Latent variable models in biology and ecology

Chapter 1: Introduction

Sophie Donnet. INRAE

Master 2 MathSV. January 10, 2023

- Classical statistical models are quite limited in practice

$$Y_i \sim \mathcal{F}_\theta(\cdot)$$

- Purpose of statistical modeling is often to retrieve some hidden process that is at work behind what is observed.
- **Purpose of this lecture**: present a series of statistical models involving hidden (also called latent) variables with application to biology and ecology.

- **Here**: strong focus on cases where the hidden process refers to some unobserved classification
- **Thus** the hidden process is supposed to have a discrete state-space
- **But**: most of the techniques presented hereafter can be generalized to continuous state-space models.

Such models are part of so-called **incomplete data models**.

Inference

- Estimation of the parameters, classification of the observations, estimation of the number of clusters
- Inference requires some specific developments
- Presented techniques : variations around the **expectation-maximization** (EM) algorithm [Dempster et al., 1977]
- Variational versions: see [Jaakkola, 2000] for an introduction or [Wainwright and Jordan, 2008] for a very complete review

Critical step

- Determination of the **conditional distribution of the hidden variables Z given the observed ones Z**

$$p(\mathbf{Z}|\mathbf{Y})$$

- Or at least the calculation of some of its moments.
- Organization of these course is based on this point.
  - The calculation is straightforward

    Mixture models
  - The calculation is not straightforward but still possible.

    Hidden Markov models
  - The calculation is not possible and approximations are needed.

    Stochastic Block models

Organization

Chapter 2 Mixture models and EM algorithm:

Chapter 3 Hidden Markov models

Chapter 4 Stochastic block models and variational EM

Chapter 5 Bayesian inference in latent variable models

Dates (séance de 3h00)

Campus Agro Palaiseau Saclay, C1.1.27

- 11 janv 2023
- 18 janv 2023
- 25 janv 2023
- 01 fev 2023
- 08 fev 2023
- 15 fev 2023

Evaluation

Examen écrit de 3h le 08/03/2023 (sans documents)

References

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977).

Maximum likelihood from incomplete data via the EM algorithm.

Jr. R. Stat. Soc. B, 39:1–38.

Jaakkola, T. S. (2000).

Tutorial on Variational Approximation Methods.

In *In Advanced Mean Field Methods: Theory and Practice*, pages 129–159. MIT Press.

Wainwright, M. J. and Jordan, M. I. (2008).

Graphical Models, Exponential Families, and Variational Inference.

Found. Trends Mach. Learn., 1(1–2):1–305.