Estimation of Time Transformation Models with Bernstein Polynomials

Alexander C. McLain* and Sujit K. Ghosh†

June 12, 2009

Institute of Statistics Mimeo Series #2625

Abstract

Time transformation models assume that the survival time is linearly related to covariates through an unknown monotonic transformation function and an error term with known distribution. In this paper the sieve method of maximum likelihood is used to estimate the unknown monotonic transformation of survival time. More specifically a suitable class of Bernstein polynomials is used to estimate the transformation function, that preserve the monotonicity and smoothness. This estimation method is less parametrically intensive than current time transformation methods. Furthermore, our method produces a smooth estimate of the time transformation and hence the survival function. We discuss the selection of the number of parameters for the polynomial asymptotically, and for practical sample sizes. The asymptotic properties of the estimators are shown, including the asymptotic normality and efficiency of the regression

*A. McLain (mclain@stat.ncsu.edu) is a VIGRE Postdoctoral Fellow, Department of Statistics, North Carolina State University, Raleigh, NC.

†S. Ghosh is Professor, Department of Statistics, North Carolina State University, Raleigh, NC.

coefficient. Simulation studies illustrate that our estimator has reasonably good empirical properties in practical sample sizes. The method is demonstrated on two data sets and compared to previous similar works.

Key Words: Censoring; Sieve maximum likelihood; Semiparametric efficiency; Survival analysis; Transformation models.

1. Introduction

In analysis of biomedical data is commonly of interest to estimate and determine the risk associated with covariates from 'time to event' or 'survival' data in light of censoring due to loss of follow up or death from unrelated causes. The proportional hazards model (Cox, 1972), the proportional odds model (Bennett, 1983), and the accelerated failure time model (Wei, 1992) are three major approaches for regression analysis of censored survival data. All of these models are included in the general family of time transformation models. Time transformation methods are similar to generalized linear models and hence provide a flexible way for accessing the effect of explanatory variables and estimating the distribution of censored data. These methods transform the scale of survival time by a linear function of covariates. The model specifies that the survival time T, with p-dimensional covariate vector $\boldsymbol{Z}$, has the form

$$H(T) = -\boldsymbol{\beta}^T \boldsymbol{Z} + \epsilon, \tag{1}$$

where ϵ is a random variable with known distribution function $F_\epsilon(\cdot)$ with $E(\epsilon) = 0$, $\boldsymbol{\beta}$ is a p-dimensional vector of regression coefficients and $H(\cdot)$ is an unknown monotonic function. The proportional hazards, and proportional odds model are special cases of the transformation model when ϵ follows an extreme-value and logistic distribution, respectively. When ϵ follows

a standard normal distribution, (1) is a generalization of the usual Box–Cox model. See Dabrowska and Doksum (1988), and Bickel et al. (1993) for more extensions and variants of transformation models.

Let $C_1, \ldots, C_n$ denote the censoring values, for the event times $T_1, \ldots, T_n$, with 'survival' functions $S_C(\cdot|\boldsymbol{Z})$ and $S_T(\cdot|\boldsymbol{Z})$ respectively. In the presence of censoring $\boldsymbol{Y} \equiv (X = C \wedge T, \Delta, \boldsymbol{Z})$ will be observed, where $\Delta = I_{(T \leq C)}$ and $I_{(\cdot)}$ is the indicator function. In the case of time independent $\boldsymbol{Z}$, Chen, Jin, and Ying (2002) based inference on $H(\cdot)$ through generalized estimating equations (GEE). GEE provide an estimate of $\boldsymbol{\beta}$ and $H(X_i)$ for all i such that $\Delta_i = 1$ through an iterative procedure (see Chen et al., 2002, for details). Zeng and Lin (2006, 2007) generalized the data framework of time transformation models to incorporate time-dependent covariates. Their method uses nonparametric maximum likelihood estimation (NPMLE) methods for estimation of $H(\cdot)$ and $\boldsymbol{\beta}$. This is a one-step procedure that views the parameters as $\boldsymbol{\beta}$ and $H(X_i)$ for all i such that $\Delta_i = 1$, and maximizes the likelihood with respect to these values. NPMLE has the benefit of asymptotic efficiency of its estimators, but is computationally more difficult than the GEE's.

Gehan (1965) presented the remission times of 42 Leukemia patients (30 observed failures), half of which were treated with the drug 6-mercaptopurine, and half were controls. The Veterans' Administration Lung Cancer (Prentice, 1973) presented survival times of 137 people (91 observed failures) along with the covariates; treatment, cell-type (grouped into small, large, adeno, and squamous), age, and many others. Letting $q_n = \sum_{i=1}^{n} \Delta_i$, the dimensionality of the parameter space of the GEE and NPMLE methods is $p + q_n$. As a result, the number of parameters is on the order $O(n)$. This creates a computational difficult task when the sample size is moderate to large. We are proposing that $H(\cdot)$ be estimated through a Sieve estimation procedure, with the use of Bernstein polynomials. This method is less parametrically intensive than the GEE and NPMLE. An advantage of using Bernstein

polynomials is that the number of parameters used for $H(\cdot)$, denoted as N, be on the order $O(n^{\kappa})$ where $\kappa = 1/3$ or $1/5$ depending on the differentiability of H (see theorem 2).

The interpretation of the β parameter in transformation models has sparked some debate in the past (see Hinkley and Runger, 1984, and comments). Notice from (1) that β_j can be interpreted as the predicted decrease in time on a transformed scale for a one unit increase in Z_j. The expected time can be expressed as, $E[T \mid \boldsymbol{Z}] \approx H^{-1}(-\boldsymbol{\beta}^T \boldsymbol{Z})$. When $H(\cdot)$ is a step function $\partial E[T \mid \boldsymbol{Z}]/\partial Z_j$ will not always exist. Therefore methods that produce a step function rely on the equivalence of the time transformation model to the proportional odds, and hazard models to interpret the coefficients. The estimate of $H(\cdot)$ in the proposed method is a smooth function. A smooth estimate of H will not need to use the error distributions that correspond to these models to have an interpretable coefficients. Hence, the proposed method simplifies the analysis of a broader range of error distributions. In our estimation procedure H^{-1} does not have closed form, but can always be estimated and is continuously differentiable.

Sieve estimation procedures have been considered for censored data with the proportional odds model by Huang and Rossini (1997), and Shen (1998). Recall, that the transformation model simplifies to the proportional odds model when F_ϵ is logistic. Huang and Rossini (1997), Shen (1998) and the work here all differ in the construction of the Sieve space. Huang and Rossini (1997) consider a piecewise linear function with knots at pre-specified locations. Shen (1998) uses B-splines and gives an algorithm that uses Fibonacci search strategies to locate the knots and the order of the splines. The Bernstein method smooths over equidistant knots from 0 to the maximum τ. Bernstein polynomials are a classical method of statistical smoothing that have long been used in nonparametric Bayesian density estimation. They are used in this paper due to their ease of implementation, continuous differentiability and theoretical properties. Furthermore, the monotonicity constraint of H can be enforced by a

simple change of variables and an unconstrained maximization procedure (i.e. `optim` in `R`) can be used.

The paper proceeds as follows; Section 2 will introduce our model, establish consistency and give the rate of convergence for our estimators of $\boldsymbol{\beta}$, $H(\cdot)$. In this section our estimator of $\boldsymbol{\beta}$ is shown to be asymptotic normal and semiparametrically efficient. Section 3 presents studies of the empirical properties of our model in practical sample sizes through simulation. Section 4 demonstrates the use of our model on the Leukemia data and the Veterans' Administration Lung Cancer study. Section 5 concludes with some future direction to the research. Appendix gives the technical details and proofs of the main theorem in section 2.

