Statistical Analysis and Data Mining (2024), **0**, 0, *pp.* 1–**??**
doi:10.1093/Statistical Analysis and Data Mining/output

Bayesian shrinkage models for integration and analysis of multi-platform high-dimensional genomics data

HAO XUE

Department of Computational Biology, Cornell University, 215 Tower Rd, Ithaca, NY 14853, USA

hx222@cornell.edu

SOUNAK CHAKRABORTY*

Department of Statistics, University of Missouri, C209F Middlebush Hall, Columbia, MO, 65211, USA

chakrabortys@missouri.edu

TANUJIT DEY

Center for Surgery & Public Health, Department of Surgery, Brigham and Women's Hospital, Harvard Medical School, 1620 Tremont Street, Suite 2-016, Boston, MA, 02120, USA

tdey@bwh.harvard.edu

Summary

With the increasing availability of biomedical data from multiple platforms of the same patients in clinical research, such as epigenomics, gene expression, and clinical features, there is a growing need for statistical methods that can jointly analyze data from different platforms to provide

*To whom correspondence should be addressed.

complementary information for clinical studies. In this paper, we propose a two-stage hierarchical Bayesian model that integrates high-dimensional biomedical data from diverse platforms to select biomarkers associated with clinical outcomes of interest.

In the first stage, we use Expectation Maximization based approach to learn the regulating mechanism between epigenomics (e.g., gene methylation) and gene expression while considering functional gene annotations. In the second stage, we group genes based on the regulating mechanism learned in the first stage. Then, we apply a group-wise penalty to select genes significantly associated with clinical outcomes while incorporating clinical features. Simulation studies suggest that our model-based data integration method shows lower false positives in selecting predictive variables compared with existing method. Moreover, real data analysis based on a Glioblastoma (GBM) dataset reveals our method's potential to detect genes associated with GBM survival with higher accuracy than the existing method. Moreover, most of the selected biomarkers are crucial in GBM prognosis as confirmed by existing literature.

Key words: Data Integration; Expectation Maximization; Glioblastoma; Hierarchical Bayesian Model; Multiomics.

1. Introduction

Cancer is caused by the accumulation of somatic DNA aberration and epigenetic alteration of transcriptions, protein products, and cell behavior (Sun *et al.*, 2018). As biomedical technologies for performing comprehensive profiling of the cancer genome have advanced, data from different platforms have been brought together on the same patient set (e.g. The Cancer Genome Atlas (TCGA) projects, https://www.cancer.gov) for cancer study. It is of consensus that borrowing information across different genomics and clinical platforms can help us better understand the mechanism of a disease and the corresponding complex biological process since each output from

other platforms could provide a different and complementary view of the whole genome and the overall disease progression (Consortium *et al.*, 2010). Hence integrating multi-platform data has become increasingly crucial in modern genomics studies for its potential in novel target gene identification and cancer clinical management improvement. However, multi-platform data integration is not trivial due to its complex association across different types of molecular features and clinical outcomes. For example, molecular features measured at the transcription level (e.g., gene expression level) affect the clinical outcome more directly than molecular features measured at the DNA/epigenetics level (e.g., methylation) (Wang *et al.*, 2013). Figure 1 represents an example of this type of fundamental biological relationship among outputs of these different platforms and their association with the clinical outcome.

Generally speaking, in a given multiplatform genomic data set, identifying important genes that have a significant association with the clinical outcome through the integration of gene expression level and other output of different types of platforms, such as DNA methylation, is of the greatest interest. In this paper, we focus on building a model that can simultaneously incorporate the multiplatform information (mRNA, methylation level, gene function annotation, clinical features) and identify the significant biomarkers associated with the clinical outcome. The objective is to learn the regulatory mechanism among different platforms (e.g., methylation level and gene expression) and exploit such information to identify disease-specific genes.

Another challenge lies in the high-dimensionality, where large scale of the different types of gene alteration is accompanied with limited number of patient samples. This challenge can be termed as "large p small n problem" (Bernardo *et al.*, 2003). Classical best subset selection procedures are usually computationally expensive and time-consuming since 2^p model probabilities are required for hypothesis testing/model selection. Thus to relieve the computational burden when p is relatively large, penalized likelihood type models can be utilized (Cole *et al.*, 2014).

Substantial consideration is given to these aforementioned challenges from various modeling

perspectives. One such method that received significant attention is the integrative Bayesian analysis of genomics data (iBAG) framework proposed by Wang *et al.* (2013) and generalized to more platforms by Jennings *et al.* (2013). iBAG tackles the data integration and variable selection challenges in two stages. The first stage, a "mechanistic model", infers the effect of methylation on gene expression. Then they divide genes into two groups, the part of expression feature modulated by methylation and the other part modulated by mechanisms other than methylation. In their second stage, a "clinical model" exploits the information from the first stage for the prediction of clinical outcomes. Even though an *ad hoc* gene selection method is provided in the iBAG framework, biomarker selection in iBAG requires choosing a parameter to control the number of selections and iBAG tends to overfit in high-dimensional setting. Therefore, a natural variable selection for detecting biomarkers in high-dimensional setting is an attractive alternative. Other models using similar two-stage modeling approaches have also been proposed recently, such as Jarquin *et al.* (2023) and Zhu *et al.* (2016). For example, Zhu *et al.* (2016) also adopted a two-stage analysis, wherein in the first stage, they developed a collective representation of gene expressions regulated by other omics measurements; in the second stage, they proposed a cancer outcome model including not only the regulated gene expressions but also residuals of gene expressions and residuals of regulators to select relevant markers. However, such model apply uniform penalty to all features, which may not be desirable since epigenetic modification such as dynamic DNA methylation is a critical mechanism of cancer initiation, maintenance, and progression (Lakshminarasimhan and Liang, 2016) and methylated genes and others should be treated differently in penalization. Recently, methods based on latent factor have received significant attention. Wong *et al.* (2018), based on a structural equation modeling framework, formed a structural model that links latent variables across platforms and a measurement model that relates latent variables to observed variables for integrative analysis in genomics studies via the Cox model. In a subsequent extension, Maity *et al.* (2020) used a two-stage model by introducing

latent variables that control mRNA expression and copy number variation measurements. The omics data from multiple platforms are integrated via a Bayesian structural equation coupled with the Bayesian AFT model, which predicts the survival of the patients. Min *et al.* (2018) use Bayesian factor analysis to find latent variables generating multiomics data and then apply k-means to clustering latent variables. Similarly, Li *et al.* (2020) use both latent variables and loading to construct biclustering. Those latent factor models are suitable for summarizing the overall behaviors of a bulk of genes in tasks like clustering or biclustering; however, introducing latent factor makes interpretation of the regulation mechanism in each latent factor complicated and difficult to verify the biological soundness of the outcome. A simpler model is required so that validation of biological significance based on existing studies is possible. Therefore, we will adapt assumption that epigenetic modifications mediate gene expression level, which directly impact phenotype, instead of resort to opaque latent factors.

In this paper, we develop a two-stage hierarchical Bayesian model to resemble the biological process illustrated in Figure 1, which in principle, arises from iBAG (Wang *et al.*, 2013) model with strong improvements in several aspects. We name them the first-stage data integration model (FSDIM) and the second-stage data integration model (SSDIM). In our approach, we incorporate the Expectation-Maximization Variable Selection (EMVS) (Ročková and George, 2014) procedure in our FSDIM and the Bayesian LASSO with modified normal-exponential-gamma (NEG) prior (Ročková and Lesaffre, 2014) in the SSDIM. The former performs efficient Bayesian variable selection in high dimensional regression settings to learn how epigenetics regulate gene expressions. This model is flexible enough to be extended to higher-order polynomials so that potential nonlinear relationships between gene expressions and regulators can be captured and prior knowledge, like gene groupings by gene functional annotations, can be embedded as the random intercept since correlation of expression level is observed in functionally similar genes (Sevilla *et al.*, 2005). The latter can incorporate epigenetic regulating information learned from

the first stage by enforcing group penalties. This group-driven shrinkage not only achieves the sparsity requirement but also quantifies group importance (explained in detail in Section 2.4), which further adds to the biological interpretability of our model.

Our proposed method offers a versatile alternative and improved framework to select by incorporating information across platforms and gene function knowledge, with the following distinct advantages over the iBAG model: Our model (i) adapts a hierarchical Bayesian approach to learn the underlying biological mechanisms among different high-throughput platforms. Our method can simultaneously integrate gene grouping information by gene function annotations along with the DNA methylation. This was not possible in the iBAG approach (Wang *et al.*, 2013). Similar to gene function similarity other type of domain knowledge can also be incorporated in our model with minor modification ; (ii) performs high-dimensional variable selection for discovering relationship across platforms and identifying clinically relevant biomarkers. Our approach which is based on Expectation-Maximization Variable Selection (EMVS) is scalable and can handle far many number of variables than what iBAG can do; (iii) Our method uses closed-form EM algorithms for variable selection, which is more efficient and practical than MCMC based approach; (iv) Our method uses special group-wise penalties based on regulatory mechanism so that genes with different types of epigentic modification are penalized differently. All software codes implemented in R could be found in our GitHub repositories `https://github.com/xvehao/EMMultiOmics`.

The rest of the paper is organized as follows. In Section 2, we illustrate the main architecture of our model. We also provide two extensions of the clinical model to fit binary as well as right censored outcomes. In Section 3, we conduct several simulation studies to evaluate the performance of our proposed models. In Section 4, we apply our model to the glioblastoma (GBM) dataset obtained from TCGA data repository to select GBM biomarkers and estimate the effect of clinical features. Section 5 provides general discussions and conclusions about our

models and a guideline for our future work. Further details about our simulation models, posterior derivations, and extension to binary outcomes are delayed to Supplementary material.

2. Methods

2.1 *Overview of the Data and Model Structures*

Assume, in total, there are N subjects. For a particular subject i, our observed data consist of (i) A clinical data set, Y_i, the continuous clinical outcome of interest (e.g.survival time), (ii) A gene expression data set, where we have $(g_{i1}, ...g_{iK})$, the measures of gene expression level for K genes, (iii) Molecular features measured at DNA/Epigenetics level (e.g. methylation, copy number, and mutation status). Here we illustrate with a DNA methylation data set, where we have $(m_{i1}, ...m_{iJ_0})$, the measures of methylation levels for J_0 probes/sites on the whole genome, (iv) a demographic and hospital record dataset which contains $(c_{i1}, ...c_{iL})$, the values of the L clinical factors (e.g. tumor stage, age, gender, and other demographic variables), (v) a partition of genes based on prior knowledge, which could be illustrated with partition using functional classification. Assume genes are partitioned into R clusters according to functional similarity. Then for a given cluster, say K_r, the size of that cluster is denoted by $|K_r|$. For ease of notation, we assume genes are rearranged such that genes in partition K_r followed by genes in K_{r+1}. If a gene belongs to more than one functional cluster, we assign the gene to an arbitrary one. These observed data sets from multiple platforms (i)-(iv) can be written in matrix notations as: $\mathbf{Y}_{N\times 1}$, $\mathbf{M}_{N\times J_0}$, $\mathbf{G}_{N\times K}$, $\mathbf{C}_{N\times L}$ respectively.

