Mixture models : Theory and applications

Thursday 21 June 2018 - Friday 22 June 2018

Paris

Recueil des résumés

Contents

Theory around mixtures / 1

Multidimensional two components Gaussian mixture detection

Auteur: Béatrice Laurent[1]

Co-auteurs: Cathy Maugis-Rabusseau [1]; Clément Marteau [2]

[1] *IMT/INSA*

[2] *ICJ*

We consider a d-dimensional i.i.d sample from a distribution with unknown density f. The problem of detection of a two-component mixture is considered. Our aim is to decide whether f is the density of a standard Gaussian random d-vector ($f = \phi_d$) against f is a two-component mixture: $f = (1-\varepsilon)\phi_d + \varepsilon\phi_d(.-\mu)$ where (ε, μ) are unknown parameters. Optimal separation conditions on ε, μ, n and the dimension d are established, allowing to separate both hypotheses with prescribed errors. Several testing procedures are proposed and two alternative subsets are considered.

Work in collaboration with C. Marteau (ICJ) and Cathy Maugis-Rabusseau (IMT/INSA)

Theory around mixtures / 2

Weakly informative reparameterisations for location-scale mixtures

Auteur: Kaniav Kamary[1]

Co-auteurs: Christian P. Robert [2]; Jeong Eun Lee [3]

[1] *Universite Paris-Dauphine / CEREMADE / INRIA, Saclay*

[2] *Universite Paris-Dauphine / University of Warwick*

[3] *Auckland University of Technology, New Zealand*

While mixtures of Gaussian distributions have been studied for more than a century, the construction of a reference Bayesian analysis of those models remains unsolved, with a general prohibition of improper priors due to the ill-posed nature of such statistical objects. This diculty is usually bypassed by an empirical Bayes resolution . By creating a new parameterisation centred on the mean and possibly the variance of the mixture distribution itself, we manage to develop here a weakly informative prior for a wide class of mixtures with an arbitrary number of components. We demonstrate that some posterior distributions associated with this prior and a minimal sample size are proper. We provide MCMC implementations that exhibit the expected exchangeability.We only study here the univariate case, the extension to multivariate location-scale mixtures being currently under study. An R package called Ultimixt is associated with this paper.

Theory around mixtures / 3

Vitesses d'estimation des paramètres d'un mélange fini

Auteur: Jonas Kahn[1]

Co-auteur: Philippe Heinrich [2]

[1] *CNRS/IMT*

[2] *Université Lille 1*

Un mélange statistique fini est une distribution de la forme $\sum_i \pi_i f(\cdot, \theta_i)$, c'est-à-dire que chaque donnée est produite
de la manière suivante: on choisit i avec probabilité π_i, et la
donnée est produite suivant la loi $f(\cdot, \theta_i)$. Les mélanges
sont donc bien adaptés à la modélisation de populations hétérogènes, ou
pour produire des distrutions complexes à partir de distributions
relativement simples.
L'estimation des paramètres π_i et θ_i du mélange sont plus
difficiles que dans les cas paramétriques lisses. Nous allons montrer
que la vitesse minimax d'estimation pour un mélange à au plus m
composantes est $n^{-1/(4m-2)}$, corrigeant ainsi le taux erronné de
$n^{-1/4}$ qui était connu.
Une part de la confusion vient sans doute du fait que les vitesses
d'estimation point par point sont différentes: en $n^{-1/2}$, mais elles
ne sont pas uniformes sur l'espace. Nous nous étendrons sur cette
différence qui n'est peut-être pas très courante.

Theory around mixtures / **4**

Estimation dans un modèle de contamination par méthode L2

Auteur: Sébastien Gadat[1]

Co-auteurs: Cathy Maugis-Rabusseau [2]; Clément Marteau [3]; Jonas Kahn [4]

[1] *TSE*

[2] *IMT/INSA*

[3] *ICJ*

[4] *CNRS/IMT*

Dans ce travail théorique, nous étudions la question de l'estimation dans un modèle de contamination par translation. On observe un échantillon iid de loi à densité dans R^d

$f^\star = (1-\lambda^\star)\phi + \lambda^\star \phi(.-\mu^\star)$

et souhaitons étudier une méthode d'estimation de la probabilité de contamination $\lambda^\star$ et son effet $\mu^\star$.

Nous proposons un critère d'estimation reposant sur une minimisation $\mathbb{L}^2$ et obtenons des résultats optimaux pour les paramètres $(\lambda^\star, \mu^\star)$. Nous utilisons pour ce-faire un raffinement astucieux et nouveau de l'inégalité de Cauchy-Schwarz pour des points sur une sphère $\mathbb{L}^2$. Enfin, nous relions nos résultats à des problèmes d'estimation en distance de Wasserstein.

