SOME ASYMPTOTIC PROPERTIES OF MODEL SELECTION CRITERIA IN THE LATENT BLOCK MODEL

Christine Keribin [1,2]

[1] Laboratoire de Mathématiques d'Orsay, Univ. Paris-Sud, CNRS, Université Paris-Saclay, 91405 Orsay, France, (e-mail: christine.keribin@math.u-psud.fr)

[2] INRIA Saclay - Île de France, Équipe CELESTE

ABSTRACT: Co-clustering designs in a same exercise a simultaneous clustering of the rows and the columns of a data array. The Latent Block Model (LBM) is a probabilistic model for co-clustering, based on a generalized mixture model. LBM parameter estimation is a difficult problem as the likelihood is numerically untractable. However, deterministic or stochastic strategies have been designed and the consistency and asymptotic normality have been recently solved when the number of blocks is known. We address model selection for LBM and propose here a class of penalized log-likelihood criteria that are consistent to select the true number of blocks for LBM.

KEYWORDS: Latent block model, co-clustering, model selection, BIC, ICL

1 Introduction

Clustering is an essential unsupervised tool to discover hidden structure from data by detecting groups of observations that are similar within a group and dissimilar from one group to another one. The challenge of modern data is to learn from observations $x_i \in R^d$ with a large number n of units observed on a large number d of variables, and the question is not only to cluster the observations, but also to cluster simultaneously the observations and the variables, leading to a tremendous parsimonious data representation.

This is called co-clustering and has many applications in many fields such as recommendation systems (to cluster simultaneously customers and goods), text mining (to co-cluster words and documents), genomics (to co-cluster genes and experimental conditions) for example. As for clustering, there are many ways to perform co-clustering, and we will focus here on the latent block model (LBM). We present the model and its asymptotical properties. In particular, we shall analyze the log-likelihood ratio under model order misspecifications, and derive a class of penalized log-likelihood criteria asymptotically consistent, results that are new for LBM.

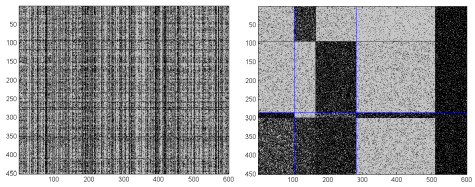

Figure 1. $n \times d = 450 \times 600$ *observations (left) and their reorganization according to the underlying structure in* 4×5 *blocks (right)*

2 The latent block model

LBM is a probabilistic model for co-clustering. Upon a data matrix $X = (x_{ij})$ of n rows and d columns, it defines a block clustering latent structure as the Cartesian product of a row partition $\mathbf{z}$ by a column partition $\mathbf{w}$ with three main assumptions:

- row assignments (or labels) $\mathbf{z}_i$, $i = 1,\ldots,n$, are independent from column assignments (or labels) $\mathbf{w}_j$, $j = 1,\ldots,d$: $p(\mathbf{z},\mathbf{w}) = p(\mathbf{z})p(\mathbf{w})$;
- row labels are independent, with a common multinomial distribution: $\mathbf{z}_i \sim \mathcal{M}(1,\pi = (\pi_1,\ldots,\pi_g))$; in the same way, column labels are i.i.d. multinomial variables: $\mathbf{w}_j \sim \mathcal{M}(1,\rho = (\rho_1,\ldots,\rho_m))$.
- conditionally to row and column assignments $(\mathbf{z}_1,\ldots,\mathbf{z}_n) \times (\mathbf{w}_1,\ldots,\mathbf{w}_d)$, the observed data X_{ij} are independent, and their (conditional) distribution $\varphi(.,\alpha)$ belongs to the same parametric family, which parameter α only depends on the given block:

$$X_{ij}|\{z_{ik}w_{j\ell} = 1\} \sim \varphi(.,\alpha_{k\ell})$$

  where z_{ik} is the indicator membership variable of whether row i belongs to row-group k and $w_{j\ell}$ is the indicator variable of whether column j belongs to column-group ℓ.

Hence, the complete parameter set is $\theta = (\pi,\rho,\alpha)$, with $\alpha = (\alpha_{11},\ldots,\alpha_{gm})$. With these assumptions, the likelihood of the *complete data* is

$$p(\mathbf{x},\mathbf{z},\mathbf{w};\theta) = p(\mathbf{z};\theta)p(\mathbf{w},\theta)p(\mathbf{x}|\mathbf{z},\mathbf{w};\theta) = \prod_{i,k}\pi_k^{z_{ik}}\prod_{j,\ell}\rho_\ell^{w_{j\ell}}\prod_{i,j,k,\ell}\varphi(x_{ij};\alpha_{k\ell})^{z_{ik}w_{j\ell}}$$

