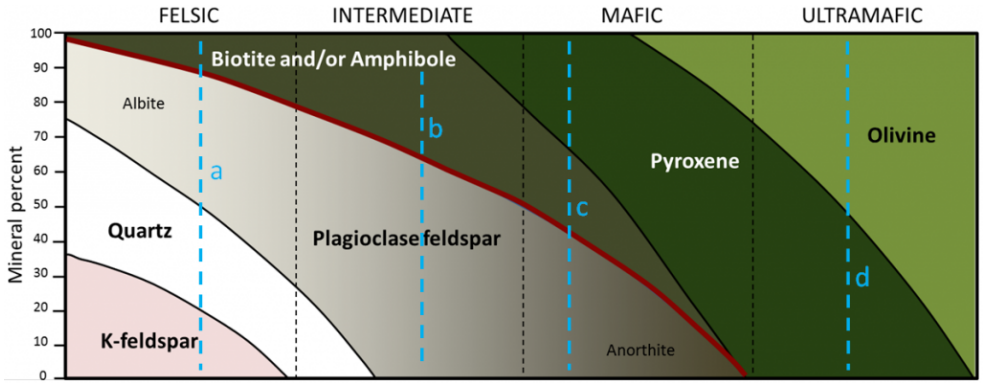
Predictive Modeling of Compositional Data with Supervised Log-Ratios

Jing Ma

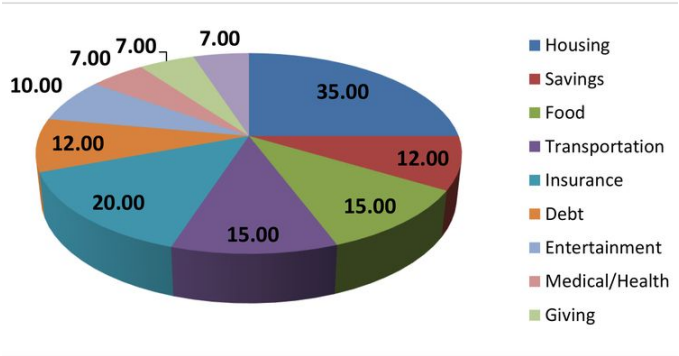
Division of Public Health Sciences
Fred Hutchinson Cancer Center

1 July 2022
ICSA China Conference

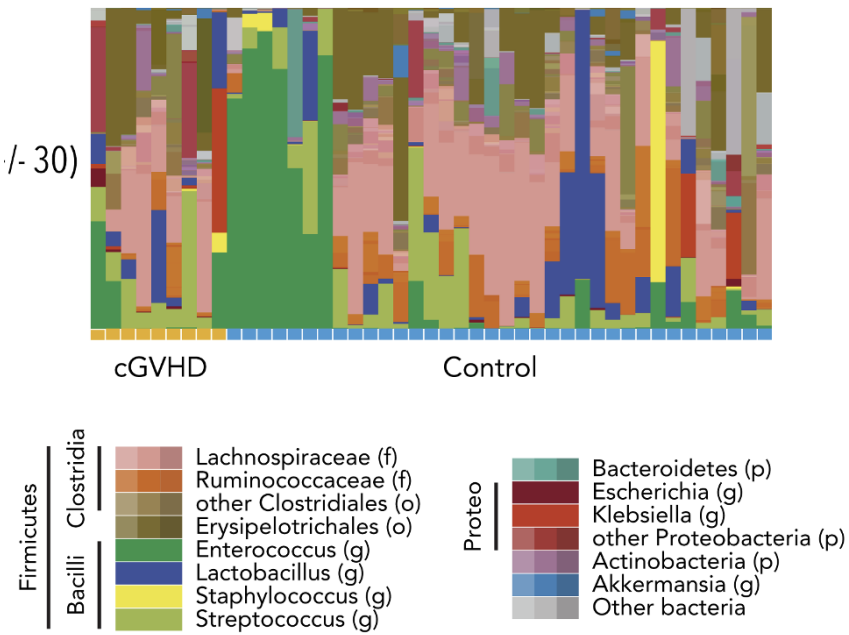
Compositional Data are Everywhere



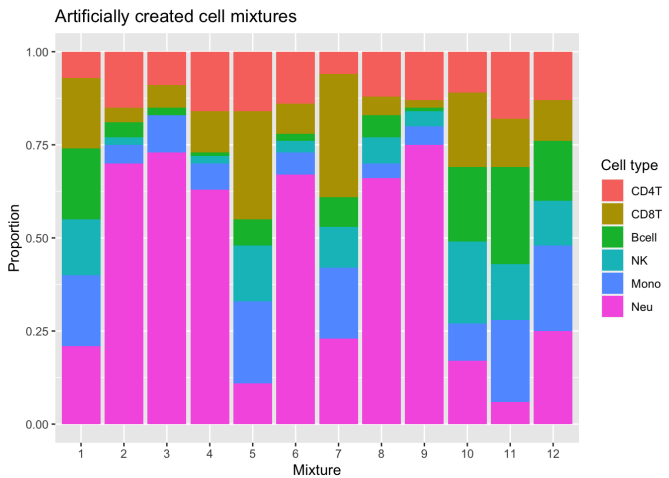
Geology



Sociology



Microbiome: Markey et al., Blood, 20'



