SIS - CLADAG

CLADAG 2015

10° Scientific Meeting of the Classification and Data Analysis Group of the Italian Statistical Society

Flamingo Resort, Santa Margherita di Pula, October 8-10, 2015

BOOK OF ABSTRACTS

Editors:
Francesco Mola, Claudio Conversano

university press
ricerche sociali

Università degli Studi di Cagliari

Fondazione
Banco di Sardegna

CUEC

CLADAG is a member of the International Federation of Classification Societies (IFCS). Among its activities, CLADAG organizes a biennial scientific meeting, schools related to classification and data analysis, publishes a newsletter, and cooperates with other member societies of the IFCS to the organization of their conferences. The scientific program comprises three Keynote Lectures, an Invited Session, 10 Specialized Sessions, 15 Solicited Sessions and 15 Contributed Sessions. All the Specialized and Solicited Sessions have been promoted by the members of the Scientific Program Committee. The organizers wish to thank them for their cooperation in contributing to the success of CLADAG 2015. The Book of Abstracts contains short papers of all the presentations scheduled in the conference program. It is organized according to type of session/lecture: Keynote Lectures, Specialized Sessions, Solicited Sessions and Contributed Sessions.

CLADAG 2015

10th Scientific Meeting of the
Classification and Data Analysis Group
of the Italian Statistical Society

Flamingo Resort, Santa Margherita di Pula, October 8-10, 2015

BOOK OF ABSTRACTS

Editors:

Francesco Mola,
Claudio Conversano

CUEC
editrice

CUEC Editrice
by Sardegna Novamedia Soc. Coop.
Via Basilicata n. 57/59
09127 Cagliari, Italy
Tel. & Fax (+39) 070 271573
www.cuec.eu
info@cuec.eu

First Electronic Edition CUEC © 2016
ISBN: 978-88-8467-949-9
Euro 6,90

eBook Design: Giovanni Caprioli www.servizi-per-editoria.it • info@servizi-per-editoria.it

Participating Organizations

International Federation
of Classifican Societies
(IFCS)

Società Italiana
di Statistica
(SIS)

SIS - CLADAG
Classification and Data Analysis Group
of the Italian Statistical Society

Fondazione
Banco di Sardegna

Università degli Studi
di Cagliari

Table of Contents

- Bayesian nonparametric clustering [*Organizer: Fabrizio Ruggeri Chair: Renata Rotondi*]

  A Baysian nonparametric Approach to Model Association between Clusters of SNPs and Disease Responses [*Raffaele Argiento, Alessandra Guglielmi, Chuhsing Kate Hsiao, Fabrizio Ruggeri, Charlotte Wang*]

  A Bayesian nonparametric Model for Clustering and Borrowing Information [*Antonio Lijoi, Bernardo Nipoti, Igor Prünster*]

  Sequential Clustering based on Dirichlet Process Priors [*Roberto Casarin, Andrea Pastore, Stefano F. Tonellato*]

- Causal Inference with Complex Data Structures [*Organizer and Chair: Alessandra Mattei*]

  Short term impact of PM10 exposure on mortality: A propensity score approach [*Michela Baccini, Alessandra Mattei, Fabrizia Mealli*]

  Identification and Estimation of Causal Mechanisms in Clustered Encouragement Designs: Disentangling Bed Nets using Bayesian Principal Stratification [*Laura Forastiere, Fabrizia Mealli, Tyler van der Weele*]

  The effects of a dropout prevention program on secondary students' outcomes [*Enrico Conti, Silvia Duranti, Alessandra Mattei, Fabrizia Mealli, Nicola Sciclone*]

- Clustering in Time Series [*Organizer and Chair: Michele La*

Rocca]

Probabilistic Boosted-Oriented Clustering of Time Series [*Antonio D'Ambrosio, Gianluca Frasso, Carmela Iorio, Roberta Siciliano*]