2. Transformation Models using Bernstein Polynomials

Assuming that model (1) holds, note that the survival distribution of T can be expressed as

$$S_T(t \mid \boldsymbol{Z}) = \bar{F}_\epsilon\{H(t) + \boldsymbol{\beta}^T \boldsymbol{Z}\}, \tag{2}$$

and $f_T(t \mid \boldsymbol{Z}) = H'(t) f_\epsilon(H(t) + \boldsymbol{\beta}^T \boldsymbol{Z})$, where $f_\epsilon(\cdot)$ denotes the density of ϵ and $\bar{F}_\epsilon(t) = 1 - F_\epsilon(t)$. This allows the likelihood, to be expressed as a function of $\bar{F}_\epsilon(\cdot)$ and $f_\epsilon(\cdot)$. Furthermore, note that

$$\bar{F}_\epsilon^{-1}\{S_T(t \mid \boldsymbol{Z})\} = H(t) + \boldsymbol{\beta}^T \boldsymbol{Z} \tag{3}$$

This relationship gives an idea of the assumption made by the transformation model, and how the transformation function is constructed. The assumption in (3) can be empirically validated by plotting $\bar{F}_\epsilon^{-1}[\tilde{S}_T(t \mid \boldsymbol{Z})]$, where $\tilde{S}$ is obtained nonparametrically, for different

values of $\boldsymbol{Z}$ and checking that the curves are equidistant in t. This is equivalent to checking the proportion hazards, or odds, assumption when $\bar{F}_\epsilon$ corresponds to these models.

It is straightforward to verify that H and $\boldsymbol{\beta}$ will identifiable when $\boldsymbol{Z}$ consists of p independent vectors. If there exists two error distributions such that (3) is satisfied for each, then the corresponding regression parameters and transformation functions will differ. That is, the true parameters will depend on the error specification.

For the remainder, assume that we have a fixed F_ϵ and that $\boldsymbol{\beta}_0$ and H_0 denote the true regression parameter and transformation for F_ϵ. Assume $\tau \equiv \inf\{t : \mathrm{pr}(T \wedge C > t) = 0\} < \infty$, $\boldsymbol{\beta} \in B \subseteq \mathbb{R}^p$ and $H \in \Gamma$ with,

$$\Gamma = \left\{ H(\cdot);\ M^- < H(t) < M^+,\ H'(t) > 0,\ t \in (0, \tau) \right\}.$$

where $H'(t) = \partial H(t)/\partial t$. The boundedness of H at 0 is required if $\mathrm{pr}(T \wedge C = 0) > 0$. Where convenient we will use the notation $\theta = (\boldsymbol{\beta}, H)$ and $\theta_0 = (\boldsymbol{\beta}_0, H_0)$, with $\theta \in \Theta = B \times \Gamma$.

$H(t)$ will be estimated by

$$\begin{aligned} H_N(t; \boldsymbol{\gamma}) \equiv H_N(t; \tau, \boldsymbol{\gamma}) &= \sum_{k=0}^{N} \gamma_k \binom{N}{k} \left(\frac{t}{\tau}\right)^k \left(1 - \frac{t}{\tau}\right)^{N-k} \\ &\equiv \sum_{k=0}^{N} \gamma_k \psi_{k,N}\left(\frac{t}{\tau}\right), \end{aligned}$$

where $\psi_{k,N}(q)$ is the probability mass function of a binomial random variable with sample size N, success probability q, evaluated at k. Bernstein polynomials have been recently used for density estimation in Brown and Chen (1999), Ghosal (2001), and Petrone and Wasserman (2002). Under the assumption that $H(\cdot)$ is continuous and bounded, the Bernstein–Weierstrass approximation theorem provides the uniform convergence of $H_N(t)$ to $H_0(t)$ as $N \to \infty$, when $\gamma_k = H_0(k/N)$ for $k = 0, \ldots, N$.

Our estimation procedure will require that $H_N(\cdot) \in \Gamma_N \subset \Gamma$ with,

$$\Gamma_N = \{H_N(\cdot;\boldsymbol{\gamma}); M^- < \gamma_0 \leq \cdots \leq \gamma_N < M^+\}, \quad N = 1, 2, \ldots,$$

and $\theta_N \in \Theta_N = B \times \Gamma_N$. The number of parameters for the transformation function will be a function of the sample size (i.e. $N \equiv N(n)$) and will be decided, in a finite sample, using information criteria (detailed in sections 3 and 4). In a limiting sense we will have $N = O(n^\kappa)$ where $0 < \kappa < 1$. The optimal choice of κ is given in theorem 2 and will depend on the smoothness of H_0.

The survival and density functions of T will be estimated by $S_N(t|\boldsymbol{Z}, \tau, \boldsymbol{\gamma}, \boldsymbol{\beta})$ and $f_N(t|\boldsymbol{Z}, \tau, \boldsymbol{\gamma}, \boldsymbol{\beta})$ with,

$$\begin{aligned} S_N(t \mid \boldsymbol{Z}, \tau, \boldsymbol{\gamma}, \boldsymbol{\beta}) &= \bar{F}_\epsilon \left\{H_N(t;\boldsymbol{\gamma}) + \boldsymbol{\beta}^T \boldsymbol{Z}\right\} \\ f_N(t \mid \boldsymbol{Z}, \tau, \boldsymbol{\gamma}, \boldsymbol{\beta}) &= f_\epsilon \left\{H_N(t;\boldsymbol{\gamma}) + \boldsymbol{\beta}^T \boldsymbol{Z}\right\} H'_N(t;\boldsymbol{\gamma}), \end{aligned} \tag{4}$$

where $H'_N(t;\boldsymbol{\gamma}) = \sum_{k=1}^N (\gamma_k - \gamma_{k-1})\psi_{k,N}(t/\tau)$. Next, $S_N(t \mid \boldsymbol{Z}, \tau, \boldsymbol{\gamma}, \boldsymbol{\beta})$ and $f_N(t \mid \boldsymbol{Z}, \tau, \boldsymbol{\gamma}, \boldsymbol{\beta})$ can be used to form the log-likelihood

$$l_N(\boldsymbol{\beta}, \boldsymbol{\gamma} \mid \tau, \boldsymbol{Y}) = \sum_{i=1}^n \Delta_i \log\{f_N(X_i \mid \boldsymbol{Z}_i, \tau, \boldsymbol{\gamma})\} + (1 - \Delta_i) \log\{S_N(X_i \mid \boldsymbol{Z}_i, \tau, \boldsymbol{\gamma})\}. \tag{5}$$

Let $\hat{\boldsymbol{\beta}}_{n,N}$ and $\hat{\boldsymbol{\gamma}}_{n,N}$ denote the values that maximize, $l_N(\boldsymbol{\beta}, \boldsymbol{\gamma})$ such that $\hat{\boldsymbol{\beta}}_{n,N} \in \mathbb{R}^p$ and $\hat{H}_N(\cdot) \equiv H_N(\cdot; \hat{\boldsymbol{\gamma}}_{n,N}) \in \Gamma_N$. We can similarly write $\hat{\theta}_{n,N} = \{\hat{\boldsymbol{\beta}}_{n,N}, \hat{H}_N\}$ maximizes

$$l_N(\theta_N \mid \tau, \boldsymbol{Y}) \text{ over } \Theta_N.$$

To verify various portions of the asymptotic properties the following conditions will be

needed.

(I) $H_0(\cdot)$ is increasing and bounded over $(0, \tau)$.

(II) Conditional on $\boldsymbol{Z} = \boldsymbol{z}$, T is independent of C, and $S_C(\cdot \mid \boldsymbol{Z})$ is ancillary for $(\boldsymbol{\beta}, H)$.

(III) For $r = 1, 2$ the rth derivative of $H_0(\cdot)$ exists, is bounded and continuous on $(0, \tau]$, and $\boldsymbol{\beta}$ lies in the interior of a compact set $B \subseteq \mathbb{R}^p$.

(IV) $\boldsymbol{Z}$ is bounded. There exists a Z^+ such that $\text{pr}(\| \boldsymbol{Z} \| \leq Z^+) = 1$, where $\| \cdot \|$ denotes the Euclidean norm in $\mathbb{R}^p$.

(V) The error distribution F_ϵ in independent of $\boldsymbol{Z}$ and defined on $\mathbb{R}$. Furthermore, $f_\epsilon(t)$ and $\bar{F}_\epsilon(t)$ are log-concave in t.

