A new statistical graph model to systematically study associations between multivariate exposome data and multivariate metabolomics data

Qiong Wu

Co-authors: Drs. Shuo Chen, Charles Ma, Donald Milton

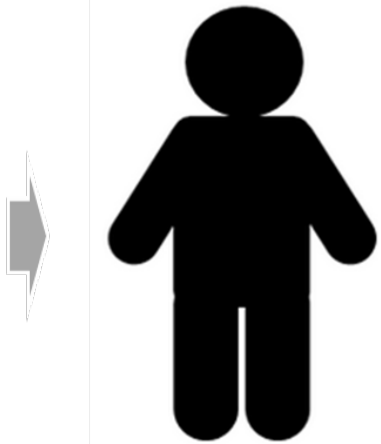
April, 2021

UNIVERSITY OF
MARYLAND

Outline

Introduction

- Association studies in multi-omics data
  - Discover the systematic association patterns between a set(s) of multivariate correlated exposure variables and a set(s) of multivariate metabolomics (or gene/protein expression data);
  - Study human responses to a mixture of environmental exposures: **Multivariate predictors and multivariate outcomes.**



Goal

- *Goal*: parsimonious multivariate-multivariate association pattern detection, to select a set(s) of exposures and a set(s) of accordingly affected outcomes.
- Challenges:
  - Univariate methods (a bag of pairwise associations)
    - false positive and negative associations can disrupt revealing the underlying systematic association patterns;
  - Dimension reduction methods (e.g., principal component analysis or canonical correlation analysis)
    - limited to identify specific exposome and metabolome variables in the correlated components;
  - Biclustering algorithms
    - miss the patterns by equally assigning variables to clusters.

Graph model setup

- We characterize the relationship as a **Bipartite Graph** $G = (U, V, E, \boldsymbol{W})$.
- Nodes U: all exposures; nodes V: all metabolites; and weighted edges $\boldsymbol{W}$ the marginal association measures ($|U| \times |V|$).
- We focus on the subgraph $G[U_0, V_0]$, $U_0 \subset U$ and $V_0 \subset V$ of significant exposures-metabolites associations concentrated.

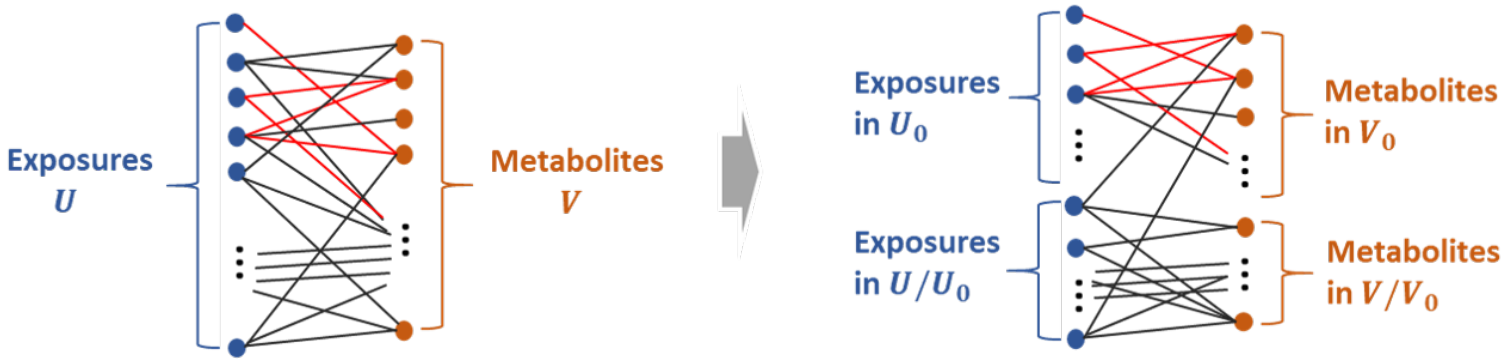


Figure 1: A demonstration of the bipartite graph with subgraph $G[U_0, V_0]$. The right subfigure indicates $G[U_0, V_0]$ in G with nodes reordered.