Copula-based fuzzy clustering of time series [*Pierpaolo D'Urso, Marta Disegna, Fabrizio Durante*]

Comparing multi-step ahead forecasting functions for time series clustering [*Marcella Corduas, Giancarlo Ragozini*]

- Multiway Analysis [*Organizer and Chair: Giuseppe Bove*]

(Interactive) visualisation of threeway data [*Casper J. Albers, John C. Gower*]

Robust fuzzy clustering of multivariate time trajectories [*Pierpaolo D'Urso, Riccardo Massari*]

Estimation procedures for avoiding degenerate solutions in Candecomp/Parafac [*Paolo Giordani*]

- Big Data Analysis [*Organizer and Chair: Donato Malerba*]

Towards a statistical framework for attribute comparison in very large relational databases [*Cesare Alippi, Elisa Quintarelli, Manuel Roveri, Letizia Tanca*]

Mining Big Data with high performance computing solutions [*Fabrizio Angiulli, Stefano Basta, Stefano Lodi, Gianluca Moro, Claudio Sartori*]

Enhancing Big Data Exploration with Faceted Browsing [*Sonia Bergamaschi, Giovanni Simonini and Song Zhu*]

- New Methodologies for Composite Indicators [*Organizer and Chair: Agostino Di Ciaccio*]

  Advances in Composite-based Path Modeling for Synthetic Indicators [*Vincenzo Esposito Vinzi, Laura Trinchera, Giorgio Russolillo*]

  Composite Indicators Modeling [*Maurizio Vichi*]

  Measuring the importance of variables in composite indicators [*William Becker, Michaela Saisana, Paolo Paruolo, Andrea Saltelli*]

- Cluster analysis software and validation [*Organizer and Chair: Christian Hennig*]

  Adaptive Choice Of Input Parameters In Robust Clustering [*Luis A. Garcìa-Escudero, Augustin Mayo-Iscar*]

  Robust Model-based Clustering with Covariance Matrix Constraints [*Pietro Coretto, Christian Hennig*]

  Flexible Implementation of Resampling Schemes for Cluster Validation [*Friedrich Leisch*]

- Selecting a mixture model with a clustering focus [*Organizer and Chair: Gilles Celeux*]

  Clustering in finite mixtures using an Integrated Completed

Likelihood criterion [*Marco Bertoletti, Nial Friel and Riccardo Rastelli*]

Estimation and Model Selection for Model-Based Clustering with the Conditional Classification Likelihood [*Jean-Patrick Baudry*]

On the different ways to compute the Integrated Completed Likelihood criterion [*Gilles Celeux*]

- Exploring relationships between blocks of variables [*Organizer and Chair: Giorgio Russolillo*]

Weighted Multiblock Clustering [*Ndéye Niang, Mory Ouattara*]

Thematic Model Exploration through Multiple Co-Structure maximisation: Method and Software [*Xavier Bry, Thomas Verron*]

A New Component-based Approach of Regularisation for Multivariate Generalised Linear Regression [*Catherine Trottier, Xavier Bry, Frederic Mortier, Guillaume Cornu*]

SOLICITED SESSION

- Advances in Density-based clustering [*Organizer and Chair: Francesca Greselin*]

A Nonparametric Clustering method for Image Segmentation [*Giovanna Menardi*]

Robust Clustering for Heterogenous Skew Data [*Luis A.*

Garcìa-Escudero, Francesca Greselin, Agustin Mayo-Iscar]

Regularizing finite mixtures of Gaussian Distributions [*Bettina Grün, Gertraud Malsiner-Walli*]

- Latent variable models for longitudinal data Part I [*Organizer and Chair: Silvia Bacci*]

A Joint Model For Longitudinal and Survival Data Based on an AR(1) Latent Process [*Silvia Bacci, Francesco Bartolucci, Silvia Pandolfi*]

Finite Mixture Models for Mixed Data: EM Algorithms and Parafac Representations [*Marco Alfò, Paolo Giordani*]

