

Statistica Sinica Preprint No: SS-2025-0230	
Title	Two-way Subgroup Structure Learning for High Dimensional Clustered Survival Data
Manuscript ID	SS-2025-0230
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202025.0230
Complete List of Authors	Ye He, Dongsheng Tu, Liu Liu and Ling Zhou
Corresponding Authors	Ling Zhou
E-mails	zhouling@swufe.edu.cn
Notice: Accepted author version.	

Two-way subgroup structure learning for high dimensional clustered survival data

Ye He¹, Dongsheng Tu², and Liu Liu³, and Ling Zhou ⁴

Abstract: Identifying patient subgroups with heterogeneous clinical outcomes and varying treatment-biomarker interactions is crucial in clinical research. These two types of subgroups are referred to as subgroup analysis (SA) and homogeneity fusion (HF), respectively. Unlike existing methods that conduct SA or HF separately, simultaneously identifying two-way subgroup structures improves estimation and clustering accuracy, particularly in the presence of cluster heterogeneity and weak signals across high-dimensional biomarkers. In this paper, we propose a novel Double Centre-Augmented Penalties Fixed Effect Model(DCAF) for the simultaneous identification of two-way subgroups based on clustered survival data. Instead of conventional pairwise penalties, we introduce a differentiable centre-augmented harmonic-type penalty to facilitate subgroup identification. We establish theoretical results on estimation and selection consistency. Numerical studies, including simulations and two real data examples, demonstrate the effectiveness of the proposed DCAF method in accurately identifying two-way subgroups, as well as in selecting important biomarkers for each subgroup.

Key words and phrases: Double-centre-augmented, Subgroup analysis, Homogeneity fusion, High dimensional biomarkers, Theoretical property

1. Introduction

Heterogeneity in outcomes and treatment effects has long been recognized in complex diseases such as cancer, reflecting differences in disease severity and treatment response. Identifying patient subgroups based on genetic and clinical characteristics, referred to as biomarkers, is crucial for prognosis and personalized treatment. Recent reviews in Pathology highlight advances in subgroup identification for solid tumors (Cooper and Tan, 2024). Failing to account for these heterogeneous effects can result in biased estimates and predictions, as demonstrated by He et al. (2023).

In many statistical applications, survival data may be correlated. For example, our recent analyses involved data from two companion CCTG clinical trials: PA.3, a randomized trial comparing erlotinib plus gemcitabine (Treatment\Trt) with gemcitabine alone (Control) in patients with advanced pancreatic cancer (Moore et al., 2007), and CO.26, a recently completed randomized trial comparing combined immune checkpoint inhibition with best supportive care alone in patients with advanced colorectal cancer (Chen et al., 2020). These studies involved patients from multiple clinical centers, where heterogeneity in outcomes across centres is well-documented (Legrand et al., 2006). To account for the clustered structure of a clinical outcome, frailty (Vaupel et al., 1979) is commonly employed to describe sim-

ilarities in baseline risk within clusters, such as clinical centres, and the variability between clinical centres in terms of this outcome. Conventionally, such cluster heterogeneity was modeled through frailty, that is, a random effect with an associated distribution that induces within-cluster dependence; common choices included parametric random frailty, which assumed a specific distribution such as Gamma or log-normal (Hougaard, 1986; Therneau et al., 2003) but may suffer from distributional misspecification (Heckman and Singer, 1984), and discrete or non-parametric discrete frailty, in which the frailty remained a random variable taking a finite number of distinct values (Caroni et al., 2010; Guo and Rodriguez, 1992; Gasperoni et al., 2020). In contrast, we do not adopt a frailty (random-effect) formulation. Instead, we treat each cluster-specific intercept as a fixed effect and introduce a novel centre-augmented penalty that enables data-driven grouping of clusters. This penalized fixed-effects approach is distribution-free, thereby avoiding the misspecification risk of random-effect frailty models, while retaining interpretability through direct cluster-specific estimates. Unlike standard fixed-effects Cox models that treat each cluster independently, our penalty borrows information across clusters to achieve parsimonious grouping.

More importantly, we focus on identifying heterogeneity effects for clustered survival data. Specifically, heterogeneity considered from two di-

mensions, center heterogeneity and biomarker heterogeneity. The former, known as subgroup analysis (SA), aims to identify subgroup structures where patients within each subgroup exhibit similar treatment effects on Overall Survival(OS). The latter, known as homogeneity fusion (HF), aims to identify subgroups where biomarkers show similar effects on OS. There is a substantial body of literature focusing on either SA or HF. In the context of SA, Eng and Hanlon (2014) proposed a finite-mixture Cox proportional hazards (PH) model to classify patients into subgroups that are not known *a priori* based on multiple biomarkers, later extended by You et al. (2018) to high-dimensional settings to select important and subgroup-specific biomarkers via a variable selection method. Recently, a concave pairwise fusion approach for SA was invented by Ma and Huang (2017) and extended to a heterogeneous Cox model by Hu et al. (2021) and to a general class of heterogeneous transformation models by Han et al. (2024). However, mixture-model-based methods such as FMCox must estimate K separate nonparametric baseline hazard functions, which introduces additional variability and may lead to misspecification when the assumed mixture structure does not match the data. Pairwise fusion methods, on the other hand, require alternating direction method of multipliers (ADMM) optimization (Boyd et al., 2010) with $O(n^2)$ pairwise comparisons, mak-

ing them computationally burdensome as the number of clusters grows (He et al., 2023) and sensitive to initialization in practice. For HF, numerous studies have been conducted across different model frameworks; see for example, Tibshirani et al. (2005); Shen and Huang (2010); Ke et al. (2015); Zhu et al. (2013), among others. Existing methods typically require prior knowledge of subgroup structures or exhaustive pairwise comparisons. The former is often impractical and can lead to misleading conclusions, see Supplementary Materials (Supp) S4.7, while the latter is computationally intensive with $O(n^2)$ pairwise comparisons. A critical limitation shared by all the above approaches is that they address heterogeneity in only one dimension—either across centres (SA) or across biomarkers (HF)—and none incorporates two-way subgroup identification in a unified framework. Furthermore, no existing method provides a unified framework for clustered survival data that jointly performs data-driven cluster-specific fixed effect grouping, center-based subgroup analysis, biomarker-based homogeneity fusion, and high-dimensional variable selection.

In this paper, we fill these gaps by proposing a unified, data-driven framework to simultaneously uncover subgroup structures along both the center and biomarker dimensions in clustered survival data. The novelty and contributions of our proposed method are fourfold. *First*, to our knowl-

edge, this is the first framework that simultaneously identifies center-based prognostic subgroups (sharing common treatment, biomarker, and cluster-specific fixed effects) and biomarker-based predictive subgroups (exhibiting shared treatment-by-biomarker interactions), addressing a key limitation of existing approaches that focus on only one source of heterogeneity. This two-way structure enables the detection of weak but biologically relevant predictive signals that would be missed by conventional one-dimensional approaches. *Second*, we develop a Double Centre-Augmented Penalties Fixed Effect Model(DCAF) for cluster survival data based on a smooth K-Power Means penalty (Zhang et al., 1999; Xu and Lange, 2019) to identify both types of subgroup structures. Compared with mixture model based methods (e.g., FMCox), which must estimate multiple nonparametric baseline hazard functions and may suffer from model misspecification, our penalty-based approach requires no distributional assumptions on either the cluster-specific fixed effects or the subgroup structure, providing greater flexibility. Compared to pairwise fusion methods, our centre-augmented penalty involves only $O(nK)$ differences, reducing the complexity to linear and offering a far more scalable alternative. *Third*, the smoothness of the K-Power Means function enables efficient optimization via classical gradient-based methods, leading to fast implementation. Unlike classical K -means-type

methods or pairwise fusion with ADMM, the proposed algorithm is less sensitive to initialization, enhancing its robustness in practice. We further provide formal convergence guarantees showing that the objective function increases monotonically at each iteration. *Fourth*, we establish rigorous theoretical properties for the DCAF estimator, including estimation consistency, oracle property, and asymptotic normality, under a framework that accommodates both fixed and diverging numbers of subgroups.

The remainder of this paper is organized as follows. In Section 2, we introduce the DCAF model and the associated algorithm for simultaneously identifying two-way subgroups based on clustered survival data. Theoretical results including consistency and oracle property are shown in Section 3. A practical inference procedure is also introduced in this Section. Sections 4 and 5 include numerical simulation studies and analyses of data from PA.3 and CO.26. Section 6 provides concluding remarks. The related theories and additional numerical results are deferred to the Supp.

2. Methods and Algorithm

Assume there are n independent clusters, for example, clinical centres of subjects enrolled in a clinical study with m_i subjects in the i -th cluster ($i \in [n]$, where $[n] := 1, \dots, n$). Let T_{ij} and C_{ij} be the true failure and censoring

time of the subject j in the cluster i ($j \in [m_i]$). Denote the observed failure time and associated censoring indicator as $\tilde{T}_{ij} = \min(T_{ij}, C_{ij})$ and $\delta_{ij} = I(T_{ij} \leq C_{ij})$, respectively. For each subject, we observe biomarkers $\mathbf{x}_{ij} \in R^p$, and treatment assignment w_{ij} . In addition, confounding variables include $\mathbf{z}_{ij} \in R^q$, such as age, gender, and ECOG performance status, and disease stage for patients in PA.3 study.

