Inference

Raphaelle Momal

April 5, 2018

Contents

1 Notation

- Y is the data table, Y has n rows and d columns
- i indexes the rows of Y. Example : Y_i is an observed sample
- j indexes the columns of Y. Example : Y^j is the j^{th} species which has been observed
- $\mathcal{T}$ is the set of all trees possibly defined on sets of edges E and nodes V
- T is a gaussian decision tree, composed of a set of edges $E_T \subset E$, and a set of vertices $V_T \subset V$
- k and l are two nodes of V. Example : k and l are two studied species
- $|\cdot|$ is the determinant

2 Technical results

2.1 Formulas related to probabilities

We define a d-dimensional gaussian random variable Y such that:

$$Y \sim \mathcal{N}(0, \Sigma).$$

The log of the density of Y can be written as follows:

$$\log(p_\theta(Y)) = -\frac{n}{2}\log(|\Sigma|) - \frac{1}{2}tr(\underbrace{\Sigma^{-1}Y^{\mathrm{T}}Y}_{nI_d}) + cst,$$

and can thus be estimated using an estimate for Σ. Classically, $\hat{\Sigma} = Y^{\mathrm{T}}Y$ is used, and we get:

$$\begin{aligned}\log(\hat{p}_\theta(Y)) &= -\frac{n}{2}\log(|\hat{\Sigma}|) - \frac{1}{2}tr(\underbrace{\hat{\Sigma}^{-1}Y^{\mathrm{T}}Y}_{nI_d}) + cst \\ &= -\frac{n}{2}\log(|\hat{\Sigma}|) - \frac{nd}{2} + cst\end{aligned}$$

In the special case where Y is standardised (the diagonal of Σ only contains ones) and $d = 2$, we obtain a compact expression for the estimation of the density of Y:

$$\log(\hat{p}_\theta(Y)) = -\frac{n}{2}\log(1 - \hat{\rho}^2) - n + cst, \tag{1}$$

where $\rho = Cor(Y_1, Y_2)$.

2.2 Graphical Models

In the following we consider undirected graphical models, or Markov random fields, which represent the conditional dependance structure between random variables.

2.2.1 Definitions

- The cliques of a graph $\mathcal{G} = (V, E)$ are all subsets of V such that all vertices are linked by an edge. A maximal clique of $\mathcal{G}$, $C_\mathcal{G}$, is a clique that cannot be strictly contained by any other clique of $\mathcal{G}$.

- A density f on random variable $Y = (Y_1, ..., Y_p)$ is said to factorize according to $\mathcal{G}$ if:

$$f(y) = \prod_{c \in C_\mathcal{G}} f_c(y_c),$$

  where f_c are positive functions which depends on Y through Y_c only.

2.2.2 Results

A probability measure P satisfies the *pairwise Markov property* relative to a graph $\mathcal{G}$ with vertex set V, if for any pair of non adjacent vertices (k, l),

$$k \perp\!\!\!\perp l | V \setminus \{k, l\}.$$

Theorem [Hammersley and Clifford, 1971]: A probability distribution P with positive and continuous density f satisfies the pairwise Markov property with respect to an undirected graph $\mathcal{G}$ if and only if it factorizes according to $\mathcal{G}$.

2.3 Trees

Trees can be defined as specific graphical models, where each child node has only one parent. This means that loops are forbidden in trees. Spanning trees are trees where every vertex is linked to at least one other vertex. This yields an interesting property of spanning trees, which is that the cliques of a spanning tree all contain exactly two nodes of the graph. Then, a density which factorizes according to a spanning tree T will be of the following form:

$$f(y) = \prod_{k,l \in V_T} f^*(y_k, y_l),$$

where f^* are positive functions.

2.4 Algebra

2.4.1 Matrix Tree Theorem

We define the Laplacian matrix of a symmetric matrix $W = [\beta_{ij}]_{1 \leq i,j \leq n}$ as follows :

$$\mathcal{Q}_{uv}(W) = \begin{cases} -\beta_{uv} & 1 \leq u < v \leq n \\ \sum_{w=1}^{n} \beta_{wv} & 1 \leq u = v \leq n. \end{cases}$$

Matrix Tree Theorem (MTT) [Kirchhoff, 1847]: for any adjacency matrix W of a graph G, any minor of the Laplacian of W is the number of spanning trees of G. Writing $Q^*_{uv}(W)$ as the $(u, v)^e$ minor of $Q(W)$, this theorem implies that:

$$|Q^*_{uv}(W)| = \sum_{T \in \mathcal{T}} \prod_{\{k,l\} \in E_T} a_{kl} := Z(W).$$

The extension of this theorem to a real-valued matrix of weights was given by Meilă *et al* [Meilă and Jaakkola, 2006]. We call this extension the GMTT.

