Statistical Papers (2024) 65:2165–2190
https://doi.org/10.1007/s00362-023-01470-9

REGULAR ARTICLE

Privacy-preserving and homogeneity-pursuit integrative analysis for high-dimensional censored data

Xin Ye[1] · Baihua He[2] · Yanyan Liu[1] · Shuangge Ma[3]

Received: 13 October 2022 / Revised: 27 June 2023 / Published online: 19 September 2023

Abstract

In the analysis of data with a censored survival outcome and high-dimensional covariates, when a single data source has a limited sample size/power, integrative analysis of data from multiple sources can effectively increase sample size and improve estimation and variable selection performance. Under certain circumstances, for example when it is desirable to preserve data privacy, only summary statistics, as opposed to raw data, can be pooled for integrative analysis. In this study, we consider summary statistics-based integrative analysis of multi-source data with a censored survival outcome and high-dimensional covariates under the Cox model. This data setting can be more challenging than many in the literature. We further consider the scenario where some (but not all) covariates have homogeneous effects, and note that properly identifying such homogeneity can lead to more efficient estimation and a deeper understanding of the underlying data generation mechanisms. To this end, we propose a privacy-preserving penalized integrative analysis method, which can simultaneously achieve regularized estimation, variable selection, and homogeneity pursuit. An effective computational algorithm is developed, and asymptotic consistency and distributional properties are rigorously established. Numerical studies, including simulation and the analysis of a bladder cancer data set, convincingly demonstrate the practical effectiveness of the proposed method.

Keywords Homogeneity pursuit · Integrative analysis · Penalized estimation · Privacy preservation

Mathematics Subject Classification 62N01 · 62N02 · 62P10

✉ Baihua He
baihua@ustc.edu.cn

[1] School of Mathematics and Statistics, Wuhan University, Wuhan, China

[2] International Institute of Finance, School of Management, University of Science and Technology of China, Hefei, China

[3] Department of Biostatistics, Yale University, New Haven, USA

1 Introduction

Data with a censored survival outcome and high-dimensional covariates are commonly encountered. In the analysis of such data, variable selection is usually needed along with estimation. Many regularization methods, based on the penalization, thresholding, Bayesian, boosting, and other techniques, have been developed (Lee et al. 2011; Wang and Wang 2011; Zhang 2010). In practice, data from a single source often have a limited sample size/power, leading to unsatisfactory estimation. For many scientific problems, there are multiple independent data sources/sets, making it possible to pool data, conduct integrative analysis, increase sample size/power, and improve estimation and selection. The first, relatively easier integrative analysis scenario is when raw data from multiple sources can be shared. For application examples and methodological and theoretical developments, we refer to Cheng et al. (2015), Huang et al. (2017), and Tang et al. (2018). In such analysis, multiple sources can have the same model or heterogeneous covariate effects.

It has been recognized that under certain circumstances, raw data cannot be shared. One consideration pertains to preserving privacy (Gomatam et al 2005; Homer et al. 2008; Wolfson et al. 2010). For example, there are often many challenges with sharing medical and genetic data. Similar concerns may also arise in the analysis of personal financial data. Additionally, when both the dimensionality of covariates and sample size are high, sharing raw data may incur considerable cost. Under these circumstances, it can be easier or even necessary to share summary statistics, which are generated from analyzing local data, as opposed to raw data. To this end, Wolfson et al. (2010) proposed a generic individual information protected integrative analysis framework, called DataSHIELD, that transfers only summary statistics from distributed local sites to a central site for pooled analysis. Several distributed approaches for high-dimensional data can be used for integrative analysis under the DataSHIELD constraint. A popular approach is to synthesize local estimators. Chen and Xie (2014) proposed to average local estimates and recover sparsity by adopting majority voting. Wang et al. (2014) adopted a similar idea to combine local Lasso estimators. Lee and Zeng (2017) and Battey et al. (2018) proposed distributed inference procedures by aggregating local debiased Lasso estimators which can be used for inference directly. They obtained sparse estimators by truncating aggregated debiased estimators. An alternative approach is to adopt sequential communications between a central site and multiple local sites and minimize a global loss function, refered to as distributed regression (Karr et al. 2007). Li et al. (2016) developed the communication procedure for ridge logistic regression. Wang et al. (2016) and Jordan et al. (2019) proposed a communication-efficient surrogate likelihood framework for high-dimensional settings.

The above and some other works assume homogeneity in covariate effects. That is, multiple data sources/sets share the same underlying model. In integrative analysis based on raw data, it has been well argued that this assumption may be too stringent (Liu et al. 2013, 2014). There are often significant differences in experimental settings, sample characteristics, data collection mechanisms, and other aspects that cannot be eliminated and can lead to different regression models. It is noted that the heterogeneity structure (in model/covariate effects) includes the homogeneity structure as a special case and is more flexible. A traditional way to deal with heterogeneous data is meta-analysis based on random effect model. However, it only concerns about the estimation of mean effect. When heterogeneity is of interest, it can not be able to detect such effects (Danieli and Moodie 2021; Walker et al. 2008). To tackle this problem as well as to protect privacy, some summary statistics-based approaches have been developed to accommodate heterogeneity in covariate effects. For example, He et al. (2016) proposed an overall surrogate loss function based on local maximum likelihood estimates and Hessian matrixs, and accommodated heterogeneity by adopting hierarchical Lasso. Cai et al. (2021) proposed an overall surrogate loss function based on local debiased Lasso estimators, gradient vectors and Hessian matrixs. They tackled the heterogeneity problem by reparameterizing coefficients into a homogeneous part and a heterogeneous part.

In most of the aforementioned and other existing works of privacy preserving, the methods has focused on completely observable data. Though some of them have generality on many problems, they are not directly appliable to survival data with censoring. The models to deal with censored data are more complex and the theories are different. Lu et al. (2015) developed a privacy preserving web service for distributed censored data. Shu et al. (2020) proposed a distributed method for inverse probability weighted Cox model under the case where there are two groups (control/treatment) and there is no individualized effect in Cox model. More recently, Li et al. (2022) developed a more general method for distributed Cox model which involves a multi-stage procedure that alternates between sites and central computer. Moodie et al. (2022) gave a summary-statistics based approach for privacy-preserving estimation of individualized treatment rule based on accelerated failure time model. However, these studies are limited to low-dimensional covariates and does not address the problem of covariate effect heterogeneity. Cheng et al. (2015) proposed an integrative approach for pooled cohort studies that can accommodate covariates with homogeneous and heterogeneous effects, and Tang et al. (2018) used a fused Lasso-type penalty to accommodate partially heterogeneous effects under the Cox model. It is noted that both studies demand raw data and thus fail to meet privacy preservation. As far as we know, the privacy preserving analysis of high-dimensional censored survival data considering heterogeneous covariate effects is still a blank.

In this article, we consider the integrative analysis of multi-source data with a censored survival outcome and high-dimensional covariates where heterogeneous effects may exist. We focus on the scenario where, with a data privacy consideration, only summary statistics can be shared. The proposed method is privacy-preserving for multi-site high-dimensional heterogeneous censored data. We assume a Cox proportional hazard model and flexibly consider the model setting where some covariates have homogeneous effects (but others do not), with the set of such covariates being unknown a priori. The proposed method is data-driven to identify homogeneous covariate effects (we call it homogeneity pursuit here) and at the same time select important variables. It can be seen that pursuing homogeneity can reduce the number of unknown parameters in the model and give more understanding of the connection among multiple datasets.

The proposed method can be viewed as a distributed regression approach since it aims to minimize a (approximated) global loss function, but it don’t request multiple communications that alternate between sites and central computer. Each site is only requested once to send summary statistics to the central computer. In this way, privacy of personal data can be preserved. Since a (approximated) global loss function is built in one computer, it will be more convenient and flexible to adopt regularization technique to deal with heterogeneous covariates effects. This study may complement the existing literature in multiple important aspects. First, the integrative analysis of censored survival data is conducted, which can be more challenging than the analysis of completely observable data. Second, the privacy-preserving analysis can complement the analysis that demands raw data. Third, the proposed approach allows for partially homogeneous covariate effects and can be more flexible than those demanding fully homogenous effects. It can data-dependently determine which covariates have homogeneous (versus heterogeneous) effects. This can lead to a deeper understanding of the underlying data generating mechanisms and "interconnections" among multiple data sources/sets. Fourth, our theoretical and computational developments can also shed broad insights into integrative analysis and high-dimensional regularized estimation. Last but not least, this study can deliver a practically useful tool for many scientific problems.

2 Methods

2.1 Integrative analysis

Suppose that there are K independent data sources with n_k subjects in the kth source, $k = 1, 2, \dots, K$. $n = \sum_{k=1}^{K} n_k$. Let $\widetilde{T}_i^{(k)}$ be the event time, and $C_i^{(k)}$ be the censoring time, for $k = 1, 2, \dots, K$ and $i = 1, 2, \dots, n_k$. For the ith subject in the kth source, the observed sample is $\left\{T_i^{(k)}, \boldsymbol{Z}_i^{(k)}, \Delta_i^{(k)}\right\}$, where $T_i^{(k)} = \widetilde{T}_i^{(k)} \wedge C_i^{(k)}$, $\Delta_i^{(k)} = I(\widetilde{T}_i^{(k)} \leq C_i^{(k)})$, and $\boldsymbol{Z}_i^{(k)} \in \mathbb{R}^p$ is the covariate vector. Assume that $\widetilde{T}_i^{(k)}$ follows the Cox model with hazard:

$$\lambda^{(k)}(t|\boldsymbol{Z}_i^{(k)}) = \lambda_0^{(k)}(t) \exp\left\{\boldsymbol{\beta}^{(k)\mathrm{T}} \boldsymbol{Z}_i^{(k)}\right\}, \tag{1}$$

where $\boldsymbol{\beta}^{(k)} \in \mathbb{R}^p$ is the coefficient vector. Let $Y_i^{(k)}(t) = I(T_i^{(k)} \geq t)$. The log partial likelihood function based on the kth local data is:

$$\ell_k(\boldsymbol{\beta}^{(k)}) = \sum_{i=1}^{n_k} \Delta_i^{(k)} \left\{\boldsymbol{\beta}^{(k)\mathrm{T}} \boldsymbol{Z}_i^{(k)} - \log\left(\sum_{i=1}^{n_k} Y_i^{(k)}(T_i^{(k)}) \exp\left(\boldsymbol{\beta}^{(k)\mathrm{T}} \boldsymbol{Z}_i^{(k)}\right)\right)\right\}, \tag{2}$$