The main objective of our model is to integrate information based on genomic data and multiple types of data from other platforms and also do scalable variable selcetion from each of the data platforms. We developed our hierarchical Bayesian model, which consists of two stages, the First Stage Data Integration Model (FSDIM) and the second Stage Data Integration Model (SSDIM). FSDIM aims to integrate the direct effects of DNA methylation on gene expression with the gene

function considered a random effect because it was reported by abundant previous studies that expression correlation should be related to annotation similarity (Sevilla *et al.*, 2005). The gene with functional similarities is assumed to have same intercepts in gene expression levels. Specifically, FSDIM is realized by fitting a linear model for each gene with a correlation structure (between the expression of genes) based on the functional classification obtained from Database for Annotation, Visualization, and Integrated Discovery (DAVID, https://david.ncifcrf.gov) (Sherman *et al.*, 2022; Huang *et al.*, 2009). The expression level of each gene is regressed on the transformation of methylation level at J_0 sites, which is obtained via a customized mapping function f (which can capture any linear or non-linear functional relationship between gene expression and DNA methylation). Thereafter, we divide the K genes into three groups: (1) genes that are modulated by only DNA methylation (M type effect); (2) genes that are modulated by aspects other than DNA methylation ($\bar{M}$ type effect); (3) genes that are modulated jointly by DNA methylation and aspects other than DNA methylation. Genes will be assigned group membership depending on the coefficient of determination, R_k^2 of FSDIM. Methylation was believed to play a crucial role in repressing gene expression, perhaps by blocking the promoters at which activating transcription factors should bind. Therefore, learning exactly which genes are regulated by methylation is important. SSDIM will fit a Bayesian Lasso by embedding the grouping information learned from the first stage with NEG-modified priors so that genes are penalized differently according to their methylation-expression regulating mechanism for biomarker selection. In this way, this group penalties could capture the mediation effect of methylation through gene expression to disease. For instance, it's possible that tumor suppressor genes were silenced due to methylation. Growing evidence suggests that the methylation of promoter regions in various genes, including tumor suppressor genes, leads to the subsequent impairment of functional protein expression. Therefore, methylation regulated genes should be given a different penalty when studying their association with the disease (Luo *et al.*, 2018). Figure 2 provides an overview of our two-stage

hierarchical model, and a detailed introduction and explanation of each component are presented in the next subsections.

In the first stage, we will combine information from methylation, gene expression, and gene function annotation data to group genes based on how their expressions are regulated by DNA methylation. In this context, the sample size of sequencing data is typically far smaller than the number of genes sequenced due to the sequencing expense, referred to as the "large P small N" problem. To mitigate this challenge, we implement the EMVS (Ročková and George, 2014) algorithm with a random intercept to get a sparse first-stage model with expression levels as response and methylation levels as predictors. EMVS is a deterministic alternative to stochastic search based on the EM algorithm with "spike-and-slab" prior, which can quickly find posterior modes. Another advantage of EMVS is its flexibility, due to which we can further incorporate the gene function classification in the random intercept of the model. To do this, DAVID is used to partition genes based on their functional annotations with a fuzzy heuristic multiple-linkage partition to agglomerate genes based on gene-gene similarity measurements. This prior information is subsequently incorporated in our model as a random intercept, where genes are assumed to be partitioned into R number clusters according to gene functional similarity. We assume for each cluster, and the intercept is different.

In total, there are $K = \sum_{r=1}^{R} |K_r|$ number of genes, which leads to K regression fittings with gene expressions as response and the measures of transformation of methylation levels for J_0 probes as covariates. Therefore, for a specific gene k in partition K_r, our model is:

$$\mathbf{g}_k = \tilde{\mathbf{M}}\boldsymbol{\omega}_k + \xi_r \mathbf{1} + \mathbf{g}_k^{\bar{\mathbf{M}}}, \mathbf{g}_k^{\bar{\mathbf{M}}} \sim N(\mathbf{0}, \sigma^2 I_N), \tag{2.1}$$

where $\tilde{\mathbf{M}}$ is the image of methylation level $\mathbf{M}$ mapped by a pre-specified function $f : \mathbb{R}^{N \times J_0} \to \mathbb{R}^{N \times J}$, J is the dimension of image dependent on the choice of f and $\boldsymbol{\xi} = (\xi_1, ..., \xi_R)$ are random intercept added to each functional group, and $\mathbf{g}_k^{\bar{\mathbf{M}}}$ can be interpreted as the portion of the expression level of k-th gene in r-th functional cluster which is modulated by factors other than

methylation. The function f captures the functional relationship between gene expression and DNA methylation, and the choice of f could be customized. For instance, we can use the identity function to capture the linear relationship between gene expression and methylation, or we can also use non-linear mappings such as quadratic polynomials, in which the i-th row of $\tilde{\mathbf{M}}_{N\times 2J_0}$ is $(m_{i1}, m_{i1}^2, ...m_{iJ_0}, m_{iJ_0}^2)$, or natural spline ($ns$ function in R by Perperoglou *et al.* (2019)) to capture the non-linear relationships between gene expression and methylation factors. The FSDIM models using identity function, quadratic polynomial, and cubic splines are respectively named as linear, quadratic, and cubic FSDIM models.

For parameter $\boldsymbol{\xi}$, we assign the priors as:

$$\pi(\boldsymbol{\xi}|\boldsymbol{\tau}) = N(\mathbf{0}, D), \text{ where } D = \text{diag}(\tau_1^2, ..., \tau_R^2), \tag{2.2}$$

$$\pi(\tau_r^2) = IG(a_0, b_0), \tag{2.3}$$

where IG is inverse gamma. To implement the EMVS method, we assign priors for each of the parameters:

$$\boldsymbol{\omega}_k|\sigma^2, \boldsymbol{\gamma}_k, \nu_0, \nu_1 \sim N(\mathbf{0}, \mathbf{D}_{\sigma,\gamma_k}), \tag{2.4}$$

where $\boldsymbol{\gamma}_k = (\gamma_{1k}, ...\gamma_{Jk})^T$ is a vector of binary latent variables, specifying the inclusion or exclusion of variables, $\gamma_{jk} \in \{0, 1\}$ and $\mathbf{D}_{\sigma,\gamma_k} = \sigma^2\text{diag}\{(1-\gamma_{jk})\nu_0 + \gamma_{jk}\nu_1\}$, $j = 1, ...J$. This spike-and-slab prior formulation is the core ingredient for yielding a sparse (with respect to the number of probe sites from the methylation data) first-stage model. For a predictive variable (which is methylation level of a certain site), $\gamma_{jk} = 1$, indicating the inclusion of the variable, the corresponding posterior of variance is large, which makes the posterior flat enough to accommodate all possible values. Whereas, for a non-predictive variable, $\gamma_{jk} = 0$, indicates exclusion of the variable, and the prior variance is designed to be small, creating a shrinkage effect towards zero. The prior for $\boldsymbol{\gamma}_k$ is the i.i.d. Bernoulli prior form:

$$\pi(\boldsymbol{\gamma}_k|\theta_k) = \theta_k^{|\boldsymbol{\gamma}_k|}(1-\theta_k)^{J-|\boldsymbol{\gamma}_k|},$$

where $|\boldsymbol{\gamma}_k| = \sum_j \gamma_{jk}$ and $\boldsymbol{\gamma}_k = (\gamma_{1k}, ...\gamma_{Jk})$. We also subsequently assign priors for σ^2 and θ respectively as:

$$\sigma^2 \sim IG(a_1, b_1) \quad \text{and} \quad \theta \sim Beta(c, d), \tag{2.5}$$

The choice of prior of σ^2 is to ensure it is relatively non-influential. The beta-binomial priors $\pi(\gamma)$ resulting from the beta prior of θ can favor parsimony. We are using EMVS method to group of genes based on how they are regulated by methylation. The information from methylation, gene expression, and gene function annotation is combined in a single linear model, where methylation data are treated as the predictor, gene expression as the response, and gene functions as random effects. It is assumed that the expression levels of some genes are sparsely regulated by methylation. We use the R^2 of the fitted model to group genes into those regulated by methylation, not regulated by methylation, and regulated by both methylation and other unknown factors. The purpose of using R^2 does not lie in accessing prediction accuracy; instead, we use the resulting R^2 as a quantification of how much variability of a certain gene expression level can be explained by the variability of methylation. If the value of R^2 is high, then there is strong evidence that the expression level of the gene of interest is mediated through the methylation-expression relation learned by EMVS. Our method can use existing biological information and infer novel dynamic relations. The rationale for grouping genes is based on the hypothesis that genes regulated differently by methylation are likely to have different extend of influence on clinical outcomes. For instance, tumor suppressor genes are often silenced in cancer cells due to hypermethylation.

2.2 *EM Based First Stage Data Integration Model Fitting*

We employ the EMVS (Ročková and George, 2014) algorithm, which iteratively optimizes an objective function Q in (2.6) as below:

$$\begin{aligned} &Q(\boldsymbol{\omega}, \boldsymbol{\theta}, \boldsymbol{\sigma}, \boldsymbol{\xi}, \boldsymbol{\tau} | \boldsymbol{\omega}^{(t)}, \boldsymbol{\theta}^{(t)}, \boldsymbol{\sigma}^{(t)}, \boldsymbol{\xi}^{(t)}, \boldsymbol{\tau}^{(t)}) = \mathbb{E}_{\boldsymbol{\gamma}|\cdot}[\log(\pi(\boldsymbol{\omega}, \boldsymbol{\theta}, \boldsymbol{\sigma}, \boldsymbol{\gamma}, \boldsymbol{\xi} | \mathbf{g}))] \\ &\propto Q_1(\boldsymbol{\omega}, \boldsymbol{\sigma}, \boldsymbol{\xi}, \boldsymbol{\tau} | \boldsymbol{\omega}^{(t)}, \boldsymbol{\theta}^{(t)}, \boldsymbol{\sigma}^{(t)}, \boldsymbol{\xi}^{(t)}, \boldsymbol{\tau}^{(t)}) + Q_2(\boldsymbol{\theta} | \boldsymbol{\omega}^{(t)}, \boldsymbol{\theta}^{(t)}, \boldsymbol{\sigma}^{(t)}, \boldsymbol{\xi}^{(t)}, \boldsymbol{\tau}^{(t)}) \end{aligned} \tag{2.6}$$

where

$$\begin{aligned} &Q_1(\boldsymbol{\omega},\boldsymbol{\sigma},\boldsymbol{\xi},\boldsymbol{\tau}|\boldsymbol{\omega}^{(t)},\boldsymbol{\theta}^{(t)},\boldsymbol{\sigma}^{(t)},\boldsymbol{\xi}^{(t)},\boldsymbol{\tau}^{(t)}) \\ &= \sum_{r=1}^{R}\sum_{k\in K_r}\Big\{-\frac{1}{2\sigma_k^2}(\mathbf{g}_k-\tilde{\mathbf{M}}\boldsymbol{\omega}-\xi_r\mathbf{1})^T(\mathbf{g}_k-\tilde{\mathbf{M}}\boldsymbol{\omega}-\xi_r\mathbf{1})-\frac{N+J+\nu+2}{2}\log(\sigma_k^2) \\ &-\frac{1}{2\sigma_k^2}\sum_{j=1}^{J}\omega_{jk}^2\mathbb{E}_{\boldsymbol{\gamma}|\cdot}[\frac{1}{(1-\gamma_{jk})\nu_0+\gamma_{jk}\nu_1}]-\frac{\nu\lambda}{2\sigma_k^2}\Big\}+\sum_{r=1}^{R}-2\log(\tau_r^2)-\frac{\xi_r^2}{2\tau_r^2}-\frac{1}{2\tau_r^2} \end{aligned} \tag{2.7}$$

and

$$\begin{aligned} &Q_2(\boldsymbol{\theta}|\boldsymbol{\omega}^{(t)},\boldsymbol{\theta}^{(t)},\boldsymbol{\sigma}^{(t)},\boldsymbol{\xi}^{(t)},\boldsymbol{\tau}^{(t)}) \\ &= \sum_{r=1}^{R}\sum_{k\in K_r}\Big\{\sum_{j=1}^{J}\mathbb{E}_{\boldsymbol{\gamma}|\cdot}(\gamma_{jk})\log(\frac{\theta_k}{1-\theta_k})+(a-1)\log(\theta_k)+(J+b-1)\log(1-\theta_k)\Big\}. \end{aligned} \tag{2.8}$$

