Groupwise cross-correlation mining in bi-view data

Finding stable groups of cross-correlated features with application to eQTL analysis

Miheer Dewaskar

UNC Chapel Hill & Duke University

1st Aug 2022
IISER Pune

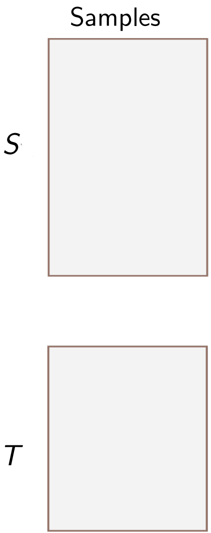
Outline

1. Bimodules: groups of significant cross-correlated features in bi-view data
    - Bi-view data and Bimodules
    - Stable bimodules and the Bimodule Search Procedure (BSP)
2. Application to genomics
    - Introduction to eQTL analysis
    - Using BSP for groupwise eQTL analysis
3. Theoretical analysis of BSP
    - Asymptotics of BSP
    - Null correlation networks

Outline

1. Bimodules: groups of significant cross-correlated features in bi-view data
   - Bi-view data and Bimodules
   - Stable bimodules and the Bimodule Search Procedure (BSP)
2. Application to genomics
   - Introduction to eQTL analysis
   - Using BSP for groupwise eQTL analysis
3. Theoretical analysis of BSP
   - Asymptotics of BSP
   - Null correlation networks

Bi-view data



Measurements of two types of features

$S = \{s_1, \dots, s_p\}$ & $T = \{t_1, \dots, t_q\}$

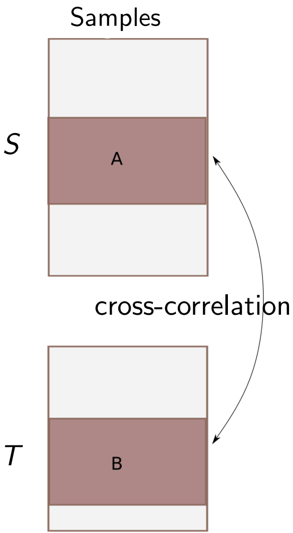
on n common samples. Typically $p, q \geq n$.

Examples

- Samples represent time and measure:
  S $=\{p$ temperature stations$\}$ and
  T $=\{q$ precipitation stations$\}$ worldwide.
- Samples represent habitats and measure
  S $=\{p$ environmental features$\}$ and
  T $=\{q$ microbial species$\}$ abundance.

How are features from S and T associated?

Exploratory problem of interest



We distinguish between two types of correlations

cross-correlation (CC) b/w features $s \in S$ and $t \in T$

intra-correlation b/w features $s, s' \in S$ or $t, t' \in T$.

Bimodule (rough definition)

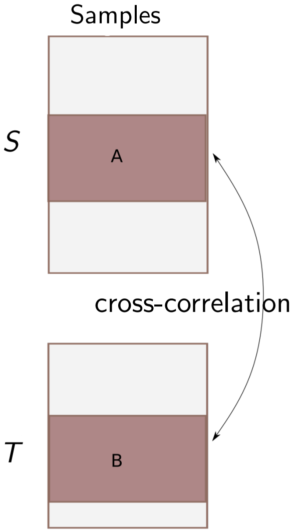
(A, B) is a bimodule if

- $A \subseteq S$ and $B \subseteq T$
- A and B have significant aggregate CC.

Motivation to aggregate CCs

- Capture complex associations between feature groups A and B
- Improve power by amplifying weak signal

Exploratory problem of interest



We distinguish between two types of correlations

cross-correlation (CC) b/w features $s \in S$ and $t \in T$

intra-correlation b/w features $s, s' \in S$ or $t, t' \in T$.

Bimodule (rough definition)

(A, B) is a bimodule if

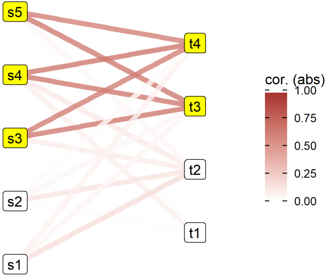
- $A \subseteq S$ and $B \subseteq T$
- A and B have significant aggregate CC.

Motivation to aggregate CCs

- Capture complex associations between feature groups A and B
- Improve power by amplifying weak signal

Network perspective

cross-correlation networks



$S = \{s_1, \ldots, s_5\},\ T = \{t_1, \ldots, t_4\}$

Weights: sample correlation (abs.)

Bimodules: communities in this network.

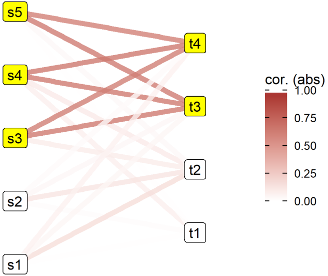
Example: $A = \{s_3, s_4, s_5\}$ and $B = \{t_3, t_4\}$.

Community (rough definition)

Nodes in a community are more correlated, on average, to nodes inside the community than to nodes outside.

Network perspective

cross-correlation networks



$S = \{s_1, \ldots, s_5\}$, $T = \{t_1, \ldots, t_4\}$

Weights: sample correlation (abs.)