Single cell transcriptomics

Compositional Data

A vector $X = (X_1, \ldots, X_p)$ representing proportions of some whole is subject to the constraint

$$X_1 + \cdots + X_p = 1$$

Compositional Data

A vector $X = (X_1, \ldots, X_p)$ representing proportions of some whole is subject to the constraint

$$X_1 + \cdots + X_p = 1$$

Predictive modeling

- Predictors $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,p})$: compositional
- Outcome y_i: continuous or binary

Compositional Data

A vector $X = (X_1, \ldots, X_p)$ representing proportions of some whole is subject to the constraint

$$X_1 + \cdots + X_p = 1$$

Predictive modeling

- Predictors $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,p})$: compositional
- Outcome y_i: continuous or binary

Scientific question:

- Define biomarker(s) using a small set of variables that predict disease risk

Compositional Data

A vector $X = (X_1, \ldots, X_p)$ representing proportions of some whole is subject to the constraint

$$X_1 + \cdots + X_p = 1$$

Predictive modeling

- Predictors $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,p})$: compositional
- Outcome y_i: continuous or binary

Scientific question:

- Define biomarker(s) using a small set of variables that predict disease risk

Challenges:

- The unit-sum constraint makes it difficult to interpret the effect of predictors on the response

Log-Contrast Models

Additive log-ratio transform:

$$\text{alr}(X) = (\log \frac{X_1}{X_p}, \ldots, \log \frac{X_{p-1}}{X_p})$$

Log-contrast regression:

$$\mathbb{E}[y_i \mid \boldsymbol{x}_i] = (\boldsymbol{\theta}^{\text{alr}})^\intercal \text{alr}(\boldsymbol{x}_i)$$

[1] Lin et al., Biometrika, 14'; Shi et al., AOAS, 16'

[2] Wang and Zhao, AOAS, 17'; Bien et al., Scientific Reports, 21'

Log-Contrast Models

Additive log-ratio transform:

$$\text{alr}(X) = (\log \frac{X_1}{X_p}, \ldots, \log \frac{X_{p-1}}{X_p})$$

Log-contrast regression:

$$\mathbb{E}[y_i \mid \boldsymbol{x}_i] = (\boldsymbol{\theta}^{\text{alr}})^\mathsf{T} \text{alr}(\boldsymbol{x}_i) = \boldsymbol{\beta}^\mathsf{T} \log(\boldsymbol{x}_i),$$
$$\text{subject to } \boldsymbol{\beta}^\mathsf{T} 1 = 0.$$

[1] Lin et al., Biometrika, 14'; Shi et al., AOAS, 16'

[2] Wang and Zhao, AOAS, 17'; Bien et al., Scientific Reports, 21'

Log-Contrast Models

Additive log-ratio transform:

$$\text{alr}(X) = (\log \frac{X_1}{X_p}, \ldots, \log \frac{X_{p-1}}{X_p})$$

Log-contrast regression:

$$\mathbb{E}[y_i \mid \boldsymbol{x}_i] = (\boldsymbol{\theta}^{\text{alr}})^\top \text{alr}(\boldsymbol{x}_i) = \boldsymbol{\beta}^\top \log(\boldsymbol{x}_i), \quad \text{subject to } \boldsymbol{\beta}^\top 1 = 0.$$

High-dimensional extensions: compositional Lasso[1], tree-guided compositional Lasso[2]

[1] Lin et al., Biometrika, 14'; Shi et al., AOAS, 16'

[2] Wang and Zhao, AOAS, 17'; Bien et al., Scientific Reports, 21'

Log-Contrast Models

Additive log-ratio transform:

$$\text{alr}(X) = (\log \frac{X_1}{X_p}, \ldots, \log \frac{X_{p-1}}{X_p})$$

Log-contrast regression:

$$\mathbb{E}[y_i \mid \boldsymbol{x}_i] = (\boldsymbol{\theta}^{\text{alr}})^\top \text{alr}(\boldsymbol{x}_i) = \boldsymbol{\beta}^\top \log(\boldsymbol{x}_i), \quad \text{subject to } \boldsymbol{\beta}^\top 1 = 0.$$

High-dimensional extensions: compositional Lasso[1], tree-guided compositional Lasso[2]

Limitation: alr coefficients need to be interpreted w.r.t. a reference variable, while constrained regression suffers from prediction accuracy.

