Interpreting Patterns of Gene Expression with Multi-Response Regression

Group 5

Huihang Liu, Jing Wu, Tianchen Xu, Ziyu Chen

University of Science and Technology of China

2019.12.23

Outline

Section 1

Introduction

Introduction

Background

Genetic variation $\rightarrow$ Gene expression level $\rightarrow$ Phenotype

Goal

Relation between genetic variations and gene expression levels.

Section 2

Yeast Gene Expression Data

[1]The yeast data can be accessed in Gene Expression Omnibus(GEO) by accession number GSE1990.

[2]The data were derived from a cross between two strains of the budding yeast: BY4716 and RM11-1a.

[3]Brem, R. B., Kruglyak, L. (2005). The landscape of genetic complexity across 5,700 gene expression traits in yeast. Proceedings of the National Academy of Sciences, 102(5), 1572-1577.

Why yeast?

- Complete genome sequence
- Share some genes with human cells

Data Description

Title	Genetic complexity in yeast transcripts
Organism	Saccharomyces cerevisiae (Baker's yeast)
Experiment type	Expression profiling by array
Data Size	112 yeast samples. Data set consists of 3244 genotypes and 6216 genes. $X \in \mathbb{R}^{3244\times 112}$, $Y \in \mathbb{R}^{6216\times 112}$.
Description	Genotype[1] is a categorical variable, and gene expression level is given by log_2(sample/BY reference).

Table 1: Information About Data

[1] eQTLs (expression Quantitative Trait Locis): some special SNPs which are associated with gene expression.

Question of Interest

Question of Interest

- How eQTLs influence gene expression levels in the yeast MAPK signaling pathways ?
- Which group of eQTLs affect certain group of genes?

Equivalent Question in Statistics

Reveal multiple distinct associations between subsets of genes (eQTLs) and subsets of genetic variants.

Section 3

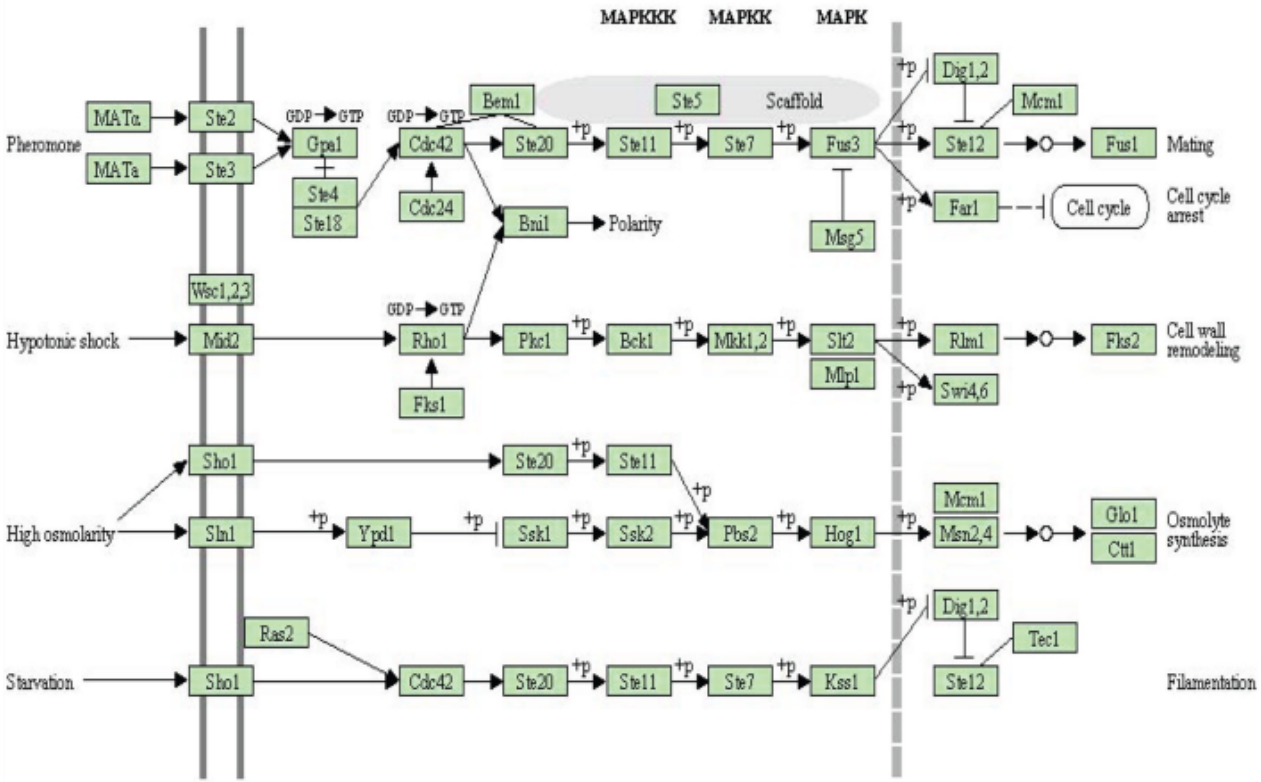
Data Preparation

Processing Genotype Data

1. Hierarchical clustering by complete distance.
   We got 949 blocks where the SNPs within a block differed by at most 1 sample.
2. Select representative SNPs.
   For each block, we choose a representative SNP with the most repetitions.
- Marginal gene-marker association analysis.
  Discussed in the following.

