

DIPARTIMENTO DI ECONOMIA E GIURISPRUDENZA
UNIVERSITÀ DI CASSINO E DEL LAZIO MERIDIONALE



CLADAG 2019

11-13 SEPTEMBER 2019
CASSINO



```
def business_model()  
  arr=[ ]  
  items="a,b,c"  
  items>>arr  
  return arr  
end
```

Book of Short Papers

Giovanni C. Porzio
Francesca Greselin
Simona Balzano
Editors



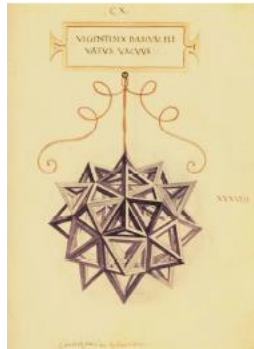
Società
Italiana di
Statistica

© CC – Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)
<https://creativecommons.org/licenses/by-nc/4.0/>

2019

Università di Cassino e del Lazio Meridionale
Centro Editoriale di Ateneo
Palazzo degli Studi Località Folcara, Cassino (FR), Italia

ISBN 978-88-8317-108-6



CLADAG 2019

Book of Short Papers

Giovanni C. Porzio

Francesca Greselin

Simona Balzano

Editors

2019

Contents

Keynotes lectures

Unifying data units and models in (co-)clustering <i>Christophe Biernacki</i>	3
Statistics with a human face <i>Adrian Bowman</i>	4
Bayesian model-based clustering with flexible and sparse priors <i>Bettina Grün</i>	5
Grinding massive information into feasible statistics: current challenges and opportunities for data scientists <i>Francesco Mola</i>	6
Statistical challenges in the analysis of complex responses in biomedicine <i>Sylvia Richardson</i>	7

Invited and contributed sessions

Model-based clustering of time series data: a flexible approach using nonparametric state-switching quantile regression models <i>Timo Adam, Roland Langrock, Thomas Kneib</i>	8
Some issues in generalized linear modeling <i>Alan Agresti</i>	12
Assessing social interest in burnout using functional data analysis through google trends <i>Ana M. Aguilera, Francesca Fortuna, Manuel Escabias</i>	16
Measuring equitable and sustainable well-being in Italian regions: a non- aggregative approach <i>Leonardo Salvatore Alaimo, Filomena Maggino</i>	20
Bootstrap inference for missing data reconstruction <i>Giuseppina Albano, Michele La Rocca, Maria Lucia Parrella, Cira Perna</i>	22
Archetypal contour shapes <i>Aleix Alcacer, Irene Epifanio, M. Victoria Ibáñez, Amelia Simó</i>	26

Random projections of variables and units <i>Laura Anderlucci, Roberta Falcone, Angela Montanari</i>	30
Sparse linear regression via random projections ensembles <i>Laura Anderlucci, Matteo Farnè, Giuliano Galimberti, Angela Montanari</i>	34
High-dimensional model-based clustering via random projections <i>Laura Anderlucci, Francesca Fortunato, Angela Montanari</i>	38
Multivariate outlier detection in high reliability standards fields using ICS <i>Aurore Archimbaud, Klaus Nordhausen, Anne Ruiz-Gazen</i>	42
Evaluating the school effect: adjusting for pre-test or using gain scores? <i>Bruno Arpino, Silvia Bacci, Leonardo Grilli, Raffaele Guetto, Carla Rampichini</i>	45
ACE, AVAS and robust data transformations <i>Anthony Atkinson</i>	49
Mixtures of multivariate leptokurtic Normal distributions <i>Luca Bagnato, Antonio Punzo, Maria Grazia Zoia</i>	53
Detecting and interpreting the consensus ranking based on the weighted Kemeny distance <i>Alessio Baldassarre, Claudio Conversano, Antonio D'Ambrosio</i>	57
Predictive principal components analysis <i>Simona Balzano, Maja Bozic, Laura Marcis, Renato Salvatore</i>	61
Flexible model-based trees for count data <i>Federico Banchelli</i>	63
Euclidean distance as a measure of conformity to Benford's law in digital analysis for fraud detection <i>Mateusz Baryła, Józef Pociecha</i>	67
The evolution of the purchase behavior of sparkling wines in the Italian market <i>Francesca Bassi, Fulvia Pennoni, Luca Rossetto</i>	71
Modern likelihood-frequentist inference at work <i>Ruggero Bellio, Donald A. Pierce</i>	75
Ontology-based classification of multilingual corpuses of documents <i>Sergey Belov, Salvatore Ingrassia, Zoran Kalinić, Paweł Lula</i>	79
Modeling heterogeneity in clustered data using recursive partitioning <i>Moritz Berger, Gerhard Tutz</i>	83