Note that the decomposition of Q via two additive components Q_1 and Q_2 is possible since the posterior distribution of $\boldsymbol{\gamma}_k$ given $(\boldsymbol{\omega}^{(t)},\boldsymbol{\theta}^{(t)},\boldsymbol{\sigma}^{(t)},\mathbf{g}_k)$ depends on $\mathbf{g}_k$ only through $(\boldsymbol{\omega}^{(t)},\boldsymbol{\theta}^{(t)},\boldsymbol{\sigma}^{(t)})$. This separation of the objective function Q yields two distinct terms, Q_1 and Q_2, which can be maximized in the M-step independently much faster and easier.

For **E-step**, we first consider the expectation involved in Q_2:

$$\mathbb{E}_{\boldsymbol{\gamma}_k|\cdot}\gamma_{jk}=P(\gamma_{jk}=1|\boldsymbol{\omega}_k^{(t)},\theta_k^{(t)},\boldsymbol{\sigma}_k^{(t)})=p_{jk},$$

where

$$p_{jk}=\frac{\pi(\omega_{jk}^{(t)}|\sigma_k^{(t)},\gamma_{jk}=1)P(\gamma_{jk}=1|\theta_k^{(t)})}{\pi(\omega_{jk}^{(t)}|\sigma_k^{(t)},\gamma_{jk}=1)P(\gamma_{jk}=1|\theta_k^{(t)})+\pi(\omega_{jk}^{(t)}|\sigma_k^{(t)},\gamma_{jk}=0)P(\gamma_{jk}=0|\theta_k^{(t)})}. \tag{2.9}$$

This is equivalent to the posterior update of mixing proportion for fitting a two-component Gaussian mixture to $\boldsymbol{\omega}_k^{(t)}$ with the conventional EM algorithm. Based on this, the other conditional expectation can be computed as:

$$\mathbb{E}_{\boldsymbol{\gamma}_k|\cdot}[\frac{1}{\nu_0(1-\gamma_{jk})+\nu_1\gamma_{jk}}]=\frac{\mathbb{E}_{\boldsymbol{\gamma}_k|\cdot}(1-\gamma_{jk})}{\nu_0}+\frac{\mathbb{E}_{\boldsymbol{\gamma}_k|\cdot}\gamma_{jk}}{\nu_1}=\frac{1-p_{jk}}{\nu_0}+\frac{p_{jk}}{\nu_1}\equiv d_{jk} \tag{2.10}$$

In the **M-step**, the $\boldsymbol{\omega}_k^{(t+1)}$ can be obtained by using the Sherman-Morrison-Woodbury formula as:

$$\boldsymbol{\omega}_k^{(t+1)}=(\tilde{\mathbf{M}}^T\tilde{\mathbf{M}}+\mathbf{D}_k)^{-1}\tilde{\mathbf{M}}^T(\mathbf{g}_k-\xi_r\mathbf{1}) \tag{2.11}$$

$$=[\mathbf{D}_k^{-1}-\mathbf{D}_k^{-1}\tilde{\mathbf{M}}^T(\mathbf{I}_N+\tilde{\mathbf{M}}\mathbf{D}_k^{-1}\tilde{\mathbf{M}}^T)^{-1}\tilde{\mathbf{M}}\mathbf{D}_k^{-1}]\tilde{\mathbf{M}}^T(\mathbf{g}_k-\xi_r\mathbf{1}), \tag{2.12}$$

where $\mathbf{D}_k = \text{diag}\,\{d_{jk}\}_{j=1}^{J}$. Note that this is equivalent to a ridge regression with penalty specified by $\mathbf{D}_k$.

To maximize $Q_1(\boldsymbol{\omega}, \boldsymbol{\sigma}, \boldsymbol{\xi}, \boldsymbol{\tau} | \boldsymbol{\omega}^{(t)}, \boldsymbol{\theta}^{(t)}, \boldsymbol{\sigma}^{(t)}, \boldsymbol{\xi}^{(t)}, \boldsymbol{\tau}^{(t)})$ with respect to σ_k, we use the simple update:

$$\sigma_k^{(t+1)} = \sqrt{\frac{||\mathbf{g}_k - \tilde{\mathbf{M}}\boldsymbol{\omega}_k^{(t+1)} - \xi_r \mathbf{1}||^2 + ||\mathbf{D}^{1/2}\boldsymbol{\omega}_k^{(t+1)}||^2 + 1}{N + J + 1}},$$

The maximization of Q_1 with respect to τ_r and ξ_r are respectively:

$$\tau_r = \sqrt{\frac{\xi_r^2 + 1}{4}} \tag{2.13}$$

and

$$\begin{aligned} \xi_r^{(t+1)} &= \arg\min_{\xi_r} \sum_{k \in K_r} \frac{1}{2\sigma_k^2} (\mathbf{g}_k - \tilde{\mathbf{M}}\boldsymbol{\omega} - \xi_r \mathbf{1})^T (\mathbf{g}_k - \tilde{\mathbf{M}}\boldsymbol{\omega} - \xi_r \mathbf{1}) + \sum_{r=1}^{R} \frac{\xi_r^2}{2\tau_r^2} \\ &= \frac{\sum_{K_r} \frac{\tau_r^2}{\sigma_k^2} (\mathbf{g}_k - \tilde{\mathbf{M}}\boldsymbol{\omega}_k)^T \mathbf{1}}{1 + \sum_{K_r} \frac{\tau_r^2}{\sigma_k^2} \mathbf{1}^T \mathbf{1}}. \end{aligned} \tag{2.14}$$

For maximization of Q_2 with respect to $\boldsymbol{\theta}_k$, we get $\theta_k^{(t+1)} = \frac{\sum_{j=1}^{J} p_{jk} + a - 1}{a + b + J - 2}$. We need to threshold the EM output for variable selection in each gene-specific model (2.1). It is reasonable that those variables with the large posterior of inclusion should be retained in the model, where $\hat{\boldsymbol{\gamma}}_k = \arg\max_{\boldsymbol{\gamma}_k} P(\boldsymbol{\gamma}_k | \hat{\boldsymbol{\omega}}_k, \hat{\theta}_k, \hat{\sigma}_k)$ and the corresponding posterior can be computed as $P(\boldsymbol{\gamma}_k | \hat{\boldsymbol{\omega}}_k, \hat{\theta}_k, \hat{\sigma}_k) = \prod_{j=1}^{J} P(\gamma_{jk} | \hat{\omega}_{jk}, \hat{\theta}_k, \hat{\sigma}_k)$, where

$$P(\gamma_{jk} | \hat{\omega}_{jk}, \hat{\theta}_k, \hat{\sigma}_k) = \frac{\pi(\hat{\omega}_{jk} | \hat{\sigma}_k, \gamma_{jk}) P(\gamma_{jk} | \hat{\boldsymbol{\theta}})}{\pi(\hat{\omega}_{jk} | \hat{\sigma}_k, \gamma_{jk} = 1) P(\gamma_{jk} = 1 | \hat{\boldsymbol{\theta}}) + \pi(\hat{\omega}_{jk} | \hat{\sigma}_k, \gamma_{jk} = 0) P(\gamma_{jk} = 0 | \hat{\theta}_k)}. \tag{2.15}$$

Then the inclusion of coefficient ω_{kj} in the model could be specified by a thresholding scheme as, $\hat{\gamma}_{jk} = 1 \iff P(\gamma_{jk} = 1 | \hat{\boldsymbol{\omega}}_k, \hat{\theta}_k, \hat{\sigma}_k) \geqslant 0.5$.

2.3 *Knowledge Discovery from FSDIM*

Here we develop a strategy to group genes based on how much they are affected by the DNA methylation levels. In that respect, here we introduce a loading matrix $\mathbf{Z}(K \times Q)$, consisting

of Q columns of dummy variables coding for group membership (for a total number of $Q-1$ groups). In our study, we group our K genes into three non-overlapping groups. They are a) genes regulated by methylation effect (M-group), b) genes regulated by aspects other than methylation ($\bar{M}$-group), and genes that are regulated by both methylation effect and non-methylation factors ($M+\bar{M}$-group). Here we also add an intercept which can be treated as a "global group" or behave like a global shrinkage on all genes. Therefore in our case, $Q=4$. For example, if a gene is in the M-group, the corresponding dummy variable is coded as $(1,1,0,0)^T$; if it is in the $\bar{M}$-group the corresponding dummy variable is coded as $(1,0,1,0)^T$, and finally if a gene is in the $M+\bar{M}$-group the corresponding dummy variable is coded as $(1,0,0,1)^T$. A toy example with 10 genes and their group membership is given in Figure 3.

To construct our loading matrix $\mathbf{Z}$, we calculate the coefficients of determination R_k^2, $k=1,...K$ for each fitted regression model in Section 2.2. The larger value of R_k^2 indicates the stronger relationship of the genes with methylation factors. Here genes with R_k^2 smaller than 0.2 are put into $\bar{M}$ effect group, larger than 0.8 are put into M effect group and the rest are put into $M+\bar{M}$ effect group. To be noted that the R^2 based cut-off is subjective to user choice.

