The Annals of Applied Statistics
2019, Vol. 13, No. 1, 444–465
https://doi.org/10.1214/18-AOAS1198

CAPTURING HETEROGENEITY OF COVARIATE EFFECTS IN HIDDEN SUBPOPULATIONS IN THE PRESENCE OF CENSORING AND LARGE NUMBER OF COVARIATES

BY FARHAD SHOKOOHI*, ABBAS KHALILI*,[1], MASOUD ASGHARIAN*,[2] AND SHILI LIN†,[3]

*McGill University** *and Ohio State University*†

The advent of modern technology has led to a surge of high-dimensional data in biology and health sciences such as genomics, epigenomics and medicine. The high-grade serous ovarian cancer (HGS-OvCa) data reported by The Cancer Genome Atlas (TCGA) Research Network is one example. The TCGA and other research groups have analyzed several aspects of these data. Here we study the relationship between Disease Free Time (DFT) after surgery among ovarian cancer patients and their DNA methylation profiles of genomic features. Such studies pose additional challenges beyond the typical big data problem due to population substructure and censoring. Despite the availability of several methods for analyzing time-to-event data with a large number of covariates but a small sample size, there is no method available to date that accommodates the additional feature of heterogeneity. To this end, we propose a regularized framework based on the finite mixture of accelerated failure time model to capture intangible heterogeneity due to population substructure and to account for censoring simultaneously. We study the properties of the proposed framework both theoretically and numerically. Our data analysis indicates the existence of heterogeneity in the HGS-OvCa data, with one component of the mixture capturing a more aggressive form of the disease, and the second component capturing a less aggressive form. In particular, the second component portrays a significant positive relationship between methylation and DFT for BRCA1. By further unearthing the negative relationship between expression and methylation for this gene, one may provide a biologically reasonable explanation that sheds light on the relationship between DNA methylation, gene expression and mutation.

1. Introduction. The advent of modern technology has led to a surge of high-dimensional data in biology and health sciences such as genomics, epigenomics

Received May 2017; revised March 2018.

[1]Supported by the Natural Sciences and Engineering Council of Canada through Discovery Grant NSERC RGPIN-2015-03805 and also by the Fonds de recherche du Québec-Nature et technologies (FQRNT 2012-146827).

[2]Supported by the Natural Sciences and Engineering Council of Canada through Discovery Grant NSERC RGPIN 217398-13.

[3]Supported in part by the National Science Foundation Grant DMS-1220772.

Key words and phrases. DNA methylation, ovarian cancer, finite mixture of AFT model, penalized regression, right censoring.

and medicine. In such applications often a large number of covariates are recorded, many of which may have no effect on a response variable, hence creating a difficult variable selection problem. In some situations, the response variable is subject to censoring, which adds an extra layer of difficulty to the variable selection problem. In addition, the underlying population may also be heterogeneous with observations potentially coming from multiple unknown subpopulations. Such heterogeneity further complicates data analysis.

Ovarian cancer is the fifth-leading cause of cancer deaths among women in the United States [The Cancer Genome Atlas Research Network (2011)]. Early-stage ovarian cancer can be treated successfully using surgery and chemotherapy. The high-grade serous ovarian cancer (HGS-OvCa) dataset [The Cancer Genome Atlas Research Network (2011)], available at the cBioPortal for Cancer Genomics website http://www.cbioportal.org [Cerami et al. (2012), Gao et al. (2013)], includes the information of DFT of ovarian cancer patients after surgery and DNA methylation of 9452 genes for 396 individuals. The relationship between DFT of ovarian cancer patients and their DNA methylation profiles of genomic features is an example of the challenges (possible heterogeneity and censoring) mentioned above.

The TCGA and other research groups have analyzed several aspects of these data [Han et al. (2016), Yang et al. (2011)]. Here we study the relationship between DFT after surgery among ovarian cancer patients and their DNA methylation profiles of genomic features. The DNA methylation profile of a genome may provide valuable information and may even be part of a genetic signature for DFT, overall survival time and several other variables after surgery. Yang et al. (2011) analyzed 316 high-grade serous ovarian cancer cases concluding that neither methylation nor mutation of the BRCA1 gene was associated with the prognosis. Bolton et al. (2012) performed a pooled analysis of 26 observational studies on the survival of women with ovarian cancer and reached a different conclusion. Their results showed that BRCA1 was a significant factor on improvement for overall survival, though they focused on mutation rather than methylation. A more recent study, on the other hand, showed that a number of genes, including BRCA1, may have different methylation profiles at different stages or in different subtypes of ovarian cancer [Koukoura et al. (2014)]. This illustrates the heterogeneous relationship between outcome variables (e.g., DFT) and methylation profiles. Our preliminary analysis of HGS-OvCa data also lends support to this hypothesis (Supplementary Figure S1 [Shokoohi et al. (2019)]). As we see from this and previous studies, inconsistency of results and observed heterogeneity are prevalent in genomic studies, which necessitates a more careful analysis where heterogeneity as well as other important aspects of the data must be taken into consideration.

Motivated by the above discussion and the fact that the issues raised are commonly seen in biomedical and genomic studies beyond the specific dataset discussed above, we propose a model that can capture heterogeneity in the population. The literature and our preliminary analysis of the data lead us to hypothesize a

mixture of survival models for DFT in the HGS-OvCa data that is subject to right censoring and potentially influenced by DNA methylation in gene promoters, a very high-dimensional problem. This leads to the problem of variable selection in finite mixture of survival models, the topic that we will address in the current paper.

Variable selection in linear and generalized linear models has been extensively studied in different settings over the past decades. Among others, Least Absolute Shrinkage and Selection Operator (LASSO) by Tibshirani (1996), Smoothly Clipped Absolute Deviation Penalty (SCAD) by Fan and Li (2001), adaptive LASSO (AdpLASSO) by Zou (2006), Smooth Integration of Counting and Absolute Deviation (SICA) by Lv and Fan (2009), and Minimax Concave Penalty (MCP) by Zhang (2010) provide modern methods for selecting variables and performing parameter estimation simultaneously. Gilbride, Allenby and Brazell (2006), Khalili and Chen (2007) and Städler, Bühlmann and van de Geer (2010) studied variable selection in finite mixture of regression (FMR) models. These methods lead to a tremendous reduction of the computational burden compared to traditional variable selection methods such as "best subset selection" and "stepwise deletion".

There is a rich literature on variable selection for lifetime models. Among others, see for example, Volinsky and Raftery (2000) for a traditional approach with modified BIC, Tibshirani (1997) on LASSO for Cox proportional hazards model, Sohn et al. (2009) for gradient lasso, Fan and Li (2002) on SCAD for Cox proportional hazards and proportional hazards frailty models, Cai et al. (2005) for multivariate failure time data with a growing number of regression coefficients, Liu et al. (2015) for multivariate varying-coefficient hazard model and Faraggi and Simon (1998), Sha, Tadesse and Vannucci (2006) and Lee, Chakraborty and Sun (2011) on the Bayesian approach for the Cox and the Accelerated Failure Time (AFT) models. The latter is among the most widely used type of parametric regression models in survival analysis [Lawless (2003), Chapter 6] and will be the model considered in this paper.

The problem of variable selection for FMR models when observations are subject to right censoring, on the other hand, has not been treated much in the literature. McLachlan and McGiffin (1994) studied maximum likelihood estimation in mixture of survival models. The only research we found on model and variable selection problems in these models was the work of Lu (2009), who adopted a two-step model selection method based on BIC and backward elimination for a mixture of normal distributions with a left censoring mechanism focusing on econometric applications. However, our simulations indicate that the BIC in the first step of Lu's method may dramatically overestimate the mixture order; see Section 3.3 for further discussion.

We propose a penalized likelihood method for variable selection in finite mixture of AFT models when observations are subject to right censoring. We believe that this work will be widely usable in many applications, especially in genomics.

Simplicity, flexibility and in particular parameter interpretability have made AFT models appealing. We study large sample properties of our proposed method theoretically. The small sample behaviour of the method has been studied using simulated data modelled after our motivating example of the ovarian cancer data.

The rest of this manuscript is organized as follows. Section 2 is devoted to modelling and methodology. In Section 2.1, we briefly review mixture of AFT models. We propose our penalized approach for variable selection in mixture of AFT models in Section 2.2. Large sample properties of the proposed approach will be studied in Section 2.3. Simulation studies are presented in Section 3 to assess the validity of the proposed approach for small to moderate size data. In Section 4, an analysis of HGS-OvCa data is provided to shed light on the relationship between DFT and genomic features of ovarian cancer. Section 5 includes some concluding remarks.