On the use of the contaminated Gaussian distribution in Hidden Markov models for longitudinal data [*Antonio Punzo, Antonello Maruotti*]

- Latent variable models for longitudinal data Part II [*Organizer and Chair: Francesco Bartolucci*]

A hidden Markov approach to the analysis of incomplete multivariate longitudinal data [*Francesco Lagona*]

Latent Markov and growth mixture models: a comparison [*Fulvia Pennoni, Isabella Romeo*]

Latent worths and longitudinal paired comparison. A Markov model of dependence [*Brian Francis, Alexandra Grand, Regina Dittrich*]

- Multivariate data analysis in environmental sciences [*Organizer: Fabrizio Ruggeri; Chair: Raffaele Argiento*]

    Multivariate downscaling for non-Gaussian data [*Daniela Cocchi, Lucia Paci, Carlo Trivisano*]

    Preliminary results on tapering multivariate spatio temporal models for exposure to airborne multipollutants in Europe [*Alessandro Fassò, Francesco Finazzi and Ferdinand Ndongo*]

    Clustering macroseismic fields by statistical data depth functions [*Claudio Agostinelli, Renata Rotondi and Elisa Varini*]

- Advanced models for tourism analysis [*Organizer and Chair: Stefania Mignani*]

    Analysing territorial heterogenety in tourist' satisfaction towards Italian destinations [*Cristina Bernini, Augusto Cerqua and Guido Pellegrini*]

    Micro-economic determinants of tourist expenditure: A quantile regression approach [*Emanuela Marrocu, Raffaele Paci and Andrea Zara*]

    Inequalities and tourism consumption behaviour: a mixture model analysis [*Cristina Bernini, Maria Francesca Cracolici, Cinzia Viroli*]

- Bayesian Networks and Graphical Models in Socio-Economic Sciences [*Organizer and Chair: Paola Vicard*]

Bayesian Networks for Firm Performance Evaluation [*Maria E. De Giuli, Pietro Gottardo, Anna M. Moisello and Claudia Tarantola*]

Graphical model using copulas for measurement error modeling [*Daniela Marella, Paola Vicard*]

- Time Series in Clustering [*Organizer and Chair: Michele La Rocca*]

  Parsimonious Clustering of Time Series [*Carmela Iorio, Antonio D’Ambrosio, Gianluca Frasso, Roberta Siciliano*]

  Dynamic Time Warping-based fuzzy clustering for spatial time series [*Pierpaolo D'Urso, Marta Disegna, Riccardo Massari*]

  Periodical Feature Based Time Series Clustering [*Francesco Giordano, Michele La Rocca and Maria Lucia Parrella*]

- Big Data Analysis [*Organizer and Chair: Donato Malerba*]

  Interactive Machine Learning with R [*Giorgio Maria Di Nunzio*]

  Workload estimation for a call center [*Pierluigi Riva and Ruggiero Scommegna*]

  Prediction in Olive Oil Trade using Regression Models on Temporal Data Network [*Corrado Loglisci , Umberto Medicamento, Arturo Casieri*]

- Advances in Ordinal and Preference Data [*Organizer and Chair: Antonio D'Ambrosio*]

  Measuring consensus in the setting of non-uniform qualitative scales [*José L. Garcìa-Lapresta, David Pérez-Romàn*]

  Accurate Algorithms for Consensus Ranking Detection [*Giulio Mazzeo, Antonio D'Ambrosio, Roberta Siciliano*]

  Logistic Regression Trees for Ordinal and Preference Data [*Thomas Rusch, Achim Zeileis, Kurt Hornik*]

- Case studies in data science from Ligurian companies [*Organizer and Chair: Delio Panaro*]

  Statistical methods for the analysis of «Ostreopsis ovata» bloom events from meteo-marine data [*Ennio Ottaviani, Valentina Asnaghi, Mariachiara Chiantore, Andrea Pedroncini, Rosella Bertolotto*]