2.4 *The Second Stage Data Integrating Model (SSDIM)*

SSDIM is a linear model that can be used to detect biomarkers associated with the clinical outcome of interest while considering the knowledge learned from FSDIM. It regresses clinical outcome of interest on relevant clinical features and gene expression, regularized with a group penalty reflecting methylation regulating mechanism and gene functional similarity learned from FSDIM. In this way, the SSDIM further integrates clinical outcomes to develop a unified way of combining and curating the set of predictor variables with the clinical responses. The SSDIM can be formulated as:

$$\mathbf{Y}=\mathbf{C}\boldsymbol{\beta}^{\mathbf{C}}+\mathbf{G}\boldsymbol{\beta}^{\mathbf{G}}+\boldsymbol{\epsilon},\quad \boldsymbol{\epsilon}\sim N(\mathbf{0},\sigma^2\mathbf{I}_{N\times N}), \tag{2.16}$$

where $\mathbf{Y}$ is the clinical outcome of interest, $\mathbf{G}$ is the gene expression level, $\mathbf{C}$ is other clinical features (see definitions in Section 2.1), $\boldsymbol{\beta}^{\mathbf{C}} = (\beta_1^{\mathbf{C}}, ...\beta_L^{\mathbf{C}})^T$ and $\boldsymbol{\beta}^{\mathbf{G}} = (\beta_1^{\mathbf{G}}, ...\beta_K^{\mathbf{G}})^T$ are regression coefficients, and $\boldsymbol{\epsilon} = (\epsilon_1, ...\epsilon_N)^T$ is the random noise (with slight notation abuse in parameters, since parameters in two stages are fitted separately, some notations appearing in FSDIM are recycled in SSDIM for brevity).

In our second-stage model (2.16), the high-dimensionality problem still arises due to the large value of K (the number of genes in our gene expression data). Here we adopt a Bayesian LASSO with Normal-Exponential-Gamma (NEG) modified prior structure (Ročková and Lesaffre, 2014; Griffin and Brown, 2011) to better handle the problem of high dimensionality. Additionally, we embed the grouping information obtained from the FSDIM (Section 2.3) as an additional regression layer in the hierarchy to penalize each group of genes differently. Then we implement the EM algorithm by using the Laplace Representation suggested by Ročková and Lesaffre (2014), which yields a LASSO model in the M-step and thus shrinks non-informative predictors into zero.

In our SSDIM, the NEG modified priors for our model (2.16) parameters are given below,

$$\beta_k^{\mathbf{G}}|\sigma^2, \tau_k \sim N(0, g\sigma^2\tau_k^2) \quad \text{and} \quad \boldsymbol{\beta}^{\mathbf{C}}|\sigma^2 \sim N(\mathbf{0}, \sigma^2 I_L). \tag{2.17}$$

In our multivariate linear regression (2.16), a large fixed values of σ^2 may increase the number of false positives, which could be problematic in high-dimensional settings. Therefore, here we add a fixed scaling factor $g = 1/N^2$ to the prior variance as $Var(\beta_j|\sigma^2, \tau_j) = g\tau_i^2\sigma^2$. The rest of hyperprior distributions are as follow,

$$\begin{aligned} \tau_k^2|\lambda_k^2 &\overset{iid}{\sim} \lambda_k^2 \exp(-\lambda_k^2\tau_k^2) I(\tau_k > 0) \\ \lambda_k^2|\mathbf{b} &\overset{iid}{\sim} \Gamma(a, h(\mathbf{Z}_k^T\mathbf{b})), \quad k = 1, ..., K \\ b_l &\overset{iid}{\sim} \pi(\boldsymbol{\theta}), \quad l = 0, 1, 2, 3, \\ \sigma^2 &\sim IG(c, d) \end{aligned} \tag{2.18}$$

The introduction of $\mathbf{Z}_k$ vector (obtained from the FSDIM in Section 2.3) in the prior distribution of λ_k^2 creates group based penalty structure. In general we have Q (here our $Q = 4$)

non-overlapping groups of genes, where $\mathbb{Q}_l$ is the set of genes in the l-th group, $l = 1, \ldots, Q$. Therefore, if the k-th gene $\in \mathbb{Q}_l$, then $E\lambda_k^2 = ah(b_0 + b_l)$. Here, $h(.)$ is the inverse link function so that b_l can be treated as the weight of the l-th group, and a larger value of b_l represents higher importance attached to the l-th group. This formulation encourages the model to keep groups with "large weights" in the model and then select the significant variables within those "important" groups. Under the usual regularity conditions (A)-(C) by Fan and Li (2001), the coefficient estimator follows variable selection consistency and asymptotic normality (Ročková and Lesaffre, 2014). In addition to that as all our assigned priors are proper an represents a scale mixture representation of the double exponential distribution which in turn guarantees the propriety and unimodality of the posterior distribution (Park and Casella, 2008). Unimodality is a very desirable criteria as it makes the EM algorithm based variable selection faster and more efficient.

As MCMC is painstakingly slow to implement when number of variables both clinical and genetic is of high dimensional we design a EM based approach for the variable selection in this Second Stage as explained in the next section.

2.5 *EM Based Second Stage Data Integration Model Fitting*

Here we implement an EM algorithm following a similar idea as in the original EMVS model given by Ročková and George (2014). We compute the objective function by taking the expectation of complete log-likelihood given the observed data and current estimation of $\boldsymbol{\beta}^{(t)}, \mathbf{b}^{(t)}, \boldsymbol{\sigma}^{(t)}$ at the t-th iteration (detailed derivation in Section 1, Supplementary material):

$$\begin{aligned} &Q(\boldsymbol{\beta}^{\mathbf{G}}, \boldsymbol{\beta}^{\mathbf{C}}, \mathbf{b}, \sigma | (\boldsymbol{\beta}^{\mathbf{G}})^{(t)}, (\boldsymbol{\beta}^{\mathbf{C}})^{(t)}, \mathbf{b}^{(t)}, \sigma^{(t)}) = \mathbb{E}_{\boldsymbol{\lambda}^2|.}[\log p(\boldsymbol{\beta}^{\mathbf{G}}, \boldsymbol{\beta}^{\mathbf{C}}, \boldsymbol{b}, \sigma | y)] \\ &= C + Q_1(\boldsymbol{\beta}^{\mathbf{G}}, \boldsymbol{\beta}^{\mathbf{C}}, \sigma | (\boldsymbol{\beta}^{\mathbf{G}})^{(t)}, (\boldsymbol{\beta}^{\mathbf{C}})^{(t)}, b^{(t)}, \sigma^{(t)}) + Q_2(\boldsymbol{b} | (\boldsymbol{\beta}^{\mathbf{G}})^{(t)}, (\boldsymbol{\beta}^{\mathbf{C}})^{(t)}, \mathbf{b}^{(t)}, \sigma^{(t)}), \end{aligned} \tag{2.19}$$

where

$$\begin{aligned}
&Q_1(\boldsymbol{\beta}^{\mathbf{G}}, \boldsymbol{\beta}^{\mathbf{C}}, \sigma | (\boldsymbol{\beta}^{\mathbf{G}})^{(t)}, (\boldsymbol{\beta})^{\mathbf{C}(t)}, b^{(t)}, \sigma^{(t)}) \\
&= -\frac{(\mathbf{Y} - \mathbf{G}\boldsymbol{\beta}^{\mathbf{G}} - \mathbf{C}\boldsymbol{\beta}^{\mathbf{C}})^T(\mathbf{Y} - \mathbf{G}\boldsymbol{\beta}^{\mathbf{G}} - \mathbf{C}\boldsymbol{\beta}^{\mathbf{C}})}{2g\sigma^2} \\
&- \frac{\sqrt{2}}{\sqrt{g}\sigma}\sum_{k=1}^{K}|\beta_k^{\mathbf{G}}|E_{\boldsymbol{\lambda}^2|\cdot}\lambda_k - \frac{N+K+2c+2}{2}\log(\sigma^2) - \frac{d}{g\sigma^2} \\
&- \frac{2e+2+L}{2}\log(\sigma_C^2) - \frac{f}{\sigma^2} - \frac{1}{2}\sum_{l=1}^{L}\frac{(\beta_l^{\mathbf{C}})^2}{\sigma^2}
\end{aligned} \tag{2.20}$$

and

$$\begin{aligned}
&Q_2(\boldsymbol{b} | (\boldsymbol{\beta}^{\mathbf{G}})^{(t)}, (\boldsymbol{\beta}^{\mathbf{C}})^{(t)}, b^{(t)}, \sigma^{(t)}) \\
&= \sum_{k=1}^{K}\{-a\log(h(Z_k'\boldsymbol{b}) - \frac{\mathbb{E}_{\boldsymbol{\lambda}^2|\cdot}\lambda_k^2}{h(Z_k'\boldsymbol{b})})\} + \sum_{l=0}^{q}[(\alpha-1)\log b_l - \gamma b_l],
\end{aligned} \tag{2.21}$$

where $E_{\boldsymbol{\lambda}^2|\cdot}(\cdot)$ denotes the conditional expectation $E_{\boldsymbol{\lambda}^2|\cdot}(\cdot|\boldsymbol{\beta}^{(t)}, \mathbf{b}^{(t)}, \boldsymbol{\sigma}^{(t)}, \mathbf{y})$. Analogously as in FSDIM, we can split the objective function of two distinct terms (Q_1 and Q_2) and optimize them separately. One difference from the FSDIM worth noting is that, from the formulation of Q_1, we can recognize that solving $\boldsymbol{\beta}$ is essentially solving an adaptive lasso (Zou, 2006).

For **E-step**, we need to calculate the $E_{\boldsymbol{\lambda}^2|\cdot}\lambda_j$ and $E_{\boldsymbol{\lambda}^2|\cdot}\lambda_j^2$. Here we use the general formula:

$$\mathbb{E}_{\boldsymbol{\lambda}^2|\cdot}\lambda_k^s = \frac{\Gamma(2a+s+1)[h(\mathbf{Z}_k^T\mathbf{b}^{(t)})]^{(s+1)/2}}{\sigma^{(t)}\sqrt{g}\Gamma(a)2^{a+s/2}p_{a,s,\sigma}(\beta_k^{\mathbf{G}})}\exp(\frac{\beta_k^{\mathbf{G}(t)2}h(\mathbf{Z}_k^T\mathbf{b}^{(t)})}{4\sigma^{(t)2}g})D_{-(2a+1+s)}(\frac{|\beta_k^{\mathbf{G}}|\sqrt{h(\mathbf{Z}_k^T\mathbf{b}^{(t)})}}{\sigma^{(t)}\sqrt{g}}), \tag{2.22}$$

where $D_\eta(x)$ is parabolic cylinder function and $p_{a,s,\sigma}(\beta_j)$ is the density of NEG distribution.