(VI) The distribution function of the censoring values $F_C(\cdot \mid \boldsymbol{Z})$ has first and second derivatives that are continuous and bounded on $(0, \tau]$. Furthermore, the bounds does not depend on $\boldsymbol{Z}$.

The second portion of assumption (V) is needed to verify that the log-likelihood in (5) is concave in H. This is a condition that is also required in M-estimation theory. If we consider the class of logarithmic error distributions given in Chen et al. (2002), where

$$\bar{F}_\epsilon(t) = e^{-G(e^t)}$$

with $G(t) = \log(1 + rt)/r$ for $r > 0$ and $G(t) = t$ for $r = 0$, then it is straightforward to show that assumption (V) is satisfied. This class of distributions includes the extreme value and logistic. Furthermore, for the class of Box–Cox type error distributions considered in Zeng and Lin (2006), where $G(t) = \{(1 + t)^\rho - 1\}/\rho$ for $\rho > 0$ and $G(t) = \log(1 + t)$ for $\rho = 0$, assumption (V) is satisfied when $\rho \leq 2$.

Convergence will be shown in terms of the metric d on $B \times \Gamma$, defined by

$$d\{(\boldsymbol{\beta}_1, H_1), (\boldsymbol{\beta}_2, H_2)\} = \| \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2 \| + \| H_1 - H_2 \|_2 \tag{6}$$

where $\| H_1 - H_2 \|_2 = \int |H_1(u) - H_2(u)| dF_X(u)$, with $F_X(u) = \text{pr}(T \wedge C \leq u)$.

Theorem 1. *(Consistency). Suppose that assumptions (I) - (IV) hold. Then*

$$d\{(\hat{\boldsymbol{\beta}}_{n,N}, \hat{H}_N), (\boldsymbol{\beta}_0, H_0)\} \to 0$$

almost surely, as $n \to \infty$.

The proof of this theorem uses the tools for consistency of approximate maximum likelihood estimators developed by Wang (1985). The appendix verifies the necessary conditions.

Recalling that $N = O(n^\kappa)$ the rate of convergence will be a function of κ. The rate at which N increases must be such that $N \to \infty$ and $N/n \to 0$ as $n \to \infty$. The theorem below gives the optimal choice for κ which is in terms of r in assumption (III).

Theorem 2. *(Rate of Convergence) Under assumptions (I) – (V), if* $\kappa = 1/(1+2r)$ *then*

$$d\{(\hat{\boldsymbol{\beta}}_{n,N}, \hat{H}_N), (\boldsymbol{\beta}_0, H_0)\} = O(n^{-r/(1+2r)}),$$

as $n \to \infty$.

As a result, the rate of convergence is $n^{1/3}$ and $n^{2/5}$ when H_0 has continuous bounded 1st and 2nd derivatives, respectively. This is the same rate of convergence achieved by Huang and Rossini (1997). The proof of this rate follows from theorem 2 of Shen and Wong (1994). The necessary conditions are verified in the appendix. This rate may be improved if a Lipshizt $\alpha > 1$ condition is imposed on H_0. Even though H_0 cannot be estimated at the

regular $n^{1/2}$ rate, β_0, as demonstrated in the following theorem, can be.

Theorem 3. *(Asymptotic Normality and Efficiency) Under assumptions (I) – (VI), if* $\kappa = 1/(1+2r)$

$$n^{1/2}(\hat{\boldsymbol{\beta}}_{n,N} - \boldsymbol{\beta}_0) = I(\boldsymbol{\beta}_0)^{-1} n^{1/2} \sum_{i=1}^{n} l^*_{\beta}(\boldsymbol{Y}, \theta_0) + o(1) \to N(0, \boldsymbol{I}^{-1}(\boldsymbol{\beta}_0))$$

in distribution, as $n \to \infty$*. The form of the efficient score function and information matrix for* $\boldsymbol{\beta}$*,* $l^*_{\beta}(\boldsymbol{Y}, \theta_0)$*, are given in the appendix (e.g. (13), and (14) respectively). Furthermore,* $\hat{\boldsymbol{\beta}}_{n,N}$ *is semiparametrically efficient in that it's asymptotic covariance achieves the semiparametric efficiency bound for* $\boldsymbol{\beta}_0$*.*

The proof of the semiparametric efficiency and asymptotic normality follow theorem 6.1 of Huang (1996). The necessary conditions are verified in the appendix, including the computation of the efficient score and information. The efficient score and information are useful for ascertaining asymptotic properties, but are difficult, and most often unnecessary, to compute in practice. The following section uses a numerical approximation to the observed Fisher information to calculate the standard error of $\hat{\boldsymbol{\beta}}$. Since our method is not very parametrically intensive this matrix is rather easy to compute.

3. Simulation Study

This section explores the accuracy of the regression parameter, and standard error estimates by use of simulation studies. These models were fit in R by use of the `optim` function. As mentioned in the previous section the standard error of $\hat{\boldsymbol{\beta}}$ is estimated via a numerical approximation to the observed Fisher information matrix. The constraint that $\gamma_i < \gamma_{i-1}$ is imposed using by reparameterizing the model with $\eta_0 = \gamma_0$ and $\exp(\eta_i) = \gamma_i - \gamma_{i-1}$ for

$i = 1, \ldots, N$. After generation of the data, $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ are estimated for four different values of N (4, 7, 10 and 13). These values were selected after some initial trial runs. Bayesian information criterion (BIC) and Akaike information criterion (AIC) shall be used to quantify the performance of each value of N, where

$$\begin{aligned} \text{BIC} &= -2l_N(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}} \mid \tau, \boldsymbol{Y}) + \log(q_n)(p + N + 1), \\ \text{AIC} &= -2l_N(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}} \mid \tau, \boldsymbol{Y}) + 2(p + N + 1). \end{aligned}$$

On a standard PC each iteration of the simulation took approximately 2.5 minutes to run for all four values of N.

The following studies simulated data from two contrived settings. In the first setting, given the covariate values Z_1 and Z_2, the event times are distributed according to an exponential distribution with rate parameter $\lambda = 0.2 \exp(-0.5Z_1 + 0.4Z_2)$. The covariates were randomly generated according to $Z_1 \sim Bernoulli(1/2)$ and $Z_2 \sim \text{N}(0, 1)$ (same for both simulation settings). The error distribution was specified as extreme value, and as a result the transformation model coincides with the Cox model. The censoring was generated from an exponential distribution with mean 20 (independent of T and Z), which resulted in around 17% of the observations being censored. For τ, we rounded up the maximum observed value.

Table 1 contains the results of the study for the Bernstein, and Cox models. Table 1 contains the average bias of the estimated regression parameters, the corresponding Monte Carlo standard error based on the $1,000$ runs, the average estimated standard error using the estimated observed Fisher information matrix, the empirical coverage probabilities of confidence intervals for regression parameters, and the percentage that a given Bernstein model was chosen using BIC and AIC.

Judging by the average bias, and the empirical coverage of $\hat{\beta} \pm 2SE\{\hat{\beta}\}$ it appears the AIC

Table 1: Simulation results using the exponential setting, given as $(\hat{\beta}_1, \hat{\beta}_2)$.

	Bernstein Method				
	N=4	7	10	13	Cox PL
AVG BIAS	(.09, -.07)	(.05,-.04)	(.03, -.03)	(.03, -.03)	(.003, -.004)
SE	(.29, .15)	(.27, .14)	(.26, .14)	(.26, .14)	(.24, .13)
AVG SE	(.23, .12)	(.23, .12)	(.23, .12)	(.23, .12)	(.23, .12)
95% CI	(.89, .88)	(.91, .92)	(.93, .93)	(.93, .94)	(.95, .95)
% AIC	0%	10%	42%	48%	–
% BIC	2%	42%	47%	10%	–

is more effective in choosing the appropriate model. In this setting the models with smaller N values are experiencing more bias, and have coverage rates well below the .95 level. The performance of Bernstein models with $N = 10$ and 13 are similar to those from the Cox model, which is included here as the gold standard. There is a slight loss of efficiency and slightly more bias in the Bernstein models.

