12-7-2006

# A Semiparametric Approach for the Nonparametric Transformation Survival Model With Multiple Covariates

Xiao Song
*University of Washington*, songx@u.washington.edu

Shuangge Ma
*Yale University*, shuangge.ma@yale.edu

Jian Huang
*University of Iowa*, jian@stat.uiowa.edu

Xiao-Hua Zhou
*University of Washington*, azhou@u.washington.edu

## 1. Introduction

Semiparametric models are widely used to characterize the relationship between survival time and covariates. Popular semiparametric models include the proportional hazards model (Cox, 1972), the additive hazards model (Aalen, 1980) and the accelerated failure time model (Kalbfleisch & Prentice, 1980). Although these semiparametric models are more flexible than parametric models, they may still be restrictive in practice. Recently, efforts have been devoted to relaxing assumptions on these models. One attractive extension is the transformation model, which assumes that an unknown monotone transformation of the survival time depends on the covariates through a linear model. Cheng et al. (1995), Fine et al. (1998) and Chen et al. (2002) studied the case when the error has a known distribution and the transformation function is unspecified, and Cai et al. (2003) explored the situation of a specific parametric transformation (Box-Cox transformation) with no parametric assumptions on the error; the former includes the proportional hazards model as a special case and the latter includes the accelerated failure time model as a special case. A more flexible form is the nonparametric transformation model that makes no parametric assumptions on either the transformation function or the error. This model includes the aforementioned models as special cases.

Several approaches have been proposed for estimating the regression parameters in the nonparametric transformation model with uncensored outcomes, including the maximum rank correlation estimator (Han 1987) and the monotone rank estimator (Cavanagh & Sherman 2001). Both estimators can be extended to the case of censored survival outcomes using the inverse censoring probability weighting technique (Khan & Tamer 2004). However, this requires that the censoring time is independent of the survival time and the covariates, and that the support of censoring time contains that of the survival time. Such restrictions may be unrealistic even for randomized clinical trials. Recently, Khan & Tamer (2004) proposed an appealing partial rank (PR) approach that relaxes these assumptions. This approach allows the censoring time to depend on the covariates as long as it is conditionally independent of the survival time given the covariates, and there is no restriction on the support of the censoring time. However, like the maximum rank correlation estimator and the monotone rank estimator, the partial rank estimator is based on maximization of a discontinuous function. It is difficult to compute this estimator when there are multiple covariates.

In this paper, based on the partial rank estimator of Khan & Tamer (2004), we propose a new estimator, called the smoothed partial rank (SPR) estimator, for estimating the regression parameters in the nonparametric transformation model. The proposed estimator maximizes a smoothed partial rank objective function. Smoothing makes it feasible to adapt the rank based approach to data with multiple covariates, without loss of asymptotic efficiency. We further propose using the weighted bootstrap for computation of the variance by analogy to that introduced in Jin et al. (2001).

The paper is organized as follows. We give the model definition in Section 2. The estimator is derived in Section 3. We show the asymptotic properties and propose the weighted bootstrap method in Section 4. The finite sample properties of the estimator is assessed by simulation studies in Section 5. We apply the approach to a real dataset in Section 6. The paper concludes with discussions in Section 7.

## 2. Model definition

Let $T$ denote the survival time, $C$ denote the censoring time, and $\boldsymbol{Z}$ be a length $d$ vector of covariates. Under right censoring, the observed survival data are $V = \min(T, C)$ and $\Delta = I(T \leq C)$. Assume that the survival time depends on the covariate through the nonparametric transformation model,

$$g(T) = \beta' \boldsymbol{Z} + e, \tag{1}$$

where $g(\cdot)$ is an unspecified monotone function, $e$ is the random error term with an unknown distribution independent of $\boldsymbol{Z}$, $\beta$ is a length $d$ regression coefficient, and $\beta'$ denotes the transpose of $\beta$. This model is very flexible and includes many of the popular models as special cases. For example, the proportional hazards model and the proportional odds model are two special cases of (1) with $e$ following the standard extreme value and logistic distributions, respectively; the accelerated failure time model is a special case of (1) with $g(\cdot) = \log(\cdot)$. However, the additive hazards model (Aalen 1980) is not a transformation model.

The nonparametric transformation model is location and scale invariant. To avoid identifiability problem and without loss of generality, the first element of $\beta$ is restricted to 1, that is, $\beta = (1, \theta')'$. Our interest focuses on estimation of $\theta$.

## 3. Estimation

Suppose that the observed data $(V_i, \Delta_i, \boldsymbol{Z}_i)$, $i = 1, \ldots, n$, are independent and identically distributed as $(V, \Delta, \boldsymbol{Z})$. The partial rank (PR) approach (Khan & Tamer 2004) is based on the partial ranks of the observed survival times, the event indicators and the linear predictors. Specifically, the partial rank estimator has the form

$$\tilde{\beta} = \arg\max_{\beta} \left\{ \tilde{O}_n(\beta) = \frac{1}{n(n-1)} \sum_{i \neq j} \Delta_j I(V_i \geq V_j) I\left(\beta' \boldsymbol{Z}_i > \beta' \boldsymbol{Z}_j\right) \right\}, \tag{2}$$

where $I(\cdot)$ is the indicator function, and $\tilde{\beta} = (1, \tilde{\theta}')'$. The objective function $\tilde{O}_n$ can be considered as generalization of Kendall's $\tau$-correlation statistic to censored data, which is a U-statistic of order 2. Loosely speaking $\tilde{\beta}$ seeks to maximize the Kendall's $\tau$-correlation between the survival time and a linear combination of covariates. Under regularity conditions similar to those given in the Appendix, Khan & Tamer (2004) proved that the partial rank estimator is $\sqrt{n}$ consistent and asymptotically normal, using the standard U-statistic theory developed in Han (1987) and Sherman (1993). However, we note that the objective function $\tilde{O}_n(\beta)$ can be viewed as a weighted sum of indicator functions and hence discontinuous. Maximization can be extremely difficult when there are multiple covariates. In the case of $d \geq 2$, a brutal search is usually needed to obtain the estimator, which is very time-consuming. Although the Nelder-Meader method (Nelder & Meader, 1965) may be an alternative faster optimization method, it may even fail to reach the local maxima, especially when $d$ is relatively large.

