Overlapping Group Logistic Regression with Applications to Genetic Pathway Selection

Yaohui Zeng and Patrick Breheny

Department of Biostatistics, University of Iowa, Iowa City, IA, USA.

ABSTRACT: Discovering important genes that account for the phenotype of interest has long been a challenge in genome-wide expression analysis. Analyses such as gene set enrichment analysis (GSEA) that incorporate pathway information have become widespread in hypothesis testing, but pathway-based approaches have been largely absent from regression methods due to the challenges of dealing with overlapping pathways and the resulting lack of available software. The R package grpreg is widely used to fit group lasso and other group-penalized regression models; in this study, we develop an extension, grpregOverlap, to allow for overlapping group structure using a latent variable approach. We compare this approach to the ordinary lasso and to GSEA using both simulated and real data. We find that incorporation of prior pathway information can substantially improve the accuracy of gene expression classifiers, and we shed light on several ways in which hypothesis-testing approaches such as GSEA differ from regression approaches with respect to the analysis of pathway data.

KEYWORDS: overlapping group lasso, penalized logistic regression, gene set enrichment analysis, pathway selection

Introduction

Since the original proposal of the lasso by Tibshirani,[1] penalized regression methods for variable selection in high-dimensional settings have attracted considerable attention in modern statistical research. These methods have been extensively studied in theory and widely applied in practice. Most of the methods focus on selecting individual explanatory variables (or predictors). In many settings, however, predictors possess a group structure. Incorporating this grouping information into the modeling process has the potential to improve both the interpretability and the accuracy of the model.

Consider first the linear regression problem with J nonoverlapping groups,

$$\mathbf{y} = \sum_{j=1}^{J} \mathbf{X}^j \beta^j + \varepsilon \quad (1)$$

where $\mathbf{y}$ is an $n \times 1$ response vector, $\varepsilon \sim N_n(0, \sigma^2 \mathbf{I})$, $\mathbf{X}^j$ is an $n \times K^j$ matrix corresponding to the jth group, K^j is the number of elements in group j, and β^j is the associated $K^j \times 1$ coefficient vector. In equation (1), we take y to be centered, thereby eliminating the need for an intercept. To perform variable selection at the group level, Yuan and Lin[2] proposed the group lasso estimator, defined as the value β minimizing

$$Q(\beta) = L(\beta \mid \mathbf{y}, \mathbf{X}) + \lambda \sum_{j=1}^{J} \sqrt{K^j} \, \|\beta^j\| \quad (2)$$

where $\|\cdot\|$ is the Euclidean (l_2) norm and $L(\beta \mid \mathbf{y}, \mathbf{X})$ is the loss function. For linear regression, the loss function is simply the residual sum of squares, that is, $\|\mathbf{y} - \mathbf{X}\beta\|^2/2n$. For other models, it can be any term that quantifies the fit of the model; for example, Meier et al.[3] extended the group lasso selection to logistic regression by using the negative log-likelihood as the loss function. The second term in equation (2) is called the group lasso penalty, and it leads to variable selection at the group level. That is, the coefficient estimates of the variables in the jth group will be all nonzero if group j is selected and all zero otherwise.

However, an obvious limitation of the group lasso is that it assumes that the groups do not overlap. This introduces a barrier to its application for many problems where variables may be included in more than one group. The application we focus on in this study is the analysis of gene expression profiles, where individual genes can be grouped into pathways in which the collective action of several genes is required for the cell to carry out a complicated function. These pathways generally overlap with each other as one gene can play a role in multiple pathways. Here, $\mathbf{X}^j$ represents the expression data for all genes in the jth pathway, K^j is the number of genes in that pathway, J is the number of pathways, and $\mathbf{y}$ is a vector of phenotypes or clinical responses that we are interested in explaining or predicting using the gene expression data.

Within the hypothesis-testing framework, a number of pathway-based approaches have been proposed for analyzing

gene expression data under the premise that weak expression changes in individual genes are coordinated and can be combined in groups to produce stronger signals.[4] Hence, by incorporating prior pathway information, these approaches aim to identify differentially expressed pathways, instead of individual genes. Compared to traditional single-gene tests, pathway-based tests often lead to higher statistical power and better biological interpretation. Among the pathway-testing approaches, gene set enrichment analysis (GSEA)[5,6] has been widely used. However, the hypothesis-testing framework has certain limitations for pathway analysis, such as the inability to account for the effect of multiple pathways simultaneously, and it is not well suited to using gene expression and pathway data to predict biological outcomes.[7]

On the other hand, pathway-based approaches have been largely absent from regression methods due to the challenges of dealing with overlapping pathways in regression models. Limited attempts have been made to build pathway-based regression models. Wei and Li[8] proposed a nonparametric pathway-based regression using gradient decent boosting. Liu et al.[9] developed a semiparametric regression framework to model the pathway effects using least-squares kernel machines. However, the former is a "black box" approach, and its results are difficult to interpret in terms of how pathways are related to the outcome, while the latter approach only works for estimating the effect of a single pathway and cannot model multiple pathways simultaneously.

In this study, we formulate the overlapping group logistic regression model based on the latent group lasso approach,[10] making it applicable to perform pathway selection under the general linear modeling framework. This approach naturally preserves the straightforward interpretation of regression coefficients and offers the ability to scale up to model hundreds of overlapping pathways simultaneously in high-dimensional settings.

We also conduct a systematic comparison of this overlapping group lasso (OGLasso) approach with both the ordinary lasso and GSEA via both simulation and real-data studies. Our aim is to demonstrate the fundamental differences between hypothesis-testing approaches and regression models with respect to their implications for pathway selection. Thus, although a variety of extensions and refinements of GSEA have been proposed, such as GSEAlm,[11] ROAST,[12] and npGSEA,[13] we restrict our attention here to GSEA, the most well-known and widely used method in this group.

