SCIENTIFIC REPORTS
natureresearch

OPEN

Meta-Analysis Based on Nonconvex Regularization

Hui Zhang[1], Shou-Jiang Li[1], Hai Zhang[1,2], Zi-Yi Yang[1], Yan-Qiong Ren[1], Liang-Yong Xia[1] & Yong Liang[1*]

The widespread applications of high-throughput sequencing technology have produced a large number of publicly available gene expression datasets. However, due to the gene expression datasets have the characteristics of small sample size, high dimensionality and high noise, the application of biostatistics and machine learning methods to analyze gene expression data is a challenging task, such as the low reproducibility of important biomarkers in different studies. Meta-analysis is an effective approach to deal with these problems, but the current methods have some limitations. In this paper, we propose the meta-analysis based on three nonconvex regularization methods, which are $L_{1/2}$ regularization (meta-Half), Minimax Concave Penalty regularization (meta-MCP) and Smoothly Clipped Absolute Deviation regularization (meta-SCAD). The three nonconvex regularization methods are effective approaches for variable selection developed in recent years. Through the hierarchical decomposition of coefficients, our methods not only maintain the flexibility of variable selection and improve the efficiency of selecting important biomarkers, but also summarize and synthesize scientific evidence from multiple studies to consider the relationship between different datasets. We give the efficient algorithms and the theoretical property for our methods. Furthermore, we apply our methods to the simulation data and three publicly available lung cancer gene expression datasets, and compare the performance with state-of-the-art methods. Our methods have good performance in simulation studies, and the analysis results on the three publicly available lung cancer gene expression datasets are clinically meaningful. Our methods can also be extended to other areas where datasets are heterogeneous.

With the rapid development of biotechnology and its wide applications, many database repositories of high-throughput gene expression data have been created and published. For example, Gene Expression Omnibus (GEO) currently has stored more than 2.76 million samples over 105,000 studies[1]. The gene expression datasets have been widely used in the prediction and diagnosis of diseases, and their application prospects are increasingly promising.

It is desirable to consider variable selection into the analysis of gene expression data due to its small sample size and high dimensionality. Variable selection not only enhances generalization by reducing overfitting, but also enhances interpretability by simplifying the model, i.e., identifying important biomarkers associated with the disease and helping to find the best solution for patients in the treatment process. For a single dataset, there exist many variable selection methods, such as Least Absolute Shrinkage and Selection Operator (LASSO)[2], $L_{1/2}$ regularization[3–5], Minimax Concave Penalty (MCP)[6], Smoothly Clipped Absolute Deviation (SCAD)[7–9], Group LASSO[10], elastic net[11], Hard Ridge[12], SCAD-L_2[13], Complex Harmonic Regularization (CHR) penalty[14] and so on. These methods are effective in discovering important biomarkers in a single dataset. However, it is well known that the analysis of gene expression data is still a challenging task due to high noise and low reproducibility of important biomarkers. There are two main reasons for this challenging task. One is that the decisive biomarkers that regulate the phenotypes are usually very sparse compared to the total number of biomarkers in the entire genome, and their effects are usually weak, therefore, the results of individual studies are not remarkable and difficult to reproduce. The other is that the different experimental datasets may come from inconsistent experimental conditions, sample preparation methods, measurement sensitivities or precision, and also from different study groups, biological sample selections. Therefore, the important genes in some studies may be not remarkable in other studies, which we call the data have the heterogeneity. The data heterogeneity reveals the complexity of gene expression data and significantly obstructs gene expression technology in clinical applications.

[1]Faculty of Information Technology & State Key Laboratory of Quality Research in Chinese Medicines, Macau University of Science and Technology, Taipa, 999078, Macau. [2]School of Mathematics, Northwest University, 710127, Xi'an, China. *email: yliang@must.edu.mo

Since many genomic databases are publicly available, meta-analysis is an effective approach to address the heterogeneity among different datasets and make full use of different datasets. Meta-analysis is a significant technique for clinical diagnosis, which plays an important role in summarizing and synthesizing scientific evidence from multiple studies. Classic meta-analysis methods, which aggregate the summary statistics from individual datasets to obtain total scores and then evaluate them based on statistical significance of all studies, including p values[15], ranks[16,17], effect sizes[18–21]. Li and Tseng[22] apply Fisher's method combining p values by summation of log-transformed p values, and the method increases the biological interpretation of meta-analysis results. Similar strategies can be applied to combine effect sizes of Random Effects Model (REM) or Fixed Effects Model (FEM) from individual studies. A comprehensive review of these methods is given in the researches[23–25]. These methods perform well in identifying differentially expressed genes, but they ignore the correlations between the covariates (genes). There are some approaches that attempt to model the preprocessed microarray datasets using latent variable-based models[26–30]. In general, latent variables are not observable in the data, but can be inferred from other observed variables. Huo *et al.*[31] use latent variable to quantify homogeneous and heterogeneous differentially expressed signals across studies to detect genes that are differentially expressed in only a subset of the combined studies. Rashid *et al.*[32] utilize a penalized Generalized Linear Mixed Model based on latent variable to select gene signatures and address between-study heterogeneity. These methods provide the potential to pool information across genes, making it possible to more clearly infer which genes are differentially expressed. Compared with the previous classical meta-analysis methods, these methods are more complex, which limit their application in practice. Recently, Zhang *et al.*[33] set different constant terms for multiple studies in the logistic regression model to measure the heterogeneity of the samples. This method assumes that the same variables in multiple studies should make the same contribution to their corresponding responses. In other words, this method conducts variable selection in an 'all-in-or-all-out' fashion. In this paper, we consider that some important genes in some studies are likely to be ineffective in other studies, and it is important to allow such flexibility.

Some researchers propose the bi-level selection methods which consider the coefficients of each variable (gene) from all datasets as a group, and simultaneously shrink these groups and the variables within these groups by the penalty function to study the correlation between variables and identify important genes. Existing bi-level selection methods include composite MCP[34], group Bridge[35] and group exponential LASSO[36], meta-SVM[37] etc. These methods of the aforementioned references generally treat the coefficients of one gene from different datasets as a group, and conduct two levels selection. The first is to determine whether a particular gene is related to the response variable in all datasets, and the second is to determine which dataset contains the identified gene related to the response variable. These methods consider both the heterogeneity and the correlation between the datasets. However, for M independent datasets $\{(\boldsymbol{X}_m, \boldsymbol{y}_m)\}_{m=1}^{M}$, each of which contains n_m samples and p variables, these methods consider to solve the problem which has the $\sum_{m=1}^{M} n_m \times Mp$ dimensional measurement matrix $\widetilde{\boldsymbol{X}} = diag(\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_M)$, the $\sum_{m=1}^{M} n_m$ dimensional response $\widetilde{\boldsymbol{y}} = (\boldsymbol{y}_1^T, \boldsymbol{y}_2^T, \cdots, \boldsymbol{y}_M^T)^T$ and the Mp dimensional unknown coefficients $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \cdots, \boldsymbol{\beta}_M^T)^T$, where the superscript T represents the transpose of the vector. Since the gene expression data has the characteristics of small sample size and high-dimensional, these methods greatly increase the variable dimension, so it may increase the difficulty of solving the problem.

Zhou and Zhu[38] propose a new group variable selection method "hierarchical LASSO" that can be used for gene-set selection. The hierarchical LASSO not only removes unimportant groups effectively, but also maintains the flexibility of selecting variables within the group. They also showed that the new method offers the potential for achieving the theoretical "oracle" property. Li *et al.*[39] propose meta-LASSO for variable selection with high-dimensional meta-analyzed data. The meta-LASSO not only improves the ability to identify important genes with the strength of multiple datasets, but also maintains the flexibility of selection between datasets to consider the data heterogeneity.

For many practical applications, LASSO often cannot find the most sparse solutions (this is extremely important for model selection), and it is inefficient when the errors in data have heavy tail distribution[2]. Zhao and Yu[40] give the Strong Irrepresentable Condition for the model selection consistency of LASSO, and show that to induce sparsity, LASSO shrinks the estimates for the nonzero coefficients too heavily. When Strong Irrepresentable Condition fails, the irrelevant covariates are correlated with the relevant covariates enough to be picked up by LASSO to compensate the over-shrinkage of the nonzero parameters. Therefore, to get universal consistency, some nonconvex regularization methods have been proposed in recent years, such as $L_{1/2}$ penalty, Minimax Concave Penalty (MCP) and Smoothly Clipped Absolute Deviation (SCAD) penalty etc. These methods achieve both selection consistency and nearly unbiasedness, which make them widely applied in signal/image processing, statistics and machine learning, such as biological feature selection[14,41–44], compressed sensing and low rank matrix completion[8,45–48], sparse signals separation and image inpainting[49,50], and dictionary learning[51] etc.

In this paper, we propose the meta-analysis based on three nonconvex regularization methods ($L_{1/2}$ regularization, MCP regularization and SCAD regularization), dubbed as meta-Half, meta-MCP and meta-SCAD respectively. Our methods combine the advantages of meta-analysis and the nonconvex regularization methods. We propose the efficient algorithms which apply the nonconvex iterative thresholding algorithms based on approximate message passing (Half-AMP, MCP-AMP and SCAD-AMP)[52,53] to solve our models. Furthermore, we apply our methods to the simulation data and three publicly available lung cancer gene expression datasets, and compare the performance of our methods with other four state-of-the-art methods, which are meta-LASSO, composite MCP, group Bridge and group exponential LASSO. The experiments results show that our methods have favorable performance.

Methodology

In this section, we study the meta-analysis based on the three nonconvex regularization methods ($L_{1/2}$ regularization, MCP regularization and SCAD regularization).

Consider M independent datasets $D = \{(\boldsymbol{X}_m, \boldsymbol{y}_m)\}_{m=1}^{M}$, each of which contains n_m samples. Denote $\boldsymbol{X}_m = (\boldsymbol{x}_{m1}, \boldsymbol{x}_{m2}, \cdots, \boldsymbol{x}_{m,n_m})^T$ and $\boldsymbol{y}_m = (y_{m1}, y_{m2}, \cdots, y_{m,n_m})^T$, where the superscript T represents the transpose of the vector, $\boldsymbol{x}_{mi} = (x_{mi,1}, x_{mi,2}, \cdots, x_{mi,p})^T$ $(i = 1, 2, \cdots, n_m)$ is ith sample in the mth dataset which contains p variables (genes), and y_{mi} is the response variable, in this paper, we consider the response variable is a binary phenotype (for example, if the ith sample of the mth dataset is a disease patient, y_{mi} is 1, and 0 otherwise). The p genes are assumed common in all datasets. We assume the conditional probability that y_{mi} takes value 1 given the gene expression vector $\boldsymbol{x}_{mi}$ follows the logistic regression model

$$\log \frac{Pr(y_{mi} = 1|\boldsymbol{x}_{mi})}{Pr(y_{mi} = 0|\boldsymbol{x}_{mi})} = \beta_{m0} + \boldsymbol{x}_{mi}^T\boldsymbol{\beta}_m,\ i = 1, 2, \cdots, n_m,\ \ m = 1, 2, \cdots, M, \tag{1}$$

where β_{m0} is an intercept and $\boldsymbol{\beta}_m = (\beta_{m1}, \cdots, \beta_{mp})^T$ is the unknown coefficients for the mth data. Due to heterogeneity between datasets, we allow β_{m0} and β_m in (1) to vary with m. We hope to find the true nonzero components of β_m for each dataset.