2. Model and method.

2.1. *A model to capture heterogeneity.* We propose a mixture of AFT models to capture heterogeneity in our data. Let X be time-to-event (for instance, DFT in our case) with support $\mathcal{X} \subset \mathbb{R}^+$, and $\boldsymbol{z} = (z_1, \ldots, z_d)^\top \subset \mathbb{R}^d$ be a d-dimensional vector of covariates that may have an effect on X. Let $T = \min\{Y, C\}$, where $Y = \log X$ and C is the logarithm of the censoring time. We further use δ to denote the censoring indicator. That is, $\delta = 0$ if the time is censored. Note that we do not observe X (or equivalently Y); the observed data are instead (T, δ). In what follows we assume that censoring (variable C) is noninformative and independent of the pair $(Y, \boldsymbol{z})$.

We say $V = (T, \delta, \boldsymbol{z})$ follows a finite mixture of AFT regression models of order K if the conditional density of (T, δ) given $\boldsymbol{z}$ has the form

$$
\begin{aligned}
f(t, \delta; \boldsymbol{z}, \boldsymbol{\Psi}) = \sum_{k=1}^{K} \pi_k \big[f_Y\big(t; \theta_k(\boldsymbol{z}), \sigma_k\big) \big]^{\delta} \big[S_Y\big(t; \theta_k(\boldsymbol{z}), \sigma_k\big) \big]^{1-\delta} \\
\times \big[f_C(t) \big]^{1-\delta} \big[S_C(t) \big]^{\delta},
\end{aligned} \tag{2.1}
$$

where the π_k ($0 < \pi_k < 1$, with $\sum_{k=1}^{K} \pi_k = 1$) are mixing probabilities, and $f_C(\cdot)$ and $S_C(\cdot)$ are the density and survival functions of C, respectively. On the other hand, $f_Y(\cdot)$ and $S_Y(\cdot)$ are respectively the density and survival functions of Y, in which $\theta_k(\boldsymbol{z}) = h(\beta_{0k} + \boldsymbol{z}^\top \boldsymbol{\beta}_k)$, where $h(\cdot)$ is a known link function, and β_{0k} and $\boldsymbol{\beta}_k = (\beta_{k1}, \beta_{k2}, \ldots, \beta_{kd})^\top$ are the intercept and regression coefficients, respectively, and σ_k is a dispersion parameter. It is worth noting that for each component of the mixture specified in (2.1), say the kth component,

$$
Y = \log X = \beta_{0k} + \boldsymbol{z}^\top \boldsymbol{\beta}_k + \sigma_k \varepsilon, \tag{2.2}
$$

where ε has a suitable distribution such as standard normal, extreme value, generalized extreme value or logistic. A common AFT model in survival analysis is

based on the Log-Normal distribution [Lawless (2003)] in which $\varepsilon \sim N(0,1)$. The vector of all parameters is

$$\boldsymbol{\Psi} = (\beta_{01}, \ldots, \beta_{0K}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K, \sigma_1, \ldots, \sigma_K, \pi_1, \ldots, \pi_{K-1})^\top,$$

which has length $d^* = K(d+3) - 1$, increasing with the order (K) of the mixture.

For a given sample when all the observations are failure times, that is, there is no censoring (i.e., $\delta_i = 1$), the density $f(t, \delta; \boldsymbol{z}, \boldsymbol{\Psi})$ is proportional to the density of a finite mixture of regression models [McLachlan and Peel (2004)]. Since the term $[f_C(t)]^{1-\delta}[S_C(t)]^{\delta}$ in (2.1) does not involve any model parameter, it is noninformative for parameter estimation and thus omitted from the conditional density function hereafter. Therefore (2.1) can be rewritten as

$$f(t, \delta; \boldsymbol{z}, \boldsymbol{\Psi}) \propto \sum_{k=1}^{K} \pi_k \big[f_Y\big(t; \theta_k(\boldsymbol{z}), \sigma_k\big)\big]^{\delta} \big[S_Y\big(t; \theta_k(\boldsymbol{z}), \sigma_k\big)\big]^{1-\delta}. \tag{2.3}$$

Suppose a large number of covariates $\boldsymbol{z} = (z_1, \ldots, z_d)^\top$ are recorded but most of them have no effect on the response variable Y; we do not, of course, know which ones. We therefore assume that the underlying model is sparse in the sense that most of the coefficients β_{kj} are zero. The maximum likelihood estimator (MLE) of $\boldsymbol{\Psi}$ will not be able to provide exact zero estimates for those coefficients that are truly zero particularly when the number of covariates is large compared to the sample size. As such the maximum likelihood approach does not provide a sparse solution. To address this problem, in this work, we study regularization methods for variable selection in finite mixture of AFT models.

2.2. *Penalized likelihood method.* In order to select the important variables in the model we propose to use a penalized likelihood approach, if necessary after some screening.

Let $(t_i, \delta_i, \boldsymbol{z}_i)$, $i = 1, 2, \ldots, n$, be a random sample from a finite mixture of AFT regression models, as given in (2.3). The conditional log-likelihood is given by

$$\ell_n(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k \big[f_Y\big(t_i; \theta_k(\boldsymbol{z}_i), \sigma_k\big)\big]^{\delta_i} \big[S_Y\big(t_i; \theta_k(\boldsymbol{z}_i), \sigma_k\big)\big]^{1-\delta_i}. \tag{2.4}$$

We propose to estimate $\boldsymbol{\Psi}$ by maximizing the penalized log-likelihood

$$\tilde{\ell}_n(\boldsymbol{\Psi}) = \ell_n(\boldsymbol{\Psi}) - \mathbf{p}_{\boldsymbol{\lambda}_n}(\boldsymbol{\Psi}) \tag{2.5}$$

with the penalty function being

$$\mathbf{p}_{\boldsymbol{\lambda}_n}(\boldsymbol{\Psi}) = n \sum_{k=1}^{K} \pi_k^{\alpha} \left\{ \sum_{j=1}^{d} p_{\lambda_{n,k}}(\beta_{kj}) \right\}, \tag{2.6}$$

where $p_{\lambda_{n,k}}(\beta_{kj})$ is a nonnegative and nondecreasing function in $|\beta_{kj}|$, $\boldsymbol{\lambda}_n = (\lambda_{n,1}, \ldots, \lambda_{n,K})$ is the vector of tuning parameters, and $\alpha \in [0,1]$ controls the

contribution of the mixing probabilities π_k in the penalty as suggested in Städler, Bühlmann and van de Geer (2010). The advantage of this method compared to the MLE is that the resulting fitted model is sparse, that is, those regression coefficients that are not significant or close to zero will be estimated as zero. Therefore, the variable selection and parameter estimation are combined and done simultaneously. This approach will reduce the computational burden substantially.

We consider several well-known penalties.

- AdpLASSO: $p_{\lambda_k}(\beta_{kj}) = \lambda_k \tilde{w}_{kj}|\beta_{kj}|$, where the $\tilde{w}_{kj}$ are some known weights.
- LASSO: $p_{\lambda_k}(\beta_{kj}) = \lambda_k|\beta_{kj}|$.
- SCAD: $p_{\lambda_k}(\beta_{kj}) = \lambda_k|\beta_{kj}|\mathbf{1}_{\{|\beta_{kj}|\le\lambda_k\}} + [-\frac{|\beta_{kj}|^2-2a\lambda_k|\beta_{kj}|+\lambda_k^2}{2(a-1)}]\mathbf{1}_{\{\lambda_k<|\beta_{kj}|\le a\lambda_k\}} + \frac{(a+1)\lambda_k^2}{2}\mathbf{1}_{\{|\beta_{kj}|>a\lambda_k\}}$ with $a > 2$ where $\mathbf{1}$ is indicator function; Fan and Li (2001) suggested $a = 3.7$ based on minimizing a Bayes risk.
- MCP: $p'_{\lambda_k}(\beta_{kj}) = \frac{1}{\gamma}(\gamma\lambda_k - |\beta_{kj}|)_+$, where $\gamma > 0$ and $p'(\cdot)$ is the first derivative with respect to β_{kj}. In our simulations, we set $\gamma = 4.5$ which is computed based on pairwise sample correlation of covariates as suggested by Zhang (2010).
- SICA: $p_{\lambda_k}(\beta_{kj}) = \lambda_k\frac{(a+1)|\beta_{kj}|}{(a+|\beta_{kj}|)}$, where $a > 0$ is chosen based on a data-dependent approach so as to achieve unbiasedness, sparsity and continuity. Meanwhile, this leads to a less complex optimization problem in terms of concavity as discussed by Lv and Fan (2009). They argued that to achieve unbiasedness and sparsity, a must be in $(0, \infty)$ and to achieve continuity $a \in [a_0, \infty)$ where $a_0 = \lambda + \sqrt{\lambda^2 + \lambda}$. They also argued that the optimal value of a should be chosen "to have the minimal maximum concavity, which is favourable from the computational point of view since the degree of concavity is related to the computational difficulty". Accordingly we set $a = 5$.

