Boosting method for nonlinear transformation models with censored survival data

WENBIN LU*, LEXIN LI
Department of Statistics, North Carolina State University, Raleigh, NC 27695, USA
lu@stat.ncsu.edu

SUMMARY
We propose a general class of nonlinear transformation models for analyzing censored survival data, of which the nonlinear proportional hazards and proportional odds models are special cases. A cubic smoothing spline–based component-wise boosting algorithm is derived to estimate covariate effects non-parametrically using the gradient of the marginal likelihood, that is computed using importance sampling. The proposed method can be applied to survival data with high-dimensional covariates, including the case when the sample size is smaller than the number of predictors. Empirical performance of the proposed method is evaluated via simulations and analysis of a microarray survival data.

Keywords: Boosting; Censored survival data; Importance sampling; Marginal likelihood; Nonlinear transformation models; Smoothing spline.

1. INTRODUCTION
The proportional hazards model (Cox, 1972) is one of the most popular semiparametric regression models for the analysis of censored survival data. The distribution of a patient’s survival time $T$ is specified through the hazard function

$$
\lambda(t|Z) = \lambda_0(t) \exp(\beta'Z),
$$

where $\lambda_0(t)$ is a completely unspecified baseline hazard function, $Z$ denotes a $p$-dimensional covariate vector, and $\beta$ is the unknown $p$-dimensional regression parameter vector. Estimation of $\beta$ is achieved by maximizing the partial likelihood function (Cox, 1975), and the resulting estimator is consistent and semiparametrically efficient (Andersen and Gill, 1982).

It has been noted by many authors that, in some medical applications, the proportional hazards model may not be the best choice. For instance, when the hazard ratio of 2 treatment groups converges to 1, the proportional odds model (Pettitt, 1982, 1984; Bennett, 1983; Dabrowska and Doksum, 1988; Murphy and others, 1997) is preferable to the proportional hazards model. More generally, a class of linear transformation models (Clayton and Cuzick, 1985; Bickel and others, 1993; Cheng and others, 1995; Fine and others, 1998) have been proposed to model the distribution of survival time as a function of covariates in the form

*To whom correspondence should be addressed.
where $H$ is an unknown monotonically increasing function and $\epsilon$ is an error term with a known continuous distribution that is independent of $Z$ and the censoring time. Both the proportional hazards model and the proportional odds model are special cases of the linear transformation model (1.2), with the proportional hazards model corresponding to an error with the extreme-value distribution and the proportional odds model to an error with the logistic distribution. In addition, (1.2) generalizes the usual Box–Cox transformation model in a natural way when $\epsilon$ follows the standard normal distribution.

The covariate effect or risk score function $F(Z)$ is assumed linear in (1.2), that is, $F(Z) = \beta'Z$. In some applications, the linearity is induced through some transformation, that is, $F(Z) = \psi(\beta'Z)$, where $\psi$ denotes a known link function. However, in practice, it may be too restrictive to assume a linear form for the risk score function $F(Z)$, and an appropriate link function $\psi$ is often unknown. This is particularly true when the dimension $p$ of the covariates $Z$ is large. Additionally, there are increasing number of applications where the number of covariates $p$ is greater than the number of study subjects $n$. For instance, Rosenwald and others (2002) reported a microarray gene expression data of 240 diffuse large B-cell lymphoma (DLBCL) patients. Patients’ survival times were recorded, and among them, 138 patients died during the follow-up period. In addition, expressions of 7399 genes were measured for each patient. The study was to predict patients’ survival using gene expression information and also to identify important genes contributing to the survival outcomes. Classical solutions would be unsatisfactory for these data where the covariate effects may be nonlinear and the sample size is smaller than the number of predictors. As such it is important to consider a general class of survival models that allow nonparametric forms of the risk score function $F$ and also to develop an estimation procedure that can be relied upon for a wide range of $(n, p)$ values.

