Clustering with Multidimensional Mixture Models: Analysis on World Development Indicators

Leonard K.M. Poon

Department of Mathematics and Information Technology
The Education University of Hong Kong, Hong Kong, China
kmpoon@eduhk.hk

Abstract. Clustering is one of the core problems in machine learning. Many clustering algorithms aim to partition data along a single dimension. This approach may become inappropriate when data has higher dimension and is multifaceted. This paper introduces a class of mixture models with multiple dimensions called pouch latent tree models. We use them to perform cluster analysis on a data set consisting of 75 development indicators for 133 countries. We further propose a method that guides the selection of clustering variables due to the existence of multiple latent variables. The analysis results demonstrate that some interesting clusterings of countries can be obtained from mixture models with multiple dimensions but not those with single dimensions.

Keywords: Multidimensional clustering, pouch latent tree models, mixture models, world development indicators, clustering variables selection

1 Introduction

Clustering [8] is a core problem in machine learning. Many clustering algorithms aim to partition data along a single dimension [2,16]. To handle data with higher dimensions, *feature selection* and *subspace clustering* approaches are often adopted. The former approach selects a subset of relevant features on data in which a clustering can be found [5,14]. The latter approach considers dense regions as clusters and tries to identify all dense subspaces (with reduced dimension) for partitioning the data [11,9]. Both approaches partition data along only a single dimension, in the sense that each data point belongs to at most one partition.

The above approach becomes inappropriate when data is multifaceted and multiple meaningful clusterings can be obtained. Suppose we want to cluster countries into different groups. We may partition them based on their land sizes and populations, systems of government, income levels, levels of freedom, etc. To obtain clusterings on different aspects, one may perform cluster analysis on data sets with different attributes. However, sometimes one may not know which aspect of data will yield to meaningful clusterings and sometimes the

attributes in different aspects are interdependent. Hence it is more appropriate to perform cluster analysis that produces clustering along multiple dimensions simultaneously.

In our previous work [12,13], we propose a class of probabilistic graphical models called *pouch latent tree models* (PLTMs) for multidimensional clustering. The models are similar to Gaussian mixture models. However, they can contain multiple latent variables and hence can produce multiple clusterings.

In this paper, we present the results of a cluster analysis on countries based on the world development indicators. The indicators are statistics provided by the World Bank about the development and human lives for different countries. The data set we used includes 75 indicators relevant to risk management in the context of development for 133 countries. The data obviously represent different aspects of countries and our study aims to show the usefulness of multidimensional clustering. Due to the existence of multiple latent variables, we propose a method that guides the selection of clustering variables. Before we show the results, we review model-based clustering and introduce PLTMs.

2 Model-Based Clustering

Gaussian mixture models (GMMs) are commonly used in model-based clustering for numeric data [10]. GMMs assume that the population is be made up from a finite number of clusters. Suppose a variable Y is used to indicate this cluster, and variables $\boldsymbol{X}$ represent the attributes in the data. The variable Y is referred to as a *latent* (or unobserved) variable, and the variables $\boldsymbol{X}$ as *manifest* (or observed) variables. The manifest variables $\boldsymbol{X}$ is assumed to follow a mixture distribution

$$P(\boldsymbol{x}) = \sum_y P(y)P(\boldsymbol{x}|y),$$

where $P(\boldsymbol{x}|y)$ is known as the *component distribution* and in GMMs is assumed to be a multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)$, with mean vector $\boldsymbol{\mu}_y$ and covariance matrix $\boldsymbol{\Sigma}_y$ conditional on the value of Y.

3 Pouch Latent Tree Models

There is only one single latent variable in GMMs and hence can produce one clustering. To allow having multiple clusterings, we previously propose *pouch latent tree models* (PLTMs) [12,13]. A PLTM is a tree-structured probabilistic graphical model, where each internal node represents a latent variable, and each leaf node represents a set of manifest variables. All the latent variables are discrete, while all the manifest variables are continuous. A leaf node, also called *pouch node*, may contain a single manifest variable or several of them. An example is shown in Figure 2.