To tackle this difficulty, we propose to use a continuous differentiable function to approximate the indicator function containing $\beta$ in (2). Then the objective function will be a smooth function of $\beta$. Specifically, we propose using the sigmoid function $s(u) = 1/\{1 + \exp(-u)\}$. For large $|u|$, $s(u)$ is a good approximation to $I(u > 0)$. However, for $u$ close to 0, this approximation is not accurate and thus may lead to biased estimator of $\beta$. An effective way to improve accuracy is to introduce a sequence of strictly positive and decreasing numbers $\sigma_n$ satisfying $\lim_{n\to\infty} \sigma_n = 0$, and using $s_n(u) = s(u/\sigma_n)$ to approximate $I(u > 0)$ in (2). Then the smoothed partial rank (SPR) estimator $\hat{\beta}$ is given by

$$\hat{\beta} = \arg\max_{\beta\in} \left\{ O_n(\beta) \equiv \frac{1}{n(n-1)} \sum_{i\neq j} \Delta_j I(V_i \geq V_j) s_n\left(\beta' \boldsymbol{Z}_i - \beta' \boldsymbol{Z}_j\right) \right\}. \tag{3}$$

Here $\hat{\beta} = (1, \hat{\theta}')'$. Following the same arguments as in Khan & Tamer (2004), we can show that $O_n(\beta)$ is also a U-statistic of order 2.

The objective function $O_n$ is continuously differentiable. So commonly used algorithms, for example the Newton-Raphson algorithm or the gradient search method, can be used to obtain $\hat{\beta}$. Compared with brutal search as needed for maximizing $\tilde{O}_n$, those algorithms are fast and relatively insensitive to the number of covariates.

A similar approach was proposed by Horowitz (1992) in the context of maximum score estimator for the binary response model. Horowitz considered a general class of distribution-like kernel functions for approximation of $I(u > 0)$. Let $K(u)$ be a differentiable distribution function on the real line such that it is non-decreasing and satisfies $\lim_{u\to-\infty} K(u) = 0$ and $\lim_{u\to+\infty} K(u) = 1$. Then $K_n(u) = K(u/\sigma_n)$ can be used to approximate $I(u > 0)$. Since the sigmoid function is also the logistic distribution function, $s_n$ is a special case of $K_n$. Because of the good approximation property and simplicity of $s_n$, we focus on $s_n$ in this paper.

## 4. Asymptotic properties

### *4.1. Consistency and asymptotic normality*

We now investigate the asymptotic properties of the proposed SPR estimator. We state the main results here and relegate the regularity conditions and the proofs to the Appendix.

THEOREM 1. Under assumptions A1—A6 given in the Appendix, if $\sigma_n \to 0$ as $n \to \infty$, then $\hat{\beta} \to \beta_0$ almost surely as $n \to \infty$.

Theorem 1 establishes the consistency of the proposed SPR estimator. We now investigate the asymptotic distribution of the SPR estimator. Recall that $\beta = (1, \theta')'$. To indicate that $\theta$ is the actual parameter, write $\beta(\theta) = (1, \theta')'$. Let $X = (V, \Delta, \boldsymbol{Z})$, $x = (v, \delta, \boldsymbol{z})$. First we define

$$\tau(x, \beta(\theta)) = E\left\{\Delta I(v \geq V) I\left(\beta'\boldsymbol{z} - \beta'\boldsymbol{Z} > 0\right) + \delta I(V \geq v) I\left(\beta'\boldsymbol{Z} - \beta'\boldsymbol{z} > 0\right)\right\}.$$

The asymptotic distribution of $\hat{\theta}$ is given in the following theorem.

THEOREM 2. Under assumptions A1–A8 given in the Appendix, if $\sigma_n \to 0$ as $n \to \infty$, $n^{1/2}(\hat{\theta} - \theta_0) \to N(0, \Sigma)$, where $\Sigma = A^{-1}B\left\{A^{-1}\right\}'$, $A = -E\{\nabla_2\tau(x, \beta(\theta_0))\}/2$, and $B =$

$E\{\nabla_1 \tau(x, \beta(\theta_0)) \nabla_1 \tau'(x, \beta(\theta_0))\}$.

Theorem 2 shows that the SPR estimator is asymptotically equivalent to the PR estimator. We refer to Khan & Tamer (2004) for the asymptotic properties of the PR estimator. The results hold generically when the sigmoid function $s(\cdot)$ is replaced by any symmetric distribution function with a continuous second-order derivative.