Boosting, since its first origination in the machine learning field (Schapire, 1990; Freund, 1995; Freund and Schapire, 1997), has received considerable interest as a very flexible and powerful tool for building predictive models. It is a sequential procedure: a fitting method or a classifier, called the base learner, is applied to repeatedly reweighted data, then a linear combination of such multiple predictions is taken as the final boosting estimator. Breiman (1999) showed that boosting can be viewed as a gradient-descent optimization algorithm in functional space, and such a view has led to emergence of many variants of boosting. Friedman and others (2000) further derived the boosting algorithm as a method for fitting an additive model in a forward stage-wise manner. Consequently, one may view boosting as a “bias”-reducing procedure which increases the flexibility of base learners by incorporating a large number of them in a jointly fitted additive expansion. See Hastie and others (2001) and Buhlmann (2003) for more introductions on boosting. More recently, Friedman (2001) proposed a general gradient boosting paradigm for a variety of loss functions. In particular, Buhlmann and Yu (2003) developed a component-wise cubic smoothing spline–based boosting procedure for $L_2$-loss function. Most of these new developments were designed for classification and regression problems where there is no censoring in the response variable. For censored survival data, Ridgeway (1999) fitted the proportional hazards model with nonlinear covariate effects using boosting and Li and Luan (2005) studied the same model but used a component-wise smoothing spline–based boosting procedure that can handle high-dimensional gene expressions. Both the methods were developed for the proportional hazards model built upon the partial likelihood.

For the linear transformation model (1.2), the conventional partial likelihood function is not available. Most estimation procedures for the regression coefficients in (1.2) are based on estimating equations (Cheng and others, 1995; Fine and others, 1998; Chen and others, 2002), which are not convenient to incorporate the boosting method when the risk score function becomes nonlinear. In this paper, we propose a gradient boosting method based on the marginal likelihood for a class of nonlinear transformation models. In Section 2, we specify the nonlinear transformation model and derive the associated marginal likelihood as well as the boosting algorithm. Numerical studies including simulations and an
Suppose a random sample of \( n \) subjects is chosen. Let \( T_i \) and \( C_i \) denote the survival time (failure time) and censoring time of subject \( i, i = 1, \ldots, n \), respectively. Define \( \tilde{T}_i = \min(T_i, C_i) \) and \( \delta_i = I(T_i \leq C_i) \). We use \( Z_{ij} \) to denote the \( j \)th covariate value for the \( i \)th individual and \( Z_i = (Z_{i1}, \ldots, Z_{ip})^T \) for the \( p \)-dimensional covariate vector of subject \( i \). The observations consist of \( (\tilde{T}_i, \delta_i, Z_i), i = 1, \ldots, n \), which are independent copies of \( (\tilde{T}, \delta, Z) \). The distribution of the survival time \( T \) is specified by the following nonlinear transformation model:

\[
H(T) = -F(Z) + \epsilon, \tag{2.1}
\]

where \( H \) and \( \epsilon \) are the same as defined in (1.2) and \( F(\cdot) \) is an unspecified smooth function with \( F(0) = 0 \) for identifiability. The nonlinear transformation model (2.1) considered in this paper is different from the one proposed by Tsodikov (2003) and Tsodikov and Garibotti (2007), where the survival function of one proposed by Tsodikov (2003) and Tsodikov and Garibotti (2007), where the survival function of \( T \) is constructed using a known nonlinear transformation generating function. In (2.1), the 2 nonparametric functions \( H \) and \( F \) need to be estimated under different constraints: the monotonicity of \( H \) and the smoothness of \( F \). Since the rank of survival times is invariant under any strictly increasing transformation, it motivates us to consider the marginal likelihood function described in Section 2.2.

