Multi-Group Covariance Estimation with applications in ‘Omics

Alexander Franks

Department of Statistics and Applied Probability

UC SANTA BARBARA

Introduction

- A key challenge in quantitative biology is identifying the relevant and irrelevant sources of variation
- Today: novel methodology for inferring covariance matrices in multiple subpopulations
- Covariance matrices can be useful for
  - Hypothesis-free and hypothesis-driven analyses
  - Variation across dimensions (e.g. subject, experimental condition)
  - Measurement error
- Case study on the metabolomics of neurodegenerative disease

Neurodegenerative Disease

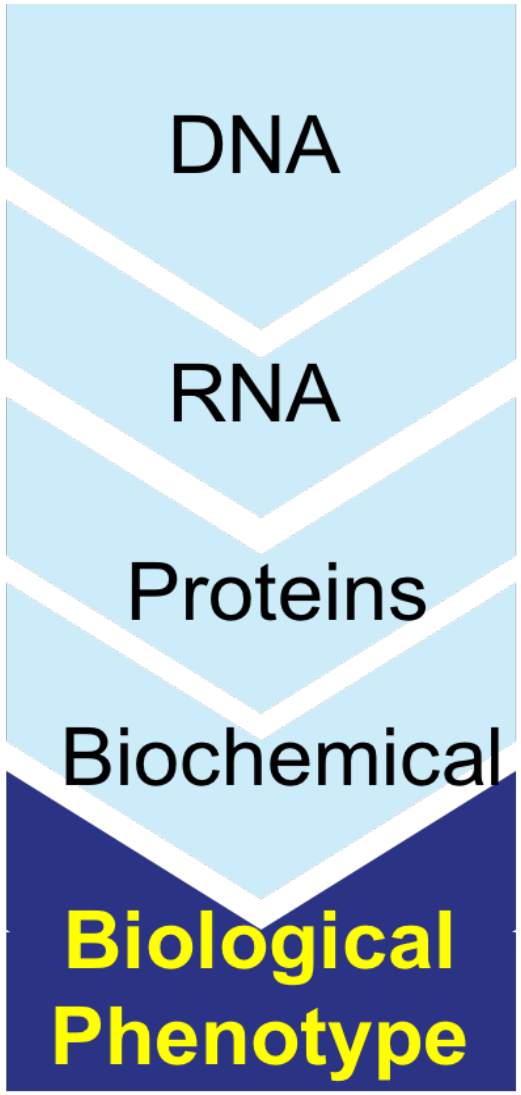
- Number of adult cases is forecast to reach 100 million worldwide in the next 35 years

- The majority of cases lack simple Mendelian genetic causes.
  - How do age, environment, and polygenic variation contribute to risk?

- Recent work suggests that the metabolome can provide a powerful tool to help us identify the mechanisms that underlie neuropathology.

Metabolomics

- Metabolites are the small molecules involved in metabolism
- Include amino acids, vitamins, sugars, drugs, etc.
- The metabolome is the complete set of metabolites in a sample



Genomics

Transcriptomics

Proteomics

Metabolomics

Metabolomics of Neurodegenerative Disease

- Links to mitochondrial dysfunction caused by the deleterious effects from oxidative stress and chronic inflammation
- AD and PD are comorbid with abnormal glucose metabolism and insulin resistance
- Initial studies suggest the possibility for predictive biomarkers
- Apolipoprotein E (**ApoE**) is a class of proteins involved in the metabolism of fats and is the largest known genetic risk factor for AD

Questions of Interests

Alzheimer's

- Can we identify biomarkers for Alzheimer's?
- What can we learn about AD mechanisms? (AD vs CO)
- Does ApoE status correlated with changes in the metabolome?

Parkinson's

- Can we identify biomarkers for Parkinson's?
- What can we learn about PD mechanisms? (PD vs CO)

Aging

- How does the metabolome change as we age (controls only)

Data

- Cerebrospinal fluid samples (CSF) from 198 individuals.
- 57 Alzheimer’s disease (AD), 56 Parkinson’s disease (PD), 85 controls
- For controls, have subjects from all ages
- Age, Sex, ApoE status

Data

- Mass Spectrometry-Based Metabolomics
  - Northwest Metabolomics Research Center (NW-MRC)
  - Targeted, approximately 100 features (ids known)
  - Untargeted, approximately 8000 features (ids unknown)

- Lipidomics
  - 1000 lipids

- Large p, small n problem
  - Only 200 observations of high-dimensional data

Model Building

- X = (disease status, age, sex…)
  - Relatively few features
- Y = (Fructose, DOPA, Creatinine, …)
  - Thousands of features

- Predict X given Y?
  - Given a metabolite measurements, does the subject have Alzheimer’s?
- Predict Y given X?
  - Given disease status, what can we say about the metabolome?

Predict disease status given metabolites

- Common framing in most machine learning problems
    - Use many features to classify (typically) into small number of categories
- If classifying disease status is the primary objective this is reasonable
    - Don’t need to model the complex interactions in the metabolome

Model the metabolome given phenotype

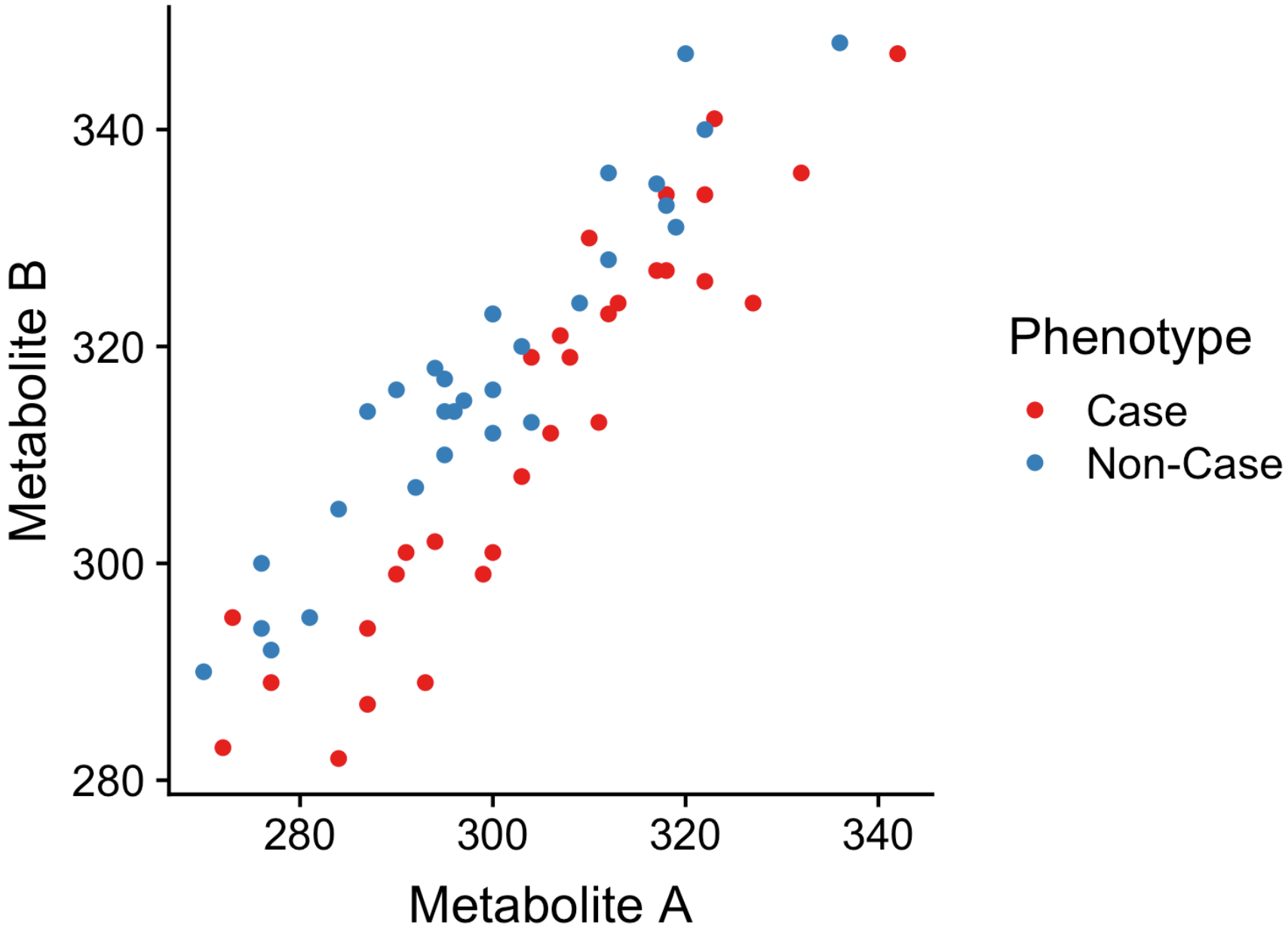
- Interested in mechanisms
  - How and why is the metabolome different in ND subjects
- More plausible causal direction?
  - Consistent with a the notion that a “disease causes symptoms”
- Measurement error and missing data in metabolite abundances

