Promoting Similarity of Sparsity Structures in Integrative Analysis with Penalization

YUAN HUANG*, QINGZHAO ZHANG*, SANGUO ZHANG,
JIAN HUANG, SHUANGGE MA

Abstract

For data with high-dimensional covariates but small sample sizes, the analysis of single datasets often generates unsatisfactory results. The integrative analysis of multiple independent datasets provides an effective way of pooling information and outperforms single-dataset and several alternative multi-datasets methods. Under many scenarios, multiple datasets are expected to share common important covariates, that is, the corresponding models have similarity in their sparsity structures. However, the existing methods do not have a mechanism to promote the similarity in sparsity structures in integrative analysis. In this study, we consider penalized variable selection and estimation in integrative analysis. We develop an L_0-penalty based method, which explicitly promotes the similarity in sparsity structures. Computationally it is realized using a coordinate descent algorithm. Theoretically it has the selection and estimation consistency properties. Under a wide spectrum of simulation scenarios, it has identification and estimation performance comparable to or better than the alternatives. In the analysis of three lung cancer datasets with gene expression measurements, it identifies genes with sound biological implications and satisfactory prediction performance.

Keywords: integrative analysis; sparsity structure; variable selection; L_0 penalization; cancer genomic data.

[1]Yuan Huang is a postdoctoral associate and Shuangge Ma is an associate professor at the Department of Biostatistics, Yale University (New Haven, CT 06520) and VA Cooperative Studies Program Coordinating Center (West Haven, CT 06516). Email: yuan.huang@yale.edu and shuangge.ma@yale.edu. Qingzhao Zhang is a postdoctoral associate at the Department of Biostatistics, Yale University (New Haven, CT 06520). Email: qingzhao.zhang@yale.edu. Sanguo Zhang is a professor at the School of Mathematical Science, University of Chinese Academy of Sciences (Beijing, China). Email: sgzhang@ucas.ac.cn. Jian Huang is a professor at the Department of Statistics and Actuarial Science, University of Iowa (Iowa City, IA 52242). Email: jian-huang@uiowa.edu. This work was supported by CA142774 and CA016359 from NIH, 13CTJ001 and 13&ZD148 from National Social Science Foundation of China, and the VA Cooperative Studies Program of the Department of Veterans Affairs, Office of Research and Development. **equal contributions from these authors*

1 Introduction

Data with high-dimensional covariates and small sample sizes are now routinely encountered. A large number of statistical methods and theories have been developed for the analysis of such data. Despite tremendous successes, a problem commonly encountered in practice is that the results from single-dataset analysis are often unsatisfactory with the estimation and identification results having low reliability and poor reproducibility (Zhao et al., 2015). Among the possible causes, the most important is likely to be the small sample sizes of individual studies (Zhao et al., 2015; Tseng et al., 2015). When there are multiple datasets from independent comparable studies, multi-datasets analysis can pool information, increase sample size, and outperform single-dataset analysis (Guerra and Goldstein, 2009). In multi-datasets analysis, integrative analysis, which jointly analyzes the raw data from multiple datasets, can be more effective than several alternatives including classic meta-analysis, which analyzes each dataset separately and then pools summary statistics (Liu et al., 2014a). For comprehensive reviews on integrative analysis and other multi-datasets methods, seeTseng et al. (2015) and references therein.

With high-dimensional covariates, variable selection is usually needed along with estimation. Two cases have been considered in integrative analysis (Liu et al., 2014a). The first is the homogeneity case, under which the models in multiple datasets share the same set of important covariates, that is, they have the same *sparsity structure*. Here only one-dimensional variable selection is needed. As suggested in Liu et al. (2014a) and others, often the homogeneity case is too restricted. As an alternative, under the heterogeneity case, the models not necessarily have identical sets of important covariates. That is, the sparsity structures may differ across datasets. The heterogeneity case includes the homogeneity case as a special example and is more flexible. With the heterogeneity in sparsity structure, two-dimensional variable selection is needed (Zhao et al., 2015). The aforementioned integrative analysis studies have been mostly focused on biomedical data. In the field of machine learning, the problem of jointly estimating models using multiple datasets has also

been studied and referred to as "multi-task learning (MTL)" (Argyriou et al., 2008; Lounici et al., 2009; Yuan et al., 2012).

In integrative analysis, although it is important to be flexible and allow for different sparsity structures across datasets, under many scenarios, it is desirable to *promote* their similarity. As the first family of examples, consider data from independent studies on the same response variable and the same set of covariates. A representative example is described in Section 5. Here with the differences in experimental setups and heterogeneity in samples, a covariate can be important in some datasets but not others. However it is still reasonable to expect that the sets of identified important covariates are similar across datasets to a large extent. As the second example, consider the integrative analysis of genetic data on different cancer types (Liu et al., 2014b). Here the similarity in sparsity structures correspond to genes associated with multiple cancer types, which represent the more essential features of cancer and can be of more interest than type-specific cancer genes.

Under the above scenarios and those alike, promoting the similarity in sparsity structures can potentially improve analysis. However, the existing integrative analysis and MTL methods lack an explicit mechanism to do so. To fix ideas, in Figure 1, we consider the heterogeneity case and its extreme, the homogeneity case. In each panel, a column represents a dataset (M1, M2, or M3), a row represents a covariate (Covariate 1, 2, ...), and a shaded rectangle represents a true or identified important covariate. Under the homogeneity case, the three datasets share the same six important covariates. Under the heterogeneity case, the three datasets share three common important covariates, and each also has three dataset-specific important covariates. The goal is to promote the identification of important covariates shared by multiple datasets (which correspond to the similarity in sparsity structures), while allowing for dataset-specific important covariates (difference in sparsity structures). Figure 1 shows that the proposed method can achieve such a goal better than the competing composite MCP. More details follow in Section 4.

In the literature, there are a few methods that also investigate the "interconnections" across

models and datasets. The most relevant is perhaps the contrasted penalization (Shi et al., 2014), which also conducts integrative analysis and applies penalty to smooth over the regression coefficients of the same covariates in multiple datasets. The contrasted penalization and other smoothing methods are concerned with the *magnitudes* of regression coefficients. Even when multiple studies measure the same response variable and covariates, as for example in Section 5, it is difficult to achieve full comparability of measurements across datasets (Guerra and Goldstein, 2009). When for example different studies are on different disease types as in Liu et al. (2014b), it is not sensible to compare the magnitudes of regression coefficients across datasets. Under these scenarios, it is appropriate to directly work with the sparsity structures but not the magnitudes of regression coefficients.

Significantly advancing from the existing ones, this study will directly address the similarity in sparsity structures in integrative analysis. The proposed method has an intuitive formulation and solid statistical basis, can be effectively realized, and numerically outperforms the alternatives. In what follows, we describe the data and model settings in Section 2. The proposed method and its computational algorithm and statistical properties are described in Section 3. Numerical studies, including simulation in Section 4 and data analysis in Section 5, demonstrate its satisfactory performance. The article concludes with discussion in Section 6. Additional technical details and numerical results are provided in a Supplementary File.

