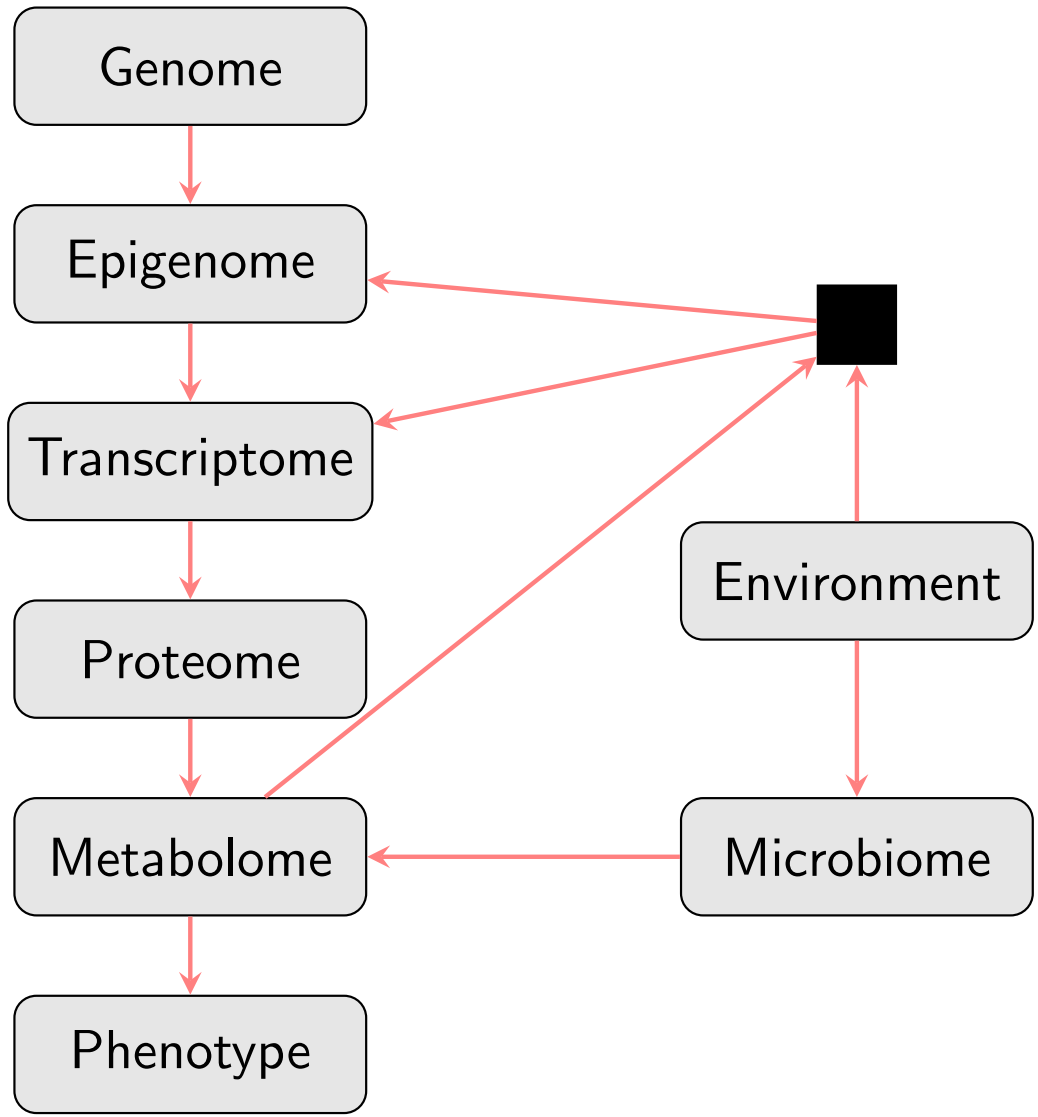
Regression Analysis of Multi-view Microbiome Data

Jing Ma

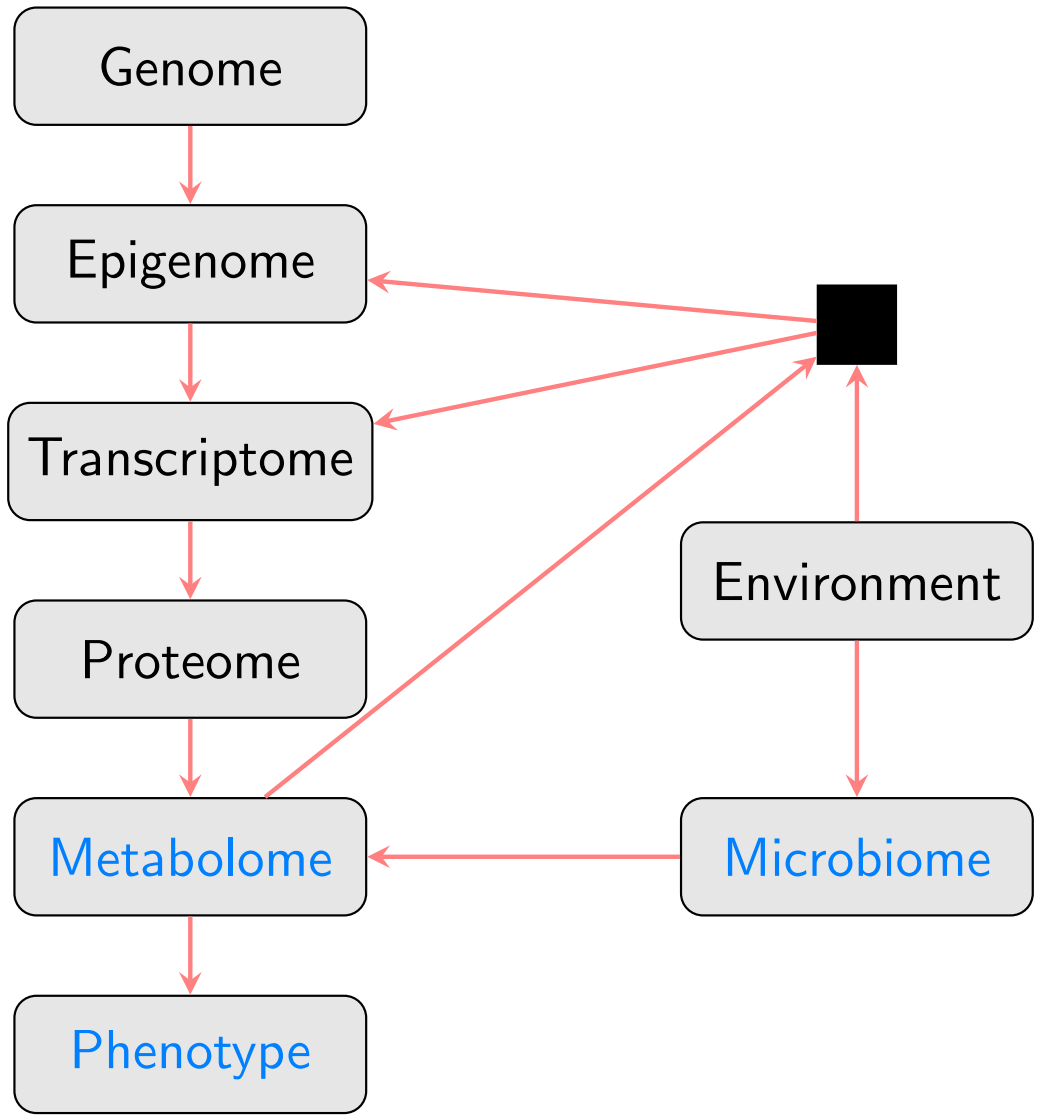
Division of Public Health Sciences
Fred Hutchinson Cancer Research Center

21 February 2022
IMSI Workshop

Systems Biology



Systems Biology



Outline

BactoCARB Study

A randomized crossover study of 80 subjects aimed at the effects of high/low glycemic load on a variety of biomarkers.