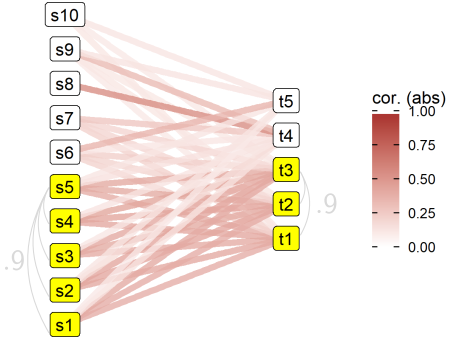
Bimodules: communities in this network.

Example: $A = \{s_3, s_4, s_5\}$ and $B = \{t_3, t_4\}$.

Community (rough definition)

Nodes in a community are more correlated, on average, to nodes inside the community than to nodes outside.

Almost cross-correlation communities: role of intra-correlations



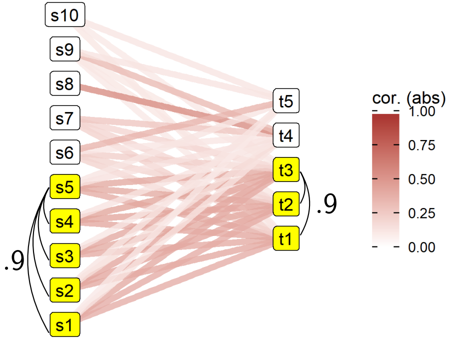
(A, B) is a community in the CC network.

Likely to see this community by chance in random data? Yes

- Depending only on CC can mislead.
- Must account for *intra-correlations* while assessing bimodule significance.

$$A = \{s_1, \ldots, s_5\}, \; B = \{t_1, t_2, t_3\}$$

Almost cross-correlation communities: role of intra-correlations



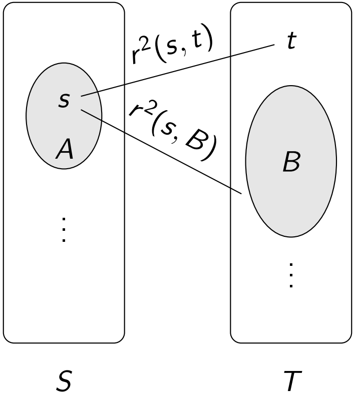
(A, B) is a community in the CC network.

Likely to see this community by chance in random data? Yes

- Depending only on CC can mislead.
- Must account for *intra-correlations* while assessing bimodule significance.

$$A = \{s_1, \ldots, s_5\},\ B = \{t_1, t_2, t_3\}$$

Stable Bimodules



$r(s,t)$: sample correlation of s, t

$r^2(A', B') \doteq \sum_{s \in A'} \sum_{t \in B'} r^2(s,t)$

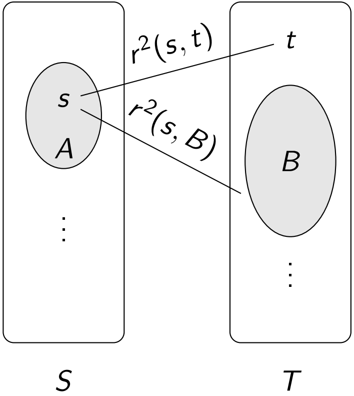
Stable bimodule (definition)

(A, B) is a *stable bimodule* if

$$A = \{s \in S \mid r^2(s,B) \text{ is significant}\}, \text{ and}$$
$$B = \{t \in T \mid r^2(A,t) \text{ is significant}\}.$$

- Recursive definition like a community based on aggregate correlations $r^2(s,B)$ & $r^2(A,t)$.
- Interest in connected stable bimodules.
- “Significance” quantified using hypothesis testing that accounts for inflation in variance of $r^2(s,B)$ due to intra-correlations.

Bimodule Search Procedure (BSP)



$r(s,t)$: sample correlation of s & t
$r^2(A,B) \doteq \sum_{s\in A} \sum_{t \in B} r^2(s,t)$

Stability is equivalent to $(A,B) = (\Gamma_S(B), \Gamma_T(A))$ where

$$\Gamma_S(B) \doteq \{s \in S \mid r^2(s,B) \text{ is significant}\}$$
$$\Gamma_T(A) \doteq \{t \in T \mid r^2(A,t) \text{ is significant}\}.$$

Hence, we can find stable bimodules by iterating

$$B_k = \Gamma_T(A_{k-1}); A_k = \Gamma_S(B_k) \quad k = 1, 2, \ldots$$

till sets don't change, starting from suitable $A_0 \subseteq S$.

Bimodule Search Procedure (BSP)

Starting from singletons $A_0 = \{s\} \subseteq S$, iterate the definition till fixed point is reached (or sets cycle).

Covergence on real data
Example of an iterative search

Quantifying significance using hypothesis testing

How to quantify Γ_T defined as:

$$\Gamma_T(A) \doteq \{t \in T \mid r^2(A,t) \text{ is significant}\}.$$

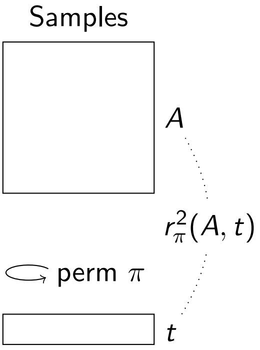
Steps

1. $\forall t \in T$ obtain p-value $p(A,t)$ from $r^2(A,t)$ (see right)
2. reject p-values using multiple-testing correction γ_α

$$\Gamma_T(A) = \{t \in T \mid p(A,t) \le \gamma_\alpha\}$$

at some level $\alpha \in (0,1)$.