Next, we propose our **Scheme EM** to estimate $\boldsymbol{\beta}^{\mathbf{C}}$. Here we update $\boldsymbol{\beta}^{\mathbf{C}}$, $\boldsymbol{\beta}^{\mathbf{G}}$ and σ alternatively by maximization in M-step:

$$\begin{aligned}
(\boldsymbol{\beta}^{\mathbf{G}})^{(t+1)} &= \arg\min_{\boldsymbol{\beta}^{\mathbf{G}}}\left\{\|\mathbf{Y} - \mathbf{G}\boldsymbol{\beta}^{\mathbf{G}} - \mathbf{C}(\boldsymbol{\beta}^{\mathbf{C}})^{(t)}\|_2 + \frac{2\sqrt{2}\sigma^{(t)}}{\sqrt{g}}\|D^{(t)}\boldsymbol{\beta}^{\mathbf{G}}\|_1\right\} \\
(\boldsymbol{\beta}^{\mathbf{C}})^{(t+1)} &= \arg\min_{\boldsymbol{\beta}^{\mathbf{C}}}\left\{\|\mathbf{Y} - \mathbf{G}(\boldsymbol{\beta}^{\mathbf{G}})^{(t+1)} - \mathbf{C}\boldsymbol{\beta}^{\mathbf{C}}\|_2 + \|\boldsymbol{\beta}^{\mathbf{C}}\|_2^2\right\}
\end{aligned}$$

where $\mathbf{D}^{(t)} = diag[E_{\boldsymbol{\lambda}^2|\cdot}\lambda_1, ..., E_{\boldsymbol{\lambda}^2|\cdot}\lambda_K]$. The solution can be obtained by applying the standard LASSO computation after reweighting the regression matrix. If we are further interested in the posterior distribution of coefficients of clinical features, we can use Gibbs sampler to sample $\boldsymbol{\beta}^{\mathbf{C}}$,

we call it **Scheme MCMC** with

$$\pi[\boldsymbol{\beta}^{\mathbf{C}}|\cdot] \propto \exp\{-\frac{1}{2\sigma^2}[(\tilde{\mathbf{Y}}-\mathbf{C}\boldsymbol{\beta}^{\mathbf{C}})^T(\tilde{\mathbf{Y}}-\mathbf{C}\boldsymbol{\beta}^{\mathbf{C}})+(\boldsymbol{\beta}^{\mathbf{C}})^T(\boldsymbol{\beta}^{\mathbf{C}})]\},$$

Then we can update $\sigma^{(t+1)}$ by:

$$\sigma^{(t+1)} = \frac{1}{2\sqrt{g}(N+K+2C+2+L)}\Big\{\sqrt{2}\|D^{(t)}(\boldsymbol{\beta}^{\mathbf{G}})^{(t+1)}\|_1 + \big[2\|D^{(t)}(\boldsymbol{\beta}^{\mathbf{G}})^{(t+1)}\|_1^2 \tag{2.23}$$

$$+ 4(N+K+2C+2+L)(\|\mathbf{Y}-\mathbf{G}(\boldsymbol{\beta}^{\mathbf{G}})^{(t+1)}-\mathbf{C}(\boldsymbol{\beta}^{\mathbf{C}})^{(t+1)}\|_2^2 + \|(\boldsymbol{\beta}^{\mathbf{C}})^{(t+1)}\|_2^2 + 2d)\big]^{\frac{1}{2}}\Big\} \tag{2.24}$$

Finally, the updates $\mathbf{b}^{(t+1)} = \arg\max_{\mathbf{b}\in\mathbb{R}^{q+1}} Q_2(\mathbf{b}|\boldsymbol{\beta}^{\mathbf{C}(t)}, \boldsymbol{\beta}^{\mathbf{C}(t)}, \mathbf{b}^{(t)}, \boldsymbol{\sigma}^{(t)})$ can be computed using box-constrained optimization routines. Here we assume $a = 1$ for computational convenience.

After obtaining posterior estimates of $\mathbf{b}$ and $\boldsymbol{\beta}^{\mathbf{G}}$, we follow a hierarchical variable selection strategy. First, we select important groups by discarding groups with b_l reaching lower bound or with small value. Thereafter, we choose variables only from the informative groups (passing selection criteria in the previous step) by thresholding values of $\beta_k^{\mathbf{G}}$.

2.6 *Handling Survival time with Right Censored Data*

Often in clinical studies, the clinical outcome is the patient survival time with right censoring. If we have censorship information, then we may use the AFT model with data augmentation approach (Sha *et al.*, 2006) to deal with the right censored situation. In this case, we observe (t_i, δ_i) for subject i, where t_i is the survival time and δ_i is the censoring status for subject i with $\delta_i = 0$ for censored, and $\delta_i = 1$ otherwise. We also use T_i to denote the real failure time for subject i. Therefore, for $i = 1, 2, ..., N$, we have

$$\begin{cases} t_n = T_n, \delta_n = 1 \\ t_n < T_n, \delta_n = 0. \end{cases}$$

Based on the AFT model our SSDIM (2.16) will now be, $\ln \mathbf{T} \sim N(\mathbf{C}\boldsymbol{\beta}^{\mathbf{C}} + \mathbf{G}\boldsymbol{\beta}^{\mathbf{G}}, \sigma^2\mathbf{I})$, where $\mathbf{T} = (T_1, ...T_N)^T$. For convenience, let $\mathbf{Y} = (Y_1, ..., Y_N)^T = (\ln t_1, ..., \ln t_N)^T$. For those censored subjects, we can use latent variables by sampling from $\mathbf{Z} =$ Truncated $N(\mathbf{C}\boldsymbol{\beta}^{\mathbf{C}}+\mathbf{G}\boldsymbol{\beta}^{\mathbf{G}}, \sigma^2\mathbf{I}_{N\times N}, Y, +\infty)$

to replace $\mathbf{Y}$ (for censored individuals), where Truncated $N(\mu, \zeta^2, A, B)$ is truncated normal with mean μ, variance ζ^2 and support $[A, B]$. Model extension for binary data is also provided in Section 2 of Supplementary material.

3. Simulation Studies

This section aims to assess the operating characteristics of our proposed FSDIM and SSDIM through synthetic numeric examples. These simulation schemes were introduced by Ročková and Lesaffre (2014) and Wang *et al.* (2013). In this section, we consider two distinct scenarios. In the first scenario, we follow the simulation study by Ročková and Lesaffre (2014) to evaluate the performance of SSDIM for a given pre-specified groupings. That means here, we do not need to employ our FSDIM to learn any grouping knowledge. In the second scenario, we adopt the similar data-generating scheme proposed by Wang *et al.* (2013) to simulate both gene and DNA methylation data for the corresponding patient set and apply both FSDIM and SSDIM described above. Under this second scenario, we are actually investigating the full potential and benefits of our two-step data integration approach. Under each simulation setting we create separate grouping scenarios. The details of simulation formula derivation and data generation for Simulation 1 and 2 along with the grouping scenarios are included in the Section 3 of the Supplementary material.

3.1 *Simulation Study 1*

The false discoveries (FD) and false negatives (FN) under two grouping scenarios (details in Section 3 of Supplementary material) in Simulation Study 1 are summarised in Table 1. From the results of Grouping Scenario 1 in Table 1, we can see increasing a induces a larger penalty, which causes the estimated weights assigned to predictive groups, $\mathbf{b}$, to increase because higher weights with inverse link function compensate for the larger penalty. As a result, FD drops and

FN increase with an increment of a. A similar trend could be observed in the result of Grouping Scenario 2 (Table 1) as a increases, false discoveries within both the non-predictive group (FD1) and within the predictive group (FD2) drop.

The estimation of $\boldsymbol{b}$ reflects the importance of groups since we choose the inverse link function in (2.18). In Grouping Scenario 1, the elements of $\mathbf{b}$ corresponding to non-relevant groups attain the lower bound we set for box-constrained optimization of $\mathbf{b}$ or a relatively small value. In contrast, those associated with the predictive group remain significantly larger than zero. This suggests the model's strong potential to detect the predictive biomarker groups as a whole. For the second grouping scenario, where there exists "within-group sparsity", the magnitude of estimated weights indicates the proportion of important variables of the corresponding group; in other words, the denser informative variables in a group l, the larger value of the estimation of corresponding group weight, b_l. This is generally true for all hyper-parameter settings in Table 1 (except $a = 3, g = 1$).

We now compare the false discovery results in Table 1 between cases when grouping information used (Grouping Scenario 1 and 2) with not used (NEG) and iBAG model. It is clearly seen that using a hierarchical selection strategy (FDH) (Section 2.5) with grouping information results in much lower false discoveries compared to the naive NEG without grouping information. Recall that in a simple NEG setting in (2.18), the inverse link function h is replaced by a fixed constant. The FDH takes into account the sequential shrinkage by first selecting the groups and then selecting the variables within the selected groups. Whereas FD only selects variables without considering any group-level shrinkage. For this reason, we see similar results comparing the FD columns from Table 1 (Grouping Scenario 1, 2, and NEG-No Grouping). But on the other hand, comparing the FDH columns from Table 1 (Grouping Scenario 1 and 2) with the FD column of NEG-No Grouping shows a distinct advantage. Moreover, suppose the grouping is good enough in the sense that predictive variables are correctly grouped all together (grouping scenario 1,

without any within-group sparsity). In that case, the improvement in model performance is dramatic. To be noted here that under Simulation 1 it is not feasible to run the iBAG model. Hence no comparison with iBAG was made in Table 1.

3.2 *Simulation Study 2*

Here we want to compare our two-stage driven grouped procedure with a non-grouped NEG-prior-based procedure and iBAG framework.

Recall that $\hat{b}_l$ are parameters weighting the importance of the l-th group; the larger $\hat{b}_l$ is, the more important the l-th group is, hence less penalty applied to that group. We expect groups with predictive variables to have larger $\hat{b}_i$. From Table 2, we find that the estimation of weight parameters $\boldsymbol{b}$ is close to the lower limit set in the optimization algorithm. This is not odd, as the three groups' predictive covariates are very low. The FDH from our two-stage model is significantly smaller than the FD of the model without the grouping information. This implies that after incorporating the grouping information provided by FSDIM, the overall performance of SSDIM in terms of variable selection is significantly improved. On the contrary, iBAG yields a large FD in high-dimensional seting where sample size is much smaller than the feature size. Thus conclusively shows the direct benefits of multi-platform data integration.

4. TCGA Glioblastoma Multiplatform Dataset

In this section, we aim to integrate gene expression data, gene functional classification database (from DAVID), DNA methylation data, and patients' clinical features from the Glioblastoma (GBM) study (Network *et al.*, 2008; Brennan *et al.*, 2013) to find genes significantly associated with the clinical outcome of interest, which is survival time in this dataset. The dataset is available on cBioPortal's website (https://www.cbioportal.org/). After removing patients with incomplete data, our final data consists of 155 patients with an expression level of 11,861 genes and a total

methylation level of 9,020 probes. This dataset's clinical features of interest are age, prior glioma record, gender, and previous treatment history.

We reduce our gene set size by choosing 1000 genes with the smallest p-values by fitting univariate Accelerated Failure Time models. Then we find the methylation sites corresponding to the 1000 genes we selected based on Entrez ID. In total, 1000 genes from 155 patients remain in our final analysis. Next, we use DAVID (https://david.ncifcrf.gov) functional classification tool (Sherman *et al.*, 2009; Huang *et al.*, 2009) to partition genes, which yields 28 categories. Then, we use quadratic FSDIM to group genes using gene expression, methylation data, and results of DAVID functional classification. Last but not least, in SDDIM, we use Scheme EM (see Section 2.5) to estimate the coefficients.

We fix $g = 1$ (non-scaling) and search the optimal a along $\{0.2, 0.5, 1, 1.5, 2, 5\}$ based on 10-fold cross-validation. To evaluate the performance of each model with different parameters, we compute the concordance index (C-index), which is defined as the proportion of all pairs of subjects whose survival time can be ordered such that the subject with the higher predicted survival is the one who survived longer (Harrell Jr, 2015). It is equivalent to $\sum_{(i,j)\in\Phi} I(\hat{t}_i, \hat{t}_j)/|\Phi|$, where $I(\hat{t}_i, \hat{t}_j) = 1$ for $\hat{t}_i > \hat{t}_j$ and 0 otherwise, $\hat{t}_i$ is the estimated survival time for patient i and Φ is the set that consists of all pairs of i, j such that $t_i > t_j$. Therefore, it serves as an effective metric for evaluating the performance of the prediction of survival data. We compute C-index for each left-out fold and take the average of the C-index across ten folds. In Table 3, we present prediction accuracy (C-index) based on Scheme EM in SSDIM.