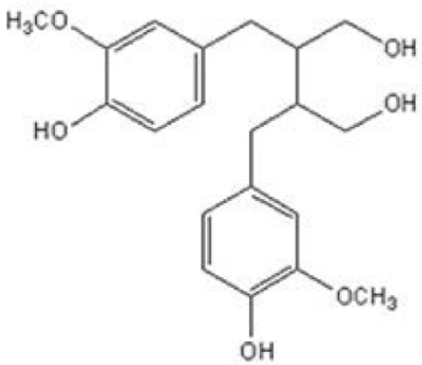
- $X^{(1)} = 144$ metabolites & bile acids (blood)
- $X^{(2)} = 134$ microbial genus abundances (stool)
- $y =$ enterolactone 'ENL' (urine), a product jointly produced by microbes and metabolites

Problem of Interest: which metabolites and bacteria are *jointly* associated with y?

BactoCARB Study

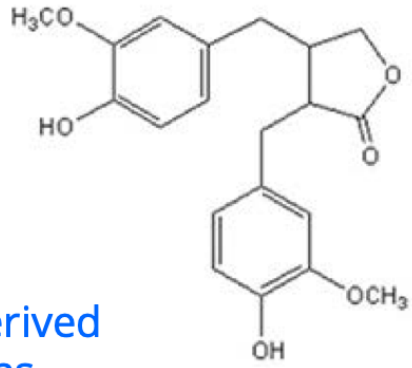
Maybe introduce a slide on how the data look like

Microbial Metabolism



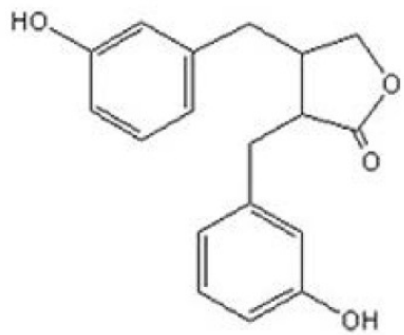
Secoisolariciresinol

plant-derived lignans



Matairesinol

etc . . .

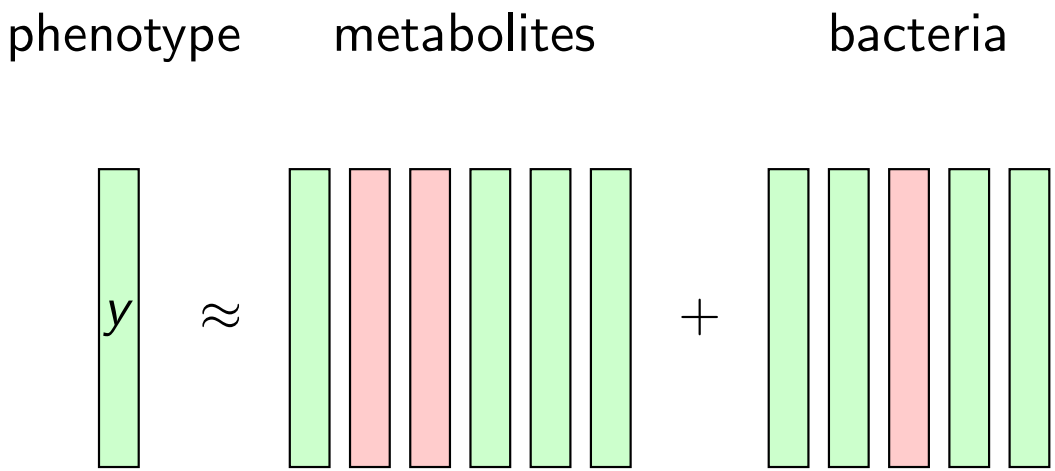


Enterolactone

more is better!

https://lpi.oregonstate.edu/mic/dietary-factors/phytochemicals/lignans

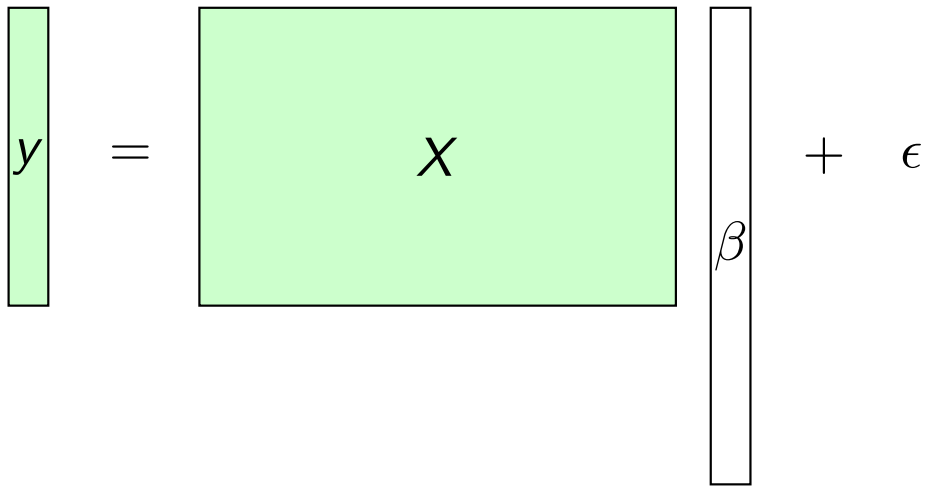
Today's Goal



Find correlated metabolites and bacteria jointly associated with y

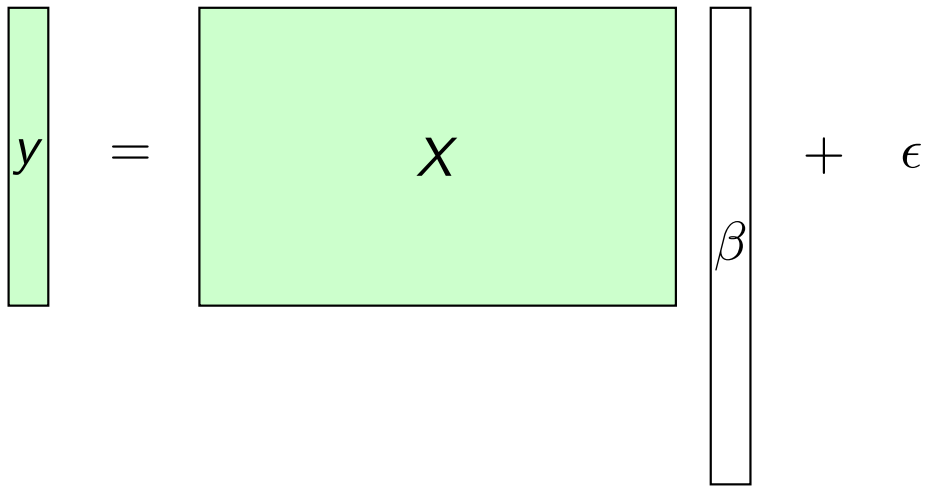
Outline

Regression Analysis of One Dataset



Problem of Interest: which variables are associated with y?

Regression Analysis of One Dataset



Problem of Interest: which variables are associated with y?

Common challenges:

- number of variables $p > n$
- correlations among variables

Ridge Regression

$$\begin{aligned}\hat{\beta}^{\text{ridge}} &= \arg\min_{\beta} \left\{ \|y - X\beta\|^2 + \lambda \|\beta\|^2 \right\} \\ &= (X^\top X + \lambda I_p)^{-1} X^\top y\end{aligned}$$

The ridge estimator $\hat{\beta}^{\text{ridge}}$ has less variance, though it introduces bias.

