The Annals of Applied Statistics
2024, Vol. 18, No. 3, 1947–1964
https://doi.org/10.1214/23-AOAS1865

OUTCOME-GUIDED DISEASE SUBTYPING BY GENERATIVE MODEL AND WEIGHTED JOINT LIKELIHOOD IN TRANSCRIPTOMIC APPLICATIONS

BY YUJIA LI[a], PENG LIU[b], WENJIA WANG[c], WEI ZONG[d], YUSI FANG[e], ZHAO REN[f], LU TANG[g], JUAN C. CELEDÓN[h], STEFFI OESTERREICH[i] AND GEORGE C. TSENG[j]

University of Pittsburgh, [a]*yul178@pitt.edu,* [b]*pel67@pitt.edu,* [c]*wew89@pitt.edu,* [d]*wez97@pitt.edu,* [e]*yuf31@pitt.edu,* [f]*zren@pitt.edu,* [g]*lutang@pitt.edu,* [h]*celedonj@pitt.edu,* [i]*oesterreichs@upmc.edu,* [j]*ctseng@pitt.edu*

With advances in high-throughput technology, molecular disease subtyping by high-dimensional omics data has been recognized as an effective approach for identifying subtypes of complex diseases with distinct disease mechanisms and prognoses. Conventional cluster analysis takes omics data as input and generates patient clusters with similar gene expression pattern. The omics data, however, usually contain multifaceted cluster structures that can be defined by different sets of genes. If the gene set associated with irrelevant clinical variables (e.g., sex or age) dominates the clustering process, the resulting clusters may not capture clinically meaningful disease subtypes. This motivates the development of a clustering framework with guidance from a prespecified disease outcome, such as lung function measurement or survival, in this paper. We propose two disease subtyping methods by omics data with outcome guidance using a generative model or a weighted joint likelihood. Both methods connect an outcome association model and a disease subtyping model by a latent variable of cluster labels. Compared to the generative model, weighted joint likelihood contains a data-driven weight parameter to balance the likelihood contributions from outcome association and gene cluster separation, which improves generalizability in independent validation but requires heavier computing. Extensive simulations and two real applications in lung disease and triple-negative breast cancer demonstrate superior disease subtyping performance of the outcome-guided clustering methods in terms of disease subtyping accuracy, gene selection and outcome association. Unlike existing clustering methods, the outcome-guided disease subtyping framework creates a new precision medicine paradigm to directly identify patient subgroups with clinical association.

1. Introduction. Many complex diseases used to be considered as a single disease entity in which all patients receive identical diagnosis and treatment regardless of individual differences. With the advances of modern omics technologies, heterogeneous disease subtypes have been identified with distinct disease mechanisms, therapeutic targets, and survival outcomes. Targeted treatment for many disease subtypes has improved prognosis toward precision medicine. One prominent example is the discovery of breast cancer subtypes of luminal A, luminal B, HER2-enriched and basal-like, first reported by Perou et al. (2000). Luminal-type (luminal A and luminal B) patients tend to have better survival and lower recurrence; HER2-enriched patients are often treated with HER2-targeted therapies, such as trastuzumab; basal-like tumors usually have a poor prognosis with rapid relapse and require a combination of surgery, radiotherapy, and chemotherapy. Precision medicine via successful molecular disease subtyping has decreased breast cancer mortality over the years (Jemal et al. (2009)).

Cluster analysis, such as K-means, Gaussian mixture models, and many others, have been widely used in disease subtyping when the dimensionality of data is low (e.g., cluster analysis

Received June 2023; revised December 2023.
Key words and phrases. Disease subtyping, omics data, high-dimensional cluster analysis, generative model, weighted joint likelihood.

by clinical variables) and when the clusters are well-separated. High-dimensional omics data, such as microarray or RNA-seq data, often have moderate sample size (e.g., 50∼500) but a larger number of genes (e.g., 2,000∼20,000), which leads to a small-n-large-p problem. It has been generally recognized that cluster structure is determined by a small fraction of features (i.e., signature or intrinsic genes) while the other genes are essentially background noise. As a result, most high-dimensional clustering methods simultaneously perform clustering and feature selection (Fop and Murphy (2018), Pan and Shen (2007), Witten and Tibshirani (2010), Zhou, Pan and Shen (2009)). For example, two popular methods for the clustering of high-dimensional omics data embed feature selection in the models: sparse K-means (Witten and Tibshirani (2010)) and penalized model-based clustering (PMBC) (Pan and Shen (2007)) (more details in Supplementary Material Section S1 (Li et al. (2024))). Sparse K-means incorporates gene-specific weights in K-means to allow feature selection by maximizing the weighted between-cluster sum-of-squares with a lasso penalty. In general, only a small set of genes is expected to characterize the clusters. The tuning parameter controlling gene sparsity is selected by gap statistic (Tibshirani, Walther and Hastie (2001)). PMBC is based on the Gaussian mixture model with a lasso penalty to facilitate feature selection, which is determined by Bayesian information criterion (BIC).

The current practice of determining the success and interest of further investigation of identified clusters is to conduct a post hoc association analysis between the clusters and clinical outcomes of interest (e.g., survival). If there is little or no association with the clinical outcome, the omics cluster analysis is considered as a failed exploration. High-dimensional omics data, however, often form multifaceted cluster structures where many are associated with clinically irrelevant variables (e.g., clusters associated with sex, age, or race characterized by their respective gene signatures), which often impede the identification of clinically relevant subtypes. For example, Supplementary Material Figure S1(A) and S1(B) show male/female and young/middle-age/old clusters defined by sex- and age-related genes in a lung disease transcriptomic dataset, which will be investigated further as the first real application in Section 5.1. These well-separated clusters are clinically irrelevant but can mask the discovery of novel disease subtypes that we want to pursue toward precision medicine. To alleviate this problem, Bair and Tibshirani (2004) attempted a two-stage semisupervised clustering method which first selects the top M genes marginally correlated with survival outcome by a Cox model (Cox (1972)), followed by conventional clustering methods such as K-means. Such a two-stage approach is ad hoc in determining M, and marginal screening alone is well-known to ignore gene dependence and can miss critical genes in clustering. Consequently, we will only consider marginal screening as an optional preprocessing step on top of our proposed methods and will assess its usefulness in gene selection.

In this paper we propose two latent class methods, namely, a generative model (abbreviated as $\text{ogClust}_{\text{GM}}$) and a weighted joint likelihood ($\text{ogClust}_{\text{WJL}}$) for outcome-guided disease subtyping. Both methods simultaneously identify disease subtypes with outcome guidance and perform gene selection. As will be shown in Section 2, $\text{ogClust}_{\text{GM}}$ contains two components: an *outcome association model* and a *gene disease subtyping model*, linked by a latent variable of cluster labels. Although $\text{ogClust}_{\text{GM}}$ incorporates outcome guidance to enhance the detection of clinically relevant disease subtypes, it lacks the flexibility to tune the relative contribution of outcome association and gene expression separation, which potentially can reduce generalizability in independent validation. As a remedy, $\text{ogClust}_{\text{WJL}}$ contains a data-driven weight parameter for adjusting the relative likelihood contribution of outcome association and gene cluster separation. We will perform extensive simulations and two real applications to evaluate these two disease subtyping methods with outcome guidance. $\text{ogClust}_{\text{WJL}}$ is expected to alleviate potential overfitting of $\text{ogClust}_{\text{GM}}$ while with an expense of heavier computing.

Both $\text{ogClust}_{\text{GM}}$ and $\text{ogClust}_{\text{WJL}}$ belong to the field of latent class models, which in our view contains at least four major categories. The first category is a set of unsupervised model-based clustering methods (e.g., Gaussian mixture models) (Dean and Raftery (2010), Lanza and Rhoades (2013)), which often identify clinically irrelevant clusters, as we discussed above. The second category links outcome with latent class variables without variables (signatures) to characterize the cluster membership, making it incapable of classifying future patients into the subtypes (Chang et al. (2020), Desantis et al. (2012)). In other words, these models rely on the outcome to classify the patients, but outcome information (e.g., survival) is often unavailable for new patients. The third set of latent class models assumes outcome follows a mixture of distributions where the probability of each component is a function of other variables (e.g., covariates or gene signatures). For a new patient, cluster membership can be determined by the predicted probability of belonging to a mixture component. These models have been widely applied in social sciences, psychology as well as public health (Desantis et al. (2008), Guo, Wall and Amemiya (2006), Houseman, Coull and Betensky (2006)). The generative model in $\text{ogClust}_{\text{GM}}$ belongs to this category. The fourth category is the widely used joint analysis of survival and longitudinal data (Furgal, Sen and Taylor (2019), Lin et al. (2002), Proust-Lima et al. (2014), Proust-Lima and Taylor (2009), Sun et al. (2019)), where survival and longitudinal data come from a common mixture of components and the cluster assignment is determined by the joint-likelihood of both parts. The weighted joint likelihood approach in $\text{ogClust}_{\text{WJL}}$ is closely related to this category, while it has two major innovations, including a data-driven weight parameter to balance the contribution from two data sources of gene expression and clinical outcome (Section 3.2.2), and a weight rescaling step to normalize the likelihood contribution between thousands of genes and one outcome variable (Section 3.2.1). The four categories of latent class models above can also be viewed as a "mixture of experts model" (Gormley and Frühwirth-Schnatter (2019)), which is a broad definition of any mixture model incorporating covariates or concomitant variables. In this paragraph above, we only discuss selected "mixture of experts models," which are relevant to the clustering/disease-subtyping problem that our model is intended to address.

The paper is structured as follows. In Section 2 and Section 3, $\text{ogClust}_{\text{GM}}$ and $\text{ogClust}_{\text{WJL}}$ will be introduced, respectively, followed by inference and tuning parameter selection, and their extension to a survival outcome will be discussed in Supplementary Material Section S6. Benchmarks for method evaluation are introduced in Section 2.2. Extensive simulations and two real applications in lung disease and triple-negative breast cancer (TNBC) are shown in Sections 4 and 5, respectively. Section 6 provides a final conclusion and discussion.

