Temporal probabilistic modeling of bacterial compositions derived from 16S rRNA sequencing

Supplemental Material

Tarmo Åijö, Christian L. Müller, and Richard Bonneau

August 1, 2017

Supplemental Notes

Synthetic data generation

We generate synthetic data that reflects realistic dynamics and distribution of abundances by matching features of the synthetic data generation with real data, including fraction of zeros, time series structure, and number of taxa. Specifically, we consider the Subject A time series from David et al. (2014) and match the relative abundances and dynamics of taxa in synthetic data using real data:

1. We calculate the proportion estimates of bacterial orders/families/genera in real data at individual time points under the multinomial model.
2. We filter the proportion estimates series using a running median filter of length 15
   \[ \hat{y}_{filt,t} = \text{median}(\{y_{t-7}, y_{t-6}, \ldots, y_{t-1}, y_t, y_{t+1}, y_{t+2}, \ldots, y_{t+6}, y_{t+7}\}) \]
   in order to reduce the amount of noise present in estimates. The filtered estimates are re-normalized to ensure that they sum up to one at each time point.
3. We discard those bacterial orders/families/genera that are lowly abundant (average proportion is less than a threshold) followed by a re-normalization step leaving us noise-free relative abundances of 34 bacterial orders (threshold $5 \times 10^{-5}$) (lines in Supplemental Figure 1 in Additional file 2), 67 bacterial families ($5 \times 10^{-5}$), 109 bacterial genera ($5 \times 10^{-5}$), and 226 bacterial species ($1 \times 10^{-5}$).
4. We transform the simplex-valued estimates to real space using the inverse softmax function to add noise and sampling zeros.
5. We add Gaussian distributed noise with zero-mean and standard deviation (SD) $\sigma_n = 0.5$ and impose a predefined number of sampling zeros by setting corresponding log odds ratios to -10, i.e., to a value that is much smaller than the other values.
6. Noisy relative abundances are obtained by projecting the values onto the simplex using the softmax function (an example with 34 bacterial orders is illustrated Supplemental Figure 1)
7. Noisy (overdispersed and zero-inflated) count data ($N_t$ is sampled from the Poisson distribution with the rate $\lambda = 100,000$ if not stated otherwise) are generated from multinomial distribution using the noisy relative abundances of the part of the Subject A time series (days from 60 to 140) that is highly dynamic David et al. (2014).

The sampling depth used in generating synthetic data is similar to the ones seen in real data (see the next subsection) Data sets with different number of time points are generated by down-sampling the series accordingly; for instance, while generating a data set with nine time points, we consider the relative abundance parameters of every ninth time point.

Longitudinal human gut microbiota data

We reanalyzed longitudinal gut microbiota 16S rRNA sequencing data sets of four individuals Caporaso et al. (2011); David et al. (2014). The individuals’ (referred as Subject A and Subject B) gut microbiota covered in the study of David et al. were sampled daily (with only a few missing time points). Subject A’s and Subject B’s gut microbiota were sampled through 359 days with 60 missing time points and 318 days with 138 missing time points, respectively. Because there are 66 missing time points between the last and the second-to-last measured time points of Subject B, we discarded the last measurement time point of Subject B. Both microbial ecosystems were exposed to severe environmental perturbations. Specifically, Subject A was exposed to novel diet between days 71 and 122 (while traveling abroad) and during this diet Subject A had two diarrheas (between days 80 and 85, and 104 and 113; shaded rectangles in SFigure 4). Subject B also experienced a large perturbation, a Salmonella infection between days 151 and 159 (shaded rectangle in SFigure 5). The
taxonomic assignment of the 16S rRNA sequencing reads was done at the order level; resulting 82 and 65 orders with at least one non-zero read count for Subject A and Subject B, respectively. The percentages of zeros in the Subject A and Subject B data sets were 75% and 78%, respectively (calculated from only the time points with measurements). Medians of number of sequencing reads assigned to bacterial orders in Subject A and Subject B samples are 156,081 and 72,111, respectively.