In PLTMs, the dependency of a discrete latent variable Y on its parent $\Pi(Y)$ is characterized by a conditional discrete distribution $P(y|\pi(y))$. Let $\boldsymbol{W}$ be the

variables of a pouch node with a parent node $Y = \Pi(\boldsymbol{W})$. We assume that, given a value y of Y, $\boldsymbol{W}$ follows the conditional Gaussian distribution $P(\boldsymbol{w}|y) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)$ with mean vector $\boldsymbol{\mu}_y$ and covariance matrix $\boldsymbol{\Sigma}_y$. Denote the sets of pouch nodes and latent nodes by $\mathcal{W}$ and $\mathcal{Y}$, respsectively. The whole model defines a joint distribution over all observed variables $\boldsymbol{X}$ and latent variables $\boldsymbol{Y}$

$$P(\boldsymbol{x}, \boldsymbol{y}) = \prod_{\boldsymbol{W} \in \mathcal{W}} P(\boldsymbol{w}|\pi(\boldsymbol{W})) \prod_{Y \in \mathcal{Y}} P(y|\pi(Y)) \tag{1}$$

Given a model structure m, the parameters can be estimated by the EM algorithm [4]. To learn the model structure, we use a greedy search that aims to maximize the BIC score [15]: $BIC(m|\mathcal{D}) = \log P(\mathcal{D}|m, \boldsymbol{\theta}^*) - \frac{d(m)}{2} \log N$, where $\mathcal{D}$ is the data set, $\boldsymbol{\theta}^*$ are the parameters estimated by the EM algorithm, $d(m)$ is the number of parameters in the model, and N is the data size. Interested readers are referred to [13] for details of the learning algorithm.

After we have learned a PLTM, we can partition data using each of the latent variables Y. Each data point $\boldsymbol{d}$ can be classified to one of the states of Y by computing the probability $P(y|\boldsymbol{d})$ based on the joint distribution (Eq. 1).

4 Analysis on World Development Indicators

Here we present the results of a cluster analysis on world development indicators using PLTM aiming to show its effectiveness for multidimensional clustering.

4.1 Data Set

In our experiment, we used the data set called World Development Report (WDR) 2014 provided by the World Bank.[1] The data set includes 75 indicators relevant to risk management in the context of development for 133 countries. The indicators are grouped into seven categories, namely key indicators of development, selected risk indicators, selected indicators related to risk management at the household level, enterprise sector level, financial sector level, macroeconomy level, respectively, and natural disasters and climate change indicators. For some indicators, the data set includes multiple values at different time periods. Some statistics are not available for some countries. In summary, the data set has 93 attributes and 133 samples with 15% of missing data.

4.2 Empirical Comparison

We included three methods based on GMMs for comparison in our experiment. The first method is `mclust` [6], which is an implementation of the parsimonious Gaussian mixture models [1]. The second method is the GS method [7]. It models the data using a collection of independent GMMs, each on a distinct subset

[1] `http://data.worldbank.org/data-catalog/world-development-report-2014`

Table 1: Comparison of methods on the World Development Report 2014 data set. The table shows the numbers of latent variables (#LV), numbers of parameters (dim), and BIC scores of the models obtained. It also shows the NMI and number of clusters (#C) of the clustering closest to the given classifcation.

Method	**#LV**	**dim**	**BIC**	**NMI**	**#C**
`mclust`	1	1109	-51965	0.41	6
GS model	40	2125	(-37422)	0.52	4
PLTM	28	1043	**-46706**	**0.62**	4

of attributes. The third method is PLTMs. The first method produces unidimensional clusterings, whereas the other two method produce multidimensional clusterings. Since `mclust` and the GS method cannot handle missing data, we impute the missing data using the R package `mice` [18] before training them.

Table 1 shows the results obtained by the three methods. The `mclust` model contains one latent variable, whereas the GS model and PTLM contains 40 and 28 latent variables, respectively. In terms of model complexity, PLTM has the lowest number of parameters. This happens even though PLTM has more latent variables than `mclust` model and it has connections between latent variables unlike GS model.

To evaluate the model quality, we compute the BIC score of the models. We use the completed data as the test data set for consistency. The parameters of PLTM were re-estimated on the complete data after learning the model structure on the incomplete data. This should not be unfair to other methods since `mclust` and GS method used the same test data set for training while PTLM optimized its the structure using a data set different from the test data set.

The BIC scores in Table 1 show that PLTM has a higher quality than the `mclust` model. The BIC of GS model is even higher. However, this was possibly due to spurious clusters [10]. Those clusters have component distributions with very small variance and hence can attain very high likelihood on data. This can be seen from the fact that although the smallest variance in the data is 0.32, the smallest scale of variance of the component distribution in the GS model is much smaller at 2.8×10^{-16}.