Compared with the variable selection of single dataset model, the variable selection of the M datasets models are distinguishing and peculiar. On the one hand, each variable has M coefficients, which belong to the same explanatory variable. Therefore, there is some correlation or similarity, which makes it impossible to make coefficient estimation and variable selection separately, otherwise this correlation will be ignored. On the other hand, the significance of variables is not identical, so we cannot simply synthesize estimation. The penalization methods with meta-analysis make full use of this particularity to study data differences. These methods conduct variable selection by maximizing,

$$\sum_{m=1}^{M}\ell_m(\beta_{m0}, \boldsymbol{\beta}_m) - P(\boldsymbol{\beta};\lambda), \tag{2}$$

where $\ell_m(\beta_{m0}, \boldsymbol{\beta}_m)$ is the log-likelihood for the mth dataset and has the following form

$$\ell_m(\beta_{m0}, \boldsymbol{\beta}_m) = \sum_{i=1}^{n_m}[y_{mi}(\beta_{m0} + \boldsymbol{x}_{mi}^T\boldsymbol{\beta}_m) - \log\{1 + \exp(\beta_{m0} + \boldsymbol{x}_{mi}^T\boldsymbol{\beta}_m)\}],$$

P is a penalty function and λ is the regularization parameter that controls the complexity of the machine.

In this paper, we focus on the three nonconvex regularization methods ($L_{1/2}$ regularization, MCP regularization and SCAD regularization), and through the hierarchical decomposition of coefficients that maintain the flexibility of variable selection as well incorporate the relationship between different datasets. We consider the following hierarchical reparameterization:

$$\beta_{mj} = h_j\xi_{mj},\ m = 1, 2, \cdots, M;\ j = 1, 2, \cdots, p. \tag{3}$$

The parameter h_j is the effect of the jth gene, and the different m for ξ_{mj} reflects the different effects of the jth gene among M datasets. If $h_j = 0$, then $\boldsymbol{\beta}_j = (\beta_{1j}, \beta_{2j}, \cdots, \beta_{Mj})^T = \boldsymbol{0}$, this indicates that the jth gene is not significant in all M datasets. If $h_j \neq 0$, then whether the β_{mj} is equal to 0 depends on whether ξ_{mj} is equal to 0. Since the M datasets may have heterogeneity (the M datasets may come from inconsistent experimental conditions, sample preparation methods, measurement sensitivities or precision, and also from different study groups, biological sample selections.), then one gene is important in some datasets may be not remarkable in other datasets. Through ξ_{mj} contral β_{mj} to keep the selection flexibility among M datasets. If the M datasets have no heterogeneity, then $h_j = \beta_{mj}$ for $m = 1, \cdots, M$ defined in (2) and $\xi_{mj} = 1$ for all j and m. With reparameterization (3), we propose a meta-analysis method based on nonconvex regularization. Our method selects important genes by solving

$$\max_{\beta_0, \boldsymbol{h}, \boldsymbol{\xi}_m}\sum_{m=1}^{M}\ell_m(\beta_{m0}, \boldsymbol{h}, \boldsymbol{\xi}_m) - \sum_{j=1}^{p}P\Big(h_j;\lambda_h\Big) - \sum_{j=1}^{p}\sum_{m=1}^{M}P\Big(\xi_{mj};\lambda_\xi\Big), \tag{4}$$

where $\ell_m(\beta_{m0}, \boldsymbol{h}, \boldsymbol{\xi}_m)$ is the likelihood function and has the following form

$$\ell_m(\beta_{m0}, \boldsymbol{h}, \boldsymbol{\xi}_m) = \sum_{i=1}^{n_m}[y_{mi}(\beta_{m0} + \boldsymbol{x}_{mi}^T(\boldsymbol{h}\cdot\boldsymbol{\xi}_m)) - \log\{1 + \exp(\beta_{m0} + \boldsymbol{x}_{mi}^T(\boldsymbol{h}\cdot\boldsymbol{\xi}_m))\}], \tag{5}$$

$\boldsymbol{h} = (h_1, h_2, \cdots, h_p)^T$, $\boldsymbol{\xi}_m = (\xi_{m1}, \xi_{m1}, \cdots, \xi_{mp})^T$, $\boldsymbol{\beta}_0 = (\beta_{10}, \beta_{20}, \cdots, \beta_{M0})^T$, $\boldsymbol{\xi} = (\boldsymbol{\xi}_1^T, \boldsymbol{\xi}_2^T, \cdots, \boldsymbol{\xi}_M^T)^T$, and $\boldsymbol{h}\cdot\boldsymbol{\xi}_m$ means the element-wise product. $P(\cdot)$ is a nonconvex penalty function. In this paper, considering the three nonconvex penalty function, the $L_{1/2}$ penalty, the MCP penalty and the SCAD penalty. The $L_{1/2}$ penalty function is $P_{L_{1/2}}(x;\lambda) = \lambda\|x\|_{1/2}^{1/2} = \sum_{i=1}^{p}|x_i|^{1/2}$. MCP penalty function has the following form

$$P_{MCP}(x;\lambda) = \lambda\int_0^x\left(1 - \frac{s}{\gamma\lambda}\right)_+ ds,$$

where $\left(1 - \frac{s}{\gamma\lambda}\right)_+ = \max\left\{1 - \frac{s}{\gamma\lambda}, 0\right\}$. The SCAD penalty function has the following form

$$P_{SCAD}(x;\ \lambda) = \lambda|x|I_{\{0\le|x|\ <\ \lambda\}} + \left(\frac{(a\ -\ 1)\lambda^2}{2} + \lambda^2\right)I_{\{|x|\ge a\lambda\}} + \left(\frac{a\lambda(|x|\ -\ \lambda)\ -\ (|x|^2\ -\ \lambda^2)/2}{a\ -\ 1} + \lambda^2\right)I_{\{\lambda\le|x|\ <\ a\lambda\}},$$

we call these three nonconvex penalties for the methods (4) as "meta-Half", "meta-MCP" and "meta-SCAD", respectively.

Algorithm

In this section, we give the efficient algorithms (Algorithm 1) to solve our models. Note that we can assume that the mean of the predictor variable is zero (through the location transformation). (4) can be decomposed into two nonconvex problems, each of which views $\boldsymbol{h}$ or $\boldsymbol{\xi}$ as fixed. We propose to iteratively solve β_0, $\boldsymbol{h}$, and $\boldsymbol{\xi}$ in (4). First, we fix β_0 and $\boldsymbol{\xi}$ in (4) to maximize $\boldsymbol{h}$. We next fix β_0 and $\boldsymbol{h}$ to maximize $\boldsymbol{\xi}$. Finally, we maximize over β_0 by fixing $\boldsymbol{h}$ and $\boldsymbol{\xi}$. Iterate these steps until the algorithm converges. Since at each step, the value of the objective function (4) decreases, the solution is guaranteed to converge. Specifically, the algorithm is described as follows

Algorithm 1. The iterative optimization algorithm for solving our meta-analysis based on nonconvex regularization models.

```
1: (Standardization) For each dataset $(\boldsymbol{X}_m, \boldsymbol{y}_m)$, the columns of $\boldsymbol{X}_m$ are standardized to zero mean and unit variance;
2: (Initialization) Initialize $\hat{\xi}_{mj}^{(0)} = 1$ for $1 \le m \le M; 1 \le j \le p$, and $\hat{\beta}_{m0}^{(0)} = 0$ for $1 \le m \le M$;
3: (Update $\hat{h}_j^{(k)}$) Let $\tilde{x}_{mi,j} = x_{mi,j}\hat{\xi}_{mj}^{(k-1)}$, where $\hat{\xi}_{mj}^{(k-1)}$ is the value of $\hat{\xi}_{mj}$ at the $(k-1)$th step and
       $\ell(\boldsymbol{h}) = \sum_{m=1}^{M}\sum_{i=1}^{n_m}[y_{mi}(\hat{\beta}_{m0}^{(k-1)} + \sum_{j=1}^{p}\tilde{x}_{mi,j}h_j) - \log\{1 + \exp(\hat{\beta}_{m0}^{(k-1)} + \sum_{j=1}^{p}\tilde{x}_{mi,j}h_j)\}].$
   Estimate $h_j$ by $\hat{h}_j^{(k)} = \arg\max_{h_j}[\ell(\boldsymbol{h}) - \sum_{j=1}^{p} P(h_j; \lambda_h)], j = 1, 2, \cdots, p$ ;
4: (Update $\hat{\xi}_{mj}^{(k)}$) Let $\breve{x}_{mi,j} = x_{mi,j}\hat{h}_j^{(k)}$ and
       $\ell(\boldsymbol{\xi}) = \sum_{m=1}^{M}\sum_{i=1}^{n_m}[y_{mi}(\hat{\beta}_{m0}^{(k-1)} + \sum_{j=1}^{p}\breve{x}_{mi,j}\xi_{mj}) - \log\{\hat{\beta}_{m0}^{(k-1)} + \sum_{j=1}^{p}\breve{x}_{mi,j}\xi_{mj}\}].$
   Estimate $\xi_{mj}$ by $\hat{\xi}_{mj}^{(k)} = \arg\max_{\xi_{mj}}[\ell(\boldsymbol{\xi}) - \sum_{j=1}^{p}\sum_{m=1}^{M} P(\xi_{mj}; \lambda_\xi)]$;
5: (Update $\hat{\beta}_{mj}^{(k)}$) Let $\hat{\beta}_{mj}^{(k)} = \hat{h}_j^{(k)}\hat{\xi}_{mj}^{(k)}$ for $m = 1, 2, \cdots, M; j = 1, 2, \cdots, p$, and
       $\ell(\boldsymbol{\beta}_0) = \sum_{m=1}^{M}\sum_{i=1}^{n_m}[y_{mi}(\beta_{m0} + \sum_{j=1}^{n_m} x_{mi,j}\hat{\beta}_{mj}^{(k)}) - \log\{1 + \exp(\beta_{m0} + \sum_{j=1}^{p} x_{mi,j}\hat{\beta}_{mj}^{(k)})\}].$
   Estimate $\beta_{m0}$ by $\hat{\beta}_{m0}^{(k)} = \arg\max_{\beta_{m0}}\ell(\boldsymbol{\beta}_0)$;
6: Repeat Steps 3-5 until $\max_{1\le m\le M, 0\le j\le p}|\beta_{mj}^{(k)} - \beta_{mj}^{(k-1)}| < \varepsilon$, the $\varepsilon$ is the predefined threshold (in this paper, we use $10^{-3}$).
```

Step 3 and step 4 are general nonconvex regularization problem. We[52,53] propose the nonconvex iterative thresholding algorithms based on approximate message passing (Half-AMP, MCP-AMP and SCAD-AMP) to solve the nonconvex regularization problem, and verified the effectiveness of the algorithms through theoretical analysis and experiment. In this paper, for the two problems in step 3 and step 4, we apply the Half-AMP algorithm, the MCP-AMP algorithm and the SCAD-AMP algorithm to solve the meta-Half, meta-MCP and meta-SCAD respectively.

