S1 Approximating a Poisson Process using Beta random variables

Consider approximating a Poisson process on $(0,1)$ with intensity $\nu(\sigma) = \alpha\sigma^{-1}(1-\sigma)^{-1/2}$ by a finite counting process formed by n iid samples drawn from $\text{Beta}(\epsilon_n, 1/2-\epsilon_n)$ where $\epsilon_n < 1/2$. Denote the Poisson process as $N(t)$ and the approximating process as $N'_n(t)$, we first calculate the probability of having m points in interval $(\delta, t]$, where $m \leq n$, $t < 1$ and $0 < \delta \ll 1$,

$$
\begin{aligned}
P\left[N((\delta,t]) = m\right] =& \frac{\left[\int_\delta^t \alpha\sigma^{-1}(1-\sigma)^{-1/2}d\sigma\right]^m}{m!}\exp\left(-\int_\delta^t \alpha\sigma^{-1}(1-\sigma)^{-1/2}d\sigma\right), \\
P\left[N'_n((\delta,t]) = m\right] =& \begin{pmatrix} n \\ m \end{pmatrix}\left(\frac{1}{\text{Beta}(\epsilon_n, 1/2-\epsilon_n)}\int_\delta^t \sigma^{-1+\epsilon_n}(1-\sigma)^{-1/2-\epsilon_n}d\sigma\right)^m \times \\
& \left(1 - \frac{1}{\text{Beta}(\epsilon_n, 1/2-\epsilon_n)}\int_\delta^t \sigma^{-1+\epsilon_n}(1-\sigma)^{-1/2-\epsilon_n}d\sigma\right)^{n-m}.
\end{aligned}
$$

The moment generating functions (MGFs) of $N((\delta,t])$ and $N'_n((\delta,t])$ are

$$
\begin{aligned}
M_N(\lambda) =& \exp\left[\left(e^\lambda - 1\right)\int_\delta^t \alpha\sigma^{-1}(1-\sigma)^{-1/2}d\sigma\right], \\
M_{N'_n}(\lambda) =& \left[\frac{e^\lambda - 1}{\text{Beta}(\epsilon_n, 1/2-\epsilon_n)}\int_\delta^t \sigma^{-1+\epsilon_n}(1-\sigma)^{-1/2-\epsilon_n}d\sigma + 1\right]^n.
\end{aligned}
$$

These two MGFs will be the same asymptotically if

$$
\lim_{n\to\infty} \frac{n}{\text{Beta}(\epsilon_n, 1/2-\epsilon_n)}\int_\delta^t \sigma^{-1+\epsilon_n}(1-\sigma)^{-1/2-\epsilon_n}d\sigma = \alpha\int_\delta^t \sigma^{-1}(1-\sigma)^{-1/2}d\sigma. \quad \text{(S1)}
$$

This will be satisfied when $\epsilon_n = \alpha/n$. Indeed, under this assumption, we have

$$
\lim_{n\to\infty} \frac{n\left(\sigma/(1-\sigma)\right)^{\epsilon_n}}{\text{Beta}(\epsilon_n, 1/2-\epsilon_n)} = \alpha.
$$

In addition, since when n is large enough, the map $n \mapsto \frac{n(\sigma/(1-\sigma))^{\epsilon_n}}{\text{Beta}(\epsilon_n, 1/2-\epsilon_n)}$ is a non-increasing function, by Lebesgue's monotone convergence theorem, we can establish the convergence of the left hand side of (S1) to the right hand side. Using this result, we can prove the weak convergence of the finite dimension distribution: $(N'(\delta,t_1],\ldots,N'(\delta,t_n]) \xrightarrow{d} (N(\delta,t_1],\ldots,N(\delta,t_n])$. This follows by a direct application of the multinomial theorem.

Now we need to verify the tightness condition, this is automatically satisfied as $N_n(t)'$ is a càdlàg process (Daley and Vere-Jones, 1988) (Theorem 11.1. VII and Proposition 11.1. VIII, iv, Volume 2). Therefore we prove the weak convergence of the process $N'_n(t)$ to the Poisson process $N(t)$ when $n \to \infty$ and $\epsilon_n = \alpha/n$.

S2 Proof of Proposition 1

We use the notation $P^j(\cdot) = \frac{\sum_i I(Z_i \in \cdot)\sigma_i Q_{i,j}^{+2}}{\sum_i \sigma_i Q_{i,j}^{+2}}$ where $Q_{i,j} = \langle \mathbf{X}_i, \mathbf{Y}^j \rangle$. Denote $((Q_{i,j}, Q_{i,j'}), i \geq 1)$ as $\mathbf{Q}$. The joint distribution of $(Q_{i,j}, Q_{i,j'})$ is a multivariate normal with mean $\mathbf{0}$ and covariance $\phi(j, j')$, and the vectors $(Q_{k,j}, Q_{k,j'})$, $k = 1, 2, \ldots,$ are independent. We derive an expression for the covariance

$$\begin{aligned}\mathrm{cov}[P^j(A), P^{j'}(A)] =& E[E[P^j(A)P^{j'}(A)|\sigma, \mathbf{Q}]] - E[P^j(A)]E[P^{j'}(A)] \\ =& (G(A) - G^2(A))E\left[\frac{\sum_i \sigma_i^2 Q_{i,j}^{+2} Q_{i,j'}^{+2}}{\sum_i \sigma_i Q_{i,j}^{+2} \sum_k \sigma_k Q_{k,j'}^{+2}}\right].\end{aligned}$$