The WDR includes a classification of countries based on four income levels, namely low, lower middle, upper middle, and high. The classification is used as a class variable for evaluating the clusterings given by the models. To evaluate the similarity between the partition given by a latent variable Y and the class variable C, we use the normalized mutual information $NMI(C;Y)$ [17]: $NMI(C;Y) = \frac{MI(C;Y)}{\sqrt{H(C)H(Y)}}$, where $MI(C;Y)$ is the mutual information between C and Y and $H(V)$ is the entropy of a variable V [3].

Table 1 shows the NMI attained by the three methods. Among the multiple clusterings given by GS method and PLTM, only the ones with the highest NMI are reported. The result shows that PLTM performed best in recovering the classification based on income levels.

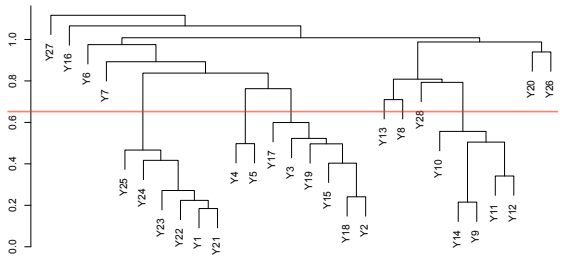


Fig. 1: Hierarchical clustering on latent variables based on the subset of attributes on which they partition the data. We cut off the tree at the red horizontal line.

4.3 Selection of Clustering Variables

Figure 2 shows the PLTM obtained from the WDR data set. The latent nodes are represented by the oval nodes. Each of them produces a partition of data. They partition the data along different facets of data as can be seen from the different attributes connected to them. For example, the latent variable Y_1 is connected to three attributes, namely gross national income per capita (`gni_pc`), PPP gross national income per capita (`ppp_gni_pc`), and worldwide government indicator (`worldwide_government_indicator`). The three observed variables are put inside a pouch node meaning that they are not independent conditionally on Y_1. The partition given by Y_1 happens to be the one closest to the classification based on income level given by WDR.

The PLTM obtained contains 28 latent variables and thus provides 28 ways to partition data. One issue arising from multidimensional clustering is how to select clustering among those available. In practice, there may not be any reference clustering for selection as we do in the previous subsection. Therefore, we propose a method for selecting clustering variables below.

Due to the model structure, each latent variable partitions data based on a different subset of attributes. To quantify this, we compute the NMI between a latent variable and each of the attributes.[2] After obtaining a vector of NMI values for each latent variable, we normalize them such that each one has unit magnitude. We then cluster the variables using hierarchical clustering.

The clustering of variables can help us look for a clustering of interest. We illustrate the idea using the PLTM obtained as an example. Figure 1 shows the hierarchical clustering result. We see that some latent variables (e.g. Y_1, Y_{21}–Y_{25}) are closer to each other, while some latent variables (e.g. Y_6, Y_7, Y_{16}, Y_{27}) are further away from the others. We cut off the tree at the red horizontal line in Figure 1. There are four groups of latent variable below the line and they are indicated by different colors in Figure 2. The grouping of variables is consistent with the model structure. It shows which latent variables partition data along a similar subset of attributes. On the other hand, the ungrouped latent variables partition data along a relatively distinct subset of attributes.

[2] The NMI can be computed using the empirical distribution after discretizing the continuous attributes.

