A semiparametric approach for the nonparametric transformation survival model with multiple covariates

XIAO SONG*, SHUANGGE MA
Department of Biostatistics, University of Washington, Box 357232, 1705 Northeast Pacific Street, Seattle, WA 98195, USA
songx@u.washington.edu

JIAN HUANG
Department of Statistics and Actuarial Science and Program in Public Health Genetics, University of Iowa, 241 Schaeffer Hall, Iowa City, IA 52242, USA

XIAO-HUA ZHOU
Department of Biostatistics, University of Washington, Box 357232, 1705 Northeast Pacific Street, Seattle, WA 98195, USA and Puget Sound Health Care System, 1660 South Columbian Way, Seattle, WA 98018, USA

SUMMARY
The nonparametric transformation model makes no parametric assumptions on the forms of the transformation function and the error distribution. This model is appealing in its flexibility for modeling censored survival data. Current approaches for estimation of the regression parameters involve maximizing discontinuous objective functions, which are numerically infeasible to implement with multiple covariates. Based on the partial rank (PR) estimator (Khan and Tamer, 2004), we propose a smoothed PR estimator which maximizes a smooth approximation of the PR objective function. The estimator is shown to be asymptotically equivalent to the PR estimator but is much easier to compute when there are multiple covariates. We further propose using the weighted bootstrap, which is more stable than the usual sandwich technique with smoothing parameters, for estimating the standard error. The estimator is evaluated via simulation studies and illustrated with the Veterans Administration lung cancer data set.

Keywords: Nonparametric transformation model; Partial rank estimator; Survival analysis; Weighted bootstrap.

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 and Prentice, 2002). Although these semiparametric models are more flexible than parametric models, they may still be
restrictive in practice. Our study is partly motivated by examples like the Veterans Administration lung cancer data set described in Kalbfleisch and Prentice (2002, pp. 71–2), which includes measurements from a clinical trial of 137 patients with advanced inoperable lung cancer. The patients were randomized to either a standard or a test chemotherapy. The primary end point for therapy comparison was time to death. There were 128 events. Possible heterogeneity may exist for some baseline covariates, such as disease extent and pathology, previous treatment of the disease, demographic background, and initial health status. It is a common practice to adjust for such covariates in assessing the treatment effect on time to death. Therneau and Grambsch (2000, Chapter 6.3) assessed the proportional hazards assumption for this data set and indicated that the proportionality might not hold. A more flexible model is needed to account for such non-proportionality.

One feasible alternative is the transformation model, which postulates that an unknown monotone increasing transformation of the survival time depends on the covariates through a linear model. Cheng and others (1995), Fine and others (1998), and Chen and others (2002) studied this model with a known error distribution and an unspecified transformation function, and Cai and others (2005) 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 and is especially preferred when, for example, the Cox model cannot be properly justified.

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 and Sherman, 1998). Both estimators can be extended to the case of censored survival outcomes using the inverse censoring probability weighting technique (Khan and Tamer, 2004). However, this requires that the censoring time is independent of the survival time and 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 and 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 and monotone rank estimators, the PR estimator is based on maximization of a discontinuous objective function. It is practically impossible to compute this estimator when there are multiple covariates.

In this paper, based on the PR estimator of Khan and 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 SPR 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 and others (2001).

Similar approximation with a rank estimator has been investigated in Ma and Huang (2005) in a binary classification study, which focuses on variable selection with microarray data. Compared to Ma and Huang (2005), our study makes the following new contributions. First, although the approximation idea has been proposed in other areas, it has not been fully explored in survival analysis—theoretical proof and finite sample performance are not known. Second, the asymptotic properties of such approximated estimators have not been established in previous studies. Moreover, the inference based on the weighted bootstrap has not been investigated.

The rest of this 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 are assessed by simulation studies in Section 5. We apply the approach to the Veterans Administration lung cancer data in Section 6. The paper concludes with a discussion in Section 7.