Mixtures of experts with flexible concomitant covariate effects: a bayesian solution <i>Marco Berrettini, Giuliano Galimberti, Thomas Brendan Murphy, Saverio Ranciati</i>	87
Sampling properties of an ordinal measure of interrater absolute agreement <i>Giuseppe Bove, Pier Luigi Conti, Daniela Marella</i>	91
Tensor analysis can give better insight <i>Rasmus Bro</i>	95
A boxplot for spherical data <i>Davide Buttarazzi, Giuseppe Pandolfo, Giovanni C. Porzio, Christophe Ley</i>	97
Machine learning models for forecasting stock trends <i>Giacomo Camba, Claudio Conversano</i>	99
Tree modeling ordinal responses: CUBREMOT and its applications <i>Carmela Cappelli, Rosaria Simone, Francesca Di Iorio</i>	103
Supervised learning in presence of outliers, label noise and unobserved classes <i>Andrea Cappelozzo, Francesca Greselin, Thomas Brendan Murphy</i>	104
Asymptotics for bandwidth selection in nonparametric clustering <i>Alessandro Casa, José E. Chacón, Giovanna Menardi</i>	108
Foreign immigration and pull factors in Italy: a spatial approach <i>Oliviero Casacchia, Luisa Natale, Francesco Giovanni Truglia</i>	112
Dimensionality reduction via hierarchical factorial structure <i>Carlo Cavicchia, Maurizio Vichi, Giorgia Zaccaria</i>	116
Likelihood-type methods for comparing clustering solutions <i>Luca Coraggio, Pietro Coretto</i>	120
Labour market analysis through transformations and robust multilevel models <i>Aldo Corbellini, Marco Magnani, Gianluca Morelli</i>	124
Modelling consumers' qualitative perceptions of inflation <i>Marcella Corduas, Rosaria Simone, Domenico Piccolo</i>	128
Noise resistant clustering of high-dimensional gene expression data <i>Pietro Coretto, Angela Serra, Roberto Tagliaferri</i>	132
Classify X-ray images using convolutional neural networks <i>Federica Crobu, Agostino Di Ciaccio</i>	136

A compositional analysis approach assessing the spatial distribution of trees in Guadalajara, Mexico <i>Marco Antonio Cruz, Maribel Ortego, Elisabet Roca</i>	140
Joining factorial methods and blockmodeling for the analysis of affiliation networks <i>Daniela D'Ambrosio, Marco Serino, Giancarlo Ragozini</i>	142
A latent space model for clustering in multiplex data <i>Silvia D'Angelo, Michael Fop</i>	146
Post processing of two dimensional road profiles: variogram scheme application and sectioning procedure <i>Mauro D'Apuzzo, Rose-Line Spacagna, Azzurra Evangelisti, Daniela Santilli, Vittorio Nicolosi</i>	150
A new approach to preference mapping through quantile regression <i>Cristina Davino, Tormod Naes, Rosaria Romano, Domenico Vistocco</i>	154
On the robustness of the cosine distribution depth classifier <i>Houyem Demni, Amor Messaoud, Giovanni C. Porzio</i>	158
Network effect on individual scientific performance: a longitudinal study on an Italian scientific community <i>Domenico De Stefano, Giuseppe Giordano, Susanna Zaccarin</i>	162
Penalized vs constrained maximum likelihood approaches for clusterwise linear regression modelling <i>Roberto Di Mari, Stefano Antonio Gattone, Roberto Rocci</i>	166
Local fitting of angular variables observed with error <i>Marco Di Marzio, Stefania Fensore, Agnese Panzera, Charles C. Taylor</i>	170
Quantile composite-based path modeling to estimate the conditional quantiles of health indicators <i>Pasquale Dolce, Cristina Davino, Stefania Taralli, Domenico Vistocco</i>	174
AUC-based gradient boosting for imbalanced classification <i>Martina Dossi, Giovanna Menardi</i>	178
How to measure material deprivation? A latent Markov model based approach <i>Francesco Dotto</i>	182
Decomposition of the interval based composite indicators by means of biclustering <i>Carlo Drago</i>	186
Consensus clustering via pivotal methods <i>Leonardo Egidi, Roberta Pappadà, Francesco Pauli, Nicola Torelli</i>	190