2.4.2 Meila and Kirshner's theorems

Meila *et al.* give a formula for the derivative of $Z(W)$, using the GMTT. Define the symmetric matrix $M(W)$ with 0 diagonal such that :

$$\begin{cases} M_{uv} = [\mathcal{Q}^{*-1}]_{uu} + [\mathcal{Q}^{*-1}]_{vv} - 2[\mathcal{Q}^{*-1}]_{uv} & u, v < n \\ M_{nv} = M_{vn} = [\mathcal{Q}^{*-1}]_{vv} & v < n \\ M_{vv} = 0. \end{cases}$$

Meila *et al.* then demonstrate that

$$\frac{\partial Z(W)}{\partial \beta_{kl}} = M_{kl} \times Z(W) \tag{2}$$

2.5 EM algorithm

We have observed data Y and unobserved data Z. The goal is to compute the likelihood of the data, $p_\theta(Y)$.

$$\log(p_\theta(Y)) = \log(p_\theta(Y, Z)) - \log(p_\theta(Z|Y)).$$

The advantage of this is to link $p_\theta(Y)$ with $p_\theta(Y, Z)$ which is easier to compute in general. We now take the expectation, conditioned on the data Y :

$$\log(p_\theta(Y)) = \mathbb{E}_\theta\left(\log(p_\theta(Y, Z))|Y\right) \underbrace{-\mathbb{E}_\theta\left(\log(p_\theta(Y|Z))|Y\right)}_{\mathcal{H}(p_\theta(Y|Z))}$$

E step : Data Y is considered fixed, leading the entropy term to be fixed as well. This step is dedicated to the computation of $\mathbb{E}_\theta\left(\log(p_\theta(Y,Z))|Y\right)$, which is the conditional expectation of the complete log-likelihood and where only the hidden part Z is varying.

M step : We consider that θ is varying and we want to maximise the expectation with respect to these parameters. This step generally uses the value computed in the previous E step.

Repeating these steps, we obtain optimised values of the parameters, which can give us some information about the hidden variable Z. We are also able to compute the likelihood of the model, but it is generally not of primary use.

3 Network inference

3.1 The model

We consider our data Y to be standardised. We suppose gaussian densities for Y:

$$Y \sim \mathcal{MVN}(\mathbf{0}, \Sigma),$$

where the diagonal of the Σ matrix is composed of ones. We will also use the model for a couple (k,l) of columns only, which is:

$$Y_{kl} \sim \mathcal{N}(0, \Sigma_{kl}),$$

where Σ_{kl} is a 2×2 square matrix with ones on its diagonal and $[\Sigma_{kl}]_{1,2} = \rho_{kl}$.

We then assume that the species under study, which make the columns of Y, are dependent on one another and that the dependance structure is shaped as a tree $T \in \mathcal{T}$.

$$\forall i \in \{1, ..., n\},\ Y_i|T \text{ iid. } \sim \mathcal{N}(\mathbf{0}, \Sigma_i)$$

The dependance tree is built as a mixture of hidden trees. In practice, each edge of E has a given weight, and the probability of the final tree is the normalised product of all these weights. In our model, we consider the weights as random.

$$\mathbb{P}(T) = \frac{1}{B} \prod_{k,l \in T} \beta_{kl} \text{ , with } B = \sum_{T \in \mathcal{T}} \prod_{k,l \in T} \beta_{kl}$$

3.2 Likelihood

$$\begin{aligned}
\mathbb{P}(Y|T) &= \mathbb{P}(Y_1|T) \prod_{j=2}^{d} \frac{\mathbb{P}(Y_j, Y_{a_j}|T)}{\mathbb{P}(Y_{a_j}|T)} \\
&= \underbrace{\prod_{j=1}^{d} \mathbb{P}(Y_j|T)}_{A} \prod_{(k,l) \in T} \underbrace{\frac{\mathbb{P}(Y_k, Y_l|T)}{\mathbb{P}(Y_k|T) \times \mathbb{P}(Y_l|T)}}_{\psi_{kl}(Y)} \\
&= A \prod_{k,l \in T} \psi_{kl}(Y)
\end{aligned}$$

Following result (1) we have:

$$\log(\mathbb{P}(Y|T)) = \log(A) + \sum_{k,l \in T} \log(\mathbb{P}(Y_k, Y_l|T)) - \log(\mathbb{P}(Y_k|T)) - \log(\mathbb{P}(Y_l|T))$$

Taking maximum likelihood estimates for parameters in the following expression, we obtain:

$$\begin{aligned} -\log(\mathbb{P}(Y_k)) - \log(\mathbb{P}(Y_l)) &= \frac{n}{2}(\log(\hat{\sigma}_k^2) + \log(\hat{\sigma}_l^2)) + \frac{1}{2}\left(\frac{1}{\hat{\sigma}_l^2}Y_k^T Y_k + \frac{1}{\hat{\sigma}_k^2}Y_l^T Y_l\right) \\ &= \frac{n}{2} \times 0 + n \end{aligned}$$