In order to solve the above two problems in step 3 and step 4, we first consider the solution of the traditional logistic regression model. Here we omit the intercept term (in fact, just rewrite the input variable as $\tilde{\boldsymbol{x}}_i = (1,\ \boldsymbol{x}_i^T)^T$), the logistic regression can be expressed as the following optimization problem

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}{\operatorname{argmin}}\ell(\boldsymbol{\beta}) = \underset{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}{\operatorname{argmin}}\left\{-\sum_{i=1}^{N}[y_i(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})\ -\ \ln\ (1\ +\ \exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}))]\right\}. \tag{6}$$

Differentiating $\ell(\beta)$ with respect to β, we can get

$$\frac{\partial\ell(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = -\sum_{i=1}^{N}\boldsymbol{x}_i(y_i\ -\ \mu(\boldsymbol{x}_i;\ \boldsymbol{\beta})), \tag{7}$$

where $\mu(\boldsymbol{x}_i;\,\boldsymbol{\beta}) = \frac{\exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})}$. To find the optimal solution $\hat{\boldsymbol{\beta}}$ of equation (6), we let $\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0$, and use the Newton-Raphson iteration algorithm which requires computing the second derivative, the Hessian matrix has the following form

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathrm{T}}} = \sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}} \mu(\boldsymbol{x}_i;\,\boldsymbol{\beta})\,(1 - \mu(\boldsymbol{x}_i;\,\boldsymbol{\beta})). \tag{8}$$

Hence, given the current estimated value β^{old} of β, the new estimated value β^{new} is updated as following

$$\boldsymbol{\beta}^{\mathrm{new}} = \boldsymbol{\beta}^{\mathrm{old}} - \left(\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathrm{T}}}\right)^{-1} \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}, \tag{9}$$

where the value of the derivative (and second derivative) is calculated at the point β^{old}. The equation (9) can be expressed by matrix form. Let $\boldsymbol{X}$ be a $N \times P$ matrix, where the i-th row is $\boldsymbol{x}_i$; W is a diagonal matrix, and the elements on the diagonal

$$w_i = \mu(\boldsymbol{x}_i;\,\boldsymbol{\beta})\,(1 - \mu(\boldsymbol{x}_i;\,\boldsymbol{\beta})). \tag{10}$$

Let $\boldsymbol{y} = (y_1, y_2, \cdots, y_N)^{\mathrm{T}}$, $\boldsymbol{\mu} = (\mu(\boldsymbol{x}_1;\,\boldsymbol{\beta}), \mu(\boldsymbol{x}_2;\,\boldsymbol{\beta}), \cdots, \mu(\boldsymbol{x}_N;\,\boldsymbol{\beta}))^{\mathrm{T}}$, then the formulas (7) and (8) can be expressed as

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{\mu}), \quad \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathrm{T}}} = \boldsymbol{X}^{\mathrm{T}} \boldsymbol{W} \boldsymbol{X}. \tag{11}$$

Therefore, Newton-Raphson iteration (9) can be expressed as

$$\begin{aligned} \boldsymbol{\beta}^{\mathrm{new}} &= \boldsymbol{\beta}^{\mathrm{old}} + (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{\mu}) \\ &= (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}(\boldsymbol{X}\boldsymbol{\beta}^{\mathrm{old}} + \boldsymbol{W}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})) \\ &= (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{z}, \end{aligned} \tag{12}$$

where

$$\boldsymbol{z} = \boldsymbol{X}\boldsymbol{\beta}^{\mathrm{old}} + \boldsymbol{W}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}). \tag{13}$$

It can be seen that each Newton-Raphson iteration actually solves the weighted least squares problem as follows

$$\hat{\boldsymbol{\beta}} = \underset{\beta}{\operatorname{argmin}} (\boldsymbol{z} - \boldsymbol{X}\boldsymbol{\beta})^{\mathrm{T}} \boldsymbol{W} (\boldsymbol{z} - \boldsymbol{X}\boldsymbol{\beta}). \tag{14}$$

Based on the solution process of the traditional logistic regression model, a similar iterative algorithm can be used to solve the logistic regression with nonconvex penalties problem, and only a slight deformation of the formula (14) is needed to obtain the iterative algorithm.

$$\hat{\boldsymbol{\beta}} = \underset{\beta}{\operatorname{argmin}} (\boldsymbol{z} - \boldsymbol{X}\boldsymbol{\beta})^{\mathrm{T}} \boldsymbol{W} (\boldsymbol{z} - \boldsymbol{X}\boldsymbol{\beta}) + P(\boldsymbol{\beta};\,\lambda), \tag{15}$$

where $P(\,\cdot\,;\cdot\,)$ is the nonconvex penalty function. It is easy to see that the minimization problem (15) is equivalent to the maximization problem (2). The minimization problem (15) can be solved by the nonconvex iterative thresholding algorithms based on approximate message passing[52,53] (which are based on linear regression $\boldsymbol{y} = \boldsymbol{X}\beta$, $\boldsymbol{y} \in R^N$, $\boldsymbol{X} \in R^{N\times p}$). The algorithms are according to the following iteration:

$$\boldsymbol{\beta}^{(k+1)} = \eta(\boldsymbol{\beta}^{(k)} + \boldsymbol{X}^{T}\boldsymbol{r}^{(k)}), \tag{16}$$

$$\boldsymbol{r}^{(k+1)} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^{(k)} + \frac{1}{\delta}\boldsymbol{r}^{(k)}\left\langle \eta'\left(\boldsymbol{\beta}^{(k-1)} + \boldsymbol{X}^{T}\boldsymbol{r}^{(k-1)}\right)\right\rangle, \tag{17}$$

where $\delta = \frac{p}{N}$ represents a measure of indeterminacy of the measurement system, in this paper, considering the case δ is fixed for $N \to \infty$. For a vector $u = (u_1, u_2, \ldots, u_N)$, $\langle u \rangle = \sum_{i=1}^{N} u_i/N$, $\eta'(x) = \frac{\partial}{\partial x}\eta(x)$. η is the thresholding function, in this paper, η represents the Half thresholding function, the MCP thresholding function and the SCAD thresholding function, respectively. The Half thresholding function is

$$\eta(u;\,\lambda) = \begin{cases} g(u;\,\lambda), & |u| > \lambda, \\ 0, & |u| \le \lambda, \end{cases} \tag{18}$$

where

$$g(u;\ \lambda) = \frac{2}{3}u\left(1 + \cos\left(\frac{2\pi}{3} - \frac{2}{3}h(u;\ \lambda)\right)\right),\ h(u;\ \lambda) = \arccos\left(\frac{\sqrt{2}}{2}\left(\frac{\lambda}{|u|}\right)^{\frac{2}{3}}\right)\right).$$

The MCP thresholding function is

$$\eta_{MCP}(u;\ \lambda) = \begin{cases} u, & |u| > \gamma\lambda, \\ \dfrac{\eta^{S}(u;\ \lambda)}{1 - \frac{1}{\gamma}}, & |u| \le \gamma\lambda, \end{cases}$$

where

$$\eta^{S}(u;\ \lambda) = \begin{cases} u - sign(u)\lambda, & |u| > \lambda, \\ 0, & |u| \le \lambda, \end{cases}$$

where $sign(u)$ is sign function,

$$sign(u) = \begin{cases} 1, & u > 0, \\ -1, & u < 0, \\ 0, & u = 0. \end{cases}$$

The SCAD thresholding function is

$$\eta_{SCAD}(u;\ \lambda) = \begin{cases} u, & |u| > a\lambda, \\ \dfrac{(a-1)u - sign(u)a\lambda}{a-2}, & 2\lambda < |u| \le a\lambda, \\ u - sign(u)\lambda, & \lambda < |u| \le 2\lambda, \\ 0, & otherwise. \end{cases}$$

Theoretical Properties

In this section, we study the theoretical properties of the meta-Half method. The meta-Half has the following uniform form

$$\max_{\beta_0,\boldsymbol{h},\boldsymbol{\xi}} \sum_{m=1}^{M}\ell_m(\beta_{m0},\ \boldsymbol{h},\ \boldsymbol{\xi}_m) - \lambda_h\sum_{j=1}^{p}|h_j|^{\frac{1}{2}} - \lambda_\xi\sum_{j=1}^{p}\sum_{m=1}^{M}|\xi_{mj}|^{\frac{1}{2}}, \tag{19}$$

there are two tuning parameters λ_h and λ_ξ in (19), we first show that the two tuning parameters can be simplified into one. Specifically, let $\lambda = \lambda_h\lambda_\xi$, we can show that (19) is equivalent to

$$\max_{\beta_0,\boldsymbol{h},\boldsymbol{\xi}} \sum_{m=1}^{M}\ell_m(\beta_{m0},\ \boldsymbol{h},\ \boldsymbol{\xi}_m) - \sum_{j=1}^{p}|h_j|^{\frac{1}{2}} - \lambda\sum_{j=1}^{p}\sum_{m=1}^{M}|\xi_{mj}|^{\frac{1}{2}}. \tag{20}$$

Lemma 1. *If* $(\widetilde{\boldsymbol{\beta}}_0,\ \widetilde{\boldsymbol{h}},\ \widetilde{\boldsymbol{\xi}})$ *is a local maximizer of* (19). *Then there exists a local maximizer* $(\widehat{\boldsymbol{\beta}}_0,\ \widehat{\boldsymbol{h}},\ \widehat{\boldsymbol{\xi}})$ *of* (20) *such that* $\widetilde{h}_j\widetilde{\xi}_{mj} = \hat{h}_j\hat{\xi}_{mj}$ *and* $\widetilde{\boldsymbol{\beta}}_0 = \hat{\boldsymbol{\beta}}_0$. *Vice versa.*

The proof is in the Supplementary. This lemma indicates that although (19) and (20) may provide different h_j and ξ_{mj}, the final fitted models from them are the same. Therefore, we only need to tune one parameter $\lambda = \lambda_h\lambda_\xi$ other than tune λ_h and λ_ξ separately in practice.

We then show that (20) can also be written in an equivalent form using the original regression coefficients β_{mj}.