2.2 The marginal likelihood

The likelihood function of the observed data is given by

\[
L_n(F, H) = \prod_{i=1}^{n} \left[ h(\tilde{T}_i) \lambda_\epsilon \{ H(\tilde{T}_i) + F(Z_i) \} \right]^{\delta_i} e^{-\Lambda_\epsilon \{ H(\tilde{T}_i) + F(Z_i) \}}, \tag{2.2}
\]

where \( h(x) = dH(x)/dx \), \( \lambda_\epsilon(x) = d\Lambda_\epsilon(x)/dx \), and \( \Lambda_\epsilon(x) \) denotes the cumulative hazard function of \( \epsilon \), that is, \( P(\epsilon > x) = \exp[-\Lambda_\epsilon(x)] \). Since the likelihood function \( L_n \) involves 2 nonparametric functions \( F \) and \( H \), it is difficult to apply the boosting method directly to \( L_n \). In addition, due to the lack of the proportionality of the hazard function, the usual partial likelihood for the proportional hazards model may not be available for the class of nonlinear transformation models.

To overcome this difficulty, we propose to use the marginal likelihood, which was also used by Lam and Kuk (1997) for the frailty model and by Lam and Leung (2001) for the proportional odds model. To be specific, let \( T_{(1)} < \cdots < T_{(K)} \) denote the ordered uncensored failure times in the sample and define \( T_{(0)} = 0, T_{(K+1)} = \infty \). For \( 0 \leq k \leq K \), let \( \mathcal{L}_k \) denote the set of labels \( i \) corresponding to those observations censored in the interval \( [T_{(k)}, T_{(k+1)}) \). Due to the censoring scheme, the complete ranking of the \( T_i \)’s is not observed. Let \( \mathbf{R} \) denote the unobserved rank vector of the \( T_i \)’s and let \( \mathcal{C} \) denote the collection of all possible rank vectors of the \( T_i \)’s consistent with the observed data \( (\tilde{T}_i, \delta_i)(i = 1, \ldots, n) \). The marginal likelihood is then defined by \( L_{n,M}(F) = P(\mathbf{R} \in \mathcal{C}) \), where the probability is with respect to the underlying uncensored version of the study. Furthermore, the event \( \{ \mathbf{R} \in \mathcal{C} \} \) can be characterized by the following domain:

\[
D = \{ T_{(k)}, T_i > T_{(k)} | i \in \mathcal{L}_k, k = 0, \ldots, K \}.
\]
It follows that

\[ L_{n,M}(F) = \int \cdots \int_D \prod_{i=1}^{n} \lambda_i(H(T_i) + F(Z_i)) \exp\{-\Lambda_i[H(T_i)+F(Z_i)]\} \prod_{i=1}^{n} dH(T_i). \]  

(2.3)

Define \( V_i = H(T_i), i = 1, \ldots, n \). Since \( H(T) \) is a strictly monotonically increasing function of \( T \), the rank vector of \( V_i \)'s is the same as that of \( T_i \)'s. Using the argument of Clayton and Cuzick (1985), the integral in (2.3) can be expressed as an integral over only the transformed failure times \( V_1, \ldots, V_K \), that is,

\[ L_{n,M}(F) = \int_{V_1 < \cdots < V_K} \prod_{i=1}^{n} [\lambda_i(V(k_i)) + F(Z_i)]^{\delta_i} \exp\{-\Lambda_i[V(k_i)+F(Z_i)]\} \prod_{k=1}^{K} dV(k), \]  

(2.4)

where \( V(k) = H(T(k)), k = 1, \ldots, K \), and \( k_i = \max\{k : T(k) \leq \tilde{T}_i, 0 \leq k \leq K\}, i = 1, \ldots, n \). Note that \( V(0) = -\infty \) and we set \([\lambda_i(V(0)) + F(Z_i)]^{\delta_i} \exp\{-\Lambda_i[V(0)+F(Z_i)]\} = 1\), for any \( i \), with \( k_i = 0 \).