### *4.2. Inference*

We can estimate the variance-covariance matrix $\Sigma$ by the plug-in estimator

$$\Sigma_n = \hat{A}_n^{-1} \hat{B}_n \{\hat{A}_n^{-1}\}',$$

with $2\hat{A}_n = -\nabla_2 O_n(\beta(\hat{\theta}))$ and $\hat{B}_n = n^{-1} \sum_{j=1}^n \left\{\nabla_1 \hat{\tau}_n(X_j, \beta(\hat{\theta})) \nabla_1 \hat{\tau}_n'(X_j, \beta(\hat{\theta}))\right\}$, where

$$\hat{\tau}_n(x, \beta(\theta)) = n^{-1} \sum_{i=1}^n \{\Delta_i I(v \geq V_i) s_n(\beta' \boldsymbol{z} - \beta' \boldsymbol{Z}_i) + \delta I(V_i \geq v) s_n(\beta' \boldsymbol{Z}_i - \beta' \boldsymbol{z})\}.$$

However, empirical studies show that this plug-in estimator can be unstable and sensitive to the choice of the tuning parameter $\sigma_n$, especially for small sample size cases. Alternatively, we consider using the following bootstrap by analogy to that used in Jin et al. (2001) and Cai et al. (2005). Specifically, consider a stochastic perturbation of $O_n(\beta(\theta))$ which has the form of

$$O_n^w(\beta(\theta)) = \frac{1}{n(n-1)} \sum_{i \neq j} h(W_i, W_j) \Delta_j I(V_i \geq V_j) s_n(\beta' \boldsymbol{Z}_i - \beta' \boldsymbol{Z}_j),$$

where $W_i$, $i = 1, \ldots, n$, are independent realizations of a positive random variable $W$, which has a known distribution, and $h$ is a known function as follows.

THEOREM 3. Assume assumptions A1–A8 hold and $\sigma_n \to 0$ as $n \to \infty$. Let $\hat{\theta}^*$ be the maximizer of $O_n^w(\beta(\theta))$. If

C1. $W$ has mean $\mu > 0$ and variance $4\mu^2$ and $h(W_i, W_j) = W_i + W_j$; or

C2. $W$ has mean 1 and variance 1 and $h(W_i, W_j) = W_i W_j$, then conditional on the data $\{(V_i, \Delta_i, \boldsymbol{Z}_i), i = 1, \ldots, n\}$, $n^{1/2}(\hat{\theta}^* - \hat{\theta}) \to N(0, \Sigma)$.

Theorem 3 implies that the distribution of $n^{1/2}(\hat{\theta} - \theta_0)$ can be approximated by the conditional distribution $n^{1/2}(\hat{\theta}^* - \hat{\theta})$. Thus in practice, we can generate a large sample of $\{W_i, i = 1, \ldots, n\}$. For each realized sample, compute $\hat{\theta}^*$. Then the asymptotic variance of $\hat{\theta}$ can be approximated by the sample variance of $\hat{\theta}^*$. To distinguish the two weighted bootstrap methods, the former (C1) is termed type I and the latter (C2) is termed type II. The validity of the type I weighted bootstrap follows directly from the U-statistic format of the objective function $O_n$, the asymptotic normality result in Theorem 2 and Proposition A3 of Jin et al. (2001). Validity of the type II weighted bootstrap can be proved using similar arguments as those in Cai et al. (2005). It can be shown that these weighted bootstrap methods can be applied to the PR estimator as well, due to the U-statistic format of $\tilde{O}_n(\beta)$ and the asymptotic normality of the PR estimator.

## 5. SIMULATION STUDIES

Extensive simulation studies are conducted to assess the performance of the SPR estimator. First, we compare the performance of the SPR estimator and the PR estimator in the case of two covariates with one estimable regression parameter. We consider this simple setting to save computational cost for the PR estimator, which demands computationally expensive brutal search. We assume that the two covariates follow a bivariate normal distribution with mean $(1, 0{\cdot}5)$, variance 1 for each covariate and covariance $-0{\cdot}2$ between the two covariates. The survival time $T$ depends on $Z_1$ and $Z_2$ through a proportional hazards model with the regression coefficients equal to $(-1, -1)$ and the baseline hazard equal to 1. This model corresponds to the transformation model (1) with $\theta = 1$. The censoring time $C$ is generated from an exponential distribution with mean 4, leading to a censoring rate of 52%. The SPR and PR estimates are computed for 100 datasets with sample size $n = 200$. The standard errors are estimated using the sandwich method and the two weighted bootstrap methods with 100 samples of $\{W_i, i = 1, \ldots, n\}$. The

95% Wald confidence intervals are calculated correspondingly. For the two weighted bootstrap methods, $W_i$ are generated as follows: $W_i/10$ follows Beta(0·125, 1·125) for the type I weighted bootstrap, and $\sqrt{2}W_i/(\sqrt{2}-1)$ follows Beta$(\sqrt{2}-1, 1)$ for the type II weighted bootstrap. The sandwich variance estimator of the PR estimator depends on the selection of the smoothing parameters, say $\xi_1$ for $A$ and $\xi_2$ for $B$, respectively (Sherman 1993). The SPR estimator depends on the choice of the smoothing parameter $\sigma_n$. To assess the impact of the smoothing parameters on the performance of the estimators, we conduct simulations with $\sigma_n = cn^{-1/2}$, $\xi_1 = cn^{-1/4}$ and $\xi_2 = cn^{-1/6}$, where $c$ takes the values 1/9, 1 and 3. The results are shown in Table 1. Both the PR and the SPR estimators show negligible biases. The weighted bootstrap methods perform well for both estimators: the standard errors track the sampling standard deviation reasonably well with better performance for the type I method, and the coverage probabilities are close to the nominal level. In contrast, the sandwich method is sensitive to the choices of the smoothing parameters: it performs well when $c = 3$, but when $c = 1/9$, the standard errors seriously underestimate the standard deviations and the coverage probabilities are well below the nominal level; the performance for $c = 1$ lies between those for $c = 1/9$ and 3.