2. Outcome-guided clustering by generative model ($\text{ogClust}_{\text{GM}}$). Throughout the paper, suppose we have n samples from K clusters, p genes, q covariates, and one outcome of interest, and each gene vector is standardized to mean zero. Denote by y_i, $\boldsymbol{g}_i = (g_{i1}, g_{i2}, \ldots, g_{ip})^T$ and $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iq})^T$ the outcome, gene features and covariates of the ith sample ($1 \le i \le n$). The generative model $\text{ogClust}_{\text{GM}}$ assigns n samples into K clinically meaningful clusters based on gene expression features $\mathbb{G} = \{\boldsymbol{g}_i, 1 \le i \le n\}$ with guidance from clinical outcome $\mathbb{Y} = \{y_i, 1 \le i \le n\}$, where the output clustering result is denoted as latent group label $\mathbb{Z} = \{z_i, 1 \le i \le n\}$, $z_i \in \{1, \ldots, K\}$, and $z_i = k$ means that sample i is assigned to cluster k. $\text{ogClust}_{\text{GM}}$ contains two components connected by the latent class $\mathbb{Z}$: *gene disease subtyping model* and *outcome association model*.

The *gene disease subtyping model* is a conventional high-dimensional discriminant analysis in which we train to characterize $\pi_{ik} = Pr(z_i = k|\boldsymbol{g}_i)$ for observation i. In this paper we consider multinomial logistic regression and sparse linear discriminant analysis (Witten and Tibshirani (2011)) for this component and compare them by simulation (see Section 4 and Supplementary Material Figure 8 for comparison). We will leave the description of

sparse linear discriminant analysis (LDA) to Supplementary Material Section S2 and describe multinomial logistic regression here: $\pi_{ik}|\boldsymbol{\gamma} = \frac{\exp(\boldsymbol{g}_i^T \boldsymbol{\gamma}_k)}{\sum_{l=1}^K \exp(\boldsymbol{g}_i^T \boldsymbol{\gamma}_l)}$, where $\boldsymbol{\gamma} = \{\boldsymbol{\gamma}_k, 1 \le k \le K\}$ and $\boldsymbol{\gamma}_k = (\gamma_{1k}, \ldots, \gamma_{pk})^T$. Since p is usually large, we assume only a small subset $\mathcal{A} \subset \{1, \ldots, p\}$ of features are effective in characterizing the clusters that affect the outcome, where its cardinality $\text{card}(\mathcal{A}) < \min(n, p)$. In other words, $\boldsymbol{\gamma}_{[j]} \neq \mathbf{0}$ if $j \in \mathcal{A}$ and $\boldsymbol{\gamma}_{[j]} = \mathbf{0}$ if $j \in \mathcal{A}^c$, where $\boldsymbol{\gamma}_{[j]} = (\gamma_{j1}, \ldots, \gamma_{jK})$. We apply lasso regularization (Tibshirani (1996)) for gene selection (see the next subsection), although group lasso or elastic net could also be considered.

In the *outcome association model*, given covariates $\mathbb{X} = \{\boldsymbol{x}_i, 1 \le i \le n\}$, we assume a mixture model $f(y_i; \boldsymbol{x_i}) = \sum_{k=1}^K \pi_{ik} f_k(y_i; \boldsymbol{x_i})$, where $f_k(y; \boldsymbol{x})$ is the density function of cluster k. We assume a continuous response $\mathbb{Y}$, where the kth mixture density $f_k(y; \boldsymbol{x}, \beta_{0k}, \boldsymbol{\beta}, \sigma)$ is parameterized by cluster specific intercept β_{0k}, common covariate effect $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)^T$, and a homogeneous error σ. Here we impose Gaussian assumption $y_i|z_i = k \sim N(\beta_{0k} + \boldsymbol{\beta}^T \boldsymbol{x_i}, \sigma^2)$ with mixture probability $\pi_{ik} = \frac{\exp(\boldsymbol{g}_i^T \boldsymbol{\gamma}_k)}{\sum_{l=1}^K \exp(\boldsymbol{g}_i^T \boldsymbol{\gamma}_l)}$, $k = 1, \ldots, K$. Denote by $\boldsymbol{\theta} = \{\boldsymbol{\beta}_0, \boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma\}$, the collection of all parameters in $\text{ogClust}_{\text{GM}}$, where $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0K})^T$. Given $\mathbb{Y}$, $\mathbb{X}$, and $\mathbb{G}$, $\boldsymbol{\theta}$ can be estimated by maximizing the following sample likelihood of the basic model: $\text{L}(\boldsymbol{\theta}) = \prod_{i=1}^n \sum_{k=1}^K \pi_{ik}(\boldsymbol{g_i}, \boldsymbol{\gamma}) f_k(y_i; \boldsymbol{x}_i, \beta_{0k}, \boldsymbol{\beta}, \sigma)$.

We note that the outcome $\mathbb{Y}$ can be extended to be a time-to-event survival outcome, and the accelerated failure time (AFT) model is used to model the outcome association (see details in Supplementary Material Section S6). We also note that the current model assumes a simplified common covariate effect $\boldsymbol{\beta}$ across all clusters. It is easy to extend to allow cluster-specific interaction term $\boldsymbol{\beta_k}$, meaning cluster-specific age or sex effects, when the sample size is sufficiently large.

2.1. *Estimation and inference*. Below we develop an EM algorithm for parameter estimation in $\text{ogClust}_{\text{GM}}$. By introducing $z_{ik}, k = 1, \ldots, K$, as missing indicator variables, the complete log-likelihood function can be written as

$$l_n^c(\boldsymbol{\theta}) = \sum_{i=1}^n \sum_{k=1}^K \{z_{ik} \log \pi_{ik} + z_{ik} \log f_k(y_i; \boldsymbol{x}_i, \beta_{0k}, \boldsymbol{\beta}, \sigma)\}, \tag{1}$$

where $z_{ik} = 1$ if sample i belongs to cluster k, and $z_{ik} = 0$ otherwise.

Since gene expression is usually high dimensional, including noninformative genes in $\mathcal{A}^c$ will introduce extra noise to the *gene disease subtyping model*. In the following we will illustrate a lasso penalty for gene selection. We define the penalized log-likelihood function as

$$\tilde{l}_n^c(\boldsymbol{\theta}) = \sum_{i=1}^n \sum_{k=1}^K \{z_{ik} \log \pi_{ik} + z_{ik} \log f_k(y_i; \boldsymbol{x}_i, \beta_{0k}, \boldsymbol{\beta}, \sigma)\} - \lambda R(\boldsymbol{\gamma}), \tag{2}$$

where λ is the regularization tuning parameter and $R(\boldsymbol{\gamma}) = \sum_{j=1}^p \sum_{k=1}^K |\gamma_{jk}|$. This procedure performs feature selection and clustering simultaneously. As alternatives, one could use group lasso penalty $R(\boldsymbol{\gamma}) = \sum_{j=1}^p \|\boldsymbol{\gamma}_{[j]}\|_2$ or, more generally, group elastic net penalty $R(\boldsymbol{\gamma}) = \sum_{j=1}^p \|\boldsymbol{\gamma}_{[j]}\|_2 + \alpha \sum_{j=1}^p \sum_{k=1}^K \gamma_{jk}^2$ (Zou and Hastie (2005)), where $\|\boldsymbol{\gamma}_{[j]}\|_2 = \sqrt{\sum_{k=1}^K \gamma_{jk}^2}$. For computational efficiency we will implement and present the lasso penalty in this paper, and $\boldsymbol{\gamma}$ is estimated following an approximation procedure of Friedman, Hastie and Tibshirani (2010). Maximization of $\tilde{l}_n^c(\boldsymbol{\theta})$ can be achieved by iteratively updating $\boldsymbol{\beta_0}$, $\boldsymbol{\beta}$, σ and $\boldsymbol{\gamma}$ in the EM algorithm, which is described in detail in Supplementary Material Section S3. The pseudo-code for fitting $\text{ogClust}_{\text{GM}}$ is given in Algorithm 1 in Supplementary Material Section

S3. Multiple initial values could be used to avoid convergence to local minimums and increase the numerical stability of clustering.

With the estimated $\hat{\beta}_{0k}$, $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\gamma}}$, $\hat{\sigma}$ by EM algorithm, the conditional probability is

$$\hat{\pi}_{ik}|\hat{\boldsymbol{\gamma}} = \frac{\exp(\boldsymbol{g}_i^T \hat{\boldsymbol{\gamma}}_k)}{\sum_{l=1}^{K} \exp(\boldsymbol{g}_i^T \hat{\boldsymbol{\gamma}}_l)}, \tag{3}$$

and the cluster assignment can be estimated as $\hat{C} = \{\hat{z}_1, \hat{z}_2, \ldots, \hat{z}_n\}$, where $\hat{z}_i = \arg\max_{1 \le k \le K} \hat{\pi}_{ik}$.

2.2. *Benchmarks for evaluation*. In terms of method evaluation for clustering, two primary benchmarks are assessed: clustering accuracy and feature selection accuracy. Additionally, two secondary benchmarks considered are outcome association and gene cluster separation (i.e., gene signature separation across clusters).

Clustering accuracy can be evaluated by adjusted Rand index (ARI) and feature selection accuracy is evaluated by Jaccard index. To evaluate the association of clustering results with a preselected continuous outcome, we calculate the R-squared value and root mean square error (RMSE) of outcome association. Specifically, the R-squared value of outcome association is defined as $R^2_{\text{outcome}}(\hat{C}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \hat{y}_i^{(1)})^2}$ (i.e., the proportion of outcome variation explained by clustering, where the clustering assignment $\hat{C} = \{\hat{z}_1, \hat{z}_2, \ldots, \hat{z}_n\}$ by $\text{ogClust}_{\text{GM}}$ is calculated from equation (3) in Section 2.1). Here $\hat{y}_i$ is predicted outcome by fitting the regression using y_i as dependent variable and cluster assignment $\hat{z}_i$ as well as $\boldsymbol{x_i}$ as independent variables (i.e., $\hat{y}_i = \sum_{k=1}^{K} I\{\hat{z}_i = k\}(\hat{\beta}_{0k} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x_i})$), while $\hat{y}_i^{(1)}$ is the predicted outcome by fitting the regression using y_i as dependent variable and only the covariates $\boldsymbol{x_i}$ as independent variable. The RMSE of outcome association is defined as $\text{RMSE}(\hat{C}) = \sqrt{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$. For gene cluster separation, we calculate average R-squared values among selected genes (i.e., the average proportion of gene expression variation explained by clustering): $R^2_{\text{genes}}(\hat{C}, \hat{G}) = \sum_{j \in \hat{G}}\{1 - \frac{\sum_{i=1}^{n}(g_{ij} - \hat{g}_{ij})^2}{\sum_{i=1}^{n}(g_{ij} - \bar{g}_{\cdot j})^2}\}/|\hat{G}|$, where $\hat{G}$ is the set of selected genes by the clustering algorithm, $\hat{g}_{ij}$ is the predicted gene expression level from the cluster assignment $\hat{z}_i$ and gene expression model, and $\bar{g}_{\cdot j}$ is the overall mean expression of gene j.

