Sparse Bayesian Group Factor Model for Feature Interactions in Multiple Count Tables Data

Shuangjie Zhang

Joint work with
Yuning Shen(Chemical and Biomolecular Engineering, UCLA),
Irene A. Chen(Chemical and Biomolecular Engineering, UCLA),
Juhee Lee(Statistics, UCSC)

September 11, 2024

Multivariate Count Data

- $\boldsymbol{y} \in \mathbb{N}^0 \Rightarrow$ represent the number of occurrences
- Various science fields: genomics (Schloissnig et al., 2013), epidemiology (Papoz, Balkau, and Lellouch, 1996), social sciences (Böhning, Dietz, and Schlattmann, 1997), and marketing (Ravishanker, Venkatesan, and Hu, 2016).
- Inferential goals: interactions among the features through covariance matrix
- Poisson distribution or negative binomial distribution

? Multivariate Poisson distribution or multivariate NB distribution

High-dimensional Count Data

- sample size n smaller than the number of variables J - ‘small n large J’ problem
- Banding sample covariance matrix or its Cholesky (Bickel and Levina, 2008b; Wu and Pourahmadi, 2003); thresholding with shrinkage (Bickel and Levina, 2008a; Rothman, Levina, and Zhu, 2009)
- Many regularization approaches but what about uncertainty?

⋆⋆ Bayesian model enters naturally

⇒ Latent continuous variables

⇒ Bayesian factor model (Bernardo et al., 2003)

Bayesian Factor model(Bernardo et al., 2003)

- Common latent factors without losing essential information

⋆⋆ the number of latent factors K smaller than dimension J

⋆⋆ the factor loadings matrix having a lot of zeros

- Spike-slab prior (Carvalho et al., 2008; Lucas et al., 2006);
  Heavy-tailed default prior (Ghosh and Dunson, 2009);
  Multiplicative gamma process shrinkage prior (Bhattacharya and Dunson, 2011)
  Dirichlet-Laplace prior (Bhattacharya et al., 2015).

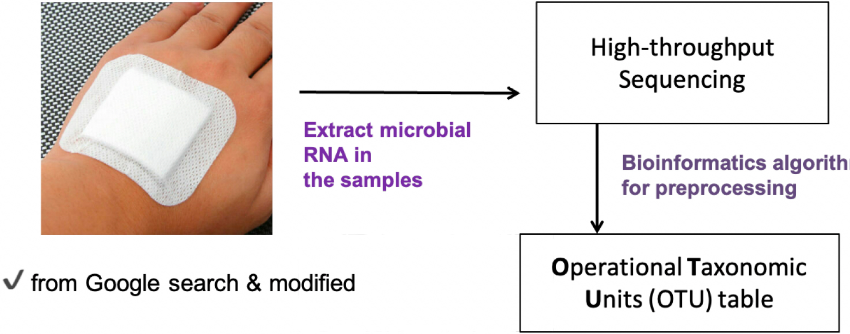
More Recent Development

- Group factor analysis (Klami et al., 2014; Virtanen et al., 2012)
- Multi-study factor analysis (De Vito et al., 2019)
- Perturbed factor analysis (Roy et al., 2021)
- Generalized factor models (Schiavon, Canale, and Dunson, 2022)

? high-dimensional multivariate count tables with added complexity

Microbiome Study

- High-throughput sequencing such as 16S rRNA gene sequencing is widely used in microbiome studies to profile microbial communities.



- Operational Taxonomic Units (OTUs) represent the information of microbial taxa.

Microbiome Count Data

subject	sample	OTU 1	OTU 2	OTU 3	$\cdots$	$\cdots$	OTU J	covariate
1	1	41	643	89	$\cdots$	0	1	x_1
1	2	0	56	24	$\cdots$	402	32	x_2
1	3	34	12	0	$\cdots$	28	17	x_3
$\vdots$	$\vdots$	$\vdots$	$\vdots$				$\vdots$	$\vdots$
S	N-2	410	601	305		106	509	x_{N-2}
S	N-1	0	0	10	$\cdots$	0	7	x_{N-1}
S	N	698	232	390	$\cdots$	131	987	x_N

Abundance changes with covariates

OTU interacts with each other

Additional Challenges

sample	OTU 1	OTU 2	OTU 3	⋯	⋯	OTU J	Total
1	41	643	89	⋯	104	1	841
2	0	56	24	⋯	10	32	908
3	34	89	0	⋯	762	17	3274
⋮	⋮	⋮				⋮	⋮
N-2	410	601	305		708	509	4210
N-1	0	0	10	⋯	0	7	590
N	698	232	390	⋯	545	987	6298

$$\frac{56}{908}$$

$$\frac{89}{3274}$$

- Compositionality; excess zeros; over-dispersion.