Ce travail est en collaboration avec Jonas Kahn (IMT), Clément Marteau (ICJ) et Cathy Maugis-Rabusseau (IMT/INSA)

Theory around mixtures / **5**

Optimal Kullback-Leibler Aggregation in Mixture Density Estimation by Maximum Likelihood

Auteur: Arnak Dalalyan[1]

Co-auteur: Mehdi Sebbar

[1] *ENSAE / CREST*

We study the maximum likelihood estimator of density of n independent observations, under the assumption that it is well approximated by a mixture with a large number of components. The main focus is on statistical properties with respect to the Kullback-Leibler loss. We establish risk bounds taking the form of sharp oracle inequalities both in deviation and in expectation. A simple consequence of these bounds is that the maximum likelihood estimator attains the optimal rate ((logK)/n)^(1/2), up to a possible logarithmic correction, in the problem of convex aggregation when the number K of components is larger than n1/2. More importantly, under the additional assumption that the Gram matrix of the components satisfies the compatibility condition, the obtained oracle inequalities yield the optimal rate in the sparsity scenario. That is, if the weight vector is (nearly) D-sparse, we get the rate (DlogK)/n. As a natural complement to our oracle inequalities, we introduce the notion of nearly-D-sparse aggregation and establish matching lower bounds for this type of aggregation.

Summary:

We study the maximum likelihood estimator of density of n independent observations, under the assumption that it is well approximated by a mixture with a large number of components. The main focus is on statistical properties with respect to the Kullback-Leibler loss. We establish risk bounds taking the form of sharp oracle inequalities both in deviation and in expectation. A simple consequence of these bounds is that the maximum likelihood estimator attains the optimal rate ((logK)/n)^(1/2), up to a possible logarithmic correction, in the problem of convex aggregation when the number K of components is larger than n1/2. More importantly, under the additional assumption that the Gram matrix of the components satisfies the compatibility condition, the obtained oracle inequalities yield the optimal rate in the sparsity scenario. That is, if the weight vector is (nearly) D-sparse, we get the rate (DlogK)/n. As a natural complement to our oracle inequalities, we introduce the notion of nearly-D-sparse aggregation and establish matching lower bounds for this type of aggregation.

Mixture modelling and applications / 6

Variable selection for latent class analysis with application to low back pain diagnosis

Auteur: Brendan Murphy[1]

Co-auteurs: Keith M. Smart [2]; Michael Fop [3]

[1] *University College Dublin*

[2] *St. Vincent's University Hospital*

[3] *University College Dublin*

The identification of most relevant clinical criteria related to low back pain disorders may aid the evaluation of the nature of pain suffered in a way that usefully informs patient assessment and treatment. Data concerning low back pain can be of categorical nature, in the form of a check-list in which each item denotes presence or absence of a clinical condition. Latent class analysis is a model-based clustering method for multivariate categorical responses, which can be applied to such data for a preliminary diagnosis of the type of pain. In this work, we propose a variable selection method for latent class analysis applied to the selection of the most useful variables in detecting the group structure in the data. The method is based on the comparison of two different models and allows the discarding of those variables with no group information and those variables carrying the same information as the already selected ones. We consider a swap-stepwise algorithm where at each step the models are compared through an approximation to their Bayes factor. The method is applied to the selection of the clinical criteria most useful for the clustering of patients in different classes. It is shown to perform a parsimonious variable selection and to give a clustering performance comparable to the expert-based classification of patients into three classes of pain.

Mixture modelling and applications / 7

The stochastic topic block model

Auteur: Pierre Latouche[1]

[1] *Université Paris 1*

Due to the significant increase of communications between individuals via social media (Facebook, Twitter, Linkedin) or electronic formats (email, web, e-publication) in the past two decades, network analysis has become a unavoidable discipline. Many random graph models have been proposed to extract information from networks based on person-to-person links only, without taking into account information on the contents. This talk will introduce the stochastic topic block model (STBM), a probabilistic model for networks with textual edges. We will address here the problem of discovering meaningful clusters of vertices that are coherent from both the network interactions and the text contents. A classification variational expectation-maximization (C-VEM) algorithm will be proposed to perform inference. Finally, we will rely on the methodology to study the Enron political and financial scandals.

Mixture modelling and applications / 8

How to use Gaussian mixture models on patches for solving image inverse problems

Auteur: Antoine Houdard[1]

Co-auteurs: Charles Bouveyron ; Julie Delon

[1] *Télécom ParisTech / MAP5*

Most patch-based methods used in image processing involve Gaussian models or Gaussian mixture models. All theses methods can be seen through the same statistical framework. The most challenging part is the parameters estimation in the high dimensional patches space. After a brief introduction on image restoration, I will present the High-Dimensional Mixture model we introduced for image denoising [HDMI], which overcomes the curse of dimensionality by estimating intrinsic dimensions for each group of the mixture model. Finally, I will present some image restoration results obtained with this method.