Note that (2.4) is independent of the nonparametric function \( H \), or it is baseline free. In addition, when the model is the proportional hazards model, that is, \( \Lambda_i(x) = \exp(x) \), (2.4) becomes the partial likelihood (Kalbfleisch and Prentice, 2002). However, in general, the integral in (2.4) has no analytical form and a Monte Carlo method is needed to approximate (2.4). Toward this, we multiply and divide the integrand in (2.4) by

\[ c \prod_{i=1}^{n} [\lambda_i(V(k_i))]^{\delta_i} \exp\{-\Lambda_i(V(k_i))\}, \]  

(2.5)

where the constant \( c \) is the total number of possible rank vectors in \( \mathcal{C} \). It can be shown that (2.5) is the density function of \( V_1, \ldots, V_K \) under progressive type II censoring (Lawless, 1982) when the underlying \( V_i (i = 1, \ldots, n) \) are independent and identically distributed according to the distribution function \( G(x) = 1 - \exp\{-\Lambda(x)\} \). Here, progressive type II censoring means that we remove \( l_k \) (the number of observations censored in the interval \([T(k), T(k+1))\) observations at random from the risk set immediately after removing the \( k \)th uncensored observation \( T(k), k = 0, \ldots, K \).

Thus, the marginal likelihood (2.5) can be expressed as

\[ L_{n,M}(F) = E\{Q(V_1, \ldots, V_K; F)\}, \]

where the expectation is with respect to the density (2.5) and

\[ Q(V_1, \ldots, V_K; F) = \frac{1}{c} \prod_{i=1}^{n} [\lambda_i(V(k_i)) + F(Z_i)]^{\delta_i} \exp\{-\Lambda_i[V(k_i)+F(Z_i)]\}. \]  

(2.6)

Now, we can use the importance sampling technique to approximate \( L_{n,M} \). Let \( G^{-1}(x) \) denote the inverse function of \( G(x) \). Then, \( L_{n,M} \) can be approximated by

\[ \hat{L}_{n,M}(F) = \frac{1}{B} \sum_{b=1}^{B} Q[G^{-1}(U_{(1)}^{b}), \ldots, G^{-1}(U_{(K)}^{b}); F], \]  

(2.7)

where \( U_{(1)}^{b}, \ldots, U_{(K)}^{b}, b = 1, \ldots, B \), represent \( B \) independent realizations of the uncensored order statistics of a random sample of size \( n \) from the uniform distribution under the above progressive type II censoring scheme.

As noted by Lam and Kuk (1997), the same set of \( (U_{(1)}^{b}, \ldots, U_{(K)}^{b}) \) is used in (2.7) regardless of the values of \( F \), which saves computing time. In addition, it implies that \( \hat{L}_{n,M}(F) \equiv \log(\hat{L}_{n,M}(F)) \) is a bona
fide function of $F$, thus its gradient with respect to $F$ can be obtained analytically. In Section 2.3, we propose a gradient boosting algorithm for estimating $F$ based on $\hat{L}_{n,M}(F)$.

### 2.3 The boosting algorithm

Boosting can deal with high-dimensional predictors including the small-$n$-large-$p$ case, can naturally incorporate nonlinear covariate effects, and can also be coupled with the marginal likelihood of the transformation model. As such we propose a boosting procedure to nonparametrically estimate the risk score function $F(Z)$ in the nonlinear transformation model (2.1). The procedure includes steps of initiation, evaluation of negative gradient of the loss function, projection of the gradient to a base learner, line search, update, and iterations. For model (2.1), we employ the negative log marginal likelihood, $-\hat{L}_{n,M}(F) = -\log(\hat{L}_{n,M}(F))$, as the loss function, and following Bühlmann and Yu (2003), we employ component-wise cubic smoothing splines as the base learner in the boosting procedure. A detailed description of the algorithm and the corresponding R implementation are given in the supplementary material, available at *Biostatistics* online.

Tuning parameters of the proposed boosting algorithm include a base learner complexity controlling parameter $\rho$ and the total number of iterations. In principle, both parameters can be tuned via cross-validation, however, that can be computationally intensive. Friedman (2001) and Bühlmann and Yu (2003) found that a base learner with a smaller complexity would often lead to a better predictive performance of the boosting algorithm. Moreover, Friedman (2001) and Li and Luan (2005) suggested that the predictive performance is not overly sensitive to the choice of $\rho$. As such we fixed $\rho$ at a small value. We also observed that the marginal likelihood profile flattens after a number of iterations in both simulations and real data analysis. A similar pattern was also observed by Li and Luan (2005) in their partial likelihood-based boosting procedure. We thus adopted the strategy of Li and Luan and stopped the iteration when the marginal likelihood flattens. Li and Luan (2005) noted that such a stopping rule gave almost the same performance as cross-validation. Our experiences also suggested that this stopping rule works pretty well in practice, which we believe is partly attributed to the well-known overfitting-resistance property of boosting. In particular, Bühlmann and Yu (2003) showed that the complexity of the fitted boosting procedure is not increased by a constant amount, but an exponentially diminishing amount as the number of iterations increases. Consequently overfitting comes in very slowly in boosting.