Processing Expression Level Data

We choose genes according to MAPK signaling pathways [2]



[2] Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M. and Tanabe, M. (2014) Data, information, knowledge and principle: Back to metabolism in KEGG. Nucleic Acids Res., 42, D199–D205.

Easy-to-use Data

Let X represent the SNPs matrix, Y represent the gene expression levels matrix, we obtain

$$X \in \mathbb{R}^{949\times 112} \quad , \quad Y \in \mathbb{R}^{53\times 112}$$

Hereinafter, we denote p as the number of explanatory variables, q as the number of response variable, n as the number of samples, E as the random error matrix, and B as the coefficient matrix, we can construct a multi-response linear model

$$Y = XB + E.$$

Section 4

Methodology

Brief Introduction

- Linear Regression Model
- Multi-response
- High Dimensional Problem

Uni-Response Regression

Let Y_j denote the j-th column of Y, represent the expression level of j-th gene.
An intuitive method is to regress each Y_j with X, and we will get q linear models. We can use LASSO to get the estimated coefficient vector $\hat{\beta}_{(j)} \in \mathbb{R}^p$.
Then combine q coefficient vectors $\hat{\beta}_{(j)}$ into a matrix $\widehat{B} \in \mathbb{R}^{p \times q}$ by column.

Result of LASSO

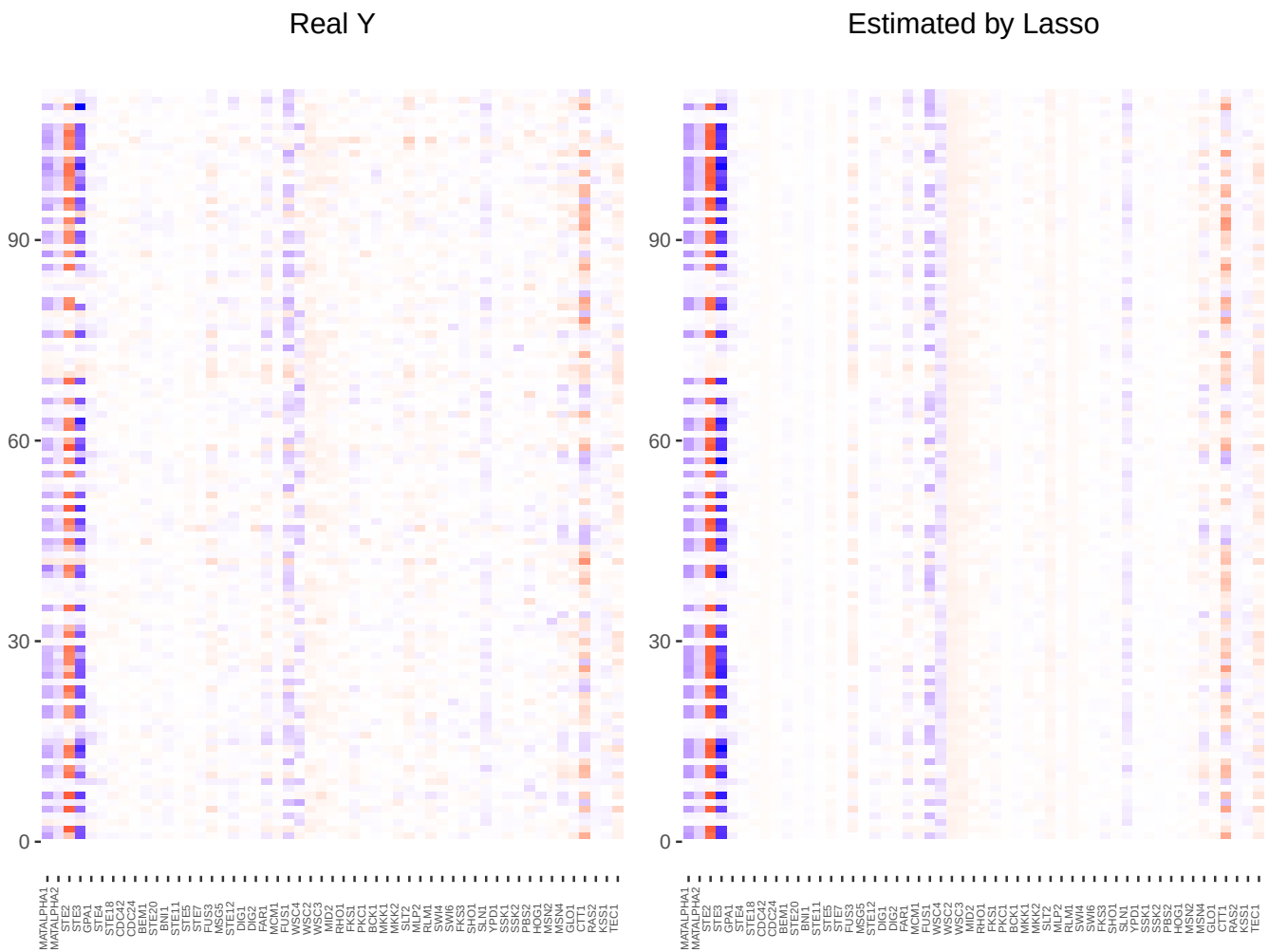


Fig. 1. Heatmaps of real Y and $\hat{Y}$ by LASSO

Result of LASSO cont.

- LASSO is a kind of shinkage estiamtion method. So a sparse $\hat{\beta}_{(j)}$ is expected for each $j \in \{1, 2, \ldots, q\}$.
- But $\widehat{B}$ may not be sparse by row.
  Actually, there are 602 nonzero rows in $\widehat{B}$ which has full column rank.

Group Sparse Linear Regression

Group sparse linear regression for multitask learning [1]

$$\widehat{B} = \arg\min_{B} \left\{ \frac{1}{2} \|Y - XB\|_F^2 + \lambda \|B\|_{(2,1)} \right\} \tag{1}$$

where $\|\cdot\|_F$ is the Frobenius norm,and where the $(2,1)$ norm in the penalty is given by $\|B\|_{(2,1)} = \sum_i \sqrt{\sum_j B_{ij}^2}$.

This penalty promotes rowwise sparsity of $\widehat{B}$.

[1] Dai, Ran, and Rina Foygel Barber. (2016)