Multiple testing correction The adaptive threshold γ_α chosen from [Benjamini and Yekutieli, 2001] controls FDR at α.

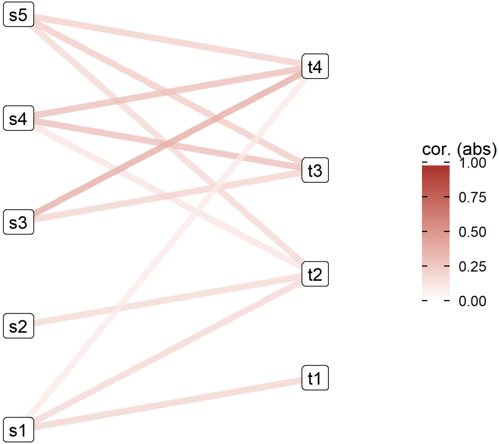


Permutation p-value

$$\mathbb{P}_\pi\left(r^2_\pi(A,t) \ge r^2_{obs}(A,t)\right)$$

Fast analytical approximation

Bimodule Search Procedure: example run

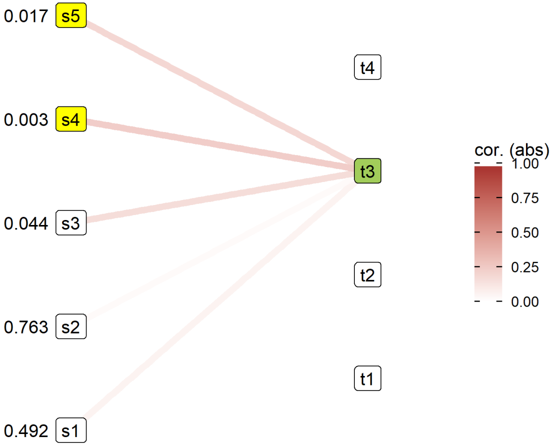


1. $B_0 = \{T_3\}$
2. $A_0 = \{S_4, S_5\}$
3. $B_1 = \{T_3, T_4\}$
4. $A_1 = \{S_3, S_4, S_5\}$
5. $B_2 = \{T_3, T_4\}$
6. $A_2 = \{S_3, S_4, S_5\}$

$(A_1, B_1) = (A_2, B_2)$

Stable bimodule found.

Bimodule Search Procedure: example run

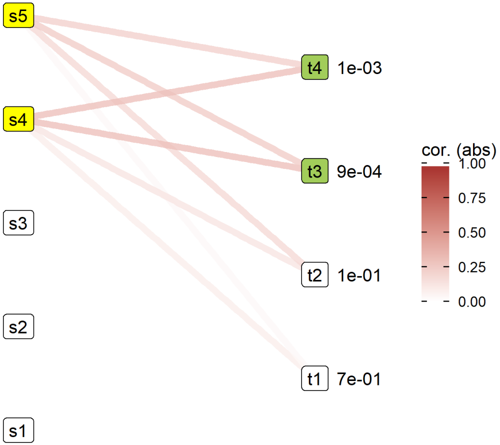


1. $B_0 = \{T_3\}$
2. $A_0 = \{S_4, S_5\}$
3. $B_1 = \{T_3, T_4\}$
4. $A_1 = \{S_3, S_4, S_5\}$
5. $B_2 = \{T_3, T_4\}$
6. $A_2 = \{S_3, S_4, S_5\}$

$(A_1, B_1) = (A_2, B_2)$

Stable bimodule found.

Bimodule Search Procedure: example run

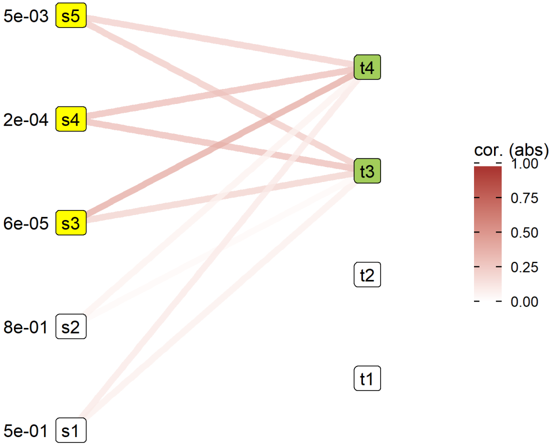


1. $B_0 = \{T_3\}$
2. $A_0 = \{S_4, S_5\}$
3. $B_1 = \{T_3, T_4\}$
4. $A_1 = \{S_3, S_4, S_5\}$
5. $B_2 = \{T_3, T_4\}$
6. $A_2 = \{S_3, S_4, S_5\}$

$(A_1, B_1) = (A_2, B_2)$

Stable bimodule found.