Association Patterns

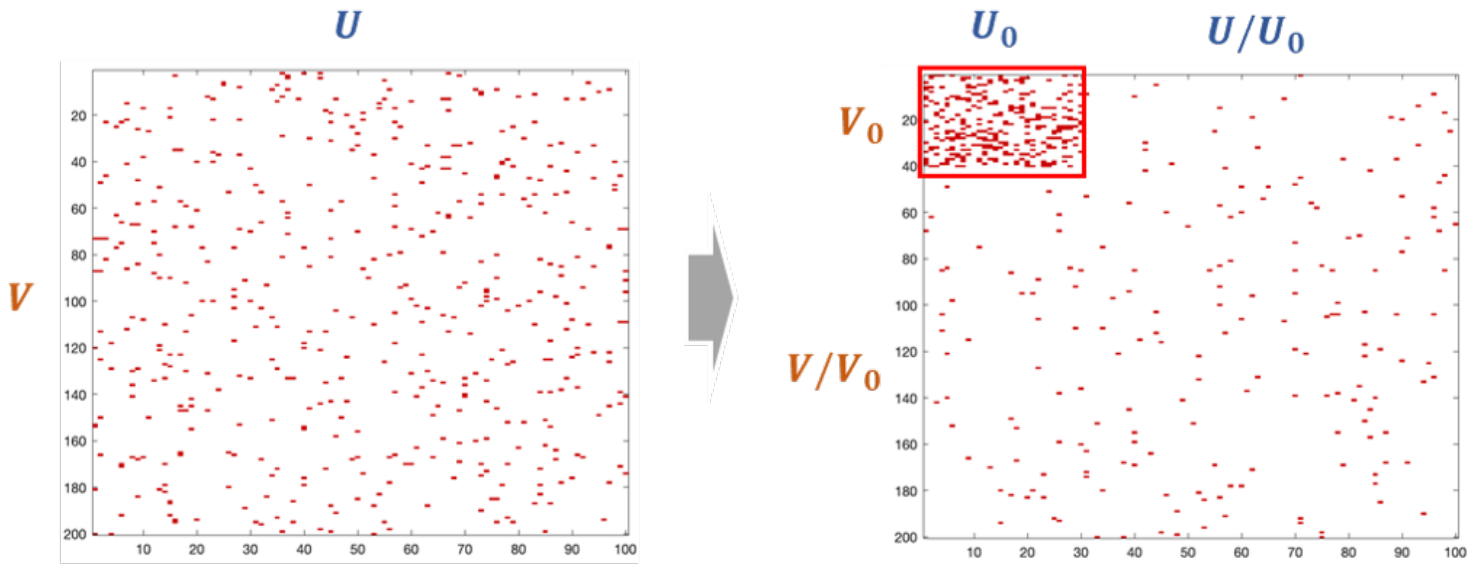


Figure 2: A demonstration of the bipartite graph with subgraph $G[U_0, V_0]$. The right subfigure indicates $G[U_0, V_0]$ in G with nodes reordered.

A false positive association edge is likely, but not the systematic pattern

- Based on graph combinatorics, the probability that a non-trivial set of exposures are highly correlated with a non-trivial set of metabolitics converges to **Zero** under the null.

