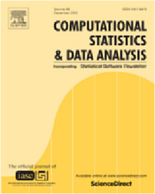
Computational Statistics and Data Analysis 105 (2017) 1–10

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda



Model-based simultaneous clustering and ordination of multivariate abundance data in ecology

Francis K.C. Hui

Mathematical Sciences Institute, The Australian National University, Canberra, 0200, Australia

ARTICLE INFO

Article history:
Received 21 October 2014
Received in revised form 13 July 2016
Accepted 14 July 2016
Available online 26 July 2016

Keywords:
Dimension reduction
Finite mixture models
Hierarchical Bayesian model
Mixtures of factor analyzers
Latent variable model

ABSTRACT

When studying multivariate abundance data, one of the main patterns ecologists are often interested in is whether the sites exhibit clustering on the low-dimensional, ordination space representing species composition. A new model-based approach called CORAL (Clustering and Ordination Regression AnaLysis) is developed for tackling this question, based on performing simultaneous clustering and ordination using latent variable regression. By drawing the latent variables from a finite mixture density, CORAL probabilistically classifies sites based on their positions on an underlying signal space. This is similar to mixtures of factor analyzers, except CORAL is designed for non-normal responses and uses species-specific rather than cluster-specific factor loadings (regression coefficients). Estimation is performed via Bayesian MCMC sampling, with code provided in the Supplementary Material. Simulations demonstrate that, by utilizing the joint information available in the data for both classification and dimension reduction, CORAL outperforms several popular, algorithm-based methods for clustering and ordination in ecology. CORAL is applied to a dataset of presence–absence records collected at sites along the Doubs River near the France–Switzerland border, with results revealing two clusters or ecological regions partly resembling the spatial separation of upstream and downstream sites.

1. Introduction

Multivariate abundance data, consisting of species occurrences (typically in the form of presence–absence responses or counts) at a set of sites, are commonly collected in ecology. One of the main goals in collecting such data is to understand how sites differ in terms of species relative abundance or composition, without explicit reference to explanatory variables. For example, a motivating dataset we consider contains presence–absence records of 27 fish species at 30 sites along the Doubs river near the France–Switzerland border (Verneaux, 1973). One of the primary aims for collecting this data is to study whether sites along the river (and potentially other European river and stream networks) could be clustered into "ecological regions" with similar species abundance and/or composition. This would provide a simplified interpretation of the river's biodiversity as a whole, while facilitating conservation planning for particular fish species by focusing efforts on the region which they dominate.

To understand how sites differ in species abundance and/or composition, ecologists commonly apply two approaches: (1) cluster analysis to classify sites based on species abundance or composition, (2) unconstrained ordination to filter the species responses and visualize the data on a low-dimensional plot (see Chapters 8 and 9, Legendre and Legendre, 2012). On the other hand, one of the main patterns of interest is whether sites exhibit any clustering on a latent, ecological signal space That is, do sites tend to group together in terms of their positions on the indirect gradient surface, reflecting the fact that they

E-mail address: fhui28@gmail.com.

http://dx.doi.org/10.1016/j.csda.2016.07.008
0167-9473/© 2016 Elsevier B.V. All rights reserved.

have similar species composition or relative abundance (see ter Braak and Prentice, 1988, for a review of indirect gradient analysis). Neither of the two approaches on its own can answer this question: cluster analysis produces site classifications on the noisy species response instead of on the underlying signal, while ordination makes no attempt to formally cluster sites and therefore any conclusion regarding the existence of clustering is exploratory. One common but *adhoc* method to get around this problem is to apply both cluster analysis and ordination separately on the same dataset, and visually combine the results in a single plot (see for instance Figure 2 in Moritz et al., 2013). However, this method fails to take advantage of any joint information from clustering and ordination, which would be useful for helping one method inform the other e.g., it is likely that the signal for clustering is more transparent on the latent or indirect signal space produced by ordination.

Most of the methods currently used for cluster analysis and ordination in ecology are algorithm-based, that is, they apply a series of algebraic operations to a matrix of pairwise dissimilarities between sites e.g., the most popular choice is the Bray–Curtis dissimilarity (Bray and Curtis, 1957). Examples of algorithm-based techniques include Ward clustering (Ward, 1963) and K-means clustering for classification, and Non-metric Multidimensional Scaling (NMDS, Kruskal and Wish, 1978) and Correspondence Analysis (CA, Hill, 1974) for ordination. The development of algorithm-based techniques for analyzing multivariate data in general remains an ongoing area of research (e.g., Polak et al., 2009; Gijbels and Omelka, 2013).

