G&I
Genomics & Informatics

Original article

eISSN 2234-0742
Genomics Inform 2022; 20(4):e48
https://doi.org/10.5808/gi.22075

Received: December 13, 2022
Accepted: December 19, 2022

*Corresponding author:
E-mail: mira@eulji.ac.kr

MP-Lasso chart: a multi-level polar chart for visualizing group Lasso analysis of genomic data

Min Song[1], Minhyuk Lee[1], Taesung Park[2], Mira Park[3]*

[1]Department of Statistics, Korea University, Seoul 02841, Korea
[2]Department of Statistics, Seoul National University, Seoul 08826, Korea
[3]Department of Preventive Medicine, Eulji University, Daejeon 34824, Korea

Penalized regression has been widely used in genome-wide association studies for joint analyses to find genetic associations. Among penalized regression models, the least absolute shrinkage and selection operator (Lasso) method effectively removes some coefficients from the model by shrinking them to zero. To handle group structures, such as genes and pathways, several modified Lasso penalties have been proposed, including group Lasso and sparse group Lasso. Group Lasso ensures sparsity at the level of pre-defined groups, eliminating unimportant groups. Sparse group Lasso performs group selection as in group Lasso, but also performs individual selection as in Lasso. While these sparse methods are useful in high-dimensional genetic studies, interpreting the results with many groups and coefficients is not straightforward. Lasso's results are often expressed as trace plots of regression coefficients. However, few studies have explored the systematic visualization of group information. In this study, we propose a multi-level polar Lasso (MP-Lasso) chart, which can effectively represent the results from group Lasso and sparse group Lasso analyses. An R package to draw MP-Lasso charts was developed. Through a real-world genetic data application, we demonstrated that our MP-Lasso chart package effectively visualizes the results of Lasso, group Lasso, and sparse group Lasso.

Keywords: group Lasso, group structure, multi-level polar chart, sparsity, variable selection, visualization

Introduction

In the analysis of high-dimensional genomics data, the least absolute shrinkage and selection operator (Lasso) method and its variants have been widely used to perform regression and model selection [1-3]. Typical Lasso performs variable selection at the individual gene level. However, genomic data can have a group structure, as in gene-expression data with pathways or in single-nucleotide polymorphism (SNP) data with genetic regions including multiple SNPs from genome-wide association studies. Better predictions can be expected if the group structure of the genes is considered. Several variants of Lasso have been developed to address the group structure [4-7]. Group Lasso ensures sparsity at the level of pre-defined groups and eliminates unimportant groups [4]. Moreover, sparse group Lasso combines Lasso and group Lasso to enable group selection as well as individual selection [5].

Visualization is used to effectively summarize the results of high-dimensional data analysis. The results of Lasso are often expressed as a tracking plot of regression coefficients.

However, few studies have explored the systematic visualization of group information. In this study, we propose a multi-level polar chart for visualizing group Lasso analysis (the MP-Lasso chart). The MP-Lasso chart is an improved version of the MP chart, which was originally developed for integrating results from multi-omics data analyses [8]. An MP-Lasso chart shows at a glance the variables selected for each variable group and their coefficient estimates produced by group-structured penalized methods. We developed a program for creating MP-Lasso charts called “MP-Lasso” using R (https://github.com/statpark/MP-Lasso).

Methods

Penalized regression model

Consider a general linear model. We have n observations, and the data consist of $n\times1$ including the response variable y and a $n\times p$ matrix X of predictors. Assume that y and X have been centered. The objective function of Lasso is

$$\min_{\beta} ||y - X\beta||_2^2 + \lambda||\beta||_1$$

where λ is a tuning parameter, $||\cdot||_1$ stands for the vector l_1-norm and $||\cdot||_2$ stands for the vector l_2-norm. This penalty shrinks any coefficients contributing to the minimization problem to 0 [1].