Lemma 2. *Suppose* $(\hat{\boldsymbol{h}},\ \hat{\boldsymbol{\xi}})$ *is a local maximizer of* (20), *for* $j = 1, 2, \cdots, p$, *let* $\hat{\beta}_{mj} = \hat{h}_j\hat{\xi}_{mj}$, $\hat{\boldsymbol{\beta}}_j = \left(\hat{\beta}_{1j},\ \hat{\beta}_{2j},\ \cdots,\ \hat{\beta}_{Mj}\right)^T$ *and* $\hat{\boldsymbol{\xi}}_j = \left(\hat{\xi}_{1j},\ \hat{\xi}_{2j},\ \cdots,\ \hat{\xi}_{Mj}\right)^T$,

(a) If $\hat{h}_j = 0$, *then* $\hat{\boldsymbol{\beta}}_j = \mathbf{0}$;

(b) If $\hat{h}_j \neq 0$, *then* $\hat{\boldsymbol{\beta}}_j \neq \mathbf{0}$ *and* $\hat{h}_j = \lambda\left\|\hat{\boldsymbol{\beta}}_j\right\|_{1/2}^{1/2}$, $\hat{\boldsymbol{\xi}}_j = \dfrac{\hat{\beta}_j}{\lambda\left\|\hat{\beta}_j\right\|_{1/2}^{1/2}}$.

The proof is in the Supplementary.

Theorem 1. *If* $(\widehat{\boldsymbol{\beta}}_0,\ \widehat{\boldsymbol{h}},\ \widehat{\boldsymbol{\xi}})$ *is a local maximizer of* (20), *then* $\hat{\boldsymbol{\beta}}$ *with* $\hat{\beta}_{mj} = \hat{h}_j\hat{\xi}_{mj}$, *is a local maximizer of*

		$\pi = 0.2$	$\pi = 0.5$	$\pi = 0.9$
meta-Half	Sensitivity	0.9693 (1.70*E* − 03)	0.9215 (4.97*E* − 03)	0.9229 (1.30*E* − 03)
	Specificity	0.9862 (2.26*E* − 04)	0.9903 (6.36*E* − 05)	0.9837 (1.40*E* − 03)
	Accuracy	0.9861 (2.24*E* − 04)	0.9901 (6.50*E* − 05)	0.9835 (1.41*E* − 03)
meta-MCP	Sensitivity	0.9651 (1.90*E* − 03)	**0.9362 (3.90E − 02)**	0.9205 (2.70*E* − 03)
	Specificity	0.9884 (4.18*E* − 05)	0.9840 (2.02*E* − 05)	0.9846 (1.15*E* − 02)
	Accuracy	0.9883 (4.16*E* − 05)	0.9838 (2.05*E* − 05)	0.9840 (1.13*E* − 02)
meta-SCAD	Sensitivity	**0.9738 (1.60E − 03)**	0.9306 (2.07*E* − 03)	0.9392 (2.50*E* − 03)
	Specificity	0.9903 (1.44*E* − 05)	0.9853 (1.20*E* − 04)	0.9505 (4.20*E* − 03)
	Accuracy	0.9903 (1.42*E* − 05)	0.9850 (1.44*E* − 05)	0.9504 (4.10*E* − 03)
meta-LASSO	Sensitivity	0.9065 (8.42*E* − 02)	0.9217 (6.60*E* − 02)	**0.9425 (6.48E − 02)**
	Specificity	0.9710 (2.72*E* − 03)	0.9869 (3.07*E* − 03)	0.9940 (1.71*E* − 03)
	Accuracy	0.9708 (2.79*E* − 03)	0.9866 (3.02*E* − 03)	0.9935 (1.69*E* − 03)
composite MCP	Sensitivity	0.8454 (1.46*E* − 01)	0.5428 (1.23*E* − 01)	0.3167 (7.60*E* − 02)
	Specificity	0.9988 (6.02*E* − 04)	0.9992 (4.85*E* − 04)	0.9984 (7.02*E* − 04)
	Accuracy	0.9985 (6.12*E* − 04)	0.9969 (1.02*E* − 03)	0.9922 (1.18*E* − 03)
group Bridge	Sensitivity	0.8734 (7.77*E* − 02)	0.6856 (1.11*E* − 01)	0.2842 (6.27*E* − 02)
	Specificity	0.9997 (2.84*E* − 04)	0.9999 (1.05*E* − 04)	0.9999 (3.49*E* − 05)
	Accuracy	0.9994 (3.42*E* − 04)	0.9983 (6.08*E* − 04)	0.9934 (6.74*E* − 04)
group exponential LASSO	Sensitivity	0.8809 (8.70*E* − 02)	0.7315 (1.67*E* − 01)	0.4661 (2.29*E* − 01)
	Specificity	0.9984 (1.10*E* − 03)	0.9981 (9.33*E* − 04)	0.9976 (1.33*E* − 03)
	Accuracy	0.9981 (1.09*E* − 03)	0.9967 (8.01*E* − 04)	0.9928 (1.37*E* − 03)

Table 1. The sensitivity, specificity and accuracy of coefficient β of the seven methods: presented values are the mean (standard error).

$$\max_{\boldsymbol{\beta}}\sum_{m=1}^{M}\ell_m(\boldsymbol{\beta}_m) - \lambda\sum_{j=1}^{p}\sum_{m=1}^{M}|\beta_{mj}|^{\frac{1}{2}}, \tag{21}$$

where $\boldsymbol{\beta} = \left(\beta_{10}, \beta_{11}, \cdots, \beta_{Mp}\right)^T$. *On the other hand, if* $\hat{\boldsymbol{\beta}}$ *is a solution of* (21), *then* $(\widehat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{h}}, \hat{\boldsymbol{\xi}})$ *is a solution of* (20), *where* $\hat{\boldsymbol{\beta}}_0 = (\hat{\beta}_{10}, \hat{\beta}_{20}, \cdots, \hat{\beta}_{M0})^T, \|\hat{\boldsymbol{\beta}}_j\|_{1/2}^{1/2} = \sum_{m=1}^{M}|\hat{\beta}_{mj}|^{\frac{1}{2}}$,

$$(\hat{\boldsymbol{h}}, \hat{\boldsymbol{\xi}}) = \begin{cases} \hat{h}_j = 0, \hat{\boldsymbol{\xi}}_j = \mathbf{0}, & if\ \hat{\boldsymbol{\beta}}_j = \mathbf{0}, \\ \hat{h}_j = \lambda\|\hat{\boldsymbol{\beta}}_j\|_{1/2}^{1/2}, \hat{\boldsymbol{\xi}}_j = \dfrac{\hat{\boldsymbol{\beta}}_j}{\lambda\left\|\hat{\boldsymbol{\beta}}_j\right\|_{1/2}^{1/2}}, & if\ \hat{\boldsymbol{\beta}}_j \neq \mathbf{0}, \end{cases}$$

where $\hat{\boldsymbol{h}} = \left(\hat{h}_1, \hat{h}_2, \cdots, \hat{h}_p\right), \hat{\boldsymbol{\xi}} = \left(\hat{\xi}_{11}, \hat{\xi}_{12}, \cdots, \hat{\xi}_{Mp}\right)^T, \hat{\boldsymbol{\beta}}_j = \left(\hat{\beta}_{1j}, \hat{\beta}_{2j}, \cdots, \hat{\beta}_{Mj}\right)^T$ *and* $\hat{\boldsymbol{\xi}}_j = \left(\hat{\xi}_{1j}, \hat{\xi}_{2j}, \cdots, \hat{\xi}_{Mj}\right)^T$.

The proof is in the Supplementary. If we regard one gene's effects among all datasets as a "group", then (21) imposes an L_1 penalty on each group and a square root penalty on individual elements within a group. The following theorem shows the theoretical properties of the meta-Half.

Theorem 2. *The meta-Half method possesses sparsity, unbiasedness and oracle properties.*

The proof is in the Supplementary.

Experiments

In this section, we analyze the performance of our methods (meta-Half, meta-MCP and meta-SCAD) by simulation and real-data analysis. We compare these three methods with other four methods, which are meta-LASSO, composite MCP, group Bridge and group exponential LASSO. The codes of our methods are available at GitHub (https://github.com/zhhui019/meta-nonconvex). The meta-LASSO is implemented by Li *et al.*[39]. The composite MCP, the group Bridge and the group exponential LASSO are implemented by Patrick Breheny and Yaohui Zeng's R package "grpreg".

Simulations.

Simulation studies are performed to compare the performance of the proposed meta-Half, meta-MCP and meta SCAD with the meta-LASSO, composite MCP, group Bridge and group exponential LASSO.

Generate simulated data. In this simulation, we use the normal distribution to generate the gene expression $\boldsymbol{x}_{mi}$ ($m = 1, 2, \cdots, M$; $i = 1, 2, \cdots, n_m$) with $M = 10$ datasets, each dataset contains $n_m = 50$ samples, and each sample contains $p = 1,000$ genes. The response y_{mi} is generated from a logistic model

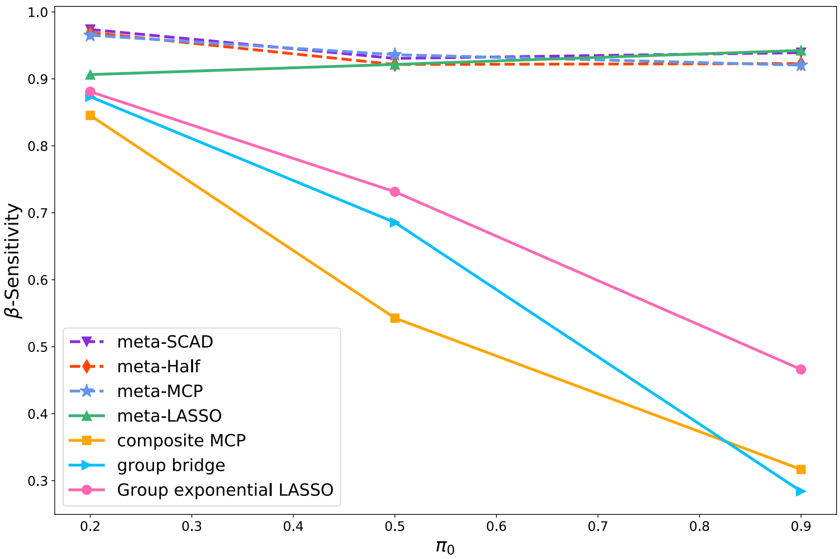


Figure 1. The sensitivity trend of coefficient β for all seven methods with the varying levels of heterogeneity.