Lemma (Under null hypothesis/random graph)

Suppose G *is observed from a random bipartite graph* $G(m, n, \pi)$. $G[S_\gamma, T_\gamma]$ *is a* γ*-quasi biclique with* $\gamma \in (\pi, 1)$. *Let* $m_0, n_0 = \Omega(\max\{m^\epsilon, n^\epsilon\})$ *for some* $0 < \epsilon < 1$. *Then for sufficiently large* m, n *with* $c(\pi, \gamma)m_0 \geq 8 \log n$ *and* $c(\pi, \gamma)n_0 \geq 8 \log m$, *we have*

$$\mathbb{P}\left(|S_\gamma| \geq m_0, |T_\gamma| \geq n_0\right) \leq 2mn \cdot \exp\left(-\frac{1}{4}c(\gamma, \pi)m_0 n_0\right),$$

where $c(a, b) = \left\{\frac{1}{(a-b)^2} + \frac{1}{3(a-b)}\right\}^{-1}$.

Search for the systematic pattern by a generalized density metric

- To detect a subgraph that includes a maximal number of association pairs (edges) with a minimal graph size, we propose a generalized density metric:

$$d_\lambda(S,T) = \frac{|\boldsymbol{W}[S,T]|}{(|S||T|)^\lambda}, \tag{1}$$

  with $\lambda \in (0,1)$.
- We output the subgraph as:

$$(\tilde{S}_\lambda, \tilde{T}_\lambda) = \arg\max_{S,T} d_\lambda(S,T)$$

Likelihood-based method for λ estimation

- The likelihood of the bipartite graph with dense subgraph has form:

$$L(\pi_1, \pi_0; S, T, \boldsymbol{A}) = \prod\nolimits_{i\in S, j\in T} \pi_1^{a_{ij}} (1-\pi_1)^{1-a_{ij}} \times \prod_{i\in U/S \text{ or } j\in V/T} \pi_0^{a_{ij}} (1-\pi_0)^{1-a_{ij}},$$

- Therefore,

$$\hat{\lambda} = \arg\max_{\lambda} L_\lambda(\pi_1, \pi_0; \tilde{S}_\lambda, \tilde{T}_\lambda, \boldsymbol{A}).$$

- For weighted adjacency matrix, we binarize as $A_{ij} = \{W(r)\}_{ij} = I(W_{ij} > r)$.
  Consider the threshold r has a support $\{r_1, ..., r_m\}$ and corresponding probability $\{g(r_1), ..., g(r_m)\}$, we integrate r out:

$$L_\lambda(\pi_1, \pi_0; \tilde{S}_\lambda, \tilde{T}_\lambda, \boldsymbol{W}) = \int L_\lambda\left(\pi_1, \pi_0; \tilde{S}_\lambda, \tilde{T}_\lambda, \boldsymbol{W}(r)\right) g(r) dr.$$

Greedy algorithm with given λ

Algorithm 1 Greedy algorithm with given λ

procedure ALGORITHM
  for $c \in \{c_1, c_2, ..., c_L\}$ **do**
    $S_1 \leftarrow U$, $T_1 \leftarrow V$
    for k=1 to $n + m - 1$ **do**
      let $i \in S_k$ with: $i = \arg\min_{i' \in S_k} \deg_X(i'; S_k, T_k)$;
      let $j \in T_k$ with: $j = \arg\min_{j' \in T_k} \deg_Y(j'; S_k, T_k)$;
      if $\sqrt{c}\deg_X(i; S_k, T_k) \leq \frac{1}{\sqrt{c}} \deg_Y(j; S_k, T_k)$ **then**
        $S_{k+1} \leftarrow S_k / \{i\}$ and $T_{k+1} \leftarrow T_k$;
      else
        $S_{k+1} \leftarrow S_k$ and $T_{k+1} \leftarrow T_k / \{j\}$;
      end if
    end for
    Output $G[S^c, T^c]$ that maximizes the density metric among $G[S_1, T_1], ..., G[S_{n+m-1}, T_{n+m_1}]$;
  end for
  Output $G[S^{c*}, T^{c*}]$ with largest density $G[S^{c_1}, T^{c_1}], ..., G[S^{c_L}, T^{c_L}]$;
end procedure

Data results

- 169 numeric exposure variables and 221 metabolites variables for 1192 subjects;
- We observe the association matrix as pairwise correlation coefficients, although other metrics can be used (e.g., -log(p) values of regression coefficients).