Suppose that the predictors consist of G groups. Let X_g be a matrix for the predictors of the g^{th} group with the corresponding coefficient vector β_g. The objective function of group Lasso is

$$\min_{\beta}(||y - \sum_{g=1}^{G} X_g\beta_g||_2^2 + \lambda\sum_{g=1}^{G}\sqrt{p_g}\,||\beta_g||_2),$$

where p_g is the corresponding weight considering group size and $||\cdot||_2$ is the Euclidean norm [4,9]. Group Lasso shrinks all β values in irrelevant groups to 0. When $\lambda=0$, this criterion is equivalent to Lasso.

Sparse group Lasso uses a more general penalty to generate sparsity at both the group and individual feature levels, allowing the selection of groups and within-group variables. The objective function of sparse group Lasso is given by

$$\min_{\beta}\ (||y - \sum_{g=1}^{G} X_g\beta_g||_2^2 +$$
$$(1-\alpha)\lambda\sum_{g=1}^{G}\sqrt{p_g}||\beta_g||_2 + \alpha\lambda||\beta||_1),$$

where $\alpha\in[0,1]$ and $\beta=(\beta_1,\ldots,\beta_G)$ [5,9]. α is a convex combination of Lasso and group Lasso. This criterion is equivalent to group Lasso if $\alpha=0$, and to Lasso if $\alpha=1$.

MP-Lasso charts

An MP-Lasso chart consists of an outer level and an inner level. The outer level shows the overall impact of each group, and the inner level represents the impact of each variable within a group. For the outer level, the circle is divided into as many sectors as the number of groups. The segments in the chart are sorted by the maximum value of the coefficients in each group. A group with a rank of 1 starts at 0°. The radius of each sector is set to be proportional to the maximum value of the coefficients. When sorting or determining the radius, the maximum value may be replaced by the average. The number of variables in each group can be distinguished by the color of each segment.

For the inner level, points representing the variables in each group are plotted in each segment. The location of variables within each group is scaled by dividing the coefficient by the radius of the sector to represent the relative size. Different symbols are used according to the sign of the coefficient, and the number of variables belonging to each group is represented by the color spectrum.

Each point is jittered slightly to avoid overlapping. Scatter plots are depicted in an interactive manner; moving the cursor on a point shows information about the variable. Fig. 1 shows an example of an MP-Lasso chart using results from group Lasso analysis with eight groups. Group 4 has the variables with the largest coefficients, followed by group 2, and so on. Group 2 contains four SNPs, among which the SNP with the largest regression coefficient is V5, with a regression coefficient of 13.9773.

Results

Implementation

We developed an R package to draw MP-Lasso charts. The program is available online (https://github.com/statpark/MP-Lasso). The MP-Lasso chart program requires two inputs: a cross-validated (CV) object and a group vector. A CV object can be obtained from the output objects of R packages, such as glmnet [10], ggLasso [11], and SGL [12]. Cv.glmnet in glmnet, cv.ggLasso in ggLasso, and cvSGL in SGL perform Lasso, group Lasso, and sparse group Lasso analyses, respectively. The output objects contain regression coefficients with respect to sequential and optimal λ values that minimize CV error. The group vector represents the group structure of variables. Group names should be in character type or integer type. The group vector should be identical to the one used when the CV object is created. MP-Lasso chart supports three methods: Lasso, group Lasso, and sparse group Lasso. Table 1 summarizes the functions of the developed package for MP-Lasso charts and related packages to obtain input data.

