Compositional Mediation Model (CMM) for Binary Outcome

Michael B. Sohn & Hongzhe Li[†]

University of Rochester

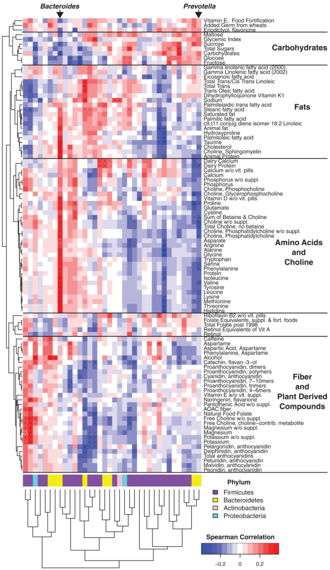
Dept. Biostatistics and Computational Biology

February 7, 2019

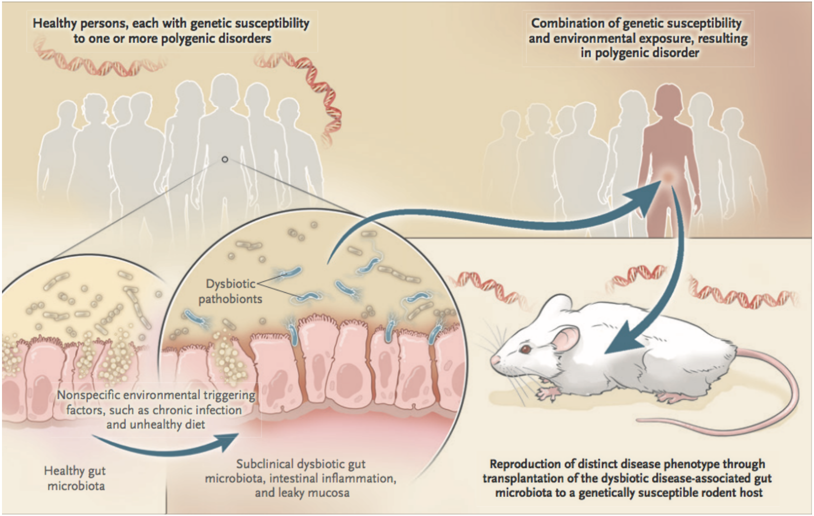
[†]University of Pennsylvania

COMBO Dataset (Wu, *et al.*, 2011 Science)

- Cross-sectional study Of diet and stool MicroBiOme composition (COMBO)
- **Data**
  - 98 healthy subjects/stool samples, not on antibiotics
  - 16S rRNA gene sequences
  - 87 genera appeared in at least one sample
  - Nutrients (FFQ diet questionnaire) & demographic data such as BMI
- **Findings**:

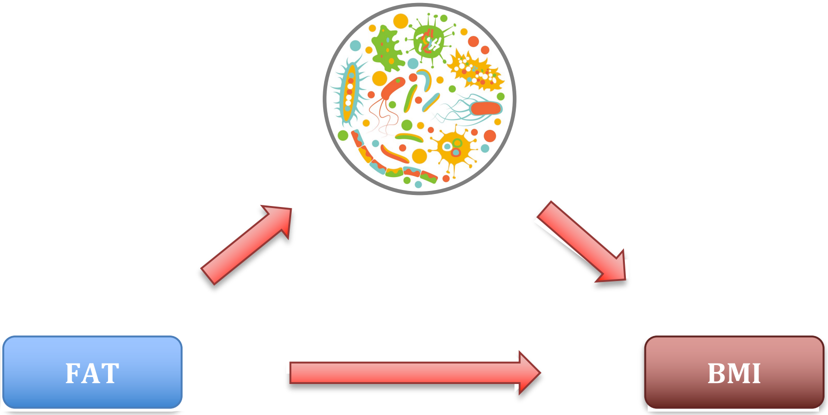


Hypothesis of Pathogenesis Caused by Dysbiosis

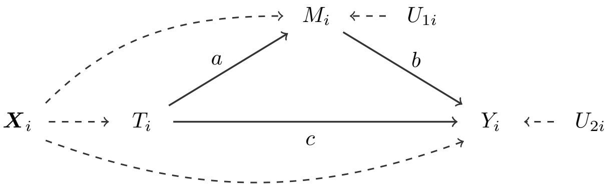


Source: N Engl J Med 2016;375:2369-79

COMBO: Mediation Effect



Mediation Analysis - Structural Equation Model (SEM)