2. MODEL DEFINITION

Let $T$ denote the survival time, $C$ denote the censoring time, and $Z$ 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,

$$g(T) = \beta' Z + e,$$

(2.1)

where $g(\cdot)$ is an unspecified monotone increasing function, $e$ is the random error term with an unknown distribution independent of $Z$, $\beta$ is a length $d$ vector of regression coefficients, 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 (2.1) with $e$ following the standard extreme value and logistic distributions, respectively; the accelerated failure time model is a special case of (2.1) 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, $\beta = (1, \theta')'$. This indicates a positive correlation between the outcome and the first covariate (the covariate can be recoded if necessary to achieve this). Our interest focuses on estimation and inference of $\theta$.

3. ESTIMATION

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

$$\hat{\beta} = \arg \max_{\beta} \left\{ \hat{O}_n(\beta) = \frac{1}{n(n-1)} \sum_{i \neq j} \Delta_j I(V_i \geq V_j) I(\beta' Z_i > \beta' Z_j) \right\},$$

(3.1)

where $I(\cdot)$ is the indicator function and $\hat{\beta} = (1, \hat{\theta}')'$. The objective function $\hat{O}_n(\beta)$ can be considered as a generalization of Kendall’s $\tau$-correlation statistic to censored data, which is a $U$-statistic of order 2. Loosely speaking $\hat{\beta}$ seeks to maximize the Kendall’s $\tau$-correlation between the survival time and a linear combination of covariates, with a proper consideration of the censoring. Under some regularity conditions similar to those given in the Appendix, Khan and Tamer (2004) proved that the PR 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 $\hat{O}_n(\beta)$ can be viewed as a weighted sum of indicator functions and hence discontinuous. Maximization is extremely difficult in practice, if not impossible, when there are multiple covariates. When $d \geq 2$, a brutal search is usually needed to obtain the estimator, which is very time consuming. Although the Nelder–Mead method (Nelder and Mead, 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 (3.1). 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 enough and thus may lead to unsatisfactory estimator of \( \beta \). An effective way motivated by the sieve estimate 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 use the family of functions \( s_n(u) = s(u/\sigma_n) \) to approximate \( I(u > 0) \) in (3.1). The SPR estimator \( \hat{\beta} \) is given by

\[
\hat{\beta} = \arg \max_{\beta} \left\{ O_n(\beta) \equiv \frac{1}{n(n-1)} \sum_{i \neq j} \Delta_j I(V_i \geq V_j) s_n(\beta' Z_i - \beta' Z_j) \right\}.
\]

Here \( \hat{\beta} = (1, \hat{\theta}')' \). Following the same arguments as in Khan and 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 to brutal search as needed for maximizing \( O_n \), 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 nondecreasing 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. Most theoretical and computational results hold for general \( K_n \) with minor modifications.

### 4. ASYMPIOTIC 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 4.1** 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 4.1 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 \( X = (V, \Delta, Z) \), \( x = (v, \delta, z) \). First, we define

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

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

**THEOREM 4.2** 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 \( \Sigma = A^{-1} B \{ A^{-1} \}' \), \( A = -E\{ \nabla_2 \tau(x, \beta(\theta_0)) \}/2 \), and \( B = E\{ \nabla_1 \tau(x, \beta(\theta_0)) \} \) \( \nabla_1 \tau'(x, \beta(\theta_0)) \).

Theorem 4.2 shows that the SPR estimator is asymptotically equivalent to the PR estimator. The proposed smoothing approach leads to a computationally affordable estimate with no loss of efficiency.
We refer to Khan and 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 $\Sigma$ with 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 $\hat{B}_n = n^{-1} \sum_{i=1}^n \{ \nabla_1 \hat{\tau}_n(X_j, \beta(\hat{\theta})) \nabla_1 \hat{\tau}_n'(X_j, \beta(\hat{\theta})) \},$ where

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

However, this plug-in estimator can be unstable and sensitive to the choice of the tuning parameter $\sigma_n$, especially for small sample size cases. Simulation studies show that the empirical coverage probability of the 95% confidence interval built with the plug-in variance estimate can be even smaller than 50% for sample size as large as 400. We refer to Section 5 for numerical studies. Alternatively, we consider using the weighted bootstrap by analogy to that used in Jin and others (2001) and Cai and others (2005). Specifically, consider a stochastic perturbation of $O_n(\beta(\theta))$ which has the form of

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

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 specified in the following theorem.

