Factorial Clustering with an Application to Plant Distribution Data

Manfred Jaeger[1], Simon P. Lyager[1], Michael W. Vandborg[1], and Thomas Wohlgemuth[2]

[1] Dept. of Computer Science, Aalborg University, Denmark
[2] Swiss Federal Research Institute WSL

Abstract. We propose a latent variable approach for multiple clustering of categorical data. We use logistic regression models for the conditional distribution of observable features given the latent cluster variables. This model supports an interpretation of the different clusterings as representing distinct, independent factors that determine the distribution of the observed features. We apply the model for the analysis of plant distribution data, where multiple clusterings are of interest to determine the major underlying factors that determine the vegetation in a geographical region.

1 Introduction

There exist a variety of different approaches to learning multiple clusterings. They can differ not only with regard to their mathematical models and algorithmic methods, but there can also be widely different intuitions and objectives with regard to the interpretation of the multiple clusterings. On the one hand, in ensemble clustering, the individual clusterings are essentially regarded as different, imperfect versions of a single underlying true clustering (e.g. [10]). In many multiple clustering methods, on the other hand, the different clusterings are intended to represent different views of the data, each providing a different insight into the structure of the data. One objective for clustering methods then is to ensure that different clusterings are in some sense independent, disparate [6], or non-redundant [8].

Probabilistic latent variable models are used for a variety of data analysis tasks (in the case of discrete data jointly referred to as *latent class analysis*), including clustering. Several authors have investigated probabilistic latent variable models for multiple clusterings [13, 4, 2]. For the case of discrete observable features, no special assumptions on the distributional form of the features given the latent variables are made in these approaches, i.e. the conditional distribution of the features follows an unconstrained multinomial distribution. Latent variable models are also commonly used for dimensionality reduction of high-dimensional numeric data. An important example is the *factor analysis* model, in which the observed data is interpreted as a noisy linear transformation of a small number of latent dimensions. While it is quite common to refer to latent class analysis

as a "categorical data analogue to factor analysis" [5][1, Chapter 13], it seems that this correspondence has not been fully exploited for clustering applications, or put into the context of multi-clustering, via the combined use of multiple latent variables, and special assumptions on the conditional distribution of the observed features.

In this paper we propose a probabilistic latent variable model for multiple clusterings. As in factor analysis, we interpret the observed (discrete) data as a noisy transformation of underlying, discrete latent dimensions. The linear mapping of factor analysis is replaced by a log-linear logistic regression model. The latent dimensions then define clusterings that can be seen as independent factors that determine the distribution of the observed features.

In contrast with several other multiple clustering methods (e.g., [4, 8]) our method is not based on an association of different clusterings with different feature subsets, even though such associations can emerge.

Our approach is partly motivated by applications to biogeographical data. Specifically, we are investigating plant distribution data. Segmentations of geographic units into floristic regions based on similarity of plant species composition were already undertaken in the 19th century. An early application of formal methods of clustering in this context is [9]. We apply our method to distribution data for 2398 plant species in Switzerland. The goal of factorial clustering for this type of data will be to obtain multiple clusterings, each of which could correspond to one of several underlying environmental, geographical, or historical factors, which jointly influence the vegetation.

2 Latent Variable Models for Clustering

Latent variable models are routinely used for clustering, both for single and multiple clustering. However, they can be used in several, slightly different ways. In order to more clearly explain our approach, we briefly review in this section possible approaches to using latent variable models for clusterings.

Throughout, we assume that the observable data $\boldsymbol{X}$ consists of n observations of k attributes, i.e. $\boldsymbol{X}$ is an $n \times k$ matrix. A latent variable model contains m additional unobserved variables, and we denote with $\boldsymbol{L}$ the $n \times m$ matrix of the latent variables in the n observations. We note that when we assume that in the n observations both the observable and latent variables are identically and independently sampled, it will be simpler and more natural to describe the model in terms of vectors $\boldsymbol{X}$, $\boldsymbol{L}$ of length n and m, respectively. However, in some applications, especially segmentation of time sequences or images, the latent variables are not independent at different data points.

A latent variable model, then, consists of a joint distribution for $\boldsymbol{X}$ and $\boldsymbol{L}$, which can be written as

$$P(\boldsymbol{X} \mid \boldsymbol{L}, \theta_{\boldsymbol{X}|\boldsymbol{L}}) P(\boldsymbol{L} \mid \theta_{\boldsymbol{L}}). \tag{1}$$

In hierarchical models, this might be extended by a distribution over $\theta_{\boldsymbol{X}|\boldsymbol{L}}, \theta_{\boldsymbol{L}}$ parametrized by hyperparameters $\boldsymbol{\lambda}$.