Multiple Count Tables Data

Cross-domain interaction

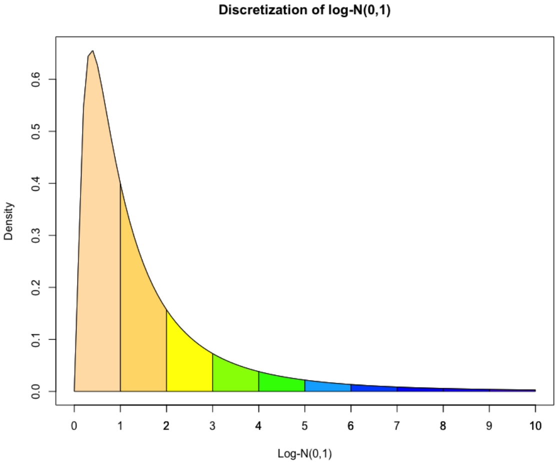
Domain 1 e.g., bacteria — **Within-domain interaction**

Domain 2 e.g., viruses — **Within-domain interaction**

subject	sample	OTU 1	OTU 2	OTU 3	⋯	OTU J_1	OTU 1	OTU 2	OTU 3	⋯	OTU J_2	covariate
1	*1*	41	643	89	⋯	1	53	79	43	⋯	83	x_1
1	*2*	0	56	24	⋯	32	120	87	42	⋯	20	x_2
1	*3*	34	12	0	⋯	17	34	0	209	⋯	73	x_3
⋮	⋮	⋮	⋮			⋮	⋮	⋮			⋮	⋮
S	*N-2*	410	601	305		509	98	158	47		92	x_{N-2}
S	*N-1*	0	0	10	⋯	7	101	934	0	⋯	82	x_{N-1}
S	*N*	698	232	390	⋯	987	389	920	0	⋯	372	x_N

Bayesian Rounded Kernel Model(Canale and Dunson, 2011)

- We propose to discretize a continuous multivariate log-normal distribution with fixed thresholds.



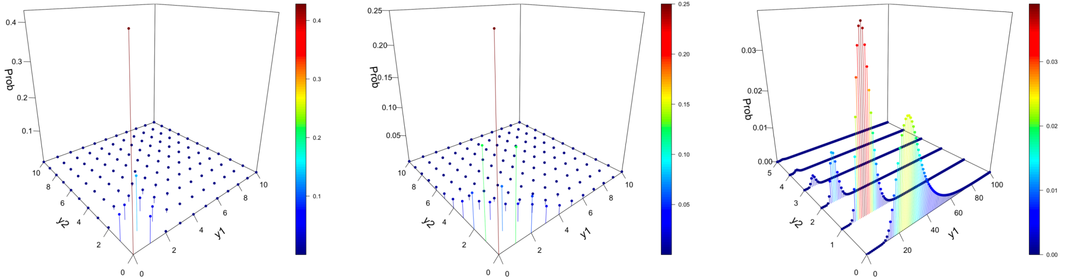
Sampling Distribution

- Sample index $i = 1, \cdots, N$, subject index $s_i = 1, \cdots, S$, group index $m = 1, \cdots, M$.
- Stack count vector from each group $\mathbf{Y}_{im}$ together as $\mathbf{Y}_i = (\mathbf{Y}_{i1}, \ldots, \mathbf{Y}_{iM})$,

$$\mathrm{P}(\mathbf{Y}_i = \mathbf{y}_i \mid \boldsymbol{\mu}_i, \Sigma) = \int_{A(\mathbf{y}_i)} \text{log-N}_J(\mathbf{y}^\star \mid \boldsymbol{\mu}_i, \Sigma) d\mathbf{y}^\star = \int_{\tilde{A}(\mathbf{y}_i)} \phi_J(\tilde{\mathbf{y}}^\star \mid \boldsymbol{\mu}_i, \Sigma) d\tilde{\mathbf{y}}^\star,$$

- where $A(\mathbf{y}_i) = \{\mathbf{y}^\star \mid y_{i1} \leq y_1^\star < y_{i1} + 1, \ldots, y_{iJ} \leq y_J^\star < y_{iJ} + 1\}$ and $\tilde{A}(\mathbf{y}_i) = \{\tilde{\mathbf{y}}^\star \mid \log(y_{i1}) \leq \tilde{y}_1^\star < \log(y_{i1} + 1), \ldots, \log(y_{iJ}) \leq \tilde{y}_J^\star < \log(y_{iJ} + 1)\}$.
- Moments of count distribution

Count distribution



(a) $\tilde{\boldsymbol{\mu}} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \tilde{\Sigma} = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}$