2 Data and Model Settings

Consider the integrative analysis of M independent datasets. In dataset $m(=1,\ldots,M)$ with n_m iid samples, denote $\boldsymbol{y}^m = (y_1^m, \cdots, y_{n_m}^m)^\top$ as the response vector and $X^m \in R^{n_m \times p_m}$ as the covariate matrix. Assume that the M datasets measure the same set of covariates. In practice, partially matched covariate sets can be easily accommodated using a rescaling approach (Liu et al., 2014a).

Thus we have $p_1 = \cdots = p_M = p$. Consider the regression models

$$\boldsymbol{y}^m = X^m \boldsymbol{\beta}^m + \boldsymbol{\epsilon}^m, \ m = 1, \cdots, M, \tag{1}$$

where $\boldsymbol{\beta}^m$ is the p-vector of regression coefficients, and $\boldsymbol{\epsilon}^m$ is the vector of random errors. Here we use linear regression for describing the proposed method. Under more generic models $\boldsymbol{y}^m \sim \phi(X^m \boldsymbol{\beta}^m)$, the proposed method can still be applicable. Assume that X^m's are normalized as $||X^{m,j}||_2 = \sqrt{n_m}$, where $X^{m,j}$ is the jth column of X^m, and $||\cdot||_q$ is the L_q norm.

Let β_j^m denote the jth component of $\boldsymbol{\beta}^m$ and $\boldsymbol{\beta}_j = (\beta_j^1, \cdots, \beta_j^M)^\top$ denote the coefficients of the jth covariate in all M datasets. Denote $\boldsymbol{\beta} = (\boldsymbol{\beta}^1, \cdots, \boldsymbol{\beta}^M) = (\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_p)^\top$ as the $p \times M$ matrix of regression coefficients. Also write $\boldsymbol{\beta} = (\beta_{jm})_{p \times M}$ with its true value $\boldsymbol{\beta}^*$, where $\beta_{jm} = \beta_j^m$. In integrative analysis with differences in for example experimental setups across datasets, β_j^m's are not equal across m. The sparsity structure of model m is measured with $\{I(\beta_j^m = 0), j = 1, \ldots, p\}$, that is, which covariates are important with nonzero coefficients. Under the homogeneity case, $I(\beta_j^m = 0) = I(\beta_j^k = 0)$ for all (m, k, j)'s. Under the heterogeneity case, it is possible that $I(\beta_j^m = 0) \neq I(\beta_j^k = 0)$ for some (m, k, j)'s.

3 Methods

We adopt penalization for selection and regularized estimation. For a review of the existing penalized integrative analysis methods, see Zhao et al. (2015) and others.

Denote $pen(\boldsymbol{\beta})$ as the penalty function. Consider the objective function

$$L(\boldsymbol{\beta}) = \sum_{m=1}^{M} \frac{1}{2n_m} ||\boldsymbol{y}^m - X^m \boldsymbol{\beta}^m||_2^2 + pen(\boldsymbol{\beta}).$$

Denote $\hat{\boldsymbol{\beta}}$ as its minimizer. A nonzero component of $\hat{\boldsymbol{\beta}}$ suggests an association between the corresponding covariate and response.

Liu et al. (2014a) proposes using the composite MCP (minimax concave penalty) under the heterogeneity case. Composite penalization provides a way of two-dimensional selection, with an alternative being sparse group penalization. The composite MCP is built on the MCP, which has been shown to have superior statistical and numerical properties. It takes the form

$$\sum_{j=1}^{p}\rho_1\left(\sum_{m=1}^{M}\rho_1(|\beta_j^m|;\lambda_1,a);0.5M\lambda_1^2,b\right). \tag{2}$$

Here $\rho_1(\cdot)$ is the MCP penalty (Zhang, 2010) defined by $\rho_1(t;\lambda,\gamma)=\lambda\int_0^{|t|}\left(1-\frac{x}{\lambda\gamma}\right)_+dx$ and its derivative $\dot\rho_1(t;\lambda,\gamma)=\lambda\,\mathrm{sgn}(t)\left(1-\frac{|t|}{\lambda\gamma}\right)_+$. a and b are the regularization parameters, and λ_1 is the tuning parameter.

We first consider the following penalty, which will be imposed in addition to (2):

$$M\sum_{i=j}^{p}\rho_2(\|\boldsymbol{\beta}_j\|_1;\lambda_2)-\sum_{m=1}^{M}\sum_{j=1}^{p}\rho_2\left(|\beta_j^m|;\lambda_2\right), \tag{3}$$

where $\rho_2(t;\lambda)=\lambda I(t\neq 0)$. Our strategy is that penalty (2) conducts the main variable selection, whereas penalty (3) accounts for the secondary model similarity structure. The proposed penalty has a form significantly different from the existing ones. Its first term counts how many rows of $\boldsymbol{\beta}$ have nonzero norms, that is, how many covariates are identified in at least one dataset. The second term counts how many individual components of $\boldsymbol{\beta}$ are nonzero. Penalty (3) is minimized if a nonzero $\|\boldsymbol{\beta}_j\|_1$ corresponds to $\beta_j^m\neq 0$ for all m, and is maximized if a nonzero $\|\boldsymbol{\beta}_j\|_1$ corresponds to only one nonzero β_j^m. Thus it can encourage the similarity of $\{I(\beta_j^m=0), j=1,\ldots,p\}$ across m. The tuning parameter λ_2 controls the degree of similarity. When $\lambda_2=0$, the proposed method reduces to the composite MCP. When $\lambda_2=\infty$, it is forced that $I(\beta_j^m=0)=I(\beta_j^k=0)$ for all (m,k,j)'s, i.e., the homogeneity case. The proposed penalty is also related to the Jaccard index of similarity (Tan et al., 2005), which is a ratio and computationally prohibitive.