Similarly, we can get the expression for the variance,

$$\mathrm{var}[P^j(A)] = (G(A) - G^2(A))E\left[\frac{\sum_i \sigma_i^2 Q_{i,j}^{+4}}{\sum_i \sigma_i Q_{i,j}^{+2} \sum_k \sigma_k Q_{k,j}^{+2}}\right].$$

It follows that

$$\mathrm{corr}[P^j(A), P^{j'}(A)] = E\left[\frac{\sum_i \sigma_i^2 Q_{i,j}^{+2} Q_{i,j'}^{+2}}{\sum_i \sigma_i Q_{i,j}^{+2} \sum_k \sigma_k Q_{k,j'}^{+2}}\right] \times \left(E\left[\frac{\sum_i \sigma_i^2 Q_{i,j}^{+4}}{\sum_i \sigma_i Q_{i,j}^{+2} \sum_k \sigma_k Q_{k,j}^{+2}}\right]\right)^{-1}.$$

Therefore the correlation is independent of the set A.

S3 Proof of Proposition 2

We follow the framework of proofs for Theorem 1 and Theorem 3 in Barrientos et al. (2012). Let $\mathcal{P}(\mathcal{Z})$ be the set of all Borel probability measures defined on $(\mathcal{Z}, \mathcal{F})$ and $\mathcal{P}(\mathcal{Z})^J$ the product space of J $\mathcal{P}(\mathcal{Z})$. Assume $\Theta \subset \mathcal{Z}$ is the support of G. To show the prior assigns strictly positive probability to the neighbourhood in Proposition 2, it is sufficient to show such neighbourhood contains certain subset-neighbourhoods with positive probability. As in Barrientos et al. (2012), we consider the subset-neighbourhoods U:

$$U(G_1, \ldots, G_J, \{A_{i,j}\}, \epsilon^*) = \prod_{i=1}^{J} \{F_i \in \mathcal{P}(\Theta) : |F_i(A_{i,j}) - G_i(A_{i,j})| < \epsilon^*, j = 1, \ldots, m_i\},$$

where G_i is a probability measure absolutely continuous w.r.t. G for $i = 1, \ldots, J$, $A_{i,1}, \ldots, A_{i,m_i} \subset \Theta$ are measurable sets with G_i-null boundary and $\epsilon^* > 0$. The existence of such subset-neighbourhoods is proved in Barrientos et al. (2012). We then define sets $B_{\nu_{1,1}\ldots\nu_{m_J,J}}$ for each $\nu_{i,j} \in \{0, 1\}$ as

$$B_{\nu_{1,1}\ldots\nu_{m_J,J}} = \bigcap_{i=1}^{J} \bigcap_{j=1}^{m_i} A_{i,j}^{\nu_{i,j}},$$

where $A_{i,j}^1 = A_{i,j}$ and $A_{i,j}^0 = A_{i,j}^c$. Set

$$J_\nu = \{\nu_{1,1} \dots \nu_{m_J,J} : G(B_{\nu_{1,1},\dots,\nu_{m_J,J}}) > 0\},$$

and let $\mathcal{M}$ be a bijective mapping from J_ν to $\{0, \dots, k\}$ where $k = |J_\nu| - 1$. We can simplify the notation using $A_{\mathcal{M}(\nu)} = B_\nu$ for every $\nu \in J_\nu$. Define a vector $\mathbf{s}_i = (w_{i,0}, \dots, w_{i,k}) = (Q_i(A_0), \dots, Q_i(A_k))$ that belongs to the $k-$simplex Δ_k. Set

$$B(\mathbf{s}_i, \epsilon) = \{(w_0, \dots, w_k) \in \Delta_k : |Q_i(A_j) - w_j| < \epsilon, j = 0, \dots, k\},$$

where $\epsilon = 2^{-\sum_{i=1}^J m_i} \epsilon^*$. The derivation in Barrientos et al. (2012) suggests a sufficient condition for assigning positive mass to $U(G_1, \dots, G_J, \{A_{i,j}\}, \epsilon^*)$ is

$$\Pi([P^i(A_0), \dots, P^i(A_k)] \in B(\mathbf{s}_i, \epsilon), i = 1, \dots, J) > 0. \tag{S2}$$

Here Π is the prior.

Now consider the following conditions

C.1 $w_{i,l} - \epsilon_0 < \sigma_{l+1} Q_{l+1,i}^{+2} < w_{i,l} + \epsilon_0$ for $i = 1, \dots, J$ and $l = 0, \dots, k$.

C.2 $0 < \sum_{l>k+1} \sigma_l Q_{l,i}^{+2} < \epsilon_0$.

C.3 $Z_{l+1} \in A_l$ for $l = 0, \dots, k$.

ϵ_0 in the above conditions satisfies the following inequality

$$\frac{w_{(i,l)} - \epsilon_0}{1 + (k+2)\epsilon_0} \geq w_{(i,l)} - \epsilon$$
$$\frac{w_{(i,l)} + 2\epsilon_0}{1 - (k+1)\epsilon_0} \leq w_{(i,l)} + \epsilon$$

for $i = 1, \dots, J$ and $l = 0, \dots, k$. This system of inequalities can be satisfied when k is large enough. If conditions $(C.1)$ to $(C.3)$ hold, it follows that $[P^i(A_0), \dots, P^i(A_k)] \in B(\mathbf{s}_i, \epsilon)$ for $i = 1, \dots, J$. Therefore, we have

$$\begin{aligned}
&\Pi([P^i(A_0), \dots, P^i(A_k)] \in B(\mathbf{s}_i, \epsilon), i = 1, \dots, J) \geq \\
&\prod_{l=0}^{k} \Pi(w_{(i,l)} - \epsilon_0 < \sigma_{l+1} Q_{l+1,i}^{+2} < w_{(i,l)} + \epsilon_0, i = 1, \dots, J) \times \\
&\Pi(\sum_{l>k+1} \sigma_l Q_{l,i}^{+2} < \epsilon_0, i = 1, \dots, J) \times \\
&\prod_{l=0}^{k} \Pi(Z_{l+1} \in A_l) \times \Pi(Z_l \in \mathcal{Z}, l = k+2, \dots).
\end{aligned}$$

