Inferring the presence of metabolites

February 14, 2023

1 Idea and concept

We seek to infer the presence or absence of M metabolites in S species. We denote by x_{sm} whether metabolite $m = 1, \ldots, M$ is present ($x_{sm} = 1$) or absent ($x_{sm} = 0$) in species $s = 1, \ldots, S$. To infer the full vector $\boldsymbol{x} = (x_{11}, \ldots, x_{1M}, \ldots, x_{SM})$, 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

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

where μ is a metabolite-specific intercept and ϵ_{sm} is normally distributed with mean 0 and co-variance $\operatorname{cov}(\epsilon_{sm}, \epsilon_{s'm'}) = \alpha\sigma_{ss'} + \beta\sigma_{mm'}$ 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 $\boldsymbol{x}$: i) Presence-absence data obtained with mass-spectrometry and ii) presence-only reports of specific metabolites in specific specie. Let $\boldsymbol{d}_{sj} = (d_{sj1}, \ldots, d_{sjM})$ be the presence-absence vector of each metabolite m obtained with mass-spectrometry run $j = 1, \ldots, J_s$ 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

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

We would have the covariance matrix such as :

$$\operatorname{cov}(\epsilon_{smt}, \epsilon_{s'm't'}) = \alpha\sigma^P_{ss'} + \beta\sigma^M_{mm'} + \gamma\sigma^E_{ss'} + \ldots \tag{1}$$

With P the phenotype between two species, E an environment factor between two species and M the TODO

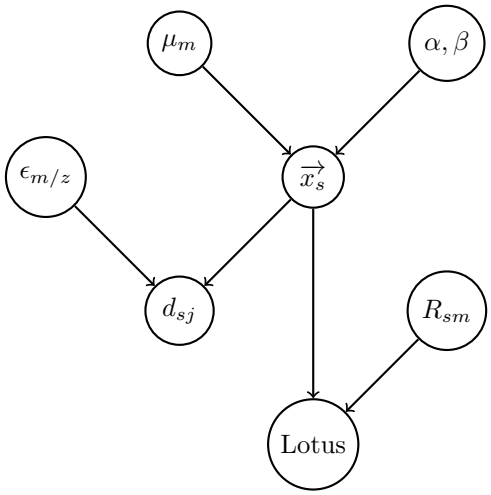
2 DAG scratch

We assume that the probability of having a molecule in a species is the average presence of that molecule across all species μ_m plus a normally distributed error that depends on certain parameters $\alpha, \beta, \ldots$.

From there, the LOTUS database and any result of MS depends on the set of molecules present in a species $\overrightarrow{x_s}$. However we still have to take into account the fact that there can be an error of analysis on the MS $\epsilon_{m/z}$ that where $\epsilon_{m/z} = f(\epsilon_{01}, \epsilon_{10})$.

Could we assume that as $j \to \infty$, then $d_{sj} \to \overrightarrow{x_s}$? TODO

According to Pierre-Marie LOTUS database is highly dependent on the research effort accorded to the specific molecule. He also said that error rate of LOTUS database is very low since most data is an actual isolation of the specific compound. Error rate of LOTUS database should then have little effect on the model.



3 Ideas scratch

$$\begin{array}{c} \\ x_{sm}=0 \\ x_{sm}=1 \end{array} \begin{array}{c} \begin{array}{cc} L_{sm}=NA & L_{sm}=1 \end{array} \\ \begin{pmatrix} 1 & 0 \\ 1-R_{sm} & R_{sm} \end{pmatrix} \end{array}$$

With x_{sm} a molecule present or not present in a specific species. L_{sm} the presence or absence of a molecule in a species that is present or not in the Lotus database. Finally, R_{sm} the research effort made for that specific molecule. R_{sm} being a function of the number of papers made on a specific molecule or species : $f(n_s, n_m)$.