We focus on sparse models to lower false discovery, among which the model with $a = 0.5$ yields the largest test C-index (larger than test C-index of iBAG or LASSO, see bolded row in Table 3). In this setting, we fit the model again with all data and detect 24 genes are selected by this optimal model, including one M-effect gene (IL8), four mixed effect genes (RANBP17, CHI3L1, SNX10 and MARS) and 19 $\bar{M}$-effect genes (see Table 4). The M-effect gene, IL8, is

reported to be expressed and released at high levels both *in vitro* and *in vivo*, and recent research indicates that it is important for glial tumor neovascularity and development (Brat *et al.*, 2005). Moreover, in Schwalbe *et al.* (2013), IL8 methylation was reported as a unique independent model high-risk biomaker in cross-validated survival models of subgroups of medulloblastoma patients. Including IL8 into existing survival models considerably foster disease risk prediction. Similarly, among mixed effect genes, the expression level of RANBP17, CHI3L1, and SNX10 are with prognosis of GBM and their methylation status are suggested to play an important role in GBM or other cancer. RANBP17 is expected to be a potential prognostic and immunotherapeutic biomarker in GBM. Furthermore, methylation levels at RANBP17 CpG sites, cg01182310, are higher in GBM and correlation between methylation of RANBP17 at this CpG location and a poor outcome in GBM patients is found in Tang *et al.* (2023). The increased level of mRNA level of CHI3L1 could be associated with poor patient survival in the GBM. The hypermethylation phenotype of CHI3L1 could lead to low CHI3L1 expression and good prognosis (Steponaitis *et al.*, 2016). SNX10 is identified as selectively expressed in glioblastoma stem cells compared with neural stem cells and essential for glioblastoma stem cells survival (Gimple *et al.*, 2023). Jiang *et al.* (2021) find that SNX10 shows abnormal expression and methylation associated with cervical cancer development. Most of $\bar{M}$-effect genes are reported differentially expressed in GBM in other studies (see Table 4 for all supporting literature), such as FPR1, whose expression has an association with the grades of the malignancy of glioma cells (Liu *et al.*, 2012), STY2L, whose overexpression are reported in numerous human cancer tissues including glioma and is involved in tumor occurrence and progression (Chuang *et al.*, 2022), and PCYOX1, which is detected as an independent predictive factor for progressive malignancy in glioma (Cao *et al.*, 2020).

From the boxplots illustrating the distribution of $\boldsymbol{\beta}^{\mathbf{C}}$ in cross-validation folds (Figure 4), we can see consistent results across folds. The coefficients of age and prior glioma are always negative, implying that older patients or patients with prior glioma histories have poorer GBM survival,

which could be confirmed by previous studies (Bozdag *et al.*, 2013; Winger *et al.*, 1989). The coefficient of pretreatment history is positive, which suggests that patients with a pretreatment history tend to have longer overall GBM survival. Pretreatment like surgical resection and the use of adjuvant radiation therapy are reported to improve survival in GBM patients (Thumma *et al.*, 2012). The finding that the coefficient of sex is negative (which means male GBM patients have slightly better survival outcome than females) contradicts previous finding (Tian *et al.*, 2018) that females GBM patients have overall better survival outcome. It's possible that such a gender discrepancy in previous study is due to unmeasured confounders such as genetic variants and hormone exposure mediated by gene expression.

5. Conclusions

In this paper, we have proposed a Bayesian two-stage model for data integration and variable selection. Compared with the original iBAG model (Wang *et al.*, 2013), our model makes improvements in several aspects. Firstly, in iBAG, Bayesian Lasso (Park and Casella, 2008) with Gibbs samplers are used in both stages, which potentially suffers from slow convergence and interpretability issues. On the contrary, our EM-based approach avoids sampling and is substantially fast. Secondly, iBAG cannot provide a direct approach for variable selection. Whereas our proposed model provides a straightforward and natural way of variable selection as we directly use gene expression as our predictors in our model. Thirdly, our model is flexible enough to explore non-linear associations between methylation and gene expressions. Fourthly, our model can also integrate prior information of genes groupings based on gene functional classification to improve the performance of the FSDIM by introducing a random intercept for genes within the same group. This prior knowledge-based random intercept reflects the net effect of the biological process on the expression level since as reported in previous study, genes with related functions having similar expression. This effect could come from the gene function database based on how

similar the function of those genes are (e.g. DAVID, https://david.ncifcrf.gov). This is a significant generalization of the problem at hand as now we can utilize the gene function database to create prior knowledge, which can be complementary information of expression and epigenetic data.

Our two-stage model is capable of integrating high-dimensional multiplatform data by first learning grouping information from the biological regulation and then performing variable selection based on grouping information learned. Our simulation studies suggest that learning this grouping information can improve the overall performance of our model in selecting relevant predictive variables. Real data analysis with TCGA GBM data reveals that our model can identify clinically and biologically important genes associated with patient GBM survival.

Our model exhibits two limitations. Firstly, the initialization of the parameters may influence the study results due to the potential multimodality of posteriors, as suggested by Park and Casella (2008). Secondly, the overlap of the true grouping structure in SSDIM could also impact the performance of our model. However, these limitations can be mitigated under specific conditions, where the external grouping structure is notably simple with minimal overlap and multiple initial values are explored. Under those conditions, our proposed two-stage hierarchical Bayesian model demonstrates considerable efficacy. For future work, different data types of clinical outcomes of interest can be considered for a generalized SSDIM, like categorical outcomes and zero-inflated count data. Moreover, information from all data platforms may not be readily available for all patients. Therefore, missing data would be a potential concern that requires significant attention and research.

6. Supplementary Material

Further details about our simulation models, posterior derivations, and extension to binary outcomes 2-5 are available in Supplementary Material.

Data Availability Statement

Software in the form of R code, together with a sample input data set and complete documentation can be found in our GitHub repositories https://github.com/xvehao/EMMultiOmics.

References

Arora, Gurpreet K, Palamiuc, Lavinia and Emerling, Brooke M. (2022). Expanding role of pi5p4ks in cancer: A promising druggable target. *FEBS letters* **596**(1), 3–16.

Bernardo, JM, Bayarri, MJ, Berger, JO, Dawid, AP, Heckerman, D, Smith, AFM and West, M. (2003). Bayesian factor regression models in the "large p, small n" paradigm. *Bayesian statistics* **7**, 733–742.

Bouchart, Christelle, Trépant, Anne-Laure, Hein, Matthieu, Van Gestel, Dirk and Demetter, Pieter. (2020). Prognostic impact of glioblastoma stem cell markers olig2 and ccnd2. *Cancer Medicine* **9**(3), 1069–1078.

Bozdag, Serdar, Li, Aiguo, Riddick, Gregory, Kotliarov, Yuri, Baysan, Mehmet, Iwamoto, Fabio M, Cam, Margaret C, Kotliarova, Svetlana and Fine, Howard A. (2013). Age-specific signatures of glioblastoma at the genomic, genetic, and epigenetic levels. *PloS one* **8**(4), e62982.

Brat, Daniel J, Bellail, Anita C and Van Meir, Erwin G. (2005). The role of interleukin-8 and its receptors in gliomagenesis and tumoral angiogenesis. *Neuro-oncology* **7**(2), 122–133.

Brennan, Cameron W, Verhaak, Roel GW, McKenna, Aaron, Campos, Benito, Noushmehr, Houtan, Salama, Sofie R, Zheng, Siyuan, Chakravarty, Debyani, Sanborn, J Zachary, Berman, Samuel H *et al.* (2013). The somatic genomic landscape of glioblastoma. *Cell* **155**(2), 462–477.

Cao, Jing-Yuan, Guo, Qing, Guan, Ge-Fei, Zhu, Chen, Zou, Cun-Yi, Zhang, Lu-Yang, Cheng, Wen, Wang, Guo-Li, Cheng, Peng, Wu, An-Hua *et al.* (2020). Elevated lymphocyte specific protein 1 expression is involved in the regulation of leukocyte migration and immunosuppressive microenvironment in glioblastoma. *Aging (Albany NY)* **12**(2), 1656.

Chuang, Chi-Cheng, Lan, Yu-Hsiang, Lu, Yu-Jen, Weng, Yu-Lun and Chen, Jyh-Ping. (2022). Targeted delivery of irinotecan and slp2 shrna with grp-conjugated magnetic graphene oxide for glioblastoma treatment. *Biomaterials Science* **10**(12), 3201–3222.

Cole, Stephen R, Chu, Haitao and Greenland, Sander. (2014). Maximum likelihood, profile likelihood, and penalized likelihood: a primer. *American journal of epidemiology* **179**(2), 252–260.

Consortium, International Cancer Genome *et al.* (2010). International network of cancer genome projects. *Nature* **464**(7291), 993.

De Larco, Joseph E, Wuertz, Beverly RK, Yee, Douglas, Rickert, Brenda L and Furcht, Leo T. (2003). Atypical methylation of the interleukin-8 gene correlates strongly with the metastatic potential of breast carcinoma cells. *Proceedings of the National Academy of Sciences* **100**(24), 13988–13993.

Fan, Jianqing and Li, Runze. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**(456), 1348–1360.

Gerber, Naamit K, Goenka, Anuj, Turcan, Sevin, Reyngold, Marsha, Makarov, Vladimir, Kannan, Kasthuri, Beal, Kathryn, Omuro, Antonio, Yamada, Yoshiya, Gutin, Phillip *et al.* (2014). Transcriptional diversity of long-term glioblastoma survivors. *Neuro-oncology* **16**(9), 1186–1195.

Gimple, Ryan C, Zhang, Guoxin, Wang, Shuai, Huang, Tengfei, Lee, Jina, Taori,

Suchet, Lv, Deguan, Dixit, Deobrat, Halbert, Matthew E, Morton, Andrew R *et al.* (2023). Sorting nexin 10 sustains pdgf receptor signaling in glioblastoma stem cells via endosomal protein sorting. *JCI insight* **8**(6).

Griffin, Jim E and Brown, Philip J. (2011). Bayesian hyper-lassos with non-convex penalization. *Australian & New Zealand Journal of Statistics* **53**(4), 423–442.

Harrell Jr, Frank E. (2015). *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer.

Heller, Sonja, Maurer, Gabriele D, Wanka, Christina, Hofmann, Ute, Luger, Anna-Luisa, Bruns, Ines, Steinbach, Joachim P and Rieger, Johannes. (2018). Gene suppression of transketolase-like protein 1 (tktl1) sensitizes glioma cells to hypoxia and ionizing radiation. *International Journal of Molecular Sciences* **19**(8), 2168.

Huang, Da Wei, Sherman, Brad T and Lempicki, Richard A. (2009). Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature protocols* **4**(1), 44–57.

