Variational Learning of Dirichlet Process Parsimonious Mixtures

1 Subject description

Cluster analysis is an essential task for the unsupervised analysis of complex data and is a central topic in statistics and machine learning as well in their connected fields including signal processing. In that context, mixture models (McLachlan and Peel., 2000) are being increasingly used to model complex and heterogeneous data and attention has been focused on mixtures for multivariate high-dimensional data and to provide a clustering of such data.

In this research, we will consider Dirichlet Process Mixtures (DPM) and Chinese Restaurant Process (CRP) mixtures which provide a principled Bayesian non-parametric (BNP) alternative to the classical model-based clustering using semi-parametric (non-)Bayesian mixtures. DPM are fully Bayesian mixture approaches that automatically infer the number of mixture components (i.e clusters) and the mixture parameters, from the data and thus avoid assuming restricted functional forms for the data and allow the complexity and accuracy of the inferred models to grow as more data is observed.

We will particularly focus on the new Dirichlet Process Parsimonious Mixtures (DPPM) (Chamroukhi et al., 2014; Bartcus, 2015) which are a BNP formulation of the finite Gaussian mixtures with and eigenvalue decomposition of the group covariance matrices (Banfield and Raftery, 1993; Celeux and Govaert, 1995; Fraley and Raftery, 2002; Bensmail et al., 1997) and have proven their flexibility in cluster analysis. DPPM allow to automatically infer the model parameters and the optimal parsimonious model structure from the data, from different models, going from simplest spherical ones to more complex general ones. Learning the DPPMs is currently performed by Markov Chain Monte Carlo (MCMC) sampling using the Gibbs sampler.

Then main objective of this research is to propose a Variational learning algorithm for DPPMs rather than MCMC Gibbs sampling, in order to make computation and convergence faster. The Variational Bayesian approaches have indeed shown their performance in improving the computation in DPMs (Blei and Jordan, 2006). The algorithm will be tested on benchmarks and on a large scale bioacoustic application of unsupervised decomposition of whale song signals.

One perspective in case of faster and better convergence is to jointly consider the problems of signal decomposition and blind source separation (Moulines et al., 1997; Attias, 1999; Hyvärinen et al., 2001) with Variational DPPMs, from stereo signals.

2 Additional Information

Supervisor: Faicel Chamroukhi http://chamroukhi.univ-tln.fr/, Maître de conférences - Habilité à Diriger les Recherches
Co-Supervisor: Hervé Glotin, http://glotin.univ-tln.fr/, Professeur
Location: The internship will be conducted within the LSIS laboratory UMR CNRS 7296, in Toulon
Required skills: Bases of statistical modeling and estimation; Programming skills in Matlab, R; Scientific English
Desired skills: Unsupervised Learning, Mixture models, Bayesian inference, Python programming
Internship gratification: 554,40 € / month for 4 to 5 months - starting by February
Possibility of a Ph.D. grant after the internship

References

Attias, H. (1999). Independent factor analysis. *Neural Computation*, 11:803–851.

Banfield, J. D. and Raftery, A. E. (1993). Model-Based Gaussian and Non-Gaussian Clustering. *Biometrics*, 49(3):803–821.

Bartcus, M. (2015). *Bayesian non-parametric parsimonious mixtures for model-based clustering*. Ph.D. thesis, Université de Toulon, Laboratoire des Sciences de l'Information et des Systèmes (LSIS).

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.

Blei, D. M. and Jordan, M. I. (2006). Variational Inference for Dirichlet Process Mixtures. *Bayesian Analysis*, 1(1):121–144.

Celeux, G. and Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition*, 28(5):781–793.

Chamroukhi, F., Bartcus, M., and Glotin, H. (2014). Bayesian non-parametric parsimonious clustering. In *Proceedings of 22nd International Conference on Pattern Recognition (ICPR)*, Stockholm.

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

Hyvärinen, A., Karhunen, J., and Oja, E., editors (2001). *Independent Component Analysis*. Wiley.

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

Moulines, E., Cardoso, J., and Gassiat, E. (1997). Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models. In *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '97, Munich, Germany, April 21-24, 1997*, pages 3617–3620. IEEE Computer Society.