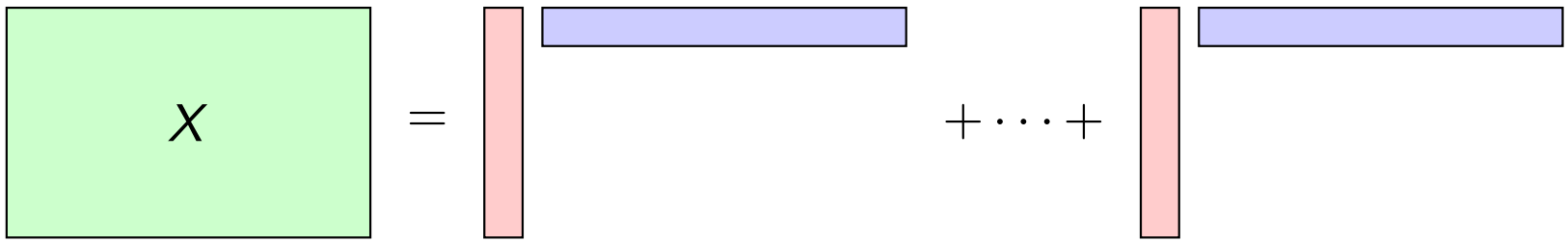
Ridge Regression

$$\begin{aligned}\hat{\beta}^{\text{ridge}} &= \arg\min_{\beta} \left\{ \|y - X\beta\|^2 + \lambda \|\beta\|^2 \right\} \\ &= (X^\top X + \lambda I_p)^{-1} X^\top y\end{aligned}$$

The ridge estimator $\hat{\beta}^{\text{ridge}}$ has less variance, though it introduces bias.

Strongly positively correlated predictors have similar effects on y.

Ridge Regression

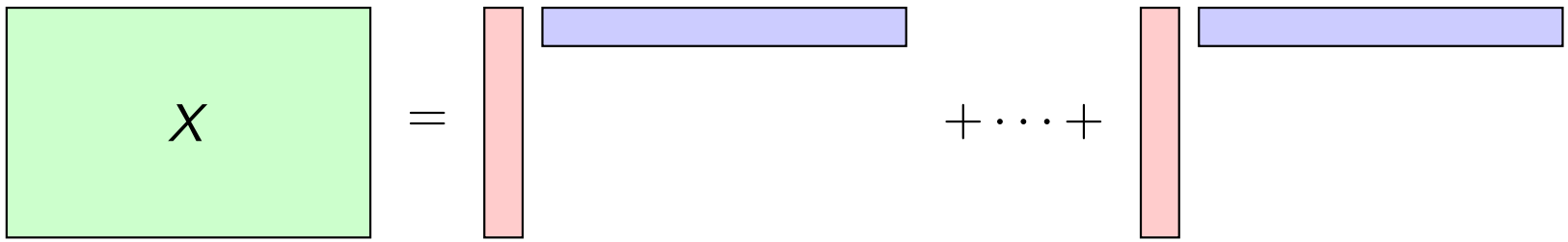


$$X = USV^\top$$

$$\hat{\beta}^{\text{ridge}} = V(S^2 + \lambda I_n)^{-1} SU^\top y$$

[1] Bühlmann. Bernoulli, 13'

Ridge Regression



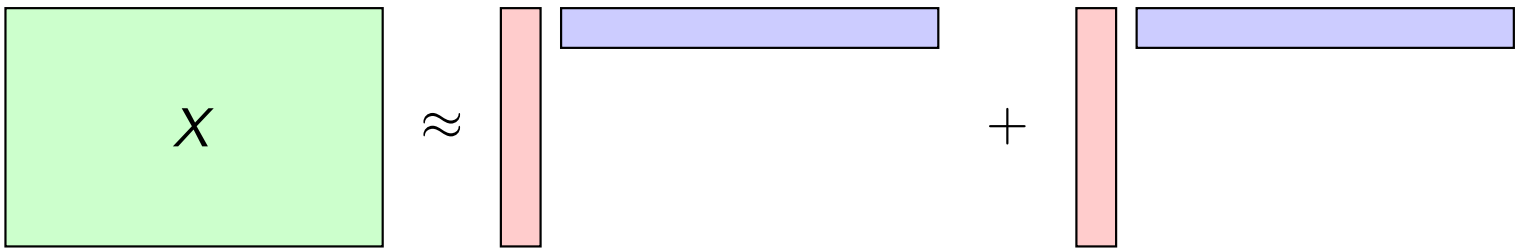
$$X = USV^\intercal$$

$$\hat{\beta}^{\text{ridge}} = V(S^2 + \lambda I_n)^{-1} SU^\intercal y$$

Statistical inference is possible under certain conditions[1].

[1] Bühlmann. Bernoulli, 13'

Principal Component Regression

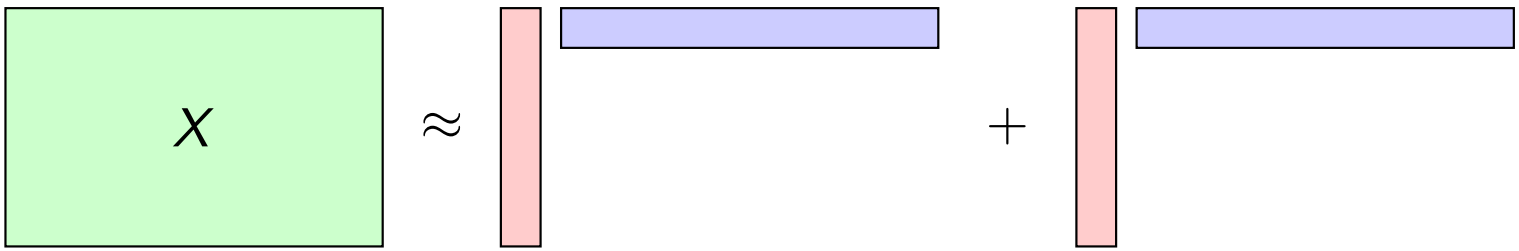


$$X \approx U_2 S_2 V_2^{\mathsf{T}}$$

$$\hat{\beta}^{\mathrm{PCR}} = V_2 S_2^{-1} U_2^{\mathsf{T}} y,$$

U_2, V_2: the top 2 left and right singular vectors.

Principal Component Regression



$$X \approx U_2 S_2 V_2^\intercal$$

$$\hat{\beta}^{\mathrm{PCR}} = V_2 S_2^{-1} U_2^\intercal y,$$

U_2, V_2: the top 2 left and right singular vectors.

PCR projects high-dimensional data into 2D, and estimates association between the 2D summary and y. The PCR estimator $\hat{\beta}^{\mathrm{PCR}}$ is effectively in 2D.

Latent Variable Model

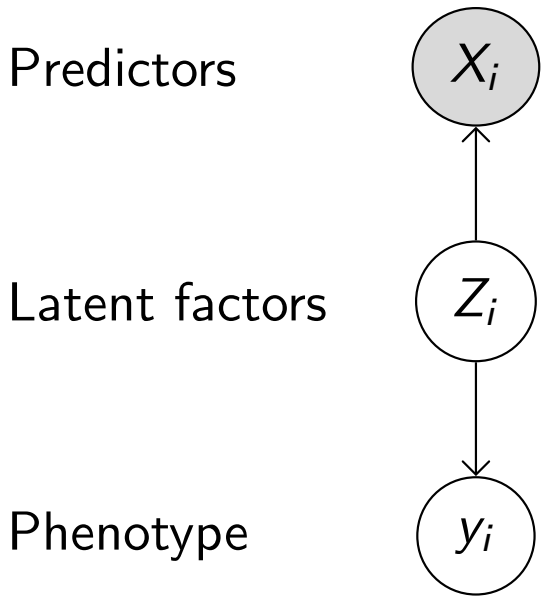


Figure: Illustration of the data generating mechanism underlying PCR. Both predictors and response are driven by the latent factors. PCR extracts latent factors via PCA of X and uses the resulting estimates $\widehat{Z}$ as the new predictor. Regressing y against $\widehat{Z}$ yields the PCR estimate, which typically enjoys smaller variance though it introduces bias.