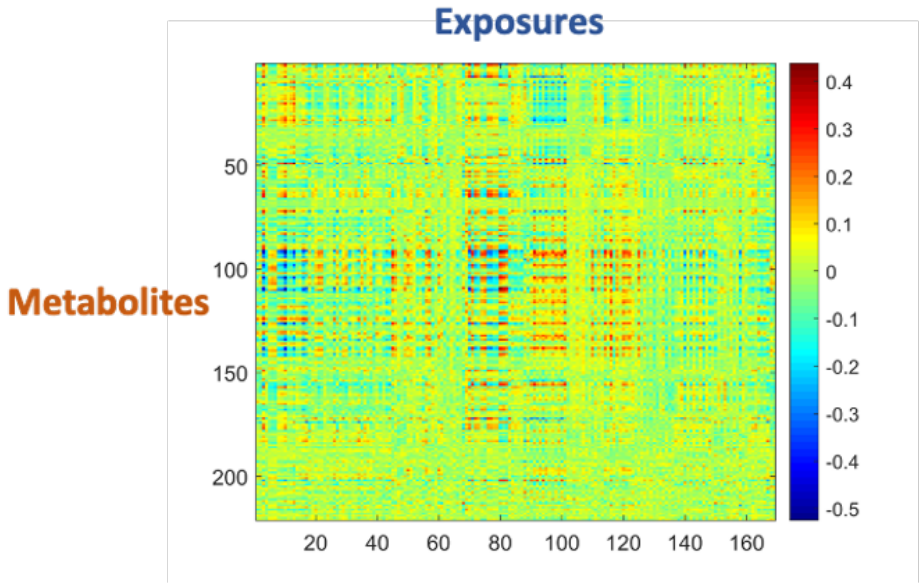


Figure 3: Association matrix between exposome and metabolites

Data results

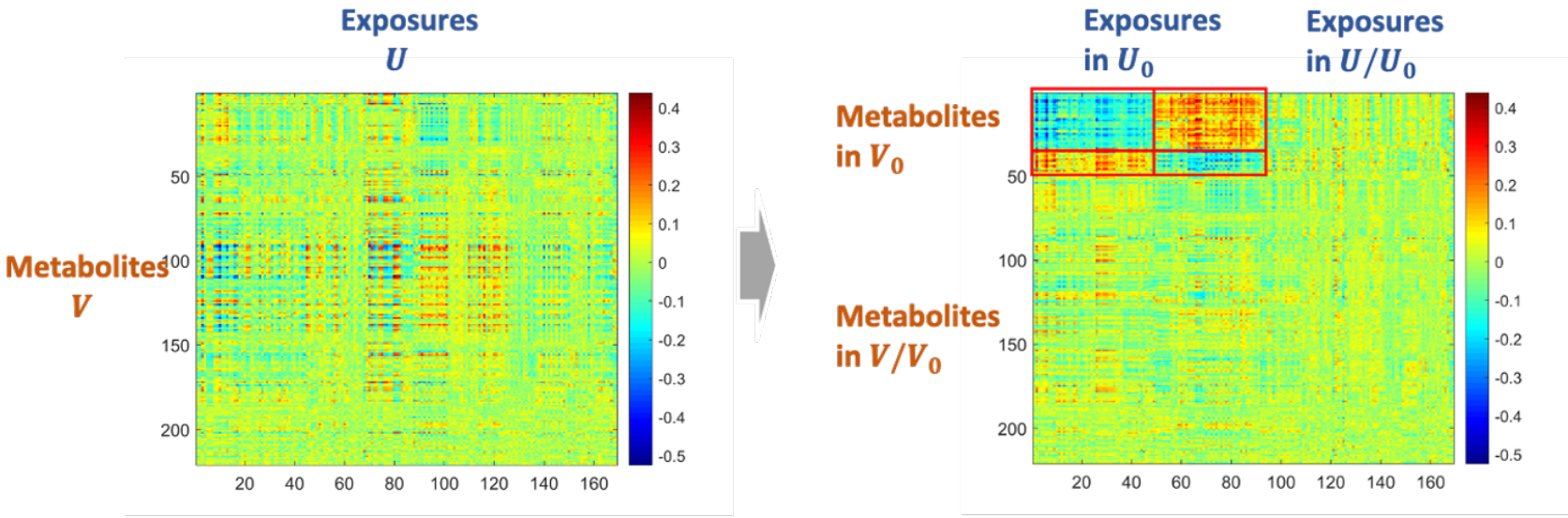


Figure 4: Detecting the systematic association patterns between multiple exposure variables and metabolomics

