A Robust Sparse Correlation Matrix Estimator for Metabolomics

CHELSY NGUYEN AND CHRIS SOBCZAK

Outline

Omics

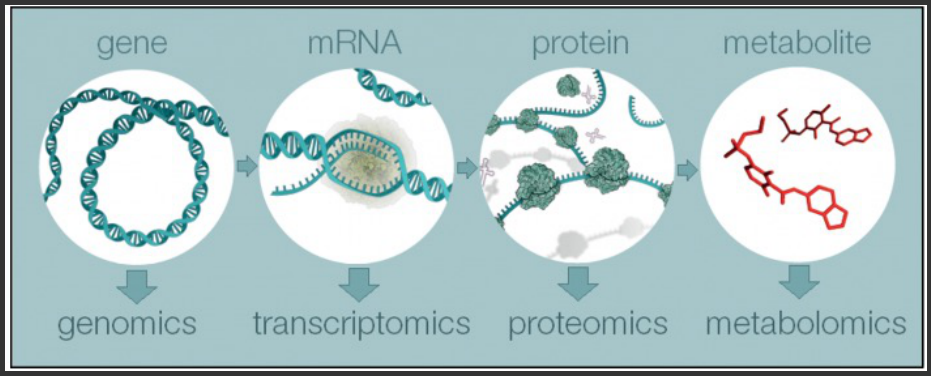

Figure 1: An overview of the research areas within *omics* (Tong and Bangzhuo 2023)

- Omics is the study of all the kinds of molecules in all processes of an organism
- The metabolome is the collection of all small molecules in an organism as well as molecules introduced from the environemnt (ex: carbohydrates, penicillin and 3PBA)

Genetics

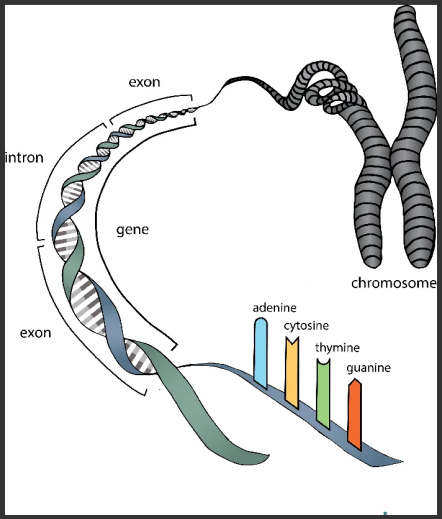

Figure 2: Diagram of genetic organization (Shaer et al. 2017)

- The genome is the collection of all (deoxyribonucleic acid) DNA in an organism
- DNA is a sequence of an alphabet of 4 molecules (ACTG)
- The human genome contains about 3 billion of those molecules
- Genes are a higher level unit of information drawn from the genome
- Genes in different individuals will be expressed slightly differently based on that individuals unique version of the gene (this variation can be measured)

Gene Expression and Metabolites

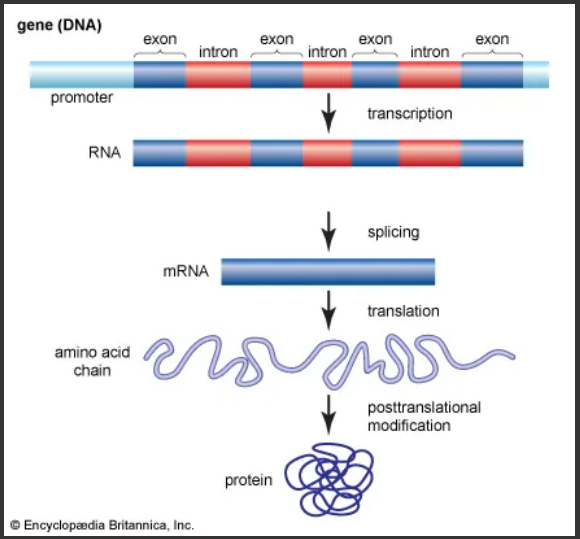

Figure 3: Diagram of transcription (Britannica 2026)