[2]Wang et al. arXiv, 21'

Latent Variable Model

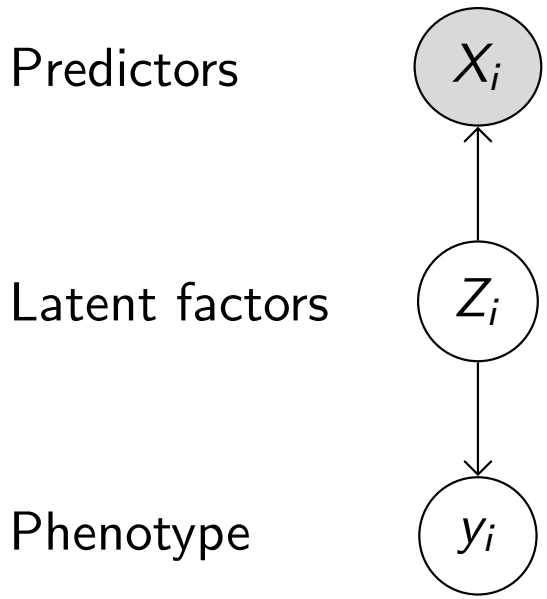


Figure: Illustration of the data generating mechanism underlying PCR. Both predictors and response are driven by the latent factors. PCR extracts latent factors via PCA of X and uses the resulting estimates $\widehat{Z}$ as the new predictor. Regressing y against $\widehat{Z}$ yields the PCR estimate, which typically enjoys smaller variance though it introduces bias.

Statistical inference is also possible under certain conditions[2].

[2] Wang et al. arXiv, 21'

L1 Regularization

L1 penalized regression projects HD data into a LD space defined by a few active variables, and estimates association between the LD summary and y.

The Lasso estimator $\hat{\beta}^{\text{lasso}}$ is effectively in LD.

Statistical inference is also possible under compatibility and sparsity conditions[3].

[3] Zhang and Zhang, JRSSB, 14'; van de Geer et al., AoS, 14'; Cai and Guo, AoS, 17'

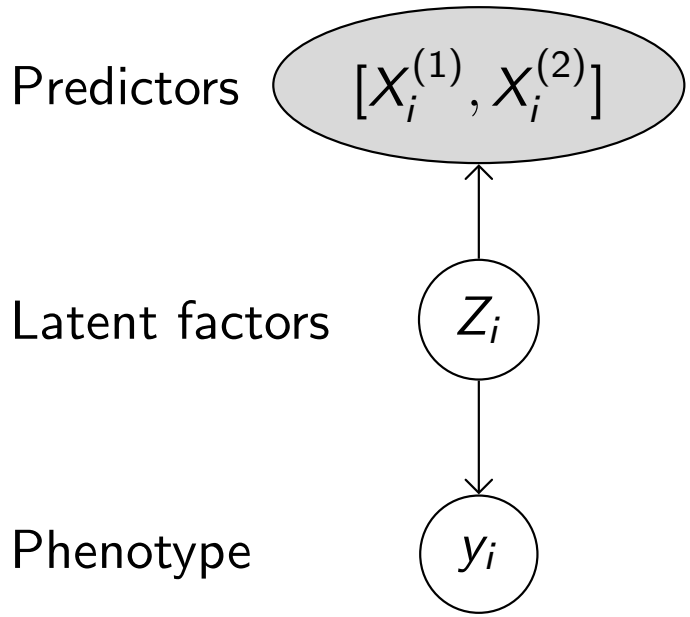
Naive Regression Analysis of Two Datasets

FRED HUTCH
CURES START HERE™

Could apply ridge regression or PCR by regressing y on $[X^{(1)}, X^{(2)}]$, but ...

Naive Regression Analysis of Two Datasets

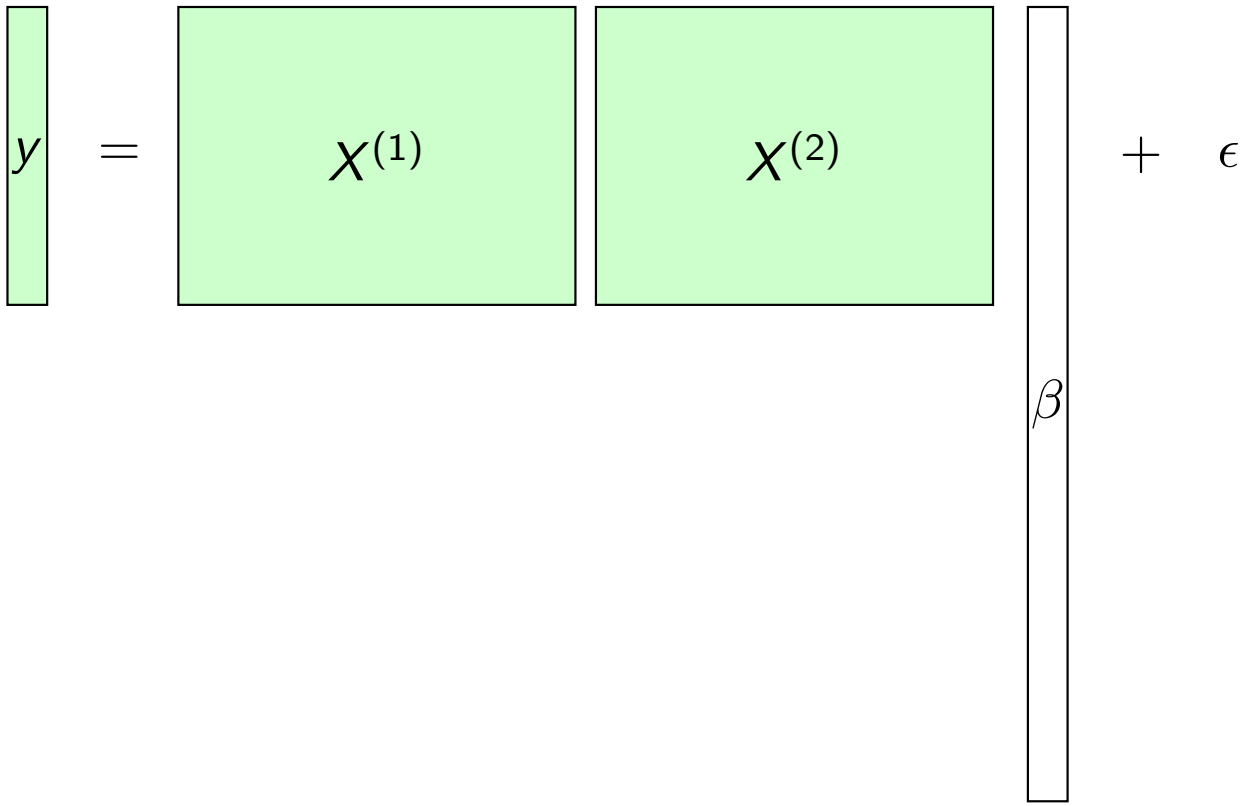
Could apply ridge regression or PCR by regressing y on $[X^{(1)}, X^{(2)}]$, but ...



unable to tell what drives the latent factors: $\mathrm{Var}(X^{(1)})$? $\mathrm{Var}(X^{(2)})$? or $\mathrm{Cov}(X^{(1)}, X^{(2)})$?

Outline

Regression Analysis of Two Datasets



Problem of Interest: which metabolites and bacteria are *jointly* associated with y?

Collaborative Regression

CoRe[4] performs supervised canonical correlation analysis by optimizing

$$b_1\|y - X^{(1)}\theta_1\|^2 + b_2\|y - X^{(2)}\theta_2\|^2 + b_3\|X^{(1)}\theta_1 - X^{(2)}\theta_2\|^2$$

[4] Gross and Tibshirani, Biostatistics, 15'

Collaborative Regression

CoRe[4] performs supervised canonical correlation analysis by optimizing

$$b_1\|y - X^{(1)}\theta_1\|^2 + b_2\|y - X^{(2)}\theta_2\|^2 + b_3\|X^{(1)}\theta_1 - X^{(2)}\theta_2\|^2$$

CoRe uncovers signal that is common to $y, X^{(1)}$ and $X^{(2)}$.

[4] Gross and Tibshirani, Biostatistics, 15'

Collaborative Regression

CoRe[4] performs supervised canonical correlation analysis by optimizing

$$b_1\|y - X^{(1)}\theta_1\|^2 + b_2\|y - X^{(2)}\theta_2\|^2 + b_3\|X^{(1)}\theta_1 - X^{(2)}\theta_2\|^2$$

CoRe uncovers signal that is common to $y, X^{(1)}$ and $X^{(2)}$.

- CoRe only uses the first canonical direction.