Finally, we provide a publicly available implementation of the OGLasso method described in this study through the R package grpregOverlap. This package serves as an extension of the R package grpreg, which provides a variety of functions for fitting penalized regression models involving grouped predictors but requires those groups to be nonoverlapping.

The rest of the study is organized as follows. In the "Methods" section, we review the OGLasso approach and construct the OGLasso model. In addition, we give a brief introduction to GSEA, along with some discussions. In the "Simulation studies" section, we first compare the ordinary lasso and OGLasso in terms of model accuracy with simulated data. We then examine the group selection accuracy of OGLasso and GSEA under different simulation settings. In addition, we provide two real-data studies in the "Real-data studies" section. We conclude the study with final discussions in "Discussion" section.

Methods

Overlapping group lasso. Suppose the p predictors $\{x_1, x_2, \ldots, x_p\}$ are assigned into J possibly overlapping groups (ie, a given predictor x_i may be included in more than one group). The group lasso estimator (2) does not necessarily select groups in this overlapping setting. For example, suppose $p = 3$ and $J = 2$, with one covariate shared between the two groups: group "A" and group "B", with group A truly related to the outcome. If group B is not selected, then all of its coefficients are zero, even though one coefficient also appears in group A. Thus, group A is only partially selected. This problem is greatly exacerbated as the groups grow in size and complexity and is described in greater detail in Jenatton et al.[14]

To select entire groups of covariates in the overlapping setting, Jacob et al.[10] proposed the OGLasso, formulated as

$$\min_{\beta} Q(\beta) = L(\beta \mid \mathbf{y}, \mathbf{X}) + \lambda \sum_{j=1}^{J} \sqrt{K^j} \left\| \gamma^j \right\|$$

$$\text{subject to } \beta = \sum_{j=1}^{J} \gamma^j, \tag{3}$$

where $\{\gamma^j\}_{j=1}^J$ are J so-called latent coefficient vectors. The collection of latent vectors $\gamma^i = \left(\gamma_1^j, \gamma_2^j, \ldots, \gamma_p^j\right)'$ satisfies $\sum_{j=1}^{J} \gamma^j = \beta$, if x_k does not belong to group j, with $\gamma_k^j \neq 0$ otherwise.

The idea of model (3) is to decompose the original coefficient vector into a sum of group-specific latent effects. This decomposition allows us to apply the group lasso penalty to the latent vectors $\{\gamma^j\}_{j=1}^J$, which do not overlap, instead of the original, overlapping coefficients. Consequently, when a latent vector γ^j is selected, all covariates in group j will be selected, even if some members of the group are also involved in unselected groups.

It is worth clarifying the exact meaning of "latent" here. It is not the case that the grouping structure is unobservable – we are considering situations in which the grouping is known in advance. For example, much is already known about how genes are organized into pathways; we want to leverage this information to produce more accurate models.

Rather, what is latent is the decomposition of the effect of each feature into the groups it belongs to. For example, suppose gene X belongs to pathways A and B. It may be that gene X's effect on the response is mediated entirely through pathway A and that its membership in pathway B is irrelevant. This parsing of the effect of the genes into pathways is the

latent aspect of the problem that cannot be observed directly – we can only observe changes in the expression of gene X, not whether expression changed in order to produce an effect in pathway A or B.

Figure 1 illustrates the coefficient decomposition mechanism described in equation (3). Suppose that there are four variables x_1, x_2, x_3, x_4 that are included in four groups, $S^1 = \{x_1, x_2\}$, $S^2 = \{x_2, x_3\}$, $S^3 = \{x_1, x_3\}$, and $S^4 = \{x_3, x_4\}$, where S^j denote the set of variables in group j. Since x_1 is in both groups 1 and 3, β_1 is thus decomposed into $\gamma_1^1 + \gamma_1^3$. Likewise, β_3 is decomposed into $\gamma_3^2 + \gamma_3^3 + \gamma_3^4$, and so on. Suppose group 1 is the sole truly nonzero group in this example. The OGLasso model can select γ^1, thereby indirectly selecting β_1 and β_2 and eliminating β_3 and β_4 since they do not appear in group 1. Note that the original group lasso cannot accomplish this – if group 3 is eliminated, then predictor 1 is eliminated as well since it belongs to group 3.

Based on the coefficient decomposition, model (3) can be transformed into a new minimization problem[15] with respect to γ:

$$\min_{\gamma} Q(\gamma) = L(\gamma \mid \mathbf{y}, \tilde{\mathbf{X}}) + \lambda \sum_{j=1}^{J} \sqrt{K^j} \, \|\gamma^j\|. \qquad (4)$$

Here, γ in principle consists of all elements of γ^j, although in practice one can leave off the zero elements as they have no effect on the objective function. The new design matrix $\tilde{\mathbf{X}}$ is constructed by duplicating the columns of overlapped variables in the raw design matrix $\mathbf{X}$, where appropriate, to match the elements of γ. The equivalence of the loss functions $L(\beta \mid \mathbf{y}, \mathbf{X})$ and $L(\gamma \mid \mathbf{y}, \tilde{\mathbf{X}})$ can be seen by observing that $\mathbf{X}\beta = \mathbf{X}\sum_j \gamma^j = \tilde{\mathbf{X}}\gamma$.

