheng et al. estimator is accurate, the estimator of its standard error tends to be too low. Tables 1 and 2 give the results of six representative simulations, each of 1000 samples. In all six simulations, $\hat{\beta} = (0, 1)$, $H(t) = t$, and $n = 100$.

Our example is from the Veterans Administration lung cancer trial (Prentice, 1973). This data set has been analyzed by many authors; in particular, Bennett (1983ab), Pettitt (1984) and Cheng et al. (1995) fit a proportional odds model. All of these authors use the subgroup of 97 patients with no prior therapy. The response is patient survival time and the four covariates are performance status (PS) and a factor with four levels, (large, adeno, small and squamous tumor types). Cheng et al. present a table comparing their estimates and estimated standard errors with the other two methods. Estimates, standard errors and likelihood ratio test p-values for the profile method are given in Table 3. The likelihood ratio tests are for the null hypothesis that the parameter is zero. Figure 1 gives a contour plot of 2 times the profile log likelihood ratio for the coefficient of small (vs. large tumor type) and PS. Each contour corresponds to a constant value of 2 times the log likelihood, maximized over the entire parameter space minus 2 times the log likelihood, maximized over H and the two remaining regression coefficients. The dark contour is at 6, the 95th percentile of a chi-squared on two degrees of freedom. As a result, the collection of covariate values corresponding to the interior of the dark contour forms a 95% confidence region for small tumor type and PS. Figure 2 gives a similar plot for adeno (vs. large tumor type) and PS. Both plots illustrate the degree to which the data indicates that the parameters are nonzero. Additionally both plots lend support to the asymptotic normality results given in the last section.

4. Appendix

We denote expectation with respect to the empirical distribution of the data by \mathbb{P}_n and denote expectation with respect to the true underlying distribution of the data by P_{H_0, β_0} or, more briefly, by P_0 . In general, for a function g of the data, X , and estimators, $(\hat{H}, \hat{\beta})$, $\mathbb{P}_n[g(X; H, \beta)]$ evaluated at $(H, \beta) = (\hat{H}, \hat{\beta})$ is written as $\mathbb{P}_n[g(X; \hat{H}, \hat{\beta})]$ and likewise for expectation with respect to P_0 . The supremum norm on the interval $[0, \tau]$ is denoted by $\|\cdot\|_\infty$.

We first show that the maximum likelihood estimator, \hat{H} , satisfies the equation,

$$(4.1) \quad \hat{H}(t) = \int_0^t \frac{1}{W_n(u; \hat{H}, \hat{\beta})} dG_n(u),$$

where $(1/W_n)(u)$ and $G_n(u)$ are nondecreasing functions in u defined by

$$W_n(u; H, \beta) = \mathbb{P}_n \left[\frac{I\{Y \geq u\}}{H(Y) + e^{-Z^T \beta}} + \frac{\delta I\{Y > u\}}{H(Y-) + e^{-Z^T \beta}} \right], \quad G_n(u) = \mathbb{P}_n \delta I\{Y \leq u\}.$$

The equations (4.1) are a reexpression of the likelihood equations for H . To derive these equations, define a path through \hat{H} , indexed by ϵ , as $dH_\epsilon(t) = (1 + \epsilon h_1(t)) d\hat{H}(t)$, where h_1 is an arbitrary nonnegative bounded function. Since the log likelihood is maximized at $(\hat{H}, \hat{\beta})$ over the whole model, it is maximized at $\epsilon = 0$ when evaluated on the submodel given by $(H_\epsilon, \hat{\beta})$. The derivative of the log likelihood with respect to ϵ evaluated at $\epsilon = 0$ yields a score function for H in the ‘‘direction’’ h_1 , given by

$$\ell_{1H\beta}(X)[h_1] = \delta h_1(Y) - \frac{\int_0^Y h_1 dH}{H(Y) + e^{-Z^T \beta}} - \frac{\delta \int_0^{Y-} h_1 dH}{H(Y-) + e^{-Z^T \beta}}.$$

The preceding argument shows that $\mathbb{P}_n \ell_{1\hat{H}, \hat{\beta}}(X)[h_1] = 0$ for all h_1 . Putting $h_1(u) = I\{u \leq t\}$ and changing the order of integration results in (4.1). In a similar fashion, define a path through $\hat{\beta}$, indexed by ϵ , as $\beta_\epsilon = \hat{\beta} + \epsilon h_2$, where h_2 is a fixed vector in R^p . Differentiation of the log likelihood with respect to ϵ and evaluation at $\epsilon = 0$ yields the score function for β in the direction h_2 ,

$$h_2^T \ell_{2H\beta}(X) = -h_2^T Z \left(1 - \frac{e^{-Z^T \beta}}{H(Y) + e^{-Z^T \beta}} - \frac{\delta e^{-Z^T \beta}}{H(Y-) + e^{-Z^T \beta}} \right).$$

As before, $\mathbb{P}_n \ell_{2\hat{H}, \hat{\beta}}(X) = 0$.

We shall repeatedly use the following lemma, taken from Chapter 2.10 of van der Vaart and Wellner (1996). The general definition of a Donsker class can be found in this reference as well. However, we shall need only the result that the class BV_M of all functions $f: [0, \tau] \rightarrow R$ that are uniformly bounded by a constant M and are of variation bounded by M is Donsker (for each fixed $M < \infty$). We note that every Donsker class \mathcal{F} with integrable envelope function $x \rightarrow \sup_{f \in \mathcal{F}} |f(x)|$, in particular a uniformly bounded class, is Glivenko-Cantelli. This means that $\sup_{f \in \mathcal{F}} |\mathbb{P}_n f - P f| \rightarrow 0$, almost surely.

LEMMA 4.1. For a Lipschitz function $\phi: \mathbb{R}^k \rightarrow \mathbb{R}$ and classes \mathcal{F}_i of functions $f_i: \mathcal{X} \rightarrow \mathbb{R}$ let $\phi(\mathcal{F}_1, \dots, \mathcal{F}_k)$, denote the set of all functions $x \rightarrow \phi(f_1(x), \dots, f_k(x))$ as f_i ranges over \mathcal{F}_i , for each i . If each class \mathcal{F}_i is Donsker with integrable envelope function, then the class $\phi(\mathcal{F}_1, \dots, \mathcal{F}_k)$ is Donsker, provided that it consists of square-integrable functions.