Bimodule Search Procedure: example run

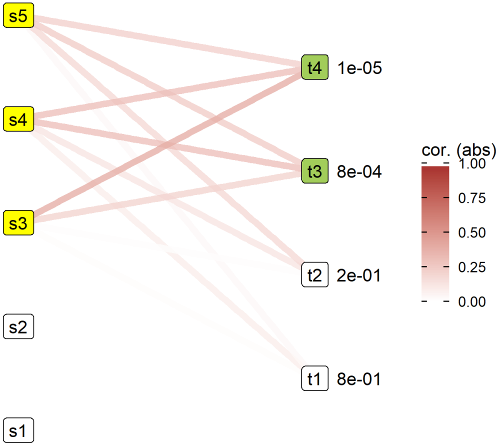


1. $B_0 = \{T_3\}$
2. $A_0 = \{S_4, S_5\}$
3. $B_1 = \{T_3, T_4\}$
4. $A_1 = \{S_3, S_4, S_5\}$
5. $B_2 = \{T_3, T_4\}$
6. $A_2 = \{S_3, S_4, S_5\}$

$$(A_1, B_1) = (A_2, B_2)$$

Stable bimodule found.

Bimodule Search Procedure: example run

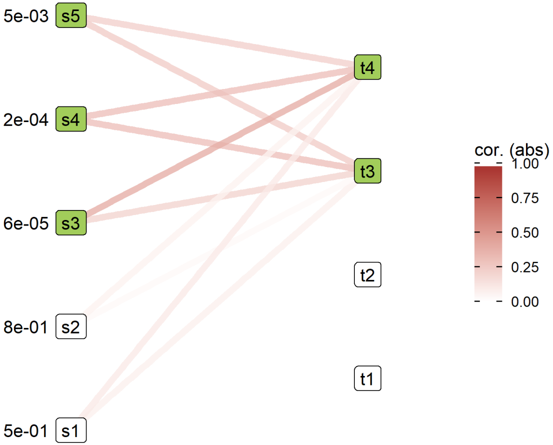


1. $B_0 = \{T_3\}$
2. $A_0 = \{S_4, S_5\}$
3. $B_1 = \{T_3, T_4\}$
4. $A_1 = \{S_3, S_4, S_5\}$
5. $B_2 = \{T_3, T_4\}$
6. $A_2 = \{S_3, S_4, S_5\}$

$$(A_1, B_1) = (A_2, B_2)$$

Stable bimodule found.

Bimodule Search Procedure: example run



1. $B_0 = \{T_3\}$
2. $A_0 = \{S_4, S_5\}$
3. $B_1 = \{T_3, T_4\}$
4. $A_1 = \{S_3, S_4, S_5\}$
5. $B_2 = \{T_3, T_4\}$
6. $A_2 = \{S_3, S_4, S_5\}$

$$(A_1, B_1) = (A_2, B_2)$$

Stable bimodule found.

Bimodule Search Procedure (pseudo code)

```
Initialize $A_0 = \{s\} \subseteq S$.

For $k = 0, \ldots, k_{max}$:

    • Calculate $p(A_k, t)$ for each $t \in T$
    • Let $B_k = \{t \in T \mid p(A_k, t) \leq \gamma_\alpha\}$ be indices rejected by BY($\alpha$).
    • Calculate $p(s, B_k)$ for each $s \in S$
    • Let $A_{k+1} = \{s \in S \mid p(s, B_k) \leq \gamma_\alpha\}$ be indices rejected by BY($\alpha$).

Output : $(A_{k_{max}}, B_{k_{max}})$ if it is non-empty ($A_{k_{max}} \neq \emptyset$) and a fixed point ($A_{k_{max}} = A_{k_{max}+1}$).
```

Bimodule Search Procedure (Software)

R package https://github.com/miheerdew/cbce.

Features

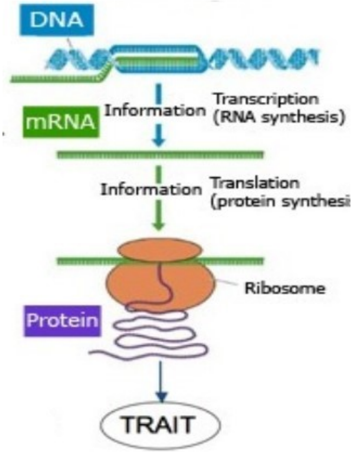
- Fast and parallel implementation (Analytical approximation to the permutation distribution + RCpp + Microsoft ROpen)
- Permutation based procedure to select primary parameter $\alpha \in (0, 1)$.
- Allows overlapping bimodules (and filtering for duplicates).
- Code tested and documented

Outline

1. Bimodules: groups of significant cross-correlated features in bi-view data
   - Bi-view data and Bimodules
   - Stable bimodules and the Bimodule Search Procedure (BSP)
2. Application to genomics
   - Introduction to eQTL analysis
   - Using BSP for groupwise eQTL analysis
3. Theoretical analysis of BSP
   - Asymptotics of BSP
   - Null correlation networks

Concepts from genomics (simplified version)

genome.gov/genetics-glossary



Gene expression Process used by cells to assemble protein molecules based on a gene.

Gene A region of the genome that encodes for a protein; ∼20K genes identified in humans.

Single nucleotide polymorphism (SNP) A location on the genome that has a nucleotide variation within the population.

Genetic basis of gene expression Millions of SNPs are identified in humans. Which ones influence traits?

Expression quantitative trait loci (eQTL)

A genomic region (e.g. SNP) that influences the expression level of one or more genes.

Data from GTEx project (v8)

from gtexportal.org

NIH funded GTEx project

A large collection of multi-tissue eQTL data from donors.