The implication of equation (4) is that the OGLasso problem is equivalent to a classical group lasso in an expanded, nonoverlapping space. This is of considerable practical convenience, as it allows us to solve equation (4) using computationally efficient algorithms that have previously been developed for the group lasso.[16]

Overlapping group logistic regression. It is relatively straightforward to extend equation (4) to models other than

Figure 1. The coefficient decomposition of overlapping group lasso.

linear regression; in this section, we describe its application to penalized logistic regression in the presence of overlapping groups. Here, $\mathbf{y}$ is the response vector of binary entries, and the intercept β_0 cannot be removed by centering $\mathbf{y}$. For convenience, we assume that the first column of the design matrix $\mathbf{X}$ is the unpenalized column of 1s for the intercept β_0 and denote $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ip})'$ for $i = 1, \ldots, n$. Correspondingly, we denote $\beta = (\beta_0, \beta_1, \ldots, \beta_p)'$. The logistic regression model is

$$\Pr\left(y_i = 1 \mid \mathbf{x}_i\right) = \pi_i = \frac{\exp(\mathbf{x}_i'\beta)}{1 + \exp(\mathbf{x}_i'\beta)}. \qquad (5)$$

The corresponding loss function is the (scaled) negative log-likelihood function,

$$L(\beta \mid \mathbf{y}, \mathbf{X}) = -\frac{1}{n}\sum_{i=1}^{n}\left\{y_i(\mathbf{x}_i'\beta) - \log\left(1 + \exp(\mathbf{x}_i'\beta)\right)\right\}.$$

We can then duplicate the columns of the overlapped covariates, expanding the design matrix to $\tilde{\mathbf{X}}$ as described previously, and construct the overlapping group logistic regression model in the same fashion as model (4), with

$$L\left(\gamma \mid \mathbf{y}, \tilde{\mathbf{X}}\right) = \frac{1}{n}\sum_{i=1}^{n}\left\{y_i(\tilde{\mathbf{x}}_i'\gamma) - \log\left(1 + \exp(\tilde{\mathbf{x}}_i'\gamma)\right)\right\}, \qquad (6)$$

where $\tilde{x}_i'$ is the ith row of the expanded design matrix $\tilde{\mathbf{X}}$, and the first element of γ is the unpenalized intercept β_0.

Gene set enrichment analysis. Among the hypothesis-testing approaches for pathway selection, GSEA stands out due to its relative simplicity and for preserving the gene–gene dependencies that occur in real biological data.[17]

The procedure of GSEA[6] starts with ranking the p genes by the correlation, r_j, between each gene and the phenotype. Then a test statistic, the enrichment score (ES), is calculated for each gene set by walking down the ranked gene list and accumulating the correlation information: increasing ES by $|r_i|^\alpha / \sum_{j \in s} |r_j|^\alpha$ if gene i is included in gene set S; decreasing ES by $1/(p - |S|)$ otherwise. Here, α is a prespecified exponent parameter. When $\alpha = 1$, ES corresponds to the normalized Kolmogorov–Smirnov statistic. Next, the significance level of the ES is assessed by a permutation test. Finally, the significance of the gene sets is determined by controlling the false discovery rate (FDR).

Though widely used, GSEA also has several limitations. First, GSEA may be biased in favor of larger gene sets by systematically assigning those gene sets higher ES[18]; second, it implicitly assumes that genes within the same gene set show coordinated (ie, either all positive or all negative) associations with the phenotype, making it less likely to detect sets in which the genes are heterogeneous with respect to the direction of association with the phenotype.[19]
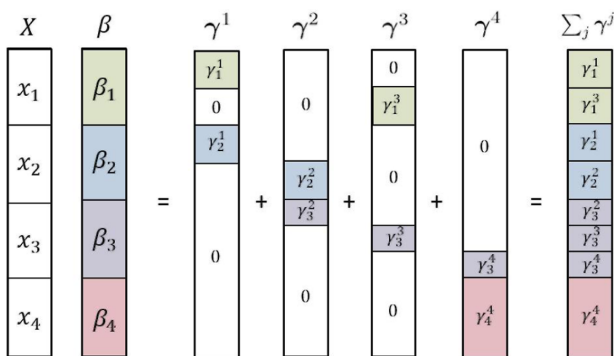
There are inherent differences between GSEA and the proposed overlapping group logistic regression method in the sense that GSEA treats the phenotype as fixed and gene expression as random, while regression-based methods do the opposite. Thus, GSEA tends to be more appropriate in settings where the phenotype can be directly manipulated by the experiment (eg, knockout mice), while regression is more appropriate in observational settings (eg, predicting patient outcomes). Nevertheless, there are many situations in which either method could reasonably be used, and therefore, it is of interest to compare the selection properties of the two approaches.

Simulation Studies

In all the simulation studies, we use the term "null group" to denote a group whose coefficients are all equal to zero in the true model and "true group" to denote a group with all non-zero coefficients in the true model. In addition, we refer to $||\gamma^j||$ as the effect size of group j and γ_k^j as the latent effect of covariate k in group j.

OGLasso versus ordinary lasso. We start by comparing the OGLasso with the ordinary lasso in terms of estimation and prediction accuracy. We use root mean squared error (RMSE) to measure estimation accuracy and misclassification error (ME) to measure prediction accuracy, defined as follows:

$$\text{RMSE} = \sqrt{\frac{1}{p}\sum_{k=1}^{p}\left(\beta_k - \hat{\beta}_k\right)^2}; \quad \text{ME} = \frac{\#\,\text{incorrectly classified}}{\text{Sample size}}$$

It should be noted that we compute ME based on a new response vector generated by the same design matrix for each replication. Specifically, given a design matrix $\mathbf{X}$, two response vectors $\mathbf{y}$ and $\mathbf{y}^*$ are simulated. The data $\{\mathbf{X}, \mathbf{y}\}$ is used to fit the model, and its prediction accuracy is tested on data $\{\mathbf{X}, \mathbf{y}^*\}$.

