MiMo 2021
Workshop on statistical methods for mixture models

PLANNING

All talks take place online via the software Zoom. The invitation link will be sent later via e-mail to participants.

Thursday, April 08

09 : 15 – 09 : 30	Welcome speech
09 : 30 – 10 : 15	Andréa Rau - INRAE *Mixture models as a useful tool for identifying co-expressed genes from RNA-seq data*
10 : 15 – 11 : 00	Vincent Brault - Université de Grenobles Alpes *Models of segmentation mixtures*
11 : 00 – 11 : 30	Coffee break (Gather Town)
11 : 30 – 12 : 15	Christophe Biernacki - INRIA, Université de Lille *Gaussian-Based Visualization of Gaussian and Non-Gaussian Model-Based Clustering*
12 : 15 – 13 : 30	Lunch Break (with Gather Town from 13 :00)
13 : 30 – 14 : 15	Pierre Vandekerkhove - Université Paris-Est Marne-la-Vallée *Semiparametric two-sample nodular distribution test*
14 : 15 – 15 : 00	Élisabeth Gassiat - Université Paris-Saclay *Efficient semi-parametric estimation for multidimensional mixtures*

Friday, April 09

09 : 30 – 10 : 15	Trung Tin Nguyen - Université de Caen Normandie *Non-asymptotic penalization criteria for the Gaussian-gated mixture-of-experts regression models*
10 : 15 – 11 : 00	Cinzia Viroli - Università di Bologna *Recent advances on deep mixture models for the analysis of textual data*
11 : 00 – 11 : 30	Coffee break (Gather Town)
11 : 30 – 12 : 15	Dimitris Karlis - Athens University of Economics and Business *Model Based Clustering through copulas : parsimonious models*
12 : 15 – 13 : 30	Lunch Break (with Gather Town from 13 :00)
13 : 30 – 14 : 15	Hajo Holzmann - Philipps Universität-Marburg *Mixture models and mixtures of regressions with nonparametric components*
14 : 15 – 15 : 00	Van Hà Hoang - Université de Rouen Normandie *Adaptive non-parametric estimation of a component density in a two-class mixture model*

ABSTRACTS

Mixture models as a useful tool for identifying co-expressed genes from RNA-seq data

Andréa RAU - INRAE

Complex studies of transcriptome dynamics are now routinely carried out using RNA sequencing (RNA-seq). A common goal in such studies is to identify groups of co-expressed genes that share similar expression profiles across several treatment conditions, time points, or tissues. These co-expression analyses serve both as an exploratory visualization tool as well as a hypothesis-generating tool for poorly annotated genes. In this talk, I will discuss some of the mixture-model based approaches we have developed in recent years for RNA-seq co-expression analysis, with a particular focus on the practical issues surrounding the use of such approaches and their implementation within the R/Bioconductor package ecosystem. Finally, as studies with matched multi-view data (i.e., at different biological levels of molecular information) are becoming increasingly common, I will also briefly discuss our recent work to integratively use multiple data views to aggregate or split existing clusters from multi-omics data.

Models of segmentation mixtures

Vincent BRAULT - Université Grenoble Alpes

(Joint work with : E. Devijver and C. Laclau)

In the literature, two different approaches exist when one can assume that observations of a phenomenon come from different distributions : (1) if the order of these observations is meaningful, as it is the case for time series for instance, then we search for change points to differentiate the distributions [1] ; (2) if the order of the observations is not relevant then mixture models [2] are a suitable approach. Furthermore, in the case of data matrices one can also search for distinct patterns between the rows and the columns, and once again, two techniques can be used depending on whether the order of the rows and/or the columns is relevant [3, 4] or not [5]. However, to the best of our knowledge, there is no existing procedure able to handle the case where the order is only important for one of the two dimensions (i.e. either the rows or the columns). Traditionally, the models introduced in this context are considering the columns without constraining the order of the obtained group, hoping that they are connected [6].
In this work, we study the possibility of a hybrid approach between the aforementioned communities of research. We will introduce different procedures, where the order can be imposed only on the rows or on the columns, and will compare to previous works that were not taking order into account.

References

[1] Edward G. Carlstein, Hans-Georg Müller and David Siegmund. Change-point problems, *Institute of Mathematical Statistics*, 1994.

[2] Geoffrey J. McLachlan. The classification and mixture maximum likelihood approaches to cluster analysis, *Handbook of Statistics*, **2**, 199-208, 1982.

[3] Vincent Brault, Julien Chiquet, Céline Lévy-Leduc and others. Efficient block boundaries estimation in block-wise constant matrices : An Application to HiC data, *Electronic journal of statistics*, **11 :1**, 1570-1599, 2017.