Since $(Q_{l,1}, \dots, Q_{l,J})$ are multivariate normal random vectors with strictly positive definite covariance matrix and σ_l are always positive, the vector $(\sigma_{l+1} Q_{l+1,i}^{+2}, i =$

$1, \ldots, J)$ has full support on $\mathbb{R}^{+J}$ and will assign positive probability to any subset of the space. If follows that

$$\Pi(w_{i,l} - \epsilon_0 < \sigma_{l+1} Q_{l+1,i}^{+2} < w_{i,l} + \epsilon_0, i = 1, \ldots, J) > 0 \text{ for } l = 0, \ldots, k.$$

Using the Gamma process argument, we know $\sum_{l>k+1} \sigma_l Q_{l,i}^{+2}$ is the tail probability mass for a well-defined Gamma process and thus will always be positive and continuous for all i. It follows that

$$\Pi(\sum_{l>k+1} \sigma_l Q_{l,i}^{+2} < \epsilon_0, i = 1, \ldots, J) > 0.$$

Since $\mathcal{Z}$ is the topological support of G, it follows that $P(Z_{i+1} \in A_i) > 0$ and $P(Z_i \in \mathcal{Z}) = 1$. Combining these facts, we prove that Equation (S2) holds.

S4 Total variation bound of Laplace approximate of $p(Q_{i,j}|\mathbf{Q}_{i,-j}, \boldsymbol{\sigma}, \mathbf{T}, \mathbf{n})$

We consider the class of densities $g(x; k, \mu, s^2)$

$$g(x; k, \mu, s^2) \propto I(x \geq 0) x^{2k} f(x; \mu, s^2), k \in \mathbb{N}^+$$

where $f(x; \mu, s^2)$ is the density function of $N(\mu, s^2)$. The Laplace approximation of $g(x; k, \mu, s^2)$ is written as $f(x; \widehat{\mu}, \widehat{s}^2)$. Here $\widehat{\mu} = \text{argmax}_x g(x; k, \mu, s^2)$ and $\widehat{s}^2 = -\left(\left(\partial^2 \log(g)/\partial x^2\right)|_{\widehat{\mu}}\right)^{-1}$. We want to calculate the total variation distance between density $f(x; \widehat{\mu}, \widehat{s}^2)$ and $g(x; k, \mu, s^2)$, denoted as $d_{TV}(f(x; \widehat{\mu}, \widehat{s}^2), g(x; k, \mu, s^2))$.

Define class of functions $V(x; k, \mu)$ for $k \in \mathbb{N}^+, \mu > 0$:

$$V(x; k, \mu) = \begin{cases} 2k\left[\log(x/\mu) - (x/\mu - 1) + \frac{1}{2}(x/\mu - 1)^2\right] & x > 0 \\ -\infty & x \leq 0 \end{cases}$$

This function is non-decreasing and when $x = \mu$, $V(x; k, \mu) = 0$, $dV/dx = 0$ and $d^2V/dx^2 = 0$.

It follows that

$$\log g(x; k, \mu, s^2) - \log f(x; \widehat{\mu}, \widehat{s}^2) = V(x; k, \widehat{\mu}) + a_0 + a_1 x + a_2 x^2.$$

Moreover, since the $\widehat{\mu}$ is the mode of both $g(x; k, \mu, s^2)$ and $f(x; \widehat{\mu}, \widehat{s}^2)$, and the second derivative of $\log g(x; k, \mu, s^2)$ and $\log f(x; \widehat{\mu}, \widehat{s}^2)$ are identical at $x = \widehat{\mu}$, we can find that $a_1 = a_2 = 0$. Hence,

$$\log g(x; k, \mu, s^2) - \log f(x; \widehat{\mu}, \widehat{s}^2) = V(x; k, \widehat{\mu}) + a_0$$

and $g(x; k, \mu, s^2) = \exp\left(V(x; k, \widehat{\mu}) + a_0\right) f(x; \widehat{\mu}, \widehat{s}^2)$.

Since $V(x;k,\widehat{\mu})$ is monotone increasing, the total variation distance between $g(x;k,\mu,s^2)$ and $f(x;\widehat{\mu},\widehat{s}^2)$ can be expressed as

$$\begin{aligned} d_{TV}(g(x;k,\mu,s^2),f(x;\widehat{\mu},\widehat{s}^2)) &= \int_{x_0}^{+\infty} \left[\exp\left(V(x;k,\widehat{\mu})+a_0\right)-1\right] f(x;\widehat{\mu},\widehat{s}^2)dx \\ &= \int_{-\infty}^{x_0} \left[1-\exp\left(V(x;k,\widehat{\mu})+a_0\right)\right] f(x;\widehat{\mu},\widehat{s}^2)dx \end{aligned}$$

where $V(x_0;k,\widehat{\mu})=-a_0$. If $a_0 \leq 0$, we have $x_0 \geq \widehat{\mu}$ and

$$\begin{aligned} &\int_{x_0}^{+\infty} \left[\exp\left(V(x;k,\widehat{\mu})+a_0\right)-1\right] f(x;\widehat{\mu},\widehat{s}^2)dx \\ \leq &\int_{x_0}^{+\infty} \left[\exp\left(V(x;k,\widehat{\mu})\right)-1\right] f(x;\widehat{\mu},\widehat{s}^2)dx \\ \leq &\int_{\widehat{\mu}}^{+\infty} \left[\exp\left(V(x;k,\widehat{\mu})\right)-1\right] f(x;\widehat{\mu},\widehat{s}^2)dx \end{aligned}$$