A penalty involving the L_0 norm is difficulty to optimize. For computational feasibility, we replace ρ_2 in (3) with the SELO (seamless-L_0) penalty (Dicker et al., 2012), which is defined by

$$\rho_2(t;\lambda_2,\tau)=\frac{\lambda_2}{\log(2)}\log\left(\frac{|t|}{|t|+\tau}+1\right),\quad \dot\rho_2(t;\lambda_2,\tau)=\frac{\lambda_2}{\log(2)}\frac{\mathrm{sgn}(t)\,\tau}{(|t|+\tau)(2|t|+\tau)}, \tag{4}$$

where τ is a small positive constant. Dicker et al. (2012) shows that the SELO can have the same asymptotic properties as the L_0 penalty while being computationally more feasible. Overall, the proposed penalized objective function is

$$L(\boldsymbol{\beta}) = \sum_{m=1}^{M} \frac{1}{2n_m} \|\boldsymbol{y}^m - X^m \boldsymbol{\beta}^m\|_2^2 + \sum_{j=1}^{p} \rho_1 \left(\sum_{m=1}^{M} \rho_1(|\beta_j^m|; \lambda_1, a); 0.5M\lambda_1^2, b \right) + \sum_{j=1}^{p} \left\{ M\rho_2(\|\boldsymbol{\beta}_j\|_1; \lambda_2, \tau) - \sum_{m=1}^{M} \rho_2(|\beta_j^m|; \lambda_2, \tau) \right\}. \quad (5)$$

3.1 Computation

Optimizing (5) is realized using a coordinate descent (CD) approach, which contains an outer loop for $\boldsymbol{\beta}_j$'s and an inner loop for β_j^m's. Denote $\hat{\boldsymbol{\beta}}_j^{(K)}$ as the estimate of $\boldsymbol{\beta}_j$ at the Kth loop. The CD algorithm proceeds as follows:

1. Initialize $K = 0$ and $\hat{\boldsymbol{\beta}}_j^{(K)} = \mathbf{0}$ for all j.

2. $K = K + 1$. For $j = 1, \ldots, p$, minimize $M(\boldsymbol{\beta}_j)$ with respect to $\boldsymbol{\beta}_j$ where

$$\begin{aligned} M(\boldsymbol{\beta}_j) &= T_{1,j} + T_{2,j} + T_{3,j}, \\ T_{1,j} &= \sum_{m=1}^{M} \frac{1}{2n_m} \|\boldsymbol{y}^m - \sum_{k<j} X^{m,k} \hat{\beta}_k^{m(K)} - \sum_{k>j} X^{m,k} \hat{\beta}_k^{m(K-1)} - X^{m,j} \beta_j^m\|_2^2, \\ T_{2,j} &= \rho_1 \left(\sum_{m=1}^{M} \rho_1(|\beta_j^m|; \lambda_1, a); 0.5M\lambda_1^2, b \right), \\ T_{3,j} &= M\rho_2(\|\boldsymbol{\beta}_j\|_1; \lambda_2, \tau) - \sum_{m=1}^{M} \rho_2 \left(|\beta_j^m|; \lambda_2, \tau \right). \end{aligned}$$

This can be achieved as follows. For $m = 1, \ldots, M$,

(a) Calculate

$$\begin{aligned} \boldsymbol{r}^m &= \boldsymbol{y}^m - \sum_{k<j} X^{m,k}\hat{\beta}_k^{m(K)} - X^{m,j}\hat{\beta}_j^{m(K-1)} - \sum_{k>j} X^{m,k}\hat{\beta}_k^{m(K-1)}, \\ T_{2,j}^m &= \dot{\rho}_1\left(\sum_{m=1}^{M}\rho_1\left(|\hat{\beta}_j^{m(K-1)}|;\lambda_1,a\right);0.5M\lambda_1^2,b\right)\times\dot{\rho}_1(|\hat{\beta}_j^{m(K-1)}|;\lambda,a), \\ T_{3,j}^m &= M\dot{\rho}_2\left(||\hat{\boldsymbol{\beta}}_j||_1^{(K-1)};\lambda_2,\tau\right) - \dot{\rho}_2\left(|\hat{\beta}_j^{m(K-1)}|;\lambda_2,\tau\right). \end{aligned}$$

(b) Update the estimate of β_j^m as

$$\hat{\beta}_j^{m(K)} \leftarrow S\left(\frac{1}{n}(X^{m,j})^\top \boldsymbol{r}^m + \hat{\beta}_j^{m(K-1)}, (T_{2,j}^m + T_{3,j}^m)\right), \tag{6}$$

where $S(z,\eta) = \text{sgn}(z)(|z|-\eta)_+$ is the soft-thresholding operator.

3. Repeat Step 2 until convergence. In numerical study, $||\hat{\boldsymbol{\beta}}^{(K)} - \hat{\boldsymbol{\beta}}^{(K-1)}||_2 \le 10^{-3}$ is used as the convergence criterion.

This algorithm uses **0** as the initial value. A "hot start" may reduce computational cost. Although seemingly complicated, the proposed algorithm is computationally affordable. Using code written in R, the analysis of one replicate (generated under the scenario described in Table 1, $M = 3$, $n_m = 100$, and $p = 1000$; 240 tuning parameter values) takes about five minutes on a regular laptop. Our numerical experiment suggests that the computer time increases almost linearly with p and slower than linearly with n and the number of tuning values (partial results are provided in the Supplementary File S2.1). For all of our simulated and practical datasets, convergence is achieved within a small number of iterations.

Tuning parameters The composite MCP involves three regularization/tuning parameters. For a and b, Breheny and Huang (2009) suggests setting them connected in a manner such that the group level penalty attains its maximum if and only if all of its individual components are at the maximums, that is, $a = b$. Following published studies (Liu et al., 2014a), we set $a = b = 6$. In the SELO penalty, τ=0.005, which follows the suggestion of setting it as a small positive number

(Dicker et al., 2012). For the values of a, b, and τ, we have conducted simulations and found that the results are not very sensitive to their values (details are provided in the Supplementary File S2.2). However, to be cautious, one may still need to examine other $a(b)$ and τ values in practice. (λ_1, λ_2) are chosen using a BIC criterion with model size for the degrees of freedom (Zhang et al., 2010).

Parameter path We simulate one replicate under the setting of "unstructured auto-regressive correlation with $\rho = 0.7$ + unclustered important covariate effects with nonzero coefficients $\sim$ unif(0.2, 1)" (see Section 4 for details). We analyze using the proposed method and composite MCP and plot the parameter paths for one dataset in Figure 9 (Supplementary File S2.3). Overall the parameter paths of the proposed method are similar to those of other penalized estimates. Compared to the composite MCP, they are "less smoother". For this simulated dataset, when the tunings are properly chosen, the proposed method can correctly identify the true positives, while the composite MCP fails to do so.

3.2 Statistical properties

The proposed penalty differs significantly from the existing ones, and the existing results and techniques are not directly applicable in establishing the statistical properties. Additional complexity is also brought by the heterogeneity across datasets.

The important covariate index sets of the M datasets are respectively labeled as $S_1, \cdots, S_M$. Let $S = \bigcup_{m=1}^{M} S_m$ denote the overall important set. Let S^c and $|S|$ denote the complement and cardinality of S, respectively. Let $\mathcal{A} = \{(i, m) : \beta_i^{m*} \neq 0\}$ and $\boldsymbol{\beta}_{\mathcal{A}}$ denote the components of $\boldsymbol{\beta}$ indexed by $\mathcal{A}$. For a $p \times 1$ vector v and index set $I \subset \{1, \cdots, p\}$, let v_I denote the components of v indexed by I. Moreover, let $X^{m,I}$ denote the $n_m \times |I|$ submatrix of X^m formed by columns with indices in I. Denote the total sample size as $n = \sum_{m=1}^{M} n_m$.