The perhaps most common use of model (1) for clustering is to perform two steps [13]: first, fit the parameters $\theta_{\boldsymbol{X}|\boldsymbol{L}}, \theta_{\boldsymbol{L}}$ by maximizing the marginal likelihood of the observed data $\boldsymbol{X} = \boldsymbol{x}$:

$$(\theta^*_{\boldsymbol{X}|\boldsymbol{L}}, \theta^*_{\boldsymbol{L}}) := \underset{\theta_{\boldsymbol{X}|\boldsymbol{L}}, \theta_{\boldsymbol{L}}}{arg\,max} \sum_{\boldsymbol{l}} P(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{L} = \boldsymbol{l}, \theta_{\boldsymbol{X}|\boldsymbol{L}}) P(\boldsymbol{L} = \boldsymbol{l} \mid \theta_{\boldsymbol{L}}). \tag{2}$$

This step is usually performed using the EM algorithm. Then, compute the most probable values of $\boldsymbol{L}$ given $\boldsymbol{X} = \boldsymbol{x}$:

$$\boldsymbol{l}^* = \underset{\boldsymbol{l}}{arg\,max}\, P(\boldsymbol{L} = \boldsymbol{l} \mid \boldsymbol{X} = \boldsymbol{x}, \theta^*_{\boldsymbol{X}|\boldsymbol{L}}, \theta^*_{\boldsymbol{L}}) \tag{3}$$

In multiple clustering, a joint configuration of the latent variables defines multiple cluster indices. For simplicity we may assume for now that each latent variable defines its own clustering, and that therefore the membership of the ith data item in the jth clustering is given by $\boldsymbol{l}^*_{i,j}$. However, in the multi-cluster case, the second step can also take a slightly different form, and the most probable latent variable values be computed component-wise. Denoting by $\boldsymbol{l}_j$ the jth column of $\boldsymbol{l}$ (i.e., $\boldsymbol{l} = (\boldsymbol{l}_1, \ldots, \boldsymbol{l}_m)$), this can be written as

$$\boldsymbol{l}^*_j = \underset{\boldsymbol{l}_j}{arg\,max} \sum_{\boldsymbol{l}_1, \ldots, \boldsymbol{l}_{j-1}, \boldsymbol{l}_{j+1}, \boldsymbol{l}_m} P(\boldsymbol{L} = \boldsymbol{l} \mid \boldsymbol{X} = \boldsymbol{x}, \theta^*_{\boldsymbol{X}|\boldsymbol{L}}, \theta^*_{\boldsymbol{L}}). \tag{4}$$

This is the (hard) clustering rule used, e.g., in [13, 14]. The clusterings obtained from (3) and (4) can differ.

If the ultimate goal is only to compute a most probable configuration of $\boldsymbol{L}$, then one may also try to simplify the combination of (2) and (3) into a single optimization:

$$\boldsymbol{l}^* := \underset{\boldsymbol{l}}{arg\,max} \underset{\theta_{\boldsymbol{X}|\boldsymbol{L}}, \theta_{\boldsymbol{L}}}{max} P(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{L} = \boldsymbol{l}, \theta_{\boldsymbol{X}|\boldsymbol{L}}) P(\boldsymbol{L} = \boldsymbol{l} \mid \theta_{\boldsymbol{L}}). \tag{5}$$

This rule can be justified by a Bayesian interpretation, for example: it amounts to finding the jointly most probably values of $\boldsymbol{l}, \theta_{\boldsymbol{X}|\boldsymbol{L}}, \theta_{\boldsymbol{L}}$, given the data $\boldsymbol{X} = \boldsymbol{x}$, and assuming a uniform prior for $\theta_{\boldsymbol{X}|\boldsymbol{L}}, \theta_{\boldsymbol{L}}$. Rule (5) may be still further simplified, if one assumes the model for the latent variables to be fixed, and not subject to optimization, i.e., $P(\boldsymbol{L} = \boldsymbol{l} \mid \theta_{\boldsymbol{L}}) = P(\boldsymbol{L} = \boldsymbol{l} \mid \theta^*_{\boldsymbol{L}})$ for fixed parameters $\theta^*_{\boldsymbol{L}}$, and the parameter optimization is only for $\theta_{\boldsymbol{X}|\boldsymbol{L}}$. If, furthermore, $P(\boldsymbol{L} = \boldsymbol{l} \mid \theta^*_{\boldsymbol{L}})$ is assumed uniform, then (5) reduces to

$$\boldsymbol{l}^* := \underset{\boldsymbol{l}}{arg\,max} \underset{\theta_{\boldsymbol{X}|\boldsymbol{L}}}{max} P(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{L} = \boldsymbol{l}, \theta_{\boldsymbol{X}|\boldsymbol{L}}). \tag{6}$$