Next we consider the case of three covariates with two estimable parameters. The covariates are generated from a multivariate normal distribution with mean (0, 1, 0·3), variance 1 for each covariate and covariance 0.2 between any two covariates. The survival time $T$ depends on the covariates through a proportional hazards model with the regression coefficients equal to (−1, 0·5, −0·5) and baseline hazard equal to 1. This model corresponds to the transformation model (1) with $\theta = (\theta_1, \theta_2) = (-0{\cdot}5, 0{\cdot}5)$. The censoring time $C$ is generated as above, leading to a censoring rate of 36%. We fit the model using only the SPR approach because it is very difficult to implement the PR method in this case. Table 2 presents the results from 100 simulated datasets with $n = 200$ and the smoothing parameter $\sigma = cn^{-1/2}$, $c = 1/9, 1, 3$. The performance of the SPR estimator is similar to that observed for one estimable parameter except that the type II bootstrap method performs worse for $c = 1/9$ and 1 with the coverage probabilities for $\theta_2$ reaching 1.

We have also conducted simulations under other survival models such as the accelerated failure time model and observed similar results. The type I bootstrap method outperforms the type II method in all the cases. We note that as sample size increases, the performance of the type II method improves and is able to provide satisfactory inference results with moderate to large sample size cases. We also note that the bootstrap distribution may be skewed if the sample size is not large and there may exist some "outliers" due to the failure of reaching the global maxima for some bootstrap samples. In contrast, the normalized median absolute deviation of the estimates obtained from the bootstrap datasets are relatively stable and close to the empirical standard deviation (see Tables 1 and 2). We suggest using the normalized median absolute deviation from the type I bootstrap method to estimate the standard error.

The accuracy of approximation depends on the tuning parameter $\sigma_n$. Theoretically speaking, the smaller the $\sigma_n$ is, the better the sigmoid approximation is. Extensive simulation studies show that as long as $\sigma_n$ is small enough, the SPR estimate is insensitive to the choice of $\sigma_n$. However, numerical studies also show that for extremely small $\sigma_n$, the maximization procedure may be unstable. In data analyses, a rule of thumb for choosing $\sigma_n$ is to guarantee a majority of $|\beta'(\boldsymbol{Z}_i - \boldsymbol{Z}_j)/\sigma_n| > 5$ (Gammerman, 1996). We propose the following approach for choosing $\sigma_n$. Initialize $\sigma_n^0 = a_n$, where $a_n$ is user-specified and data-independent, satisfying $a_n \to 0$ as $n \to \infty$. In our data analysis, we use $a_n = 1/\sqrt{n}$. Construct the SPR estimate $\hat{\beta}$ with $\sigma_n^0$, under the identifiability constraint. Theoretically speaking, the estimator with $\sigma_n^0 = a_n$ is consistent. Denote $\sigma_n^1$ as the largest constant such that 95% of the $|\hat{\beta}'(\boldsymbol{Z}_i - \boldsymbol{Z}_j)/\sigma_n|$ is greater than 5. Set $\sigma_n = min(\sigma_n^0, \sigma_n^1)$. With the proposed procedure, the asymptotic requirement $\sigma_n = o(1)$ is met; meanwhile the rule of thumb for choosing $\sigma_n$ is also satisfied. Extensive simulation studies show that the estimation and inference results are relatively not sensitive to the choice of $a_n$, as long as it is small enough.

## 6. Real data example

As illustration, we apply the proposed approach to the Veterans Administration lung cancer data described in Kalbfleisch & Prentice (2002, pp.71–2), which includes data from a clinical trial of 137 patients with advanced inoperable lung cancer. The patients were randomized to either a standard or test chemotherapy. The survival time was time to death. There were 128 events. We consider five covariates: $Z_1$ = age/100, $Z_2$ = diagtime/100, $Z_3 = I$(treatment = test chemotherapy), $Z_4$ = karno/10, and $Z_5$ = prior/10. The nonparametric transformation model is fitted using the SPR approach with the regression parameter of $Z_1$ fixed at 1. The standard errors are computed using the type I resampling method as suggested at the end of Section 5. Since the sample size is small, the standard errors are large. For the purpose of comparison, we also fit the Cox proportional hazards model and the accelerated failure time model with the same covariates. If the Cox model fit the data well, we would expect the ratios of the corresponding coefficients from the two models to be close to a constant. However, the results given in table 3 indicate that the proportional hazards assumption may not hold. This is further confirmed by the testing method in Therneau & Grambsch (Ch. 6.2, 2000) with the p-value being 0.0008. The results from the accelerated failure time model are also given for comparison.

Note that for the Cox model, the accelerated failure time model and the nonparametric transformation model, the estimated linear combination of the covariates $\beta' \boldsymbol{Z}$ may be used as a predictor for survival. The predictive capacity may be assessed using the areas under the survival ROC curve (AUC) proposed by Heagerty et al. (2000). Figure 1 shows the AUCs for the estimated linear combinations for these three models over a range of times. The AUCs from the nonparametric transformation model are greater than those from the Cox model and the accelerated failure time model except at a few times before 30 days. Thus the nonparametric transformation model appears to provide a better fit to this dataset.