Define the hazard function of the true failure time T_{ij} for a subject with biomarkers $\mathbf{x}_{ij}$, the treatment w_{ij} , confounding variables $\mathbf{z}_{ij}$ as $\lambda(t|\mathbf{x}_{ij}, w_{ij}, \mathbf{z}_{ij})$. To account for cluster effect and heterogeneous cluster effects of treatments and biomarkers and identify biomarkers, we consider the following heterogeneous Cox model:

$$\lambda(t|\mathbf{x}_{ij}, w_{ij}, \mathbf{z}_{ij}) = \lambda_0(t)u_i \exp(\boldsymbol{\zeta}_i^\top \mathbf{x}_{ij} + v_i w_{ij} + \mathbf{g}^\top \mathbf{x}_{ij} w_{ij} + \boldsymbol{\gamma}^\top \mathbf{z}_{ij}), \quad (2.1)$$

where $\lambda_0(t)$ is an unknown baseline hazard function, u_i , $\boldsymbol{\zeta}_i$ and v_i represent respectively the cluster specific fixed effects, the heterogeneous effects of biomarkers, and the heterogeneous treatment effects, all across clusters, on the hazard of T_{ij} . Here each u_i is treated as a cluster-specific fixed intercept rather than a random variable drawn from a distributional family, and we do not impose any distribution on u_i . Accordingly, the similarity among individuals within the same cluster arises from they share a common fixed intercept, rather than from a stochastic dependence structure induced

by a shared latent random variable. This avoids distributional misspecification while preserves interpretability through direct cluster-specific estimates. $\mathbf{g} = (g_1, \dots, g_p)^\top$ is the interaction effect between the biomarkers $\mathbf{x}_{ij}$ and the treatment assignment w_{ij} , which is assumed to be constant across clusters and have grouping structures across biomarkers. That is, biomarkers in the same subgroup share the same interaction effects on the survival outcome. It is important to note that the interaction effects $\mathbf{g}$ are not assumed to be identical across all biomarkers; instead, they possess their own grouping structure via the predictive subgroups $\mathcal{D}_1, \dots, \mathcal{D}_{K_2}$. The assumption reflect intrinsic pharmacological mechanisms—for instance, whether a biomarker modifies the response to erlotinib depends on the molecular pathway the drug targets, not on the clinical centre where the patient was enrolled. In addition, allowing cluster-specific interactions would introduce $n \times p$ additional parameters, leading to severe identifiability and estimation challenges. $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_q)^\top$ with γ_h as a fixed effect of the h -th confounding variable, which is assumed to be a constant across clusters. The risk of a subject in the i -th cluster increases if $u_i \geq 1$ and decreases otherwise. To avoid imposing a positive restriction on u_i , we write $\eta_i = \ln(u_i)$. Then the model (2.1) can be rewritten as follows: $\lambda(t|\mathbf{x}_{ij}, w_{ij}, \mathbf{z}_{ij}) = \lambda_0(t) \exp(\eta_i + \boldsymbol{\zeta}_i^\top \mathbf{x}_{ij} + v_i w_{ij} + \mathbf{g}^\top \mathbf{x}_{ij} w_{ij} + \boldsymbol{\gamma}^\top \mathbf{z}_{ij})$. Assume that T_{ij} and C_{ij} are conditionally

independent given $\mathbf{x}_{ij}$, w_{ij} and $\mathbf{z}_{ij}$, indicating that the censoring mechanism is noninformative. Denote $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top, \dots, \boldsymbol{\theta}_n^\top)^\top$, $\boldsymbol{\theta}_i = (\boldsymbol{\zeta}_i^\top, \eta_i, v_i)^\top$, and $\mathcal{R}(T_{ij})$ as the risk set of patients at time T_{ij} . Then, the log partial likelihood of the observed data $(\tilde{T}_{ij}, \delta_{ij}, \mathbf{x}_{ij}, w_{ij}, \mathbf{z}_{ij})$ ($i \in [n]$, $j \in [m_i]$) can be written as: $L_n(\boldsymbol{\theta}, \mathbf{g}, \boldsymbol{\gamma}) = \sum_{i=1}^n \sum_{j=1}^{m_i} \delta_{ij} \left[\eta_i + v_i w_{ij} + \boldsymbol{\zeta}_i^\top \mathbf{x}_{ij} + \mathbf{g}^\top \mathbf{x}_{ij} w_{ij} + \boldsymbol{\gamma}^\top \mathbf{z}_{ij} - \log \left\{ \sum_{\tilde{T}_{rs} \in \mathcal{R}(T_{ij})} \exp(\eta_r + v_r w_{rs} + \boldsymbol{\zeta}_r^\top \mathbf{x}_{rs} + \mathbf{g}^\top \mathbf{x}_{rs} w_{rs} + \boldsymbol{\gamma}^\top \mathbf{z}_{rs}) \right\} \right]$. To simultaneously identify two-way subgroup structures of biomarkers, we assume that there are K_1 prognostic subgroups and K_2 predictive subgroups. We identify the former by clustering the parameters $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n$, which yields subgroup centers $\boldsymbol{\alpha}_k^{sa} := (\boldsymbol{\alpha}_{1,k}^\top, \alpha_{2,k}, \alpha_{3,k})^\top$ for $k \in [K_1]$. Similarly, we identify the latter by clustering $g_1, \dots, g_p$, resulting in centers $\alpha_{4,k}$ for $k \in [K_2]$. In other words, let $\mathcal{G} = (\mathcal{G}_1, \dots, \mathcal{G}_{K_1})$ be a partition of $\{1, \dots, n\}$ and $\mathcal{D} = (\mathcal{D}_1, \dots, \mathcal{D}_{K_2})$ be a partition of $\{1, \dots, p\}$, respectively. We have $\boldsymbol{\theta}_i = \boldsymbol{\alpha}_k^{sa}$ for all $i \in \mathcal{G}_k, k \in [K_1]$, and $g_j = \alpha_{4,k}$ for all $j \in \mathcal{D}_k, k \in [K_2]$. We propose to use the following penalized log partial likelihood function for identifying these two-way subgroups and the important confounding variables that are related to the hazards of T_{ij} :

$$S_n(\boldsymbol{\Theta}, \boldsymbol{\alpha}) = L_n(\boldsymbol{\theta}, \mathbf{g}, \boldsymbol{\gamma}) - p_{\lambda_1}(\boldsymbol{\theta}, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\alpha}_3) - p_{\lambda_2}(\mathbf{g}, \boldsymbol{\alpha}_4), \quad (2.2)$$

where $\boldsymbol{\Theta} = (\boldsymbol{\theta}^\top, \mathbf{g}^\top, \boldsymbol{\gamma}^\top)^\top$, $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^\top, \boldsymbol{\alpha}_2^\top, \boldsymbol{\alpha}_3^\top, \boldsymbol{\alpha}_4^\top)^\top$, $\boldsymbol{\alpha}_1 = (\boldsymbol{\alpha}_{1,1}^\top, \dots, \boldsymbol{\alpha}_{1,K_1}^\top)^\top$, $\boldsymbol{\alpha}_2 = (\alpha_{2,1}, \dots, \alpha_{2,K_1})^\top$, $\boldsymbol{\alpha}_3 = (\alpha_{3,1}, \dots, \alpha_{3,K_1})^\top$, and $\boldsymbol{\alpha}_4 = (\alpha_{4,1}, \dots, \alpha_{4,K_2})^\top$,

$p_{\lambda_1}(\boldsymbol{\theta}, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\alpha}_3) = \lambda_1 \sum_{i=1}^n (1/K_1 \sum_{k=1}^{K_1} \|\boldsymbol{\theta}_i - \boldsymbol{\alpha}_k^{sa}\|^{2s})^{1/s}$, and $p_{\lambda_2}(\mathbf{g}, \boldsymbol{\alpha}_4) = \lambda_2 \sum_{j=1}^p (1/K_2 \sum_{k=1}^{K_2} \|g_j - \alpha_{4,k}\|^{2s})^{1/s}$. Since each g_j is a scalar, $\|g_j - \alpha_{4,k}\|^{2s}$ reduces to $|g_j - \alpha_{4,k}|^{2s}$.

The penalty term in (2.2) is centre-based and encourages the subjects to shrink toward their own centres. A similar idea has been employed in the literature, including grouping covariates in high-dimensional regression (Witten et al., 2014), grouping observations in the network community detection problem (Zhang et al., 2021), individualized multidirectional variable selection (Tang et al., 2020), and centre-augmented l_2 -type regularization (He et al., 2023). In all these applications, regularized K -means clustering was considered and, specifically, the following penalty function was used: $p_\lambda(\boldsymbol{\alpha}) = \lambda \sum_{i=1}^n \min \{d(\boldsymbol{\beta}_i, \boldsymbol{\alpha}_1), \dots, d(\boldsymbol{\beta}_i, \boldsymbol{\alpha}_K)\}$, where $\boldsymbol{\beta}_i$ denote individual-specific parameters, $\boldsymbol{\alpha}_k$ the center parameters for subgroup k , and d is the l_2 or l_1 distance. In our specific model, $\boldsymbol{\beta}_i$ corresponds to $\boldsymbol{\theta}_i$ for the prognostic subgroups and to g_j for the predictive subgroups, while $\boldsymbol{\alpha}_k$ corresponds to $\boldsymbol{\alpha}_k^{sa}$ and $\alpha_{4,k}$, respectively. The nonconvexity of the minimum function in this penalty requires the development of special algorithms for the related optimization. The proposed penalty term in (2.2) includes many common means, usually termed as power means. It is known that $s = 1$ and $s = -1$ correspond to the arithmetic and harmonic means,

respectively, and $s \rightarrow -\infty$ and $s \rightarrow \infty$ lead to the minimum and maximum functions, respectively. In practice, we may take $s = -1$, which is related to the K-Harmonic Means approach (KHM), because it has been shown to be less sensitive to the initialization of the centres (Zhang et al., 1999; Xu and Lange, 2019). More importantly, it is a continuous function, and thus, a conventional gradient-based algorithm can be used.

In high-dimensional cluster analysis, the clustering structure often lies in the same/different low dimensional subspace, and the majority of the variables may be redundant in capturing the structure. We therefore propose adding a third penalty term, which encourages the cluster centre $\alpha_{1,k}$ to be sparse for $k = [K_1]$, to $S_n(\Theta, \alpha)$:

$$S_{\text{DCAF}}(\Theta, \alpha) := S_n(\Theta, \alpha) - \sum_{k=1}^{K_1} \sum_{j=1}^p p_{\lambda_3}(\alpha_{1,kj}), \quad (2.3)$$

where $\alpha_{1,kj}$ is the j -th element of $\alpha_{1,k}$. In this article, we choose the adaptive LASSO penalty function (Zou, 2006) for its favorable asymptotic properties and ease of computation. The estimators obtained by maximizing $S_{\text{DCAF}}(\Theta, \alpha)$ in expression (2.3) are denoted as DCAF estimators, and the proposed method is referred to as the DCAF method.

