Published in final edited form as:
*J Am Stat Assoc*. 2010 ; 105(489): 302–311. doi:10.1198/jasa.2009.tm08459.

# Semiparametric Efficient Estimation for a Class of Generalized Proportional Odds Cure Models

**Meng Mao, Ph.D. [candidate in Biostatistics]** and
University of California, Davis, CA 95616 (mmao@wald.ucdavis.edu).

**Jane-Ling Wang [Professor]**
Department of Statistics, University of California, Davis, CA 95616 (wang@wald.ucdavis.edu).

## Abstract

We present a mixture cure model with the survival time of the “uncured” group coming from a class of linear transformation models, which is an extension of the proportional odds model. This class of model, first proposed by Dabrowska and Doksum (1988), which we term “generalized proportional odds model,” is well suited for the mixture cure model setting due to a clear separation between long-term and short-term effects. A standard expectation–maximization algorithm can be employed to locate the nonparametric maximum likelihood estimators, which are shown to be consistent and semiparametric efficient. However, there are difficulties in the M-step due to the nonparametric component. We overcome these difficulties by proposing two different algorithms. The first is to employ an majorize-minimize (MM) algorithm in the M-step instead of the usual Newton–Raphson method, and the other is based on an alternative form to express the model as a proportional hazards frailty model. The two new algorithms are compared in a simulation study with an existing estimating equation approach by Lu and Ying (2004). The MM algorithm provides both computational stability and efficiency. A case study of leukemia data is conducted to illustrate the proposed procedures.

### Keywords

Expectation–maximization algorithm; Logistic regression; Majorize-minimize algorithm; Newton–Raphson algorithm; Nonparametric maximum likelihood estimator; Transformation model

## 1. INTRODUCTION

A typical assumption in survival analysis is that, had there been no censoring, the event would eventually be observed for every subject when the followup time is sufficiently long. However, in some experiments a substantial portion of subjects do not experience the event by the end of the experiment. Farewell (1982) thus assumed that some of these patients will never experience the event. Subjects whose endpoints will not occur are usually referred to as “cured” or “long-term survivors.” An ad hoc way to identify data with a fraction of “cured” subjects is to look at the Kaplan–Meier estimate of the data, where the Kaplan–Meier estimate would have a flat long right tail that is apparently above zero. There are two basic methods to model data with a “cured” fraction. The first is a mixture model proposed by Farewell (1982) and subsequently studied by: Kuk and Chen (1992), Maller and Zhou (1992), Taylor (1995), Peng, Dear, and Denham (1998), Fine (1999), Sy and Taylor (2000), Peng and Dear (2000), Li and Taylor (2002), Lu and Ying (2004), and Fang, Li, and Sun (2005). It assumes that the population consists of a “cured” and an “uncured” group, where the “cured” group will never experience the event or, in other words, the event time of individuals in the “cured” group is infinity. On the other hand, the event time of individuals in the “uncured” group is finite, so classical survival models are applicable.

Let $T^*$ and $C$ denote the event and censoring time, respectively, and $T = T^* \wedge C$ and $\delta = I(T^* \leq C)$ be the observed variable under the standard random censoring scheme. We introduce a binary variable $Y$ as the indicator that an individual will experience the event eventually, that is, belongs to the “uncured” group ($Y = 1$). Otherwise, $Y = 0$ for those cured. Notice that $Y = 0$ implies $\delta = 0$, and $\delta = 1$ implies $Y = 1$. However, the value of $Y$ is missing for censored observations because we do not know whether they will experience the event eventually. The population survival function $S(\cdot)$ for the event time $T^*$ can thus be written as

$$S(\cdot)=pS(\cdot|Y=1)+(1-p) \tag{1}$$

under the mixture model, where $p = \Pr(Y = 1)$ and $S(\cdot| Y = 1)$ is the underlying survival function of the uncured group. Let $\mathbf{X}$ be a $(q + 1)$-dimensional covariate associated with the cure probability, with one as the first component of $\mathbf{X}$. A logistic regression model is often used to model the cure probability $p$, for a subject with covariate $\mathbf{X}$, through

$$p(\mathbf{X};\alpha)=\frac{\exp(\mathbf{X}^T\alpha)}{1+\exp(\mathbf{X}^T\alpha)}. \tag{2}$$

The survival function for the uncured population could be modeled as in standard survival analysis without the cure factor. In this paper, we utilize a general class of models indexed by a known transformation parameter $\rho$. Given a $q$-dimensional covariate vector $\mathbf{Z}$, the survival function for the uncured event time $T^*$ has the form of

$$S_{\mathbf{z}}(\cdot;\rho)=\frac{\exp(-\mathbf{Z}^T\beta/\rho)}{[\exp(-\mathbf{Z}^T\beta)+\rho H_e(\cdot)]^{1/\rho}}, \tag{3}$$

where $H_e(\cdot)$ is an unknown monotone increasing function with $H_e(0) = 0$. Note that when $\rho = 1$, (3) becomes a standard proportional odds (PO) model and when $\rho$ approaches zero, it tends to a proportional hazards model. Besides, similar to PO models, the ratio of $S_{\mathbf{Z}}(\cdot; \rho)^{\rho}$ and $1 - S_{\mathbf{Z}}(\cdot; \rho)^{\rho}$ in (3) at different covariate values is proportional over time. This class of model was mentioned in Zeng and Lin (2007). Since it originates from the work of Dabrowska and Doksum (1988), we adopt a similar term as “generalized proportional odds models.” The covariate vector $\mathbf{Z}$ in (3) and the logistic regression covariate vector $\mathbf{X}$ could share some or no common variables. When $\mathbf{X}^T = (1, \mathbf{Z}^T)$, the survival probability for the uncured group and the population cure probability share the same covariates. It is possible, as in the numerical study in Section 3 of the leukemia data, that the treatment have opposite effects on the cure rate and the survival probability for the uncured group. This contradictory treatment effect, which was unveiled only after we fitted the mixture cure model, demonstrates the advantage of this type of cure model.

An alternative to the mixture cure model is the well-studied bounded cumulative hazard model proposed and studied by Yakovlev and Tsodikov (1996), Tsodikov (1998), and Zeng, Yin, and Ibrahim (2006). The basic difference of this model from the mixture model is that it focuses on modeling the survival function of the combined (cured and uncured) population rather than the survival function of the uncured group.

For known transformation parameter ρ, the generalized PO models (3) belongs to the linear transformation models (Cheng, Wei, and Ying 1995), where $T^*$ is linearly linked to **Z** through an unknown monotone increasing function $H$

$$H(T^*) = -\mathbf{Z}^T\beta + \varepsilon.$$

ε is a random error with a known distribution. With $\Lambda_0(\cdot)$ denoting the cumulative hazard function for ε, the survival function of $T^*$ can be written as

$$S_{\mathbf{z}}(\cdot) = \exp\{-\Lambda_0[\mathbf{Z}^T\beta + H(\cdot)]\}. \quad (4)$$

The generalized PO models (3) can be expressed in the form (4) by equating $H_e(\cdot)$ as $\exp[H(\cdot)]$ and setting $\Lambda_0(\cdot) = \Lambda_0(\cdot;\rho) = \frac{1}{\rho}\log[1 + \rho\exp(\cdot)]$.

Fine (1999) and Lu and Ying (2004) have studied, through estimating equation approaches, the mixture cure model with the survival function for the uncured population coming from the linear transformation models (4). Although these approaches have broad applications, their efficiencies can be improved through a likelihood-based approach. In this paper we focus on the nonparametric maximum likelihood estimator (NPMLE) in the sense of Kiefer and Wolfowitz (1956). Similar to the estimation equation approaches, the NPMLE approach also faces computational challenges due to the presence of the nonparametric component $H_e(\cdot)$. Direct maximization of the likelihood function through the Newton–Raphson approach is unstable and computationally demanding. A solution to overcome this computational burden can be attained by considering the standard linear transformation model (4) as a proportional hazard frailty model. This was cleverly observed in Tsodikov (2003) and exploited in Zeng and Lin (2007) to locate the NPMLE of a linear transformation model without a cure group. An expectation–maximization (EM) algorithm treating the frailty parameter as missing data was proposed in Zeng and Lin (2007), and it provides stable estimates of the nonparametric component $H_e(\cdot)$, through a Breslow-type estimate in the M-step. This EM algorithm can be extended to incorporate long-term survivors, as demonstrated in Section 3.1. However, our numerical experience, including the simulation studies reported in Section 4, reveals that it may encounter difficulties as a Breslow-type estimate for $H_e(\cdot)$ is involved in the algorithm, which increases slowly near the tail, causing the estimate of the survival function $S(\cdot|Y = 1)$ to stay away from zero and indistinguishable from the cure probability. A suggestion by Sy and Taylor (2000) and Lu and Ying (2004) to improve the estimate for the estimate for the cure rate parameter is to set the last jump size of $H_e$ estimate an extremely large value. However, our experience has been that this modification increases the bias of the estimate for **β** in the frailty model-based algorithm and the choice of the last jump size is illusive. We propose an alternative approach that bypass the frailty framework and tackles the computational challenge in the EM algorithm directly using a minimization–maximization (MM) algorithm to perform the maximization step in the M-step. This approach also requires this suggested tip. However, numerical evidence shows that the estimate of $H_e(\cdot)$ in the MM approach can increase much faster than that in the frailty approach (even when more extreme value is used for the latter) and the resulting estimate for **β** is less sensitive to the bias issue associated with the frailty approach.

Our proposal of the generalized PO cure models is motivated by an insightful observation in Lu and Ying (2004) that, compared to the proportional hazards model, a more clear separation between the short-term and long-term covariate effects exists under the proportional odds model, where the hazard functions at different covariate values all converge to the same limit. In other words, in the presence of a cure subpopulation, a proportional odds model for the uncured group is better suited to tackle the identifiability problem in comparison to a proportional hazards model with or without frailty. The generalized PO cure model maintains this property while allowing more flexibility. In this paper, we develop procedures to locate the NPMLE under the mixture cure model [cf. (1) and (2)] with $S(\cdot|Y=1)$ modeled by (3), and establish the semiparametric efficiency of the estimates for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Because a closed-form solution exists during each EM-step to update the nonparametric baseline function, the computation of the NPMLE is actually more stable and conceptually simpler than the algorithm in Lu and Ying (2004), as their approach involves sequential use of the Newton–Raphson algorithm to locate the solutions of a series of estimating equations. It is a pleasant surprise that the proposed estimator is not only efficient in terms of its statistical accuracy but also effective as a numerical procedure.

The rest of the paper is organized as follows. The NPMLE and its asymptotic properties are presented in Section 2. Two algorithms to compute the NPMLEs are described in Section 3. The numerical performance of the estimators are examined in Section 4 through a data example and simulation studies. The findings of this paper are summarized in Section 5 with the proofs relegated to the Appendix.