## 7. DISCUSSION

We have proposed a smoothed partial rank estimator for the nonparametric transformation model for survival data with no parametric assumptions on both the transformation function and the error distribution. The proposed estimator is asymptotically equivalent to the partial rank estimator, but is much easier to compute in the case of multiple covariates. The idea of smoothing the objective function can be extended to estimate the transformation function and will be investigated in our future research.

Using a smooth function as an approximation of the indicator function has been extensively used in machine learning and neural network studies (Gammerman, 1996). Gammerman (1996) noted that such an approximation may lead to multiple local maxima in neural network studies. The global maximum may be detected by varying starting values. Although empirical studies show this does not pose a serious problem for our study, future investigation will be pursued.

A more general model is $T = g\{h(\beta' Z, e)\}$, where $h$ is a strictly increasing function of each of its components. This model included the additive hazards model as a special case, which does not belong to the framework of model (1). The estimation and inference procedure in this paper can be extended to this general model.

## ACKNOWLEDGMENTS

This research is supported in part by grants from the National Institute of Health and the National Heart, Lung and Blood Institute, Department of Veterans Affairs Health Administration HS&D, and career development funds from the University of Washington Department of Biostatistics.

## Appendix

### *Technical Details*

We note that the only special properties of the sigmoid function we use in the proof of Theorems 1 and 2 are its symmetry, $s(u) + s(-u) = 1$, and that it is smooth and has a continuous second derivative. Therefore, the proofs below are valid when we use any scaled symmetric distribution function with a continuous second derivative as an approximation to the indicator function $I(u > 0)$.

*Regularity Conditions.* Let $R$ be the support operator, for example, $R(\boldsymbol{Z})$ denotes the support of $\boldsymbol{Z}$. Denote $\beta_0 = (1, \theta_0')'$ as the true value of $\beta$. Let $\chi$ denote the last $d-1$ components of the covariate vector $Z$, let $g^0(w|r)$ denote the conditional density of $W = \beta_0' Z$ given $\chi = r$, and let $p^0(\delta, y, w|r)$ denote the conditional density of $(\Delta, Y, W)$ given $\chi = r$. We assume the following conditions.

A1. The set $\{\boldsymbol{Z} \in R(\boldsymbol{Z}) : Pr(\Delta = 1|\boldsymbol{Z}) > 0\}$ has a positive measure.

A2. The random error $\epsilon$ is independent of $C$ and $\boldsymbol{Z}$.

A3. The first component of $\boldsymbol{Z}$ has everywhere positive Lebesgue density, conditional on other components.

A4. The parameter space $\mathbb{B}$ containing $\beta_0$ is a compact subset of $\mathbb{R}^d$.

A5. $R(\boldsymbol{Z})$ is not contained in any proper linear subspace of $\mathbb{R}^d$.

A6. $T$ and $C$ are conditionally independent given $\boldsymbol{Z}$.

A7. (a) For each $x$, the function $\tau(x, \beta(\theta))$ is twice differentiable with respect to $\theta$ in a neighborhood of $\theta_0$ with the $k^{th}$ derivative $\nabla_k \tau(x, \beta(\theta))$, $k = 1, 2$. The second derivative $\nabla_2 \tau(x, \beta(\theta))$ satisfies the Lipschitz condition. (b) The partial derivatives of $g^0(w|r)$ and $p^0(\delta, y, w|r)$ with respect to $t$ exist and are bounded.

A8. $E\| \nabla_1 \tau(x, \beta(\theta_0))\|^2$ and $E\| \nabla_2 \tau(x, \beta(\theta_0))\|$ are finite, and $E\{\nabla_2 \tau(x, \beta(\theta_0))\}$ is nonsingular.

Assumptions A1-A6 are relatively mild and close to their counterparts in Khan & Tamer (2004). Particularly, assumptions A1, A2 and A6 are usually made for semiparametric models with censored survival data; assumptions A3–A5 are needed for identifiability. Assumptions A7 and A8 are made to guarantee the $\sqrt{n}$ consistency and asymptotic normality of the partial rank estimator in Khan & Tamer (2004).

*Proof of Theorem 1.* Since it has been proved in Khan & Tamer (2004) that the partial rank estimator is consistent, it suffices to prove that $\sup_{\beta \in \mathbf{B}} |O_n(\beta) - \tilde{O}_n(\beta)| \rightarrow_p 0$.

For any $\eta > 0$, we have

$$|O_n(\beta) - \tilde{O}_n(\beta)| = \frac{1}{n(n-1)} \left| \sum_{i \neq j} \Delta_j I(V_i \geq V_j) \{ I(\beta'(\boldsymbol{Z}_i - \boldsymbol{Z}_j) > 0) - s_n(\beta'(\boldsymbol{Z}_i - \boldsymbol{Z}_j)) \} \right| \\ \leq T_{n1} + T_{n2},$$

where

$$T_{n1} = \frac{1}{n(n-1)} \sum_{i \neq j} \Delta_j I(V_i \geq V_j) \, |I(\beta'(\boldsymbol{Z}_i - \boldsymbol{Z}_j) > 0) - s_n(\beta'(\boldsymbol{Z}_i - \boldsymbol{Z}_j))| \\ I(\beta'(\boldsymbol{Z}_i - \boldsymbol{Z}_j) \geq \eta),$$

$$T_{n2} = \frac{1}{n(n-1)} \sum_{i \neq j} \Delta_j I(V_i \geq V_j) \, |I(\beta'(\boldsymbol{Z}_i - \boldsymbol{Z}_j) > 0) - s_n(\beta'(\boldsymbol{Z}_i - \boldsymbol{Z}_j))| \\ I(\beta'(\boldsymbol{Z}_i - \boldsymbol{Z}_j) < \eta).$$

