On Consistent Variable Selection in Semiparametric Mixture Cure Models

Zhaoyin Zhu* Yongzhao Shao†

Abstract

Semiparametric mixture cure models have been increasingly used for analyzing time-to-event data in many applications including cancer research where there are two latent groups of patients, those who could eventually experience events and those who become immune or cured after certain treatments. For many cancer patients, a large number of variables are routinely collected in many clinical and medical examinations including those with significant predictive or prognostic importance. The successful implementation and application of mixture cure models highly depend on the identification of important risk factors that affect the cure probability and/or the survival distributions among uncured subjects. However, there is a lack of rigorous justification for variable selection in this context due to the challenges related to unknown cure status and heavy data censoring. The aim of this paper is to establish validity and asymptotic optimal properties of the penalized likelihood based methods for variable selection in mixture cure models with adaptive lasso and other penality functions. The finite sample properties of several different types of penalties and tuning parameter selection criteria are compared in simulation studies. The performance of the variable selection methods is also illustrated using a cohort of melanoma patients at New York University Cancer Center.

Key Words: Mixture cure model, Variable selection, Consistency, Adaptive LASSO, SCAD, Oracle property.

1. Introduction

With the recent advance of immunotherapies and targetted therapies for melanoma and other cancers, the curability of a significant portion of advanced as well as early-stage cancers is becoming a reality. The data of these studies typically have heavy censoring at the end of the follow-up period, and a standard survival analysis that ignoring the cure fraction would not always be appropriate. In such situations where there is sufficient evidence of a cured subpopulation, the mixture cure model can be used as a powerful statistical tool for analyzing such studies (Maller and Zhou, 1996; Yilmaz *and others*, 2013; Zhang and Shao, 2018). A binary distribution model, e.g. logistic regression model, can be assumed for the incidence probability and a failure time model, e.g. Cox's regression model, can be used to model the latency (Kuk and Chen, 1992). Importantly, Sy and Taylor (2000) and Peng and Dear (2000) proposed some EM algorithms for numerical computation fo rthe maximum likelihood method for the joint estimation of the incidence and latency regression parameters in this model.

In many cancer studies, a large set of variables are collected as potential predictors, and how to efficiently select a subset of significant variables becomes an essential task for the successful implementation of the mixture cure mode in these applcations. Moreover, because of the heavy censoring data and mixture structures of cure models, how to identify the covariates that affect the cure probability and/or the survival distribution of those who are not cured respectively is theoretically tricky (Maller and Zhou, 1996; Liu and Shao, 2003). To the best of our knowledge, no rigorously justified variable selection methods have been designed for this mixture cure model.

Using the regularized procedures to simultaneously select important variables and estimate unknown parameters has received great attention in recent years. Tibshirani (1996) proposed Least Absolute Shrinkage and Selection Operator (LASSO), Fan and Li (2001, 2002) proposed Smoothly Clipped Absolute Deviation (SCAD) and Zou (2006) and Zhang and Lu (2007) proposed adaptive LASSO. However, it is well known

*Department of Biostatistics, Incyte Corporation, Wilmington, DE, 19803, USA

†Division of Biostatistics, NYU Grossman School of Medicine, New York, NY, 10016, E-mail: yongzhao.shao@nyulangone.org

that the LASSO shrinkage produces biased estimates for the large coefficients, and thus it could be sub-optimal in terms of estimation risk, while SCAD and adaptive LASSO have been proved to satisfy oracle properties (Fan and Li, 2001; Zou, 2006).

The aim of this paper is to account for a cure fraction in the global population by assuming a mixture cure model, and investigate methods for consistent variable selection based on penalized likelihood among multiple potential covariates. There has not been a rigorous study systematically investigate whether the penalized likelihood procedures will work in the mixture cure model. Importantly, Liu *and others* (2012) and Scolas *and others* (2016) provided simulation results in the semiparametric logistic/Cox model and the parametric logistic/AFT cure models, respectively. However, there is no theoretical justification about the existence, consistency and oracle properties of the maximum penalized likelihood estimators for semiparametric cure models. Therefore, a rigorous study of the asymptotic optimal properties is needed to provide theoretical foundation for variable selection in semiparametric mixture cure model.

The layout of the remainder of this paper is as follows. In Section 2, we will describe our proposed penalized likelihood-based variable selection method and a modified EM algorithm for numerical computation. Asymptotic optimal properties of our proposed methods are given in Section 3. Some simulation studies are conducted to demonstrate the performance of the proposed method in Section 4. In Section 5, we apply the proposed method to a cohort of melanoma patients at the Cancer Center at in New York University Langone Medical Center. Some discussion are given in Section 6. Technical proofs are relegated to the Supplementary Material.

2. Consistent Variable Selection via Penalized Likelihood

Let T^* denote the failure time and C denote the censoring time. The observed time is $T = \min(T^*, C)$ and the censoring indicator is $\Delta = \mathbb{I}(T^* \leq C)$. Assume that T^* and C are conditionally independent. Let u be an uncure indicator, $u = 1$ if the subject is not cured and $u = 0$ otherwise. The indicator u is partially missing because if $\delta = 1$, then $u = 1$, but if $\delta = 0$, u is not observable. Let $S_u(t|\boldsymbol{x})$ be the survival function of T^* for the uncured patients given covariate vector $\boldsymbol{x}$, and $\pi(\boldsymbol{z})$ be the probability of being uncured given a covariate vector $\boldsymbol{z}$. The survival function $S(t|\boldsymbol{x}, \boldsymbol{z})$ of an individual from the entire population is

$$S(t|\boldsymbol{x}, \boldsymbol{z}) = \pi(\boldsymbol{z})S_u(t|\boldsymbol{x}) + 1 - \pi(\boldsymbol{z}),$$

where $\pi(\boldsymbol{z}) = P(u = 1|\boldsymbol{z})$ can be modeled by a logistic regression $\pi(\boldsymbol{z}) = \frac{\exp(\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{z})}{1+\exp(\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{z})}$, and $S_u(t|\boldsymbol{x})$ can be modeled by Cox's proportional hazards (PH) regression $S_u(t|\boldsymbol{x}) = S_0(t)^{\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x})}$. $S_0(t)$ is an unspecified baseline survival function, and $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_q)^T$, $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \cdots, \gamma_p)^T$ are unknown vectors of regression coefficients. For simplicity of exposition, we consider the case $\boldsymbol{x}$ is a baseline predictor. Our proof also applies to the case where $\boldsymbol{x} = \boldsymbol{x}(t)$ is a time-dependent predictor, and the Cox PH model is a Cox model with time-dependent covariate.