- Genes eventually lead to a certain amounts and types of proteins
- Proteins then act as fuel for the chemical reactions changing levels of metabolites
- Some genes act alone, and others require a partner
- The estimate correlation matrix gives us clues about the joint expression of genes and then their function
- To measure gene expression, microarray experiments are conducted

Measuring Gene Expression

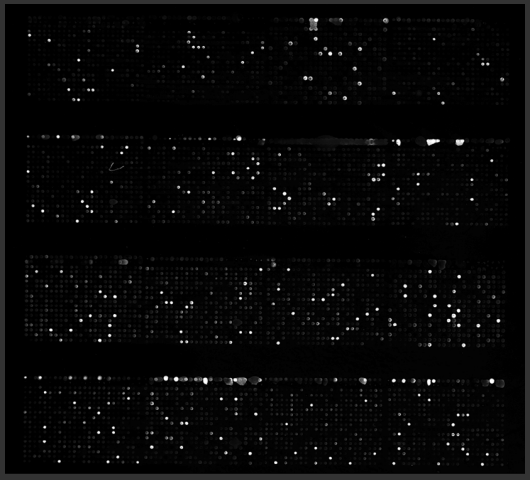

Figure 4: Image of a microarray chip associated with one experimental unit (Joseph and P S 2023)

- The intensity of the each dot in Figure 4 is related to the gene expression of the specific gene associated with that location on the array chip
- The number of dots is p
- One of these microarray experiments represents one of n individuals in a study
- So this results in a data matrix $X_{n\times p}$
- The correlation between these gene intensity patterns and a trait or disease is the usual area of interest in genomics

Correlation in Genomics and Metabolomics

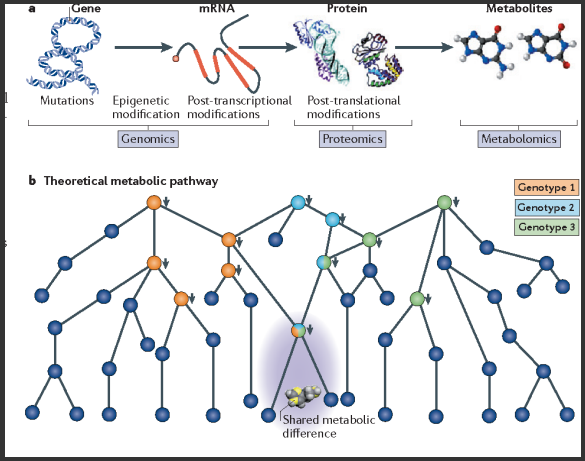

Figure 5: The relationship between genes and metabolites (Yanes and Siuzdak 2012)

- In genomics usually $p >> n$, there are heavy tails and outliers and the true correlation matrix may be sparse
- The same challenges exist for estimating correlations between metabolites
- Microarray experiments measure levels of p metabolites
- Estimating the correlation between metabolites helps to understand the organic processes leading to disease or longevity

Pearson Correlation Coefficient

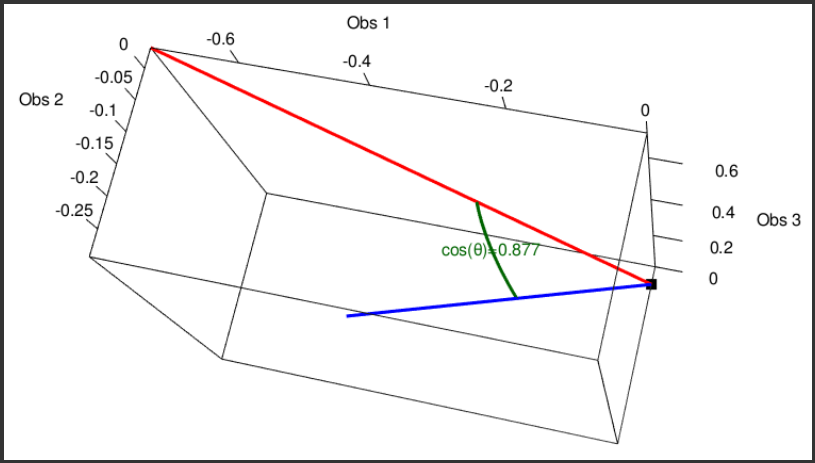

Figure 6: Pearson correlation coefficent represented as the cosine of the angle formed between two random vectors in $\mathbb{R}^n$ (A. 1924)