Denote $\hat{\boldsymbol{\beta}}_{\mathcal{A}} = (\hat{\boldsymbol{\beta}}^1_{S_1}, \cdots, \hat{\boldsymbol{\beta}}^M_{S_M})$ as the minimizer of

$$Q(\boldsymbol{\beta}_{\mathcal{A}}) = \sum_{m=1}^{M} \frac{1}{2n_m} \|\boldsymbol{y}^m - X^{m,S_m}\boldsymbol{\beta}^m_{S_m}\|^2 + M \sum_{i \in S} \rho_2(\|\boldsymbol{\beta}_i\|_1; \lambda_2, \tau) - \sum_{(i,m)\in\mathcal{A}} \rho_2(|\beta^m_i|; \lambda_2, \tau), \tag{7}$$

which is the oracle counterpart of objective function (5).

Define $\overline{c}^m = \lambda_{\max}\{n_m^{-1} X^{m,S_m\top} X^{m,S_m}\}$ and $\underline{c}^m = \lambda_{\min}\{n_m^{-1} X^{m,S_m\top} X^{m,S_m}\}$. The following conditions are assumed:

Condition 1. The n_m components of $\boldsymbol{\epsilon}^m$ are i.i.d. and sub-Gaussian with noise level σ_m. That is, for all vector $\boldsymbol{\nu}$ with $\|\boldsymbol{\nu}\|_2 = 1$ and any $t \geq 0$, $P(|\boldsymbol{\nu}^\top \boldsymbol{\epsilon}^m| \geq t) \leq 2\exp\left(-\frac{t^2}{2\sigma_m^2}\right)$.

Condition 2. Let $\delta_m = \min_i |\beta^{m*}_i|$. Then $\delta_m \sqrt{n_m/|S_m|} \to \infty$.

Condition 3. $\max_m \left\{ \frac{\lambda_2 \tau}{\delta_m^3 \underline{c}^m} \sqrt{\frac{n_m}{|S_m|}} \right\} \to 0$.

The sub-Gaussian condition is common in the field of high-dimensional variable selection. See Huang et al. (2011); Fan and Lv (2011) and others. Condition 2 provides a lower bound on the size of the smallest signal. The lower bound is allowed to vanish asymptotically, given that it does not do so faster than $\sqrt{n_m/|S_m|}$. Similar conditions can be found in Fan and Peng (2004); Dicker et al. (2012) and others. Condition 3 restricts the rate of the tuning parameter.

Lemma 1. *Under Conditions 1-3, there exists a local minimizer $\hat{\boldsymbol{\beta}}_{\mathcal{A}}$ such that for any constant $C > 0$,*

$$\Pr\left\{ \|\hat{\boldsymbol{\beta}}^m_{S_m} - \boldsymbol{\beta}^{m*}_{S_m}\|_2 \leq \left(\frac{\lambda_2 \tau}{\delta_m^2 \underline{c}^m} + C \right) \sqrt{\frac{|S_m|}{n_m}}, m = 1, \cdots, M \right\} \geq 1 - \eta_1,$$

with $\eta_1 = \sum_{m=1}^{M} \exp\left(-\frac{|S_m|(\lambda_2\tau + C\delta_m^2 \underline{c}^m)^2}{32 \overline{c}^m \sigma_m^2 \delta_m^2} \right)$.

Proof is provided in the Supplementary File S1. This lemma establishes estimation consistency when the true sparsity structures are known.

Remark 1. *If* $\lambda_2 = 0$, $\|\hat{\boldsymbol{\beta}}^m_{S_m} - \boldsymbol{\beta}^{m*}_{S_m}\|_2 = O_p(\sqrt{|S_m|/n_m})$. *This coincides with the results in He and Shao (2000).*

Consider the oracle estimator $\hat{\boldsymbol{\beta}}^o$ with $\hat{\boldsymbol{\beta}}^o_{\mathcal{A}} = \hat{\boldsymbol{\beta}}_{\mathcal{A}}$ and $\hat{\boldsymbol{\beta}}^o_{\mathcal{A}^c} = 0$. The theorem below provides sufficient conditions to ensure that $\hat{\boldsymbol{\beta}}^o$ is a local minimizer of $L(\boldsymbol{\beta})$ with a high probability. Define

$$\psi_1^m = \left\| X^{m,S-S_m\top} X^{m,S_m} \left(X^{m,S_m\top} X^{m,S_m} \right)^{-1} \right\|_\infty, \quad \psi_2^m = \left\| X^{m,S^c\top} X^{m,S_m} \left(X^{m,S_m\top} X^{m,S_m} \right)^{-1} \right\|_\infty.$$

The following additional conditions are assumed.

Condition 4. There exists $\ell \in (0,1)$, such that $\lambda_2 M/(\lambda_1^2\tau) \le \ell$, $\psi_1^m < 2\delta_m^2(1-\ell)/(\tau^2 M\ell)$, and $\psi_2^m < 2\delta_m^2(1+\ell)/(\tau^2\ell)$.

Condition 5. $b \ge a$ and $a^{-1}\lambda_1^{-1}\min \delta_m^2 \to \infty$.

Theorem 1. *Under Conditions 1-5, with probability at least* $1-\epsilon$, $\hat{\boldsymbol{\beta}}^o$ *is a local minimizer of (5), where*

$$\begin{aligned}\epsilon \;=\; & \sum_{m=1}^{M} \exp\left(-\frac{|S_m|[(M-1)\lambda_2\tau + C\delta_m^2 \underline{c}^m]^2}{32\bar{c}^m\sigma_m^2\delta_m^2} \right) + 2|S| \sum_{m=1}^{M} \exp\left\{ -\frac{n_m(\lambda_1^2\tau - \lambda_2 M)^2}{8\sigma_m^2 M^2\tau^2(1+\psi_1^m)^2} \right\} \\ & + 2(p-|S|) \sum_{m=1}^{M} \exp\left\{ -\frac{n_m(\lambda_1^2\tau + \lambda_2(M-1))^2}{8\sigma_m^2\tau^2(1+\psi_2^m)^2} \right\}.\end{aligned}$$

Model parameters can be varied to achieve $\epsilon \to 0$. This theorem establishes the oracle selection and estimation consistency properties of the proposed method.