Writing $\boldsymbol{\Psi} = \{\boldsymbol{\beta}, \boldsymbol{\gamma}, S_0\}$, the log observed likelihood function can be written as

$$\ell^o(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \{\delta_i \log[\pi(\boldsymbol{z_i})h_u(t_i|\boldsymbol{x_i})S_u(t_i|\boldsymbol{x_i})] + (1 - \delta_i) \log[1 - \pi(\boldsymbol{z_i}) + \pi(\boldsymbol{z_i})S_u(t_i|\boldsymbol{x_i})]\}. \quad (1)$$

And the corresponding log complete likelihood function is given by

$$\ell(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \log\left[\pi(\boldsymbol{z_i})^{u_i}\{1 - \pi(\boldsymbol{z_i})\}^{1-u_i}\right] + \sum_{i=1}^{n} u_i \log\{h_u(t_i|\boldsymbol{x_i})^{\delta_i} S_u(t_i|\boldsymbol{x_i})\}. \quad (2)$$

The penalized log complete likelihood can be formulated as

$$\ell_p(\boldsymbol{\Psi}) = \ell(\boldsymbol{\Psi}) - p_n(\boldsymbol{\Psi}), \quad (3)$$

where $p_n(\boldsymbol{\Psi}) = n\sum_{j=1}^{p} p_{\lambda_1}(|\gamma_j|) + n\sum_{k=1}^{q} p_{\lambda_2}(|\beta_k|)$; $p_{\lambda_1}(|\cdot|)$ and $p_{\lambda_2}(|\cdot|)$ are penalty functions for $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$; λ_1 and λ_2 are tuning parameters for regression coefficients $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$, respectively. The penalty function can take several forms, such as LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001, 2002), and Adaptive LASSO (Zou, 2006; Zhang and Lu, 2007).

The EM algorithm is a widely used procedure to estimate the parameters in mixture cure models without regularizations (Taylor, 1995; Sy and Taylor, 2000; Peng and Dear, 2000). Importantly, Cai *and others* (2012) has developed an R package *smcure* which can be implemented if the variables have been properly selected. Regarding the penalized likelihood methods for variable selection, Liu *and others* (2012) provided simulation results in logistic/Cox's semiparametric cure models and illustrated its usefulness. However, there is no theoretical justification about the existence, consistency of selection and oracle properties of the maximum penalized likelihood estimators. Therefore, we will provide theoretical foundation for variable selection in semiparametric mixture cure model, we will propose a modified EM algorithm as well as an publicly available R package *VSsmcure*. The penalized log complete likelihood in equation 3 can be rewritten into two parts: penalized likelihood for logistic regression and penalized partial likelihood for Cox's regression, thus, the parameters can be easily updated at each iteration. The details of our modified EM algorithm and the selection of tunning parameters can be found in the Section 7 Supplementary Material.

3. Asymptotic Optimal Properties

Note that, if we know or observe the cure status, then we can prove selection consistency for the Cox PH model part using Zhang and Lu (2007), and selection consistency of the logistic model using Zou (2006) directly. However, the cure status is only partially observed and the EM algorithm the infer the cure status involves the baseline hazard which is an infinite dimensional nuisance parameter. That makes the inference not straightforward. Prior to establish the asymptotic optimal properties, we are going to demonstrate that the EM algorithm which based on penalized complete likelihood gives the desired estimators from maximizing the penalized observed likelihood. In ordinary mixture problems, Dempster *and others* (1977) proved that the EM algorithm would increase the observed likelihood in each iteration step. Here we introduce Lemma 1 to show that the modified EM algorithm also works in penalized maximum likelihood problems.

Consider the complete data $Y = (X, Z)$ and observed data X. Let $f(y;\theta)$ be the density function of complete data Y, $f(x;\theta)$ be the density function of observed data X and $f(y|x;\theta)$ be the conditional density function of Y given X. Denote the log likelihood function of observed data x as $\ell(\theta, x)$ and the log likelihood function of complete data y as $\ell(\theta, y)$. Similarly, denote the penalized log likelihood function of observed data x as $\ell_p(\theta, x) = \ell(\theta, x) - p_\lambda(\theta)$, where λ is a tuning parameter and $p(\theta)$ is a penalty function.

Lemma 1. Suppose $\theta^{(m)}$ and $\theta^{(m+1)}$ are the estimates in mth and $(m+1)$th step of the modified EM algorithm under the above settings, then $\ell_p(\theta^{(m+1)}, \boldsymbol{x}) \geq \ell_p(\theta^{(m)}, \boldsymbol{x})$. The proof of Lemma 1 can be found in the Section 7 Supplementary Material.

Now we begin to establish the asymptotic optimal properties for our penalized likelihood estimatosr based on the observed likelihood function. Let

$$\boldsymbol{\gamma}_0 = (\gamma_{10}, \gamma_{20}, \cdots, \gamma_{p0})^{\mathrm{T}} = (\boldsymbol{\gamma}_{10}^{\mathrm{T}}, \boldsymbol{\gamma}_{20}^{\mathrm{T}})^{\mathrm{T}},$$
$$\boldsymbol{\beta}_0 = (\beta_{10}, \beta_{20}, \cdots, \beta_{q0})^{\mathrm{T}} = (\boldsymbol{\beta}_{10}^{\mathrm{T}}, \boldsymbol{\beta}_{20}^{\mathrm{T}})^{\mathrm{T}}.$$

Without loss of generality, we assume $\boldsymbol{\beta}_{20}$ and $\boldsymbol{\gamma}_{20}$ are zero coefficients, and $\boldsymbol{\beta}_{10}$ and $\boldsymbol{\gamma}_{10}$ are non-zero coefficients. For notation simplicity, we denote $\boldsymbol{\Psi}^{\mathrm{T}} = (\boldsymbol{\gamma}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}}, S_0) = (\boldsymbol{\Psi}_1^{\mathrm{T}}, \boldsymbol{\Psi}_2^{\mathrm{T}})$, where $\boldsymbol{\Psi}_1$ contains all non-zero effects and $\boldsymbol{\Psi}_2$ contains all zero effects. In addition, denote $\boldsymbol{\Psi}_0^{\mathrm{T}} = (\boldsymbol{\Psi}_{01}^{\mathrm{T}}, \boldsymbol{\Psi}_{02}^{\mathrm{T}})$ as true parameters.

The log observed likelihood function and the penalized log observed likelihood function are given as

$$\ell^o(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \left\{ \delta_i \log[\pi(\boldsymbol{z_i}) h_u(t_i|\boldsymbol{x_i}) S_u(t_i|\boldsymbol{x_i})] + (1-\delta_i) \log[1 - \pi(\boldsymbol{z_i}) + \pi(\boldsymbol{z_i}) S_u(t_i|\boldsymbol{x_i})] \right\},$$

$$\ell_p^o(\mathbf{\Psi}) = \ell^o(\mathbf{\Psi}) - p_n(\mathbf{\Psi}). \tag{4}$$

where $p_n(\mathbf{\Psi}) = n\sum_{j=1}^{p} p_{\lambda_1}(|\gamma_j|) + n\sum_{k=1}^{q} p_{\lambda_2}(|\beta_k|)$. Let

$$a_n = \max\{p'_{\lambda_{1_n}}(|\gamma_{j0}|), p'_{\lambda_{2_n}}(|\beta_{j0}|) : \gamma_{j0} \neq 0, \beta_{j0} \neq 0\},$$

$$b_n = \max\{p''_{\lambda_{1_n}}(|\gamma_{j0}|), p''_{\lambda_{2_n}}(|\beta_{j0}|) : \gamma_{j0} \neq 0, \beta_{j0} \neq 0\}.$$

Similar to Zhang and Lu (2007), the following conditions are needed on the penalty functions:

(P_1): $p_\lambda(\theta) \geq 0, p_\lambda(0) = 0$. $p_\lambda(\cdot)$ is nondecreasing and twice differentiable with at most a few exceptions.