Furthermore, if the preselected outcome is survival (right-censored), we benchmark the outcome association by the log-rank test statistic using the partial F statistics of C from fitting the proportional hazard regression $y_i \sim \hat{z}_i + \boldsymbol{x_i}$. We also calculate another outcome association benchmark by the adjusted C-index. The C-index (Pencina and D'Agostino (2004)) is a concordance index to measure the consistency between $\hat{y}_i$ and y_i for time-to-event data, which can be calculated by R package "survcomp" (Schröder et al. (2011)). To avoid skewed interpretation where the C-index is always positive, even in random data, we calculate the adjusted C-index by $\text{cindex}_{\text{adjust}} = \frac{\text{cindex} - E(\text{cindex})}{1 - E(\text{cindex})}$, where $E(\text{cindex})$ is calculated by permutation analysis under the null hypothesis (i.e., the average C-index from repeated permutation of $\hat{y}_i$). This adjustment is similar to the adjusted Rand index (Hubert and Arabie (1985)).

We will benchmark the $R^2_{\text{genes}}(\hat{C}, \hat{G})$, $R^2_{\text{outcome}}(\hat{C})$, $\text{RMSE}(\hat{C})$, log-rank test statistics, and adjusted C-index in both training and testing cohorts. We note that clustering accuracy and feature selection accuracy can be evaluated in simulations where the underlying truths of informative genes and sample cluster labels are known, but they are not feasible in real applications.

2.3. *Tuning parameter selection.* There are two tuning parameters of ogClust$_{\text{GM}}$: K and λ. Many methods have already been proposed and evaluated for determining K (e.g., gap statistic, BIC, or resampling evaluation), which can also be applied here. In this section we focus on the selection of λ, when K is prespecified, and determining λ based on the new criterion described below.

Given N candidate values of λ: $\vec{\lambda} = \{\lambda_1, \ldots, \lambda_N\}$, for each λ we perform 10-fold cross-validation. Specifically, within each fold of the training/testing split, we fit ogClust$_{\text{GM}}$ using the training set, predict the cluster labels in the testing fold, and then calculate the average R-squared value of selected genes as $R^2_{\text{genes}}(\hat{C}, \hat{G}; \lambda_r)$ and the R-squared value of outcome association as $R^2_{\text{outcome}}(\hat{C}; \lambda_r)$ (both are defined in Section 2.2). We determine $\hat{\lambda}$ by the geometric mean of R^2_{outcome} and R^2_{genes}: $\hat{\lambda} = \arg\max_{1\le r\le N}(\sqrt{(R^2_{\text{outcome}}(\hat{C}; \lambda_r) \cdot R^2_{\text{genes}}(\hat{C}, \hat{G}; \lambda_r))}$. Note that, compared to arithmetic mean, the geometric mean avoids a low value in either R^2_{outcome} or R^2_{genes} (e.g., R^2's $= (0.95, 0.25)$ vs. $(0.6, 0.6)$ gives the same arithmetic mean, but the latter is more balanced and preferred).

3. Outcome-guided clustering by weighted joint likelihood (ogClust$_{\text{WJL}}$). In ogClust$_{\text{GM}}$, the generative information flows from the *gene disease subtyping model* to *outcome association model* through the latent class variable. Although its full model-based approach is appealing, the likelihood contribution from outcome association tends to dominate since *gene disease subtyping model* only contributes indirectly through mixture probabilities π_{ik}. This often generates overfitting in the training data with seemingly higher outcome association, while the outcome association can greatly reduce in the testing study. To circumvent this issue, we develop a weighted joint likelihood approach ogClust$_{\text{WJL}}$, motivated from a category of methods for joint analysis of survival and longitudinal data (see the fourth category of existing methods in Section 1). We propose to link observed data $\{y_i, \boldsymbol{x_i}, \boldsymbol{g_i}\}$ by a weighted penalized joint log-likelihood,

$$(4)\qquad L(\boldsymbol{\theta_1}) = \sum_{i=1}^{n}\big[(1-w)\cdot\log\big(f_g(\boldsymbol{g}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma})\big) + w\cdot\log\big(f_o\big(y_i|\boldsymbol{x_i}, \boldsymbol{\beta}_0, \boldsymbol{\beta}, \sigma^2\big)\big)\big] - \lambda h(\boldsymbol{\mu}),$$

where $f_g(\boldsymbol{g}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{k=1}^{K}\pi_k \cdot [\prod_{j=1}^{p} N(g_{ij}|\mu_{jk}, \sigma_j^2)]$, $f_o(y_i|\boldsymbol{x_i}, \boldsymbol{\beta}_0, \boldsymbol{\beta}, \sigma^2) = \sum_{k=1}^{K}\pi_k \cdot [N(y_i|\beta_{0k} + \boldsymbol{\beta}^T\boldsymbol{x_i}, \sigma^2)]$, $\boldsymbol{\theta_1} = \{\boldsymbol{\beta}_0, \boldsymbol{\beta}, \{\boldsymbol{\mu}_k\}_{k=1}^{K}, \boldsymbol{\Sigma}, \sigma\}$, and $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0K})^T$. Here f_g denotes the density of gene features $\boldsymbol{g_i}$, which is a mixture of distributions of K components with probability π_k for the kth component ($1 \le k \le K$), and cluster-specific density $f_{gk} = \prod_{j=1}^{p} N(g_{ij}|\mu_{jk}, \sigma_j^2)$ is constructed by the multivariate normal density with cluster-specific mean vector $\boldsymbol{\mu_k} = (\mu_{1k}, \mu_{2k}, \ldots, \mu_{pk})^T$ and common diagonal covariance matrix $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_p^2)$ across all clusters, assuming independence of gene features. Same as ogClust$_{\text{GM}}$, outcome association is modelled by f_o, which follows a normal distribution with mean $\beta_{0k} + \boldsymbol{\beta^T x_i}$ for cluster k and homogeneous standard deviation σ. That is, $y_i|z_i = k \sim N(\beta_{0k} + \boldsymbol{\beta}^T\boldsymbol{x_i}, \sigma^2)$, where z_i is the latent cluster membership for the ith sample. Similar ogClust$_{\text{GM}}$, ogClust$_{\text{WJL}}$ can be extended to accommodate time-to-event outcome by the accelerated failure time (AFT) model (details described in Supplementary Material Section S6). To achieve feature selection, we similarly utilize lasso penalty $h(\boldsymbol{\mu}) = \sum_{j=1}^{p}\sum_{k=1}^{K}|\mu_{jk}|$, ensuring that only a small fraction of genes contribute to categorizing the clusters. Since each gene vector is standardized to a zero mean, if $\mu_{jk} = 0$ for all k ($1 \le k \le K$), then the jth gene does not contribute to the clustering of omics data. λ is a tuning parameter controlling gene selection, and w is a weight to determine the relative contribution of outcome association and gene clustering pattern.

The weight w is a critical parameter in ogClust$_{\text{WJL}}$. Intuitively, $f_g(\boldsymbol{g}_i|\boldsymbol{\theta_1})$ pursues cluster separation based on gene expression data, and $f_o(y_i|\boldsymbol{x_i}, \boldsymbol{\theta_1})$ helps seek clusters with outcome

association. If $w = 0$, the model is reduced to unsupervised clustering based on omics data, and the identified clusters may be irrelevant to clinical outcome y_i, as discussed in Section 1. If w is close to 1, the model mostly emphasizes on the outcome association and may obtain weak gene pattern in clusters, which weakens generalizability in test studies. Data-driven parameter selection of w will be discussed in Section 3.2 in detail.

3.1. *Estimation and inference.* Maximization of equation (4) can be achieved by using the EM algorithm. By introducing latent variable $z_{ik} = 1$ if the ith sample belongs to the kth cluster and 0 otherwise, the problem becomes maximization of the following complete penalized log-likelihood:

$$
\begin{aligned}
L_c(\boldsymbol{\theta_2}) = \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\Bigg\{ & \log(\pi_k) + (1-w)\sum_{j=1}^{p}\log\big(N\big(g_{ij}|\mu_{jk},\sigma_j^2\big)\big) \\
& + w\log\big(N\big(y_i|\beta_{0k} + \boldsymbol{\beta}^T\boldsymbol{x_i},\sigma^2\big)\big)\Bigg\} - \lambda\sum_{j=1}^{p}\sum_{k=1}^{K}|\mu_{jk}|,
\end{aligned} \tag{5}
$$

where $\boldsymbol{\theta_2} = \{\boldsymbol{\theta_1}, z\}$ and $z = \{z_{ik}\}$ $(1 \le i \le n$ and $1 \le k \le K)$. Optimization of equation (5) is achieved by EM algorithm, which is described in detail in Supplementary Material Section S4. With the estimated $\hat{\beta}_{0k}$, $\hat{\pi}_k$, $\hat{\sigma}^2$, $\hat{\boldsymbol{\beta}}$, $\hat{\mu}_{jk}$, and $\hat{\sigma}_j^2$, the final cluster membership probability is