Robust model-based clustering with mild and gross outliers <i>Alessio Farcomeni, Antonio Punzo</i>	194
Gaussian processes for curve prediction and classification <i>Sara Fontanella, Lara Fontanella, Rosalba Ignaccolo, Luigi Ippoliti, Pasquale Valentini</i>	198
A new proposal for building immigrant integration composite indicator <i>Mario Fordellone, Venera Tomaselli, Maurizio Vichi</i>	199
Biodiversity spatial clustering <i>Francesca Fortuna, Fabrizio Maturo, Tonio Di Battista</i>	203
Skewed distributions or transformations? Incorporating skewness in a cluster analysis <i>Michael Gallaughier, Paul McNicholas, Volodymyr Melnykov, Xuwen Zhu</i>	207
Robust parsimonious clustering models <i>Luis Angel Garcia-Escudero, Agustin Mayo-Isacar, Marco Riani</i>	208
Projection-based uniformity tests for directional data <i>Eduardo García-Portugués, Paula Navarro-Esteban, Juan Antonio Cuesta-Albertos</i>	212
Graph-based clustering of visitors' trajectories at exhibitions <i>Martina Gentilin, Pietro Lovato, Gloria Menegaz, Marco Cristani, Marco Minozzo</i>	214
Symmetry in graph clustering <i>Andreas Geyer-Schulz, Fabian Ball</i>	218
Bayesian networks for the analysis of entrepreneurial microcredit: evidence from Italy <i>Lorenzo Giammei, Paola Vicard</i>	222
The PARAFAC model in the maximum likelihood approach <i>Paolo Giordani, Roberto Rocci, Giuseppe Bove</i>	226
Structure discovering in nonparametric regression by the GRID procedure <i>Francesco Giordano, Soumendra Nath Lahiri, Maria Lucia Parrella</i>	230
A microblog auxiliary part-of-speech tagger based on bayesian networks <i>Silvia Golia, Paola Zola</i>	234
Recent advances in model-based clustering of high dimensional data <i>Isobel Claire Gormley</i>	238
Tree embedded linear mixed models <i>Anna Gottard, Leonardo Grilli, Carla Rampichini, Giulia Vannucci</i>	239