2.1 Algorithm

Given the latent subgroups K_1 , K_2 tuning parameters λ_1 , λ_2 , and λ_3 , the algorithm flow is as follows.

STEP 1. Initiation $\hat{\boldsymbol{\theta}}^{(0)}$, $\hat{\mathbf{g}}^{(0)}$, $\hat{\boldsymbol{\gamma}}^{(0)}$ as the ridge estimators; Initiation $\hat{\boldsymbol{\alpha}}^{(0)}$ can be obtained by K -means method.

STEP 2. Update the $\hat{\boldsymbol{\theta}}^{(m)}$ by minimizing: $-L_n(\boldsymbol{\theta}, \hat{\mathbf{g}}^{(m-1)}, \hat{\boldsymbol{\gamma}}^{(m-1)}) + \lambda_1 \sum_{i=1}^n (1/K_1 \sum_{k=1}^{K_1} \|\boldsymbol{\theta}_i - \hat{\boldsymbol{\alpha}}_k^{sa(m-1)}\|^{2s})^{1/s}$.

STEP 3. Update $\hat{\boldsymbol{\alpha}}_1^{(m)}$, $\hat{\boldsymbol{\alpha}}_2^{(m)}$ and $\hat{\boldsymbol{\alpha}}_3^{(m)}$ by minimizing: $\lambda_1 \sum_{i=1}^n (1/K_1 \sum_{k=1}^{K_1} \|\hat{\boldsymbol{\theta}}_i^{(m)} - \boldsymbol{\alpha}_k^{sa}\|^{2s})^{1/s} + \sum_{k=1}^{K_1} \sum_{j=1}^p p_{\lambda_3}(\alpha_{1,kj})$.

STEP 4. Update $\hat{\mathbf{g}}^{(m)}$ by minimizing $-L_n(\hat{\boldsymbol{\theta}}^{(m)}, \mathbf{g}, \hat{\boldsymbol{\gamma}}^{(m-1)}) + \lambda_2 \sum_{j=1}^p (1/K_2 \sum_{k=1}^{K_2} \|g_j - \hat{\alpha}_{4,k}^{(m-1)}\|^{2s})^{1/s}$.

STEP 5. Update $\hat{\boldsymbol{\alpha}}_4^{(m)}$ by minimizing $\sum_{j=1}^p (1/K_2 \sum_{k=1}^{K_2} \|\hat{g}_j^{(m)} - \alpha_{4,k}\|^{2s})^{1/s}$.

STEP 6. Update $\hat{\boldsymbol{\gamma}}^{(m)}$ by maximizing $L_n(\hat{\boldsymbol{\theta}}^{(m)}, \hat{\mathbf{g}}^{(m)}, \boldsymbol{\gamma})$.

STEP 7. Iterate STEP 2-6 until convergence.

The above algorithm, called the KHM-GA procedure, is used to find the solution that maximizes the objective function (2.3). Particularly, In STEP 3, when $s < 0$, the first term is the concave function of $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$ and $\boldsymbol{\alpha}_3$, and we use an majorize-minimize (MM) algorithm (Lange, 2016) to estimate the subgroup centres $\boldsymbol{\alpha}_{1,k}$, $\alpha_{2,k}$ and $\alpha_{3,k}$ ($k = [K_1]$). Specifically, define $\boldsymbol{\alpha}^{sa} = ((\boldsymbol{\alpha}_1^{sa})^\top, \dots, (\boldsymbol{\alpha}_{K_1}^{sa})^\top)^\top$ and $\text{KHM}(s; \hat{\boldsymbol{\theta}}_i^{(m)}, \boldsymbol{\alpha}^{sa}) =$

2.1 Algorithm

$(1/K_1 \sum_{k=1}^{K_1} \|\hat{\theta}_i^{(m)} - \alpha_k^{sa}\|^{2s})^{1/s}$. We approximate $\text{KHM}(s; \hat{\theta}_i^{(m)}, \alpha^{sa})$ through a first-order Taylor expression at $\alpha^{sa(m-1)}$, namely, $\text{KHM}(s; \hat{\theta}_i^{(m)}, \alpha^{sa(m-1)}) + \{\text{KHM}(s; \hat{\theta}_i^{(m)}, \alpha^{sa(m-1)})\}^{(1/s-1)} \times 1/K_1 \sum_{k=1}^{K_1} \|\hat{\theta}_i^{(m)} - \alpha_k^{sa(m-1)}\|^{2(s-1)} (\|\hat{\theta}_i^{(m)} - \alpha_k^{sa}\|^2 - \|\hat{\theta}_i^{(m)} - \alpha_k^{sa(m-1)}\|^2)$. Then, we estimate α_1 , by minimizing the following objective function:

$$\arg \min_{\alpha_1} \lambda_1/K_1 \sum_{k=1}^{K_1} \sum_{i=1}^n (\zeta_i - \alpha_{1,k})^\top W_{i,k} (\zeta_i - \alpha_{1,k}) + \sum_{k=1}^{K_1} \sum_{j=1}^p p_{\lambda_3}(\alpha_{1,kj}), \quad (2.4)$$

where $W_{i,k} = \text{diag}(w_{i,k}, \dots, w_{i,k})$, and $w_{i,k} = \{\text{KHM}(s; \hat{\theta}_i^{(m)}, \alpha^{sa(m-1)})\}^{(1/s-1)} \|\hat{\theta}_i^{(m)} - \alpha_k^{sa(m-1)}\|^{2(s-1)}$. Note that the objective function (2.4) can be solved by using the function *ncvfit* in the R package **ncvreg** (Breheny and Huang, 2011). We estimate α_2 and α_3 by minimizing the following objective function: $\arg \min_{\alpha_2, \alpha_3} \sum_{k=1}^{K_1} \sum_{i=1}^n w_{i,k} \{(\eta_i - \alpha_{2,k})^2 + (v_i - \alpha_{3,k})^2\}$, leading to closed forms as $\hat{\alpha}_{2,k} = \sum_{i=1}^n w_{i,k} \eta_i / \sum_{i=1}^n w_{i,k}$ and $\hat{\alpha}_{3,k} = \sum_{i=1}^n w_{i,k} v_i / \sum_{i=1}^n w_{i,k}$, $k = [K_1]$. To obtain subgroup membership, we can use the function *cluster* in

the R package **Ckmeans.1d.dp** in STEP 3. Similarity, in STEP 5, we use an majorize-minimize (MM) algorithm (Lange, 2016) to estimate the subgroup centres $\alpha_{4,k}$, $k = [K_2]$, $\text{KHM}(s; \hat{g}_j^{(m)}, \alpha_4) = \text{KHM}(s; \hat{g}_j^{(m)}, \alpha_4^{(m-1)}) + \sum_{k=1}^{K_2} \phi_{j,k} \{(\hat{g}_j^{(m)} - \alpha_{4,k})^2 - (\hat{g}_j^{(m)} - \alpha_{4,k}^{(m-1)})^2\}$, where $\phi_{j,k} = 1/K_2 \{\text{KHM}(s; \hat{g}_j^{(m)}, \alpha_4^{(m-1)})\}^{(1/s-1)} (\hat{g}_j^{(m)} - \alpha_{4,k}^{(m-1)})^{2(s-1)}$. We estimate α_4 by minimizing the following objective function: $\arg \min_{\alpha_4} \sum_{k=1}^{K_2} \sum_{j=1}^p \phi_{j,k} (\hat{g}_j^{(m)} - \alpha_{4,k})^2$, leading to closed forms as $\hat{\alpha}_{4,k} = \sum_{j=1}^p \phi_{j,k} \hat{g}_j^{(m)} / \sum_{j=1}^p \phi_{j,k}$. In STEP 6, γ is obtained

2.2 Convergence of the Algorithm

by using the *coxph* of the R package **survival** with an additional covariate

$\eta_i^{(m)} + v_i^{(m)} w_{ij} + \zeta_i^{(m)\top} \mathbf{x}_{ij} + \mathbf{g}^{(m)\top} \mathbf{x}_{ij} w_{ij}$ with fixed coefficient 1.

2.2 Convergence of the Algorithm

We next consider the convergence properties of the KHM-GA algorithm.

Proposition 1. *In KHM-GA algorithm, the sequence $\{S_{DCAF}(\boldsymbol{\Theta}^{(m)}, \boldsymbol{\alpha}^{(m)})\}_{m \geq 0}$ is monotonically non-decreasing at each iteration; that is, $S_{DCAF}(\boldsymbol{\Theta}^{(m)}, \boldsymbol{\alpha}^{(m)}) \leq S_{DCAF}(\boldsymbol{\Theta}^{(m+1)}, \boldsymbol{\alpha}^{(m+1)})$ for all $m \geq 0$.*

Proposition 1 establishes that the KHM-GA algorithm produces a monotonically non-decreasing sequence of penalized log-likelihood values; since the objective is bounded above, convergence is guaranteed. However, because the objective function is non-convex, convergence to a global optimum cannot be ensured. To address initialization sensitivity, we first estimate cluster-specific parameters via ridge regression, then apply K -means to obtain initial subgroup centres and memberships. Simulations confirm this strategy yields stable and accurate results.