$$
\hat{z}_{ik} = \frac{\hat{\pi}_k\prod_{i=1}^{n}\{\prod_{j=1}^{p}N(g_{ij}|\hat{\mu}_{jk},\hat{\sigma}_j^2)\}^{(1-w)}\{N(y_i|\hat{\beta}_{0k} + (\hat{\boldsymbol{\beta}})^T\boldsymbol{x_i},\hat{\sigma}^2)\}^w}{\sum_{l=1}^{K}\hat{\pi}_l\prod_{i=1}^{n}\{\prod_{j=1}^{p}N(g_{ij}|\hat{\mu}_{jl},\hat{\sigma}_j^2)\}^{(1-w)}\{N(y_i|\hat{\beta}_{0l} + (\hat{\boldsymbol{\beta}})^T\boldsymbol{x_i},\hat{\sigma}^2)\}^w}, \tag{6}
$$

and the cluster assignment can be estimated as $\hat{C} = \{\hat{z}_1, \hat{z}_2, \ldots, \hat{z}_n\}$, where $\hat{z}_i = \arg\max_{1\le k\le K}\hat{z}_{ik}$. One of the most important goals of disease subtyping is to predict the subtype label of new patients without knowing the outcome value, which can be predicted by

$$
\hat{z}_{ik,\text{validation}} = \frac{\hat{\pi}_k\prod_{i=1}^{n}\prod_{j=1}^{p}N(g_{ij,\text{validation}}|\hat{\mu}_{jk},\hat{\sigma}_j^2)}{\sum_{l=1}^{K}\hat{\pi}_l\prod_{i=1}^{n}\prod_{j=1}^{p}N(g_{ij,\text{validation}}|\hat{\mu}_{jl},\hat{\sigma}_j^2)} \tag{7}
$$

using only the gene expression data. Similarly, we can conclude the predicted label of the validation samples as $\hat{z}_{\text{i,validation}} = \arg\max_{1\le k\le K}\hat{z}_{ik,\text{validation}}$.

To evaluate $\text{ogClust}_{\text{WJL}}$, we assess the same benchmarks in Section 2.2, and the clustering assignment $\hat{C} = \{\hat{z}_1, \hat{z}_2, \ldots, \hat{z}_n\}$ here is calculated from equation (6) and equation (7).

3.2. *Tuning parameter selection.* $\text{ogClust}_{\text{WJL}}$ has three tuning parameters: K, w, and λ. Similar to $\text{ogClust}_{\text{GM}}$, in the subsections below we focus on the selection of w and λ when K is prespecified.

3.2.1. *Reparameterization of weight w.* In equation (5), since we have p genes but only one outcome, the magnitude of gene expression likelihood $\sum_{j=1}^{p}\log(N(g_{ij}|\mu_{jk},\sigma_j^2))$ is much larger than that of outcome log-likelihood $\log(N(y_i|\beta_{0k} + \boldsymbol{\beta}^T\boldsymbol{x_i},\sigma^2))$, which can be an issue in grid search of w. Denote by s_0 the true number of informative genes, which is an unknown quantity. Equation (5) can be written as

$$
\begin{aligned}
L_c(\boldsymbol{\theta_2}) = \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\Bigg\{ & \log(\pi_k) + [(1-w)s_0]\frac{\sum_{j=1}^{p}\log(N(g_{ij}|\mu_{jk},\sigma_j^2))}{s_0} \\
& + w\log\big(N\big(y_i|\beta_{0k} + \boldsymbol{\beta}^T\boldsymbol{x_i},\sigma^2\big)\big)\Bigg\} - \lambda\sum_{j=1}^{p}\sum_{k=1}^{K}|\mu_{jk}|.
\end{aligned}
$$

We reparametrize $w_1 = \frac{w}{w+(1-w)s_0}$ (or, equivalently, $w = \frac{s_0 \cdot w_1}{s_0 \cdot w_1 + 1 - w_1}$). Since $\frac{\sum_{j=1}^{p} \log(N(g_{ij}|\mu_{jk}, \sigma_j^2))}{s_0}$ and $\log(N(y_i|\beta_{0k} + \boldsymbol{\beta}^T \boldsymbol{x_i}, \sigma^2))$ are roughly in the same scale and comparable, uniform grid search of $0 \le w_1 \le 1$ is more effective than that of the original parameterization $0 \le w \le 1$. As shown in Table S1, when $s_0 = 500$, to search w_1 from 0.04 to 1, the effective searching space of w is from 0.95 to 1, which brings difficulty for searching optimal w in practice. As a result, for determining optimal w, we perform a uniform grid search for w_1, transform it back to w, and then solve equation (5), as we previously discussed. This approach does not change the optimization process (i.e., one-to-one mapping between w and w_1). Since the true s_0 is unknown, we need to estimate and plug in $\hat{s}_0$. Table S1 indicates that, even if we cannot exactly estimate $\hat{s}_0$, searching for w_1 evenly between 0 and 1 provides certain robustness for wrongly estimated $\hat{s}_0$. For an extreme example, even if we incorrectly estimate $\hat{s}_0$, denoting its estimate as $\hat{s}_0 = 50$ when the true value is $s_0 = 500$, searching for $w_1(\hat{s}_0 = 50)$ equally between 0 to 1 can still cover most of the search space of $w_1(s_0 = 500)$ (see Supplementary Material Table S1(B)). In practice, we plug in $\hat{s}_0 = 0.1 \times p$. In the rest of the paper, we discuss the selection of weight w_1 instead of the original weight w since they are equivalent.

We note that the weight w has monotonic interpretation in that larger weight means heavier contribution of outcome association in the joint likelihood. The magnitude of weight, however, does not have a direct interpretation since the likelihood of outcome association and that of gene clustering are not directly comparable. For example, $w = 0.5$ or $w_1 = 0.5$ does not mean equal contribution from outcome association and gene clustering in a theoretical sense. The reparameterization here is mainly for computational purpose for effective grid search of the weight.

3.2.2. *Determining optimal w_1 and λ.* We conduct a two-dimensional search for w_1 and λ. For each (w_1, λ) pair, we perform 10-fold cross-validation. Specifically, within each fold of training/testing split, we fit $\text{ogClust}_{\text{WJL}}$ using the training set and predict cluster label in the testing fold using equation (7). We calculate the average R-squared value of selected genes as $R^2_{\text{genes}}(\hat{C}, \hat{G}; w_1, \lambda)$, which intuitively measures the degree of cluster separation in gene expression of selected genes, with larger values being better. The R-squared value of outcome fitness conditional on the covariates can be calculated as $R^2_{\text{outcome}}(\hat{C}; w_1, \lambda)$, which intuitively measures whether the predicted cluster can separate the outcome well in the cross-validation, conditional on the covariates. Both R^2_{genes} and R^2_{outcome} are defined in Section 2.2.

We note that the roles and importance of w_1 and λ in clustering differ. w_1 determines the relative contribution between gene likelihood and outcome likelihood and has a more critical impact on the final clustering result. In contrast, λ controls gene selection, and a slight change of λ usually does not significantly alter the clustering result. Hence, instead of a naive two-dimensional grid search, we propose a two-stage approach by first focusing on whether predicted subtypes by gene expression data have good outcome separation (i.e., $R^2_{\text{outcome}}(\hat{C}; w_1, \lambda)$). Specifically, we obtain $\hat{w}_1$ by $\hat{w}_1 = \arg\max_{w_1}(\max_{\lambda} R^2_{\text{outcome}}(\hat{C}; w_1, \lambda))$. After $\hat{w}_1$ is selected, we determine $\hat{\lambda}$ by the geometric mean of $R^2_{\text{outcome}}(\hat{C}; \hat{w}_1, \lambda)$ and $R^2_{\text{genes}}(\hat{C}, \hat{G}; \hat{w}_1, \lambda)$, the same criteria as $\text{ogClust}_{\text{GM}}$: $\hat{\lambda} = \arg\max_{\lambda}(\sqrt{(R^2_{\text{outcome}}(\hat{C}; \hat{w}_1, \lambda) \cdot R^2_{\text{genes}}(\hat{C}, \hat{G}; \hat{w}_1, \lambda))}$.

As mentioned in Section 3.2.1, searching for w_1 evenly between 0 and 1 is reasonable. Suppose we generate grids of T weights $\{w_1^1, w_1^2, \ldots, w_1^T\}$. The next issue lies in designing the grids of $\vec{\lambda} = \{\lambda_1, \ldots, \lambda_N\}$, where N is the number of λ's to be searched for each weight. We note that if $\vec{\lambda}$ is not properly designed, most of the computing time will be wasted since many λ values select almost identical number of genes. Consequently, for each w_1 the proper grid $\vec{\lambda}$ should be designed differently, denoted as $\vec{\lambda}_{w_1} = (\lambda_{w_1,1}, \ldots, \lambda_{w_1,N})$. Inspired

by the efficient search algorithm by Li et al. (2022), we develop a similar bisection search algorithm for each weight w_1, which makes the corresponding numbers of selected genes of $\vec{\lambda}_{w_1}$ roughly equally spaced in log2 scale. The detailed bisection search algorithm is shown in Supplementary Material Section S5. In practice, to reduce computing time, it is recommended to use a small N (e.g., 10–20) when selecting $\hat{w}_1$ in the first stage. In the second stage of selecting $\hat{\lambda}$, a more dense search with a large N (e.g., 50–100) can be used.

4. Simulations.

4.1. *Simulation settings*. We conduct two simulations to evaluate the performance of clustering, feature selection, and outcome association of $\text{ogClust}_{\text{GM}}$ and $\text{ogClust}_{\text{WJL}}$ while comparing with existing methods sparse K-means and PMBC. Note that both sparse K-means and PMBC only take $\boldsymbol{g}_i$ as input to identify clusters, while clusters by $\text{ogClust}_{\text{GM}}$ and $\text{ogClust}_{\text{WJL}}$ are characterized by $\boldsymbol{g}_i$ with guidance from outcome y_i. In this section we describe and show simulation results with guidance from a continuous outcome, while leaving simulations for survival outcome in Supplementary Material Section S7. In terms of the simulation setting with a continuous outcome, we simulate $n = 99$ samples with $K = 3$ clusters, one outcome variable, two covariates, and 2000 genes in total where 50 true signal genes (p_1), 50 genes that define random clusters (p_2), 50 genes that are correlated with covariates (p_1), and the rest is noise genes (p_4). We vary the effect size for outcome separation $c_1 \in \{2, 3\}$ and the effect size for gene cluster signal $\mu \in \{0.9, 1.2, 1.5, 1.8\}$ to evaluate the performances under different levels of outcome association and gene cluster separation. For each setting we repeat 50 times. Detailed simulation procedures are described in Supplementary Material Section S7.1.