[4] Vincent Brault, Sarah Ouadah, Laure Sansonnet and Céline Lévy-Leduc. Nonparametric multiple change-point estimation for analyzing large Hi-C data matrices, *Journal of Multivariate Analysis*, **165**, 143-165, 2018.

[5] Gérard Govaert and Mohamed Nadif. Clustering with block mixture models, *Pattern Recognition*, **36 :2**, 463-473, 2003.

[6] Marco Corneli, Charles Bouveyron, Pierre Latouche and Fabrice Rossi. The dynamic stochastic topic block model for dynamic networks with textual edges, *Statistics and Computing*, **29 :4**, 677-695, 2019.

Gaussian-Based Visualization of Gaussian and Non-Gaussian Model-Based Clustering

Christophe BIERNACKI - INRIA, Université de Lille

(Joint work with : M. Marbac and V. Vandewalle)

A generic method is introduced to visualize in a "Gaussian-like way", and onto R^d, results of Gaussian or non-Gaussian model-based clustering. The key point is to explicitly force a spherical Gaussian mixture visualization to inherit from the within cluster overlap which is present in the initial clustering mixture. The result is a particularly user-friendly draw of the clusters, allowing any practitioner to have a through overview of the potentially complex clustering result. An entropic measure allows to inform of the quality of the drawn overlap, in comparison to the true one in the initial space. The proposed method is illustrated on four real data sets of different types (categorical, mixed, functional and network) and is implemented on the r package ClusVis.

Semiparametric two-sample nodular distribution test

Pierre VANDEKERKHOVE - Université Paris-Est Marne-la-Vallée

In this talk we present essentially a semiparametric testing approach to answer if the parametric family allocated to the unknown density of a two-component mixture model with one known component is correct or not. Based on a semiparametric estimation of the Euclidean parameters of the model (free from the null assumption), our method compares pairwise the Fourier's type coefficients of the model estimated directly from the data with the ones obtained by plugging the estimated parameters into the mixture model. These comparisons are incorporated into a sum of square type statistic which order is controlled by a penalization rule. We will show that, under mild conditions, our test statistic is asymptotically χ^2_1-distributed and study its behavior, both numerically and theoretically,

under different types of alternatives including contiguous nonparametric alternatives. At this stage of the talk we will also present how the above methodology can be adapted to test if the unknown components in a two samples problem are equal. We will then discuss the counterintuitive, from the practitioner point of view, lack of power of the maximum likelihood version of our test in a neighborhood of challenging non-identifiable situations. Several level and power studies are numerically conducted on models close to those considered in the literature, such as in McLachlan *et al.* [1], to validate the suitability of our approach. We also implement our testing procedure on the Carina galaxy real dataset which low luminosity mixes with the one of its companion Milky Way.

References

[1] G.J. McLachlan, R.W. Bean and L. Ben-Tovim Jones. A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays, *Bioinformatics*, **22**, 1608-1615, 2016.

Efficient semi-parametric estimation for multidimensional mixtures

Élisabeth GASSIAT - Université Paris-Saclay

(Joint work with : J. Rousseau and E. Vernet)

We consider nonparametric multidimensional finite mixture models and we are interested in the semiparametric estimation of the population weights. Here, the i.i.d. observations are assumed to have at least three components which are independent given the population. We approximate the semiparametric model by projecting the conditional distributions on step functions associated to some partition. Our first main result is that if we refine the partition slowly enough, the associated sequence of maximum likelihood estimators of the weights is asymptotically efficient, and the posterior distribution of the weights, when using a Bayesian procedure, satisfies a semiparametric Bernstein-von Mises theorem. We then propose a cross-validation like method to select the partition in a finite horizon. Our second main result is that the proposed procedure satisfies an oracle inequality. Numerical experiments on simulated data illustrate our theoretical results.

Non-asymptotic penalization criteria for the Gaussian-gated mixture-of-experts regression models

Trung Tin NGUYEN - Université de Caen Normandie

(Joint work with : F. Chamroukhi, H. D. Nguyen and G. J. McLachlan)

Mixture-of-experts (MoE) models are a popular framework for modeling heterogeneity in data, for both regression and classification problems in statistics and machine learning, due to their flexibility and the abundance of statistical estimation and model choice tools. Such flexibility comes from allowing the mixture weights (i.e, the gating functions) in the MoE model to depend on the explanatory variables, along with the experts (i.e, the component

densities). This permits the modeling of data arising from more complex data generating processes, compared to the classical finite mixtures and finite mixtures of regression models, whose mixing parameters are independent of the covariates. Furthermore, recently, we proved that Gaussian-gated mixtures of location-scale distributions can approximate arbitrary probability density function up to any desired level of accuracy provided the number of mixture components is sufficiently large in Lebesgue spaces.
The use of MoE models in a high-dimensional setting, when the number of explanatory variables can be much larger than the sample size, is challenging from the computational and theoretical points of view, where the literature still lacks results for dealing with the curse of dimensionality, in both the statistical estimation and feature selection problems. We aim at estimating the number of components of this mixture, as well as the complexity of the regression relationship using a penalized maximum likelihood approach. To this end, we provide both a weak oracle inequality and an l_1-oracle inequality for the Gaussian gated MoE regression models.