[1] Lin et al., Biometrika, 14'; Shi et al., AOAS, 16'

[2] Wang and Zhao, AOAS, 17'; Bien et al., Scientific Reports, 21'

Log-Ratio Regression

Pairwise log-ratios[3]

$$\mathbb{E}[y_i \mid \boldsymbol{x}_i] = \sum_{1 \le j < k \le p} \theta^{\mathrm{plr}}_{j,k} \log \frac{x_{i,j}}{x_{i,k}}$$

The log-contrast coefficient $\boldsymbol{\beta} = C^\top \boldsymbol{\theta}^{\mathrm{plr}}$ where for $p = 4$

$$C^\top = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & 1 & 0 \\ 0 & -1 & 0 & -1 & 0 & 1 \\ 0 & 0 & -1 & 0 & -1 & -1 \end{pmatrix}.$$

[3] Bates and Tibshirani, Biometrics, 19'

Log-Ratio Regression

Pairwise log-ratios[3]

$$\mathbb{E}[y_i \mid \boldsymbol{x}_i] = \sum_{1 \le j < k \le p} \theta^{\text{plr}}_{j,k} \log \frac{x_{i,j}}{x_{i,k}}$$

The log-contrast coefficient $\boldsymbol{\beta} = C^\top \boldsymbol{\theta}^{\text{plr}}$ where for $p = 4$

$$C^\top = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & 1 & 0 \\ 0 & -1 & 0 & -1 & 0 & 1 \\ 0 & 0 & -1 & 0 & -1 & -1 \end{pmatrix}.$$

Limitation: this model is not identifiable due to co-linearity of predictors, e.g.

$$\log \frac{X_1}{X_2}, \quad \log \frac{X_1}{X_3}, \quad \log \frac{X_2}{X_3}$$

[3] Bates and Tibshirani, Biometrics, 19'

Balance Regression

Balance is the log-ratio between two geometric means

$$B(X; I_+, I_-) = \log \frac{g(X_{I_+})}{g(X_{I_-})} = \frac{\sum_{j \in I_+} \log X_j}{|I_+|} - \frac{\sum_{j \in I_-} \log X_j}{|I_-|}$$

Balance regression searches for the best subsets I_+ and I_-:

$$\mathbb{E}[y_i \mid \boldsymbol{x}_i] = \theta_0 + \theta_1 B(\boldsymbol{x}_i; I_+, I_-)$$

[4] Rivera-Pinto, mSystems, 18'

Balance Regression

Balance is the log-ratio between two geometric means

$$B(X; I_+, I_-) = \log \frac{g(X_{I_+})}{g(X_{I_-})} = \frac{\sum_{j \in I_+} \log X_j}{|I_+|} - \frac{\sum_{j \in I_-} \log X_j}{|I_-|}$$

Balance regression searches for the best subsets I_+ and I_-:

$$\mathbb{E}[y_i \mid \boldsymbol{x}_i] = \theta_0 + \theta_1 B(\boldsymbol{x}_i; I_+, I_-)$$

selbal[4] uses greedy search to find the best subsets by adding one variable at a time from the best k-part balances for $k \geq 2$.

[4] Rivera-Pinto, mSystems, 18'

Balance Regression

Balance is the log-ratio between two geometric means

$$B(X; I_+, I_-) = \log \frac{g(X_{I_+})}{g(X_{I_-})} = \frac{\sum_{j \in I_+} \log X_j}{|I_+|} - \frac{\sum_{j \in I_-} \log X_j}{|I_-|}$$

Balance regression searches for the best subsets I_+ and I_-:

$$\mathbb{E}[y_i \mid \boldsymbol{x}_i] = \theta_0 + \theta_1 B(\boldsymbol{x}_i; I_+, I_-)$$

selbal[4] uses greedy search to find the best subsets by adding one variable at a time from the best k-part balances for $k \geq 2$.

selbal prioritizes sparse models, but exhaustive search is time consuming.

[4] Rivera-Pinto, mSystems, 18'

Balance Regression

CoDaCoRe[5] uses continuous relaxation to find the best subsets. For a vector of assignment weights $\boldsymbol{w}$, let

$$\widetilde{\boldsymbol{w}} = \frac{2}{1 + \exp(-\boldsymbol{w})} - 1.$$

Let $\widetilde{\boldsymbol{w}}^+ = \mathrm{ReLU}(\widetilde{\boldsymbol{w}})$ and $\widetilde{\boldsymbol{w}}^- = \mathrm{ReLU}(-\widetilde{\boldsymbol{w}})$. The relaxed balance is

$$\widetilde{B}(X; \boldsymbol{w}) = \frac{\sum_j \widetilde{w}_j^+ \log X_j}{\sum_j \widetilde{w}_j^+} - \frac{\sum_j \widetilde{w}_j^- \log X_j}{\sum_j \widetilde{w}_j^-}$$

[5] Gordon-Rodriguez et al., Bioinformatics, 22'

