Selection probability of multivariate regularization to identify pleiotropic variants in genetic association studies

Kipoong Kim\textsuperscript{a}, Hokeun Sun\textsuperscript{1,a}

\textsuperscript{a}Department of Statistics, Pusan National University, Korea

Abstract

In genetic association studies, pleiotropy is a phenomenon where a variant or a genetic region affects multiple traits or diseases. There have been many studies identifying cross-phenotype genetic associations. But, most of statistical approaches for detection of pleiotropy are based on individual tests where a single variant association with multiple traits is tested one at a time. These approaches fail to account for relations among correlated variants. Recently, multivariate regularization methods have been proposed to detect pleiotropy in analysis of high-dimensional genomic data. However, they suffer a problem of tuning parameter selection, which often results in either too many false positives or too small true positives. In this article, we applied selection probability to multivariate regularization methods in order to identify pleiotropic variants associated with multiple phenotypes. Selection probability was applied to individual elastic-net, unified elastic-net and multi-response elastic-net regularization methods. In simulation studies, selection performance of three multivariate regularization methods was evaluated when the total number of phenotypes, the number of phenotypes associated with a variant, and correlations among phenotypes are different. We also applied the regularization methods to a wild bean dataset consisting of 169,028 variants and 17 phenotypes.

Keywords: pleiotropy, regularization methods, selection probability, high-dimensional data

1. Introduction

In recent genetic association studies, advanced technologies and large-scale study design have provided not only high-throughput genomic data, but also a large number of phenotype data. Multiple-phenotype studies could enable us to understand underpinning of disease comorbidity and to prevent multiple diseases simultaneously. They are also expected to discover the missing heritability of complex diseases. Pleiotropy is a phenomenon where a variant or a genetic region affects multiple traits or diseases. In studies of pleiotropy, the main goal is to identify pleiotropic variants that are associated with multiple phenotypes. Solovieff \textit{et al.} (2013) described three different types of pleiotropy. The first one is biological pleiotropy where a variant has a direct effect on each trait. The second is mediated pleiotropy where multiple phenotypes are correlated with each other so that a variant can be considered to be associated with more than one phenotype. The last is spurious pleiotropy which is caused by study designs, confounding, and strong linkage disequilibrium among variants.

Many statistical methods have been proposed to identify biologically or mediately pleiotropic variants that simultaneously affect multiple traits (Bhattacharjee \textit{et al.}, 2012; van der Sluis \textit{et al.}, 2013; Schaid \textit{et al.}, 2016; Broadaway \textit{et al.}, 2016; Wu and Pankow, 2016; Lin and Lin, 2018). Most of the methods are based on individuals tests where a single variant or a group of variants is tested

\textsuperscript{1}Corresponding author: Department of Statistics, Pusan National University, Busandaehak-ro 63beon-gil, Geumjeong-Gu, Busan 46241, Korea. E-mail: hsun@pusan.ac.kr

Published 30 September 2020 / journal homepage: http://csam.or.kr
© 2020 The Korean Statistical Society, and Korean International Statistical Society. All rights reserved.
one at a time. These methods have a limitation that they fail to account for relations among correlated variants. Accommodating correlations among variants can make an important contribution to analysis of pleiotropy.

Alternatively, regularization methods have been widely applied to analysis of high-dimensional genomic data (Wu et al., 2009; Zhou et al., 2010; Sun and Wang, 2012, 2013; Sun et al., 2017; Choi et al., 2018). They have a great advantage over individual tests since genetic effects of all variants in analysis can be computed in one regression framework. However, the regularization methods are designed to select genetic variants associated with a single phenotype outcome, so they cannot be directly applied to identify pleiotropic variants associated with multiple phenotype outcomes.