Jarquin, Diego, Roy, Arkaprava, Clarke, Bertrand and Ghosal, Subhashis. (2023). Combining phenotypic and genomic data to improve prediction of binary traits. *Journal of Applied Statistics*, 1–27.

Jennings, Elizabeth M, Morris, Jeffrey S, Carroll, Raymond J, Manyam, Ganiraju C and Baladandayuthapani, Veerabhadran. (2013). Bayesian methods for expression-based integration of various types of genomics data. *EURASIP Journal on Bioinformatics and Systems Biology* **2013**(1), 13.

Ji, Pengxiang, Shan, Xueshi, Wang, Jian, Zhang, Ping and Cai, Zhan. (2022). Inte-

grative analysis of cbr1 as a prognostic factor associated with idh-mutant glioblastoma in the chinese population. *American Journal of Translational Research* **14**(8), 5394.

Jiang, Pinping, Cao, Ying, Gao, Feng, Sun, Wei, Liu, Jinhui, Ma, Ziyan, Xie, Manxin and Fu, Shilong. (2021). Snx10 and ptgds are associated with the progression and prognosis of cervical squamous cell carcinoma. *BMC cancer* **21**(1), 1–14.

Lakshminarasimhan, Ranjani and Liang, Gangning. (2016). The role of dna methylation in cancer. *DNA Methyltransferases-Role and Function*, 151–172.

Li, Ziyi, Chang, Changgee, Kundu, Suprateek and Long, Qi. (2020). Bayesian generalized biclustering analysis via adaptive structured shrinkage. *Biostatistics* **21**(3), 610–624.

Liu, Mingyong, Zhao, Jianhua, Chen, Keqiang, Bian, Xiuwu, Wang, Chunyan, Shi, Ying and Wang, Ji Ming. (2012). G protein-coupled receptor fpr1 as a pharmacologic target in inflammation and human glioblastoma. *International immunopharmacology* **14**(3), 283–288.

Liu, Yuqiao, Meng, Yiting, Zhang, Tian and Alachkar, Houda. (2021). Deregulation of apolipoprotein c2 gene in cancer: A potential metabolic vulnerability. *Clinical and Translational Medicine* **11**(6).

Luo, Chongyuan, Hajkova, Petra and Ecker, Joseph R. (2018). Dynamic dna methylation: In the right place at the right time. *Science* **361**(6409), 1336–1340.

Maity, Arnab Kumar, Lee, Sang Chan, Mallick, Bani K and Sarkar, Tapasree Roy. (2020). Bayesian structural equation modeling in multiple omics data with application to circadian genes. *Bioinformatics* **36**(13), 3951–3958.

Min, Eun Jeong, Chang, Changgee and Long, Qi. (2018). Generalized bayesian factor analysis for integrative clustering with applications to multi-omics data. In: *2018 IEEE 5th*

International Conference on Data Science and Advanced Analytics (DSAA). IEEE. pp. 109–119.

Network, Cancer Genome Atlas Research *et al.* (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**(7216), 1061.

Panagopoulos, Alexandros Theodoros, Gomes, Renata Nascimento, Almeida, Fernando Gonçalves, da Costa Souza, Felipe, Veiga, José Carlos Esteves, Nicolaou, Anna and Colquhoun, Alison. (2018). The prostanoid pathway contains potential prognostic markers for glioblastoma. *Prostaglandins & Other Lipid Mediators* **137**, 52–62.

Park, Trevor and Casella, George. (2008). The bayesian lasso. *Journal of the American Statistical Association* **103**(482), 681–686.

Perperoglou, Aris, Sauerbrei, Willi, Abrahamowicz, Michal and Schmid, Matthias. (2019). A review of spline function procedures in r. *BMC medical research methodology* **19**(1), 1–16.

Pham, Kien, Luo, Defang, Liu, Che and Harrison, Jeffrey K. (2012). Ccl5, ccr1 and ccr5 in murine glioblastoma: immune cell infiltration and survival rates are not dependent on individual expression of either ccr1 or ccr5. *Journal of neuroimmunology* **246**(1-2), 10–17.

Polano, Maurizio, Fabbiani, Emanuele, Andreuzzi, Eva, Cintio, Federica Di, Bedon, Luca, Gentilini, Davide, Mongiat, Maurizio, Ius, Tamara, Arcicasa, Mauro, Skrap, Miran *et al.* (2021). A new epigenetic model to stratify glioma patients according to their immunosuppressive state. *Cells* **10**(3), 576.

Qian, Jiawen, Luo, Feifei, Yang, Jiao, Liu, Jun, Liu, Ronghua, Wang, Luman, Wang, Chen, Deng, Yuting, Lu, Zhou, Wang, Yuedi *et al.* (2018). Tlr2 promotes glioma immune

evasion by downregulating mhc class ii molecules in microgliatlr2 downregulates microglial mhc class ii in glioma. *Cancer Immunology Research* **6**(10), 1220–1233.

Ročková, Veronika and George, Edward I. (2014). Emvs: The em approach to bayesian variable selection. *Journal of the American Statistical Association* **109**(506), 828–846.

Ročková, Veronika and Lesaffre, Emmanuel. (2014). Incorporating grouping information in bayesian variable selection with applications in genomics. *Bayesian Analysis* **9**(1), 221–258.

Schwalbe, Edward C, Williamson, Daniel, Lindsey, Janet C, Hamilton, Dolores, Ryan, Sarra L, Megahed, Hisham, Garami, Miklós, Hauser, Peter, Dembowska-Baginska, Bożena, Perek, Danuta *et al.* (2013). Dna methylation profiling of medulloblastoma allows robust subclassification and improved outcome prediction using formalin-fixed biopsies. *Acta neuropathologica* **125**, 359–371.

Sevilla, Jose L, Segura, Victor, Podhorski, Adam, Guruceaga, Elizabeth, Mato, Jose M, Martinez-Cruz, Luis A, Corrales, Fernando J and Rubio, Angel. (2005). Correlation between gene expression and go semantic similarity. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **2**(4), 330–338.

Sha, Naijun, Tadesse, Mahlet G and Vannucci, Marina. (2006). Bayesian variable selection for the analysis of microarray data with censored outcomes. *Bioinformatics* **22**(18), 2262–2268.

Sherman, Brad T, Hao, Ming, Qiu, Ju, Jiao, Xiaoli, Baseler, Michael W, Lane, H Clifford, Imamichi, Tomozumi and Chang, Weizhong. (2022). David: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res*.

Sherman, Brad T, Lempicki, Richard A *et al.* (2009). Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature protocols* **4**(1), 44.

Steponaitis, Giedrius, Skiriutė, Daina, Kazlauskas, Arunas, Golubickaitė, Ieva, Stakaitis, Rytis, Tamašauskas, Arimantas and Vaitkienė, Paulina. (2016). High chi3l1 expression is associated with glioma patient survival. *Diagnostic pathology* **11**(1), 1–8.

Sun, Wei, Bunn, Paul, Jin, Chong, Little, Paul, Zhabotynsky, Vasyl, Perou, Charles M, Hayes, David Neil, Chen, Mengjie and Lin, Dan-Yu. (2018). The association between copy number aberration, dna methylation and gene expression in tumor samples. *Nucleic acids research* **46**(6), 3009–3018.

Tang, Hai, Liu, Xuemei, Li, Huahan, Huang, Huacong, Ran, Chao, Wen, Baoying, Liu, Qundi, Lu, Huaisun, Jing, Meilian and Zhou, Lingqi. (2023). Comprehensive analysis reveals ranbp17 as a potential biomarker for prognosis and immunotherapy in glioblastoma.

Thumma, Sudheer R, Fairbanks, Robert K, Lamoreaux, Wayne T, Mackay, Alexander R, Demakas, John J, Cooke, Barton S, Elaimy, Ameer L, Hanson, Peter W and Lee, Christopher M. (2012). Effect of pretreatment clinical factors on overall survival in glioblastoma multiforme: a surveillance epidemiology and end results (seer) population analysis. *World journal of surgical oncology* **10**, 1–12.

Tian, Minjie, Ma, Wenying, Chen, Yueqiu, Yu, Yue, Zhu, Donglin, Shi, Jingping and Zhang, Yingdong. (2018). Impact of gender on the survival of patients with glioblastoma. *Bioscience reports* **38**(6).

Wang, Jianjiao, Yang, Yan, Cao, Yuandong and Tang, Xinyu. (2019). mir-342 inhibits glioma cell proliferation by targeting gprc5a. *Molecular Medicine Reports* **20**(1), 252–260.

Wang, Wenting, Baladandayuthapani, Veerabhadran, Morris, Jeffrey S, Broom,

Bradley M, Manyam, Ganiraju and Do, Kim-Anh. (2013). ibag: integrative bayesian analysis of high-dimensional multiplatform genomics data. *Bioinformatics* **29**(2), 149–159.

Winger, Michael J, Macdonald, David R and Cairncross, J Gregory. (1989). Supratentorial anaplastic gliomas in adults: the prognostic importance of extent of resection and prior low-grade glioma. *Journal of neurosurgery* **71**(4), 487–493.

Wong, Kin Yau, Zeng, Donglin and Lin, DY. (2018). Efficient estimation for semiparametric structural equation models with censored data. *Journal of the American Statistical Association* **113**(522), 893–905.

Zhu, Ruoqing, Zhao, Qing, Zhao, Hongyu and Ma, Shuangge. (2016). Integrating multidimensional omics data for cancer outcome. *Biostatistics* **17**(4), 605–618.

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

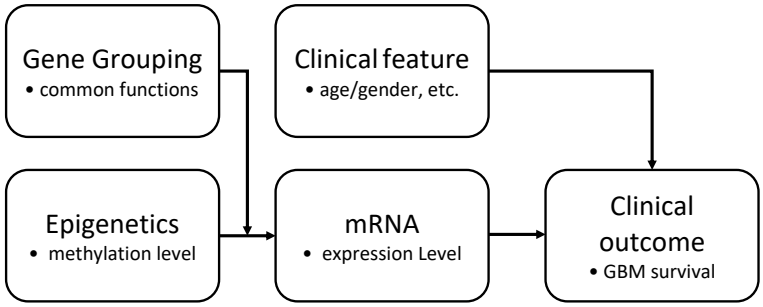

Fig. 1. A schematic diagram of multiplatform data sets and their mutual relationships. Functionally related genes have similar genomic expressions. Methylation mediates gene expression and then influences clinical outcome together with clinical features.