As discussed in Sect. 1, under the context of privacy preservation, raw data are not directly available. Accordingly, we propose the overall goodness-of-fit:

$$L(\boldsymbol{\beta}) = \frac{1}{2} \sum_{k=1}^{K} \frac{n_k}{n} \left\{\boldsymbol{\beta}^{(k)\mathrm{T}} \widehat{\mathbf{H}}^{(k)} \boldsymbol{\beta}^{(k)} - 2\widehat{\boldsymbol{g}}^{(k)\mathrm{T}} \boldsymbol{\beta}^{(k)}\right\},$$

where, for $k = 1, 2, \ldots, K$, $\widehat{\mathbf{H}}^{(k)} = -n_k^{-1}\nabla^2\ell_k\big(\widehat{\boldsymbol{b}}^{(k)}\big)$, $\widehat{\boldsymbol{g}}^{(k)} = \widehat{\mathbf{H}}^{(k)}\widehat{\boldsymbol{b}}^{(k)} + n_k^{-1}\nabla\ell_k(\widehat{\boldsymbol{b}}^{(k)})$, $\nabla\ell_k(\boldsymbol{\beta})$ and $\nabla^2\ell_k(\boldsymbol{\beta})$ are the first and second derivatives with respect to $\boldsymbol{\beta}$, and $\widehat{\boldsymbol{b}}^{(k)}$ is the local Lasso estimator with tuning parameter $a_k > 0$, that is,

$$\widehat{\boldsymbol{b}}^{(k)} = \operatorname{argmin}_{\boldsymbol{\beta}} \left\{-n_k^{-1}\ell_k\big(\boldsymbol{\beta}\big) + a_k\big\|\boldsymbol{\beta}\big\|_1\right\}, \tag{3}$$

and $\|\cdot\|_1$ denotes the ℓ_1 norm.

Remark 2.1 $L\big(\boldsymbol{\beta}\big)$ is built on the Taylor expansion of the overall negative log-likelihood function $-\sum_{k=1}^K \ell_k(\boldsymbol{\beta}^{(k)})/n$ at the local debiased Lasso estimators (Yu et al. 2021) defined as:

$$\widetilde{\boldsymbol{b}}^{(k)} = \widehat{\boldsymbol{b}}^{(k)} + \widehat{\boldsymbol{\Theta}}^{(k)} n_k^{-1}\nabla\ell_k\big(\widehat{\boldsymbol{b}}^{(k)}\big), \tag{4}$$

where $\widehat{\boldsymbol{\Theta}}^{(k)}$ is a sparse estimator approximating the inverse of $\widehat{\mathbf{H}}^{(k)}$. If $\sum_{k=1}^K \nabla\ell_k\big(\widetilde{\boldsymbol{b}}^{(k)}\big)/n = o_p(1)$, and $\widehat{\boldsymbol{\Theta}}^{(k)}\widehat{\mathbf{H}}^{(k)} = \mathbf{I} + o_p(1)$, we have

$$\begin{aligned}
-\frac{1}{n}\sum_{k=1}^K \ell_k\big(\boldsymbol{\beta}^{(k)}\big) &= -\frac{1}{n}\sum_{k=1}^K \ell_k\big(\widetilde{\boldsymbol{b}}^{(k)}\big) + \frac{1}{2}\sum_{k=1}^K \frac{n_k}{n}\big(\boldsymbol{\beta}^{(k)} - \widetilde{\boldsymbol{b}}^{(k)}\big)^{\mathrm{T}}\widehat{\mathbf{H}}^{(k)}\big(\boldsymbol{\beta}^{(k)} - \widetilde{\boldsymbol{b}}^{(k)}\big) + o_p(1) \\
&= \frac{1}{2}\sum_{k=1}^K \frac{n_k}{n}\left\{\boldsymbol{\beta}^{(k)\mathrm{T}}\widehat{\mathbf{H}}^{(k)}\boldsymbol{\beta}^{(k)} - 2\widehat{\boldsymbol{g}}^{(k)\mathrm{T}}\boldsymbol{\beta}^{(k)}\right\} + C\big(\widetilde{\boldsymbol{b}}\big) + o_p(1),
\end{aligned} \tag{5}$$

where $C\big(\widetilde{\boldsymbol{b}}\big) = -n^{-1}\sum_{k=1}^K \ell_k\big(\widetilde{\boldsymbol{b}}^{(k)}\big) + \widetilde{\boldsymbol{b}}^{(k)\mathrm{T}}\widehat{\mathbf{H}}^{(k)}\widetilde{\boldsymbol{b}}^{(k)}$, which leads to the proposed $L\big(\boldsymbol{\beta}\big)$. A similar strategy also based on local debiased Lasso estimators has been adopted in Cai et al. (2021).

Remark 2.2 As has been argued in the literature and can be easily seen, $L\big(\boldsymbol{\beta}\big)$ contains the key first- and second-order information of the overall negative log-likelihood $-n^{-1}\sum_{k=1}^K \ell_k\big(\boldsymbol{\beta}^{(k)}\big)$ on $\boldsymbol{\beta}$. As we are only interested in estimation and inference (not higher-order properties), all sites only need to share $\left\{\widehat{\mathbf{H}}^{(k)}, \widehat{\boldsymbol{g}}^{(k)}\right\}$, $k = 1, 2, \ldots, K$, as opposed to raw data. Additionally, it is noted that we make use of the local debiased Lasso estimators without calculating $\widehat{\boldsymbol{\Theta}}^{(k)}$, $k = 1, 2, \ldots, K$, which is time-consuming and can only be estimated well under strong conditions (Yu et al. 2021).

To accommodate heterogeneity, we reparameterize the regression parameters as $\boldsymbol{\alpha}^{(k)} = \boldsymbol{\beta}^{(k)} - \boldsymbol{\mu}$ for $k = 1, 2, \ldots, K$, where $\boldsymbol{\mu} = K^{-1}\sum_{k=1}^K \boldsymbol{\beta}^{(k)}$. Here, $\boldsymbol{\mu}$ represents the average effects, and $\boldsymbol{\alpha}^{(k)}$'s represent the "deviances" of effects in source k from the average. Let $\boldsymbol{\alpha}_j = \big(\alpha_j^{(1)}, \cdots, \alpha_j^{(K)}\big)^T$, and $\|\cdot\|_2$ be the Euclidean norm. Then covariates can be classified into three mutually exclusive categories: (1) homogeneous effects if $\mu_j \neq 0$ and $\|\boldsymbol{\alpha}_j\|_2 = 0$; (2) heterogeneous effects if $\|\boldsymbol{\alpha}_j\|_2 \neq 0$; and (3) null effects if $\mu_j = 0$ and $\|\boldsymbol{\alpha}_j\|_2 = 0$.

For simultaneous estimation, variable selection, and homogeneity pursuit, we propose the following penalized objective function:

$$Q(\boldsymbol{\mu}, \boldsymbol{\alpha}) = L(\boldsymbol{\mu} + \boldsymbol{\alpha}) + \widetilde{\rho}(\boldsymbol{\mu}, \boldsymbol{\alpha}; \lambda_1, \lambda_2),$$

where

$$\widetilde{\rho}(\boldsymbol{\mu}, \boldsymbol{\alpha}; \lambda_1, \lambda_2) = \sum_{j=1}^{p} \rho_{\mathrm{MCP}}(|\mu_j|; \lambda_1, \gamma) + \sum_{j=1}^{p} \rho_{\mathrm{MCP}}(\|\boldsymbol{\alpha}_j\|_2; \lambda_2, \gamma), \tag{6}$$

$\rho_{\mathrm{MCP}}(t; \lambda, \gamma) = \lambda \int_0^t \left(1 - x/(\gamma\lambda)\right)_+ dx$ is the MCP penalty function (Zhang 2010) with tuning parameters $\lambda > 0$, $\gamma > 1$. The proposed estimator is defined as:

$$\left(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\alpha}}\right) = \mathrm{argmin}_{\boldsymbol{\mu}, \boldsymbol{\alpha}} Q(\boldsymbol{\mu}, \boldsymbol{\alpha}), \quad \text{subject to } \sum_{k=1}^{K} \boldsymbol{\alpha}^{(k)} = \mathbf{0}. \tag{7}$$

The final estimates of the coefficients are $\widehat{\boldsymbol{\beta}}^{(k)} = \widehat{\boldsymbol{\mu}} + \widehat{\boldsymbol{\alpha}}^{(k)}$, $k = 1, 2, \ldots, K$.

Remark 2.3 With the proposed approach, the determination of homogeneity versus heterogeneity is fully data-driven. For a specific covariate, if homogeneity is concluded, then, intuitively, its estimate is based on information from all data. As such, more accuracy (than individual estimates) can be expected. By flexibly allowing for heterogeneity, the proposed approach can avoid the risk of making wrong homogeneity assumption and generating biased estimation. Null effects can be viewed as a special case of homogeneous effects with $\widehat{\mu}_j = 0$. As such, the proposed approach may also better identify null (versus nonzero) effects.

It is possible to replace MCP with another penalty. Cai et al. (2021) adopted Lasso type penalty. We conduct some simulation tests to compare the performances under different type of penalty. In our problems and simulation settings, as shown in our numerical analysis, MCP will gain more accuracy.

2.2 Computational algorithm

The estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\alpha}$ can be obtained by solving the following constrained optimizing problem:

$$\begin{aligned} &\min_{\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\omega}} \left\{ L(\boldsymbol{\mu} + \boldsymbol{\alpha}) + \sum_{j=1}^{p} J_{\lambda_1}(|\mu_j|) + \lambda_1 \|\boldsymbol{\mu}\|_1 + \sum_{j=1}^{p} \rho_{\mathrm{MCP}}\left(\|\boldsymbol{\omega}_j\|_2; \lambda_2\right) \right\}, \\ &\text{subject to} \quad \sum_{k=1}^{K} \boldsymbol{\alpha}^{(k)} = \mathbf{0}, \quad \boldsymbol{\alpha} = \boldsymbol{\omega}, \end{aligned} \tag{8}$$

where $J_{\lambda_1}(|\mu_j|)$ is the concave part derived from $\rho_{\text{MCP}}(\mu_j;\lambda_1)$:

$$\begin{aligned}J_{\lambda_1}(|\mu_j|) &= \rho_{\text{MCP}}(\mu_j;\lambda_1) - \lambda_1|\mu_j| \\ &= -\frac{\mu_j^2}{2\gamma}I(0 \le |\mu_j| < \gamma\lambda_1) + \left(\frac{\gamma\lambda_1^2}{2} - \lambda_1|\mu_j|\right)I(|\mu_j| \ge \gamma\lambda_1).\end{aligned}$$

We adopt the Augmented Lagrangian Multiplier (ALM) technique to solve (8).