  Data mining for optimal gambling [*Gabriele Torre, Fabrizio Malfanti*]

  A Fraud Detection Algorithm for Online Banking [*Delio Panaro, Eva Riccomagno, Fabrizio Malfanti*]

  Does directors' background matter? Firm performance, board features and financial reporting reliability [*Delio Panaro, Silvia Ferramosca and Sara Trucco*]

- Modeling ordinal data [*Organizer and Chair: Maurizio Carpita*]

Posterior predictive model checks for assessing the goodness of fit of Bayesian multidimensional IRT models [*Mariagiulia Matteucci, Stefania Mignani*]

International tourism in Italy: a Bayesian Network approach [*Federica Cugnata, Giovanni Perucca*]

Clustering upper level units in multilevel models for ordinal data [*Leonardo Grilli, Agnese Panzera, Carla Rampichini*]

- Functional data analysis for environmental data [*Organizer and Chair: Tonio Di Battista*]

Clustering Spatially dependent Functional Data: a method based on the concept of spatial dispersion function of a curve [*Elvira Romano, Antonio Balzanella, Rosanna Verde*]

Two case studies on object oriented spatial statistics [*Piercesare Secchi, Simone Vantini, Valeria Vitelli*]

Inference on functional biodiversity tools [*Tonio Di Battista, Francesca Fortuna, Fabrizio Maturo*]

- Advances in quantile regression [*Organizer and Chair: Cristina Davino*]

M-quantile regression: diagnostics and parametric representation of the model [*Annamaria Bianchi, Enrico Fabrizi, Nicola Salvati, Nikos Tzavidis*]

Quantile Regression: a Bayesian Robust Approach [*Marco Bottone, Mauro Bernardi, Lea Petrella*]

A comparison among estimators for linear regression methods [*Marilena Furno, Domenico Vistocco*]

Handling heterogeneity among units in Quantile Regression [*Cristina Davino, Domenico Vistocco*]

- Directional Data [*Organizer and Chair: Giovanni C. Porzio*]

Small biased circular density estimation [*Marco Di Marzio, Stefania Fensore, Agnese Panzera, Charles C. Taylor*]

A depth-based classifier for circular data [*Giuseppe Pandolfo*]

Nonparametric estimates of the mode for directional data [*Thomas Kirschstein, Steffen Liebscher, Giovanni C. Porzio, Giancarlo Ragozini*]

- Recent developments in statistical analysis of network data [*Organizer and Chair: Domenico De Stefano*]

Game Theory and Network Models for the Reconstruction of Archaeological Networks [*Viviana Amati, Ulrik Brandes*]

A model for clustering a spatial network with application to Local Labour System identification [*Francesco Pauli, Nicola Torelli, Susanna Zaccarin*]

On the sampling distributions of the ML estimators in Network Effect Models [*Michele La Rocca, Giovanni C. Porzio, Maria Prosperina Vitale, Patrick Doreian*]

Correspondence Analysis with Doubling for Two-Mode Valued Networks [*Giancarlo Ragozini, Domenico De Stefano Daniela D'Ambrosio*]

- Current challenges in clustering and classification of biomedical data [*Organizer and Chair: Adalbert F.X. Wilhelm*]

Semantic multi classifier systems for the detection of aging related processes [*Hans A. Kestler, Ludwig Lausser, Lyn-Rouven Schirra, Florian Schmid*]

Emotion recognition in human computer interaction using multiple classifier systems [*Friedhelm Schwenker*]

Ensemble of selected classifiers [*Berthold Lausen, Asma Gul, Zardad Khan and Osama Mahmoud*]

CONTRIBUTED PAPERS

A generalized distance for inference on functional data [*Andrea Ghiglietti, Anna M. Paganoni*]

Long gaps in multivariate spatio-temporal data: an approach based on Functional Data Analysis [*Mariantonietta Ruggieri, Antonella Plaia and Francesca Di Salvo*]

