Inferring the presence of metabolites

March 22, 2023

1 v2

We seek to infer the presence or absence of metabolites in group of samples compartmentalized by T of discrete axes such as e.g. species, tissue or environmental conditions. For any compartment c, let $\tau_t(c) = 1, \ldots, n_t$ indicate the compartment index along axis $t = 1, \ldots, T$. For convenience, let us further denote by $\tau_{\mathcal{M}}(c)$ and $\tau_{\mathcal{S}}(c)$ the metabolite and species 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 compartment c and let $\boldsymbol{x} = (x_1, \ldots, x_C)$ be the full vector x_c across all compartments $c = 1, \ldots, C$ with $C = \prod_t n_t$.

We will assume that similarities across any of the axis of compartmentalization 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 $\tau_{\mathcal{M}}(c)$ is present in compartment c is given by

$$\operatorname{logit} \mathbb{P}(x_c = 1|\boldsymbol{\mu}_c, \epsilon_c) = \sum_{t=1}^{T} \mu^{(t)}_{\tau_t(c)} + \epsilon_c \tag{1}$$

where $\boldsymbol{\mu}_c = (\mu^{(1)}_{\tau_1(c)}, \ldots, \mu^{(T)}_{\tau_T(c)})$ is a vector of axis specific intercepts and ϵ_c is normally distributed with mean 0 and co-variance

$$\operatorname{cov}(\epsilon_c, \epsilon_{c'}) = \sum_t \beta^{(t)}_{\tau_t(c)} + \sum_t \beta^{(t)}_{\tau_t(c')} + \sum_t \sum_{f=1}^{F_t} \alpha_{tf} \sigma_{tf}\Big(\tau_t(c), \tau_t(c')\Big). \tag{2}$$

Here, the $\beta^{(t)}_{\tau_t(c)}$ are positive intercepts specific for the compartment index $\tau_t(c)$ along axis t, the $\sigma_{tf}, f = 1, \ldots, F_t$, are the F_t known covariance matrices between entries along axis t, and the α_{tf} are positive scalars.

1.1 Emission probabilities

We consider several different types of data to inform about $\boldsymbol{x}$. This data may be of different dimensionality, e.g. may only discriminate along a subset of the axes or at a higher scale along some axes. For a particular data set $d = 1, \ldots, D$, let $\boldsymbol{\xi}_d = \{\xi_{d1}, \ldots, \xi_{du}\}$ denote the sets of distinguished compartments. We then define the presence of ($\boldsymbol{x}(\xi_{du}) = 1$) or absence ($\boldsymbol{x}(\xi_{du}) = 0$) in set $\xi_{du}, u = 1 \ldots, U$, as

$$\boldsymbol{x}(\xi_{du}) = \min\left(1, \sum_{c \in \xi_{du}} x_c\right).$$

1.1.1 LOTUS

The LOTUS database [1] lists known occurrences of metabolites in species. Let $L_{ms} = 1$ denote a known occurrence of metabolite m in species s, while $L_{ms} = 0$ denotes that no evidence for such an occurrence has

been reported, either because the metabolite m is truly absent in species s or because of a lack of research effort.

Let us denote by R_{sm} the probability of discovery of metabolite m in species s such that

$$\mathbb{P}(L_{ms}|\boldsymbol{x}(\xi(m,s)), R_{ms}) = \begin{cases} 0 & \text{if } \boldsymbol{x}(\xi(m,s)) = 0 \text{ and } L_{ms} = 1, \\ 1 & \text{if } \boldsymbol{x}(\xi(m,s)) = 0 \text{ and } L_{ms} = 0, \\ R_{ms} & \text{if } \boldsymbol{x}(\xi(m,s)) = 1 \text{ and } L_{ms} = 1, \\ 1 - R_{ms} & \text{if } \boldsymbol{x}(\xi(m,s)) = 1 \text{ and } L_{ms} = 0, \end{cases}$$

where $\xi(m,s)$ is the set of compartments relevant for metabolite m and species s, i.e. all compartments c for which $\tau_{\mathcal{M}}(c) = m$ and $\tau_{\mathcal{S}}(c) = s$.

To quantify the research effort R_{ms} of a particular entry L_{ms}, we will rely on two measures, the total number of relevant papers published for metabolite m (P_m) and for species s (Q_s), such that

$$R_{ms} = 1 - e^{-\gamma P_m - \delta Q_s}$$

with positives scalars γ and δ.

1.2 Prior distributions

- $\mu_{ti} \sim \mathcal{N}(0, \sigma_\mu^2), t = 1, \ldots, T, i = 1 \ldots, n_t$ with $\sigma_\mu^2 = 1$
- $\alpha_i^{(t)} \sim \text{Exp}(\lambda_\alpha), t = 1, \ldots, T, i = 1 \ldots, n_t$
- $\beta_i^{(t)} \sim \text{Exp}(\lambda_\beta), t = 1, \ldots, T, i = 1 \ldots, n_t$
- $\gamma \sim \text{Exp}(\lambda_\gamma)$
- $\delta \sim \text{Exp}(\lambda_\delta)$

1.3 Simulations

- Simulate P_m and Q_s from a Poisson distribution
- Simulate σ_{tf} using a Wishart distribution (?)

References

[1] Adriano Rutz, Maria Sorokina, Jakub Galgonek, Daniel Mietchen, Egon Willighagen, Arnaud Gaudry, James G Graham, Ralf Stephan, Roderic Page, Jiří Vondrášek, Christoph Steinbeck, Guido F Pauli, Jean-Luc Wolfender, Jonathan Bisson, and Pierre-Marie Allard. The LOTUS initiative for open knowledge management in natural products research. *eLife*, 11:e70780, May 2022. doi:10.7554/eLife.70780.