Result of GLasso

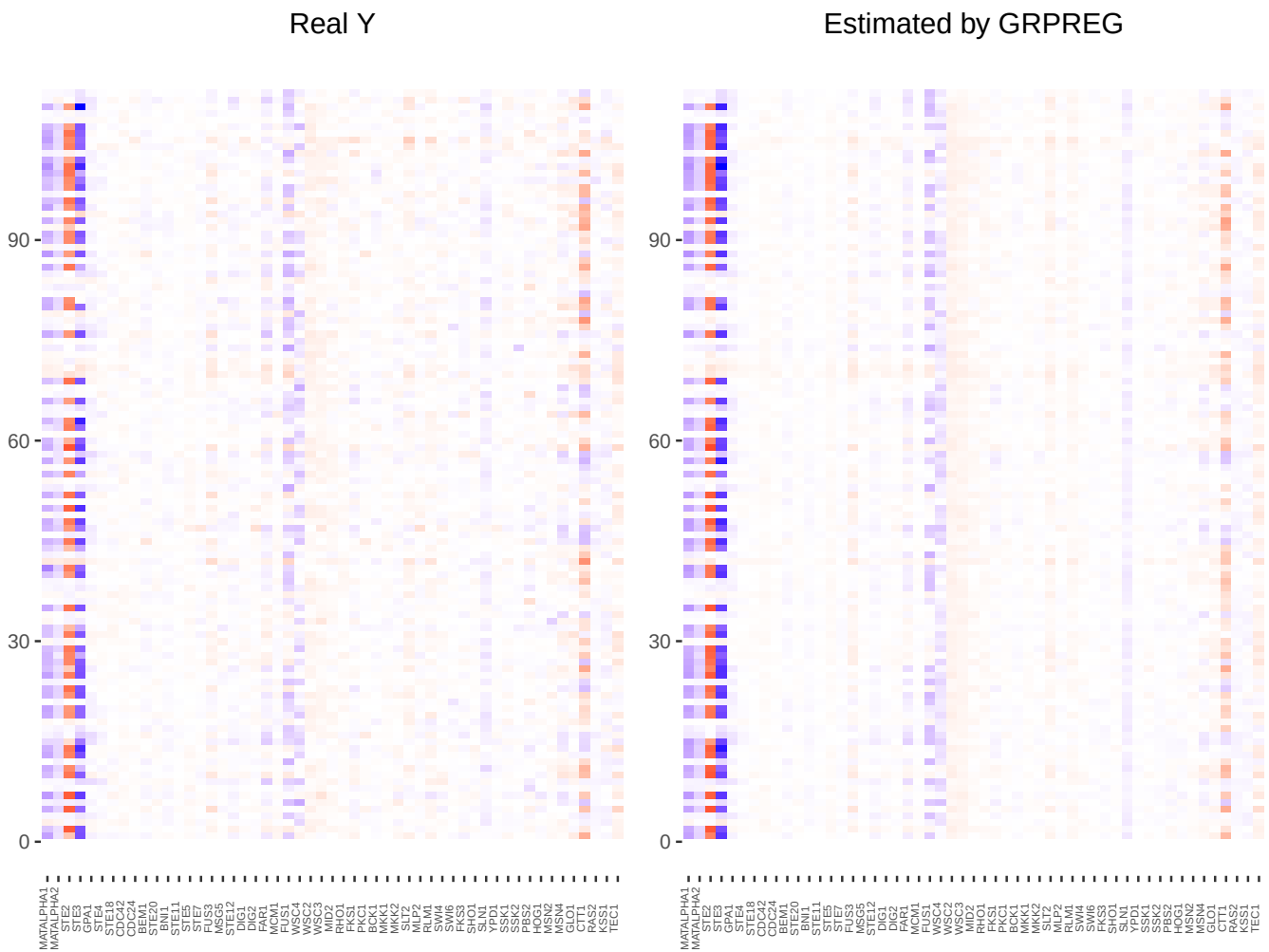


Fig. 2. Heatmaps. $\widehat{B}$ has full column rank and 201 non-zero rows.

Biological Discovery

Previous biological research has revealed some facts which could be the guidelines for us to choose suitable method of data analysis.

- Biological finding:
  Each signaling pathway involves only a subset of genes[1], which are regulated by only a few genetic variants.
- Corresponding characteristics in statistics:
  The association structure between the eQTLs and the gene is of low rank and sparsity.

[1] Gustin, M. C., Albertyn, J., Alexander, M. and Davenport, K. (1998) Map kinase pathways in the yeast saccharomyces cerevisiae. Microbiology and Molecular Biology Reviews, 62, 1264–1300.

Multi-Response Regression

SOFAR [1] uses the SVD decomposition $B = UDV^T$ and then impose penalties into U, D and V respectively.

$$
\begin{aligned}
&(\widehat{D}, \widehat{U}, \widehat{V}) \\
=&\underset{D,U,V}{\arg\min} \left\{ \frac{1}{2} \left\| X - UDV^T \right\|_F^2 + \lambda_d \|D\|_1 + \lambda_a \rho_a(UD) + \lambda_b \rho_b(VD) \right\} \qquad (2) \\
&\text{subject to } U^T U = \mathbf{I}_m, \quad V^T V = \mathbf{I}_m
\end{aligned}
$$

[1] Uematsu, Y., Fan, Y., Chen, K., Lv, J., & Lin, W. (2019). SOFAR: large-scale association network learning. IEEE Transactions on Information Theory.