Effects on curve clustering of different transformations of chronological textual data [*Matilde Trevisani and Arjuna Tuzzi*]

A note on the reliability of a classifier [*Luca Frigau*]

Robustified classification of multivariate functional data [*Francesca Ieva, Anna M. Paganoni*]

Size Control of Robust Regression Estimators [*Silvia Salini, Andrea Cerioli, Fabrizio Laurini, Marco Riani*]

The Movements of Emotions: an Exploratory Classification on Affective Movement Data [*Pasquale Dente, Arvid Kappas, Adalbert F.X. Wilhelm*]

Electre Tri-Machine Learning Approach to the Record Linkage Problem [*Valentina Minnetti, Renato De Leone*]

Quality of Classification approaches for the quantitative analysis of international conflict [*Adalbert F.X. Wilhelm*]

The rtclust Procedure for Robust Clustering [*Francesco Dotto, Alessio Farcomeni, Luis Angel Garcìa-Escudero, Agustin Mayo-Iscar*]

What are the true clusters? [*Christian Hennig*]

A novel model-based clustering approach for massive datasets of spatially registered time series. With application to sea surface temperature remote sensinf data [*Francesco Finazzi, Marian Scott*]

Big Data Classification: Simulations in the many features case [*Claus Weihs*]

From Big Data to information: statistical issues through examples [*Silvia Biffignandi, Serena Signorelli*]

Big data meet pharmaceutical industry: an application on social media data [*Caterina Liberati, Paolo Mariani*]

Defining the subjects distance in hierarchical cluster analysis by copula approach [*Andrea Bonanomi, Marta Nai Ruscone, Silvia Angela Osmetti*]

Supervised classification of defective crankshafts by image analysis [*Beatriz Remeseiro, Javier Tarrìo-Saavedra, Mario Francisco-Fernàndez, Manuel G. Penedo, Salvador Naya, Ricardo Cao*]

Archetypal Analysis for Data-Driven Prototype Identification [*Giancarlo Ragozini, Francesco Palumbo, Maria R. D'Esposito*]

Principal Component Analysis of Complex Data and Application to Climatology [*Sergio Camiz and Silvia Creta*]

Sparse exploratory multidimensional IRT models [*Lara Fontanella, Sara Fontanella, Pasquale Valentini, Nickolay Trendafilov*]

Iterative Factor Clustering for Categorical data Reconsidered [*Alfonso Iodice D'Enza, Angelos Markos, Francesco Palumbo*]

Testing Antipodal Symmetry of Circular Data [*Giovanni Casale, Giuseppe Pandolfo, Giovanni C. Porzio*]

How to define deviance residuals in multinomial regression [*Giovanni Romeo, Mariangela Sciandra, Marcello Chiodi*]

Diagnostic tools for GAMLSS fitted objects [*Andrea Marletta, Mariangela Sciandra*]

Bayesian Regression Analysis with Linked and Duplicated Data [*Andrea Tancredi, Rebecca Steorts, Brunero Liseo*]

A semi-parametric FayHerriot-type model with unknown sampling variances [*Silvia Polettini*]

Posterior Distributions from Optimally B-Robust Estimating functions and Approximate Bayesian Computation [*Ivan Luciano Danesi, Fabio Piacenza, Erlis Ruli, Laura Ventura*]

MCA Based Community Detection [*Carlo Drago*]

Classifying social roles by network structures [*Simona Gozzo, Venera Tomaselli*]

A multilevel Heckman model to investigate financial assets among old people in Europe [*Omar Paccagnella, Chiara Dal Bianco*]

Optimal Pricing Using Bayesian Semiparametric Price Response Models [*Winfried J. Steiner, Anett Weber, Stefan Lang and Peter Wechselberger*]

Monetary transmission models for banking interest rates [*Laura Parisi, Paolo Giudici, Igor Gianfrancesco and Camillo Giliberto*]