$\text{Cor}(y_1, y_2) = 0.835$

(b) $\tilde{\boldsymbol{\mu}} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \tilde{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

$\text{Cor}(y_1, y_2) = 0$

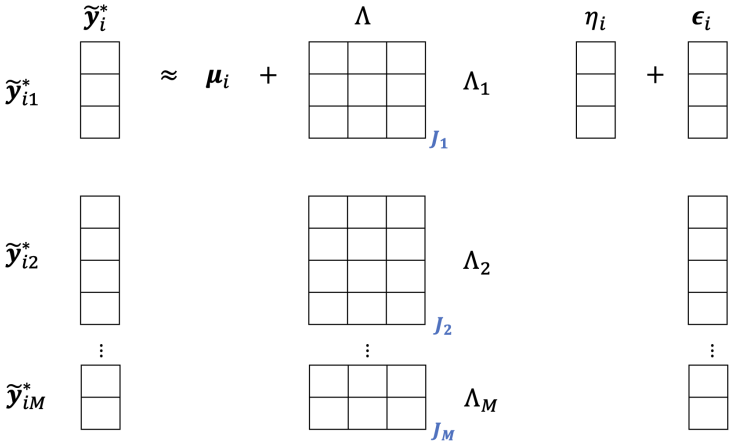
(c) $\tilde{\boldsymbol{\mu}} = \begin{bmatrix} 3 \\ 0 \end{bmatrix}, \tilde{\Sigma} = \begin{bmatrix} 0.5^2 & -0.9 \times 0.5^2 \\ -0.9 \times 0.5^2 & 0.5^2 \end{bmatrix}$

$\text{Cor}(y_1, y_2) = -0.643$

Figure: [Distribution of Counts of a Pair of OTUs] The joint distribution of counts of a pair of OTUs is computed for a rounded kernel method with bivariate log-normals, log-$N_2(\tilde{\boldsymbol{\mu}}, \tilde{\Sigma})$.

Covariance: Group Factor Model

Group factor model: extends traditional factor model to infer joint variability between two or more multivariate responses (Klami et al., 2014; Virtanen et al., 2012; Zhao et al., 2016).

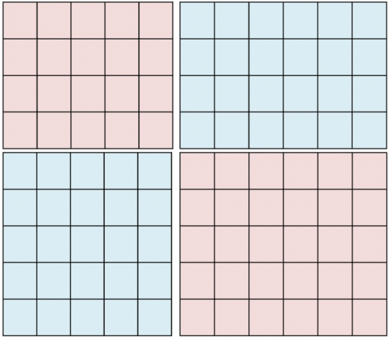


Covariance: Group Factor Model

$$\Sigma = \Lambda_{J\times K}\Lambda' + V,\ \Lambda = [\Lambda_1', \ldots, \Lambda_M']'$$

where $\Lambda_m = [\lambda_{mjk}]$ is a $J_m \times K$ matrix, and V is a J-dim diagonal matrix, with diagonal submatrices $V^{mm} = v_m^2 \mathbf{I}_{J_m}$.

$$\Sigma \approx \Lambda\Lambda' + V$$



$$\Sigma^{11} = \Lambda_1\Lambda_1' + v_1^2 \mathrm{I}_{J_1}$$

$$\Sigma^{21} = \Lambda_1\Lambda_2'$$

$$\Sigma^{12} = \Sigma^{21,\prime}$$

$$\Sigma^{22} = \Lambda_2\Lambda_2' + v_2^2 \mathrm{I}_{J_2}$$

Dirichlet-Horseshoe (Dir-HS) prior

- Construct a Dirichlet-Horseshoe (Dir-HS) prior for columns $\boldsymbol{\lambda}_k$ of Λ to efficiently induce joint sparsity;

★★ For each k, $k = 1, \ldots, K$,

$$\begin{aligned}
\lambda_{mjk} \mid \phi_{mjk}, \tau_k, \zeta_{mjk} &\overset{indep}{\sim} \mathrm{N}(0, \zeta_{mjk}^2 \phi_{mjk} \tau_k), \\
\zeta_{mjk} &\overset{iid}{\sim} \mathrm{C}^+(0, 1), \\
\boldsymbol{\phi}_k = (\phi_{11k}, \ldots, \phi_{MJ_Mk}) \mid a_\phi &\overset{iid}{\sim} \mathrm{Dir}(a_\phi, \ldots, a_\phi), \\
\tau_k \mid a_\tau, b_\tau &\overset{iid}{\sim} \mathrm{Ga}(a_\tau, b_\tau / J)
\end{aligned}$$