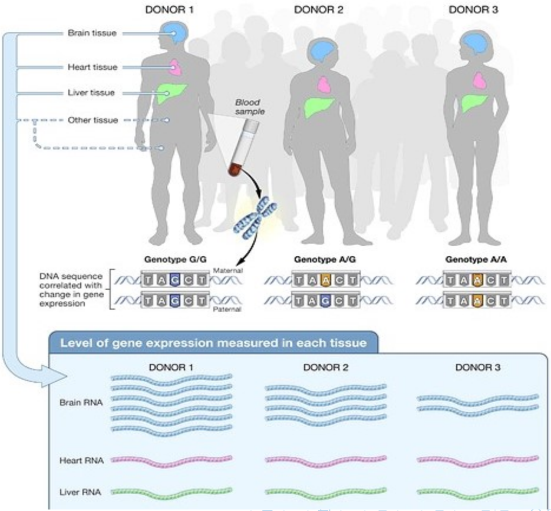
Individuals densely genotyped

Measurements for 4.9 million SNPs encoded as $\{0, 1, 2\}$ (MAF).

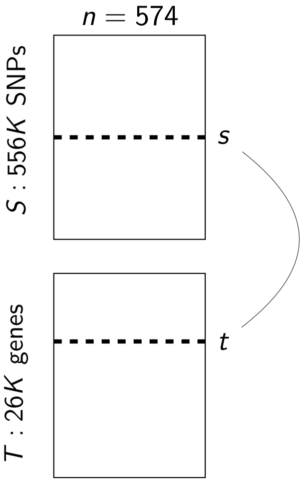
Expression measured in multiple tissues

RNA sequencing used to measure expression of genes.

Normalization, quality control, and covariate correction performed.



eQTL analysis for Thyroid data



Thyroid expression data from $n = 574$ donors for

$T = \{26K \text{ genes}\}$

$S = \{556K \text{ representative SNPs}\}$ selected using LD-pruning

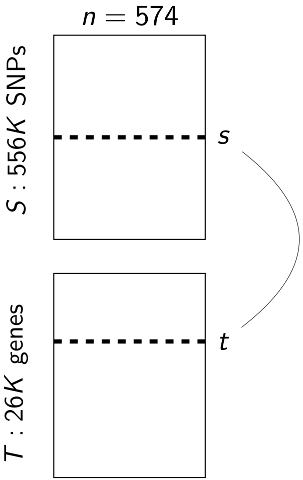
standard eQTL analysis

Find pairs $s \in S$ and $t \in T$ for which $r^2(s,t)$ is significant after *accounting for multiple-testing (MT)*.

Analysis-type	Pairs considered	MT correction
cis-analysis	local only	substantial
trans-analysis	all pairs ($\sim 10^{10}$)	huge

Distal eQTLs are harder to detect because of smaller effect size and huge MT burden.

eQTL analysis for Thyroid data



Thyroid expression data from $n = 574$ donors for

$T = \{26K \text{ genes}\}$

$S = \{556K \text{ representative SNPs}\}$ selected using LD-pruning

standard eQTL analysis

Find pairs $s \in S$ and $t \in T$ for which $r^2(s, t)$ is significant after *accounting for multiple-testing (MT)*.

Analysis-type	Pairs considered	MT correction
cis-analysis	local only	substantial
trans-analysis	all pairs ($\sim 10^{10}$)	huge

Distal eQTLs are harder to detect because of smaller effect size and huge MT burden.

Groupwise eQTL: find SNP-gene bimodules

Instead of pairs, search for SNP-gene bimodules, i.e.
bimodule (A, B), where $A \subseteq$ SNPs and $B \subseteq$ Genes are correlated.

Motivation:

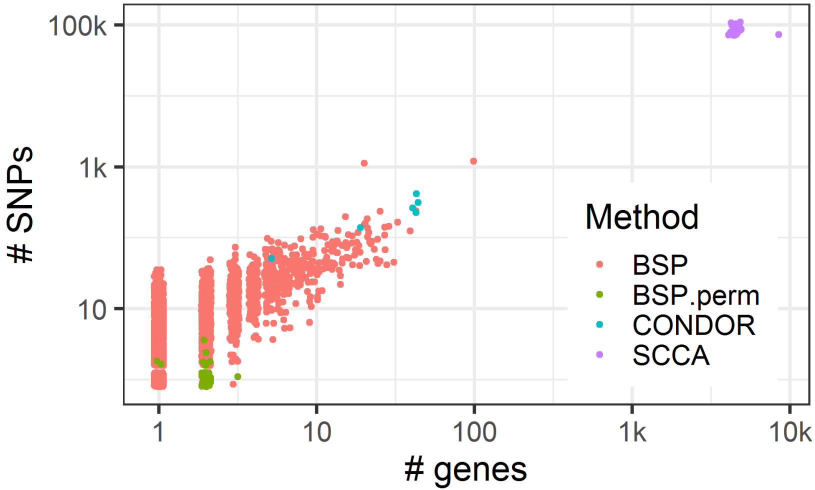
- Platig et al. (2016) find SNP-gene bimodules by community detection on a bipartite graph obtained from standard eQTL analysis.
- They show that bimodules may represent a group of SNPs that disrupt the functioning of gene regulatory networks and contribute to diseases
- Find bimodules using BSP by *aggregating effects* and *accounting for intra-correlations*.