On the set $\{|u| > \eta\}$, we have $|s_n(u) - I(u > 0)| \leq \exp(-|u|/\sigma_n) < \exp(-\eta/\sigma_n)$. Thus when $\sigma_n \rightarrow 0$, $s_n(u) \rightarrow I(u > 0)$ uniformly on the set $\{|u| > \eta\}$. Therefore, $T_{n1}$ converges to 0 uniformly over $\Theta$. The second term $T_{2n} \leq \{n(n-1)\}^{-1} \sum_{i \neq j} I(|\beta'(\boldsymbol{Z}_i - \boldsymbol{Z}_j)| < \eta)$. Since the class of indicator functions $\{I(|\beta'(\boldsymbol{Z}_i - \boldsymbol{Z}_j)| < \eta) : \beta \in \mathbb{B}\}$ is manageable, by uniform convergence of U-processes (Theorem 7, Nolan & Pollard 1987), the right-hand side converges almost surely to $P(|\beta'(\boldsymbol{Z}_i - \boldsymbol{Z}_j)| < \eta)$ over $\mathbb{B}$. However, under assumption (A2), it can be proved in a similar way as in Lemma 4 of Horowitz (1992), $P(|\beta'(\boldsymbol{Z}_i - \boldsymbol{Z}_j)| < \eta)$ converges to 0 uniformly over $\mathbb{B}$ as $\eta \rightarrow 0$. This completes the proof of consistency.

*Proof of Theorem 2.* Theorem 2 can be proved following the method of Sherman (1993). Using the same notations as those in Section 4.1, we first define

$$f(x_1, x_2, \beta(\theta)) = f^*(x_1, x_2, \beta(\theta)) - f^*(x_1, x_2, \beta(\theta_0)),$$

$$f_n(x_1, x_2, \beta(\theta)) = f_n^*(x_1, x_2, \beta(\theta)) - f_n^*(x_1, x_2, \beta(\theta_0)),$$

where

$$f^*(x_1, x_2, \beta(\theta)) = \delta_2 I(v_1 \geq v_2) I\left(\beta' \boldsymbol{z}_1 - \beta' \boldsymbol{z}_2 > 0\right) + \delta_1 I(v_2 \geq v_1) I\left(\beta' \boldsymbol{z}_2 - \beta' \boldsymbol{z}_1 > 0\right),$$

$$f_n^*(x_1, x_2, \beta(\theta)) = \delta_2 I(v_1 \geq v_2) s_n\left(\beta' \boldsymbol{z}_1 - \beta' \boldsymbol{z}_2\right) + \delta_1 I(v_2 \geq v_1) s_n\left(\beta' \boldsymbol{z}_2 - \beta' \boldsymbol{z}_1\right).$$

Note that $f_n$ is the smooth approximation of the function $f$. In Section 4.1, we defined $\tau(x, \beta(\theta)) = E\{f^*(x, X, \beta(\theta))\}$. We now consider its smoothed counterpart $E\{\tau_n(x, \beta(\theta))\}$ with $\tau_n(x, \beta(\theta)) = E\{f_n^*(x, X, \beta(\theta))\}$ and the expectation is over $X$. Denote $P_n$ as the empirical measure and $U_n$ as the U-process operator as in Sherman (1993). Consider the function $\Gamma_n(\theta) = O_n(\theta) - O_n(\theta_0)$.

First, we have

$$\nabla_1 \tau_n(x, \beta(\theta)) = \nabla_1 \tau(x, \beta(\theta)) + O(\sigma_n) \quad \text{and} \quad \nabla_2\, \tau_n(x, \beta(\theta)) = \nabla_2 \tau(x, \beta(\theta)) + O(\sigma_n). \qquad \text{(A.1)}$$

These equations can be proved based on condition (A7) and by noting that $s_n(u) + s_n(-u) = 1$, rewriting $\tau_n$ as an integral, changing variables in the integral, and using the Taylor expansion.

We can write $\Gamma_n(\theta) = \Gamma_{n0}(\theta) + P_n g_n(\cdot, \theta) + U_n h_n(\cdot, \cdot, \theta)$, where

$$\Gamma_{n0}(\theta) = E f_n(\cdot, \cdot, \theta),$$

$$g_n(x, \theta) = E f_n(x, \cdot, \theta) + E f_n(\cdot, x, \theta) - 2\Gamma_{n0}(\theta) = \tau_n(x, \theta) - \tau_n(x, \theta_0) - 2\Gamma_{n0}(\theta), \text{ and}$$

$$h_n(x_1, x_2, \theta) = f_n(x_1, x_2, \theta) - E f_n(x_1, \cdot, \theta) - E f_n(\cdot, x_2, \theta) + \Gamma_{n0}(\theta).$$

Next we show that

$$\Gamma_{n0}(\theta) = \frac{1}{2}(\theta - \theta_0)' A (\theta - \theta_0) + O(\sigma_n |\theta - \theta_0|) + o(|\theta - \theta_0|^2), \qquad \text{(A.2)}$$

$$P_n g_n(\cdot,\theta) = n^{-1/2}(\theta-\theta_0)' G_n + O(\sigma_n|\theta-\theta_0|) + o(|\theta-\theta_0|^2), \tag{A.3}$$

where $G_n = \sqrt{n} P_n \bigtriangledown_1 \tau(\cdot,\theta_0) \to_d N(0,B)$, and

$$U_n h_n(\cdot,\cdot,\theta) = o_p(n^{-1}) \tag{A.4}$$

uniformly in an $o_p(1)$ neighborhood of $\theta_0$.