## 2. MAXIMUM LIKELIHOOD ESTIMATION AND MAIN THEOREMS

Assume the mixture cure model (1) and suppose that for the $i$th subject, we observe $(t_i, \delta_i, \mathbf{x}_i, \mathbf{z}_i)$, where $t_i$ is the minimum of the event and censoring time, $\mathbf{x}_i$ and $\mathbf{z}_i$ are the covariate vectors for the logistic model (2) and the survival model (3) for $S(\cdot|Y=1)$, respectively. The transformation parameter $\rho$ is fixed throughout this section and Section 3. The observed likelihood function can be written as

$$L(\eta)=\prod_i\left\{p(\mathbf{x}_i;\alpha)\times\frac{e^{-\mathbf{z}_i^T\beta/\rho}dH_e(t_i)}{[\rho H_e(t_i)+e^{-\mathbf{z}_i^T\beta}][\rho H_e(t_i^-)+e^{-\mathbf{z}_i^T\beta}]^{1/\rho}}\right\}^{\delta_i}\times\left\{1-p(\mathbf{x}_i;\alpha)+p(\mathbf{x}_i;\alpha)\times\frac{e^{-\mathbf{z}_i^T\beta/\rho}}{[\rho H_e(t_i^-)+e^{-\mathbf{z}_i^T\beta}]^{1/\rho}}\right\}^{1-\delta_i}, \quad (5)$$

where $\eta = (H_e(\cdot), \boldsymbol{\beta}, \boldsymbol{\alpha})$. However, this likelihood is unbounded if $H_e(\cdot)$ is unrestricted. We therefore turn to the NPMLE approach. It can be shown that

*Lemma 2.1.a.* Model (1) is identifiable with $p(\mathbf{x})$ modeled as (2) and $S(\cdot|Y=1)$ modeled as (3) when the support for $\mathbf{x}$ and $\mathbf{z}$ is finite.

*Lemma 2.1.b.* The nonparametric maximum likelihood estimation in the sense of Kiefer and Wolfowitz (1956) for $H_e(t)$ is an increasing stepwise function with jumps at $t_{(i)}$, $i = 1, \ldots, m$, where $t_{(1)} < t_{(2)} < \cdots < t_{(m)}$ are the uncensored observation times.

Let $\tau$ denote the time at the end of the study. Following common practice, we assume in this paper that the true parameter $H_{e0}(\tau)$ is bounded by a constant and $(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0)$ lies in a compact set $B_{\boldsymbol{\beta}} \times B_{\boldsymbol{\alpha}}$.

*Theorem 2.2.* Assume $\rho$ is fixed. When the NPMLE for $H_e$ is bounded on $[0, \tau]$ almost surely, the NPMLE $\hat{\eta} = (\hat{H}_e, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}})$ is strongly consistent, that is, $\|\hat{H}_{en}(\cdot) - H_e(\cdot)\|_\infty$, $\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|$, and $\|\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0\|$ all converge strongly to 0, where $\|\cdot\|_\infty$ denotes the supremum norm on $[0, \tau]$.

*Remark.* An assumption is needed on the NPMLE for $H_e$ to guarantee the consistency in Theorem 2.2. This is a technical assumption that ideally should be satisfied in applications but a proof is lacking at this point. Our numerical experience suggests that it is actually beneficial to make the estimate for $H_e$ larger than the NPMLE at the last uncensored observation.

*Theorem 2.3.* If ($\boldsymbol{\beta}_0$, $\boldsymbol{\alpha}_0$) is an interior point in $B_{\boldsymbol{\beta}} \times B_{\boldsymbol{\alpha}}$ and the conditions of Theorem 2.2 hold, then $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}}$ are semiparametric efficient, where $\sqrt{n}(\hat{\beta}_n - \beta_0)$ and $\sqrt{n}(\hat{\alpha}_n - \alpha_0)$ converge separately in distribution to some multivariate normal distributions with mean **0** and variances $\Sigma_{\boldsymbol{\beta}}$ and $\Sigma_{\boldsymbol{\alpha}}$, respectively.

### Standard Error Estimation

The profile likelihood approach in Murphy, Rossini, and van der Vaart (1997) could be employed to obtain consistent estimates for $\Sigma_\beta^{-1}$ and $\Sigma_\alpha^{-1}$ and take the inverse to get the variance estimates. For example, to estimate $\Sigma_\beta^{-1}$, let $1_i^q$ be the $q$-dimensional vector with one at the $i$th position and zero elsewhere. Define the profile likelihood of $\boldsymbol{\beta}$ as

$$PL_n(\beta)=\max_{\alpha,H_e} L_n(\beta, \alpha, H_e).$$

Calculate

$$-\frac{1}{nh_ih_j}[PL_n(\hat{\beta}+h_i1_i^q+h_j1_j^q) - PL_n(\hat{\beta}+h_i1_i^q) - PL_n(\hat{\beta}+h_ja_j^q)+PL_n(\hat{\beta})] \equiv \hat{\Sigma}_\beta^{-1}\{ij\}$$

for $i \neq j$, and

$$-\frac{1}{nh_i^2}[PL_n(\hat{\beta}+h_i1_i^q) - 2PL_n(\hat{\beta}) - PL_n(\hat{\beta} - h_i1_i^q)] \equiv \hat{\Sigma}_\beta^{-1}\{ii\}.$$

Usually, we can use $h_i=\max(|\hat{\beta}_i|, 1)\ \mathrm{sign}(\hat{\beta}_i)/\sqrt{n}$, where $\boldsymbol{\beta}_i$ is the $i$th element in vector $\boldsymbol{\beta}$. Simulation results in Section 4 support the accuracy of these standard error estimates.

## 3. TWO ALGORITHMS

Lemma 2.1 tells us that the likelihood funtion (5) involves a large number of parameters. Therefore, direct maximization of (5) through the Newton–Raphson approach will be unstable. We thus revert to the EM algorithm, treating the uncured status $Y$ as missing data when it is not available for censored subjects. However, a complication arises in the M-step, due to high-dimensional nonlinear maximization involving the nonparametric function $H_e(\cdot)$. Below we describe two algorithms which avoid inversion of high-dimensional matrices and focus instead on estimating $H_e(\cdot)$ as a step function with jumps at $t_i$ of size $H_e\{t_i\}$.

### 3.1 Proportional Hazards Frailty (PHF) Models Approach

The first approach is not restricted to the generalized PO cure models and can be applied to the general linear transformation cure models in (4). Writing $G(\cdot) = \Lambda_0[\log(\cdot)]$, the survival function can be written as

$$S_{\mathbf{z}}(\cdot)=\exp\{-G[\exp(\boldsymbol{Z}^T\beta)H_e(\cdot)]\}.$$

A similar idea as in Tsodikov (2003) and Zeng and Lin (2007) is to rewrite the above linear transformation model as a proportional hazards frailty model

$$S_{\mathbf{z}}(\cdot)=\exp\{-G[\exp(\mathbf{Z}^T\beta)H_e(\cdot)]\}=E\{\{\exp[-\exp(\mathbf{Z}^T\beta)H_e(\cdot)]\}^U\}, \qquad (6)$$

where the density of the frailty random variable $U$ is the inverse Laplace transformation of $\exp[-G(\cdot)]$. With such a presentation, a standard EM algorithm can be employed and a Breslow type estimate for $H_e(\cdot)$ is available in the E-step, which also provides a profile likelihood estimate for $\boldsymbol{\beta}$. Details of the algorithm are provided in the Appendix.

### 3.2 Minimization–Maximization MM Approach

The complete likelihood for generalized PO model (3) with cure fraction, had $Y$ been observable, is equal to

$$L_c=\prod_i\{p(\mathbf{x}_i;\alpha)^{Y_i}[1-p(\mathbf{x}_i;\alpha)]^{1-Y_i}\}\times\prod_i\left\{\left[\frac{dH_e\{t_i\}}{\rho H_e(t_i)+e^{-\mathbf{z}_i^T\beta}}\right]^{\delta_i}\left[\frac{e^{-\mathbf{z}_i^T\beta}}{\rho H_e(t_i^-)+e^{-\mathbf{z}_i^T\beta}}\right]^{Y_i/\rho}\right\}.$$

For uncensored observations, $Y_i$ is observable and equals to one, so obviously

$$\pi_i\equiv E[Y_i|(T_i,\delta_i=0)]=1.$$

For censored observations,

$$\pi_i\equiv E[Y_i|(T_i,\delta_i=1)]=\frac{p(\mathbf{x}_i;\alpha)S_{\mathbf{z}}(t_i|Y_i=1)}{1-p(\mathbf{x}_i;\alpha)+p(\mathbf{x}_i;\alpha)S_{\mathbf{z}}(t_i|Y_i=1)}.$$

In the M-step, the target function, $E[\log(L_C)|(T_i,\delta_i)]$ is equal to

$$\begin{aligned}&\sum_i\{\pi_i\log[p(\mathbf{x}_i;\alpha)]+(1-\pi_i)\log[1-p(\mathbf{x}_i;\alpha)]\}\\&\quad+\sum_i\left\{\delta_i\{\log(H_e\{t_i\})-\log[\exp(-\mathbf{z}_i^T\beta)+\rho H_e(t_i^-)]\}-\frac{\pi_i}{\rho}\{\mathbf{z}_i^T\beta+\log[\exp(-\mathbf{z}_i^T\beta)+\rho H_e(t_i)]\}\right\}\\&\quad=l_{M1}(\alpha)\\&\quad+l_{M2}(\beta,H_e),\end{aligned}$$

where $l_{M1}$ and $l_{M2}$ can be maximized separately. A simple one-step Newton–Raphson method can be applied to update $\boldsymbol{\alpha}$ in $l_{M1}$, since it is of low dimension. However, the maximization of $l_{M2}$ suffers similar difficulties as many other semiparametric maximization problems in the sense that: (i) the parametric part $\boldsymbol{\beta}$ and the nonparametric part $H_e(\cdot)$ are tangled together; (ii) $H_e(\cdot)$ is high dimensional, of the order $n$. To overcome these difficulties, we utilize a MM algorithm, as detailed below, to maximize $l_{M2}$.

The "minimization–maximization" algorithm was first proposed by Ortega and Rheinboldt (1970, p. 253) and termed MM algorithm by Hunter and Lange (2000). The algorithm we use below is an extension of Hunter and Lange (2002), who introduced maximum likelihood estimation for the standard proportional odds model without a cure group. The maximization of both $\boldsymbol{\beta}$ and $H_e$ is quite stable because it avoids inverting high-dimensional matrices, which is the main attraction for us.