T_i - Treatment, M_i - Mediator, Y_i - Outcome, $\boldsymbol{X}_i$ - Pretreatment Variables

$$M_i = a_0 + aT_i + \boldsymbol{h}^\top \boldsymbol{X}_i + U_{1i} \tag{1}$$

$$Y_i = c_0 + cT_i + bM_i + \boldsymbol{g}^\top \boldsymbol{X}_i + U_{2i} \tag{2}$$

By combining Eq. (1) and (2),

$$\begin{aligned} Y_i &= c_0 + cT_i + b\left(a_0 + aT_i + \boldsymbol{h}^\top \boldsymbol{X}_i + U_{1i}\right) + \boldsymbol{g}^\top \boldsymbol{X}_i + U_{2i} \\ &= c_0^* + ({\color{red}c} + {\color{red}ab})T_i + \boldsymbol{g}^{*\top} \boldsymbol{X}_i + U_i^* \end{aligned}$$

Mediation Analysis - Potential Outcomes Framework

Let T_i represent the binary treatment variable

Causal Direct Effect: $\zeta(t) = \mathbb{E}\left[Y_i(1, M_i(t)|\boldsymbol{X}_i) - Y_i(0, M_i(t)|\boldsymbol{X}_i)\right]$

Causal Indirect Effect: $\delta(t) = \mathbb{E}\left[Y_i(t, M_i(1)|\boldsymbol{X}_i) - Y_i(t, M_i(0)|\boldsymbol{X}_i)\right]$

Necessary Assumptions:

- Stable Unit Treatment Value Assumption (SUTVA)
- Sequential Ignorability Assumption,

$$\{Y_i(t', m), M_i(t)\} \perp\!\!\!\perp T_i|\boldsymbol{X}_i = \boldsymbol{x},$$

$$Y_i(t', m) \perp\!\!\!\perp M_i(t)|T_i = t, \boldsymbol{X}_i = \boldsymbol{x},$$

where $0 < Pr(T_i = t|\boldsymbol{X}_i = \boldsymbol{x})$ and $0 < Pr(M_i(t) = m|T_i = t, \boldsymbol{X}_i = \boldsymbol{x})$ for $t = 0, 1$.

With the necessary assumptions,

$$\zeta(t) = \int \mathbb{E}(Y_i|M_i, T_i, \boldsymbol{X}_i)\left[dF_{M_i|T_i=1,\boldsymbol{X}_i}(m) - dF_{M_i|T_i=0,\boldsymbol{X}_i}(m)\right] dF_{\boldsymbol{X}_i}(\boldsymbol{x})$$

$$\delta(t) = \int \left[\mathbb{E}(Y_i|M_i, T_i = 1, \boldsymbol{X}_i) - \mathbb{E}(Y_i|M_i, T_i = 0, \boldsymbol{X}_i)\right] dF_{M_i|T_i,\boldsymbol{X}_i}(m) dF_{\boldsymbol{X}_i}(\boldsymbol{x})$$

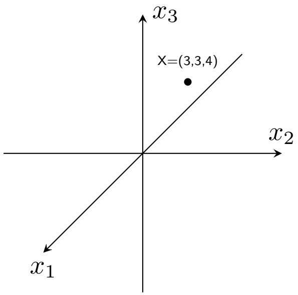
Compositional Data Analysis

Compositional data:

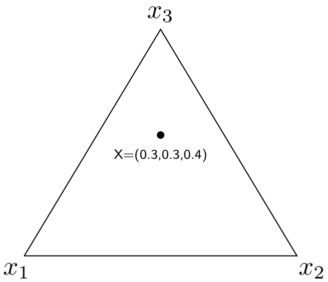
- Relative information
- Proportions or percentages of a whole

Unit-sum constraint: sum of proportions = 1

Euclidean Space



Simplex Space



Subcompositional Coherence

Principle of subcompositional coherence: analysis concerning a subset of components must not depend on excluded components

Example: Scientists A and B record the composition of soil samples:

A records **animal**, **vegetable**, **mineral**, and **water**.

B records **animal**, **vegetable**, and **mineral** after drying the sample.