**Theorem 4.3** 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

1. $W$ has known mean $\mu > 0$ and variance $4\mu^2$ and $h(W_i, W_j) = W_i + W_j$; or
2. $W$ has mean 1 and variance 1 and $h(W_i, W_j) = W_i W_j$,

then conditional on the data \{$(V_i, \Delta_i, Z_i), i = 1, \ldots, n$\}, $n^{1/2}(\hat{\theta}^* - \hat{\theta}) \to N(0, \Sigma)$.

Theorem 4.3 implies that the unconditional distribution of $n^{1/2}(\hat{\theta} - \theta_0)$ can be approximated by the conditional distribution of $n^{1/2}(\hat{\theta}^* - \hat{\theta})$. 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 and others (2001). Validity of the type II weighted bootstrap can be proved using similar arguments as those in Cai and others (2005).

Inference based on the PR estimator is not considered in detail in Khan and Tamer (2004). It can be shown that the weighted bootstrap methods in Theorem 4.3 can be applied to the PR estimator as well, due to the $U$-statistic format of objective function $\tilde{O}_n(\beta)$ and the asymptotic normality of the PR estimator. See Section 5 for numerical results.
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 the dimension of \( \theta \) equal to 1. We consider this simple setting to save computational cost for the PR estimator, which demands a computationally expensive brutal search. We assume that the two covariates follow a bivariate normal distribution with mean \((1, 0.5)\), variance 1 for each covariate, and correlation \(-0.2\) 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 (2.1) 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 data sets 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 \((\sqrt{2} - 1)W_i / \sqrt{2} \) 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 \( \xi_1 \) and \( \xi_2 \) for \( A \) and \( B \), 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 \( \sigma_n = cn^{-1/2}, \xi_1 = cn^{-1/4}, \text{ and } \xi_2 = cn^{-1/6} \), where \( c \) takes the values \( 1/9, 1, \text{ and } 3 \). 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 the dimension of \( \theta \) equal to 2. The covariates are generated from a multivariate normal distribution with mean \((0, 1, 0.3)\), variance 1 for each covariate, and correlation 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 (2.1) with \( \theta = (\theta_1, \theta_2) = (-0.5, 0.5) \). 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

<table>
<thead>
<tr>
<th>( c )</th>
<th>( B )</th>
<th>SD</th>
<th>Sandwich SE</th>
<th>Sandwich CP</th>
<th>Resampling I MAD</th>
<th>Resampling I SE</th>
<th>Resampling I CP</th>
<th>Resampling II MAD</th>
<th>Resampling II SE</th>
<th>Resampling II CP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/9</td>
<td>PR</td>
<td>0.031</td>
<td>0.186</td>
<td>0.036</td>
<td>0.23</td>
<td>0.189</td>
<td>0.213</td>
<td>0.96</td>
<td>0.203</td>
<td>0.229</td>
</tr>
<tr>
<td></td>
<td>SPR</td>
<td>0.031</td>
<td>0.184</td>
<td>0.013</td>
<td>0.09</td>
<td>0.176</td>
<td>0.208</td>
<td>0.96</td>
<td>0.198</td>
<td>0.226</td>
</tr>
<tr>
<td>1</td>
<td>PR</td>
<td>0.031</td>
<td>0.186</td>
<td>0.155</td>
<td>0.89</td>
<td>0.189</td>
<td>0.213</td>
<td>0.97</td>
<td>0.203</td>
<td>0.226</td>
</tr>
<tr>
<td></td>
<td>SPR</td>
<td>0.030</td>
<td>0.170</td>
<td>0.172</td>
<td>0.87</td>
<td>0.176</td>
<td>0.202</td>
<td>0.96</td>
<td>0.218</td>
<td>0.218</td>
</tr>
<tr>
<td>3</td>
<td>PR</td>
<td>0.031</td>
<td>0.186</td>
<td>0.182</td>
<td>0.92</td>
<td>0.189</td>
<td>0.214</td>
<td>0.97</td>
<td>0.203</td>
<td>0.231</td>
</tr>
<tr>
<td></td>
<td>SPR</td>
<td>0.077</td>
<td>0.177</td>
<td>0.191</td>
<td>0.94</td>
<td>0.187</td>
<td>0.199</td>
<td>0.97</td>
<td>0.196</td>
<td>0.212</td>
</tr>
</tbody>
</table>