Whether it is justified to assume a fixed distribution $P(\boldsymbol{L} \mid \theta^*_{\boldsymbol{L}})$ can depend on two considerations: first, assuming that (1) actually represents the generative process for the data, one might have sufficient background knowledge to identify the distribution of $\boldsymbol{L}$ a-priori. $\boldsymbol{L}$ being an unobserved variable, whose existence is essentially hypothesized, and for which it is typically even unclear how many

states it has, this is a rather unlikely case in practice, however. Second, clustering being an exploratory data-analysis tool, one may also consider what settings of $P(\boldsymbol{L} \mid \theta^*_{\boldsymbol{L}})$ may lead via (5) to interesting insights into the data, regardless of whether the underlying probabilistic model is accurate as a generative model.

For example, in the single clustering case, when the data is generated by a mixture model where one mixture component has a much higher prior probability than the others, then clustering via (3) can easily lead to only obtaining a single cluster. If, on the other hand, one eliminates the influence of the prior distribution by assuming (incorrectly) a uniform distribution over the mixture components, then clustering via (6) can reveal the mixture structure of the data.

3 The Factorial Logistic Model

In the factor analysis model, both $\boldsymbol{X}$ and $\boldsymbol{L}$ are numerical, the rows in $\boldsymbol{X}$ and $\boldsymbol{L}$ are iid, and the model (1) is given by distribution

$$P(\boldsymbol{L}_i) \sim N(\boldsymbol{0}, \boldsymbol{\Sigma_L})$$
$$P(\boldsymbol{X}_i \mid \boldsymbol{L}_i) \sim N(\boldsymbol{W}\boldsymbol{L}_i + \boldsymbol{\mu}, \boldsymbol{\Sigma_X})$$

where $\boldsymbol{\Sigma_L}$ is an arbitrary covariance matrix, $\boldsymbol{W}$ is a $k \times m$ matrix, $\boldsymbol{\mu}$ a m-dimensional mean vector, and $\boldsymbol{\Sigma_X}$ a diagonal covariance matrix. Thus, data is assumed to be generated by sampling from a lower (k) dimensional Gaussian distribution, linearly mapped into the higher (m) dimensional space, and independent Gaussian noise added to each coordinate.

The logistic regression model for the distribution of a binary variable X conditional on numeric latent variables $\boldsymbol{L}$ is given by

$$\log P(X = 1 \mid \boldsymbol{L})/P(X = 0 \mid \boldsymbol{L}) = w_0 + \boldsymbol{w}\boldsymbol{L}, \tag{7}$$

where $\boldsymbol{w} = (w_1, \ldots, w_k)$ is a k-vector of real weights. We write $X \sim LR(w_0, \boldsymbol{w}\boldsymbol{L})$ if X follows (7). This model also applies when the latent variables $\boldsymbol{L}$ are *ordinal*, i.e. each L_j codes by an integer $\{0, \ldots, r_j - 1\}$ one of r_j different, ordered categories. To accommodate *nominal* predictor variables (i.e., unordered categorical variables) in the logistic regression model, one encodes a nominal variable L_j with r states by binary indicator variables $L_{j,1}, \ldots, L_{j,r}$, i.e. $L_{j,h} = 1$, $L_{j,h'} = 0$ ($h' \neq h$) means that L_j is in its hth state.

We will consider both ordinal and nominal latent variables for clustering. An ordinal latent variable defines an ordered clustering, i.e. the cluster indices define an ordering of the clusters. Whether such an ordering is meaningful and interpretable is application dependent. For biogeographical data ordinal latent variables and ordered clusterings are often natural, since data patterns are often determined by underlying continuous variables. We will, thus, assume that $\boldsymbol{L}$ is a vector of m latent variables that define c different clusterings. Furthermore, we assume that one of the following two cases applies: (1) all L_j in $\boldsymbol{L}$ are ordinal; in this case $c = m$, and the jth clustering consists of r_j distinct cluster. (2) $\boldsymbol{L}$ is an encoding by binary indicator variables of c distinct nominal variables

with $r_1, \ldots, r_c$ distinct states, respectively. In this case $m = \sum_{i=1}^{c} r_i$. We refer to model (1) as the $(r_1 o, \ldots, r_k o)$ model, and (2) as the $(r_1 n, \ldots, r_c n)$ model. One could also consider models combining ordinal and nominal latent variables, but we will here focus on "pure" models.