Both are absolutely accurate. [adapted from Aitchison, 2005]

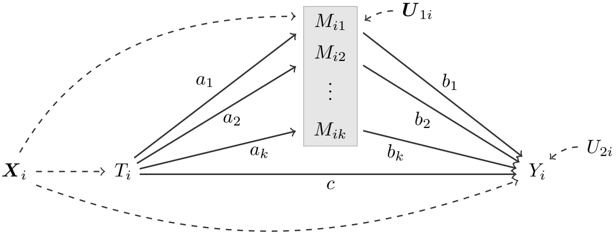
Sample A	x_1	x_2	x_3	x_4
1	0.1	0.2	0.1	0.6
2	0.2	0.1	0.2	0.5
3	0.3	0.3	0.1	0.3

Sample B	x_1	x_2	x_3
1	0.25	0.50	0.25
2	0.40	0.20	0.40
3	0.43	0.43	0.14

Corr A	x_1	x_2	x_3	x_4
x_1	1.00	0.50	0.00	-0.98
x_2		1.00	-0.87	-0.65
x_3			1.00	0.19

Corr B	x_1	x_2	x_3
x_1	1.00	-0.57	-0.05
x_2		1.00	-0.79
x_3			1.00

Compositional Mediation Model (Sohn and Li, AOAS, accepted)



Compositional operators (Aitchison, 1986; Billheimer, *et al.* 2001):

$$\boldsymbol{m} \oplus \boldsymbol{a} = \left(\frac{m_1 a_1}{\sum_{j=1}^k m_k a_k}, \cdots, \frac{m_k a_k}{\sum_{j=1}^k m_k a_k} \right)^\top; \quad \boldsymbol{m}^z = \left(\frac{m_1^z}{\sum_{j=1}^k m_k^z}, \cdots, \frac{m_k^z}{\sum_{j=1}^k m_k^z} \right)^\top$$

Compositional mediation model:

$$\boldsymbol{M}_i = \left(\boldsymbol{m}_0 \oplus \boldsymbol{a}^{T_i} \bigoplus_{r=1}^{n_x} \boldsymbol{h}_r^{X_{ri}} \right) \oplus \boldsymbol{U}_{1i}$$

$$Y_i = c_0 + c\,T_i + \boldsymbol{b}^\top (\log \boldsymbol{M}_i) + \boldsymbol{g}^\top \boldsymbol{X}_i + U_{2i}, \quad \text{subject to } \mathbf{1}_k^\top \boldsymbol{b} = 0$$

Identification for Compositional Mediation Model

Necessary assumptions:

- SUTVA
- Sequential Ignorability Assumption

Under the potential outcomes framework

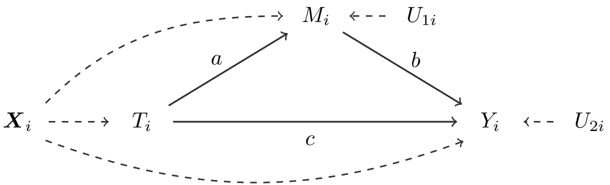
- Expected Causal Direct Effect

$$\begin{aligned}\zeta(t) &= \mathbb{E}[Y_i(1, \log \boldsymbol{M}_i(t) | \boldsymbol{X}_i(t)) - Y_i(0, \log \boldsymbol{M}_i(t) | \boldsymbol{X}_i(t))] \\ &= c\end{aligned}$$

- Expected Causal Indirect Effect

$$\begin{aligned}\delta(t) &= \mathbb{E}[Y_i(t, \log \boldsymbol{M}_i(1) | \boldsymbol{X}_i(t)) - Y_i(t, \log \boldsymbol{M}_i(0) | \boldsymbol{X}_i(t))] \\ &= (\log \boldsymbol{a})^\top \boldsymbol{b}\end{aligned}$$

Binary Outcome Under SEM



T_i - Treatment, M_i - Mediator, Y_i - Binary outcome, $\boldsymbol{X}_i$ - Pretreatment Variables

$$M_i = a_0 + aT_i + \boldsymbol{h}^\top \boldsymbol{X}_i + U_{1i} \tag{3}$$

$$Y_i = \mathbf{1}\{Y_i^* > 0\}, \text{ where } Y_i^* = c_0 + cT_i + bM_i + \boldsymbol{g}^\top \boldsymbol{X}_i + U_{2i} \tag{4}$$