Correlation Matrix Estimation

- Existing correlation estimators perform poorly when:
  1. $p >> n$
  2. Data has heavy tails and outliers
  3. True correlations may be *rare* relative to p
- Pearson (represented in Figure 6)
  - The sample correlation between two covariates is the *cosine of the angle formed between the vectors m and l in* $\mathbb{R}^n$ (A. 1924)
  - Measuring angles between too many vectors (p) in too few dimensions (n)
- Spearman: Angle between rank vectors, more robust
- Kendall: Proportion of agreeing pairs
- Serra et al. 2017 proposes an estimator that is robust to outliers and avoids spurious correlations due to high-dimensionality, therefore a *robust sparse correlation* (RSC) estimator

RSC Solution

- RSC improves the performance of correlation estimation in 2 ways:
  - Improves robustness by using the median absolute deviation as the measure of variability (Pasman and Shevlyakov 1987)
  - Removes spurious correlations by adaptive thresholding

Median Absolute Deviance (MAD)

- y is a set of k measurements of some gene and $\text{med}(y)$ is the median of those measurements
- X_l is the n measurements for gene l
- X_m is the m measurements for gene m
- $\text{R}_{\text{MAD}}(X_l, X_m)$ is the MAD correlation between genes l and m, the robust correlation

$$\text{MAD}(y) = \text{med}(|y_l - \text{med}(y)|, l = 1, 2, \ldots, k) \tag{1}$$

$$u = \frac{X_l - \text{med}(X_l)}{\sqrt{2}\text{MAD}(X_l)} + \frac{X_m - \text{med}(X_m)}{\sqrt{2}\text{MAD}(X_m)} \tag{2}$$

$$v = \frac{X_l - \text{med}(X_l)}{\sqrt{2}\text{MAD}(X_l)} - \frac{X_m - \text{med}(X_m)}{\sqrt{2}\text{MAD}(X_m)} \tag{3}$$

$$\text{R}_{\text{MAD}}(X_l, X_m) = \frac{\text{MAD}^2(u) - \text{MAD}^2(v)}{\text{MAD}^2(u) + \text{MAD}^2(v)} \tag{4}$$

Adaptive Thresholding

- Hard thresholding is applied to $\mathrm{R}_{\mathrm{MAD}}$ to dampen the effects induced by high dimensionality. The hard thresholding operator is:

$$T_h(\mathrm{R}_{\mathrm{MAD}}) = [\mathrm{R}_{\mathrm{MAD}}(X_l, X_m)\mathbb{I}\{|\mathrm{R}_{\mathrm{MAD}}(X_l, X_m)| \geq h\}]_{l,m} \tag{5}$$

- That is, any correlation between genes l and m under threshold h drops to 0

The optimal threshold is chosen using cross validation (CV) to minimize a risk function.

Steps to Hard-Threshold R_{MAD}

Given a sample correlation matrix R_{MAD}, we follow the steps to find the "expected risk" of a given threshold h:

- Step 1: Split the sample X randomly into two subsets of size n_1 and n_2, repeating K times.
- Step 2: Calculate $R_{MAD}^{(1,k)}$, the robust correlation matrix based on the n_1 samples, and $R_{MAD}^{(2,k)}$, based on the n_2 samples.
- Step 3: Threshold $R_{MAD}^{(1,k)}$ using h
- Step 4: Calculate the loss of the k-th fold using squared Frobenius Norm.
- Step 5: Average the loss over all k's to get the expected risk.