We consider two simulations with different settings described as follows.

Setting 1: Synthetic data. We begin with synthetic data where there are 15 groups of covariates. All covariate values are simulated independently from a standard Gaussian distribution. The group sizes and overlap structure are presented below.

```
ID:    1  2   3   4   5   6   7   8   9   10  11  12   13  14   15
Size:  10 10  10  10  10  10  10  10  10  10  10  10   10  10   10
          ⌣         ⌣         ⌣         ⌣          ⌣
          3         3         3         3          3
```

The number underneath the brace is the number of members shared between those two groups. For example, group 1 contains 10 members, as does group 2, but the two groups contain only 17 unique predictors, as three predictors are present in both groups. As a result, the total dimension in this setting is $p = 135$. By design, groups 1, 4, 7, 10, and 13 are set to be true groups. The sample size is set to be $n = 50$ to be consistent with that in Setting 2 as below.

Setting 2: Real data. For this simulation, a real gene expression profile data set in the p53 study[7] is used as the design matrix to mimic the complicated correlation and overlapping structures in real biomedical applications. This design matrix is fixed for each independent replication. Here, the sample size $n = 50$, the number of genes $p = 4301$, and the number of pathways (groups) is 308; a more detailed description of the study is given in the "Real-data studies" section. We chose five pathways, with sizes 15, 16, 20, 26, and 40, to represent the true groups in this simulation. The number of overlaps between the five pathways ranges from 0 to 9.

In both of the above-mentioned two settings, the true group effect sizes of each of the five true groups are set to be equal, and the latent effects are also set to be equal within each true group. In this way, the true coefficient vector is uniquely specified. Then given the design matrix, the responses are generated according to equation (5) for each independent replication. The true group effect size is varied from 1 to 5 to simulate different magnitudes of signals.

Figure 2 illustrates the estimation and prediction accuracy of the proposed grouped variable selection method, as compared to the ordinary lasso, for both settings. The top two panels show results for the synthetic data simulation, while the bottom two panels are for the real-data simulation. The left panels illustrate the median RMSE relative to ordinary lasso over 500 replications, while the right panels compare the methods in terms of ME. OGLasso consistently achieves a lower median RMSE than that of the lasso in both synthetic and real-data simulations. As expected, the ME by both methods decreases as the coefficient magnitude increases. More interestingly, the ME by OGLasso can be substantially lower than that of ordinary lasso. In the synthetic data simulation, for example, the ME by OGLasso is 7% lower than that of ordinary lasso when the group effect size is 5. The two methods are more similar in terms of predictive accuracy on the real data, where the dimensionality is much higher and correlation structure more complicated. Nevertheless, the prediction accuracy can still be improved by around 2% with OGLasso compared to ordinary lasso.

OGLasso versus GSEA. In this section, we use simulated data to compare the selection properties of the OGLasso against GSEA in a variety of different settings. Because OGLasso and GSEA do not estimate the same quantities and GSEA does not produce predictions, the only way to compare them is with respect to selection accuracy. To ensure a fair comparison, we use each method to select a fixed number of groups. We then evaluate the group selection accuracy by the true discovery rate (TDR):

$$\text{TDR} = \frac{\#\,\text{of true groups selected}}{\#\,\text{of groups selected}},$$

where the # of groups selected was fixed at 5 (ie, each method was used to identify the five most important-looking
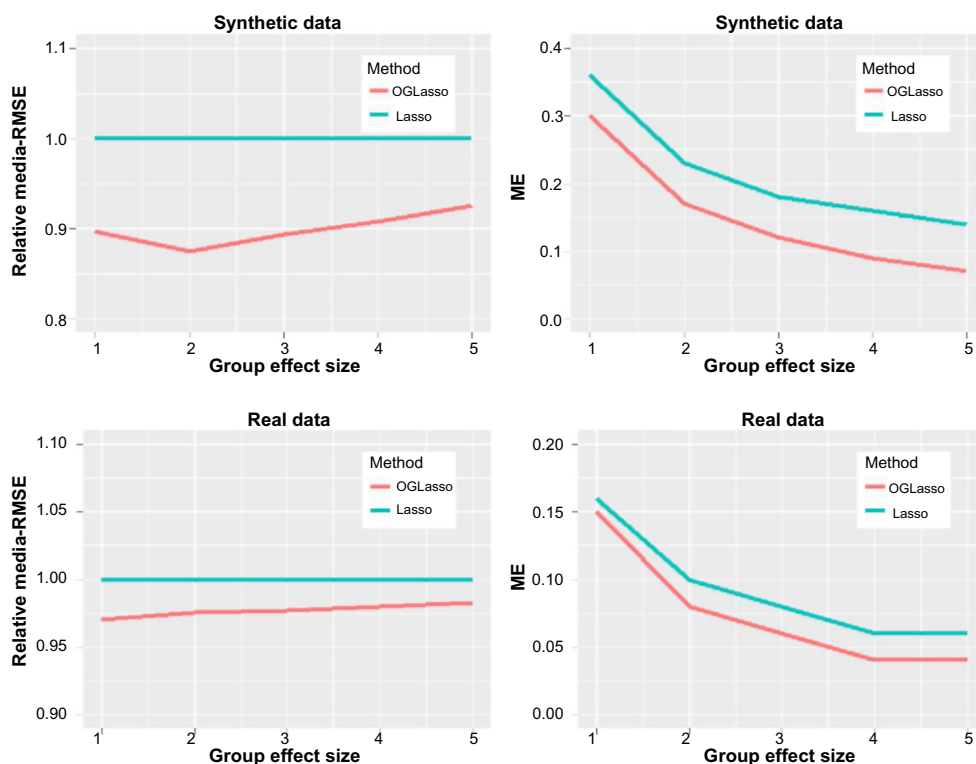
Figure 2. Accuracy of OGLasso and ordinary lasso with respect to the magnitude of the group effect size. Top two panels summarize results for the synthetic data simulation, while bottom two panels are for the real-data simulation.
Notes: Left panels: Median RMSE relative to ordinary lasso over 500 replications. Right panels: Median ME over 500 replications.

groups). In each of the following simulations, the results are based on the sample size $n = 100$ and averaged over 500 independent replications.