Result of SOFAR

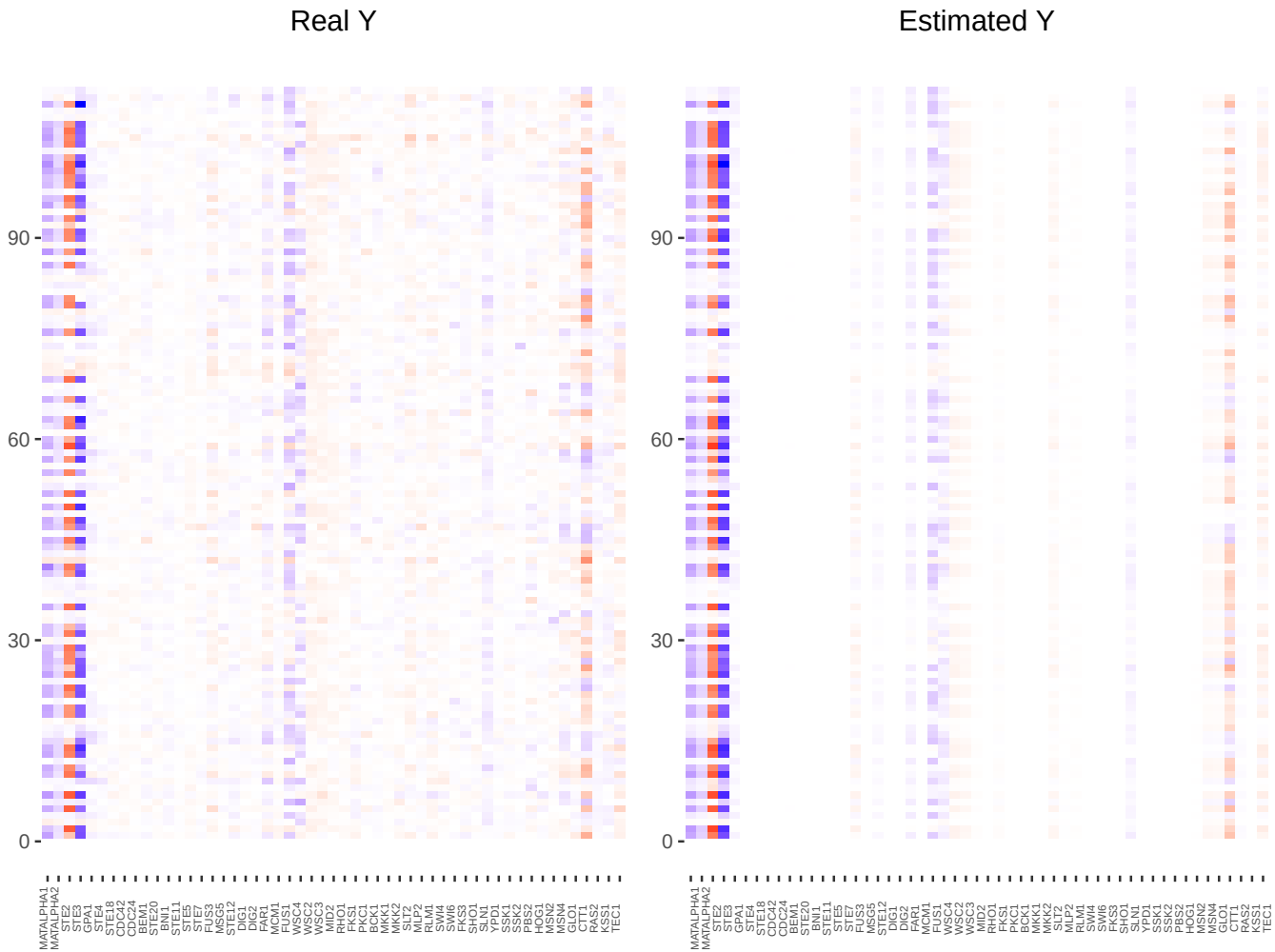


Fig. 3. Heatmaps of real Y and $\hat{Y}$ by SOFAR

Result of SOFAR cont.