[*Received xx, 2023; revised xx, 2023; accepted for publication xx, 2023*]

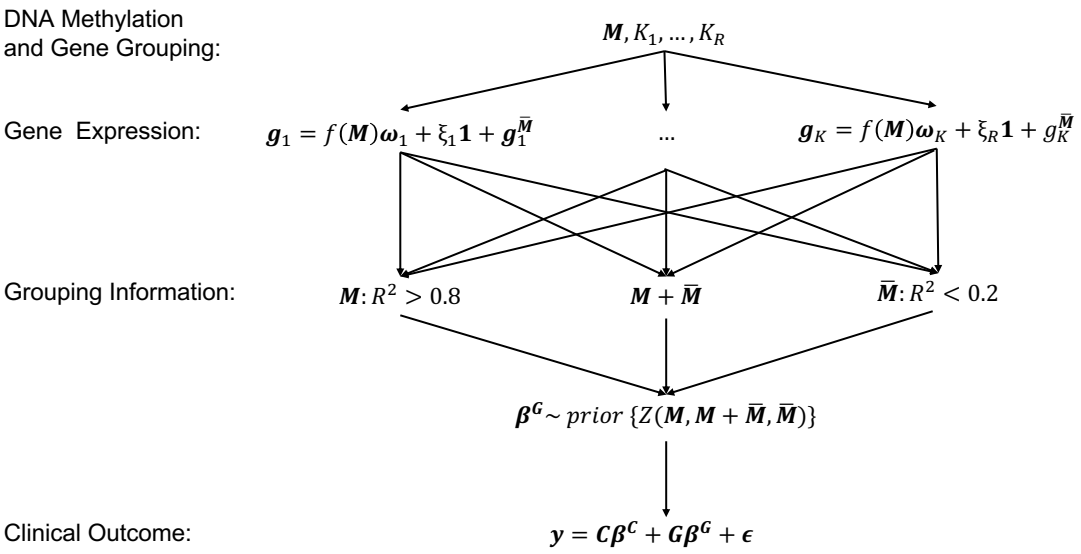

Fig. 2. Schematic of our proposed model. In the first stage, gene expression levels are regressed on methylation levels with gene functions as random intercept. In the second stage, clinical outcomes are regressed on gene expression levels and clinical features, with grouped penalty based on how gene expression are regulated by methylation, as learned from the first stage.

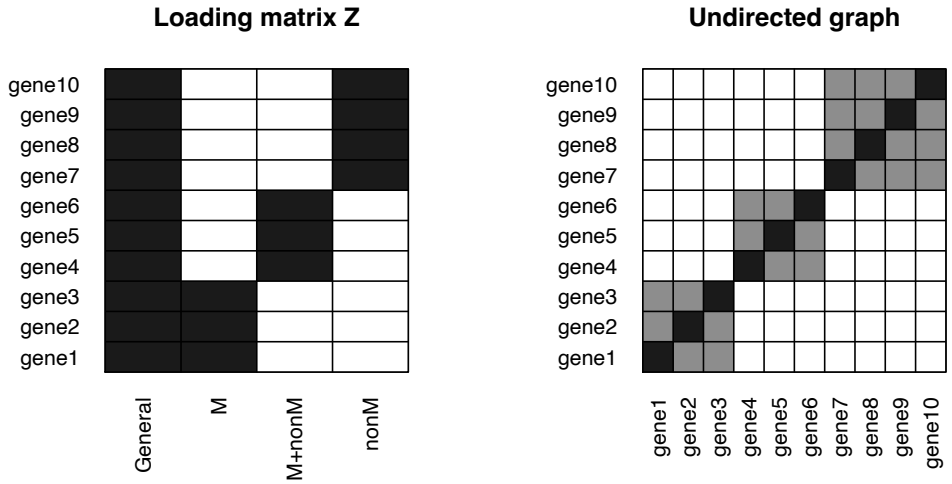

Fig. 3. Loading matrix and undirected graph

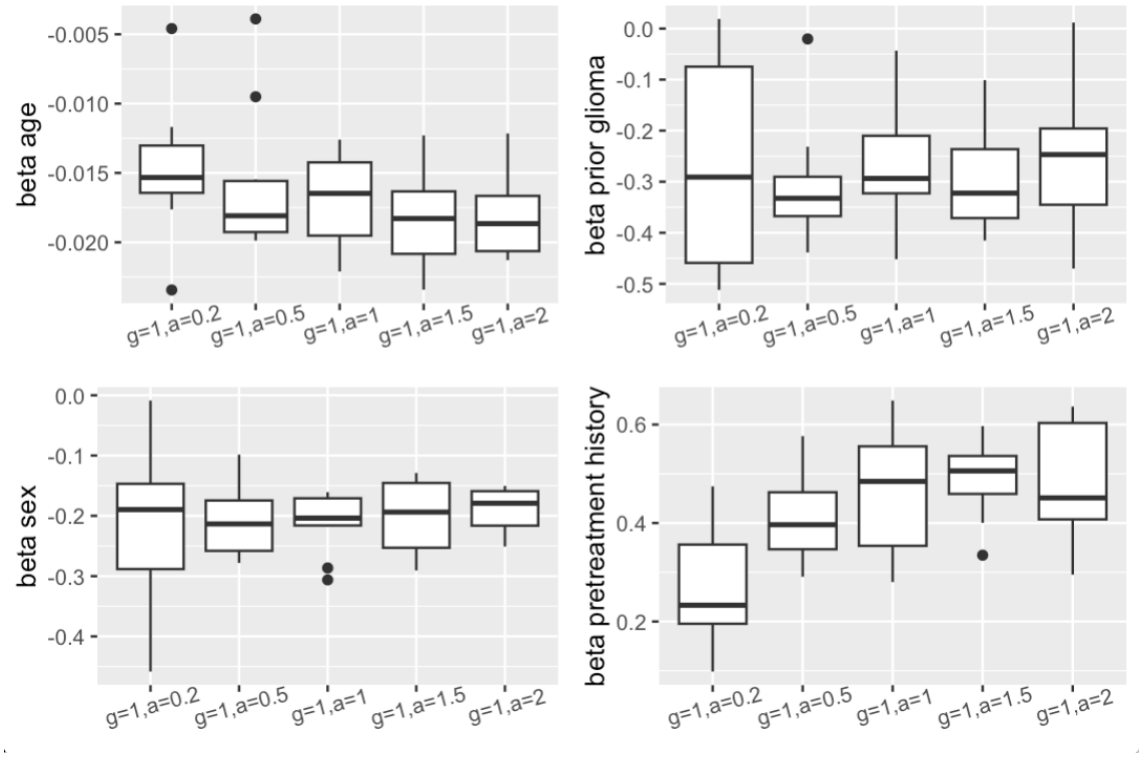

Fig. 4. Boxplot of Clinical Coefficients Estimation across 10-folds in Quadratic FSDIM and SSDIM with Scheme EM

Table 1. *Summarized results of Simulation Study 1. FD/FD1/FD2/FDH refer to the number of false discoveries overall/within the non-predictive group/within the predictive group/overall after hierarchical selection. FN refers to the number of false negatives overall after variable selection. The size and sparsity mean, respectively, the number and the proportion of predictive variables in each group. Refer to the Supplementary document for full description of all row and column headings.*

	Grouping 1								Grouping 2									NEG		
Size					5	15	5	975						10	20	30	940			10
Sparsity					1	0	1	0						1/2	1/4	0	0			1/100
	FD	FDH	FN	$\hat{b}_0$	$\hat{b}_1$	$\hat{b}_2$	$\hat{b}_3$	$\hat{b}_4$	FD1	FD2	FDH	FN	$\hat{b}_0$	$\hat{b}_1$	$\hat{b}_2$	$\hat{b}_3$	$\hat{b}_4$	FD	FN	$\hat{b}_0$
	No scaling $g = 1$																			
a=0.5	20	0	0	1e-04	2.44175	1e-04	2.43162	1e-04	42	0	0	0	1e-04	0.0029	1e-04	1e-04	1e-04	39	0	0.00010
a=1	17	0	0	1e-04	4.62672	1e-04	4.60571	1e-04	26	2	0	0	1e-04	1e-04	1e-04	1e-04	1e-04	38	0	0.00118
a = 3	0	0	0	0.00048	6.09301	0.00051	6.0889	0.00048	0	16	6	0	1e-04	12.9122	27.6935	1e-04	0.0017	0	2	0.00422
	Rescaled prior $g = 1/N^2$																			
a =0.5	7	0	0	1e-04	2.4998	1e-04	2.4998	1e-04	14	0	0	0	1e-04	0.0019	0.0012	1e-04	1e-04	14	0	0.00103
a =1	1	0	0	1e-04	4.9989	1e-04	4.99894	1e-04	4	0	0	0	1e-04	0.2292	0.0017	1e-04	1e-04	4	0	0.00141
a = 3	0	0	2	0.00048	6.09301	0.00051	6.0889	0.00048	0	0	0	2	1e-04	9.0551	6.6783	1e-04	1e-04	0	2	0.00616

Table 2. *Summarized results of Simulation Study 2: (1) FD = False Discovery in gene selection; (2) FDH = False Discovery after hierarchical selection; (3) FN = False Negative in gene selection. There are* 15 *predictive genes in total. To control FD, we apply scaling* $g = 1/(N^2)$ *as recommended by Ročková and Lesaffre (2014). Refer to the Supplementary document for full description of all row and column headings.*

	Grouping							NEG			iBAG	
size					200	100	300				600	
sparsity					2.5%	5%	1.67%				2.5%	
	FD	FDH	FN	$\hat{b}_0$	$\hat{b}_1$	$\hat{b}_2$	$\hat{b}_3$	FD	FN	$\hat{b}_0$	FD	FN
a = 2	13	6	0	0.0013	7e-04	0.0015	0.0019	13	0	0.0028		
a = 3	3	0	0	0.0018	9e-04	0.0045	1e-04	3	0	0.0076	33	14
a = 4	1	0	0	0.0043	0.0341	0.0916	0.0029	1	0	0.0256		

Table 3. *Average performance of Quadratic FSDIM and SSDIM Model with Scheme EM and* $g = 1$ *across 10 folds*

	a	# selected variables	train C-index	test C-index
	0.2	38.7	0.819	0.574
	0.5	18.0	0.755	**0.681**
SSDIM	1	13.6	0.752	0.665
	1.5	11.3	0.745	0.617
	2	9.9	0.735	0.639
	5	6.0	0.695	0.669
iBAG			0.962	0.590
LASSO		2.1	0.649	0.670

Table 4. *All 24 genes selected by the optimal model followed by supporting literature. Those without citations are new genes needed to be further studied.*

M-effect:
IL8 (De Larco *et al.*, 2003)
Mixed-effect:
SNX10 (Gimple *et al.*, 2023; Jiang *et al.*, 2021),
MARS,
CHI3L1 (Tang *et al.*, 2023; Steponaitis *et al.*, 2016),
RANBP17 (Tang *et al.*, 2023)
$\bar{M}$-effect:
JMJD1A,
OLIG2 (Bouchart *et al.*, 2020),
CBR1 (Ji *et al.*, 2022),
FPRL1 (Liu *et al.*, 2012),
SYTL2 (Chuang *et al.*, 2022),
P2RY6,
CCR1 (Pham *et al.*, 2012),
TLR2 (Qian *et al.*, 2018),
F3 (Gerber *et al.*, 2014),
SECTM1 (Polano *et al.*, 2021),
GPRC5A (Wang *et al.*, 2019),
PIP4K2C (Arora *et al.*, 2022),
APOC2 (Liu *et al.*, 2021),
TKTL1 (Heller *et al.*, 2018),
ZNF135,
PCYOX1 (Cao *et al.*, 2020),
LSP1,
ZCCHC2,
HPGD (Panagopoulos *et al.*, 2018)