Balance Regression

CoDaCoRe[5] uses continuous relaxation to find the best subsets. For a vector of assignment weights $\boldsymbol{w}$, let

$$\widetilde{\boldsymbol{w}} = \frac{2}{1 + \exp(-\boldsymbol{w})} - 1.$$

Let $\widetilde{\boldsymbol{w}}^+ = \mathrm{ReLU}(\widetilde{\boldsymbol{w}})$ and $\widetilde{\boldsymbol{w}}^- = \mathrm{ReLU}(-\widetilde{\boldsymbol{w}})$. The relaxed balance is

$$\widetilde{B}(X; \boldsymbol{w}) = \frac{\sum_j \widetilde{w}_j^+ \log X_j}{\sum_j \widetilde{w}_j^+} - \frac{\sum_j \widetilde{w}_j^- \log X_j}{\sum_j \widetilde{w}_j^-}$$

Hard thresholding

$$\hat{I}_+ = \{j : \widetilde{w}_j^+ > \tau\}, \quad \hat{I}_- = \{j : \widetilde{w}_j^- < -\tau\}$$

[5] Gordon-Rodriguez et al., Bioinformatics, 22'

Balance Regression

CoDaCoRe[5] uses continuous relaxation to find the best subsets. For a vector of assignment weights $\boldsymbol{w}$, let

$$\widetilde{\boldsymbol{w}} = \frac{2}{1 + \exp(-\boldsymbol{w})} - 1.$$

Let $\widetilde{\boldsymbol{w}}^+ = \mathrm{ReLU}(\widetilde{\boldsymbol{w}})$ and $\widetilde{\boldsymbol{w}}^- = \mathrm{ReLU}(-\widetilde{\boldsymbol{w}})$. The relaxed balance is

$$\widetilde{B}(X; \boldsymbol{w}) = \frac{\sum_j \widetilde{w}_j^+ \log X_j}{\sum_j \widetilde{w}_j^+} - \frac{\sum_j \widetilde{w}_j^- \log X_j}{\sum_j \widetilde{w}_j^-}$$

Hard thresholding

$$\hat{I}_+ = \{j : \widetilde{w}_j^+ > \tau\}, \quad \hat{I}_- = \{j : \widetilde{w}_j^- < -\tau\}$$

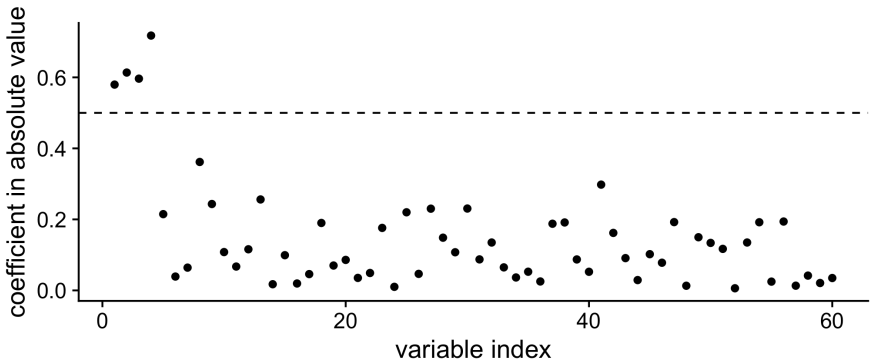
CoDaCoRe is efficient, but tends to select too many variables.

[5] Gordon-Rodriguez et al., Bioinformatics, 22'

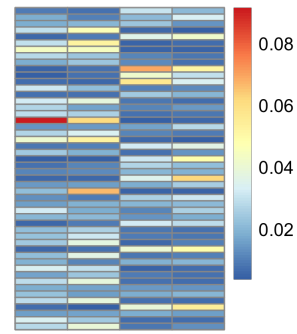
Our Framework: Supervised Log-Ratios

Input: $(\boldsymbol{x}_i, y_i)$ for $i = 1, \ldots, n$.

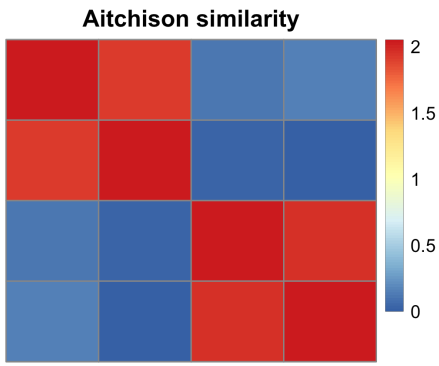
Step 1: Screen



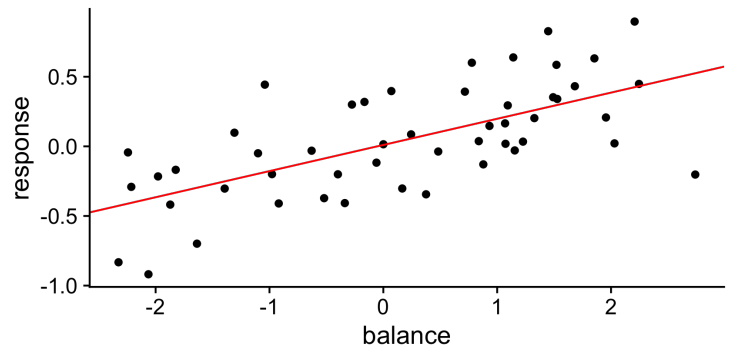
Step 2: Reduce



Step 3: Cluster



Step 4: Predict



Output: two subsets I_+, I_- of variables for defining the balance.