3. Theoretical Properties

For the groupings on the prognostic variables and predictive variables, assume $\mathcal{G} = (\mathcal{G}_{0,1}, \dots, \mathcal{G}_{0,K_1})$ and $\mathcal{D} = \{\mathcal{D}_{0,1}, \dots, \mathcal{D}_{0,K_2}\}$ be the true partitions

of $\{1, \dots, n\}$ and $\{1, \dots, p\}$, respectively. Denote $\mathbf{D} = \{\tilde{\mathbf{D}}_{ij}, i \in [n], j \in [m_i]\}$, $\tilde{\mathbf{D}}_{ij} = (\mathbf{D}_{1,ij}^\top, \mathbf{D}_{2,ij}^\top)^\top$, $\mathbf{D}_{1,ij} = (\mathbf{x}_{ij}^\top, 1, w_{ij})^\top$, $\mathbf{D}_{2,ij} = (\mathbf{z}_{ij}^\top, \sum_{l \in \mathcal{D}_{0,1}} x_{ij,l} w_{ij}, \dots, \sum_{l \in \mathcal{D}_{0,K_2}} x_{ij,l} w_{ij})^\top$. Denote $\tilde{\boldsymbol{\theta}}_i = (\boldsymbol{\theta}_i^\top, \boldsymbol{\gamma}^\top, \mathbf{g}^\top)^\top$, $\tilde{\boldsymbol{\theta}} = (\boldsymbol{\theta}_1^\top, \dots, \boldsymbol{\theta}_n^\top, \boldsymbol{\gamma}^\top, \mathbf{g}^\top)^\top$, $\tilde{\boldsymbol{\alpha}}^{(k)} = ((\boldsymbol{\alpha}_k^{sa})^\top, \boldsymbol{\gamma}^\top, \boldsymbol{\alpha}_4^\top)^\top$, and $\tilde{\boldsymbol{\alpha}} = ((\boldsymbol{\alpha}^{sa})^\top, \boldsymbol{\gamma}^\top, \boldsymbol{\alpha}_4^\top)^\top$. Denote $N_k = \sum_{i \in \mathcal{G}_{0,k}} m_i$, $N = \sum_{k=1}^{K_1} N_k$.

Let $\tilde{\boldsymbol{\alpha}}_0$ is the true value of $\tilde{\boldsymbol{\alpha}}$, $\tilde{\boldsymbol{\alpha}}_{0,1}$ is the true value of $\tilde{\boldsymbol{\alpha}}_1$, where $\tilde{\boldsymbol{\alpha}}_1 := (\boldsymbol{\alpha}_{1,1,1}^\top, \alpha_{2,1}, \alpha_{3,1}, \dots, \boldsymbol{\alpha}_{1,K_1,1}^\top, \alpha_{2,K_1}, \alpha_{3,K_1}, \boldsymbol{\gamma}^\top, \boldsymbol{\alpha}_4^\top)^\top$, $\tilde{\boldsymbol{\alpha}}_2 := (\boldsymbol{\alpha}_{1,1,2}^\top, \dots, \boldsymbol{\alpha}_{1,K_1,2}^\top)^\top$, $\boldsymbol{\alpha}_{1,k,1}$ is the important variable with dimension $p_{s,k}$ for the k -th prognostic group, $\boldsymbol{\alpha}_{1,k,2}$ is the unimportant variable with dimension $p - p_{s,k}$ for the k th prognostic group, $\boldsymbol{\alpha}_{0,1,k} := (\boldsymbol{\alpha}_{0,1,k,1}^\top, \boldsymbol{\alpha}_{0,1,k,2}^\top)^\top \equiv (\boldsymbol{\alpha}_{0,1,k,1}^\top, \mathbf{0}^\top)^\top$, and $\boldsymbol{\alpha}_{0,1,k}$ is the true value of $\boldsymbol{\alpha}_{1,k} := (\boldsymbol{\alpha}_{1,k,1}^\top, \boldsymbol{\alpha}_{1,k,2}^\top)^\top$ for $k = [K_1]$. $(\alpha_{0,2,1}, \alpha_{0,3,1}, \dots, \alpha_{0,2,K_1}, \alpha_{0,3,K_1}, \boldsymbol{\gamma}_0^\top, \boldsymbol{\alpha}_{0,4}^\top)^\top$ is the true value of $(\alpha_{2,1}, \alpha_{3,1}, \dots, \alpha_{2,K_1}, \alpha_{3,K_1}, \boldsymbol{\gamma}^\top, \boldsymbol{\alpha}_4^\top)^\top$. Denote $\tilde{\boldsymbol{\theta}}_0 = \boldsymbol{\psi} \tilde{\boldsymbol{\alpha}}_0$, where $\boldsymbol{\psi} = \{\psi_{i,j}\}$ is a $(pn + 2n + q + p) \times (pK_1 + 2K_1 + q + K_2)$ dimensional 0/1 membership matrix, which is defined in Supp S1.1

To establish the asymptotic properties of the proposed estimators $\hat{\tilde{\boldsymbol{\theta}}}$, we first establish the asymptotic properties of the oracle estimators $\hat{\tilde{\boldsymbol{\theta}}}^{or}$, which is defined as the maximum penalized partial likelihood estimators(MPLE) of $\hat{\tilde{\boldsymbol{\theta}}}$ with true grouping membership known in advance, namely $\hat{\tilde{\boldsymbol{\alpha}}}^{or}$, by maximizing: $\sum_{k=1}^{K_1} \sum_{i \in \mathcal{D}_{0,k}} \sum_{j=1}^{m_i} \delta_{ij} \left[\tilde{\mathbf{D}}_{ij}^\top \tilde{\boldsymbol{\alpha}}^{(k)} - \log \left\{ \sum_{k=1}^{K_1} \sum_{r \in \mathcal{D}_{0,k}} \sum_{\tilde{T}_{rs} \in \mathcal{R}(T_{ij})} \exp(\tilde{\mathbf{D}}_{rs}^\top \tilde{\boldsymbol{\alpha}}^{(k)}) \right\} \right] - \sum_{k=1}^{K_1} \sum_{j=1}^p p_{\lambda_3}(\alpha_{1,kj})$. Then prove that $\hat{\tilde{\boldsymbol{\theta}}}$ converges to

$\hat{\boldsymbol{\theta}}^{or}$ with a probability tending to 1. Supp S1.2 lists a set of mild regularity conditions. In what follows, it is assumed that $m_i = m$ for $i \in [n]$.

Theorem 1. *Assume that $(\mathbf{x}_{ij}, w_{ij}, \mathbf{z}_{ij})$ are independent random samples for $i = [n]$, $j = [m]$, and that T_{ij} and C_{ij} are conditionally independent given $\mathbf{x}_{ij}$, w_{ij} and $\mathbf{z}_{ij}$. Under conditions (C1)–(C5) listed in the Supp S1.2, if $N_{\min} := \min_k N_k \rightarrow \infty$, $\sqrt{N_{\min}}\lambda_3 = O_p(1)$, then there exists a local maximizer $\hat{\boldsymbol{\alpha}}^{or}$ and such that $\|\hat{\boldsymbol{\alpha}}^{or} - \tilde{\boldsymbol{\alpha}}_0\| = O_p(1/\sqrt{N_{\min}})$,*

We next show that $\hat{\boldsymbol{\alpha}}^{or}$ has the oracle property.

Theorem 2. *Under conditions (C1)–(C5) listed in the Supp S1.2, if $N < N_{\min}^2$, $K_1 = O(N^{1/2-\delta})$ for some $0 < \delta \leq 1/2$, $\sqrt{N_{\min}}\lambda_3 \rightarrow 0$ and $N_{\min}\lambda_3 \rightarrow \infty$, then the oracle estimator $\hat{\boldsymbol{\alpha}}^{or}$ must satisfies satisfies $P(\hat{\boldsymbol{\alpha}}_{1,k,2}^{or} = \mathbf{0}) \rightarrow 1$ for all $k \in [K_1]$, and $\sqrt{N_k}(\hat{\boldsymbol{\alpha}}_{1,k,1}^{or} - \boldsymbol{\alpha}_{0,1,k,1}, \hat{\boldsymbol{\alpha}}_{2,k}^{or} - \alpha_{0,2,k}, \hat{\boldsymbol{\alpha}}_{3,k}^{or} - \alpha_{0,3,k}) \rightarrow \mathcal{N}(\mathbf{0}, \tilde{\Sigma}_{11,k}(\tilde{\boldsymbol{\alpha}}_{0,1}, \tau))$, and $\sqrt{N}(\hat{\boldsymbol{\gamma}}^{or} - \boldsymbol{\gamma}_0, \hat{\boldsymbol{\alpha}}_4^{or} - \boldsymbol{\alpha}_{0,4}) \rightarrow \mathcal{N}(\mathbf{0}, C^{-1}(\tilde{\boldsymbol{\alpha}}_{0,1}, \tau))$. where $\tilde{\Sigma}_{11,k}(\tilde{\boldsymbol{\alpha}}_{0,1}, \tau)$ and $C^{-1}(\tilde{\boldsymbol{\alpha}}_{0,1}, \tau)$ are defined in Supp.*

Based on Theorem 2, the convergence rate of $\hat{\boldsymbol{\alpha}}_{1,k,1}^{or}$, $\hat{\boldsymbol{\alpha}}_{2,k}^{or}$ and $\hat{\boldsymbol{\alpha}}_{3,k}^{or}$ is $O_p(N_k^{-1/2})$ and $(\hat{\boldsymbol{\gamma}}^{or}, \hat{\boldsymbol{\alpha}}_4^{or})$ converge to their true values at the rate $O_p(N^{-1/2})$. The condition $N < N_{\min}^2$ and $K_1 = O(N^{1/2-\delta})$ for some $0 < \delta \leq 1/2$ are to ensure parameters $(\hat{\boldsymbol{\gamma}}^{or}, \hat{\boldsymbol{\alpha}}_4^{or})$ converge to their corresponding true values at the rate $O_p(N^{-1/2})$. When K_1 is a constant and $N_{\min} \rightarrow \infty$, if $N_k \equiv M$,

the condition $N < N_{\min}^2$ holds automatically, and $(\hat{\alpha}_{1,k,1}^{or}, \hat{\alpha}_{2,k}^{or}, \hat{\alpha}_{3,k}^{or})$ and $(\hat{\gamma}^{or}, \hat{\alpha}_4^{or})$ have the same convergence rate. When $K_1 \rightarrow \infty$ and $N_k \equiv M$, the condition $N < N_{\min}^2$ implies $K_1 < M$, and $M = O(N^{1/2})$.