Dataset	No. of Probs	Classes (Class 0/Class 1)	No. of samples (Class 0/Class 1)	Affymetrix Platform
GSE10072	22284	Normal/ Lung Cancer	107 (49/58)	U133A
GSE19188	54675	Normal/ Lung Cancer	156 (65/91)	U133 Plus 2.0
GSE19804	54676	Normal/ Lung Cancer	120 (60/60)	U133 Plus 2.0

Table 2. The description of three publicly available lung cancer gene expression datasets.

$$Pr\,(y_{mi}\;=\;1|\boldsymbol{x}_{mi})\;=\;\frac{\exp(\boldsymbol{x}_{mi}^{T}\boldsymbol{\beta}_{m}^{*})}{[1\;+\;\exp(\boldsymbol{x}_{mi}^{T}\boldsymbol{\beta}_{m}^{*})]},$$

where $\boldsymbol{\beta}_m^* = \left(\beta_{m1}^*, \beta_{m2}^*, \cdots, \beta_{mp}^*\right)$ *and we suppose that the intercept term* $\beta_{m0}^* = 0$. *We let* $\beta_{mj}^* = \alpha_{mj}\theta_{mj}$ *simulate possible data heterogeneity, for m* = *1, 2* ⋯, *M; j* = *1, 2* ⋯, *10,* α_{mj} *are generated from N(3, 0. 5*^{2}*) and* θ_{mj} *are generated from Bernoulli(*π_0*), for m* = *1, 2* ⋯, *M; j* = *11, 12* ⋯, *1000, let* $\beta_{mj}^* = 0$*. This means that the first 10 genes of each dataset are important to the response with probability* π_0*. The value* α_{mj} *demonstrates whether the jth gene is important in the mth dataset, and the value* θ_{mj} *demonstrates different levels of heterogeneity among different datasets, in this simulation, considering* $\pi_0 = 0.9, 0.5, 0.2$ *to represent the low, medium and high heterogeneity. We run 30 replicates and report the average measurement.*

For the all methods, the tuning parameters are selected by minimizing the BIC:

$$BIC(\lambda)\;=\;\sum_{m=1}^{M}\{-2\ell_m(\hat{\boldsymbol{\beta}}_{m,\lambda})\;+\;S_m\log\,(n_m)\},\tag{22}$$

where $\hat{\boldsymbol{\beta}}_{m,\lambda}$ is the estimated coefficients in the mth dataset, λ is the tuning parameter, S_m is the number of non-zero elements of $\hat{\boldsymbol{\beta}}_{m,\lambda}$, $\ell_m(\hat{\boldsymbol{\beta}}_{m,\lambda})$ is the log-likelihood for the mth dataset and has the form (5).

Analysis of simulation. The variable selection performance of the seven methods is evaluated using the selection sensitivity, specificity and accuracy of coefficient **β**. The sensitivity is the proportion of non-zero β_{mj}^*'s that are correctly estimated as non-zero, the specificity is the proportion of zero β_{mj}^*'s that are correctly estimated as zero and the accuracy is the proportion of β_{mj}^*'s that are correctly estimated.

The simulation results are summarized in Table 1 (The variable selection performance of the seven methods are evaluated using the selection sensitivity, specificity and accuracy of coefficient **β**). Table 1 shows that the specificity and accuracy of the coefficients **β** of all seven methods are similar. The sensitivity trend of coefficient **β** for all seven methods with the varying levels of heterogeneity is shown in Fig. 1. Figure 1 shows that the sensitivity (the proportion of non-zero β_{mj}^*'s that are correctly estimated as non-zero) of the composite MCP, the group Bridge and the group exponential LASSO dramatically decreases as π_0 increases, while the sensitivity of meta-Half, meta-MCP, meta-SCAD and meta-LASSO remains above 0.9 for $\pi_0 = 0.2, 0.5, 0.9$. When $\pi = 0.2$, the sensitivity of meta-Half, meta-MCP and meta-SCAD are 0.9693, 0.9651 and 0.9738, respectively, which are significantly higher than other methods. This result shows that our proposed meta-Half, meta-MCP and meta-SCAD

Methods	Training data			Testing data		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
meta-Half	0.9766 (2.69E-02)	0.9673 (9.10E-03)	0.9903 (1.54E-05)	0.9449 (2.16E-02)	0.9464 (7.93E-03)	0.9437 (1.66E-05)
meta-MCP	**0.9768** (5.37E-04)	**0.9677** (1.70E-03)	0.9903 (8.56E-05)	0.9452 (1.99E-03)	**0.9466** (6.94E-03)	0.9439 (1.35E-04)
meta-SCAD	0.9727 (3.13E-03)	0.9608 (1.77E-03)	0.9903 (2.19E-02)	**0.9528** (4.64E-03)	0.9464 (4.80E-03)	0.9577 (2.50E-02)
meta-LASSO	0.9309 (1.01E-02)	0.8722 (1.87E-02)	**0.9994** (2.15E-03)	0.8953 (2.11E-02)	0.8291 (3.58E-02)	**0.9792** (1.25E-02)
composite MCP	0.9353 (1.45E-02)	0.9221 (2.05E-02)	0.9519 (1.85E-02)	0.8656 (1.94E-02)	0.8283 (2.23E-02)	0.9060 (2.85E-02)
group Bridge	0.6410 (1.88E-02)	0.4039 (2.59E-02)	0.9508 (2.23E-02)	0.6255 (2.06E-02)	0.3240 (3.15E-02)	0.9317 (2.98E-02)
group exponential Lasso	0.9385 (9.78E-03)	0.9155 (1.64E-02)	0.9655 (1.48E-02)	0.8942 (1.85E-02)	0.8432 (2.47E-02)	0.9589 (2.05E-02)

Table 3. Performance comparisons of different methods in three lung cancer datasets. Presented values are the average (standard error).

have the superior performance when data heterogeneity is strong (π_0 is small). With the weakening of data heterogeneity($\pi = 0.5, 0.9$), the performance of the four meta methods (meta-Half, meta-MCP, meta-SCAD and meta-LASSO) tends to be comparable. The specificity and accuracy of the coefficients for all seven methods are similar.

Real-Data analysis.

In this section, we apply our methods (meta-Half, meta-MCP and meta-SCAD) to three publicly available lung cancer gene expression datasets, and compare our three methods with other four methods including meta-LASSO, composite MCP, group Bridge and group exponential LASSO.

Lung cancer datasets. The three publicly available lung cancer microarray datasets come from disparate platforms and can be download from GEO (https://www.ncbi.nlm.nih.gov/gds/). The three datasets are described as follows:

GSE10072 dataset. The dataset is gene expression signature of cigarette smoking, it contains 107 final expression samples from 58 tumors and 49 non-tumor tissues from 20 never smokers, 26 former smokers, and 28 current smokers, each sample has 22283 genes. The original gene expression data is provided by Landi *et al.*[54].

GSE19188 dataset. The dataset is expression data for early stage non-small-cell lung cancer (NSCLC), it contains 156 samples from 91 tumor tissues and 65 adjacent normal lung tissue samples, each sample has 54675 genes. The more information can be found in Hou *et al.*[55].

GSE19804 dataset. The dataset is non-smoking female lung cancer in Taiwan, it contains 120 samples from 60 tumors and 60 normal tumor tissues, each sample has 54675 genes. The more information can be found in Lu *et al.*[56].

Each dataset is divided into two parts, about 70 percent of the datasets as training samples and the other 30 percent as testing samples. Table 2 lists the details of the three datasets.

The original Affymetrix data was first normalized and log-transformed by a robust multi-array average (RMA) method[57]. After that, downloading and installing the appropriate custom chip definition files (CDFs) packages according to the type of microarray platform. The CDF package is necessary for probe annotation for Affymetrix data. The probes of the normalized data can be successfully mapped to Entrez Gene IDs by annotation packages in Bioconductor[58]. If multiple probes match a single Entrez ID, we calculated the median of values of those probes as the expression value for this gene.

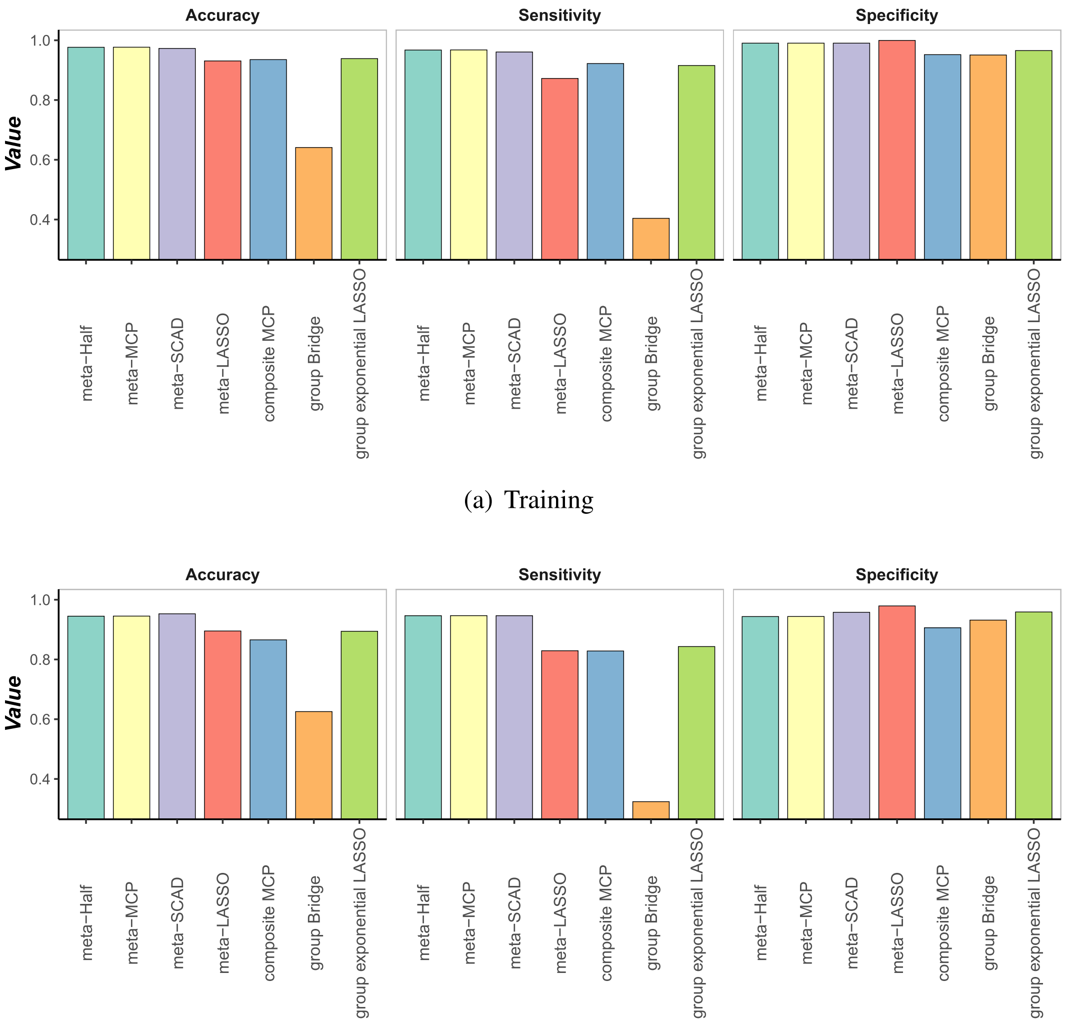
We extract common genes from the three gene expression datasets as the merged set of genes. There are 13515 common genes in three datasets and our analysis is based on those 13515 genes. We use a random partition in three lung cancer datasets, and apply aforementioned seven methods to select important genes, with the optimal tuning parameters chosen by the BIC as discussed above. We repeat this procedure 30 times and report the average measurement and standard error.

Evaluating the classification performance. Table 3 demonstrates the prediction performance of the seven methods in three lung cancer datasets. The sensitivity, specificity and accuracy of training and testing predictions for all seven methods are shown in Fig. 2.