One use of Lemma 4.1 is in the analysis of random step functions of the form

$$\int_0^\cdot \frac{1}{W_n(u; H_n, \beta_n)} dG_n(u),$$

where H_n and β_n could be random. The functions $u \rightarrow W_n(u; H, \beta)$ are contained in $BV_{M'}$ for some M' and uniformly bounded away from zero as H ranges over the set \mathcal{H} of nonnegative, nondecreasing elements of BV_M and β_n ranges over \mathcal{B} . Suppose that H_n is contained in \mathcal{H} and converges pointwise to a function H^* almost surely and β_n converges to an element β^* of \mathcal{B} almost surely. Then two applications of Lemma 4.1 yield

$$\left\| \int_0^\cdot \frac{1}{W_n(u; H_n, \beta_n)} d(G_n(u) - G(u)) \right\|_\infty \xrightarrow{\text{a.s.}} 0$$

and

$$\sup_{u \in [0, \tau], H \in \mathcal{H}, \beta \in \mathcal{B}} |W_n(u; H, \beta) - W(u; H, \beta)| \xrightarrow{\text{a.s.}} 0,$$

where $W(u; H, \beta)$ and G are the expectations of $W_n(u; H, \beta)$ and G_n , respectively, evaluated at H_0 and β_0 . The above displays, combined with the dominated convergence theorem, yield

$$(4.2) \quad \left\| \int_0^\cdot \frac{1}{W_n(u; H_n, \beta_n)} dG_n(u) - \int_0^\cdot \frac{1}{W(u; H^*, \beta^*)} dG(u) \right\|_\infty \xrightarrow{\text{a.s.}} 0.$$

Note that $H_0(t) = \int_0^t W(u; H_0, \beta_0)^{-1} dG(u)$.

LEMMA 4.2. Both H_0 and β_0 are identifiable. If H is absolutely continuous with respect to H_0 , $H(0) = 0$ and $p_{H, \beta}(x) = p_{H_0, \beta_0}(x)$ almost everywhere under P_{H_0, β_0} , then $H = H_0$ and $\beta = \beta_0$.

Proof. Considering the densities $p_{H, \beta} = p_{H_0, \beta_0}$ on $\delta = 1$, we see that

$$(4.3) \quad \frac{dH}{dH_0}(y) = e^{z^T(\beta - \beta_0)} \frac{(H(y) + e^{-z^T \beta})(H(y-) + e^{-z^T \beta})}{(H_0(y) + e^{-z^T \beta_0})(H_0(y-) + e^{-z^T \beta_0})},$$

for almost every (y, z) such that $P(C \geq y | Z = z) > 0$.

If H_0 has a jump at its left endpoint $t^* = \inf\{t: H_0(t) > 0\}$, then we can insert $y = t^*$ and obtain that

$$\frac{\Delta H(t^*)}{\Delta H_0(t^*)} = \frac{\Delta H(t^*) + e^{-z^T \beta}}{\Delta H_0(t^*) + e^{-z^T \beta_0}},$$

for F_Z -almost every z . This implies that $e^{z^T(\beta_0 - \beta)} = \Delta H(t^*)/\Delta H_0(t^*)$ and hence that the variable $z^T(\beta_0 - \beta)$ is degenerate. Thus, $\beta = \beta_0$ under our assumption that $\text{cov } Z$ is nondegenerate.

If H_0 does not jump at t^* , then (4.3) is valid for every y_m in a sequence $y_m \downarrow t^*$, for almost every z in the set $A_m = \{z: P(C \geq y_m | Z = z) > 0\}$. Then $A_m \uparrow A = \{z: P(C > t^* | Z = z) > 0\}$, which has probability 1 under F_Z by assumption. The limit as $m \rightarrow \infty$ of the right hand side of (4.3) when evaluated at y_m is $e^{z^T(\beta_0 - \beta)}$. Apparently, the left-hand side converges as well, to a limit that is not dependent on z . Again we obtain that $\beta = \beta_0$.

Now on $\delta = 1$ and the set of z for which $P(C \geq \tau | z) > 0$,

$$\frac{1}{(H(t) + e^{-z^T \beta_0})(H(t-) + e^{-z^T \beta_0})} (dH/dH_0)(t) = \frac{1}{(H_0(t) + e^{-z^T \beta_0})(H_0(t-) + e^{-z^T \beta_0})},$$

for almost every $t \in [0, \tau]$ under H_0 . By assumption, the set of z for which this holds has positive measure and hence is nonempty. Integrating both sides from 0 to u , yields

$$\frac{1}{H(u) + e^{-z^T \beta_0}} = \frac{1}{H_0(u) + e^{-z^T \beta_0}}$$

for all $u \leq \tau$. Therefore $H(u) = H_0(u)$ for all $u \leq \tau$. ■

Proof of Consistency in Theorem 2.2. First we show that the sequence $\hat{H}(\tau)$ is bounded, almost surely. Then, using (4.1) we show that the sequence $(\hat{H}, \hat{\beta})$ is relatively compact. Finally, using the above lemma on identifiability, we prove that any convergent subsequence of $(\hat{H}, \hat{\beta})$ must converge to (H_0, β_0) .

Define a random step function

$$\tilde{H}(t) = \int_0^t \frac{1}{W_n(u; H_0, \beta_0)} dG_n(u).$$

Note that \tilde{H} relates to G_n in a similar manner as \hat{H} , but with W_n evaluated at (H_0, β_0) , rather than at $(\hat{H}, \hat{\beta})$. Since H_0 is bounded on $[0, \tau]$ and β is restricted to a compact set, the functions $u \rightarrow W_n(u, H_0, \beta_0)$ are uniformly bounded away from zero and infinity and of uniformly bounded variation. By the argument in the beginning of the appendix, we see that $\|\tilde{H} - H_0\|_\infty$ converges almost surely to zero.

Since $(\hat{H}, \hat{\beta})$ maximizes the log likelihood,

$$(4.4) \quad \mathbb{P}_n \left[\log \text{lik}(X; \hat{H}, \hat{\beta}) - \log \text{lik}(X; \tilde{H}, \beta_0) \right] \geq 0.$$

By the definition of the likelihood, the left side of this display can be rewritten as

$$(4.5) \quad -\mathbb{P}_n \delta \log \left(\frac{\Delta \tilde{H}(y) \hat{H}(y) + e^{-z^T \hat{\beta}} \hat{H}(y-) + e^{-z^T \hat{\beta}}}{\Delta \hat{H}(y) \tilde{H}(y) + e^{-z^T \beta_0} \tilde{H}(y-) + e^{-z^T \beta_0}} \right) \\ + \mathbb{P}_n z^T (\beta_0 - \hat{\beta}) - \mathbb{P}_n (1 - \delta) \log \frac{\hat{H}(y) + e^{-z^T \hat{\beta}}}{\tilde{H}(y) + e^{-z^T \beta_0}}.$$