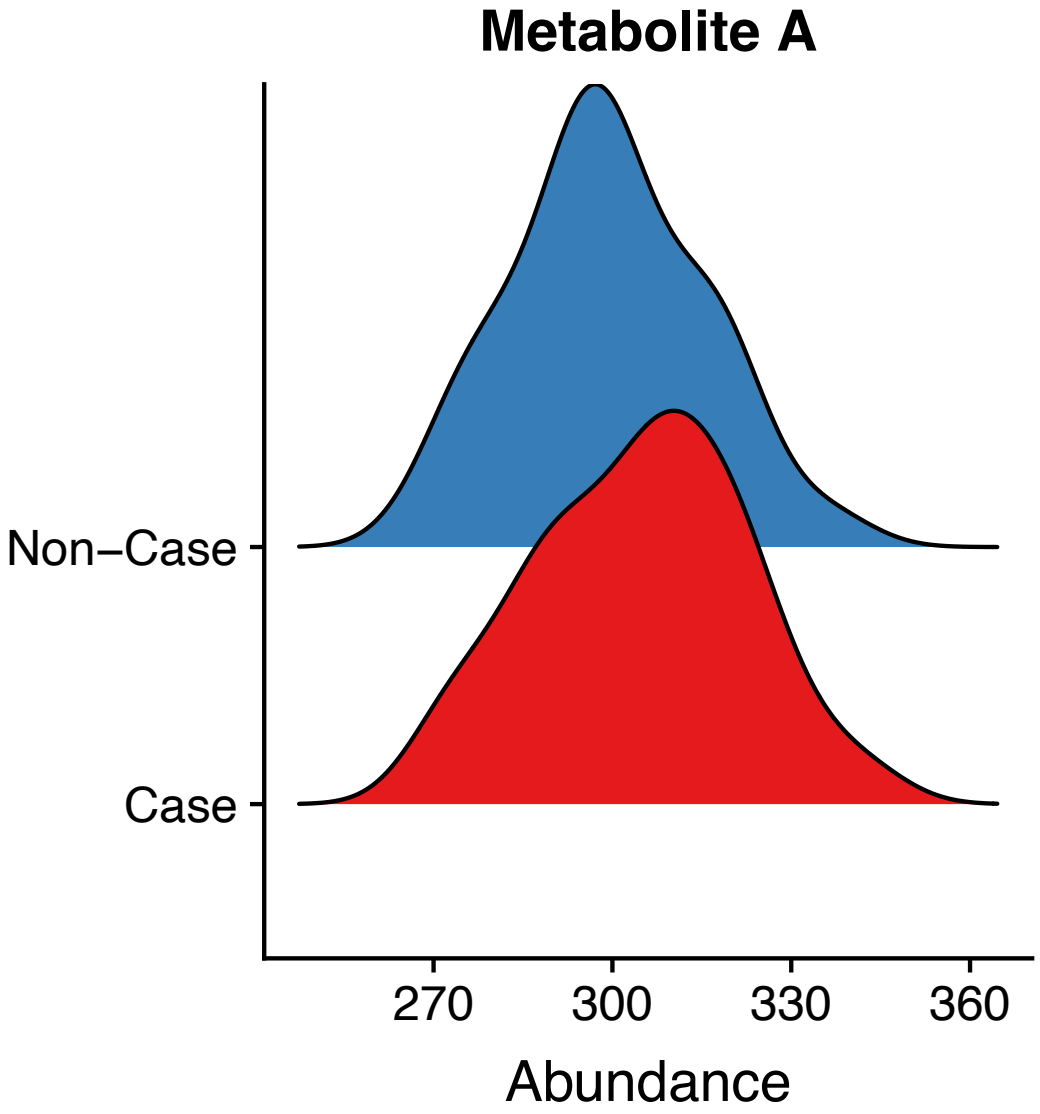
Statistical Challenges

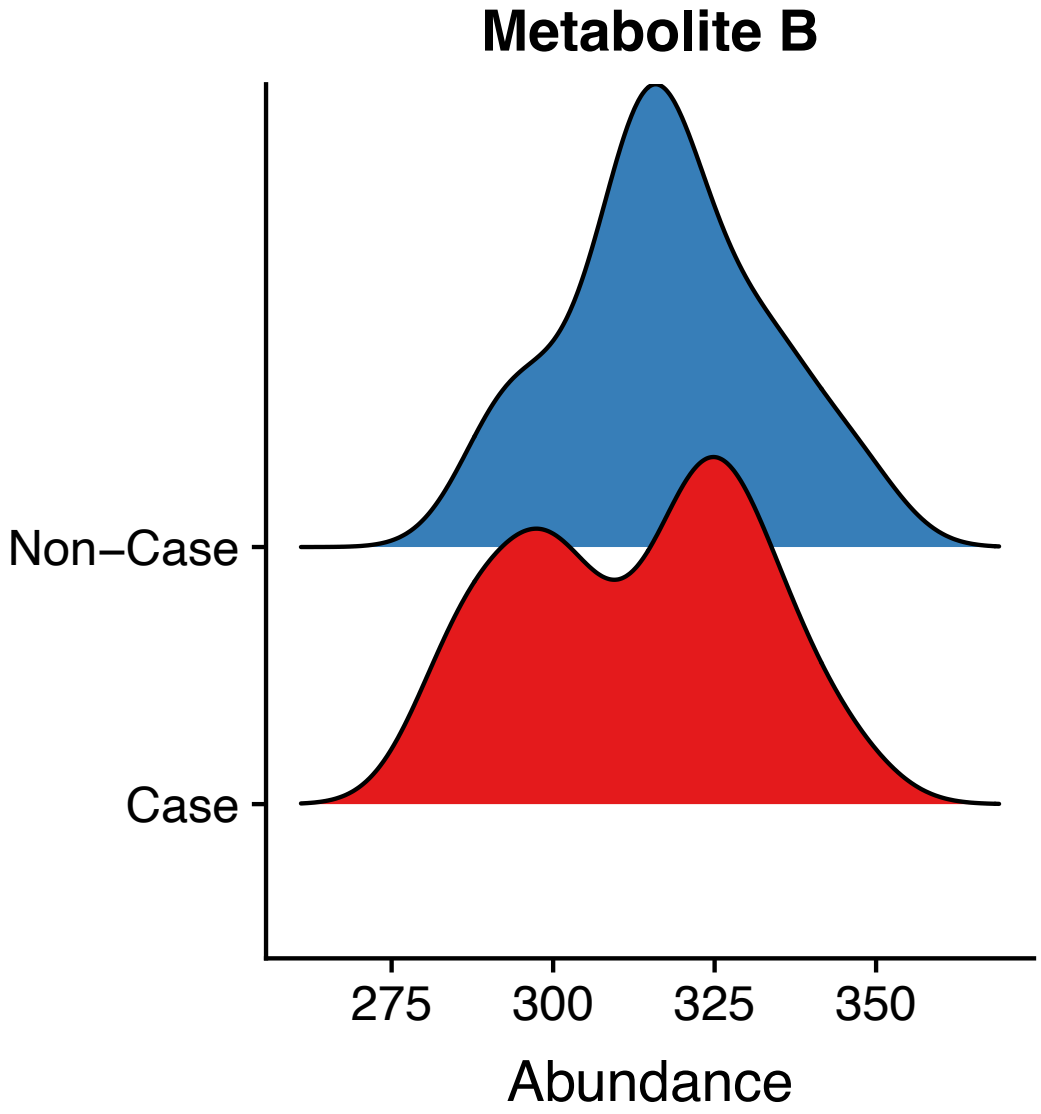
- Identify mean level differences (useful for identifying biomarkers)
- This talk: focus on inferring covariance matrices across groups
  - Relevant for learning about *mechanisms*
- Approx. 200 samples to learn about thousands of features!
  - Number of correlations on the order of 8000 squared (untargeted)
  - Need significant regularization and/or correction for multiple comparisons

Why Covariance Estimation?

- Mean level differences are often small relative to sample variability
- Covariance estimation can improve estimates of mean level differences
- Correlations are indicative of functional groups in the metabolome
- Correlations between metabolites are driven by unobserved variables
    - Disease progression or severity
    - Genetics
    - Important unmeasured molecules (e.g. metabolic enzymes)
    - Diet / extrinsic factors

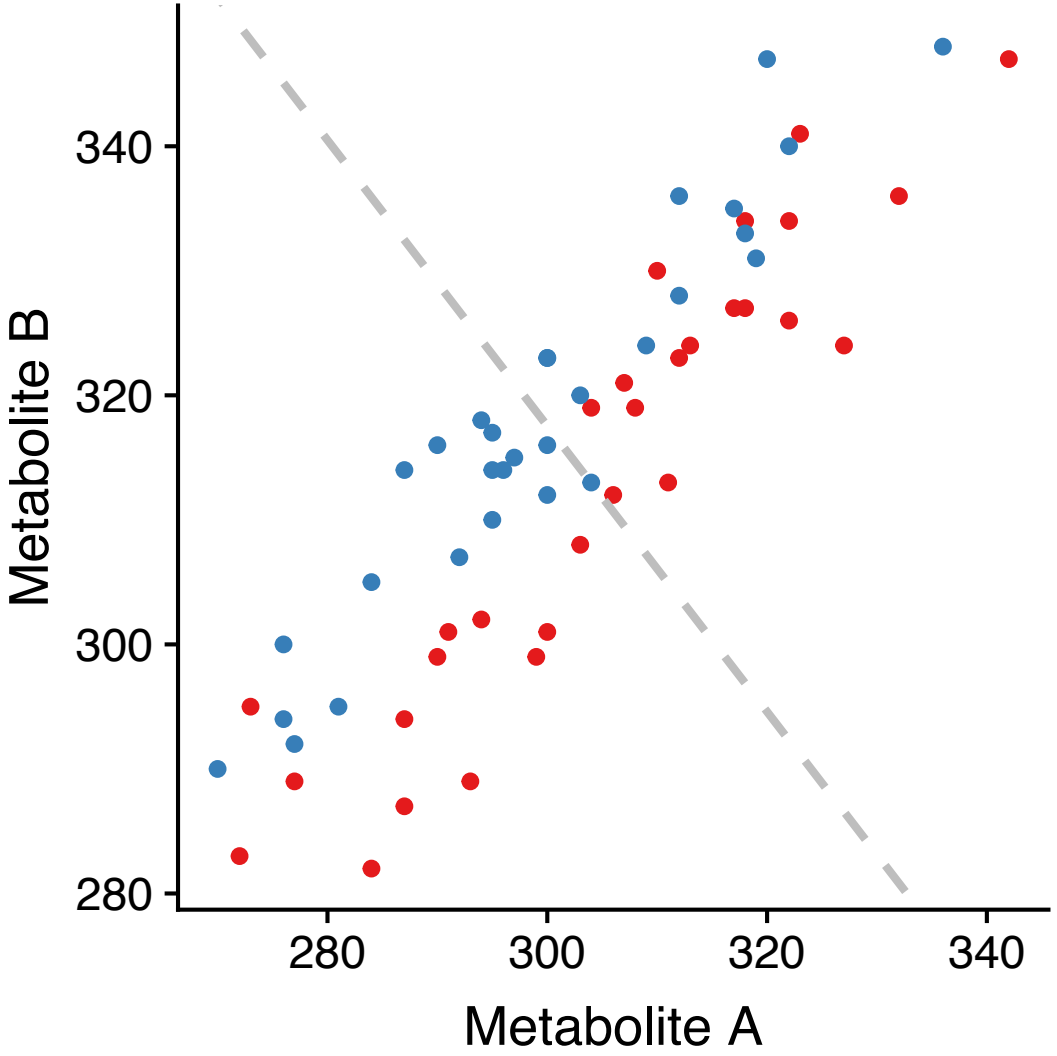

Metabolite B
340
320
300
280
280
300
320
340
Metabolite A
Phenotype
Case
Non-Case