In contrast to algorithm-based methods, clustering and ordination can be approached from a model-based framework. Latent variables models, which include finite mixture models and factor analysis as particular cases, are widely used for clustering or dimension reduction outside of ecology (Fraley and Raftery, 2002). One method capable of simultaneously performing both is Mixtures of Factor Analyzers (MFAs, McLachlan and Peel, 2000) models, which use a finite mixture of multivariate normal densities to represent clusters on a latent space (see also McNicholas and Murphy, 2008; Murray et al., 2014, for examples of extensions). Despite this, using MFA models or other model-based methods for ordination or clustering has received little attention in ecology, and only in recent years has there been a push to bridge the gap between methodology and application (e.g., Dunstan et al., 2013; Hui et al., 2015a,b; Warton et al., 2015).

Motivated by the Doubs river dataset, we propose a hierarchical framework called CORAL (Clustering and Ordination Regression AnaLysis) for performing simultaneous classification and ordination. CORAL is based on a two level hierarchical model: at the first level, a Generalized Linear Model (GLM) is used to regress a non-normal species response against a small number of latent variables. At the second level, the latent variables are drawn from a finite mixture of normal densities. This model allows CORAL to probabilistically classify sites on a latent signal space, from which a simultaneous clustering and ordination plot can be constructed. CORAL can be regarded as an extension of the standard MFA model in two ways. First, multivariate abundance data are non-normal, meaning the marginal likelihood does not have a closed form. Second, the factor loadings in MFA models are cluster-specific, whereas in CORAL they are species-specific and used to describe the species' mean responses. We used Markov Chain Monte Carlo methods to estimate CORAL models, with code provided in the Supplementary Material. Simulations show that CORAL outperformed popular, algorithm-based techniques for clustering and ordination in ecology. Applying CORAL to the Doubs river dataset reveals two ecological regions corresponding roughly to a spatial separation of upstream and downstream sites, although the exact groupings depend on whether the ordination is performed in terms of species composition or relative abundance.

2. Model-based clustering and ordination using CORAL

Multivariate abundance data are represented by a $n \times s$ response matrix $\boldsymbol{Y}$, with observations collected at $i = 1, \ldots, n$ sites for $j = 1, \ldots, s$ species. We focus on the two most commonly observed response types: presence–absence, where $Y_{ij} = 1$ if species j is present at site i and $Y_{ij} = 0$ otherwise, and count data. Two features are typically seen with multivariate abundance data. First, $\boldsymbol{Y}$ is high-dimensional with s/n being a non-negligible ratio. Second, most species are rarely observed, leading to $\boldsymbol{Y}$ being sparse. Both these features are seen in the Doubs river dataset, which has $n = 30$, $s = 27$, and 54% of the entries in $\boldsymbol{Y}$ are equal to zero. The multi-species nature of $\boldsymbol{Y}$ motivates using ordination to visualize the relationship between sites on a low-dimensional plot representing the underlying signal of species abundance or composition. The sparsity of $\boldsymbol{Y}$ is a contributing factor to the strong mean–variance relationship in the response, and must be properly accounted for to ensure subsequent analyses do not produce misleading results; see the worked example in Section 5.

CORAL uses a two level hierarchical model to perform simultaneous clustering and ordination. At the first level, the mean response for species j at site i, denoted as μ_{ij}, is regressed against $d \ll s$ latent variables, $\boldsymbol{u}_i = (u_{i1}, \ldots, u_{id})$, using a GLM,

$$[y_{ij}|\boldsymbol{u}_i, \boldsymbol{\Psi}_1] \sim f(y_{ij}|\mu_{ij}, \phi_j); \qquad g(\mu_{ij}) = \eta_{ij} = \beta_i + \beta_{0j} + \boldsymbol{u}_i^T \boldsymbol{\beta}_j, \tag{1}$$

where $g(\cdot)$ is the link function, $f(y_{ij}|\mu_{ij}, \phi_j)$ is the distributional assumption for element y_{ij} with mean μ_{ij} and dispersion parameter ϕ_j. The species-specific intercepts and regression coefficients are denoted by β_{0j} and $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jd})$ respectively. A site effect, β_i, can also be incorporated to adjust for differences in site total abundance; see Section 2.1 for further discussion. In factor analysis, $\boldsymbol{u}_i$ and $\boldsymbol{\beta}_j$ are known as factor scores and factor loadings respectively, although with CORAL we prefer to call them latent variables and species-specific coefficients. We use a Bernoulli distribution with logit link for presence–absence data, while for count data we use either the Poisson or negative binomial distribution with $\text{Var}(y_{ij}) = \mu_{ij} + \phi_j \mu_{ij}^2$ and the log link. Let $\boldsymbol{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_s)^T$ be the $s \times d$ matrix of regression coefficients, $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0s})$, and $\boldsymbol{\Psi}_1 = \{\beta_1, \ldots, \beta_n, \boldsymbol{\beta}_0, \phi_1, \ldots, \phi_s, \text{vec}(\boldsymbol{B})\}$ denote all the parameters in the first level of CORAL.