[4] Gross and Tibshirani, Biostatistics, 15'

Collaborative Regression

CoRe[4] performs supervised canonical correlation analysis by optimizing

$$b_1\|y - X^{(1)}\theta_1\|^2 + b_2\|y - X^{(2)}\theta_2\|^2 + b_3\|X^{(1)}\theta_1 - X^{(2)}\theta_2\|^2$$

CoRe uncovers signal that is common to $y, X^{(1)}$ and $X^{(2)}$.

- CoRe only uses the first canonical direction.
- Statistical inference is not available for high-dimensional datasets.

[4] Gross and Tibshirani, Biostatistics, 15'

Canonical Variate Regression

CVR[5] considers multiple phenotypes by optimizing

$$b_1 \| Y - X^{(1)} W_1 \theta_1 \|_F^2 + b_2 \| Y - X^{(2)} W_2 \theta_2 \|_F^2 + b_3 \| X^{(1)} W_1 - X^{(2)} W_2 \|_F^2$$

CVR uncovers signal that is common to $Y, X^{(1)}$ and $X^{(2)}$.

[5] Luo et al., Biostatistics, 16'

Canonical Variate Regression

CVR[5] considers multiple phenotypes by optimizing

$$b_1 \| Y - X^{(1)} W_1 \theta_1 \|_F^2 + b_2 \| Y - X^{(2)} W_2 \theta_2 \|_F^2 + b_3 \| X^{(1)} W_1 - X^{(2)} W_2 \|_F^2$$

CVR uncovers signal that is common to $Y, X^{(1)}$ and $X^{(2)}$.

- Statistical inference is not available for high-dimensional datasets.

[5] Luo et al., Biostatistics, 16'

Latent Variable Model

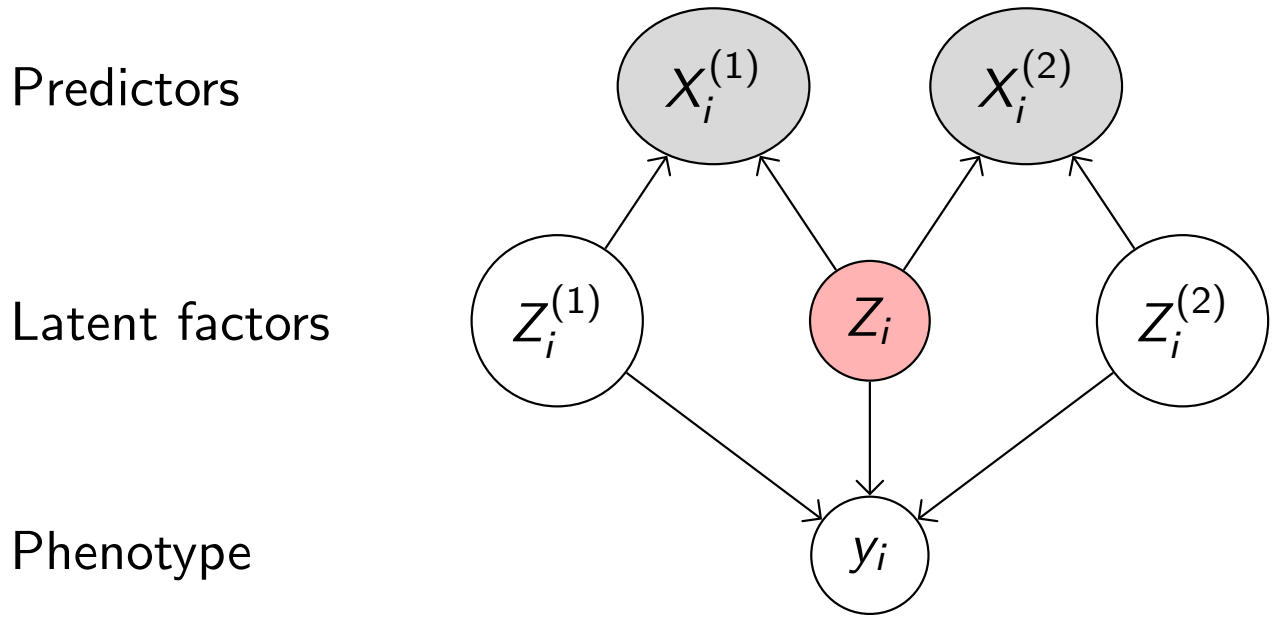


Figure: Illustration of the data generating mechanism underlying JMDR. Two sets of predictors are driven by latent factors that are shared (Z) and dataset-specific ($Z^{(d)}$). All latent factors can contribute to y, though we are mainly interested in the shared information.

Understanding the Latent Factors

Our Hypothesis: Information shared between datasets may be more causal.

Understanding the Latent Factors

Our Hypothesis: Information shared between datasets may be more causal.

Shared factors Z_i capture joint variation that is more likely due to the underlying biological mechanism.

Understanding the Latent Factors

Our Hypothesis: Information shared between datasets may be more causal.

Shared factors Z_i capture joint variation that is more likely due to the underlying biological mechanism.

Individual factors $Z_i^{(1)}$ and $Z_i^{(2)}$ capture dataset-specific variation due to, e.g. batch effects, technical artifacts, host factors, etc.

Understanding the Latent Factors

FRED HUTCH
CURES START HERE™

Our Hypothesis: Information shared between datasets may be more causal.

Shared factors Z_i capture joint variation that is more likely due to the underlying biological mechanism.

Individual factors $Z_i^{(1)}$ and $Z_i^{(2)}$ capture dataset-specific variation due to, e.g. batch effects, technical artifacts, host factors, etc.

Separating individual from joint factors may improve estimation.

Understanding the Latent Factors

Add a slide on biological interpretations of the latent factors

Joint Matrix Decomposition Regression

$$
\begin{aligned}
X^{(1)} &= Z W_1^\top + Z^{(1)} V_1^\top + E^{(1)} \\
X^{(2)} &= Z W_2^\top + Z^{(2)} V_2^\top + E^{(2)} \\
y &= X^{(1)} \beta_1^* + X^{(2)} \beta_2^* + \epsilon
\end{aligned}
$$

Joint Matrix Decomposition Regression

$$
\begin{aligned}
X^{(1)} &= Z W_1^\top + Z^{(1)} V_1^\top + E^{(1)} \\
X^{(2)} &= Z W_2^\top + Z^{(2)} V_2^\top + E^{(2)} \\
y &= X^{(1)} \beta_1^* + X^{(2)} \beta_2^* + \epsilon
\end{aligned}
$$

- W_1 and W_2 are low-rank loading matrices for *joint variation*.
- V_1 and V_2 are low-rank loading matrices for dataset-specific variation.

Joint Matrix Decomposition Regression

$$
\begin{aligned}
X^{(1)} &= Z W_1^\top + Z^{(1)} V_1^\top + E^{(1)} \\
X^{(2)} &= Z W_2^\top + Z^{(2)} V_2^\top + E^{(2)} \\
y &= X^{(1)} \beta_1^* + X^{(2)} \beta_2^* + \epsilon
\end{aligned}
$$

- W_1 and W_2 are low-rank loading matrices for *joint variation*.
- V_1 and V_2 are low-rank loading matrices for dataset-specific variation.
- Latent factors are independent.
- The errors ϵ and $E^{(d)}$ have mean zero, and are independent of each other.

Joint Matrix Decomposition Regression

$$
\begin{aligned}
X^{(1)} &= Z W_1^\top + Z^{(1)} V_1^\top + E^{(1)} \\
X^{(2)} &= Z W_2^\top + Z^{(2)} V_2^\top + E^{(2)} \\
y &= X^{(1)} \beta_1^* + X^{(2)} \beta_2^* + \epsilon
\end{aligned}
$$

- W_1 and W_2 are low-rank loading matrices for *joint variation*.
- V_1 and V_2 are low-rank loading matrices for dataset-specific variation.
- Latent factors are independent.
- The errors ϵ and $E^{(d)}$ have mean zero, and are independent of each other.
- Not interested in β_1^* and $\beta_2^* \rightarrow$ Need to define target parameter

Joint Matrix Decomposition Regression

Letting $W^\intercal = [W_1^\intercal, W_2^\intercal]$,

$$y \approx \left(Z W^\intercal + Z^{(1)} V_1^\intercal + Z^{(2)} V_2^\intercal \right) \begin{pmatrix} \beta_1^* \\ \beta_2^* \end{pmatrix}$$

- The quantity

$$\alpha = W^\intercal \begin{pmatrix} \beta_1^* \\ \beta_2^* \end{pmatrix} \in \mathbb{R}^k$$

reflects the association between the *joint variation* and y.