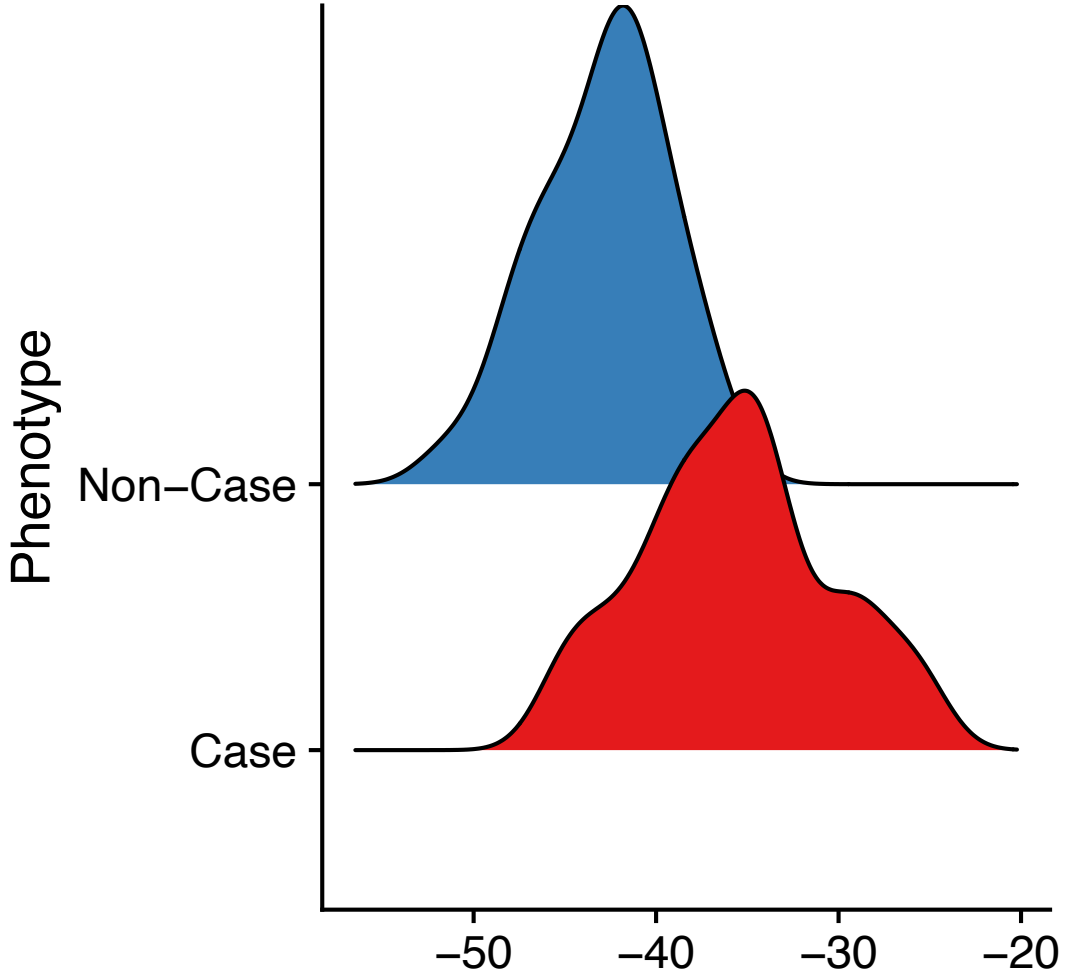




P-value: 0.20

P-value: 0.28

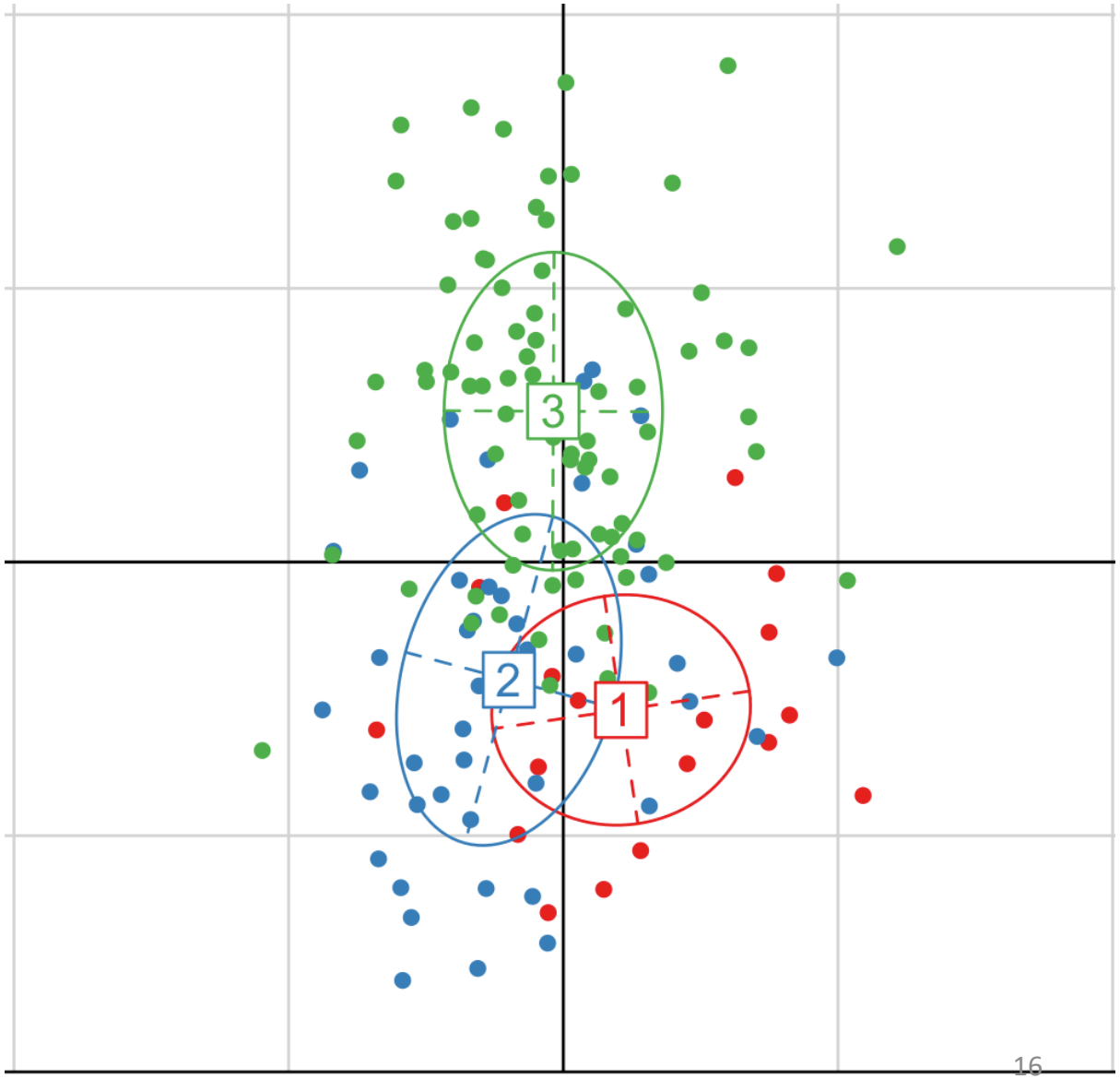

Metabolite B
340
320
300
280
280
300
320
340
Metabolite A


Phenotype
Non–Case
Case
–50
–40
–30
–20

P-value: < 1e-9

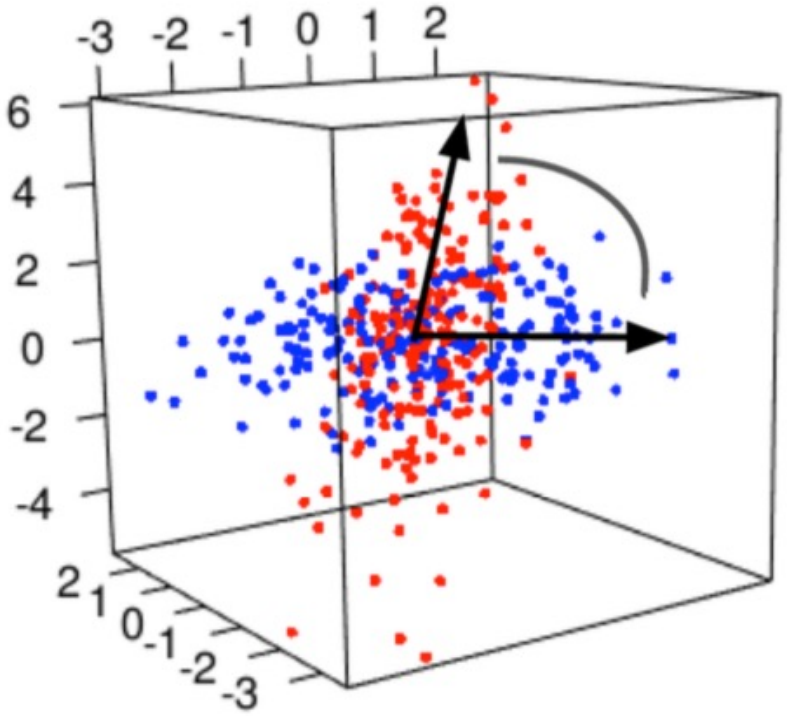
Principal Component Analysis

- PCA one of the most common dimension reduction techniques
- Latent factors explain data
- Run single PCA for all data
- Often used identify mean differences
- Unsupervised learning



Multi-group PCA

- Group by phenotype
- Do correlations differ by group?
- Infer different PCs for different groups
- Shared subspace models
  - Large p, small n
  - Share information across groups



Identifying Relevant Dimensions of Variability

- Find a subspace of variability that is invariant to changes in X
  - “Nuisance variability”
- Find the smallest subspace of variability that *not* invariant to X
  - Find all of the variation in Y that changes with X
- Requires inference for a *subspace*
  - Characterizes differences in mean and covariances in metabolites for different phenotypes

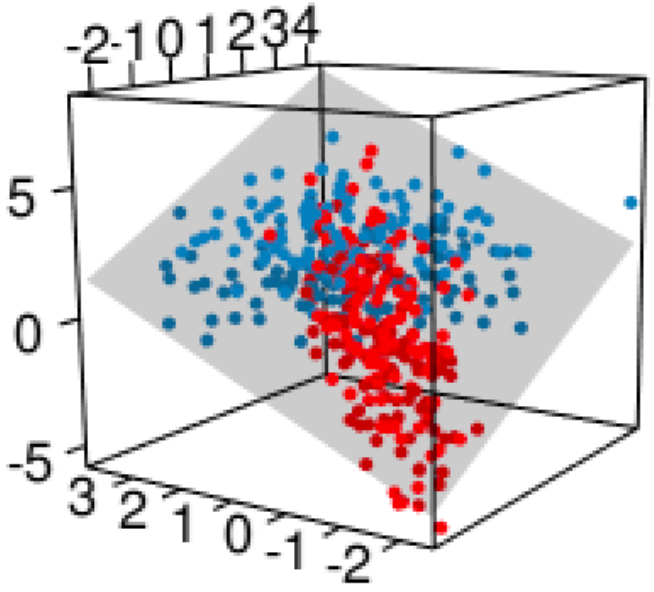
Shared Subspace Assumption

Data from similar sources often share similar structure.

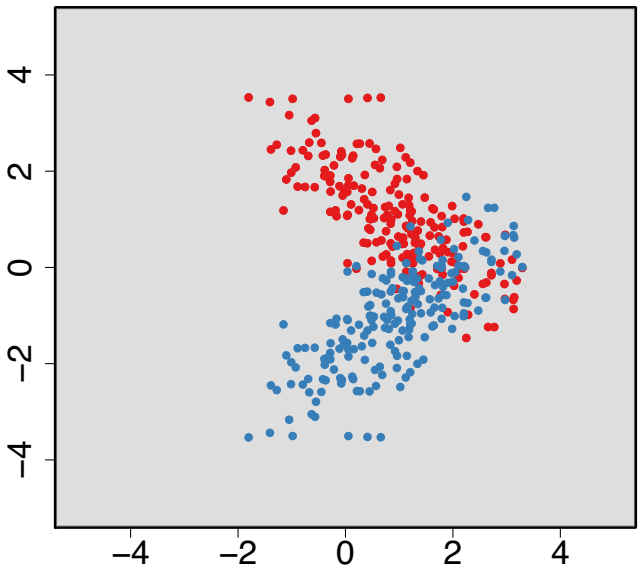
- Effective dimensionality related to number of regulatory modules
- Most structure is common across groups
- Suggests that differences between groups are on a lower dimensional shared subspace

A Shared Subspace Model

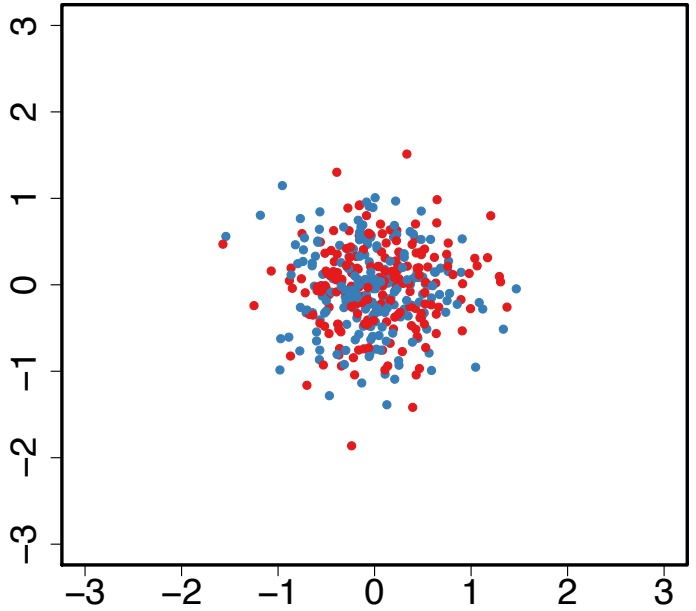
Assumption: Differences between groups are on a shared subspace.



Projection in $\mathbb{R}^3$



Subspace projection



Orthogonal projection

A Shared Subspace Model

Data from group k is multivariate normal:

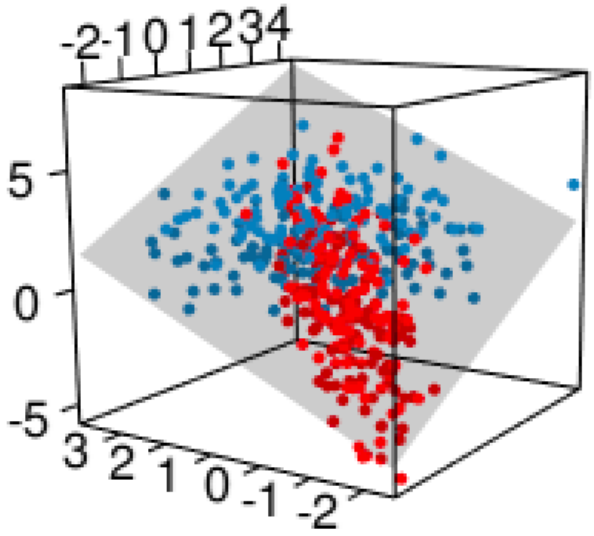
$$Y_k \sim N\left(\mu_k, \Sigma_k \otimes I\right)$$

with covariance

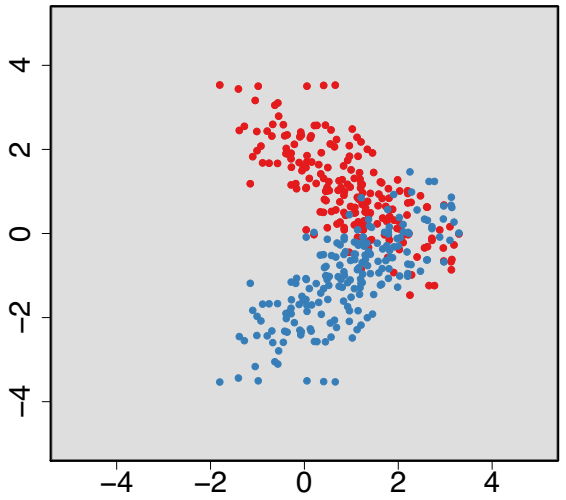
$$\boldsymbol{\Sigma_k = V \Psi_k V^T + \sigma_k^2 I}$$

- $\boldsymbol{V}$ is a p x s orthogonal matrix
- $\mathrm{span(V)}$ corresponds to the s-dimensional shared subspace of
- $\boldsymbol{\Psi_k + \sigma_k^2 I}$ are the rank s covariance matrices of projected data $\mathbb{R}^p$

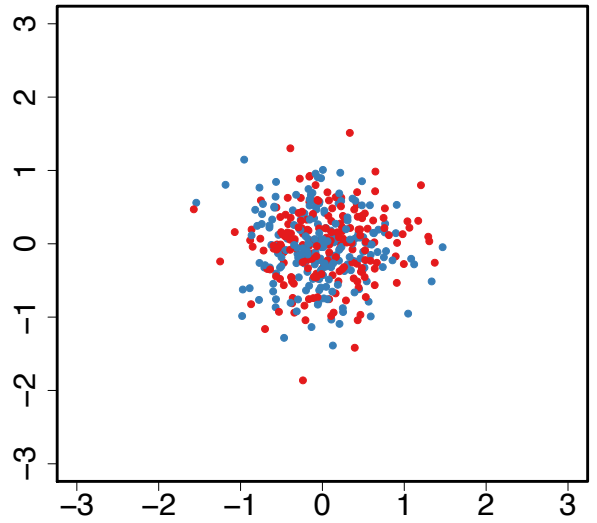
A Shared Subspace Model



Projection in $\mathbb{R}^3$



$Y_k V$



$Y_k V_\perp$

- span(V) is represented by the gray plane with s = 2
- Differences in Ψ_k and μ_k reflected in the span(V)
- No differences between groups on span($V_\perp$)

A Shared Subspace Model

- Find the “best” shared subspace of fixed dimension s
- Infer heterogeneity of the projected covariance matrices, Ψ_k
- Quantify uncertainty about differences in covariances
- Full Bayesian inference is hard
  - V is high dimensional
  - Orthogonality constraints means sampling on a manifold

Shared Subspace Objective Function

$$\ell\left(V, \Psi_k, \sigma_k^2\right) = \sum_k \operatorname{tr}\left(\left(\frac{1}{\sigma_k^2} V V^T - V\left(\Psi_k + \sigma_k^2 I\right)^{-1} V^T\right) S_k / 2\right)$$

- Maximize over $V \in \mathcal{V}_{p,s}$ (Stiefel manifold)
- $VV^T \in \mathcal{G}_{p,s}$ is called the Grassmanian manifold

For comparison, the PCA objective is $\ell(V) = \operatorname{tr}(V^T S V)$

Empirical Bayes Inference

$$\ell\left(V, \Psi_k, \sigma_k^2\right) = \sum_k \mathrm{tr}\left(\left(\frac{1}{\sigma_k^2} VV^T - V\left(\Psi_k + \sigma_k^2 I\right)^{-1} V^T\right) S_k/2\right)$$

- "Integrate out" Ψ_k and σ_k^2 to maximize marginal log-likelihood, $\ell(V)$
- Expectation Maximization algorithm to estimate V
- Bayesian inference for Ψ_k given the inferred subspace V.