Estimating the effect of prenatal care on birth outcomes [*Emiliano Sironi, Massimo Cannas*]

Recursive partitioning: an approach based on the weighted Kemeny distance [*Mariangela Sciandra, Antonella Plaia, Veronica Picone*]

Why to study abroad? An example of clustering [*Valeria Caviezel, Anna M. Falzoni*]

A graphical copula-based tool for detecting tail dependence [*Roberta Pappadà, Fabrizio Durante, Nicola Torelli*]

Classification models as tool of bankruptcy prediction – Polish experience [*Józef Pociecha, Barbara Pawełek, Mateusz Baryła, Sabina Augustyn*]

The relationship between individual price response of beer consumers and their demographic/psychographic characteristics [*Friederike Paetz*]

The Ensemble Conceptual Clustering of Symbolic Data for Customer Loyalty Analysis [*Marcin Pełka*]

Consumers' perceptions of Corporate Social Responsibilities and willingness to pay: A Partial Least Squares [*Karsten Lübke, Christian Hose, Thomas Obermeier*]

Inspecting the quality of Italian wine through causal reasoning [*Eugenio Brentari, Maurizio Carpita, Silvia Golia*]

Exploring socio-economic factors associated with adherence to the Mediterranean diet: a multilevel approach [*Tiziana Laureti, Luca Secondi*]

Big data and ‘social’ reputation: a financial example [*Paola Cerchiello*]

Bayesian Networks for Stock Picking [*Alessandro Greppi, Maria Elena De Giuli, Claudia Tarantola*]

Portfolio selection with Lasso algorithm [*Riccardo Bramante, Silvia Facchinetti, Diego Zappa*]

Sunspot in Economic Models with Externalities [*Beatrice Venturi and Alessandro Pirisinu*]

Sequential Clustering based on Dirichlet Process Priors

Roberto Casarin, Andrea Pastore and Stefano F. Tonellato[1]

[1] Department of Economics, Ca' Foscari University of Venice, (e-mail: stone@unive.it)

Abstract: This paper proposes a new sequential clustering method based on the sequential estimation of the random partition induced by the Dirichlet process. Our approach relies on Sequential Importance Resampling (SIR) and on the estimation of the posterior probabilities that each pair of individuals are generated by the same mixture component. Such estimates do not require the identification of mixture components, and therefore are not affected by label switching. Then, a dissimilarity matrix can be easily built, allowing for the implementation of agglomerative clustering methods.

Keywords: Dirichlet process, sampling importance resampling, agglomerative clustering.

1 Dirichlet process mixture

A very important class of models in Bayesian nonparametrics is based on the Dirichlet process and is known as Dirichlet process mixture (Antoniak, 1974). In this model, the observable random variables, X_i, $i = 1,\dots, n$, are assumed to be exchangeable and generated by the following hierarchical model:

$$\begin{aligned} X_i|\theta_i &\overset{ind}{\sim} p(\cdot|\theta_i),\ \theta_i \in \Theta \\ \theta_i|G &\overset{iid}{\sim} G \\ G &\sim DP(\alpha, G_0), \end{aligned}$$

where $DP(\alpha, G_0)$ denotes a Dirichlet process (DP) with base measure G_0 and precision parameter $\alpha > 0$. Since the DP generates almost surely discrete random measures on the parameter space Θ, ties among the parameter values have positive probability, leading to a batch of clusters of the parameter vector $\theta = [\theta_1,\ldots, \theta_n]^T$. Exploiting the Polya urn representation of the DP, the model can be rewritten as

$$\begin{aligned} X_i|s_i, \theta^*_{s_i} &\overset{iid}{\sim} p(\cdot|\theta^*_{s_i}),\ \theta^*_{s_i} \in \Theta & (1) \\ \theta^*_{s_i} &\overset{iid}{\sim} G_0 & (2) \\ p(s_i = j|\mathbf{s}_{<i}) &= \begin{cases} \frac{\alpha}{\alpha+i-1} & j = k \\ \frac{n_j}{\alpha+i-1} & j \in \{k-1\}, \end{cases} & (3) \\ s_i &\perp \theta^*_j \qquad \forall i, j, & (4) \end{aligned}$$