Here, by the definition of \tilde{H} and (4.1),

$$(4.6) \quad \frac{\Delta \tilde{H}(y)}{\Delta \hat{H}(y)} = \frac{W_n(y; \hat{H}, \hat{\beta})}{W_n(y; H_0, \beta_0)} \geq \frac{W_n(y; \hat{H}, \hat{\beta})}{m + o(1)}, \quad \text{a.s.},$$

uniformly in y , for a sufficiently small constant $m > 0$, because $W_n(y; H_0, \beta_0)$ converges almost surely to $W(y; H_0, \beta_0)$, which is bounded away from zero on $[0, \tau]$. Similarly, many of the other terms in (4.5) are bounded as well, by the uniform convergence of \tilde{H} to H_0 and the assumption of compact ranges for Z and β . For instance, $\tilde{H}(y) + e^{-z^T \beta_0}$ is bounded above (and below), uniformly in y , and $\hat{H}(y) + e^{-z^T \hat{\beta}}$ is bounded below by a positive constant, uniformly in y . It follows that, for a sufficiently small $m > 0$ and a sufficiently large M , the expression (4.5) is bounded above by

$$(4.7) \quad -\mathbb{P}_n \delta \log \left(W_n(y; \hat{H}, \hat{\beta}) (\hat{H}(y) + m) (\hat{H}(y-) + m) \right) \\ - \log(\hat{H}(\tau) + m) \mathbb{P}_n(1 - \delta) \{y = \tau\} + M + o(1), \quad \text{a.s.}$$

Since $G_n f(y) = \mathbb{P}_n \delta f(Y)$ by definition, and G_n has a density $W_n(y; \hat{H}, \hat{\beta})$ with respect to \hat{H} by (4.1), the first term can be rewritten as

$$- \int \log \left(W_n(y; \hat{H}, \hat{\beta}) (\hat{H}(y) + m) (\hat{H}(y-) + m) \right) W_n(y; \hat{H}, \hat{\beta}) d\hat{H}(y) \\ \leq \sup_x (-x \log x) \int \frac{d\hat{H}(y)}{(\hat{H}(y) + m) (\hat{H}(y-) + m)},$$

which is bounded above. Since the expression in (4.7) is bounded below by zero and the sequence $\mathbb{P}_n(1 - \delta) \{y = \tau\}$ converges almost surely to a positive number under our assumptions, we conclude that the sequence $\hat{H}(\tau)$ is almost surely bounded as $n \rightarrow \infty$.

For β ranging over \mathcal{B} and for H_1 and H_2 ranging over the set of bounded (by a constant M) monotone step functions on $[0, \tau]$ all of which satisfy $y \rightarrow \Delta H_1 / \Delta H_2(y) \geq 1$ with variation bounded by M , 2) bounded above by M , and 3) bounded below by a constant $m > 0$, the set of functions $x \rightarrow \log \text{lik}(x; H_1, \beta) - \log \text{lik}(x; H_2, \beta_0)$ is Glivenko-Cantelli. Since \hat{H} and \tilde{H} satisfy all these constraints as $n \rightarrow \infty$, almost surely, (4.4) implies that

$$(4.8) \quad P_0 \left[\log \text{lik}(X; \hat{H}, \hat{\beta}) - \log \text{lik}(X; \tilde{H}, \beta_0) \right] \geq -o(1), \quad \text{a.s.}$$

Now fix an ω in the underlying probability space such that $\hat{H}(\tau)$ is bounded by a constant M for every n , such that $G_n f - Gf \rightarrow 0$ uniformly in the functions of variation bounded by M and such that the preceding display is valid for this ω . The functions $u \rightarrow W_n(u; \hat{H}, \hat{\beta})$ are decreasing and bounded away from zero and infinity. By Helly's lemma and the compactness of \mathcal{B} , every subsequence of n has a further subsequence along which $\hat{\beta} \rightarrow \beta^*$ for some β^* and $W_n(u; \hat{H}, \hat{\beta}) \rightarrow W^*(u)$ for every u and some monotone

function W^* . Then, uniformly in t ,

$$\begin{aligned}\hat{H}(t) &= \int_0^t \frac{1}{W_n(u; \hat{H}, \hat{\beta})} dG_n(u) = \int_0^t \frac{1}{W_n(u; \hat{H}, \hat{\beta})} dG(u) + o(1) \\ &\rightarrow \int_0^t \frac{1}{W^*(u)} dG(u) =: H^*(t),\end{aligned}$$

by the dominated convergence theorem. Hence $\hat{H} \rightarrow H^*$ uniformly on $[0, \tau]$ and H^* is absolutely continuous with respect to G and thus also H_0 . Consequently (see the expressions (4.5) and (4.6)),

$$\log \text{lik}(x; \hat{H}, \hat{\beta}) - \log \text{lik}(x; \tilde{H}, \tilde{\beta}_0) \rightarrow \log \text{lik}(x; H^*, \beta^*) - \log \text{lik}(x; H_0, \beta_0), \quad \text{every } x.$$

By the dominated convergence theorem and (4.8), we conclude that the last function has a nonnegative mean under P_0 , which is the Kullback-Leibler divergence of the measures P_{H^*, β^*} and P_0 . Hence $H^* = H_0$ and $\beta^* = \beta_0$ by Lemma 4.2.

Since every subsequence of n contains a further subsequence for which $(\hat{H}, \hat{\beta})$ converges uniformly to (H_0, β_0) , we have convergence for the entire sequence. Uniform convergence of \hat{H} follows from a lemma stated in Chung (1974, pg. 133), since, for any u a jump point of H_0 , we have $\Delta \hat{H}(u)$ converges to $\Delta H_0(u)$ (note that $\Delta \hat{H}(u) = W_n(u; \hat{H}, \hat{\beta})^{-1} \Delta G_n(u)$).

■

Let h_1 be a bounded function of bounded variation and h_2 a p -dimensional vector. For $h = (h_1, h_2)$, denote the information operator by $\sigma(h) = (\sigma_1(h), \sigma_2(h))$ where $\sigma_1(h)$ is a bounded function of bounded variation and $\sigma_2(h)$ is a p dimensional vector. The name information operator comes from the fact that $P_0[\ell_{1H_0, \beta_0}(X)[h_1] + h_2^T \ell_{2H_0, \beta_0}(X)]^2 = \int \sigma_1(h)(u) h_1(u) dH_0 + h_2^T \sigma_2(h)$. The forms are:

$$\sigma_1(h)(u) = h_1(u)W(u; H_0, \beta_0) - P_0 \left[\frac{I\{Y \geq u\} \int_0^Y h_1 dH_0}{(H_0(Y) + e^{-Z^T \beta_0})^2} + \frac{\delta I\{Y > u\} \int_0^{Y-} h_1 dH_0}{(H_0(Y-) + e^{-Z^T \beta_0})^2} \right] +$$