Similarly, if $a_0 \geq 0$, we have

$$\int_{-\infty}^{x_0} \left[1-\exp\left(V(x;k,\widehat{\mu})+a_0\right)\right] f(x;\widehat{\mu},\widehat{s}^2)dx \leq \int_{-\infty}^{\widehat{\mu}} \left[1-\exp\left(V(x;k,\widehat{\mu})\right)\right] f(x;\widehat{\mu},\widehat{s}^2)dx$$

To summarize, we have

$$\begin{aligned} d_{TV}(g(x;k,\mu,s^2),f(x;\widehat{\mu},\widehat{s}^2)) \leq \max\Bigg(&\int_{\widehat{\mu}}^{+\infty} \left[\exp\left(V(x;k,\widehat{\mu})\right)-1\right] f(x;\widehat{\mu},\widehat{s}^2)dx, \\ &\int_{-\infty}^{\widehat{\mu}} \left[1-\exp\left(V(x;k,\widehat{\mu})\right)\right] f(x;\widehat{\mu},\widehat{s}^2)dx\Bigg) \end{aligned}$$

As we have shown in Equation (12) of the main manuscript, $\widehat{s}^2 = \left(\frac{2k}{\widehat{\mu}^2}+C\right)^{-1}$, where $C>0$. This suggests that $\widehat{s} \leq \widehat{\mu}/\sqrt{2k}$. Therefore

$$\begin{aligned} d_{TV}(g(x;k,\mu,s^2),f(x;\widehat{\mu},\widehat{s}^2)) \leq \max\Bigg(&\int_{\widehat{\mu}}^{+\infty} \left[\exp\left(V(x;k,\widehat{\mu})\right)-1\right] f(x;\widehat{\mu},\widehat{\mu}/2k)dx, \\ &\int_{-\infty}^{\widehat{\mu}} \left[1-\exp\left(V(x;k,\widehat{\mu})\right)\right] f(x;\widehat{\mu},\widehat{\mu}/2k)dx\Bigg) \end{aligned}$$

Since $V(x;\mu,s^2)$ and $f(x;\mu,s^2)$ are location-scale families, the above expression can be made free of $\widehat{\mu}$ and thus μ and s^2:

$$\begin{aligned} d_{TV}(g(x;k,\mu,s^2),f(x;\widehat{\mu},\widehat{s}^2)) \leq \max\Bigg(&\int_{1}^{+\infty} \left[\exp\left(V(x;k,1)\right)-1\right] f(x;1,1/2k)dx, \\ &\int_{-\infty}^{1} \left[1-\exp\left(V(x;k,1)\right)\right] f(x;1,1/2k)dx\Bigg) \end{aligned} \tag{S3}$$

This upper bound on the total variation distance decreases as k increases and it goes to 0 as $k \to \infty$. This suggests the convergence of the approximating normal distribution to the density family g in total variation sense. We also plot this upper bound as a function of k to verify the conclusion. It is shown in the supplemental Figure S1.

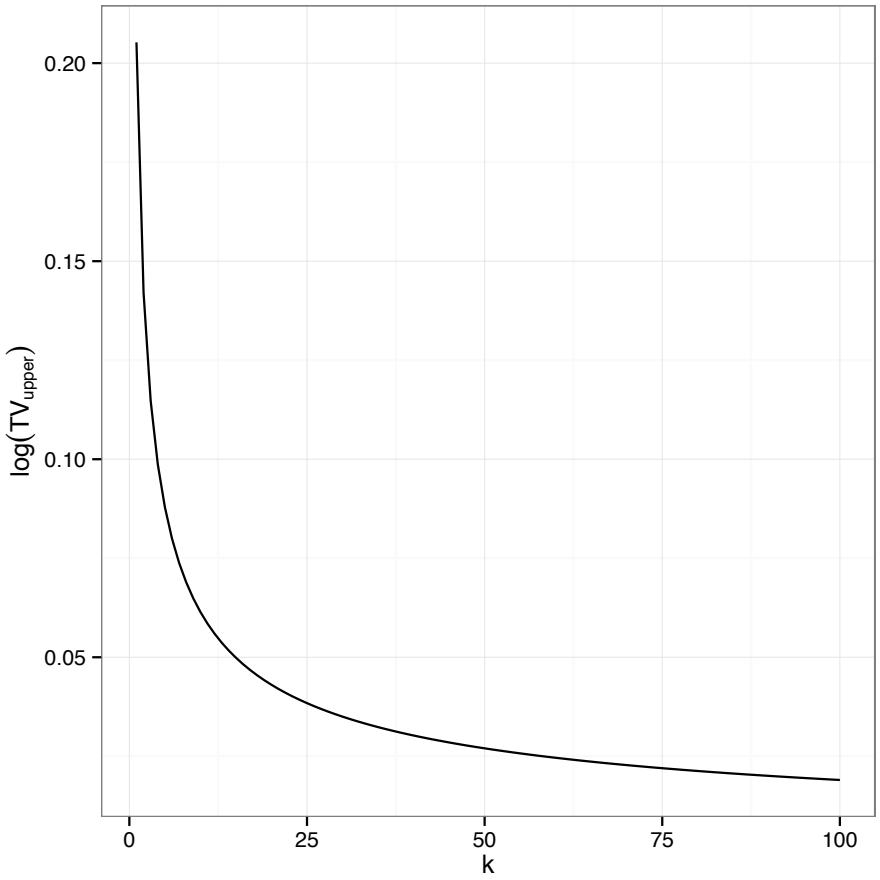

Figure S1: Upper bound of the total variation distance of Laplace approximation in (12) to the density in (11) as given in (S3) when frequency k increases.