Following our algorithm, at each iteration a function of a single predictor is added. When the iteration stops, an additive model of all the selected predictors is obtained as a final estimate of $F(Z)$. As a consequence, one can easily visualize the individual covariate effects by the marginal plots. It is also interesting to note that the proposed algorithm can perform variable selection. This is because some predictors may never be chosen in the iterative model building procedure, and as such these predictors have no effect on the risk score function. To assess the contribution of those selected predictors, we employ the relative influence measure (Friedman, 2001), which is defined for individual covariate $Z_j$ as

$$I_j = \left( E_Z \left[ \frac{\partial \hat{F}(Z)}{\partial Z_j} \right]^2 \frac{\text{Var}_Z[Z_j]}{\text{Var}_Z[Z_j]} \right)^{1/2}, \quad j = 1, \ldots, p,$$

(2.8)

where $E_Z$ is expectation with respect to the marginal distribution of $Z$ and can be computed through sample average. The larger value of $I_j$ indicates a more important effect of the covariate $Z_j$, with $I_j = 0$ corresponding to no predictor effect. The order of $I_j$, $j = 1, \ldots, p$, provides a useful ordering of all the individual covariates according to their relative importance. Finally, we note that the proposed boosting algorithm works for both $n > p$ and $n < p$ cases, since at each iteration, only a univariate smoothing is performed.
3. SIMULATIONS

In this section, we present the simulation studies demonstrating the effectiveness of the proposed method. We first examined a nonlinear transformation model with $p = 3$. The survival time $T$ was generated from an exponential distribution $T = \exp(-F(Z)+\epsilon)$, with the risk score function $F(Z) = f_1(Z_1) + f_2(Z_2) + f_3(Z_3)$, where $f_1(Z_1) = 1.25Z_1$, $f_2(Z_2) = 1.6Z_2^2 - 0.8Z_2$, and $f_3(Z_3) = \sin(3Z_3)$. We considered both the proportional hazards model, where $\epsilon = \log(-\log(1-u))$, and the proportional odds model, where $\epsilon = \log(u/(1-u))$. In both cases, $u$ is a uniform $(0, 1)$ random variable. The censoring variable $C$ followed uniform $(0, c_0)$, and the constant $c_0$ was selected to control the censoring proportion equal to about 20% and 40%, respectively. The predictors $Z = (Z_1, Z_2, Z_3)^T$ were sampled from independent uniform $(-1, 1)$, with sample size $n = 100$.

Figure 1 shows the simulation results with 20% censoring. In each panel, the solid line represents the true functional form of individual predictor effect $f_i, i = 1, 2, 3$. The dashed curve is the average of the estimated function forms based on 50 data replications, and within each run $B = 1000$ sets of uniform random variables $U_i$’s are sampled for approximating the marginal likelihood function. The top row of panels shows the results for the proportional hazards model and the middle row shows those for the proportional odds model. It is clearly seen from the plot that the proposed boosting estimation method successfully captured both linear and nonlinear predictor effects. We also investigated the robustness of...
the boosting method assuming the proportional hazards model in regardless of the real data generation. For this purpose, we generated data from the proportional odds model but fitted the data using the hazards model. The bottom row of Figure 1 shows the results. It is noted that the estimates are more biased in this case than the previous 2. This indicates the necessity of introducing the boosted odds model. A numerical discrepancy measure $D = \sum_{i=1}^{n}(\hat{f}_{ij} - f_{ij})^2/n$, where $\hat{f}_{ij}$ the estimate of $f_{ij}$, $i = 1, \ldots, n$, $j = 1, 2, 3$, are reported in Figure 1 as well. It reinforces the above findings based on Figure 1. The simulation results for 40% censoring exhibit similar pattern and are thus omitted.