References :
[HDMI] Antoine Houdard, Charles Bouveyron, Julie Delon. High-Dimensional Mixture Models For Unsupervised Image Denoising (HDMI). 2018.

Web page:
houdard.wp.imt.fr.

Mixture modelling and applications / 9

Issues, Challenges and Models for Document Clustering

Auteur: Mohamed Nadif[1]

[1] *LIPADE / University of Paris Descartes*

In recent years, document clustering or text clustering techniques have been receiving more and more attentions as a fundamental and efficient tool for organization and summarization of huge

volumes of text documents. In this talk, I provide a detailed survey of the problem of document clustering. I discuss a number of recent advances in this area and in the clustering and co-clustering contexts. I review different approaches: spectral, nonnegative matrix factorization, mixture model and latent block model.

Biological and medical applications / 11

The Latent Block Model: a useful model for high dimensional data

Auteur: Christine Keribin[1]

Co-auteurs: Gilles Celeux [2]; Valérie Robert

[1] *Université Paris Sud*

[2] *INRIA Saclay Ile-de-France*

The Latent Block Model (LBM) designs in a same exercise a clustering of the rows and the columns of a data array. Typically the LBM is expected to be useful to analyze huge data sets with many observations and many variables. But it encounters several numerical issues with big data set: maximum likelihood is jeopardized by spurious maxima and selecting a proper model is challenging since there are a lot of models in competition. In this talk, we analyze these issues. In particular, we make use of Bayesian inference to avoid spurious solutions and propose an efficient way to scan the model set. Moreover, we advocate the exact Integrated Completed Likelihood (ICL) criterion to select a proper and consistent LBM. The methods and algorithms will be illustrated with pharmacovigilance data involving large arrays of data.

Biological and medical applications / 12

C-mix: a high dimensional mixture model for censored durations, with applications to genetic data

Auteur: Agathe Guilloux[1]

[1] *Université d'Évry Val d'Essonne*

We introduce a supervised learning mixture model for censored durations (C-mix) to simultaneously detect subgroups of patients with different prognosis and order them based on their risk. Our method is applicable in a high-dimensional setting where datasets contain a large number of biomedical covariates.
To address this difficulty, we penalize the negative log-likelihood by the Elastic-Net, which leads to a sparse parameterization of the model and automatically pinpoints the relevant covariates for the survival prediction. Inference is achieved using an efficient Quasi-Newton Expectation Maximization (QNEM) algorithm. The statistical performance of the method is illustrated on three publicly available genetic cancer datasets with high-dimensional covariates.

Biological and medical applications / 13

Model-based clustering for cytometry

Auteur: Jean-Patrick Baudry[1]

[1] *LSTA*

High-dimensional flow and mass cytometry allow to measure the expression of several proteins on tens of thousands of immune cells of a patient. A common task is to predict patients disease status. This can be done based on characteristics of the cells clusters of each patient. Hence the need for clustering methods.
Some constraints make this problem challenging. The clusters of cells need to be interpretable as biologically meaningful profiles. Also, interesting groups of cells are typically rare populations.
We propose a procedure relying on model-based clustering and merging of clusters.

Biological and medical applications / **15**

Clustering of co-expressed genes

Auteur: Cathy Maugis-Rabusseau[1]

Co-auteurs: Andréa Rau [2]; Antoine Godichon-Baggioni [3]

[1] *IMT / INSA Toulouse*

[2] *INRA Jouy-en-Josas*

[3] *INSA Rouen*

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 can in fact serve a double purpose: (1) as an exploratory tool to visualize cluster-specific profile trajectories; and (2) as a hypothesis-generating tool for poorly annotated genes, as co-expression clusters may correspond to genes involved in similar biological processes or that are candidates for co-regulation.

Although a large number of clustering algorithms have been proposed in the past to identify groups of co-expressed genes from microarray data, the question of if and how such methods may be applied to RNA-seq data has only recently been addressed. During the MixStatSeq project, we have proposed several methods to solve this gene clustering problem. After a first procedure based on a Poisson mixture model (Rau et al, 2015) on the raw counts of sequenced reads for each gene, the problem was reformulated as the clustering of normalized expression profiles, which represent compositional data. Data transformations in conjunction with Gaussian mixture models were considered as an effective strategy to identify RNA-seq co-expression clusters in Rau and Maugis-Rabusseau (2017). Some related strategies were investigated in Godichon-Baggioni et al. (2018) using K-means. All of these procedures are implemented in the R/Bioconductor package coseq.

