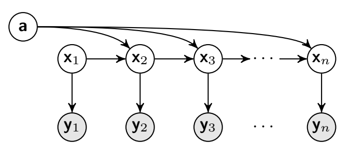
Supplementary Material: Robust and Scalable Models of microbiome Dynamics

Travis E. Gibson Georg K. Gerber

A. Extended discussion regarding constraining dynamics

We present an analysis of a naive model that directly constrains dynamics to be non-negative, to illustrate the issues this causes for the posterior distribution. Consider a dynamical process with latent state $\mathbf{x}$, measurements $\mathbf{y}$, and dynamical interaction coefficients $\mathbf{a}$:



(5)

generated by the following

$$\begin{aligned}\mathbf{x}_{k+1,i} \mid \mathbf{x}_k, \mathbf{a} &\sim \mathtt{Normal}_{\geq 0}(\mathbf{a}_i{}^\top f(\mathbf{x}_k), \sigma^2_{\mathbf{x}_i}) \\ \mathbf{y}_{k,i} \mid \mathbf{x}_{k,i} &\sim \mathtt{Normal}_{\geq 0}(\mathbf{x}_{k,i}, \sigma^2_{\mathbf{y}_i}) \\ \mathbf{a}_i &\sim \mathtt{Normal}(0, \sigma^2_{\mathbf{a}_i}).\end{aligned} \tag{6}$$

The dynamics in (6) are precisely the dynamics one obtains via adding a truncated normal measurement model to the discrete gLV dynamics presented in (1).[1] For ease of exposition let us assume for now that there is only 1 microbial species ($i = 1$ and thus index i can be dropped for this brief exposition) and all of the variance terms in (6) are equal to σ^2. Performing full Bayesian inference for $\mathbf{a}$ requires constructing the posterior $p_{\mathbf{a}|\mathbf{x}} \propto p_{\mathbf{x}|\mathbf{a}} p_{\mathbf{a}}$. Noting that the likelihood of $\mathbf{x}$ satisfies the following proportionality $p_{\mathbf{x}|\mathbf{a}} \propto \prod_k p_{\mathbf{x}_{k+1}|\mathbf{a},\mathbf{x}_k}$ and expanding this given our model in (6) we have

$$p_{\mathbf{x}|\mathbf{a}}(x \mid a) \propto \prod_k \frac{\mathrm{e}^{-\frac{1}{2\sigma^2}(x_{k+1} - a^\top f(x_k))^2}}{\sigma\sqrt{2\pi}\left(\Phi(\infty) - \Phi\left(-\frac{a^\top f(x_k)}{\sigma}\right)\right)} \tag{7}$$

where Φ is the Cumulative Distribution Function (CDF) for standard Normal distribution. Using the likelihood in (7) and the prior for $\mathbf{a}$ in (6), the posterior of $\mathbf{a}$ takes the form

$$\begin{aligned}&p_{\mathbf{a}|\mathbf{x}}(a \mid x) \\ &\propto \prod_k \frac{\mathrm{e}^{-\frac{1}{2\sigma^2}(x_{k+1} - a^\top f(x_k))^2}}{\sigma\sqrt{2\pi}\left(\Phi(\infty) - \Phi\left(-\frac{a^\top f(x_k)}{\sigma}\right)\right)} \frac{\mathrm{e}^{-\frac{1}{2\sigma^2}a^\top a}}{(\sigma^2 2\pi)^{n_a/2}}\end{aligned}$$

where n_a is the dimension of the column vector a. Having the variable a appear in the normalization constant means we cannot directly Gibbs sample $\mathbf{a}$, and also makes constructing an efficient proposal distribution in a Metropolis Hastings (MH) setting challenging too, as the proposal will in turn be affecting the scaling factor of the target distribution. A similar issue is encountered when trying to sample from the latent state $\mathbf{x} \mid \mathbf{a}, \mathbf{y}$ (filtering).

[1] Note that this is the most direct way one can enforce a hard non-negativity constraint on the dynamics, and is indeed the first direction we took before realizing the challenges it imposes.