The distribution assumptions in the second setting were chosen to correspond with the proportional odds regression model. That is, the data were distributed according the the distribution function $F(\cdot \mid Z)$ with, $logit\{F(t \mid Z)\} = \log\{t^2\} - (0.8Z_1 - 1.6Z_2)$ where $logit(x) = \log\{x/(1-x)\}$. The censoring variable was generated from an exponential distribution with mean 5 (independent of T and Z), which resulted in around 22% censored observations. Table 2 contains results from the study for the Bernstein model with $N = 4$, 7, 10, and 13, and from a full parametric maximum likelihood procedure.

The results for this simulation are much more consistent across N than in the previous setting. Note that both information criteria choose models that use less Bernstein parameters than in the first setting, even though they are ran with the same sample size. A reason for this is that distribution in the first setting is more right skewed. The per iteration τ were on average much larger in the first simulation ($\bar{\tau} = 23.91$ compared to $\bar{\tau} = 8.59$). The increase in right skewness causes the Bernstein function to be spread out over a longer range of data,

Table 2: Simulation results for the Proportional Odds Model

	Bernstein Method				
	N=4	7	10	13	MLE
AVG BIAS	(.01, -.02)	(.04, -.10)	(.05, -.11)	(.04, -.11)	(-.01, -.004)
SE	(.38, .40)	(.39, .41)	(.40, .42)	(.39, .42)	(.27, .35)
AVG SE	(.38, .36)	(.38, .37)	(.38, .37)	(.38, .37)	(.27, .34)
95% CI	(.95, .94)	(.95, .93)	(.95, .93)	(.95, .93)	(.96, .94)
% AIC	2%	41%	41%	16%	–
% BIC	10%	60%	26%	4%	–

hence the need for more parameters. So the robustness to the value of N in table 2 is due to distribution of the data. Table 1 shows the consequences of a model that is being under-fit for $N = 4$ and 7, and table 2 shows an example where the model is being slightly over-fit for $N = 13$. The conclusion to draw from this is that slight over-fitting appears to be better than under-fitting. For this reason we recommend that AIC be used to choose N.

4. Examples

This section demonstrates the use of the Bernstein method with two sets of data. In both examples we implement the following simple step-up strategy for choosing the best value of N. We start by fitting a Bernstein model with the initial value $N_0 = \lfloor n^{1/3} \rfloor$ (the lower of the asymptotic limits), let AIC_0 denote the AIC value of the initial model. We continue by fitting a models with $N = N_0 + 1, \ldots, N_0 + K_0$ with AIC values $\text{AIC}_1, \ldots, \text{AIC}_{K_0}$ respectively. The procedure stops at $K_0 = \min\{k : \text{AIC}_k - \text{AIC}_m > 3\}$ where AIC_m is the minimum AIC for $N = N_0, \ldots, N_0 + K_0$. The best value of N is then chosen as $N_0 + m$. If $m = 0$ then proceed with a step-down procedure to find the best AIC for $N \leq N_0$. When n is large (say > 100) bigger steps (two or three) will be preferable.

The first example uses data taken from Freireich (Cox, 1972) which is the remission times

Table 3: Results comparing estimates of the treatment effect from the leukaemia data set. This data set was fit using the Bernstein method with both Extreme Value and Logistic error distribution. The results from Cox (1972) and Cheng et al. (1995) are also included.

		Extreme Value Errors			
	N=3	4	5	Cox (1972)	Cheng et al. (1995)
Treatment	*-1.63(.41)*	-1.66 (0.42)	-1.68(0.43)	-1.51 (0.41)	-1.74 (0.41)
AIC	*229*	229	230	-	-
BIC	*238*	240	241	-	-
		Logistic Errors			
	N=2	3	4		
TRT	-2.52 (0.67)	*-2.42 (0.65)*	-2.41 (0.65)		
AIC	231	*230*	230		
BIC	238	*239*	241		

of 42 Leukaemia patients. This data has been used by many authors including Cox (1972), Cox and Oakes (1984), Cheng, Wei, and Ying (1995). This is a case/control design where half were treated with the drug 6-mercaptopurine and the rest are controls. The step-up procedure started with $N_0 = 3$, which ended up being the best value of N for both extreme value and logistic error distributions. Table 3 displays the estimates and standard errors for the regression parameter for the top three values of N according to AIC. The estimates and standard errors for the extreme value are similar to those from Cox (1972) and Cheng et al. (1995). Furthermore, the estimates are comparable across N.

Figure 1 contains plots of the estimated survival curves for the Leukaemia data. Each plot contains a the Kaplan–Meier survival curve separated by treatment, and the Bernstein estimate given by $S_N(t|Z, \hat{\tau}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}})$ in equation (4), for $Z = 0$ and 1. The Bernstein estimated smooth survival curves give similar estimates to the Kaplan–Meier curves. This data set contains 30 observed failure times, so the previous transformation models estimate 31 parameters. In contrast, the Bernstein models displayed in figure 1 estimate 5 parameters.

It appears from figure 1 that the extreme value error distribution fits the Kaplan–Meier

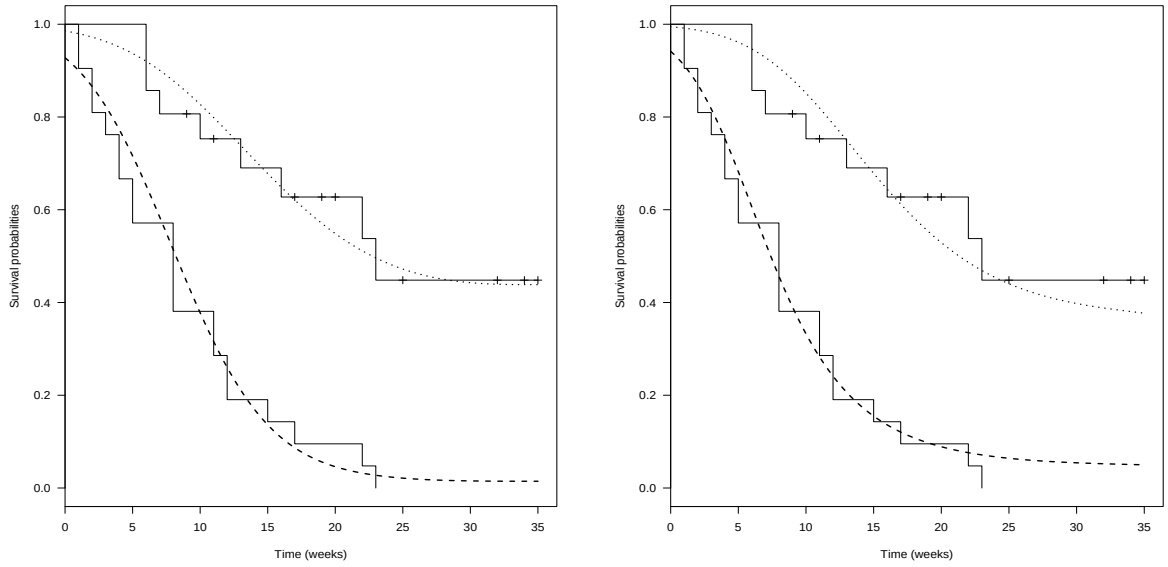

Figure 1: Comparison of Bernstein (N=3) and Kaplan–Meier Survival Curves for Leukemia Data using extreme value (left), and logistic (right) error distributions. The survival estimates for the control (−−) and treatment ($\cdots$) groups closely resemble the Kaplan–Meier step functions.

survival curve, and hence satisfies (3), slightly better than the logistic error distribution does. Furthermore, basing a choice of error distribution on AIC, BIC or deviance gives a small advantage to the extreme value (for the best values of N). However, the estimated survival curves are similar and both estimates of the treatment effect are approximately four standard deviations away from zero. As a result, one would surmise that similar conclusions would be drawn regardless of the choice of error distribution.