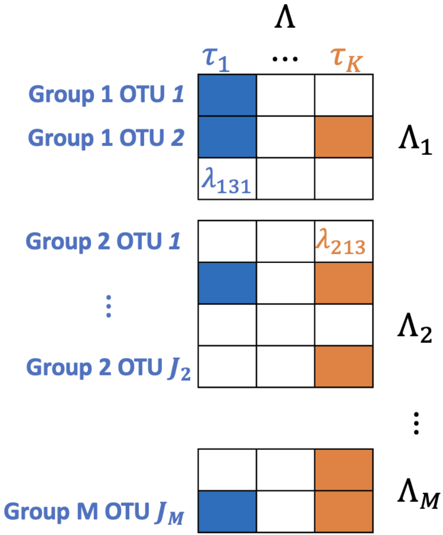
where $\mathrm{C}^+(0, 1)$ represents the half-Cauchy distribution for $\mathbb{R}_+$ with location and scale parameters 0 and 1.

Prior for Λ (contd)

$$\lambda_{mjk} \overset{indep}{\sim} \mathrm{N}(0, \zeta_{mjk}^2 \phi_{mjk} \tau_k)$$
$$\zeta_{mjk} \overset{iid}{\sim} \mathrm{C}^+(0, 1)$$
$$\Downarrow$$
$$\lambda_{mjk} \overset{indep}{\sim} \mathrm{HS}(\phi_{mjk} \tau_k)$$

with

$$\phi_k \overset{iid}{\sim} \mathrm{Dir}(a_\phi, \ldots, a_\phi)$$
$$\tau_k \overset{iid}{\sim} \mathrm{Ga}(a_\tau, b_\tau / J)$$



Dirichlet-Horseshor Prior

Theorem

Let $J = 2$. Assume $\phi_1 \sim Be(a_\phi, a_\phi)$ and let $\phi_2 = 1 - \phi_1$. Assume the Dir-HS distribution as a joint distribution for $\boldsymbol{\lambda} = (\lambda_1, \lambda_2) \in \mathrm{R}^2$ given τ. Without loss of generality, let $\tau = 1$. The marginal density $\Pi_{Dir\text{-}HS}(\lambda_1)$ of λ_1 satisfies: (a) $\lim_{\lambda_1 \to 0} \Pi_{Dir\text{-}HS}(\lambda_1) = \infty$. (b) For $\lambda_1 \neq 0$,

$$
\begin{aligned}
& 2^{2a_\phi - \frac{5}{2}} \pi^{-2} \frac{\Gamma^2(a_\phi + 1/2)}{\Gamma(2a_\phi + 1/2)} \frac{4}{\lambda_1^2} {}_3F_2\left(1, 1, a_\phi + 1/2; 2, 2a_\phi + 1/2; -\frac{4}{\lambda_1^2}\right) \\
& < \quad \Pi_{Dir\text{-}HS}(\lambda_1) \quad < \\
& 2^{2a_\phi - \frac{3}{2}} \pi^{-2} \frac{\Gamma^2(a_\phi + 1/2)}{\Gamma(2a_\phi + 1/2)} \frac{2}{\lambda_1^2} {}_3F_2\left(1, 1, a_\phi + 1/2; 2, 2a_\phi + 1/2; -\frac{2}{\lambda_1^2}\right),
\end{aligned}
$$

where ${}_pF_q$ is the generalized hypergeometric function, ${}_pF_q(\alpha_1, \ldots, \alpha_p; \beta_1, \ldots, \beta_q; x) = \sum_{t=0}^{\infty} \frac{(\alpha_1)_t \ldots (\alpha_p)_t}{(\beta_1)_t \ldots (\beta_q)_t} \frac{x^t}{t!}$.

Dirichlet-Horseshor Prior

Theorem

Especially when $a_\phi = \frac{1}{2}$,

$$\frac{1}{\sqrt{2\pi^5}}\left\{\sinh^{-1}(2/|\lambda_1|)\right\}^2 < \Pi_{Dir\text{-}HS}(\lambda_1) < \sqrt{\frac{2}{\pi^5}}\left\{\sinh^{-1}(\sqrt{2}/|\lambda_1|)\right\}^2$$

where the inverse hyperbolic sine function $\sinh^{-1}(x) = \log(x + \sqrt{x^2+1})$.