Recently, Simon et al. (2013a) proposed a multi-response elastic-net method, which basically extends an elastic-net model (Zou and Hastie, 2005) to multivariate outcome variables. Li et al. (2015) proposed a multivariate sparse group lasso that extends a sparse group lasso model (Simon et al., 2013b) to multivariate outcome variables, imposing an arbitrary group structure within outcome variables. Although they can be used to identify pleiotropic variants, they suffer a problem of tuning parameter selection which often results in either too many false positives or too small true positives. That is, variable selection results are often unstable when a selected tuning parameter is numerically too large or too small.

In this article, we applied selection probability (Meinshausen and Bühlmann, 2010) to three multivariate regularization methods in order to identify pleiotropic variants associated with multiple phenotypes. They are individual elastic-net (ENET), unified elastic-net (UNET), and multi-response elastic-net (MNET) regularization methods. In simulation studies, selection performance of three multivariate regularization methods was evaluated when the total number of phenotypes, the number of phenotypes associated with a variant, and correlations among phenotypes are different. For further validation of the regularization methods, we also applied them to a wild bean dataset consisting of 169,028 variants and 17 phenotypes.

2. Method

Let us assume that we have \( n \) observations on \( q \) phenotype outcomes and \( p \) genetic variants. Then, we consider an ordinary multivariate linear regression framework,

\[
Y = X\beta + E,
\]

where \( X \) is an \( n \times p \) matrix of genetic variants, \( Y = (Y_1, \ldots, Y_q) \) is an \( n \times q \) matrix of phenotype outcomes, \( \beta = (\beta_1, \ldots, \beta_q) \) is a \( p \times q \) coefficient matrix, and \( E = (e_1, \ldots, e_q)^T \) is an \( n \times q \) matrix of errors. For the \( k^{th} \) phenotype, \( Y_k = (y_{1k}, \ldots, y_{nk})^T \) and \( \beta_k = (\beta_{1k}, \ldots, \beta_{pk})^T \). We further assume that the error vector of the \( i^{th} \) individual \( e_i = (e_{i1}, \ldots, e_{iq})^T \) follows a multivariate normal distribution \( N_q(0, \Sigma) \), where \( \Sigma \) is a covariance matrix representing correlations among \( q \) outcomes. If the outcomes are independent with each other, \( \Sigma \) is simply an identity matrix.

Multivariate regularization methods aim to estimate a regression coefficient matrix \( \beta \) while a penalty function constrains the parameter space of \( \beta \). In a simple way, we can apply the elastic-net method (Zou and Hastie, 2005) to each phenotype outcome, where \( \hat{\beta}_k \) is individually estimated for \( k = 1, \ldots, q \). For the \( k^{th} \) phenotype outcome, the estimate \( \hat{\beta}_k \) minimizes

\[
||Y_k - X\hat{\beta}_k||_2^2 + \lambda_k \left( \frac{1 - \alpha_k}{2} ||\beta_k||_2^2 + \alpha_k ||\beta_k||_1 \right),
\]

where \( || \cdot ||_1 \) and \( || \cdot ||_2 \) are \( l_1 \) and \( l_2 \) norms, respectively. The tuning parameter \( \lambda_k > 0 \) controls the sparsity of the elastic-net, and \( \alpha_k \in [0, 1] \) is a mixing proportion between ridge and lasso penalties.
which make the elastic-net taking advantages of ridge and lasso. Thus, the elastic-net model can not only encourage a grouping effect in which highly correlated variables are included together in the model, but also perform variable selection due to a sparse model. However, this method completely ignores correlations among phenotype outcomes.

In order to consider correlations among outcomes, a multi-response elastic-net method was proposed by Simon et al. (2013a). It minimizes

$$
\|Y - \beta X\|_F^2 + \lambda \left( \frac{1 - \alpha}{2} \|\beta\|_F^2 + \alpha \sum_{j=1}^{p} \|\beta_j\|_2 \right),
$$