Setting 3: Unequal group size. First, we investigate the performance of the two approaches when group sizes are unequal. In this simulation, the design matrix consists of 15 groups with all covariate values simulated independently from a standard Gaussian distribution. The group sizes and overlap structure are shown below.

```
ID:    1  2  3   4  5  6   7  8  9   10  11  12   13  14  15
Size : 3  3  3   6  6  6   9  9  9   15  15  15   24  24  24
          ⎵          ⎵          ⎵           ⎵             ⎵
          1          2          3           5             8
```

The overlap here is designed to be one-third of the size of overlapped groups. As a result, the total dimension in this setting is $p = 152$. Moreover, groups 1, 4, 7, 10, and 13 are set to be true groups with $||\gamma^j|| = 5$, and the others are null groups with $j = 0$. The latent effects are again set to be equal within each true group.

Table 1 summarizes the mean TDR and size of selected groups for the OGLasso and GSEA over 500 replications. The two methods are comparable in terms of TDR, while the average size of selected groups from GSEA is slightly larger.

The proportion of each group selected is depicted in Figure 3. OGLasso tends to favor groups with smaller size,

Table 1. The mean (standard error) of TDR and average size of selected groups of OGLasso and GSEA over 500 replications.

METHOD	TDR	AVERAGE SIZE
OGLasso	0.77 (0.01)	8.8 (0.1)
GSEA	0.79 (0.01)	11.0 (0.1)

while GSEA has roughly an equal probability of selecting a true group regardless of its size. This is understandable, as regression-based methods have a built-in mechanism for encouraging parsimony, unlike GSEA. Whether this preference for smaller groups is desirable depends on the application and the scientific goals of the study.

Setting 4: Heterogeneous gene effects. Previous studies have shown that GSEA is less likely to detect sets of genes containing both positive and negative associations with the phenotype.[19] This is because, by pooling together correlations, GSEA assumes that the genes in a set have a coordinated effect – that is, that they all act in the same direction. In this simulation, we examine this aspect of GSEA further and demonstrate that the exhibition of heterogeneous effects among genes in a set deteriorates the statistical power of GSEA.

We employ the same configuration as in Setting 1 of the "OGLasso versus ordinary lasso" section for the design matrix (except that the sample size here is $n = 100$) but specify the
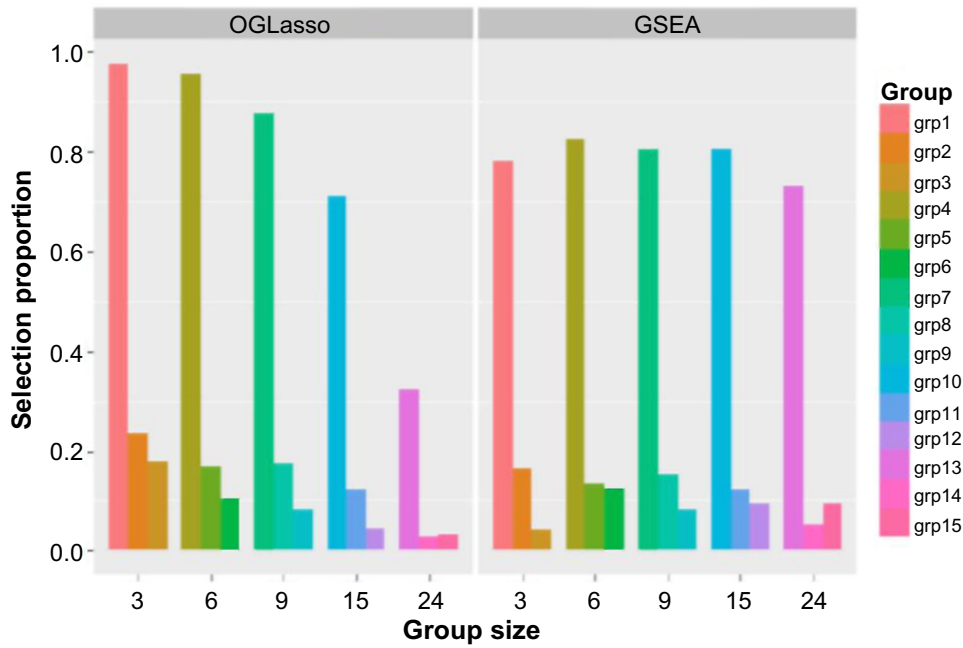
Figure 3. Comparison of the proportion of each group being selected over 500 replications for Setting 3. In each panel, the 15 vertical bars from left to right correspond to group 1 to group 15, with groups of the same dimension clustered together. The height of the bar represents the proportion of replications in which the associated group was selected by the method.

true coefficient values in a different manner. Specifically, we draw the true latent coefficients γ_k^j for each true group from a $\text{Unif}(\mu - \sigma, \mu + \sigma)$ distribution. Here, σ is a parameter that controls the degree of heterogeneity (or variability) of the gene effects. The larger σ is, the more heterogeneous the effects are. In this simulation, we vary σ to examine the effect of heterogeneity on the TDR of each method.