Additionally, we considered longitudinal gut microbiota 16S rRNA sequencing data sets of two other individuals (referred as M3 and F4) (Caporaso et al., 2011). M3’s and F4’s gut microbiota were samples through 443 days with 111 missing time points and 186 days with 56 missing time points, respectively. We used the taxonomic assignment of the original study (Caporaso et al., 2011) leading to 54 and 42 orders with at least one non-zero read count for M3 and F4, respectively. The percentages of zeros in the M3 and F4 data sets at the order level were 70% and 66%, respectively (calculated from only the time points with measurements). Medians of number of sequencing reads assigned to bacterial orders in M3 and F4 samples are 42,878 and 37,431, respectively.

Multinomial and Dirichlet-multinomial models

We compare our approach with the commonly used Dirichlet-multinomial and multinomial models. The maximum likelihood estimator of \( \Theta_t \) of the multinomial model (Equation 3) is

\[
\Theta_{t}^{ML} = \frac{1}{\sum x^{(i)}_t} x_t.
\]  

(1)

The hierarchical Dirichlet-multinomial model can be stated as

\[
\Theta_t | \alpha \sim \text{Dirichlet}(\alpha),
\]

\[
X_t | \Theta_t \sim \text{Multinomial}(\Theta_t, N_t),
\]  

(2)

where \( \alpha \) is the Dirichlet prior parameter (in this study, we assume \( \alpha_i = 1, i = 1, 2, \ldots, M \)). The Dirichlet distribution is the conjugate prior of the multinomial distribution; the posterior distribution of \( \Theta_t \) is

\[
\Theta_t | x_t, \alpha \sim \text{Dirichlet}(x_t + \alpha),
\]  

(3)

and the posterior mean of \( \Theta_t \) is

\[
E[\Theta_t | x_t, \alpha] = \frac{1}{\sum x^{(i)}_t + \alpha_i} (x_t + \alpha).
\]  

(4)

Posterior function

The full posterior distribution function up to a proportion of our model is

\[
p(\Theta^M, \Theta_G, \Theta_F, \sigma, \eta^2, \rho^2, \beta | T, \theta) \propto \prod_{t \in T} p(x_t | \Theta_t^M, N_t) \left( \prod_{t \in T} \prod_{i=1}^M p(\beta^{(i)}_t | \theta_\beta) \right) \left( \prod_{t \in T} \prod_{i=1}^{M-1} p(\beta^{(i)}_t | \sigma) \right) \left( \prod_{t \in T} \prod_{i=1}^{M-1} p(\beta^{(i)}_t | \rho) \right)
\]  

(5)

\[
= \prod_{t \in T} \prod_{i=1}^M p(x_t | \Theta_t^M, N_t) \left( \prod_{t \in T} \prod_{i=1}^M p(\beta^{(i)}_t | \theta_\beta) \right) \left( \prod_{i=1}^{M-1} p(\sigma | \theta_\sigma) \right) \left( \prod_{i=1}^{M-1} p(\rho | \theta_\rho) \right),
\]

where \( \sigma = \{\sigma_i | i = 1, 2, \ldots, M-1\} \), \( \eta^2 = \{\eta^{2}_{(i)} | i = 1, 2, \ldots, M-1\} \), \( \rho^2 = \{\rho^{2}_{(i)} | i = 1, 2, \ldots, M-1\} \), and \( \theta = \{\theta_\eta, \theta_\rho, \theta_\sigma, \theta_\beta\} \).
The probability density functions are (time point and taxa indices are omitted for simplicity)

\[ p(x|\Theta, N) = \frac{N!}{\prod_{m=1}^{M} m} \prod_{m=1}^{M} \Theta_{m}^{x_{m}} \]

\[ p(F|G, \sigma^2) = \frac{1}{(2\pi)^{T/2}} \frac{1}{\sqrt{|\sigma^2|}} \exp\left(-\frac{1}{2\sigma^2}(F - G)^{T}(G - F)\right) \]

\[ p(G|T, \eta^2, \rho^2) = \frac{1}{(2\pi)^{T/2}} \frac{1}{\sqrt{|\eta^2|}} \exp\left(-\frac{1}{2}(F - G)^{T}K_{T,T}(\eta^2, \rho^2)(G - F)\right) \]

\[ p(\sigma|\theta_\sigma) \propto \frac{1}{\sqrt{2\pi}\theta_{\sigma,2}} \exp\left(-\frac{1}{2}\left(\frac{\sigma - \theta_{\sigma,1}}{\theta_{\sigma,2}}\right)^2\right), \sigma > 0 \]

