Modèles de mélange tronqués pour l'écologie microbienne. Estimation du nombre d'espèces manquantes.

Sebastien Li-Thiao-Te, Jean-Jacques Daudin, Stéphane Robin, Emilie Lebarbier

Modèles de mélange tronqués pour l'écologie microbienne. Estimation du nombre d'espèces manquantes.

Sébastien Li-Thiao-Té & Stéphane Robin & Jean-Jacques Daudin & Émilie Lebarbier

Équipe Statistique et Génome,
UMR 518 AgroParisTech / INRA
16, rue Claude Bernard
75231 Paris CEDEX 05

Résumé français

Dans le modèle d'échantillonnage présenté par Fisher en 1943 [1], chaque espèce apporte à l'échantillon un nombre d'individus distribué selon une loi de Poisson. La moyenne de cette loi est spécifique à l'espèce. Cependant, de nombreuses espèces apportent zéro individus et ne sont donc pas observées dans les données.

Nous utilisons des modèles de mélanges tronqués pour l'abondance des espèces et estimons les paramètres par maximum de vraisemblance [2]. À cause des espèces manquantes, l'algorithme EM n'est pas applicable directement. Nous appliquons l'algorithme EM à un autre modèle, le mélange de distributions tronquées [3], et corrigeons le résultat.

Pour obtenir des intervalles de confiance, nous utilisons l'approche variationnelle dans un cadre Bayésien [4]. De même que précédemment, il est plus facile d'appliquer l'algorithme pour un mélange de distributions tronquées et d'en déduire la solution du problème.

Un cas d'application de ces méthodes consiste en l'étude de la diversité de communautés microbiennes en métagénomique. Comme de nombreuses souches ne sont pas cultivables en laboratoire, on recense les espèces bactériennes par séquençage ADN, et l'on dénombre pour chaque espèce le nombre de fragments ADN observés. Les méthodes présentées permettent d'évaluer le nombre d'espèces présentes mais non observées et de calculer des courbes de raréfaction pour la planification des expériences [5].

Résumé anglais

In the sampling model presented by Fisher in 1943 [1], species contribute a Poisson-distributed number of individuals to the dataset, and the intensity parameters are species-specific. Those species that contributed zero individuals are not recorded in the dataset.

We model the species abundance distribution using truncated mixture models, and estimate the parameters by maximum likelihood [2]. Due to the missing species, straightforward algorithms like EM are not applicable. However, the solution can be obtained by applying the EM algorithm to a different model, i.e. a mixture of truncated distributions, then correcting the output [3].

To obtain confidence intervals on the estimates, we replace the EM algorithm with the Variational Bayes EM method [4]. Likewise, it is easier to work with mixtures of truncated distributions and deduce the confidence intervals.

We estimate the number of species in microbial diversity surveys conducted in the metagenomics approach. As most bacterial strains cannot be cultured in laboratory conditions, DNA sequencing is applied directly to the biological sample and the number of DNA fragments are recorded for each strain. We then derive the expected number of species that were not observed by DNA sequencing, as well as rarefaction curves for plannification of future experiments [5].

Mots clés

Biostatistique, Méthodes bayésiennes

Bibliographie

[1] Fisher, RA and Corbet, A.S. and Williams, CB (1943) The relation between the number of species and the number of individuals in a random sample of an animal population, *Journal of Animal Ecology*, 12, 1, 42–58.
[2] Bunge, J. and Barger, K. (2008) Parametric models for estimating the number of classes, *Biometrical Journal*, 50, 5.
[3] Bohning, D. and Kuhnert, R. (2006) Equivalence of truncated count mixture distributions and mixtures of truncated count distributions, *Biometrics*, 62, 4, 1207–1215.
[4] Beal, M. J. and Ghahramani, Z. (2003) The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures, *Bayesian statistics 7: proceedings of the seventh Valencia International Meeting, June 2-6, 2002*, 453.
[5] Mao, C.X. and Lindsay, B.G. (2007), Estimating the number of classes, *Annals of Statistics*, 35, 2, 917.