Step 1: Screen

Let $\boldsymbol{z}_i$ denote the clr-transformed version of $\boldsymbol{x}_i$, where

$$Z = \left(\log \frac{X_1}{g(X)}, \ldots, \log \frac{X_p}{g(X)} \right)^\top$$

Let $\boldsymbol{z}^{(j)}$ denote the vector of observations from the j-th variable. Variables are screened by thresholding their univariate effect on $\boldsymbol{y}$:

$$\left| \frac{(\boldsymbol{y} - \bar{\boldsymbol{y}})^\top (\boldsymbol{z}^{(j)} - \bar{\boldsymbol{z}}^{(j)})}{\| \boldsymbol{z}^{(j)} - \bar{\boldsymbol{z}}^{(j)} \|^2} \right| > \tau$$

The threshold τ is chosen by cross-validation.

Step 3: Cluster

Let C_τ be the collection of indices containing selected variables.

The Aitchison variation on the reduced data matrix is defined as

$$\hat{A}(\tau)_{j,k} = \frac{1}{n}\sum_{i=1}^{n}(\log\frac{x_{i,j}}{x_{i,k}} - \frac{1}{n}\sum_{i'=1}^{n}\log\frac{x_{i',j}}{x_{i',k}})^2, \quad j,k \in C_\tau.$$

The Aitchison similarity is

$$\hat{S}(\tau)_{j,k} = \max_{j',k'}\left\{\hat{A}(\tau)_{j',k'}\right\} - \hat{A}(\tau)_{j,k}, \quad j,k \in C_\tau.$$

Clustering returns two subsets of variables for defining the balance.

A Latent Variable Model

$$\log \frac{X_j}{X_p} = \alpha_{0,j} + \alpha_{1,j} U + \epsilon_j, \quad j = \{1, \ldots, p\} \backslash \{p\} \tag{1}$$

$$y = \beta_0 + \beta_1 U + \varepsilon, \tag{2}$$

where for $c_1, c_2 > 0$ the coefficients $\alpha_{1,j}$ satisfy

$$\begin{aligned} \alpha_{1,j} &= 0, \quad j \notin I_+ \cup I_-, \\ \alpha_{1,j} &= c_1, \quad j \in I_+, \\ \alpha_{1,j} &= -c_2, \quad j \in I_-, \end{aligned}$$

$$\sum_{j=1}^{p} \alpha_{1,j} = 0.$$

Here p is an inactive variable that belongs to $I_0 = \{1, \ldots, p\} \backslash \{I_+ \cup I_-\}$.

Connection with Balance

$$B(X; I_+, I_-) = \tilde{\alpha}_0 + (c_1 + c_2)U + \tilde{\epsilon},$$

where

$$\tilde{\alpha}_0 = \frac{1}{|I_+|}\sum_{j \in I_+} \alpha_{0,j} - \frac{1}{|I_-|}\sum_{j \in I_-} \alpha_{0,j}, \quad \tilde{\epsilon} = \frac{1}{|I_+|}\sum_{j \in I_+} \epsilon_j - \frac{1}{|I_-|}\sum_{j \in I_-} \epsilon_j.$$

The response y is also linear in $B(X; I_+, I_-)$

$$y = \beta_0 - \tilde{\alpha}_0 \frac{\beta_1}{c_1 + c_2} + \frac{\beta_1}{c_1 + c_2} B(X; I_+, I_-) + \varepsilon - \frac{\beta_1}{c_1 + c_2}\tilde{\epsilon}.$$

How It Works

Let $Z_j = \log(X_j) - \log g(X)$ denote the clr-transformed data. Then

$$Z_j - \mathbb{E}[Z_j] = \alpha_{1,j} U + \frac{1}{p} \sum_{k=1}^{p} (\epsilon_j - \epsilon_k)$$

$\Rightarrow$ univariate regression can distinguish active from inactive variables

How It Works

Let $Z_j = \log(X_j) - \log g(X)$ denote the clr-transformed data. Then

$$Z_j - \mathbb{E}[Z_j] = \alpha_{1,j} U + \frac{1}{p} \sum_{k=1}^{p} (\epsilon_j - \epsilon_k)$$

$\Rightarrow$ univariate regression can distinguish active from inactive variables

