Bayesian inference of Block mixture models for clustering

1 Subject description

The problem of complex data analysis is a central topic of modern statistical and computer and information sciences, and is connected to both theoretical and applied parts of theses sciences, as well as to several application domains, including pattern recognition, signal processing, bio-informatics data mining, complex systems modeling, etc. The analysis of complex data, in general, implies the development of statistical models and autonomous learning algorithms that aim at acquiring knowledge from raw data for analysis, interpretation and to make accurate decisions and predictions for future data. Cluster analysis of complex data is one essential task in statistical machine learning and pattern recognition. One of the most popular approaches in cluster analysis is the one based on mixture models (Titterington et al., 1985; McLachlan and Peel., 2000), known as model-based clustering (McLachlan and Basford, 1988; Celeux and Govaert, 1993; Banfield and Raftery, 1993; Fraley and Raftery, 2002). The problem of clustering therefore becomes the one of estimating the parameters of the supposed mixture model. The model estimation can be performed by maximizing the observed-data likelihood by the expectation-maximization (EM) algorithm (Dempster et al., 1977; McLachlan and Krishnan, 1997) or extensions such as Classification EM (CEM) (Celeux and Govaert, 1992), or stochastic extensions, namely (Celeux et al., 1996; Celeux and Diebolt, 1985). This approach is referred to as the maximum likelihood estimation (MLE) approach. However, the MLE approach may fail due to singularities or degeneracies (e.g. see (Stephens, 1997; Fraley and Raftery, 2007) for namely Gaussian mixtures). The Bayesian approach of mixture models (Stephens, 1997; Robert, 1994; Jean-Michel Marin and Robert, 2005; Fraley and Raftery, 2007; Bensmail et al., 1997; Richardson and Green, 1997) avoids the problems associated with the maximum likelihood described previously. by replacing the MLE by a maximum a posterior (MAP) estimation. This is namely achieved by adding regularization over the model parameters via prior parameter distributions, which are assumed to be uniform in the case of MLE. The Bayesian formulation has recently took extensive research namely from a non-parametric prospective.

The standard model-based clustering techniques (Bayesian and non-Bayesian) aim at automatically providing a partition of the data into homogeneous groups of individuals, or possibly in variables. **Model-based co-clustering** (Govaert and Nadif, 2003, 2008, 2013), also called bi-clustering or block clustering, aim at automatically and simultaneously co-clustering the data into homogeneous blocks, a block being a simultaneous association of individuals and variables. They in rely on 'block' mixture models (Govaert and Nadif, 2013) and have been developed for binary data (Govaert and Nadif, 2003, 2008; Keribin et al., 2012), categorical data (Keribin et al., 2014), contingency table (Govaert and Nadif, 2003, 2006, 2008) and continuous data (Lomet, 2012; Govaert and Nadif, 2013). The block-mixture can estimated by a block CEM for maximum classification likelihood and hard co-clustering (Govaert and Nadif, 2003, 2006, 2008) or a block (variational) EM for maximum likelihood estimation and fuzzy co-clustering (Govaert and Nadif, 2006). These interesting and quite recent block mixture models have then been examined from a Bayesian prospective to deal with some problems encountered in the MLE approach. Namely, Keribin et al. (2010) proposed a stochastic technique for the latent block model for binary data, by associating a stochastic EM with Gibbs sampling. Recently, in Keribin et al. (2012, 2014), the authors proposed for the Bayesian formulation of the latent block mixture, for respectively binary data and categorical data, a variational Bayesian inference and Gibbs sampling technique.

The model selection, in model-based co-clustering, which in general consists in selecting the best number of blocks (co-clusters) is central and can be performed by approximated penalized log-likelihood criteria such as approximated ICL or BIC-like criteria as in (Lomet et al., 2012b,a; Lomet, 2012). Keribin et al. (2012) also proposed a Bayesian sampling algorithm to derive ICL and BIC criteria for model selection in the context of binary data. Then, Keribin et al. (2014) devoloped a Bayesian inference technique using MCMC for the latent block model for categorial data, and a exact ICL for model selection.

2 Scientific objectives

The scientific objectives of this training are three-fold:

1. to implement the Bayesian block mixture of Keribin et al. (2014) (for categorical data), and test it on a text mining application,
2. then, to develop an extension to the case of multivariate data by using Gaussian distributions rather than multinomials,
3. and finally, to formulate the block mixture model into a Bayesian non-parametric context, by using a Chinese Restaurant Process as a prior (Samuel and Blei, 2012).