The maximum penalized likelihood estimator (MPLE) of $\boldsymbol{\Psi}$ is then

$$\widehat{\boldsymbol{\Psi}}_n = \arg\max_{\boldsymbol{\Psi}} \tilde{\ell}_n(\boldsymbol{\Psi}). \tag{2.7}$$

With an appropriate choice of the tuning parameter $\boldsymbol{\lambda}_n$, many of the estimated regression coefficients, especially those with no effects on the outcome variable, will be zero and hence their corresponding variables do not appear in the fitted model.

The penalties in the above list have been widely used in regression problems regardless of whether the covariates z's are correlated or not. In general, as long as the regularity condition R.5 [Shokoohi et al. (2019), Supplementary Material S2] on the positive definiteness of the information matrix holds, the MPLE has the sparsity and asymptotic normality properties stated in Theorem 1 below.

Numerical implementation of the proposed method, including a modified EM algorithm, selection for the tuning parameters ($\boldsymbol{\lambda}$), and estimation of the number of mixture components (K), are all provided in Supplementary Material S1 [Shokoohi et al. (2019)]. In the next section we focus on large sample properties of the MPLE.

2.3. *Large sample properties of the estimators*. We assume that the observed data is a sample from a sparse finite mixture of AFT regression models with K components and a corresponding true parameter vector

$$\boldsymbol{\Psi}^0 = (\beta_{01}^0, \ldots, \beta_{0K}^0, \boldsymbol{\beta}_1^0, \ldots, \boldsymbol{\beta}_K^0, \sigma_1^0, \ldots, \sigma_K^0, \pi_1^0, \ldots, \pi_{K-1}^0)^\top.$$

We further assume that $\boldsymbol{\Psi}^0$ is an interior point of the parameter space $\boldsymbol{\Theta} \subset \mathbb{R}^{d^*}$ where $d^* = K(d+3) - 1$. In the sparse model many of the true regression coefficients β_{kj}^0 are zero, although we do not know which ones. We aim to study the so-called oracle property of the MPLE of $\boldsymbol{\Psi}^0$ as defined in Fan and Li (2001).

Without loss of generality, for each $1 \le k \le K$, consider the partitioning $\boldsymbol{\beta}_k^0 = (\boldsymbol{\beta}_{k1}^0, \boldsymbol{\beta}_{k2}^0)$ such that $\boldsymbol{\beta}_{k1}^0$ contains the true nonzero regression coefficients and $\boldsymbol{\beta}_{k2}^0 = \mathbf{0}$. We assume that $\dim(\boldsymbol{\beta}_{k1}^0) = d_1$. Naturally, we partition the true parameter vector as $\boldsymbol{\Psi}^0 = (\boldsymbol{\Psi}_1^0, \boldsymbol{\Psi}_2^0)$ so that $\boldsymbol{\Psi}_1^0$ contains the nonzero vectors $\boldsymbol{\beta}_{k1}^0$, and $\boldsymbol{\Psi}_2^0 = \mathbf{0}$. The intercepts β_{0k}^0, variances σ_k^0 and the mixing probabilities π_k^0 are also included in $\boldsymbol{\Psi}_1^0$. A similar partitioning is considered for any candidate parameter vector $\boldsymbol{\Psi} = (\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2) \in \boldsymbol{\Theta}$. More specifically, we assess the performance of the estimator $\widehat{\boldsymbol{\Psi}}_n$ through its sub-vectors $\widehat{\boldsymbol{\Psi}}_{n1}$ and $\widehat{\boldsymbol{\Psi}}_{n2}$, where $\dim(\widehat{\boldsymbol{\Psi}}_{n1}) = \dim(\boldsymbol{\Psi}_1^0) = d_1^*$ and $\dim(\widehat{\boldsymbol{\Psi}}_{n2}) = \dim(\boldsymbol{\Psi}_2^0) = d^* - d_1^*$, such that an oracle would know d_1^*. We investigate the following properties for the estimator $\widehat{\boldsymbol{\Psi}}_n$. As $n \to \infty$,

(i) Sparsity: $\widehat{\boldsymbol{\Psi}}_{n2} = \mathbf{0}$, with probability tending to 1.

(ii) Asymptotic normality: $\sqrt{n}(\widehat{\boldsymbol{\Psi}}_{n1} - \boldsymbol{\Psi}_1^0)$ has the same asymptotic normal distribution as the oracle estimator which estimates $\boldsymbol{\Psi}_1^0$ with the knowledge that $\dim(\boldsymbol{\Psi}_1^0) = d_1^*$ and $\boldsymbol{\Psi}_2^0 = \mathbf{0}$.

For the finite mixture of AFT regression models considered in the current paper, under the regularity conditions R.1–R.6 given in Supplementary Material S2 and Lemma S1 and Lemma S2 as established in Supplementary Material S3 [Shokoohi et al. (2019)], the MPLE $\widehat{\boldsymbol{\Psi}}_n$ in (2.7) has the oracle property described by (i) and (ii). These results particularly apply to the MPLE obtained using penalties such as the SCAD, MCP, SICA and partially for the LASSO. The proofs are aligned with those in Khalili and Chen (2007) and thus omitted. Establishing similar results for the AdpLASSO, a data-adaptive penalty, requires a different proof presented below. Consider the penalized log-likelihood

$$\tilde{\ell}_n(\boldsymbol{\Psi}) = \ell_n(\boldsymbol{\Psi}) - n \sum_{k=1}^{K} \pi_k^{\alpha} \sum_{j=1}^{d} \lambda_{n,k}(\tilde{w}_{kj}|\beta_{kj}|), \tag{2.8}$$

where the weights $\tilde{w}_{kj}$ are chosen as

$$\tilde{w}_{kj} = |\tilde{\beta}_{kj}|^{-\gamma} \tag{2.9}$$

for any $\sqrt{n}$-consistent estimator $\tilde{\beta}_{kj}$ of β_{kj}^0, and some constant $\gamma > 0$. The AdpLASSO-based MPLE is then

$$\widehat{\boldsymbol{\Psi}}_n = \arg\max_{\boldsymbol{\Psi}} \tilde{\ell}_n(\boldsymbol{\Psi}). \tag{2.10}$$

We now state our main result.

THEOREM 1. *Let $V_i = (T_i, \delta_i, \mathbf{Z}_i)$, $i = 1, 2, \ldots, n$, be a random sample from a finite mixture of AFT regression model in (2.3) that satisfies conditions* R.1–R.6 *in Supplementary Material* S2 [*Shokoohi et al.* (2019)]. *As $n \to \infty$,*

(a) *if $\sqrt{n}\lambda_{n,k} \to 0$, there exists a local maximizer $\widehat{\boldsymbol{\Psi}}_n$ of $\tilde{\ell}_n(\boldsymbol{\Psi})$ such that*

$$\|\widehat{\boldsymbol{\Psi}}_n - \boldsymbol{\Psi}^0\| = O_p(n^{-1/2}).$$

(b) (*oracle property*) *for any $\sqrt{n}$-consistent estimator $\widehat{\boldsymbol{\Psi}}_n = (\widehat{\boldsymbol{\Psi}}_{n1}, \widehat{\boldsymbol{\Psi}}_{n2})$ of $\boldsymbol{\Psi}^0$, if $\sqrt{n}\lambda_{n,k} \to 0$ and $n^{(\gamma+1)/2}\lambda_{n,k} \to \infty$,*

(i) *Sparsity*: $P(\widehat{\boldsymbol{\Psi}}_{n2} = \mathbf{0}) \to 1$.
(ii) *Asymptotic normality*:

$$\sqrt{n}\big(\widehat{\boldsymbol{\Psi}}_{n1} - \boldsymbol{\Psi}_1^0\big) \to_d N\big(\mathbf{0}, \Im_1^{-1}(\boldsymbol{\Psi}_1^0)\big),$$

where $\Im_1(\boldsymbol{\Psi}_1^0)$ is the Fisher information matrix under the true model where all covariates with zero effects are removed.

PROOF. See Supplementary Material S4 [Shokoohi et al. (2019)]. □

According to Theorem 1, asymptotic properties of the MPLE in (2.10) depend on the choices of $\tilde{w}_{ik}$, γ and $\lambda_{n,k}$. The MLE can be a good candidate for $\tilde{\beta}_{kj}$ in (2.9). Having specified γ, if we choose $\lambda_{n,k} \propto n^{-1/2-\zeta}$, for a $0 < \zeta < \gamma/2$, then they satisfy the conditions required by the theorem. A common choice of γ is 1, hence $0 < \zeta < 1/2$. Note that the $\sqrt{n}$-consistency of $\tilde{\beta}_{kj}$ used in $\tilde{w}_{ik}$ can be weakened to just having consistent estimators [Zou (2006)].