According to the definition of ψ , it is clear to see that $\hat{\theta}^{or} = \psi \hat{\alpha}^{or}$. Define the separation level between different groups as $b_1 = \min_{k \neq k'} \|(\alpha_{0,1,k,1}^\top, \alpha_{0,2,k}, \alpha_{0,3,k}) - (\alpha_{0,1,k',1}^\top, \alpha_{0,2,k'}, \alpha_{0,3,k'})\|_2$, $b_2 = \min_{k \neq k'} |\alpha_{0,4,k} - \alpha_{0,4,k'}|$. Consider the neighborhood of $(\theta_0, \gamma_0, g_0)$, that is, $\Theta_{\phi_n} = \{\theta \in R^{n(p+2)}, \gamma \in R^q, g \in R^p : \|\gamma - \gamma_0\|_2 \leq \phi_n, \max_i \|\theta_i - \theta_{0,i}\|_2 \leq \phi_n, \max_j |g_j - g_{0,j}| \leq \phi_n\}$. The following Theorem 3 indicates that the performance of the proposed estimator is asymptotically equivalent to the oracle estimator.

Theorem 3. Assume that $\phi_n \ll \min(b_1, b_2)$ and $\lambda_1 > c_1 K_1^{1/s} \phi_n^{-1}$ and $\lambda_2 > c_1 K_2^{1/s} \phi_n^{-1}$ with c_1 defined in condition (C4), then for $s < 0$, under the conditions of Theorem 2, there exists a local minimizer $\hat{\theta} \in \Theta_{\phi_n}$ satisfying $P(\hat{\theta} = \hat{\theta}^{or}) \rightarrow 1$.

Theorem 3 shows that when the separation level is small, a better initial value is needed to ensure the oracle property, ridge estimator as initial for our proposed DCAF method, which is confirmed to work well by the simulation studies in Section 4. In practice, it is preferable to conduct inference for the parameters, and we use the data splitting strategy to recalculate the standard errors and the p -values (Fei et al., 2021), with

further details provided in Supp S.2.

We propose to simultaneously select K_1 , K_2 , λ_1 , λ_2 , and λ_3 by minimizing the BIC. This part is provided in the Supp S.3.

4. Simulation Studies

In the numerical simulation studies, we evaluated the performance of the proposed DCAF from the following five aspects: (i) A comparison of the proposed method with two commonly employed methods: mixture model-based methods(FMCox, You et al. (2018)) and regularization-based methods(Pairwise, Hu et al. (2021)); (ii) The performance of the proposed method with continuous and discrete biomarkers under balanced and unbalanced subgroup sizes, respectively; (iii) The robustness of the proposed method when there are no clusters; (iv) A confirmatory semi-simulation using the covariates from the CO.26 study; (v) The performance of the proposed data splitting strategy to obtain standard errors and p -values with detailed procedures listed in the Supp S2. In addition, the proposed DCAF method is compared with two other commonly used methods. The first method, Rig-ref, employs a subgroup Cox PH model to identify subgroups through ridge estimation followed by K-means clustering. The second method, Oracle, is the estimator for which the true nonzero covariates

4.1 Simulation Study 1: Comparison with existing methods

and the true subgroup membership are known. The performance of these estimators was assessed using the bias (Bias), standard deviation (SD), root mean square error (RMSE), and Rand index (RI)(Rand, 1971), which is between 0 and 1 and the definition of RI can be found in Supp S4.8. Higher values of the RI indicate better clustering accuracy of the algorithm. To measure the performance of variable selection, we reported both Sensitivity and Specificity, representing the proportion of correctly identified zero and nonzero coefficients, respectively.

4.1 Simulation Study 1: Comparison with existing methods

FM Cox method addresses heterogeneity in scenarios with high-dimensional predictors, while Pairwise method focuses solely on treatment heterogeneity. We considered the following two different settings which align with the scopes of these two methods, respectively. In particular, we assumed individuals share the same coefficients within each cluster.

Setting 1 (Fixed dimensional biomarkers) The survival time for the j th subject in the i th cluster is $T_{ij} = \varepsilon_{ij} / \exp(\beta_i^\top \mathbf{x}_{ij})$, $i \in [30]$, $j \in [10]$, where $\varepsilon_{ij} \sim \text{Exponential}(1)$, $\mathbf{x}_{ij}$ was generated from the multivariate normal distribution with mean $\mathbf{0}_p$ and covariance matrix $(s_{ij})_{p \times p}$, where $s_{ij} = 0.5^{|i-j|}$. We considered $K = 3$, $\beta = (\beta_1, \dots, \beta_n)^\top = (\underbrace{\alpha_{1,1}, \dots, \alpha_{1,1}}_{n_1}, \underbrace{\alpha_{1,2}, \dots, \alpha_{1,2}}_{n_2}, \dots)$

4.1 Simulation Study 1: Comparison with existing methods

$\underbrace{\boldsymbol{\alpha}_{1,3}, \dots, \boldsymbol{\alpha}_{1,3}}_{n_3}^\top$. $\boldsymbol{\alpha}_{1,1} = (-2, -3, 0, \mathbf{0}_{p-3}^\top)^\top$, $\boldsymbol{\alpha}_{1,2} = (0, -1, -1.5, \mathbf{0}_{p-3}^\top)^\top$, and $\boldsymbol{\alpha}_{1,3} = (2, 1, 1.5, \mathbf{0}_{p-3}^\top)^\top$. We also assumed $p = 3$ and explored two situations: (S1) three subgroups with balanced sizes $(n_1, n_2, n_3) = (10, 10, 10)$, and (S2) three subgroups with unbalanced sizes $(n_1, n_2, n_3) = (15, 7, 8)$. The censoring time was generated from $C_{ij} \sim \text{Uniform}(0, c)$, where c was selected to obtain the desired censoring rate of 15% and 30%.

Setting 2 (High dimensional Biomarkers) The survival time for the j th subject in the i th cluster is $T_{ij} = \varepsilon_{ij} / \exp(\eta_i + \boldsymbol{\beta}_i^\top \mathbf{x}_{ij})$, $i \in [30]$, $j \in [15]$, $\boldsymbol{\eta} = (\eta_1, \dots, \eta_n)^\top$ with centers being $\alpha_{2,1} = -1$, $\alpha_{2,2} = 0$, and $\alpha_{2,3} = 1$. The other assumptions are similar with that in Setting 1 except $p = 50$.

Under both settings, the ratio of the true number of the subgroups selected (Accuracy) is presented. Figures 1 and 2 summarize the key results for Setting 1. Complete numerical results for Settings 1 and 2 are provided in Tables S1–S2 of Supp S4.1. Three main findings emerge from the graphical comparisons. (i) *Near-oracle performance of DCAF*. The proposed DCAF estimator achieves RMSE and RI values that are comparable to those of the Oracle method(which assumes the true subgroup structure and active variables are known), and consistently outperforms all competing methods across both balanced and unbalanced subgroup configurations and both censoring rates (Figures 1 and 2). Under Setting 2 with

4.1 Simulation Study 1: Comparison with existing methods

high-dimensional biomarkers (Table S2), the DCAF estimator similarly exhibits the oracle property: in comparison with Rig-ref, the proposed DCAF method showed a significant enhancement in RI reaching up to 31% improvement, while the RMSE decreased by an average of 76% in the setting with 30% censoring rate and balanced group sizes. *(ii) Robustness to initialization.* Regardless of the censoring rate, the KHM-GA algorithm consistently identified the correct number of subgroups and recovered the true memberships with high probability (Accuracy = 1 across all configurations in Figure 2). In contrast, the Pairwise method exhibited notable sensitivity to the choice of initial values, resulting in substantially lower accuracy in selecting the correct number of groups (Accuracy ranging from 0.71 to 0.81). *(iii) Advantage over FMCox.* FMCox exhibited systematically higher RMSE than DCAF across all settings (Figure 1), likely because it must estimate K_1 separate nonparametric baseline hazard functions, introducing additional variability into the parameter estimates.

In addition to estimation accuracy, we compared the computational time among DCAF, Pairwise, and FMCox across varying sample sizes, as summarized in Figure 2. In the fixed-dimensional setting (Setting 1, $p = 3$, Figure 2(c)), the Pairwise method is consistently the slowest, with its running time increasing sharply as n grows—a direct consequence of its $O(n^2)$

4.1 Simulation Study 1: Comparison with existing methods

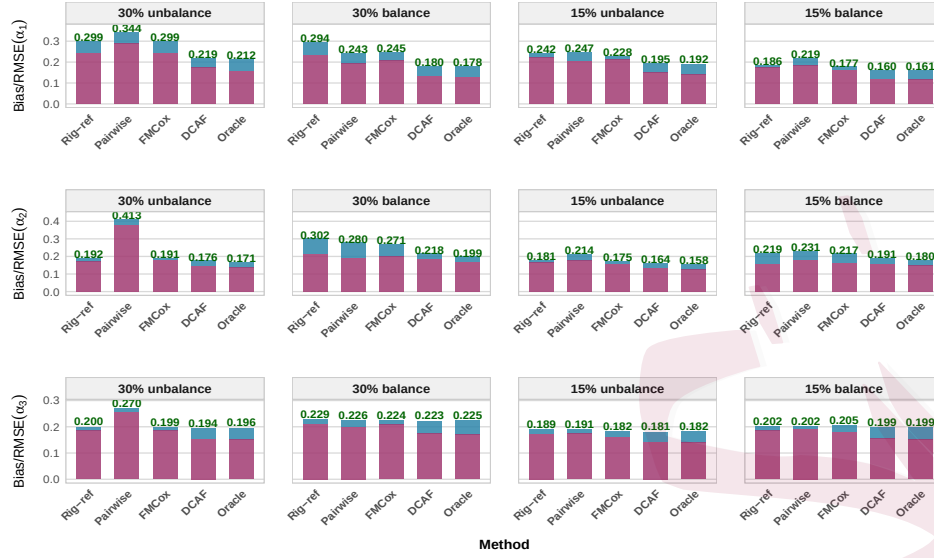


Figure 1: Bias(blue part), and RMSE(total height) of the subgroup centre estimates α under Setting 1 of Simulation Study 1, for balanced $(n_1, n_2, n_3) = (10, 10, 10)$ and unbalanced $(n_1, n_2, n_3) = (15, 7, 8)$ subgroup configurations with 15% and 30% censoring rates.