We also conducted simulations for fitting nonlinear transformation models with the number of covariates $p$ as large as 50. Again the boosting method was found to perform well. Moreover, the relative influence function (2.8) was found to provide useful guidance to select important predictors. Details of additional simulations can be found in the supplementary material, available at Biostatistics online.

4. Example: DLBCL data

We next applied the boosting method to the microarray gene expression data of Rosenwald and others (2002) introduced in Section 1. Following Li and Luan (2005), we concentrated on the top 50 genes based upon univariate Cox scores. The results of Li and Luan (2005) and ours indicate that the model does not critically depend on the preliminary gene filtering. The data were randomly divided into a training set with 160 patients and a testing set with 80 patients, and a nonlinear proportional hazards model was fitted to the training data. The log marginal likelihood flattened after 546 iterations. A total of 30 genes were found important, and each one entered the model more than twice. Among them, 15 genes show clearly nonlinear effects. Figure 2 shows the marginal plots of 4 selected important genes and their estimated effects. The 2 on top exhibit clear nonlinear effects, whereas the 2 on bottom show linear trends. It is also

![Gene Effects](https://example.com/image.png)

Fig. 2. Marginal plots of 4 selected genes demonstrating individual covariate effect.
interesting to note that the set of genes identified by our method contain all but one gene found important by Li and Luan (2005). In addition, Rosenwald and others (2002) reported 4 gene signature groups that are believed to be related to the survival of DLBCL, and 17 genes found by our method belong to those signature groups.

We further evaluated the prediction performance of the model built with the training set on both the training and the testing data sets. Figure 3 shows the Kaplan–Meier estimates of survival curves for the high-risk and low-risk group of patients, defined by the predicted risk scores $F(Z)$. The cutoff value was determined by the median of the estimated scores from the training set, and the same cutoff was applied to the testing data. The log-rank test of difference between the 2 survival curves is also reported. It is seen that the model both fitted the training data and predicted the testing data pretty well, achieving a good separation of the 2 risk groups. The $p$ values of the log-rank test were 0 and 0.0056 for the training and testing data sets, respectively.

5. Discussion

In this article, we have proposed a general class of nonlinear transformation models and developed an unified boosting estimation procedure based on the marginal likelihood function. Both the proportional hazards and the proportional odds models are special cases of the models under inquiry, while we demonstrated via simulations that fitting the data with an appropriate model would gain considerable accuracy. The proposed method is flexible enough to permit nonlinear covariate effects, meanwhile it can also facilitate model diagnosis if the covariate effect is indeed linear. In addition to the microarray gene data application, we also applied the proposed method to the mouse leukemia data and the lung cancer data (Kalbfleish and Prentice, 2002). Details of those analyses can be found in the supplementary material, available at Biostatistics online. In summary, we believe that the proposed methodology offers a useful addition to survival data analysis, in particular for modern data applications with high-dimensional covariates.

Theoretical properties of the proposed estimates are difficult to derive, mainly due to 2 reasons: first, the marginal likelihood generally does not have an analytic form and needs to be estimated using importance sampling; second, the covariate effects are nonparametric. Recently, Gu and others (2005) studied the linear transformation models for interval censored data based on the marginal likelihood and gave a heuristic argument showing that the marginal likelihood–based score function is a mean zero martingale.
However, as discussed by the authors, a rigorous proof of the asymptotic results is extremely difficult. So far, no theoretical properties of the proposed estimates are obtained yet, but it definitely warrants our future investigation.

ACKNOWLEDGMENTS

We would like to thank the referees and editors for helpful comments that led to improvements in this manuscript. Conflict of Interest: None declared.

FUNDING

National Science Foundation Grant (DMS-0504269) to W.L. (DMS-0706919) to L.L.

REFERENCES


[Received July 27, 2007; revised January 18, 2008; accepted for publication February 7, 2008]