$\Rightarrow a_\phi = \frac{1}{4}$, *lower bound* $\propto \frac{\sqrt[4]{x^2+4}}{\sqrt{|x|}} - 1$

$\Rightarrow a_\phi = 1$, *lower bound* $\propto \log(\frac{4}{x^2}+1) + x\arctan(\frac{2}{x}) - 2$

- Dir-HS has an infinite spike at 0
- $\lim_{\lambda_1\to\infty} \Pi_{\text{Dir-HS}}(\lambda_1) = O(\frac{1}{\lambda_1^2})$

Dir-HS prior

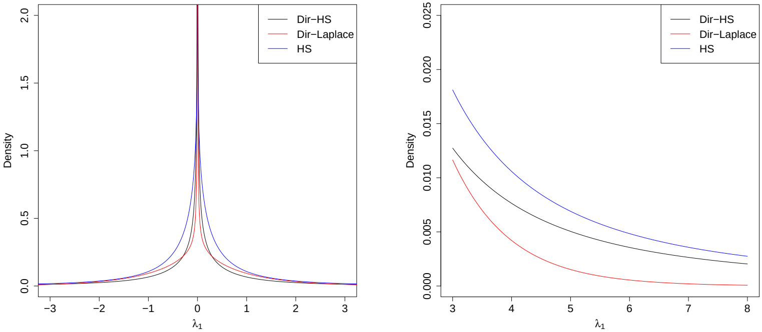


Figure: Marginal densities of λ_1 under $a_\phi = 1/20$.

- Dir-HS diverges faster than HS at 0 when $a_\phi = \frac{1}{4}, \frac{1}{2}$; the same rate when $a_\phi = \frac{3}{4}, 1$
- Dir-HS has a heavier tail than Dir-Laplace

Mean

- $$\mu_{imj} = r_{im} + \alpha_{s_i,mj} + \boldsymbol{\beta}'_{mj}\boldsymbol{x}_i$$

★★ r_{im}: sample scale factor for OTUs of group m in sample i.

Mean

subject	sample	OTU 1	OTU 2	⋯	OTU J_1	Total
1	1	μ_{imj}		⋯		5698
1	2			⋯		2312
1	3			⋯		9872
⋮	⋮		⋮		⋮	
S	N-2			⋯		598
S	N-1			⋯		532
S	N			⋯		2808

subject	sample	OTU 1	OTU 2	⋯	OTU J_2	Total
1	1			⋯		2102
1	2			⋯		743
1	3			⋯		3648
⋮	⋮		⋮		⋮	
S	N-2			⋯		2832
S	N-1			⋯		389
S	N			⋯		2231

r_{im}

Mean

- $$\mu_{imj} = r_{im} + \alpha_{s_i,mj} + \boldsymbol{\beta}'_{mj}\boldsymbol{x}_i$$

★★ r_{im}: sample scale factor for OTUs of group m in sample i.

$\Rightarrow$ Under log-N, $\text{Median}(y_{imj}) = \exp(\mu_{imj})$

$\frac{\text{Median}(y_{imj})}{\exp(r_{im})} = \exp(\alpha_{s_i,mj}) + \exp(\boldsymbol{\beta}'_{mj}\boldsymbol{x}_i)$

★★ $\alpha_{s_i,mj}$: normalized baseline abundance of OTU j belonging to group m for all samples from subject s_i.

Mean

subject	sample	OTU 1	OTU 2	⋯	OTU J_1	Total	OTU 1	OTU 2	⋯	OTU J_2	Total
1	1	μ_{imj}		⋯		5698			⋯		2102
1	2			⋯		2312			⋯		743
1	3			⋯		9872			⋯		3648
⋮	⋮		⋮		⋮			⋮		⋮	
S	N-2			⋯		598			⋯		2832
S	N-1			⋯		532			⋯		389
S	N			⋯		2808			⋯		2231

	r_{im}
$\alpha_{s_i,mj}$	

Mean

- Let

$$\mu_{imj} = r_{im} + \alpha_{s_i,mj} + \boldsymbol{\beta}'_{mj}\boldsymbol{x}_i$$

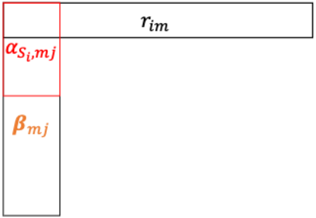
★★ r_{im}: sample scale factor for OTUs of group m in sample i.

★★ $\alpha_{s_i,mj}$: normalized baseline abundance of OTU j belonging to group m for all samples from subject s_i.

★★ $\boldsymbol{\beta}'_{mj}$: regression coefficients for OTU j of group m.