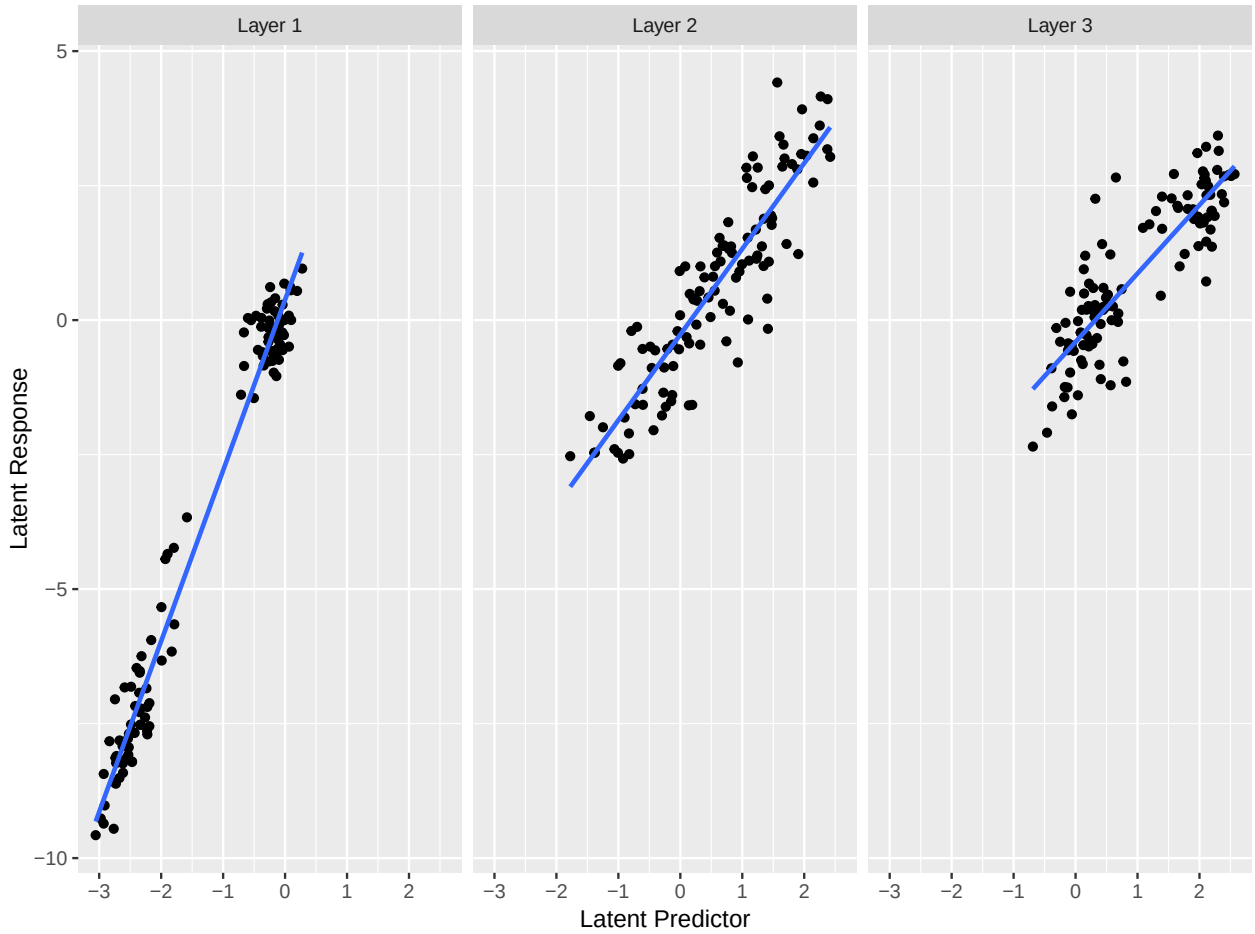


Fig. 4. Scatter plots of the latent responses versus the latent predictors in three SVD layers for the yeast data estimated by the SOFAR method

Further Reduce Dimension of X

We performed a marginal gene-marker association analysis to identify SNPs that are associated with the expression levels of at least two genes with a p-value less than 0.05, resulting in a total of $p = 776$ variables.
Reduce additional 18.2% Xs.
Result: rank 3 with 228 non-zero rows in the estimation of U and 25 non-zero rows in the estimation of V.