By combining Eq. (3) and (4),

$$\begin{aligned} Y_i^* &= c_0 + cT_i + b\left(a_0 + aT_i + \boldsymbol{h}^\top \boldsymbol{X}_i + U_{1i}\right) + \boldsymbol{g}^\top \boldsymbol{X}_i + U_{2i} \\ &= c_0^* + ({\color{red}c} + {\color{red}ab})T_i + \boldsymbol{g}^{*\top} \boldsymbol{X}_i + U_i^* \end{aligned}$$

Binary Outcome Under Potential Outcomes Framework

Assumptions: SUTVA, Sequential ignorability

Estimation of Causal Direct ζ and Indirect Effects δ:

- Logit Model
  - Complex numerical integration required
  - Odds ratios with rare outcomes: $\log OR_\zeta \approx c; \quad \log OR_\delta \approx ab$

- Probit Model

$$\zeta = \mathbb{E}\left\{\Phi\left(\frac{c+f_1}{\sqrt{\sigma^2 b^2+1}}\right) - \Phi\left(\frac{f_1}{\sqrt{\sigma^2 b^2+1}}\right)\right\}$$

$$\delta = \mathbb{E}\left\{\Phi\left(\frac{ab+f_2}{\sqrt{\sigma^2 b^2+1}}\right) - \Phi\left(\frac{f_2}{\sqrt{\sigma^2 b^2+1}}\right)\right\}$$

where $f_1 = c_0 + a_0 b + b(\boldsymbol{h}+\boldsymbol{g})^\top \boldsymbol{X}_i$ and $f_2 = c_0 + c + a_0 b + b(\boldsymbol{h}+\boldsymbol{g})^\top \boldsymbol{X}_i$

CMM for Binary Outcome (Probit Model)

Compositional mediation model for the binary outcome:

$$\boldsymbol{M}_i = \left(\boldsymbol{m}_0 \oplus \boldsymbol{a}^{T_i} \bigoplus_{r=1}^{n_x} \boldsymbol{h}_r^{X_{ri}}\right) \oplus \boldsymbol{U}_{1i}$$

$$Y_i = 1\{c_0 + cT_i + \boldsymbol{b}^\top(\log \boldsymbol{M}_i) + \boldsymbol{g}^\top \boldsymbol{X}_i + U_{2i} > 0\}, \quad \text{subject to } \mathbf{1}_k^\top \boldsymbol{b} = 0$$

where $\boldsymbol{U}_{1i} \sim LN(0, \Sigma)$ and $U_{2i} \sim N(0,1)$.

Expected Causal Direct and Indirect Effects:

$$\zeta(\tau) = \mathbb{E}\left\{\Phi\left(\frac{ct + f_\zeta(\tau, \boldsymbol{X}_i)}{\sqrt{\boldsymbol{b}_{-k}^\top \Sigma \boldsymbol{b}_{-k} + 1}}\right) - \Phi\left(\frac{ct' + f_\zeta(\tau, \boldsymbol{X}_i)}{\sqrt{\boldsymbol{b}_{-k}^\top \Sigma \boldsymbol{b}_{-k} + 1}}\right)\right\}$$

$$\delta(\tau) = \mathbb{E}\left\{\Phi\left(\frac{(\log \boldsymbol{a})^\top \boldsymbol{b}\, t + f_\delta(\tau, \boldsymbol{X}_i)}{\sqrt{\boldsymbol{b}_{-k}^\top \Sigma \boldsymbol{b}_{-k} + 1}}\right) - \Phi\left(\frac{(\log \boldsymbol{a})^\top \boldsymbol{b}\, t' + f_\delta(\tau, \boldsymbol{X}_i)}{\sqrt{\boldsymbol{b}_{-k}^\top \Sigma \boldsymbol{b}_{-k} + 1}}\right)\right\}$$

where $f_\zeta(\tau, \boldsymbol{x}) = c_0 + \boldsymbol{b}^\top(\log \boldsymbol{m}_0 + \tau \log \boldsymbol{a} + \sum_{r=1}^{n_x} x_r \log \boldsymbol{h}_r) + \boldsymbol{g}^\top \boldsymbol{x}$;

$f_\delta(\tau, \boldsymbol{x}) = c_0 + c\tau + \boldsymbol{b}^\top(\log \boldsymbol{m}_0 + \sum_{r=1}^{n_x} x_r \log \boldsymbol{h}_r) + \boldsymbol{g}^\top \boldsymbol{x}$.