pairwise penalty structure. By contrast, DCAF scales more favorably due to its $O(nK)$ centre-augmented penalty. Although DCAF is moderately slower than FMCox in this low-dimensional setting, it achieves substantially better estimation accuracy(lower RMSE), clustering quality(higher RI), and subgroup selection accuracy. In the high-dimensional setting (Setting 2, $p = 50, 100$, Figure 2(d)), the computational cost of both DCAF

4.1 Simulation Study 1: Comparison with existing methods

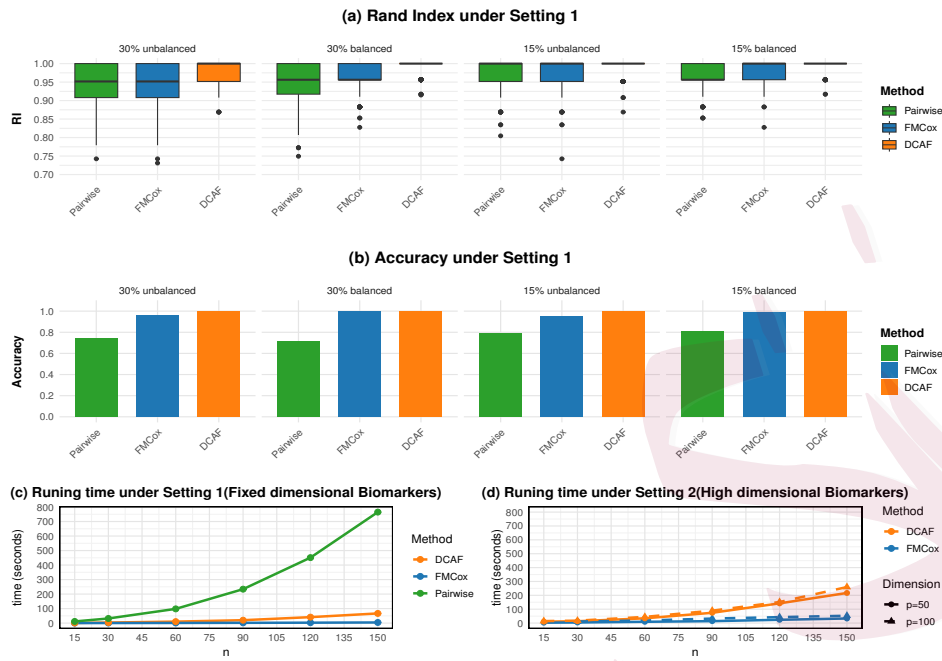


Figure 2: (a) Rand Index and (b) Accuracy for selecting the correct number of subgroups under Setting 1 of Simulation Study 1. Computational time (in seconds) of three methods under Simulation Study 1, with balanced cluster sizes and a 15% censoring rate, for $n \in \{15, 30, 60, 90, 120, 150\}$ for (c) Setting 1 ($p = 3$) and (d) Setting 2 ($p = 50, 100$), respectively.

and FMCox increases with n and p , as expected. While DCAF incurs a higher cost than FMCox, the additional computation yields consistently superior statistical performance-lower RMSE and higher RI values across all configuration (Table S2 in Supp S4.1). Moreover, in the most challenging

4.2 Simulation Study 2: Performance of the proposed method

scenarios (e.g., 15 centres with 100 biomarkers), FMCox fails to produce valid results, whereas DCAF remains robust and delivers reliable estimates. When the number of candidate biomarkers is very large (e.g., thousands of genes), we recommend a two-stage strategy: first apply a sure independence screening (SIS) method (Fan and Lv, 2008) to a moderate scale, then apply DCAF on the screened variables. Additional simulation studies evaluating the performance of this strategy are included in Supp S4.6. This approach keeps the computational cost of DCAF manageable while preserving its statistical advantages.

4.2 Simulation Study 2: Performance of the proposed method

We considered the following two substudies with continuous biomarkers as observed in PA.3 study(Substudy 2.1) and binary biomarkers as observed in CO.26 study(Substudy 2.2): the setting of these two substudies are seen in Supp S4.2. In Substudy 2.1, we considered two cases: (Case 1) m_i were randomly sampled from 5 to 50; and (Case 2) m_i were randomly sampled from 5 to 30. In each case, we considered two scenarios: (S1) three subgroups with balanced size $(n_1, n_2, n_3) = (10, 10, 10)$, and (S2) three subgroups with unbalanced size $(n_1, n_2, n_3) = (15, 7, 8)$.

The estimation results of Substudy 2.1 for Cases 1–2 are summarized in

4.2 Simulation Study 2: Performance of the proposed method

Table S3 of Supp S4.2. We can see that the RMSE of the proposed DCAF is comparable to that of Oracle estimators, which assumes the number, and membership of subgroups and significant variables are known in advance, indicating that the proposed DCAF estimators exhibit the oracle property. In comparison with Rig-ref estimators, the improvement in RI by the proposed DCAF method was as high as 34.6% and the RMSE was down by an average of 80% under Case 2 with unbalanced subgroup size and small cluster size. The performance of all of the estimators gets better as the cluster size m_i increases. The results for Substudy 2.2 are summarized in Table S4 of the Supp S4.2, from which we can draw the same conclusion as Substudy 2.1 with continuous biomarkers. Specifically, compared to Rig-ref estimators with small clinical centres $n = 30$, the improvement in DCAF was as high as 16% and was down by an average of 94% in terms of the RI and RMSE of the predictive subgroups, respectively. The performance of all of the estimators gets better as n increases.

In the Supp S4.3–S4.5, we also displayed the performance of the BIC criterion proposed in Supp S3 and further verified that the robustness of the proposed DCAF method when there are no clusters(Simulation Study 3); In addition, a confirmatory semi-simulation using the covariates from the CO.26 study was studied(Simulation Study 4), and we may conclude that

newly calculated standard errors are able to provide meaningful inference results(Simulation Study 5).

5. Case Studies

In this section, we applied the proposed DCAF method to two clinical trials, PA.3 and CO.26. To save space, CO.26 has been moved to the Supp S6.

Between October 2001 and January 2003, 569 patients with locally advanced or metastatic pancreatic cancer from 17 countries were enrolled in a randomized trial which assigned patients randomly to receive Treatment\Trt or Control(Moore et al., 2007) as mentioned in the introduction. In a companion study, the expression levels of 50 biomarkers, which are relative concentrations of key antibody proteins, were assessed from the pretreatment plasma samples of the patients enrolled in the trial using the PLA, with the aim of identifying the biomarkers that were prognostic of survival as well as predictive for a response to erlotinib with respect to survival. Since patients in this trial were enrolled from a large number of countries in four continents, there was a potential heterogeneity in the association between the biomarkers and the outcome among the countries (Cheng et al., 2015; Gasperoni et al., 2020; Liu et al., 2024). In addition, we did an analysis of variance (ANOVA) to test the differences in biomarker means among

centres, i.e. comparing the variance (or variation) between centres with the variation within each center. Approximately 40% of the biomarkers exhibit p-values below 0.05, indicating that these biomarkers have a high between-centre variance and a low within-centre variance. Here, the prognostic and predictive subgroups correspond to the subgroups across centers and biomarkers, respectively. We applied the models and inference procedures proposed in this paper to identify prognostic and predictive subgroups simultaneously based on the countries and multiple biomarkers after taking into account the following potential confounding factors: age, sex(1 for male vs. 0 for female), Eastern Cooperative Oncology Group(ECOG), performance status (PF)(0 or 1 vs 2), extent of disease(DIS)(1 for distant metastatic vs. 0 for locally advanced). After removing patients with missing covariates, 478 patients (233 in the group treated with erlotinib and gemcitabine) from 14 countries were included in the analysis. From the descriptive analysis of the confounding variables shown in Table S12 of the Supp S7, it is known that they were balanced between two treatment groups. Minimization of BIC defined in Supp S3 led to $K_1 = 2$, $K_2 = 7$ and $\lambda_1 = 0.1$, $\lambda_2 = 0.5$, $\lambda_3 = 0.00042$, implying two major prognostic subgroups and seven predictive subgroups.

Table 1 presents partial results for estimating parameters related to

the prognostic groups, including the corresponding hazard ratios, standard errors, and p -values from the proposed model and procedures, as well as three other models: the heterogeneous FMCox (You et al., 2018), which accounts for heterogeneity in the association between treatment, biomarkers, and outcomes across countries solely; a homogeneous Cox PH model with the same predictive effects (Homo-Lasso); and a homogeneous Cox PH model with different predictive effects (Cox-Lasso). The complete results can be found in Table S9 of the Supp S5, and the prognostic subgroup($\mathcal{G}_1$ and $\mathcal{G}_2$) identified by FMCox and DCAF are presented in Supp S5.

From Table 1 and Table S9 in Supp S5, we can see that: (a) All three methods (DCAF, Homo-Lasso, Cox-Lasso) found that patients with locally advanced disease and lower performance status had longer survival; (b) The treatment effect is not significant from either the Homo-Lasso (hr= 0.828, p -value=0.252) or Cox-Lasso method (hr=0.824, p -value=0.264). Additionally, FMcox method did not find any meaningful subgroup, even after considering the heterogeneity of the centres($\mathcal{G}_1$, hr=0.821, p -value=0.118; $\mathcal{G}_2$, hr=0.821, p -value=0.174). In contrast, our DCAF method successfully identified two prognostic subgroups, with the treatment being effective for the first subgroup (hr=0.717, p -value=0.019). The C-index for FMcox method is 0.651, which is lower than that (0.695) for DCAF method, sug-

gesting that FMCox may have misclassified some cases; (c) All four models indicated that elevated levels of HIF1A and IL6.3B are associated with lower survival rates. Our DCAF method identified five additional prognostic biomarkers: CEA, IGF2.6B, PDK1, EGFR, and PF4, which were not detected by the Homo-Lasso and Cox-Lasso methods. Some of these biomarkers have also been identified as prognostic by previous studies. For instance, Shultz et al. (2016) reported that higher levels of CEA and HIF1A are associated with reduced survival; (d) For the two identified subgroups of countries $\mathcal{G}_1$ and $\mathcal{G}_2$, the fixed effect in $\mathcal{G}_2$ was set to 1 for identification purposes, while the estimated fixed effect in $\mathcal{G}_1$ is 1.317 (p -value=0.199), indicating no significant difference between the two subgroups in terms of heterogeneity of the clinical outcomes.