(P_2): As $n \to \infty$, $b_n \to 0$.

(P_3): For $N_n = \left\{\theta : 0 < \theta < n^{-1/2}\log n\right\}$, $\lim\limits_{n\to\infty} \inf\limits_{\theta\in N_n} \sqrt{n}p'_\lambda(\theta) = \infty$.

Theorem 1. Suppose the observed data $O_i, i = 1, 2, \cdots, n$ are independent and identically distributed and the conditions $(P_1), (P_2)$ on penalty functions are satisfied. If $\int_0^\infty P(C > t)dH_0(t) < \infty$, then the maximum penalized likelihood estimator $\hat{\mathbf{\Psi}}_n = \arg\sup \ell_p^o(\mathbf{\Psi})$ satisfied that

$$||\hat{\mathbf{\Psi}}_n - \mathbf{\Psi}_0|| = O_p(n^{-1/2} + a_n),$$

where $||\cdot||$ represents the Euclidean norm.

Proof of Theorem 1. Let $\alpha_n = n^{-1/2} + a_n$. First, we demonstrate that with large probability, there exists a local maximum in the ball $\{\mathbf{\Psi}_0 + \alpha_n \boldsymbol{v} : ||\boldsymbol{v}|| \leq C\}$ and this local maximizer, says $\hat{\mathbf{\Psi}}_n$, satisfies $||\hat{\mathbf{\Psi}}_n - \mathbf{\Psi}_0|| = O_p(\alpha_n)$. It suffices that for any given $\epsilon > 0$, there exists a constant C such that

$$\lim_{n\to\infty} P\left\{\sup_{||\boldsymbol{v}||=C} \ell_p^o(\mathbf{\Psi}_0 + \alpha_n\boldsymbol{v}) < \ell_p^o(\mathbf{\Psi}_0)\right\} \geq 1 - \epsilon. \tag{A.1}$$

Let $D_n(\boldsymbol{v}) = \ell_p^o(\mathbf{\Psi}_0 + \alpha_n\boldsymbol{v}) - \ell_p^o(\mathbf{\Psi}_0) = [\ell^o(\mathbf{\Psi}_0 + \alpha_n\boldsymbol{v}) - \ell^o(\mathbf{\Psi}_0)] - [p_n(\mathbf{\Psi}_0 + \alpha_n\boldsymbol{v}) - p_n(\mathbf{\Psi}_0)]$. Since $p_\lambda(\theta) \geq 0, p_\lambda(0) = 0$, we have $p_n(\mathbf{\Psi}_0) = p_n(\mathbf{\Psi}_{01})$ and $p_n(\mathbf{\Psi}_0 + \alpha_n\boldsymbol{v}) \geq p_n(\mathbf{\Psi}_{01} + \alpha_n\boldsymbol{v}_1)$, where $\mathbf{\Psi}_{01}$ is the parameter vector with non-zero effects and $\boldsymbol{v}_1$ is a subvector of $\boldsymbol{v}$ with corresponding components. Thus,

$$D_n(\boldsymbol{v}) \leq [\ell^o(\mathbf{\Psi}_0 + \alpha_n\boldsymbol{v}) - \ell^o(\mathbf{\Psi}_0)] - [p_n(\mathbf{\Psi}_{01} + \alpha_n\boldsymbol{v}_1) - p_n(\mathbf{\Psi}_{01})]. \tag{A.2}$$

By Taylor expansion,

$$\ell^o(\mathbf{\Psi}_0 + \alpha_n\boldsymbol{v}) - \ell^o(\mathbf{\Psi}_0) = n^{-1/2}(1 + a_n)S(\mathbf{\Psi}_0)^{\mathrm{T}}\boldsymbol{v} - \frac{(1 + a_n)^2}{2}\boldsymbol{v}^{\mathrm{T}}I(\mathbf{\Psi}_0)\boldsymbol{v}[1 + o(p)], \tag{A.3}$$

where $S(\mathbf{\Psi}_0)$ is the score function of observed likelihood and $I(\mathbf{\Psi}_0)$ is the Fisher information matrix of observed likelihood. The explicit forms were given in Theorem 2 and Lemma 1 of FANG *and others* (2005).

By Taylor expansion and triangular inequality,

$$|p_n(\mathbf{\Psi}_{01} + \alpha_n\boldsymbol{v}_1) - p_n(\mathbf{\Psi}_{01})| \leq \sqrt{d}a_n(1 + a_n)||\boldsymbol{v}_1|| + b_n(1 + a_n)^2||\boldsymbol{v}_1||^2, \tag{A.4}$$

where $d = \max\{d_1, d_2\}$ and d_1, d_2 are the numbers of true non-zero coefficients in $\boldsymbol{\gamma}_0, \boldsymbol{\beta}_0$ respectively. The order comparison of A.3 and A.4 implies that the penalty term is negligible when n is large, and the

negative term $-\frac{(1+a_n)^2}{2}\boldsymbol{v}^{\mathrm{T}}I(\boldsymbol{\Psi}_0)\boldsymbol{v}[1+o(p)]$ is the sole leading term in A.2. Hence, for $\forall\epsilon>0$, by choosing a sufficiently large C, A.1 holds.

Additionally, by Theorem 2 and 3 of FANG *and others* (2005), when $\int_0^\infty P(C>t)dH_0(t)<\infty$, the log likelihood function $\ell^0(\boldsymbol{\Psi})$ is concave, and there exists a unique $\sqrt{n}$-consistent estimator over the whole parameter space. For adaptive LASSO, the penalty term $p_n(\boldsymbol{\Psi})$ is convex (Zou, 2006; Zhang and Lu, 2007), thus the log penalized likelihood function $\ell^0(\boldsymbol{\Psi})-p_n(\boldsymbol{\Psi})$ is concave and thus the maximizer $\hat{\boldsymbol{\Psi}}_n$ is the maximum penalized likelihood estimator. The SCAD penalty is nonconcave, similar procedure can be done following Fan and Li (2001). This completes the proof of Theorem 1.

From Theorem 1, we know that there exists a $\sqrt{n}$-consistent maximum penalized likelihood estimator. And in the following part, we will exam the oracle properties (Fan and Li, 2001) which includes sparsity and asymptotic normality.

Theorem 2. Assume the conditions given in Theorem 1 and $(P_1)-(P_3)$ on penalty functions are satisfied, we have the following:

a. Sparsity: $P(\hat{\boldsymbol{\gamma}}_{20}=\mathbf{0})\to 1$ and $P(\hat{\boldsymbol{\beta}}_{20}=\mathbf{0})\to 1$ as $n\to\infty$.

b. Asymptotic normality: Let $I_1(\boldsymbol{\Psi}_{01})$ be the Fisher information when all zero effects are removed, $\sqrt{n}\left\{[I_1(\boldsymbol{\Psi}_{01})-p_n''(\boldsymbol{\Psi}_{01})/n](\hat{\boldsymbol{\Psi}}_1-\boldsymbol{\Psi}_{01})+p_n'(\boldsymbol{\Psi}_{01})/n\right\}\to_d N(\mathbf{0},I_1(\boldsymbol{\Psi}_{01}))$.