\[ p(\eta^2|\theta_\eta) = \frac{\theta_{\eta,1}^{\eta^2}}{\Gamma(\theta_{\eta,1})} \eta^{\eta^2-1} \exp\left(-\theta_{\eta,2}\eta^2\right) \]

\[ p(\rho^2|\theta_\rho) \propto \frac{1}{\sqrt{2\pi}\theta_{\rho,2}} \exp\left(-\frac{1}{2}\left(\frac{\rho^2 - \theta_{\rho,1}}{\theta_{\rho,2}}\right)^2\right), \rho^2 > 0 \]

\[ p(\beta|\theta_\beta) = \frac{1}{B(\theta_{\beta,1}, \theta_{\beta,2})} \beta^{\theta_{\beta,1}-1}(1 - \beta)^{\theta_{\beta,2}-1}, \]

where \( B(\cdot, \cdot) \) is the beta function and \( \Gamma(\cdot) \) is the gamma function.

**Window-based analysis of long time series**

To allow time-varying length-scale parameters and to reduce computational cost through distributed computing, we analyzed the time series using sliding windows of length 60 (number of time points with measurements) (SFigure 3). Except in the cases of the first and the last window positions, only the estimates of the middle of the windows are taken into account (SFigure 3).

**Supplemental Tables**

**Supplemental Table 1:** The means and the corresponding standard deviations (SD) (mean±SD) of the estimation errors depicted in Figure 2a.

<table>
<thead>
<tr>
<th>Number of time points (36 taxa)</th>
<th>6</th>
<th>9</th>
<th>14</th>
<th>27</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temporal</td>
<td>0.15±0.09</td>
<td>0.13±0.07</td>
<td>0.10±0.06</td>
<td>0.09±0.06</td>
</tr>
<tr>
<td>Dirichlet-multinomial</td>
<td>0.17±0.09</td>
<td>0.17±0.11</td>
<td>0.17±0.10</td>
<td>0.18±0.11</td>
</tr>
</tbody>
</table>

**Supplemental Table 2:** Proportion of sampling zeros and total zeros in the four scenarios considered in Figure 2b.

<table>
<thead>
<tr>
<th>36 taxa</th>
<th>Number of sampling zeros</th>
<th>Proportion</th>
<th>10</th>
<th>20</th>
<th>40</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sampling zeros</td>
<td>1%</td>
<td>2%</td>
<td>4%</td>
<td>10%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total zeros</td>
<td>24%</td>
<td>25%</td>
<td>26%</td>
<td>30%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Supplemental Figures**

**References**

Supplemental Figure 1: Prior distributions. (a) The prior probability density function of $\rho^2$ is illustrated on top. To illustrate the effect of $\rho^2$ on the covariance function defined in Equation 9, we illustrate at bottom its behaviour with four different length-scale values as a function of the difference in days (signal variance is set $\eta^2 = 1$). (b) The prior probability density functions of $\eta^2$ (left), $\sigma^2$ (middle), and $\beta$ (right) are illustrated.

Supplemental Figure 2: Synthetic data generation. The noise-free relative abundances of 36 bacterial orders are illustrated with black lines. Example sets of noisy relative abundances are illustrated with blue points. Sampled noisy relative abundances are used to generate count data.