To prove (A.2), we first note that

$$\Gamma_{n0}(\theta) = \frac{1}{2}(\theta-\theta_0)' A_n(\theta-\theta_0) + O(\sigma_n|\theta-\theta_0|) + o(|\theta-\theta_0|^2), \tag{A.5}$$

where $A_n = -\frac{1}{2}E\{\bigtriangledown_2 \tau_n(x,\beta(\theta))\}$. This can be proved by following Theorem 4 of Sherman (1993). We then conclude (A.2) by noting (A.1) and the assumption that $\sigma_n \to 0$.

Equation (A.3) can also be proved following Theorem 4 of Sherman (1993) and (A.1).

To prove (A.4), consider the function

$$\begin{aligned} & h(x_1, x_2, \beta(\theta), \sigma) \\ & \quad = f(x_1,x_2,\beta(\theta),\sigma) - Ef(x_1,\cdot,\beta(\theta),\sigma) - Ef_n(\cdot,x_2,\beta(\theta),\sigma) + Ef_n(\cdot,\cdot,\beta(\theta),\sigma)\cdot \end{aligned}$$

Since $\sigma_n \to 0$, it suffices to show that $U_n h(\cdot,\cdot,\beta(\theta),\sigma) = o_p(n^{-1})$ uniformly over an $o_p(1)$ neighborhood of $(\beta_0, 0)$. First, by assumption A3 and using the dominated convergence theorem,

$$Eh_n^2(x,X,\theta) \to 0 \text{ as } (\theta,\sigma_n) \to (\theta_0,0)\cdot$$

Consider $\mathcal{S} = \{h(x,X,\beta(\theta),\sigma_n) : \beta \in \mathbb{B}, \sigma_n \in (0,1]\}$. Note that $\mathcal{S}$ is Euclidean by Lemma 22 (ii) of Nolan & Pollard (1987) and it is bounded by 4. Therefore, (A.4) follows from Theorem 3 of Sherman (1993) or Corollary 8 of Sherman (1993).

Finally, based on (A.2), (A.3), (A.4), and the assumption that $\sigma_n \to 0$, by Theorem 1 of Sherman (1993), $\hat\theta - \theta_0 = O_p(n^{-1/2})$. The asymptotic normality now follows from Theorem 2 of Sherman (1993). See also Theorem 3.2.16 of Van der Vaart & Wellner (Ch. 3.2, 1996).

References


Aalen, O.O. (1980). A model for regression analysis of counting processes. *Lecture Notes in Statistics, 2.* New York: Springer-Verlag.

Cai, T., Tian, L. & Wei, L.J. (2005). Semi-parametric Box-Cox power transformation models for censored survival observations. *Biometrika* **92**, 619–32.

Cavanagh, C. & Sherman, R.P. (1998). Rank estimators for monotonic index models. *J. Economet.* **84**, 351–81.

Chen, K., Jin, Z. & Ying Z. (2002). Semiparametric analysis of transformation models with censored data. *Biometrika* **89**, 659–68.

Cheng, S.C., Wei, L.J. & Ying Z. (1995). Analysis of transformation models with censored data. *Biometrika* **82**, 835–45.

Cox, D.R. (1972). Regression models and life tables (with discussion). *J. R. Statist. Soc.***B** **34**, 187–200.

Fine, J.P., Ying, Z. & L.J. Wei (1998). On the linear transformation model with censored data. *Biometrika* **85**, 980–86.

Gammerman, A. (1996) *Computational Learning and Probabilistic Reasoning.* New York, Wiley.

Han, A.K. (1987) Non-parametric Analysis of a Generalized Regression Model. *J. Economet.* **35**, 303–16.

Heagerty, P.J., Lumley, T. & Pepe, M.S. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker, *Biometrics* **56**, 337–44.

Horowitz, J. L. (1992) A Smoothed Maximum Score Estimator for the Binary Response Model. *Econometrica* **60**, 505–31.

Jin Z., Ying Z. & Wei L.J. (2001) A simple resampling method by perturbing the minimand. *Biometrika* **88**, 381–90.

Kalbfleisch, J.D. & Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data.* New York, Wiley.

Khan, S. & Tamer, E. (2004). Partial rank estimation of duration models with general forms of censoring. Submitted.

Nelder, J. A. & Mead, R. (1965). A simplex algorithm for function minimization. *Comp. J.* **7**, 308–13.

Nolan, D. & Pollard, D. (1987) U-processes: rates of convergence. *Ann. Statist.* **15**, 780–99.

Sherman, R. (1993). The limiting distribution of the maximum rank correlation estimator. *Econometrica* **61**: 123–37.

Therneau, T. & Grambsch, P. (2000) *Modelling Survival Data.* Springer, New York.

Van der Vaart, A. W. & Wellner, J. A. (1996) *Weak Convergence and Empirical Processes: With Applications to Statistics.* Springer, New York.