Proof of Theorem 2: By the definition,

$$\ell_p^o\{(\boldsymbol{\Psi}_1,\boldsymbol{\Psi}_2)\}-\ell_p^o\{(\boldsymbol{\Psi}_1,\mathbf{0})\}=[\ell^o\{(\boldsymbol{\Psi}_1,\boldsymbol{\Psi}_2)\}-\ell^o\{(\boldsymbol{\Psi}_1,\mathbf{0})\}]-[p_n\{(\boldsymbol{\Psi}_1,\boldsymbol{\Psi}_2)\}-p_n\{(\boldsymbol{\Psi}_1,\mathbf{0})\}].$$

By the mean value theorem, for some $||\boldsymbol{\xi}||\le||\boldsymbol{\Psi}_2||=O(n^{-1/2})$,

$$\ell^o\{(\boldsymbol{\Psi}_1,\boldsymbol{\Psi}_2)\}-\ell^o\{(\boldsymbol{\Psi}_1,\mathbf{0})\}=\left[\frac{\partial\ell^o\{(\boldsymbol{\Psi}_1,\boldsymbol{\xi})\}}{\partial\boldsymbol{\Psi}_2}\right]^{\mathrm{T}}\boldsymbol{\Psi}_2. \tag{B.1}$$

Moreover, by the mean value theorem again,

$$\begin{aligned}
&||\frac{\partial\ell^o\{(\boldsymbol{\Psi}_1,\boldsymbol{\xi})\}}{\partial\boldsymbol{\Psi}_2}-\frac{\partial\ell^o\{(\boldsymbol{\Psi}_{01},\mathbf{0})\}}{\partial\boldsymbol{\Psi}_2}||\\
\le&||\frac{\partial\ell^o\{(\boldsymbol{\Psi}_1,\boldsymbol{\xi})\}}{\partial\boldsymbol{\Psi}_2}-\frac{\partial\ell^o\{(\boldsymbol{\Psi}_1,\mathbf{0})\}}{\partial\boldsymbol{\Psi}_2}||+||\frac{\partial\ell^o\{(\boldsymbol{\Psi}_1,\mathbf{0})\}}{\partial\boldsymbol{\Psi}_2}-\frac{\partial\ell^o\{(\boldsymbol{\Psi}_{01},\mathbf{0})\}}{\partial\boldsymbol{\Psi}_2}||\\
\le&[||\xi||+||\boldsymbol{\Psi}_1-\boldsymbol{\Psi}_{01}||]O_p(n)=O_p(n^{1/2}).
\end{aligned}$$

By the order of score function, we know $\frac{\partial\ell^o\{(\boldsymbol{\Psi}_{01},\mathbf{0})\}}{\partial\boldsymbol{\Psi}_2}=O_p(n^{1/2})$, so $\frac{\partial\ell^o\{(\boldsymbol{\Psi}_1,\boldsymbol{\xi})\}}{\partial\boldsymbol{\Psi}_2}\le O_p(n^{1/2})$ and B.1 becomes

$$\ell^o\{(\boldsymbol{\Psi}_1,\boldsymbol{\Psi}_2)\}-\ell^o\{(\boldsymbol{\Psi}_1,\mathbf{0})\}\le O_p(n^{1/2})[\sum_{i=d_1+1}^{p}|\gamma_i|+\sum_{j=d_2+1}^{q}|\beta_j|].$$

And

$$p_n\{(\boldsymbol{\Psi}_1,\boldsymbol{\Psi}_2)\}-p_n\{(\boldsymbol{\Psi}_1,\mathbf{0})\}=n\sum_{i=d_1+1}^{p}p_{\lambda_1}(|\gamma_i|)+n\sum_{j=d_2+1}^{q}p_{\lambda_2}(|\beta_j|).$$

Thus,

$$\ell_p^o\{(\boldsymbol{\Psi}_1,\boldsymbol{\Psi}_2)\}-\ell_p^o\{(\boldsymbol{\Psi}_1,\mathbf{0})\}\le\sum_{i=d_1+1}^{p}\left[O_p(n^{1/2})|\gamma_i|-np_{\lambda_1}(|\gamma_i|)\right]+\sum_{j=d_2+1}^{q}\left[O_p(n^{1/2})|\beta_j|-np_{\lambda_2}(|\beta_j|)\right].$$

By condition P_3, in a shrinkage neighborhood of 0, $O_p(n^{1/2})|\gamma_i| < np_{\lambda_1}(|\gamma_i|)$ and $O_p(n^{1/2})|\beta_j| < np_{\lambda_2}(|\beta_j|)$. Thus, we have

$$\ell_p^o(\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2) - \ell_p^o(\boldsymbol{\Psi}_1, \mathbf{0}) < 0.$$

Let $(\hat{\boldsymbol{\Psi}}_1, \mathbf{0})$ be the maximizer of the penalized likelihood function $\ell_p^o\{(\boldsymbol{\Psi}_1, \mathbf{0})\}$. We have

$$\ell_p^o(\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2) - \ell_p^o(\hat{\boldsymbol{\Psi}}_1, \mathbf{0}) = \left[\ell_p^o(\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2) - \ell_p^o(\boldsymbol{\Psi}_1, \mathbf{0})\right] + \left[\ell_p^o(\boldsymbol{\Psi}_1, \mathbf{0}) - \ell_p^o(\hat{\boldsymbol{\Psi}}_1, \mathbf{0})\right].$$

By the property of maximum likelihood estimator, $\ell_p^o(\boldsymbol{\Psi}_1, \mathbf{0}) - \ell_p^o(\hat{\boldsymbol{\Psi}}_1, \mathbf{0}) \leq 0$. Therefore, $\ell_p^o(\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2) - \ell_p^o(\hat{\boldsymbol{\Psi}}_1, \mathbf{0}) < 0$ and this implies $\hat{\boldsymbol{\gamma}}_{20} = \mathbf{0}$ and $\hat{\boldsymbol{\beta}}_{20} = \mathbf{0}$ with probability 1 as $n \to \infty$.