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.
the results from 100 simulated data sets 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 data sets is 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 performance of the SPR approach depends on the shape of the objective function \( O_n(\beta) \), if the gradient search or the Newton methods are used. One concern is that there may exist some local maximizers which impede the optimization procedure. A simple solution is to search over a wide range of starting values (Gammerman, 1996). Theoretically speaking, the empirical objective function of the SPR estimate converges to its expectation when \( n \to \infty \), which converges to the expectation of the PR objective function as \( \sigma_n \to 0 \). The expectation of the PR objective function has a unique maximizer and has a negative definite Hessian matrix in a neighborhood of the true parameter value (Khan and Tamer, 2004). Thus, as long as the sample size is large enough, the empirical objective function of the SPR estimate has a unique maximizer with probability converging to 1 in a neighborhood of the true parameter value. In our empirical studies, local maximizers do not cause a serious problem. When there are two estimable parameters, the objective function generally has a shape similar to that shown in Figure 1, which is obtained from a simulated data set with sample size 200 and has a unique maximizer and the maximizer can be achieved with the gradient search or the Newton methods. This ensures that the good performance of the SPR approach.

The accuracy of the sigmoid 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 \( |\beta'(Z_i - Z_j)/\sigma_n| > 5 \) (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 $|\hat{\beta}'(Z_i - Z_j)/\sigma_n|$ 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.

An alternative, possibly more efficient, way is to update the tuning parameter in a manner similar to the above approach at each iteration of the gradient search or the Newton methods. However, this demands extra computation by searching $\sigma_n$ that satisfies $|\hat{\beta}'(Z_i - Z_j)/\sigma_n| > 5$ at each iteration. So the total computational efficiency does not differ much from the one-step approach above. In our study, we use the one-step approach.

Compared to most smoothing methods, the SPR has a weaker requirement on the smoothing parameter $\sigma_n$: we only require asymptotically $\sigma_n \to 0$. No special rate restriction is posed. So for the SPR, the tuning parameter selection is less crucial. However, we do note that if the initial choice of $\sigma_n$ is too small or too large, then the proposed tuning parameter selection may not work well. This can usually be detected if there are many extremely large or small $|\hat{\beta}'(Z_i - Z_j)/\sigma_n|$, which warrant a modification of the initial guess $a_n$.

6. VETERANS ADMINISTRATION LUNG CANCER STUDY

As illustration, we apply the proposed approach to the Veterans Administration lung cancer data described in Section 1. As in Therneau and Grambsch (2000, Chapter 6.3), we consider the effect of treatment, karnofsky score (karno), time in months from diagnosis to randomization (diagtime), age in years (age), prior therapy (prior, with 0 for no therapy and 10 otherwise), and histological type of the tumor (celltype: squamous, small cell, adeno, large cell) on the time to death. Specifically, $d = 8$ covariates are included, which are denoted by $Z_1 = \text{karno}/10$, $Z_2 = \text{diagtime}/100$, $Z_3 = I(\text{treatment} = \text{test chemotherapy})$, $Z_4 = \text{age}/100$, $Z_5 = \text{prior}/10$, $Z_6 = I(\text{cell type} = \text{small})$, $Z_7 = I(\text{celltype} = \text{squamous})$, and $Z_8 = I(\text{celltype} = \text{large})$. The nonparametric transformation model is fitted using the SPR approach. An important issue in fitting the nonparametric model is to choose the covariate with the coefficient fixed at 1.
Table 3. Results for the Veterans Administration data