We repeat the steps for a list of threshold, and select h with minimum expected risk.

Cross Validation

$$\mathrm{CV}(h) = \frac{1}{K}\sum_{k=1}^{K} LOSS\left[T_h\left(\mathrm{R}_{\mathrm{MAD}}^{(1,k)}\right) - \mathrm{R}_{\mathrm{MAD}}^{(2,k)}\right] \tag{6}$$

- Set a grid of all h values to try
- For each h, we threshold $R_{MAD}^{(1,k)}$ and measure the distance from the unthresholded $R_{MAD}^{(2,k)}$
- We are trying to find the level of R_{MAD} that is associated with spurious correlations
- Here, $LOSS() = ||A||_F^2 = \sum_l \sum_m a_{l,m}^2$
- However, there may be alternative $LOSS$ functions that would perform better in identifying the level of spurious correlations

Implementation in R

- For this project, we implemented a simplified version of RSC in R and proposed 2 alternative hard-thresholding functions.
- The purpose is to examine the weaknesses of RSC, its performance on genetics and metabolomics data, and motivate further improvement.
- We will compare our implemented RSC (2 versions) with the sample Pearson correlation estimation and the packaged RSC version (RSC CRAN).

RSC Weakness - Cross Validation process

The main weakness we have discovered lies in the adaptive thresholding process.

- Cross-validation is computationally demanding.
- Each fold k, RSC requires:
  - Splitting the data in half (randomly)
  - Calculating $\mathrm{R}_{\mathrm{MAD}}^{(1,k)}$ and $\mathrm{R}_{\mathrm{MAD}}^{(2,k)}$. Each will have size $p \times p$.
  - Thresholding $\mathrm{R}_{\mathrm{MAD}}^{(1,k)}$
  - Calculating the squared Frobenius distance.
- The steps are repeated for each threshold h.

RSC Weakness - Loss function

With squared Frobenius loss:

- The original RSC uses the squared Frobenius loss as the cross-validation criterion.
- Squaring the differences means a single entry where R_1 and R_2 disagree strongly can dominate the total loss and pull $h*$ toward a suboptimal value.

RSC Weakness - Loss function

Can we improve the efficiency and robustness of RSC through alternative loss functions?

- Rank-based loss function
- L1-based loss function

Data

To evaluate the performance of the alternative thresholding methods, we use 3 different datasets:

1. Simulated data based on a partially sparse correlation matrix with an $AR(1)$ block, seen in Figure 7
2. Genomics data related to expression-based drug screening of individuals with schizophrenia (Readhead et al. 2018)
3. Metabolomics data related to a study targeting the early diagnosis of multiple sclerosis (Herman et al. 2018)

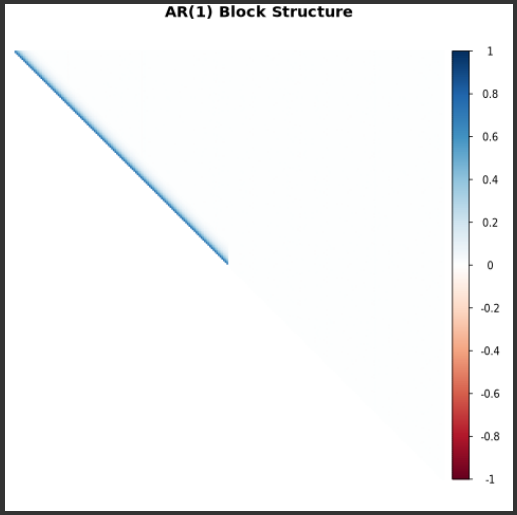

Figure 7: Correlation structure of simulated data

Alternative 1. Rank-based (1)

- When $p >> n$, there may be many lucky alignments (reference Figure 6)
- RSC still produces many false negatives with the thresholding method
- Cross validation is computationally demanding and selects a *global sparsity level*
- We are interested in the stability of a given edge
- We target stability across subsamples instead of correlation magnitude