To prove part b, by the property of maximum penalized likelihood estimator, we have

$$\frac{\partial \ell_p^o\{(\hat{\boldsymbol{\Psi}}_1, \mathbf{0})\}}{\partial \boldsymbol{\Psi}_1} = \frac{\partial \ell^o\{(\hat{\boldsymbol{\Psi}}_1, \mathbf{0})\}}{\partial \boldsymbol{\Psi}_1} - \frac{\partial p_n\{(\hat{\boldsymbol{\Psi}}_1, \mathbf{0})\}}{\partial \boldsymbol{\Psi}_1} = 0. \tag{C.1}$$

By Taylor expansion,

$$\frac{\partial \ell^o\{(\hat{\boldsymbol{\Psi}}_1, \mathbf{0})\}}{\partial \boldsymbol{\Psi}_1} = \frac{\partial \ell^o(\boldsymbol{\Psi}_{01})}{\partial \boldsymbol{\Psi}_1} + \left[\frac{\partial^2 \ell^o(\boldsymbol{\Psi}_{01})}{\partial \boldsymbol{\Psi}_1 \partial \boldsymbol{\Psi}_1^{\mathrm{T}}} + o_p(n)\right](\hat{\boldsymbol{\Psi}}_1 - \boldsymbol{\Psi}_{01}),$$

and

$$\frac{\partial p_n\{(\hat{\boldsymbol{\Psi}}_1, \mathbf{0})\}}{\partial \boldsymbol{\Psi}_1} = p_n'(\boldsymbol{\Psi}_{01}) + [p_n''(\boldsymbol{\Psi}_{01}) + o_p(n)](\hat{\boldsymbol{\Psi}}_1 - \boldsymbol{\Psi}_{01}).$$

Then C.1 can be written as

$$\left[\frac{\partial^2 \ell^o(\boldsymbol{\Psi}_{01})}{\partial \boldsymbol{\Psi}_1 \partial \boldsymbol{\Psi}_1^{\mathrm{T}}} - p_n''(\boldsymbol{\Psi}_{01}) + o_p(n)\right](\hat{\boldsymbol{\Psi}}_1 - \boldsymbol{\Psi}_{01}) = \frac{\partial \ell^o(\boldsymbol{\Psi}_{01})}{\partial \boldsymbol{\Psi}_1} - p_n'(\boldsymbol{\Psi}_{01}).$$

Furthermore, by the asymptotic properties of score function and Fisher information matrix, we have

$$\frac{1}{n}\frac{\partial^2 \ell^o(\boldsymbol{\Psi}_{01})}{\partial \boldsymbol{\Psi}_1 \partial \boldsymbol{\Psi}_1^{\mathrm{T}}} = I_1(\boldsymbol{\Psi}_{01}) + o_p(1),$$

$$\frac{1}{\sqrt{n}}\frac{\partial \ell^o(\boldsymbol{\Psi}_{01})}{\partial \boldsymbol{\Psi}_1} \to_d N(\mathbf{0}, I_1(\boldsymbol{\Psi}_{01})).$$

By Slutsky's theorem, the asymptotic normality in Theorem 2 holds. For simplicity of exposition, we considered the case $\boldsymbol{x}$ is a baseline predictor. The above proof also applies to the case where $\boldsymbol{x} = \boldsymbol{x}(t)$ is a time-dependent predictor, and the Cox PH model is a Cox model with time-dependent covariate.

4. Simulation Studies

Let $S_u(t|\boldsymbol{x})$ be the survival function of T^* for the uncured patients given covariate vector $\boldsymbol{x}$, and $\pi(\boldsymbol{z})$ be the probability of being uncured given a covariate vector $\boldsymbol{z}$. In our simulation study, $\boldsymbol{x}$ and $\boldsymbol{z}$ are two 8-dimensional random vectors. The first 4 components of $\boldsymbol{x}$ and $\boldsymbol{z}$ are generated from an independent Bernoulli distribution with success probability 0.5. The last four components are multivariate normal random variables with mean 0 and covariance matrix $\Sigma = \rho^{|i-j|}$, in which $\rho = 0.5$. The survival time T^* is generated from a Cox's PH model with parameters $\boldsymbol{\beta} = (-0.693, 0, 1, 0, -0.5, 0.75, 0, 0)$ and covariates $\boldsymbol{x}$. Censoring time C follows an exponential distribution with mean 10 and independent of the survival time. We observe $T + \min(T^*, C)$. The partially observed cure indicators are generated from a logistic regression model with parameters $\boldsymbol{\gamma} = (0.5, 0, 0.25, -0.75, 0, 0, -0.5, 0.3, 0)$ and covariates $\boldsymbol{z}$. Under this simulation setting,

the cure rate is around 44% and the censoring rate is around 49%. 3 types of penalties, LASSO, SCAD and Adaptive LASSO will be used in the proposed penalized likelihood procedures. Tuning parameters are searched among $\{10^{-2+4k/99}; k = 0, 1, \cdots, 99\}$ and BIC will be used to select the optimal tuning parameters. For comparison, we set sample size N = 100, 300 and 500. To evaluate the performance of our proposed method, we use the commonly used notation: TNV: True Negative Value, FNV: False Negative Value and F: F measure (Powers, 2011) will be reported with 500 replications.

Table 1: Simulation Results of Variable Selection with Tuning Parameters Selected by BIC

		γ (**Cure**)			β (**Survival**)		
N	Penalty	TNV	FNV	F	TNV	FNV	F
100	LASSO	2.674	2.566	0.556	2.890	0.536	0.808
	SCAD	2.578	2.394	0.577	2.798	0.516	0.802
	ALASSO	2.888	2.406	0.593	2.992	0.524	0.819
300	LASSO	2.594	1.540	0.705	2.978	0.012	0.885
	SCAD	2.498	1.360	0.718	2.896	0.008	0.900
	ALASSO	3.270	1.718	0.728	3.312	0.022	0.918
500	LASSO	2.662	0.870	0.789	3.154	0.004	0.904
	SCAD	2.534	0.726	0.796	3.082	0.000	0.897
	ALASSO	3.516	1.080	0.834	3.688	0.014	0.961

As shown in Table 1, as the sample size increases, the true negative values increase toward the optimal number 4, the false negative values decrease toward the optimal number 0 and F measures increase towards the optimal number 1 for all the 3 types of penalties. Compared with LASSO and SCAD, Adaptive LASSO tends to have higher TNV, as well as higher FNV with respect to both γ and β. With regard to F as the measure of overall performance, LASSO and SCAD have similar results, while Adaptive LASSO consistently outperformed the other two penalties. These simulation results demonstrate the variable selection methods work well at moderate sample size.

5. Analysis of an NYU Melanoma Dataset

In this section, we analyze a cohort of prospectively-accrued, cutaneous melanoma patients at New York University Medical Center. The cohort included 1,164 patients prospectively enrolled in the Interdisciplinary Melanoma Cooperative Group (IMCG) between 2002 and 2009, with follow-up until 2013 (Wich *and others*, 2009). It is well known that many early-stage melanomas can be cured by surgery and never experience cancer recurrence. The Kaplan-Meier curves of recurrence for the study cohort is given in Figure 1. It is clear that the Kaplan-Meier curve for recurrence-free survival levels off at a value substantially greater than 0 after 5 years follow-up, which indicates that some of the patients have been cured and will not experience a recurrence after the treatments. Thus, mixture cure models are preferred than other commonly used time-to-event models which do not take the proportion of cured patients into consideration.

The semiparametric logistic-Cox mixture cure model is utilized to analyze this cohort. Demographic and clinicopathologic information were collected for all melanoma patients including the following 9 variables: age at pathological diagnosis (in years), gender (1 for male and 2 for female), histopathological features (0 for de novo melanoma and 1 for nevus-associated melanoma), primary tumor thickness (mm), ulceration status (1 for present and 0 for absent), mitosis (1 for present and 0 for absent), histological subtype (1 for Nodular and 0 for others), anatomic site (1 for Axial/Head/Neck and 2 for Extremity) and clinical stage (0

for Stage I and 1 for others).