Mean

subject	sample	OTU 1	OTU 2	⋯	OTU J_1	Total	OTU 1	OTU 2	⋯	OTU J_2	Total
1	1	μ_{imj}		⋯		5698			⋯		2102
1	2			⋯		2312			⋯		743
1	3			⋯		9872			⋯		3648
⋮	⋮		⋮		⋮			⋮		⋮	
S	N-2			⋯		598			⋯		2832
S	N-1			⋯		532			⋯		389
S	N			⋯		2808			⋯		2231



Under log-N,

$$\text{Median}(y^*_{imj}) = \exp(r_{im} + \alpha_{S_i,mj} + \beta'_{mj}\, x_i)$$

Mean prior

- A mean-constrained prior with a mixture of mixture of normals on r_{im} and $\alpha_{s_i mj}$: (Li et al., 2017; Shuler et al., 2021)

$$r_{im} \overset{indep}{\sim} \sum_{l=1}^{\infty} \psi^r_{ml} \left\{ \omega^r_{ml} \mathrm{N}(\xi^r_{ml}, u_r^2) + (1-\omega^r_{ml}) \mathrm{N}\left(\frac{\nu^r_m - \omega^r_{ml}\xi^r_{ml}}{1-\omega^r_{ml}}, u_r^2 \right) \right\}$$

- Normalized baseline abundance of OTU j of group m in samples taken from subject s_i:

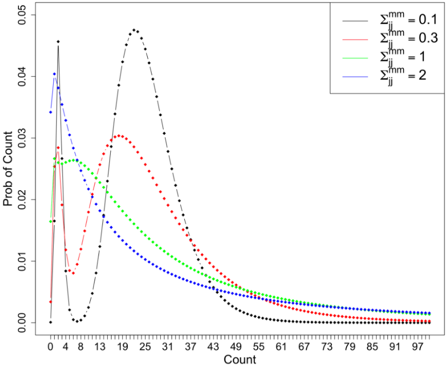
$$\begin{aligned}
\boldsymbol{\alpha}_{s_i} \mid G &\overset{iid}{\sim} G(\boldsymbol{\alpha}), \qquad s_i \in \{1, \ldots, S\}. \\
G(\boldsymbol{\alpha}) &= \prod_{m=1}^{M} \prod_{j=1}^{J_m} G_{mj}(\alpha_{mj}) \\
&= \prod_{m=1}^{M} \prod_{j=1}^{J_m} \left[\sum_{l=1}^{\infty} \psi^{\alpha}_{ml} \left\{ \omega^{\alpha}_{ml} \delta_{\xi^{\alpha}_{mjl}} + (1-\omega^{\alpha}_{ml}) \delta_{\left(\frac{\nu^{\alpha}_{mj} - \omega_{ml}\xi^{\alpha}_{mjl}}{1-\omega^{\alpha}_{ml}} \right)} \right\} \right]
\end{aligned}$$

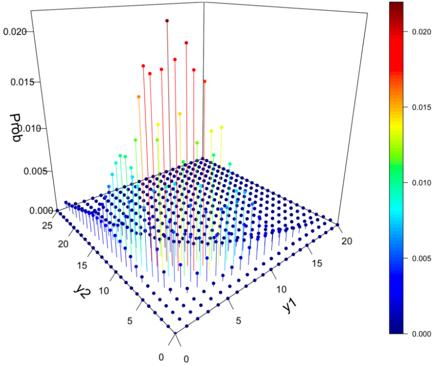
- location $\xi \overset{iid}{\sim}$ N, inner weights $\omega \overset{iid}{\sim}$ Be, outer weights $\psi \sim$ a stick-breaking process

Model for the Mean (contd)

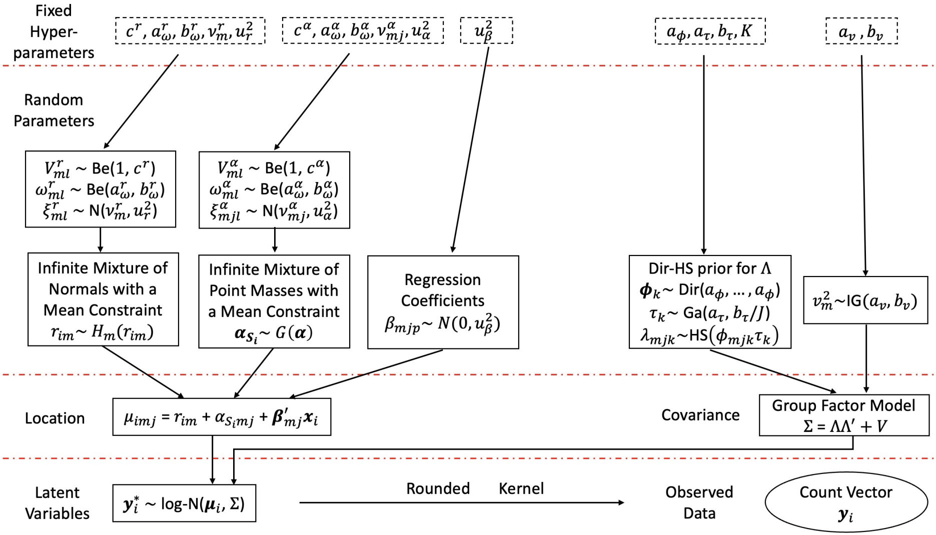
- For sample i from subject s_i, we assume

$$\mathbf{y}_i^\star \mid \boldsymbol{r}_i, \boldsymbol{\alpha}_{s_i}, \boldsymbol{\beta}, \Sigma \overset{indep}{\sim} \int \text{log-N}_J(\boldsymbol{r}_i + \boldsymbol{\alpha}_{s_i} + \boldsymbol{\beta}\mathbf{x}_i, \Sigma) dG(\alpha)$$