Supplemental Figure 3: Our sliding window approach. The sliding analysis windows of length 60 are illustrated (rows). Except in the cases of the first and the last window, only the estimates at the 20 middle positions of the windows are taken into account.
Supplemental Figure 4: Estimation of sampling zeros and their effect on composition estimates. (a) As in Figure 4a but here the sampling depth is lower ((N ∼ Poisson(10000) vs. Poisson(10000))). (b) The prior probability density functions of θ1,1 (top) and θ2,1 (bottom) used in the sensitivity analysis are shown. Red lines depict the original values θ1,1 and θ2,1. (c) The results from the sensitivity analysis of the results depicted in (Figure 4a) with respect to the prior distributions of η and ρ are illustrated. (d) Box plots illustrate the variation of β values of taxa (proportions ≥ 1e-4) and time points with added sampling zero. The cases of 14 time points with either 10, 20, 40, or 120 added sampling zeros are considered. (e) As in (d) but here the focus is on the lowly abundant taxa (proportions < 1e-4). (f) Box plots illustrate the estimation error of the temporal (orange) and DM (green) models at the time points with induced sampling zeros. The case of 27 time points is considered. The cases of 71 (left column), 102 (middle column), and 160 (right column) species are considered. The cases of 120 (top row), 240 (middle row), and 480 (bottom row) sampling zeros are considered. (g) As in (f) but here the sampling depth is lower (N ∼ Poisson(10000) vs. Poisson(100000)). (h) Filled curves illustrate the distributions θβ,1 ∼ Beta(16, 4) (shaded in green) and θβ,2 ∼ Beta(8, 12) (shaded in dark gray) used in the sensitivity analysis of β. Red lines depict the original values θβ,1 = 0.8 and θβ,2 = 0.4. (d) Box plots illustrate the estimation error of the temporal (orange) and DM (green) models at the time points with induced sampling zeros. The case of 27 time points is considered. The cases of 71 (left column), 102 (middle column), and 160 (right column) species are considered. The cases of 120 (top row), 240 (middle row), and 480 (bottom row) sampling zeros are considered. (i) The sensitivity of the estimation of β with respect to the prior of β is studied in the scenario of (d). The estimated β values of taxa (proportions ≥ 1e-4) and time points with induced sampling zeros are compared by calculating the difference between the estimates obtained under the original prior (β ∼ Beta(0.8, 0.4)) and perturbed prior (β ∼ Beta(θβ,1, θβ,2)). The cases of 14 time points with either 10, 20, 40, or 120 sampling zeros are considered. Estimation error is defined to be the Euclidean distance between the first M-1 of the simplex-valued proportions vectors. Each box plot is calculated from 100 simulations. The two-sided p-values from the Wilcoxon signed-rank tests are listed. Outliers are not depicted.
Supplemental Figure 5: **Dynamics of bacterial orders in human gut microbiota of Subject A** (a) The microbial composition of Subject A’s gut microbiota (David et al., 2014) at order level over time. The black curves are the posterior mean estimates, $\Theta_G$, from the temporal analysis. The filled regions show the 5% and 95% credible intervals. The semi-transparent circles depict the maximum likelihood estimates under the multinomial model. The orange curve is the LOWESS ($\alpha = 0.05$, which corresponds approximately to 20 days) estimate calculated from the maximum likelihood estimates. The time period where the subject was abroad and suffered from diarrheas are illustrated using the three shaded rectangles. (b) Number of reads (log10) per time point in the Subject A data set are denoted using black points. Time points without measurements are depicted using shaded rectangles.
Supplemental Figure 6: **Dynamics of bacterial orders in human gut microbiota of Subject B** (a) The microbial composition of Subject B’s gut microbiota ([David et al., 2014]) at order level over time. The black curves are the posterior mean estimates, Θ̂_G, from the temporal analysis. The filled regions show the 5% and 95% credible intervals. The semi-transparent circles depict the maximum likelihood estimates under the multinomial model. The orange curve is the LOWESS (α = 0.05, which corresponds approximately to 20 days) estimate calculated from the maximum likelihood estimates. The time period where the subject suffered from the Salmonella infection is illustrated using the shaded rectangle. (b) Number of reads (log10) per time point in the Subject B data set are denoted using black points. Time points without measurements are depicted using shaded rectangles.
Supplemental Figure 7: Dynamics of bacterial orders in human gut microbiota of M3 (a) The microbial composition of M3’s gut microbiota (from (Caporaso et al., 2011)) at order level over time. The black curves are the posterior mean estimates, $\Theta^G$, from the temporal analysis. The filled regions show the 5% and 95% credible intervals. The semi-transparent circles depict the maximum likelihood estimates under the multinomial model. The orange curve is the LOWESS ($\alpha = 0.05$, which corresponds approximately to 20 days) estimate calculated from the maximum likelihood estimates. (b) Number of reads (log10) per time point in the M3 data set are denoted using black points. Time points without measurements are depicted using shaded rectangles.
Supplemental Figure 8: Dynamics of bacterial orders in human gut microbiota of F4 (a) The microbial composition of F4’s gut microbiota (from (Caporaso et al., 2011)) at order level over time. The black curves are the posterior mean estimates, $\Theta_G$, from the temporal analysis. The filled regions show the 5% and 95% credible intervals. The semi-transparent circles depict the maximum likelihood estimates under the multinomial model. The orange curve is the LOWESS ($\alpha = 0.05$, which corresponds approximately to 20 days) estimate calculated from the maximum likelihood estimates. (b) Number of reads (log10) per time point in the F4 data set are denoted using black points. Time points without measurements are depicted using shaded rectangles.
Supplemental Figure 9: **Jensen-Shannon distance matrices.** Pairwise Jensen-Shannon distances between the composition estimates of different time points are visualized. The different comparisons correspond to the cases: Subject A (first row, first column), Subject B (first row, second column), M3 (second row, first column), F4 (second row, second column), 2X-downsampled Subject A (third row, first column), 3X-downsampled Subject A (third row, second column), Subject A with the Matérn covariance function (fourth row, first column), and shuffled Subject A (fourth row, second column).
Subject A (Matern covariance function)

