Modèle à blocs latents pour l'analyse de données métagénomiques

Julie Aubert [1,2] & Trung Ha [1,2] & Tristan MaryHuard [1,2,3,4,5]

[1] *INRA, UMR 518 MIA, F-75005 Paris, France*
[2] *AgroParisTech, UMR 518 MIA, F-75005 Paris, France*
[3] *INRA, UMR 320 GV, F-91190 Gif-sur-Yvette, France*
[4] *Univ Paris Sud, UMR 8120 GV, F-91190 Gif-sur-Yvette, France*
[5] *CNRS, UMR 8120 GV, F-91190 Gif-sur-Yvette, France*
julie.aubert@agroparistech.fr, trung.ha@agroparistech.fr,
tristan.mary-huard@agroparistech.fr

Résumé. Les modèles à blocs latents fournissent un cadre probabiliste pour la double classification de lignes et colonnes d'une matrice de données. Dans cet article nous considérons un modèle à blocs latents pour des données de comptage surdispersées. Les variables latentes ne sont pas indépendantes conditionnellement aux variables observées ce qui rend l'inférence classique par maximum de vraisemblance impossible. Nous présenterons un algorithme d'inférence basé sur une approche variationnelle. Nous appliquerons ce modèle sur des données de métagénomique pour étudier les interactions entre les bactéries présentes dans la rhizosphère et les plantes.

Mots-clés. modèle de mélange, classification non supervisée, données de comptage, métagénomique

Abstract. Latent block models provide a probabilistic framework for the biclustering of lines and columns of a data matrix. In this paper we consider a latent block model for overdispersed count data. Latent variables are not independent conditional on observed ones and therefore the classical maximum likelihood inference is impossible. We will present an inference algorithm based on a variational approach. We will apply this model to metagenomics data in order to study the interactions between bacteria and plants in the rhizosphere.

Keywords. mixture model, clustering, count data, metagenomics

1 Introduction

Les matrices de données binaires ou de comptage sont présentes dans de nombreuses disciplines et notamment en écologie. La métagénomique, qui étudie le matériel génétique récupéré directement à partir d'échantillons environnementaux, fournit des matrices d'abondance

où les lignes représentent des espèces bactériennes et les colonnes des échantillons biologiques. Dans l'exemple motivant notre travail nous disposons de données correspondant aux abondances de bactéries présentes dans la rhizosphère de différentes plantes. La rhizosphère est la région du sol directement formée et influencée par les racines et les micro-organismes associés. Une case de la matrice contient ainsi l'abondance d'une bactérie dans un échantillon de sol. L'objectif est de trouver des associations entre des communautés microbiennes et des plantes. D'un point de vue plus général, il s'agit de mettre en lumière des relations priviligiées entre des groupes de plantes et des groupes de bactéries, ces groupes étant à découvrir. Les modèles à blocs latents introduits par Govaert et Nadif (2010) fournissent un cadre probabiliste à ce problème de double classification non supervisée.

Nous présenterons l'adaptation d'un modèle à blocs latents pour analyser des données de métagénomique du sol et discuterons de différents choix de modélisation.

2 Modèle

Soient respectivement (Z_i) et (W_j) les labels inconnus (latents) des colonnes et des lignes. Les variables latentes sont indépendantes et respectivement distribuées selon des lois multinomiales :

$$(Z_i) \sim \mathcal{M}(1, \pi = (\pi_1, \ldots, \pi_g)) \quad \text{et} \quad (W_j) \sim \mathcal{M}(1, \rho = (\rho_1, \ldots, \rho_m)).$$

Les observations (X_{ij}),$i = 1, \ldots, n$ et $j = 1, \ldots, d$ sont supposées indépendantes conditionnellement aux variables latentes :

$$X_{ij}|(Z_{ik} = 1, W_{jl} = 1) \sim f(.; \alpha_{kl}).$$

On note $\alpha = (\alpha_{kl})$ le vecteur des paramètres des lois d'émission.

Keribin et al (2013) ont montré que ce modèle est identifiable pour des données binaires et catégorielles. Nous étendons ce résultat sous certaines conditions aux distributions paramétriques associées à des données discrètes.

Les données de métagénomique correspondent à des nombres de séquences attribuées à une bactérie dans un échantillon de sol. Il semble naturel dans un premier temps de considérer que les (X_{ij}) sont distribuées selon une loi de Poisson. Ce cas traité par Govaert et Nadif (2010) constituera notre premier modèle. Les études récentes dont White et al (2009) montrent que ces données sont surdispersées et les comptages à 0 souvent surreprésentés. Les distributions Poisson Gamma ou Poisson avec sur-abondance de zéros sont souvent envisagées comme alternatives. Nous proposerons donc l'utilisation d'une loi zero-inflated Poisson (ZIP) pour notre deuxième modèle.

$$X_{ij}|(Z_{ik} = 1, W_{jl} = 1) \sim ZIP(.; \alpha_{kl})$$

$$\begin{cases} P(X_{ij}|(Z_{ik}=1,W_{jl}=1)=0) = \delta_{kl} + (1-\delta_{kl})e^{-\lambda_{kl}} \\ P(X_{ij}|(Z_{ik}=1,W_{jl}=1)=x_{ij}) = (1-\delta_{kl})\frac{\lambda_{kl}^{x_{ij}}e^{-\lambda_{kl}}}{x_{ij}!} \text{ pour } x_{ij} \geq 1 \end{cases}$$

Le vecteur des paramètres des lois d'émission devient $\alpha = (\alpha_{kl})$ avec $\alpha_{kl} = (\delta_{kl}, \lambda_{kl})$, où δ_{kl} correspond à la probabilité d'excès en zéros dans le bloc kl.

Nous utiliserons pour effectuer l'inférence une stratégie basée sur une approche variationnelle (Govaert et Nadif (2010), Wainwright et Jordan (2008)).

3 Discussion

Nous discuterons de l'extension de ces modèles à la prise en compte de covariables. En effet les colonnes peuvent être caractérisées par différents facteurs tels qu'un groupe de traitement ou un protocole expérimental susceptible d'influencer la variable observée. Nous considérerons comme troisième modèle un mélange de régression de Poisson :

$$X_{ij}|(Z_{ik}=1,W_{jl}=1) \sim \mathcal{P}(\lambda_{kl}e^{\beta^\intercal y_{ij}})$$

où y_{ij} est le vecteur des covariables.

Nous appliquerons les différents modèles proposés à des données de métagénomique de la rhizosphère.

Bibliographie

[1] Govaert, G. et Nadif, M. (2010), *Latent Block Model for contingency table*, Communications in Statistics - Theory and Methods, 39, 3, 416–425.
[2] White, J. et Nagarajan, N. et Pop, M. (2009), *Statistical Methods for Detecting Differentially Abundant Features in Clinical Metagenomic Samples*, PLoS Computational Biology, 5, 4.
[3] Keribin, C. et Brault, V. et Celeux, G. (2013), *Estimation and selection for the latent block model on categorical data*, INRIA Research report 8264.
[4] Wainwright, M. J. and Jordan, M. I. (2008), *Graphical models, exponential families, and variational inference. Found.*, Trends Mach. Learn.,1, 1–305.