Result of SOFAR after Reduction

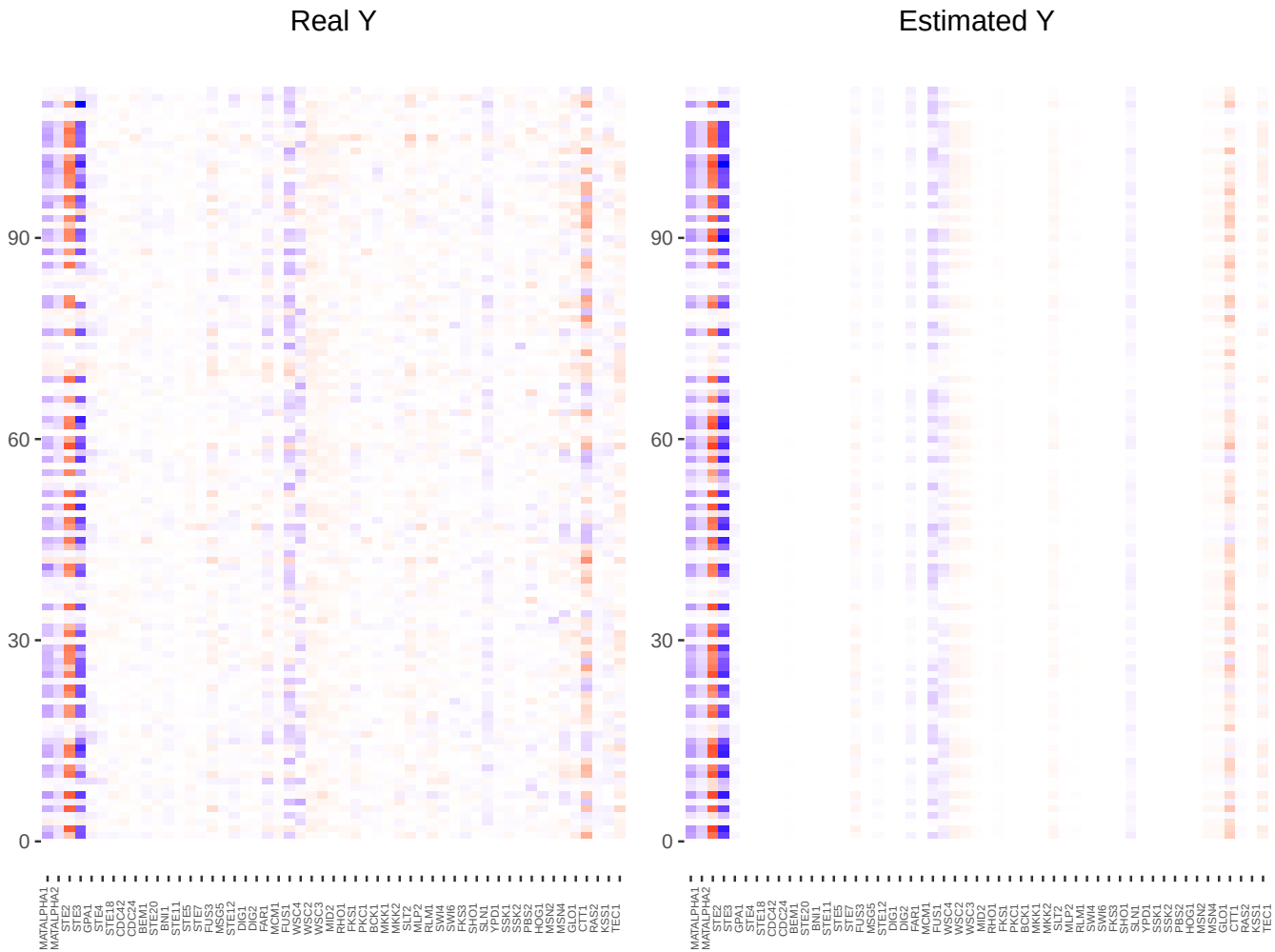


Fig. 5. Heatmaps of real Y and $\hat{Y}$ by SOFAR

Result of SOFAR after Reduction