(2.2)

where $\beta_j = (\beta_{j1}, \ldots, \beta_{jq})^T$ is the $q$-dimensional vector for the $j^{th}$ variant, and $\|\cdot\|_F$ indicates the Frobenius norm. The $l_2$-norm of the coefficients of the $j^{th}$ variant plays a role of a group lasso (Yuan and Lin, 2006) where the coefficient estimates of the $j^{th}$ variant have all zero values when no association with $q$ phenotypes is detected. That is to say, the $j^{th}$ variant can be selected only if at least one phenotype is associated with it. The squared Frobenius norm of $\beta$ is essentially applying the elastic-net model to all of $q$ phenotype outcomes where the penalty function induces simultaneous shrinkage of coefficients of correlated variants. Therefore, the multi-response elastic-net method takes into account correlations among variants as well as correlations among phenotypes. Similar to (2.1), the tuning parameters $\lambda > 0$ is a parameter to control sparsity and $\alpha \in [0, 1]$ is a mixing proportion between group lasso and elastic-net penalties.

In order to estimate the regression coefficient matrix $\beta$ in (2.1) and (2.2), the tuning parameters should be first determined. The most common way to choose the optimal tuning parameter is cross-validation, where samples are randomly split into training sets and validation sets in order to measure estimation errors for each of candidate values of the tuning parameters. Unfortunately, variable selection result based on cross-validation is often unstable due to random split in high-dimensional genomic data, where the sample size is much smaller than the number of variants.

Alternatively, selection probability has been proposed for stable variable selection by Meinshausen and Bühlmann (2010). It computes a selection proportion of individual variants after repeatedly estimating the coefficients from resampling, so selection probability essentially measures relative selection frequency of each variant. We can rank variants from highest to lowest based on their selection probability. Ranking of variants is rarely changed. Therefore, computation of selection probability does not require to choose the optimal tuning parameter. For this reason, selection probability has been widely used to identify top ranked genetic variants or genes that are associated with a phenotype outcome for analysis of high-dimensional genomic data (Alexander and Lange, 2011; Sun and Wang, 2013; Sun et al., 2017; Choi et al., 2018; Kim and Sun, 2019). However, selection probability has never been applied to multivariate regularization methods.

Let $I_m$ be the $m^{th}$ random subsample of an index set $\{1, 2, \ldots, n\}$ with a size of $\lfloor n/2 \rfloor$ without replacement, where $\lfloor n/2 \rfloor$ denotes the largest integer not greater than $n/2$. Suppose that we denote the $q$-dimensional coefficient vector of the $j^{th}$ variant by $\beta_j = (\beta_{j1}, \ldots, \beta_{jq})^T$. For a given value of tuning parameters $\lambda$ and $\alpha$, let us denote the estimated coefficient vector of the $j^{th}$ variant based on $I_m$ subsample by $\hat{\beta}_j(I_m; \lambda, \alpha)$. Then, the selection probability of the $j^{th}$ variant can be defined as

$$
SP_j = \max_{\lambda, \alpha} \frac{1}{M} \sum_{m=1}^{M} I(\hat{\beta}_j(I_m; \lambda, \alpha) \neq \emptyset),
$$

(2.3)
where \( M = 100 \) is a total number of resamplings and \( I(\cdot) \) is an indicator function. Note that the \( q \)-dimensional vector \( \hat{\beta}_j(I_m; \lambda, \alpha) \) has either all zero values or all nonzero values because of a group lasso penalty in (2.2).

Similarly, selection probability of the individual elastic-net method in (2.1) can be computed and details are described by (Sun and Wang, 2013; Sun et al., 2017; Choi et al., 2018; Kim and Sun, 2019). However, the selection probability of the \( j^{th} \) variant of the elastic-net can have up to \( q \) different values since the coefficient vector is estimated from each of \( q \) phenotype outcomes. Therefore, the selection performance of individual elastic-net cannot be fairly compared with that of multi-response elastic-net. So, we developed new selection strategy that unifies individual elastic-net models for each outcomes, using selection probability. Let us denote the estimated coefficient of the \( j^{th} \) variant and the \( k^{th} \) outcome in the elastic-net model by \( \hat{\beta}_j(I_m; \lambda_k, \alpha_k) \) when the estimation is based on \( I_m \) subsample with fixed values of \( \alpha_k \) and \( \lambda_k \). We then select the \( j^{th} \) variant only if \( \hat{\beta}_j(I_m; \lambda_k, \alpha_k) \) has nonzero values for at least one \( k \). Specifically, an indicator value of the \( j^{th} \) variant can be defined as