Running BSP on Thyroid data

Highlights

- $\alpha = 0.03$ chosen using permutation.
- Most iterations lead to a (often empty) fixed point. search details
- Effective number of bimodules: 3305.
- Runtime 4.7 hrs (20-core/2.4 GHz).
- Bimod size range: 1-1000 SNPs & 1-100 genes. plot
- Median sizes: 7 SNPs and 1 gene.

Sizes of bimodules discovered by various methods



Obtaining networks from bimodules

A SNP-gene bimodule (A, B) has significant aggregate correlation between A and B.

But which edges $(s, t) \in A \times B$ are significant?

Threshold at $\tau \in (0, 1)$**:** $E_\tau(A, B) = \{(s, t) \mid r^2(s, t) \geq \tau^2,\ s \in A, t \in B\}$

How to choose τ?

Conservative estimate of strongest edges

Since a bimodule must be connected, choose the largest $\tau^* \in (0, 1)$ so that $(A \sqcup B, E_{\tau^*}(A, B))$ is a connected graph.

$E_{\tau^*}(A, B)$ are called *essential-edges* of the bimodule.

Thyroid network statistics

Obtaining networks from bimodules

A SNP-gene bimodule (A, B) has significant aggregate correlation between A and B.

But which edges $(s, t) \in A \times B$ are significant?

Threshold at $\tau \in (0, 1)$: $E_\tau(A, B) = \{(s, t) \mid r^2(s, t) \geq \tau^2,\ s \in A, t \in B\}$

How to choose τ?

Conservative estimate of strongest edges

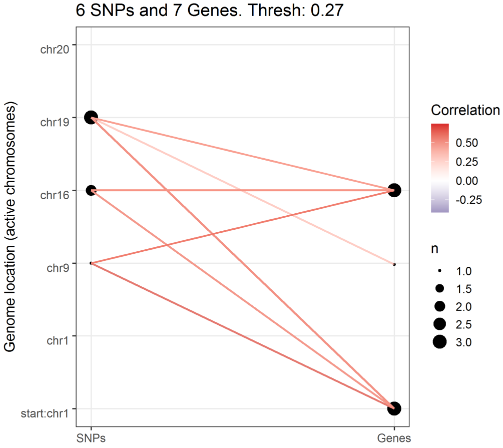
Since a bimodule must be connected, choose the largest $\tau^* \in (0, 1)$ so that $(A \sqcup B, E_{\tau^*}(A, B))$ is a connected graph.

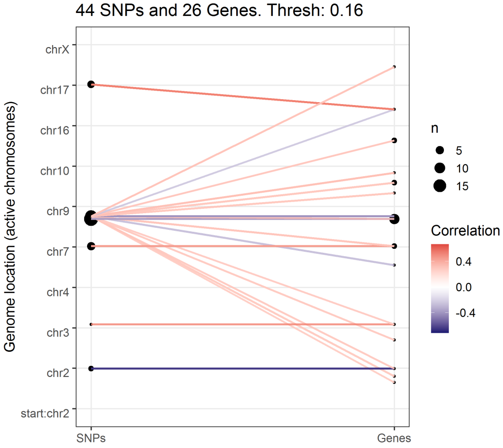
$E_{\tau^*}(A, B)$ are called *essential-edges* of the bimodule.

Thyroid network statistics

Essential-edge networks in GTEx thyroid data

examples from two bimodules





Comparing bimodules to standard eQTL analysis

Standard eQTL analysis performed using MatrixEQTL ($\alpha = 0.05$).

Bimodules find most standard eQTLs

84% of eQTLs from trans-analysis, and 51% of eQTLs from cis-analysis. But note

- bimodules find SNP-gene networks not just pairs, and
- cis-analysis improves power by restricting to local pairs.

New potential eQTLs from bimodules

Essential-edges from bimodules reveal 300 local and 8.8k distal SNP-gene pairs that

- are not detected by standard analysis,
- but show significance at the network level.

Comparing bimodules to standard eQTL analysis

Standard eQTL analysis performed using MatrixEQTL ($\alpha = 0.05$).

Bimodules find most standard eQTLs

84% of eQTLs from trans-analysis, and 51% of eQTLs from cis-analysis. But note

- bimodules find SNP-gene networks not just pairs, and
- cis-analysis improves power by restricting to local pairs.

New potential eQTLs from bimodules

Essential-edges from bimodules reveal 300 local and 8.8k distal SNP-gene pairs that

- are not detected by standard analysis,
- but show significance at the network level.

Analysis of genomic locations of bimodules

Recall BSP does not use genomic locations of SNPs and Genes. Nevertheless

Proximity of SNPs and genes within the bimodule.

- Almost all (99.3%) bimodules have at least one local SNP-gene pair.
- In addition, almost half of the larger bimodules found gene and SNPs that had distal effects.

Chromosomal locations of SNPs and genes from bimodules.

- Bimodule SNPs and Genes distributed across all 23 chromosomes.
- Most small bimodules (95%) were restricted to single chromosome.
- Nearly half of the larger bimodules spanned 2-11 chromosomes each.

Enrichment of known gene sets in bimodules

The GO database (http://geneontology.org/) contains collection of gene sets known to be associated with biological functions.

- Consider our 145 bimodules that have 7 or more genes.
- We used Fisher's test to assess overlap of gene sets from these bimodules with GO sets.
- Gene sets from 18 bimodules had significant overlap with gene sets associated to known biological processes.
- But the associated function did not seem thyroid relevant.