$$(4.9) \quad h_2^T P_0 \left[\left(\frac{I\{Y \geq u\}}{(H_0(Y) + e^{-Z^T \beta_0})^2} + \frac{\delta I\{Y > u\}}{(H_0(Y-) + e^{-Z^T \beta_0})^2} \right) e^{-Z^T \beta_0} Z \right]$$

and

$$\sigma_2(h) = P_0 \left[\left(\frac{\int_0^Y h_1 dH_0}{(H_0(Y) + e^{-Z^T \beta_0})^2} + \frac{\delta \int_0^{Y-} h_1 dH_0}{(H_0(Y-) + e^{-Z^T \beta_0})^2} \right) e^{-Z^T \beta_0} Z \right] +$$

$$(4.10) \quad P_0 \left[\left(\frac{H_0(Y)}{(H_0(Y) + e^{-Z^T \beta_0})^2} + \frac{\delta H_0(Y-)}{(H_0(Y-) + e^{-Z^T \beta_0})^2} \right) e^{-Z^T \beta_0} Z Z^T \right] h_2.$$

Define the space BV to be the set of uniformly bounded functions of bounded variation, equipped with the norm $\|\cdot\|_{\text{BV}}$, which is defined as the maximum of the supremum norm and the total variation norm both on the interval $[0, \tau]$. We equipped the space $\text{BV} \times R^p$ with the norm $\|\cdot\|$ equal to the maximum of $\|\cdot\|_{\text{BV}}$ and the Euclidean norm.

LEMMA 4.3. *The linear operator, $\sigma: \text{BV} \times R^p \rightarrow \text{BV} \times R^p$ is onto and continuously invertible.*

Proof. The operator σ can be written as the sum of $A + K$ of two operators, where $Ah = (h_1 W(\cdot; H_0, \beta_0)^{-1}, h_2)$. The operator $A(h)$ is continuously invertible (with inverse $A^{-1}(h) = (h_1 W(\cdot; H_0, \beta_0)^{-1}, h_2)$), since W is bounded away from zero. Since we can write σ in the form $A(I + A^{-1}K)$, it suffices by Theorem 4.25 in Rudin (1973) to show that $A^{-1}K$ is compact and that $A + K$ is one-to-one. The first is true if K is compact.

Consider K . Since a bounded linear operator with finite-dimensional range is compact, we need only show that the operator $K_1: \text{BV} \rightarrow \text{BV}$, given by

$$K_1(h_1)(u) = P_0 \left[\frac{I\{Y \geq u\} \int_0^Y h_1 dH_0}{(H_0(Y) + e^{-Z^T \beta_0})^2} + \frac{\delta I\{Y > u\} \int_0^{Y-} h_1 dH_0}{(H_0(Y-) + e^{-Z^T \beta_0})^2} \right]$$

is compact. Thus, given a sequence of functions h_{1n} with $\|h_{1n}\|_{\text{BV}} \leq 1$, we must show that there exists a subsequence and an element $g \in \text{BV}$ such that $\|K_1 h_{1n} - g\|_{\text{BV}} \rightarrow 0$. Now K_1 is a linear operator with $\|K_1 h_1\|_{\text{BV}} \leq C \int |h_1| dH_0$ for every h_1 , and a fixed constant C . Hence, it suffices to show that there exists a subsequence of h_{1n} that converges in $L_1(H_0)$. This is an easy consequence of Helly's lemma. We can split h_{1n} in its positive and negative variation, and select a subsequence along which both parts converges pointwise. Then h_{1n} converges to the difference of the limits in $L_1(H_0)$ by the dominated convergence theorem.

To prove that σ is one-to-one, we prove that if $\|\sigma(h)\| = 0$ for an h with a bounded norm then $\|h\| = 0$. Suppose that $\|\sigma(h)\| = 0$, then $\int h_1 \sigma(h) dH_0 + h_2^T \sigma_2(h) = 0$, and both $\sigma_1(h)$ and $\sigma_2(h)$ are identically zero. As in the proof of Lemma 4.2, consider two cases, corresponding to $\Delta H_0(t^*)$ equal to zero or positive, in order to prove that h_2 is zero (in Lemma 4.2 the goal was to prove that $\beta = \beta_0$). We omit the details, as they are very similar to the method given in Lemma 4.2. Since $0 = \sigma_1(h_1, 0) = P_0[\ell_{1H_0\beta_0}(X)[h_1]]$, we have that when $P_0(C = \tau|Z) > 0$,

$$\frac{\int_0^\tau h_1 dH_0}{H_0(\tau) + e^{Z^T \beta_0}} = 0.$$

Since $P_0(C = \tau|Z) > 0$ on a set of positive probability, we may use the above along with a proof that h_1 is either nonnegative or nonpositive, to infer that h_1 is almost everywhere (*a.e.*) equal to zero. This plus the form of equation (4.9) and $h_2 = 0$ implies that $h_1(u)W(u; H_0, \beta_0)$ is identically equal to zero and thus h_1 is identically equal to zero.

To prove that h_1 is either nonnegative or nonpositive, we use $\sigma_1(h_1, 0) = 0$. By restricting ourselves to the portion of the sample space for which $\delta = 1$ and Z is in a set A of positive probability for which $P[C = \tau|Z] > 0$, we have that

$$h_1(T) = \frac{\int_0^T h_1 dH_0}{(H_0(T) + e^{-Z^T \beta_0})^2} + \frac{\int_0^{T-} h_1 dH_0}{(H_0(T-) + e^{-Z^T \beta_0})^2}$$

for almost every $T \in [0, \tau]$ and $Z \in A$. Rearranging terms results in,

$$h_1(T) = \frac{\int_0^{T-} h_1 dH_0}{H_0(T-) + e^{-Z^T \beta_0}} \left(1 + \frac{H_0(T) + e^{-Z^T \beta_0}}{H_0(T-) + e^{-Z^T \beta_0}} \right).$$

Replace h_1 by a function g which has both right and left hand limits ($g = h_1$ a.e. H_0), given by,

$$g(t) = \frac{\int_0^{t-} h_1 dH_0}{H_0(t-) + e^{-z_0^T \beta_0}} \left(1 + \frac{H_0(t) + e^{-z_0^T \beta_0}}{H_0(t-) + e^{-z_0^T \beta_0}} \right)$$

for an arbitrary z_0 in A . The function g can be expressed as a sum of integrals with respect to H_0 ,

$$g(t) = \int_0^t \frac{\int_0^{u-} g dH_0}{(H_0(u-) + e^{-z_0^T \beta_0})^2} dH_0(u) + \int_0^{t-} \frac{\int_0^{u-} g dH_0}{(H_0(u-) + e^{-z_0^T \beta_0})^2} dH_0(u).$$

A proof by contradiction suffices to show that g must be either nonnegative or nonpositive. Since g is a.e. equal to h_1 , we have that h_1 is a.e. (H_0) nonnegative or nonpositive. ■

Write the inverse of σ as $\tilde{\sigma}$. So $\tilde{\sigma}(h) = (\tilde{\sigma}_1(h), \tilde{\sigma}_2(h))$ with first component a function in BV and second component a p dimensional vector. Denote the p dimensional vector with a one in the i location and zeros elsewhere by e_i . Define the $p \times p$ matrix,

$$(4.11) \quad \Sigma^{-1} = (\tilde{\sigma}_2(0, e_1), \dots, \tilde{\sigma}_2(0, e_p)).$$

The matrix is symmetric and Lemma 4.3 implies that $h_2^T \Sigma^{-1} h_2 = h_2^T \tilde{\sigma}_2(0, h_2) > 0$ for h_2 nonzero; hence Σ^{-1} is invertible.

