Inferring the presence of metabolites

March 21, 2023

We seek to infer the presence or absence of metabolites in group of samples compartimentalized by an arbitrary number of discrete axis such as e.g. species, tissue or environmental conditions. Let $\boldsymbol{\tau} = \{\tau_1, \dots, \tau_T\}$ denote the set of T axis of compartimentalization and let, without loss of generality, $\tau_1 = \mathcal{M}$ be the axis of metabolites. For any compartment c, let $\tau_t(c) = 1, \dots, n_t$ indicate the compartment index along axis τ_t with $\tau_{\mathcal{M}}(c) = \tau_1(c)$ indicating the metabolite of that compartment.

Let x_c denote the presence ($x_c = 1$) or absence ($x_c = 0$) of a metabolite $\tau_{\mathcal{M}}(c)$ in comparment c and let $\boldsymbol{x} = (x_1, \dots, x_C)$ be the full vector x_c across all compartments $c = 1, \dots, C$.

We will assume that similarities across any of the axis of compartimentalization is reflected in the patterns of presences and absences in $\boldsymbol{x}$. For instance, closely related species may share a similar set of metabolites and metabolites related in their synthesis may share a similar distribution across species. To model such similarities, we assume that the probability $\mathbb{P}(x_c = 1|\boldsymbol{\mu}_c, \epsilon_c)$ with which metabolite $\tau_{\mathcal{M}}(c) = m$ is present in compartment c is given by

$$\operatorname{logit} \mathbb{P}(x_c = 1|\boldsymbol{\mu}_c, \epsilon_c) = \sum_t \mu_t + \epsilon_c,$$

where $\boldsymbol{\mu}_c = (\mu_1, \dots, \mu_T)$ is a vector of axis specific intercepts and ϵ_c is normally distributed with mean 0 and co-variance

$$\operatorname{cov}(\epsilon_c, \epsilon_{c'}) = \sum_t \beta^{(\tau_t)}_{\tau_t(c)} + \sum_t \beta^{(\tau_t)}_{\tau_t(c')} + \sum_f \alpha_f \sigma_f \Big(\tau_{\tau(f)}(c), \tau_{\tau(f)}(c') \Big). \tag{1}$$

Here, the $\beta^{\tau_t}_{\tau_t(c)}$ are intercepts specific for the compartment index $\tau_t(c)$ along axis τ_t, the $\sigma_f, f = 1, \dots, F$, are known covariances between entries along axis $\tau(f)$ and the α_f are positive scalars.

We consider two sets of data informative about $\boldsymbol{x}$: i) Presence-absence data obtained with mass-spectrometry and ii) presence-only reports of specific metabolites in specific specie. Let $\boldsymbol{d}_{sj} = (d_{sj1}, \dots, d_{sjM})$ be the presence-absence vector of each metabolite m obtained with mass-spectrometry run $j = 1, \dots, J_s$ performed on species s. Assuming a false-positive and false-negative error rates ϵ_{01} and ϵ_{10}, respectively, we have

$$\mathbb{P}(\boldsymbol{d}_{sj}|\boldsymbol{x}, \epsilon_{01}, \epsilon_{10}) = \prod_m \left[x_{sm} \left(\epsilon_{10}^{1-d_{sjm}} (1-\epsilon_{10})^{d_{sjm}} \right) + (1 - x_{sm}) \left(\epsilon_{01}^{d_{sjm}} (1-\epsilon_{01})^{1-d_{sjm}} \right) \right].$$

To model the presence only data, it must be put in relation to the expected research effort. Let p_{sm} denote the known number of presence-only reports for metabolite m in species s and n_{sm} the unknown number of research projects that aimed at discovering metabolite m in species s. Assuming a false-positive and false-negative error rates π_{01} and π_{10}, respectively, we have

$$\mathbb{P}(p_{sm}|n_{sm}, \pi_{01}, \pi_{10}) =$$