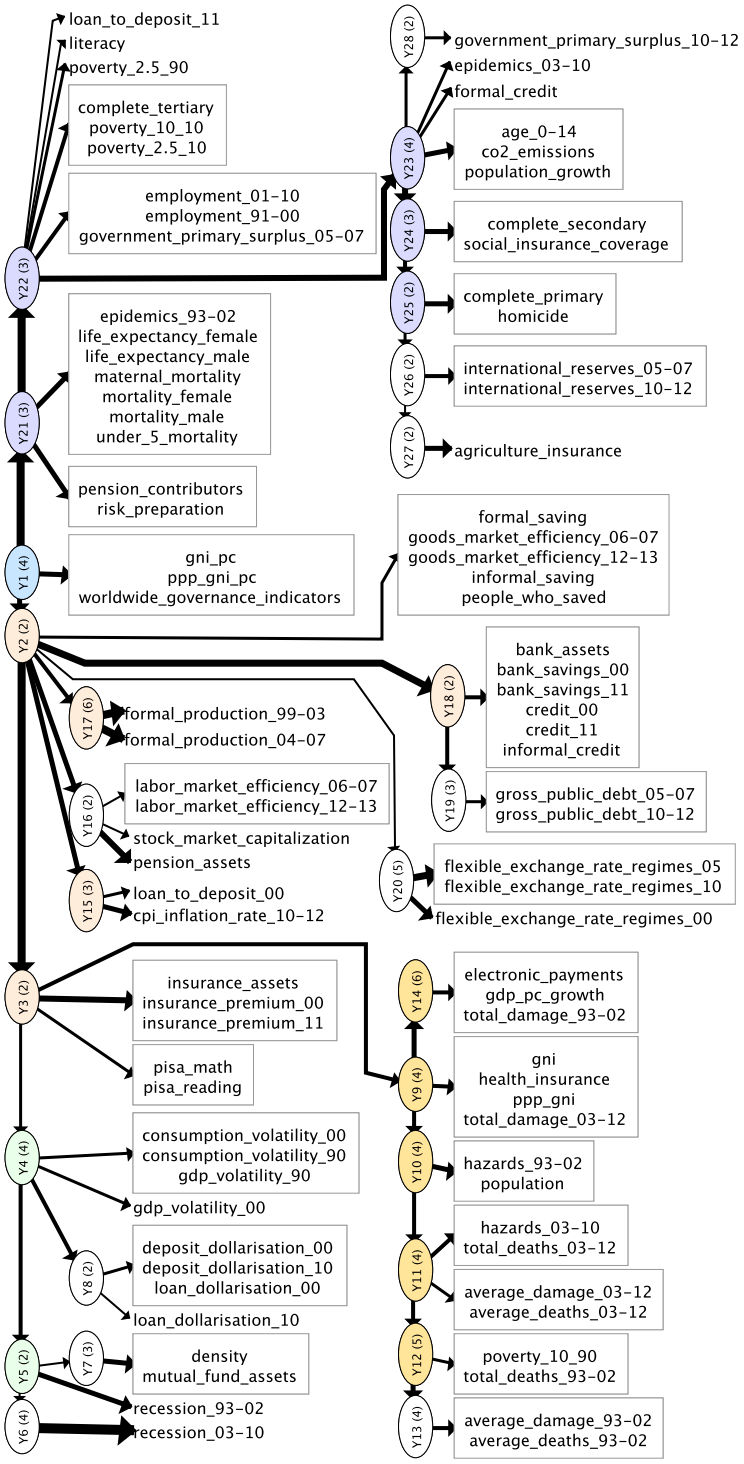
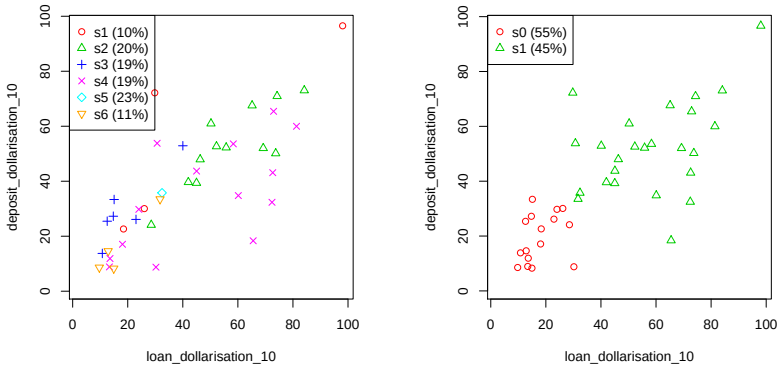


Fig. 2: Pouch latent tree models obtained from the World Development Report 2014 data set. Latent nodes are represented by oval nodes. Each of them produces a partition of data. Pouch nodes with multiple observed variables are shown as text in rectangular border, whereas those with single observed variables are shown as text without borders. The width of an edge indicates the strength of probabilistic dependency in terms of NMI between two nodes. The latent variable Y_1 (blue) yields a clustering of countries closest to the income level classification given by the report. The colors of latent nodes indicate their grouping.



(a) `mclust` (b) PLTM (Y_8)

Fig. 3: Clusterings along two dollarization attributes obtained from unidimensional clustering method `mclust` and multidimensional clustering method PLTM. PLTM partitions data neatly along this facet while `mclust` does not.