As in the factor analysis model, we assume that $P(\boldsymbol{X} \mid \boldsymbol{L}) \sim \prod_{i=1}^{n} P(\boldsymbol{X}_i \mid \boldsymbol{L}_i) \sim \prod_{i=1}^{n} \prod_{j=1}^{k} P(\boldsymbol{X}_{i,j} \mid \boldsymbol{L}_i)$. Assuming that each $X_{i,j}$ follows a logistic regression model (7) with parameters $w_{j,0}, \boldsymbol{w}_j$, one obtains the model for the ith data item:

$$P(\boldsymbol{X}_i \mid \boldsymbol{L}_i) \sim \prod_{j=1}^{k} LR(w_{j,0}, \boldsymbol{w}_j \boldsymbol{L}_i). \tag{8}$$

This conditional model for $\boldsymbol{X}$ may be combined with various models for $P(\boldsymbol{L})$, with or without an iid assumption for the rows of $\boldsymbol{L}$. We refer to multiple clustering based on (8) as *factorial logistic (FL)* clustering.

4 Learning

We apply the simple learning rule (6) for clustering with the logistic regression model. Thus, we assume that $\boldsymbol{L}$ is uniformly distributed, which implies, in particular, independence over rows: $P(\boldsymbol{L}) \sim \prod_i P(\boldsymbol{L}_i)$. In case of $\boldsymbol{L}$ encoding nominal variables, the uniform distribution, of course, is conditional on "legal" states of $\boldsymbol{L}$, i.e. at most one indicator variable for any particular nominal variable being equal to 1.

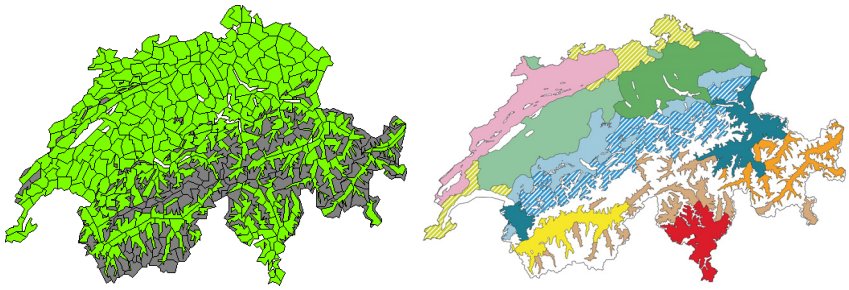
For the optimization of (6) we then use the obvious iterative procedure, where after a random initialization $\boldsymbol{L} := \boldsymbol{l}_0$ two steps are alternated:

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

Step **i** is performed in our implementation using the SPSS method of fitting logistic regression models, which supports both ordinal and nominal predictor variables. Due to the factorization (8), the optimization reduces to k independent optimizations for the parameters $(w_{j,0}, \boldsymbol{w}_j)$ $(j = 1, \ldots, k)$. It is thus linear in k. It also is linear in n, since the likelihood only depends on the counts $|\{i \mid \boldsymbol{X}_{i,j} = 1, \boldsymbol{L}_i = \hat{\boldsymbol{l}}\}|$ for fixed configurations $\hat{\boldsymbol{l}}$ of the latent variables.

For step **ii** we have $P(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{L} = \boldsymbol{l}, \theta_{\boldsymbol{X}|\boldsymbol{L}_t}) = \prod_i P(\boldsymbol{X}_i = \boldsymbol{x}_i \mid \boldsymbol{L}_i = \boldsymbol{l}_i, \theta_{\boldsymbol{X}|\boldsymbol{L}_t})$, so that the problem decomposes into n distinct optimizations for the $\boldsymbol{l}_i$. It can be naively performed by computing $P(\boldsymbol{X}_i = \boldsymbol{x}_i \mid \boldsymbol{L}_i = \boldsymbol{l}_i, \theta_{\boldsymbol{X}|\boldsymbol{L}_t})$ for each candidate $\boldsymbol{l}_i$, which gives a procedure that is still linear in n and k, but exponential in c.

Overall, we obtain a learning method that is linear in the number of data items and the observable attributes, and exponential in the number of clusterings.