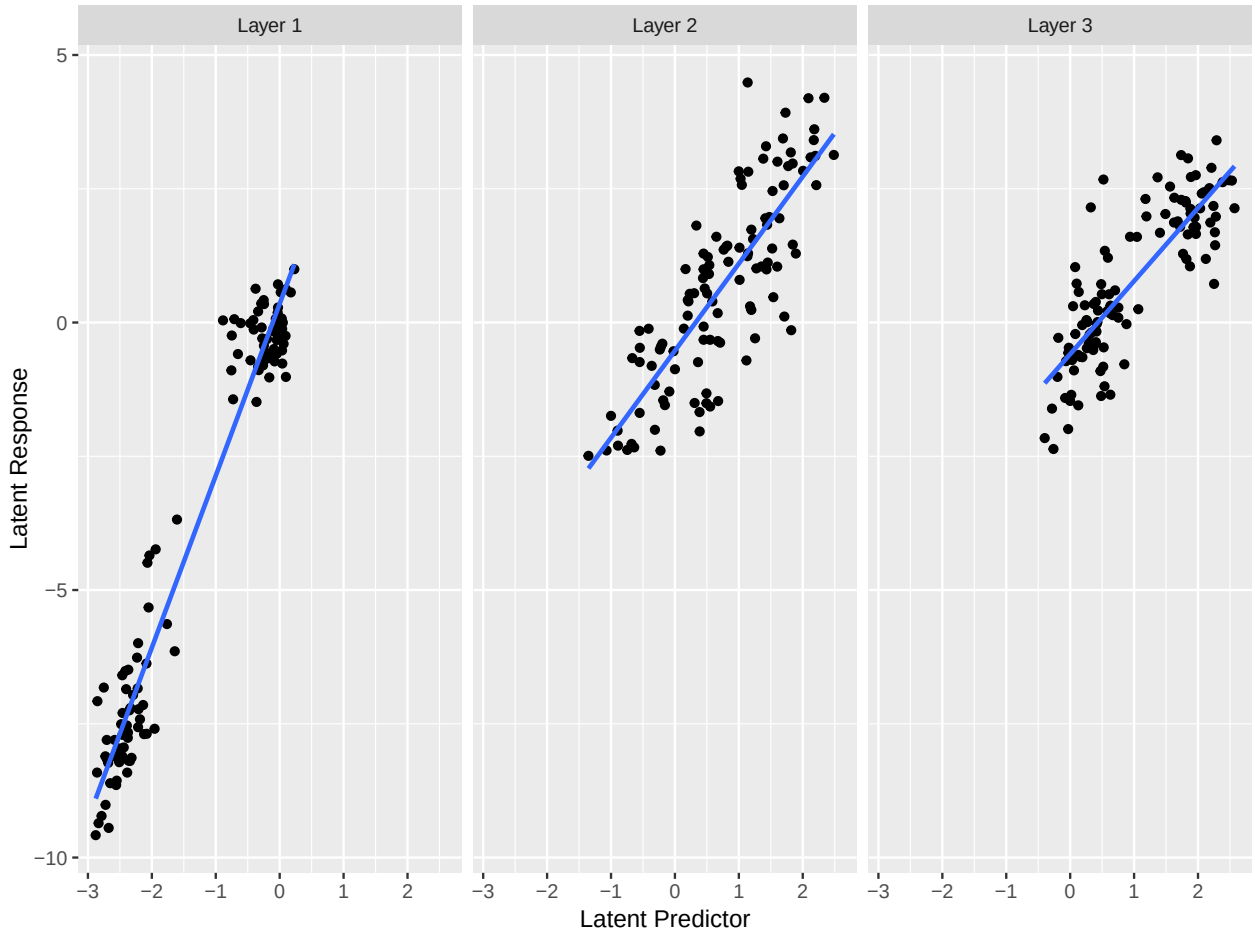


Fig. 6. Scatter plots of the latent responses versus the latent predictors in three SVD layers for the yeast data estimated by the SOFAR method