4 Simulation study

We set $M = 3$, $n_m = 100$, and $p = 1000$. $X^m \sim N_p(0, \Sigma)$, with all diagonal elements of Σ equal to 1. Consider the following correlation structures: (a) structured correlation. Covariates form groups of size ten. Covariates in the same group are correlated, and different groups are uncorrelated;

(b) unstructured correlation. All covariates form one big group; and (c) no correlation (independence). Under the structured and unstructured correlations, consider the (a) auto-regressive correlation, where the correlation coefficient between covariates j and k is $\rho^{|j-k|}$ with $\rho = 0.3$ and 0.7, representing weak and strong correlations, respectively, and (b) banded correlation, where the correlation coefficient between covariates j and k is $\max(0, 1-|j-k|\rho)$ with $\rho = 0.2$ and 0.33. The random errors are independently generated from standard normal.

Each dataset has six important covariates. Thus across the three datasets, there are eighteen important covariates. For the sparsity structures, consider three scenarios: (a) All overlapping, where the three datasets have identical sparsity structure, as exemplified in the left-upper panel of Figure 1; (b) Half overlapping, where the three datasets share three common important covariates, and each also has three dataset-specific important covariates, as exemplified in the left-lower panel of Figure 1; and (c) None overlapping, where there is no important covariate shared by any two datasets. We also consider different scenarios for the positions of important covariates, which may affect performance when covariates are correlated: (a) clustered, where the important covariates are "nearby", and (b) unclustered, where the important covariates are far apart. More details are provided in Table 4 (Supplementary File). For the important covariates, their regression coefficients are (a) set all equal to 0.6, and (b) generated randomly from unif(0.2, 1).

To better gauge performance of the proposed method, we also compare with the competing alternatives. The first is the contrasted gBridge (group bridge) method (Shi et al., 2014), which promotes similarity in the magnitudes of regression coefficients across datasets. Under the present settings, important covariates have similar coefficients in multiple datasets, which favors this method. The second is an MTL method. Under the sparsity condition, many MTL studies have adopted the group Lasso to achieve a common set of features for multiple datasets (Argyriou et al., 2008; Lounici et al., 2009; Liu et al., 2009; Yuan et al., 2012). As the MCP-based group selection outperforms the Lasso-based (Ma et al., 2011), we adopt the group MCP for MTL analysis. The third method for comparison is the composite MCP, which can provide direct insights into the newly

developed penalty.

When evaluating the proposed method and comparing with the alternatives, we are the most interested in variable selection performance, which is measured using the numbers of TP (true positive) and FP (false positive). In addition, we are interested in prediction performance evaluated using PRE defined as $\sum_m(\hat{\boldsymbol{\beta}}^m - \boldsymbol{\beta}^m)^\top\Sigma(\hat{\boldsymbol{\beta}}^m - \boldsymbol{\beta}^m)$ and estimation performance evaluated using EMSE defined as $\sum_m \|\hat{\boldsymbol{\beta}}^m - \boldsymbol{\beta}^m\|_2^2$.

The summary identification results under the auto-regressive correlation are presented in Table 1. Those under the banded correlation and independence are presented in Tables 5 and 6 (Supplementary File). The summary prediction and estimation results are presented in Tables 7, 8, and 9 (Supplementary File).

The general patterns of the performance of different methods are similar across data settings. As a representative example, consider Table 1 and the setting with unstructured correlation and clustered important covariates. Under the all-overlapping scenario, performance of the proposed method is slightly inferior to that of the contrasted gBridge and multi-task but superior to that of the composite MCP. For example with $\rho = 0.3$ and the nonzero coefficients ∼unif(0.2, 1), the (TP, FP) dual is equal to (17.0, 0.0) under the contrasted gBridge, (17.8, 0.0) under the multi-task, (14.7, 0.1) under the composite MCP, and (16.5, 0.0) under the proposed method. Under the half- and none-overlapping scenarios, the proposed method significantly outperforms the alternatives. For example with $\rho = 0.7$ and the nonzero coefficients all equal to 0.6, under the half-overlapping scenario, the (TP, FP) dual is equal to (14.4, 5.9) under the contrasted gBridge, (13.4, 11.2) under the multi-task, (13.0, 0.7) under the composite MCP, and (15.9, 1.4) under the proposed method. Similar observations are made in the prediction and estimation evaluation. Consider for example Table 7, which has the same settings as Table 1. With $\rho = 0.3$ and the nonzero coefficients ∼unif(0.2, 1), under the all-overlapping scenario, the (PRE, EMSE) dual is equal to (0.28, 0.24) under the contrasted gBridge, (0.26, 0.22) under the multi-task, (0.63, 0.52) under the composite MCP, and (0.35, 0.29) under the proposed method. With $\rho = 0.7$ and the nonzero coefficients all

equal to 0.6, under the half-overlapping scenario, the (PRE, EMSE) dual is equal to (2.68, 0.80) under the contrasted gBridge, (3.31, 0.96) under the multi-task, (3.39, 1.02) under the composite MCP, and (1.58, 0.50) under the proposed method. When further examining the results under the half-overlapping scenario, we find that the proposed method has improvement for both common and dataset-specific important covariates. As expected, there is more improvement for common important covariates (details are presented in the Supplementary File S2.5).

The overall observation is that the proposed method is comparable to the alternatives under a few scenarios but significantly outperforms them under many others. Thus it provides a "safe" choice for practical data analysis. It is interesting to observe superior performance of the proposed method under the none-overlapping scenario. This observation is also reasonable. The three simulated datasets, although having no common important covariate, share a large number of covariates with zero effects. The proposed method can also promote the zero effects to be consistent across datasets and thus reduce false positives and improve variable selection.

5 Analysis of lung cancer data

Lung cancer is the leading cause of cancer death for both men and women in the U.S. Non-small-cell lung cancer (NSCLC) is the most common type of lung cancer, constituting approximately 85% of the cases. Gene profiling studies have been widely conducted on lung cancer, searching for markers associated with prognosis. Following Xie et al. (2011), we collect data from three independent studies. The DFCI (Dana-Farber Cancer Institute) study had a total of 78 patients, among whom 35 died during followup. The median followup time was 51 months. The HLM (Moffitt Cancer Center) study had a total of 76 patients, among whom 59 died during followup. The median followup time was 39 months. The MSKCC (Memorial Sloan-Kettering Cancer Center) study had a total of 102 patients, among whom 38 died during followup. The median followup time was 43.5 months. Affymetrix U133 plus 2.0 arrays were used to measure gene expressions. After processing, data on 22,283 probe sets are available for analysis. The quantile-quantile method is applied for normalization. Although the proposed method can analyze all the probes, we conduct

prescreening following Liu et al. (2013), remove noises, and analyze data on 2,000 probes.

The response is survival time, for which we adopt the AFT (accelerated failure time) model, following Liu et al. (2013). Under high-dimensional settings, this model can be preferred with its lucid interpretation and low computational cost. In the Supplementary File S3, we provide details on the estimation procedure. The objective function has a weighted least squares form, and the proposed method and computational algorithm are directly applicable.