Data results

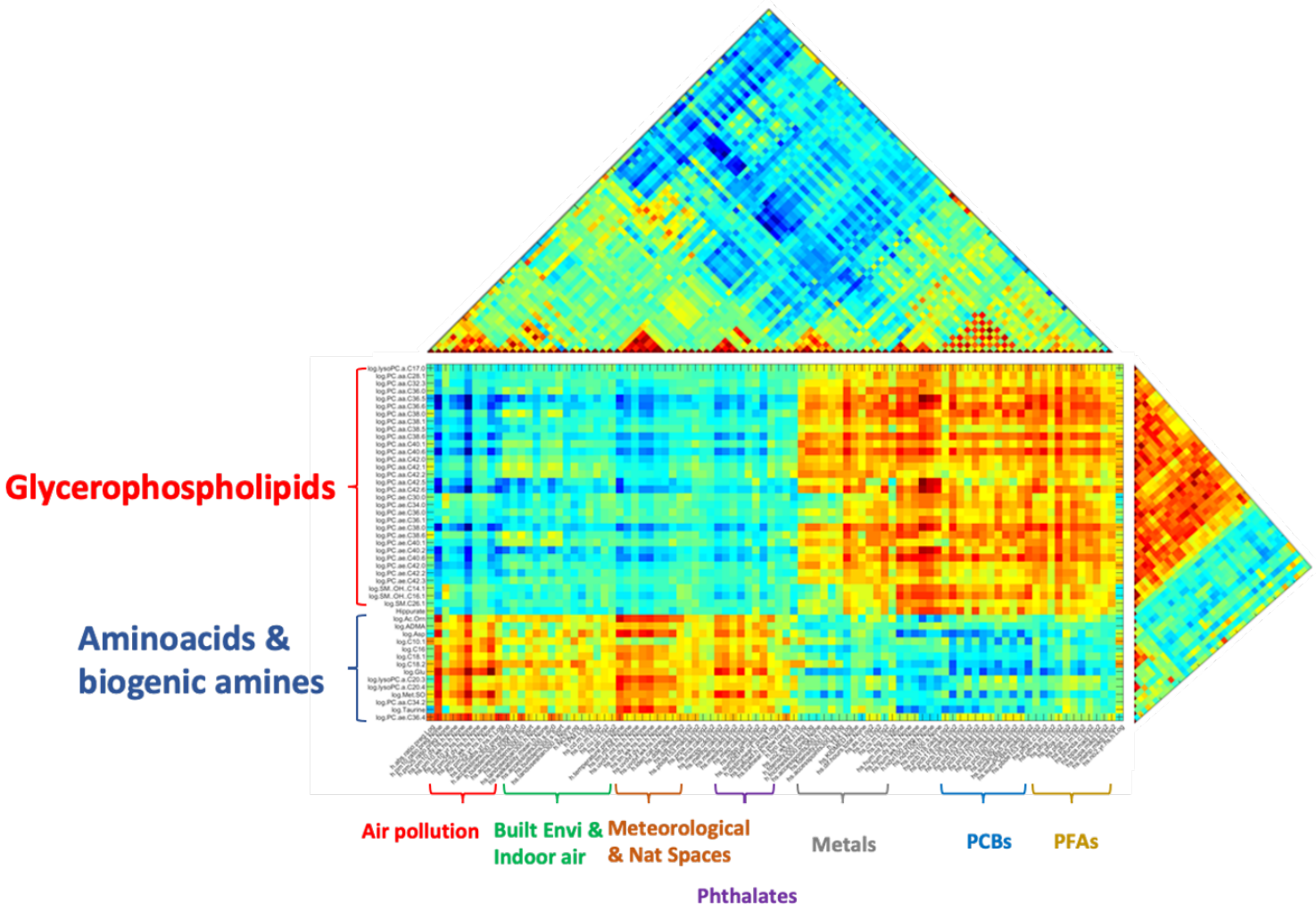


Figure 5: Zoomed association patterns between multiple exposure variables and metabolomics