Repeating above process with randomly chosen gene sets of the similar sizes did not detect significant association.

Outline

1. Bimodules: groups of significant cross-correlated features in bi-view data
   - Bi-view data and Bimodules
   - Stable bimodules and the Bimodule Search Procedure (BSP)
2. Application to genomics
   - Introduction to eQTL analysis
   - Using BSP for groupwise eQTL analysis
3. Theoretical analysis of BSP
   - Asymptotics of BSP
   - Null correlation networks

Recipe for BSP asymptotics & population stable bimodules

- Suppose columns of the data matrix $D_n = \begin{bmatrix} \mathbb{X} \\ \mathbb{Y} \end{bmatrix}$ consist of i.i.d. realizations of a random vector $(X, Y)^t$ distributed as $\mathcal{N}_{p+q}(0, \Sigma)$.
- This defines a (random) BSP update function

$$\Gamma_n : 2^{S \cup T} \to 2^{S \cup T}$$

  whose fixed points are stable bimodules.
- How to establish of BSP asymptotics as $n \to \infty$?

Recipe:

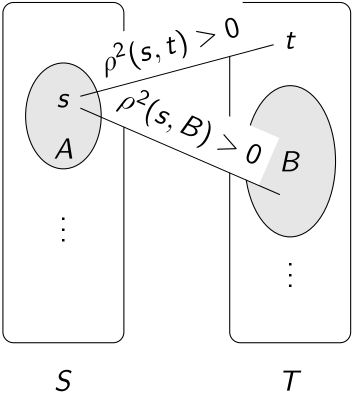
1. Identify $\Gamma : 2^{S \cup T} \to 2^{S \cup T}$ so that $\Gamma_n \xrightarrow{P} \Gamma$ pointwise.
2. Identify fixed points of Γ, and show they are reached by iterating Γ for k steps.

Lemma (BSP Asymptotics)

Assuming 1 & 2 above, with high probability as $n \to \infty$, the BSP on D_n will

- *find a stable bimodule within k iterations,*
- *and all the stable bimodules will be fixed points of Γ (population stable bimodules).*

Population picture in the large sample regime $n \gg \min(p, q)^2$



Population cross-correlation network with edge (s, t) if $\rho(s, t) \neq 0$.

In this regime:

- $\Gamma : 2^{S \cup T} \to 2^{S \cup T}$ is the neighborhood relation in the population cross-correlation network (PCCN).
- The (minimal & non-empty) fixed points of Γ are exactly the connected components of the PCCN.

Theorem (Dewaskar and Nobel, 2022)

When $n \gg \max(p, q)^2$, with high probability as $n \to \infty$, the BSP iterations starting from singleton set $\{s\} \subseteq S$ will reach a stable bimodule, which is a (non-trivial) connected component of the PCCN.

Null correlation network

Consider i.i.d. observations $X_1, \ldots, X_n \in \mathbb{R}^p$ of $\mathcal{N}_p(\mu, \Sigma_p)$.

Denote for $i, j \in \{1, \ldots, p\}$

$$S_n(i,j) : \text{sample covariance, and}$$
$$R_n(i,j) : \text{sample correlation.}$$

High-dimensional covariances

$S_n \to \Sigma_p$ as $n \to \infty$ for fixed p.

But global consistency may fail

$$\lambda(S_n) \not\to \lambda(\Sigma_p)$$

when $p \geq n$ (Jonstone, 2001).

Consider $\Sigma_p = I_p$ and sample correlation network

$$\mathcal{G}_{n,p} \doteq (V_p = \{1, \ldots, p\}, W_{n,p} = R_n).$$

Problem: Study asymptotic properties of $\mathcal{G}_{n,p}$.

Applications to *Correlation Network Mining*. E.g. methods that detect (in networks derived from Σ_p)

- Edges [Cai, 2017]
- Hubs [Hero and Rajaratnam, 2011]
- Cliques [Devroye, György, Lugosi, Udina 2011]
- Communities [Arias-Castro, Bubeck, Lugosi].

Null correlation network

Consider i.i.d. observations $X_1, \ldots, X_n \in \mathbb{R}^p$ of $\mathcal{N}_p(\mu, \Sigma_p)$.

Denote for $i, j \in \{1, \ldots, p\}$

$$\begin{aligned} S_n(i,j) &: \text{sample covariance, and} \\ R_n(i,j) &: \text{sample correlation.} \end{aligned}$$

High-dimensional covariances

$S_n \to \Sigma_p$ as $n \to \infty$ for fixed p.

But global consistency may fail

$$\lambda(S_n) \not\to \lambda(\Sigma_p)$$

when $p \geq n$ (Jonstone, 2001).

Consider $\Sigma_p = I_p$ and sample correlation network

$$\mathcal{G}_{n,p} \doteq (V_p = \{1, \ldots, p\}, W_{n,p} = R_n).$$

Problem: Study asymptotic properties of $\mathcal{G}_{n,p}$.

Applications to *Correlation Network Mining*. E.g. methods that detect (in networks derived from Σ_p)

- Edges [Cai, 2017]
- Hubs [Hero and Rajaratnam, 2011]
- Cliques [Devroye, György, Lugosi, Udina 2011]
- Communities [Arias-Castro, Bubeck, Lugosi].