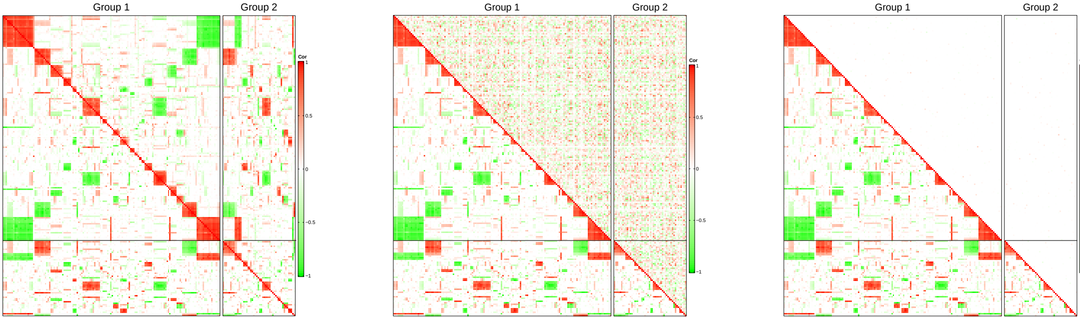
Sp-BGFM Model



Simulation Studies

- $M = 2$ groups, $N = 20$ samples, $J_1 = 150$, $J_2 = 50$ OTUs, baseline abundance level: -5, N(4, 1), N(10, 1)
- Sim 1: $\lambda_{mjk}^{\mathrm{tr}} \sim \mathrm{N}(0, 1) \pm 1$
- Sim 2,3 vine method ((Lewandowski, Kurowicka, and Joe, 2009) random correlation w & w/o covariates
- Sim 4 multinomial distribution without any associations
- Sim 5 SpiecEasi(Kurtz et al., 2015) without cross-domain association

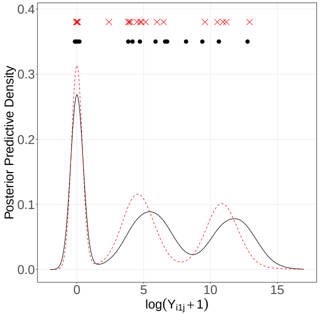
Simulation 2



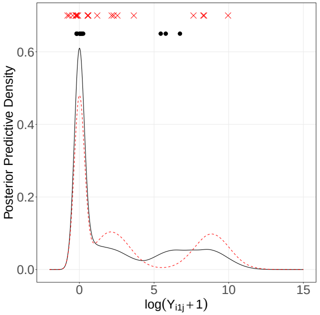
(a) Sp-BGFM (b) MOFA(Argelaguet et al., 2018) (c) SPIEC-EASI

Figure: [Simulation 2] The upper right and lower left triangles of a heatmap illustrate estimates $\hat{\rho}_{jj'}^{mm'}$ of correlations and their truth, respectively. The horizontal and vertical lines are to divide the groups. The estimates in panels (a)-(c) are from Sp-BGFM, MOFA and SPIEC-EASI, respectively.

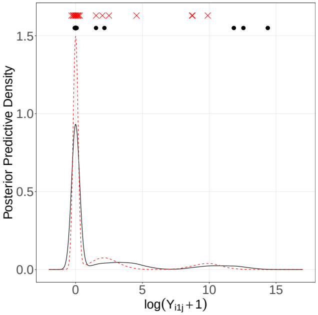
Posterior Predictive Checking



(a) Group 1 OTU 12



(e) Group 1 OTU 32

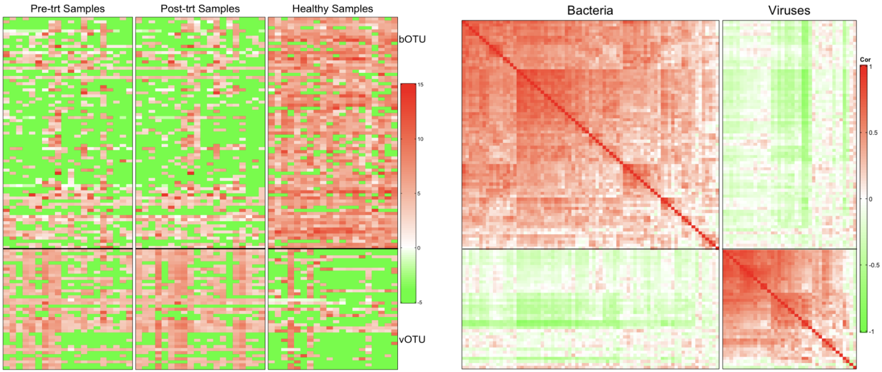


(f) Group 2 OTU 11

Multi-domain Skin Microbiome Data Analysis

- $S = 20$ patients at wound care clinic
- Three samples: pre- and post-treatment, and a control site on the healthy skin