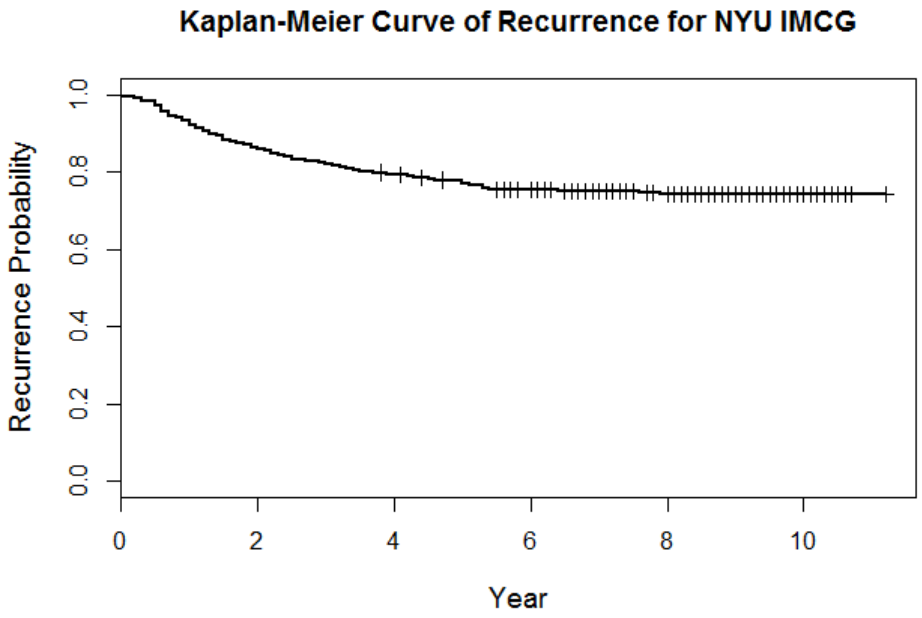

Figure 1: Kaplan-Meier Curve of Recurrence for NYU IMCG

Mixture cure model is utilized to analyze this cohort and all the 9 covariates are included as candidate risk factors for both probability of cure and time of recurrence. There are 1110 Patients with complete observations on all the covariates. The results are listed in Table 2. In each table, MLE are the maximum likelihood estimates without penalization and ALASSO are the penalized likelihood estimates with Adaptive LASSO penalty, in which the tuning parameters are selected based on BIC. We list the estimated coefficients and corresponding P-values in the full model and the final model selected by adaptive LASSO (aLasso).

Table 2: Estimation Result in NYU Interdisciplinary Melanoma Cooperative Group (IMCG)

	γ (**Cure**)				β (**Survival**)			
Covariate	MLE	P-value	aLasso	P-value	MLE	P-value	aLasso	P-value
(Intercept)	-0.895	0.147	-0.861	0.002				
Age	0.006	0.450			0.003	0.639		
Gender	-0.390	0.056			0.134	0.503		
Histopathology	-0.667	0.008	-0.734	0.009	-0.065	0.815		
Log Thickness	0.983	< 0.001	1.043	< 0.001	0.174	0.240	0.224	0.047
Ulceration	0.355	0.235	0.282	0.032	0.410	0.105	0.550	< 0.001
Anatomic Site	0.058	0.663			-0.078	0.520		
Histo Subtype	-0.001	0.996			0.103	0.625		
Mitosis	0.087	0.779			0.224	0.502		
Clinical Stage	0.659	0.056	0.766	0.018	0.255	0.402		

There are several findings in Table 2. After variable selection via penalized likelihood method with

Adaptive LASSO penalty, all the covariates remaining in the mixture cure models are statistically significant, and the results are consistent with the naive maximum likelihood estimates in general. The positive coefficients of Thickness indicate that the patients with thicker tumors are less likely to be cured, and even among the patients who have not been cured, the patients with thicker tumors have higher hazards to recurrence of melanoma. Furthermore, the patients with ulceration are less likely to be immune from recurrence and have higher hazards as well. In addition, from the model, we can see that clinical stage is a significant risk factor to predict cure status. The positive coefficient of clinical stage is consistent with the clinical findings that the patients with stage I melanoma have more than 90% survival rate (Dickson and Gershenwald, 2011) and will never experience recurrence.

6. Discussion

The classic survival analysis model is not appropriate for analyzing the failure time data in which there is a proportion of patients who could never experience the event of interest and mixture cure models are useful when there is sufficient evidence of a non-susceptible population (Maller and Zhou, 1996; Yilmaz *and others*, 2013; Sy and Taylor, 2000; Peng and Dear, 2000; Zhang and Shao, 2018). In this paper, we studied the variable selection method for semiparametric mixture cure models through different penalized likelihood approaches and provided theoretical justifications of these methods. This paper has met the demand of identification significant predictors in mixture cure models for many cancer studies and has filled the vacancy of theoretical supports for the proposed method including existence, sparsity and asymptotic normality. The proof provided here can also be used in the context when the Cox PH model in the mixture cure models is replaced by a stratified Cox PH model or by Cox models with time-dependent covariates. Time-dependent covariates are useful for predicting survival outcomes. Finally, we will write an R package to perform variable selection insemiparametric mixture cure models.

7. Supplementary Material

A modified EM algorithm. Let $O_i = \{T_i, \Delta_i, \boldsymbol{x_i}, \boldsymbol{z_i}\}$, $i = 1, 2, \cdots, n$ be the observed data for ith patient. Denote $\Psi = \{\boldsymbol{\beta}, \boldsymbol{\gamma}, S_0\}$ as the unknown parameters. The log complete likelihood function in can be written as $\ell(\boldsymbol{\beta}, \boldsymbol{\gamma}, S_0|O, \mathbf{u}) = \ell_1(\boldsymbol{\gamma}|O, \mathbf{u}) + \ell_2(\boldsymbol{\beta}, S_0|O, \mathbf{u})$, where

$$\ell_1(\boldsymbol{\gamma}|O, \mathbf{u}) = \sum_{i=1}^{n} \{u_i \log[\pi(\boldsymbol{z_i})] + (1 - u_i) \log[1 - \pi(\boldsymbol{z_i})]\},$$

$$\ell_2(\boldsymbol{\beta}, S_0|O, \mathbf{u}) = \sum_{i=1}^{n} \{u_i \delta_i \log[h_u(t_i|\boldsymbol{x_i})] + u_i \log[S_u(t_i|\boldsymbol{x_i})]\}.$$

The penalized log complete likelihood function in can be expressed as $\ell_p(\boldsymbol{\beta}, \boldsymbol{\gamma}, S_0|O, \mathbf{u}) = \ell_{p1}(\boldsymbol{\gamma}|O, \mathbf{u}) + \ell_{p2}(\boldsymbol{\beta}, S_0|O, \mathbf{u})$, where

$$\ell_{p1}(\boldsymbol{\gamma}|O, \mathbf{u}) = \sum_{i=1}^{n} \{u_i \log[\pi(\boldsymbol{z_i})] + (1 - u_i) \log[1 - \pi(\boldsymbol{z_i})]\} - n \sum_{j=1}^{p} p_{\lambda_1}(|\gamma_j|), \tag{5}$$

$$\ell_{p2}(\boldsymbol{\beta}, S_0|O, \mathbf{u}) = \sum_{i=1}^{n} \{u_i \delta_i \log[h_u(t_i|\boldsymbol{x_i})] + u_i \log[S_u(t_i|\boldsymbol{x_i})]\} - n \sum_{k=1}^{q} p_{\lambda_2}(|\beta_k|). \tag{6}$$