On a technical note, it must be pointed out that varying σ will also change the group effect, $||\gamma^j||$. To suppress this possibly confounding effect, we adjust μ along with σ so that the (root mean square) group effect size remains constant. Specifically, choosing $\mu = \sqrt{\frac{5}{2} - \frac{1}{3}\sigma^2}$ results in a constant $\sqrt{\mathbb{E}\left(|| \gamma^j ||^2\right)} = 5$ for all values of σ.

Figure 4 compares OGLasso and GSEA in terms of TDR as a function of σ. OGLasso is essentially unaffected by heterogeneity: it detects approximately four of the five true groups regardless of the magnitude of heterogeneity. In contrast, the TDR of GSEA decreases as σ increases. This effect is apparent even when all genes in a group have a consistent direction, although the effect is much more significant for $\sigma > 1.37$, at which point it is possible for genes within a true group to have opposite directions.

Setting 5: Correlation among genes. In this simulation, we assess how correlation among genes affects group selection. We use the same settings for the groups and overlap structure as in Setting 1, where $p = 135$. The true coefficients are fixed so that the group effect $||\gamma^j|| = 5$ for each true group and that all latent effects γ_k^j within a true group j are equal.

In this setting, covariates are no longer independent but are instead simulated from a multivariate Gaussian distribution with mean 0 and variance Ω. We impose a block-diagonal covariance structure with five compound-symmetric blocks, as shown below:

$$\Omega = \begin{bmatrix} \Sigma & & & & \\ & \Sigma & & & \\ & & \Sigma & & \\ & & & \Sigma & \\ & & & & \Sigma \end{bmatrix}, \text{ where } \Sigma = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}_{27 \times 27}$$

Figure 5 compares OGLasso and GSEA in terms of TDR as a function of pairwise correlation ρ. As expected, TDR of both methods decreases as the correlation among genes increases. However, GSEA is much more strongly affected by correlation than the OGLasso. For example, as ρ increases from 0 to 0.1, the TDR of GSEA drops from around 0.8 to 0.45. This ability – to adjust for correlation between pathways – is one of the primary potential advantages of a regression-based approach over a hypothesis-testing approach, which is limited to considering a single pathway at a time.

Real-Data Studies

In this section, we analyze the data from two gene expression studies reported in Subramanian et al.[6], one involving the
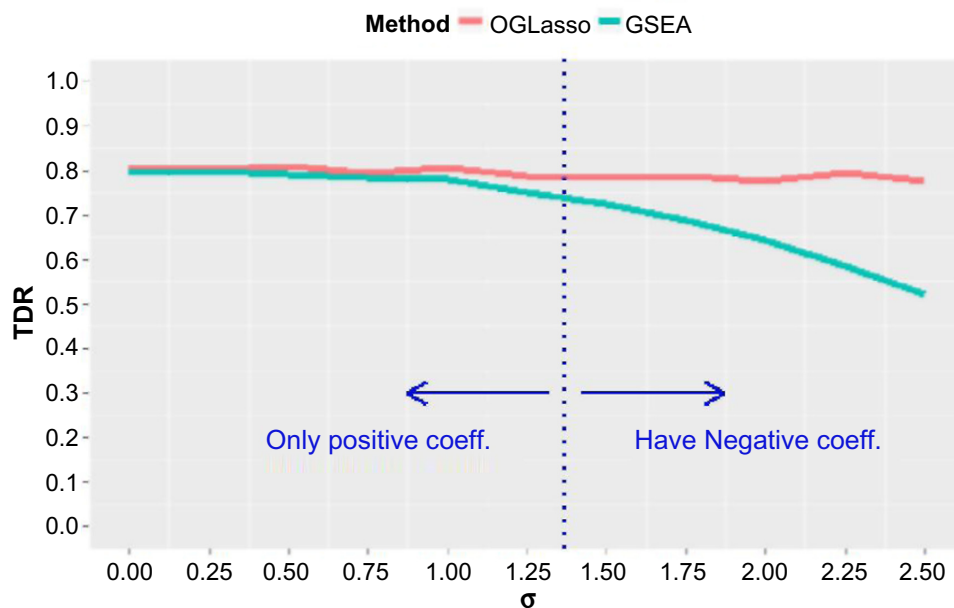
Figure 4. Comparison of OGLasso and GSEA in terms of TDR as a function of heterogeneity parameter σ. The blue dotted line indicates $\sigma = 1.37$, after which negative coefficients can occur by design. The mean values over 500 replications are displayed.

mutational status of p53 in cell lines and the other involving the prognosis of lung cancer patients.

The p53 study aims to identify pathways that correlated with the mutational status of the gene p53, which regulates gene expression in response to various signals of cellular stress. The p53 data[20] consist of 50 cell lines, 17 of which are classified as normal and 33 of which carry mutations in the p53 gene. To be consistent with the analysis in Subramanian et al.[7], 308 gene sets that have size between 15 and 500 are included in our analysis. These gene sets contain a total of 4301 genes.

The lung cancer data[21] contains gene expression profiles in 86 tumor samples, of which 24 are classified as "poor" outcome and the remaining as "good" outcome. The data sets are preprocessed in the same fashion as in the p53 study, resulting in 258 gene sets that contain a total of 3256 genes. Compared to the p53 data, the lung cancer data show much weaker signals: no individual gene is found to be significant in a conventional single-gene analysis.