Supplemental Figure 10: Effect of the covariance function. As in SFigure 5a but here the analysis is done using the Matern covariance function ($\nu = 3/2$) instead of the squared exponential covariance function.
Supplemental Figure 11: **Control experiment with shuffled Subject A data.** (a) As in SFigure 5a but here the time points are shuffled before the analysis. (b) As in SFigure 5b but for the shuffled data.
Supplemental Figure 12: **Experiment with downsampled (2X) Subject A data.** (a) As in SFigure 5a but here the data is downsampled before the analysis by taking into account only every second measurement. The used analysis window is length of 30 measurements (10+10+10) to approximately match the window used on the original data. (b) As in SFigure 5b but for the downsampled data used in (a).
Supplemental Figure 13: **Control experiment with downsampled (3X) Subject A data.** (a) As in SFigure 5a but here the data is downsampled before the analysis by taking into account only every third measurement. The used analysis window is length of 15 measurements (5+5+5) to approximately match the window used on the original data. (b) As in SFigure 5b but for the downsampled data used in (a).
Supplemental Figure 14: Temporal variation of bacterial orders. (a) Autocorrelation functions of relative abundances of abundant (average > 1e − 3) bacterial orders. TGP-CODA (black curves) and maximum likelihood relative abundance estimates (red curves) are considered. Missing maximum likelihood estimates are imputed using piecewise linear interpolation. In the case of RF39 only the first 155 time points are considered because it disappears after 155 days. (b) Generalized Hurst exponents (q = 1) of the maximum likelihood and temporal relative abundance estimates are listed. The same bacterial orders as in (a) are considered.
Supplemental Table 4: Proportion of sampling zeros and total zeros in the four scenarios considered in SFigure 4f.

<table>
<thead>
<tr>
<th>Number of sampling zeros</th>
<th>71 taxa</th>
<th>102 taxa</th>
<th>160 taxa</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sampling zeros</td>
<td>Total zeros</td>
<td>Sampling zeros</td>
</tr>
<tr>
<td>240</td>
<td>6%</td>
<td>27%</td>
<td>4%</td>
</tr>
<tr>
<td>480</td>
<td>13%</td>
<td>31%</td>
<td>9%</td>
</tr>
<tr>
<td>960</td>
<td>25%</td>
<td>39%</td>
<td>17%</td>
</tr>
</tbody>
</table>

Supplemental Table 5: Proportion of sampling zeros and total zeros in the four scenarios considered in SFigure 4g. Here the sampling depth is lower (N Poisson(10,000)), which increases the percentages of zeros.

<table>
<thead>
<tr>
<th>Number of sampling zeros</th>
<th>71 taxa</th>
<th>102 taxa</th>
<th>160 taxa</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sampling zeros</td>
<td>Total zeros</td>
<td>Sampling zeros</td>
</tr>
<tr>
<td>240</td>
<td>6%</td>
<td>60%</td>
<td>4%</td>
</tr>
<tr>
<td>480</td>
<td>13%</td>
<td>61%</td>
<td>9%</td>
</tr>
<tr>
<td>960</td>
<td>25%</td>
<td>65%</td>
<td>17%</td>
</tr>
</tbody>
</table>

Supplemental Figure 15: **Power spectrum of bacterial orders abundances.** The power spectrum of relative abundances of abundant (average > 1e-3) bacterial orders. TGP-CODA (black curves) and maximum likelihood relative abundance estimates (red curves) are considered. Missing maximum likelihood estimates are imputed using piecewise linear interpolation. In the case of RF39 only the first 155 time points are considered because it disappears after 155 days. Lines are fitted to the TGP-CODA power spectrums (yellow curves; coefficients are listed).