The second example uses another well known data set, the Veterans' Administration Lung Cancer study (Prentice, 1973). This data set has been fit with the proportional odds model in Bennett (1983), Pettitt (1984) and Murphy et al. (1997). Furthermore, it has appeared in the transformation model papers Cheng et al. (1995), Chen et al. (2002) and Zeng and Lin (2006). Similar to the previous works we analyze the data only for patients that did not receive any prior therapy. The covariates included in our model are performance

Table 4: Results comparing estimates of the treatment effect from the Veterans' Administration Lung Cancer study.

	Extreme Value Errors			Logistic Errors		
	N=9	10	Cox Model	N= 6	7	8
Small Cell	*.58(.32)*	.58(.32)	.55(.32)	1.54(.54)	*1.51(.53)*	1.48(.53)
Adeno	*.91(.35)*	.91(.35)	.86(.35)	1.44(.57)	*1.41(.56)*	1.37(.56)
Squamous	*-.22(.34)*	-.21(.34)	-.22(.35)	-.16(.62)	*-.16(.60)*	-.18(.60)
PS	*-.03(.006)*	-.03(.006)	-.02(.006)	-.056(.01)	*-.058(.01)*	-.058(.01)
AIC	*1058*	1059	-	1041	*1039*	1039
BIC	*1093*	1096	-	1068	*1069*	1071

	Cheng et al. (1995)	Murphy et al. (1997)	Chen et al. (2002)	Zeng and Lin (2006)
Small Cell	1.50(0.5)	1.44(.53)	1.23 (0.48)	1.38(.52)
Adeno	1.56(.41)	1.34(.56)	1.50 (0.53)	1.31(.55)
Squamous	-.006(.57)	-.22(.59)	-0.47 (0.61)	-.18(.59)
PS	-.055(.01)	-.055(.01)	-0.044 (0.01)	-.053(.01)

status, which is measured on a scale of 0 – 100, and cell type. Cell type is broken into four categories; large cell (chosen as the baseline group), small cell, adeno, and squamous.

The step-up procedure started with $N_0 = 4$, and found the best value to be $N = 9$ for the extreme value and $N = 7$ for the logistic. Table 4 summarizes the results from our model (top three AIC values) and those from Cheng et al. (1995), Murphy et al. (1997), Chen et al. (2002) and Zeng and Lin (2006). The results in Cheng et al. (1995), Chen et al. (2002) and Zeng and Lin (2006) contains estimates using the logistic and extreme value error distributions, but for brevity we only display those from the logistic. Their results using the extreme value are similar to the results from our model contained in the table. When focusing on the results from the logistic error distribution the estimates of the regression parameters for the Bernstein method appear to be comparable to those in the previous works.

The results are comparable for all values of N for both error distributions. All of the models for the logistic error distribution have a better fit than the best model for the extreme value distribution in terms of AIC, BIC and deviance. For these reasons the logistic error distribution with $N = 7$ is the recommended model. This model estimates a total of 12

parameters. Since the data set that has 91 observed failures, the previous transformation model methods estimate 95 parameters.

5. Conclusion

We have displayed a simple semiparametric transformation model. This method is less computationally intensive than the previous transformation models and appears to give similar results. Furthermore, this method outputs a smooth estimate of the time transformation, and as a result the survival curve. We have shown the estimate of the regression parameter to be asymptotically normal, and efficient. This method does require that the value of N be chosen, but does not appear to be sensitive to the choice. Furthermore, the AIC appears to do a sufficient job of picking a good model.

In the future it would be interesting to give a more general data framework that could be incorporated. This framework might include recurrent events, time dependent covariates, and dependent censoring.

Acknowledgement

The authors would like to thank the National Science Foundation for their support of this work through the Vertical Integration of Research and Education (VIGRE) program.

Appendix

Asymptotic properties

For all of the following proofs assume that we have a fixed F_ϵ, such that the model is correctly specified, and that H_0 and $\boldsymbol{\beta}_0$ denote the 'true' transformation and regression parameter for the chosen F_ϵ.

Consistency. The proof of theorem 1, which follows from theorem 3.1 of Wang (1985). For simplicity we shall show consistency for the case where $\boldsymbol{Z} \in \mathbb{R}$, the extension to $\mathbb{R}^p$ is straightforward. The proof will consist of showing that,

$$\int_{-\infty}^{\infty} \mid F_\epsilon\{\hat{H}_N(u) + \hat{\beta} Z\} - F_\epsilon\{H_0(u) + \beta_0 Z\} \mid dF_X(u \mid Z) \to 0 \tag{7}$$

almost surely as $n \to \infty$. Theorem 1 follows from the boundedness of $H_0(\cdot)$ on $(0, \tau]$, the continuity of F_ϵ, and dominated convergence theorem.

The fact that $\hat{\theta}_N$ is an approximate maximum likelihood estimator follows from the Bernstein–Weierstrass approximation theorem. This gives us that $\hat{\theta}_N$ maximizes the likelihood in (5) as $N \to \infty$. For convenience let $\pi(\hat{\theta}_N, t, Z) = F_\epsilon\{\hat{H}_N(u) + \hat{\beta} Z\}$, with $\pi(\theta_0, t, Z) = F_T(t \mid Z)$.

To show (7) from theorem 3.1 of Wang (1985), we must verify the necessary conditions. The first condition if verified by the definition of our metric in (7), and the assumption that B is a bounded subset of $\mathbb{R}$. This gives us that (Θ, d) is a separable compact metric space. Define $V_r(\theta_0)$, $r \geq 1$ as a decreasing sequence of basic neighborhoods of θ_0 with radius r^{-1}. Furthermore, define $A_\varphi(\theta)$ be the following,

$$A_\varphi(\theta) = \varphi \pi(\theta, t, Z) + (1 - \varphi)\pi(\theta_0, t, Z).$$

For all $\theta \in \Theta$, let $l\{\boldsymbol{Y}, A_\varphi(\theta)\} = \Delta_i \log\{A'_\varphi(\theta)\} + (1 - \Delta_i) \log\{A_\varphi(\theta)\}$ with $A'_\varphi(\theta) = \partial A_\varphi(\theta)/\partial t$. Notice that for every $r \geq 1$ there exists a $\varphi \in (0, 1]$ with $\varphi \leq r^{-1}$ such that $A_\varphi(\theta) \in V_r(\theta_0)$. Under assumption (IV) we will need to show the following to verify condi-

tions 2 and 3:

$$E_{(X,\Delta)}\left[l\{\boldsymbol{Y}, A_\varphi(\theta)\}/l(\boldsymbol{Y},\theta) \mid Z\right] > 0. \tag{8}$$

Notice that if $l(X,\Delta,Z,\theta) = 0$ for any X or Z in the range of X and Z, (8) holds. As a result the following assumes that for all X and Z, $l(X,\Delta,Z,\theta) > 0$. It is straightforward to show that the marginal density of X is,

$$f_X(x \mid Z) \;\; = \;\; S_C(x) f_t(x \mid Z) + S_T(x) f_c(x \mid Z).$$

Then, letting $\pi' = (\partial/\partial t)\pi$, note that

$$\int \frac{\pi'(\theta,u,Z)}{\pi'(\theta_0,u,Z)} S_C(u) f_T(u \mid Z) du = \int \pi'(\theta,u,Z) S_C(u) du < \int \pi'(\theta,u,Z) du \leq 1,$$

and

$$\int \frac{\pi(\theta,u,Z)}{\pi(\theta_0,u,Z)} S_T(u) f_C(u \mid Z) du = \int \pi(\theta,u,Z) f_C(u) du < \int f_C(u) du = 1.$$

Using Jensens inequality we get that

$$\int \log\left\{\frac{\pi'(\theta,u,Z)}{\pi'(\theta_0,u,Z)}\right\} S_C(u) f_T(u \mid Z) du \;\; < \;\; 0$$

and similarly when using π. Combining this we get,

$$\begin{aligned}
E_X\left[\frac{l(X,1,z,A_\varphi(\theta))}{l\{X,1,z,\theta\}} \mid Z = z, \Delta = 1\right] &= E_X\left[\log\left\{\frac{A_\varphi(\theta)}{\pi'(\theta,X,Z)}\right\} \mid Z, \Delta = 1\right] \\
&\geq E_X\left[(1-\varphi)\log\left\{\frac{\pi'(\theta_0,X,Z)}{\pi'(\theta,X,Z)}\right\} \mid Z, \Delta = 1\right] \\
&= (1-\varphi) E_X\left[(1-\varphi)\log\left\{\frac{\pi'(\theta_0,X,Z)}{\pi'(\theta,X,Z)}\right\} \mid Z, \Delta = 1\right] > 0
\end{aligned}$$

and similarly for $\Delta = 0$ using π. This verifies (8). Conditions 4 and 5 are verified by the continuity of $l\{\boldsymbol{Y}, \theta\}$. This proves the theorem.