Estimators for Compositional Parameters

Optimization problem in a simplex space:

$$(\widehat{\boldsymbol{m}}_0, \widehat{\boldsymbol{a}}, \widehat{\boldsymbol{h}}_1, \ldots, \widehat{\boldsymbol{h}}_{n_x}) = \underset{\boldsymbol{m}_0, \boldsymbol{a}, \boldsymbol{h}_r \in \mathbb{S}^{k-1}}{\operatorname{argmin}} \sum_{i=1}^{n} \left\| \boldsymbol{M}_i \ominus \left(\boldsymbol{m}_0 \oplus \boldsymbol{a}^{T_i} \bigoplus_{r=1}^{n_x} \boldsymbol{h}_r^{X_{ri}} \right) \right\|^2$$

where

$$\boldsymbol{m} \ominus \boldsymbol{a} = \left(\frac{m_1 a_1^{-1}}{\sum_{j=1}^{k} m_k a_k^{-1}}, \frac{m_2 a_2^{-1}}{\sum_{j=1}^{k} m_k a_k^{-1}}, \cdots, \frac{m_k a_k^{-1}}{\sum_{j=1}^{k} m_k a_k^{-1}} \right)$$

$$\|\boldsymbol{m}\| = \langle \boldsymbol{m}, \boldsymbol{m} \rangle^{1/2} = \mathrm{alr}(\boldsymbol{m})^\top \boldsymbol{\mathcal{N}}^{-1} \mathrm{alr}(\boldsymbol{m})$$

$$\mathrm{alr}(\boldsymbol{m}) = \left(\log \frac{m_1}{m_k}, \log \frac{m_2}{m_k}, \ldots, \log \frac{m_{k-1}}{m_k} \right)^\top$$

$$\boldsymbol{\mathcal{N}}^{-1} = \boldsymbol{\mathcal{I}}_{k-1} - \frac{1}{k} \mathbf{1}_{k-1} \mathbf{1}_{k-1}^\top$$

Estimator for Parameters in Probit Regression

Let $\eta_i = 2y_i - 1$, $\boldsymbol{z}_i = (1, t_i, \log(\boldsymbol{m}_i)^\top, \boldsymbol{x}_i^\top)^\top$, and $\boldsymbol{\beta} = (c_0, c, \boldsymbol{b}^\top, \boldsymbol{g}^\top)^\top$

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \log \Phi(\eta_i \boldsymbol{z}_i^\top \boldsymbol{\beta}) + \lambda \left\| \boldsymbol{\beta} \right\|_1 \right\}, \quad \text{s.t. } \mathbf{1}_k^\top \boldsymbol{b} = 0 \tag{5}$$

Alternative optimization problem:

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \frac{1}{2n} \| \Xi^{1/2} (\boldsymbol{u} - Z\boldsymbol{\beta}) \|_2^2 + \lambda \left\| \boldsymbol{\beta} \right\|_1 \right\}, \quad \text{s.t. } \mathbf{1}_k^\top \boldsymbol{b} = 0, \tag{6}$$

where Ξ is the $n \times n$ diagonal matrix with $\Xi_{ii} = \xi_i(\eta_i \boldsymbol{z}_i^\top \boldsymbol{\beta}^*)[\boldsymbol{z}_i^\top \boldsymbol{\beta}^* + \xi_i(\eta_i \boldsymbol{z}_i^\top \boldsymbol{\beta}^*)]$, $\xi_i(\eta_i \boldsymbol{z}_i^\top \boldsymbol{\beta}) = \eta_i \phi(\eta_i \boldsymbol{z}_i^\top \boldsymbol{\beta}) / \Phi(\eta_i \boldsymbol{z}_i^\top \boldsymbol{\beta})$, $\boldsymbol{u} = Z\boldsymbol{\beta}_0 + (\Xi)^{-1} \boldsymbol{\xi}$, $Z = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)^\top$, and $\boldsymbol{\xi} = (\xi_1(\eta_1 \boldsymbol{z}_1^\top \boldsymbol{\beta}_0), \ldots, \xi_1(\eta_n \boldsymbol{z}_n^\top \boldsymbol{\beta}_0))^\top$