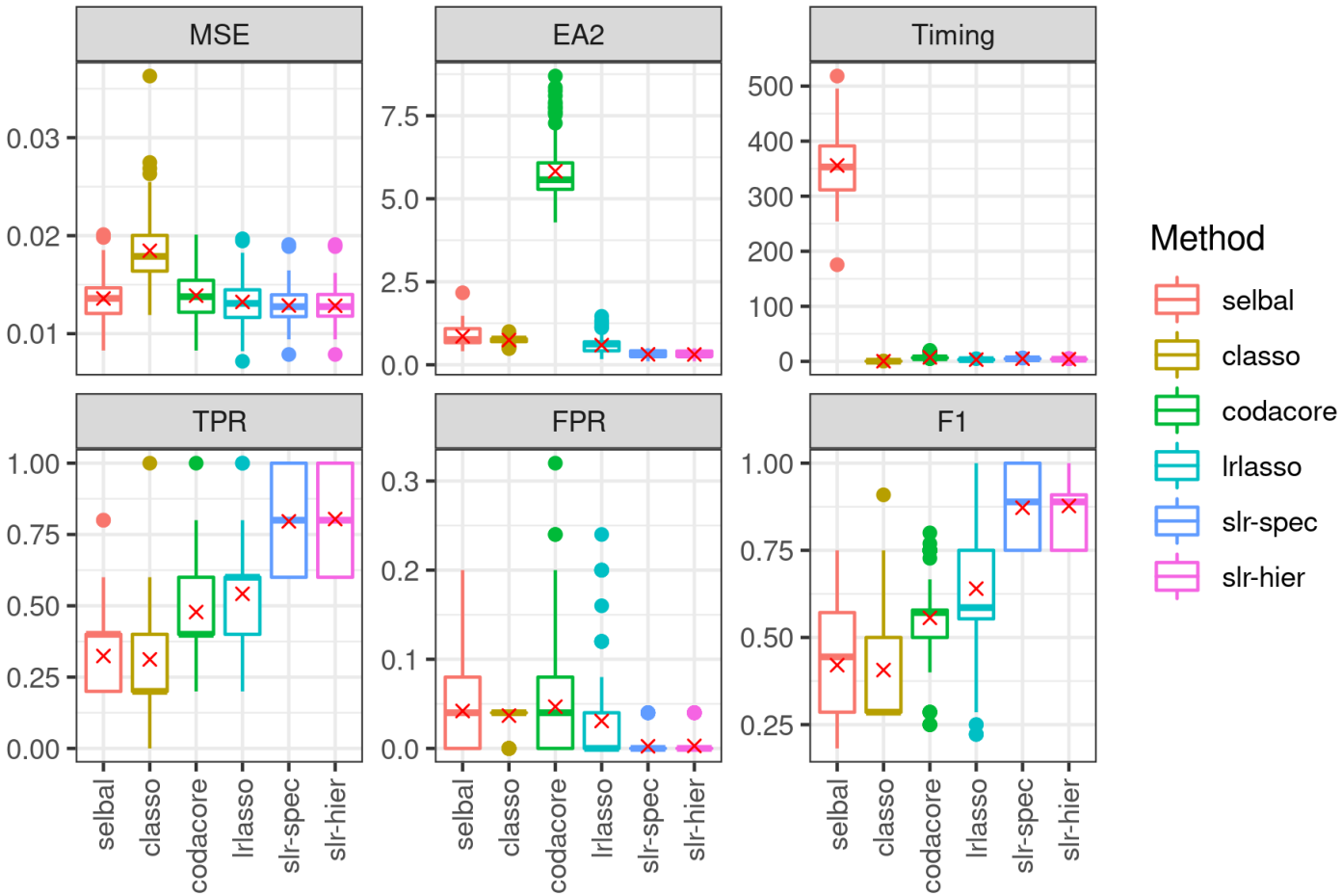
Aitchison Variation

$$\mathrm{Var}(\log \frac{X_j}{X_k}) = \begin{cases} 2\sigma_\epsilon^2 & j \in I_+, k \in I_+ \\ (c_1 + c_2)^2 \sigma_U^2 + 2\sigma_\epsilon^2 & j \in I_+, k \in I_- \\ 2\sigma_\epsilon^2 & j \in I_-, k \in I_-. \end{cases}$$

