Statistical modeling of species co-occurrence for clinical and microbial data

Joe Mihaljevic, Sylvia Ranjeva, Max Joseph, Greg Dwyer, Sarah Cobey

February 2, 2016

Introduction

Here we will apply established statistical methods of species co-occurrence to clinical and microbial data sets and assess the model's feasibility given the number of interacting 'species'. The model will estimate pairwise correlations between the species among sampling units, as well as pairwise effects of species on establishment and persistence.

Model Specification

Here we will outline the statistical model, based off of SebastiÃąn-GonzÃąlez2010. This is a longitudinal model which accounts for the effects of each species' occurrence in the previous observation on the occurrence of each species in the current observation (i.e. pairwise effects on persistence and colonization). The model also accounts for among sampling unit correlations in species' occurrence probabilities (e.g. a patient being a sampling unit for pathogens). In this case I number of sampling units are surveyed for J number of species over T number of temporally spaced observations. The model structure is as follows:

$$y_{ijt} \sim Bern(\psi_{ijt})$$

$$logit(\psi_{ijt}) = \alpha_{0j} + \alpha_{ij} + \mathbf{B}_{jk}\mathbf{X}_{t-1}$$

$$\alpha_{ij} \sim N(0, \Sigma)$$

In this model, the occurrence of species j in sampling unit i during observation t is Bernouli distributed with occurrence probability ψ_{ijt}. The logit-transformed occurrence probability is then linearly related to a species-specific intercept, α_{0j}, a random effect of sampling unit, α_{ij}, and a number of covariates with fixed effects. $\mathbf{X}_{t-1}$ is a covariate matrix with J columns and $T-1$ rows. This matrix is filled with the occurrence of each species at the previous observation $t-1$. Then, $\mathbf{B}_{jk}$ represents a vector of the covariate effects of species k on species j's occurrence in observation t, where $k = 1, \ldots, J$. Thus, $\mathbf{B}'$ is a JxJ matrix of covariate effects that represent the effect of the occurrence of species k in observation $t-1$ on species i in the current observation, t.

The random effect of sampling unit, α_{ij}, follows a multivariate normal distribution, where Σ represents the species-species covariance matrix. In other words, the model estimates the among-sampling unit correlation between each species. Thus, the variances along the diagonal of this matrix, σ_j^2, captures the variability in each species' occurrence probabilities among sampling units, and ρ_{ij} represents the correlation between species i and j. Concretely, for a simple, two species scenario:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$