(a) (b)

Fig. 1. Mapping areas with mountain - valley division (a), and previous segmentation of valley areas into floristic regions [12] (b)

5 Experiments

We apply FL-clustering to geobotanical data. In our experiments we use the source data for the "Swiss Web Flora" [3] [11]. The dataset contains information on the distribution of 2697 plant species in Switzerland, which has been divided into 565 mapping areas. We reduced slightly more detailed species abundance information in the original data to simple binary presence/absence data. We also in this process deleted plants with a very sparse and uncertain distribution. This left us with 2398 species in our data.[4] We view each plant species as an observable attribute, and the mapping areas as independent observations. Thus, $n = 565$ and $k = 2398$ in the notation of Section 2. Figure 1 shows the division of Switzerland into the mapping areas. Apart from the species occurrence data, only a single additional variable is recorded for each area: a binary variable that indicates whether the area is a mountain area (above timberline), or a valley area (below timberline). The value of this variable is shown if Figure 1 by a green color for valley, and grey color for mountain areas.

Conventional (single-) clusterings of the data lead to a segmentation of Switzerland into *floristic regions*. Figure 1 (b) shows a result obtained by agglomerative hierarchical clustering of the valley areas only [12] (thus, the white part of the figure does not correspond to a computed cluster; it comprises areas not included in the clustering).

5.1 Synthetic Data

In order to obtain an initial evaluation of the feasibility of our approach, we first conduct an experiment with synthetic data. For this we constructed two artificial segmentations of Switzerland based on the same mapping regions as in the real

[3] `www.wsl.ch/land/products/webflora/welcome-en.ehtml`

[4] The data is available at http://www.wsl.ch/info/mitarbeitende/wohlgemu/lehre_EN/

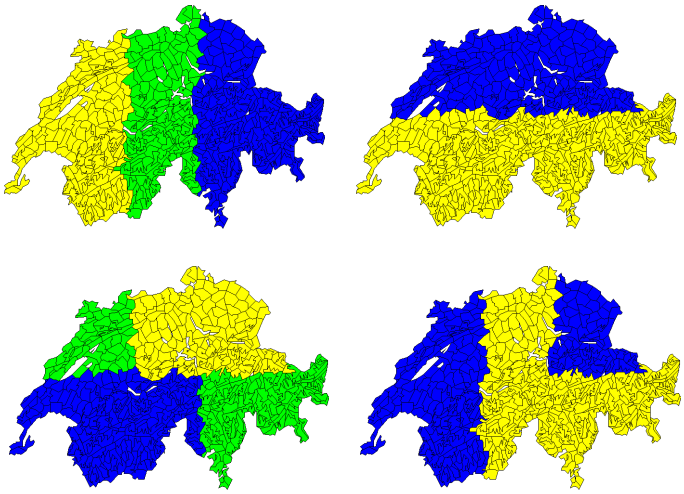


Fig. 2. Artificial segmentations (top); "Wrong" clusterings (bottom)

data. These segmentations are shown in Figure 2 (top), and henceforth referred to as "vertical" and "horizontal" segmentation, respectively. For each combination of a vertical and a horizontal segment, we defined a species distribution type by a nominal logistic regression model that expresses a preference of the species for the selected vertical and horizontal segment. The logistic regression weights were adjusted so as to obtain conditional probability distributions for the presence of a species of the following form (here showing the case of preference for the first segment in both segmentations):

	Vertical		
Horizontal	yellow	green	blue
blue	0.98	0.5	0.5
yellow	0.5	0.02	0.02

(9)

According to each distribution type we created 15 synthetic plant species, and randomly sampled an occurrence variable for the species at each of the mapping areas.

We then performed FL clustering based on the 90 synthetic species using the (3o,2o) model (it is not our ambition at this point to detect the "right" number of segmentations and segments per segmentation). In approximately 1 out of 3 random restarts the algorithm terminated with the correct segmentations of Figure 2, or solutions that differed from the correct one in cluster assignments for 2-3 regions. In the remaining restarts the algorithm terminated at local optima, a representative example of which is shown in Figure 2 (bottom). However, the (almost) correct solutions were identified by higher log-likelihood scores (between -19912 and -19783) than that of the wrong solutions (between -23474 and -21677).