EM Inference in the Shared Subspace

$$\ell\left(V, \Psi_k, \sigma_k^2\right) = \sum_k \mathbf{tr}\left(\left(\frac{1}{\sigma_k^2} VV^T - V\left(\Psi_k + \sigma_k^2 I\right)^{-1} V^T\right) S_k/2\right)$$

- E-step:

$$\mathcal{M}_t^{-1} = E\left[\left(\Psi_k + \sigma_k^2/\right)^{-1} | V_{(t-1)}\right] = n_k \left(V_{(t-1)}^\top S_k V_{(t-1)}\right)^{-1}$$

$$\tau_t = E\left[\frac{1}{\sigma_k^2} | V_{(t-1)}\right] = \frac{n_k(p-s)}{\mathrm{tr}\left[\left(1 - V_{(t-1)} V_{(t-1)}^\top\right) S_k\right]}$$

- M-step:

$$V_t = \underset{V \in \mathcal{V}_{p,s}}{\arg\max} \sum_k \mathrm{tr}\left(-\left(V \mathcal{M}_t V^\top + \tau_t V V^\top\right) S_k/2\right)$$

Inference in the Shared Subspace Model

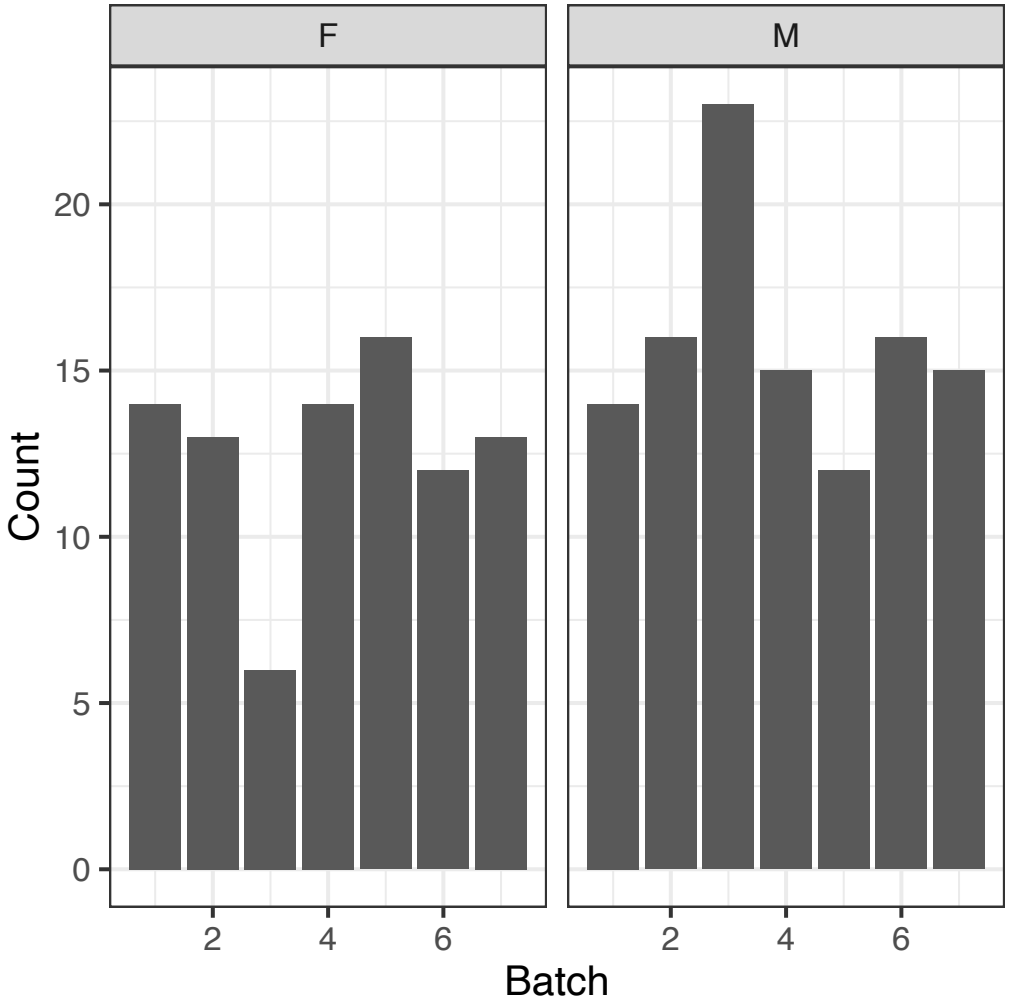
- Optimization on the Stiefel Manifold
  - Computational complexity dominated by s, not p (Wen and Yin, 2013)
  - Efficient for 10k+ features if subspace dimension when s is moderate
  - Implemented in the R package *rstiefel* (Hoff and Franks)
- Bayesian inference for the projected data covariance matrices
  - Low dimensional and tractable, facilitates uncertainty quantification

Analysis of Metabolomics Data

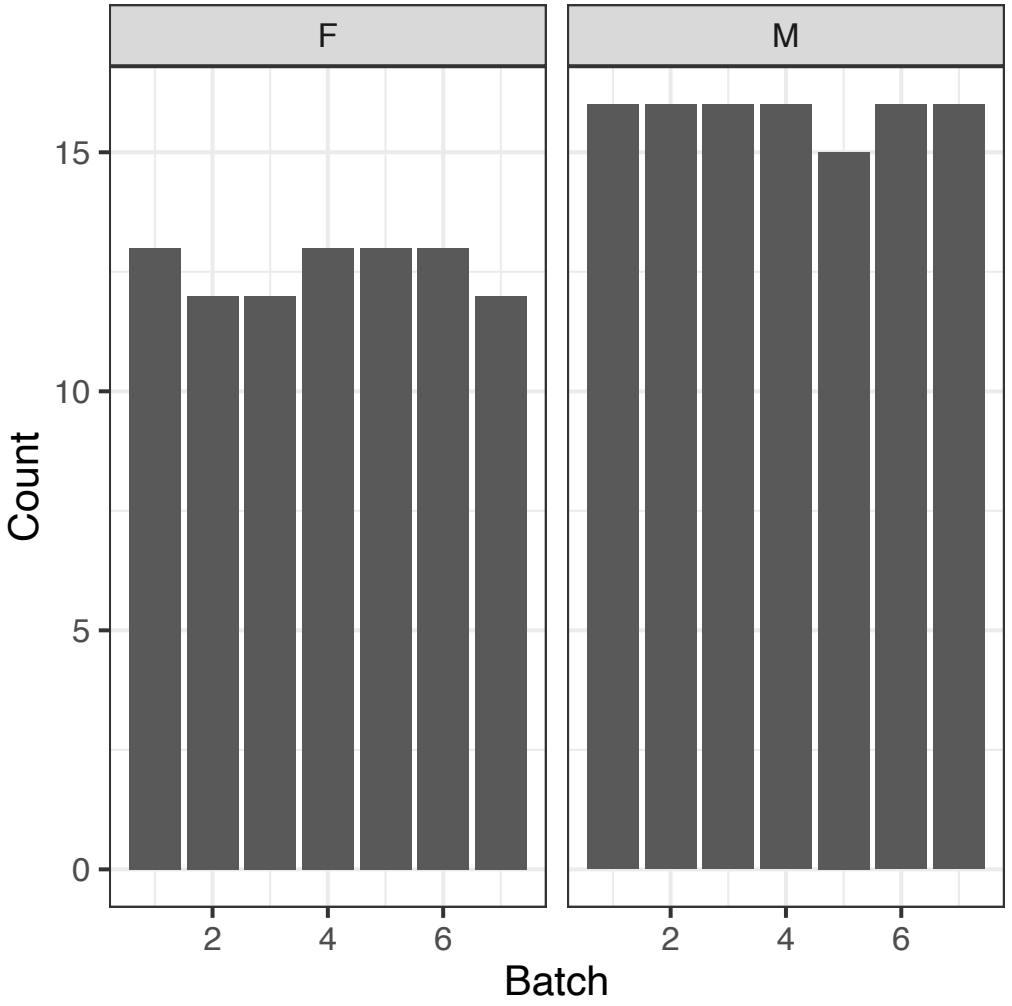
- Batch effects and drift can be large and obscure signals
- Samples prepped in 7 batches of about 30 subjects each
- At the very least, randomize the samples
- Can do better: explicitly maximize balance of features across batches