4th Letent Pattern

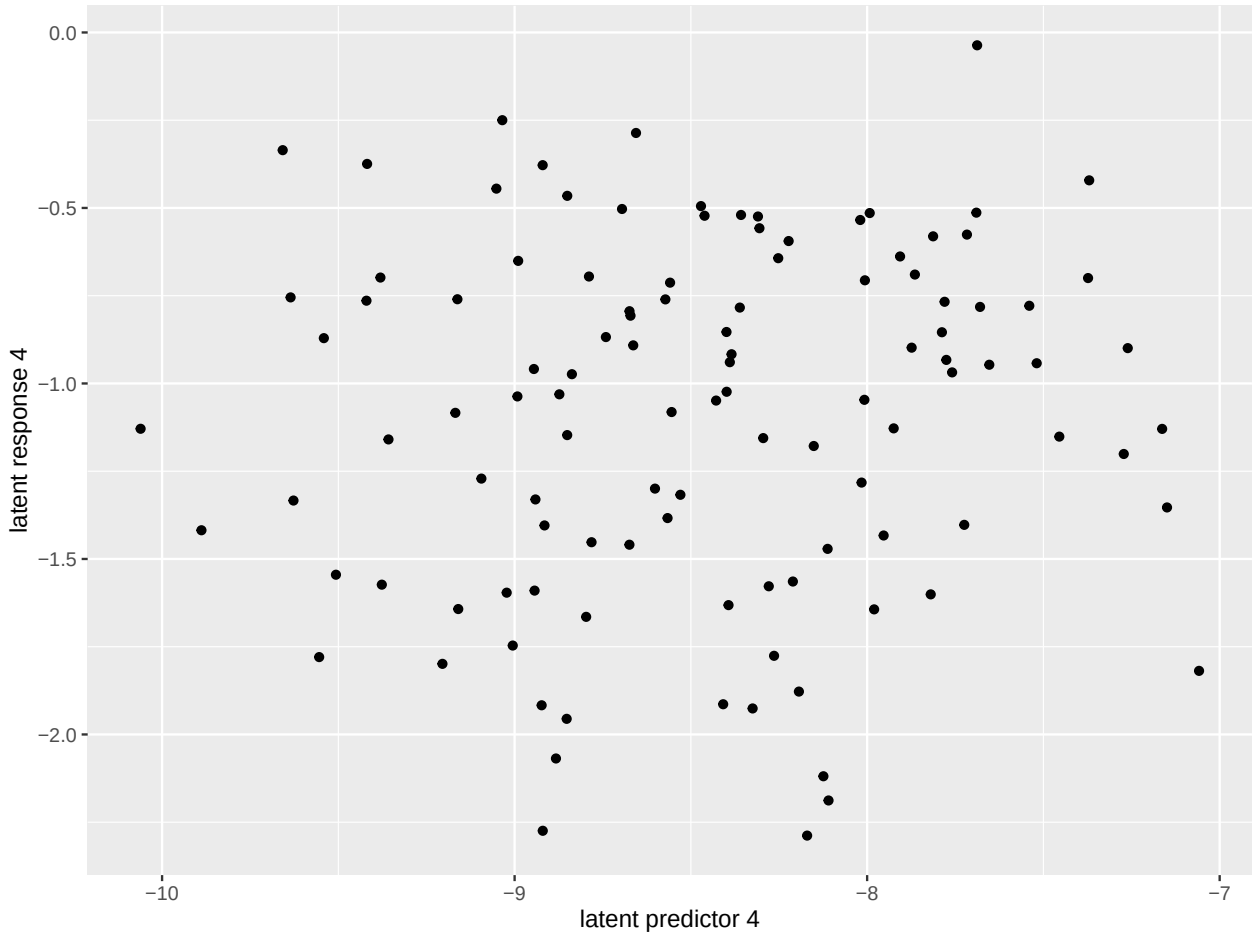


Fig. 7. Scatter plots of the 4th latent responses versus the latent predictors for the yeast data estimated by the SOFAR method

Section 5

Summary

Comparison of Different Methods

- **Entrywise Sparse**: Lasso, SCAD, etc.
- **Rowwise Sparse**: Group sparse linear regression.
- **Sparse & Low Rank**: SRRR, SOFAR, SEED, etc.

Biological Implications.

Biological interpretation and significance of our results

- Results suggest: certain linear combination of eQTLs have effect on a subset of genes, and the groups are orthogonal may offer new information about structure of the genetic variants and gene expressions.
- Results indicate: there may be only 3 types of patterns in MAPK signal pathways.

Future Work

- Find some other genes...
- Try to control FDR...

Thank you !