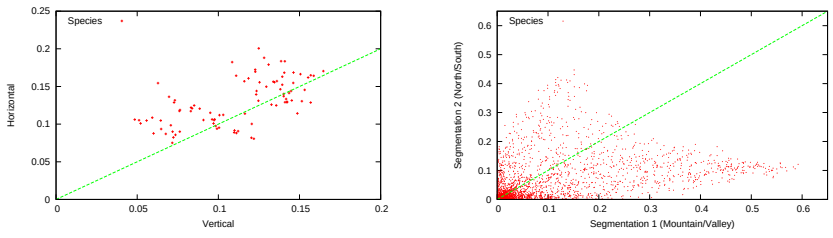
For comparison, we also performed an experiment where the logistic regression model for $P(\boldsymbol{X} \mid \boldsymbol{L})$ was replaced by a full multinomial model, i.e. for each species we fit a conditional probability table of the form (9) with 6 independent parameters. In this case, almost all restarts terminated with wrong solutions as in Figure 2, and, more importantly, the correct solutions could not be distinguished by a higher likelihood score: in the multinomial model, any pair of segmentations whose combination identifies the 6 different combinations of vertical and horizontal segments achieves the same, optimal, likelihood score.

We also use this synthetic data experiment to demonstrate that in FL-clustering there is not necessarily a correlation between clusterings and feature-subsets. Figure 3 (a) shows for each of the 90 synthetic species the mutual information between the species occurrence feature and the two clusterings of Figure 2 (top). The plot shows that there is no strong association of individual species features with one or the other of the two segmentations.

5.2 Real Data

We now perform experiments with the real data consisting of the actual 2398 species. Again, we do not try at this point to automatically detect an appropriate number of segmentations, or segments per segmentation. We run the learning algorithm with a few selected ordinal and nominal logistic models. In all cases we perform 20 runs of the algorithm with different random initializations of the latent variables $\boldsymbol{L}$. The results shown in the following are the segmentations that achieved the highest likelihood score (6) within the 20 restarts.

Figure 4 shows the result of clustering with the (3o,3o) and (3n,3n) logistic models. We use different colors to represents segments computed by nominal logistic models, and greyscale values for ordinal logistic models. The greyscale values then show the ordering of the segments according to their (ordinal) index. One first observes that both models have produced one segmentation in which the mountain areas are identified as one segment: there is an almost perfect correspondence between the mountain attribute illustrated in Figure 1, and the dark grey, respectively yellow, segments in the first segmentations of Figure 4



(a) (b)

Fig. 3. Mutual information: synthetic data (a), real data (b)

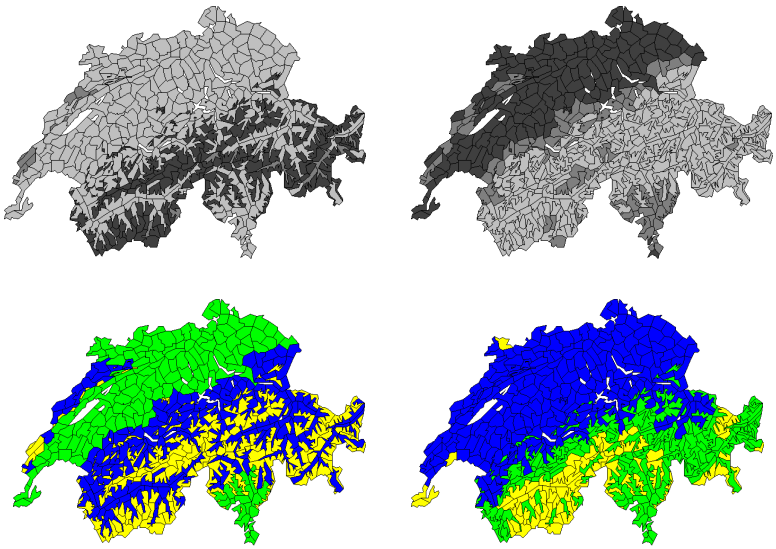


Fig. 4. Clustering result using (3o,3o) (top) and (3n,3n) (bottom) logistic models

(note that our method does not entail an ordering of the different segmentations; in particular, in Figure 4 we have just for convenience vertically aligned similar segmentations, and arbitrarily put the ones containing the mountain segment first).

Apart from the mountain/valley attribute our data does not contain "hidden class variables" that could be used for interpreting the segmentations, and therefore one has to look for additional, external data sources, and expert knowledge. As previously mentioned, we expect that the different segmentations to some extent correspond to ecological factors that determine plant growth. A difficulty we now encounter is that many such candidate factors (e.g., average annual temperature, average precipitation) are highly correlated with the mountain/valley division, and often show a secondary gradient in north-south direction. The second segmentations of Figure 4 are somewhat dominated by a north-south stratification, and also exhibit some of the patterns visible in Figure 1 (b). However, it seems impossible to identify these north-south segmentations with any particular ecological factor. Instead, it can only be taken as aggregating the north-south dependency of several factors. Moreover, whereas one clustering showing the mountain/valley division was quite consistently produced in the random restarts, there were larger variations observed in the north-south clustering.