In the evaluation and for each simulated dataset, we randomly split samples into three folds where two folds serve as training/discovery data and the remaining one for external testing/validation. We assume the true number of informative genes is unknown and use our proposed approach (see Section 2.3 and 3.2) to select w_1 and λ for $\text{ogClust}_{\text{WJL}}$ and λ for $\text{ogClust}_{\text{GM}}$, while sparse K-means and PMBC use gap statistic and BIC, respectively, to select genes (Pan and Shen (2007), Witten and Tibshirani (2010)). In addition, to allow a fair comparison, we perform an additional comparison when each method selects the number of genes close to the underlying truth $p_1 = 50$ (i.e., by performing a grid search of λ and selecting the λ that generates the number of genes closest to the underlying truth).

4.2. *Simulation results*. Figure 1 shows the result at $c_1 = 2$ when the true number of signal genes ($p_1 = 50$) is unknown. The result demonstrates a general superior performance of outcome-guided clustering methods ($\text{ogClust}_{\text{GM}}$ and $\text{ogClust}_{\text{WJL}}$) over traditional clustering without outcome guidance (sparse K-means and PMBC) in terms of disease subtyping accuracy, feature selection, and outcome association. As expected, $\text{ogClust}_{\text{WJL}}$ improves the overfitting of $\text{ogClust}_{\text{GM}}$ and shows better performance in some independent validation scenarios.

In Figure 1 with continuous outcome guidance and an unknown number of informative genes, the average number of selected genes from 50 simulations (standard error in parenthesis) by each method are shown in Figure 1(A) for varying μ and the boxplots of corresponding Jaccard indexes are shown in Figure 1(B). The result shows inadequate feature selection by sparse K-means, PMBC and $\text{ogClust}_{\text{GM}}$ and that although the numbers of selected features are close to the truth $p_1 = 50$, Jaccard indexes are less than 0.3, even when gene signal is strong at $\mu = 1.8$. $\text{ogClust}_{\text{WJL}}$ performs the best with increasingly accurate feature selection in strong signal cases (selects about 47 genes and Jaccard index ≈ 0.9 when $\mu = 1.8$). In terms of disease subtyping performance, Figure 1(C) shows clustering accuracy by ARI

(A)

Method \ μ	0.9	1.2	1.5	1.8
$ogClust_{GM}$	41 (1.5)	35 (1.2)	34 (1.1)	33 (1.3)
$ogClust_{WJL}$	31 (2.6)	43 (2.4)	44 (1.7)	47 (1.1)
sparse K-means	53 (4.8)	66 (3.5)	63 (0.8)	62 (0.7)
PMBC	54 (2.3)	64 (4.1)	56 (1.1)	57 (2.8)

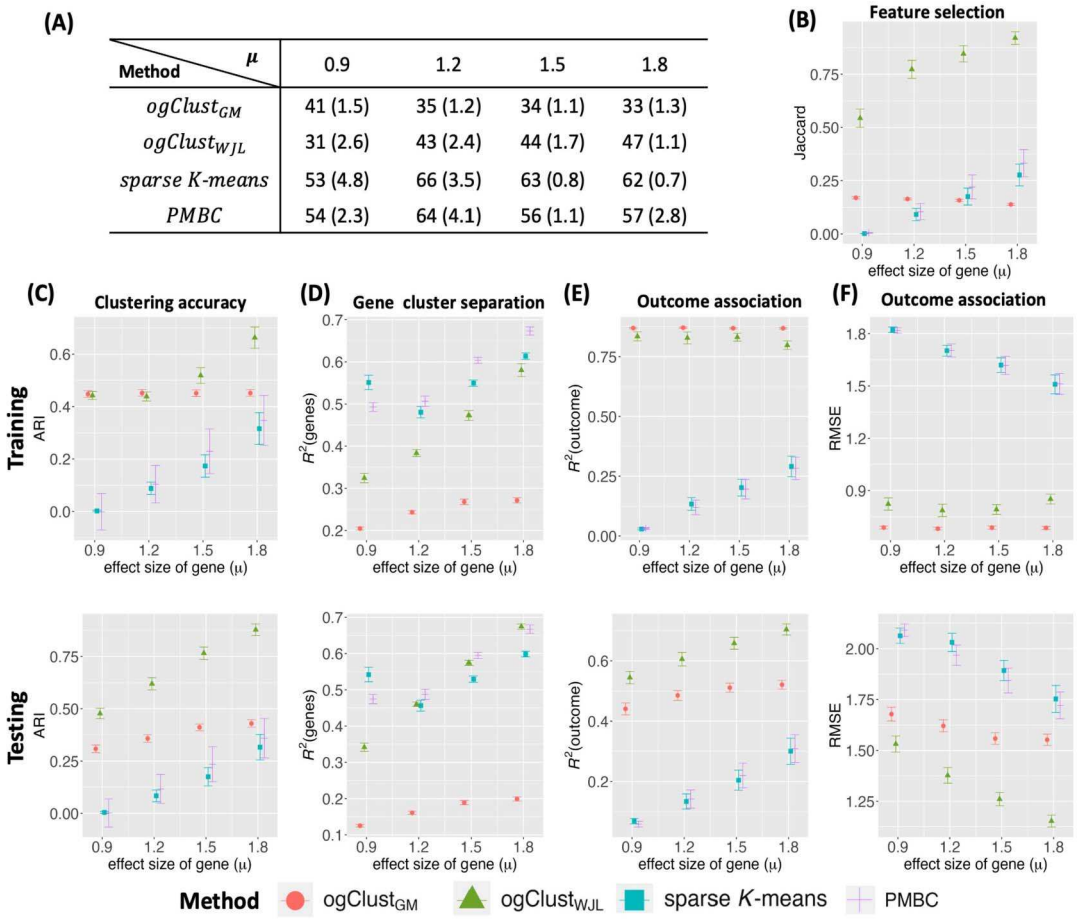

FIG. 1. *Result for Simulation I under outcome association effect size* $c_1 = 2$, *assuming the true number of informative features is unknown for all four methods. Tuning parameters of* $\text{ogClust}_{\text{GM}}$ *and* $\text{ogClust}_{\text{WJL}}$ *are determined by the approach in Section* 2.3 *and Section* 3.2 *while tuning parameters of sparse K-means and PMBC are selected by gap statistic and BIC, respectively. Table* (A) *shows the average number of features selected over* 50 *replications with standard error in parenthesis by each method under different values of* μ. *Figure* (B) *shows the distribution of the Jaccard index from* 50 *replications for benchmarking feature selection accuracy. Figures* (C–F) *show ARI* (*clustering accuracy*), R^2_{genes} *of selected genes* (*gene cluster separation*), R^2_{outcome} (*outcome association*), *and RMSE* (*outcome association*) *in the training* (*upper*) *and testing* (*lower*) *data, respectively. The error bar represents the standard error across* 50 *replications.*

of each method. Sparse K-means and PMBC both do not perform well (ARI $= 0 - 0.4$) in training and testing. $\text{ogClust}_{\text{GM}}$ has high clustering accuracy (ARI $= 0.4 - 0.5$) in training but shows overfitting with lower accuracy in testing. In contrast, $\text{ogClust}_{\text{WJL}}$ has the best performance with the highest ARI in both training and testing data. Figure 1(D) shows gene cluster separation by R^2_{genes} of selected genes. The result presents an improved performance of $\text{ogClust}_{\text{WJL}}$, compared to $\text{ogClust}_{\text{GM}}$, indicating stronger gene signatures in identified disease subtypes. Sparse K-means and PMBC have comparable or sometimes higher R^2_{genes} because they might have selected genes from clinically irrelevant clusters (p_2 genes) or with covariate association (p_3 genes), which does not mean better performance. Figure 1(E) and 1(F) illustrate outcome association of identified clusters by R^2_{outcome} and RMSE, respectively. The two outcome-guided clustering methods outperform sparse K-means and PMBC by a large margin. $\text{ogClust}_{\text{GM}}$ shows stronger overfitting drawback and often has worse performance than $\text{ogClust}_{\text{WJL}}$ in testing data. Similar results hold when $c_1 = 3$ in Supplementary Material Figure S2 where both $\text{ogClust}_{\text{GM}}$ and $\text{ogClust}_{\text{WJL}}$ have better performance, and $\text{ogClust}_{\text{WJL}}$ in

particular shows almost perfect feature selection (Jaccard index ≈ 1) and clustering accuracy (ARI ≈ 1) in strong gene signal.

Since Figure 1 corresponds to an unknown number of informative genes, the performance of different methods in Figure 1(C)–(F) may depend on different numbers of selected genes in Figure 1(A). Alternatively, we perform another set of evaluations, assuming that $p_1 = 50$ is given and all methods select as close to 50 genes as possible. The corresponding results are included in Supplementary Material Figure S3 with almost identical conclusions as shown in Figure 1(C)–(F).

To evaluate the impact of marginal screening (i.e., selecting top outcome-associated genes before clustering), Supplementary Material Figure S4 shows the result of preselecting the top 500 outcome-associated genes out of 2000 genes. Performance of feature selection, disease subtyping, and outcome association improve significantly for sparse K-means and PMBC. The improvement is of a much smaller magnitude for ogClust$_{\mathrm{GM}}$ and ogClust$_{\mathrm{WJL}}$, demonstrating their strength with embedded outcome guidance. Since the decision of marginal screening parameter (i.e., the number of prescreened genes) is usually subjective, we recommend marginal screening as an optional step in our software package and generally do not recommend it for ogClust$_{\mathrm{GM}}$ and ogClust$_{\mathrm{WJL}}$.

The simulation setting above contains an equal sample size for all three clusters. To evaluate the performance of detecting a rare disease subtype (cluster) with a relatively smaller sample size, Supplementary Material Figure S9 shows results with sample size ratio 3:3:1, 2:2:1, 3:3:2, 1:1:1 with identical total sample size (i.e., the smallest cluster has 1/7, 1/5, 1/4, and 1/3 of the total sample size). Similar to previous results, ogClust$_{\mathrm{WJL}}$ has the best performance of the four methods in the testing data. As expected, performance decreases when the smallest cluster has fewer samples in ogClust$_{\mathrm{GM}}$ and ogClust$_{\mathrm{WJL}}$.