As shown in Table 3 and Fig. 2, for the training dataset and testing dataset, the sensitivity and accuracy of meta-Half, meta-MCP, meta-SCAD are consistently higher than the other four methods, and the specificity of all methods are similar. This result shows that our three methods are more effectively distinguish whether an individual is a disease patient compared to the other four methods. Therefore, our three methods have superior performance than the other four methods in the prediction and diagnosis of diseases.

Analysis of the selected genes. Table 4 gives the names of genes selected in each dataset. We focus on the gene WIF1 which is bolded in the Table 4. WIF1, a secreted Wnt antagonist, is a downstream gene of the Wnt/β-catenin pathway, which exerts inhibition through direct binding to Wnt proteins[59]. WIF1 was found to be silenced by methylation in various human carcinomas including lung[60], oral[61], nasopharyngeal[62], esophageal[63], breast[64] and colon cancer[65] etc.

As shown in Table 4, our three methods (meta-Half, meta-MCP and meta-SCAD) all select gene WIF1 on both datasets GSE10072 and GSE19188, but not select gene WIF1 on dataset GSE19804. As we known that



(a) Training

(b) Testing

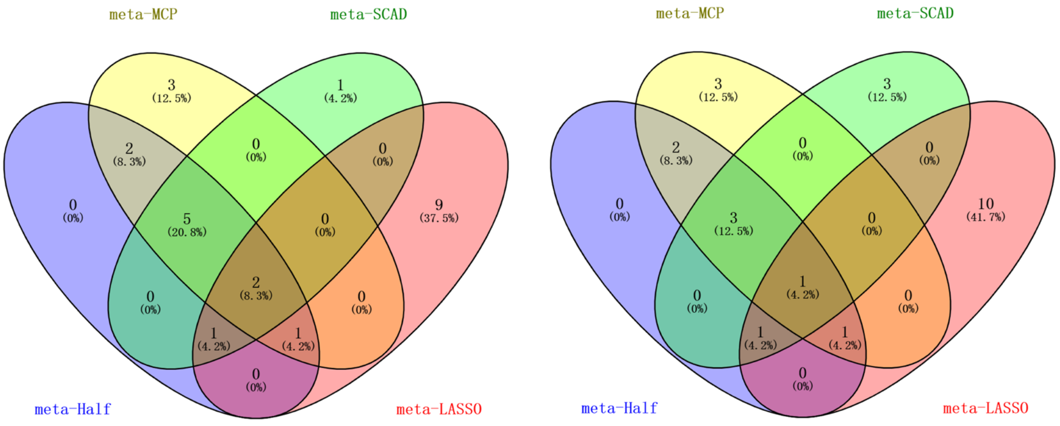
Figure 2. Training and testing prediction performance of different methods on lung cancer datasets. (**a**) Training. (**b**) Testing.

GSE10072 dataset is gene expression signature of cigarette smoking, GSE19188 dataset is expression data for early stage non-small-cell lung cancer (NSCLC) and GSE19804 dataset is non-smoking female lung cancer in Taiwan. Huang *et al*. showed that WIF1 is significantly associated with the smoking behavior in NSCLC patients[66]. It shows that our three methods can more realistically identify the important biomarkers from different datasets which have heterogeneity. The meta-LASSO selects gene WIF1 in all three lung cancer datasets, and the genes selected by meta-LASSO in the three lung cancer datasets are the same. The other three methods (composite MCP, group Bridge and group exponential LASSO) cannot select gene WIF1 in all three lung cancer datasets. Therefore, our three methods are superior to the other four methods when applied in the heterogeneity datasets.

The number of genes selected by meta-Half, meta-MCP, meta-SCAD and meta-LASSO are 11, 13, 9 and 13 respectively. Figure 3 shows the overlap of commonly selected genes across the four different methods (meta-Half, meta-MCP, meta-SCAD, meta-LASSO) in three lung cancer datasets. The other three methods (composite MCP, group Bridge and group exponential LASSO) select fewer genes, so we don't show the genes they selected in Fig. 3. As shown in Fig. 3(a), for the datasets GSE10072 and GSE19188, seven common genes are selected by meta-Half, meta-MCP and meta-SCAD, which are CXCL13, COL11A1, SPP1, MMP12, AGER, WIF1 and FCN3. Two common genes are selected by meta-Half, meta-MCP, meta-SCAD and meta-LASSO, which are SPP1 and WIF1. Figure 3(b) shows that for dataset GSE19804, four common genes selected by meta-Half, meta-MCP and meta-SCAD are CXCL13, SPP1, MMP12 and AGER. One common genes are selected by meta-Half, meta-MCP, meta-SCAD and meta-LASSO, which is SPP1. More unique non-overlapping sets of genes are selected by our three methods and meta-LASSO. In addition, some of the aforementioned genes have been reported in the literature. COL11A1 is collagen type XI alpha 1 chain. The over-expression of COL11A1 reportedly correlates with lymph node metastasis and poor prognosis in non-small cell lung cancer and ovarian cancer[67]. Zhang *et al*. suggest that SPP1 and AGER are risk factors for lung adenocarcinoma, and these two genes may be utilized in the

	GSE10072			GSE19188			GSE19804		
meta-Half	CXCL13	MMP12	COL11A1	CXCL13	MMP12	COL11A1	CXCL13	SPINK1	AGER
	TOX3	SPINK1	FCN3	TOX3	SPINK1	FCN3	SPP1	COL10A1	GPM6A
	SPP1	COL10A1	**WIF1**	SPP1	COL10A1	**WIF1**	MMP12	TOX3	
	GPM6A	AGER		GPM6A	AGER				
meta-MCP	CXCL13	SPP1	COL11A1	CXCL13	SPP1	COL11A1	CXCL13	SPP1	
	AGER	MMP12	FCN3	AGER	MMP12	FCN3	AGER	MMP12	
	PPAP2C	COL10A1	**WIF1**	PPAP2C	COL10A1	**WIF1**	PPAP2C	COL10A1	
	GPM6A	TOX3		GPM6A	TOX3		GPM6A	TOX3	
	TOP2A	TMEM100		TOP2A	TMEM100		TOP2A	TMEM100	
meta-SCAD	CXCL13	MMP12	COL11A1	CXCL13	MMP12	COL11A1	CXCL13	MMP12	COL11A1
	CYP4B1	SPINK1	FCN3	CYP4B1	SPINK1	FCN3	CYP4B1	SPINK1	FCN3
	SPP1	AGER	**WIF1**	SPP1	AGER	**WIF1**	SPP1	AGER	
meta-LASSO	PPBP	SFTPC	SPP1	PPBP	SFTPC	SPP1	PPBP	SFTPC	SPP1
	CLDN10	AKR1B10	SFTPD	CLDN10	AKR1B10	SFTPD	CLDN10	AKR1B10	SFTPD
	UPK3B	APOLD1	XIST	UPK3B	APOLD1	XIST	UPK3B	APOLD1	XIST
	SPINK1	COL10A1	**WIF1**	SPINK1	COL10A1	**WIF1**	SPINK1	COL10A1	**WIF1**
	HLA-DQA1 /// LOC100509457			HLA-DQA1 /// LOC100509457			HLA-DQA1 /// LOC100509457		
composite MCP	SOSTDC1	COL11A1		SYNE1			SOSTDC1	COL11A1	
group Bridge	P2RY14			P2RY14	ATP1A2		P2RY14		
group	GDF10	FABP4	COL11A1	GDF10	FABP4	COL11A1	GDF10	FABP4	COL11A1
exponential									
LASSO									

Table 4. Gene selections of seven methods in three lung cancer datasets.

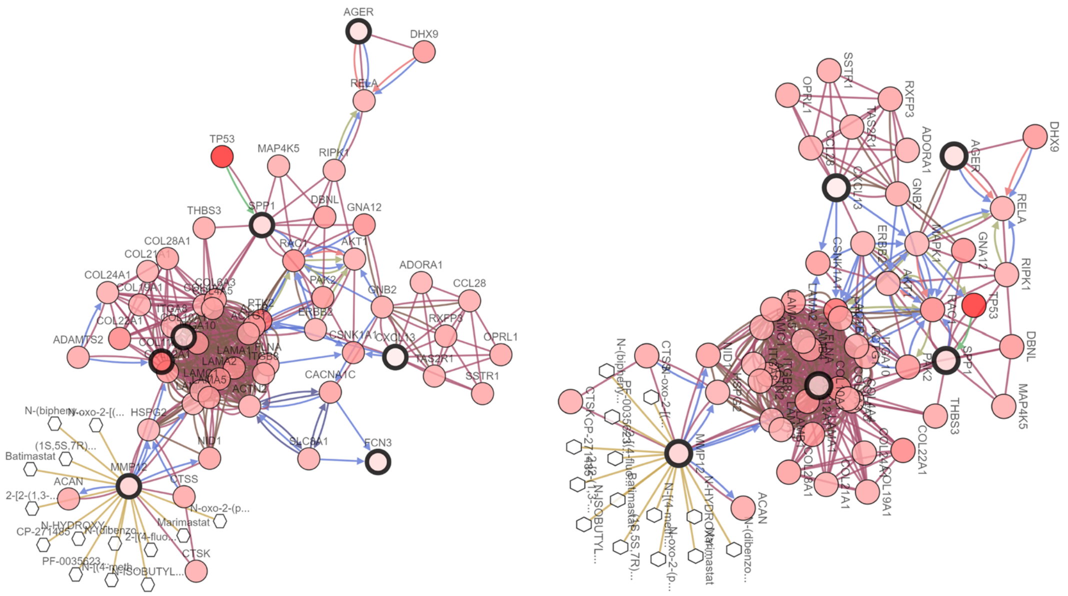


(a) GSE10072 and GSE19188 (b) GSE19804

Figure 3. Overlap of commonly selected genes across the different methods in lung cancer datasets. (**a**) GSE10072 and GSE19188. (**b**) GSE19804.