A MM algorithm usually proceeds in two steps. Let $L(\theta)$ be the target function to be maximized and $\theta^{(k)}$ be the current value of iteration. The first step is to construct a surrogate function $R(\theta|\theta^{(k)})$, which is a function of $\theta$ and satisfies: (i) $R(\theta^{(k)}|\theta^{(k)}) = L(\theta^{(k)})$; (ii) $R(\theta|\theta^{(k)}) \le L(\theta)$ for all $\theta$ and $\theta^{(k)}$. In other words, for every fixed $\theta^{(k)}$, the function $R(\theta|\theta^{(k)})$ is always beneath $L(\theta)$ and they meet at $\theta^{(k)}$. Then we maximize or simply increase $R(\theta|\theta^{(k)})$ with respect to $\theta$ to obtain $\theta^{(k+1)}$. This results in $R(\theta^{(k+1)}|\theta^{(k)}) > R(\theta^{(k)}|\theta^{(k)})$. Now one can see that, combing (i) and (ii) leads to $L(\theta^{(k+1)}) \ge R(\theta^{(k+1)}|\theta^{(k)}) > R(\theta^{(k)}|\theta^{(k)}) = L(\theta^{(k)})$. So the target function $L(\theta)$ has been increased.

In our setting, it is more convenient to reparametrize the step function $H_e(\cdot)$ by the logarithm of its jump sizes so that the maximization is unrestricted. Let $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \dots, \boldsymbol{\gamma}_m)$ be the corresponding log of jump sizes of $H_e(\cdot)$ at $t_{(1)}, \dots, t_{(m)}$. For each individual $i$, introduce $d_i = \max\{j : t_{(j)} \le t_i\}$, the index of the latest jump point before $t_i^+$. Thus, $H_e(\cdot)$ is converted to $\boldsymbol{\gamma}$ with $H_e(t_i)=\sum_{j=1}^{d_i}\exp(\gamma_j)$ and

$$l_{M2}(\beta,\gamma)=\sum_i\left\{\delta_i\left[\gamma_{d_i}-\log\left(\rho\sum_{j=1}^{d_i-1}e^{\gamma_j}+e^{-\mathbf{z}_i^T\beta}\right)\right]-\frac{\pi_i}{\rho}\left[\mathbf{z}_i^T\beta+\log\left(\rho\sum_{j=1}^{d_i}e^{\gamma_j}+e^{-\mathbf{z}_i^T\beta}\right)\right]\right\},$$

which is our target function to be maximized. As pointed out in Hunter and Lange (2000), for a convex function $f(v)$, the inequality based on the first order Taylor expansion

$$f(v)\le f\left(v^{(k)}\right)+f'\left(v^{(k)}\right)\left(v-v^{(k)}\right)$$

provides a natural surrogate function which is linear in $v$. A surrogate function for $l_{M2}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ can thus be developed by applying this inequality to the convex function $-\log(\cdot)$ there. This leads to

$$R(\theta|\theta^{(k)})=\sum_{i=1}^{n}\left\{\delta_i\gamma_{d_i}-\frac{\delta_i(\rho\sum_{j=1}^{d_i-1}e^{\gamma_j}+e^{-\mathbf{z}_i^T\beta})}{\rho\sum_{j=1}^{d_i-1}e^{\gamma_j^{(k)}}+e^{-\mathbf{z}_i^T\beta^{(k)}}}-\frac{\pi_i}{\rho}\left[\mathbf{z}_i^T\beta+\frac{\rho\sum_{j=1}^{d_i}e^{\gamma_j}+e^{-\mathbf{z}_i^T\beta}}{\rho\sum_{j=1}^{d_i}e^{\gamma_j^{(k)}}+e^{-\mathbf{z}_i^T\beta^{(k)}}}\right]+C_i(\theta^{(k)})\right\},$$

where $C_i(\theta^{(k)})=\pi_i/\rho+\delta_i-\pi_i\log(\rho\sum_{j=1}^{d_i}e^{\gamma_j^{(k)}}+e^{-\mathbf{z}_i^T\beta^{(k)}})/\rho-\delta_i\log(\rho\sum_{j=1}^{d_i-1}e^{\gamma_j^{(k)}}+e^{-\mathbf{z}_i^T\beta^{(k)}})$ and $\theta = (\boldsymbol{\beta}, \boldsymbol{\gamma})$. Since $C_i(\theta^{(k)})$ does not depend on $\theta$, we can ignore this term and reduce $R(\theta|\theta^{(k)})$ to

$$\begin{aligned}\sum_{i=1}^{n}&\left[-\frac{\pi_i}{\rho}\mathbf{z}_i^T\beta-e^{-\mathbf{z}_i^T\beta}\left(\frac{\delta_i}{\rho\sum_{j=1}^{d_i-1}e^{\gamma_j^{(k)}}+e^{-\mathbf{z}_i^T\beta^{(k)}}}+\frac{\pi_i/\rho}{\rho\sum_{j=1}^{d_i}e^{\gamma_j^{(k)}}+e^{-\mathbf{z}_i^T\beta^{(k)}}}\right)\right]\\&+\sum_{j=1}^{m}\left[u_j\gamma_j-e^{\gamma_j}\left(\sum_{i:d_i>j}\frac{\delta_i\rho}{\rho\sum_{j=1}^{d_i-1}e^{\gamma_j^{(k)}}+e^{-\mathbf{z}_i^T\beta^{(k)}}}+\sum_{i:d_i\geq j}\frac{\pi_i}{\rho\sum_{j=1}^{d_i}e^{\gamma_j^{(k)}}+e^{-\mathbf{z}_i^T\beta^{(k)}}}\right)\right]\\&=\sum_{i=1}^{n}f_i(\beta|\theta^{(k)})+\sum_{j=1}^{m}g_j(\gamma_j|\theta^{(k)}),\end{aligned}\qquad(7)$$

where $u_j$ represents the number of uncensored observation at $t_{(j)}$.

One advantage of this expression is that one can maximize or increase the functions $\sum_{i=1}^{n}f_i(\beta|\theta^{(k)})$ and $\sum_{j=1}^{m}g_j(\gamma_j|\theta^{(k)})$ separately. We use a one-step Newton–Raphson method to update $\boldsymbol{\beta}$ in each iteration. As for the high-dimensional $\boldsymbol{\gamma}$ parameter, there exists a closed-form solution to the equation $g_j^{'}(\gamma_j|\theta^{(k)})=0$. So one can update $\boldsymbol{\gamma}$ by simply setting

$$\gamma_j^{(k+1)}=\log(u_j)-\log\left(\sum_{i:d_i>j}\frac{\delta_i\rho}{\rho\sum_{j=1}^{d_i-1}e^{\gamma_j^{(k)}}+e^{-\mathbf{z}_i^T\beta^{(k)}}}+\sum_{i:d_i\geq j}\frac{\pi_i}{\rho\sum_{j=1}^{d_i}e^{\gamma_j^{(k)}}+e^{-\mathbf{z}_i^T\beta^{(k)}}}\right).$$

The main advantage of the MM algorithm in the M-step is its stability. However, as the algorithm slows down near the end, we thus suggest switching to the Newton–Raphson method near the end of the algorithm. The rationale is that the maximization steps would have stabilized after a few iterations, so applying Newton–Raphson at the end generally does not cause computational complications. This hybrid algorithm is not sensitive to initial values. Meanwhile, it is substantially more stable than applying the Newton–Raphson algorithm in the M-step throughout the iteration. Additional savings in computing time can be achieved by doing only a one-step iteration, instead of a full iteration, when implementing the MM and Newton–Raphson algorithms during the EM iteration. This results in substantial saving in computing time. There is essentially no loss in accuracy, since major updating of the estimates occurs during the main loop of the EM algorithm, and therefore one-step iteration in the intermediate loop suffices. This one-step shortcut is conceptually different from the sequential algorithm in Lu and Ying (2004) based on estimating equations. In the simulation studies presented in Section 4, we demonstrate the advantages of the MM-based EM algorithm as compared to the estimating equation method and the proportional hazards frailty models approach in Section 3.1.

## 4. NUMERICAL STUDIES

We applied both approaches to the leukemia data in Klein and Moeschberger (1997, p. 10, table 1.4). This dataset consists of 101 patients with advanced acute myelogenous leukemia reported to the International Bone Marrow Transplant Registry. Fifty patients received an allogeneic bone marrow transplant from an HLA (Histocompatibility Leukocyte Antigen) matched sibling. The other patients had an autologous bone marrow transplant where their own marrow was reinfused by a high dose of chemotherapy. The leukemia-free survival times (in months) were recorded. Let the covariate $Z$ be a one-dimensional indicator of the autologous transplants and $\mathbf{X}$ = (1, $\mathbf{Z}$) for the logistic regression model. We can see from the Kaplan–Meier plots (Figure 1) that the survival curves of both groups level off after 25 months and stayed above 0.3 with a long tail till the end of study, which is 50 months. This suggests a cure fraction in both populations based on the recommendation in Maller and Zhou (1992).

To apply the two algorithms we need to provide a value for the transformation parameter ρ. It is in general very difficult to estimate a transformation parameter accurately unless the sample size is extremely large. This was also observed by Zeng, Yin, and Ibrahim (2006) for another cure-transformation model. We thus propose an ad hoc model selection method to select the desirable ρ over a grid of candidate values, and choose the one that maximizes the likelihood function. This method is fairly robust in the sense that the survival functions will look very similar if one jitters the value of ρ a little.

Figure 2 provides the estimated likelihood function over values of the transformation parameter ρ obtained from the MM approach (solid) and proportional hazard frailty (PHF) approach (dashed). We can see that the estimated likelihood function of PHF approach is always under the one from MM approach. The resulting maximum likelihood estimate for ρ is 2.1 from MM and 2.3 from PHF. We use the nearest integer 2 for the transformation parameter. The NPMLEs of the regression coefficients from the MM approach are $\hat{\boldsymbol{\beta}} = -1.7171$ and $\hat{\boldsymbol{\alpha}} = (-0.1066, 0.5511)$. Thus, for the uncured group, the ratio of $S^2/(1 - S^2)$ between the autologous and the allogeneic group is $e^{1.7171} = 5.5684$. The estimated cure probability for allogeneic group is $1 - \hat{p}_{allo} = 0.5266$, and $1 - \hat{p}_{auto} = 0.3907$ for the autologous group. This implies that the allogeneic group has a higher cure probability. However, in the uncured group, autologous transplants have a higher survival probability throughout the study. The estimations from the PHF approach are $\hat{\beta} = -0.9834$ and $\hat{\boldsymbol{\alpha}} = (-0.0348, 0.5529)$. The estimated $H_e(\cdot)$ function from both approaches are plotted in Figure 3 up to the third largest event time. The solid curve (MM) increases much faster than the dashed one (PHF) even though both have been set to the same value at the largest event time following our suggestion at the end of Section 3. The estimated survival curves in Figure 1 also reveals that the MM approach (dashed curve) provides a better fit than the PHF approach (dotted curve) when compared to the Kaplan–Meier curves (solid curve). Overall, the MM aproach leads to a satisfactory fit of the generalized PO cure model.