5. Real applications.

5.1. *Lung disease transcriptomic application.* In this subsection we evaluate ogClust$_{\mathrm{GM}}$ and ogClust$_{\mathrm{WJL}}$ on two lung disease transcriptomic studies from the Lung Genomics Research Consortium (for details, see Supplementary Material Section S8), which contains the datasets of 12,958 common genes over 218 COPD (chronic obstructive pulmonary disease) patients and 249 ILD (interstitial lung disease) patients in total. Currently, COPD and ILD are classified purely by clinical features and, due to somewhat shared symptoms between these two diseases, the classification is often debatable. Forced expiratory score (FEV1) measures the amount of air that can be forced from lungs in one second, which is a key clinical feature and is used as the outcome to guide disease subtyping, while age and sex are utilized as relevant covariates. In data preprocessing we first filter out half of the genes with low average expression, and among the remaining half, we further keep the top 50% genes of large standard deviation (i.e., filter out nonexpressed genes by mean and noninformative genes by standard deviation), resulting in 3,240 remaining genes. In addition, the downloaded FEV1 outcome is normalized to zero mean and unit standard deviation, although the standardization does not impact likelihood in the methods.

We treat data from the study of 143 COPD and 188 ILD patients as the discovery study and the remainder as a validation study. We perform disease subtyping using the discovery cohort with the number of clusters $K = 3$ by ogClust$_{\mathrm{GM}}$, ogClust$_{\mathrm{WJL}}$, sparse K-means, and PMBC, and evaluate the four methods by performing clustering in the training cohort and apply the estimated clustering parameters to determine clustering in the testing cohort. We first apply tuning parameter selection criteria in each method for selecting λ in ogClust$_{\mathrm{GM}}$, w_1 and λ in ogClust$_{\mathrm{WJL}}$ and feature selection in sparse K-means (by gap statistic) and PMBC (by BIC). It turns out that the four methods select very different numbers of genes (197, 300, 3240 and

(A)

Method \ # Genes	Tune	Fix ~ 300
$ogClust_{GM}$	197	301
$ogClust_{WJL}$	300	308
sparse K-means	3240	304
PMBC	2674	295

(B)

Method \ Criteria	Outcome association				Gene cluster separation	
	$R^2_{outcome}$		*RMSE*		R^2_{genes}	
	Train	Test	Train	Test	Trian	Test
$ogClust_{GM}$	0.83	0.24	0.40	0.97	0.03	0.04
$ogClust_{WJL}$	0.80	0.31	0.57	0.84	0.07	0.32
sparse K-means	0.02	0.01	0.97	0.99	0.55	0.53
PMBC	0.04	0.01	0.95	0.98	0.53	0.51

(C)

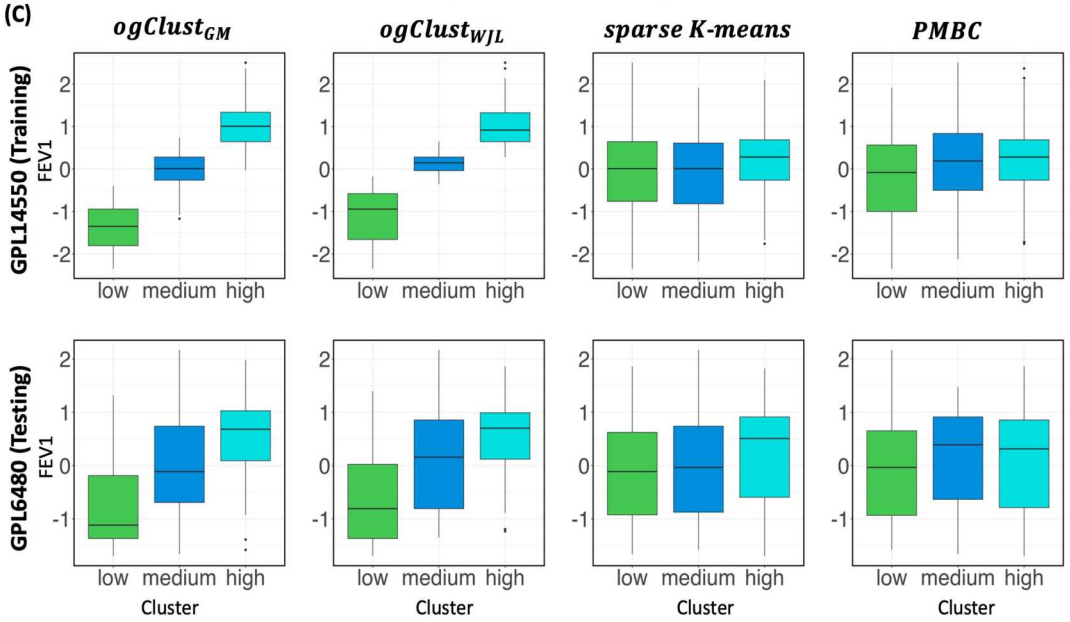

FIG. 2. *Results of independent validation evaluation of* $\text{ogClust}_{\text{GM}}$, $\text{ogClust}_{\text{WJL}}$, *sparse K-means, and PMBC in the lung disease application. Table* (*A*) *shows the different numbers of genes selected for clustering after parameter tuning of each method as well as the number of genes used in clustering when selecting as close to* 300 *genes as possible for all the methods. With about* 300 *genes selected by each method, Figure* (*B*) *shows* R^2_{outcome} *and RMSE, which reflects the outcome association, as well as* R^2_{genes} *for gene cluster separation by each method in training and testing data. Figure* (*C*) *is the boxplot of the outcome observations FEV*1 *in three predicted clusters by each method in the training* (*upper*) *and testing* (*lower*) *study, respectively.*

2674 genes in Figure 2(A)). For a fair comparison without impact from varying gene selection, we select roughly 300 genes in all four methods and the result is shown in Figure 2(B) and 2(C). To evaluate the sensitivity of choosing different numbers of selected genes, Supplementary Material Figure S6(II) shows the method comparison of selecting about 400 genes. For each method, the cluster label is reordered to "low, medium," and "high" according to the median of FEV1 value in the cluster from discovery/training cohort analysis.

Unlike in simulation, the true informative genes and true disease subtyping are unknown in real applications, so evaluations by ARI and Jaccard index are not applicable. Figure 2(B) shows results of outcome association (by R^2_{outcome} and RMSE) and gene cluster separation (by R^2_{genes}) from applying the four methods in the training and validating in the testing cohort. Similar to the simulation result, both outcome-guided clustering methods generate sample clusters with better outcome association (larger R^2_{outcome} and smaller RMSE) than sparse K-means and PMBC. Clustering from $\text{ogClust}_{\text{WJL}}$ has slightly better outcome association and less overfitting in the testing study result than that from $\text{ogClust}_{\text{GM}}$. The boxplots in Figure 2(C) further show outcome association between identified clusters and normalized

FEV1 in the learning in training and validation in the testing study. We find that differential FEV1 values across clusters are clear in the $\text{ogClust}_{\text{GM}}$ and $\text{ogClust}_{\text{WJL}}$ results but not in sparse K-means and PMBC, showing failure of the two traditional methods to identify clinically relevant clusters. In terms of gene cluster separation, $\text{ogClust}_{\text{WJL}}$ slightly outperforms $\text{ogClust}_{\text{GM}}$ (larger R^2_{genes}) in Figure 2(B) for both training and testing studies. Sparse K-means and PMBC appear to identify well-separated cluster patterns in gene expression ($R^2_{\text{genes}} \approx 0.5$) but with almost no clinical association with FEV1. All the findings and conclusions are similar in Supplementary Material Figure S6(I) and S6(II), where the number of informative genes is learned from the algorithm or assumed at about 400 genes, respectively.

We next conduct pathway enrichment analysis for the $\sim$300 genes selected by each method using Ingenuity Pathway Analysis (IPA). Supplementary Material Table S2 shows the number of detected pathways under different cutoffs of p-value and false discovery rate (FDR). At FDR$=5\%$ threshold, $\text{ogClust}_{\text{WJL}}$ identifies 40 enriched pathways, while the other three methods do not detect any. To provide a fair presentation, Supplementary Material Table S3 shows the union of the top 10 pathways detected by the four methods and their enrichment p-values. $\text{ogClust}_{\text{WJL}}$ identifies many pathways with highly significant p-values associated with immune responses and organismal injury while the top pathways from the other three methods are less relevant to lung disease. For example, $\text{ogClust}_{\text{WJL}}$ identifies IL-10 signaling and IL-6 signaling pathway, which has important ramifications for the diseases of asthma and chronic obstructive pulmonary disease (Lin et al. (2021), Ogawa, Duru and Ameredes (2008)), pathogen-induced cytokine storm signaling pathway, where anticytokine therapy has potentials in chronic obstructive pulmonary disease (Chung (2001)), role of chondrocytes in rheumatoid arthritis signaling pathway, and many other important pathways.

5.2. *Triple-negative breast cancer transcriptomic application.* Triple-negative breast cancer (TNBC) accounts for about 10–15% of all breast cancers, which is categorized by lack of ER, PR and HER2 protein expression. TNBC lacks targeted therapy and usually has the worst prognosis compared with ER, PR, or HER2 positive breast cancers. In this subsection we investigate two large-scale breast cancer studies (for details, see Supplementary Material Section S8) with transcriptomic data of 275 and 151 TNBC samples, respectively, from Illumina HT-12 microarray and Illumina HiSeq to evaluate the disease subtyping in TNBC of the four methods discussed. We take the study of 275 samples as the training cohort while the other as the testing cohort. Overall survival is used as the outcome variable for subtyping guidance, and age at diagnosis is utilized as a relevant covariate. In total, the two datasets have 18,964 genes in common. Similar to the lung disease application, we filter out 75% of genes based on mean and standard deviation, retaining 4741 genes. As we will see later, successful training in microarray data and independent validation in RNA-seq data is a strength in this application.