Rate of Convergence. The proof follows from verifying the four condition in theorem 2 of Shen and Wong (1994). For the remainder let $E(g) = E_Y(g)$. Define the function,

$$H_{0N}(t) = \sum_{k=0}^{N} H_0(k/N)\psi_{k,N}(t/\tau)$$

as the true Bernstein function. Under assumption (III), we can apply (B) of section 1.6 in Lorentz (1986) to find that,

$$|H_0(t) - H_{0N}(t)| \leq CN^{-r}$$

for some constant $C > 0$, which gives

$$\| H_0 - H_{0N} \|_2 = O(n^{-2r\kappa})$$

where $\| \cdot \|_2$ is defined in (6) of our paper, and

$$E\{l(Y, \theta_0) - l(Y, \theta_{0N})\} = O(n^{-2r\kappa}).$$

Define the metric

$$\rho^2(\theta, \theta_0) = E\{|\beta - \beta_0|^T \dot{l}_\beta(\boldsymbol{Y}, \theta) + \dot{l}_H(\boldsymbol{Y}, \theta)|H - H_0|\}^2$$

where $\dot{l}_\beta$ is the vector of score functions for β, and $\dot{l}_H$ is the score operator for H. Huang and Rossini (1997) verify that if $\rho(\theta, \theta_0) = O(r_n)$ then $d(\theta, \theta_0) = O(r_n)$. Condition 1 is verified from assumption (IV) and by the use of this metric. This metric bounds the third-order

Taylor–series expansion of $l(Y,\theta)$ about $l(Y,\theta_0)$. That is, it can be shown that

$$\inf_{\{\rho(\theta,\theta_0)\geq\delta,\theta\in\Theta_N\}} E\{l(\boldsymbol{Y},\theta_0)-l(\boldsymbol{Y},\theta)\}\geq 2C_1\delta^2.$$

for some constant $C_1>0$. Now that we have established that the mean difference in log-likelihoods is bounded below for $\rho(\theta,\theta_0)\geq\delta$ we will look to establish that the variance is decreasing with δ. This can be shown by using lemma 1 and a Taylor–series expansion, which yields that

$$\sup_{\{\rho(\theta,\theta_0)\leq\delta,\theta\in\Theta_N\}} E\{l(\boldsymbol{Y},\theta_0)-l(\boldsymbol{Y},\theta)\}^2\leq C_2\rho^2(\theta,\theta_0).$$

for some constant $C_2>0$. Condition 2 follows from the property,

$$\sup_{\{\theta\in\Theta_N\}} E\{l(\boldsymbol{Y},\theta_0)-l(\boldsymbol{Y},\theta)\}\leq C_3,$$

for some constant $C_3>0$. This property holds on f_X since $H\in\Gamma$ is bounded on f_X, f_ϵ is defined on $\mathbb{R}$, B is a bounded subset, and $|Z|$ is bounded by Z^+. This boundedness also verifies condition 4.

To show the final condition we begin by defining $B(\boldsymbol{\beta}_0,\delta)$ as a ball centered at $\boldsymbol{\beta}_0$ with radius δ with $B(\boldsymbol{\beta}_0,\delta)\subseteq B$ (this is guaranteed since B is compact with $\boldsymbol{\beta}_0$ not on the boundary). Similarly we define

$$\begin{aligned} B(H_0,\delta) &= \{H:\parallel H-H_0\parallel<\delta\} \\ B(\theta_0,\delta) &= \{\theta: d\{(\boldsymbol{\beta},\boldsymbol{\beta}_0),(H,H_0)\}<\delta\} \end{aligned}$$

Notice that if $\theta\in B(\theta_0,\delta)$ then $H\in B(H_0,\delta)$ and $\boldsymbol{\beta}\in B(\boldsymbol{\beta}_0,\delta)$.

Let $L_2(F_X) = \{f : \int f^2 dF_X < \infty\}$, if for $\mathcal{F} \subset L_2(F_X)$ there exists $S(\varepsilon, m) = \{f_1^l, f_1^u, \ldots, f_m^l, f_m^u\} \subset L_2(F_X)$ with $\| f_j^u - f_j^l \|_2 \leq \varepsilon$ for all $j = 1, \ldots, m$ and $f_j^u \leq f \leq f_j^l$ almost everywhere for any $f \in \mathcal{F}$, then $S(\varepsilon, m)$ is the bracketing ε-covering of $\mathcal{F}$ with respect to $\| \cdot \|_2$. Let $N_{[]}(\varepsilon, \mathcal{F}, L_2(F_X))$ denote the minimum m such that $S(\varepsilon, m)$ is a bracketing ε-covering of $\mathcal{F}$. Furthermore, let $H_{[]}(\varepsilon, \mathcal{F}, L_2(F_X)) = \log N_{[]}(\varepsilon, \mathcal{F}, L_2(F_X))$ denote the bracketing $\| \cdot \|_2$ entropy of $\mathcal{F}$.

From example 3 in Shen and Wong (1994), since $H \in \Gamma$,

$$H_{[]}(\varepsilon, B(H_0, \delta), L_2(F_X)) \leq C_4 N \log(\delta/\varepsilon) \leq C_5 n^{\kappa} \log(\delta/\varepsilon).$$

Then from Huang (1996),

$$H_{[]}(\varepsilon, B(\boldsymbol{\beta}_0, \delta), L_2(F_X)) \leq C_6 p \log(1/\varepsilon) \tag{9}$$

Putting these results together we find that

$$H_{[]}(\varepsilon, B(\theta_0, \delta), L_2(F_X)) \leq C_6 p \log(1/\varepsilon) + C_5 n^{\kappa} log(\delta/\varepsilon),$$

For large n the second term dominates so we have

$$H_{[]}(\varepsilon, B(\theta_0, \delta), L_2(F_X)) \leq C_5 n^{\kappa} \log(\delta/\varepsilon),$$

condition 4 follows from lemma 2.1 of Ossiander (1987) on the bracketing entropy of a bounded function.

From the verified conditions and the result on the convergence rate of H_{0N} to H_0 we have

$$\rho(\hat{\theta}_N, \theta_0) = O\{\max(n^{-\alpha}, n^{-\kappa r})\}$$

where $\alpha = (1-\kappa)/2 - \log 2/(2\log n)$. This rate is maximized by setting $\kappa = 1/(1+2r)$. The rate stated in the theorem 2 is for large n.

Efficient Score and Information Calculation. This calculation of the efficient score function for $\boldsymbol{\beta}$ follows Bickel et al. (1993) page 147.