Joint Matrix Decomposition Regression

Letting $W^\intercal = [W_1^\intercal, W_2^\intercal]$,

$$y \approx \left(Z W^\intercal + Z^{(1)} V_1^\intercal + Z^{(2)} V_2^\intercal \right) \begin{pmatrix} \beta_1^* \\ \beta_2^* \end{pmatrix}$$

- Each coordinate in

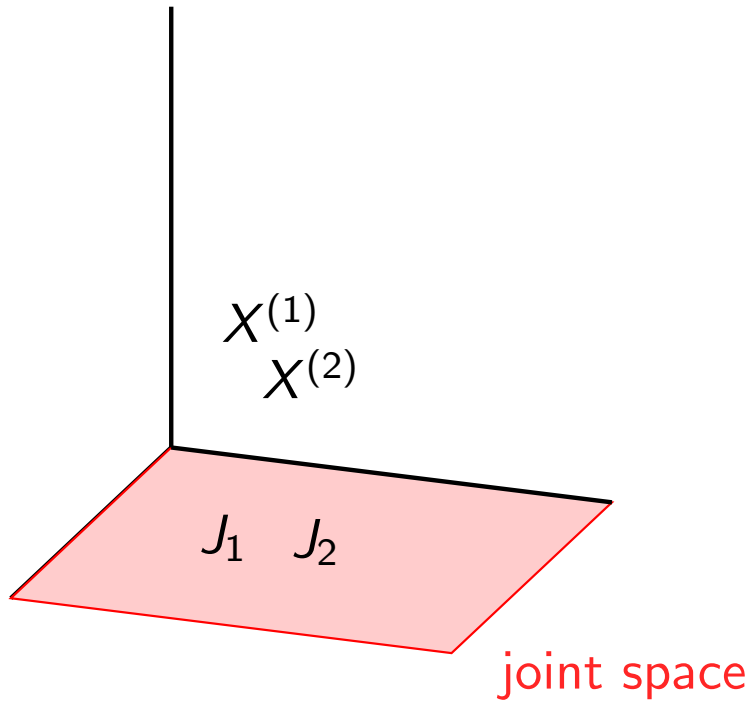
$$\theta = W W^\intercal \begin{pmatrix} \beta_1^* \\ \beta_2^* \end{pmatrix} \in \mathbb{R}^{p_1 + p_2}$$

  reflects the association between each variable and y through *joint variation*.
- We use θ to determine which metabolites and bacteria are *jointly* associated with y.

Model Identifiability

- α is identifiable up to rotations
- θ is invariant to rotations

Understanding Joint Association



$X^{(1)} \approx J_1, X^{(2)} \approx J_2$ in the joint space, but J_1 and J_2 are special low-rank matrices

θ is effectively low-dimensional

Comparison to Linear Regulatory Module

LRM[6] considers the relationship $X_i^{(1)} \to X_i^{(2)} \to y_i$

$$\begin{aligned} X^{(2)} W_2 &= X^{(1)} W_1 + E \\ y &= X^{(2)} W_2 \beta_0 + \mathrm{res}_1 \cdot \beta_1 + \mathrm{res}_2 \cdot \beta_2 + \epsilon \end{aligned}$$

Columns in W_1 and W_2 are assumed to be sparse, reflecting a small number of active regulation.

[6] Zhu et al. Biostatistics, 16'

Comparison to Linear Regulatory Module

LRM[6] considers the relationship $X_i^{(1)} \to X_i^{(2)} \to y_i$

$$X^{(2)} W_2 = X^{(1)} W_1 + E$$
$$y = X^{(2)} W_2 \beta_0 + \text{res}_1 \cdot \beta_1 + \text{res}_2 \cdot \beta_2 + \epsilon$$

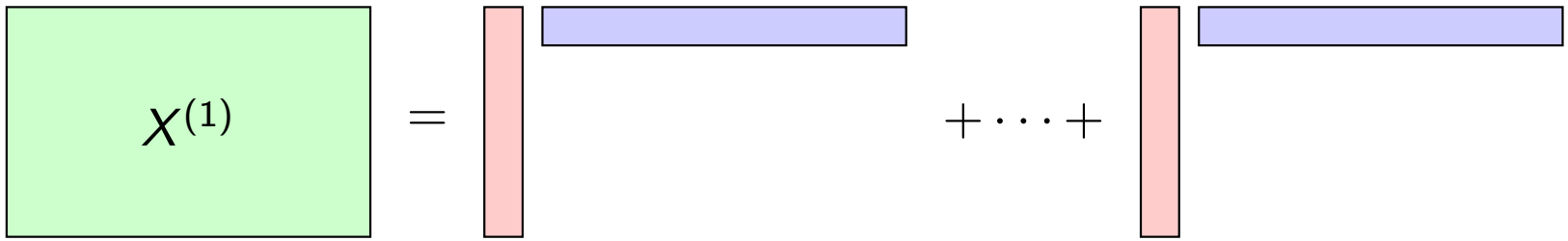
Columns in W_1 and W_2 are assumed to be sparse, reflecting a small number of active regulation.

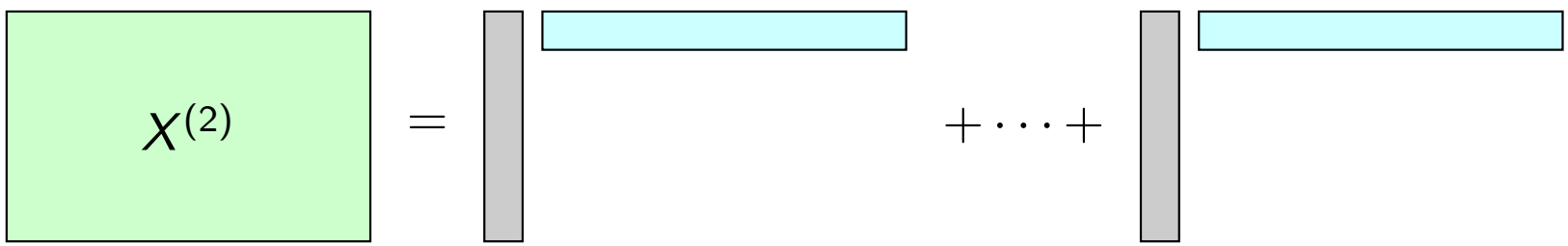
LRM does not provide statistical inference for selected variables.

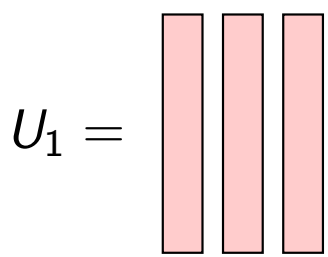
[6] Zhu et al. Biostatistics, 16'

How JMDR works (briefly)

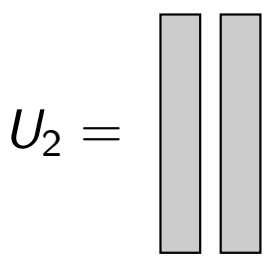
JMDR Algorithm







n sample scores from $X^{(1)}$



n sample scores from $X^{(2)}$

JMDR Algorithm

We use $U_1^T U_2$ to project $X^{(1)}$ and $X^{(2)}$ onto a *joint subspace*.

JMDR Algorithm

We use $U_1^\top U_2$ to project $X^{(1)}$ and $X^{(2)}$ onto a *joint subspace*.

- Decompose $U_1^\top U_2 = RDQ^\top$ and define a projection

$$J_1 = (U_1 R R^\top U_1^\top) X^{(1)}$$
$$J_2 = (U_2 Q Q^\top U_2^\top) X^{(2)}$$

JMDR Algorithm

We use $U_1^T U_2$ to project $X^{(1)}$ and $X^{(2)}$ onto a *joint subspace*.