where $\{k\} = \{1,\ldots, k\}$, $\mathbf{s}_{<i} = \{s_j, j \in \{i-1\}\}$ (in the rest of the paper, the subscript $< i$ will refer to those quantities that involve all the observations $X_{i'}$ such that $i' < i$), $s_j \in \{k\}$ for $j \in \{k-1\}$, and n_j is the number of θ_1's equal to θ^*_j. In this model representation, the parameter θ can be expressed as $(\mathbf{s}, \theta^*)$, with $\mathbf{s} = \{s_i : s_i \in \{k\}, i \in \{n\}\}$, $\theta^* = [\theta^*_1,\ldots,\theta^*_k]^T$ with $\theta^*_j \sim^{\text{iid}} G_0$ and $\theta_i = \theta^*_{s_i}$. Consequently, the marginal distribution of X_i is a mixture with k components, where k is an unknown random integer.

In a parametric non Bayesian approach, it would be quite straightforward to cluster the data by maximising the probability of the allocation of each datum to one of the k clusters (with k fixed and known), conditionally on the observed sample

(McLachlan & Peel, 2000). Unfortunately, under the assumptions we made, such computations are not feasible even numerically, due to the well known label switching problem (Frühwirth-Schnatter, 2006). Nevertheless, equations (1)-(4) will be very helpful in building a hierarchical clustering algorithm based on a Bayesian nonparametric model specification.

2 Sampling importance resampling

Under the assumptions we introduced above, following the arguments of MacEachern *et al.*, 1999, we can write the conditional posterior distribution of s_i given $x_1,\ldots, x_i$, as

$$p(s_i = j|\mathbf{s}_{<i},\theta^*,\mathbf{x}^{(j)}_{<i},x_i) = \begin{cases} \frac{n_j}{\alpha+i-1} p(x_i|\theta^*_j,\mathbf{s}_{<i},\mathbf{x}^{(j)}_{<i}) & j \in \{k\} \\ \frac{\alpha}{\alpha+i-1} p(x_i|\theta^*_{k+1}) & j = k+1, \end{cases}$$

where $\mathbf{x}^{(j)}_{<i} = \{x_{i'} : i' < i, s_{i'} = j\}$, $j = 1,\ldots,k$, and $\mathbf{x}^{(k+1)}_{<i} = \emptyset$, since $\forall i' < i$, $s_{i'} \in \{k\}$.

We can marginalise the conditional posterior of s_i with respect to θ^*, obtaining

$$p(s_i = j|\mathbf{s}_{<i},\mathbf{x}^{(j)}_{<i},x_i) = \begin{cases} \frac{n_j}{\alpha+i-1} p(x_i|s_i = j,\mathbf{s}_{<i},\mathbf{x}^{(j)}_{<i}) & j \in \{k\} \\ \frac{\alpha}{\alpha+i-1} p(x_i|s_i = k+1,\mathbf{s}_{<i},\mathbf{x}_{<i}) & j = k+1, \end{cases}$$

where

$$p(x_i|s_i = j,\mathbf{s}_{<i},\mathbf{x}_{<i}) = \int_\Theta p(x_i|\theta, s_i = j,\mathbf{s}_{<i},\mathbf{x}^{(j)}_{<i}) p(\theta|s_i = j,\mathbf{s}_{<i},\mathbf{x}^{(j)}_{<i}) d\theta \tag{5}$$

and

$$p(x_i|s_i = k+1,\mathbf{s}_{<i},\mathbf{x}_{<i}) = \int_\Theta p(x_i|\theta) dG_0(\theta). \tag{6}$$

Notice that when G_0 is a conjugate prior for (1), the computation of (5) and (6) is often straightforward.