With the proposed method, thirteen genes are identified in at least one dataset, among which eleven are identified in all three datasets. The estimation results are presented in Table 2. Literature search suggests that some of the identified genes have important implications. Specially, a few represent the hallmark of cancer, such as signaling and cell adhesion. The protein encoded by gene CSF1 is a cytokine that controls the production, differentiation, and function of macrophages. This gene has been associated with the development of giant cell tumors. Chemokine (C-X-C motif) ligand 3 (CXCL3) is a small cytokine belonging to the CXC chemokine family that is also known as GRO3 oncogene (GRO3). CXCL3 controls migration and adhesion of monocytes. Its down-regulation has been implicated in cancer development. Gene MTO1 encodes a mitochondrial protein involved in mitochondrial tRNA modification. It has been implicated in cancer development (www.genecards.org). The protein encoded by gene RFXANK, along with regulatory factor X-associated protein and regulatory factor-5, forms a complex that binds to the X box motif of certain MHC class II gene promoters and activates their transcription. Major histocompatibility (MHC) class II molecules are transmembrane proteins that have a central role in development and control of the immune system. Gene LGALS8 encodes a member of the galectin family. The galectins have been implicated in many essential functions including development, differentiation, cell-cell adhesion, cell-matrix interaction, growth regulation, apoptosis, and RNA splicing. This gene is widely expressed in tumoral tissues and involved in integrin-like cell interactions. The protein encoded by gene SOCS6 belongs to the cytokine-induced STAT inhibitor (CIS) protein family. Differential expression of this gene has been implicated in the development of leukemia and other cancers. The protein encoded by gene PTPLA belongs to the protein tyrosine phosphatases (PTPs) family. Members of the PTP family are known to be signaling molecules that regulate a variety of cellular processes.

Beyond the proposed, the three alternative methods are also applied. The summary comparison is presented in Table 3. The alternatives identify 28, 35, and 68 genes in at least one dataset. Detailed results are available from the authors. The contrasted gBridge and multi-task methods identify the same sets of genes across datasets. The numbers of identified genes are much larger than that of the proposed, which leads to unfocused hypothesis for downstream study. In addition, the three studies were independently conducted and had considerable differences, and the result of identical gene sets may be too strong (Xie et al., 2011). The composite MCP identifies largely different gene sets across datasets. More specifically, the numbers of identified genes for the three datasets are 35, 18, and 18, respectively, and any two datasets share only one common gene. There is a lack of objective measure which set of identified genes is "biologically more meaning". Prediction evaluation is conducted, which may provide some insights on the identified genes, using a random sampling approach (Liu et al., 2014b). Specifically, each dataset is randomly split into a training and a testing set, with sizes 2:1. Estimates are generated using the training data and used to make prediction for the testing set samples. The logrank statistic is used to evaluate prediction. With 100 random splittings, the average logrank statistics are 7.77 (contrasted gBridge), 6.76 (multi-task), 3.32 (composite MCP), and 10.45 (proposed), respectively. The proposed method has the best prediction performance.

6 Discussion

For data with high-dimensional covariates but small sample sizes, integrative analysis provides an effective way of pooling information and increasing power and outperforms single-dataset and several multi-datasets methods. This study advances from the existing integrative analysis studies by explicitly promoting the similarity in model sparsity structures across multiple datasets. A novel penalization method has been developed, which uses the composite MCP for selection and introduces a new penalty to address the sparsity structures. We rigorously establish that the proposed method has the selection and estimation consistency properties. In simulation, the proposed method has slightly inferior performance when multiple datasets are highly similar but significantly outperforms the alternatives under other scenarios. It provides a safe choice in practice when the degree of similarity of model sparsity structures is unknown. In data analysis, it identifies genes with sound biological interpretations and superior prediction performance.

Smoothing over covariate effects has been pursued in the literature. See for example Huang et al. (2011), the work on fused Lasso, and others. The proposed method significantly advances from them by directly addressing the sparsity structures, which, under many scenarios, is more sensible than smoothing over the magnitudes of regression coefficients. The lack of full comparability across datasets and covariate effects has been well acknowledged (Guerra and Goldstein, 2009; Zhao et al., 2015), and thus the proposed technique can have broad applications. On the other hand, the advancement inevitably brings drawbacks. Compared to some alternatives, the proposed method is computationally more expensive. With the present algorithm and R code, the computer time can pose a hurdle if p is of the order 10^5 or higher. Under the sparsity condition, the number of relevant covariates is usually small. The application of screening and other statistical techniques can significantly reduce the dimensionality so that the proposed method is applicable. In addition, adopting more advanced computing hardware and other programming languages may also reduce computer time. The additional complexity of proposed method also brings challenges to the study of convergence. In all of our numerical analyses, convergence is satisfactorily achieved. However, examining the literature suggests that the existing techniques may not be directly applicable to establishing the convergence properties. Studies such as Dicker et al. (2012) suggest that theoretical convergence study for non-convex minimization is extremely challenging. More fundamental studies may be needed in the future.

References

Argyriou, A., Evgeniou, T., and Pontil, M. (2008), "Convex multi-task feature learning," *Machine Learning*, 73(3), 243–272.

Breheny, P., and Huang, J. (2009), "Penalized methods for bi-level variable selection," *Statistics and its Interface*, 2(3), 369–380.

Dicker, L., Huang, B., and Lin, X. (2012), "Variable selection and estimation with the seamless-l_0 penalty," *Statistica Sinica*, 23, 929–962.

Fan, J., and Lv, J. (2011), "Nonconcave penalized likelihood with NP-dimensionality," *IEEE Transactions on Information Theory*, 57, 5467–5484.

Fan, J., and Peng, H. (2004), "Nonconcave penalized likelihood with a diverging number of parameters," *The Annals of Statistics*, 32, 928–961.

Guerra, R., and Goldstein, D. (2009), *Meta-analysis and combining information in genetics and genomics*, CRC Press.

He, X., and Shao, Q. (2000), "On parameters of increasing dimensions," *Journal of Multivariate Analysis*, 73, 120–135.

Huang, J., Breheny, P., and Ma, S. (2012), "A selective review of group selection in high-dimensional models," *Statistical science: a review journal of the Institute of Mathematical Statistics*, 27(4), 481–499.

Huang, J., Ma, S., Li, H., and Zhang, C.-H. (2011), "The sparse Laplacian shrinkage estimator for high-dimensional regression," *Annals of statistics*, 39(4), 2021–2046.

Liu, J., Huang, J., and Ma, S. (2014a), "Integrative analysis of cancer diagnosis studies with composite penalization," *Scandinavian Journal of Statistics*, 41(1), 87–103.

Liu, J., Huang, J., Zhang, Y., Lan, Q., Rothman, N., Zheng, T., and Ma, S. (2013), "Identification of gene–environment interactions in cancer studies using penalization," *Genomics*, 102(4), 189–194.