\[
S_j(I_m) = \begin{cases} 
0, & \text{if } \max_{\lambda, \alpha} \prod_{k=1}^q |\hat{\beta}_j(I_m; \lambda_k, \alpha_k)| = 0, \\
1, & \text{otherwise},
\end{cases}
\]

(2.4)

where \( \lambda = (\lambda_1, \ldots, \lambda_q) \) and \( \alpha = (\alpha_1, \ldots, \alpha_q) \). Finally, the selection probability of the \( j^{th} \) variant in the unified elastic-net model can be computed as

\[
SP_j = \frac{1}{M} \sum_{m=1}^M S_j(I_m).
\]

(2.5)

Note that this unified approach shares information of variants associated with other phenotype outcomes unlike the individual elastic-net method, where variant selection by one phenotype outcome is not affected by selection from other outcomes.

For tuning parameter values, both \( \alpha \) for the multi-response elastic-net and \( \alpha_k \) for the individual elastic-net were fixed as 0.1, 0.2, \ldots, 0.9. For each fixed \( \alpha \) value, we generated 10 different \( \lambda \) values such that

\[
\lambda_{[\max]} = \lambda_{[1]} > \lambda_{[2]} > \cdots > \lambda_{[10]},
\]

where \( \lambda_{[\max]} \) is the smallest \( \lambda \) value for which the entire estimated vector \( \beta \) is equal to 0. In order to obtain a fine grid of \( \lambda \), we equally spaced \( \lambda \) values such that

\[
\lambda_{[u+1]} - \lambda_{[u]} = \lambda_{[u]} - \lambda_{[u-1]}
\]

for \( u = 2, \ldots, 9 \), and \( \lambda_{[10]} = \epsilon \lambda_{[\max]} \) for \( 0 < \epsilon < 1 \). For each fixed \( \alpha_k \) value, we generated 10 different \( \lambda_k \) values for \( k = 1, \ldots, q \) in the same way.

### 3. Simulation studies

We conducted simulation studies to evaluate selection performances of three regularization methods. They are individual elastic-net, unified elastic-net, and multi-response elastic-net, respectively. To obtain single nucleotide polymorphism (SNP) data, we applied the same process used in Choi et al. (2018). Specifically, we first generated two \( p \)-dimensional vectors \( (a_1, \ldots, a_p)^T \) and \( (b_1, \ldots, b_p)^T \) from a multivariate normal distribution with a mean vector of 0 and \( \Delta AR(1) \) covariance \( \Sigma \), which has
The number of selected variants

True Positive Rates

Methods

ENET UNET MNET

γ = 1 γ = 2 γ = 3 γ = 4 γ = 5

q = 2 q = 3 q = 4 q = 5

Figure 1: Weakly correlated case with ρ = 0.2: Averaged true positive rates (TPR) of individual elastic-net (ENET), unified elastic-net (UNET) and multi-response elastic-net (MNET) along with the number of selected variants ranked by their selection probabilities. The total number of phenotypes and the number of phenotypes that have an association with a variant are denoted by q and γ, respectively.