Table 1. *Simulation results in the case of one estimable regression parameter.*

| | | | | Sandwich | | Resampling I | | | Resampling II | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $c$ | | B | SD | SE | CP | MAD | SE | CP | MAD | SE | CP |
| 1/9 | PR | 0·031 | 0·186 | 0·036 | 0·23 | 0·189 | 0·213 | 0·96 | 0·203 | 0·229 | 0·96 |
| | SPR | 0·031 | 0·184 | 0·013 | 0·09 | 0·176 | 0·208 | 0·96 | 0·198 | 0·226 | 0·96 |
| 1 | PR | 0·031 | 0·186 | 0·155 | 0·89 | 0·189 | 0·213 | 0·97 | 0·203 | 0·226 | 0·96 |
| | SPR | 0·030 | 0·170 | 0·172 | 0·87 | 0·176 | 0·202 | 0·96 | 0·218 | 0·218 | 0·98 |
| 3 | PR | 0·031 | 0·186 | 0·182 | 0·92 | 0·189 | 0·214 | 0·97 | 0·203 | 0·231 | 0·96 |
| | SPR | 0·077 | 0·177 | 0·191 | 0·94 | 0·187 | 0·199 | 0·97 | 0·196 | 0·212 | 0·95 |

Resampling I, type I resampling method with $h(W_i, W_j) = W_i + W_j$; Resampling II, type II resampling method with $h(W_i, W_j) = W_i W_j$; B, bias; SD, empirical standard deviation across simulated data sets; SE, average of estimated standard errors; CP, coverage probability of the 95% Wald confidence interval; MAD, normalized median absolute deviation.

Table 2. *Simulation results in the case of two estimable regression parameters.*

| | | | | Sandwich | | Resampling I | | | Resampling II | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $c$ | | B | SD | SE | CP | MAD | SE | CP | MAD | SE | CP |
| 1/9 | $\theta_1$ | −0·015 | 0·132 | 0·009 | 0·11 | 0·125 | 0·139 | 0·96 | 0·135 | 0·148 | 0·98 |
| | $\theta_2$ | 0·008 | 0·131 | 0·011 | 0·07 | 0·147 | 0·154 | 0·98 | 0·151 | 0·167 | 1·00 |
| 1 | $\theta_1$ | −0·012 | 0·123 | 0·119 | 0·80 | 0·120 | 0·130 | 0·95 | 0·129 | 0·139 | 0·97 |
| | $\theta_2$ | 0·013 | 0·122 | 0·129 | 0·88 | 0·138 | 0·145 | 0·98 | 0·141 | 0·157 | 1·00 |
| 3 | $\theta_1$ | −0·035 | 0·122 | 0·125 | 0·96 | 0·122 | 0·129 | 0·96 | 0·128 | 0·135 | 0·96 |
| | $\theta_2$ | 0·054 | 0·127 | 0·140 | 0·95 | 0·139 | 0·145 | 0·95 | 0·144 | 0·152 | 0·97 |

Resampling I, type I resampling method with $h(W_i, W_j) = W_i + W_j$; Resampling II, type II resampling method with $h(W_i, W_j) = W_i W_j$; B, bias; SD, empirical standard deviation across simulated data sets; SE, average of estimated standard errors; CP, coverage probability of the 95% Wald confidence interval; MAD, normalized median absolute deviation.

Table 3. *Results for the Veterans Administration data.*

| | age/100 | | diagtime/100 | | treatment | | karno/10 | | prior/10 | |
|---|---|---|---|---|---|---|---|---|---|---|
| Model | Est | SE | Est | SE | Est | SE | Est | SE | Est | SE |
| Cox | −0·388 | 0·925 | 0·172 | 0·900 | 0·193 | 0·186 | −0·341 | 0·053 | −0·078 | 0·222 |
| AFT | 0·479 | 1·104 | 0·142 | 1·018 | −0·133 | 0·203 | 0·364 | 0·054 | 0·065 | 0·248 |
| NT | 1 | — | 3·627 | 13·549 | −1·874 | 4·930 | 8·428 | 6·927 | 6·343 | 7·490 |

AFT, accelerated failure time model; NT, nonparametric transformation model. protime, prothrombin time.

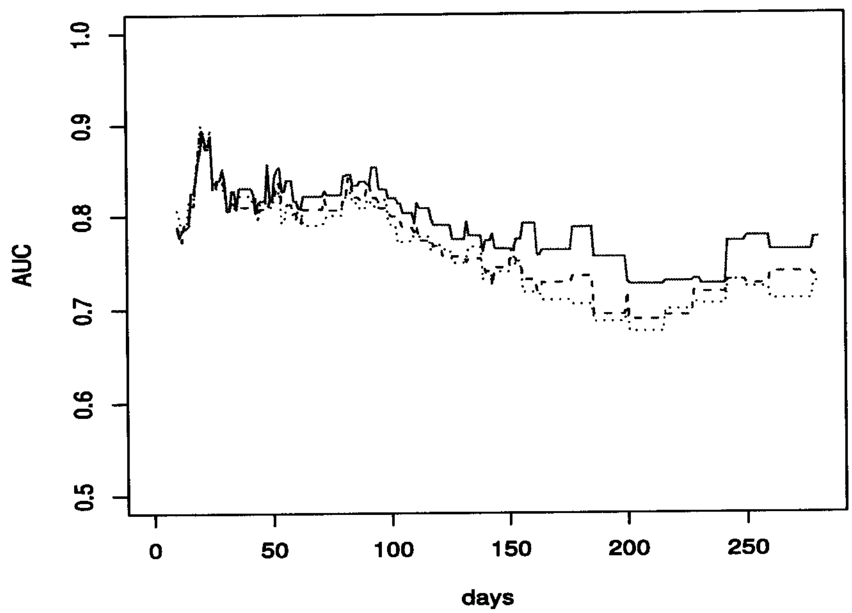


Fig. 1. Estimated AUC for the Veterans Administration data. Solid line, nonparametric transformation model; dotted line, Cox model; dashed line, accelerated failure time model.