Data results

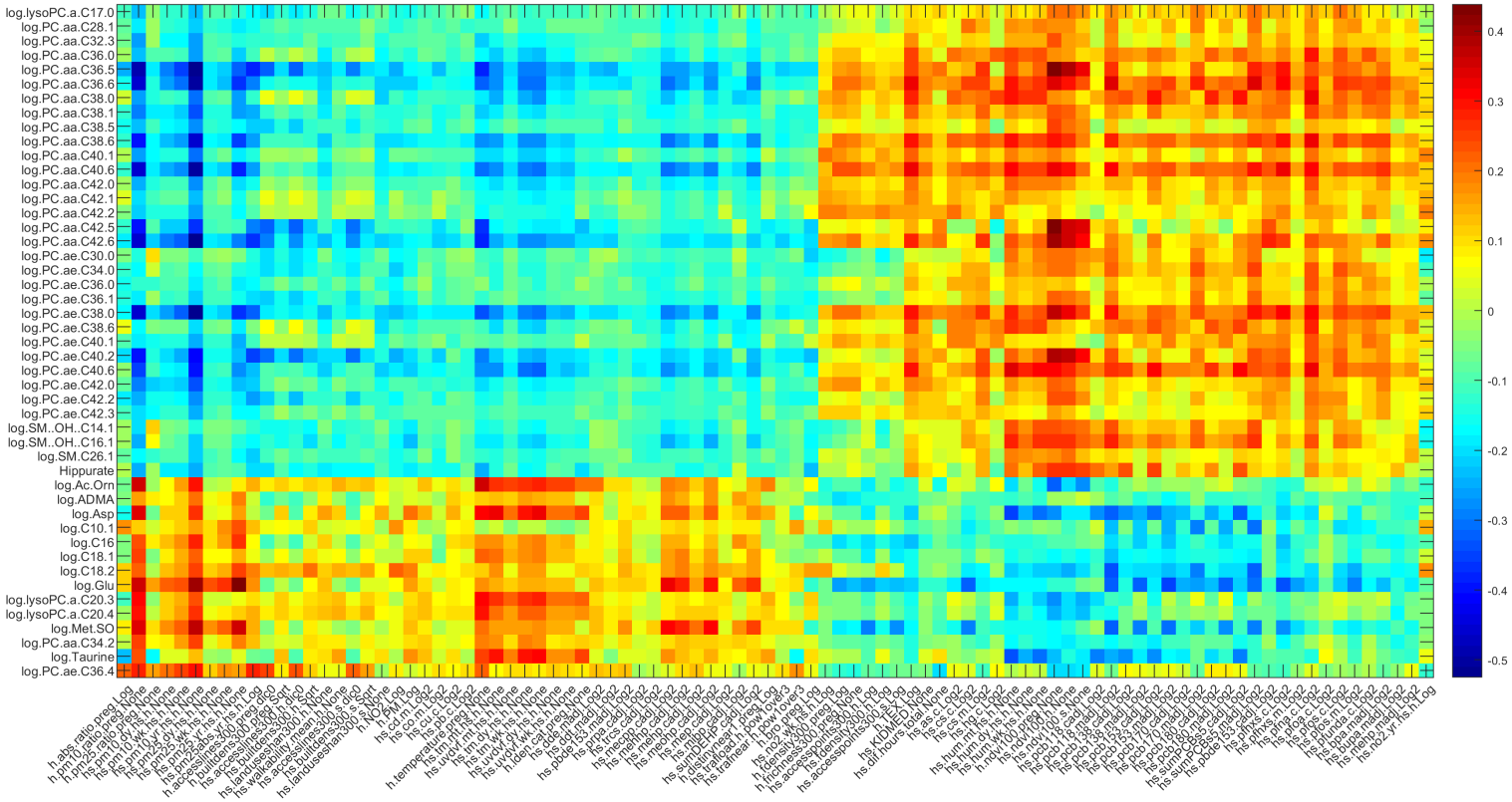


Figure 6: Zoomed association patterns between multiple exposure variables and metabolomics

Selected metabolites can better explained by our selected exposomes than all exposomes

- Full models:

$$\boldsymbol{y}_v \sim \boldsymbol{X}_v \boldsymbol{\beta}_v = \sum_{u \in U} x_{uv} \beta_{uv}, \quad \forall v \in \tilde{T}_{\hat{\lambda}}$$

where U represents the set of all exposures, $\tilde{T}_{\hat{\lambda}}$ is the set of metabolites within the detected subgraph.

- Reduced models:

$$\boldsymbol{y}_v \sim \boldsymbol{X}_v^{\text{sub}} \boldsymbol{\beta}_v^{\text{sub}} = \sum_{u \in \tilde{S}_{\hat{\lambda}}} x_{uv} \beta_{uv}, \quad \forall v \in \tilde{T}_{\hat{\lambda}}$$

where $\tilde{S}_{\hat{\lambda}}$ is the set of exposures selected in the subgraph.

	Number of predictors	R^2
Full models	169	0.317 (0.080)
Reduced models	92	0.261 (0.085)

Cross-validation performance

- Evaluate the performance using cross-validation (5-folds):
  - 80% data are used to fit the model (estimates $\hat{\boldsymbol{\beta}}_{\text{train}}$);
  - 20% are testing data to predict $\hat{\boldsymbol{y}}_{\text{test}} = \boldsymbol{X}_{\text{test}}\hat{\boldsymbol{\beta}}_{\text{train}}$