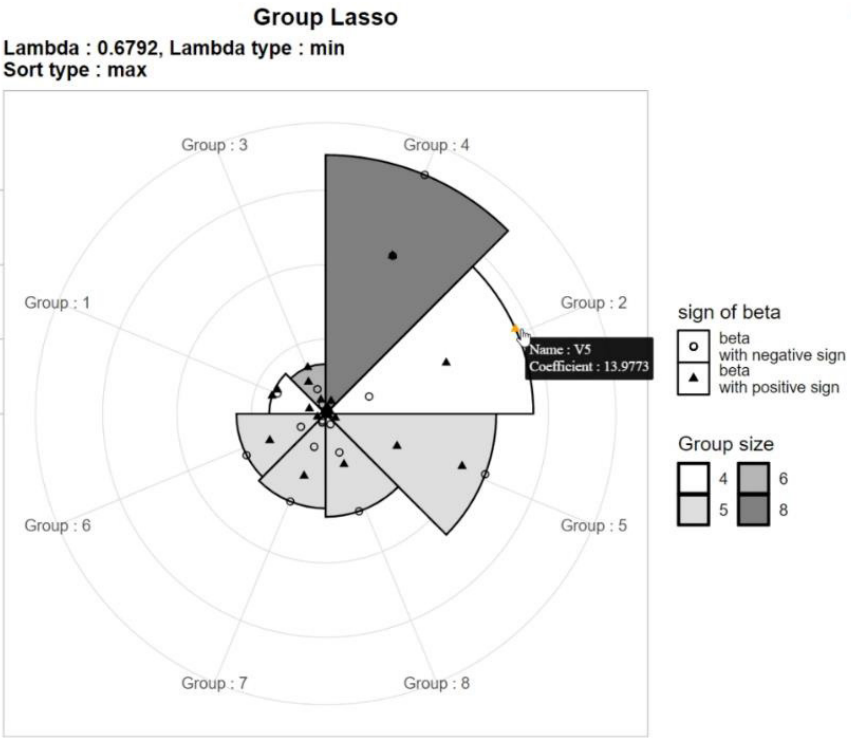


Fig. 1. Example of multi-level polar least absolute shrinkage and selection operator (Lasso) chart using group Lasso analysis.

Table 1. R packages and functions for MP-Lasso charts

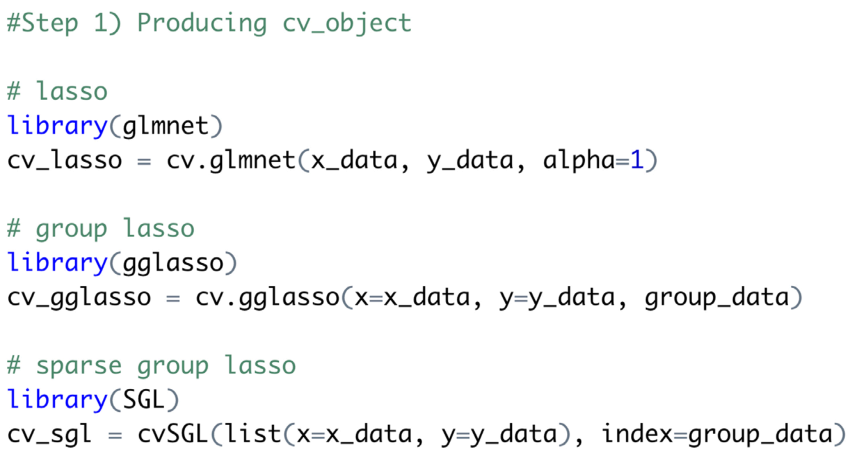
Method	MP-Lasso chart function	Related package	
		Package	CV object function
Lasso	MP_Lasso()	glmnet	cv.glmnet()
Group Lasso	MP_gLasso()	ggLasso	cv.ggLasso()
Sparse group Lasso	MP_SGL()	SGL	cvSGL()

MP-Lasso, multi-level polar least absolute shrinkage and selection operator; CV, cross-validated.

Fig. 2 shows an example of code to obtain a CV object from the data. The group Lasso and sparse group Lasso methods require three inputs: predictor variables (x_data), dependent variables (y_data) and group vector data (group_data). The user can choose the number of folds k. It should be noted that the cvSGL function for sparse group Lasso takes input in list type only. For the Lasso model, α is set to 1 in the cv.glmnet function.

An example of code for an MP-Lasso chart is shown in Fig. 3. The required libraries should be loaded before using the MP-Lasso R code. The function is required to match the Lasso method that creates the CV object. The same group vector used to create a CV object is also taken as input.