LEMMA 4.4. *The efficient score function for the estimation of β is the vector*

$$\tilde{\ell}(X) = \ell_{2H_0\beta_0}(X) - \ell_{1H_0\beta_0}(X)[g^*]$$

where $\ell_{1H_0\beta_0}$ is applied component wise to the p dimensional vector of functions g^* and,

$$g^* = -\Sigma \begin{pmatrix} \tilde{\sigma}_1(0, e_1) \\ \vdots \\ \tilde{\sigma}_1(0, e_p) \end{pmatrix}.$$

Proof. We must show that $\tilde{\ell}$ is orthogonal to the score for H , given by $\ell_{1H_0\beta_0}(X)[g_1]$ for any bounded function g_1 . Consider $e_i^T \Sigma^{-1} P_0 \tilde{\ell}_{1H_0\beta_0}(X)[g_1]$ which is equal to,

$$P_0 [e_i^T \Sigma^{-1} \ell_{2H_0\beta_0}(X) - \ell_{1H_0\beta_0}(X)[e_i^T \Sigma^{-1} g^*]] \ell_{1H_0\beta_0}(X)[g_1]$$

Now $e_i^T \Sigma^{-1} = \tilde{\sigma}_2(0, e_i)^T$ so the above is equal to

$$P_0 [\tilde{\sigma}_2(0, e_i)^T \ell_{2H_0\beta_0}(X) + \ell_{1H_0\beta_0}(X)[\tilde{\sigma}_1(0, e_i)]] \ell_{1H_0\beta_0}(X)[g_1].$$

This is equal to $\int g_1 \sigma_1(\tilde{\sigma}(0, e_i)) dH_0 = 0$. ■

Proof of Asymptotic Normality in Theorem 2.2. In the following, “uniform boundedness” refers to the norm $\|\cdot\|$ on $BV \times R^p$ equal to the maximum of $\|\cdot\|_{BV}$ and the Euclidean norm. Let $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P_0)$ be the empirical process of the observations.

We prove that

$$(4.12) \quad \begin{aligned} & \sqrt{n} \left(\int h_1 d(\hat{H} - H_0) + h_2^T (\hat{\beta} - \beta_0) \right) \\ &= \mathbb{G}_n \left((\tilde{\sigma}_2(h))^T \ell_{2H_0\beta_0}(X) + \ell_{1H_0\beta_0}(X) [\tilde{\sigma}_1(h)] \right) + o_P(1), \end{aligned}$$

where the remainder term converges to zero in probability uniformly over $h \in BV_M \times R^p$. Since Lemma 4.3 implies that $\tilde{\sigma}(h)$ is uniformly bounded for all $h \in BV \times R^p$ which are uniformly bounded, setting $h_1 = 0$ results in

$$\sqrt{n} h_2^T (\hat{\beta} - \beta_0) = \mathbb{G}_n \left(\tilde{\sigma}_2(0, h_2)^T \ell_{2H_0\beta_0}(X) + \ell_{1H_0\beta_0}(X) [\tilde{\sigma}_1(0, h_2)] \right) + o_P(1),$$

which is asymptotically normal with mean zero and variance $h_2^T \tilde{\sigma}_2(0, h_2) = h_2^T \Sigma^{-1} h_2$ for each h_2 in R_p . Indeed, we may set h_2 equal to each of the unit vectors, e_i to get that $\sqrt{n}(\hat{\beta} - \beta_0) = \mathbb{G}_n \Sigma^{-1/2} \tilde{\ell}(X) + o_P(1)$, where $\tilde{\ell}$ is the efficient score function as given in Lemma 4.4.

Alternatively, we may put $h_2 = 0$ in (4.12) to get

$$\begin{aligned} \sqrt{n} \|\hat{H} - H_0\|_\infty &\leq \sqrt{n} \sup_{\|h_1\|_{BV} \leq 1} \left| \int h_1 d(\hat{H} - H_0) \right| \\ &= \sup_{\|h_1\|_{BV} \leq 1} \left| \mathbb{G}_n \left(\tilde{\sigma}_2(h_1, 0)^T \ell_{2H_0\beta_0}(X) + \ell_{1H_0\beta_0}(X) [\tilde{\sigma}_1(h_1, 0)] \right) \right| + o_P(1), \end{aligned}$$

which is bounded in probability, since the integrand with h_1 varying over BV_1 belongs to a Donsker class.

Setting $h_2^* = \tilde{\sigma}_2(h)$ and $h_1^* = \tilde{\sigma}_1(h)$, we have that

$$\begin{aligned} 0 &= \mathbb{P}_n \left((h_2^*)^T \ell_{2\hat{H}\hat{\beta}}(X) + \ell_{1\hat{H}\hat{\beta}}(X) [h_1^*] \right) \\ 0 &= P_0 \left((h_2^*)^T \ell_{2H_0\beta_0}(X) + \ell_{1H_0\beta_0}(X) [h_1^*] \right). \end{aligned}$$

Furthermore, simple algebra using the consistency of $(\hat{H}, \hat{\beta})$, the assumption that Z has bounded support and the fact that $h = \sigma(h^*)$ shows that

$$(4.13) \quad \begin{aligned} & P_0 \left((h_2^*)^T \ell_{2\hat{H}\hat{\beta}}(X) + \ell_{1\hat{H}\hat{\beta}}(X) [h_1^*] \right) - P_0 \left((h_2^*)^T \ell_{2H_0\beta_0}(X) + \ell_{1H_0\beta_0}(X) [h_1^*] \right) \\ &= - \int h_1 d(\hat{H} - H_0) - h_2^T (\hat{\beta} - \beta_0) + O_P(\|\hat{H} - H_0\|_\infty^2 + \|\hat{\beta} - \beta_0\|_2^2). \end{aligned}$$