$\Rightarrow$ clustering can distinguish variables in I_+ from those in I_-

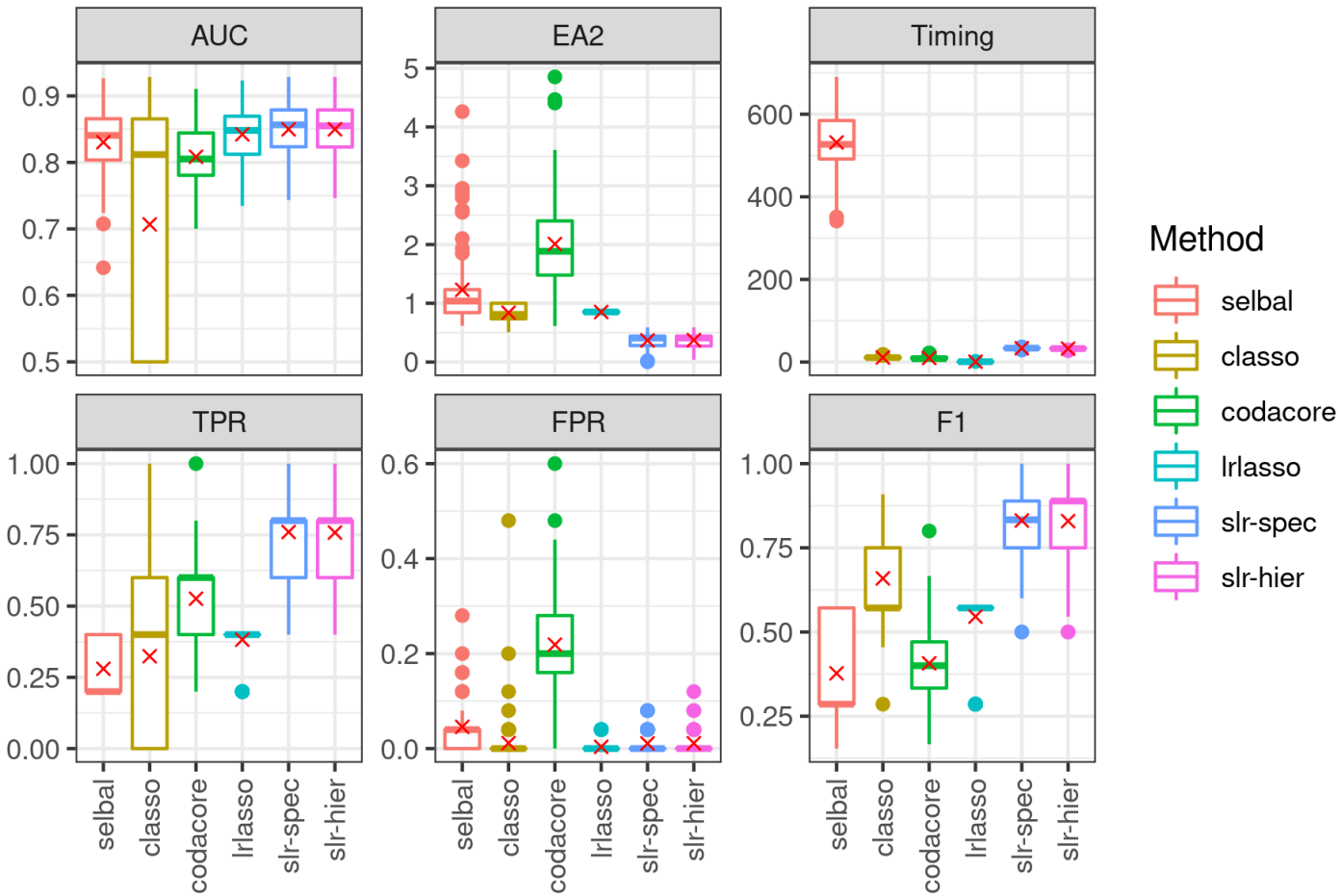
Simulation with Continuous Response

$n = 100, p = 30;\ I_+ = \{1, 2, 3, 4\}, I_- = \{5\}$



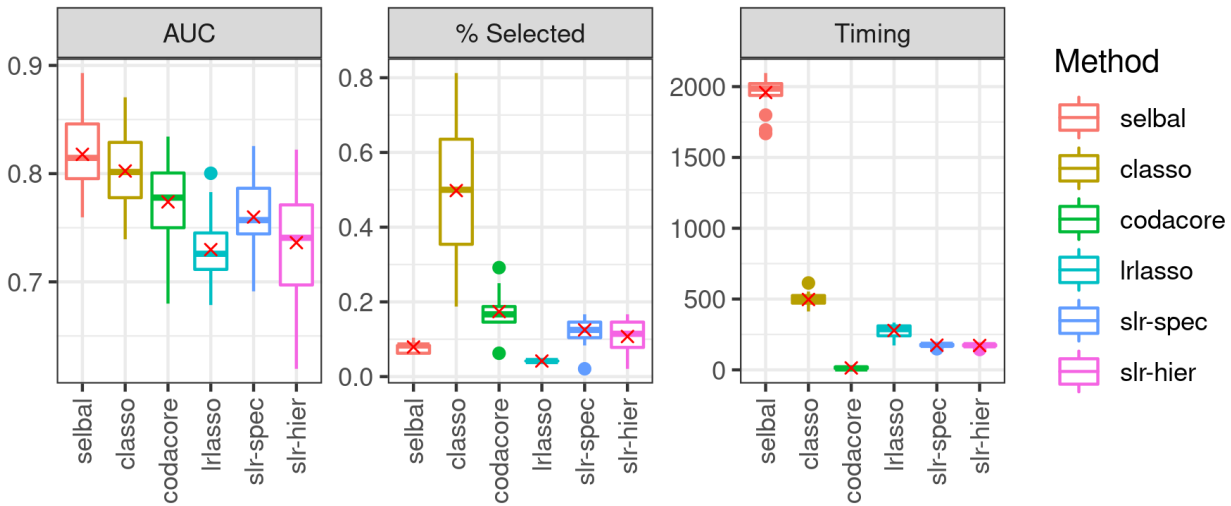
Simulation with Binary Response

$n = 100, p = 30;\ I_+ = \{1, 2, 3, 4\}, I_- = \{5\}$



Classification of Crohn's Disease

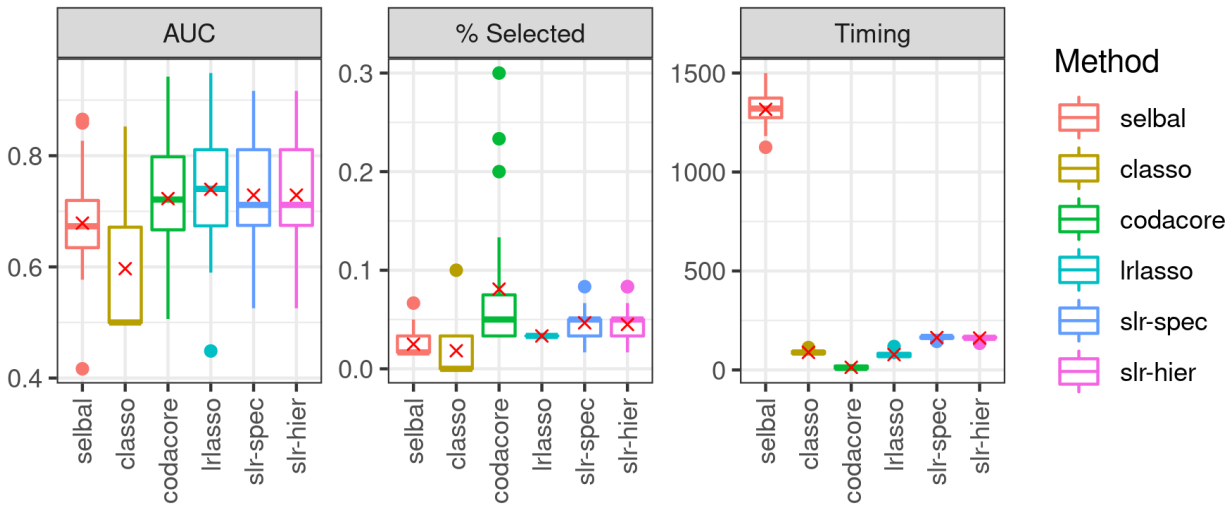
$n = 975; p = 48$ genera; y is binary with 662 cases



- Selbal is the most accurate and also the most time consuming.
- classo does well in AUC, but returns a non-sparse model.
- SLR with spectral clustering and CoDaCoRe are comparable.

Classification of HIV Status

$n = 155; p = 60$ genera; y is binary with 128 cases



- SLR selects a sparser model than CoDaCoRe.
- selbal is the most time consuming.
- classo do not perform well. lrlasso is the most sparse.

Microbiome and sCD14 Inflammation

$n = 151; p = 60$ genera; y is continuous

Taxa	selbal	codacore-1	lrlasso-1	lrlasso-2	slr-spec	slr-hier
"g_Faecalibacterium"					+	+
"f_Ruminococcaceae_g_unclassified"					+	+
"g_Subdoligranulum"	+	+	+		+	+
"g_Thalassospira"	+	+			+	+
"f_Defluviitaleaceae_g_Incertae_Sedis"		+			+	+
