Factorial Clustering of Species Distribution Data

Manfred Jaeger

Institut for Datalogi, Aalborg University
Selma-Lagerlöfs Vej 300, 9220 Aalborg Ø, Denmark
jaeger@cs.aau.dk

Introduction

Based on long-term data collection, quite extensive and detailed data on the distribution of plant species is available for many parts of the world (for North America e.g. the USDA plants database plants.usda.gov; for Europe the Atlas Florae Europaeae http://www.luomus.fi/english/botany/afe). The analysis of species distribution patterns has led to the identification of vegetation zones which divide a given geographic region into zones of relatively homogeneous species composition. Based on manual analysis, vegetation zones were already studied in the 19th century. Algorithmic methods of hierarchical clustering were developed by Orloci [2], and applied to Swiss distribution data by Wohlgemuth [3].

Traditional methods of floristic analysis lead to a single division (or possibly a hierarchical division, providing views at different levels of granularity) of the geographic region. However, the division so obtained is in fact a combination of multiple environmental, geographic or historical factors that influence the distribution of a plant species. Each of these factors (or combinations of some of them) defines a division of its own of the underlying geographic area. The identification and analysis of the most relevant factors underlying species distribution patterns in a particular geographic region is necessary for an understanding of observed vegetation zones and biological diversity, and the protection of endangered habitats.

Multiple clustering is a recently emerging area in Machine Learning that investigates techniques for computing multiple clusterings of a dataset, where the different clusterings provide alternative views of the structure in the data. We propose a specific method for multiple clustering that is well-suited for high-dimensional 0/1-valued data, and is designed to create clusterings which can be interpreted as causal factors that influence the values in each dimension. We apply this *factorial clustering* method to the distribution data of 2398 plant species over a division of Switzerland into 565 mapping areas. First results show that the method constructs clusterings corresponding to interpretable and meaningful distribution factors.

Factorial Clustering

Our factorial clustering approach [1] is based on a probabilistic model for the observed 0/1-valued species occurrence variables $\boldsymbol{X} = X_1, \ldots, X_k$ ($k = 2398$ in

our data) as a conditional distribution $P(\boldsymbol{X} \mid \boldsymbol{L})$ dependent on m latent variables $\boldsymbol{L} = L_1, \ldots, L_m$. Each latent variable L_j has r_j different states. L_j can be *ordinal*, in which case its states are the integers $0, 1, \ldots, r_j - 1$, or *nominal*, in which case the states are unordered labels $l_0, l_1, \ldots, l_{j-1}$. After fitting the conditional distribution model $P(\boldsymbol{X} \mid \boldsymbol{L})$, one obtains m clusterings by computing for the species occurrence vector $\boldsymbol{x}_i$ of the ith mapping area the most likely cluster-label vector $\boldsymbol{l}_i^* := arg\,max_{\boldsymbol{l}}\, P(\boldsymbol{X} = \boldsymbol{x}_i \mid \boldsymbol{L} = \boldsymbol{l})$. The value of L_j in $\boldsymbol{l}_i^*$ is the cluster index of $\boldsymbol{x}_i$ in the jth clustering. If L_j is an ordinal variable, then the clusters in the jth clustering are ordered.

We assume that the species variables X_h are independent given the latent class variables $\boldsymbol{L}$, and conditional distributions $P(X_h \mid \boldsymbol{L})$ follow a logistic regression model, which for ordinal L_j takes the form

$$\log P(X_h = 1 \mid \boldsymbol{L}) \sim w_{h,0} + \sum_j w_{h,j} L_j$$

with coefficients $w_{h,0}, \ldots, w_{h,m}$. As usual, a nominal L_j is encoded by r_j 0/1-valued indiciator variables for its states.

The fitting of the model parameters $w_{h,j}$ is performed by alternating until convergence a parameter fitting and a cluster label imputation step:

i $\boldsymbol{w}_t := arg\,max_{\boldsymbol{w}}\, P(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{L} = \boldsymbol{l}_t, \boldsymbol{w})$
ii $\boldsymbol{l}_{t+1} := arg\,max_{\boldsymbol{l}}\, P(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{L} = \boldsymbol{l}, \boldsymbol{w}_t)$

The complexity of this method is $O(nk2^m)$, where n is the number of datapoints (i.e., mapping regions in our data). It is therefore well-suited for high-dimensional data, but limited to a relatively small number of clusterings.

Results

We have implemented the factorial clustering method in R using the nnet package for fitting logistic regression models.

To illustrate the intuitions underlying our approach, and to test the algorithmic feasibility, we first perform an experiment with synthetic data. For this, the 565 mapping areas of the Swiss floral data are artificially divided in two alternative ways representing two different hypothetical ecological factors as shown in Figure 1 (a),(b).

Based on these two factors, we obtain distribution types for plant species by assuming that each species has a preference for one specific segment of each factor, and that it grows with high probability in areas belonging to both prefered segments, with lower probability in areas belonging to only one of the prefered segments, and with lowest probability in areas belonging to no prefered segment. Figure 1 (c) illustrates the distribution type of species with a preference for the green segment in both factors. Darker color represents a higher probability of occurrence in the given areas (the concrete occurrence probabilities we used were 0.98, 0.5, and 0.018 for the three different colors). For each of the resulting 6 different distribution types we sampled 15 random species distribution maps.