3. Simulation studies. In this section, we assess the performance of the proposed methods via simulations. All the results are obtained using our open-source `fmrs` package implemented in R which is available at https://CRAN.R-project.org/package=fmrs.

3.1. *Scenarios and settings.* We considered mixtures of Log-Normal AFT models, where $f_Y(\cdot) = \phi_Y(\cdot)$ and $S_Y(\cdot) = \Phi_Y(\cdot)$ in (2.3) are respectively the probability density and cumulative distribution of Normal with mean $\theta_k(\boldsymbol{z}) = \beta_{0k} + \boldsymbol{z}^\top \boldsymbol{\beta}_k$ and variance σ_k^2. We consider mixture of AFT models of orders $K = 2, 3$ with intercepts, dispersion parameters and mixing probabilities chosen

as in Table S1 in Supplementary Material S5 [Shokoohi et al. (2019)]. For this mixture model, the signal-to-noise (SNR) ratio is computed as

$$\mathrm{SNR} = \frac{\sum_{k=1}^{K} \pi_k \boldsymbol{\beta}_k^\top \Sigma \boldsymbol{\beta}_k}{\sum_{k=1}^{K} \pi_k \sigma_k^2},$$

where Σ is the covariance matrix of the vector of covariates $\mathbf{z}$, which is assumed to follow a multivariate Normal distribution. For the parameter settings given in Supplementary Table S1 [Shokoohi et al. (2019)], the SNR varies between 5 and 10.

Specifically let n be the sample size. The vectors of covariates $\boldsymbol{z}_i$, $i = 1, \ldots, n$, are generated from a d-multivariate Normal distribution with mean $\mathbf{0}$ and a variance-covariance matrix Σ whose (l, m)th element is $\rho^{|l-m|}$ for $\rho = 0.5$ and 0.75. Given the sparse vector $\boldsymbol{\beta}_k$ of the regression coefficients in Supplementary Table S1 [Shokoohi et al. (2019)], the pair-wise correlation between the signal and noise covariates varies from (close to) zero to 0.75. Given (n, d), the generated design matrix $(\boldsymbol{z}_i,\ i = 1, \ldots, n)$ is kept fixed throughout the simulations. We consider uniform censoring on $(0, U_M)$, where U_M is a pre-specified value chosen in such a way that the average censoring percentage is equal to 10, 30, 40 and 60 for each of the sample sizes $n = 200$, 300 and 700. Given $(K, d, n, \rho, \boldsymbol{\pi}, \boldsymbol{\sigma}, \beta_{01}, \boldsymbol{\beta}_1, \ldots, \beta_{0K}, \boldsymbol{\beta}_K)$ and the design matrix, the data (t_i, δ_i), $i = 1, 2, \ldots, n$, are generated from the mixture of AFT regression models (2.3) as follows. For instance for $K = 2$,

(i) we generate u_i from Uniform(0, 1), $i = 1, \ldots, n$. If $u_i \le \pi_1$ (1st element of $\boldsymbol{\pi}$), then y_i is generated from $N(\beta_{10} + \boldsymbol{z}_i^\top \boldsymbol{\beta}_1, \sigma_1^2)$; otherwise y_i is generated from $N(\beta_{20} + \boldsymbol{z}_i^\top \boldsymbol{\beta}_2, \sigma_2^2)$.

(ii) the censoring time is generated as $c_i \sim \log[\mathrm{Uniform}(0, U_M)]$ with U_M pre-specified to achieve the desired level of right censoring. We then set $t_i = \min\{y_i, c_i\}$ and $\delta_i = I_{\{y_i < c_i\}}$.

In our simulations for each scenario and setting, we generated $R = 500$ data sets as described above. Given a mixture order K, for each simulated sample, we use the BIC_k given in Supplementary Section S1.A4 [Shokoohi et al. (2019)], for tuning parameter (λ_k) selection. Computationally, each set of simulation ($R = 500$ replicates) took between 22 seconds (for a balanced 2-component mixture of AFT model with 200 samples and 10 covariates) and 96 minutes (for an unbalanced 3-component mixture of AFT model with 700 samples and 50 covariates) to be analyzed on a MacBook Pro 15-inch with 2.5 GHz Intel Core i7 Processor and 16 GB DDR3 Memory, making it possible to carry out extensive simulation studies.

3.2. *Results*. In the discussion below, let CIZ = # Correctly Identified Zero, CIN = # Correctly Identified Nonzero, IIZ = # Incorrectly Identified Zero and IIN = # Incorrectly Identified Nonzero regression coefficients. The sensitivity (SE)

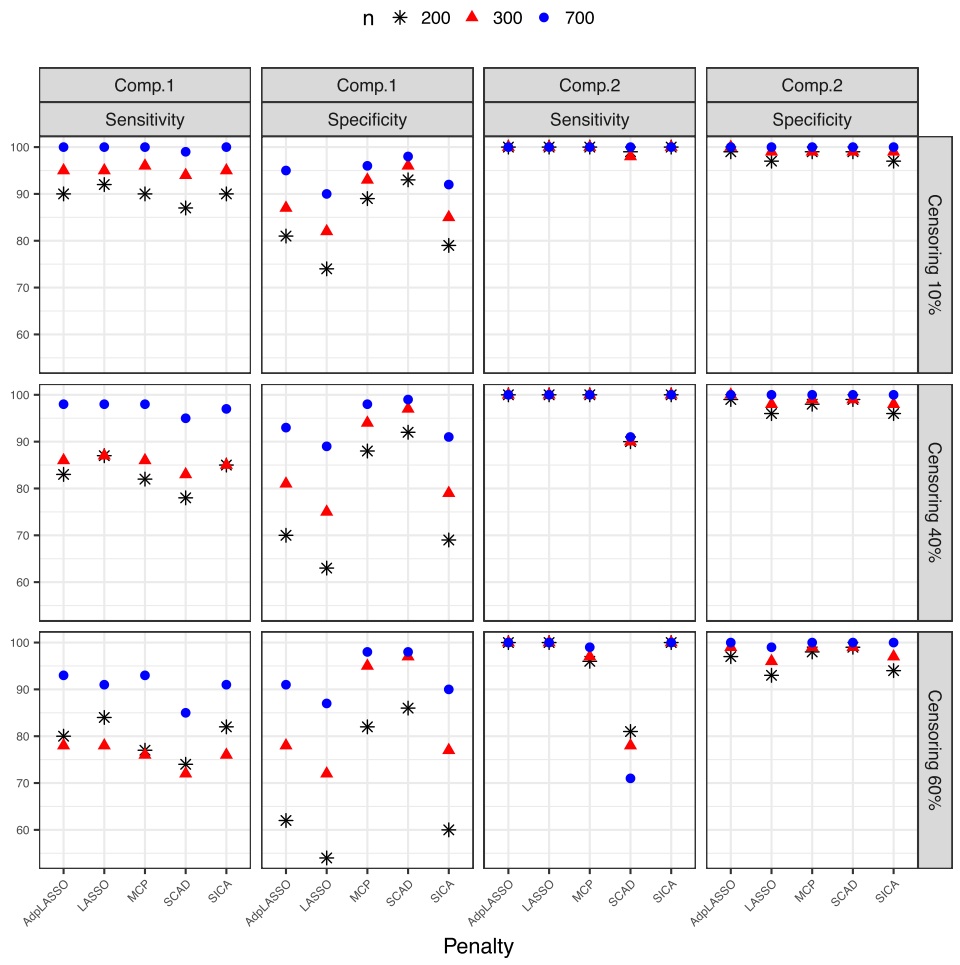


FIG. 1. *Average specificity (SP) and sensitivity (SE) for an unbalanced mixture with two components for three different censoring levels and three sample sizes n.*

and specificity (SP) are respectively defined as $\mathrm{SE} = \frac{\mathrm{CIN}}{\mathrm{CIN}+\mathrm{IIZ}}$ and $\mathrm{SP} = \frac{\mathrm{CIZ}}{\mathrm{CIZ}+\mathrm{IIN}}$. Also we denote component-wise L_2-loss of the regression coefficient estimates as

$$L_{2k} = \sqrt{\sum_{j=1}^{d} (\hat{\beta}_{kj} - \beta_{kj})^2}, \qquad k = 1, \ldots, K.$$

Figure 1 compares the average SE and SP for the LASSO, AdpLASSO, MCP, SCAD and SICA penalties over $R = 500$ simulated data sets for one of the scenarios described above. The corresponding numerical values with standard deviations among 500 replicates are given in Supplementary Table S2 [Shokoohi et al. (2019)]. The results for the other scenarios, which show similar trends as that in

Figure 1, are presented in Supplementary Material S5 [Shokoohi et al. (2019), Tables S3–S13].