Remark 2.4 The original objective function (7) is not easy to solve due to the concavity of the penalty. In (8), we divide $\rho_{\text{MCP}}(\mu_j;\lambda_1)$ into two parts: convex Lasso and a non-convex but differentiable residual. Then, the optimization with respect to $\boldsymbol{\mu}$ involves a differentiable part $L_{\text{PHI}}(\boldsymbol{\mu},\boldsymbol{\alpha}) + \sum_{j=1}^{p} J_{\lambda_1}(|\mu_j|)$ as well as a convex and nonsmooth part $\lambda_1\|\boldsymbol{\mu}\|_1$, which can be solved with the proximal gradient (PG) method. A similar decomposition strategy has been used in Wang et al. (2013). Additionally, we consider the reparameterization $\boldsymbol{\omega}$ with the constraint $\boldsymbol{\alpha} = \boldsymbol{\omega}$. This way, the optimization with respect to both $\boldsymbol{\alpha}$ and $\boldsymbol{\omega}$ has explicit solutions. A similar strategy has been adopted with group MCP in He et al. (2020).

The augmented Lagrangian function of (8) with tuning parameter $\psi > 0$ is

$$\begin{aligned}\mathcal{L}_{\psi}(\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\omega},\boldsymbol{V},\boldsymbol{\Lambda}) = \Bigg\{ & L(\boldsymbol{\mu}+\boldsymbol{\alpha}) + \sum_{j=1}^{p} J_{\lambda_1}(|\mu_j|) + \lambda_1\|\boldsymbol{\mu}\|_1 + \sum_{j=1}^{p}\rho_{\text{MCP}}\big(\|\boldsymbol{\omega}_j\|_2;\lambda_2\big) \\ & + \frac{\psi}{2}\|\sum_{k=1}^{K}\boldsymbol{\alpha}^{(k)} + \frac{1}{\psi}\boldsymbol{V}\|_2^2 + \frac{\psi}{2}\sum_{j=1}^{p}\|\boldsymbol{\alpha}_j - \boldsymbol{\omega}_j + \frac{1}{\psi}\boldsymbol{\Lambda}_j\|_2^2\Bigg\}.\end{aligned}$$

Overall, we propose an iterative algorithm. With the local Lasso estimates as the initial value, the proposed algorithm iterates between the following two steps until convergence (which is concluded when the absolute difference between the estimates from two consecutive iterations is smaller than a predefined cutoff). Here we use superscript $[m]$ to denote the mth iteration.

Step 1: Given $\boldsymbol{V}^{[m]}$ and $\boldsymbol{\Lambda}^{[m]}$, solve

$$\left\{\boldsymbol{\mu}^{[m+1]}, \boldsymbol{\alpha}^{[m+1]}, \boldsymbol{\omega}^{[m+1]}\right\} = \text{argmin}_{\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\omega}}\mathcal{L}_{\psi}(\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\omega},\boldsymbol{V}^{[m]},\boldsymbol{\Lambda}^{[m]}).$$

Step 2: Given $\boldsymbol{\alpha}^{[m+1]}$, $\boldsymbol{\omega}^{[m+1]}$, $\boldsymbol{V}^{[m]}$, and $\boldsymbol{\Lambda}^{[m]}$, update $\boldsymbol{V}$ and $\boldsymbol{\Lambda}$ as follows:

$$\begin{aligned}\boldsymbol{V}^{[m+1]} = \boldsymbol{V}^{[m]} + \psi\sum_{k=1}^{K}\boldsymbol{\alpha}^{(k)[m+1]}, \quad & \boldsymbol{\Lambda}^{[m+1]} \\ & = \boldsymbol{\Lambda}^{[m]} + \psi\left(\boldsymbol{\alpha}^{[m+1]} - \boldsymbol{\omega}^{[m+1]}\right).\end{aligned}$$

More details of Step 1 are provided in the Appendix.

The proposed approach is computationally affordable. In all of our numerical studies, convergence is achieved within a small to moderate number of iterations.

Convergence of the proposed algorithm can be established by similar arguments as in Wang et al. (2014) and He et al. (2020).

For choosing tuning parameters λ_1 and λ_2, we consider the generalized information criterion (GIC) (Wang and Leng 2007) defined as:

$$\mathrm{GIC}(\lambda_1, \lambda_2) = L(\widehat{\boldsymbol{\beta}}) + \frac{\nu_n}{2}\mathrm{DF}(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\alpha}}),$$

where $\mathrm{DF}(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\alpha}})$ is the overall number of nonzero components of $\widehat{\boldsymbol{\mu}}$ and $\boldsymbol{\alpha}^{(k)}$, $k \in \{2, \dots, K\}$. As discussed in Cai et al. (2021), ν_n can be $2/n$ for AIC, $\log n/n$ for BIC, and $\log(\log p) \log n/n$ for modified BIC. In our numerical studies, we adopt the modified BIC, which leads to favorable results. Since we consider the scenario where raw data can not be shared, adopting cross validation (CV) to select tuning parameters will result in additional communications. Therefore here we don't suggest using CV for tuning parameter selection.

2.3 Theoretical properties

Here we examine theoretical properties of the proposed estimator under high dimensional settings. Assumptions (A1)–(A4) are described in the Appendix, and we note that they are sensible and comparable to those in the literature.

For a set $\mathcal{A}$, $|\mathcal{A}|$ denotes its size. For a vector $\boldsymbol{q} = (q_j, j = 1, 2, \dots, p)$ and an index set $\mathcal{A} \subset \{1, 2, \dots, p\}$, $\boldsymbol{q}_{\mathcal{A}} = (q_j, j \in \mathcal{A})$. Let $\boldsymbol{\beta}_0^{(k)}$, $\boldsymbol{\mu}_0$, and $\boldsymbol{\alpha}_0^{(k)}$, $k = 1, 2, \dots, K$, be the true parameter values. Define $\boldsymbol{\varphi}^{\mathrm{T}} = (\boldsymbol{\mu}^{\mathrm{T}}, \boldsymbol{\alpha}_1^{(-1)\mathrm{T}}, \dots, \boldsymbol{\alpha}_p^{(-1)\mathrm{T}})$, where $\boldsymbol{\alpha}_j^{(-1)} = (\alpha_j^{(k)}, k = 2, 3, \dots, K)^{\mathrm{T}}$, and denote its true value as $\boldsymbol{\varphi}_0$. Let the support set of $\boldsymbol{\mu}_0$ be $\mathcal{S}_\mu = \left\{j = 1, 2, \dots, p \middle| \mu_{0j} \neq 0\right\}$ and that of $\boldsymbol{\alpha}_0$ be $\mathcal{S}_\alpha = \left\{j = 1, 2, \dots, p \middle| \|\boldsymbol{\alpha}_{0j}\|_2 \neq 0\right\}$. Define $\mathcal{S}_\mu^c = \left\{j = 1, 2, \dots, p \middle| j \notin \mathcal{S}_\mu\right\}$ and $\mathcal{S}_\alpha^c = \left\{j = 1, 2, \dots, p \middle| j \notin \mathcal{S}_\alpha\right\}$. Let $\mathcal{S}_{\mu,\alpha} = \mathcal{S}_\mu \cup \mathcal{S}_\alpha$ and $\mathcal{S} = \left\{j = 1, 2, \dots, pK \middle| \boldsymbol{\varphi}_0 \neq 0\right\}$.

In our theoretical developments, we take a popular two-step strategy. In particular, we first consider the oracle estimator, for which the true sparsity structure is known. Then we establish that the estimate is equal to the oracle one with a high probability. Specifically, the oracle estimator is defined as:

$$\begin{gathered}\left(\widehat{\boldsymbol{\mu}}^{\mathrm{or}}, \widehat{\boldsymbol{\alpha}}^{\mathrm{or}}\right) = \operatorname{argmax}_{\boldsymbol{\mu},\boldsymbol{\alpha}} \left\{\frac{1}{n}\sum_{k=1}^{K} \ell_k(\boldsymbol{\mu} + \boldsymbol{\alpha}^{(k)})\right\}, \\ \text{subject to } \boldsymbol{\mu}_{\mathcal{S}_\mu^c} = \mathbf{0}, \boldsymbol{\alpha}_j = \mathbf{0}, j \in \mathcal{S}_\alpha^c, \sum_{k=1}^{K} \boldsymbol{\alpha}^{(k)} = \mathbf{0},\end{gathered} \tag{9}$$

and the oracle estimator $\widehat{\boldsymbol{\varphi}}^{\mathrm{or}}$ can be defined accordingly. Let $q_0 = |\mathcal{S}_\mu| + |\mathcal{S}_\alpha|$, $\lambda^* = \max(\lambda_1, \lambda_2)$, $\lambda_* = \min(\lambda_1, \lambda_2)$, $\mu_* = \min_{j \in \mathcal{S}_\mu} |\mu_{0j}|$, and $\|\boldsymbol{\alpha}_*\|_2 = \min_{j \in \mathcal{S}_\alpha} \|\boldsymbol{\alpha}_{0j}\|_2$. We first establish the following results.

Theorem 2.1 (Estimation and selection consistency) *Suppose that* (**A1**)*,* (**A2**)*,* (**A3**) *and* (**A4**) *described in the Appendix hold. Then,*

(i) $\|\widehat{\boldsymbol{\varphi}}^{or} - \boldsymbol{\varphi}_0\|_2 = O_p\left(\sqrt{|\mathcal{S}|/n}\right)$.

(ii) *If there exists a sequence* $\lambda_n \to \infty$ *as* $n \to \infty$, *such that* $\min(\mu_* - \gamma\lambda_1, \|\boldsymbol{\alpha}_*\|_2 - \gamma\lambda_2) \geq \lambda_n\sqrt{\log(\max(q_0, 1)/n)}$, *and* $\lambda_* \geq \lambda_n\sqrt{q_0\log(2p - q_0)/n}$, *then* $\mathbb{P}\left(\widehat{\boldsymbol{\varphi}} \neq \widehat{\boldsymbol{\varphi}}^{or}\right) \to 0$.