Softwares / **16**

MASSICCC: A SaaS Platform for Clustering Mixed Data

Auteur: Christophe Biernacki[1]

[1] *Université Lille 1/INRIA*

The “Big Data” paradigm involves large and complex data sets where the clustering task plays a central role for data exploration. For this purpose, model-based clustering has demonstrated many theoretical and practical successes in a various number of fields. In this context, user-friendly software are essential for speeding up diffusion of such academic advance inside the applicative world. MASSICCC (massive clustering in cloud computing) is a user-friendly SaaS platform which hosts three software specialized in different clustering tasks and written in C++. This platform allows to

manipulate complex data with very light computing tools (as a smartphone), including also some dynamical graphical outputs. However, it offers also the possibility to export the results into a R data format for further more expert tasks. The three embedded software are Mixmod, Mixtcomp and Blockcluster. Mixmod (Lebret et al. 2015) is dedicated to clustering of continuous, categorical and a mixing of continuous and categorical data. Mixtcomp (Biernacki 2015) adds the possibility to cluster totally mixed data (continuous, categorical, count, ordinal, rank, functional), potentially including missing or partially missing (like interval) data. Blockcluster (Bhatia et al. 2017) is dedicated to co-clustering of large data sets composed of different kinds of data like continuous, categorical and count ones. In this talk, we will make a focus on both the Mixmod and MixtComp software.
MASSICCC is freely available at https://massiccc.lille.inria.fr

References:
P. Bhatia, S. Iovleff & G. Govaert (2017). Blockcluster: An R Package for Model-Based Co-Clustering. Journal of Statistical Software, 76:9.
C. Biernacki (2015). Model-based clustering with mixed/missing data using the new software MixtComp. 8th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics 2015), University of London, UK, 12-14 December.
R. Lebret, S. Iovleff, F. Langrognet, C. Biernacki, G. Celeux & G. Govaert (2015). Rmixmod: The R Package of the Model-Based Unsupervised, Supervised and Semi-Supervised Classification Mixmod Library. Journal of Statistical Software, 67:6.

Softwares / **17**

”Blockcluster” and ”simerge” : Two R packages for Latent Block Models and Latent Block Models with co-variables implemented in C++

Auteur: Serge Iovleff[1]

[1] *CNRS / Laboratoire Paul Painlevé*

The basic idea of Latent Block Model (LBM) consists in making permutations of individuals (rows) and variables (columns) in order to draw a correspondence structure between individuals and variables. The R package “blockcluster” implements generative LBMs for binary, contingency, continuous and categorical data sets.
In order to estimate the parameters, it implements BEM, BCEM algorithms. The R package “simerge” is a work in progress and allows to estimate LBM when additional information is available. It implements BEM algorithm.
Both packages used C++ implementation and benefits from advanced C++ structures implemented by STK++ library and rtkore package (the port of STK++ to R). In this talk we will outline the theory LBM (with and without co-variables) and present some showcases examples. In a second part we will focus on implementation and explain how packages take advantages from C++ for large tables.

Softwares / **24**

Packages R pour la réduction de dimension en clustering

Auteur: Mohamed Sedki[1]

[1] *Université Paris-Sud*

Les méthodes de clustering ne sont pas en reste quand il s’agît de regrouper des données de grande dimension.
L’échec dû à la grande dimension a incité la communauté des statisticiens à développer des procédures de sélection

de variables contenant l'information discriminante. Une grande partie de ces techniques sont mises à disposition
sous forme de packages R. Cette présentation est une tentative de revue à travers des exemples, des packages R consacrés à la sélection de modèle en clustering par les modèles de mélange en priorité et aux méthodes basées sur la transformation des variables dans un second temps si le temps le permet.

Softwares / 25

Co-expression analyses of RNA-seq data in practice with the R/Bioconductor package coseq

Auteur: Andréa Rau[1]

Co-auteurs: Antoine Godichon-Baggioni [2]; Cathy Maugis-Rabusseau [3]

[1] *INRA Jouy-en-Josas*

[2] *INSA Rouen*

[3] *IMT/INSA*

In this talk, I will present some of the features of the R/Bioconductor package coseq, which provides a straightforward wrapper to identify groups of co-expressed genes from RNA sequencing data using Poisson mixture models (Rau et al., 2015), Gaussian mixture models (Rau et al., 2017), or the K-means algorithm (Godichon-Baggioni et al., 2018) in conjunction with appropriately chosen data transformations. In particular, I will focus on our efforts to facilitate use of coseq within standard RNA-seq analysis pipelines. I will also highlight some successful recent biological applications of coseq at INRA in a variety of organisms, including the chicken (Endale Ahanda et al., 2014), tomato (Sauvage et al., 2017), and a parasite of the honeybee (Mondet et al., 2018). Finally, I will briefly discuss some of our recent efforts to integratively make use of multiple data views (i.e., biological levels of molecular information) to identify biologically relevant and interpretable clusters from multi-omics data.