Combination of the three equations yields that

$$\begin{aligned} \sqrt{n} \left(\int h_1 d(\hat{H} - H_0) + h_2^T (\hat{\beta} - \beta_0) \right) &= \mathbb{G}_n \left((h_2^*)^T \ell_{2\hat{H}\hat{\beta}}(X) + \ell_{1\hat{H}\hat{\beta}}(X)[h_1^*] \right) \\ &\quad + \sqrt{n} O_P \left(\|\hat{H} - H_0\|_\infty^2 + \|\hat{\beta} - \beta_0\|_2^2 \right). \end{aligned}$$

The set \mathcal{F} of functions of the type $x \rightarrow h_2^T \ell_{2H\beta}(x) + \ell_{1H\beta}(x)[h_1]$, with h_2 and β varying in a compact set in R^p , h_1 varying over BV_M , and H varying in the set of nonnegative nondecreasing functions with $H(\tau) \leq 2H_0(\tau)$, is Donsker and uniformly bounded. The function $x \rightarrow (h_2^*)^T \ell_{2\hat{H}\hat{\beta}}(x) + \ell_{1\hat{H}\hat{\beta}}(x)[h_1^*]$ belongs to \mathcal{F} with probability converging to one. Furthermore,

$$\rho(\hat{H}, \hat{\beta}, h; H_0, \beta_0, h) = P_0 \left(h_2^T \ell_{2\hat{H}\hat{\beta}}(X) + \ell_{1\hat{H}\hat{\beta}}(X)[h_1] - h_2^T \ell_{2H_0\beta_0}(X) - \ell_{1H_0\beta_0}(X)[h_1] \right)^2$$

converges to zero almost surely, uniformly over bounded h , by the consistency of $(\hat{H}, \hat{\beta})$ and the dominated convergence theorem. Hence, by the asymptotic uniform equicontinuity of the empirical process,

$$\begin{aligned} \sqrt{n} \left(\int h_1 d(\hat{H} - H_0) + h_2^T (\hat{\beta} - \beta_0) \right) &= \mathbb{G}_n \left((h_2^*)^T \ell_{2H_0\beta_0}(X) + \ell_{1H_0\beta_0}(X)[h_1^*] \right) \\ &\quad + o_P(1) + \sqrt{n} o_P \left(\|\hat{H} - H_0\|_\infty + \|\hat{\beta} - \beta_0\|_2 \right). \end{aligned}$$

Take the norm left and right, and conclude first that $\sqrt{n}\|\hat{H} - H_0\|_\infty$ and $\sqrt{n}\|\hat{\beta} - \beta_0\|_2$ are bounded in probability, and next that the remainder term in the display is $o_P(1)$. ■

Proof of Theorem 2.4. For a p -dimensional vector, denote the first entry by the subscript of 1 and the vector of the remaining $p - 1$ entries by the subscript, (1). We prove Theorem 2.4 with $j = 1$ (i.e., $H_0: \beta_1 = \beta_{10}$). The unconstrained MLE of β is given by $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_{(1)}^T)^T$ and the constrained MLE of β is $\hat{\beta}_0 = (\beta_{10}, \hat{\beta}_{(1)0}^T)^T$. Partition the matrix Σ into

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{1(1)} \\ \Sigma_{(1)1} & \Sigma_{(1)(1)} \end{pmatrix}.$$

We verify the conditions of Theorems 3.1 and 3.2 of Murphy and van der Vaart (1995) with

$$\begin{aligned} \psi_t^U &= (t, \beta_{(1)t}(\hat{\beta}, \hat{H}), H_t(\hat{\beta}, \hat{H})) \\ \psi_t^L &= (t, \beta_{(1)t}(\hat{\beta}_0, \hat{H}_0), H_t(\hat{\beta}_0, \hat{H}_0)) \end{aligned}$$

where $\beta_{(1)t}(\beta, H) = \beta_{(1)} + (\beta_1 - t)k_{(1)}^*$ and $H_t(\beta, H) = H + (\beta_1 - t) \int_0^\cdot k^* dH$. The $p - 1$ dimensional vector $k_{(1)}^*$ is given by $-\tilde{\sigma}_2(0, e_1)_1^{-1} \tilde{\sigma}_2(0, e_1)_{(1)}$ and the bounded function of bounded variation $k^*(u)$ is $-\tilde{\sigma}_2(0, e_1)_1^{-1} \tilde{\sigma}_1(0, e_1)(u)$. Put

$$\ell(x; t, (\beta, H)) = \log \text{lik}(x; H_t(\beta, H), (t, \beta_{(1)t}(\beta, H))).$$

By differentiating with respect to t we get the $\dot{\ell}$ in their equation (3.16)

$$\begin{aligned} \dot{\ell}(x; t, (\beta, H)) &= \tilde{\sigma}_2(0, e_1)_1^{-1} \left[\tilde{\sigma}_2(0, e_1)^T \ell_{2H_t(\beta, H)(t, \beta_{(1)t}(\beta, H))}(x) \right. \\ &\quad \left. + \ell_{1H_t(\beta, H)(t, \beta_{(1)t}(\beta, H))}(x) [\tilde{\sigma}_1(0, e_1)] \right] + \frac{(\beta_1 - t)k^*(y)^2}{1 - (\beta_1 - t)k^*(y)}. \end{aligned}$$

It can be shown that $k_{(1)}^* = \Sigma_{(1)(1)}^{-1} \Sigma_{(1)1}$ and that $k^*(u) = (1, -\Sigma_{1(1)} \Sigma_{(1)(1)}^{-1}) g^*(u)$. Therefore, $\dot{\ell}(x; \beta_{10}, (\beta_0, H_0))$ is the efficient score for the estimation of β_1 given by,

$$(1, -\Sigma_{1(1)} \Sigma_{(1)(1)}^{-1})(\tilde{\ell}),$$

where $\tilde{\ell}$ is the efficient score for the estimation of β as given in lemma 4.4. This agrees with the intuition given in Section 3 of Murphy and van der Vaart (1995).

Consideration of the form of $\ell(x; t, (\beta, H))$ along with an application of Lemma 4.1 suffices to verify the conditions of Murphy and van der Vaart's Lemma 3.3.