The range of likelihood values obtained in 20 restarts was $-288 \cdot 10^3$ to $-278 \cdot 10^3$ for (3o,3o) clustering, and $-308 \cdot 10^3$ to $-296 \cdot 10^3$ for (3n,3n) clustering (since the former model fits more independent parameters, higher likelihood scores are to be expected).

Figure 3 shows the mutual information values for the plant features and the (3o,3o) segmentation of Figure 4. This plot shows a relatively strong cor-

relation of some species with the mountain/valley clustering, and a somewhat less pronounced primary correlation of some other species with the north/south segmentation. The large number of species with very small mutual information values for either segmentation is largely made up of species that occur in only a few areas.

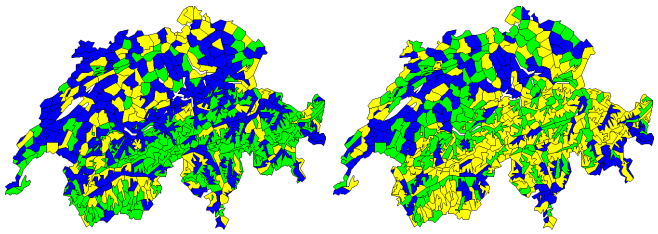


Fig. 5. Clustering result using multinomial $P(\boldsymbol{X} \mid \boldsymbol{L})$

In analogy to the experiment with synthetic data, we also perform with the real data an experiment with full multinomial species distribution models instead of the logistic ones. The result is shown in Figure 5. While the mountain/valley pattern is also partly visible in some of the segments, there is no single segment or segmentation corresponding to this division, and the overall segmentation result is clearly less useful than the one obtained with the logistic models.

The poor performance of the multinomial model may in part be due to this model's inability to isolate in its different clusterings several independent explanatory factors, as illustrated by the synthetic data experiments. In addition, the multinomial model suffered from severe problems of convergence to local optima: even though the global likelihood maximum of the multinomial model must be at least as high as the logistic optimum, the likelihood values found in 20 restarts were significantly lower for the multinomial than for the logistic models (range $-433 \cdot 10^3$ to $-405 \cdot 10^3$).

The average time consumption of a single run (restart) of (3o,3o) or (3n,3n) clustering was approximately 3 hours, with an average of approximately 15 iterations until convergence. This increased to approximately 6 hours for (3o,3o,3o) or (3n,3n,3n) clusterings. The time is consumed almost entirely in fitting in each iteration the 2398 logistic regression models for all the plants. For comparison, a single run with the multinomial model (taking approximately 8 iterations on average until convergence) takes only about 1 minute, since the multinomial model is easily fit by taking simple counts.

6 Discussion and Future Work

Our experiments have shown that using FL-clustering we can find multiple meaningful clusterings of categorical data. The objective in our approach is explana-

tory (identify underlying factors that determine the overall data patterns) rather than descriptive (provide the user with multiple views of the data).

For our purpose, it is clearly essential to use a conditional model $P(\boldsymbol{X} \mid \boldsymbol{L})$ of a restricted functional form, rather than an unconstrained multinomial model. Logistic regression models are a canonical choice, and can be seen as a categorical data analogue to the linear mappings between latent and observed dimensions in the factor analysis model.

A common objective in multiple clustering is that different clusterings are in some sense orthogonal or complementary. We are not yet able to say in which sense, or to what extent, FL-clustering satisfies such an objective. Empirically, a bias towards learning complementary clusterings was difficult to verify with our data, since most natural candidate segmentations based on hidden environmental variables would exhibit rather similar patterns (and not at all resemble the segmentations in Figure 2). Theoretically, one can note that a multi-clustering $\boldsymbol{L} = \boldsymbol{l}$ in which two clusterings are identical can not be a local maximum of the likelihood (6) (except for some degenerate, noise-free, data sets). FL-clustering, thus, is biased away from returning multiple identical clusterings. How this can be strengthened into a formal result linking likelihood gain and complementarity of different clusterings is a subject for future work.

In our experiments we have used data with a spatial structure on the data instances. Within this paper, we have used the spatial structure only for the visualization of the clustering (i.e., segmentation) results. The model can equally be used for other categorical data, and is especially suited for high-dimensional binary data (such as text document data).