Liu, J., Huang, J., Zhang, Y., Lan, Q., Rothman, N., Zheng, T., and Ma, S. (2014b), "Integrative analysis of prognosis data on multiple cancer subtypes," *Biometrics*, 70(3), 480-488.

Liu, J., Ji, S., and Ye, J. (2009), Multi-task feature learning via efficient $l_{2,1}$-norm minimization, in *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, AUAI Press, pp. 339–348.

Lounici, K., Pontil, M., Tsybakov, A. B., and van de Geer, S. (2009), "Taking advantage of sparsity in multi-task learning," *Proceedings of 22nd Annual Conference on Learning Theory, COLT 2009. Lecture Notes in Artificial Intelligence*, Springer.

Ma, S., Huang, J., Wei, F., Xie, Y., and Fang, K. (2011), "Integrative analysis of multiple cancer prognosis studies with gene expression measurements," *Statistics in medicine*, 30(28), 3361–3371.

Shi, X., Liu, J., Huang, J., Zhou, Y., Shia, B., and Ma, S. (2014), "Integrative analysis of high-throughput cancer studies with contrasted penalization," *Genetic Epidemiology*, 38(2), 144–151.

Stute, W. (1996), "Distributional convergence under random censorship when covariables are present," *Scandinavian Journal of Statistics*, 461–471.

Tan, P., Steinbach, M., and Kumar, V. (2005), *Introduction to Data Mining* Addison-Wesley.

Tseng, G., Ghosh, D., and Zhou, J. (2015), *Integrating Omics Data* Cambridge University Press.

Xie, Y., Xiao, G., Coombes, K. R., Behrens, C., Solis, L. M., Raso, G., Girard, L., Erickson, H. S., Roth, J., Heymach, J. V., Moran, C., Danenberg, K., Minna, J. D., and Wistuba, I. I. (2011), "Robust gene expression signature from formalin-fixed paraffin-embedded samples predicts prognosis of non-small-cell lung cancer patients," *Clin Cancer Res*, 17, 5705–5714.

Yuan, X.-T., Liu, X., and Yan, S. (2012), "Visual classification with multitask joint sparse representation," *Image Processing, IEEE Transactions on*, 21(10), 4349–4360.

Zhang, C.-H. (2010), "Nearly unbiased variable selection under minimax concave penalty," *The Annals of Statistics*, 38(2), 894–942.

Zhang, Y., Li, R., and Tsai, C.-L. (2010), "Regularization parameter selections via generalized information criterion," *Journal of the American Statistical Association*, 105(489), 312–323.

Zhang, Q., Zhang, S., Liu, J., Huang, J., and Ma, S. (2015), "Penalized integrative analysis under the accelerated failure time model," *Statistica Sinica, in press*.

Zhao, Q., Shi, X., Huang, J., Liu, J., Li, Y., and Ma, S. (2015), "Integrative analysis of "-omics" data using penalty functions," *Wiley Interdisciplinary Reviews: Computational Statistics*, 7, 99–108.

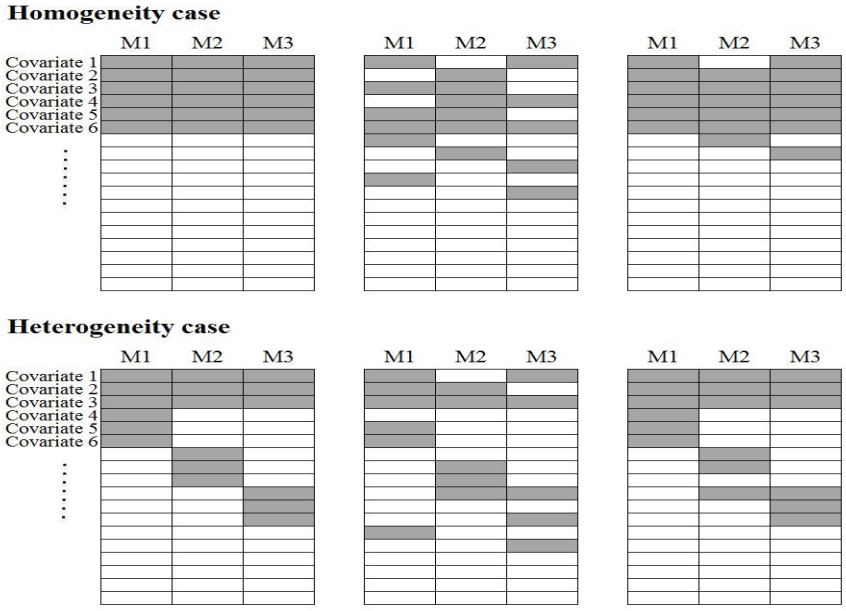

Figure 1: Analysis scheme: shaded rectangles represent the true and identified important covariates. Right: true sparsity structures. Middle: sparsity structures identified by composite MCP. Right: sparsity structures identified by the proposed method.

Table 1: Simulation: summary identification results under the auto-regressive correlation, based on 500 replicates. In each cell, mean(sd).