Some properties of the random correlation network $\mathcal{G}_{n,p}$

Correlations and uniform points on the sphere. With $U_1, \ldots, U_p$ i.i.d. Unif$(\mathbb{S}^{n-2})$,

$$(R_n(i,j) : i,j \in [p]) \overset{d}{=} (\langle U_i, U_j \rangle : i,j \in [p]) \,.$$

Related work

1. $(n^{-1} \log p \to \beta)$ Max & min angle between points $\{U_i\}_{i=1}^{p}$ [Cai, Fan, Jiang, 2013]
2. $(n^{-1} \log p \to 0)$ Dense geometric graph formed by the points $\{U_i\}_{i=1}^{p}$ behaves like an ER random graph (e.g. [Basak, Bhamidi, Chakraborty, and Nobel, 2016]).

My interest

- Understand *stable modules* in $\mathcal{G}_{n,p}$. If $A \subseteq [p]$ is a stable module then $\{U_i\}_{i \in A}$ cluster around the mean $\bar{U}_A$.
- Study sizes of maximal clusters of $\{U_i\}_{i=1}^{p}$ to provide false discovery guarantees for the Module Search Procedure.

Some properties of the random correlation network $\mathcal{G}_{n,p}$

Correlations and uniform points on the sphere. With $U_1, \dots, U_p$ i.i.d. $\mathrm{Unif}(\mathbb{S}^{n-2})$,

$$(R_n(i,j) : i,j \in [p]) \overset{d}{=} (\langle U_i, U_j \rangle : i,j \in [p]) \,.$$

Related work

1. $(n^{-1} \log p \to \beta)$ Max & min angle between points $\{U_i\}_{i=1}^{p}$ [Cai, Fan, Jiang, 2013]
2. $(n^{-1} \log p \to 0)$ Dense geometric graph formed by the points $\{U_i\}_{i=1}^{p}$ behaves like an ER random graph (e.g. [Basak, Bhamidi, Chakraborty, and Nobel, 2016]).

My interest

- Understand *stable modules* in $\mathcal{G}_{n,p}$. If $A \subseteq [p]$ is a stable module then $\{U_i\}_{i \in A}$ cluster around the mean $\bar{U}_A$.
- Study sizes of maximal clusters of $\{U_i\}_{i=1}^{p}$ to provide false discovery guarantees for the Module Search Procedure.

Conclusion

We looked at

- **Bimodules:** statistically significant communities in bipartite correlation networks derived from multi-view data.
- **BSP**: iterative testing procedure to find *stable* bimodules.
- **Application to eQTL analysis**: using BSP to detect SNP-gene sub-networks and potentially new eQTLs.
- **Related theoretical problems**: Asymptotics as $n, (p \vee q) \to \infty$:
  1. BSP asymptotics via its update function.
  2. Asymptotics of the null correlation network via properties of uniformly distributed points on the sphere.

Thank you

Manuscript `https://arxiv.org/pdf/2009.05079.pdf`

Software `https://github.com/miheerdew/cbce`.

Collaborators

- John Palowitch (Google)
- Mark He (Columbia University)
- Andrew Nobel (UNC Statistics and Operations Research)
- Michael Love (UNC Biostatistics)

Supporting Grants

- NIH R01 HG009125-01
- NSF DMS-1613072

We don't perform permutation: moment based analytical approximation

Permutation p-values Permuting the sample labels of t using π, define the p-value

$$p(A,t) \doteq \mathbb{P}_{\pi}\left(r_{\pi}^{2}(A,t) \geq r^{2}(A,t)\right),$$

which conditions on correlations in A.

Monte-Carlo estimation too slow. For faster analytical approximation to the null distribution of $T = r_{\pi}^{2}(A,t)$:

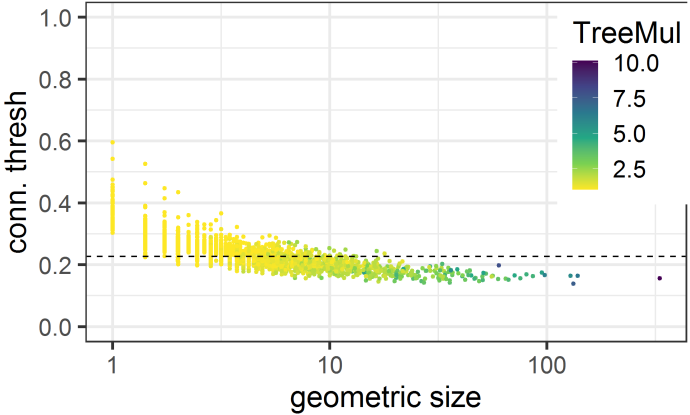
- Approximate the first three moments of T based on the eigenvalues of matrix X_A [Zhou, Gallins and Wright, 2019].
- Fit a shifted gamma distribution determined by the first three moments of T

BSP Thyroid search details

Search details

- 304K attempted searches.
- Majority (277K) give empty set in the first iteration.
- Few (20) did not terminate within 20 iterations.
- Remaining reached a fixed point in 20 iterations.
- 92.3% of these fixed points contained the seed singleton.

Network statistics from Thyroid



Smaller bimods are connected mainly by strong local associations (large τ^*). E_{τ^*} is tree-like. **Larger bimods** are connected by strong local + weak distal associations (small τ^*). E_{τ^*} has upto 10x more edges than a tree.