## Simulation

We conducted two simulation settings to examine the performance of the proposed procedure. The first setting reported in Table 1 mimics the setting of the leukemia samples with a transformation parameter ρ = 2, where $\beta_0$ and $\boldsymbol{\alpha}_0$ were set to be −1.7 and (−0.10, 0.55). As for the target baseline function $H_{e0}(\cdot)$, we approximated the NPMLE of $H_e(\cdot)$ from the leukemia study with a continuous function

$$H_{e0}(t)=\begin{cases} (t/2.5)^2, & t \le 15.5 \\ 160(t-15.5)+38.4, & \text{otherwise}, \end{cases}$$

and use it as the true $H_{e0}$-function.

The censoring variable is generated uniformly from [0, 50], providing it with a very similar shape to the Kaplan–Meier estimate of the censoring survival function for the leukemia data. Thus, all target parameters/functions resemble their counterparts estimated from the leukemia data. We aimed at comparing the two EM algorithms with the estimating equation (EE) approach in Lu and Ying (2004). However, we could not get any sensible results for the EE method due to its low convergence rate. Table 1 thus only contains the results of the two EM algorithms. The last jump size of $\hat{H}_e(\cdot)$ is set to be 5 ×$10^3$ for both approaches in Table 1 and $10^3$ for Table 2. The convergence rate for the MM approach is 98.6% when sample size = 100 and 97.4% when sample size = 200; for the PHF approach, the convergence rate is 98.8% when sample size = 100 and 97.8% when sample size = 200. The mean curve of the estimated $H_e(\cdot)$ for $n = 100$ is plotted in Figure 4, where the MM approach is much less biased than the PHF approach near the end of study.

In a second simulation, we modified the setting in Simulation 1 to $\rho = 1$, $\beta_0 = -1.3$, $\boldsymbol{\alpha}_0 = (-0.12, 0.56)$, and

$$H_{e0}(t)=\begin{cases}(t/4.5)^2, & t \leq 16.5\\ 15.3(t-16.5)+13.4, & \text{otherwise}.\end{cases}$$

Both simulations are based on samples of sizes $n = 100$ and 200, each with half of the patients assigned to the autologous group and the other half to the allogeniec group. For initial estimates, we use a rough estimate of $\boldsymbol{\alpha}$ from the end point of the Kaplan–Meier estimate and an estimate from the Cox proportional hazards model as the initial value for $\boldsymbol{\beta}$. Our experience shows that a "reasonable" initial value helps to improve the convergence rate for the EE approach, however, the two EM algorithms are not sensitive to the initial values.

The results for all three approaches are reported in Table 2. While both EM algorithms converge over 95% of the time (sample size = 100, MM convergence rate = 98.4%, PHF convergence rate = 98.4%; sample size = 200, MM convergence rate = 96.8%, PHF convergence rate = 97.2%), the EE method was very unstable with a substantial divergence rate (the convergence rates for the EE method for sample sizes = 100 and 200 are 72.2% and 21%, respectively). We thus can only compare our procedure with those Monte Carlo samples where the EE method converges.

In general, the MM-based EM algorithm has better performance, especially for the estimation of $\beta$. As expected, the precision of the two EM algorithms (MM and PHF) improves as sample size increases, and a fairly large sample size is needed to estimate $\boldsymbol{\alpha}$ satisfactorily. This is a prevalent concern for all cure models. Li, Taylor, and Sy (2001) have provided a comprehensive discussion about this concern. In general, due to a potential lack of data near the end of the study, a sufficient long followup time with large sample size and many censored observations beyond the typical event time is needed to gain additional precision.

## 5. DISCUSSION

In this paper, we investigate the NPMLE for a class of "generalized proportional odds" models with a cure fraction and propose two EM algorithms (PHF-based and MM-based) to locate the NPMLE. One of the EM algorithms (PHF-based) can be also applied to general linear transformation models with a cure fraction, but may not be as efficient as the other version (MM-based), which is specifically designed for the class of generalized PO models. In reality, one needs to specify the value of the transformation parameter $\rho$ but direct estimation is impractical and would magnify the near nonidentifiability problem in cure model because the estimation of $\rho$ and the cumulative hazard function $H_e(\cdot)$ might affect each other. We thus propose a model selection procedure in lieu of estimation in Section 4.

The use of the EM algorithm is quite natural in our setting, since the indicator of cure status can be treated as a missing variable. The main challenge lies in the M-step, due to the high-dimensional parameters of the order $n$. Traditional Newton–Raphson methods would be very unstable here as they involve inverse of high-dimensional matrices. To avoid this problem, we use a much more stable maximizing algorithm, the MM algorithm, which provides a closed-form solution to updating high-dimensional parameters. Our experience suggests that for mixture cure models, the stability of algorithms is crucial, due to the nearly nonidentifiability feature that arises at the end of a study, although some of the algorithms work well when the cure probability is not involved. We also want to clarify that although both the MM and PHF algorithms maximize the same objective function and theoretically they should lead to the same result, the numerical results could be different in practice. The simulation studies demonstrate that the MM algorithm suits better to the "cure" model settings than the latter.

We also proved the semiparametric efficiency of the NPMLE, which might be extended to other classes of linear transformation models when the invertibility of the information operator can be verified, given a specific form of the link function. The advantages of NPMLE go beyond the usual semiparametric efficiency and extend to computational efficiency. The latter is a particular attractive feature of the NPMLE approach, as it also has computational advantages over existing estimators based on estimating equations.

## Acknowledgments

This research is partially supported by NIH R01AGD25218-01A2.

## APPENDIX

## A.1 EM Algorithm Based on Proportional Hazards Frailty Models

The complete log-likelihood, had $Y$ and $U$ been observable, is equal to

$$l_c=\sum_i \{Y_i \log p(\mathbf{x}_i;\alpha)+(1-Y_i)\log[1-p(\mathbf{x}_i;\alpha)]\}+\sum_i \left\{\delta_i \left[\log(U_i)+\mathbf{z}_i^T\beta+\log(H_e\{t_i\})\right]+Y_i[-U_i \exp(\mathbf{z}_i^T\beta)H_e(t_i)]\right\}.$$

In the E-step, we have

$$\pi_i=\frac{p(\mathbf{x}_i;\alpha)S_{\mathbf{z}}(t_i|Y_i=1)}{1-p(\mathbf{x}_i;\alpha)+p(\mathbf{x}_i;\alpha)S_{\mathbf{z}}(t_i|Y_i=1)},$$

for censored observations. In addition, the E-step also involves calculation of the conditional expectation of $Y_iU_i$. Note that

$$E(Y_iU_i|T_i,\delta_i)=E[E(Y_iU_i|Y_i,T_i,\delta_i)]=E[Y_iE(U_i|Y_i=1,T_i,\delta_i)]=E(Y_i|T_i,\delta_i)E(U_i|Y_i=1,T_i,\delta_i)].$$

Denoting $E(U_i| Y_i = 1, T_i, \delta_i)$ by $\hat{U}_i$, it can be shown that

$$\hat{U}_i=\begin{cases} G'[\exp(\mathbf{z}_i^T\beta)H_e(t_i)], & \delta_i=0\\ -\frac{G''[\exp(\mathbf{z}_i^T\beta)H_e(t_i)]}{G'[\exp(\mathbf{z}_i^T\beta)H_e(t_i)]}+G'(\exp(\mathbf{z}_i^T\beta)H_e(t_i)), & \delta_i=1, \end{cases}$$

completing the E-step.

In the M-step, after plugging in $\boldsymbol{\pi}_i$ and $\hat{U}_i$, the target function becomes

$$\sum_i \{\pi_i \log[p(\mathbf{x}_i;\alpha)]+(1-\pi_i)\log[1-p(\mathbf{x}_i;\alpha)]\}+\sum_i \left\{\delta_i \left[\mathbf{z}_i^T\beta+\log(H_e\{t_i\})\right]-\pi_i\hat{U}_i \exp(\mathbf{z}_i^T\beta)H_e(t_i)\right\}=l_{M1}(\alpha)+l_{M2}(\beta,H_e),$$

where $l_{M1}$ and $l_{M2}$ can be maximized separately. A Breslow type estimate for $H_e(\cdot)$ is available when maximizing $l_{M2}$:

$$\hat{H}_e(t_i)=\frac{\delta_i}{\sum_{j\geq i}\pi_i\hat{U}_i \exp(\mathbf{z}_i^T\beta)},$$

which also provides a profile likelihood estimate for **β**.

## A.2 Proofs of the Main Results in Section 2

*Proof of Lemma 2.1.a.* Without loss of generality, we assume the logistic model (2) and the generalized PO model (3) share the same covariate variables and write $p(x)$ as $p(z)$ and $\exp(-z^T\beta)$ as $r(z)$ for convenience. The proof will be similar to that of theorem 2 in Li, Taylor, and Sy (2001) for the identifiability of the proportional hazards mixture cure models. If there exist two different sets of functions $(p(x), r(x), H_e(t))$ and $(p^*(x), r^*(x), H_e^*(t))$ which yield the same survival function for the mixture population, that is, for any $z$ and $t$,

$$p(z)\frac{r(z)^{1/\rho}}{[r(z)+\rho H_e(t)]^{1/\rho}}=p^*(z)\frac{r^*(z)^{1/\rho}}{[r^*(z)+\rho H_e^*(t)]^{1/\rho}},$$

we will have

$$\frac{p(z)}{p^*(z)}=\left(1-\frac{r^*(z)^{1/\rho}}{[r^*(z)+\rho H_e^*(t)]^{1/\rho}}\right)\Big/\left(1-\frac{r(z)^{1/\rho}}{[r(z)+\rho H_e(t)]^{1/\rho}}\right)\equiv c(z), \tag{A.1}$$

where $c(z)$ is not depending on the time $t$. Setting $z = 0$ and solve for $H_e^*(t)$ in (A.1) gives

$$H_e^*(t)=\frac{1}{\rho}\left\{\left[1-c(0)+c(0)(1+\rho H_e(t))^{-1/\rho}\right]^{-\rho}-1\right\}. \tag{A.2}$$

Solving for $r^*(z)$ in (A.1) and plugging in (A.2), we have

$$r^*(z)=\left\{\left[1-c(0)+\frac{c(0)}{(1+\rho H_e(t))^{1/\rho}}\right]^{-\rho}-1\right\}\times\left\{\left[1-c(z)+\frac{c(z)r(z)^{(1/\rho)}}{(r(z)+\rho H_e(t))^{1/\rho}}\right]^{\rho}\right\}\Big/\left(1-\left[1-c(z)+\frac{c(z)r(z)^{(1/\rho)}}{(r(z)+\rho H_e(t))^{1/\rho}}\right]^{\rho}\right). \tag{A.3}$$

Rewrite $r^*(z)$ as $r^*(z) = g(c(z), H_e(t), c(0), r(z))$, for appropriately defined function $g$. Let $\Delta c(z) = c(z) - c(0)$, expand $g$ around $c(0)$ by the first-order Taylor expansion:

$$r^*(z)=g(c(0),H_e(t),c(0),r(z))+\Delta cg'(c^*,H_e(t),c(z),r(z)), \tag{A.4}$$

where $c^*$ is somewhere between $c(0)$ and $c(z)$. Since $r^*(z)$ should be independent of $H_e(t)$, it requires that $\Delta c(z) = 0$ and $c(0) = 1$ in (A.4). Thus, from (A.1) and (A.3), we have $p(z) = p^*(z)$ and $r^*(z) = r(z)$ and hence $H_e^*(t)=H_e(t)$.