We first compare the OGLasso to the ordinary lasso in terms of prediction accuracy. For each method, 10-fold cross-validation was used to choose the regularization parameter λ.
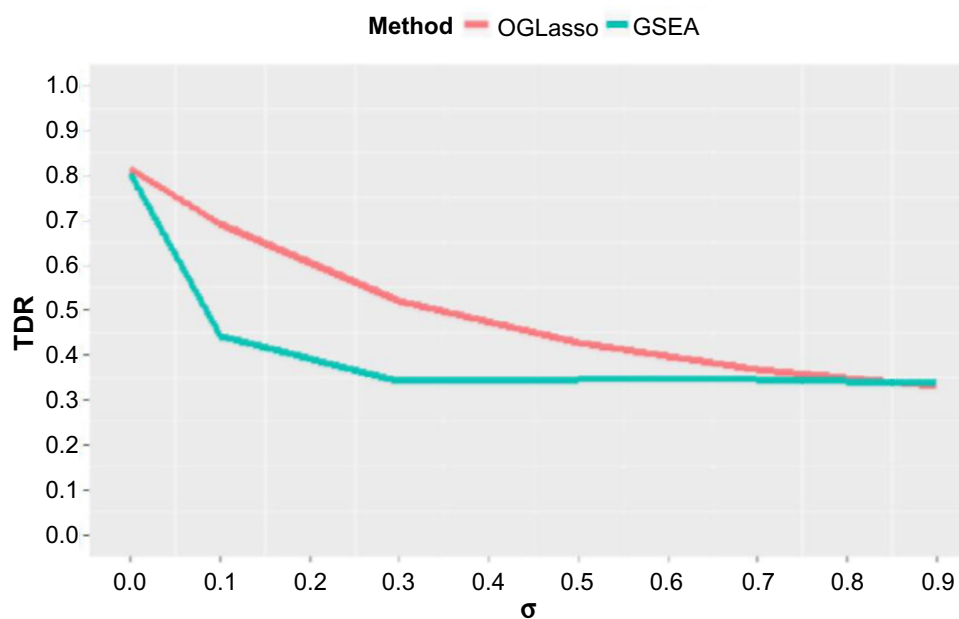

Figure 5. Comparison of OGLasso and GSEA in terms of TDR as a function of pairwise correlation ρ. The mean values over 500 replications are displayed.

Indeed, as shown in Table 2, the incorporation of pathway information into the regression model produces more accurate predictions in both studies. In the p53 study, where the signals are relatively strong, the ME of the ordinary lasso is 8% lower than that of the intercept-only model. However, the OGLasso can further lower the error by an additional 6%. In the lung cancer study, due to a small signal-to-noise ratio, the ordinary lasso performed even worse than the intercept-only model. However, the OGLasso was able to improve on the predictions of the intercept-only model, albeit only slightly.

We now turn to comparing the pathways selected by OGLasso and GSEA. Again, 10-fold cross-validation is used to select A for OGLasso, while a FDR cutoff of 0.25 was used to select pathways with GSEA. Table 3 lists the number of pathways, the number of total genes, and the number of unique genes in those selected pathways by OGLasso and GSEA. In both studies, GSEA selects more pathways than OGLasso, especially in the lung cancer study (21 vs. 3). Moreover, in agreement with our earlier simulation results, GSEA selects substantially larger pathways than OGlasso. For example, in the lung cancer study, the average pathway size for GSEA is 820/21 = 39 genes, while the average size for OGLasso is only 51/3 = 17 genes.

Table 4 presents a summary of pathway selection results in the p53 study that sheds light on the nature of the pathways selected by each approach; an equivalent table for the lung cancer study is included in the Supplementary Table 1. Naturally, both approaches identify the "p53 Pathway" as being associated with p53 mutation status. However, GSEA also selects pathways "radiation_sensitivity", which shares nine genes with "p53 Pathway", "p53hypoxiaPathway" (seven shared genes), and "P53_UP" (five shared genes). From a regression perspective, these four pathways are largely redundant, and the three unselected pathways carry no additional useful information beyond that already contained in the p53 pathway. On the other hand, OGLasso selects one pathway, "cklPathway", not

identified by GSEA. Although the ckl pathway has a weaker marginal relationship with p53 mutation status than the hsp27 and p53 pathways, the information it contains is largely independent of the other pathways included in the model (no overlaps with the hsp27 and p53 pathways), potentially shedding light on novel p53 relationships that would not be apparent from the GSEA approach.

The biological interpretations of the pathways selected in the lung cancer study are less clear due to the weaker signals and more complicated biological outcome. Nevertheless, there are some interesting similarities and differences here as well. Of the three gene sets selected by OGLasso, one pathway (ceramide) is also selected by GSEA. The other two gene sets, although not selected by GSEA, contain a fair amount of overlap with GSEA-selected sets. For example, OGLasso selects the Fas pathway, while GSEA selects the p53 pathway. However, both pathways are involved in apoptosis, and six genes are shared between the two pathways. The simulation studies of the "OGLasso versus GSEA" section suggest that differences in the size, heterogeneity, or correlation patterns of these pathways provide an explanation for why OGLasso prefers the Fas pathway to the p53 pathway.

Discussion

Pathway-based approaches for analyzing gene expression data have become increasingly popular in recent years. Most methods have approached the problem from a multiple hypotheses testing perspective. However, the latent group lasso approach proposed by Jacob et al.[10] allows the incorporation of pathway information into regression models as well.

Table 4. The p53 study: pathways selected by OGLasso and GSEA with FDR ≤ 0.25.

PATHWAY LABEL	SIZE	FDR Q VALUE	GSEA	OGLASSO
hsp27 pathway	15	<0.001	✓	✓
p53hypoxia pathway	20	<0.001	✓	–
p53 pathway	16	<0.001	✓	✓
Radiation sensitivity	26	0.078	✓	–
p53 UP	40	0.013	✓	–
rasPathway	22	0.171	✓	–
ck1Pathway	15	0.500	–	✓

Table 2. Real-data studies: 10-fold cross-validated ME for different models. "Baseline" is the intercept-only model.