Alternative 1. Rank-based (2)

- If the true correlation matrix was sparse:
  - On each split, we expect a given l and m pair to appear amoung the top q percent equally (with probability q)
  - The correlations should appear in random permutations
- Over K splits:

$$K \cdot \pi_{l,m} \sim \text{Binomial}(K, q) \tag{7}$$

- Where $\pi_{l,m}$ is the fraction of subsamples $r_{l,m}$ appears among the top part of the ranking

Alternative 1. Rank-based (3)

Rank-frequency stability:

- On each of k splits, compute R_{MAD} and rank correlations by magnitude
- For a top-rank fraction q, record whether each pair is in the top q percent
- For each pair l, m, compute its rank-frequency:

$$\pi_{l,m} = \frac{1}{K} \sum_{k=1}^{K} \mathbb{I}\{\text{rank}_k(l, m) \leq q\} \tag{8}$$

- If $\pi_{l,m} \approx q$, this is likely noise
- If $\pi_{l,m} \approx 1$, this is likely structure

Alternative 1. Rank-based (4)

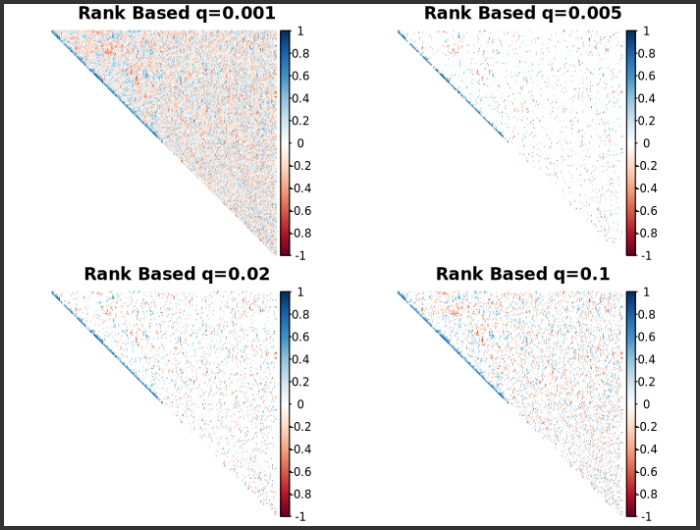

Figure 8: Rank based regularization with different q

Alternative 1. Rank-based (4)

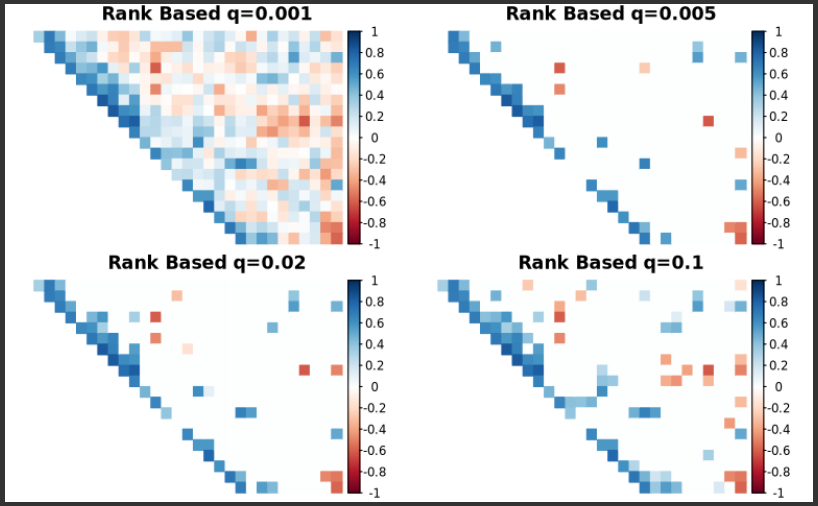

Figure 9: Rank based regularization with different q, zoomed in

Alternative 2. L1-based (1)

- L1 loss is the matrix-level analog of replacing the mean with the median - a classical principle in robust statistics (Huber, 1964).
- The squared loss function results in an arithmetic mean-unbiased estimator, while the absolute-value loss function results in a median-unbiased estimator.