Proof is presented in the Appendix. It is noted that, although the theoretical developments share some similar spirit with the existing literature, with the censored survival outcome and homogeneity pursuit, the existing results cannot be directly applied, and nontrivial developments are needed. With the estimation and selection consistency results, we then move on to studying asymptotic normality. Define the log partial likelihood function based on all data, with the constraint $\sum_{k=1}^{K}\boldsymbol{\alpha}^{(k)} = 0$ taken into account, as $\ell(\boldsymbol{\varphi}) = \ell(\boldsymbol{\mu}, \boldsymbol{\alpha}^{(-1)}) = \sum_{k=1}^{K}\ell_k(\boldsymbol{\mu}, \boldsymbol{\alpha}^{(-1)})/n$, where ℓ_k is defined in (2). Accordingly, for the penalty function, we write $\widetilde{\rho}(\boldsymbol{\varphi})$ instead. Use $\nabla_{\mathcal{S}}\ell_k(\boldsymbol{\varphi})$ and $\nabla^2_{\mathcal{S}}\ell_k(\boldsymbol{\varphi})$ to denote the first and second order derivatives with respect to $\boldsymbol{\varphi}_{\mathcal{S}}$. Let $\Sigma_{\mathcal{S}} = \mathbb{E}\nabla^2_{\mathcal{S}}\ell(\boldsymbol{\varphi}_0)$, $a_{1n} = \nabla_{\mathcal{S}}\widetilde{\rho}(\boldsymbol{\varphi}_0)$, and $a_{2n} = \nabla^2_{\mathcal{S}}\widetilde{\rho}(\boldsymbol{\varphi}_0)$.

Theorem 2.2 (Asymptotic normality) *Suppose that the results in Theorem 2.1 hold. Assume that assumptions* (**A1**) *and* (**A2**) *described in the Appendix hold. Assume* $|\mathcal{S}|^2\sqrt{\log p/n} = o(1)$, $\sqrt{n}\max\left(\|a_{1n}\|_2, \|a_{2n}\|_2\right) = o(1)$, *and the eigenvalues of* $\Sigma_{\mathcal{S}}$ *are finite and bounded away from zero. Then for any* $\boldsymbol{q} \in \mathbb{R}^{|\mathcal{S}|}$, *if* $\boldsymbol{q}^{\mathrm{T}}\Sigma_{\mathcal{S}}^{-1}\boldsymbol{q} \to \sigma^2 < \infty$ *with* $\|\boldsymbol{q}\|_2 < \infty$, *it holds that* $\sqrt{n}\boldsymbol{q}^{\mathrm{T}}\left(\widehat{\boldsymbol{\varphi}}_{\mathcal{S}} - \boldsymbol{\varphi}_{0\mathcal{S}}\right) \to_d N(0, \sigma^2)$.

Proof is presented in the Appendix. It is noted that, in practice, with a fixed p, we can approximate the variance of $\widehat{\boldsymbol{\varphi}}_{\mathcal{S}}$ by $\frac{1}{n}\left\{\nabla^2_{\mathcal{S}}\ell(\widehat{\boldsymbol{\varphi}})\right\}^{-1}$.

3 Simulation

We conduct simulation to examine performance of the proposed approach (referred to as PMCP) and gauge against the following alternatives: (i) Oracle estimator, which is defined in (9); (ii) Local Lasso estimator (LasLo), defined in (3); (iii) averaged debiased Lasso estimator (dLasDC) from Lee and Zeng (2017), which takes the mean of local debiased Lasso estimators and recovers sparsity by thresholding; (iv) sparse meta analysis (SMA) method from He et al. (2016), which is based on local MLE estimators (if dimension is high, then marginal screening is first conducted), and solves for:

$$\widehat{\boldsymbol{\beta}}_{SMA} = \arg\min_{\boldsymbol{\beta}}\sum_{k=1}^{K}(\boldsymbol{\beta}^{(k)} - \widehat{\boldsymbol{\beta}}^{(k)}_{MLE})^{\mathrm{T}}\widehat{\mathbf{H}}^{(k)}(\boldsymbol{\beta}^{(k)} - \widehat{\boldsymbol{\beta}}^{(k)}_{MLE}) + \lambda\sum_{j=1}^{p}\left(\sum_{k=1}^{K}w_{jk}|\beta^{(k)}_j|\right)^{1/2},$$

where $\widehat{\boldsymbol{\beta}}^{(k)}_{MLE}$ is the local MLE estimator, $\widehat{\mathbf{H}}^{(k)}$ is the same as defined in Sect. 2.1, w_{jk} is the predetermined weight taking value $w_{jk} = \left(\widehat{\boldsymbol{\beta}}^{(k)}_{MLE,j}\right)^{-1}$ under homogeneity, or $w_{jk} = \left(\sum_{k=1}^{K}\widehat{\boldsymbol{\beta}}^{(k)}_{MLE,j}/K\right)^{-1}$, $k = 1, 2, \ldots, K$, under heterogeneity and λ is the tuning parameters selected by a modified information criterion (MIC): $MIC(\lambda) =$

$\sum_{k=1}^{K} \|\widehat{\boldsymbol{\beta}}^{(k)}_{MLE} - \widehat{\boldsymbol{\beta}}^{(k)}_{MLE}\|_2^2 + \sum_{k=1}^{K} \frac{\log n_k}{n_k} DF_k$, with DF_k being the number of the nonzero components of $\widehat{\boldsymbol{\beta}}^{(k)}_{SMA}$; (v) median selection of local Lasso estimators from Wang et al. (2014) (LasVote), which is based on local Lasso estimators $\widehat{\boldsymbol{b}}^{(k)}$ and recovers sparsity by selecting the median values of $\widehat{b}^{(k)}_j, k = 1, 2, \ldots, K$, for every $j = 1, 2, \ldots, p$. Additionally, to "re-establish" the advantage of MCP, we also consider an alternative that applies Lasso (as opposed to MCP) in the proposed method. This alternative is referred to as PLAS. For the alternative methods, tuning parameters are selected in a way similar to the proposed method. We acknowledge that there may be other applicable alternatives. The above may be the most relevant.

To evaluate estimation performance, we adopt absolute estimation error (AEE). To evaluate variable selection performance, we calculate true positive rate (TPR) with $\widehat{\boldsymbol{\beta}}^{(\cdot)}$ (which is defined as the ratio of nonzero coefficients that can be correctly identified) and false discovery rate (FDR, which is defined as the rate of mistaken positive identification). To evaluate the identification of covariate effects structure, we calculate the number of correct/incorrect discovery of homogeneity/heterogeneity. To be specific, correct discovery of homogeneity is defined as the number of elements in $\left\{j : \mu_{0j} \neq 0, \boldsymbol{\alpha}_{0j} = \mathbf{0}, \widehat{\mu}_j \neq 0, \widehat{\boldsymbol{\alpha}}_j = \mathbf{0}\right\}$, referred to as HomoCorrect. Incorrect discovery of homogeneity is defined as the number of elements in $\left\{j : \boldsymbol{\alpha}_{0j} \neq \mathbf{0}, \widehat{\mu}_j \neq 0, \widehat{\boldsymbol{\alpha}}_j = \mathbf{0}\right\} \cup \left\{j : \mu_{0j} = 0, \widehat{\mu}_j \neq 0, \widehat{\boldsymbol{\alpha}}_j = \mathbf{0}\right\}$, referred to as HomoIncorrect. Correct discovery of heterogeneity is defined as the number of elements in $\left\{j : \boldsymbol{\alpha}_{0j} \neq \mathbf{0}, \widehat{\boldsymbol{\alpha}}_j \neq \mathbf{0}\right\}$, referred to as HeteCorrect. Incorrect discovery of heterogeneity is defined as the number of elements in $\left\{j : \boldsymbol{\alpha}_{0j} = \mathbf{0}, \widehat{\boldsymbol{\alpha}}_j \neq \mathbf{0}\right\}$, referred to as HeteIncorrect.

We set the number of studies $K = 4$ and total sample size $n = 800$. We consider a balanced design with all individual sample sizes = 200 and an unbalanced design with individual sample sizes = 50, 150, 300, 300. For the dimension of covariates, we consider $p \in \{400, 800\}$. The covariates $\boldsymbol{Z}^{(k)}_i$ are generated from multivariate normal distributions with marginal means 0, variances 1, and correlations $\text{cov}(Z^{(k)}_{ij}, Z^{(k)}_{il}) = 0.5^{|j-l|}$ for $i = 1, 2, ..., n_k, k = 1, 2, \ldots, K$, and $j = 1, 2, \ldots, p$. The survival times are generated from the Cox model with hazard:

$$\lambda^{(k)}(t|\boldsymbol{Z}^{(k)}_i) = (t - 0.5)^2 \exp\left\{\boldsymbol{\beta}^{(k)\mathrm{T}} \boldsymbol{Z}^{(k)}_i\right\}. \tag{10}$$

The censoring times $C^{(k)}_i$ are generated from uniform distributions on $[c/2, c]$, where $c > 0$ is adjusted to achieve censoring rate (CR) 30% and 60%. For the covariate effects, we consider three scenarios:

(S1) 12 covariates have homogeneous effects, and the rest have zero effects. Specifically, $\boldsymbol{\beta}^{(1)}_0 = \boldsymbol{\beta}^{(2)}_0 = \boldsymbol{\beta}^{(3)}_0 = \boldsymbol{\beta}^{(4)}_0 = (0.5 \times \mathbf{1}_{1\times6}, -0.5 \times \mathbf{1}_{1\times6}, \mathbf{0}_{1\times(p-12)})^{\mathrm{T}}$;

(S2) 6 covariates have homogeneous effects, 6 covariates have heterogeneous effects, and the rest have zero effects. Specifically, $\boldsymbol{\beta}^{(1)}_0 = \boldsymbol{\beta}^{(2)}_0 = (0.5 \times \mathbf{1}_{1\times6}, -0.5 \times \mathbf{1}_{1\times6}, \mathbf{0}_{1\times(p-12)})^{\mathrm{T}}$, $\boldsymbol{\beta}^{(3)}_0 = \boldsymbol{\beta}^{(4)}_0 = (0.5 \times \mathbf{1}_{1\times12}, \mathbf{0}_{1\times(p-12)})^{\mathrm{T}}$;

(S3) 12 covariates have heterogenous effects, and the rest have zero effects. Specifically, $\boldsymbol{\beta}^{(1)}_0 = (0.5 \times \mathbf{1}_{1\times6}, -0.5 \times \mathbf{1}_{1\times6}, \mathbf{0}_{1\times(p-12)})^{\mathrm{T}}$, $\boldsymbol{\beta}^{(2)}_0 = (-0.5 \times$

$\mathbf{1}_{1\times 6}, 0.5\times\mathbf{1}_{1\times 6}, \mathbf{0}_{1\times(p-12)})^{\mathrm{T}}$, $\boldsymbol{\beta}_0^{(3)} = (0.5\times\mathbf{1}_{1\times 6}, -0.5\times\mathbf{1}_{1\times 6}, \mathbf{0}_{1\times(p-12)})^{\mathrm{T}}$, $\boldsymbol{\beta}_0^{(4)} = (-0.5\times\mathbf{1}_{1\times 6}, 0.5\times\mathbf{1}_{1\times 6}, \mathbf{0}_{1\times(p-12)})^{\mathrm{T}}$.