$\text{ogClust}_{\text{GM}}$, $\text{ogClust}_{\text{WJL}}$, sparse K-means, and PMBC are applied for disease subtyping with $K = 2$ in the training cohort and identified disease subtypes are validated in the testing cohort. Similar to the lung disease application, sparse K-means and PMBC fail in feature selection by selecting 4166 and 2273 genes in Figure 3(A) while $\text{ogClust}_{\text{GM}}$ and $\text{ogClust}_{\text{WJL}}$ select 195 and 429 genes. To provide a fair comparison, we select roughly 400 genes, according to the $\text{ogClust}_{\text{WJL}}$ result for all four methods, and leave the result of selecting 300 genes to Supplementary Material Figure S7(II).

Figure 3(B) shows results of outcome association (by log-rank test statistic and adjusted C-index) and gene cluster separation (by R^2_{genes}) from applying the four methods in the training study and validating in the testing cohort. Similar to the simulation and lung disease application results, $\text{ogClust}_{\text{WJL}}$ generates patient clusters with better outcome association (larger log-rank test statistic and larger adjusted C-index) than sparse K-means, PMBC and

(A)

Method \ # Genes	Tune	Fix ~ 400
$ogClust_{GM}$	195	378
$ogClust_{WJL}$	429	433
sparse K-means	4166	336
PMBC	2273	417

(B)

Method \ Criteria	Outcome association				Gene cluster separation	
	log-rank test statistic		Adjusted C-index		R^2_{genes}	
	Train	Test	Train	Test	Trian	Test
$ogClust_{GM}$	324.47	3.96	0.76	0.41	0.01	0.02
$ogClust_{WJL}$	323.17	6.27	0.73	0.54	0.03	0.25
sparse K-means	0.26	4.73	0.12	0.46	0.32	0.35
PMBC	3.93	3.27	0.17	0.49	0.30	0.31

(C)

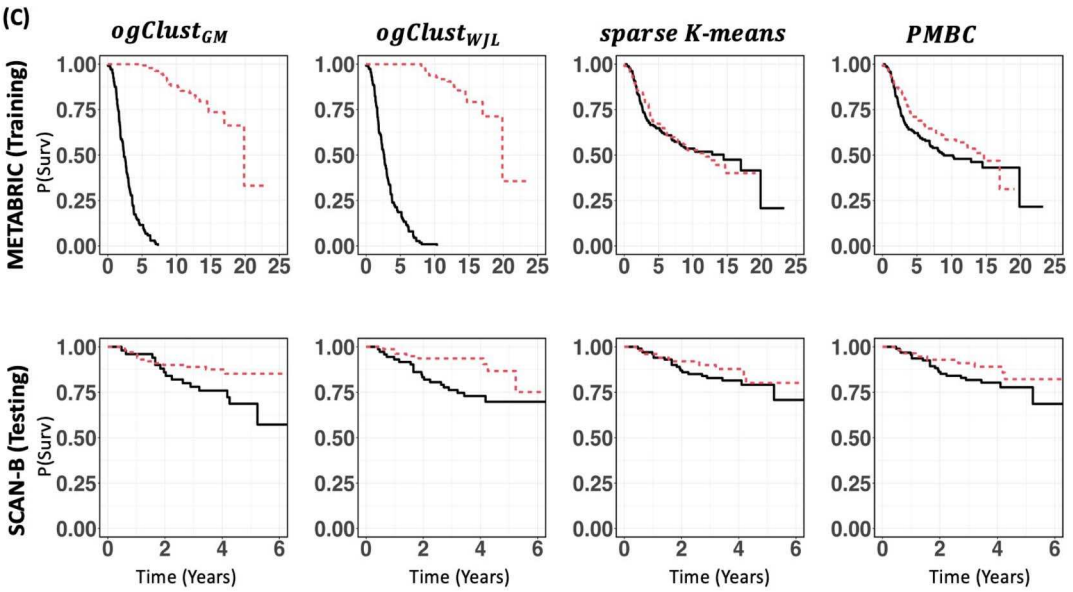

FIG. 3. *Independent validation evaluation of* $\text{ogClust}_{\text{GM}}$, $\text{ogClust}_{\text{WJL}}$, *sparse K-means, and PMBC in the triple-negative breast cancer application. Table* (*A*) *shows the different numbers of genes selected for clustering after parameter tuning of each method as well as the number of genes used in clustering when selecting as close to* 400 *genes as possible for all the methods. With about* 400 *genes selected for each method, Figure* (*B*) *shows* log*-rank test statistic and adjusted C-index, which reflects the outcome association, as well as* R^2_{genes} *for gene cluster separation, by each method in training and testing data. Figure* (*C*) *shows Kaplan–Meier survival curves of identified clusters from each method in the training* (*upper*) *and testing* (*lower*) *cohort, respectively.*

$\text{ogClust}_{\text{GM}}$ in the testing cohort. $\text{ogClust}_{\text{GM}}$ shows overfitting with high performance in training but much worse in testing. Figure 3(C) shows Kaplan–Meier survival curves of identified clusters from each method, and the result further confirms the conclusion above. Results of gene cluster separation by R^2_{genes} are similar to the lung disease application that sparse K-means and PMBC can identify well-separated clusters (large R^2_{genes}) while the clusters are clinically irrelevant with no outcome association. All results are similar in Supplementary Material Figure S7(I) and S7(II), where the number of informative genes is learned from the algorithm or assumed at about 300 genes, respectively.

Next, we conduct pathway enrichment analysis using IPA. Supplementary Material Table S4 shows the number of enriched pathways under different p-values and FDR cutoffs. $\text{ogClust}_{\text{WJL}}$ detects more enriched pathways than the other methods and $\text{ogClust}_{\text{GM}}$ detects almost none (e.g., at FDR=5%, $\text{ogClust}_{\text{WJL}}$ has 13 enriched pathways, sparse K-means and PMBC have six, and $\text{ogClust}_{\text{GM}}$ detects zero). The union of the top 10 pathways from each method is shown in Supplementary Material Table S5, where many immune-related pathways are identified. These include Th1 and Th2 pathways, allograft rejection signaling pathways,

PD1 and PD-L1 pathways, and many T cell and B cell pathways, which are known to be related to breast cancer treatment and prognosis (Oshi et al. (2021), Planes-Laine et al. (2019), Zhao et al. (2019)). Immune response has been recognized as critical prognosis biomarkers in breast cancer and many targeted therapies function as immune checkpoint inhibitors (Wang, Wu and Sun (2022)).

6. Conclusion and discussion. In this paper we propose a novel outcome-guided clustering framework with two proposed methods by generative model $\text{ogClust}_{\text{GM}}$ and weighted joint likelihood $\text{ogClust}_{\text{WJL}}$ for performing disease subtyping by transcriptomic data with guidance from a clinical outcome. In $\text{ogClust}_{\text{GM}}$ the generative model utilizes gene expression to characterize disease subtypes by a latent class variable, which further associates with outcome. The likelihood target function in $\text{ogClust}_{\text{GM}}$, however, is not flexible for balancing contribution from gene clustering and outcome association. In view of this drawback, $\text{ogClust}_{\text{WJL}}$ uses a weighted joint likelihood approach for integrating information from gene features and clinical outcome. The weight w determines relative likelihood contribution of outcome association and gene cluster separation. By tuning the weight, $\text{ogClust}_{\text{WJL}}$ can identify clusters with distinct gene cluster signatures and high outcome association simultaneously so that the omics subtyping model obtained from the training study can generalize to the testing cohort. Compared to the two proposed outcome-guided clustering methods, conventional cluster analysis (e.g., sparse K-means and PMBC) purely based on transcriptomic data often identifies clusters irrelevant to clinical outcomes. In extensive simulations and two real applications, $\text{ogClust}_{\text{WJL}}$ generally has slightly better performance than $\text{ogClust}_{\text{WJL}}$.

For the lung disease application ($n = 331$ patients in the training cohort, $p = 3240$ genes, and $q = 2$ covariates) in Section 5.1, $\text{ogClust}_{\text{GM}}$ requires about 25 minutes for given a fixed λ, using the proposed multinomial logistic regression, and reduces to 30 seconds if changed to sparse linear discriminant analysis. With the additional tuning step of w, $\text{ogClust}_{\text{WJL}}$ requires about six hours in computing. Although computing is relatively heavy, they are affordable in general omics applications. Since the optimization is through EM algorithm with a closed-form solution in the iterative updates, the methods are almost linearly scalable and can be paralleled in computing. For example, searching for parameters using 50 grids of w_1 and 20 grids of λ with 10-fold cross-validation in Section 3.2 can be finished within one hour by 64 computing threads (Intel Xeon Gold with 384GB RAM) and can be further improved with GPU computing. Overall, $\text{ogClust}_{\text{WJL}}$ is preferred in general applications when computing is not an issue. But when computing is a top concern (e.g., in cross-validation or repeated subsampling evaluation), $\text{ogClust}_{\text{GM}}$ is an effective method with fast implementation, especially by using the sparse linear discriminant analysis option.

Our proposed algorithm and evaluations in simulations and real applications assume the number of clusters is given. When we used a larger K in the two applications, our methods have inferior clustering performance (almost empty cluster or clusters with worse validation). Estimating the number of clusters is widely studied in the literature (e.g., by elbow plot or resampling evaluation (Li et al. (2022))) and is not the focus of this paper. While our model only incorporates one cohort and considers one type of omics data (i.e., transcriptomic data) to characterize disease subtypes, it can be extended to incorporate multiple data sources, such as multiomics data (i.e., vertical integration) or multiple transcriptomic studies (i.e., horizontal meta-analysis) to enhance the signal, which will be a future direction. Another future direction is cluster analysis guided by multiple potentially relevant outcomes. In complex diseases the subtypes are often relevant to multiple clinical outcomes. Take lung disease application as an example; forced vital capacity (FVC) measures the amount of air that can be forcibly exhaled after taking the deepest breath possible. In clinical practice, FEV1, FVC, and FEV1/FVC are all critical outcomes, and how to utilize multiple outcomes for clustering

guidance is also a future direction. Finally, current applications summarize RNA-seq data to a continuous expression level. Direct modeling of count data, similar to Li et al. (2023), is expected to improve the performance and is a potential future direction.