<table>
<thead>
<tr>
<th>Model</th>
<th>Kano/10</th>
<th>Diagtime/100</th>
<th>Treatment</th>
<th>Age/100</th>
<th>Prior/10</th>
<th>Squamous</th>
<th>Small</th>
<th>Large</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Est</td>
<td>SE</td>
<td>Est</td>
<td>SE</td>
<td>Est</td>
<td>SE</td>
<td>Est</td>
<td>SE</td>
</tr>
<tr>
<td>Cox</td>
<td>-0.328</td>
<td>0.055</td>
<td>0.008</td>
<td>0.914</td>
<td>0.295</td>
<td>0.208</td>
<td>-0.871</td>
<td>0.930</td>
</tr>
<tr>
<td>AFT</td>
<td>0.320</td>
<td>0.054</td>
<td>-0.168</td>
<td>0.199</td>
<td>0.568</td>
<td>1.017</td>
<td>-0.091</td>
<td>0.231</td>
</tr>
<tr>
<td>NT</td>
<td>1</td>
<td>—</td>
<td>-0.752</td>
<td>3.351</td>
<td>-0.839</td>
<td>0.530</td>
<td>0.827</td>
<td>1.989</td>
</tr>
</tbody>
</table>

AFT, accelerated failure time model; NT, nonparametric transformation model; Est, estimate; SE, standard error.
One possible way is to fit a proportional hazards model with all \( d \) covariates and choose the continuous covariate with the smallest \( p \)-value. Alternatively, for each continuous covariate, we can fit a proportional hazards model with the covariate as the single one in the model and choose the covariate with the smallest \( p \)-value. We take the first approach for this specific data set and the coefficient for \( Z_1 = \text{karno}/10 \) is 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. The only significant covariate besides karnofsky score is \( I(\text{celltype} = \text{large}) \).

For the purpose of comparison, we also give in Table 3 the results from the Cox proportional hazards model and the accelerated failure time model with the same covariates. If the Cox model fits the data well, we would expect the ratios of the corresponding coefficients from the two models to be close to a constant, which is not the case for this data set. This indicates that the proportional hazards assumption may not hold, which conforms to the results in Therneau and Grambsch (2000, Chapter 6.3). For similar reasons, the accelerated failure time model may not fit the data well.

Note that for the Cox model, the accelerated failure time model, and the nonparametric transformation model, the estimated linear combination of the covariates \( \beta'Z \) 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 and others (2000). Figure 2 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 after 150 days. Thus, the nonparametric transformation model appears to provide a better fit to this data set. However, no significant difference is detected due to the large variances of the AUCs. A larger sample size may be needed for more reliable comparison of the models.

For practical use, the above results only provide an estimate of “relative risk,” i.e. we only have the relative risk score \( \beta'Z \). To get a more comprehensive description of the conditional hazard or the survival

![Fig. 2. Estimated AUC for the Veterans Administration data. Solid line, nonparametric transformation model; dotted line, Cox model; dashed line, accelerated failure time model.](https://academic.oup.com/biostatistics/article-abstract/8/2/197/229993)
time itself, we may need to estimate the link function \( g \) and/or the error distribution. This may be achieved by methods similar to those in the absence of censoring (e.g. Horowitz, 1996; Chen, 2002), but is beyond the scope of our study. For a large number of medical studies, especially those targeting at evaluation of specific covariate effects, we may be more interested in the relative risk or the relative effects of different covariates and the SPR estimate itself is enough in such cases.

We also note that with the PR and SPR estimates, one estimated coefficient needs to be kept fixed. So it is not feasible to evaluate the significance of its corresponding covariate. A possible \textit{ad hoc} solution is to fix another coefficient for a relatively important continuous covariate, and redo the estimation and inference. By rotating the fixed coefficient, we can evaluate the significance of all covariates. Simulation study shows that as long as the chosen covariate is moderately associated with the outcome, the effect of choosing different covariates for identifiability is ignorable.