Numerical Algorithm

Proposed method: IRLS-CDMM

$$\widehat{\boldsymbol{\beta}}^{(\ell)} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \frac{1}{2n} \left\| \Xi^{(\ell-1)1/2} (\boldsymbol{u}^{(\ell-1)} - Z\boldsymbol{\beta}) \right\|_2^2 + \lambda \left\| \boldsymbol{\beta} \right\|_1 \right\}, \text{ s.t. } \mathbf{1}_k^\top \boldsymbol{b}^{(\ell)} = 0,$$

$$= \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \frac{1}{2n} \left\| \Xi^{(\ell-1)1/2} (\boldsymbol{u}^{(\ell-1)} - \widetilde{Z}\boldsymbol{\beta}) \right\|_2^2 + \lambda \left\| \boldsymbol{\beta} \right\|_1 \right\}, \text{ s.t. } \boldsymbol{\iota}^\top \boldsymbol{\beta}^{(\ell)} = 0,$$

where $\widetilde{Z} = Z(I_p - \boldsymbol{\iota}\boldsymbol{\iota}^\top / k)$ and $\boldsymbol{\iota} = (0, 0, 1, \ldots, 1, 0, \ldots, 0)^\top$.

Algorithm: IRLS-CDMM with Augmented Lagrangian Method

Step 1. Initialize $\boldsymbol{\beta}^{(0)}$, $\alpha^{(0)}$

Step 2. Update $\beta_j^{(\ell+1)}$ until convergence

Step 3. Update $\Xi^{(\ell+1)}$ and $\boldsymbol{u}^{(\ell+1)}$ by minimizing $\sum_{i=1}^n q(\eta_i \boldsymbol{z}_i^\top \boldsymbol{\beta})$

Step 4. Update $\alpha^{(k+1)}$

Debiased Estimator & Its Asymptotic Convergence

Debiased Estimator of (6):

$$\widehat{\boldsymbol{\beta}}_{db} = \widehat{\boldsymbol{\beta}} + \frac{1}{n}\widetilde{M}\widetilde{Z}^\top \Xi(\boldsymbol{u} - \widetilde{Z}\widehat{\boldsymbol{\beta}}),$$

where $\widetilde{M} = (I_p - \boldsymbol{\iota}\boldsymbol{\iota}^\top/k)M$ and $M = (\boldsymbol{m}_1, \ldots, \boldsymbol{m}_p)^\top$ is a solution of the convex problem (Javanmard and Montanari, 2014; Shi, *et. al.*, 2016):

$$\min \boldsymbol{m}_j^\top \widehat{\Sigma} \boldsymbol{m}_j \ \text{ s.t. } \|\widehat{\Sigma}\boldsymbol{m}_j - (I_p - \boldsymbol{\iota}\boldsymbol{\iota}^\top/k)\boldsymbol{e}_j\|_\infty \le \gamma, \ j = 1, \ldots, p,$$

where $\boldsymbol{e}_j$ is the j^{th} natural basis and γ is some constant.

Theorem: *For an s-sparse $\boldsymbol{\beta}$, under some regularity conditions,*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_{db} - \boldsymbol{\beta}) = R + \Delta, \qquad \mathbb{E}(R|Z) = \mathbf{0}, \qquad \|\Delta\|_\infty \to 0$$

Inference

Null Hypothesis for the expected causal direct and indirect effects:

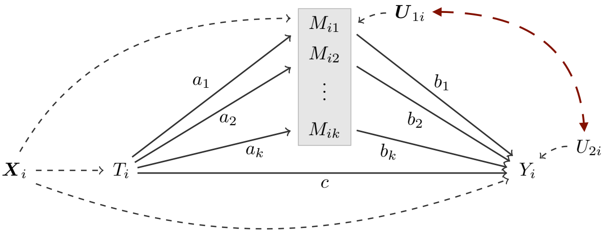
$$H_0 : \zeta(\tau) = 0 \quad \text{vs.} \quad H_1 : \zeta(\tau) \neq 0.$$

$$H_0 : \delta(\tau) = 0 \quad \text{vs.} \quad H_1 : \delta(\tau) \neq 0.$$

Testing Procedure (Non-parametric Bootstrap):