All that remains is to verify equations (3.15) and (3.17) of Murphy and van der Vaart. Equation (3.15) can be verified with the help of Lemma 4.1. To verify (3.17) we must prove that

$$\sqrt{n}P_0 \left(\tilde{\sigma}_2(0, e_1)^T \ell_{2\hat{H}_0\hat{\beta}_0}(x) + \ell_{1\hat{H}_0\hat{\beta}_0}(x) [\tilde{\sigma}_1(0, e_1)] \right)$$

converges to zero in probability. From the above we subtract

$$\sqrt{n}P_0 \left(\tilde{\sigma}_2(0, e_1)^T \ell_{2H_0\beta_0}(x) + \ell_{1H_0\beta_0}(x) [\tilde{\sigma}_1(0, e_1)] \right)$$

which is equal to zero and we subtract $\sqrt{n}(\hat{\beta}_0 - \beta_0)^T \sigma_2(\tilde{\sigma}(0, e_1)) + \sqrt{n} \int \sigma_1(\tilde{\sigma}(0, e_1)) d(\hat{H} - H_0)$ which is equal to $(\hat{\beta}_0 - \beta_0)^T e_1$ and is also zero. Simple algebra suffices to show that the combination of the three terms is $\sqrt{n}O_P(1)(\|\hat{H}_0 - H_0\|_\infty^2 + \|\hat{\beta}_0 - \beta_0\|_2^2)$. Theorem 2.2 implies that this converges to zero in probability. ■

Table 1: Average CPU Time for the Profile Estimator and the Ratio of Mean Square Errors

Design		CPU Time (Profile) ^a	MSE (Profile)/MSE (Cheng)
Uniform Z_2 and,	β_1	33	1.00
10% independent censoring	β_2		1.01
Uniform Z_2 and,	β_1	24	1.03
20% independent censoring	β_2		1.04
Uniform Z_2 and,	β_1	34	1.01
10% dependent censoring(1)	β_2		1.01
Uniform Z_2 and,	β_1	38	1.02
10% dependent censoring(2)	β_2		1.03
Exponential Z_2 and,	β_1	36	0.99
10% dependent censoring(1)	β_2		0.87
Exponential Z_2 and,	β_1	38	0.97
10% dependent censoring(2)	β_2		0.90

^a Average CPU Time in Seconds on a Sun Sparc Station 10, 32mb RAM, 264mb swap space

Table 2: Error rates of Likelihood Ratio Test^a
and Wald Confidence Intervals (Profile, Cheng et al. (1995))

Design		LRT (Profile)	95% CI (Profile)	95% CI (Cheng)
Uniform Z_2 and, 10% independent censoring	β_1	0.051	0.050	0.055
	β_2		0.053	0.069*
Uniform Z_2 and, 20% independent censoring	β_1	0.067*	0.064*	0.068*
	β_2		0.066*	0.074*
Uniform Z_2 and, 10% dependent censoring(1)	β_1	0.058	0.054	0.060
	β_2		0.053	0.063
Uniform Z_2 and, 10% dependent censoring(2)	β_1	0.046	0.045	0.052
	β_2		0.058	0.062
Exponential Z_2 and, 10% dependent censoring(1)	β_1	0.059	0.056	0.060
	β_2		0.057	0.078*
Exponential Z_2 and, 10% dependent censoring(2)	β_1	0.044	0.042	0.048
	β_2		0.065*	0.076*

^aType I error is .05

*Error rates significantly different from .05

Table 3: Analysis of the Veteran Administration Lung Cancer Data

Covariate	Estimator	Estimated Standard Error	LRT p-value
PS	-0.055	0.010	2.2×10^{-8}
adeno vs. large	1.339	0.556	0.015
small vs. large	1.440	0.525	0.006
squamous vs. large	-0.217	0.589	0.715