MP-Lasso chart has three options to determine the details of a plot. The lambda.type option decides which λ to use for each method, and it can take two values (“min” and “1se”), with “min” as default. The “min” option chooses the λ value that minimizes the CV loss. The “1se” chooses the largest λ with a CV error not 1 standard error further from the minimum CV loss. The “1se” option chooses fewer variables. The sort.type option determines which numeric feature represents the coefficients of variables in each group. Two choices are available for the sort.type option, “max” and “mean.” The “max” option uses the maximum absolute coefficient in each group as the feature of the group, while the “mean” option uses the mean of the absolute coefficients in each group. The last option is the max.shown option. When the number of chosen group is large, the chart can be too crowded with segments, making the chart difficult to interpret. By choosing max.shown, the user can decide the maximum number of segments shown on the chart. In the resulting MP-Lasso chart, interactive features are used. Moving the cursor over a point in the inner level displays information about that vari-



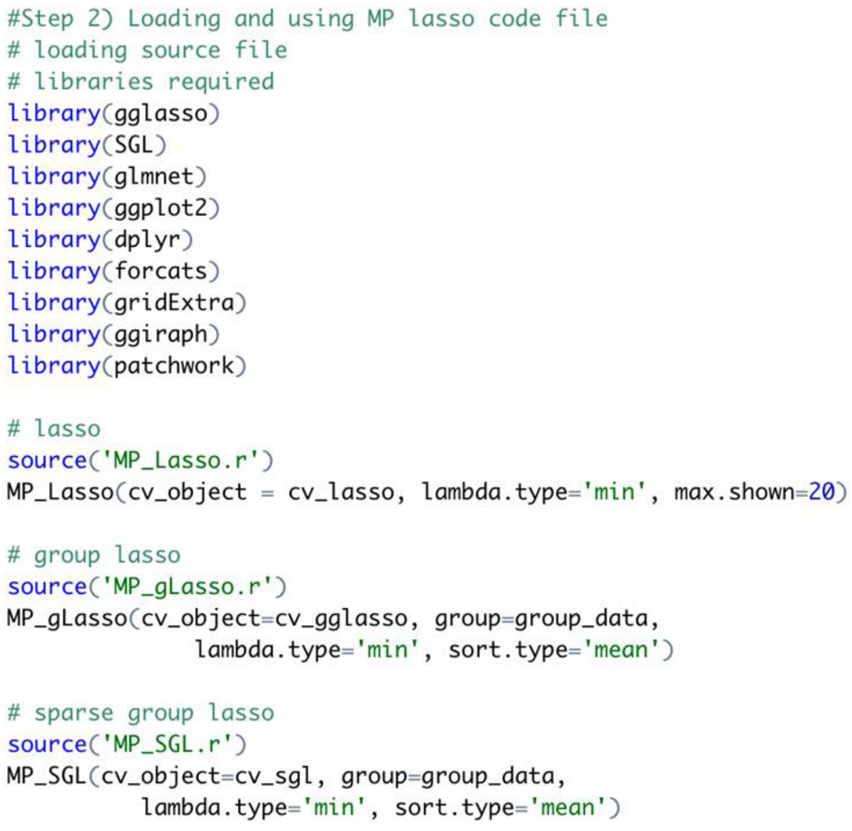
```
#Step 1) Producing cv_object

# lasso
library(glmnet)
cv_lasso = cv.glmnet(x_data, y_data, alpha=1)

# group lasso
library(gglasso)
cv_gglasso = cv.gglasso(x=x_data, y=y_data, group_data)

# sparse group lasso
library(SGL)
cv_sgl = cvSGL(list(x=x_data, y=y_data), index=group_data)
```

Fig. 2. Example R code for creating cross-validated object.



```
#Step 2) Loading and using MP lasso code file
# loading source file
# libraries required
library(gglasso)
library(SGL)
library(glmnet)
library(ggplot2)
library(dplyr)
library(forcats)
library(gridExtra)
library(ggiraph)
library(patchwork)

# lasso
source('MP_Lasso.r')
MP_Lasso(cv_object = cv_lasso, lambda.type='min', max.shown=20)

# group lasso
source('MP_gLasso.r')
MP_gLasso(cv_object=cv_gglasso, group=group_data,
              lambda.type='min', sort.type='mean')

# sparse group lasso
source('MP_SGL.r')
MP_SGL(cv_object=cv_sgl, group=group_data,
          lambda.type='min', sort.type='mean')
```

Fig. 3. Example R code for multi-level polar least absolute shrinkage and selection operator chart.

able. For Lasso, moving the mouse over the sector area displays the corresponding information.