1. Randomly select n samples from the original n samples with replacement
2. Estimate $\zeta_b(\tau)$ and $\delta_b(\tau)$
3. Repeat Steps 1 and 2 to construct sampling distributions of $\zeta_b(\tau)$ and $\delta_b(\tau)$
4. Construct percentile bootstrap confidence intervals for $\zeta_b(\tau)$ and $\delta_b(\tau)$

Sensitivity Analysis for Binary Outcome



Probit regression: $Y_i = 1\{\tilde{c}_0 + \tilde{c}T_i + \tilde{\boldsymbol{g}}^\top \boldsymbol{X}_i + U_{0i} > 0\}$, where $U_{0i} \sim N(0,1)$

Expected causal indirect effect given $\boldsymbol{\rho} = \text{corr}(\text{alt}(\boldsymbol{U}_{1i}), U_{2i})$:

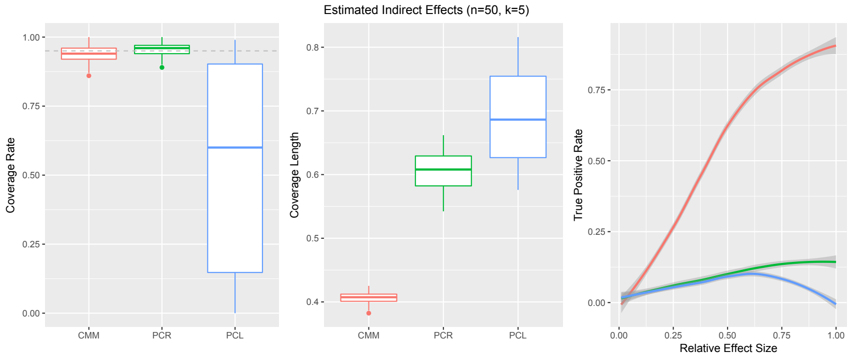
$$\delta_\rho(\tau) = \mathbb{E}\left\{\Phi\left(\tilde{f}_\delta(\tau) + \frac{(\log \boldsymbol{a})^\top \boldsymbol{b}_\rho(t-\tau)}{\Psi(\boldsymbol{\rho}, \boldsymbol{b}_\rho, \Sigma)}\right) - \Phi\left(\tilde{f}_\delta(\tau) + \frac{(\log \boldsymbol{a})^\top \boldsymbol{b}_\rho(t'-\tau)}{\Psi(\boldsymbol{\rho}, \boldsymbol{b}_\rho, \Sigma)}\right)\right\},$$

where $\tilde{f}_\delta(\tau) = \tilde{c}_0 + \tilde{c}\tau + \tilde{\boldsymbol{g}}^\top \boldsymbol{x}_i$, $\Psi(\boldsymbol{\rho}, \boldsymbol{b}_\rho, \Sigma) = \left[(\boldsymbol{b}_\rho)_{-k}^\top \Sigma\, (\boldsymbol{b}_\rho)_{-k} + 2\boldsymbol{\rho}^\top D\, (\boldsymbol{b}_\rho)_{-k} + 1\right]^{1/2}$, and D is a diagonal matrix with diag$(\Sigma^{1/2})$.

Performance Evaluation I ($\alpha = 0.05$)

Binary treatment ($t = 1,\ t' = 0$); $\quad \boldsymbol{a} = (20, 10, 5, 2, \mathbf{1}_{k-4}^\top)^\top / (20, 10, 5, 2, \mathbf{1}_{k-4}^\top)^\top \mathbf{1}_k$;

$\boldsymbol{b} = (0.5, -0.5, 0.5, -0.5, \mathbf{0}_{k-4}^\top)^\top$; $\quad (\log \boldsymbol{a})^\top \boldsymbol{b} \times$ Effect Size