A comparison of the baseline characteristics of patients from these two subgroups reveals that a lower percentage of patients from $\mathcal{G}_1$ are white compared to those from $\mathcal{G}_2$ (86% vs. 97%). Additionally, no Oriental patients were enrolled from $\mathcal{G}_2$, while 8% of patients in $\mathcal{G}_1$ were Oriental. As shown in the Kaplan-Meier curves in Figure 3, there was a marginally significant treatment effect among all patients included in the analyses (log-rank test, p -value=0.073). However, patients in $\mathcal{G}_1$ treated with erlotinib and gemcitabine had significantly longer survival than those treated with placebo

plus gemcitabine (log-rank test, p -value=0.0028). In contrast, there was no significant difference in survival between the two treatment groups for patients in $\mathcal{G}_2$ (log-rank test, p -value=0.100). This discrepancy may be attributed to the different racial distributions in the two subgroups. This racial distribution difference is consistent with known ethnic variations in pancreatic cancer biology and drug metabolism (Ballman, 2015), providing a plausible biological explanation for the differential treatment effects observed across the two subgroups.

The results for the estimates of parameters related to the predictive subgroups with p -values are summarized in Table 2. A few weak predictive biomarkers were identified by the proposed method, leading to seven predictive subgroups. Significant interaction with treatment was found in six of these subgroups, which indicates they are meaningful predictive subgroups. In contrast, no subgroup with a significant interaction with the treatment was found by the Homo-Lasso approach, and two biomarkers IGF2.6B and TGF.BETA.10B (TG.10B) were found to have a weak interaction with the treatment by the Cox-Lasso method. These two biomarkers were clustered into the same predictive subgroup in our proposed DCAF model with significant negative effects. Furthermore, among other seven biomarkers that were included in the Cox-Lasso model but did not have significant interac-

tions with the treatment because of their small effects, our proposed DCAF procedure clustered CXCL6 and VEGFD.7B(VE.7B) into the same predictive subgroup with significant negative interaction effect. CA.19.9 was clustered into one and IL8.4B with MMP7.10.B(MMP7) into another predictive subgroups with significant positive interaction effects. The remaining two biomarkers REG.IV.4B(RE.4B) and HER2 were clustered into the same predictive group with non-significant effects. These result showed that our proposed DCAF outperformed conventional variable selection methods in terms of identifying weak signals. This advantage arises from the centre-augmented penalty, which enables the detection of weak but biologically relevant predictive signals by grouping biomarkers with similar interaction effects—for example, clustering CXCL6 and VE.7B into the same predictive subgroup with significant negative interaction effects.

6. Discussion

In this paper, we proposed a novel DCAF method to simultaneously identify two-way subgroup structures in prognostic and predictive for clustered survival data. A primary assumption of DCAF is that observations within the same center belong to the same cluster, and that biomarkers within the same subgroup exhibit consistent interaction effects with treatment across

Table 1: The estimated coefficients(est), hazard ratio(hr), standard errors(se) and p -values for prognostic biomarkers identified by the proposed DCAF, the FMCox, the Homo-Lasso, and the Cox-Lasso method based on PA.3.

Method		Prognostic Variables										
(C-index)		$\mathcal{G}_1$	Intercept	Trt	CEA	HIF1A	IL6.3B	IGF2.6B	PDK1	EGFR	AXL	PF4
DCAF	est	0.275	-0.332	0.215	0.162	0.176	-0.284	-0.238	-0.209	-0.177	-0.311	
	hr	1.317	0.717	1.239	1.176	1.192	0.753	0.788	0.811	0.838	0.733	
	se	0.216	0.142	0.082	0.077	0.077	0.105	0.091	0.084	0.103	0.122	
	p-value	0.199	0.019	0.008	0.034	0.022	0.007	0.009	0.013	0.084	0.011	
	(0.695)	$\mathcal{G}_2$	-	Trt	CEA	HIF1A	IL6.3B	IGF2.6B	PDK1	EGFR	AXL	PF4
FMCox	est	-	0.391	-0.172	-0.007	0.295	-0.216	-0.249	-0.235	-0.152	-0.146	
	hr	-	1.479	0.842	0.993	1.343	0.805	0.780	0.791	0.859	0.864	
	se	-	0.285	0.176	0.193	0.165	0.189	0.239	0.210	0.157	0.218	
	p-value	-	0.167	0.321	0.950	0.073	0.248	0.292	0.259	0.329	0.493	
	Homo-Lasso	$\mathcal{G}_1$	-	Trt	CEA	HIF1A	IL6.3B	IGF2.6B	IL8.4B	EGFR	AXL	PF4
est		-	-0.198	0.151	0.186	0.230	-0.198	0.143	-0.218	-0.231	-0.079	
hr		-	0.821	1.163	1.204	1.259	0.820	1.155	0.804	0.794	0.924	
se		-	0.127	0.067	0.080	0.081	0.110	0.084	0.094	0.086	0.070	
(0.651)		p-value	-	0.118	0.023	0.020	0.005	0.070	0.086	0.020	0.007	0.256
Cox-Lasso	$\mathcal{G}_2$	-	Trt	CEA	HIF1A	IL6.3B	IGF2.6B	IL8.4B	EGFR	AXL	PF4	
	est	-	-0.197	0.449	0.610	0.265	-0.192	0.484	-0.066	-0.224	0.003	
	hr	-	0.821	1.567	1.841	1.303	0.826	1.622	0.936	0.799	1.003	
	se	-	0.147	0.120	0.118	0.093	0.122	0.131	0.109	0.097	0.189	
	(0.694)	p-value	-	0.174	< 0.001	< 0.001	0.004	0.114	< 0.001	0.533	0.020	0.969
Cox-Lasso	CXCL6	Trt	CEA	HIF1A	IL6.3B	IGF2.6B	PDK1	EGFR	AXL	IL8.4B		
	est	-0.100	-0.189	0.167	0.269	0.261	-0.243	-0.180	-0.129	-0.123	0.165	
	hr	0.905	0.828	1.181	1.308	1.298	0.785	0.835	0.879	0.885	1.179	
	se	0.124	0.167	0.091	0.088	0.088	0.126	0.140	0.141	0.107	0.148	
	(0.702)	p-value	0.413	0.252	0.066	0.002	0.003	0.053	0.196	0.354	0.248	0.261
Cox-Lasso	CXCL6	Trt	CEA	HIF1A	IL6.3B	IGF2.6B	PDK1	EGFR	AXL	IL8.4B		
	est	-0.116	-0.194	0.165	0.261	0.257	-0.198	-0.187	-0.158	-0.164	0.171	
	hr	0.891	0.824	1.179	1.298	1.293	0.820	0.830	0.854	0.849	1.187	
	se	0.138	0.175	0.092	0.088	0.088	0.148	0.141	0.128	0.110	0.161	
	(0.702)	p-value	0.395	0.264	0.072	0.003	0.003	0.178	0.180	0.214	0.134	0.283

all patients. The method performs well when each center has a moderate number of patients, but may be less effective when centers contain very few patients or when covariates are high-dimensional.

Several directions remain for future work. First, additional steps would be needed to generate classification rules for new patients from centers not included in the clustering dataset; cluster-then-predict approaches (Peikari

Table 2: The estimated coefficients(est), hazard ratios(hr), standard errors(se) and p -values for predictive biomarkers for the response to erlotinib plus gemcitabine (Treatment) identified by the proposed DCAF, the FM-Cox, the Homo-Lasso, and the Cox-Lasso method.

DCAF		Predictive Biomarkers			
	variable	est	hr	se	p -value
$\mathcal{D}_1$	PF4	-0.640	0.527	0.190	0.001
$\mathcal{D}_2$	IGF2.10B, IL7.9B,	-0.282	0.754	0.100	0.005
$\mathcal{D}_3$	IGF2.6B, EGFR, TR.2, CPA1.6B, CXCL6, VE.7B, MMP2, CXCL1, GAS6, OPN.10B, TG.10B,	-0.180	0.835	0.037	< 0.001
$\mathcal{D}_4$	BMP.2.4, MMMP3, MMP1.3, SPARC.3B, IL7.4B, RE.4B, SPARC.6B, TG.9B, MMP1.6, CXCL6.6B, IL12.7B, RE.5B, CXCL9.3B, HER2, VE.8B, IL8,IP10, MMP7, AXL,CEA	0.009	1.009	0.019	0.618
$\mathcal{D}_5$	IL6.3B, PDK1, IL8.4B, MESO.8B, IL12.8B, MMP7	0.126	1.135	0.045	0.005
$\mathcal{D}_6$	CA.19.9, OPN.4B, HIF1A, PDGFA, IGFB.2	0.208	1.232	0.074	0.005
$\mathcal{D}_7$	IGFBP3, MUC1,	0.412	1.509	0.123	0.001
Homo-Lasso		est	hr	se	p -value
$\mathcal{D}$		-0.008	0.992	0.009	0.345
Cox-Lasso		est	hr	se	p -value
	CXCL6	-0.125	0.882	0.179	0.474
	CA19.9	0.161	1.174	0.151	0.281
	IL8.4B	0.204	1.227	0.236	0.378
	RE.4B	0.146	1.157	0.154	0.337
	IGF2.6B	-0.314	0.731	0.172	0.066
	VE.7B	-0.153	0.858	0.186	0.403
	HER2	-0.161	0.851	0.136	0.232
	MMP7	0.158	1.171	0.151	0.288
	TG.10B	-0.210	0.811	0.119	0.077

et al., 2018; Werner et al., 2023) offer a promising avenue. Second, like most penalized survival methods, DCAF assumes non-informative censoring; extensions via inverse probability of censoring weighting (Robins and Rotnitzky, 1992) or joint modeling of event and censoring processes are important directions. Third, the assumption of constant treatment-by-