S5 Details of self-consistent estimates in Section 3.1

First we estimate $\boldsymbol{\sigma}$ and then we transform the data $n_{i,j}$ into $\sqrt{n_{i,j}/\sigma_i}$. If $n_{i,j}$ is representative and $\boldsymbol{\sigma}$ is estimated accurately, we have $\sqrt{n_{i,j}/\sigma_i} = c_j Q^+_{i,j}$. If the covariance matrix of $\mathbf{Q}_i$ is $\boldsymbol{\Sigma}$, then the covariance matrix of $(\sqrt{n_{i,j}/\sigma_i}, j = 1, \ldots, J)$ will be $\widetilde{\boldsymbol{\Sigma}} = \boldsymbol{\Lambda}\boldsymbol{\Sigma}\boldsymbol{\Lambda}$ where $\boldsymbol{\Lambda} = diag\{c_1, \ldots, c_J\}$.

It is obvious that $(\sqrt{n_{i,j}/\sigma_i}, j = 1, \ldots, J)$ is MVN and the correlation matrix will

be the same as the induced correlation matrix from $\mathbf{\Sigma}$. Methods on identifying the covariance matrix using this truncated dataset are abundant and well-studied. One way to do it is the EM algorithm. This estimated covariance matrix will by no means to be the same as $\mathbf{\Sigma}$, but the induced correlation matrix will be very close to the true correlation matrix induced by $\mathbf{\Sigma}$. Hence if our interest is on estimating correlation matrix, we can just treat $(\sqrt{n_{i,j}/\sigma_i}, j = 1, \ldots, J)$ as the truncated version of the true $\mathbf{Q}_i$ and proceed.

The EM algorithm should then be derived for the following settings. Let $\mathbf{Q}_i \overset{iid}{\sim} MVN(\mathbf{0}, \mathbf{\Sigma})$. Instead of observing I independent $\mathbf{Q}_i$, we only observe the positive entries in each $\mathbf{Q}_i$ and know the rest of the entries are negative. Denote the observed data vector as $\widetilde{\mathbf{Q}}_i$. We want to estimate $\mathbf{\Sigma}$ from the data $\widetilde{\mathbf{Q}}_i, i = 1, \ldots, I$. A standard EM algorithm can be easily formulated as following:

E-step Get the conditional expectation of full data log likelihood, given the observed data. Define two index sets, $\mathcal{A}_i = \{j|\widetilde{Q}_{i,j} > 0\}$ and $\mathcal{B}_i = \{j|\widetilde{Q}_{i,j} = 0\}$. For an arbitrary index set $\mathcal{I}$, denote $Q_{\mathcal{I}} = (Q_{i,j}|j \in \mathcal{I})$. Denote $\mathcal{A} = \{(i,j)|j \in \mathcal{A}_i, i = 1, \ldots, I\}$ and $\mathcal{B} = \{(i,j)|j \in \mathcal{B}_i, i = 1, \ldots, I\}$. The E-step function at $t+1$ iteration is,

$$L(\Sigma|\Sigma_t) = \mathbb{E}\left[-\frac{I}{2}\log|\mathbf{\Sigma}| - \frac{1}{2}\mathrm{Tr}(\mathbf{\Sigma}^{-1}\sum_i \mathbf{Q}_i\mathbf{Q}_i')|\mathbf{\Sigma}_t, Q_{\mathcal{A}} = \widetilde{Q}_{\mathcal{A}}, Q_{\mathcal{B}} < 0\right].$$

Notice this expectation is not easy to calculate in general. We use instead Monte Carlo method to approximate it. We sample K copies of $\mathbf{Q}_i$ from the conditional distribution $(\mathbf{Q}_i|Q_{\mathcal{A}_i} = \widetilde{Q}_{\mathcal{A}_i}, Q_{\mathcal{B}_i} < 0)$ where $\mathbf{Q}_i \sim MVN(\mathbf{0}, \mathbf{\Sigma}_t)$. The conditional distribution is a truncated multivariate normal distribution and we use the R package `tmvtnorm` (Wilhelm, 2015) to sample from it. If we denote by $\mathbf{Q}_i^1, \ldots, \mathbf{Q}_i^K$ the K samples of Q_i, L can be approximated as

$$\widehat{L}(\Sigma|\Sigma_t) = -\frac{1}{K}\sum_{k=1}^{K}\left[\mathrm{Tr}(\mathbf{\Sigma}^{-1}\sum_i \mathbf{Q}_i^k(\mathbf{Q}_i^k)')\right] - \frac{I}{2}\log|\mathbf{\Sigma}|.$$

M-step We seek to maximize $\widehat{L}$ with respect to $\mathbf{\Sigma}$. Due to a well-known fact on the maximum likelihood estimate of covariance matrix of multivariate normal, it is straightforward to get

$$\mathbf{\Sigma}_{t+1} = \frac{1}{IK}\sum_{i,k}\mathbf{Q}_i^k(\mathbf{Q}_i^k)'.$$

We applied this algorithm to the simulated datasets generated for Figure 3(a) to estimate the normalized Gram matrix $\mathbf{S}$. A summary of the RV-coefficients between the estimates from the above algorithm and the truth is shown in Figure S2. We also compared the estimates from this algorithm with those from MCMC simulations in Figure S2. The estimates of $\mathbf{S}$ from MCMC simulation are always better than those given by the self-consistent algorithm but both perform very well.