Real data analysis

To illustrate the proposed MP-Lasso charts, we used T-cell and B-cell acute lymphocytic leukemia (ALL) data from the Ritz Laboratory [13]. The data consisted of microarray assays for 11,683 genes with 8,776 groups from 128 individuals with B-cell or T-cell ALL (https://bioconductor.org/packages/release/data/experiment/html/ALL.html). We conducted Lasso, group Lasso, and sparse group Lasso analysis for a binary phenotype using the glmnet, ggLasso, and SGL packages. Each method is depicted using the λ value that yields the minimum 10-fold CV loss, and each group is represented by maximum absolute coefficients. For Lasso analysis, we set *max.shown*=30 for better representation. A summary of the results is presented in Table 2.

Fig. 4 shows the resulting MP-Lasso chart sorted by maximum coefficients. Without group information, the Lasso analysis in Fig. 4A shows variables with the largest coefficients. The variables in the CD3D, TNNI3, and ACAP1 groups had the largest coefficients, in descending order. For group Lasso in Fig. 4B, the CD3D group had much larger maximum coefficients than the other groups and the next two groups (HLA-DPB1 and TRDC) had similar maximum coefficients to each other. Two variables in the HLA-DPB1 group had very similar coefficient values, which can be read from the position of points in the HLA-DPB1 sector. Fig. 4C shows the results from sparse group Lasso. Unlike in group Lasso, no single group dominated, and several significant groups

Table 2. Top three groups with the highest maximum absolute coefficients

Method	Ranking	No. of variables in group	Group	Maximum absolute coefficients
Lasso	1	1	CD3D	0.121
	2	1	TNNI3	0.107
	3	1	ACAP1	0.097
Group Lasso	1	1	CD3D	–0.225
	2	2	HLA-DPB1	0.038
	3	1	TRDC	0.037
Sparse group Lasso	1	1	CD7	–1.280
	2	1	CD3D	–1.230
	3	1	BLNK	0.905

ALL, acute lymphocytic leukemia; Lasso, least absolute shrinkage and selection operator.