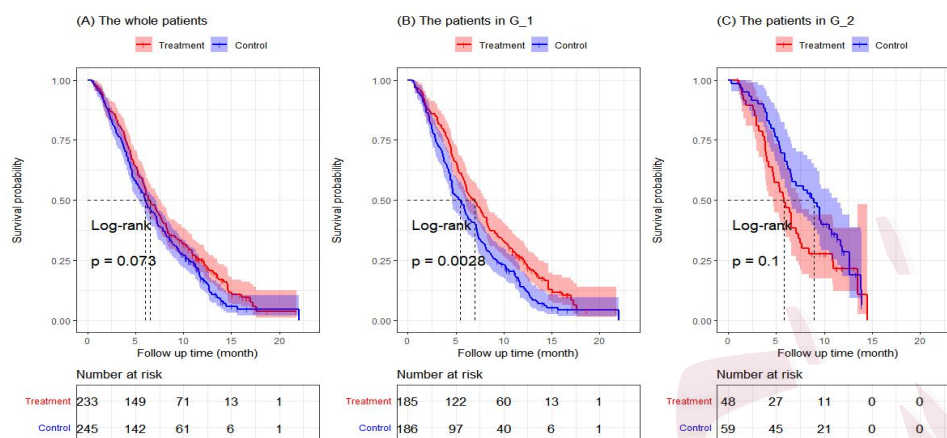


Figure 3: Kaplan-Meier curves for (A) The whole patients (assume the 17 centres are homogeneous); (B) The patients in G_1 (the first prognostic subgroup determined by proposed DCAF model); and (C) The patients in G_2 (the second prognostic subgroup determined by proposed DCAF model) received Treatment or Control based on PA.3.

biomarker interactions across clusters may be restrictive in some settings, and extending DCAF to accommodate cluster-heterogeneous interactions is a worthwhile area for future research.

Acknowledgements

The authors gratefully acknowledge National Key R&D Program of China (No. 2022YFA1003702), National Natural Science Foundation of China (No. 12426309) and Guanghai Talent Project of SWUFE, Natural Sciences

REFERENCES

and Engineering Research Council of Canada, and Sichuan Sichuan Science and Technology Program (Grant No. 2024NSFSC0040, 2024ZYD0115).

Supplementary Materials

The proofs and more simulation results are presented in the Supp.

References

- Ballman, K. (2015). Biomarker: Predictive or prognostic? J Clin Oncol, 33(33):3968–3971.
- Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2010). Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends in Machine Learning, 3(1):1–122.
- Breheny, P. and Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. Ann Appl Stat, 5(1):232–253.
- Caroni, C., Crowder, M., and Kimber, A. (2010). Proportional hazards models with discrete frailty. Lifetime Data Analysis, 16:374–384.
- Chen, E. X., Jonker, D. J., Loree, J. M., Kennecke, H. F., Berry, S. R., Couture, F., et al. (2020). Effect of Combined Immune Checkpoint Inhibition vs Best Supportive Care Alone in Patients With Advanced Colorectal Cancer: The Canadian Cancer Trials Group CO.26 Study. Jama Oncol, 6(6):831–838.
- Cheng, X., Lu, W., and Liu, M. (2015). Identification of homogeneous and heterogeneous

REFERENCES

- variables in pooled cohort studies. Biometrics, 71(2):397–403.
- Cooper, W. A. and Tan, P. H. (2024). Predictive and prognostic biomarkers in solid tumours. Pathology, 56(2):145–146.
- Eng, K. H. and Hanlon, B. M. (2014). Discrete mixture modeling to address genetic heterogeneity in time-to-event regression. Bioinformatics, 30(12):1690–1697.
- Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion). J R Stat Soc Ser B (Stat Methodol), 70:849–911.
- Fei, Z., Zheng, Q., Hong, H. G., and Li, Y. (2021). Inference for high-dimensional censored quantile regression. J Am Stat Assoc, 118(542):898–912.
- Gasperoni, F., Ieva, F., Paganoni, A. M., Jackson, C., and Sharples, L. (2020). Non-parametric frailty cox models for hierarchical time-to-event data. Biostatistics, 21(3):531–544.
- Guo, G. and Rodriguez, G. (1992). Estimating a multivariate proportional hazards model for clustered data using the em algorithm, with an application to child survival in guatemala. J Am Stat Assoc, 87(420):969–976.
- Han, M., Lin, Y., Liu, W., and Wang, Z. (2024). Robust inference for subgroup analysis with general transformation models. Journal of Statistical Planning and Inference, 229.
- He, Y., Zhou, L., Xia, Y., and Lin, H. (2023). Center-augmented ℓ_2 -type regularization for subgroup learning. Biometrics, 79(3):2157–2170.
- Heckman, J. and Singer, B. (1984). A method for minimizing the impact of distributional assumptions in econometric models for duration data. Econometrica, 52(2):271–320.

REFERENCES

- Hougaard, P. (1986). A class of multivariate failure time distributions. Biometrika, 73(3):671–678.
- Hu, X., Huang, J., Liu, L., Sun, D., and Zhao, X. (2021). Subgroup analysis in the heterogeneous Cox model. Stat Med, 40(3):739–757.
- Ke, Z. T., Fan, J., and Wu, Y. (2015). Homogeneity pursuit. J Am Stat Assoc, 110(509):175–194.
- Lange, K. (2016). MM optimization algorithm. SIAM.
- Legrand, C., Duchateau, L., Sylvester, R., Janssen, P., van der Hage, J. A., van de Velde, C. J., et al. (2006). Heterogeneity in disease free survival between centers: lessons learned from an EORTC breast cancer trial. Clin Trials, 3(1):10–18.
- Liu, L., He, K., Wang, D., Ma, S., Qu, A., Luan, Y., et al. (2024). Health care provider clustering using fusion penalty in quasi-likelihood. Biometrical J, 66(2).
- Ma, S. and Huang, J. (2017). A concave pairwise fusion approach to subgroup analysis. J Am Stat Assoc, 112(517):410–423.
- Moore, M. J., Goldstein, D., Hamm, J., Figer, A., Hecht, J. R., Gallinger, S., et al. (2007). Erlotinib plus gemcitabine compared with gemcitabine alone in patients with advanced pancreatic cancer: a phase III trial of the National Cancer Institute of Canada Clinical Trials Group. J Clin Oncol, 25(15):1960–1966.
- Peikari, M., Salama, S., Nofech-Mozes, S., and Martel, A. L. (2018). A cluster-then-label semi-supervised learning approach for pathology image classification. Sci Rep-uk, 8(1):7193.

REFERENCES

-
- Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. J Am Stat Assoc, 66(336):846–850.
- Robins, J. M. and Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. pages 297–331. Birkhäuser, Boston, MA.
- Shen, X. and Huang, H. C. (2010). Grouping pursuit through a regularization solution surface. J Am Stat Assoc, 105(490):727–739.
- Shultz, D., Pai, J., Chiu, W., Ng, K., Hellendag, M., Heestand, G., et al. (2016). A novel biomarker panel examining response to gemcitabine with or without erlotinib for pancreatic cancer therapy in ncic clinical trials group pa.3. PLoS One, 11(1).
- Tang, X., Xue, F., and Qu, A. (2020). Individualized multi-directional variable selection. J Am Stat Assoc, 116(535):1280–1296.
- Therneau, T. M., Grambsch, P. M., and Pankratz, V. S. (2003). Penalized survival models and frailty. J Comput Graph Stat, 12(1):156–175.
- Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. J R Stat Soc Ser B (Stat Methodol), 67(1):91–108.
- Vaupel, J. W., Manton, K. G., and Stallard, M. E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. Demography, 16(3):439–454.
- Werner, E., Clark, J. N., Hepburn, A., Bhamber, R. S., Ambler, M., Bourdeaux, C. P., et al. (2023). Explainable hierarchical clustering for patient subtyping and risk prediction. Exp Biol Med, 248(244):2457–2559.

REFERENCES

- Witten, D. M., Shojaie, A., and Zhang, F. (2014). The cluster elastic net for high-dimensional regression with unknown variable grouping. Technometrics, 56(1):112–122.
- Xu, J. and Lange, K. (2019). Power k-means clustering. Proceedings of Machine Learning Research, 97:6921–6931.
- You, N., He, S., Wang, X., Zhu, J., and Zhang, H. (2018). Subtype classification and heterogeneous prognosis model construction in precision medicine. Biometrics, 74(3):814–822.
- Zhang, B., Hsu, M., and Dayal, U. (1999). K-harmonic means - a data clustering algorithm. Hewlett-Packard Technical Report.
- Zhang, J., He, X., and Wang, J. H. (2021). Directed community detection with network embedding. J Am Stat Assoc, 117(540):1809–1819.
- Zhu, Y., Shen, X., and Pan, W. (2013). Simultaneous grouping pursuit and feature selection over an undirected graph. J Am Stat Assoc, 108(502):713–725.
- Zou, H. (2006). The adaptive lasso and its oracle properties. J Am Stat Assoc, 101(476):1418–1429.

¹School of Mathematics Science, Sichuan Normal University, China, E-mail: (hey@sicnu.edu.cn)

²Canadian Cancer Trials Group, Queen’s University, Kingston, Canada, E-mail: (dtu@ctg.queensu.ca)

³School of Mathematics Science, Chengdu University of Technology, China, E-mail: (liuli-ums@cdut.edu.cn)

⁴Center of Statistical Research and School of Statistics and Data Science, Southwestern University of Finance and Economics, China, E-mail: (zhouling@swufe.edu.cn)