The E-step computes the conditional expectation of penalized log complete likelihood with respect to $\mathbf{u}$ given the observed data O and the current estimate of parameters $\Psi^{(m)} = \{\boldsymbol{\beta}^{(m)}, \boldsymbol{\gamma}^{(m)}, S_0^{(m)}\}$. Since both (5) and (6) are linear functions of $\mathbf{u}$. The expectation $w_i = E(u_i|O, \Psi^{(m)})$ can be written as

$$w_i = E(u_i|O,\Psi) = P(u_i = 1|O,\Psi) = \delta_i + (1-\delta_i)\frac{\pi(\boldsymbol{z_i})S_u(t_i|\boldsymbol{x_i})}{1-\pi(\boldsymbol{z_i})+\pi(\boldsymbol{z_i})S_u(t_i|\boldsymbol{x_i})}.$$

Therefore the E-step in the $(m+1)$th iteration is

$$E(\ell_{p1}) = \sum_{i=1}^{n} w_i^{(m)} \log[\pi(\boldsymbol{z_i})] + (1-w_i^{(m)}) \log[1-\pi(\boldsymbol{z_i})] - n\sum_{j=1}^{p} p_{\lambda_1}(|\gamma_j|), \tag{7}$$

$$\begin{aligned} E(\ell_{p2}) &= \sum_{i=1}^{n} \Big\{ \delta_i \log[w_i^{(m)} h_u(t_i|\boldsymbol{x_i})] + w_i^{(m)} \log[S_u(t_i|\boldsymbol{x_i})] \Big\} - n\sum_{k=1}^{q} p_{\lambda_2}(|\beta_k|) \\ &= \sum_{i=1}^{n} \Big\{ \delta_i \log[h_0(t_i)\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x_i} + \log w_i^{(m)})] + \log[S_0(t_i)^{\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x_i} + \log w_i^{(m)})}] \Big\} - n\sum_{k=1}^{q} p_{\lambda_2}(|\beta_k|). \end{aligned} \tag{8}$$

In order to proceed the E-step in the EM algorithm, the baseline survival can be updated by the Breslow-type estimator. Let $t_{(1)} < t_{(2)} < \cdots < t_{(N)}$be the distinct uncensored failure times, and $d_{t_{(j)}}$ be the number of events and $R_{(t_{(j)})}$ denote the risk set at time $t_{(j)}$. Then the baseline survival function $S_0(t)$ in mth iteration is given by

$$\hat{S}_0^{(m)}(t) = \exp\left(- \sum_{t_{(j)} \le t} \frac{d_{t_{(j)}}}{\sum_{i \in R_{(t_{(j)})}} w_i^{(m)} \exp(\boldsymbol{\beta^{(m)}}^{\mathrm{T}}\mathbf{x})} \right). \tag{9}$$

The M-step in the $(m+1)$th iteration is to maximize (7) and (8) to obtain $\Psi^{(m+1)}$. The attractive feature of the EM algorithm for this problem is that the two components can be maximized separately. Note that (7) is the penalized log likelihood function for ordinal logistic regression model and (8) is the penalized log likelihood for Cox proportional hazards model with the additional offset variable $\log(w_i^{(m)})$. In addition, the Breslow-type estimator for the baseline hazard h_0 will be updated at each iteration by given $w_i^{(m)}$.

In practice, we suggest that the estimates obtained from the mixture cure model without penalizations would be good initial values of parameters $\Psi^{(0)} = \{\boldsymbol{\beta}^{(0)}, \boldsymbol{\gamma}^{(0)}, S_0^{(0)}\}$ for EM algorithm. Furthermore, $\boldsymbol{\beta}^{(0)}$ and $\boldsymbol{\gamma}^{(0)}$ can be used as initial estimates in adaptive LASSO penalty as well.

The methods and algorithms to maximize the penalized likelihood function of logistic and Cox's PH models with different penalties (e.g. LASSO, SCAD, Adaptive LASSO) already existed (Fan and Li, 2001,Fan and Li, 2002,Zou, 2006,Zhang and Lu, 2007). Several R packages *glmnet* (Friedman *and others*, 2010), *Coxnet* (Simon *and others*, 2011) and *ncvreg* (Breheny and Huang, 2011) are available to solve these problems.

Tuning Parameter Selection. As is well known, the penalized likelihood procedures highly rely on tuning parameters, and the oracle properties will not be obtained if λs are not selected appropriately. Zhang *and others* (2010) found that the resulting model selected by Generalized Cross-Validation (GCV) has potential to lose efficiency, while Bayesian information criterion (BIC) is able to identify the finite-dimensional true

models consistently. In this paper, the tuning parameters λ_1 and λ_2 are selected respectively via a grid search among possible values and we take the combination with the smallest BIC.

$$\hat{\lambda} = (\hat{\lambda_1}, \hat{\lambda_2}) = \arg\min BIC(\lambda),$$

where $BIC(\lambda) = -\ell_p^o(\hat{\mathbf{\Psi}}_n) + \log(n)df_\lambda$, and df_λ is the number of non-zero coefficients in $\hat{\mathbf{\Psi}}_n$. This procedure allows λ_1 to be different from λ_2. Therefore, different sets of variables in the cure probability part and in the survival distribution part can be reached.

Proof of Lemma 1. The conditional density function of y given X can be written as

$$f(y|x;\theta) = \frac{f(y;\theta)}{f(x;\theta)}.$$

Then we have

$$\log f(y|x;\theta) = \ell(\theta, y) - \ell(\theta, x).$$

Taking conditional expectation given $X = x$ under current estimates $\theta^{(m)}$, we get

$$\begin{aligned} E_{\theta^{(m)}}[\log f(y|x;\theta)|X = x] &= E_{\theta^{(m)}}[\ell(\theta, y|X = x)] - E_{\theta^{(m)}}[\ell(\theta, x|X = x)] \\ &= E_{\theta^{(m)}}[\ell(\theta, y|X = x)] - \ell(\theta, x). \end{aligned}$$

Thus

$$\ell_p(\theta, x) = E_{\theta^{(m)}}[\ell(\theta, y|X = x)] - p_\lambda(\theta) - E_{\theta^{(m)}}[\log f(y|x;\theta)|X = x].$$