Alternative 2. L1-based (2)

Pairwise L1 loss:

$$L_{L1}(h) = \sum_{l<m} |T_h(\hat{R}_{1,lm}) - \hat{R}_{2,lm}|$$

$$L_{L1}(h) = \sum_{l=1}^{m} |r_{2,l}| + \sum_{|r_{1,l}\geq h|} (|r_{1,l} - r_{2,l}| - |r_{2,l}|)$$

- When r_1 and r_2 agree ($|r_{1,l} - r_{2,l}| \approx 0$), keeping the entry reduces loss L by $|r_{2,l}|$.
- If r_1 and r_2 disagree, keeping the entry will increase loss.

Alternative 2. L1-based (3)

RSC comparison for $AR(1)$ (n=50, x=250) with K = 25.

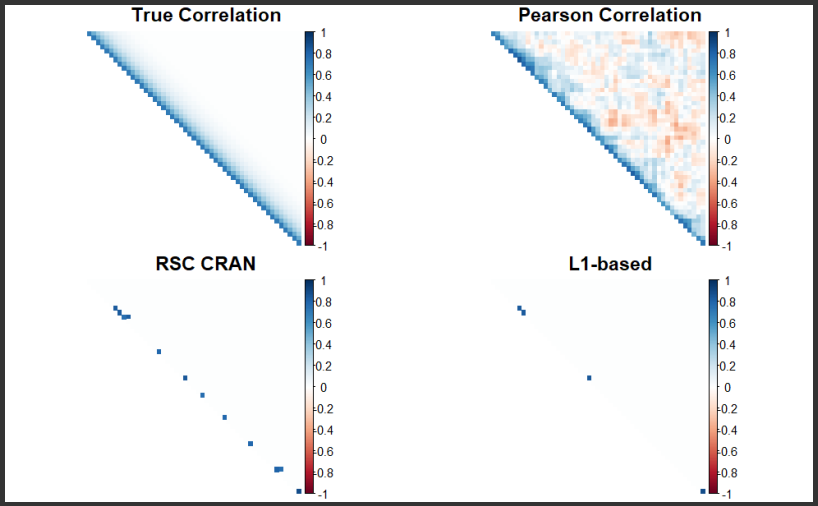

Figure 10: Sample correlation of the first 50 columns

Alternative 2. L1-based (4)

Comparing thresholds with simulated data (x=100, K = 10, n=seq(20, 300, by = 10))

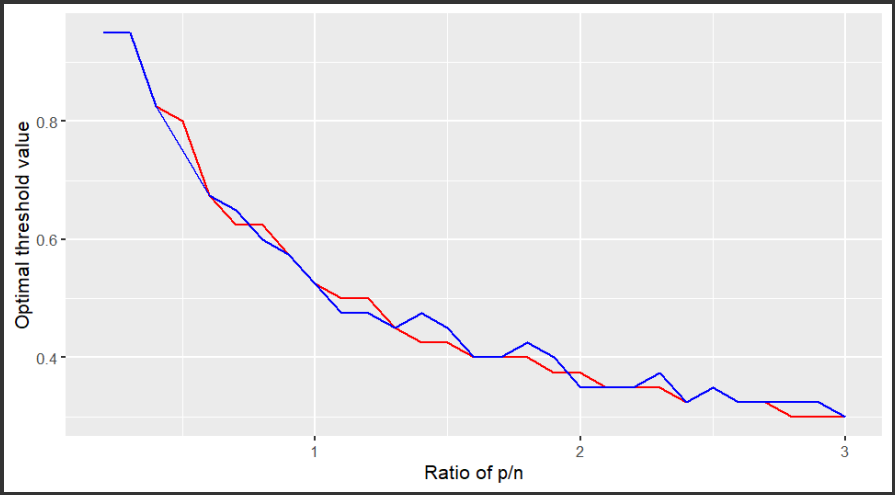

Figure 11: Comparison of optimal threshold (red=RSC, blue=L1-based RSC)

Alternative 2. L1-based (4)

Eigenvalues comparison for $AR(1)$ (n=50, x=250) with K = 25.