We note that, unlike $\text{ogClust}_{\text{GM}}$, $\text{ogClust}_{\text{WJL}}$ is a nonstandard handling of the likelihood function and is not fully model-based under a probabilistic framework. To circumvent overfitting of $\text{ogClust}_{\text{GM}}$, a fully generative model can be developed similar to Bayesian consensus clustering (Lock and Dunson (2013)), where outcome-fitting clustering and gene-fitting clustering adhere loosely to an overall consensus clustering. Although such a full model is mathematically appealing, the required Bayesian computing and inference is expected to be prohibitive for thousands of genes.

An R package "ogClust", together with data and programming code used in the simulations and real applications, are available on GitHub https://github.com/wenjiaking/ogClust.

Acknowledgments. The authors are partially supported by NIH R21LM012752, R01LM014142, and NSF DMS-2113568. The authors would also like to acknowledge many insightful and constructive suggestions from the Associate Editor and two reviewers in the review process.

SUPPLEMENTARY MATERIAL

Web-based supplementary materials for "Outcome-guided disease subtyping by generative model and weighted joint likelihood in transcriptomic applications" (DOI: 10.1214/23-AOAS1865SUPP; .pdf). A zip file that contains a pdf file for supplementary method descriptions, algorithms, figures and tables, as well as the code and data sets to replicate the analysis shown in the paper.

REFERENCES

BAIR, E. and TIBSHIRANI, R. (2004). Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biol*. **2** E108. https://doi.org/10.1371/journal.pbio.0020108

CHANG, W., WAN, C., ZANG, Y., ZHANG, C. and CAO, S. (2020). Supervised clustering of high-dimensional data using regularized mixture modeling. *Brief. Bioinform*. **22** bbaa291.

CHUNG, K. F. (2001). Cytokines in chronic obstructive pulmonary disease. *Eur. Respir. J*. **18** 50s–59s.

COX, D. R. (1972). Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B* **34** 187–220. MR0341758

DEAN, N. and RAFTERY, A. E. (2010). Latent class analysis variable selection. *Ann. Inst. Statist. Math*. **62** 11–35. MR2577437 https://doi.org/10.1007/s10463-009-0258-9

DESANTIS, S. M., HOUSEMAN, E. A., COULL, B. A., NUTT, C. L. and BETENSKY, R. A. (2012). Supervised Bayesian latent class models for high-dimensional data. *Stat. Med*. **31** 1342–1360. MR2925053 https://doi.org/10.1002/sim.4448

DESANTIS, S. M., HOUSEMAN, E. A., COULL, B. A., STEMMER-RACHAMIMOV, A. and BETENSKY, R. A. (2008). A penalized latent class model for ordinal data. *Biostatistics* **9** 249–262. https://doi.org/10.1093/biostatistics/kxm026

FOP, M. and MURPHY, T. B. (2018). Variable selection methods for model-based clustering. *Stat. Surv*. **12** 18–65. MR3794323 https://doi.org/10.1214/18-SS119

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw*. **33** 1.

FURGAL, A. K. C., SEN, A. and TAYLOR, J. M. G. (2019). Review and comparison of computational approaches for joint longitudinal and time-to-event models. *Int. Stat. Rev*. **87** 393–418. MR3994765 https://doi.org/10.1111/insr.12322

GORMLEY, I. C. and FRÜHWIRTH-SCHNATTER, S. (2019). Mixture of experts models. In *Handbook of Mixture Analysis. Chapman & Hall/CRC Handb. Mod. Stat. Methods* 271–307. CRC Press, Boca Raton, FL. MR3889697

GUO, J., WALL, M. and AMEMIYA, Y. (2006). Latent class regression on latent factors. *Biostatistics* **7** 145–163. https://doi.org/10.1093/biostatistics/kxi046

HOUSEMAN, E. A., COULL, B. A. and BETENSKY, R. A. (2006). Feature-specific penalized latent class analysis for genomic data. *Biometrics* **62** 1062–1070. MR2297677 https://doi.org/10.1111/j.1541-0420.2006.00566.x

HUBERT, L. J. and ARABIE, P. (1985). Comparing partitions. *J. Classification* **2** 193–218.

JEMAL, A., SIEGEL, R., WARD, E., HAO, Y., XU, J. and THUN, M. J. (2009). Cancer statistics, 2009. *CA Cancer J. Clin.* **59** 225–249.

LANZA, S. T. and RHOADES, B. L. (2013). Latent class analysis: An alternative perspective on subgroup analysis in prevention and treatment. *Prev. Sci.* **14** 157–168. https://doi.org/10.1007/s11121-011-0201-1

LI, Y., LIU, P., WANG, W., ZONG, W., FANG, Y., REN, Z., TANG, L., CELEDÓN, J. C, OESTERREICH, S. and TSENG, G. C (2024). Supplement to "Outcome-guided disease subtyping by generative model and weighted joint likelihood in transcriptomic applications." https://doi.org/10.1214/23-AOAS1865SUPP

LI, Y., RAHMAN, T., MA, T., TANG, L. and TSENG, G. C. (2023). A sparse negative binomial mixture model for clustering RNA-seq count data. *Biostatistics* **24** 68–84. MR4522704 https://doi.org/10.1093/biostatistics/kxab025

LI, Y., ZENG, X., LIN, C.-W. and TSENG, G. C. (2022). Simultaneous estimation of cluster number and feature sparsity in high-dimensional cluster analysis. *Biometrics* **78** 574–585. MR4450577 https://doi.org/10.1111/biom.13449

LIN, B., BAI, L., WANG, S. and LIN, H. (2021). The association of systemic interleukin 6 and interleukin 10 levels with sarcopenia in elderly patients with chronic obstructive pulmonary disease. *Int. J. Gen. Med.* **14** 5893–5902.

LIN, H., TURNBULL, B. W., MCCULLOCH, C. E. and SLATE, E. H. (2002). Latent class models for joint analysis of longitudinal biomarker and event process data: Application to longitudinal prostate-specific antigen readings and prostate cancer. *J. Amer. Statist. Assoc.* **97** 53–65. MR1947272 https://doi.org/10.1198/016214502753479220

LOCK, E. F. and DUNSON, D. B. (2013). Bayesian consensus clustering. *Bioinformatics* **29** 2610–2616. https://doi.org/10.1093/bioinformatics/btt425

OGAWA, Y., DURU, E. A. and AMEREDES, B. T. (2008). Role of IL-10 in the resolution of airway inflammation. *Curr. Mol. Med.* **8** 437–445. https://doi.org/10.2174/156652408785160907

OSHI, M., LE, L., ANGARITA, F. A., TOKUMARU, Y., YAN, L., MATSUYAMA, R., ENDO, I. and TAKABE, K. (2021). Association of allograft rejection response score with biological cancer aggressiveness and with better survival in triple-negative breast cancer (TNBC).

PAN, W. and SHEN, X. (2007). Penalized model-based clustering with application to variable selection. *J. Mach. Learn. Res.* **8** 1145–1164.

PENCINA, M. J. and D'AGOSTINO, R. B. (2004). Overall C as a measure of discrimination in survival analysis: Model specific population value and confidence interval estimation. *Stat. Med.* **23** 2109–2123.

PEROU, C. M., SØRLIE, T., EISEN, M. B., VAN DE RIJN, M., JEFFREY, S. S., REES, C. A., POLLACK, J. R., ROSS, D. T., JOHNSEN, H. et al. (2000). Molecular portraits of human breast tumours. *Nature* **406** 747.

PLANES-LAINE, G., ROCHIGNEUX, P., BERTUCCI, F., CHRÉTIEN, A.-S., VIENS, P., SABATIER, R. and GONÇALVES, A. (2019). PD-1/PD-L1 targeting in breast cancer: The first clinical evidences are emerging—a literature review. *Cancers* **11** 1033.

PROUST-LIMA, C., SÉNE, M., TAYLOR, J. M. G. and JACQMIN-GADDA, H. (2014). Joint latent class models for longitudinal and time-to-event data: A review. *Stat. Methods Med. Res.* **23** 74–90. MR3190688 https://doi.org/10.1177/0962280212445839

PROUST-LIMA, C. and TAYLOR, J. M. G. (2009). Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment PSA: A joint modeling approach. *Biostatistics* **10** 535–549. https://doi.org/10.1093/biostatistics/kxp009

SCHRÖDER, M. S., CULHANE, A. C., QUACKENBUSH, J. and HAIBE-KAINS, B. (2011). Survcomp: An R/bioconductor package for performance assessment and comparison of survival models. *Bioinformatics* **27** 3206–3208.

SUN, J., HERAZO-MAYA, J. D., MOLYNEAUX, P. L., MAHER, T. M., KAMINSKI, N. and ZHAO, H. (2019). Regularized latent class model for joint analysis of high-dimensional longitudinal biomarkers and a time-to-event outcome. *Biometrics* **75** 69–77. MR3953708 https://doi.org/10.1111/biom.12964

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

TIBSHIRANI, R., WALTHER, G. and HASTIE, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **63** 411–423. MR1841503 https://doi.org/10.1111/1467-9868.00293

WANG, D.-R., WU, X.-L. and SUN, Y.-L. (2022). Therapeutic targets and biomarkers of tumor immunotherapy: Response versus non-response. *Signal Transduct. Targeted Ther.* **7**.

WITTEN, D. M. and TIBSHIRANI, R. (2010). A framework for feature selection in clustering. *J. Amer. Statist. Assoc.* **105** 713–726. MR2724855 https://doi.org/10.1198/jasa.2010.tm09415

WITTEN, D. M. and TIBSHIRANI, R. (2011). Penalized classification using Fisher's linear discriminant. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 753–772. MR2867457 https://doi.org/10.1111/j.1467-9868.2011.00783.x

ZHAO, X., LIU, J., GE, S., CHEN, C., LI, S., WU, X., FENG, X., WANG, Y. and CAI, D. (2019). Saikosaponin A inhibits breast cancer by regulating Th1/Th2 balance. *Frontiers in Pharmacology* **10** 624.

ZHOU, H., PAN, W. and SHEN, X. (2009). Penalized model-based clustering with unconstrained covariance matrices. *Electron. J. Stat.* **3** 1473–1496. MR2578834 https://doi.org/10.1214/09-EJS487

ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 301–320. MR2137327 https://doi.org/10.1111/j.1467-9868.2005.00503.x