- Decompose $U_1^T U_2 = RDQ^T$ and define a projection

$$J_1 = (U_1 R R^T U_1^T) X^{(1)}$$
$$J_2 = (U_2 Q Q^T U_2^T) X^{(2)}$$

Remark: Use principal angles to determine the dimensions of $U_1^T U_2$ that can be considered *joint*.

JMDR Algorithm

We use $U_1^\top U_2$ to project $X^{(1)}$ and $X^{(2)}$ onto a *joint subspace*.

- Decompose $U_1^\top U_2 = RDQ^\top$ and define a projection

$$J_1 = (U_1 R R^\top U_1^\top) X^{(1)}$$
$$J_2 = (U_2 Q Q^\top U_2^\top) X^{(2)}$$

  Remark: Use principal angles to determine the dimensions of $U_1^\top U_2$ that can be considered *joint*.

- Estimate the latent factors by $[J_1, J_2] = \widehat{Z}\widehat{\Lambda}\widehat{W}^\top$.

JMDR Algorithm

We use $U_1^\top U_2$ to project $X^{(1)}$ and $X^{(2)}$ onto a *joint subspace*.

- Decompose $U_1^\top U_2 = RDQ^\top$ and define a projection

$$J_1 = (U_1 R R^\top U_1^\top) X^{(1)}$$
$$J_2 = (U_2 Q Q^\top U_2^\top) X^{(2)}$$

  Remark: Use principal angles to determine the dimensions of $U_1^\top U_2$ that can be considered *joint*.

- Estimate the latent factors by $[J_1, J_2] = \widehat{Z}\widehat{\Lambda}\widehat{W}^\top$.
- Regress y on $\widehat{Z}$.

Statistical Inference

Global test:

$$H_0 : \alpha = 0$$

Element-wise test:

$$H_{0,j} : \theta_j = 0$$

JMDR inference accounts for subspace recovery error.

Results

BactoCARB Study

$X^{(1)}$ = 144 metabolites & bile acids (blood)

$X^{(2)}$ = 134 microbial genus abundances (stool)

y = enterolactone 'ENL' (urine)

BactoCARB Study

$X^{(1)}$ = 144 metabolites & bile acids (blood)

$X^{(2)}$ = 134 microbial genus abundances (stool)

y = enterolactone 'ENL' (urine)

Rank of joint space: 4

Rank of individual space: 11 for metabolites and 4 for microbiome

BactoCARB Study

$X^{(1)}$ = 144 metabolites & bile acids (blood)

$X^{(2)}$ = 134 microbial genus abundances (stool)

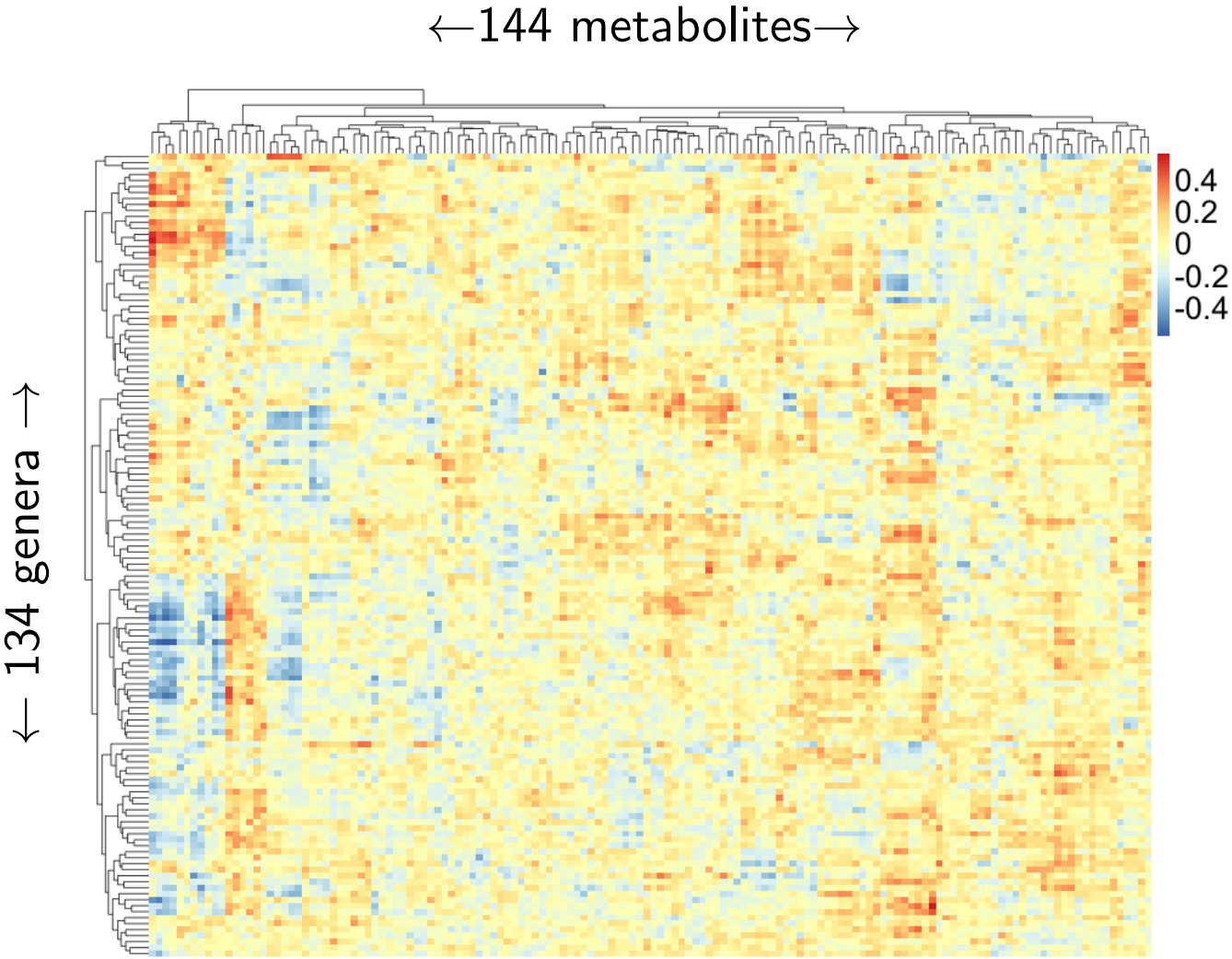
y = enterolactone 'ENL' (urine)

Rank of joint space: 4

Rank of individual space: 11 for metabolites and 4 for microbiome

A global test for association between ENL and *joint* signal:
p-value $< 10^{-5}$