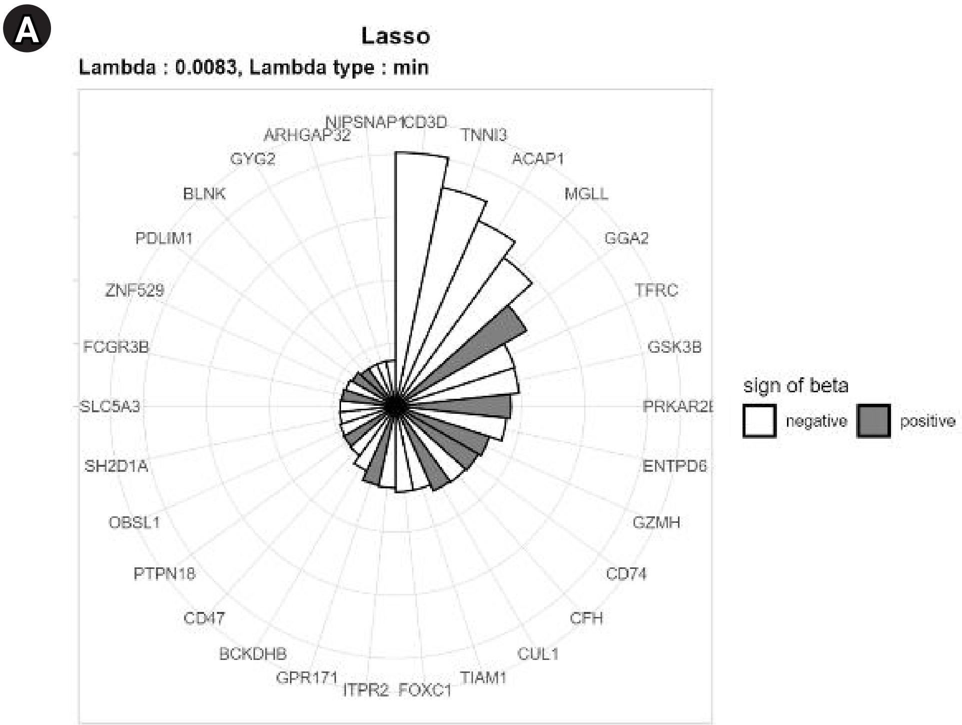


Fig. 4. Multi-level polar least absolute shrinkage and selection operator (Lasso) chart for Lasso (A), group Lasso (C), and sparse group Lasso analysis of acute lymphocytic leukemia data. (Continued to the next page)

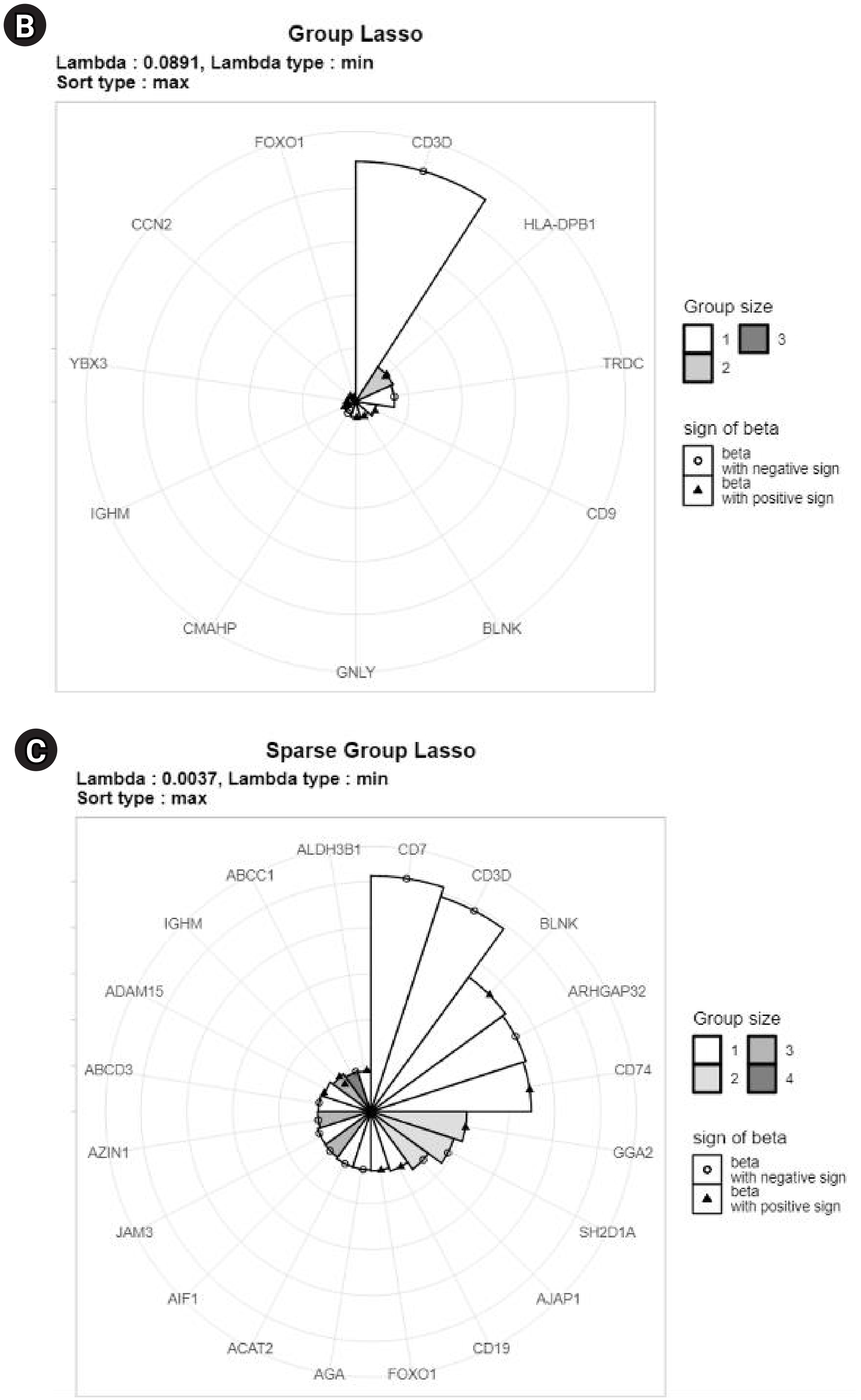


Fig. 4. (Continued from the previous page)