	contrasted gBridge		multi-task		composite MCP		Proposed	
	TP	FP	TP	FP	TP	FP	TP	FP
Structured correlation								
ρ=0.3, nonzero coefs=0.6								
All	18.0(0.0)	0.0(0.0)	18.0(0.0)	0.0(0.3)	17.2(1.2)	0.7(1.0)	18.0(0.1)	0.0(0.1)
Half	17.2(1.1)	2.6(2.4)	17.4(0.8)	17.2(1.9)	17.0(1.2)	0.8(1.2)	17.7(0.6)	0.4(0.8)
None	16.3(2.0)	6.9(4.1)	16.1(1.9)	36.2(6.5)	17.0(1.1)	1.1(1.5)	17.5(0.7)	0.9(1.0)
ρ=0.3, nonzero coefs~unif(0.2,1)								
All	17.0(1.1)	0.0(0.2)	17.9(0.7)	0.0(0.3)	14.4(1.9)	0.4(0.7)	16.7(1.2)	0.0(0.3)
Half	15.0(1.6)	1.4(1.7)	15.5(1.4)	13.4(2.7)	14.0(1.7)	0.4(1.0)	15.3(1.6)	0.2(0.5)
None	12.6(2.2)	3.7(3.4)	13.1(2.2)	27.7(5.5)	13.8(1.8)	0.4(0.9)	14.6(1.8)	0.4(0.6)
ρ=0.7, nonzero coefs=0.6								
All	17.9(0.3)	0.0(0.1)	17.0(1.9)	0.2(1.3)	12.4(2.9)	0.6(1.3)	17.8(0.5)	0.0(0.1)
Half	14.2(1.5)	1.7(2.3)	13.6(3.0)	12.1(3.4)	12.3(2.6)	0.5(1.0)	16.6(1.4)	0.7(1.4)
None	9.6(1.6)	2.2(2.3)	11.4(2.2)	23.1(4.7)	12.3(2.2)	0.5(0.7)	14.8(1.8)	0.7(1.0)
ρ=0.7, nonzero coefs~unif(0.2,1)								
All	16.3(1.5)	0.0(0.0)	15.8(2.4)	0.1(0.7)	11.6(2.2)	0.4(0.8)	15.5(1.6)	0.0(0.0)
Half	13.6(1.8)	1.8(2.3)	13.5(2.0)	11.8(2.6)	11.6(2.2)	0.4(0.8)	15.5(1.6)	0.0(0.0)
None	9.5(1.5)	2.7(2.4)	10.9(1.6)	22.4(3.9)	11.9(1.6)	0.4(1.0)	13.1(1.7)	0.3(0.6)
Unstructured correlation + Clustered important covariates								
ρ=0.3, nonzero coefs=0.6								
All	18.0(0.0)	0.0(0.1)	18.0(0.0)	0.0(0.0)	17.5(0.7)	0.4(0.7)	18.0(0.0)	0.0(0.3)
Half	17.1(1.0)	2.2(1.9)	17.4(0.8)	17.3(2.0)	17.2(1.0)	0.7(1.0)	17.7(0.5)	0.5(0.8)
None	16.3(1.9)	6.1(4.1)	16.8(1.4)	36.1(4.8)	17.2(0.9)	0.4(0.8)	17.7(0.5)	0.3(0.6)
ρ=0.3, nonzero coefs~unif(0.2,1)								
All	17.0(1.0)	0.0(0.1)	17.8(0.7)	0.0(0.0)	14.7(1.9)	0.1(0.3)	16.5(1.4)	0.0(0.1)
Half	15.1(1.7)	2.0(1.9)	15.4(1.6)	13.2(3.1)	14.2(1.8)	0.2(0.6)	15.4(1.7)	0.2(0.5)
None	12.8(2.2)	3.9(3.2)	13.3(2.2)	27.6(5.2)	14.4(1.8)	0.3(0.9)	15.3(1.4)	0.3(0.7)
ρ=0.7, nonzero coefs=0.6								
All	17.6(0.5)	0.0(0.0)	16.1(2.2)	0.1(0.7)	12.1(2.0)	0.2(0.7)	17.3(1.2)	0.0(0.3)
Half	14.4(1.2)	5.9(2.2)	13.4(2.0)	11.2(2.2)	13.0(2.3)	0.7(1.0)	15.9(1.5)	1.4(1.4)
None	8.4(1.8)	6.9(6.1)	11.2(1.5)	22.6(3.2)	12.7(1.6)	0.4(0.8)	14.2(1.8)	0.8(0.9)
ρ=0.7, nonzero coefs~unif(0.2,1)								
All	16.0(1.7)	0.0(0.0)	14.6(2.7)	0.1(0.4)	11.7(1.7)	0.3(0.7)	15.2(1.7)	0.0(0.1)
Half	13.3(2.1)	6.1(3.1)	12.6(2.0)	10.5(1.9)	11.8(2.1)	0.5(0.9)	13.8(1.9)	0.8(1.2)
None	8.2(1.9)	7.4(6.0)	10.2(1.4)	20.4(2.8)	11.3(2.0)	0.3(0.8)	12.7(1.8)	0.7(1.3)
Unstructured correlation + Unclustered important covariates								
ρ=0.3, nonzero coefs=0.6								
All	18.0(0.0)	0.0(0.0)	18.0(0.0)	0.0(0.3)	15.9(2.1)	3.7(3.3)	18.0(0.3)	0.9(1.7)
Half	16.9(1.4)	3.9(2.3)	15.8(2.0)	17.6(8.1)	15.3(2.4)	3.4(3.3)	16.6(1.5)	2.7(2.4)
None	15.7(2.4)	8.7(3.6)	12.2(4.0)	39.6(26.3)	14.6(2.6)	4.0(3.1)	15.6(2.2)	3.7(3.1)
ρ=0.3, nonzero coefs~unif(0.2,1)								
All	16.3(1.6)	0.1(0.3)	17.8(0.8)	0.1(0.6)	12.5(2.3)	1.5(2.5)	15.9(2.4)	0.4(1.0)
Half	13.5(2.2)	1.6(2.0)	14.3(2.2)	12.5(5.2)	11.8(2.4)	1.6(2.0)	13.2(2.4)	1.1(1.6)
None	11.0(2.5)	4.4(3.6)	10.5(2.5)	27.6(14.0)	11.7(2.1)	1.1(1.3)	12.2(2.2)	1.1(1.2)
ρ=0.7, nonzero coefs=0.6								
All	18.0(0.0)	0.0(0.1)	18.0(0.0)	0.2(0.7)	16.6(1.2)	2.8(2.1)	18.0(0.2)	1.6(1.6)
Half	16.9(1.2)	3.4(2.5)	16.9(1.2)	21.3(4.7)	15.8(1.8)	4.2(2.5)	16.8(1.1)	3.3(2.4)
None	15.4(2.4)	7.9(3.8)	13.8(3.0)	43.8(11.6)	15.5(1.9)	4.4(2.4)	16.2(1.4)	4.2(2.5)
ρ=0.7, nonzero coefs~unif(0.2,1)								
All	16.8(1.2)	0.1(0.3)	17.7(0.9)	0.3(1.1)	13.6(2.2)	1.5(1.4)	16.1(1.6)	1.0(1.2)
Half	14.1(2.0)	2.2(2.3)	14.9(1.9)	15.7(5.8)	13.0(2.2)	1.9(1.7)	13.8(2.2)	1.6(1.7)
None	11.7(2.5)	5.3(4.2)	11.9(2.6)	34.6(11.2)	12.8(2.2)	2.2(1.8)	13.3(2.1)	2.2(1.9)

Table 2: Analysis of the lung cancer data: genes identified using the proposed method.

Gene	Data 1	Data 2	Data 3
PEX11B	0.0223	0.0154	-0.0166
GABRG2	0.0083	–	0.0263
CSF1	0.0014	-0.0003	0.0366
ARL4D	-0.0122	-0.0344	-0.0135
LMBRD1	0.0009	0.0324	-0.0035
CXCL3	-0.0453	–	0.0038
MTO1	0.0259	0.0028	-0.0508
CCDC64	-0.0152	0.0291	0.0133
RFXANK	-0.0479	0.0217	-0.0208
PSMB7	-0.0112	-0.0230	0.0347
LGALS8	0.0103	0.0205	0.0070
SOCS6	-0.0274	-0.0202	-0.0087
PTPLA	-0.0686	-0.0126	-0.0111

Table 3: Analysis of the lung cancer data: numbers of overlapped genes.

	contrasted gBridge	multi-task	composite MCP	proposed
contrasted gBridge	28	12	24	4
multi-task		35	18	4
composite MCP			68	8
proposed				13