the value of ρ|u−v| for the u\text{th} row and v\text{th} column in the matrix for 1 \leq u, v \leq p. The correlation coefficient of AR(1) covariance was fixed as r = 0.6. Next, minor allele frequency (MAF) of each variant was generated from a truncated chi-square distribution with a range between 0.01 and 0.45, and the MAF of the j\text{th} variant is denoted by MAFj. Finally, the simulated SNP of the i\text{th} individual and the j\text{th} variant was computed as

\[ x_{ij} = I(q_{a_{ij}} < \text{MAF}_j) + I(q_{b_{ij}} < \text{MAF}_j), \]

where \( q_{a_{ij}} = (1/n) \sum_{i' = 1}^{n} I(a_{i'j} > a_{ij}) \) and \( q_{b_{ij}} = (1/n) \sum_{i' = 1}^{n} I(b_{i'j} > b_{ij}) \) are the reversed quantile of \( a_{ij} \) and \( b_{ij} \), respectively. Our SNP data generated as above attained Hardy-Weinberg equilibrium among n samples as well as linkage disequilibrium between neighboring variants (Foulkes, 2009).

After we obtain SNP data, we could generate multivariate phenotype outcomes based on correlated errors. For the i\text{th} individual, we generated \( e_i = (e_{i1}, \ldots, e_{iq})^T \) from a multivariate normal distribution \( N_q(0, \Sigma_p) \), where \( \Sigma_p \) is a compound symmetric covariance with a correlation coefficient of ρ. We considered both weak correlation ρ = 0.2 and strong correlation ρ = 0.8. If k\text{th} phenotype outcome is associated with a variant, then the k\text{th} phenotype outcome value for the i\text{th} individual was computed as

\[ y_{ik} = \sum_{j=1}^{p} x_{ij} \beta_{jk} + e_{ik}, \]
The number of selected variants 

True Positive Rates 

Methods

ENET
UNET
MNET

Figure 2: Strongly correlated case with $\rho = 0.8$: Averaged true positive rates (TPR) of individual elastic-net (ENET), unified elastic-net (UNET) and multi-response elastic-net (MNET) along with the number of selected variants ranked by their selection probabilities. The total number of phenotypes and the number of phenotypes that have an association with a variant are denoted by $q$ and $\gamma$, respectively.

while $y_{ik} = e_{ik}$ if the $k^{th}$ phenotype does not have any association with a variant. Therefore, the phenotype outcomes are always correlated with each other, regardless of association. We located only one outcome-related variant among $p = 1,000$. Its effect size was fixed as $\beta_{jk} = 0.25$ if the $j^{th}$ variant is associated with the $k^{th}$ phenotype, and $\beta_{jk} = 0$ otherwise. We considered 2 phenotypes to 5 phenotypes, i.e., $q \in \{2, 3, 4, 5\}$. Also, we control the total number of phenotypes that have an association with a variant, and it is denoted by $\gamma \in [1, q]$. The sample size was fixed as $n = 100$, and the number of simulation replications was 100.

Selection performance of three regularization methods was evaluated by averaged true positive rate, which is equivalent to the simulation proportion detected the outcome-related variant among 100 replications since we have only one outcome-related variant. For fair comparison, we compared true positive rates (TPR) of three methods when they select the same number of variants. Each method ranks 1,000 variants by their selection probabilities from highest to lowest, and then we selected top 10, 20, . . . , 50 variants. Note that ENET was applied to each phenotype so TPR of ENET was computed for each phenotype. For this reason, the number of TPR lines of ENET is equivalent the number of phenotypes. The selection results are displayed in Figure 1 for week correlation and Figure 2 for strong correlation.

When phenotype outcomes are weakly correlated with each other, UNET and MNET seem to have no difference in TPR. Since ENET individually applied each phenotype outcome, the number of TPR lines of ENET are equal to $q$, where only $\gamma$ TPR lines are relatively high because they are correspond-
Figure 3: Averaged true positive rates (TPR) of individual elastic-net (ENET), unified elastic-net (UNET) and multi-response elastic-net (MNET) along with the number of selected variants ranked by their selection probabilities are displayed for weakly correlated case with $\rho = 0.2$ and strongly correlated case with $\rho = 0.8$. The total number of phenotypes and the number of phenotypes that have an association with a variant are denoted by $q$ and $\gamma$, respectively.

