Extending the supervised component-based generalized linear regression to joint response modelling for species-rich ecosystems.

Context

Abrupt climate change or land-use changes are triggering species extinctions, range shifts and biotic homogenization. The development of models able to predict the future of biodiversity as well as ecosystem services is now a critical endeavor, as stated by the recently published report of IPBES. Recently, tremendous progress has been made in extending species distribution models (SDMs) to Joint Species Distribution Models (JSDMs). These advances allow the integration of ecological mechanisms more efficiently by considering how species dependencies shape community composition. On the other hand, modelling the abundances of species requires taking into account a large number of potential explanatory covariates for which some sort of dimension-reduction and regularization must be performed.

Methods and objectives

Bry et al. (2013, 15, 18) proposed a technique, called *supervised component-based generalized linear regression* (SCGLR), which bridges the multivariate generalized linear model estimation with component-based dimension reduction of the explanatory space. SCGLR optimizes a trade-off criterion between the goodness-of-fit of the model and some structural strength or relevance of directions with respect to the explanatory variables. Doing so, this technique not only finds interpretable and strong explanatory dimensions, but also produces regularized predictors, something much needed in a high-dimensional framework. SCGLR extends many dimension reduction data analysis techniques, such as PLS regression (Wold, 2001), PCA on instrumental variables, canonical correspondence analysis (Ter Braak, 1987), and other empirical methods related to generalized linear modelling. SCGLR is designed to analyze and model responses of different types simultaneously (Poisson and/or Bernoulli and/or Gaussian…). SCGLR has recently been extended (Chauvet et al. 2018, 2019) to take into account random effects in the GLM (i.e. to perform supervised component-based generalized linear mixed effects modelling). This extension allows to model repeated measurements and panel-data with an autoregressive temporal random effect. However, SCGLR still assumes that all responses (species counts) depend on the same set of explanatory dimensions and are independent conditional on these common explanatory dimensions.

The first issue is to identify communities from the data, finding clusters of species such that species in a cluster can be modelled using the same explanatory dimensions, while being possibly different across clusters. One approach would be to consider mixtures of regressions. While mixtures have commonly been used to cluster observations, few works have proposed to cluster outcomes (Monni & Tadesse, 2009, Mortier et al. 2015, Tadesse et al 2016…). The second issue will consist to extend SCGLR and its clustering generalization to overcome conditional independence adapting graphical lasso to SCGLR (Friedman et al., 2007), thereby allowing the estimation of dependencies between outcomes.

Data and case studies

The PhD student will develop methods to address these issues and will apply them to species, traits and ecosystem services data from a comprehensive collection of datasets shared with GAMBAS' partners from different ecosystems (terrestrial and aquatic, temperate and tropical). This will allow the student to collaborate with all GAMBAS' partners.

Practical Information

The PhD fellowship will start on October 1st 2019. It will be funded by the GAMBAS project involving six partners (CIRAD, IRSTEA, CNRS, MNHN, Univ de Montpellier and Paris-Sud Orsay). The PhD student will be based at Université de Montpellier.

PhD Director:
Catherine Trottier, Université de Montpellier, UMR IMAG, Equipe Probabilités et Statistique, Place Eugene Bataillon, 34095 Montpellier Cedex 5
Tel. 04 67 14 41 64
Co-supervisors:
Xavier Bry (Université de Montpellier, UMR IMAG) and *Frédéric Mortier* (CIRAD, Forêts et Sociétés).

Candidate profile

The candidate should have a training in applied mathematics with a strong background in statistics and probability. He/she should also be interested in applications of mathematical tools to the applied field of ecology. The candidate should have good skills in scientific English (reading, speaking and writing) and work organization (autonomous and good reporting skills). In addition, the candidate must have good computational skills and will be expected to develop R packages, such as incorporating the new methodological developments in the SCGLR package.

Candidates selection

A first phase of selection will be organized in mid 2019. Candidates have to send:

- a CV,
- a motivation letter,
- two letters of recommendation,
- a Master's thesis report

by e-mail to Catherine Trottier (catherine.trottier@umontpellier.fr) before June 15th 2019. Selected candidates will be interviewed between late June and early July by the researchers involved in the project.

References

1. Brown P., Vannucci M. and Fearn T. (1998): *Multivariate Bayesian variable selection and prediction*. Journal of the Royal Statistical Society, Series B.
2. Bry X., Trottier C., Mortier F. and Cornu G. (2018): *Component-based regularisation of a multivariate GLM with a thematic partioning of the explanatory variables*. Statistical Modelling.
3. Bry X., Trottier C., Verron T. and Mortier F. (2013): *Supervised component generalized linear regression using a PLS-extension of the Fisher scoring algorithm*. Journal of Multivariate Analysis.
4. Chauvet J., Trottier C. and Bry X. (2019): *Component-based regularisation of multivariate generalised linear mixed models*, Journal of Computational and Graphical Statistics (in press).
5. Cornu G., Mortier F., Trottier C. and Bry X. (2018): *SCGLR: Supervised Component Generalized Linear Regression. R package version 3.0*. https://CRAN.R-project.org/package=SCGLR
6. Friedman, J., Hastie, T., and Tibshirani, R. (2007): *Sparse inverse covariance estimation with the graphical LASSO*. Biometrics.
7. Monni S. and Tadesse M.G. (2009): *A stochastic partitioning method to associate high-dimensional responses and covariates*. Bayesian Analysis.

8. Mortier F. , Ouédraogo D.-Y., Claeys F., Tadesse M.G., Cornu G., Baya F., Benedet F., Freycon V., Gourlet-Fleury S. and Picard N. (2015): *Mixture of inhomogeneous matrix models for species-rich ecosystems*. Environmetrics.
9. Tadesse M.G., Mortier F. and Monni S. (2016): *Uncovering cluster structure and group-specific associations: variable selection in multivariate mixture regression models*. Interdisciplinary Mathematical Research and Applications. Ed. B. Toni : Springer.