We begin by letting $\epsilon = H_0(T) + \boldsymbol{\beta}_0^T \boldsymbol{Z}$, $\zeta = H_0(C) + \boldsymbol{\beta}_0^T \boldsymbol{Z}$, with ϵ_i and ζ_i being based on $(T_i, C_i, \boldsymbol{Z}_i)$. If we let $\boldsymbol{Y}^\circ = (T, \boldsymbol{Z})$ denote the unobserved true data, then log-likelihood based on $\boldsymbol{Y}^\circ$ is,

$$l(\boldsymbol{Y}^\circ, \theta) = f_\epsilon\{H(T) + \boldsymbol{\beta}^T \boldsymbol{Z}\} H'(T).$$

Furthermore, the score for $\boldsymbol{\beta}$ and H based on $\boldsymbol{Y}^\circ$ are

$$\begin{aligned} \dot{l}_\beta(\boldsymbol{Y}^\circ, \theta) &= \boldsymbol{Z} \frac{f'_\epsilon\{H(T) + \boldsymbol{\beta}^T \boldsymbol{Z}\}}{f_\epsilon\{H(T) + \boldsymbol{\beta}^T \boldsymbol{Z}\}} \equiv Z\phi(\epsilon) \qquad (10) \\ \dot{l}_H(\boldsymbol{Y}^\circ, \theta)[a] &= a\{H(T) + \boldsymbol{\beta}^T \boldsymbol{Z}\} \qquad (11) \end{aligned}$$

where $\phi(t) = \partial \log f_\epsilon(t)/\partial t$ and $a \in L_2^0(F_T) = \{f : \int f dF_T = 0 \text{ and } \int f^2 dF_T < \infty\}$. From Bickel et al. (1993) and Groeneboom and Wellner (1992), the score for $\boldsymbol{\beta}$ and H based on $\boldsymbol{Y}$ are

$$\begin{aligned} \dot{l}_\beta(\boldsymbol{Y}, \theta) &= E\{\dot{l}_\beta(\boldsymbol{Y}^\circ, \theta) \mid \boldsymbol{Y}\} \\ \dot{l}_H(\boldsymbol{Y}, \theta)[a] &= E\{\dot{l}_H(\boldsymbol{Y}^\circ, \theta)[a] \mid \boldsymbol{Y}\}. \end{aligned}$$

Let,

$$M(t) = I_{\{(\epsilon \wedge \zeta) \le t\}} - \int_{-\infty}^{t} I_{\{(\epsilon \wedge \zeta) \le s\}} d\Lambda_\epsilon(s)$$

where $\Lambda_\epsilon(\cdot)$ denotes the cumulative hazard function of ϵ. Furthermore define the R function as,

$$R_{F_\epsilon} a(t) = a(t) - E_{F_\epsilon}\{a(\epsilon) \mid \epsilon > t\}$$

where we will use the simplified notation $R_{F_\epsilon} a(t) \equiv Ra(t)$. It can be shown (cf. Bickel et al., 1993, page 435) that

$$\begin{aligned} \dot{l}_\beta(\boldsymbol{Y}, \theta) &= Z \int R\phi(u) dM(u) \\ \dot{l}_H(\boldsymbol{Y}, \theta)[a] &= \int Ra(u) dM(u). \end{aligned}$$

To find the efficient score function we want to find an a^* such that

$$E[\{\dot{l}_\beta(\boldsymbol{Y}, \theta) - \dot{l}_H(\boldsymbol{Y}, \theta)[a^*]\}\dot{l}_H(\boldsymbol{Y}, \theta)[a]] = 0 \tag{12}$$

for all $a \in L_2^0(F_T)$. To find a^* note that the left hand side can be written as,

$$\begin{aligned} & E(E[\{\dot{l}_\beta(\boldsymbol{Y}, \theta) - \dot{l}_H(\boldsymbol{Y}, \theta)[a^*]\}\dot{l}_H(\boldsymbol{Y}, \theta)[a] \mid Z, \Delta]) \\ &= E(E[\int I_{\{(\epsilon \wedge \zeta) \geq \cdot\}}\{ZR\phi - Ra^*\} Rad\Lambda_\epsilon \mid Z, \Delta]) \\ &= E\{\int I_{(\zeta \geq \cdot)}(ZR\phi - Ra^*) RadF_\epsilon\} \\ &= \int \left[E\{ZI_{(\zeta \geq \cdot)}\}R\phi - E\{I_{(\zeta \geq \cdot)}\}Ra^*\right] RadF_\epsilon \end{aligned}$$

since ϵ is independent of Z and ζ. As a result, if we let

$$L\{b(t)\} = b(t) - \int_{-\infty}^{t} bd\Lambda_\epsilon$$

then (12) is satisfied by setting

$$\begin{aligned} Ra^*(s) &= \frac{E\{ZI_{(\zeta \geq s)}\}}{E\{I_{(\zeta \geq s)}\}} R\phi(s) \\ a^*(s) &= L\{E(Z \mid \zeta \geq s) R\phi(s)\}. \end{aligned}$$

Therefore, the efficient score function for $\boldsymbol{\beta}$ is,

$$\begin{aligned} l^*_\beta(\boldsymbol{Y}, \theta) &= \dot{l}_\beta(\boldsymbol{Y}, \theta) - \dot{l}_H(\boldsymbol{Y}, \theta)[a^*] \\ &= \int \{Z - E(Z \mid H_0(X) + \boldsymbol{\beta}_0^T Z \geq s)\} R\phi(s) dM(s), \end{aligned} \tag{13}$$

since Z is independent of ϵ and $I_{(\epsilon \wedge \zeta \geq s)} = I_{\{H_0(X) + \beta_0^T Z \geq s\}}$. Since $M(\cdot)$ is a mean zero martingale $l^*_\beta(\boldsymbol{Y}, \theta)$ will be a zero mean martingale (cf. Fleming and Harrington, 1991). The predictable quadratic variation of $l^*_\beta(\boldsymbol{Y}, \theta)$ can then be used to obtain an expression for the expected information. As a result, the efficient information for $\boldsymbol{\beta}$ is

$$\begin{aligned} I(\boldsymbol{\beta}) &= E\{l^*_\beta(\boldsymbol{Y}, \theta)^2\} \\ &= \int I_{\{(\epsilon \wedge \zeta) \geq s\}} [\{Z - E(Z \mid H_0(X) + \boldsymbol{\beta}_0^T Z \geq s)\} R\phi(s)]^2 d\Lambda_\epsilon(s). \end{aligned} \tag{14}$$

Note that $I(\boldsymbol{\beta})$ is positive definite.

Efficiency and Asymptotic Normality. As mentioned in theorem 3, the proof of this theorem follows theorem 6.1 of Huang (1996). We will continue to use the notation, $E(g) = E_Y(g)$, and for the score for $\boldsymbol{\beta}$ and H that were introduced in (10) and (11). Furthermore, let

$$\begin{aligned} S_{\beta n}(\hat{\theta}) &\equiv P_n \dot{l}_\beta(\boldsymbol{Y}, \hat{\theta}) \\ S_{Hn}(\hat{\theta})[a] &\equiv P_n \dot{l}_H(\boldsymbol{Y}, \theta)[a] \end{aligned}$$

where P_n is the empirical measure of the observations $Y_1, \ldots, Y_n$. For the remainder we will use linear functional notation, thus $Pf = \int f dP$. Before checking conditions 1–5 of Huang (1996), we will verify (i) of theorem 6.1 which is that $S_{\beta n}(\hat{\theta})$ and $S_{Hn}(\hat{\theta})[a^*]$ are $o(n^{-1/2})$. Using theorem 1, and the compactness of B, $S_{\beta n}(\hat{\theta}) = 0$ with probability 1 for large n. Note that since we are carrying out the optimization over the constrained set Γ_N, we will not have $S_{Hn}(\hat{\theta})[a^*] = 0$ for large n.