*Proof of Lemma 2.1.b.* The NPMLE must be a step function with positive jumps at observation times because of the term $dH_e(t)$ in the likelihood function. Next we show that the NPMLE should only assign positive weights to uncensored observations. This can be seen through the following example. Let $H_{e1}(t)$ be any step function with positive weight at all uncensored observations plus one censored observation time $t^*$ with $t_{(j)} < t^* < t_{(j+1)}$, where $t_{(0)} < t_{(1)} < t_{(2)} < \cdots < t_{(m)} < t_{(m+1)}$ are ordered uncensored observation times with $t_{(0)} = 0$ and $t_{(m+1)} = \infty$. Define $H_{e2}(t) = H_{e1}(t)$ everywhere, except that $H_{e2}(t) = H_{e1}(t_{(j)})$, for $t^* \leq t < t_{(j+1)}$, that is, the weight of $t^*$ in $H_{e1}(t)$ is assigned to $t_{(j+1)}$ in $H_{e2}(t)$. When we compare the two likelihood functions $L(H_{e1}(\cdot), \boldsymbol{\beta}, \boldsymbol{\alpha})$ and $L(H_{e2}(\cdot), \boldsymbol{\beta}, \boldsymbol{\alpha})$, the contribution of each individual would be the same except for those subjects whose observation times fall between $t^{*-}$ and $t_{(j+1)}$ (inclusive). For censored observations in this interval, it is easy to see that

$$\left[1-p(\mathbf{x};\alpha)+p(\mathbf{x};\alpha)\,\frac{e^{-\mathbf{z}^T\beta/\rho}}{\rho H_{e1}(t^-)+e^{-\mathbf{z}^T\beta}}\right]<\left[1-p(\mathbf{x};\alpha)+p(\mathbf{x};\alpha)\frac{e^{-\mathbf{z}^T\beta/\rho}}{\rho H_{e2}(t^-)+e^{-\mathbf{z}^T\beta}}\right].$$

For uncensored observations at $t_{(j+1)}$, we have

$$\left\{p(\mathbf{x};\alpha)\frac{e^{-\mathbf{z}^T\beta/\rho}dH_{e1}(t)}{[\rho H_{e1}(t)+e^{-\mathbf{z}^T\beta}][\rho H_{e1}(t^-)+e^{-\mathbf{z}^T\beta}]}\right\}<\left\{p(\mathbf{x};\alpha)\frac{e^{-\mathbf{z}^T\beta/\rho}dH_{e2}(t)}{[\rho H_{e2}(t)+e^{-\mathbf{z}^T\beta}][\rho H_{e2}(t^-)+e^{-\mathbf{z}^T\beta}]}\right\},$$

because $dH_{e1}(t) < dH_{e2}(t)$ and $H_{e1}(t^-) > H_{e2}(t^-)$ at $t_{(j+1)}$. So $L(H_{e1}(\cdot), \boldsymbol{\beta}, \boldsymbol{\alpha})$ will always be smaller than $L(H_{e2}(\cdot), \boldsymbol{\beta}, \boldsymbol{\alpha})$. This simple illustration holds in general.

*Proof of Theorem 2.2.* From Lemma 2.1.a, we replace $H_e(\cdot)$ by a step function with jump size $H_e\{t_{(i)}\}$ at $t_{(i)}$ in the likelihood function (5). The relationship between NPMLE $\hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{\beta}}$, and $\hat{H}_e$ can be derived by solving the score equation $\frac{\partial \log L(\eta)}{\partial H_e\{t_{(i)}\}}=0$ and this leads to

$$\hat{H}_e\{t_{(i)}\}=\delta_i\left[\sum_{k=1}^{n}\frac{I(t_k\geq t_i)}{\rho\hat{H}_e(t_k)+e^{-\mathbf{z}_k^T\hat{\beta}}}+\frac{(\delta_k-1)I(t_k\geq t_i)}{\rho\hat{H}_e(t_k)+e^{-\mathbf{z}_k^T\hat{\beta}}+e^{-\mathbf{z}_k^T\hat{\beta}/\rho+\mathbf{x}_k^T\hat{\alpha}}}+\frac{\delta_k I(t_k>t_i)}{\rho\hat{H}_e(t_k^-)+e^{-\mathbf{z}_k^T\hat{\beta}}}\right]^{-1}.$$

We introduce $P_n$ and $P_0$ to denote the expectation with respect to the empirical distribution of the data and with respect to the distribution under the true parameter $\eta_0$. Define

$$W_n(u;\eta)=P_n\left[\frac{I(T\geq u)}{\rho H_e(T)+e^{-\mathbf{z}^T\beta}}+\frac{(\delta-1)I(T\geq u)}{\rho H_e(T)+e^{-\mathbf{z}^T\beta}+e^{-\mathbf{z}^T\beta/\rho+\mathbf{x}^T\alpha}}+\frac{\delta I(T>u)}{\rho H_e(T^-)+e^{-\mathbf{z}^T\beta}}\right]$$

and $G_n(u) = P_n\delta I(T \leq u)$, then the NPMLE $\hat{H}_e(\cdot)$ satisfies the equation

$$\hat{H}_e(t)=\int_0^t\frac{dG_n(u)}{W_n(u;\hat{\eta})}.$$

The consistency of the NPMLE can be derived following a similar framework as the proof of theorem 2.2 in Murphy, Rossini, and van der Vaart (1997).

*Proof of Theorem 2.3.* First we derive the score operator and information operator as defined in van der Vaart (1998, p. 371). To do so, we introduce a family of submodels of the form

$\varepsilon \mapsto \eta_\varepsilon \equiv \eta+\varepsilon(\int_0^{\cdot} h_{H_e}\, dH_e, h_\beta, h_\alpha)$, where, for any fixed direction $h \equiv (h_{H_e}(\cdot), h_\beta, h_\alpha)$, $h_{H_e}(\cdot)$ is an arbitrary nonnegative bounded function, and $h_\beta$ and $h_\alpha$ are arbitrary $p$ and $(p + 1)$-dimensional vectors. By taking the derivative of $l(T, \delta; \eta_\varepsilon)$ with respect to $\varepsilon$ and evaluating it at 0, we obtain a score operator

$$S(\eta)(h)=S_{H_e}(\eta)(h_{H_e})+h_\beta^T S_\beta(\eta)+h_\alpha^T S_\alpha(\eta), \tag{A.5}$$

where

$$S_{H_e}(\eta)(h_{H_e})=\delta h_{H_e}(T)-\frac{\int_0^T \rho h_{H_e}\, dH_e}{\rho H_e(T)+e^{-\mathbf{Z}^T\beta}}-\frac{\delta\int_0^{T-} \rho h_{H_e}\, dH_e}{\rho H_e(T^-)+e^{-\mathbf{Z}^T\beta}}+\frac{(1-\delta)\int_0^T \rho h_{H_e}\, dH_e}{\rho H_e(T)+e^{-\mathbf{Z}^T\beta}+e^{-\mathbf{Z}^T\beta/\rho+\mathbf{X}^T\alpha}},$$

$$S_\beta(\eta)=-\mathbf{Z}\left[\delta/\rho-\frac{e^{-\mathbf{Z}^T\beta}}{\rho H_e(T)+e^{-\mathbf{Z}^T\beta}}-\frac{\delta e^{-\mathbf{Z}^T\beta}}{\rho H_e(T^-)+e^{-\mathbf{Z}^T\beta}}\cdot+\frac{(1-\delta)e^{-\mathbf{Z}^T\beta}}{\rho H_e(T)+e^{-\mathbf{Z}^T\beta}+e^{-\mathbf{Z}^T\beta/\rho+\mathbf{X}^T\alpha}}\right],$$

$$S_\alpha(\eta)=\mathbf{X}\left[\delta-\frac{e^{\mathbf{X}^T\alpha}}{1+e^{\mathbf{X}^T\alpha}}+\frac{(1-\delta)e^{\mathbf{X}^T\alpha}}{\rho H_e(T)+e^{-\mathbf{Z}^T\beta}+e^{-\mathbf{Z}^T\beta/\rho+\mathbf{X}^T\alpha}}\right].$$

The information operator

$$\sigma(h)=(\sigma_{H_e}(h), \sigma_\beta(h), \sigma_\alpha(h))$$

is the solution to

$$P_0[S^2(\eta_0)(h)]=\int \sigma_{H_e}(h)(u)h_{H_e}(u)\, dH_{e0}(u)+h_\beta^T\sigma_\beta(h)+h_\alpha^T\sigma_\alpha(h). \tag{A.6}$$

In our setting,

$$\sigma_{H_e}(h)(u)=h_{H_e}(u)P_0(A_1)+P_0(A_2^2-2A_3)-h_\beta' P_0[(A_1+A_2e^{-\mathbf{Z}^T\beta_0}-A_3)\mathbf{Z}]-h_\beta' P_0(A_1e^{-\mathbf{Z}^T\beta_0}\mathbf{Z})A_5(u)-h_\alpha' P_0\left[\left(A_1+A_1A_4-A_3-\frac{e^{\mathbf{X}^T\alpha_0}}{1+e^{\mathbf{X}^T\alpha_0}}\right)\mathbf{X}\right],$$

$$\begin{aligned}\sigma_\beta(h)=P_0\{[&B_1(T)\\ &+B_1^2(T)e^{-\mathbf{Z}^T\beta_0}\\ &-2\delta B_1(T)\\ &/\rho^2]\times e^{-\mathbf{Z}^T\beta_0}ZZ^T\}h_\beta-P_0\{[B_2(T)+B_1(T)B_2(T)e^{-\mathbf{Z}^T\beta_0}\\ &-\delta B_2(T)\\ &/\rho-\delta B_1(T)e^{-\mathbf{Z}^T\beta_0}h_{H_e}(T)/\rho]\mathbf{Z}\},\end{aligned}$$

$$\sigma_{\alpha}(h)=P_0\ \left\{\left[C_1(T)+C_1^2(T)-\frac{2\delta e^{\mathbf{X}^T\alpha_0}}{1+e^{\mathbf{X}^T\alpha_0}}\right]\mathbf{X}\mathbf{X}^T\right\}h_{\alpha}-P_0\left\{\left[B_2(T)+C_1(T)B_2(T)-\delta B_2(T)-\frac{\delta e^{\mathbf{X}^T\alpha_0}}{1+e^{\mathbf{X}^T\alpha_0}}h_{H_e}(T)\right]\mathbf{X}\right\},$$