The labels are usually unobserved, and the *observed likelihood* is obtained by marginalization over all the label configurations:

$$p(\mathbf{x};\theta) = \sum_{\mathbf{z}\in\mathcal{Z},\mathbf{w}\in\mathcal{W}} \left(\prod_{i,k} \pi_k^{z_{ik}} \prod_{j,\ell} \rho_\ell^{w_{j\ell}} \prod_{i,j,k,\ell} \varphi(x_{ij};\alpha_{k\ell})^{z_{ik}w_{j\ell}} \right)$$

LBM deals with matrix of homogeneous data, such as binary (Govaert & Nadif, 2008), Gaussian (Lomet, 2012), categorical (Keribin *et al.*, 2015) or count (Govaert & Nadif, 2010) data. It involves a double missing data structure $\mathbf{z}$ for rows and $\mathbf{w}$ for columns, and the observed likelihood can not be factorized as a product of the mixing density as for simple mixture models. This implies that the likelihood is rapidly not tractable numerically even for few observations and few blocks, as the marginalization involves $k^n \times d^m$ terms. The estimation can however be performed either with numerical approximations (such as variational methods) or with Bayesian approaches (VBayes algorithm or Gibbs sampling).

3 Asymptotic properties

The double missing structure also leads to a very challenging and interesting study to state the asymptotic behavior of the maximum likelihood (MLE) and variational (VE) estimators. This question was first studied on the Stochastic Block Model (SBM) which is a LBM with the same statistical units in rows and columns, used to model graph adjacency matrices. In this case, there is only one set of latent variables $\mathbf{z}$. Celisse *et al.*, 2012 first proved that under the true parameter value, the conditional distribution of the assignments of a binary SBM converges to a Dirac of the real assignments. Assuming the existence of an estimator of α converging at rate at least n^{-1}, they obtained the consistency of MLE and VE. Mariadassou & Matias, 2015 presented a unified framework for LBM and SBM for observations coming from an exponential family, but cannot get rid off the previous assumption to prove consistency. Using a different approach, Bickel *et al.*, 2013 showed for binary SBM (i) the consistency and aymptotic normality of the MLE in the complete model where the labels are known (ii) these properties can be transferred to the MLE of the observed model. Recently, Brault *et al.*, 2017 solved the consistency and the asymptotic normality of the MLE and VE for LBM observations coming from an exponential family.

These results were obtained when the true order $(K \times L)$ of the model is known. The question of the choice of K and L is crucial, and well-posed in the probability framework of LBM. Let K' (resp. L') be misspecifications of the number of row (resp. column) clusters. In this talk, we will study the

likelihood ratio statistics

$$D_{KK',LL'} = \log \frac{\sup_{\theta \in \Theta_{K',L'}} p(\mathbf{x};\theta)}{\sup_{\theta \in \Theta_{K,L}} p(\mathbf{x};\theta)}$$

for $K' \neq K$ or $L' \neq L$ or both. Extending Wang *et al.*, 2017 methodology for SBM, we deal with the LBM double asymptotic in row and column to provide an appropriate penalty term and define a class of selection criteria asymptotically consistent.

References

BICKEL, PETER, CHOI, DAVID, CHANG, XIANGYU, ZHANG, HAI, *et al.* 2013. Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *The Annals of Statistics*, **41**(4), 1922–1943.

BRAULT, VINCENT, KERIBIN, CHRISTINE, & MARIADASSOU, MAHENDRA. 2017. Consistency and asymptotic normality of latent blocks model estimators. *arXiv preprint arXiv:1704.06629*.

CELISSE, ALAIN, DAUDIN, JEAN-JACQUES, PIERRE, LAURENT, *et al.* 2012. Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electronic Journal of Statistics*, **6**, 1847–1899.

GOVAERT, GÉRARD, & NADIF, MOHAMED. 2008. Block clustering with bernoulli mixture models: Comparison of different approaches. *Computational Statistics & Data Analysis*, **52**(6), 3233–3245.

GOVAERT, GÉRARD, & NADIF, MOHAMED. 2010. Latent block model for contingency table. *Communications in Statistics—Theory and Methods*, **39**(3), 416–425.

KERIBIN, CHRISTINE, BRAULT, VINCENT, CELEUX, GILLES, & GOVAERT, GÉRARD. 2015. Estimation and selection for the latent block model on categorical data. *Statistics and Computing*, **25**(6), 1201–1216.

LOMET, AURORE. 2012. *Sélection de modèles pour la classification de données continues*. Ph.D. thesis, Université Technologique de Compiègne.

MARIADASSOU, MAHENDRA, & MATIAS, CATHERINE. 2015. Convergence of the groups posterior distribution in latent or stochastic block models. *Bernoulli*, **21**(1), 537–573.

WANG, YX RACHEL, BICKEL, PETER J, *et al.* 2017. Likelihood-based model selection for stochastic block models. *The Annals of Statistics*, **45**(2), 500–528.