Recent advances on deep mixture models for the analysis of textual data

Cinzia VIROLI - Università di Bologna

Deep learning methods are receiving an exponentially increasing interest in the last years as powerful tools to learn complex representations of data. When the aim is uncovering unknown classes in an unsupervised classification perspective, methods of deep learning can be developed along the lines of mixture modeling, because of their ability to decompose a heterogeneous collection of units into a finite number of sub-groups with homogeneous structures (Viroli and McLachlan, 2019 [1]). In textual data analysis, Mixtures of Unigrams (Nigam et al., 2000 [2]) are one of the simplest and most efficient tools for clustering textual data, as they assume that documents related to the same topic have similar distributions of terms, naturally described by Multinomials. When the classification task is particularly challenging due to sparse and high-dimensional document-term matrices, a more composite representation can provide better insight on the grouping structure. In this talk, a deep version of mixtures of Unigrams for the unsupervised classification of very short documents with a large number of terms is presented. Simulation studies and real data analysis prove that going deep in clustering such data highly improves the classification accuracy with respect to more 'shallow' methods.

References

[1] Cinzia Viroli and Geoffrey J. McLachlan. Deep gaussian mixture models, *Statistics and Computing*, **29 :1**, 43-51, 2019.

[2] Kamal Nigam, Andrew Kachites McCallum, Sebastian Thrun and Tom Mitchell. Text classification from labeled and unlabeled documents using EM, *Machine learning*, **39 :2-3**, 103-134, 2000.

Model Based Clustering through copulas : parsimonious models

Dimitris KARLIS - Athens University of Economics and Business

(Joint work with : F. Panagou and I. Kosmidis)

In a recent paper (Kosmidis and Karlis, 2016 [1]) proposed model based clustering based on multivariate distributions defined through copulas. This approach offers a number of advantages over existing methods mainly due to the flexibility to define appropriate models in certain different circumstances. In this talk we exploit the ideas of extending the approach for higher dimensions. The central idea is to use a Gaussian copula and implement the correlation matrix of the Gaussian copula through certain parsimonious representations giving rise to models of different complexity. We use two different approaches, the first makes use of factor analyzers based on the factor decomposition of the correlation matrix and the second is based on Choleski type decompositions. Application with real and simulated data will be also described.

References

[1] Ioannis Kosmidis and Dimitris Karlis. Model-based clustering using copulas with applications, *Statistics and computing*, **26 :5**, 1079-1099, 2016.

Mixture models and mixtures of regressions with nonparametric components

Hajo HOLZMANN - Philipps Universität-Marburg

(Joint work with : H. Werner and P. Vandekerkhove)

Recently there has been some interest in mixture models and mixtures of regressions in which at least some components are not parametrically, but rather semi- or non- parametrically specified. In the first part of the talk we give an overview of the literature and in particular of some recent contributions to this subject. Then we investigate in detail a flexible two-component semiparametric mixture of regressions model, in which one of the conditional component distributions of the response given the covariate is unknown but assumed symmetric about a location parameter, while the other is specified up to a scale parameter. The location and scale parameters together with the proportion are allowed to depend nonparametrically on covariates. After settling identifiability, we provide local M-estimators for these parameters which converge in the sup-norm at the optimal rates over Hölder-smoothness classes. We also introduce an adaptive version of the estimators based on the Lepski-method. We investigate the finite-sample behaviour of our method in a simulation study, and give an illustration to a real data set from bioinformatics.

Adaptive non-parametric estimation of a component density in a two-class mixture model

Van Hà HOANG - Université de Rouen Normandie

(Joint work with : G. Chagny, A. Channarond and A. Roche)

We consider a mixture model of two probability distributions where one component is the uniform distribution on $[0, 1]$. We adresse here the problem of nonparametric and adaptive estimation of the unknown probability density of the second component in the mixture, assuming that we have at our disposal an estimator of the proportion of each class. This problem appears, for instance, in the procedures of control the false discovery rate in a multiple testing context. We propose a kernel estimator with a fully data-driven bandwidth selection method. We obtain an oracle inequality and optimal rates of convergence over Hölder classes. Our theoretical results are illustrated by some numerical simulations.