"f_Lachnospiraceae_g_Incertae_Sedis"	+			+		+
"g_Dorea"	+					
"g_Dialister"		+				
"f_Lachnospiraceae_g_unclassified"	-	-	-		+	+
'g_Catenibacterium"		-			-	-
"g_Mitsuokella"		-			-	-
"g_Bifidobacterium"	-	-			-	-
"g_Collinsella"	-	-		-	-	-
"g_Lachnospira"	-	-				
"k_Bacteria_g_unclassified"		-				
"g_Ruminococcus"		-				
"g_Megasphaera"		-				
"g_Sutterella"		-				
"o_Clostridiales_g_unclassified"		-				

Summary

- Supervised log-ratio can efficiently predict health outcomes from compositional data.
- SLR leads to interpretable biomarker selection.
- SLR can be extended to semi-supervised settings.

Summary

- Supervised log-ratio can efficiently predict health outcomes from compositional data.
- SLR leads to interpretable biomarker selection.
- SLR can be extended to semi-supervised settings.
- SLR requires proper zero handling.

Summary

- Supervised log-ratio can efficiently predict health outcomes from compositional data.
- SLR leads to interpretable biomarker selection.
- SLR can be extended to semi-supervised settings.
- SLR requires proper zero handling.
- Selection of more than one balance?

Acknowledgment

Kristyn Pantoja @TAMU

David Jones @Google

Thank You!

https://drjingma.com