The results presented in Figure 1 and those from the tables in Supplementary Material S5 [Shokoohi et al. (2019)] lead to the following general observations. Setting $\alpha = 0$, that is, removing π_k from the penalty in (2.6), improves the results of AdpLASSO, LASSO and SICA but worsens the performance of MCP and SCAD [Shokoohi et al. (2019), Table S3 and S4]. In some scenarios, for instance Table S13 [Shokoohi et al. (2019)], which correspond to a 3-components mixture ($K = 3$), we observe low SE. Given the small sample size and heavy censoring in the components of the mixture of AFT model, such performance is of course expected. Ignoring censoring in analyzing the data causes a considerable loss of performance in terms of SE and SP even in the presence of a moderate censoring, as can be seen from Table S3 and S5, where Table S5 contains the results for the same scenario as in Table S3 but with censoring ignored. Furthermore, when the correlation factor ρ increases from 0.5 to 0.75, the performance of the penalization method SP and SE declines, mainly in the smaller component of the mixture model. Finally, our study shows that each penalty has its unique merits; none of them can universally dominate its competitors.

As some methods perform better in terms of SE while at the same time worse in terms of SP, we further use the ROC curve to combine SE and SP into a single comparable statistics, the area under the curve (AUC), to compare the performance of different penalty functions in the variable selection problem. We consider a scenario with $K = 2$, $d = 50$ and $\rho = 0.75$. Figure S2 in Supplementary Material S6 [Shokoohi et al. (2019)] depicts the upper left part of ROC curves for different penalties under different sample sizes and censoring percentages. From this figure we observe that the AdpLASSO is the dominant method in many scenarios and has consistent behaviour. The other penalties also perform well in terms of AUC, but they show inconsistent behaviour. The AUCs were greater than 90%.

Since the method with all penalties perform well under the settings in Table S1, we further carried out an additional simulation to investigate the performance of the method in a setting in which the SNR was set to be approximately 1, a 5-fold or more reduction of the SNR. The results are presented in Table S14. This setting differs from the one whose results are presented in Table S2 only in the coefficients of covariates and the variances of the mixture components. By comparing the results in Table S2 with those in Table S14, we can see that overall the performance loss of the method (in terms of the averages sensitivity and specificity) for the setting with a much smaller SNR is approximately between 2 to 20%. Therefore, as expected, the weaker the signal-to-noise ratio, the harder the task of variable selection. Nevertheless, it is remarkable to see that even a 5–10 fold reduction in SNR still leads to considerable sensitivity and specificity of the method.

With the good performance of the method using various penalties documented above, one obvious question is whether a conventional, nonregularized method would perform equally well. To investigate this, we compared the component-wise

L_2-loss of MLE of the regression coefficients using the procedure of McLachlan and McGiffin (1994) with those of MPLE from our method considering various penalties. The results are presented in Tables S15 and S16 in Supplementary Material S5 [Shokoohi et al. (2019)] for $d = 10$ and 50, respectively. Overall, the L_2-loss of the MLE is larger than that of the MPLE (among different penalties) even with a small number of covariates. As the number of covariates increases, the performance of the MLE compared to the MPLE deteriorates dramatically.

We also examined the performance of the variable selection method under model mis-specification. We simulated $R = 500$ data sets from a finite mixture of Weibull AFT models and then fitted a mixture of Log-Normal AFT models to the data. Although Weibull and Log-Normal can be similar depending on the parameters, we have verified that the simulating models are quite different from the fitted models for our simulation settings. The results are presented in Supplementary Table S17 [Shokoohi et al. (2019)], which shows that all the penalties are reasonably robust to model mis-specification. In the worst case scenario the SE and SP are down by at most 15%. As the sample size increases the loss decreases rapidly.

3.3. *Order selection.* For the results presented above, we assumed that the order K of a mixture model is known a priori. In practice, however, the order usually needs to be estimated. Information criteria such as BIC have been studied thoroughly for order selection in many situations. In this section, we study the performance of the BIC*, described in Supplementary Material S1.A5 [Shokoohi et al. (2019)], for selecting the order of a finite mixture of AFT models when observations are subject to right censoring. The simulation set-up is as follows.

We generated $R = 500$ right-censored data sets from a mixture of Log-Normal AFT models with $K = 2$ as described in the previous section. We then fit a mixture of Log-Normal AFT models of orders $K = 1, 2, \ldots, 5$, to each simulated data set, and choose the optimal $\widehat{K}$ as the one that minimizes the BIC*. We also considered using AIC* as the criterion, but its performance was seen to be worse than BIC* and was thus not pursued.

Figure 2 shows the average proportion of times that the true order $K = 2$ is selected by the BIC* computed based on the MPLE using different penalties. More details are available in Supplementary Table S18 [Shokoohi et al. (2019)], which contains the average proportions of times that an order $K = 1, \ldots, 5$, is selected by BIC*. Table S19 in Supplementary Material S5 [Shokoohi et al. (2019)] presents SE and SP when the order of the mixture of AFT model is selected. These results are based on matching the closest (in terms of $\hat{\beta}_{kj}$) estimated components to the true components when the order of the mixture is overestimated. Note that the scenario of underestimation was never seen in this set of simulations.

From Figure 2, we observe that depending on the sample size, censoring level and the choice of penalty, the BIC* selects the true order of the mixture of AFT models between 60% to 96% of the times. More specifically, the BIC* based on

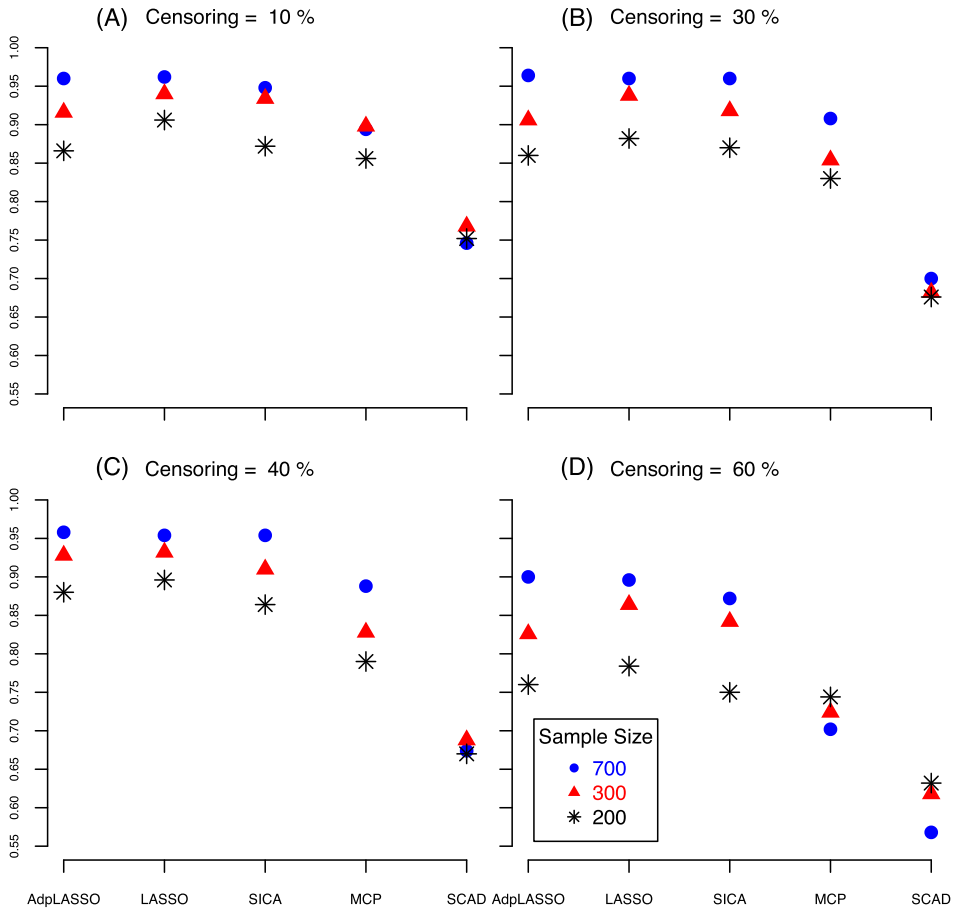


FIG. 2. *Comparison of performance of* BIC* *based on different penalties in selecting the true order* ($K = 2$) *of the underlying mixture of AFT model. The* BIC* *based on AdpLASSO, LASSO and SICA outperforms the* BIC* *based on SCAD and MCP. The performance is consistent with low* (A) *and moderate* (B & C) *censoring. Both SCAD and MCP show inconsistent behaviour when there is heavy censoring* (D).