Then:

$$\begin{aligned} \log(\mathbb{P}(Y|T)) &= \log(A) + \sum_{k,l \in T} -\frac{n}{2}\log(|\Sigma_{kl}|) - \frac{1}{2}tr(\hat{\Sigma}_{kl}^{-1} Y_{kl}^{\mathrm{T}} Y_{kl}) + n \\ &= -\frac{n}{2}\sum_k (\log(\hat{\sigma}_k^2) + 1) + \sum_{k,l \in T}\left(-\frac{n}{2}\log(1-\hat{\rho}_{kl}^2) - \frac{2n}{2} + n\right) \end{aligned}$$

$$\boxed{\log(\mathbb{P}(Y|T)) \propto \sum_{k,l \in T} \underbrace{-\frac{n}{2}\log(1-\hat{\rho}_{kl}^2)}_{\log(\hat{\psi}_{kl})} + \sum_k \underbrace{-\frac{n}{2}\log(\hat{\sigma}_k^2)}_{\log(\hat{A})}}$$

3.3 EM algorithm

3.3.1 E step

$$\mathbb{P}(Y,T) = \mathbb{P}(T) \times \mathbb{P}(Y|T)$$

$$\begin{aligned} \log(\mathbb{P}(Y,T)) &= \sum_{(k,l) \in E_T} [\log(\beta_{kl}) + \log(\psi_{kl})] - \log(B) + \log(A) \\ &= \sum_{k,l \in V} \mathbb{1}_{\{(k,l) \in E_T\}}(\log(\beta_{kl}) + \log(\psi_{kl})) - \log(B) + \log(A) \end{aligned}$$

Conditional expectation :

$$\mathbb{E}_\theta[\log(\mathbb{P}(Y,T))|Y] = \sum_{k,l \in V} \mathbb{P}((k,l) \in E_T|Y)(\log(\beta_{kl}) + \log(\psi_{kl})) - \log(B) + \log(A)$$

Computation of conditional probability : using Bayes theorem, we consider the proportion of trees which contain an edge between the nodes k and l.

$$\begin{aligned} \mathbb{P}((k,l) \in T|Y) &= \sum_{T \in \mathcal{T}:(k,l) \in T} \mathbb{P}(T|Y) \\ &= \frac{\sum_{(k,l) \in T} \mathbb{P}(T)\mathbb{P}(Y|T)}{\sum_T \mathbb{P}(T)\mathbb{P}(Y|T)} \\ &= \frac{\sum_{(k,l) \in T} \prod_{uv} \beta_{uv}\psi_{uv}(Y)}{\sum_T \prod_{uv} \beta_{uv}\psi_{uv}(Y)} \end{aligned}$$

This conditional probability is computed using Kirshner's theorem on the matrix $K = [\beta_{ij}\psi_{ij}]_{1 \le i,j \le d}$.

3.3.2 M step

Moving to the M step, the quantity $\tau_{kl} = \mathbb{P}((k,l) \in E_T|Y)$ has been computed and is now considered as fixed. Parameters of variance have already been optimised thanks to the use of the ML estimators in the formulas above. Here we maximize the conditional expectation with respect to parameters β_{kl}, the edge probabilities.

$$\arg\max_{\beta_{kl}} \left\{ \sum_{k,l \in V} \tau_{kl}(\log(\beta_{kl}) + \log(\psi_{kl})) - \log(B) - n \times cst \right\}$$

We take the derivative with respect to β_{kl}:

$$\frac{\partial \mathbb{E}_\theta[\log(\mathbb{P}(Y,T))|Y]}{\partial \beta_{kl}} = \frac{1}{\beta_{kl}}\tau_{kl} - \frac{1}{B}\frac{\partial B}{\partial \beta_{kl}} \tag{3}$$

Using result (2) from Meila et al and setting equation 3 to 0 we obtain :

$$\boxed{\hat{\beta}_{kl} = \frac{\tau_{kl}}{M_{kl}}}$$

References

[Hammersley and Clifford, 1971] Hammersley, J. M. and Clifford, P. (1971). Markov fields on finite graphs and lattices.

[Kirchhoff, 1847] Kirchhoff, G. (1847). Ueber die auflösung der gleichungen, auf welche man bei der untersuchung der linearen vertheilung galvanischer ströme geführt wird. *Annalen der Physik*, 148(12):497–508.

[Meilă and Jaakkola, 2006] Meilă, M. and Jaakkola, T. (2006). Tractable Bayesian learning of tree belief networks. *Statistics and Computing*, 16(1):77–92.