The number of selected variants

True Positive Rates

Methods

| $\gamma$ | ENET | UNET | MNET |
|----------|------|------|------|
| $\gamma = 1$ | ![Graph](image1) | ![Graph](image2) | ![Graph](image3) |
| $\gamma = 2$ | ![Graph](image4) | ![Graph](image5) | ![Graph](image6) |
| $\gamma = 3$ | ![Graph](image7) | ![Graph](image8) | ![Graph](image9) |
| $\gamma = 4$ | ![Graph](image10) | ![Graph](image11) | ![Graph](image12) |
| $\gamma = 5$ | ![Graph](image13) | ![Graph](image14) | ![Graph](image15) |

When $\gamma = 1$, TPR of ENET can be higher than that of MNET, while MNET is always higher than ENET when $\gamma > 1$. However, we cannot know which phenotype is truly associated with a variant in practice, so ENET has a limitation to identify pleiotropic variants in real data analysis. Also, we can noticed that TPR of MNET is gradually increased as the number of phenotypes associated with a variant increases. In contrast, TPR of ENET is not influenced by the value of $\gamma$. For strong correlated phenotypes, MNET clearly outperforms ENET and UNET, although three methods have almost the same TPR when all phenotypes are associated with a variant, i.e., $\gamma = q$. It appears that ENET is rarely affected by correlations among phenotypes, even though UNET shows slightly better selection performance than ENET. It is noticeable that TPR of MNET was dropped down when all phenotypes are associated with a variant, and correlations among phenotypes are relatively high. When correlations among phenotypes are low, an association signal to each phenotype can be cumulative. However, the association signal from each phenotype can be disrupted when correlations among phenotypes are high.

Finally, we conducted an additional simulation study where the number of outcome-related variants was increased as 5 and all other simulation settings were the same. In Figure 3, averaged true positive rates of three regularization methods are compared with each other when the total number of phenotypes are fixed as $q = 5$, and the number of associated phenotypes is $\gamma = 1, 2, 3, 4, 5$ for both weakly correlated case with $\rho = 0.2$ and strongly correlated case with $\rho = 0.8$. It appears that the simulation result is very similar to the results of the the previous simulation studies. Consequently, UNET and MNET have almost identical TPR in the weakly correlated case while TPR of MNET is overall higher than that of UNET in the strongly correlated case. MNET is designed to perform group selection where all of $q$ regression coefficients of the $j^{th}$ variant corresponding to $q$ phenotypes are likely to have nonzero values if at least one phenotype is associated with the $j^{th}$ variant. When pheno-
types are highly correlated with each other, association signals of MNET can be boosted because the regression coefficients are simultaneously estimated based on the group penalty. In contrast, UNET is indifferent to correlated phenotypes since the selection probability of UNET is computed based on association signals between an individual phenotype and variants.

4. Real data analysis

We applied both MNET and UNET to a wild bean data set from National Institute of Crop Science, Rural Development Administration, Republic of Korea. The wild beans dataset consists of genotype data and phenotype data for 439 samples and 169,028 SNPs. There are 17 quantitative phenotypes representing essential amino acids, i.e., \( q = 17 \). We first imputed the missing values of SNPs, using GAPIT (Lipka et al., 2012). We then excluded samples that have any missing values for phenotypes, and variants with MAF < 0.01. Finally, we ended up with \( n = 370 \) samples and \( p = 160,035 \) SNPs. In our simulation studies, we learned that selection performance of both UNET and MNET can depend on correlations among phenotype outcomes. Therefore, we separated 17 phenotypes into two different groups, where a highly correlated group \((\rho > 0.76)\) has 13 phenotypes and a weakly correlated group \((0.12 < \rho < 0.49)\) has 4 phenotypes. Hierarchical clustering analysis for the correlation matrix of 17 phenotypes are summarized as a heatmap in Figure 4. For each group, we computed selection probabilities of entire SNPs, using MNET and UNET. Then, we made two manhattan plots, where \( p\)-values on the y-axis were replaced by selection probabilities of variants. Figure 5 and Figure 6 display the manhattan plots of MNET and UNET for the highly correlated group and the weakly correlated group, respectively.