and

$$A_1=A_1(u;T,\delta,\eta_0)=\frac{I(T\geq u)}{\rho H_{e0}(T)+e^{-\mathbf{Z}^T\beta_0}}+\frac{\delta I(T>u)}{\rho H_{e0}(T^-)+e^{-\mathbf{Z}^T\beta_0}}-\frac{(1-\delta)I(T\geq u)}{\rho H_{e0}(T)+e^{-\mathbf{Z}^T\beta_0}+e^{-\mathbf{Z}^T\beta_0/\rho+\mathbf{X}^T\alpha_0}},$$

$$A_2=A_2(u;T,\delta,\eta_0)=\frac{I(T\geq u)\int_0^T\rho h_{H_e}\ dH_{e0}}{\rho H_{e0}(T)+e^{-\mathbf{Z}^T\beta_0}}+\frac{\delta I(T>u)\int_0^{T-}\rho h_{H_e}\ dH_{e0}}{\rho H_{e0}(T^-)+e^{-\mathbf{Z}^T\beta_0}}-\frac{(1-\delta)I(T\geq u)\int_0^T\rho h_{H_e}\ dH_{e0}}{\rho H_{e0}(T)+e^{-\mathbf{Z}^T\beta_0}+e^{-\mathbf{Z}^T\beta_0/\rho+\mathbf{X}^T\alpha_0}},$$

$$A_3=A_3(u;T,\delta,\eta_0)=\frac{\delta I(T\geq u)}{\rho H_{e0}(T)+e^{-\mathbf{Z}^T\beta_0}}+\frac{\delta I(T>u)}{\rho H_{e0}(T^-)+e^{-\mathbf{Z}^T\beta_0}},$$

$$A_4=A_4(u;T,\delta,\eta_0)=\frac{I(T\geq u)e^{\mathbf{X}^T\alpha_0}}{1+e^{\mathbf{X}^T\alpha_0}}-\frac{(1-\delta)e^{-\mathbf{X}^T\alpha_0}I(T\geq u)}{\rho H_{e0}(T)+e^{-\mathbf{Z}^T\beta_0}+e^{-\mathbf{Z}^T\beta_0/\rho+\mathbf{X}^T\alpha_0}},$$

$$A_5(u)=A_5(u;\eta_0)=\frac{1}{\rho H_{e0}(u)+e^{-\mathbf{Z}^T\beta_0}}+\frac{1}{\rho H_{e0}(u^-)+e^{-\mathbf{Z}^T\beta_0}},$$

$$B_1(T)=\frac{1}{\rho H_{e0}(T)+e^{-\mathbf{Z}^T\beta_0}}+\frac{\delta}{\rho H_{e0}(T^-)+e^{-\mathbf{Z}^T\beta_0}}-\frac{1-\delta}{\rho H_{e0}(T)+e^{-\mathbf{Z}^T\beta_0}+e^{-\mathbf{Z}^T\beta_0/\rho+\mathbf{X}^T\alpha_0}},$$

$$B_2(T)=\frac{\int_0^T\rho h_{H_e}\ dH_{e0}}{\rho H_{e0}(t)+e^{-\mathbf{Z}^T\beta_0}}+\frac{\delta\int_0^{T-}\rho h_{H_e}\ dH_{e0}}{\rho H_{e0}(T^-)+e^{-\mathbf{Z}^T\beta_0}}-\frac{(1-\delta)\int_0^T\rho h_{H_e}\ dH_{e0}}{\rho H_{e0}(T)+e^{-\mathbf{Z}^T\beta_0/\rho}+e^{-\mathbf{Z}^T\beta_0+\mathbf{X}^T\alpha_0}},$$

$$C_1(T)=\frac{e^{\mathbf{X}^T\alpha_0}}{1+e^{\mathbf{X}^T\alpha_0}}-\frac{(1-\delta)e^{\mathbf{X}^T\alpha_0}}{\rho H_{e0}(T)+e^{-\mathbf{Z}^T\beta_0}+e^{-\mathbf{Z}^T\beta_0/\rho}+\mathbf{X}^T\alpha_0}.$$

This can be shown from $\boldsymbol{\sigma}H_e\,(h)(u)=I_1-I_4-I_5$, $\boldsymbol{\sigma}_{\boldsymbol{\beta}}(h)=I_2-I_4-I_6$ and $\boldsymbol{\sigma}_{\boldsymbol{\alpha}}(h)=I_3-I_5-I_6$, where

$$P_0[S\,(\eta_0)(h)]^2=I_1+I_2+I_3-2I_4-2I_5-2I_6,$$

and

$$I_1=P_0[S^2_{H_e}(\eta)(h_{H_e})], I_2=P_0[h^T_\beta S_\beta(\eta)S^T_\beta(\eta)h_\beta],$$

$$I_3=P_0[h^T_\alpha S_\alpha(\eta)S^T_\alpha(\eta)h_\alpha], I_4=P_0[h^T_\beta S_\beta(\eta)S_{H_e}(\eta)\,(h_{H_e})],$$

$$I_5=P_0[h^T_\alpha S_\alpha(\eta)S_{H_e}(\eta)\,(h_{H_e})], I_6=P_0[h^T_\beta S_\beta(\eta)S^T_\alpha(\eta)h_\alpha].$$

Adopting a similar framework as in the proof of theorem 2.3 in Murphy, Rossini, and van der Vaart (1997), it suffices to show that the information operator $\sigma$ is onto and continuously invertible. As pointed out by Murphy, Rossini, and van der Vaart (1997) in the proof of lemma A.3, this can be done by writing $\sigma$ as the sum of two operators $A$ and $K$, where $A(h) = (h_{H_e}W(\cdot;\eta_0), h_\beta, h_\alpha)$, then proving $A^{-1}K$ is compact and $\sigma$ is one to one. The first part is relatively straightforward and is omitted here. Hence we only need to show the second part. Note from (A.6) that $\|\sigma(h)\| = 0$ implies for almost every $\mathbf{z}$, that the score operator $S(\eta_0)(h)$ is almost surely equal to zero with respect to $(T, \delta)$. When $\delta = 1$, this becomes, for almost every $\mathbf{z}$ and $t$,

$$h_{H_e}(t)-h^T_\beta\mathbf{z}-\frac{1}{1+e^{\mathbf{x}^T\alpha_0}}h^T_\alpha\mathbf{x}-\left[\frac{1}{\rho H_{e0}(t)+e^{-\mathbf{z}^T\beta_0}}+\frac{1}{\rho H_{e0}(t^-)+e^{-\mathbf{z}^T\beta_0}}\right]\times\left[\int_0^t\rho h_{H_e}\,dH_{e0}-h^T_\beta ze^{-\mathbf{z}^T\beta_0}\right]=0. \quad (A.7)$$

Define $t^* = \inf\{t: H_{e0}(t) > 0\}$. By the continuity of $H_{e0}(\cdot)$, we have $H_{e0}(t^*) = 0$ and $H_{e0}(t) > 0$ for $t > t^*$. When evaluated at $t^*$, (A.7) becomes

$$h_{H_e}(t^*)+h^T_\beta\mathbf{z}/\rho-\frac{1}{1+e^{\mathbf{x}^T\alpha_0}}h^T_\alpha\mathbf{x}=0. \quad (A.8)$$

Subtracting (A.8) from (A.7), we have for almost every $\mathbf{z}$ and $t > t^*$, that

$$h_{H_e}(t)-h_{H_e}(t^*)=\left[\frac{1}{\rho H_{e0}(t)+e^{-\mathbf{Z}^T\beta_0}}+\frac{1}{\rho H_{e0}(t^-)+e^{-\mathbf{Z}^T\beta_0}}\right]\times\left[\int_0^t\rho h_{H_e}\,dH_{e0}-h^T_\beta ze^{-\mathbf{z}^T\beta_0}\right]+2h^T_\beta z/\rho. \quad (A.9)$$

Note that the left-hand side of (A.9) does not depend on $\mathbf{z}$, which implies $h_{H_e}(t) = 0$ and $h_\beta = 0$, and hence $h_\alpha = 0$.

## REFERENCES

Cheng SC, Wei LJ, Ying Z. Analysis of Transformation Models With Censored Data. Biometrika. 1995; 87:867–878. [303].

Dabrowska DM, Doksum KA. Estimation and Testing in the Two-Sample Generalized Odds-Rate Model. Journal of the American Statistical Association. 1988; 83:744–749. [302].

Fang H, Li G, Sun J. Maximum Likelihood Estimation in a Semiparametric Logistic/Proportional-Hazards Mixture Model. Scandinavian Journal of Statistics. 2005; 32:59–75. [302].

Farewell VT. The Use of Mixture Models for the Analysis of Survival Data With Long-Term Survivors. Biometrics. 1982; 43:181–192. [302].

Fine JP. Analysing Competing Risks Data With Transformation Models. Journal of the Royal Statistical Society, Ser. 1999; 61:817–830. [302,303].

Hunter DR, Lange K. Rejoinder to Optimization Transfer Using Surrogate Objective Functions. Journal of Computational and Graphical Statistics. 2000; 9:52–59. [305].

Hunter DR, Lange K. Computing Estimates in the Proportional Odds Model. Annals of the Institute of Statistical Mathematics. 2002; 54:155–168. [305].

Kiefer J, Wolfowitz J. Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters. The Annals of Mathematical Statistics. 1956; 27:887–906. [303,304].

Klein, JP.; Moeschberger, ML. Survival Analysis: Techniques for Censored and Truncated Data. New York: Springer; 1997. [306]

Kuk AYC, Chen C. A Mixture Model Combining Logistic Regression With Proportional Hazards Regression. Biometrika. 1992; 79:531–541. [302].

Li C-S, Taylor JMG. A Semi-Parametric Accelerated Failure Time Cure Model. Statistics in Medicine. 2002; 21:3235–3247. [302]. [PubMed: 12375301]

Li C-S, Taylor JMG, Sy JP. Identifiability of Cure Models. Statistics & Probability Letters. 2001; 54:389–395. [307,309].

Lu W, Ying Z. On Semiparametric Transformation Cure Models. Biometrika. 2004; 91:331–343. [302,303,306,307].

Maller RA, Zhou S. Estimating the Proportion of Immunes in a Censored Sample. Biometrika. 1992; 79:731–739. [302,306].

Murphy SA, Rossini AJ, van der Vaart AW. Maximum Likelihood Estimation in the Proportional Odds Model. Journal of the American Statistical Association. 1997; 92:968–976. [304,310,311].