We now take a look at the four groups of latent variables. The first group, Y_1 and Y_{21}–Y_{25}, depends on attributes related to general development level, such as GNI, life expectancy, mortality and poverty. The second group, Y_2, Y_3, Y_{15}, Y_{17} and Y_{18}, depends mainly on attributes related to the financial sector. The third group, Y_9–Y_{12} and Y_{14}, is mainly related to deaths and damages from natural disasters. The fourth group of variables Y_4 and Y_5 partition data based on number of recessions and GDP and household volatility.

The ungrouped variables also partition data based on meaningful subsets of attributes. For example, Y_8 partitions data based on dollarization, Y_{19} partitions based on flexibility of exchange rate, Y_{20} partitions based on gross public debt, and Y_{26} partitions based on international reserves.

The grouping of latent variables allows us to have an overview on the aspects of attributes from which we obtain a clustering. We can look for the attributes in which we are interested and select the latent variable connected to it to partition the data. As an example, suppose we are interested in dollarization. We can use the latent variable Y_8 in PLTM for clustering. Figure 3(b) shows the different countries projected on two dollarization attributes. The countries are classified into two groups neatly by Y_8. For comparison, we show the clustering obtained by `mclust` in Figure 3(a). The comparison shows that the multidimensional clustering method PLTM can obtain some meaningful clusterings that cannot be obtained by the unidimensional clustering method `mclust`.

5 Conclusion

In this paper, we introduce a class of multidimensional mixture models called pouch latent tree models and use them for cluster analysis on world development indicators. PLTM is shown to recover the given classification better. It is also

shown to produce meaningful clusterings that another unidimensional method cannot. We illustrate how to use hierarchical clustering on latent variables to guide the selection of clustering variables. The source code of algorithms for PLTMs can be found online: `https://github.com/kmpoon/pltm-east`.

References

1. Banfield, J.D., Raftery, A.E.: Model-based Gaussian and non-Gaussian clustering. Biometrics 49(3), 803–821 (1993)
2. Bouveyrona, C., Brunet-Saumard, C.: Model-based clustering of high-dimensional data: A review. Computational Statistics and Data Analysis 71, 52–78 (2014)
3. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley (2006)
4. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological) 39(1), 1–38 (1977)
5. Dy, J.G., Brodley, C.E.: Feature selection for unsupervised learning. Journal of Machine Learning Research 5, 845–889 (2004)
6. Fraley, C., Raftery, A.E., Murphy, T.B., Scrucca, L.: mclust version 4 for R: Normal mixture modeling for model-based clustering, classification, and density estimation. Tech. rep., Department of Statistics, University of Washington (2012)
7. Galimberti, G., Soffritti, G.: Model-based methods to identify multiple cluster structures in a data set. Computational Statistics and Data Analysis 52, 520–536 (2007)
8. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: A review. ACM Computing Surveys 31(3), 264–323 (1999)
9. Kriegel, H.P., Kröger, P., Zimek, A.: Clustering high dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. ACM Transactions on Knowledge Discovery from Data 3(1), 1–58 (2009)
10. McLachlan, G.J., Peel, D.: Finite Mixture Models. Wiley, New York (2000)
11. Parsons, L., Haque, E., Liu, H.: Subspace clustering for high dimensional data: A review. ACM SIGKDD Explorations Newsletter 6(1), 90–105 (2004)
12. Poon, L.K.M., Zhang, N.L., Chen, T., Wang, Y.: Variable selection in model-based clustering: To do or to facilitate. In: Proceedings of the 27th International Conference on Machine Learning. pp. 887–894 (2010)
13. Poon, L.K.M., Zhang, N.L., Liu, T., Liu, A.H.: Model-based clustering of high-dimensional data: Variable selection versus facet determination. International Journal of Approximate Reasoning 54(1), 196–215 (2013)
14. Raftery, A.E., Dean, N.: Variable selection for model-based clustering. Journal of American Statistical Association 101(473), 168–178 (2006)
15. Schwarz, G.: Estimating the dimension of a model. The Annals of Statistics 6(2), 461–464 (1978)
16. Shirkhorshidi, A.S., Aghabozorgi, S., Wah, T.Y., Herawan, T.: Big data clustering: A review. In: Computational Science and Its Applications – ICCSA 2014, pp. 707–720. Springer (2014)
17. Strehl, A., Ghosh, J.: Cluster ensembles — a knowledge reuse framework for combining multiple partitions. Journal of Machine Learning Research 3, 583–617 (2002)
18. van Buuren, S., Groothuis-Oudshoorn, K.: mice: Multivariate imputation by chained equations in R. Journal of Statistical Software 45(3), 1–67 (2011)