Correlations btw Metabolites and Microbes



Correlations in the Joint Space

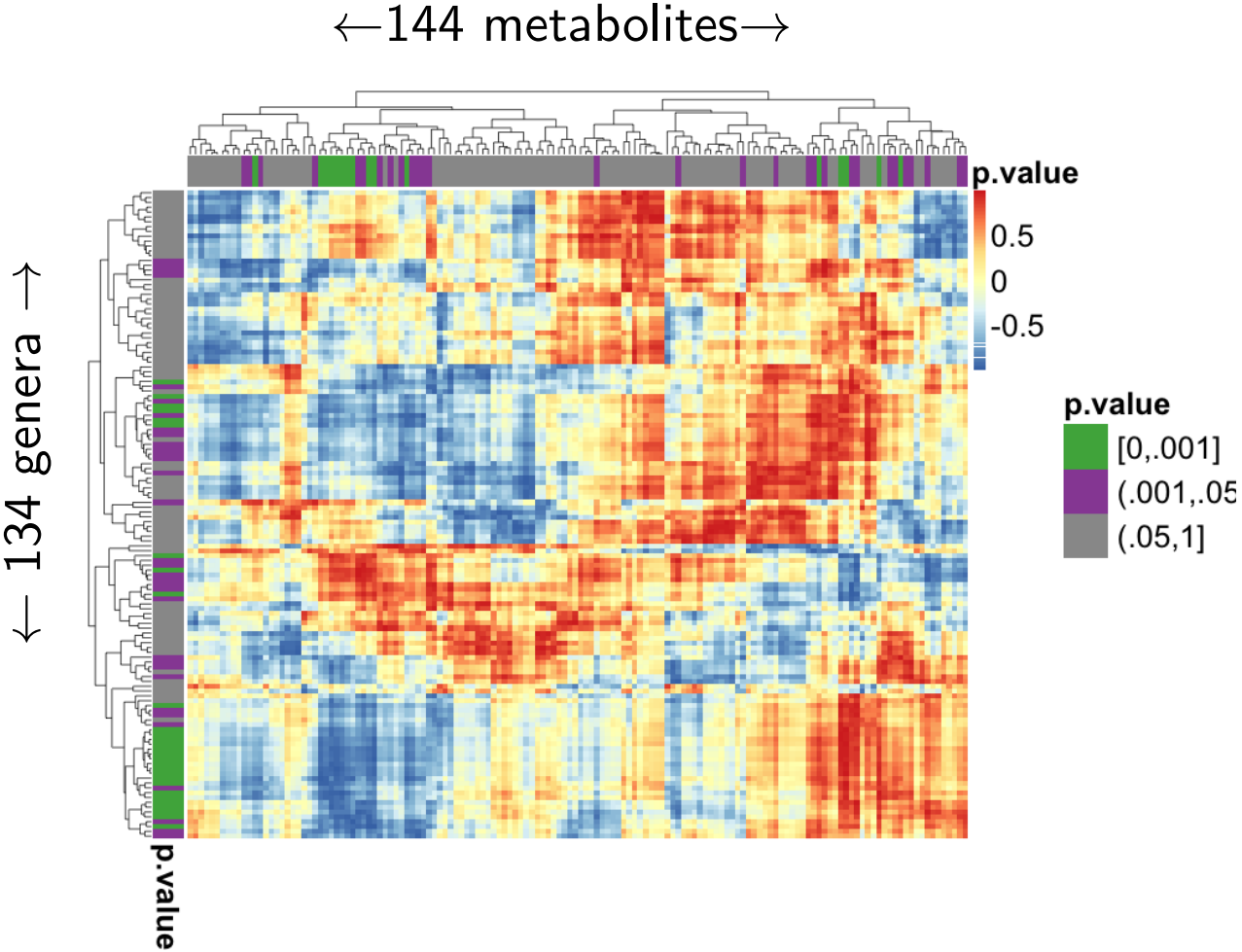
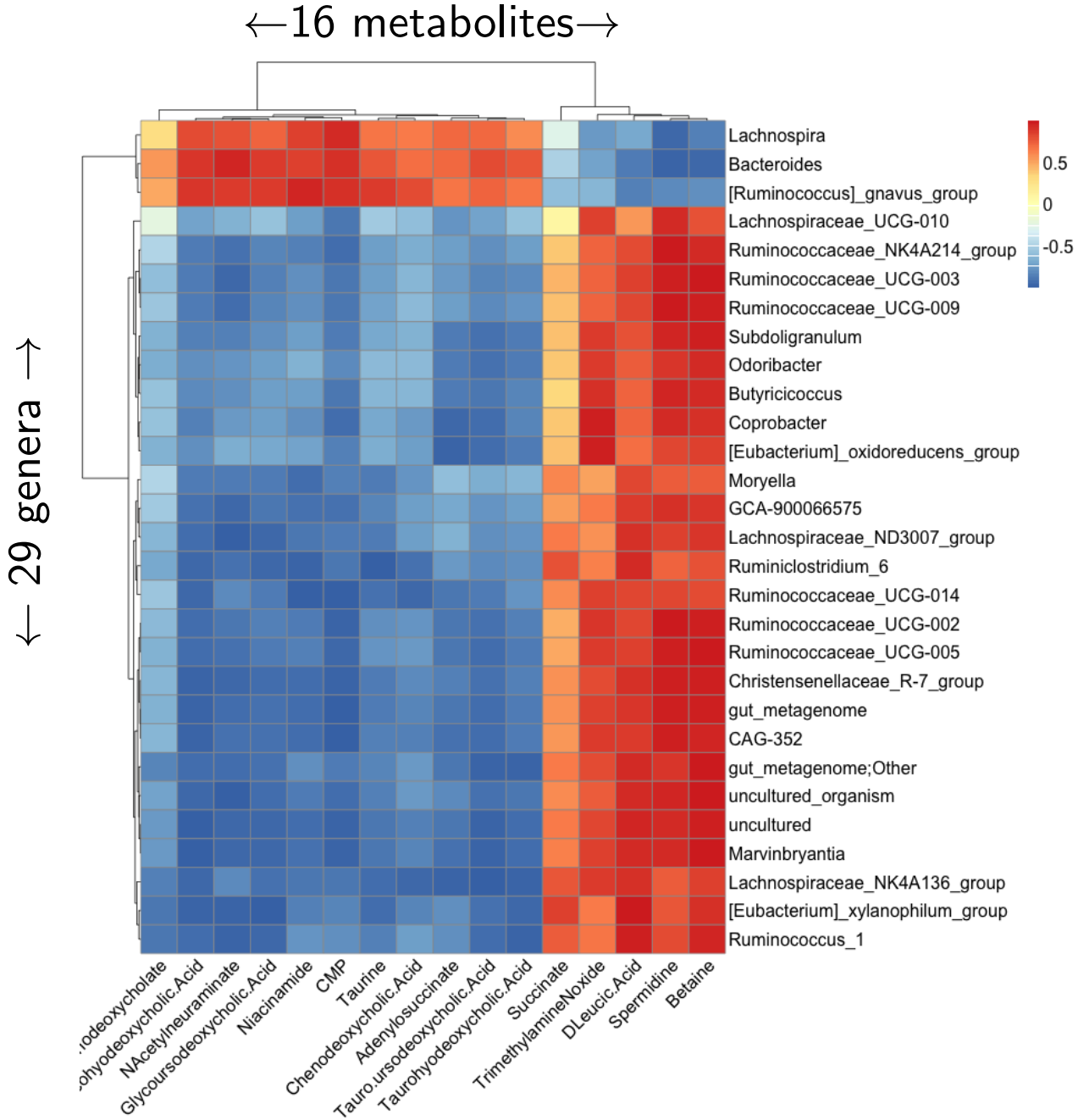


Figure: Pairwise correlations in the 4-dim joint space

p-values are from element-wise test in JMDR model

$$\text{ENL} \sim \text{metabolites} + \text{genera}$$

Significant Variables



Summary

- Given 2 (or more) datasets, strip off individual/separate variation and represent data using only shared/joint variation.

Summary

FRED HUTCH
CURES START HERE™

- Given 2 (or more) datasets, strip off individual/separate variation and represent data using only shared/joint variation.
- *Joint Matrix Decomposition Regression* identifies variables in each dataset that are jointly associated with a phenotype.

Summary

FRED HUTCH
CURES START HERE™

- Given 2 (or more) datasets, strip off individual/separate variation and represent data using only shared/joint variation.
- *Joint Matrix Decomposition Regression* identifies variables in each dataset that are jointly associated with a phenotype.
- JMDR also works for dichotomous outcomes.

Dog Aging Project

FRED HUTCH
CURES START HERE™

An epidemiological cohort $n \geq 1000$ of companion dogs aimed at identifying the genetic and environmental factors of aging

- $X^{(1)}$ = metabolites (blood)
- $X^{(2)}$ = microbial species/gene abundances (stool)
- $X^{(3)}$ = epigenetic markers (PBMCs)
- y = cognitive score, aging score

Problem of Interest: which metabolites, bacteria, and epigenetic markers are *jointly* associated with y?

Future Work

FRED HUTCH
CURES START HERE™

- Extend to more than 2 datasets and partially shared joint signals
- Extend from linear to nonlinear models
- Bayesian modeling
- Applications to uncovering biomarkers related to cognition and aging (Dog Aging Project, HCHS/SOL)

Aknowledgement

FRED HUTCH
CURES START HERE™

Yue Wang @ASU

Ali Shojaie @UW

Tim Randolph @Fred Hutch

Thank You!

https://drjingma.com