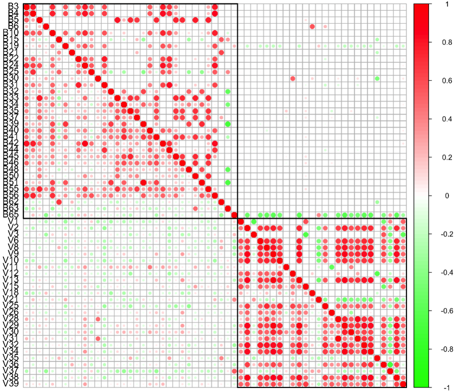
$\Rightarrow$ $N = 60$ samples from $S = 20$ subjects with a categorical covariate



(a) Log-transformed normalized OTU counts

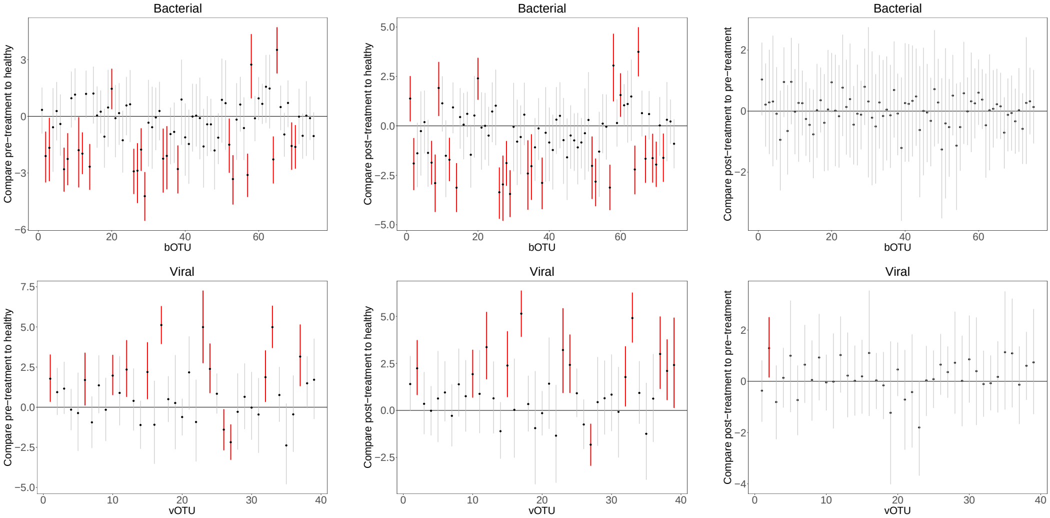
(b) Empirical correlation estimates

Multi-domain skin microbiome data



- bOTU 65(Staphylococcus aureus), a prominent skin pathogen
- *Pseudomonas* (b 59) and *Pseudomonas* phage (v 18) $\hat{\rho} = 0.38$

$\hat{\beta}_{mjp} - \hat{\beta}_{mjp'}$



Ongoing and future work

- Heteroskedasticity $\Sigma(x_i)$
- Longitudinal microbiome analysis
- Future work: tree-evolving count tables

Selected Reference

- Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2015). Dirichlet–Laplace Priors for Optimal Shrinkage. *Journal of the American Statistical Association*, 110:1479–1490.
- Canale, A. and Dunson, D. B. (2011). Bayesian Kernel Mixtures for Counts. *Journal of the American Statistical Association*, 106:1528–1539.
- Cao, Y., Lin, W., and Li, H. (2019). Large Covariance Estimation for Compositional Data via Composition-Adjusted Thresholding. *Journal of the American Statistical Association*, 114:759–772.
- Chung, H. C., Gaynanova, I., and Ni, Y. (2022). Phylogenetically Informed Bayesian Truncated Copula Graphical Models for Microbial Association Networks. *The Annals of Applied Statistics*, 16:2437–2457.
- Denti, F., Camerlenghi, F., Guindani, M., and Mira, A. (2023). A Common Atoms Model for the Bayesian Nonparametric Analysis of Nested Data. *Journal of the American Statistical Association*, 118(541):405–416. PMID: 37089274.