MacEachern *et al.*, 1999, introduced the following importance sampler.

SIS algorithm. For $i = 1,\ldots, n$, repeat steps (A) and (B)
(A) Compute

$$g(x_i|\mathbf{s}_{<i}, \mathbf{x}_{<i}) \propto \sum_{j=1}^{k+1} \frac{n_j}{\alpha + i - 1} p(x_i|s_i = j, \mathbf{s}_{<i}, \mathbf{x}_{<i}^{(j)}),$$

with $n_{k+1} = \alpha$.
(B) Generate s_i from the multinomial distribution with

$$p(s_i = j|\mathbf{s}_{<i}, \mathbf{x}_{<i}^{(j)}, x_i) \propto \frac{n_j}{\alpha + i - 1} p(x_i|s_i = j, \mathbf{s}_{<i}, \mathbf{x}_{<i}^{(j)}).$$

Taking R independent replicas of this algorithm we obtain $s_i{}^{(r)}$, $i = 1,\ldots,n$, $r = 1,\ldots,R$, and $\theta^*_j \sim p(\theta|\mathbf{x}^{(j)})$, with $\mathbf{x}^{(j)} = \{x_i : i \in \{n\}, s_i = j\}$, and compute the importance weights

$$w_r \propto \prod_{i=1}^{n} g(x_i|\mathbf{s}_{<i}, \mathbf{x}_{<i})$$

such that $\sum^{R}_{r=1} w_r = 1$. Should the variance of the importance weights be too small, the efficiency of the sampler could be improved by resampling as follows (Cappé *et al.*, 2005). Compute $N_{\text{eff}} = (\sum^{R}_{r=1} w^2{}_r)^{(-1)} = 1$. If $N_{\text{eff}<R/2}$, draw R particles from the current particle set with probabilities equal to their weights, replace the old particle with the new ones and assign them constant weights $w_r = 1/R$.

3 Pairwise dissimilarities and hierarchical clustering

Intuitively, we can state that two individuals, i and j, are similar if x_i and x_j are generated by the same mixture component, i.e. if $s_i =$

s_j. Label switching prevents us from identifying mixture components, but not from assessing similarities among individuals. In fact, the algorithm introduced in the previous section may help us in estimating dissimilarities between individuals. The posterior probability that x_i and x_j are generated by the same component, i.e. the posterior probability of the event $\{s_i = s_j\}$, can be estimated as

$$\hat{p}_{ij} = \sum_{r=1}^{R} w_r I\left(s_i^{(r)}, s_j^{(r)}\right),$$

where $I(x, y) = 1$ if $x = y$ and $I(x, y) = 0$ otherwise. We can then define a dissimilarity matrix D with $i\,j$-th element $d_{ij} = 1 - p\hat{}_{ij}$, allowing us to use standard agglomerative hierarchical clustering methods based on posterior evidence.

4 Discussion

The flexibility of Bayesian nonparametric models improves robustness of classification with respect to finite mixture models. Sampling importance resampling algorithms allow for efficient computations, particularly when the base measure is conjugate to model likelihood. No restrictions on the parameters or post processing of the posterior simulations are required.

References

Antoniak, C.E. 1974. Mixtures of Dirichlet processes with applications to Bayesian non parametric problems. *Annals of Statistics*., 2, 1152-1174.

Cappé, O., Moulines, E., & T., Rydén. 2005. *Inference in Hidden Markov Models*. New York: Springer.

Frühwirth-Schnatter, S. 2006. *Finite Mixture and Markov Switching Models*. Berlin: Springer.

MacEachern, S.N., Clyde, M., & Liu, J.S. 1999. Sequential importance sampling for nonparametric Bayes models: The next generation. *The Canadian Journal of Statistics*, 27, 251-267.

McLachlan, G., & Peel, D. 2000. *Finite Mixture Models*. New York: Wiley.