We focused top 50 variants ranked by selection probabilities of either MNET or UNET. The blue reference line of each plot indicates the selection probability of top 50 ranked variant. Among 50
Figure 5: Manhattan plots of selection probabilities for multi-response elastic-net (MNET) and unified elastic-net (UNET) are displayed for the highly correlated group. Blue circles represent variants selected by both MNET and UNET, and red circles represent variants uniquely selected by either MNET or UNET. The solid reference line is the selection probability of top 50 ranked variant.

Figure 6: Manhattan plots of selection probabilities for multi-response elastic-net (MNET) and unified elastic-net (UNET) are displayed for the weakly correlated group. Blue circles represent variants selected by both MNET and UNET, and red circles represent variants uniquely selected by either MNET or UNET. The solid reference line is the selection probability of top 50 ranked variant.

variants, blue circles represent variants selected by both MNET and ENET, and red circles represent variants uniquely selected by either MNET or UNET. That is, red circles of top panel are selected by only MNET, and red circles of bottom panel are selected by only UNET. Note that selection
probabilities of some variants are exactly the same so top 50 variant list can have more than 50 variants due to ties. For the highly correlated group, 33 variants were selected by both MNET and UNET, while 18 variants were uniquely selected by either MNET or UNET. For the weakly correlated group, 21 variants were selected by both MNET and UNET, while 37 variants were selected by MNET only and 34 variants were selected by UNET only. We also found that selection probabilities of variants selected by both MNET and UNET are quite similar to each other except a few variants in the highly correlated group. In contrast, selection probabilities of MNET and UNET have some discrepancy in the weakly correlated group. Figure 7 displays a scatter plot of selection probabilities computed by MNET on the $x$-axis and UNET on the $y$-axis for the highly and weakly correlated groups. It seems that correlations of selection probabilities between MNET and UNET are relatively high in the highly correlated group, although there are some outliers. The sample correlation coefficients are 0.9084 and 0.8554 for the highly correlated group and the weakly correlated groups, respectively. In our simulation studies, we learned that selection performance of UNET and MNET become almost identical in the weakly correlated case. Also, they are very similar when all of the phenotypes are associated with a variant in the strongly correlated case. Therefore, we can expect that there exist some variants associated with almost all of the 13 phenotypes.

5. Conclusion

We applied selection probability to three different multivariate regularization methods in order to identify pleiotropic variants associated with multiple phenotype outcomes. Selection probability has been popularly used to stabilize variant selection of regularization methods in analysis of high-dimensional genomic data. However, it has not been applied to multivariate regularization methods yet as far as we know. We also found that selection performance of multivariate regularization can rely on the number of phenotypes that associated with variants and correlations among phenotypes. In particular, when phenotypes are strongly correlated with each other, MNET shows the best selection performance if the number of phenotypes associated with a variant is less than the total number of phenotypes. In real data analysis, the number of pleiotropic variants selected by both MNET and ENET were relatively high in highly correlated phenotypes.

Although recent genetic association studies pay much attention to identification of pleiotropic

Figure 7: Scatter plots of selection probabilities computed by MNET and UNET for the highly correlated group in the left and the weakly correlated group in the right.
variants, most of existing statistical methods have mainly focused on individual tests that consider only correlations among phenotypes, ignoring correlations among variants. Multivariate regularization methods can be a good alternative method to identify pleiotropic variants, but there are currently only a few regularization methods that can be directly applied to multiple phenotype outcomes. We developed a unified elastic-net method based on selection probability, and it enables individual elastic-net to be applied to multiple phenotype outcome. For association with a single phenotype outcome, there are relatively many univariate regularization methods such as group lasso, sparse group lasso, and network-constrained regularization. Selection performance can be improved when a prior knowledge like biological network are utilized into genetic association studies. Therefore, we can easily extend our unified approach based on selection probability to other types of univariate regularization methods. In addition, multivariate regularization methods are limited to quantitative outcomes while univariate regularization methods can be applied qualitative outcome. We can also extend the unified approach to these methods for future research.

