

Inferring the presence of metabolites

February 13, 2023

We seek to infer the presence or absence of M metabolites in S species. We denote by x_{sm} whether metabolite $m = 1, \dots, M$ is present ($x_{sm} = 1$) or absent ($x_{sm} = 0$) in species $s = 1, \dots, S$. To infer the full vector $\mathbf{x} = (x_{11}, \dots, x_{1M}, \dots, x_{SM})$, we assume that related species share a similar set of metabolites and that metabolites related in their synthesis share a similar distribution across species. Let $\mathbb{P}(x_{sm} = 1 | y_{sm}) = y_{sm}$ be the probability with which metabolite m is present in species s . We then assume that

$$\text{logit } y_{sm} = \mu_m + \epsilon_{sm}$$

where μ is a metabolite-specific intercept and ϵ_{sm} is normally distributed with mean 0 and co-variance $\text{cov}(\epsilon_{sm}, \epsilon_{s'm'}) = \alpha\sigma_{ss'} + \beta\sigma_{mm'}$ between each combination of species and metabolite. Here, $\sigma_{ss'}$ and $\sigma_{mm'}$ are known measures of covariance between species s and s' and between metabolites m and m' , respectively, and α and β are positive scalars.

We consider two sets of data informative about \mathbf{x} : i) Presence-absence data obtained with mass-spectrometry and ii) presence-only reports of specific metabolites in specific specie. Let $\mathbf{d}_{sj} = (d_{sj1}, \dots, d_{sjM})$ 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

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

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

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