Based on 800 replicates, we summarize the AEE, TPR, and FDR results in Table 4. The results of homogeneity pursuit using the proposed method are summarized in Table 5. Additionally, to assess the asymptotic normality results, we calculate standard error (SE) and standard deviation (SD) for the estimated parameters. Noting that the numbers of parameters under different homogeneity settings are different, we show the average SE and SD values of all parameters in Table 6. Additionally, to demonstrate that estimation gets better with a larger sample size, we also consider scenario (S3) with $p = 800$, $K = 8$, a balanced sample size setting (with all individual sample sizes 200), and an unbalanced sample size setting (with two individual sample sizes 50, two 150, and four 300). Beyond the tables, we also show the AEE, FDR, and TPR results graphically in Figs. 1, 2, and 3, respectively.

We make the following observations. (i) In Table 4, for AEE, it can be observed that the proposed method is the closest to the oracle. Under the proposed analysis, Lasso has inferior performance compared to MCP, because of its inherent biasedness. When all covariate effects are homogeneous and the sample sizes are balanced, dLasDC has performance comparable to the proposed. However, it performs badly when the sample sizes are unbalanced, which can be explained by its dependence on local estimators. The proposed method gets benefit from its overall surrogate loss function which approximates the overall likelihood function. Thus it is not so sensitive to the unbalanced design. In general, AEE gets larger when CR is higher, sample sizes are unbalanced, or heterogeneity level gets higher. (ii) In Table 4, for variable selection, it can be seen that the proposed method performs well under all settings, with high TPR and low FDR. The proposed analysis with the alternative Lasso penalty has good TPR but relatively high FDR. SMA, dLasDC, and LasVote perform unsatisfactorily when the sample sizes are unbalanced. It is reasonable as the penalty weights of SMA significantly depend on local estimators, and the other two methods are sensitive to local estimators. In general, performance gets worse when CR is higher, heterogeneity level gets higher, or sample sizes are unbalanced. (iii) For structure identification, in Table 5, it can be seen that under all scenarios, both PMCP and PLAS behave well and can effectively recover the homogeneity structure. PMCP outperforms with smaller false discovery. (iv) From Figs. 1, 2, and 3, it can be seen that performance gets better when there are more datasets/samples. (v) It can be seen from Table 6 that the proposed SE is close to SD. Overall, we conclude competitive performance of the proposed method. We have also examined a few other settings and made similar observations.

To evaluate the robustness of the proposed approach to the model misspecification, we conduct the following numerical experiment. We generate the data from the hazard models as follow:

$$\lambda^{(k)}\left(t \middle| \boldsymbol{Z}^{(k)}\right) = \lambda_0^{(k)}(t)\exp\left(\boldsymbol{g}(\boldsymbol{Z}^{(k)})^{\mathrm{T}}\boldsymbol{\beta}^{(k)}\right), k = 1, 2, \ldots, K,$$

where $\boldsymbol{g}(\boldsymbol{Z}) = \left(g_j\left(Z_j\right), j = 1, 2, \ldots, p\right)$, $\lambda_0^{(k)}(t) = (t-0.5)^2$ and

Setting 1: $g_j(x) = -1 + 2/\left(1+\exp\left(-2x\right)\right), j = 1, 2, \ldots, p$,

Setting 2: $g_j(x) = x^3/3, j = 1, 2, \ldots, p$,

Table 1 Data analysis using the proposed method: estimated coefficient and stability

Genes	Estimation	Stability/%	Genes	Estimation	Stability/%	Genes	Estimation	Stability/%
TESC	– 0.68	67	FOXD1	3.44	38	HOXA1	2.28	22
SNX31	– 1.16	61	RRAGD	2.05	38	MCOLN3	– 0.35	22
TM4SF1	0.51	60	ABCC4	4.06	37	APEX2	– 11.2	21
CRIP1	0.38	58	IKZF2	2.45	37	OLFML2B	1.07	21
ABCA5	4.65	52	MSX2	– 3.21	37	KLHL13	– 3.23	20
HSPA7///HSPA6	0.94	50	CXCR4	– 1.65	36	POC1B	– 1.07	20
POSTN	0.54	49	NABP2	– 1.88	35	TLE4	– 0.33	20
EMP1	3.27	48	FAM149A	– 2.42	34	NOD2	0.22	17
SALL4	1.12	48	HMGCS1	3.37	34	THUMPD1	– 6.59	17
TAGLN	– 2.56	48	CRISPLD2	0.80	33	API5	– 3.02	16
BAIAP2L2	3.27	46	CNN3	2.56	31	SLC3A2	7.87	15
CDC25B	0.46	44	NME5	– 2.26	30	COL8A2	5.59	13
SNCAIP	0.92	44	POF1B	5.37	30	ARHGAP26	– 2.59	12
FZD2	– 0.23	43	COL3A1	1.79	29	PMPCB	– 3.26	10
SLC51A	– 4.93	43	TROAP	1.82	26	RBM24	– 2.61	9
SYT17	4.60	41	C7orf10	1.04	24			
ZBTB7C	1.60	40	CDK4	9.03	23			

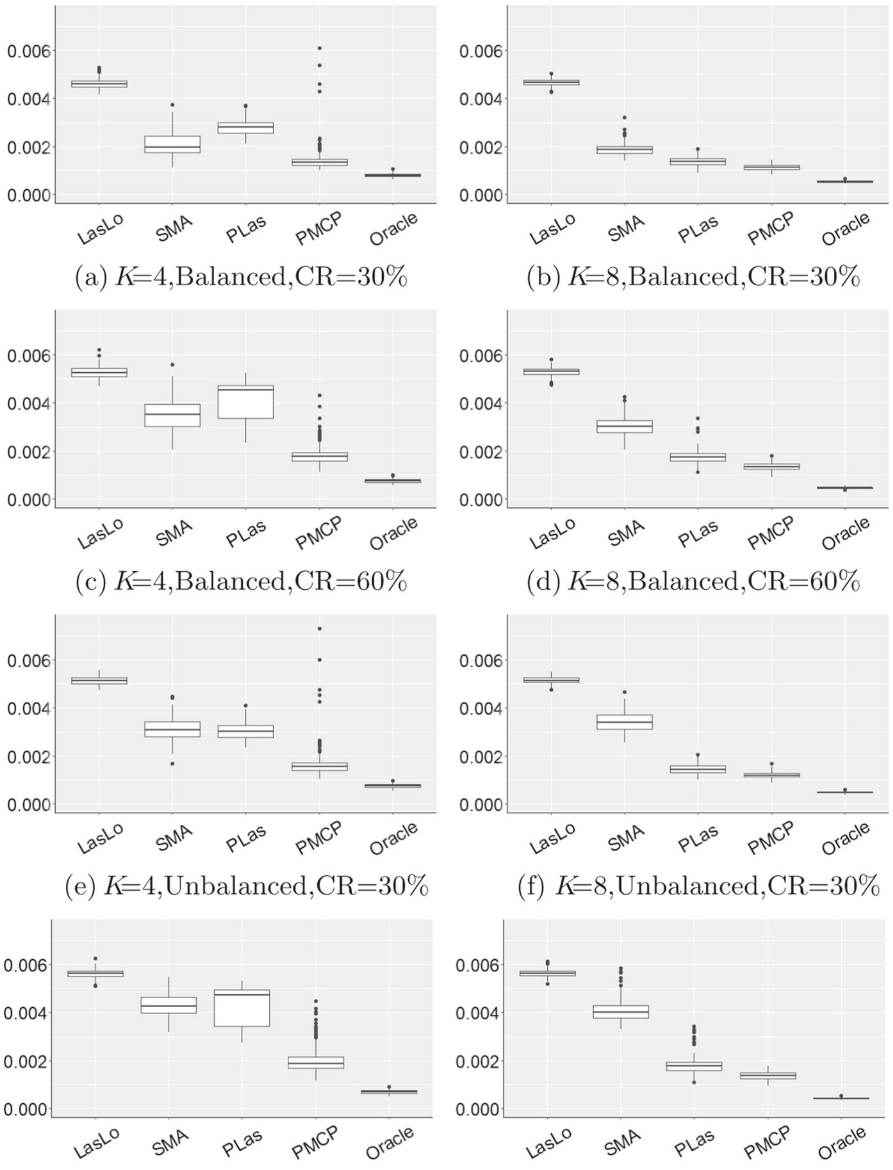
Table 2 Data analysis using SMA: estimated coefficient and stability

Gene	GSE13507		GSE31684		GSE32894	
	Estimate	Stability/%	Estimate	Stability/%	Estimate	Stability/%
AAR2	10.4	16	− 14.9	13	14.5	16
ABCA5	0	–	3.71	19	0	–
ALDH1L1-AS2	− 11.8	12	0	–	0	–
AP2S1	0	–	11.9	35	14.3	33
C7orf26	3.55	1	−3.89	8	0	–
COL3A1	0	–	2.89	10	6.44	11
DHCR24	3.65	22	0.54	15	0	–
FHIT	0	–	8.09	3	0	–
MAF1	− 5.52	0	− 1.78	1	0	–
NABP2	− 1.02	4	0	–	9.95	5
RNF24	6.59	17	0	–	16.8	4
TBC1D1	− 9.14	16	− 0.07	2	0	–
TLE4	0	–	− 3.30	5	0	–
UBE2R2	− 6.41	14	8.66	13	− 29.4	1
ZNF2	− 21.9	29	0	–	0	–

Table 3 Estimation and stability of pooled data analysis

Genes	Estimation	Stability/%	Genes	Estimation	Stability/%	Genes	Estimation	Stability/%
CDK4*	4.08	100	KLHL13*	– 0.89	84	RBM24	– 0.57	67
COL1A1	0.75	100	CRISPLD2*	0.24	81	CA12	– 0.74	66
IKZF2*	1.40	100	MCOLN3	– 0.46	81	ARL14	– 0.52	64
TM4SF1*	1.10	100	FOXD1*	0.91	80	CALD1	0.11	62
ABCC4*	– 0.84	99	C7orf10*	0.66	78	CRIP1*	0.30	60
DHCR24	1.16	99	FAM149A*	– 0.90	78	PTRF	0.14	60
NME5*	– 1.03	99	HOXA1*	0.82	78	SNX31*	– 0.42	57
S100A8	0.20	99	OLFML2B*	0.85	78	HMGCS1*	0.90	56
COL3A1*	0.49	98	CPED1	– 0.87	77	TESC*	– 0.24	56
POF1B*	1.26	98	SLC51A*	– 1.10	77	NOD2*	0.58	55
LDLR	0.42	97	SNCAIP*	0.81	76	TROAP*	0.54	54
POSTN*	0.77	97	ABCA5*	1.07	74	CDC25B*	0.30	51
TMEM154	– 1.16	97	PLEKHA7	0.57	73	FSTL1	0.40	50
COL5A1	0.24	95	ZBED2	0.37	73	ZBTB7C*	0.51	48
COL5A2	0.38	95	TLE4*	– 0.77	71	APEX2*	– 0.73	45
LRRC49	– 0.34	95	DOCK2	– 0.41	70	TRIP10	0.31	38
SYT17*	0.94	94	SMCO4	– 0.55	70	MRAS	– 0.46	35
PDGFRB	0.15	87	FZD2	– 0.80	69	SPOPL	– 0.55	34
TNFRSF1B	– 0.91	87	HMHA1	– 0.84	68	EMP1*	0.14	32