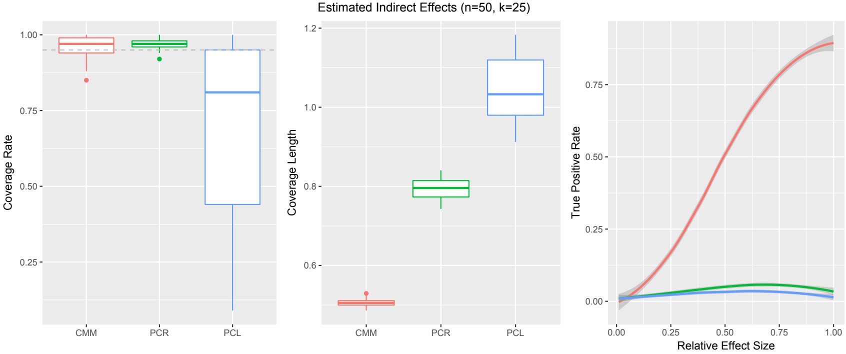


CMM: Proposed compositional mediation model

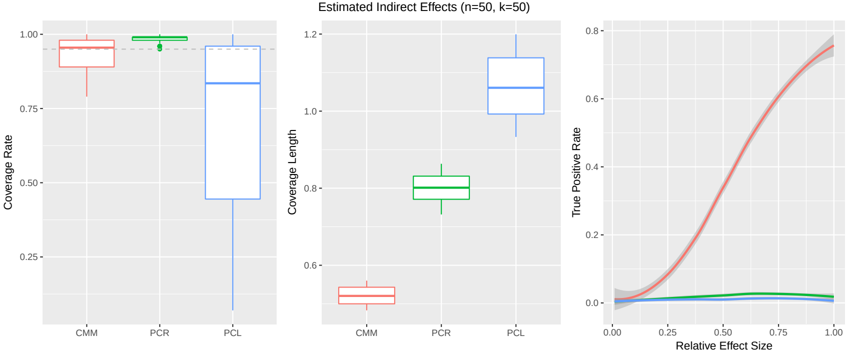
PCR: Principal components of compositional variables under POF

PCL: Principal components of compositional variables under SEM

Performance Evaluation II ($\alpha = 0.05$)

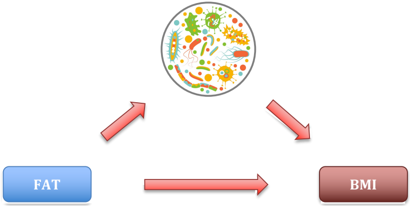


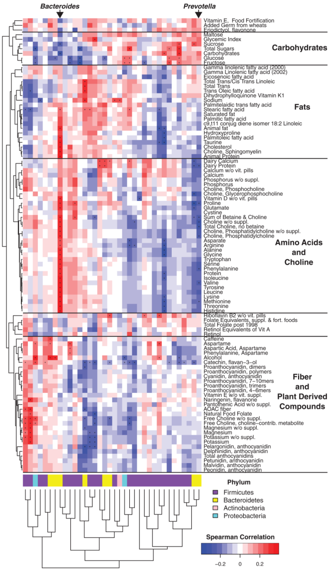
Performance Evaluation III ($\alpha = 0.05$)



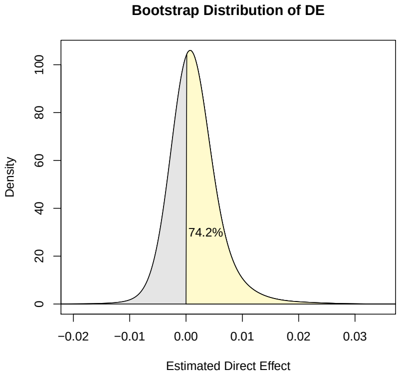
COMBO Dataset

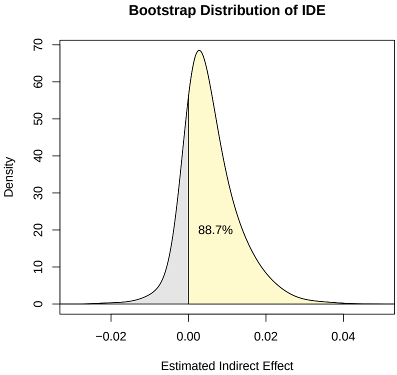
- **Data**
  - 98 healthy subjects
  - Fat intake as treatment
  - 45 genera as compositional mediators
  - Dichotomized BMI at 25
- **Interest**:





Fat intake, Microbiome, and Obesity (COMBO)





DE	TIDE
0.002 (-0.003, 0.011)	0.008 (-0.005, 0.023)
20%	80%

$\Rightarrow$ About 80% of effect of fat intake on dichotomized BMI is mediated through gut microbiome

Acknowledgement

Many thanks to

- Hongzhe Li, PhD