On the other hand, our work was also specifically motivated by spatial data, and the relationship in this case of multiple clustering with factorial hidden Markov models [3] and factorial Markov random fields [7]. For spatial data one can impose a Markov random field structure on the latent variables $\boldsymbol{L}$, i.e., the assumption of a uniform distribution for $\boldsymbol{L}$ which we used to derive (6) is replaced, e.g., by the assumption that $P(\boldsymbol{L} = \boldsymbol{l} \mid \theta^*_{\boldsymbol{L}})$ is a Gibbs distribution with fixed parameters $\theta^*_{\boldsymbol{L}}$. Learning in such a setting proceeds in the same way as described in Section 4, only that $P(\boldsymbol{L} = \boldsymbol{l} \mid \theta^*_{\boldsymbol{L}})$ has to be added as a likelihood factor. The optimization in step ii will then usually not be possible precisely, and require an approximate solution. In this paper we did not employ a Markov random field model, since this would usually be used to enforce some smoothness and contiguity properties of the learned segments, which, for our data, seems unwarranted (considering, e.g., the rugged outline of the mountain areas).

In this paper we focused on the core of a probabilistic (multi-) clustering model, i.e., the joint distribution of latent and observable variables. In this model, the number of clusterings, and the number of clusters in each clustering is fixed. We remark, however, that either model selection techniques like BIC or MDL scoring, or a nonparametric Bayesian 'wrapper' around the core model can be used to also learn the model structure.

7 Conclusion

We proposed a latent variable model for multiple clustering of categorical data based on a logistic regression model for the conditional distribution of the observed features. We believe that in analogy to successful techniques for dimensionality reduction, a restricted distributional form for the noisy transformation between the latent and the observed features can be instrumental for revealing relevant patterns in the latent feature space.

For clustering based on a latent variable model we have suggested a simple optimization of the conditional likelihood of the data given the latent variables, with a fixed marginal distribution for the latent variables. This leads to a learning procedure that is linear in the number of observed features, and enables us to experiment with high-dimensional biogeographical data. Our preliminary results from these experiments demonstrate the ability of the method to discover clusterings that represent meaningful explanatory factors for the data. However, further work is needed to consolidate the results returned for this data, and to investigate their potential biological meaning.

References

1. A. Agresti. *Categorical Data Analysis*. Wiley, 2002.
2. T. Chen, N. Zhang, T. Liu, K. M. Poon, and Y. Wang. Model-based multidimensional clustering of categorical data. *Artificial Intelligence*, 2011. To appear.
3. Z. Ghahramani and M. Jordan. Factorial hidden markov models. *Machine Learning*, 29:245–273, 1997.
4. Y. Guan, J. G. Dy, D. Niu, and Z. Ghahramani. Variational inference for nonparametric multiple clustering. In *KDD10 Workshop on Discovering, Summarizing and Using Multiple Clusterings*, 2010.
5. J. A. Hagenaars. *Loglinear Models with Latent Variables*. Number 94 in Quantitative Applications in the Social Sciences. Sage Publications, 1993.
6. P. Jain, R. Meka, and I. Dhillon. Simultaneous unsupervised learning of disparate clusterings. In *SIAM Int. Conf. on Data Mining*, pages 858–869, 2008.
7. J. Kim and R. Zabih. Factorial markov random fields. In *Proc. of ECCV 2002*, number 2352 in LNCS, pages 321–334, 2002.
8. D. Niu, J. G. Dy, and M. I. Jordan. Multiple non-redundant spectral clustering views. In *Proc. of the 27th Int. Conf. on Machine Learning (ICML-10)*, 2010.
9. L. Orloci. An agglomerative method for classification of plant communities. *The Journal of Ecology*, 55(1):193–206, 1967.
10. H. Wang, H. Shan, and A. Banerjee. Bayesian cluster ensembles. *Statistical Analysis and Data Mining*, 4(1):54–70, 2011.
11. T. Wohlgemuth. Biogeographical regionalization of switzerland based on floristic data: How many species are needed? *Biodiversity Letters*, 3(6):180–191, 1996.
12. T. Wohlgemuth. Ein floristischer ansatz zur biogeographischen gliederung der schweiz. *Botanica Helvetica*, 106:227–260, 1996.
13. N. L. Zhang. Hierarchical latent class models for cluster analysis. *Journal of Machine Learning Research*, 5:697–723, 2004.
14. N. L. Zhang, Y. Wang, and T. Chen. Discovery of latent structures: Experience with the COIL challenge 2000 data set. *Journal of Systems Science and Complexity*, 21:172–183, 2008.