Appendices

A Tau-Path Algorithm

The method presented in Algorithm 1 is a modified version of the tau path algorithm presented in (Zhang et al, 2018).

Algorithm 1 Tau-Path Method for Association

Input: Abundance measure: $\mathcal{A}$ and Prevalence measure: $\mathcal{P}$ corresponding to all B bacteria types and $\pi : \{1, \ldots, B\} \to \{1, \ldots, B\}$ to be a permutation function.

Output: Tau-path $(t_1, ..., t_{B-1})$ of re-ordered pairs with strong association pulled up to the top of the list.

1. **for** r in $1, 2, \ldots, B$

    for s in $1, 2, \ldots, B$ **do**

    Define the concordance matrix $C_0 = \{c_{[r,s]}\}$ as,

$$c_{[r,s]} = \begin{cases} 1, & \text{if } (\mathcal{A}_r, \mathcal{P}_r), (\mathcal{A}_s, \mathcal{P}_s) \text{ are concordant} \\ -1, & \text{if } (\mathcal{A}_r, \mathcal{P}_r), (\mathcal{A}_s, \mathcal{P}_s) \text{ are discordant} \\ 0, & \text{if } r = s \end{cases}$$

2. **for** j in $1, 2, \ldots, B-1$ **do**

    (a) Define m_j denote the observation corresponding to the smallest row/column sum of C_{j-1}

    (b) **if** m_j is unique **then**
    Set $\pi(B+1-j) = m_j$

    else for l in $1, 2, \ldots, k$ where k of the m_j lead to the same sum **do**

    Calculate s_{jl}, the sum of the tau-path $t_l = \dfrac{\sum\sum_{1 \le r \le s \le l} c_{[r,s]}}{\binom{l}{2}}$ and set $\pi(B+1-j) = m_{jl^*}$ where $l^* = \underset{l}{\operatorname{argmax}}\{s_{jl}\}$

    (c) Obtain C_j to be the matrix formed by removing row and column m_j of C_{j-1}

3. Set $\pi(B)$ to be the last remaining observation number.

B LDA Variational Algorithm

The LDA model allows us to estimate the model parameters and compute the posterior distribution for inference. Although the posterior distribution is intractable for exact inference in LDA, we can estimate it with an approximate inference algorithm known as the Variational Inference algorithm described in Algorithm 2. The parameter estimation is done using Variational EM algorithm shown in Algorithm 3. Both these algorithms are adapted from (Blei et al, 2003).

Variational Inference (VI) Algorithm

Since the posterior distribution is intractable for exact inference, we approximate the posterior by considering a tractable family of lower bound indexed by *variational* parameters given by,

$$q(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{\gamma}, \boldsymbol{\phi}) = q(\boldsymbol{\theta}|\boldsymbol{\gamma}) \prod_{n=1}^{N} q(z_n|\phi_n) \tag{1}$$

where the symmetric Dirichlet parameter γ and multinomial parameters $(\phi_1, \ldots, \phi_N)$ are variational parameters. The optimizing values of the variational parameters are found by minimizing the Kullback- Leibler (KL) divergence between the variational distribution and the true posterior. Based on this optimization problem, the pair of update equations are derived are described in Algorithm 2. Once the variational parameters are estimated, the variational distribution is computed as an approximation to the true posterior.

Algorithm 2 Variational Inference (VI) Algorithm

Input: $\phi^0_{n_i} = \dfrac{1}{T}$ and $\gamma_i = \alpha_i + \dfrac{N}{T}$ $\forall i = 1, 2, \ldots, T$ and $n = 1, 2, \ldots, N$

repeat

for n in $1, 2, \ldots N$

for i in $1, 2, \ldots T$

$\phi^{(t+1)}_{n_i} = \beta_i w_n \exp[\psi(\gamma^{(t)}_i)]$

Normalize $\phi^{(t+1)}_n$ to sum to 1.

$\gamma^{(t+1)} = \alpha + \sum_{n=1}^{N} \phi^{(t+1)}_n$

until convergence

Variational Expectation-Maximization (VEM) Algorithm

In order to estimate the model parameters α and $\boldsymbol{\beta}$, we need to maximize the log-likelihood of the data, which is intractable. Therefore, an alternative variational EM procedure is considered that maximizes the lower bound with respect to the variational parameters γ and $\boldsymbol{\phi}$ on each subject, and then, for fixed values of the variational parameters, maximizes the lower bound with respect to the model parameters α and $\boldsymbol{\beta}$ as described in Algorithm 3.

Algorithm 3 Variational EM (VEM) Algorithm

---

repeat until convergence

E-Step: For each subject, obtain the optimizing variational parameters $\{\gamma_d, \phi_d : d \in \{1, 2, \dots M\}\}$. This is done using the Variational inference algorithm described in Algorithm 2.

M-Step: Maximize the lower bound on the log-likelihood with respect to α and $\boldsymbol{\beta}$

- $\beta_{i,j}^{(t+1)} = \sum_{d=1}^{M} \sum_{n}^{N_d} \phi_{d_{n_i}}^{(t+1)} w_{d_n}^{j}$
- Update $\alpha^{(t+1)}$ using linear time Newton-Raphson method.

References

Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022

Zhang Y, Ravishanker N, Ivan J, Mamun S (2018) An Application of the Tau-Path Method in Highway Safety. Journal of the Indian Society for Probability and Statistics 20