remained. It can also be read that the first two groups had similar maximum coefficients. In this example, each group contained a small number of features. However, even if there are many variables in the groups, the color spectra are automatically adjusted so that the number of groups can be distinguished. Therefore, the user can recognize that a brighter segment indicates a smaller group, while a darker segment corresponds to a larger group. In conclusion, an MP-Lasso chart illustrates the different choices of groups and variables made by each model in detail.

Discussion

We proposed a simple and efficient graph called an MP-Lasso chart for visualizing results from a group-penalized model. We also developed a corresponding R package. An MP-Lasso chart provides a clear representation of each group's information and the relative importance of each variable within a group. Using our package, one can identify important groups and variables at a glance without having to check tables containing thousands of coefficients. It also facilitates model interpretation and comparisons of multiple models.

ORCID

Min Song: https://orcid.org/0000-0002-9554-4793
Minhyuk Lee: https://orcid.org/0000-0003-0705-7766
Taesung Park: https://orcid.org/0000-0002-8294-590X
Mira Park: https://orcid.org/0000-0003-3827-9089

Authors' Contribution

Conceptualization: MP. Data curation: MS, ML. Funding acquisition: MP. Methodology: MP, TP. Writing - original draft: ML. Writing - review & editing: MP, TP.

Conflicts of Interest

Taesung Park serves as an editor of the Genomics and Informatics, but has no role in the decision to publish this article. All remaining authors have declared no conflicts of interest.

Acknowledgments

This research was supported by a National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (NRF-2021R1A2C1007788).

References

1. Tibshirani R. Regression shrinkage and selection via the Lasso. J R Stat Soc Series B Methodol 1996;58:267-288.
2. Wu TT, Chen YF, Hastie T, Sobel E, Lange K. Genome-wide association analysis by Lasso penalized logistic regression. Bioinformatics 2009;25:714-721.
3. Ogutu JO, Piepho HP. Regularized group regression methods for genomic prediction: Bridge, MCP, SCAD, group bridge, group Lasso, sparse group Lasso, group MCP and group SCAD. BMC Proc 2014;8:S7.
4. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. J R Stat Soc Series B Stat Methodol 2006; 68:49-67.
5. Simon N, Friedman J, Hastie T, Tibshirani R. A sparse-group Lasso. J Comput Graph Stat 2013;22:231-245.
6. Zhao P, Rocha G, Yu B. The composite absolute penalties family for grouped and hierarchical variable selection. Ann Stat 2009;37:3468-3497.
7. Wang H, Leng C. A note on adaptive group Lasso. Comput Stat Data Anal 2008;52:5277-5286.
8. Park M, Kim D, Moon K, Park T. Integrative analysis of multi-omics data based on blockwise sparse principal components. Int J Mol Sci 2020;21:8202.
9. Friedman J, Hastie T, Tibshirani R. A note on the group Lasso and a sparse group Lasso. Preprint at https://doi.org/10.48550/arXiv.1001.0736 (2010).
10. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw 2010;33:1-22.
11. Yang Y, Zou H. Package 'ggLasso': group Lasso penalized learning using a unified BMD algorithm. Vienna: R Project for Statistical Computing, 2020.
12. Simon F, Friedman J, Hastie T, Tibshirani R. Package 'SGL': fit a GLM (or Cox Model) with a combination of Lasso and group Lasso regularization. R package version 1.3. Vienna: R Project for Statistical Computing, 2019.
13. Li X. ALL: a data package. Vienna: R Project for Statistical Computing, 2022.