7. DISCUSSION

We have proposed a SPR 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 PR estimator, but is much easier to compute in the case of multiple covariates. Combined with the flexibility of the transformation model, the proposed approach provides a valuable alternative for analyzing censored survival data. The idea of smoothing the objective function can be extended to estimate the transformation function and will be investigated in our future research. The proposed SPR estimate can be easily achieved using existing software. R code is available upon request.

This paper advances previous publications, especially Ma and Huang (2005), by comprehensively investigating the sigmoid approximation approach in survival studies under the transformation model. The asymptotic properties and the inference based on the weighted bootstrap are established. The proposed approach makes the transformation model more applicable in medical studies. Several important issues, like the merits of the transformation model and when to use it, have been investigated in the literature (see Khan and Tamer, 2004, and the references therein).

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 includes the additive hazards model as a special case, which does not belong to the framework of model (2.1). The estimation and inference procedure in this paper can be extended to this general model and will be pursued in a separate study.

ACKNOWLEDGMENTS

We thank the Editors, the Associate Editor, and a referee for helpful comments. This research is supported in part by the National Institutes of Health grants N01-HC-95159 (Ma) and HL72288 (Huang), and U.S. Department of Veterans Affairs, Veterans Affairs Health Administration, HSR&D grant ECI-03-206 (Zhou and Song). This paper presents the findings and conclusions of the authors. It does not necessarily represent those of Veterans Administration HSR&D Service. Conflict of Interest: None declared.

A. APPENDIX

A.1 Technical details

We note that the only special properties of the sigmoid function we use in the proof of Theorems 4.1 and 4.2 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 \( I(u > 0) \).

### A.2 Regularity conditions

Let \( R \) be the support operator, for example, \( R(Z) \) denotes the support of \( Z \). 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 \( \{ Z \in R(Z); \Pr(\Delta = 1|Z) > 0 \} \) has a positive measure.

A2. The random error \( \epsilon \) is independent of \( C \) and \( Z \).

A3. The first component of \( Z \) has everywhere positive Lebesgue density, conditional on other components.

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

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

A6. \( T \) and \( C \) are conditionally independent given \( Z \).

A7. (a) For each \( x \), the function \( r(x, \beta(\theta)) \) is twice differentiable with respect to \( \theta \) in a neighborhood of \( \theta_0 \) with the \( k \)th derivative \( \nabla_k r(x, \beta(\theta)) \), \( k = 1, 2 \). The second derivative \( \nabla_2 r(x, \beta(\theta)) \) 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. \( E \| \nabla_1 r(x, \beta(\theta_0)) \|^2 \) and \( E \| \nabla_2 r(x, \beta(\theta_0)) \| \) are finite, and \( E\{ \nabla_2 r(x, \beta(\theta_0)) \} \) is nonsingular.

Assumptions A1–A6 are relatively mild and close to their counterparts in Khan and 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 PR estimator in Khan and Tamer (2004).

### A.3 Proof of Theorem 4.1

Since it has been proved in Khan and Tamer (2004) that the PR estimator is consistent, it suffices to prove that \( \sup_{\beta \in \mathbb{R}} |O_n(\beta) - \tilde{O}_n(\beta)| \to_p 0 \).