Figure 1. 2*Profile Log Likelihood Ratio for PS and Small Tumor Type

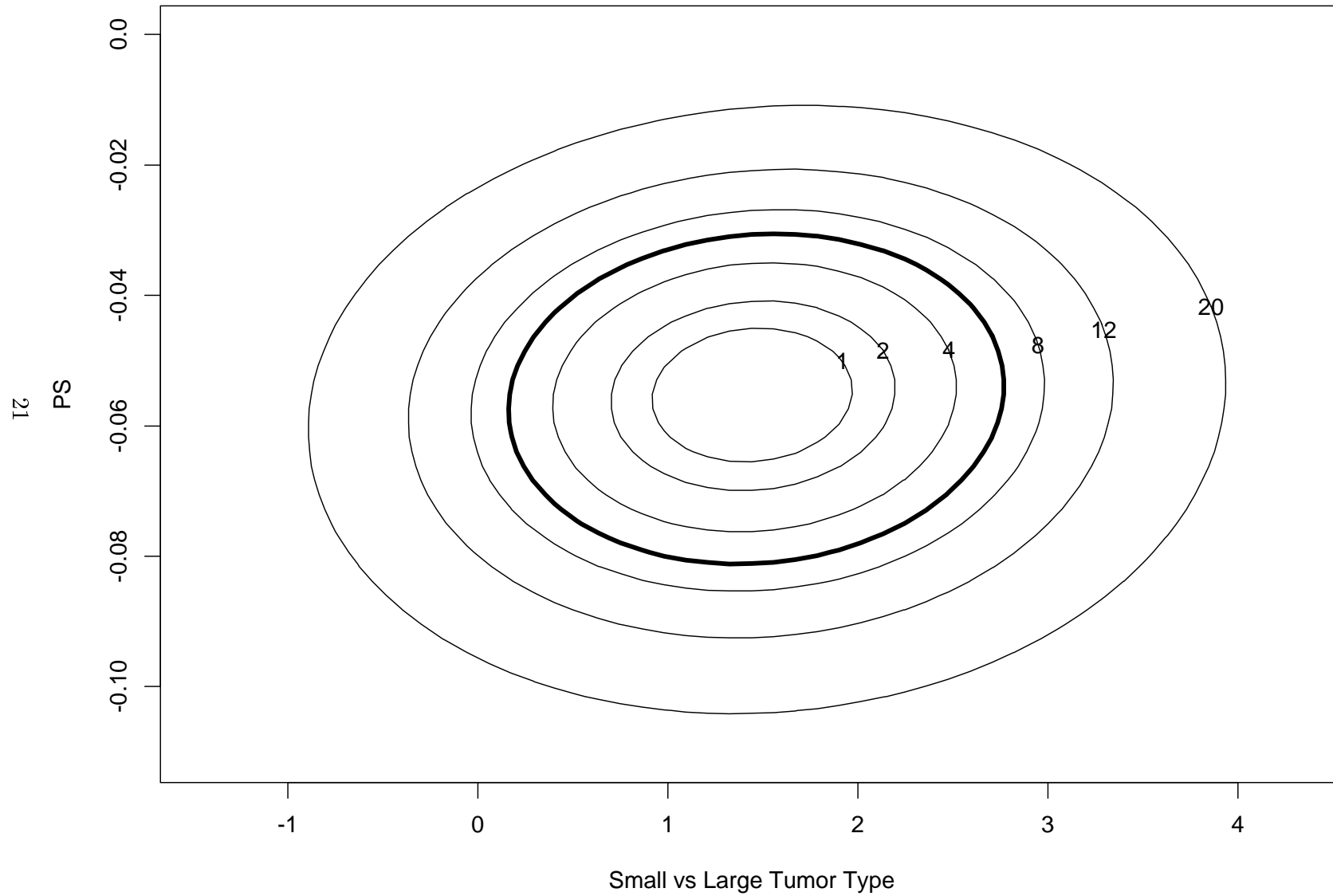
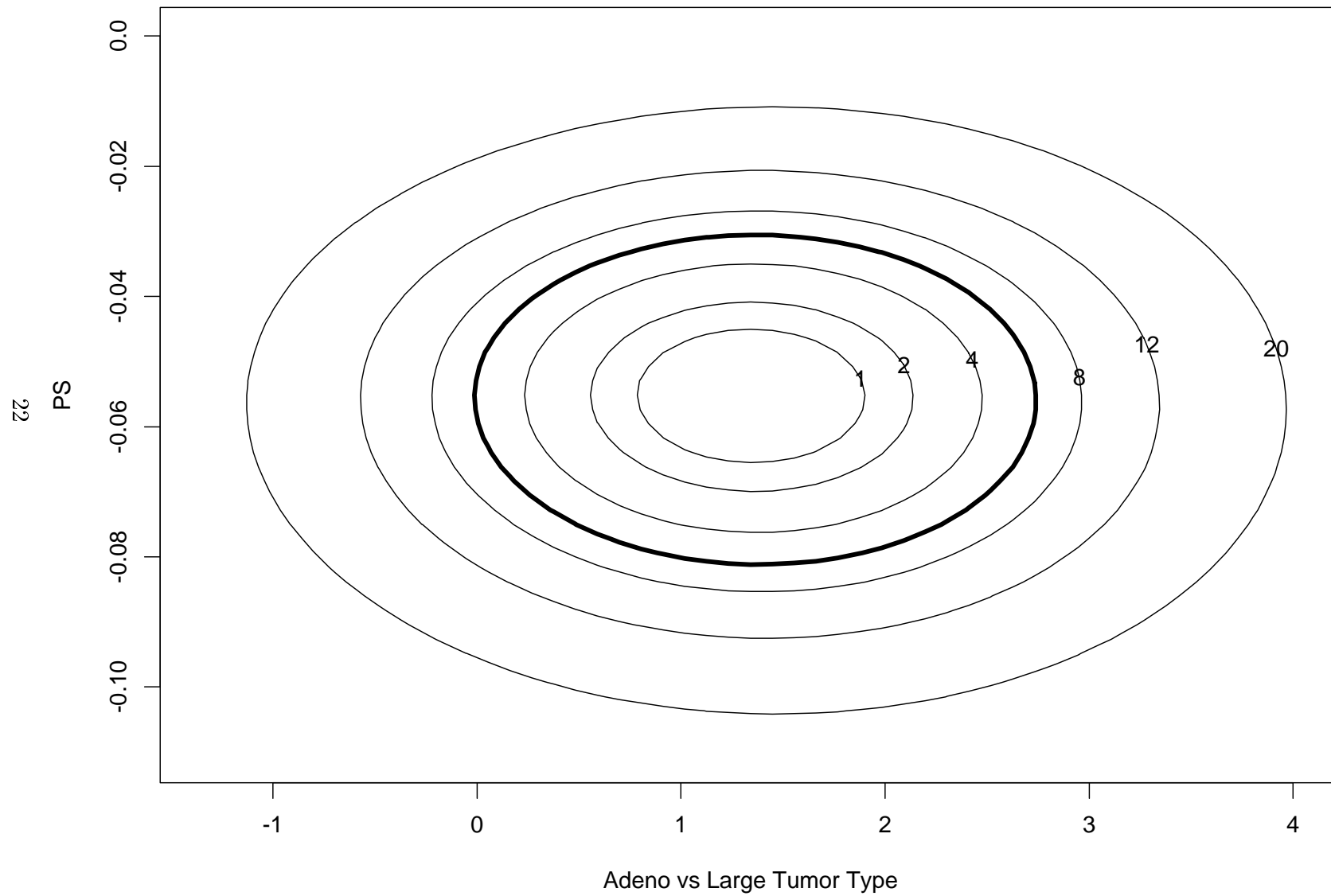


Figure 2: 2*Profile Log Likelihood Ratio for PS and Adeno Tumor Type



REFERENCES

- Bennett, S., (1983). Analysis of survival data by the proportional odds model. *Statistics in Medicine* **2**, 273-277.
- Bennett, S., (1983). Log-logistic regression models for survival data. *Applied Statistics* **32**, 165-171.
- Cheng, S.C., Wei, L.J. and Ying, Z., (1995). Analysis of transformation models with censored data. *preprint*.
- Chung, K.L., (1974). *A Course in Probability Theory*. Academic Press, Inc, San Diego.
- Cuzick, J., (1988). Rank regression. *Annals of Statistics* **16**, 1369-1389.
- Dabrowska, D.M. and Doksum, K.A., (1988). Estimation and testing in the two-sample generalized odds-rate model. *Journal of American Statistical Association* **83**, 1-23.
- Huang, J. and Rossini, A.J., (1995). Sieve estimation for the proportional odds failure-time regression model with interval censoring. *preprint*.
- Murphy, S.A. and Van der Vaart, A.W., (1995). Semiparametric likelihood ratio inference. *preprint*.
- Murphy, S.A. and Van der Vaart, A.W., (1996). Observed information in semiparametric models. *preprint*.
- Pettitt, A.N., (1984). Proportional odds models for survival data and estimates using ranks. *Applied Statistics* **33**, 169-175.
- Prentice, R.L., (1973). Exponential survivals with censoring and explanatory variables. *Biometrika* **60**, 279-288.
- Rudin, W., (1973). *Functional Analysis*. McGraw-Hill, New York.
- Shen, X., (1995). Proportional odds regression and universal sieve maximum likelihood estimation. *preprint*.
- Van der Vaart, A.W. and Wellner, J.A., (1996). *Weak Convergence and Empirical Processes*. Springer Verlag, New York.
- Wu, C.O., (1995). Estimating the real parameter in a two-sample proportional odds model.. *Annals of Statistics* **23**, 376-395.

S. A. Murphy
Department of Statistics
Pennsylvania State University
326 Classroom Building
University Park, PA 16802

USA

A. J. Rossini
Department of Statistics
University of South Carolina
Columbia, SC 29208
USA

A.W. van der Vaart
Department of Mathematics
Free University
De Boelelaan 1081a
1081 HV Amsterdam
Netherlands