The genes which are also selected by the proposed method are marked with ’*’



(a) K=4,Balanced,CR=30%

(b) K=8,Balanced,CR=30%

(c) K=4,Balanced,CR=60%

(d) K=8,Balanced,CR=60%

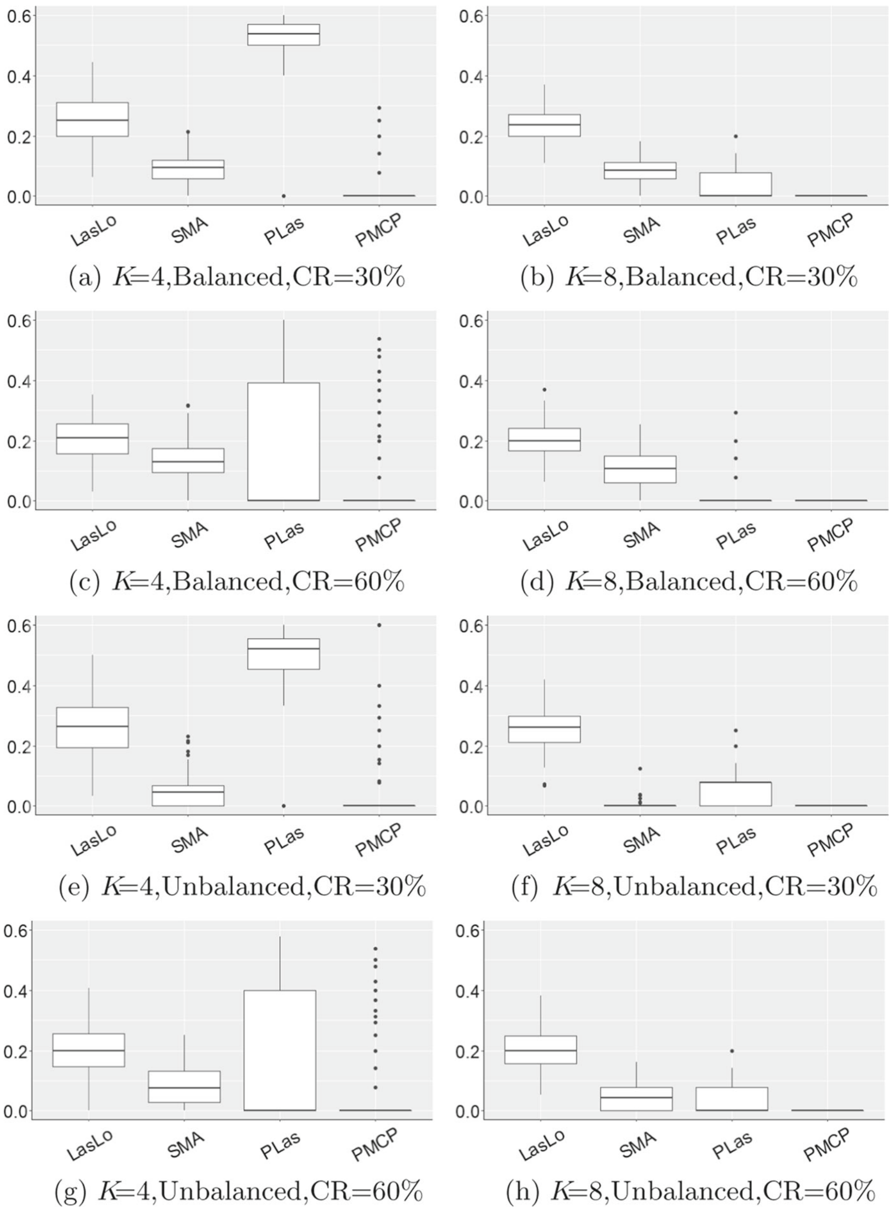
(e) K=4,Unbalanced,CR=30%

(f) K=8,Unbalanced,CR=30%

(g) K=4,Unbalanced,CR=60%

(h) K=8,Unbalanced,CR=60%

Fig. 1 Boxplots of AEEs based on simulation replicates when $p = 800$ in scenario (S3). The x-axis displays the listed methods. The y-axis is the value of AEE



(a) K=4,Balanced,CR=30%

(b) K=8,Balanced,CR=30%

(c) K=4,Balanced,CR=60%

(d) K=8,Balanced,CR=60%

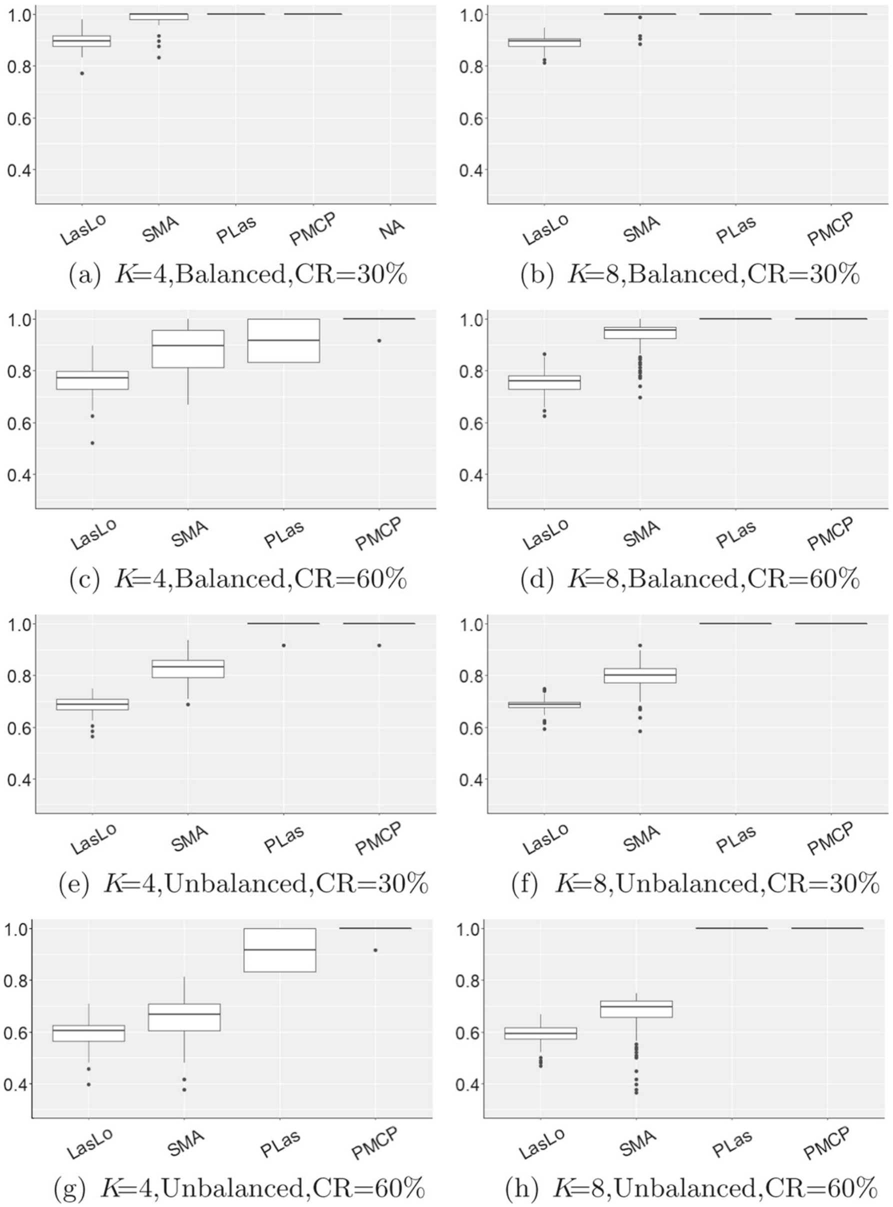
(e) K=4,Unbalanced,CR=30%

(f) K=8,Unbalanced,CR=30%

(g) K=4,Unbalanced,CR=60%

(h) K=8,Unbalanced,CR=60%

Fig. 2 Boxplots of FDRs based on simulation replicates when $p = 800$ in scenario (S3). The x-axis displays the listed methods. The y-axis is the value of FDR



(a) K=4,Balanced,CR=30%

(b) K=8,Balanced,CR=30%

(c) K=4,Balanced,CR=60%

(d) K=8,Balanced,CR=60%

(e) K=4,Unbalanced,CR=30%

(f) K=8,Unbalanced,CR=30%

(g) K=4,Unbalanced,CR=60%

(h) K=8,Unbalanced,CR=60%

Fig. 3 Boxplots of TPRs based on simulation replicates when $p = 800$ in scenario (S3). The x-axis displays the listed methods. The y-axis is the value of TPR

Setting 3: $g_j(x) = x^3/3$, $j = 1, 2, \ldots, p$, $g_j(x)$ be randomly selected from

$$\left\{-1 + \frac{2}{(1 + \exp(-2x))}, \frac{x^3}{3}, -1 + \frac{(x-1)^2}{4}, |Z| - 1\right\}$$

for each $j = 1, 2, \ldots, p$. We set the number of studies be $K = 4$, total sample size be $n = 800$, dimension of covariates be $p = 400$, and censoring rate be around 30%. We consider senario (S3) and consider a balanced/unbalanced design same as that in the preceding part. We still use the (10) for estimation (i.e., it serves as working model). The simulation results are shown in Table 7. It can be seen that the model we use can handle Setting 1 and 2 well. The result of Setting 3 is tolerable. It can be concluded that our method will be affected if model is mis-specified. However, it shows some robustness if the deviations of correct model specification are not extreme.

4 Data analysis

High-throughput profiling is now common, and genetic measurements have been extensively associated with disease survival. Published integrative analysis studies suggest that, for many common survival problems, there are multiple independent data sources. Although there has been significant effort in sharing raw genetic data, this is still not routine, with concerns on violating privacy (Qin et al. 2020). It has been well argued that it is critical to preserve privacy in genetic data analysis (Erlich and Narayanan 2014).

