Modèles à variables latentes pour l'écologie et la biologie

Examen

Note that a few useful formula are given at the end of the document.

We are interested in modeling the count of individuals of the same species on several sites. Let i be the geographical site ($i = 1, \dots, N$). Each site i is visited n_i times. Y_{ij} is the number of individuals observed at site i observation j.

Part 1. Frequentist estimation

We consider the following model:

$$\left\{\begin{array}{ccl} Y_{ij}|Z_i & \sim_{ind} & \mathcal{P}(e^{Z_i}) \\ Z_i & \sim_{i.i.d} & \mathcal{N}(\mu, \sigma^2). \end{array}\right. \tag{1}$$

(Z_i) introduces a variability due to the heterogeneity of the sites.

In what follows, we will use the following notations:

$$\begin{array}{rcl} \boldsymbol{Y} & := & (Y_{ij})_{i=1,\dots,N, j=1,\dots,n_i} \\ \boldsymbol{Z} & := & (Z_i)_{i=1,\dots,N} \\ \theta & := & (\mu, \sigma^2) \end{array}$$

1. Give the expression of the complete log likelihood $\log p(\boldsymbol{Y}, \boldsymbol{Z}; \theta)$.
2. Are you able to give a close form expression of the likelihood of the observations $\boldsymbol{Y}$ and of the conditional distribution $p(\boldsymbol{Z}|\boldsymbol{Y};\theta)$? (Explain why)
3. Recall the general principle of the Variational EM algorithm (lower bound, VE and M step, algorithm...) [Question de cours]

We propose to approximate the conditional distributions $p(Z_i|\boldsymbol{Y})$ in the Gaussian family:

$$\widetilde{q}(\boldsymbol{z}) = \prod_{i=1}^{N} \widetilde{q}_i(z_i) \quad \text{where} \quad \widetilde{q}_i(z_i) = f_{\mathcal{N}(\widetilde{\mu}_i, \widetilde{\omega}_i^2)}(z_i) \tag{2}$$

4. Prove that the entropy of $\widetilde{q}$ is

$$\mathcal{H}(\widetilde{q}) = \sum_{i=1}^{N} \left(\frac{1}{2}\log(2\pi) + \frac{1}{2}\log \widetilde{\omega}_i^2 + 1\right)$$

5. Derive the expression of the lower bound as a function of θ and $\tau = (\widetilde{\mu}_i, \widetilde{\omega}_i^2)_{i=1,\dots,N}$

6. Give the expression of $\widehat{\theta}$ solution of the M-step
7. Give the equation verified by $\widehat{\tau}$ solution of the VE-step. In practice, how do you propose to solve it?
8. Propose an (efficient) initialisation of your VEM algorithm.

Part 2. Bayesian estimation

Assume that each site is described by a collection of p environmental covariates. Let $x_i \in \mathbb{R}^p$ the covariates for site $i = 1, \dots, N$.

We set the following model

$$\left\{\begin{array}{ccl} Y_{ij}|Z_i & \sim_{ind} & \mathcal{P}(e^{Z_i}) \\ Z_i & \sim_{ind} & \mathcal{N}(x_i^T \boldsymbol{\beta}, \sigma^2). \end{array}\right. \tag{3}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$. In this model, a part of the variability between sites is explained by the covariates. We assume that the covariate $x_1, \dots, x_N$ are such that the $n \times p$ matrix $X = \begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix}$ is of full rank. For the sake of simplicity, <u>σ^2 is assumed to be known</u>.

We consider a Bayesian inference for β et set the following prior distribution on β:

$$\boldsymbol{\beta} \sim \mathcal{N}_p(0_p, \omega^2 \mathbf{I}_p) \tag{4}$$

where 0_p is the null vector of size p, $\mathbf{I}_p$ is the identity matrix of size p. $\omega^2 \in \mathbb{R}^{+*}$.

9. What is the role of ω^2 in the prior distribution (4)?
10. Give the expression of $p(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\beta}; \omega^2, \sigma^2)$ for Model (3) and prior distribution on $\boldsymbol{\beta}$ (4).

The aim is to propose a method to obtain a sample from $p(\boldsymbol{\beta}, \boldsymbol{Z}|\boldsymbol{Y})$. Let h be the iteration number. We propose to sample iteratively:

a. $\boldsymbol{\beta}^{(h)} \sim p(\boldsymbol{\beta}|\boldsymbol{Z}^{(h-1)}, \boldsymbol{Y}; \omega^2, \sigma^2)$
b. $\boldsymbol{Z}^{(h)} \sim p(\boldsymbol{Z}|\boldsymbol{\beta}^{(h)}, \boldsymbol{Y}; \omega^2, \sigma^2)$

11. [Question de cours] What is the name of the algorithm? Give quickly its properties.
12. Show that the simulation at step [a.] can be performed exactly (give the distribution of $\boldsymbol{\beta}$ given the latent variable $\boldsymbol{Z}$ and the data $\boldsymbol{Y}$).
13. Are you able to simulate explicitely $\boldsymbol{Z}^{(h)} \sim p(\boldsymbol{Z}|\boldsymbol{\beta}^{(h)}, \boldsymbol{Y}; \omega^2, \sigma^2))$?
14. [Question de cours] Recall the principle of the Metropolis-Hastings algorithm.

15. How can you apply it here at step [b.] of the algorithm?
16. How can you tune the algorithm to reach the adequate convergence rate?

Useful formulae

- $\mathbb{E}[Z^2] = \mathbb{V}(Z) + (\mathbb{E}[Z])^2$
- $\mathcal{H}(q) = \mathbb{E}_{Z\sim q}[\log q(Z)]$
- Assume that $Z \sim \mathcal{N}(\mu, \sigma^2)$ then
  - ⋆ the density is $f(z) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(z-\mu)^2}$
  - ⋆ $\mathbb{E}(e^Z) = e^{\frac{\sigma^2}{2}+\mu}$
- **Gaussian vector** Let X be a full rank matrix of size $n \times p$ ($n < p$). If

$$\begin{array}{rcl} \boldsymbol{Z} = (Z_1, \dots, Z_n) & \sim & \mathcal{N}_n(X\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n) \\ \text{and} \qquad \beta & \sim & \mathcal{N}_p(0_p, \omega^2 \mathbf{I}_p) \end{array}$$

then

$$\begin{array}{rcl} \beta|\boldsymbol{Z} & \sim & \mathcal{N}_p(\mu^{post}, \Omega^{post}) \\ \text{with} \quad \Omega^{post} & = & \left(X^T X + \dfrac{\sigma^2}{\omega^2}\mathbf{I}_p\right)^{-1} \\ \text{and} \quad \mu^{post} & = & \Omega^{post} X^T \boldsymbol{Z} \end{array}$$

where X^T is the transposed matrix of X.