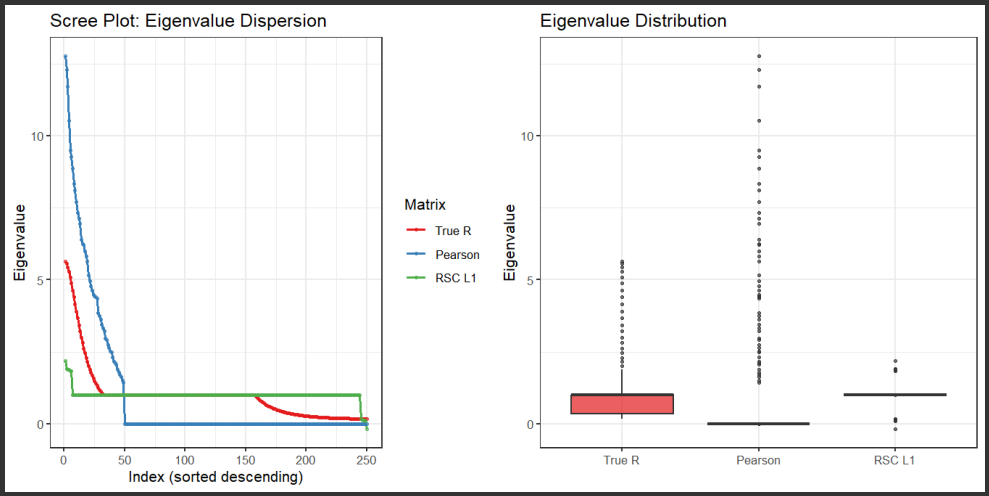

Figure 12: Comparison of eigenvalues obtained using different sample correlation

Alternative 2. L1-based (6)

Eigenvalues comparison for $AR(1)$ (n=50, x=250) with K = 25.

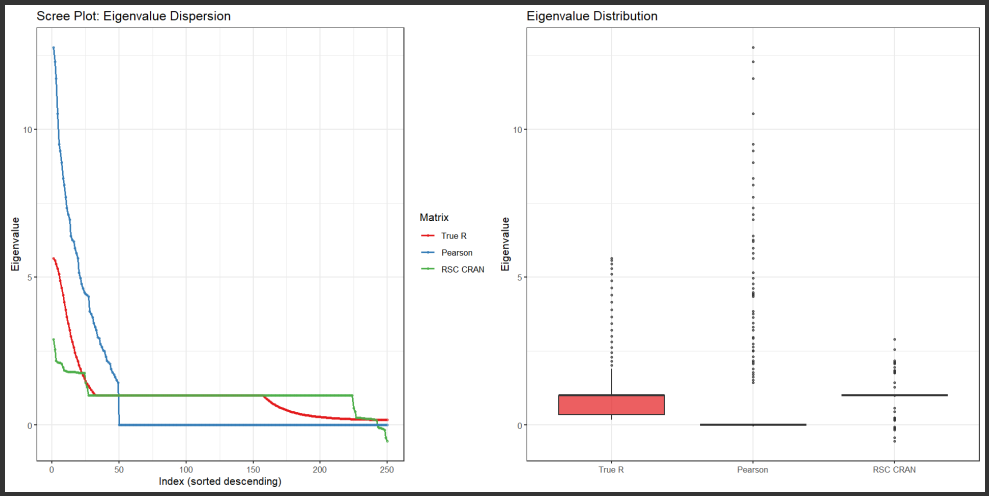

Figure 13: Comparison of eigenvalues obtained using different sample correlation

Final Results

Sample correlation matrix estimated with 4 methods - $AR(1)$

- R_{MAD} **(unthresholded)**: Generally results in more noise than Pearson, but robust to outliers
  - The noise is removed with thresholding
- **RSC**: The original RSC estimator does well in removing false positives, but at the cost of false negatives.
- **RSC Rank**: R_{MAD} with rank based regularization performs better than Pearson, removing false positives but retaining more than necessary that RSC removes
- **RSC L1**: R_{MAD} with thresholding set by L1-based loss function provides similar results to original RSC