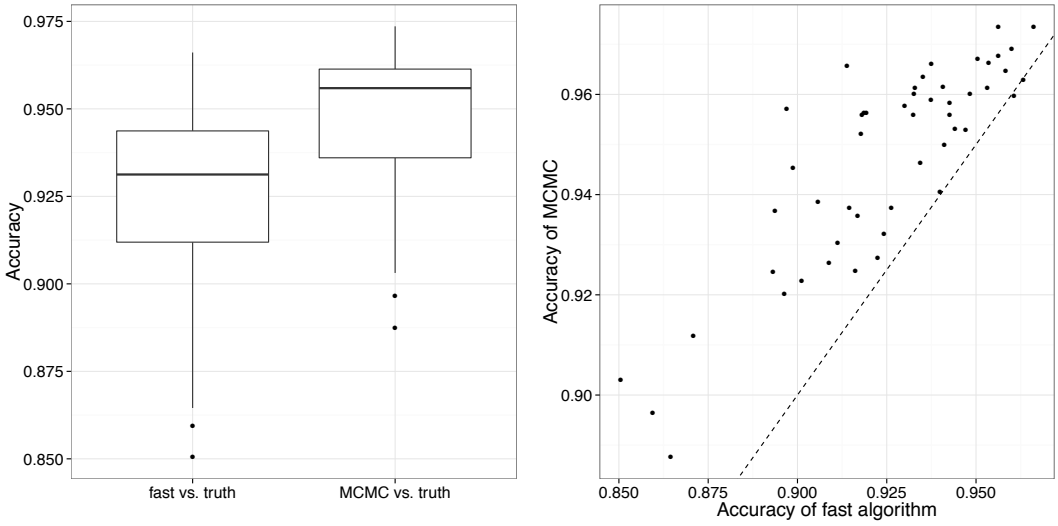

Figure S2: (**Left**) Box-plots compare the distributions of RV-coefficients between estimates from our self-consistent algorithm and between estimates from MCMC simulation and truth. (**Right**) Scatter plot to show per simulation comparison of RV coefficients for the self-consistent algorithm and MCMC sampling. Dashed line indicates where the two algorithms have identical accuracy.

S6 Standard PCoA for ordination of simulated dataset, Global Patterns dataset and Ravel's vaginal microbiome dataset

In this section, we include three sets of ordination figures generated using the standard PCoA method in microbiome studies. We first calculate the dissimilarity matrix of biological samples by applying Bray-Curtis dissimilarity metric on the empirical microbial distributions. We then perform classic Multi-dimensional Scaling (MDS) to ordinate biological samples based on the dissimilarity matrix. In Figure S3, we show the PCoA result for the simulated dataset generated for Figure 3(f). In Figure S4 and S5, we illustrate the PCoA results for the Global Patterns dataset and Ravel's vaginal microbiome dataset respectively. To be consistent with the main results, we show the ordination results based on the first three principal coordinates for the Global Patterns dataset and Ravel's vaginal microbiome dataset.

S7 Benchmarking the MCMC sampler

In this section, we focus on evaluating the computational performance of our MCMC sampler. We first consider the computational time of the sampler under different scenarios. We then illustrated a convergence diagnosis to check whether the sampler has reached mixing in the setting of our simulation study in the main manuscript. In

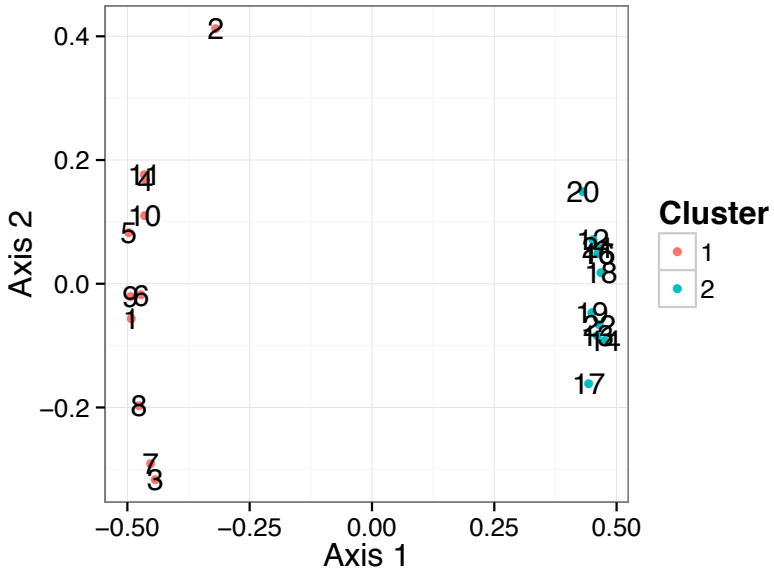

Figure S3: PCoA result for the simulated dataset generated for Figure 3(f).