AdpLASSO, LASSO and SICA outperforms the BIC* based on SCAD and MCP. The latter overestimated the order of the mixture of AFT models in most of the scenarios (Table S20). One possible explanation based on our simulations is that SCAD and MCP resulted in more sparse models (higher SP) with lower SE (see, e.g., Table S21 in Supplementary Material S5 [Shokoohi et al. (2019)]), thus forcing the BIC* to add extra component(s) to the mixture of AFT model to account for the lack of fit. As a comparison, included in Table S20 are also order selection results by the BIC computed using the MLE as described in Lu (2009). It is clearly seen that our method outperforms the comparison method by a large margin, espe-

cially when the sample size is smaller, as the MLE-based BIC often overestimates the mixture order.

4. Analysis of HGS-OvCa data. As discussed earlier, the main motivation for our work is to understand the relationship between DFT of ovarian cancer patients and their DNA methylation profiles. The two main challenges are heterogeneity caused by hidden subpopulations and right censuring of DFT. We use the methodology developed in the previous section to analyze the TCGA HGS-OvCa dataset described in Section 1. Specifically, this dataset contains nonmissing DFT data for 396 patients after surgery, in which 28% were right censored (i.e., patients remained disease free). The median of DFT was 14 months (range: 0–180 months), with 49% remaining alive at the time of the last follow-up. For this group of patients, methylation levels for 9452 gene promoters were also available. The correlations between these genes (before screening) range from -0.90 to 0.96. In addition, there are also other covariates, including age at diagnosis, stage and tumor grade [The Cancer Genome Atlas Research Network (2011)].

One of the questions of interest, especially for prognosis, is how methylation levels of gene promoters and other variables are related to DFT after surgery. Although several studies have analyzed this same dataset already, their conclusions were inconsistent or different from each other. One possible explanation is that disease and patient heterogeneity were not taken into account in the previous works.

Our own preliminary analysis of the response variable (DFT) shows that the log of response variable can be modelled using a two-components mixture [Shokoohi et al. (2019), Figure S1 in Supplementary Material S6], leading to a mixture model hypothesis.

We apply a log-transformation on all covariates. The minimum observed nonzero methylation level was added to avoid taking the logarithm of zero, if needed. The transformed values were then centred. A single-variable screening scheme led to the selection of 18 genes whose methylation levels at the promoter regions had the greatest influence on DFT based on our data. We further included the BRCA1 and BRCA2 genes based on information from the literature. The correlation between these 20 selected genes ranges from -0.34 to 0.46. These 20 genes are listed in segment (A) of Table 1.

We then used the regularization method of Section 2.2 and fitted mixtures of Log-Normal and Weibull AFT models each of orders $K = 1, 2, 3, 4$, to the data. Our results (with the optimal penalty parameters selected for each method) show that a two-component Log-Normal mixture model based on SCAD and a penalty on the variances of the components of the mixture [Chen and Tan (2009)] minimized the BIC. The BIC results are given in Table S23 in Supplementary Material S8 [Shokoohi et al. (2019)]. Based on these results and those provided in Supplementary Material S7 [Shokoohi et al. (2019)], the Log-Normal model with 2 components has the best fit.

TABLE 1

The fitted mixture of AFT models with 2 *components*; *HGS-OvCa data. The active genes and the corresponding nonzero estimated coefficients* (A) *and other parameters* (B) *are presented*

(A) Active genes & estimated coefficients

Gene	**BRCA1**	BRCA2	NPPA	**GHSR**	TUBB6
Component 1	2.5×10^{-13}	1.6×10^{-14}	3.8×10^{-14}	9.5×10^{-14}	2.2×10^{-5}
Component 2	**0.51**	7.3×10^{-13}	-2.6×10^{-13}	**0.48**	1.6×10^{-12}

Gene	IL1R2	C11orf9	ALOX12	**MMRN1**	**ASIP**
Component 1	4.0×10^{-13}	7.1×10^{-14}	1.4×10^{-13}	-9.6×10^{-14}	**0.74**
Component 2	8.3×10^{-13}	-1.4×10^{-12}	4.5×10^{-13}	**−2.18**	8.4×10^{-15}

Gene	**TWSG1**	HOXC11	KL	**NFRKB**	PPYR1
Component 1	3.8×10^{-13}	1.2×10^{-9}	2.5×10^{-4}	**1.5**	4.9×10^{-13}
Component 2	**−1.22**	-2.5×10^{-12}	6.2×10^{-12}	-5.8×10^{-15}	1.2×10^{-12}

Gene	**LCN2**	**SNUPN**	PTEN	LGALS1	**SFTPD**
Component 1	1.7×10^{-14}	-2.4×10^{-14}	4.9×10^{-13}	-3.5×10^{-14}	5.2×10^{-14}
Component 2	**1.46**	**−0.88**	-4.2×10^{-13}	1.4×10^{-9}	**0.51**

(B) Other parameters

Parameter	**Intercept**	σ	π	λ^{SCAD}
Component 1	2.70	0.50	0.67	0.07
Component 2	3.62	0.79	0.33	0.11

LogLik: −345.01, BIC: 743.86.

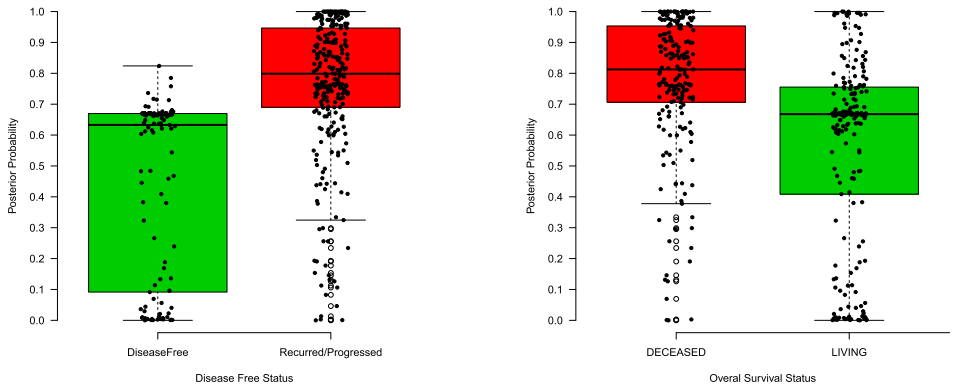
Under this selected model, we can see that the promoter methylation levels of genes ASIP and NFRKB are active in Component 1 while genes BRCA1, GHSR, MMRN1, TWSG1, LCN2, SNUPN and SFTPD are active in Component 2 (Table 1). The mixing proportions for Components 1 and 2 are 67% an 33%, respectively, with Component 1 having a smaller dispersion (sd = 0.5) compared to that of Component 2 (sd = 0.79). The full results are given in Supplementary Material S7 [Shokoohi et al. (2019)]. Our bioinformatics analysis reveals that all genes identified in either component have been previously studied, and several of them have been implicated in cancer and other diseases. We have summarized the genes that have been implicated in various types of cancer (BRCA1, ASIP and SFTPD)

TABLE 2
Cancer-associated genes identified in the HGS-OvCa data analysis

Gene	Information	Summary
ASIP	Full name	Agouti-Signaling Protein; Protein Coding gene
	Cancer	Associated with nonmelanoma skin cancers. Single nucleotide polymorphisms in ASIP is associated with basal cell carcinoma (cancer). An important role for exon 17b of ASIP in cancer cells was identified; alternative splicing isoforms, hASIP-sa, hASIP-sb, had different effects on cell growth and Fas/FasL-mediated apoptosis in BEL-7404 human hepatoma cells.
BRCA1	Full name	BRCA1, DNA Repair Associated; Protein Coding gene
	Cancer	Breast-Ovarian Cancer (Familial), and Pancreatic Cancer
SFTPD	Full name	Surfactant Protein D; Protein Coding gene
	Cancer	chest wall lymphoma cancer

in Table 2. Full details, including the genes that have been identified to be involved in other disorders in the literature (NFRKB, GHSR, MMRN1 and LCN2) are provided in Table S24 of Supplementary Material S9 [Shokoohi et al. (2019)].