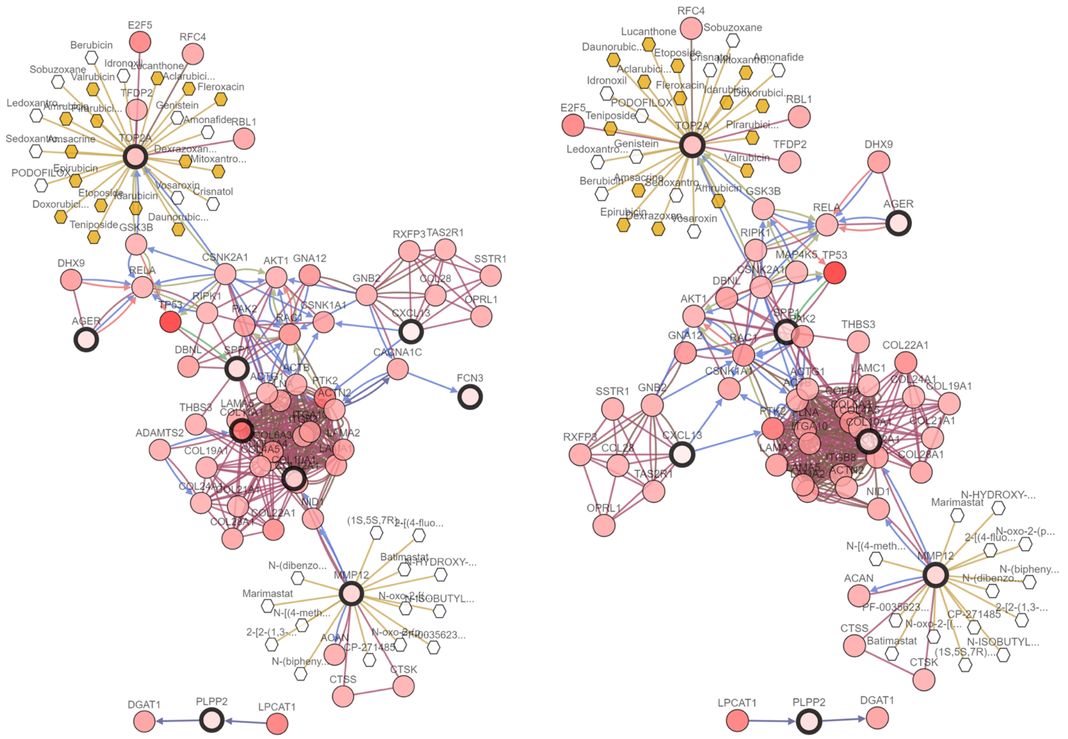
prognostic evaluation of patients with lung adenocarcinoma[68]. The advanced glycosylation end-product specific receptor (AGER) belongs to the immunoglobulin superfamily, whose abnormal expression has been detected in lung cancer[69]. MMP12 is matrix metallopeptidase 12 and may play a role in aneurysm formation and mutations in this gene are associated with lung function and chronic obstructive pulmonary disease (COPD)[70]. WIF1 was found to be silenced by methylation in lung[60]. Lea *et al*. shows that the Ficolin-3, encoded by the FCN3 gene and expressed in the lung and liver, is a recognition molecule in the lectin pathway of the complement system[71]. The aforementioned genes CXCL13, MMP12, AGER and FCN3 are only selected by our three methods, and the gene COL11A1 is selected by our three methods and group exponential LASSO.

To make it easier to demonstrate the interplay between the selected genes from the different methods, we construct a network of interactions among the genes using the cBioPortal[72,73]. Figures 4, 5 and 6 show the interactive network of the genes selected by our three methods in three lung cancer datasets. Most of the genes selected by our three methods are linked to the frequently altered neighbor genes from the TCGA lung adenocarcinoma dataset. The expression of SPP1 is controlled by TP53. TP53 is tumor protein p53, this gene encodes a tumor suppressor protein containing transcriptional activation, DNA binding, and oligomerization domains. Mutations



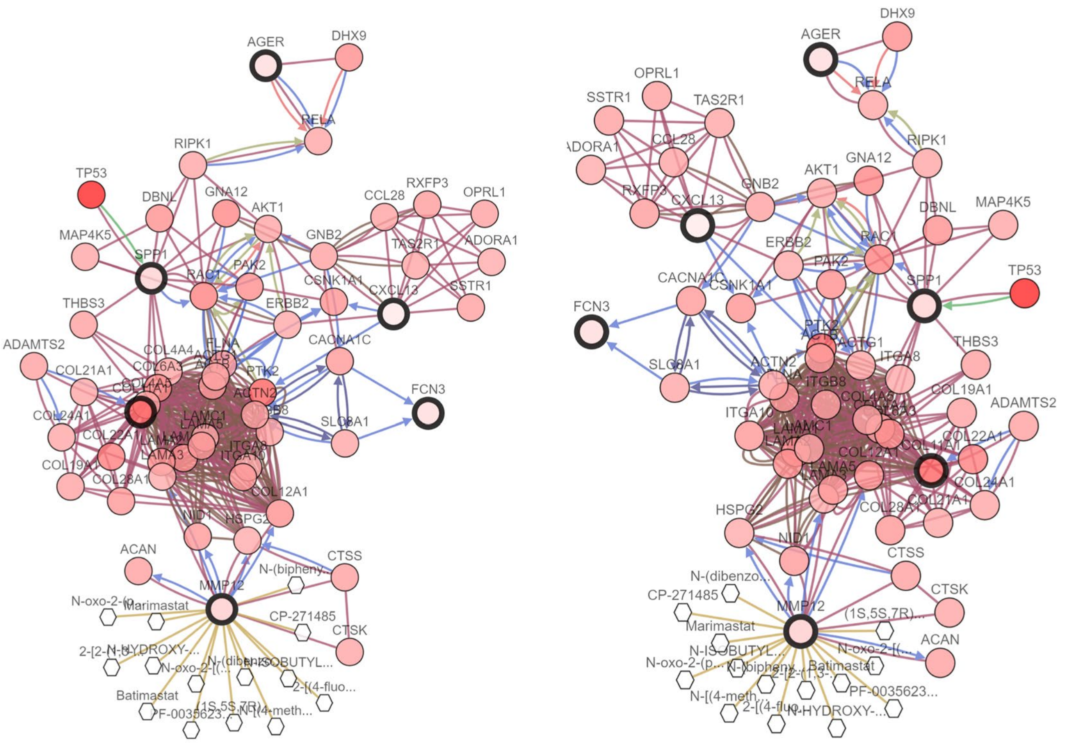
(a) GSE10072 and GSE19188 (b) GSE19804

Figure 4. Network view of the genes selected from meta-Half in lung cancer datasets. The genes corresponding to the selected variables are highlighted by a thicker black outline. The rest of the nodes correspond to the genes that are frequently altered and are known to interact with the highlighted genes (based on publicly available interaction data). The nodes are gradient color-coded according to the alteration frequency based on microarray data derived from the TCGA lung cancer dataset via cBioPortal. (**a**) GSE10072 and GSE19188. (**b**) GSE19804.



(a) GSE10072 and GSE19188 (b) GSE19804

Figure 5. Network view of the genes selected from meta-MCP in lung cancer datasets. (**a**) GSE10072 and GSE19188. (**b**) GSE19804.



(a) GSE10072 and GSE19188 (b) GSE19804

Figure 6. Network view of the genes selected from meta-SCAD in lung cancer datasets. (**a**) GSE10072 and GSE19188. (**b**) GSE19804.

in this gene are associated with a variety of human cancers[74]. MMP12 and TOP2A are targeted by certain cancer drugs, and are only selected by our three methods.

In this part, we analyze the genes selected by the four methods (meta-Half, meta-MCP, meta-SCAD and meta-LASSO) in three lung cancer datasets. According to the network of interactions between genes, among the genes selected by our three methods, we find that some genes are connected to other frequently altered genes in publicly available datasets, and some genes are targeted by certain cancer drugs. Some functions may also need to be verified in the future. Results demonstrate that our three methods have good performance in the high-dimensionality gene expression data with heterogeneity.

Conclusion

With the rapid development of biotechnology and its wide applications, a large number of publicly available gene expression datasets have been produced. However, due to the gene expression datasets have the characteristics of small sample size, high dimensionality and high noise, the application of biostatistics and machine learning methods to analyze gene expression data is a challenging task, such as the low reproducibility of important biomarkers in different studies. The low reproducibility of important biomarkers is mainly caused by the heterogeneity of the different datasets. These problems reveal the complexity of gene expression data and significantly obstruct biotechnology in clinical applications. Meta-analysis is an effective approach to deal with these problems. It plays an important role in summarizing and synthesizing scientific evidence from multiple studies, and provides a more comprehensive understanding of the biological systems, but the current methods have some limitations. The nonconvex regularization method is an effective approach for variable selection developed in recent years. In this paper, we combine the advantages of meta-analysis and the nonconvex regularization method, and propose three novel methods, dubbed as meta-Half, meta-MCP and meta-SCAD, respectively. Through the hierarchical decomposition of coefficients, our methods not only consider the data heterogeneity to maintain the flexibility in selecting variables on different datasets, but also consider the correlation between multiple datasets to improve the ability of identifying important biomarkers. We give the efficient algorithms which apply the nonconvex iterative thresholding algorithms based on approximate message passing (Half-AMP, MCP-AMP and SCAD-AMP) to solve our models and study the theoretical property of meta-Half. The theoretical property analysis of MCP-AMP and SCAD-AMP are the future work. We prove meta-Half possesses sparsity, unbiasedness and oracle properties. Furthermore, we apply our methods to the simulation data and three publicly available lung cancer gene expression datasets, and compare the performance of our methods with other four methods, which are meta-LASSO, composite MCP, group Bridge and group exponential LASSO. Simulation studies demonstrate our methods have the superior performance when data heterogeneity is strong. In the three publicly available lung cancer gene expression datasets, the analysis results show that our three methods have good performance in the gene expression data of small sample size and high dimensionality from different sources (heterogeneity), and the selected important biomarkers have clinical significance. Our methods can also be extended to other areas where datasets are heterogeneous.

Received: 22 August 2019; Accepted: 6 March 2020;
Published online: 01 April 2020