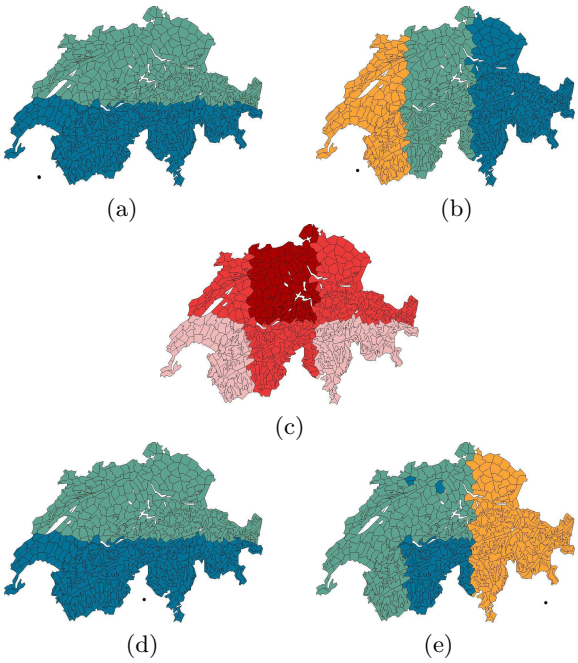

Fig. 1. Experiment with Synthetic Data:

Since the factorial clustering algorithm starts with a random initialization $\boldsymbol{L} = \boldsymbol{l}_0$ for the cluster assignments, one obtains some variations in the results from different runs of the algorithm. For our synthetic data, the result coincided exactly with the underlying segmentations (a),(b) in about one out of three runs of the algorithm. In other runs clusterings as shown in Figure 1 (d),(e) were produced. However, these sub-optimal clusterings were distinguished from the correct one by lower likelihood scores, so that multiple restarts of the algorithm with final selection of the clustering with maximal likelihood score robustly leads to the correct solution.

We next apply our method to the source data for the "Swiss Web Flora" [1] [3]. Figure 2 shows the result of factorial clustering with two nominal latent variables with 3 states each. A comparison with a division of the mapping areas into mountain regions (above timberline) and valley regions (below timberline) shows a very close correspondence of the first clustering with the altitude factor. The second clustering (Figure 2 right) largely corresponds to a basic north-south division defined by the alp mountains. There are multiple ecological factors that correlate with this division, including temperature gradients and geological

[1] `www.wsl.ch/land/products/webflora/welcome-en.ehtml`.
The data is available at
`www.wsl.ch/staff/thomas.wohlgemuth/datasets/swf-all.mrf.gz`

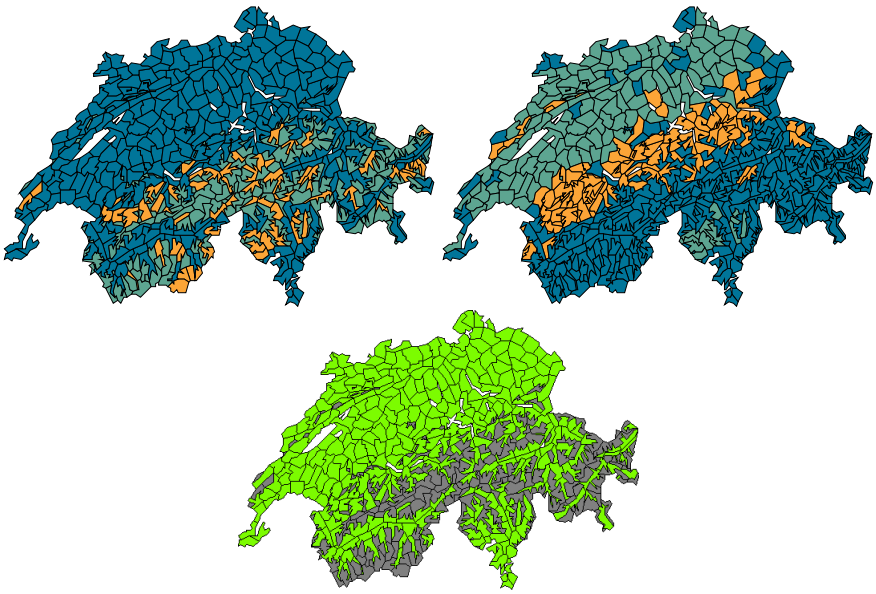

Fig. 2. Clustering result with two nominal latent variables (top), and actual mountain/valley division of areas (bottom)

factors, so that an interpretation of this clustering in terms of a single dominating factor is probably impossible.

References

1. M. Jaeger, S. P. Lyager, M. W. Vandborg, and T. Wohlgemuth. Factorial clustering with an application to plant distribution data. In *Proceedings of the 2nd MultiClust Workshop: Discovering, Summarizing and Using Multiple Clusterings*, 2011. Online proceedings http://dme.rwth-aachen.de/en/MultiClust2011.
2. L. Orloci. An agglomerative method for classification of plant communities. *The Journal of Ecology*, 55(1):193–206, 1967.
3. T. Wohlgemuth. Biogeographical regionalization of switzerland based on floristic data: How many species are needed? *Biodiversity Letters*, 3(6):180–191, 1996.