We computed the posterior probability that an observation belongs to Component 1. Stratified according to several clinical variables (Disease Free Status, Overall Survival Status, Tumor Stage, Grade and Platinum Status), we plotted these probabilities as a box plot for each stratum within each variable. As we can see from Figure 3(A), those who were disease free tended to have a smaller proba-



(A) *Disease Free Time* (B) *Overall Survival Status*

FIG. 3. *Posterior probability versus clinical variables. On average, patients who have a better prognosis, for instance those who were disease free* (A) *and those who were still living* (B), *have a smaller probability of belonging to Component* 1.

bility of belonging to Component 1. From Figure 3(B), we observed a smaller average probability of belonging to Component 1 for those who were still living under the clinical variable Overall Survival Status. Similar observations are made from the other clinical variables: on average, those who have a better prognosis have a smaller probability of belonging to Component 1 [Shokoohi et al. (2019), Figures S3–S5 in Supplementary Material S6]. Taken together, the results seem to indicate that Component 2 captures the less aggressive form of the disease whereas Component 1 is the opposite.

It is interesting to see that the active genes for Components 1 and 2 do not overlap. This suggests that the genes selected based on their methylation level may be further studied as potential biomarkers. In particular, since hypermethylation may lead to downregulation [Earp and Cunningham (2015), Kong et al. (2015), Kwon and Shin (2011), Kwong et al. (2006), Schöndorf et al. (2016)], it is of interest to investigate the relationship between gene expression and DNA methylation for the HSG-OvCa dataset. Gene expression measures are available for eight of the nine genes that were deemed to significantly influence the response (expression values for gene ASIP is not available). We computed and tested Kendall's tau correlation for each of the two components. For Component 1, significant negative correlations between expression and methylation were detected for genes BRCA1, LCN2, SFTPD and TWSG1, taking multiple testing into consideration. The scatterplots with nominal p-values are provided in Figures S6–S13 in Supplementary Material S6 [Shokoohi et al. (2019)].

For Component 2, significant negative correlations were detected only for genes LCN2 and SFTPD. For the rest of the genes, the relationships are inconclusive. This is not completely surprising given that a component composed of patients having the less aggressive form of the disease may have hidden heterogeneity. As such, it is plausible that there exists a more homogeneous subset of individuals in Component 2 for which there is a clear relationship between the gene expression and methylation. To investigate this, we employed a method that is capable of detecting positive or negative association that exists only in a subsample coming from an underlying cryptic heterogeneous population. Since the subset, if it does exist, is unknown a priori, we used the tau-path method with the CEMCtp algorithm [Zhang, Ding and Lin (2017)] to probe the presence of such a hidden relationship. Using this method, we confirmed that the two genes that were found to have significant negative correlations with the full set analysis once again showed a significant negative correlation with almost all the observations contributing to the results [Shokoohi et al. (2019), Figures S14 and S15 in Supplementary Material S6]. More interestingly, of the other genes that were found not to be significant for the full set of observations, the expression and DNA methylation of genes BRCA1 and GHSR are now found to be significantly negatively correlated (see Figure 4 for BRCA1 and Supplementary Figure S16 for GHSR [Shokoohi et al. (2019)]) for a subset of observations only, suggesting further heterogeneity

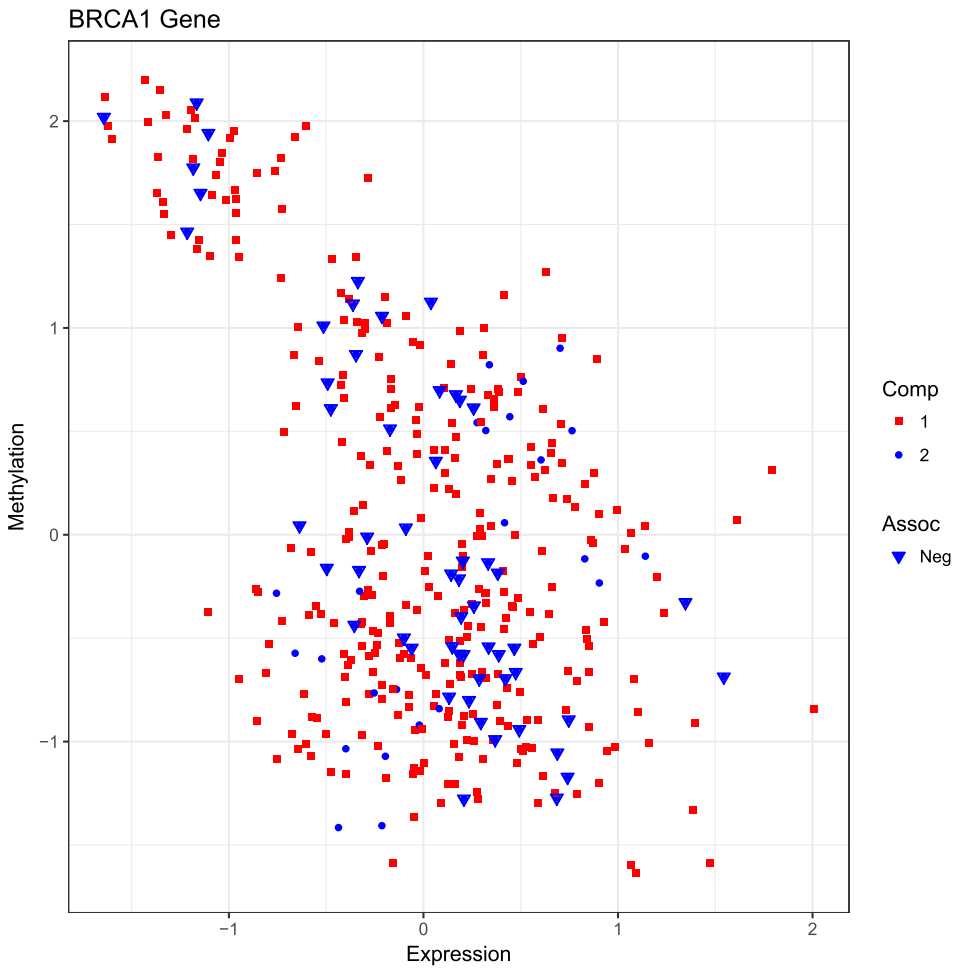


FIG. 4. *Scatter plot of gene expression vs methylation of BRCA*1. *Patients classified into Component* 1 *are shown in red, while those in Component* 2 *are in blue. The subset of patients in Component* 2 *that exhibit a negative association between gene expression and methylation are depicted in upside-down deltas.*

within Component 2. In summary, although the full-set analysis failed to find a relationship between gene expression and methylation, yet visual inspection seems to indicate that the relationships are similar for the two components, using taupath, we were able to not only uncover a negative relationship, but also identify the subset of observations that contributed to such a relationship.

5. Closing remarks and conclusions. There have been many studies on variable selection in different settings, but only a few that focus on FMR models. There also exist studies that consider variable selection in an AFT model with censoring.

However, to the best of our knowledge, none of these studies have considered variable selection in FMR when the observations are subject to a censoring mechanism. The ovarian cancer study, among many others, is an example of a real data set with both heterogeneity and censoring where there are numerous genomic features that may possibly have an effect on the response variable. We have developed a regularization method for variable selection in mixture of AFT models when observations are subject to right censoring. A natural extension of the proposed methods is to consider a mixture of semi-parametric AFT models without specifying the random error distributions. The main challenge in this direction is the issue of model identification. Hunter, Wang and Hettmansperger (2007) provided identifiability conditions for semi-parametric finite mixture models with $K = 2, 3$ components. However, model identification in the presence of covariates and censoring is yet to be formally studied. Another interesting research problem is hypothesis testing of the number of mixture components (K). Kasahara and Shimotsu (2015) and Zhu and Zhang (2004) provided statistical approaches for testing K in mixture of regression models. These results, however, are not applicable to our setting due to censoring and variable selection.

We studied the statistical properties of our proposed method theoretically. Our simulation results show that the proposed method based on different penalties performs reasonably well in finite samples. None of the penalties can universally dominate the others. The AdpLASSO shows a more consistent behaviour.

In our analysis of the ovarian cancer data, a pre-screening step as we had taken in the current paper was indispensable, because the dimension of the predictor variables should be smaller than the sample size. Such a pre-screening step is even more imperative for mixture of AFT models as the number of parameters is multiplied by the order of the mixture model. The results of our analysis point to the existence of both disease and patient heterogeneity in the HGS-OvCa data. Specifically, we uncovered two components in our optimal mixture of AFT model, with Component 1 capturing a more aggressive form of the disease and Component 2, a less aggressive form. This conclusion was reached based on several prognostic variables in conjunction with the posterior probability of each patient belonging to each of the components. In Component 2, the significant positive relationship between methylation and DFT for the gene BRCA1 is especially noteworthy, since BRCA1 has been found to play an important role in ovarian cancer, either based on mutation, gene expression or methylation. Our finding that, for a majority of patients in Component 2, their methylation level is negatively associated with the expression of BRCA1 is particularly interesting and biologically relevant. It is likely that the BRCA1 mutation is kept at base by the epigenetic mechanism for patients in Component 2, leading to their longer DFT. This finding provides additional insights into the disease mechanism, and may potentially help with sorting out inconsistent results seen in the literature.