Denote $Q(\theta;\theta^{(m)}) = E_{\theta^{(m)}}[\ell(\theta, y|X = x)] - p_\lambda(\theta)$ and $H(\theta;\theta^{(m)}) = E_{\theta^{(m)}}[\log f(y|x;\theta)|X = x]$, then

$$\ell_p(\theta, x) = Q(\theta;\theta^{(m)}) - H(\theta;\theta^{(m)}).$$

In the M-step of EM algorithm, we are looking for a $\theta^{(m+1)}$ which maximizes $Q(\theta;\theta^{(m)})$, so it is obvious that

$$Q(\theta^{(m+1)};\theta^{(m)}) \geq Q(\theta^{(m)};\theta^{(m)}).$$

Furthermore,

$$\begin{aligned} H(\theta^{(m+1)};\theta^{(m)}) - H(\theta^{(m)};\theta^{(m)}) &= E_{\theta^{(m)}}\left[\log \frac{f(y|x;\theta^{(m+1)})}{f(y|x;\theta^{(m)})}|X = x\right] \\ &\leq \log E_{\theta^{(m)}}\left[\frac{f(y|x;\theta^{(m+1)})}{f(y|x;\theta^{(m)})}|X = x\right] \qquad \text{By Jensen's Inequality} \\ &= \log \int \frac{f(y|x;\theta^{(m+1)})}{f(y|x;\theta^{(m)})} f(y|x;\theta^{(m)})dy \\ &= \log \int f(y|x;\theta^{(m+1)})dy \\ &= \log 1 = 0. \end{aligned}$$

$$\begin{aligned} \ell_p(\theta^{(m+1)}, x) - \ell_p(\theta^{(m)}, x) &= [\ell(\theta^{(m+1)}, x) - p_\lambda(\theta^{(m+1)})] - [\ell(\theta^{(m)}, x) - p_\lambda(\theta^{(m)})] \\ &= [Q(\theta^{(m+1)};\theta^{(m)}) - Q(\theta^{(m)};\theta^{(m)})] - [H(\theta^{(m+1)};\theta^{(m)}) - H(\theta^{(m)};\theta^{(m)})] \\ &\geq 0. \end{aligned}$$

Acknowledgments

This research is partially supported by the NIH grants P50CA225450, P30CA016087, P30AG066512, P30ES000260, P01AG060882. The authors would like to thank Dr. Xiang Liu for useful communication. The authors also thank the IMCG investigators at NYU Langone Health for the melanoma dataset used for illustration of proposed method.

Conflict of Interest: None declared.

References

BREHENY, PATRICK AND HUANG, JIAN. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The annals of applied statistics* **5**(1), 232.

CAI, CHAO, ZOU, YUBO, PENG, YINGWEI AND ZHANG, JIAJIA. (2012). smcure: An r-package for estimating semiparametric mixture cure models. *Computer methods and programs in biomedicine* **108**(3), 1255–1260.

CHANG, GREGORY, TADEPALLI, JS, SHAO, YONGZHAO AND POLSKY, DAVID. (2016). Sensitivity of plasma brafmutant and nrasmutant cell-free dna assays to detect metastatic melanoma in patients with low recist scores and non-recist disease progression. *Molecular oncology* **10**(1), 157–165.

DEMPSTER, ARTHUR P, LAIRD, NAN M AND RUBIN, DONALD B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, 1–38.

DICKSON, PAXTON V AND GERSHENWALD, JEFFREY E. (2011). Staging and prognosis of cutaneous melanoma. *Surgical oncology clinics of North America* **20**(1), 1–17.

FAN, JIANQING AND LI, RUNZE. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association* **96**(456), 1348–1360.

FAN, JIANQING AND LI, RUNZE. (2002). Variable selection for cox's proportional hazards model and frailty model. *Annals of Statistics*, 74–99.

FANG, HONG-BIN, LI, GANG AND SUN, JIANGUO. (2005). Maximum likelihood estimation in a semiparametric logistic/proportional-hazards mixture model. *Scandinavian Journal of Statistics* **32**(1), 59–75.

FRIEDMAN, ERICA, SHULIAN, SHANG, SHAO, YONGZHAO AND OSMAN, IMAN. (2012). Serum micrornas as biomarkers for recurrence in melanoma. *Journal of Translational Medicine* **10**(155), PMCID–PMC3479021.

FRIEDMAN, JEROME, HASTIE, TREVOR AND TIBSHIRANI, ROB. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software* **33**(1), 1.

KUK, ANTHONY YC AND CHEN, CHEN-HSIN. (1992). A mixture model combining logistic regression with proportional hazards regression. *Biometrika* **79**(3), 531–541.

LIU, MENGLING, LU, WENBIN AND SHAO, YONGZHAO. (2006). Interval mapping of quantitative trait loci for time-to-event data with the proportional hazard mixture cure model. *Biometrics* **62**(4), 1053–1061.

LIU, XIANG, PENG, YINGWEI, TU, DONGSHENG AND LIANG, HUA. (2012). Variable selection in semi-parametric cure models based on penalized likelihood, with application to breast cancer clinical trials. *Statistics in medicine* **31**(24), 2882–2891.

LIU, XIN AND SHAO, YONGZHAO. (2003). Asymptotics for likelihood ratio tests under loss of identifiability. *Annals of statistics* **31**(3), 807–832.

MALLER, ROSS A AND ZHOU, XIAN. (1996). *Survival analysis with long-term survivors*, Volume 16. John Wiley & Sons.

PENG, YINGWEI AND DEAR, KEITH BG. (2000). A nonparametric mixture model for cure rate estimation. *Biometrics* **56**(1), 237–243.

POWERS, DAVID MARTIN. (2011). Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *International Journal of Machine Learning Technology* **2**(1), 37–63.

SCOLAS, SYLVIE, EL GHOUCH, ANOUAR, LEGRAND, CATHERINE AND OULHAJ, ABDERRAHIM. (2016). Variable selection in a flexible parametric mixture cure model with interval-censored data. *Statistics in medicine* **35**(7), 1210–1225.

SIMON, NOAH, FRIEDMAN, JEROME, HASTIE, TREVOR AND TIBSHIRANI, ROB. (2011). Regularization paths for coxs proportional hazards model via coordinate descent. *Journal of statistical software* **39**(5), 1.

SY, JUDY P AND TAYLOR, JEREMY MG. (2000). Estimation in a cox proportional hazards cure model. *Biometrics* **56**(1), 227–236.

TAYLOR, JEREMY MG. (1995). Semi-parametric estimation in failure time mixture models. *Biometrics*, 899–907.

TIBSHIRANI, ROBERT. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

WICH, LINDSAY G, HAMILTON, HEATHER K, SHAPIRO, RICHARD L, PAVLICK, ANNA, BERMAN, RUSSELL S, POLSKY, DAVID, GOLDBERG, JUDITH D, HERNANDO, EVA, MANGA, PRASHIELA, KROGSGAARD, MICHELLE *and others*. (2009). Developing a multidisciplinary prospective melanoma biospecimen repository to advance translational research. *Am J Transl Res* **1**(1), 35–43.

YILMAZ, YILDIZ E, LAWLESS, JERALD F, ANDRULIS, IRENE L AND BULL, SHELLEY B. (2013). Insights from mixture cure modeling of molecular markers for prognosis in breast cancer. *Journal of Clinical Oncology* **31**(16), 2047–2054.

ZHANG, HAO HELEN AND LU, WENBIN. (2007). Adaptive lasso for cox's proportional hazards model. *Biometrika* **94**(3), 691–703.

ZHANG, YIYUN, LI, RUNZE AND TSAI, CHIH-LING. (2010). Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association* **105**(489), 312–323.

ZHANG, YILONG AND SHAO, YONGZHAO. (2018). Concordance measure and discriminatory accuracy in transformation cure models. *Biostatistics* **19**(1), 14–26.

ZOU, HUI. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association* **101**(476), 1418–1429.