In our analysis, we consider genetic data from three independent sources: GSE13507 collected in South Korea, GSE31684 collected in America, and GSE32894 collected in Sweden. The information about these datasets can be found on the website of Gene Expression Omnibus (GEO). In all the three studies, the goal is to identify gene expressions that are associated with all-cause mortality following a diagnosis of bladder cancer. In these three datasets, the index "sample type" records patient's type of illness. They contain 165, 93, and 308 samples with sample type be "tumor", respectively. The overall proportion of missing data is 14.8%. Since the missing rate is relatively small, we exclude the incomplete samples. There are 14,025 gene expressions commonly measured in all studies. The censoring rates are 58%, 30%, and 89% respectively. To obtain more reliable estimation (and also reduce computational cost), we conduct a screening in each dataset and select $n/log(n)$ candidate genes as Fan and Lv (2008) recommended. Then we take the union of these genes from each datasets as final candidates. For the specific datasets we use, we finally select 300 genes for downstream analysis.

The estimation results of the proposed method and SMA are shown in Tables 1 and 2, respectively. Analysis using the proposed method suggests that the covariate effects are homogeneous. Genes ABCA5, COL3A1, NABP2, and TLE4 are selected by both the proposed method and SMA. The other genes selected by the two methods are different. To test the fitting effect of these methods, we follow the procedure: for the kth dataset, firstly, we divide it into two subsets by the values of $\left\{\widehat{\boldsymbol{\beta}}^{(k)\mathrm{T}} \boldsymbol{Z}_i^{(k)}, i = 1, 2, \ldots, n_k\right\}$.

Then we apply the logrank test on the survival observations of these two subsets. The reason for this procedure is that the estimate with better fitting effect to the survival data should seperate the survival curves of these two subsets well, that is, with more significant p-value in logrank test. With the proposed method, the logrank p-values are 5.21×10^{-6}, 1.45×10^{-10}, and 3.18×10^{-8}, respectively, for the three datasets. In comparison, the SMA p-values are 2.55×10^{-7}, 1.13×10^{-3}, and 1.70×10^{-7}, respectively. It can be seen that the proposed method has better fitting effect, especially for GSE31684 which contains the least number of samples among these three datasets.

A quick literature search suggests that the identified genes can be biologically sensible. Low expression of gene TESC was found to have negtive effect on overall survival in cancer patients (Zhang et al. 2022). Gene SNX31 has been associated with bladder urothelium (Vieira et al. 2014). TM4SF1 gene has been reported to be overexpressed in breast, ovarian, lung, pancreatic, prostate, and colon carcinomas (Wang et al. 2015). Its overexpression was also associated with poor survival in patients with glioma (Wang et al. 2015). Abnormal expression of CRIP1 has been identified in numerous solid tumors, and its overexpression was found to be related with shorter overall survival (Ma et al. 2020). Upregulation of HSPA7 was related to poor survival in colon cancer (Guan et al 2021). Overexpression of COL3A1 confers a poor prognosis in human bladder cancer (Yuan et al. 2017).

With real data, it is difficult to evaluate identification performance. Since we actually have access to all datasets, the results from pooled data may provide support for identification. In addition, we randomly select 90% of the samples from each dataset, conduct the proposed estimation, and record whether the genes are selected. This procedure is repeated 100 times, and we calculate the probability of each gene being identified across replicates. Higher probabilities can suggest more stable identification. The results from pooled data are shown in Table 3. The stability results of the listed methods are provided in Tables 1, 2 and 3, respectively. In Table 3, the genes which are also selected by the proposed method are marked with ‘*’. It can be seen that the results from pooled data analysis are more stable and the selected genes share great similarity with that from the proposed method. The results from the proposed method are less stable than that from pooled data analysis, but more stable than SMA. These results can further support the proposed method.

5 Discussion

In this article, we have developed a new integrative analysis method for data with a censored survival outcome and high-dimensional covariates that can preserve data privacy and automatically determine the homogeneity (versus heterogeneity) structure in covariate effects. The proposed method has an intuitive formulation and can be easily extended to other survival models and models for other types of outcomes. It has been shown to have the well-desired consistency properties under high-dimensional settings, and it is noted that our theoretical developments may also shed insights into other integrative analyses. Simulation has demonstrated the practical advantage of the proposed method, and the analysis of gene expression data on bladder cancer has demonstrated its practical utility.

Table 4 Simulation results of parameter estimation

Scenario	p	Method	Balanced						Unbalanced					
			CR = 30%			CR = 60%			CR = 30%			CR = 60%		
			AEE	TPR	FDR	AEE	TPR	FDR	AEE	TPR	FDR	AEE	TPR	FDR
S1	400	Oracle	1.31	–	–	1.53	–	–	1.33	–	–	1.54	–	–
		PMCP	1.90	100.00	0.02	2.67	99.29	0.12	1.99	100.00	0.09	2.64	99.10	0.09
		PLAS	4.05	100.00	7.91	5.59	99.88	5.87	4.15	100.00	8.24	5.51	99.96	5.99
		SMA	3.81	95.61	1.87	5.54	89.23	2.69	5.18	86.47	1.70	7.93	68.83	10.76
		dLasDC	2.24	100.00	0.08	2.89	99.25	1.47	7.68	86.58	10.47	5.65	85.04	7.36
		LasVote	8.41	95.83	0.83	9.74	86.59	0.39	9.69	96.86	0.33	10.64	87.41	0.32
		LasLo	8.71	91.69	23.82	10.02	79.54	19.70	9.89	70.07	23.45	10.82	62.04	20.47
	800	Oracle	0.65	–	–	0.74	–	–	0.66	–	–	0.76	–	–
		PMCP	1.04	100.00	0.14	1.43	99.41	0.04	1.03	100.00	0.13	1.35	99.43	0.02
		PLAS	2.13	100.00	8.51	2.82	100.00	6.52	2.13	100.00	8.51	2.84	100.00	9.10
		SMA	2.09	93.38	2.29	2.88	87.48	3.02	3.13	79.26	2.20	4.08	67.51	14.21
		dLasDC	1.15	99.83	0.08	1.58	97.75	2.15	3.95	87.25	12.75	2.84	84.29	8.08
		LasVote	4.55	92.02	0.26	5.19	84.24	0.13	5.08	94.51	0.19	5.57	84.88	0.14
		LasLo	4.67	88.90	22.84	5.31	75.56	19.98	5.16	68.43	24.49	5.64	59.16	19.62

Table 4 continued

Scenario	p	Method	Balanced						Unbalanced					
			CR = 30%			CR = 60%			CR = 30%			CR = 60%		
			AEE	TPR	FDR	AEE	TPR	FDR	AEE	TPR	FDR	AEE	TPR	FDR
S2	400	Oracle	1.47	–	–	1.51	–	–	1.41	–	–	1.46	–	–
		PMCP	2.20	99.99	0.20	2.90	99.55	0.33	2.60	99.92	0.33	3.17	99.44	0.38
		PLAS	5.64	100.00	5.39	7.58	99.33	4.95	6.13	100.00	6.82	7.86	99.63	5.26
		SMA	3.35	99.84	2.13	5.08	96.03	3.57	5.12	88.73	2.43	7.53	73.07	1.08
		LasLo	8.08	95.54	22.75	9.29	87.00	19.76	9.63	70.79	22.49	10.51	64.21	19.92
	800	Oracle	0.72	–	–	0.75	–	–	0.70	–	–	0.72	–	–
		PMCP	1.14	100.00	0.24	1.54	99.43	0.16	1.34	99.99	0.79	1.63	99.54	0.30
		PLAS	2.92	100.00	5.85	3.88	99.50	6.91	3.02	100.00	8.62	3.92	99.85	7.88
		SMA	1.69	99.84	2.66	2.60	95.43	4.27	2.59	88.38	2.96	3.87	72.16	1.95
		LasLo	4.34	94.12	21.91	4.96	83.69	19.12	5.03	69.35	22.94	5.47	61.83	19.03
S3	400	Oracle	1.58	–	–	1.49	–	–	1.47	–	–	1.37	–	–
		PMCP	2.75	100.00	2.61	3.61	100.00	3.51	3.41	100.00	5.05	4.09	100.00	5.32
		PLAS	4.85	100.00	44.76	7.82	92.92	15.55	5.47	100.00	42.61	8.26	94.17	15.38
		SMA	3.93	98.44	7.03	6.51	91.03	11.10	5.82	85.61	4.76	8.28	66.75	5.71
		LasLo	8.72	91.76	23.53	10.01	79.64	19.71	9.88	70.14	23.30	10.83	61.99	20.02
	800	Oracle	0.79	–	–	0.74	–	–	0.74	–	–	0.68	–	–
		PMCP	1.36	100.00	0.65	1.80	99.96	1.64	1.62	99.98	1.81	1.97	99.95	2.10
		PLAS	2.78	100.00	53.56	4.16	91.83	15.66	3.04	99.92	51.62	4.28	93.17	17.52
		SMA	2.11	97.66	9.31	3.49	88.50	13.39	3.21	82.75	4.72	4.41	65.28	8.33
		LasLo	4.67	89.02	23.54	5.29	75.80	20.29	5.16	68.37	24.57	5.64	58.99	19.82

The unit of AEE is 10^{-3}, and the unit TPR and FDR is %. CR represents censoring rate

Table 5 Simulation results of homogeneity/heterogeneity discovery

Scenario	Type	Method	Balanced				Unbalanced			
			CR = 30%		CR = 60%		CR = 30%		CR = 60%	
			p = 400	800	p = 400	800	p = 400	800	p = 400	800
S1	HomoCorrect(12)	PMCP	12.00	12.00	11.92	11.93	12.00	12.00	11.89	11.88
		PLAS	12.00	12.00	11.91	11.90	12.00	12.00	11.90	11.87
	HomoIncorrect(0)	PMCP	0.00	0.02	0.02	0.01	0.01	0.02	0.01	0.00
		PLAS	1.11	1.19	0.81	0.90	1.16	1.74	0.82	1.29
	HeteIncorrect(0)	PMCP	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		PLAS	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
S2	HomoCorrect(6)	PMCP	6.00	6.00	5.95	5.94	6.00	6.00	5.95	5.95
		PLAS	6.00	6.00	6.00	6.00	6.00	6.00	6.00	6.00
	HomoIncorrect(0)	PMCP	0.02	0.03	0.03	0.02	0.03	0.06	0.02	0.03
		PLAS	0.75	0.80	0.67	0.95	0.94	1.21	0.71	1.09
	HeteCorrect(6)	PMCP	6.00	6.00	6.00	5.99	5.99	6.00	5.99	6.00
		PLAS	6.00	6.00	5.92	5.94	6.00	6.00	5.96	5.98
	HeteIncorrect(0)	PMCP	0.01	0.01	0.02	0.01	0.02	0.06	0.03	0.02
		PLAS	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
S3	HomoIncorrect(0)	PMCP	0.01	0.00	0.04	0.03	0.02	0.02	0.03	0.04
		PLAS	0.15	0.07	0.04	0.01	0.08	0.06	0.01	0.01
	HeteCorrect(12)	PMCP	11.65	11.58	10.79	10.46	11.31	11.10	10.42	10.18
		PLAS	12.00	12.00	11.15	11.02	12.00	11.99	11.30	11.18
	HeteIncorrect(0)	PMCP	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00
		PLAS	10.15	14.39	2.95	3.48	9.37	13.36	2.89	3.72