  - Predictive R^2, RMSE and MAE are calculated based on $\hat{\boldsymbol{y}}_{\text{test}}$ and $\boldsymbol{y}_{\text{test}}$

	Predictive R^2	RMSE [1]	MAE [2]
Full models	0.116 (0.028)	0.373 (0.015)	0.295 (0.012)
Reduced models	0.146 (0.033)	0.356 (0.013)	0.281 (0.010)

[1] RMSE: Root Mean Squared Error

[2] MAE: Mean Absolute Error

Summary

- Identify systematic associations between multivariate exposome and multivariate metabolome variables, which can be generalized to other multi-omics data;
- Parsimonious multivariate-multivariate association pattern extraction via detecting subgraphs in a bipartite graph with concentrated association pairs via a computationally efficient algorithm;
- Distinguish systematic negative and positive association blocks;
- Exposures within the detected subgraph can explain the population variance of selected metabolites comparing to the whole set of exposures.

Github Files

- Matlab functions:
  - greedy_bipar.m (maximize the density metric)
  - greedy_lik_fun.m (select the tunning parameter via likelihood)
  - NICE.m (refine the association pattern)
- Data analysis:
  - prepare_data.R
  - code.m
  - Output_figures.m
  - compare_models.R
- Data results:
  - meta_merge.mat
  - expos_merge.mat
  - expos_meta_res.mat
  - Output_figures.html

Thank you for your attention!