Inferring the presence of metabolites

July 18, 2023

1 Markov Random Field

We seek to infer the presence or absence of metabolites in a group of samples compartmentalized by several discrete dimensions such as e.g. species, tissue or environmental conditions. We assume that the pattern of presence and absence of these metabolites is modulated by similarities within each dimensions. 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 adopt a Markov Random Field approach.

Let D denote the total number of dimensions, of which, without loss of generality, the first shall be the metabolite. Each dimension $d = 1, \dots, D$ consist of a set $\mathcal{E}_d$ of discrete entries (e.g. individual species along the species dimension). We model similarities between the entries of dimension d using a Markov process along a known tree $\mathcal{T}_d$ consisting of $\mathcal{N}_d = \mathcal{E}_d \cup \mathcal{R}_d \cup \mathcal{I}_d$ nodes, of which the entries $\mathcal{E}_d$ are leaves, connected to the set of roots $\mathcal{R}_d$ through a set $\mathcal{I}_d$ of internal nodes; $\mathcal{E}_d \cap \mathcal{R}_d = \varnothing$, $\mathcal{E}_d \cap \mathcal{I}_d = \varnothing$ and $\mathcal{R}_d \cap \mathcal{I}_d = \varnothing$. For every node $n \in \mathcal{N}_d, n \notin \mathcal{R}_d$ that is not a root, let $p(n) \in \mathcal{N}_d$ denote its parent node and $b(n) \geq 0$ the length of the branch connecting it to its parent.

Let $\mathcal{X}$ denote a Markov Random Field of which each variable $x \in \mathcal{X}$ represents a combination of nodes from each of the D dimensions and indicates the presence($x = 1$) or absence ($x = 0$) of metabolite for that combination of nodes. Let $\delta_d(x) \in \mathcal{N}_d$ reflect the node of x in dimension d with $\delta_1(x)$ indicating the metabolite of x, and let $\delta(x) = (\delta_1(x), \dots, \delta_D(x))$. We only consider two sets of variables: 1) the set $\mathcal{Y}$ of variables representing an entry in each dimension such that for a variable $y \in \mathcal{Y}$, $\delta_d(y) \in \mathcal{E}_d$ for all $d = 1, \dots, D$, and 2) the set $\mathcal{Z}$ of variables representing leaves in all dimensions except one such that for a variable $z \in \mathcal{Z}$, $\delta_k(z) \in \mathcal{I}_k$ and $\delta_d(z) \in \mathcal{E}_d$ for all $d \neq k$. We then have $\mathcal{X} = \mathcal{Y} \cup \mathcal{Z}$ and $\mathcal{Y} \cap \mathcal{Z} = \varnothing$.

We suppose that the joint density of $\mathcal{X}$ can be factorized over a set of cliques $\mathcal{C}$. Each clique $c \in \mathcal{C}$ consist of a set of variables $x_1, x_2, \dots \in \mathcal{X}$ that represent the same leaves in all but one dimension k. Specifically, for all $x \in c$, $\delta_d(x) \in \mathcal{E}_d$ for all $d \neq k$ and $\delta_k(x) \in \mathcal{N}_k$, and for all $x_i, x_j \in c$, $\delta_{-k}(x_i) = \delta_{-k}(x_j)$, where $\delta_{-k}(x)$ denotes the vector of nodes of x in all dimensions but k. For such a clique, we will refer to the dimension $\nu(c) = k$ as its *variable* dimension and will denote by $\delta_{-\nu(c)}(c)$ the vector of nodes in the *fixed* dimensions. By definition, $\delta_{-\nu(c)}(c) = \delta_{-\nu(c)}(x)$ for every $x \in c$.

We will further denote by $\mathcal{C}_k \subset \mathcal{C}$ the subset of cliques that share the variable dimension k, i.e. $\nu(c) = k$ for all $c \in \mathcal{C}_k$. Note that each clique is in exactly one subset ($\mathcal{C}_k \cap \mathcal{C}_d = \varnothing$ for all $k \neq d$) and cliques of the same subset do not share any variables ($c_1 \cap c_2 = \varnothing$ for all $c_1, c_2 \in C_k$). However, each variable $x \in \mathcal{Y}$ will be part of exactly one clique from each subset: the clique $c \in \mathcal{C}_k$ for which $\delta_{-k}(c) = \delta_{-k}(x)$. In contrast, each variable $x \in \mathcal{Z}$ will be part of exactly one clique: the clique $c \in \mathcal{C}$ for which $\delta_{-\nu(c)}(c) = \delta_{-\nu(c)}(x)$ and $\delta_{\nu(c)}(x) \in \mathcal{I}_{\nu(c)}$.

The joint density of $\mathcal{X}$ factorizes as

$$\mathbb{P}(\mathcal{X}) = \prod_{d=1}^{D} \prod_{c \in \mathcal{C}_d} \phi(c),$$

where we model the clique functions $\phi(c)$ using a Markov model along tree $\mathcal{T}_d$. Let

$$\boldsymbol{\Lambda}_c = \begin{pmatrix} -\mu_{c1} & \mu_{c1} \\ \mu_{c0} & -\mu_{c0} \end{pmatrix}$$

be the rate matrix for changes between states 0 and 1 along the tree. For each node $n \in \mathcal{N}_d, n \notin \mathcal{R}_d$ that is not a root, the transition probabilities between parent node $p(n)$ and n are then given by

$$\boldsymbol{P}(n) = \exp(\boldsymbol{\Lambda}_c b(n)).$$

We assume the root state probabilities are given by the stationary distribution of the Markov chain:

$$\boldsymbol{P}_\infty = \left(\frac{\mu_{c0}}{\mu_{c0} + \mu_{c1}}, \frac{\mu_{c1}}{\mu_{c0} + \mu_{c1}} \right).$$

The clique function $\phi(c)$

$$\phi(c) = \prod_{x \in c,} \left(\mathrm{I}(x \in \mathcal{R}_{\nu(c)})[\boldsymbol{P}_\infty]_x + \mathrm{I}(x \notin \mathcal{R}_{\nu(c)})[\boldsymbol{P}(\delta_{\nu(c)}(x))]_{p_c(x),x} \right)$$

where we used the shorthand $x \in \mathcal{R}_{\nu(c)}$ for $\delta_{\nu(c)}(x) \in \mathcal{R}_{\nu(c)}$ to indicate whether the node in the variable dimension of c of x is a root and $p_c(x)$ to identify the variable $z \in c$ for which $\delta_{\nu(c)}(z) = p(\delta_{\nu(c)}(x))$.

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 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 $\tau_{\S}(c) = s$.

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 (?)