The number in the parenthesis is the ideal result where $\widehat{\boldsymbol{\mu}} = \boldsymbol{\mu}_0, \widehat{\boldsymbol{\alpha}} = \boldsymbol{\alpha}_0$. CR represents censoring rate

Table 6 Simulation results of average SE and SD for all estimated coefficients

$p = 400$					$p = 800$				
Scenario	Balanced	CR	SE×10^3	SD×10^3	Scenario	Balanced	CR	SE×10^3	SD×10^3
S1	Yes	30%	56.70	52.30	S1	Yes	30%	56.68	50.71
		60%	75.56	76.54			60%	75.18	68.33
	No	30%	56.75	54.62		No	30%	56.68	52.53
		60%	75.57	79.36			60%	75.26	70.07
S2	Yes	30%	87.93	84.21	S2	Yes	30%	87.34	79.49
		60%	116.12	106.03			60%	115.87	105.23
	No	30%	87.89	88.58		No	30%	87.24	81.52
		60%	116.79	105.38			60%	116.44	102.19
S3	Yes	30%	101.21	93.62	S3	Yes	30%	101.10	92.72
		60%	136.14	121.98			60%	135.58	121.22
	No	30%	101.43	95.36		No	30%	101.13	94.23
		60%	136.38	121.17			60%	136.14	122.49

CR represents censoring rate

Table 7 Simulation results under deviations of correct model specification

Setting	Balanced			Unbalanced		
	1	2	3	1	2	3
AEE/10^{-3}	6.70	6.45	8.48	7.00	6.69	8.75
TPR/10^{-2}	94.90	94.95	80.55	95.33	94.89	79.38
FDR/10^{-2}	2.07	3.64	3.10	5.16	5.92	4.21
HomoCorrect(6)	5.44	5.43	4.10	5.46	5.41	4.01
HomoIncorrect(0)	0.00	0.00	0.05	0.00	0.00	0.02
HeteCorrect(6)	5.94	5.95	5.33	5.92	5.94	5.15
HeteIncorrect(0)	0.05	0.15	0.30	0.38	0.46	0.52

This study can be potentially extended in multiple directions. An “easy” extension is to consider alternative models, e.g., accelerate failure time model, transformed model. This study has considered the scenario where all data sources only provide summary statistics, while some published studies have considered the other extreme with all raw data available. A more realistic scenario is to have summary statistics from some sources while raw data from other sources. In this case, the overall loss function should be the combination of both summary-statistics-based and log-likelihood-based loss functions. In our and published studies, all available data sources have been integrated. When some data sources are overly different from the others, it may be desirable not to integrate them. In a sense, it may be of interest to identify heterogeneity at the data source level. A hypothesis test to rule out significantly different datasets should be considered before integrative analysis.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00362-023-01470-9.

Acknowledgements We thank the editor, associate editor, and reviewers for their careful review and insightful comments, which have led to a significant improvement of the article. Yanyan Liu gratefully acknowledges research support from the National Nature Science Foundation of China [11971362]. Shuangge Ma was financially supported by the Natural Science Foundation [2209685].

References

Battey H, Fan J, Liu H, Lu J et al (2018) Distributed testing and estimation under sparse high dimensional models. Ann Stat 46(3):1352–1382

Cai T, Liu M, Xia Y (2021) Individual data protected integrative regression analysis of high-dimensional heterogeneous data. J Am Stat Assoc 117:2105–2119

Chen X, Xie M (2014) A split-and-conquer approach for analysis of extraordinarily large data. Stat Sin 24:1655–1684

Cheng X, Lu W, Liu M (2015) Identification of homogeneous and heterogeneous variables in pooled cohort studies. Biometrics 71:397–403

Danieli C, Moodie E (2021) Preserving data privacy when using multi-site data to estimate individualized treatment rules. Stat Med 41:1627–1643

Erlich Y, Narayanan A (2014) Routes for breaching and protecting genetic privacy. Nat Rev Genet 15:409–421

Jordan M, Lee J, Yang Y (2019) Communication-efficient distributed statistical inference. J Am Stat Assoc 114(526):668–681

Fan J, Peng H (2004) Nonconcave penalized likelihood with a diverging number of parameters. Ann Stat 32(3):928–961

Fan J, Lv J (2008) Sure independence screening for ultrahigh dimensional feature space. J R Stat Soc B 32(3):928–961

Gomatam S, Karr A, Reiter J, Sanil A (2005) Data dissemination and disclosure limitation in world without microdata: a risk-utility framework for remote access analysis servers. Stat Sci 20:163–177

Guan Y, Zhu X, Liang J, Wei M et al (2021) Upregulation of HSPA1A/HSPA1B/HSPA7 and downregulation of HSPA9 were related to poor survival in colon cancer. Front Oncol 11:749673

He Q, Zhang H, Avery C, Lin D (2016) Sparse meta-analysis with high-dimensional data. Biostatistics 17:205–220

He B, Zhong T, Huang J, Liu Y et al (2020) Histopathological imaging-based cancer heterogeneity analysis via penalized fusion with model averaging. Biometrics 77:1397–1408

Huang Y, Liu J, Yi H, Shia B et al (2017) Promoting similarity of model sparsity structures in integrative analysis of cancer genetic data. Stat Med 36:509–559

Homer N, Szelinger S, Redman M, Duggan D et al (2008) Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genet 4(8):e1000167

Karr A, Fulp W, Vera F, Young S et al (2007) Secure, privacy-preserving analysis of distributed databases. Technometrics 49(3):335–345

Li W, Liu H, Yang P, Xie W (2016) Supporting regularized logistic regression privately and efficiently. PLoS ONE 11(6):1037–1057

Lee K, Chakraborty S, Sun J (2011) Bayesian variable selection in semiparametric proportional hazards model for high dimensional survival data. Int J Biostat 7(1):21

Lee J, Liu Q, Sun Y, Taylor J (2017) Communication-efficient sparse regression. J Mach Learn Res 18:1–30

Li D, Lu W, Shu D, Toh S et al (2022) Distributed Cox proportional hazards regression using summary-level information. Biostatistics 24:776–794

Liu J, Huang J, Xie Y, Ma S (2013) Sparse group penalized integrative analysis of multiple cancer prognosis datasets. Genet Res 95:68–77

Liu J, Ma S, Huang J (2014) Integrative analysis of cancer diagnosis studies with composite penalization. Scand J Stat 41:87–103

Lu C, Wang S, Ji Z, Wu Y et al (2015) WebDISCO: a web service for distributed cox model learning without patient-level data sharing. J Am Med Inform Assoc 22(6):1212–1219

Ma B, Zhang T, Wang C, Xu Z et al (2020) Methylation-independent CRIP1 expression is a potential biomarker affecting prognosis in cytogenetically normal acute myeloid leukemia. Am J Transl Res 12(9):4840–4852

Moodie E, Coulombe J, Danieli C, Renoux C et al (2022) Privacy-preserving estimation of an optimal individualized treatment rule: a case study in maximizing time to severe depression-related outcomes. Life Time Data Anal 28(3):512–542

Qin S, Zhou F, Zhang Z, Xu Z et al (2020) Privacy-preserving substring search on multi-source encrypted gene data. IEEE Access 99:50472–50484

Shu D, Yoshida K, Fireman B, Toh S (2020) Inverse probability weighted Cox model in multi-site studies without sharing individual-level data. Stat Methods Med Res 29(6):1668–1681

Tang L, Zhou L, Song P (2018) Fusion learning algorithm to combine partially heterogeneous Cox models. Comput Stat 34(1):395–414

Vieira N, Deng F, Liang F, Liao Y et al (2014) SNX31: a novel sorting nexin associated with the uroplakin-degrading multivesicular bodies in terminally differentiated urothelial cells. PLoS ONE 9(6):e99644

Walker E, Hernandez A, Kattan M (2008) Meta-analysis: its strengths and limitations. Clevel Clin J Med 75(6):431–439

Wang H, Leng C (2007) Unified lasso estimation by least squares approximation. J Am Stat Assoc 102:1039–1048

Wang Z, Wang C (2011) Buckley–James boosting for survival analysis with high-dimensional biomarker data. Stat Appl Genet Mol Biol 9(1):24

Wang L, Kim Y, Li R (2013) Calibrating nonconvex penalized regression in ultra-high dimension. Ann Stat 41(5):2505–2536

Wang X, Peng P, Dunson D (2014) Median selection subset aggregation for parallel inference. In: 28th conference on neural information processing systems (NIPS)

Wang P, Bao W, Zhang G, Deng Y et al (2015) Clinical significance of TM4SF1 as a tumor suppressor gene in gastric cancer. Neuroreport 26(8):455–461

Wang J, Kolar M, Zhang T (2016) Efficient distributed learning with sparsity. arXiv:1605.07991

Wolfson M, Wallace S, Masca N, Rowe G et al (2010) DataSHIELD: resolving a conflict in contemporary bioscience-performing a pooled analysis of individual-level data without sharing the data. Int J Epidemiol 39:1372–1382

Yang G, Huang J, Zhou Y (2014) Concave group methods for variable selection and estimation in high-dimensional varying coefficient models. Sci China-Math 31(1):243–267

Yu Y, Bradic J, Samworth R (2021) Confidence intervals for high-dimensional Cox models. Stat Sin 31(1):243–267

Yuan L, Shu B, Chen L, Qian K et al (2017) Overexpression of COL3A1 confers a poor prognosis in human bladder cancer identified by co-expression analysis. Oncotarget 8(41):70508–70520

Zhang C (2010) Nearly unbiased variable selection under minimax concave penalty. Ann Stat 38:894–942

Zhang Z, Huang L, Li J, Wang P (2022) Bioinformatics analysis reveals immune prognostic markers for overall survival of colorectal cancer patients: a novel machine learning survival predictive system. BMC Bioinform 23:124

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.