Ortega, JM.; Rheinboldt, WC. Iterative Solution of Nonlinear Equations in Several Variables. Orlando: Academic Press; 1970. [305]

Peng Y, Dear KBG. A Nonparametric MixtureModel for Cure Rate Estimation. Biometrics. 2000; 56:237–243. [302]. [PubMed: 10783801]

Peng Y, Dear KBG, Denham JW. A Generalized F Mixture Model for Cure Rate Estimation. Statistics in Medicine. 1998; 17:813–830. [302]. [PubMed: 9595613]

Sy JP, Taylor JMG. Estimation in a Cox Proportional Hazards Cure Model. Biometrics. 2000; 56:227–236. [302,303]. [PubMed: 10783800]

Taylor JMG. Semi-Parametric Estimation in Failure Time Mixture Models. Biometrics. 1995; 51:899–907. [302]. [PubMed: 7548707]

Tsodikov AD. A Proportional Hazards Model Taking Account of Long-Term Survivors. Biometrics. 1998; 54:1508–1516. [303]. [PubMed: 9883549]

Tsodikov AD. Semiparametric Models: A Generalized Self-Consistency Approach. Journal of the Royal Statistical Society, Ser. 2003; 65:759–774. [303,304].

van der Vaart, AW. Asymptotic Statistics. Cambridge: U.K.: Cambridge University Press; 1998. [310]

Yakovlev, AY.; Tsodikov, AD. Stochastic Models of Tumor Latency and Their Biostatistical Applications. Singapore: World Scientific; 1996. [303]

Zeng D, Lin DY. Maximum Likelihood Estimation in Semi-parametric Regression Models With Censored Data. Journal of the Royal Statistical Society, Ser. 2007; 69:507–564. [302–304].

Zeng D, Yin G, Ibrahim JG. Semiparametric Transformation Models for Survival Data With a Cure Fraction. Journal of the American Statistical Association. 2006; 101:670–684. [303,306].

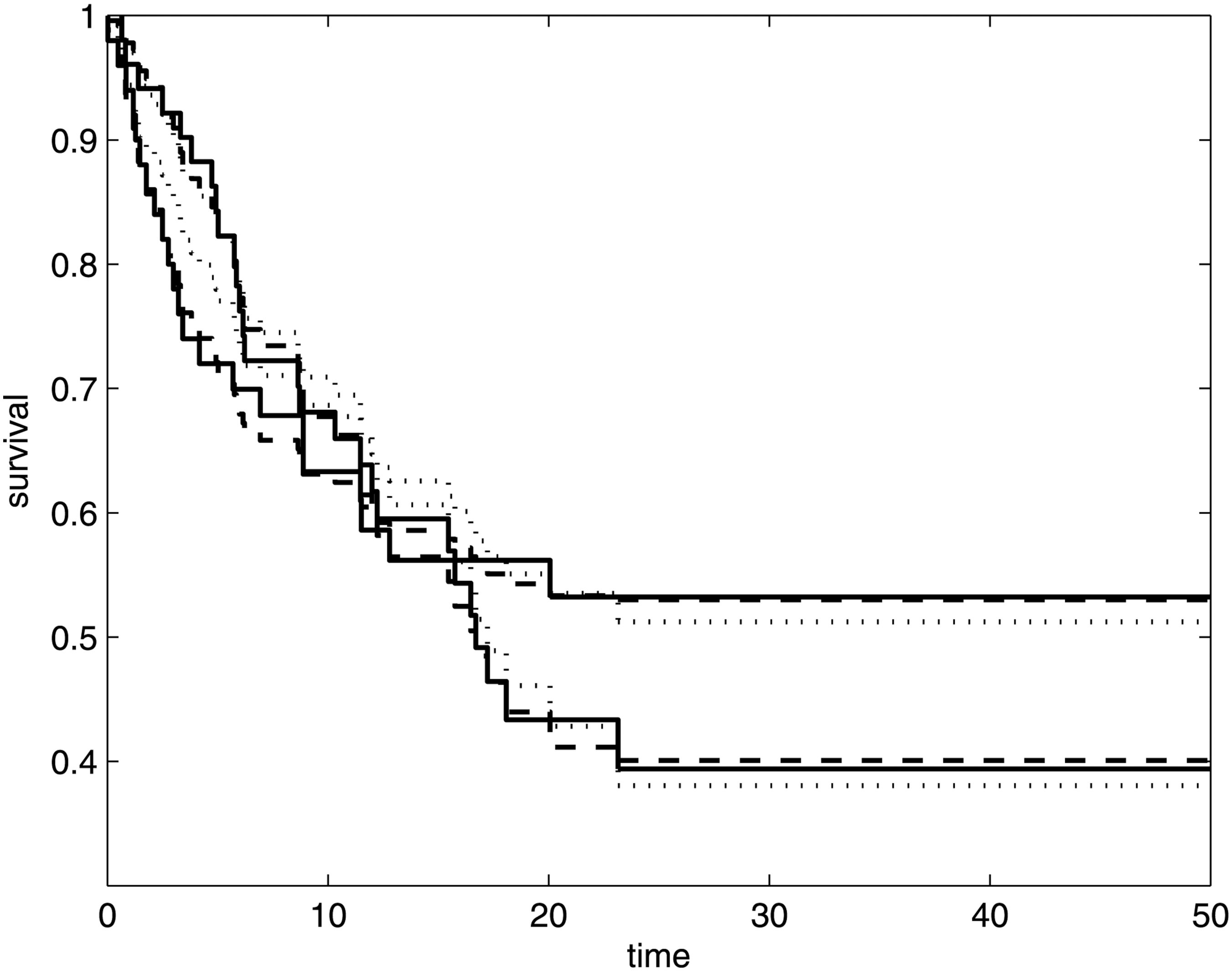

**Figure 1.**

Estimated survival curve for the generalized PO model with ρ = 2 by the Kaplan–Meier estimates (solid), MM approach (dashed) and PHF approach (dotted).

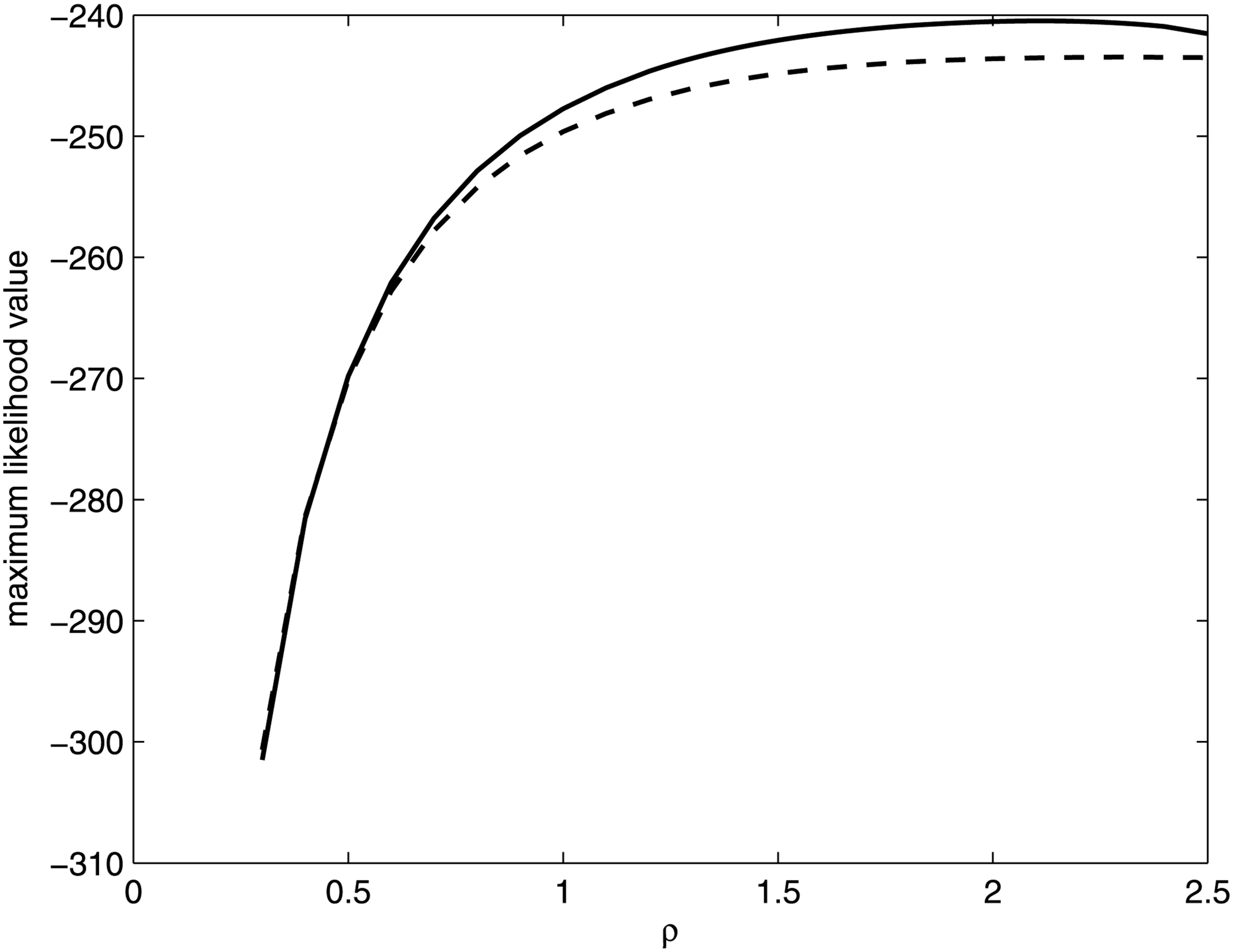

**Figure 2.**
Maximum likelihood function of the transformation parameter ρ by the MM approach (solid) and PHF approach (dashed).