Future Directions to Improve RSC

- Replace cross-validation with a different method to find optimal threshold.
  - Expand on rank based method
  - Further develop the rank stability criteria
  - Test on various correlation structures and data contexts
- Experiment with other robust correlation estimators based on Pasman and Shevlyakov 1987.
- Integrate a process to handle missing values in data matrix.
- Further testing of RSC on metabolomics data, identifying the unique challenges there that do not exist in genomics correlation estimation.

Summary

- Robust and Sparse Correlation matrix estimation (RSC) has many advantages over Pearson correlation for high-dimensional data.
- RSC consists of 2 steps - estimating R_{MAD} and adaptive thresholding.
- RSC can be further improved to be more robust, computationally efficient, and suitable for metabolomics data.

Bibliography I

- A., FISHER R. (1924). "The distribution of the partial correlation coefficient". In: Metron 3, pp. 329–332. URL: https://cir.nii.ac.jp/crid/1572824500366447104.
- Pasman, VR and Georgy Leonidovich Shevlyakov (1987). "Robust methods of estimating the correlation coefficient". In: Avtomatika i Telemekhanika 3, pp. 70–80.
- Yanes, Óscar and Gary Siuzdak (2012). "Metabolomics: the apogee of the omics trilogy". In: URL: https://api.semanticscholar.org/CorpusID:17912313.
- Serra, Angela et al. (Oct. 2017). "Robust and sparse correlation matrix estimation for the analysis of high-dimensional genomics data". In: Bioinformatics 34.4, pp. 625–634. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btx642. eprint: https://academic.oup.com/bioinformatics/article-pdf/34/4/625/48913355/bioinformatics_34_4_625.pdf. URL: https://doi.org/10.1093/bioinformatics/btx642.

Bibliography II

- Shaer, Orit et al. (Jan. 2017). "Communicating Personal Genomic Information to Non-experts: A New Frontier for Human-Computer Interaction". In: Foundations and Trends® in Human-Computer Interaction 11, pp. 1–62. DOI: 10.1561/1100000067.
- Herman, Stephanie et al. (2018). "Integration of magnetic resonance imaging and protein and metabolite CSF measurements to enable early diagnosis of secondary progressive multiple sclerosis". In: Theranostics 8.16, pp. 4477–4490. ISSN: 1838-7640. DOI: 10.7150/thno.26249. URL: http://dx.doi.org/10.7150/thno.26249.
- Readhead, Benjamin et al. (Oct. 2018). "Expression-based drug screening of neural progenitor cells from individuals with schizophrenia". In: Nature Communications 9.1. ISSN: 2041-1723. DOI: 10.1038/s41467-018-06515-4. URL: http://dx.doi.org/10.1038/s41467-018-06515-4.

Bibliography III

- Joseph, Steffy and Sathidevi P S (Apr. 2023). "Microarray Image Lossless Compression Using General Entropy Coders and Image Compression Standards". In: Circuits, Systems, and Signal Processing 42, pp. 1–28. DOI: 10.1007/s00034-023-02347-w.
- Tong, Hua Zou and Bangzhuo (Dec. 2023). Data Analysis in Metabolomics. URL: https://xbiomeanalysis.github.io/Metabolomics_Aanlysis/.
- Britannica (Mar. 2026). Gene | definition, structure, expression, & facts | britannica. URL: https://www.britannica.com/science/gene.