Sex

Randomized

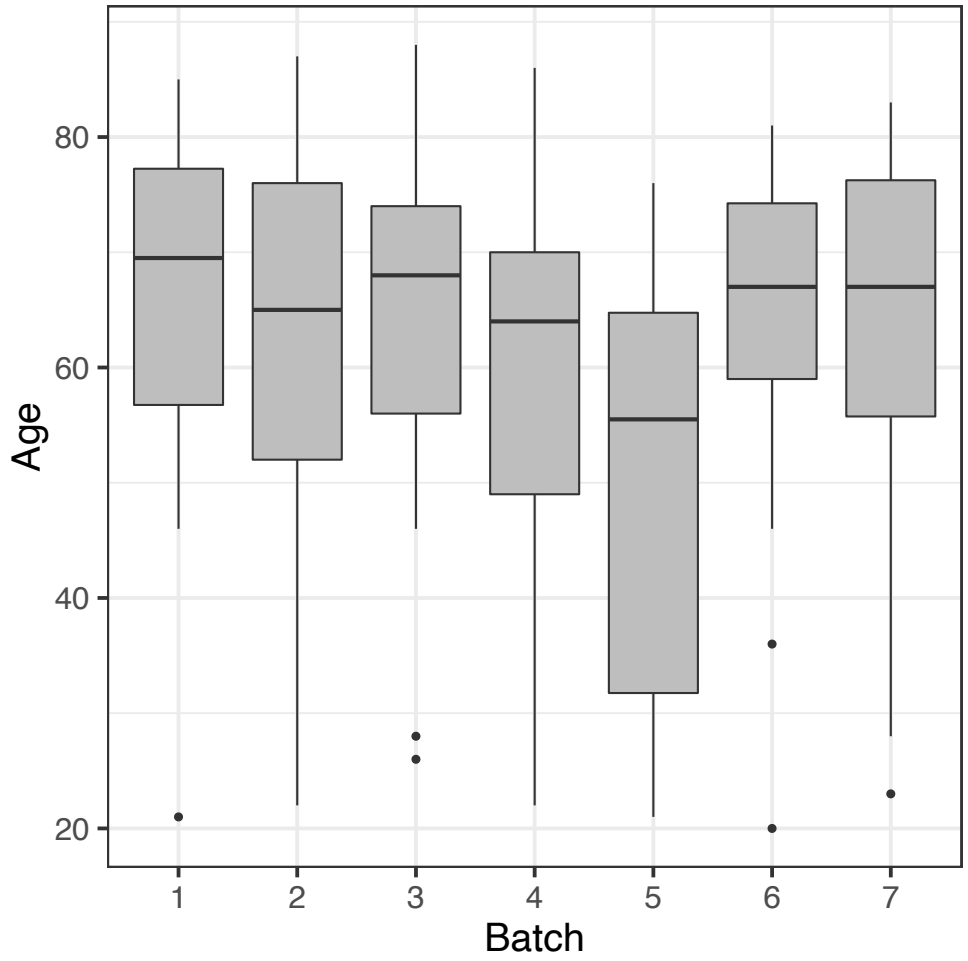


Optimized

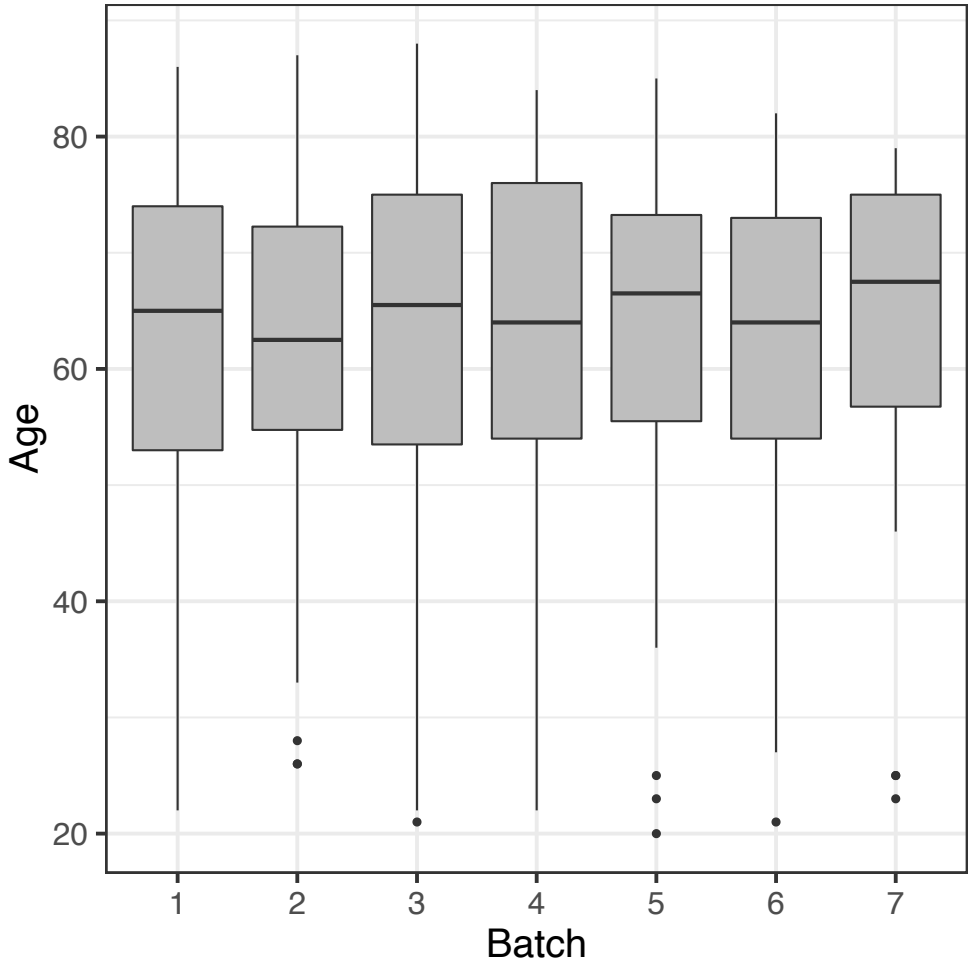


Age

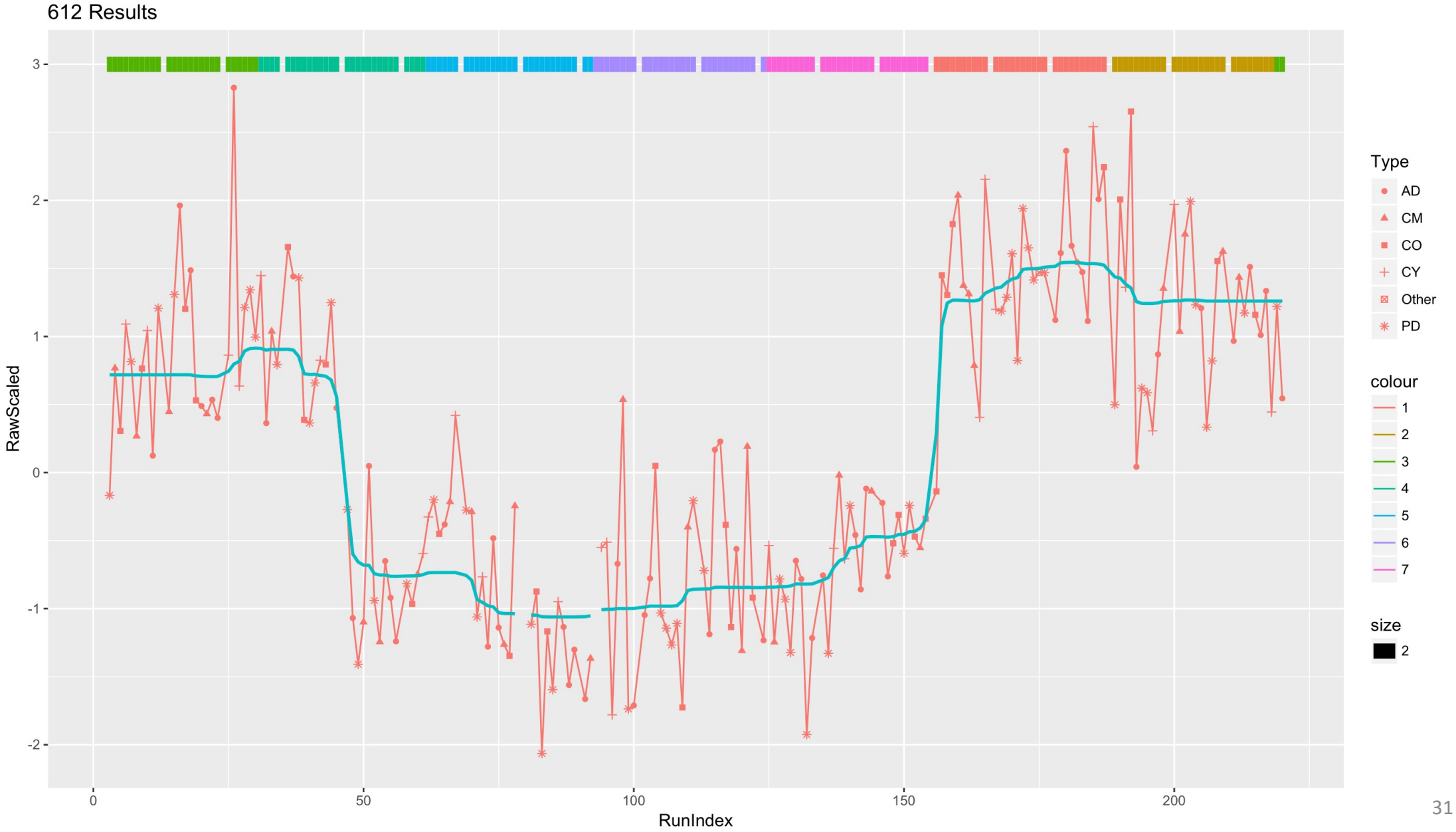
Randomized

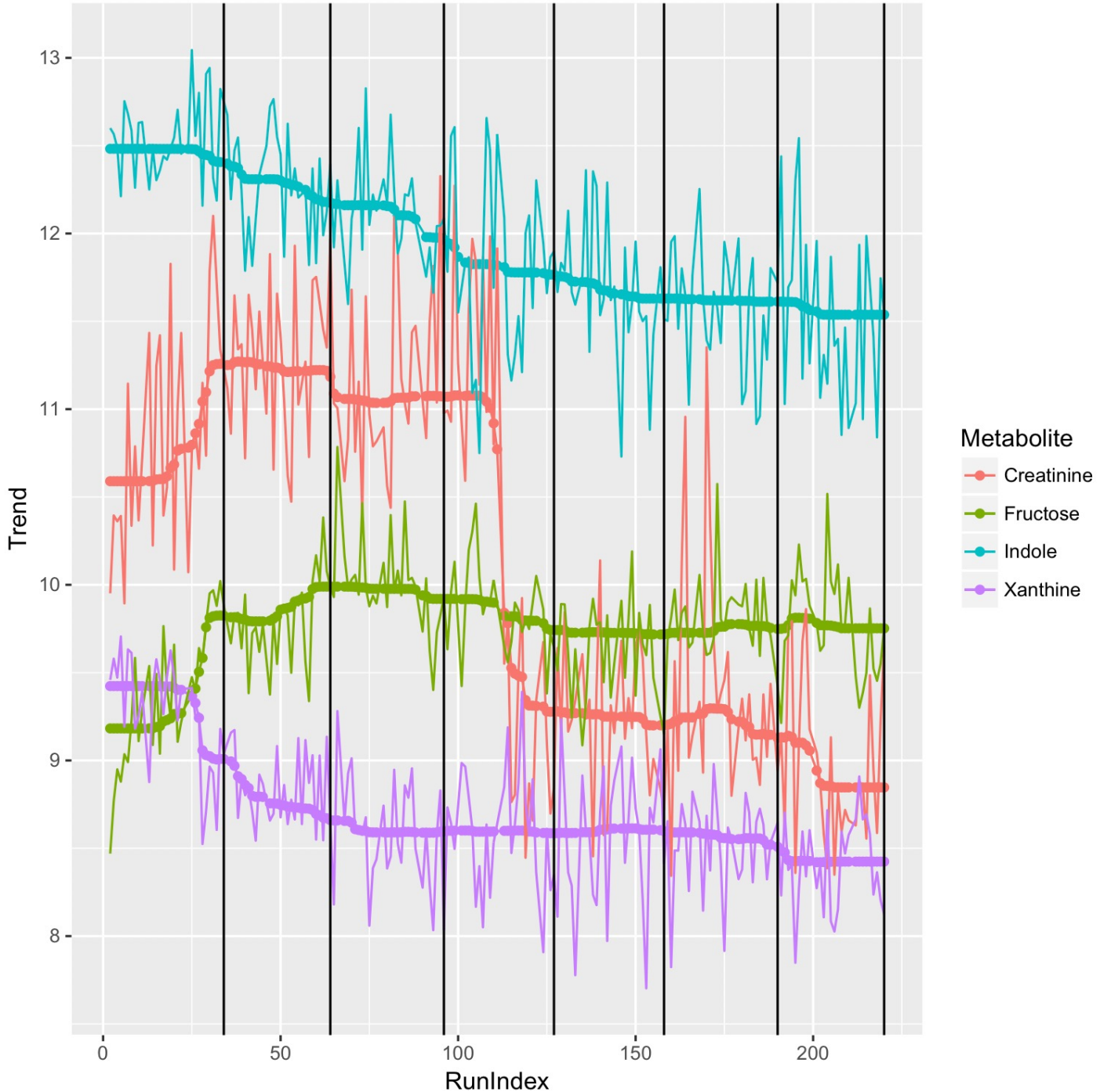


Balance Optimized



Correcting for drift in metabolites abundances




13
12
11
10
9
8
Trend
0
50
100
150
200
RunIndex
Metabolite
Creatinine
Fructose
Indole
Xanthine

Analysis of Metabolomic Data

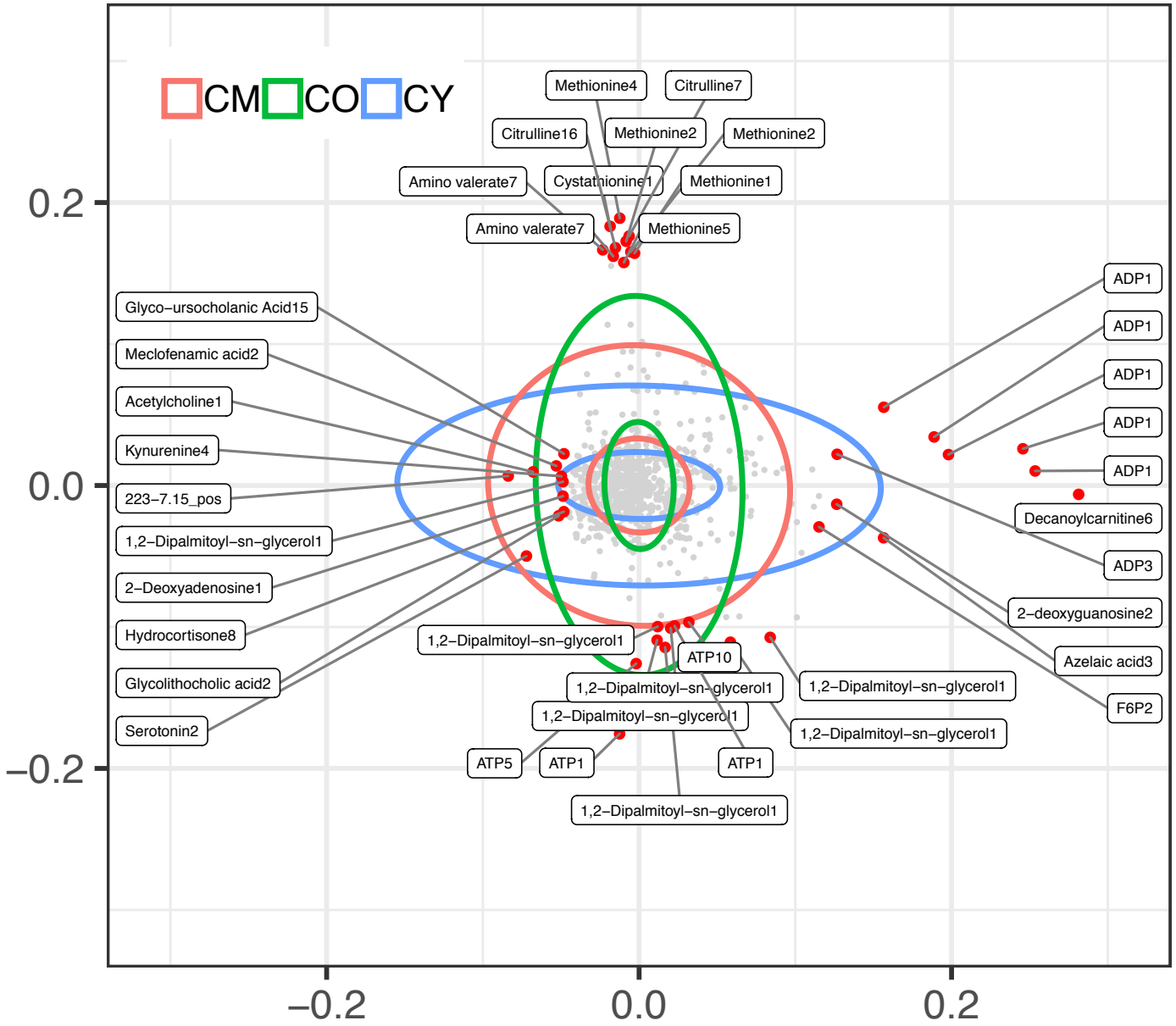
- Cerebrospinal fluid samples (CSF) from 198 individuals. Samples from
  - 57 Alzheimer's disease (AD)
  - 56 Parkinson's disease (PD)
  - 85 controls split by age (young (CY), middle age (CM) and old age (CO))

- De-trend drift using non-parametric regression (boosted trees)

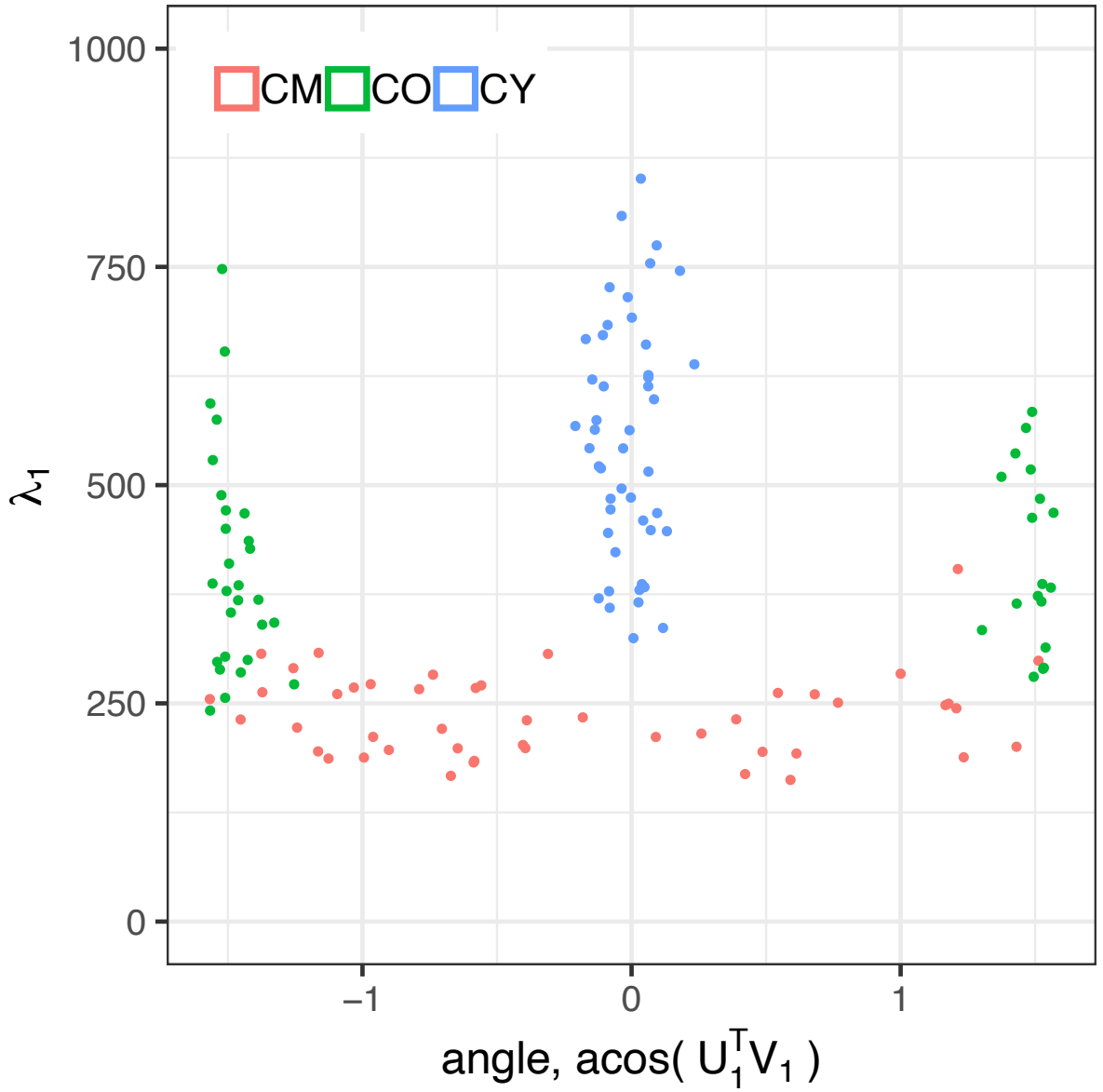
- Fit shared subspace model, explore differences in correlations

Can we detect heterogeneity across correlation matrices?