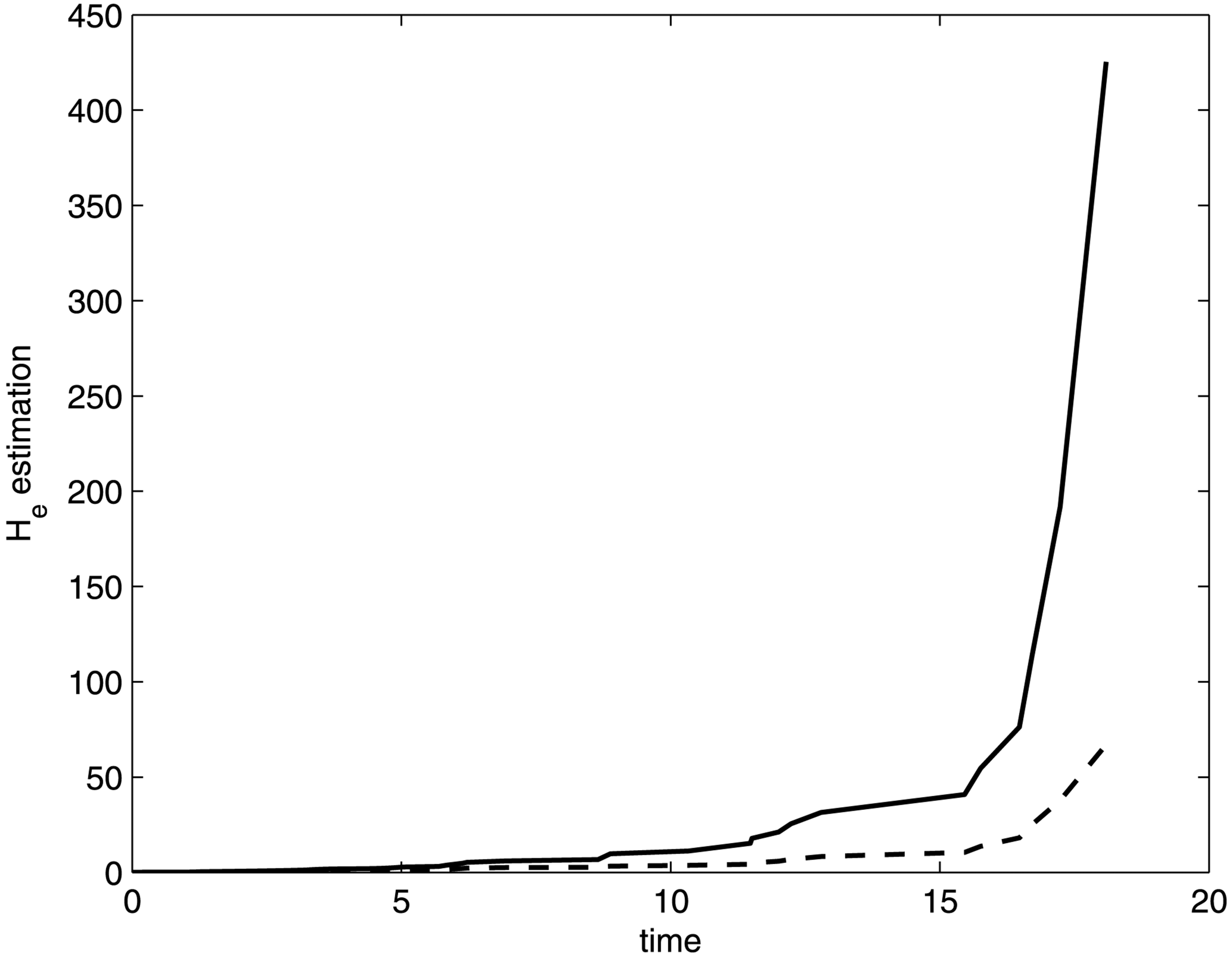

**Figure 3.**

Estimated $H_e(\cdot)$ by the MM approach (solid) and PHF approach (dashed).

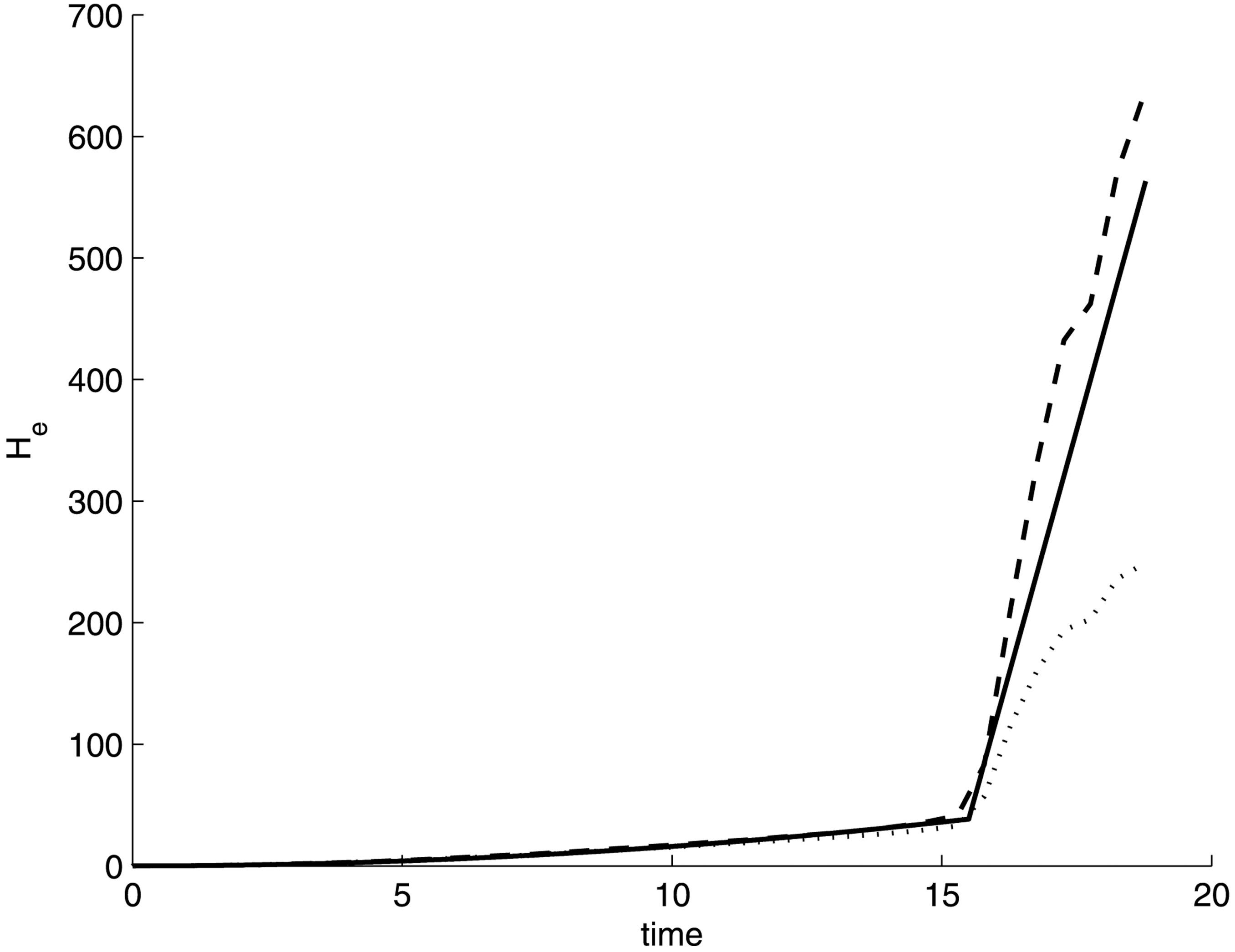

**Figure 4.**
Mean of the estimated $H_e(\cdot)$ by the MM approach (dashed) and PHF approach (dotted) compared with $H_{e0}(\cdot)$ (solid).

### Table 1

500 Monte Carlo samples of sample size 100 and 200 for two EM algorithms when ρ = 2. The estimated standard deviation is listed on the third column and the corresponding Monte Carlo standard deviation is listed in column 4

| | | True | Estimation | Est SD | MC SD | MSE |
|---|---|---|---|---|---|---|
| $n = 100$ | $\beta(MM)$ | −1.7 | −1.9129 | 0.7565 | 0.7861 | 0.6633 |
| | $(PHF)$ | −1.7 | −2.0157 | 0.7642 | 0.7846 | 0.7153 |
| | $\alpha_1(MM)$ | −0.1 | −0.1085 | 0.3103 | 0.3325 | 0.1106 |
| | $(PHF)$ | −0.1 | −0.0940 | 0.3602 | 0.3371 | 0.1137 |
| | $\alpha_2(MM)$ | 0.55 | 0.5915 | 0.4598 | 0.4919 | 0.2437 |
| | $(PHF)$ | 0.55 | 0.6121 | 0.4814 | 0.5103 | 0.2643 |
| $n = 200$ | $\beta(MM)$ | −1.7 | −1.7481 | 0.4558 | 0.4859 | 0.2384 |
| | $(PHF)$ | −1.7 | −1.8226 | 0.4577 | 0.4856 | 0.2508 |
| | $\alpha_1(MM)$ | −0.1 | −0.1057 | 0.2185 | 0.2338 | 0.0547 |
| | $(PHF)$ | −0.1 | −0.1167 | 0.2355 | 0.2360 | 0.0560 |
| | $\alpha_2(MM)$ | 0.55 | 0.5763 | 0.3251 | 0.3453 | 0.1199 |
| | $(PHF)$ | 0.55 | 0.5836 | 0.3523 | 0.3618 | 0.1320 |

## Table 2

360 ($n$ = 100) and 105 ($n$ = 200) Monte Carlo samples that converge under the EE method when $\rho = 1$. The estimated standard deviation of the EM algorithms is listed on the third column and the corresponding Monte Carlo standard deviation of both estimates are listed in column 4. The last column gives the MSE ratio of EE/MM and EE/PHF

| | | True | Estimation | Est SD | MC SD | MSE | MSE ratio |
|---|---|---|---|---|---|---|---|
| $n$ = 100 | $\beta$(*MM*) | −1.3 | −1.3994 | 0.5745 | 0.5737 | 0.3429 | 141.03% |
| | (*PHF*) | −1.3 | −1.5150 | 0.5655 | 0.5700 | 0.3711 | 130.32% |
| | (*EE*) | −1.3 | −1.6066 | | 0.6242 | 0.4836 | |
| | $\alpha_1$(*MM*) | −0.12 | −0.1808 | 0.3065 | 0.2804 | 0.0823 | 102.43% |
| | (*PHF*) | −0.12 | −0.1318 | 0.3160 | 0.3124 | 0.0977 | 86.28% |
| | (*EE*) | −0.12 | −0.1862 | | 0.2827 | 0.0843 | |
| | $\alpha_2$(*MM*) | 0.56 | 0.5076 | 0.4510 | 0.4353 | 0.1922 | 100.10% |
| | (*PHF*) | 0.56 | 0.6096 | 0.4852 | 0.4588 | 0.2130 | 110.65% |
| | (*EE*) | 0.56 | 0.5284 | | 0.4375 | 0.1924 | |
| $n$ = 200 | $\beta$(*MM*) | −1.3 | −1.3376 | 0.3967 | 0.4104 | 0.1699 | 119.54% |
| | (*PHF*) | −1.3 | −1.3933 | 0.3993 | 0.4035 | 0.1715 | 118.43% |
| | (*EE*) | −1.3 | −1.4575 | | 0.4222 | 0.2031 | |
| | $\alpha_1$(*MM*) | −0.12 | −0.1544 | 0.2174 | 0.2094 | 0.0450 | 103.11% |
| | (*PHF*) | −0.12 | −0.1217 | 0.2226 | 0.2218 | 0.0492 | 94.31% |
| | (*EE*) | −0.12 | −0.1562 | | 0.2123 | 0.0464 | |
| | $\alpha_2$(*MM*) | 0.56 | 0.5076 | 0.3210 | 0.3242 | 0.1079 | 101.85% |
| | (*PHF*) | 0.56 | 0.5798 | 0.3408 | 0.3436 | 0.1185 | 92.74% |
| | (*EE*) | 0.56 | 0.5270 | | 0.3299 | 0.1099 | |