METHOD	p53 STUDY	LUNG CANCER STUDY
Baseline	0.34	0.28
Lasso	0.26	0.30
OGLasso	0.20	0.27

Table 3. Real-data studies: number of selected pathways (# pathways), number of total genes (# total genes), and number of unique genes (# unique genes) in selected pathways by OGLasso and GSEA.

p53 STUDY				LUNG CANCER STUDY		
METHOD	# PATHWAYS	# TOTAL GENES	# UNIQUE GENES	# PATHWAYS	# TOTAL GENES	# UNIQUE GENES
OGLasso	3	46	44	3	51	50
GSEA	6	139	105	21	820	629

Regression models offer two distinct advantages in this setting. First, they provide a direct method for using the entirety of the pathway information to predict biological responses. Second, they make no assumptions about the distribution of the expression data itself. For this reason, the methods we develop here can be applied to any gene expression study, regardless of the technology used for quantification (qPCR, microarrays, RNA-Seq, etc.).

In this study, we present evidence that the incorporation of pathway information can substantially improve the accuracy of gene expression classifiers. Furthermore, we provide open-source software, publicly available at cran.r-project.org, for fitting the OGLasso models described in this study. By retaining the underlying framework of regression modeling, this approach can be applied to both continuous and binary outcomes, and it is straightforward to extend the idea to Cox proportional hazards models for time-to-event outcomes.

Finally, this study provides, to our knowledge, the only systematic comparison of OGLasso methods with the GSEA approach. There is a fundamental difference between the two methods: GSEA carries out independent tests of each gene set, while the OGLasso is a regression method that considers the effect of all pathways simultaneously. We show that, while there is broad agreement between the two, substantial differences between the approaches may arise with respect to pathway size, heterogeneity of gene effects, and correlations between gene sets. These factors, along with the goals and design of the study, should be carefully considered when deciding upon an approach to data analysis.

Author Contributions

Conceived and designed the experiments: YZ, PB. Analyzed the data: YZ. Wrote the first draft of the manuscript: YZ. Contributed to the writing of the manuscript: YZ, PB. Agree with manuscript results and conclusions: YZ, PB. Jointly developed the structure and arguments for the paper: YZ, PB. Made critical revisions and approved final version: YZ, PB. Both authors reviewed and approved of the final manuscript.

Supplementary Material

Supplementary Table 1. The lung cancer study: pathways selected by OGLasso and GSEA with FDR $\leq$ 0.25.

REFERENCES

1. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc B Stat Methodol*. 1996;58(1):267–88.
2. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. *J R Stat Soc Series B Stat Methodol*. 2006;68(1):49–67.
3. Meier L, van de Geer S, Bhlmann P. The group lasso for logistic regression. *J R Stat Soc B Stat Methodol*. 2008;70(1):53–71.
4. Nam D, Kim SY. Gene-set approach for expression pattern analysis. *Brief Bioinform*. 2008;9(3):189–97.
5. Mootha VK, Lindgren CM, Eriksson KF, et al. Pgc-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet*. 2003;34(3):267–73.
6. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005;102(43):15545–50.
7. Goeman JJ, Bühlmann P. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*. 2007;23(8):980–7.
8. Wei Z, Li H. Nonparametric pathway-based regression models for analysis of genomic data. *Biostatistics*. 2007;8(2):265–84.
9. Liu D, Lin X, Ghosh D. Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. *Biometrics*. 2007;63(4):1079–88.
10. Jacob L, Obozinski G, Vert JP. Group lasso with overlap and graph lasso. In: Proceedings of the 26th Annual International Conference on Machine Learning. ACM, Montreal, Canada; 2009:433–40.
11. Oron AP, Jiang Z, Gentleman R. Gene set enrichment analysis using linear models and diagnostics. *Bioinformatics*. 2008;24(22):2586–91.
12. Wu D, Lim E, Vaillant F, Asselin-Labat ML, Visvader JE, Smyth GK. Roast: rotation gene set tests for complex microarray experiments. *Bioinformatics*. 2010;26(17):2176–82.
13. Larson JL, Owen AB. Moment based gene set tests. *BMC Bioinformatics*. 2015;16(1):1.
14. Jenatton R, Audibert JY, Bach F. Structured variable selection with sparsity-inducing norms. *J Mach Learn Res*. 2011;12:2777–824.
15. Obozinski G, Jacob L, Vert JP. Group Lasso with Overlaps: the Latent Group Lasso approach. [Research Report] 2011, pp.60. <inria-00628498>.
16. Breheny P, Huang J. Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Stat Comput*. 2015;25(2):173–87.
17. Tamayo P, Steinhardt G, Liberzon A, Mesirov JP. The limitations of simple gene set enrichment analysis assuming gene independence. *Stat Methods Med Res*. 2016;25(1):472–87.
18. Damian D, Gorfine M. Statistical concerns about the GSEA procedure. *Nat Genet*. 2004;36(7):663.
19. Dinu I, Potter JD, Mueller T, et al. Improving gene set analysis of microarray data by SAM-GS. *BMC Bioinformatics*. 2007;8(1):242.
20. Olivier M, Eeles R, Hollstein M, Khan MA, Harris CC, Hainaut P. The iarc tp53 database: new online mutation analysis and recommendations to users. *Hum Mutat*. 2002;19(6):607–14.
21. Beer DG, Kardia SL, Huang CC, et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat Med*. 2002;8(8):816–24.