For any \( \eta > 0 \), we have

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

where

\[
T_{n1} = \frac{1}{n(n-1)} \sum_{i \neq j} \Delta_j I(V_i \geq V_j)(I(\beta'(Z_i - Z_j) > 0) - s_n(\beta'(Z_i - Z_j))) I(\beta'(Z_i - Z_j) \geq \eta),
\]

\[
T_{n2} = \frac{1}{n(n-1)} \sum_{i \neq j} \Delta_j I(V_i \geq V_j)(I(\beta'(Z_i - Z_j) > 0) - s_n(\beta'(Z_i - Z_j))) I(\beta'(Z_i - Z_j) < \eta).
\]

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

### A.4 Proof of Theorem 4.2

Theorem 4.2 can be proved following the method of Sherman (1993). Using the same notations as those in Section 4.1, we first define

\[
\begin{align*}
f(x_1, x_2, \beta(\theta)) &= f^*(x_1, x_2, \beta(\theta)) - f^*(x_1, x_2, \beta(\theta_0)), \\
f_n(x_1, x_2, \beta(\theta)) &= f^*_n(x_1, x_2, \beta(\theta)) - f^*_n(x_1, x_2, \beta(\theta_0)),
\end{align*}
\]

where

\[
\begin{align*}
f^*(x_1, x_2, \beta(\theta)) &= \delta_2 I(v_1 \geq v_2)I(\beta'z_1 - \beta'z_2 > 0) + \delta_1 I(v_2 \geq v_1)I(\beta'z_2 - \beta'z_1 > 0), \\
f^*_n(x_1, x_2, \beta(\theta)) &= \delta_2 I(v_1 \geq v_2)s_n(\beta'z_1 - \beta'z_2) + \delta_1 I(v_2 \geq v_1)s_n(\beta'z_2 - \beta'z_1).
\end{align*}
\]

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

\[
\nabla_1 \tau_n(x, \beta(\theta)) = \nabla_1 \tau(x, \beta(\theta)) + O(\sigma_n) \quad \text{and} \quad \nabla_2 \tau_n(x, \beta(\theta)) = \nabla_2 \tau(x, \beta(\theta)) + O(\sigma_n).
\]

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

\[
\Gamma_{n0}(\theta) = Ef_n(\cdot, \cdot, \theta), \\
g_n(x, \theta) = Ef_n(x, \cdot, \theta) + Ef_n(\cdot, x, \theta) - 2\Gamma_{n0}(\theta) = \tau_n(x, \theta) - \tau_n(x, \theta_0) - 2\Gamma_{n0}(\theta), \quad \text{and}
\]

\[
h_n(x_1, x_2, \theta) = f_n(x_1, x_2, \theta) - Ef_n(x_1, \cdot, \theta) - Ef_n(\cdot, x_2, \theta) + \Gamma_{n0}(\theta).
\]

Next we show that

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

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

where \( G_n = \sqrt{n}P_n \nabla_1 \tau(\cdot, \theta_0) \to_d N(0, B) \), and

\[
U_n h_n(\cdot, \cdot, \theta) = o_p(n^{-1}) \quad \text{(A.4)}
\]

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

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

\[
\Gamma_{n0}(\theta) = \frac{1}{2}(\theta - \theta_0)'A_n(\theta - \theta_0) + O(\sigma_n|\theta - \theta_0|) + o(|\theta - \theta_0|^2), \quad \text{(A.5)}
\]
where $A_n = -\frac{1}{2}E\{\nabla_2 \tau_n(x, \beta(\theta))\}$. 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

$$h(x_1, x_2, \beta(\theta), \sigma) = f(x_1, x_2, \beta(\theta), \sigma) - Ef_n(\cdot, x_2, \beta(\theta), \sigma) + Ef_n(x_1, \cdot, \beta(\theta), \sigma).$$

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,

$$Eh^2_n(x, X, \theta) \to 0 \quad \text{as} \quad (\theta, \sigma_n) \to (\theta_0, 0).$$

Consider $S = \{h(x, X, \beta(\theta), \sigma_n): \beta \in \mathbb{B}, \sigma_n \in (0, 1]\}$. Note that $S$ is Euclidean by Lemma 22 (ii) of Nolan and 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), $\bar{\theta} - \theta_0 = O_p(n^{-1/2})$. The asymptotic normality now follows from Theorem 2 of Sherman (1993). See also Theorem 3.2.16 of Van der Vaart and Wellner (1996, Chapter 3.2).

**References**


[Received December 20, 2005; revised April 24, 2006; accepted for publication April 28, 2006]