In conclusion, we believe that our analysis of the ovarian cancer data using the proposed methodology can serve as a catalyst for the analyses of other data

with similar features; two particular characteristics are heterogeneity and censoring, which appear to be common in biomedical and genomics studies.

Acknowledgments. We would like to thank the Editor, the Associate Editor and the two anonymous referees for their constructive comments and suggestions, which have led to improved presentation and substance in the revised manuscript.

SUPPLEMENTARY MATERIAL

Supplement to "Capturing heterogeneity of covariate effects in hidden subpopulations in the presence of censoring and large number of covariates" (DOI: 10.1214/18-AOAS1198SUPP; .pdf). Supplementary Materials referenced in Section 2–4, including regularity conditions, proofs, numerical approaches, supplementary tables and figures, and the `fmrs` output are available with this paper at the Annals of Applied Statistics website.

REFERENCES

BOLTON, K. L., CHENEVIX-TRENCH, G., GOH, C., SADETZKI, S., RAMUS, S. J., KARLAN, B. Y. et al. (2012). Association between BRCA1 and BRCA2 mutations and survival in women with invasive epithelial ovarian cancer. *JAMA* **307** 382–389.

CAI, J., FAN, J., LI, R. and ZHOU, H. (2005). Variable selection for multivariate failure time data. *Biometrika* **92** 303–316. MR2201361

CERAMI, E., GAO, J., DOGRUSOZ, U., GROSS, B. E., SUMER, S. O., AKSOY, B. A. et al. (2012). The cBio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2** 401.

CHEN, J. and TAN, X. (2009). Inference for multivariate normal mixtures. *J. Multivariate Anal.* **100** 1367–1383. MR2514135

EARP, M. A. and CUNNINGHAM, J. M. (2015). DNA methylation changes in epithelial ovarian cancer histotypes. *Genomics* **106** 311–321.

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

FAN, J. and LI, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *Ann. Statist.* **30** 74–99. MR1892656

FARAGGI, D. and SIMON, R. (1998). Bayesian variable selection method for censored survival data. *Biometrics* **54** 1475–1485. MR1671590

GAO, J., AKSOY, B. A., DOGRUSOZ, U., DRESDNER, G., GROSS, B., SUMER, S. O. et al. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* **6** pl1.

GILBRIDE, T. J., ALLENBY, G. M. and BRAZELL, J. D. (2006). Models for heterogeneous variable selection. *J. Mark. Res.* **43** 420–430.

HAN, S. W., CHEN, G., CHEON, M.-S. and ZHONG, H. (2016). Estimation of directed acyclic graphs through two-stage adaptive lasso for gene network inference. *J. Amer. Statist. Assoc.* **111** 1004–1019. MR3561925

HUNTER, D. R., WANG, S. and HETTMANSPERGER, T. P. (2007). Inference for mixtures of symmetric distributions. *Ann. Statist.* **35** 224–251. MR2332275

KASAHARA, H. and SHIMOTSU, K. (2015). Testing the number of components in normal mixture regression models. *J. Amer. Statist. Assoc.* **110** 1632–1645. MR3449060

KHALILI, A. and CHEN, J. (2007). Variable selection in finite mixture of regression models. *J. Amer. Statist. Assoc.* **102** 1025–1038. MR2411662

KONG, B., LV, Z.-D., WANG, Y., JIN, L.-Y., DING, L. and YANG, Z.-C. (2015). Down-regulation of BRMS1 by DNA hypermethylation and its association with metastatic progression in triple-negative breast cancer. *Int. J. Clin. Exp. Pathol.* **8** 11076–11083.

KOUKOURA, O., SPANDIDOS, D. A., DAPONTE, A. and SIFAKIS, S. (2014). DNA methylation profiles in ovarian cancer: Implication in diagnosis and therapy (review). *Mol. Med. Rep.* **10** 3–9.

KWON, M. J. and SHIN, Y. K. (2011). Epigenetic regulation of cancer-associated genes in ovarian cancer. *Int. J. Mol. Sci.* **12** 983–1008.

KWONG, J., LEE, J.-Y., WONG, K.-K., ZHOU, X., WONG, D. T. W., LO, K.-W. et al. (2006). Candidate tumor-suppressor gene DLEC1 is frequently downregulated by promoter hypermethylation and histone hypoacetylation in human epithelial ovarian cancer. *Neoplasia* **8** 268–278.

LAWLESS, J. F. (2003). *Statistical Models and Methods for Lifetime Data*, 2nd ed. Wiley-Interscience, Hoboken, NJ. MR1940115

LEE, K. H., CHAKRABORTY, S. and SUN, J. (2011). Bayesian variable selection in semiparametric proportional hazards model for high dimensional survival data. *Int. J. Biostat.* **7** Art. 21. MR2787411

LIU, J., ZHANG, R., ZHAO, W. and LV, Y. (2015). Variable selection in semiparametric hazard regression for multivariate survival data. *J. Multivariate Anal.* **142** 26–40. MR3412736

LU, Z.-H. (2009). Covariate selection in mixture models with the censored response variable. *Comput. Statist. Data Anal.* **53** 2710–2723. MR2665920

LV, J. and FAN, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *Ann. Statist.* **37** 3498–3528. MR2549567

MCLACHLAN, G. J. and MCGIFFIN, D. C. (1994). On the role of finite mixture models in survival analysis. *Stat. Methods Med. Res.* **3** 211–226.

MCLACHLAN, G. and PEEL, D. (2004). *Finite Mixture Models*. Wiley, New York.

SCHÖNDORF, T., EBERT, M. P., HOFFMANN, J., BECKER, M., MOSER, N., PUR, Ş. et al. (2016). Hypermethylation of the PTEN gene in ovarian cancer cell lines. *Cancer Lett.* **207** 215–220.

SHA, N., TADESSE, M. G. and VANNUCCI, M. (2006). Bayesian variable selection for the analysis of microarray data with censored outcomes. *Bioinformatics* **22** 2262–2268.

SHOKOOHI, F., KHALILI, A., ASGHARIAN, M. and LIN, S. (2019). Supplement to "Capturing heterogeneity of covariate effects in hidden subpopulations in the presence of censoring and large number of covariates." DOI:10.1214/18-AOAS1198SUPP.

SOHN, I., KIM, J., JUNG, S.-H. and PARK, C. (2009). Gradient Lasso for Cox proportional hazards model. *Bioinformatics* **25** 1775–1781.

STÄDLER, N., BÜHLMANN, P. and VAN DE GEER, S. (2010). L_1-penalization for mixture regression models. *TEST* **19** 209–256. MR2677722

THE CANCER GENOME ATLAS RESEARCH NETWORK (2011). Integrated genomic analyses of ovarian carcinoma. *Nature* **474** 609–615.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

TIBSHIRANI, R. (1997). The Lasso method for variable selection in the Cox model. *Stat. Med.* **16** 385–395.

VOLINSKY, C. T. and RAFTERY, A. E. (2000). Bayesian information criterion for censored survival models. *Biometrics* **56** 256–262.

YANG, D., KHAN, S., SUN, Y., HESS, K., SHMULEVICH, I., SOOD, A. K. et al. (2011). Association of BRCA1 and BRCA2 mutations with survival, chemotherapy sensitivity, and gene mutator phenotype in patients with ovarian cancer. *JAMA* **306** 1557–1565.

ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. MR2604701

ZHANG, F., DING, J. and LIN, S. (2017). Testing for associations of opposite directionality in a heterogeneous population. *Statist. Biosci.* **9** 137–159.

ZHU, H.-T. and ZHANG, H. (2004). Hypothesis testing in mixture regression models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **66** 3–16. MR2035755

ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469

F. SHOKOOHI
A. KHALILI
M. ASGHARIAN
DEPARTMENT OF MATHEMATICS
AND STATISTICS
MCGILL UNIVERSITY
805 SHERBROOKE STREET WEST
BURNSIDE HALL
MONTREAL, QUEBEC H3A 0B9
CANADA
E-MAIL: shokoohi@icloud.com
farhad.shokoohi@mcgill.ca
farhad.shokoohi@concordia.ca
abbas.khalili@mcgill.ca
masoud.asgharian-dastenei@mcgill.ca
URL: https://goo.gl/AVFyCz
http://www.math.mcgill.ca/khalili/
http://www.math.mcgill.ca/asgharian/

S. LIN
DEPARTMENT OF STATISTICS
OHIO STATE UNIVERSITY
COLUMBUS, OHIO 43210
USA
E-MAIL: shili@stat.osu.edu
URL: http://www.stat.osu.edu/~shili/