References

1. Barrett, T. *et al*. Ncbi geo: archive for functional genomics data sets—update. *Nucleic acids research* **41**, D991–D995 (2012).
2. Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58**, 267–288 (1996).
3. Fu, W. Penalized regressions: The bridge versus the lasso. *Journal of Computational and Graphical Statistics* **7**, 397–416 (1998).
4. Xu, Z. B., Zhang, H., Wang, Y., Chang, X. Y. & Liang, Y. $l_{1/2}$ regularization. *Science China Information Sciences* **53**, 1159–1169 (2010).
5. Liang, Y. *et al*. Sparse logistic regression with a $l_{1/2}$ penalty for gene selection in cancer classification. *BMC bioinformatics* **14**, 198 (2013).
6. Zhang, C. H. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics* **38**, 894–942 (2010).
7. Fan, J. Q. & Li, R. Z. Statistical challenges with high dimensionality: Feature selection in knowledge discovery,proceeding of the international congress of mathematicians. *European Mathematical Society* 595–622 (2006).
8. Zhang, H., Liang, Y., Xu, Z. & Chang, X. Compressive sensing with noise based on scad penalty. *Acta Mathematica Sinica (in Chinese)* **56**, 767–776 (2013).
9. Zhang, H., Zhang, H. & Gou, M. Convergence analysis of compressive sensing based on scad iterative thresholding algorithm. *Chinese Journal of engineering mathematics* **33**, 243–258 (2016).
10. Yuan, M. & Lin, Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**, 49–67 (2006).
11. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)* **67**, 301–320 (2005).
12. She, Y. *et al*. Thresholding-based iterative selection procedures for model selection and shrinkage. *Electronic Journal of statistics* **3**, 384–415 (2009).
13. Zeng, L. & Xie, J. Group variable selection via *scad*-l_2. *Statistics* **48**, 49–66 (2014).
14. Liu, X.-y. *et al*. Novel regularization method for biomarker selection and cancer classification. *IEEE/ACM transactions on computational biology and bioinformatics* (2019).
15. Rhodes, D. R., Barrette, T. R., Rubin, M. A., Ghosh, D. & Chinnaiyan, A. M. Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer research* **62**, 4427–4433 (2002).
16. DeConde, R. P. *et al*. Combining results of microarray experiments: a rank aggregation approach. *Statistical applications in genetics and molecular biology***5** (2006).
17. Zintzaras, E. & Ioannidis, J. P. Meta-analysis for ranked discovery datasets: theoretical framework and empirical demonstration for microarrays. *Computational biology and chemistry* **32**, 39–47 (2008).
18. Choi, J. K., Yu, U., Kim, S. & Yoo, O. J. Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics* **19**, 84–90 (2003).
19. Grützmann, R. *et al*. Meta-analysis of microarray data on pancreatic cancer defines a set of commonly dysregulated genes. *Oncogene* **24**, 5079 (2005).
20. Han, B. & Eskin, E. Interpreting meta-analyses of genome-wide association studies. *PLoS genetics* **8**, e1002555 (2012).
21. Bhattacharjee, S. *et al*. A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits. *The American Journal of Human Genetics* **90**, 821–835 (2012).
22. Li, J. *et al*. An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies. *The Annals of Applied Statistics* **5**, 994–1019 (2011).
23. Ramasamy, A., Mondry, A., Holmes, C. C. & Altman, D. G. Key issues in conducting a meta-analysis of gene expression microarray datasets. *PLoS medicine* **5** (2008).
24. Hong, F. & Breitling, R. A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. *Bioinformatics* **24**, 374–382 (2008).
25. Tseng, G. C., Ghosh, D. & Feingold, E. Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic acids research* **40**, 3785–3799 (2012).
26. Shen, R., Ghosh, D. & Chinnaiyan, A. M. Prognostic meta-signature of breast cancer developed by two-stage mixture modeling of microarray data. *BMC genomics* **5**, 94 (2004).
27. Conlon, E. M., Song, J. J. & Liu, J. S. Bayesian models for pooling microarray studies with multiple sources of replications. *BMC bioinformatics* **7**, 247 (2006).
28. Choi, H., Shen, R., Chinnaiyan, A. M. & Ghosh, D. A latent variable approach for meta-analysis of gene expression data from multiple microarray experiments. *BMC bioinformatics* **8**, 364 (2007).
29. Scharpf, R. B., Tjelmeland, H., Parmigiani, G. & Nobel, A. B. A bayesian model for cross-study differential gene expression. *Journal of the American Statistical Association* **104**, 1295–1310 (2009).
30. Fan, X. *et al*. Bayesian meta-analysis for identifying periodically expressed genes in fission yeast cell cycle. *The Annals of applied statistics* **4**, 988–1013 (2010).
31. Huo, Z., Song, C. & Tseng, G. Bayesian latent hierarchical model for transcriptomic meta-analysis to detect biomarkers with clustered meta-patterns of differential expression signals. *The annals of applied statistics* **13**, 340 (2019).
32. Rashid, N. U., Li, Q., Yeh, J. J. & Ibrahim, J. G. Modeling between-study heterogeneity for improved replicability in gene signature selection and clinical prediction. *Journal of the American Statistical Association* 1–14 (2019).
33. Zhang, K., Geng, W. & Zhang, S. Network-based logistic regression integration method for biomarker identification. *BMC systems biology* **12**, 135 (2018).
34. Breheny, P. & Huang, J. Penalized methods for bi-level variable selection. *Statistics and its interface* **2**, 369 (2009).
35. Huang, J., Ma, S., Xie, H. & Zhang, C.-H. A group bridge approach for variable selection. *Biometrika* **96**, 339–355 (2009).
36. Breheny, P. The group exponential lasso for bi-level variable selection. *Biometrics* **71**, 731–740 (2015).
37. Kim, S., Jhong, J.-H., Lee, J. & Koo, J.-Y. Meta-analytic support vector machine for integrating multiple omics data. *BioData mining* **10**, 2 (2017).
38. Zhou, N. & Zhu, J. Group variable selection via a hierarchical lasso and its oracle property. *arXiv preprint arXiv:1006.2871* (2010).
39. Li, Q., Wang, S., Huang, C.-C., Yu, M. & Shao, J. Meta-analysis based variable selection for gene expression data. *Biometrics* **70**, 872–880 (2014).
40. Zhao, P. & Yu, B. On model selection consistency of lasso. *Journal of Machine learning research* **7**, 2541–2563 (2006).
41. Chai, H., Li, Z.-n., Meng, D.-y., Xia, L.-y. & Liang, Y. A new semi-supervised learning model combined with cox and sp-aft models in cancer survival analysis. *Scientific Reports***7**, 13053.
42. Fan, J. *et al*. Local partial-likelihood estimation for lifetime data. *The Annals of Statistics* **34**, 290–325 (2006).

43. Breheny, P. & Huang, J. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The annals of applied statistics* **5**, 232 (2011).
44. Fan, J. & Li, R. Variable selection for cox's proportional hazards model and frailty model. *Annals of Statistics* **30**, 74–99 (2002).
45. Jin, Z.-F., Wan, Z., Jiao, Y. & Lu, X. An alternating direction method with continuation for nonconvex low rank minimization. *Journal of Scientific Computing* **66**, 849–869 (2016).
46. Wen, F., Pei, L., Yang, Y., Yu, W. & Liu, P. Efficient and robust recovery of sparse signal and image using generalized nonconvex regularization. *IEEE Transactions on Computational Imaging* **3**, 566–579 (2017).
47. Cui, Z.-X. & Fan, Q. A nonconvex nonsmooth regularization method for compressed sensing and low rank matrix completion. *Digital signal processing* **62**, 101–111 (2017).
48. Huang, X. & Yan, M. Nonconvex penalties with analytical solutions for one-bit compressive sensing. *Signal Processing* **144**, 341–351 (2018).
49. Wen, F. *et al*. Nonconvex regularization-based sparse recovery and demixing with application to color image inpainting. *IEEE Access***5**, 11513–11527.
50. You, J., Jiao, Y., Lu, X. & Zeng, T. A nonconvex model with minimax concave penalty for image restoration. *Journal of Scientific Computing* **78**, 1063–1086 (2019).
51. Li, Z. *et al*. Manifold optimization-based analysis dictionary learning with an $l_{1/2}$-norm regularizer. *Neural Networks* **98**, 212–222 (2018).
52. Zhang, H. & Zhang, H. Approximate message passing algorithm for $l_{1/2}$ regularization. *Science China Information Sciences (in Chinese)* **47**, 58–72 (2017).
53. Zhang, H., Zhang, H., Liang, Y., Yang, Z.-Y. & Ren, Y. Approximate message passing algorithm for nonconvex regularization. *IEEE Access* **7**, 9080–9090 (2019).
54. Landi, M. T. *et al*. Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival. *PloS one* **3**, e1651 (2008).
55. Hou, J. *et al*. Gene expression-based classification of non-small cell lung carcinomas and survival prediction. *PloS one* **5**, e10312 (2010).
56. Lu, T.-P. *et al*. Identification of a novel biomarker, sema5a, for non–small cell lung carcinoma in nonsmoking women. *Cancer Epidemiology and Prevention Biomarkers* **19**, 2590–2597 (2010).
57. Irizarry, R. A. *et al*. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–264 (2003).
58. Gentleman, R. C. *et al*. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology* **5**, R80 (2004).
59. Reguart, N. *et al*. Cloning and characterization of the promoter of human wnt inhibitory factor-1. *Biochemical and biophysical research communications* **323**, 229–234 (2004).
60. Wissmann, C. *et al*. Wif1, a component of the wnt pathway, is down-regulated in prostate, breast, lung, and bladder cancer. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland* **201**, 204–212 (2003).
61. Pannone, G. *et al*. Wnt pathway in oral cancer: epigenetic inactivation of wnt-inhibitors. *Oncology reports* **24**, 1035–1041 (2010).
62. Lin, Y.-C. *et al*. Wnt signaling activation and wif-1 silencing in nasopharyngeal cancer cell lines. *Biochemical and biophysical research communications* **341**, 635–640 (2006).
63. Clément, G. *et al*. Epigenetic alteration of the wnt inhibitory factor-1 promoter occurs early in the carcinogenesis of barrett's esophagus. *Cancer science* **99**, 46–53 (2008).
64. Ai, L. *et al*. Inactivation of wnt inhibitory factor-1 (wif1) expression by epigenetic silencing is a common event in breast cancer. *Carcinogenesis* **27**, 1341–1348 (2006).
65. Park, S. Y. *et al*. Promoter cpg island hypermethylation during breast cancer progression. *Virchows Archiv* **458**, 73–84 (2011).
66. Huang, T. *et al*. Meta-analyses of gene methylation and smoking behavior in non-small cell lung cancer patients. *Scientific reports* **5**, 8897 (2015).
67. Chong, I.-W. *et al*. Great potential of a panel of multiple hmth1, spd, itga11 and col11a1 markers for diagnosis of patients with non-small cell lung cancer. *Oncology reports* **16**, 981–988 (2006).
68. Zhang, W. *et al*. Spp1 and ager as potential prognostic biomarkers for lung adenocarcinoma. *Oncology letters* **15**, 7028–7036 (2018).
69. Pan, Z. *et al*. Long non-coding rna ager-1 functionally upregulates the innate immunity gene ager and approximates its anti-tumor effect in lung cancer. *Molecular carcinogenesis* **57**, 305–318 (2018).
70. Hunninghake, G. M. *et al*. Mmp12, lung function, and copd in high-risk populations. *New England Journal of Medicine* **361**, 2599–2608 (2009).
71. Munthe-Fog, L. *et al*. Immunodeficiency associated with fcn3 mutation and ficolin-3 deficiency. *New England Journal of Medicine* **360**, 2637–2644 (2009).
72. Gao, J. *et al*. Integrative analysis of complex cancer genomics and clinical profiles using the cbioportal. *Sci. Signal.* **6**, pl1–pl1 (2013).
73. Cerami, E. *et al*. The cbio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data (2012).
74. Oros Klein, K. *et al*. Gene coexpression analyses differentiate networks associated with diverse cancers harboring tp53 missense or null mutations. *Frontiers in genetics* **7**, 137 (2016).

Acknowledgements

This work is supported by the Macau Science and Technology Development Funds Grant No. 0055/2018/A2 from the Macau Special Administrative Region of the People's Republic of China.

Author contributions

Hui Zhang and Yong Liang propose the novel methods (meta-Half, meta-MCP, meta-SCAD) and give the efficient algorithms for the novel methods. Hui Zhang and Hai Zhang proved the theoretical property for the proposed methods. Hui Zhang, Shou-Jiang Li and Zi-Yi Yang conceived and conducted the experiment. Yan-Qiong Ren and Liang-Yong Xia provided the real data and analysis the information of biology. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-62473-2.

Correspondence and requests for materials should be addressed to Y.L.

Reprints and permissions information is available at www.nature.com/reprints.

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