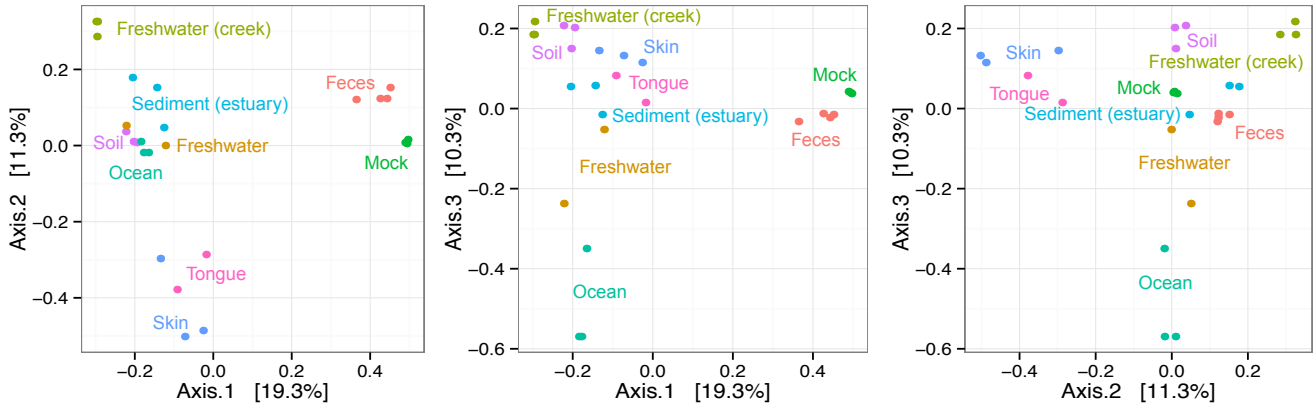

Figure S4: PCoA results for the Global Patterns dataset. We show the three two-dimensional representations of the ordination given by the first three principal coordinates.

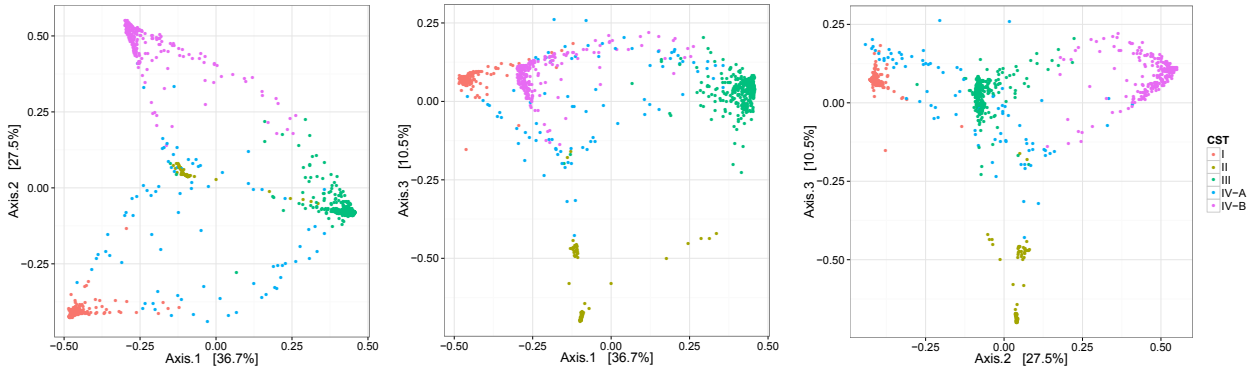

Figure S5: PCoA results for Ravel's vaginal microbiome dataset. We show the three two-dimensional representations of the ordination given by the first three principal coordinates.

addition, we created two larger datasets to verify the number of iterations needed to reach mixing will not be compromised if the underlying latent structure remains low dimensional.

S7.1 Computation time of the MCMC sampler

In Table S1 we listed the elapsed time in seconds for the MCMC sampler to finish $1,000$ iterations under different scenarios. All the scenarios are run with a single thread on a MacBook Pro with 2.7GHz Intel Core i5 and 8 GB 1867 MHz DDR3 RAM. In particular, we evaluated the effect of the number of biological samples (J), the number of species (I), the dimension of the latent factors (m), and the total counts per biological sample (n^j).

Table S1: Computation time (in seconds) of 1,000 iterations for the MCMC sampler

		$I = 68$			$I = 500$			$I = 1000$		
		$m = 5$	$m = 10$	$m = 20$	$m = 5$	$m = 10$	$m = 20$	$m = 5$	$m = 10$	$m = 20$
$J = 22$	$n^j = 10^3$	2.3	2.8	2.4	5.7	5.8	7.0	11.4	10.4	12.6
	$n^j = 10^4$	1.3	1.6	1.9	5.7	5.5	6.4	8.7	8.8	11.3
	$n^j = 10^5$	1.1	1.4	1.5	4.7	3.9	6.3	7.2	8.2	11.5
$J = 100$	$n^j = 10^3$	3.6	3.7	5.5	11.5	14.6	17.1	21.8	21.0	30.2
	$n^j = 10^4$	3.3	3.7	5.4	11.5	12.1	20.4	18.1	21.1	29.5
	$n^j = 10^5$	3.4	4.0	5.5	12.3	18.9	17.8	19.2	21.5	31.1
$J = 1000$	$n^j = 10^3$	31.4	34.3	49.6	121.2	118.4	152.1	152.1	173.8	251.0
	$n^j = 10^4$	28.2	33.4	53.1	96.3	144.3	159.7	143.7	164.8	254.2
	$n^j = 10^5$	40.1	38.2	52.2	129.1	111.5	138.2	163.2	171.7	246.0

Increasing the total number of reads per biological sample (n^j) does not affect the computation time. On the other hand, there is a weak effect associated with the dimension of the latent factors (m). In general, the computation time tends to

increase with m. The number of species (I) and the number of biological samples (J) affect the speed of computation significantly. These results illustrate that the MCMC sampler can finish $50,000$ iterations for a dataset with 100 samples and 1000 species in less than 20 minutes.