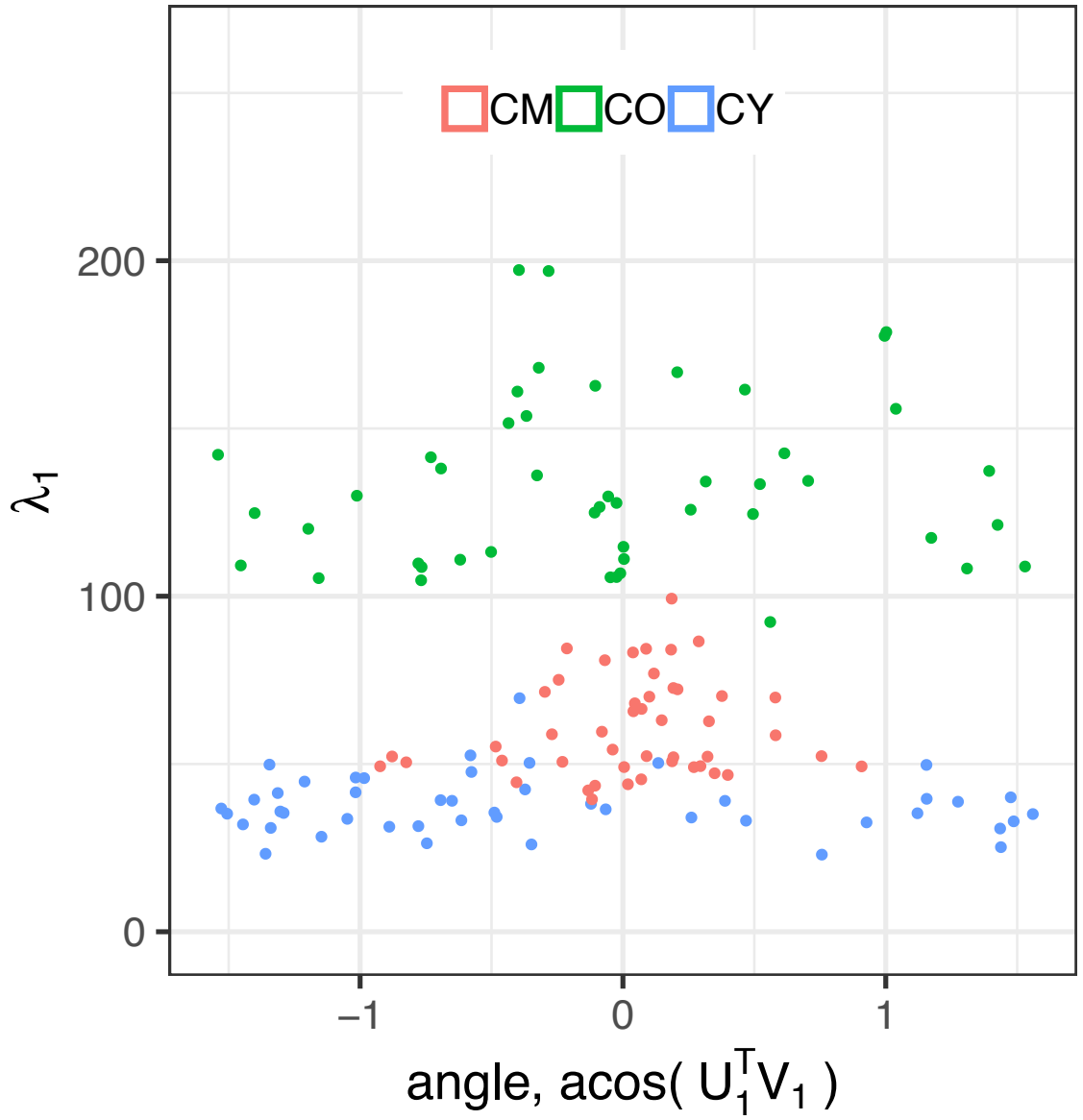
Metabolite Correlations Change with Age

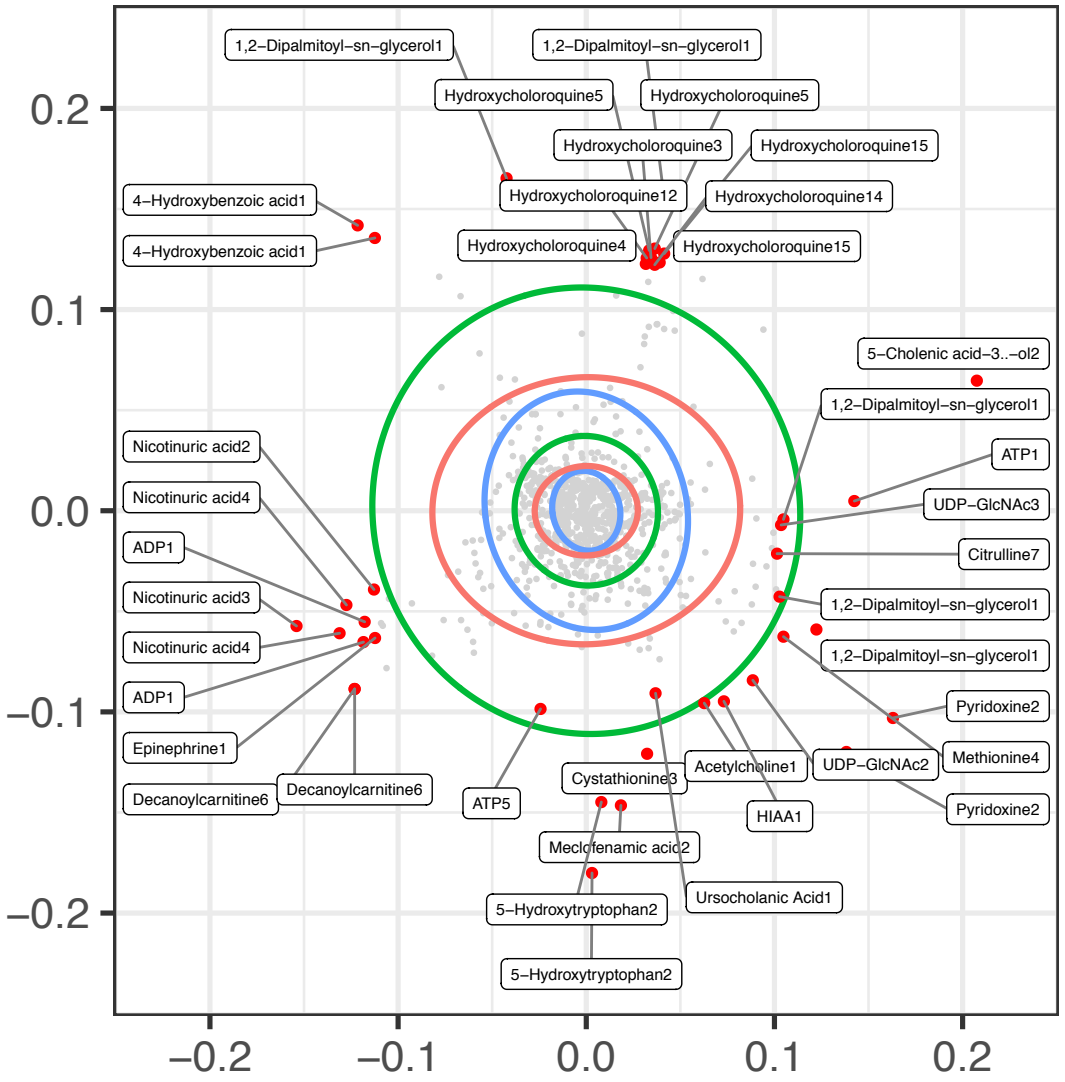


Posterior Uncertainty

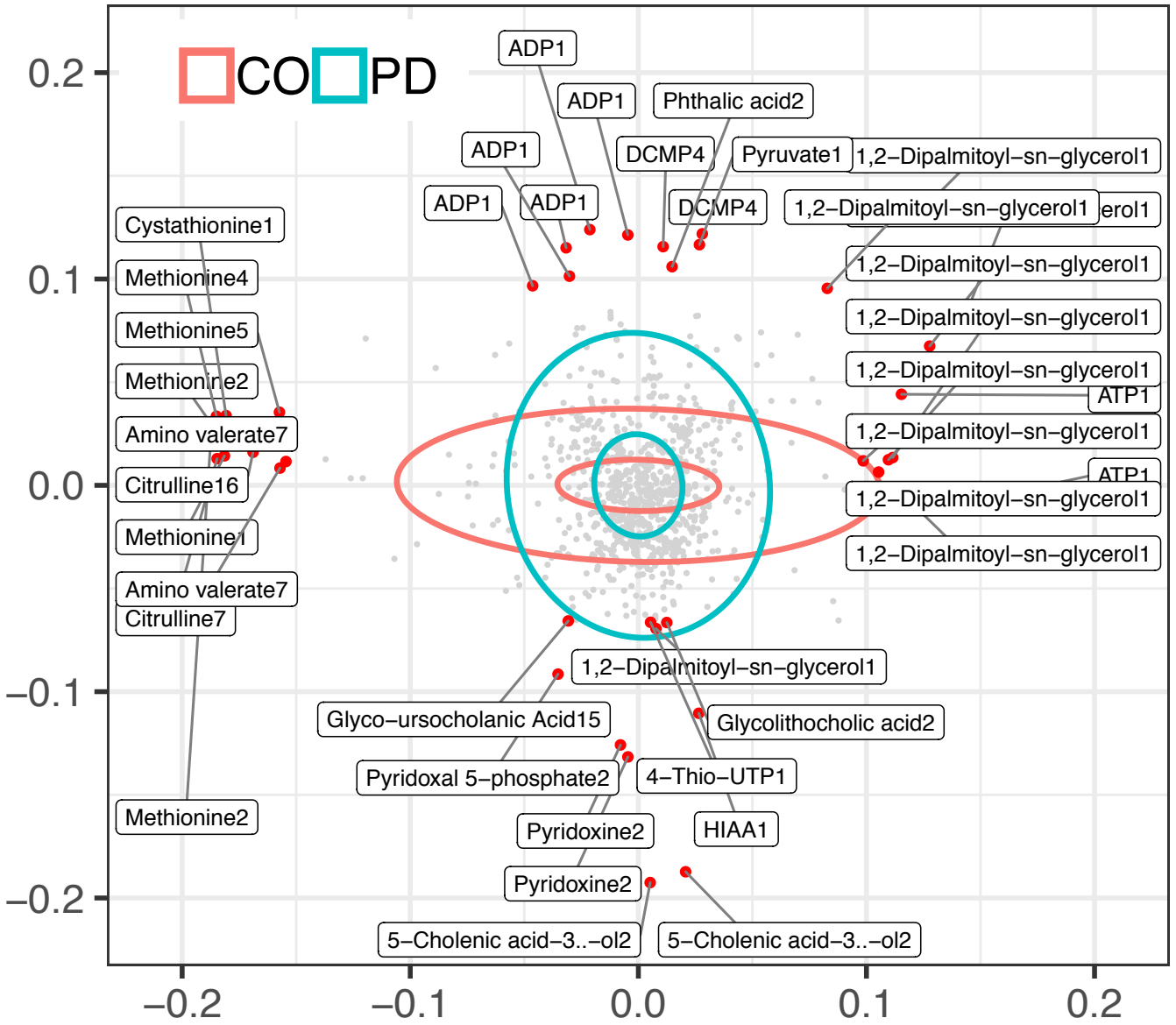


Metabolite Correlations Change with Age

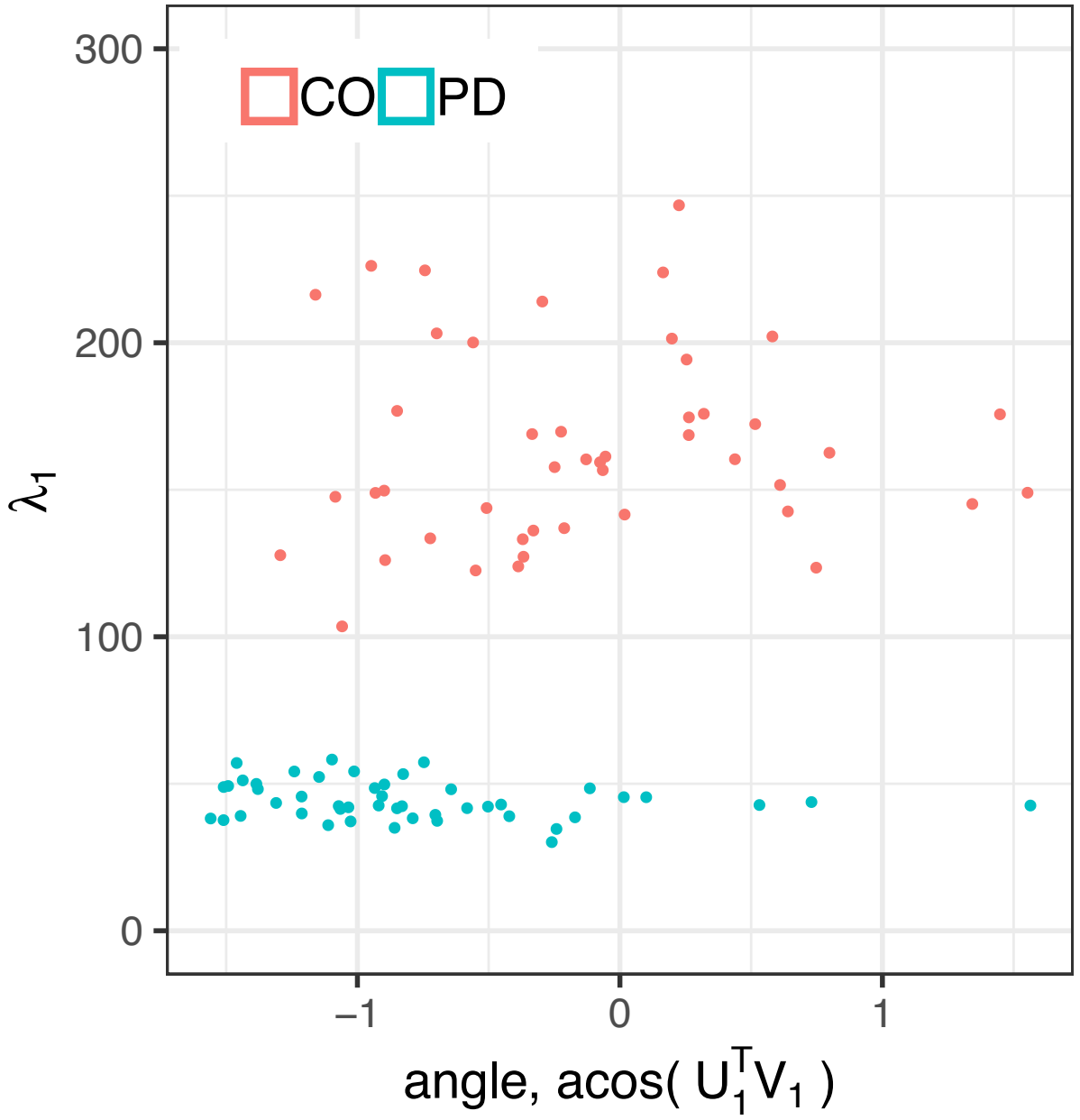


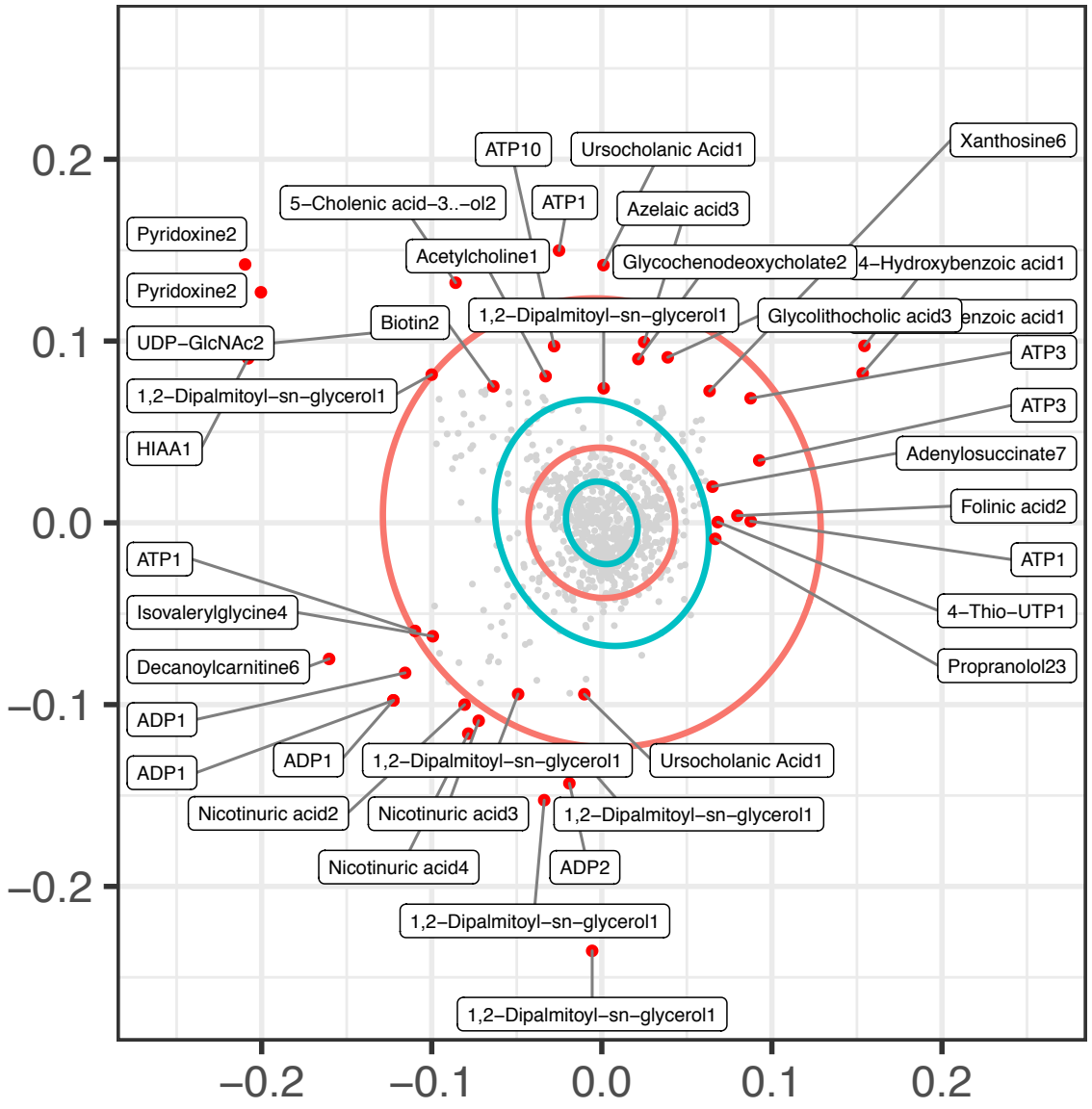


Metabolite Correlations in Parkinson's

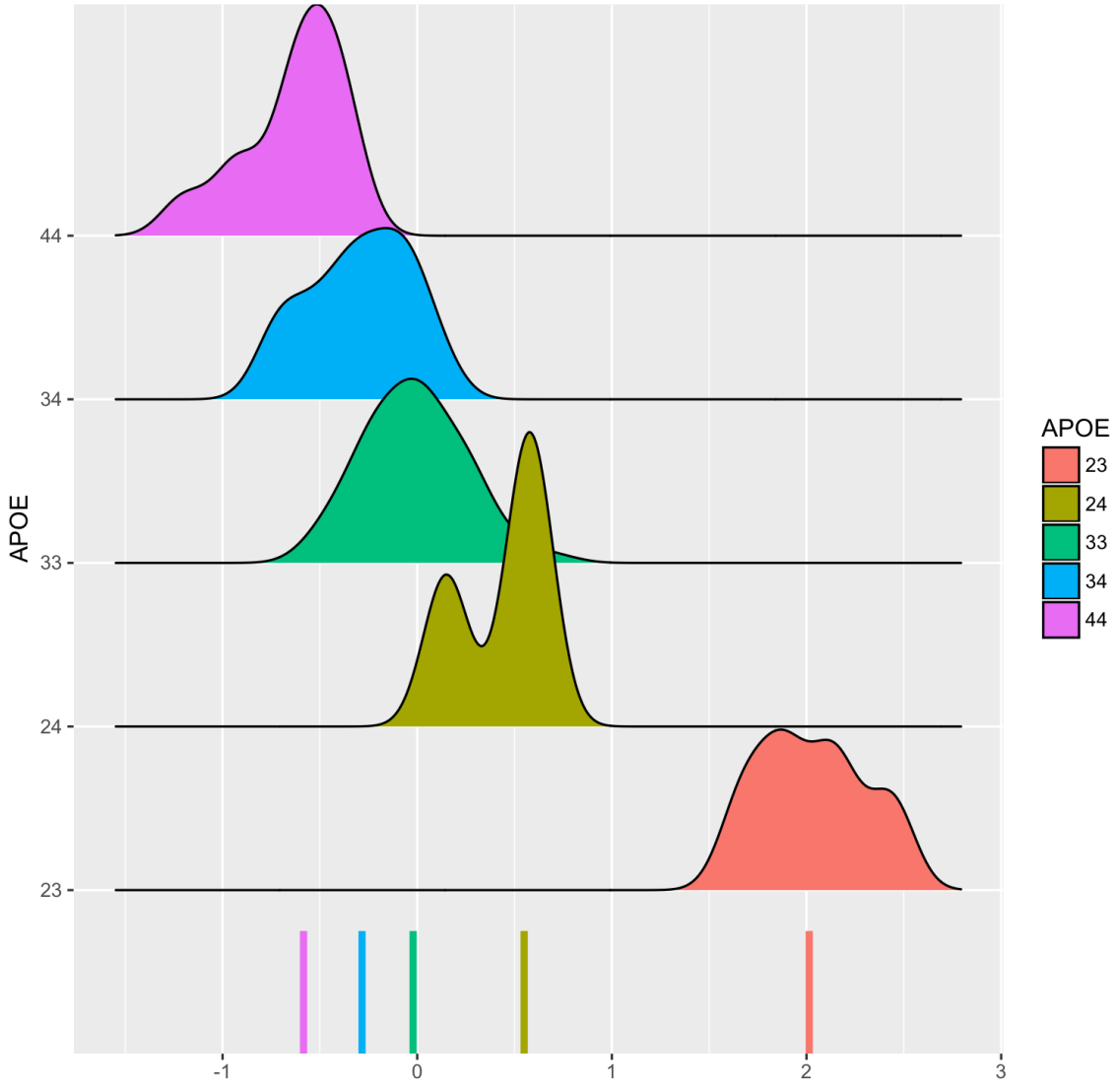


Metabolite Correlations in Parkinson's





Multivariate Analysis of ApoE



Next Steps

- Pathway analysis and interpretation
  - Enrichment