3 Additional Information

Supervisor: Faicel Chamroukhi http://chamroukhi.univ-tln.fr/, Maître de conférences
Location: The internship will be conducted within the LSIS laboratory UMR CNRS 7296, in Toulon
Required skills: Bases of statistical modeling and estimation; *Strong programming skills* in Matlab, R or Python; Scientific English
Desired skills: Unsupervised Learning, Mixture models, EM algorithms, Bayesian inference
Internship gratification: 508.20 € / month for 4 to 5 months
Possibility of a phd position after the internship
How to apply: Send your CV + transcripts + reference letter(s), in **A SINGLE .pdf** file, to chamroukhi@univ-tln.fr

References

Banfield, J. D. and Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49(3):803–821.

Bensmail, H., Celeux, G., Raftery, A. E., and Robert, C. P. (1997). Inference in model-based cluster analysis. *Statistics and Computing*, 7(1):1–10.

Celeux, G., Chauveau, D., and Diebolt, J. (1996). Stochastic versions of the em algorithm: an experimental study in the mixture case. *Journal of Statistical Computation and Simulation*, 55(4):287–314.

Celeux, G. and Diebolt, J. (1985). The SEM algorithm a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quarterly*, 2(1):73–82.

Celeux, G. and Govaert, G. (1992). A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics and Data Analysis*, 14:315–332.

Celeux, G. and Govaert, G. (1993). Comparison of the mixture and the classification maximum likelihood in cluster analysis. *Journal of Statistical Computation and Simulation*, 47:127–146.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of The Royal Statistical Society, B*, 39(1):1–38.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97:611–631.

Fraley, C. and Raftery, A. E. (2007). Bayesian regularization for normal mixture estimation and model-based clustering. *Journal of Classification*, (2):155–181.

Govaert, G. and Nadif, M. (2003). Clustering with block mixture models. *Pattern Recognition*, 36(2):463 – 473. Biometrics.

Govaert, G. and Nadif, M. (2006). Fuzzy clustering to estimate the parameters of block mixture models. *Soft Computing*, 10(5):415–422.

Govaert, G. and Nadif, M. (2008). Block clustering with bernoulli mixture models: Comparison of different approaches. *Computational Statistics & Data Analysis*, 52(6):3233 –3245.

Govaert, G. and Nadif, M. (2013). *Co-Clustering*. Computer engineering series. Wiley. 256 pages.

Jean-Michel Marin, K. M. and Robert, C. P. (2005). Bayesian modelling and inference on mixtures of distributions. *Bayesian Thinking - Modeling and Computation*, (25):459–507.

Keribin, C., Brault, V., Celeux, G., and Govaert, G. (2012). Model selection for the binary latent block model. In *Proceedings of COMPSTAT*.

Keribin, C., Brault, V., Celeux, G., and Govaert, G. (2014). Estimation and selection for the latent block model on categorical data. *Statistics and Computing*, pages 1–16.

Keribin, C., Govaert, G., and Celeux, G. (2010). Estimation d'un modèle à blocs latents par l'algorithme SEM. In *42èmes Journées de Statistique*, Marseille.

Lomet, A. (2012). *Sélection de modèle pour la classification croisée de données continues*. Ph.D. thesis, Université de Technologie de Compiègne.

Lomet, A., Govaert, G., and Grandvalet, Y. (2012a). An approximation of the integrated classification likelihood for the latent block model. In *ICDM Workshops*, pages 147–153.

Lomet, A., Govaert, G., and Grandvalet, Y. (2012b). Model selection in block clustering by the integrated classification likelihood. In *20th International Conference on Computational Statistics (COMPSTAT)*, pages 519–530.

McLachlan, G. and Basford, K. (1988). *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, New York.

McLachlan, G. J. and Krishnan, T. (1997). *The EM algorithm and extensions*. New York: Wiley.

McLachlan, G. J. and Peel., D. (2000). *Finite mixture models*. New York: Wiley.

Richardson, S. and Green, P. J. (1997). On bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society*, 59(4):731–792.

Robert, C. P. (1994). *The Bayesian choice: a decision-theoretic motivation*. Springer-Verlag.

Samuel, J. G. and Blei, D. M. (2012). A tutorial on bayesian non-parametric model. *Journal of Mathematical Psychology*, 56:1–12.

Stephens, M. (1997). *Bayesian Methods for Mixtures of Normal Distributions*. PhD thesis, University of Oxford.

Titterington, D., Smith, A., and Makov, U. (1985). *Statistical Analysis of Finite Mixture Distributions*. John Wiley & Sons.