Acknowledgement
This work was supported by a 2-Year Research Grant of Pusan National University.

References
Alexander D and Lange K (2011). Stability selection for genome-wide association, Genetic Epidemiology, 35, 722–728.
Bhattacharjee S, Rajaraman P, Jacobs KB, et al. (2012). A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits, American Journal of Human Genetics, 90, 821–835.
Broadaway KA, Cutler DJ, Duncan R, et al. (2016). A statistical approach for testing cross-phenotype effects of rare variants, American Journal of Human Genetics, 98, 525–540.
Choi J, Kim K, and Sun H (2018). New variable selection strategy for analysis of high-dimensional DNA methylation data, Journal of Bioinformatics and Computational Biology, 16, 1850010.
Foulkes AS (2009). Applied Statistical Genetics with R, Springer-Verlag, New York.
Kim K and Sun H (2019). Incorporating genetic networks into case-control association studies with high-dimensional DNA methylation data, BMC Bioinformatics, 20, 510.
Li Y, Nan B, and Zhu J (2015). Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure, Biometrics, 71, 354–363.
Lin Z and Lin X (2018). Multiple phenotype association tests using summary statistics in genome-wide association studies, Biometrics, 74, 165–175.
Lipka AE, Tian F, Wang Q, et al. (2012). GAPIT: genome association and prediction integrated tool, Bioinformatics, 28, 2397–2399.
Meinshausen N and Bühlmann P (2010). Stability selection, Journal of the Royal Statistical Society Series B, 72, 417–473.
Schaid DJ, Tong X, Larrabee B, Kennedy RB, Poland GA, and Sinnwell JP (2016). Statistical methods for testing genetic pleiotropy, Genetics, 204, 483–497.
Simon N, Friedman J, and Hastie T (2013a). A blockwise descent algorithm for group-penalized multiregression and multinomial regression, arXiv preprint arXiv:1311.6529.
Simon N, Friedman J, Hastie T, and Tibshirani R (2013b). A sparse-group lasso, Journal of Computational and Graphical Statistics, 22, 231–245.
Solovieff N, Cotsapas C, Lee PH, Purcell SM, and Smoller JW (2013). Pleiotropy in complex traits:
challenges and strategies, *Nature Reviews Genetics*, **14**, 483–495.

Sun H and Wang S (2012). Penalized logistic regression for high-dimensional DNA methylation data analysis with case-control studies, *Bioinformatics*, **28**, 1368–1375.

Sun H and Wang S (2013). Network-based regularization for matched case-control analysis of high-dimensional DNA methylation data, *Statistics in Medicine*, **32**, 2127–2139.

Sun H, Wang Y, Chen Y, Li Y, and Wang S (2017). pETM: a penalized Exponential Tilt Model for analysis of correlated high-dimensional DNA methylation data, *Bioinformatics*, **33**, 1765–1772.

van der Sluis S, Posthuma D, and Dolan CV (2013). TATES: efficient multivariate genotype-phenotype analysis for genome-wide association studies, *PLoS Genetics*, **9**, e1003235.

Wu B and Pankow JS (2016). Sequence kernel association test of multiple continuous phenotypes, *Genetic Epidemiology*, **40**, 91–100.

Wu T, Chen Y, Hastie T, Sobel E, and Lange K (2009). Genome-wide association analysis by lasso penalized logistic regression, *Bioinformatics*, **25**, 714–721.

Yuan M and Lin Y (2006). Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society Series B*, **68**, 49–67.

Zou H and Hastie T (2005). Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society Series B*, **67**, 301–320.

Zhou H, Sehl M, Sinsheimer J, and Lange K (2010). Association screening of common and rare genetic variants by penalized regression, *Bioinformatics*, **26**, 2375–2382.

Received April 26, 2020; Revised June 12, 2020; Accepted July 12, 2020