The table illustrates that it is possible to apply our model to microbiome datasets with comparable numbers of biological samples. It is rare to have datasets with more than a thousand confidently assigned OTUs (Callahan et al., 2016).

S7.2 Convergence diagnosis of the MCMC sampler

We evaluate the convergence of the MCMC sampler in the setting of Section 5 (simulation study). The number of biological samples is fixed at $J = 22$. We ran three parallel chains for three scenarios $I = 68$, $I = 500$ and $I = 1,000$. For each different I, we obtain the posterior samples of the first three eigenvalues of the normalized Gram matrix $\mathbf{S}$ in all three chains and use $\widehat{R}$ statistics (Gelman and Rubin, 1992) to check if the chains reached mixing. We chose to visualize the eigenvalues of $\mathbf{S}$ since in our model $\mathbf{S}$ is identifiable. The results are shown in Figure S6.

The $\widehat{R}$ statistics are all close to one supporting good MCMC mixing after 20,000 iterations, so our choice of 50,000 total iterations seems reasonable for providing posterior inference.

References

Barrientos, A. F., A. Jara, F. A. Quintana, et al. (2012). On the support of maceacherns dependent dirichlet processes and extensions. *Bayesian Analysis 7*(2), 277–310.

Callahan, B. J., P. J. McMurdie, M. J. Rosen, A. W. Han, A. J. A. Johnson, and S. P. Holmes (2016). Dada2: High-resolution sample inference from illumina amplicon data. *Nature methods 13*(7), 581–583.

Daley, D. J. and D. Vere-Jones (1988). An introduction to the theory of point processes.

Gelman, A. and D. B. Rubin (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 457–472.

Wilhelm, G. S. with contributions from Manjunath, B. (2015, August). tmvtnorm: Truncated Multivariate Normal and Student t Distribution.

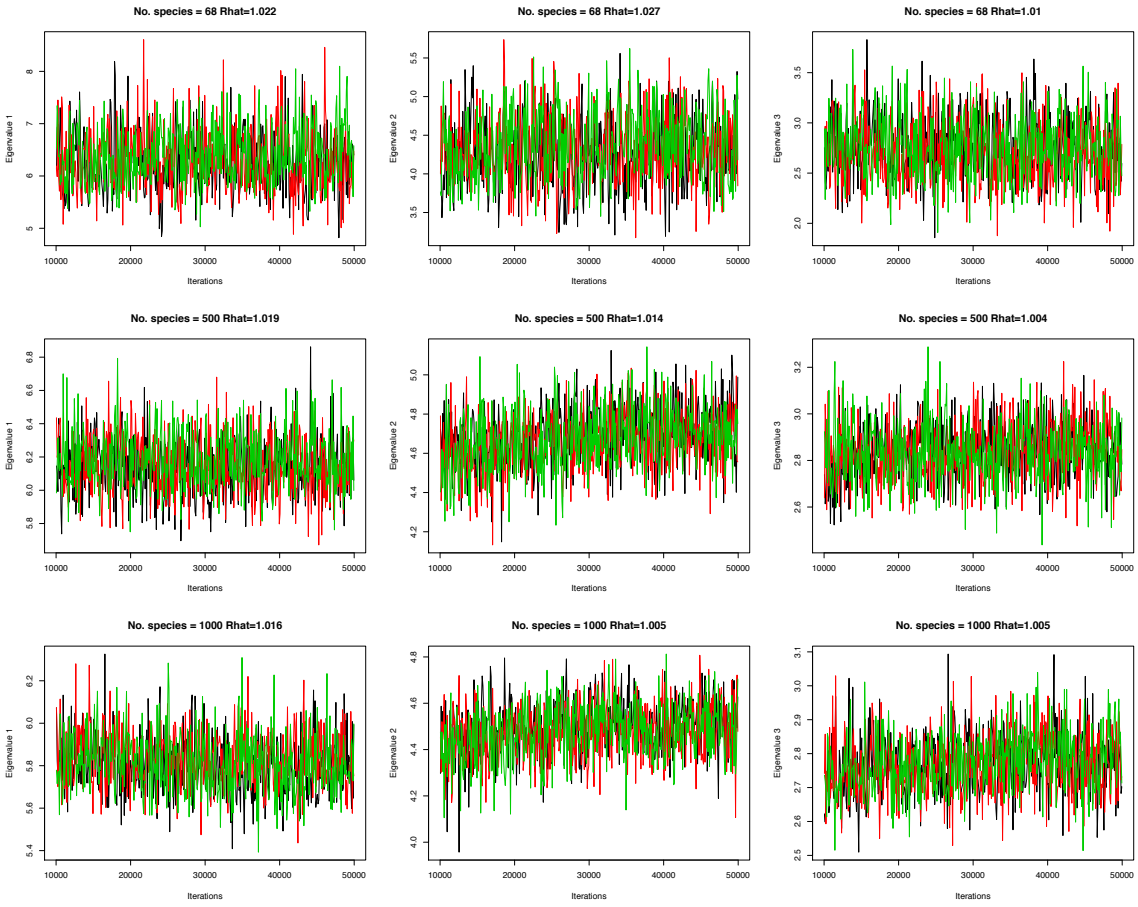

Figure S6: Traceplots for the posterior samples of the first three eigenvalues of $\mathbf{S}$. Each row corresponds to a different I and each column to a different eigenvalue. The $\widehat{R}$ statistics are shown in the title of each figure.