The verification that $S_{Hn}(\hat{\theta})[a^*] = o(n^{-1/2})$ follows similar steps to those given in Huang and Rossini (1997). First we construct a function a_n^* that approximates a^* such that,

$$\int_0^\tau |a_n^*(t) - a^*(t)| dt \leq O(\parallel \hat{H}_N - H_0 \parallel_2)$$

Using assumption (V), it can be verified that the likelihood in $l(\hat{\boldsymbol{\beta}}, H)$ is concave in H. Seeing that H_N is a linear function of $\gamma_0, \ldots, \gamma_N$, $l(\hat{\boldsymbol{\beta}}, H)$ is therefore concave in $\gamma_0, \ldots, \gamma_N$. Then by the Lagrangian duality for constrained concave maximization, Rockafellar (1970),

$$\dot{l}_{\gamma_k}(\hat{\beta}, \hat{H}_N) = (\partial/\partial\gamma_k) l(\hat{\boldsymbol{\beta}}, H_N)|_{H_N = \hat{H}_N} = 0.$$

This shows that $S_{Hn}(\hat{\theta})[a_n^*] = o(n^{-1/2})$. Letting P be the joint measure of $\boldsymbol{Y}$, notice that

$$\begin{aligned} S_{Hn}(\hat{\theta})[a^*] - S_{Hn}(\hat{\theta})[a_n^*] &= S_{Hn}(\hat{\theta})[a^* - a_n^*] = P_n \dot{l}_H(\boldsymbol{Y}, \hat{\theta})[a^* - a_n^*] \\ &= (P_n - P)\{\dot{l}_H(\boldsymbol{Y}, \hat{\theta})[a^* - a_n^*]\} \\ &\quad + P\{\dot{l}_H(\boldsymbol{Y}, \hat{\theta})[a^* - a_n^*] - \dot{l}_H(\boldsymbol{Y}, \theta_0)[a^* - a_n^*]\} \end{aligned}$$

since $P\{\dot{l}_H(\boldsymbol{Y}, \hat{\theta})[a^* - a_n^*]\} = 0$. Using a similar argument to lemma 1, $\dot{l}_H(\boldsymbol{Y}, \hat{\theta})[a^* - a_n^*]$ is a Donsker class of functions. As a result, the first portion of the last line is asymptotical equicontinuous, and by the Arzelà–Ascoli theorem is $o(n^{-1/2})$. The second portion of the

last line can be bounded by the Cauchy–Schwartz inequality, that is, if $\kappa = 1/(1+2r)$ this term is,

$$O(n^{-r/(1+2r)})^2 = O(n^{-2r/(1+2r)}),$$

and $o(n^{-1/2})$ for $r = 1$, 2. Consequently, $S_{Hn}(\hat{\theta})[a^*] = o(n^{-1/2})$.

The remaining regularity conditions are rather straightforward to confirm. Condition 1 is a consequence of Theorem 2. The positive definiteness of $I(\beta)$ shown in the previous section verifies condition 2. If $\kappa = 1/(1+2r)$ the rate of convergence of $\hat{H}_N$ is faster than $n^{1/4}$, and hence condition 4 can be verified with $\alpha = 2$. A Taylor expansion shows condition 4. Condition 5 follows from assumption (V) on the properties of F_ϵ, which guarantees that $I(\boldsymbol{\beta}) < \infty$.

Condition 3 follows from the following result from lemma 7.1 of Huang (1996).

Lemma 1. *For $\delta > 0$ the define classes of functions,*

$$\begin{aligned}
\Psi_\beta(\delta) &= \{\dot{l}_\beta(\boldsymbol{Y}, \theta) - \dot{l}_\beta(\boldsymbol{Y}, \theta_0); |\theta - \theta_0| \leq \delta, \parallel H - H_0 \parallel \leq \delta\} \\
\Psi_H(\delta) &= \{\dot{l}_H(\boldsymbol{Y}, \theta)[a^*] - \dot{l}_H(\boldsymbol{Y}, \theta_0)[a^*]; |\theta - \theta_0| \leq \delta, \parallel H - H_0 \parallel \leq \delta\}.
\end{aligned}$$

Then for ε close to zero, using the $H_{[]}$ bracketing notation introduced earlier,

$$H_{[]}(\varepsilon, \Psi_\beta, L_2(F_X)) \leq C_1 1/\varepsilon, \ \& \ H_{[]}(\varepsilon, \Psi_H, L_2(F_X)) \leq C_2 1/\varepsilon.$$

for some constants C_1 and C_2. Hence, Ψ_β and Ψ_H are Donsker classes of functions.

Proof: From van de Geer (1993), the L_2 covering number for the class of uniformly bounded functions Γ is on the order of $O(e^{1/\delta})$. Furthermore, from Huang (1996) the covering number for β is $O(1/\delta^p)$. As a result, for δ close to 0, the entropy number for the class of

functions indexed by β and H such that $|\theta - \theta_0| \leq \delta$, and $\| H - H_0 \| \leq \delta$ is on the order $O(1/\delta)$. The lemma is obtained by adding assumption (IV).

References

Bennett, S. (1983), "Analysis of survival data by the proportional odds model," *Statistics in Medicine*, 2, 273–277.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993), *Efficient and adaptive estimation for semiparametric models*, Johns Hopkins Series in the Mathematical Sciences, Baltimore, MD: Johns Hopkins University Press.

Brown, B. M. and Chen, S. X. (1999), "Beta-Bernstein Smoothing for Regression Curves with Compact Support," *Scandinavian Journal of Statistics*, 26, 47–59.

Chen, K., Jin, Z., and Ying, Z. (2002), "Semiparametric Analysis of Transformation Models with Censored Data," *Biometrika*, 89, 659–668.

Cheng, S. C., Wei, L. J., and Ying, Z. (1995), "Analysis of Transformation Models with Censored Data," *Biometrika*, 82, 835–845.

Cox, D. R. (1972), "Regression models and life-tables," 34, 187–220.

Cox, D. R. and Oakes, D. (1984), *Analysis of survival data*, Monographs on Statistics and Applied Probability, London: Chapman & Hall.

Dabrowska, D. M. and Doksum, K. A. (1988), "Partial Likelihood in Transformation Models with Censored Data," *Scandinavian Journal of Statistics*, 15, 1–23.

Fleming, T. R. and Harrington, D. P. (1991), *Counting processes and survival analysis*, Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, New York: John Wiley & Sons Inc.

Gehan, E. A. (1965), "A Generalized Two-Sample Wilcoxon Test for Doubly Censored Data," *Biometrika*, 52, 650–653.

Ghosal, S. (2001), "Convergence Rates for Density Estimation with Bernstein Polynomials," *The Annals of Statistics*, 29, 1264–1280.

Groeneboom, P. and Wellner, J. A. (1992), *Information bounds and nonparametric maximum likelihood estimation*, vol. 19 of *DMV Seminar*, Basel: Birkhäuser Verlag.

Hinkley, D. V. and Runger, G. (1984), "The Analysis of Transformed Data," *Journal of the American Statistical Association*, 79, 302–309.

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

Huang, J. and Rossini, A. J. (1997), "Sieve estimation for the proportional-odds failure-time regression model with interval censoring," *J. Amer. Statist. Assoc.*, 92, 960–967.

Lorentz, G. G. (1986), *Bernstein polynomials*, New York: Chelsea Publishing Co., 2nd ed.

Murphy, S. A., Rossini, A. J., and Vaart, A. W. v. d. (1997), "Maximum Likelihood Estimation in the Proportional Odds Model," *Journal of the American Statistical Association*, 92, 968–976.

Ossiander, M. (1987), "A Central Limit Theorem Under Metric Entropy with L2 Bracketing," *The Annals of Probability*, 15, 897–919.

Petrone, S. and Wasserman, L. (2002), "Consistency of Bernstein Polynomial Posteriors," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 64, 79–100.

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.

Rockafellar, R. T. (1970), *Convex analysis*, Princeton Mathematical Series, No. 28, Princeton, N.J.: Princeton University Press.

Shen, X. (1998), "Proportional odds regression and sieve maximum likelihood estimation," *Biometrika*, 85, 165–177.

Shen, X. and Wong, W. H. (1994), "Convergence rate of sieve estimates," *Ann. Statist.*, 22, 580–615.

van de Geer, S. (1993), "Hellinger-consistency of certain nonparametric maximum likelihood estimators," *Ann. Statist.*, 21, 14–44.

Wang, J.-L. (1985), "Strong consistency of approximate maximum likelihood estimators with applications in nonparametrics," *Ann. Statist.*, 13, 932–946.

Wei, L. J. (1992), "The accelerated failure time model: A useful alternative to the cox regression model in survival analysis," *Statist. Med.*, 11, 1871–1879.

Zeng, D. and Lin, D. Y. (2006), "Efficient estimation of semiparametric transformation models for counting processes," *Biometrika*, 93, 627–640.

— (2007), "Semiparametric transformation models with random effects for recurrent events," *J. Amer. Statist. Assoc.*, 102, 167–180.