  - Network models (de novo reconstruction)
- Improved metabolite identification
- Robust inference
  - Extend methodology to heavy-tailed distributions
  - Multivariate t or laplace distributions

Some Remarks

- Papers
  - Shared Subspace Models for Multi-Group Covariance Estimation https://arxiv.org/abs/1607.03045

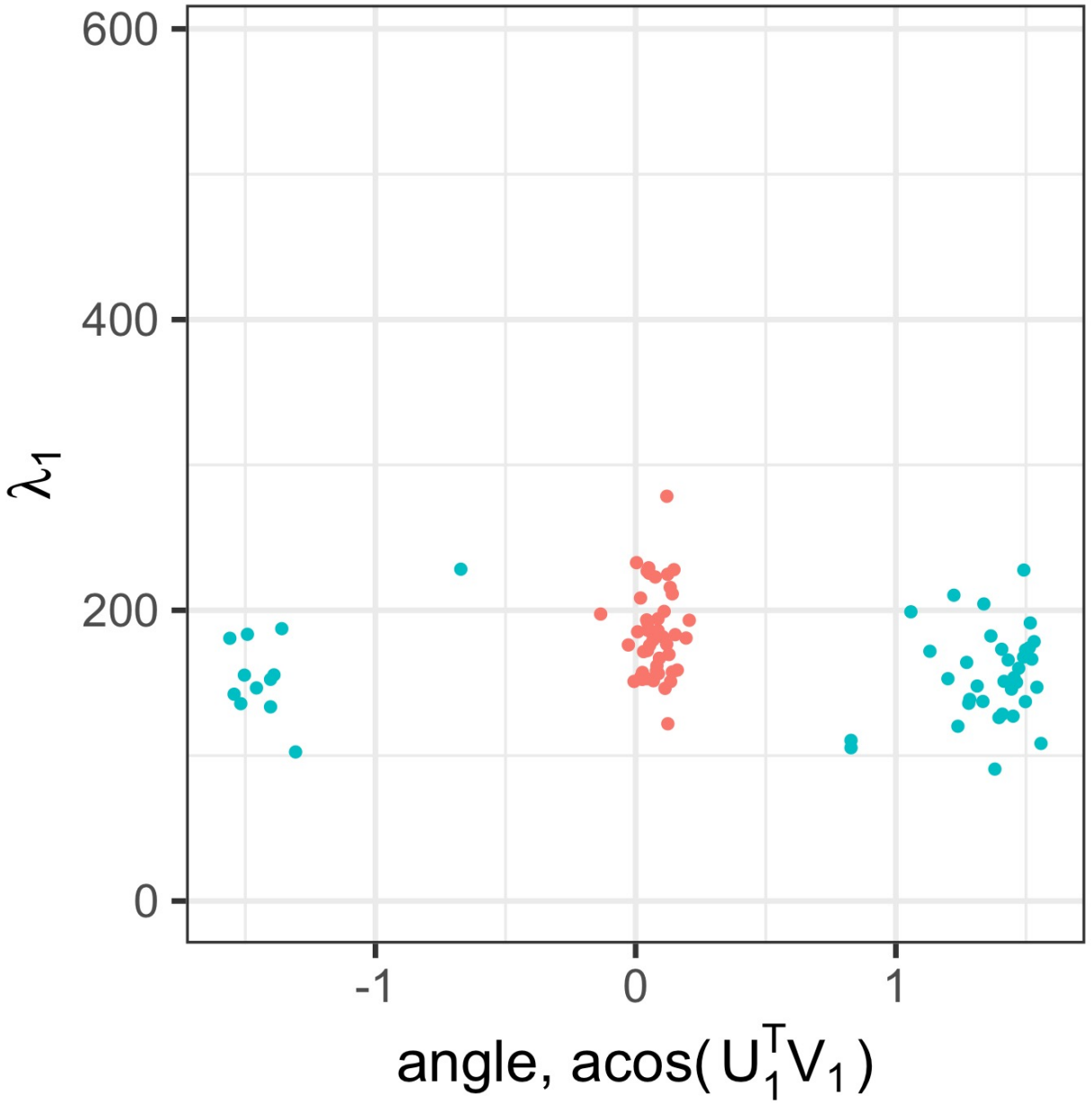
- Software
  - https://github.com/afranks86/shared-subspace
  - *rstiefel*: R package for optimization on the Stiefel manifold (w/ Peter Hoff)
  - *mgCov*: Forthcoming R package multi-group covariance

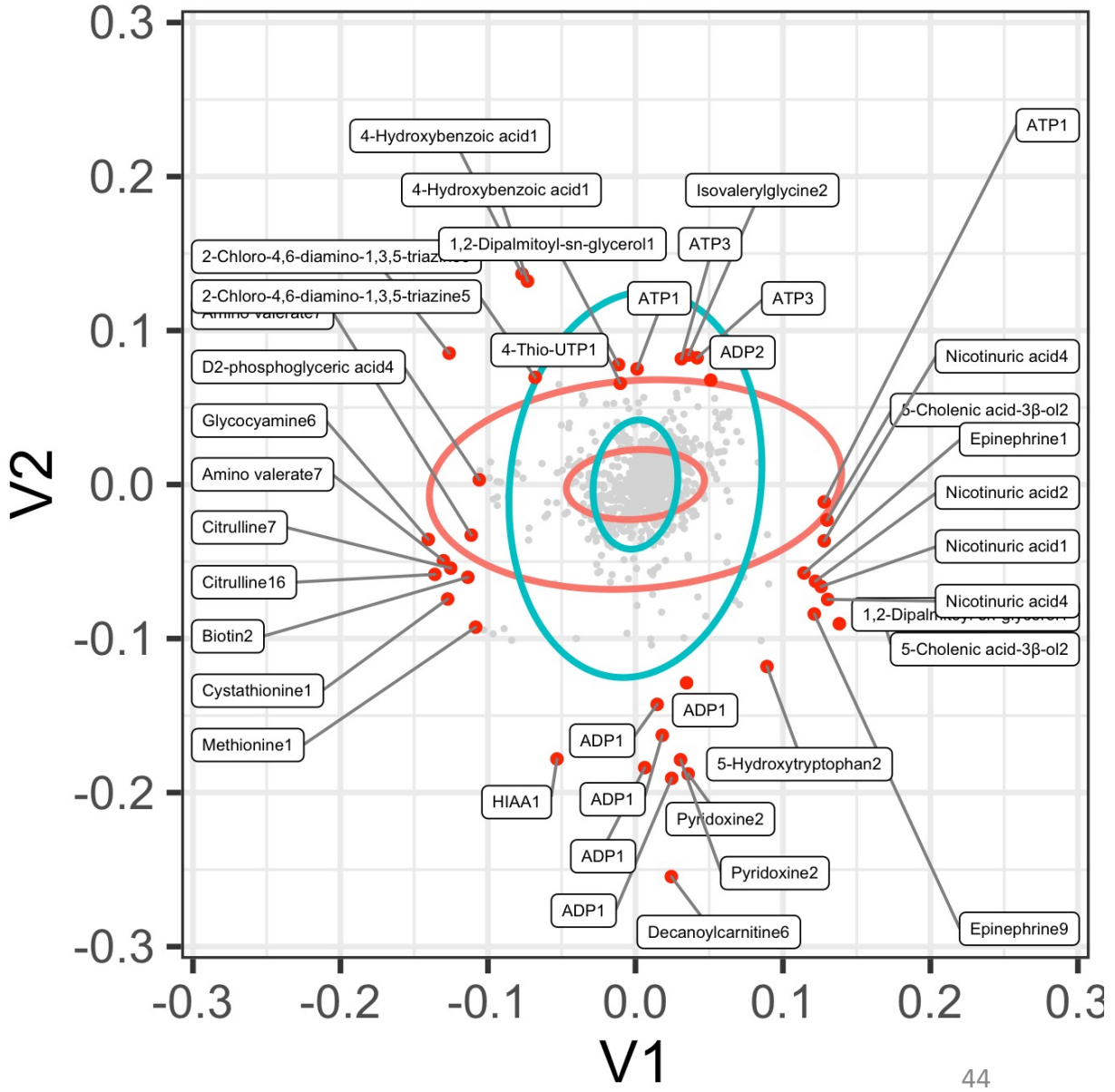
Acknowledgements:

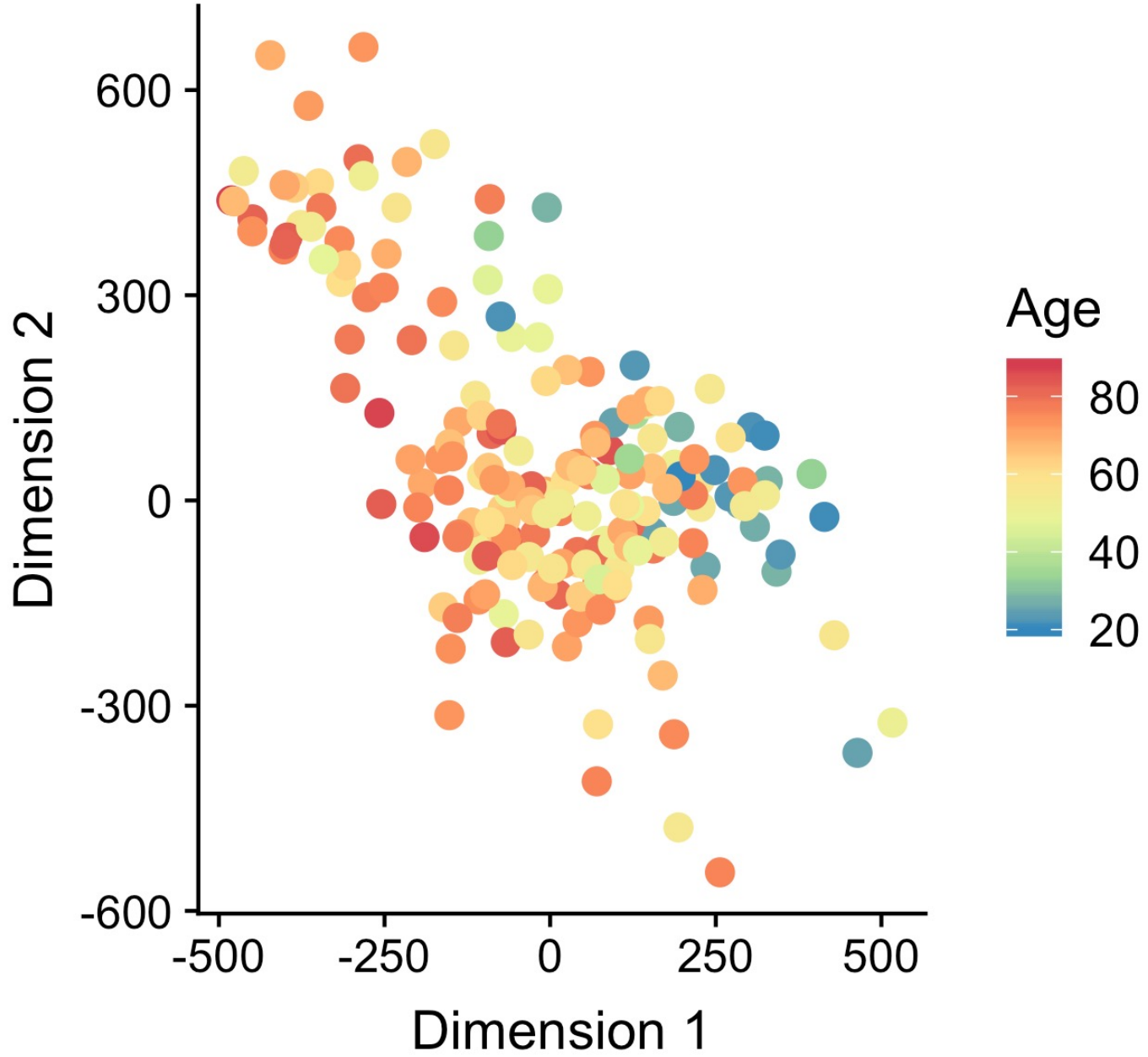
- Daniel Promislow (University of Washington, Pathology)
- Peter Hoff (Duke University, Statistical Science)
- Daniel Raftery (Northwest Metabolomics Research Center)
- Marie Davis (University of Washington, Neurology)
- Cyrus Zabetian (University of Washington, Neurology)
- Elaine Peskind (University of Washington, Psychiatry and Behavioral Sciences)

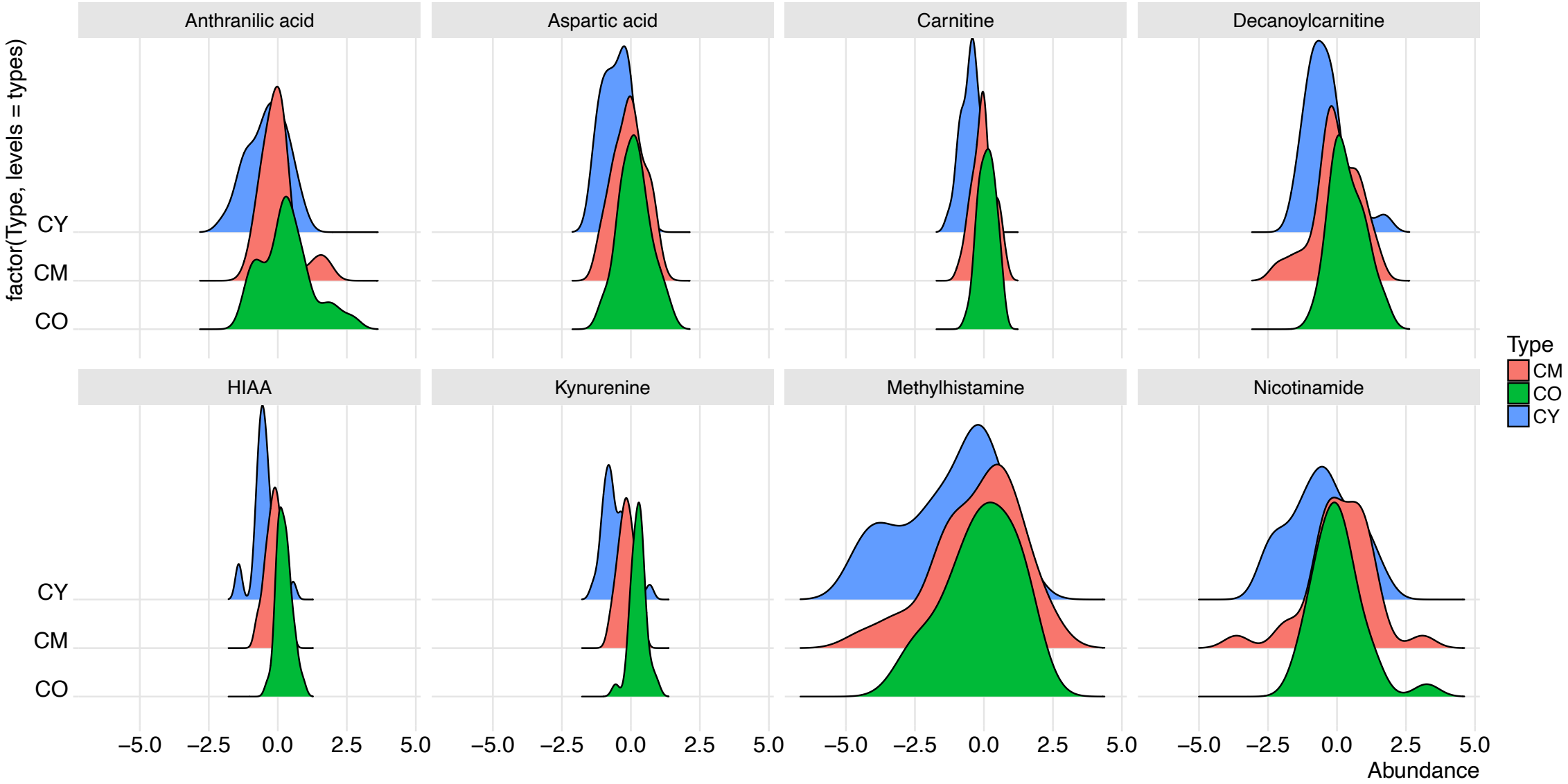
Thank you!

Metabolite Correlations in Alzheimer's






600
300
0
-300
-600
Dimension 2
-500
-250
0
250
500
Dimension 1
Age
80
60
40
20


Anthranilic acid
Aspartic acid
Carnitine
Decanoylcarnitine
HIAA
Kynurenine
Methylhistamine
Nicotinamide
factor(Type, levels = types)
CY
CM
CO
−5.0
−2.5
0.0
2.5
5.0
Abundance
Type
CM
CO
CY