Weighted likelihood estimation of mixtures <i>Luca Greco, Claudio Agostinelli</i>	243
A canonical representation for multiblock methods <i>Mohamed Hanafi</i>	247
An adequacy approach to estimating the number of clusters <i>Christian Hennig</i>	251
Classification with weighted compositions <i>Karel Hron, Julie Rendlova, Peter Filzmoser</i>	255
MacroPCA: an all-in-one PCA method allowing for missing values as well as cellwise and rowwise outliers <i>Mia Hubert, Peter J. Rousseeuw, Wannes Van den Bossche</i>	256
Marginal effects for comparing groups in regression models for ordinal outcome when uncertainty is present <i>Maria Iannario, Claudia Tarantola</i>	258
A multi-criteria approach in a financial portfolio selection framework <i>Carmela Iorio, Giuseppe Pandolfo, Roberta Siciliano</i>	262
Clustering of trajectories using adaptive distances and warping <i>Antonio Irpino, Antonio Balzanella</i>	266
Sampling and learning Mallows and generalized Mallows models under the Cayley distance: short paper <i>Ekhine Irurozki, Borja Calvo, Jose A. Lozano</i>	270
The gender parity index for the academic students progress <i>Aglaia Kalamatianou, Adele H. Marshall, Mariangela Zenga</i>	274
Some asymptotic properties of model selection criteria in the latent block model <i>Christine Keribin</i>	278
Invariant concept classes for transcriptome classification <i>Hans Kestler, Robin Szekely, Attila Klimmek, Ludwig Lausser</i>	282
Clustering of ties defined as symbolic data <i>Luka Kronegger</i>	283
Application of data mining in the housing affordability analysis <i>Viera Labudová, Ľubica Sipková</i>	284
Cylindrical hidden Markov fields <i>Francesco Lagona</i>	288

Comparing tree kernels performances in argumentative evidence classification <i>Davide Liga</i>	292
Recent advancement in neural network analysis of biomedical big data <i>Pietro Liò, Giovanna Maria Dimitri, Chiara Sopegno</i>	296
Bias reduction for estimating functions and pseudolikelihoods <i>Nicola Lunardon</i>	297
Large scale social and multilayer networks <i>Matteo Magnani</i>	301
Uncertainty in statistical matching by BNs <i>Daniela Marella, Paola Vicard, Vincenzina Vitale</i>	305
Evaluating the recruiters' gender bias in graduate competencies <i>Paolo Mariani, Andrea Marletta</i>	309
Dynamic clustering of network data: a hybrid maximum likelihood approach <i>Maria Francesca Marino, Silvia Pandolfi</i>	313
Stability of joint dimension reduction and clustering <i>Angelos Markos, Michel Van de Velden, Alfonso Iodice D'Enza</i>	317
Hidden Markov models for clustering functional data <i>Andrea Martino, Giuseppina Guatteri, Anna Maria Paganoni</i>	321
Composite likelihood inference for simultaneous clustering and dimensionality reduction of mixed-type longitudinal data <i>Antonello Maruotti, Monia Ranalli, Roberto Rocci</i>	325
Bivariate semi-parametric mixed-effects models for classifying the effects of Italian classes on multiple student achievements <i>Chiara Masci, Francesca Ieva, Tommaso Agasisti, Anna Maria Paganoni</i>	329
Multivariate change-point analysis for climate time series <i>Gianluca Mastrantonio, Giovanna Jona Lasinio, Alessio Pollice, Giulia Capotorti, Lorenzo Teodonio, Carlo Blasi</i>	333
A dynamic stochastic block model for longitudinal networks <i>Catherine Matias, Tabea Rebafka, Fanny Villers</i>	337
Unsupervised fuzzy classification for detecting similar functional objects <i>Fabrizio Mauro, Francesca Fortuna, Tonio Di Battista</i>	339
Mixture modelling with skew-symmetric component distributions <i>Geoffrey McLachlan</i>	343

New developments in applications of pairwise overlap <i>Volodymyr Melnykov, Yana Melnykov, Domenico Perrotta, Marco Riani, Francesca Torti, Yang Wang</i>	344
Modelling unobserved heterogeneity of ranking data with the bayesian mixture of extended Plackett-Luce models <i>Cristina Mollica, Luca Tardella</i>	346
Issues in nonlinear time series modeling of European import volumes <i>Gianluca Morelli, Francesca Torti</i>	350
Gaussian parsimonious clustering models with covariates and a noise component <i>Keefe Murphy, Thomas Brendan Murphy</i>	352
Illumination in depth analysis <i>Stanislav Nagy, Jiří Dvořák</i>	353
Copula-based non-metric unfolding on augmented data matrix <i>Marta Nai Ruscone, Antonio D'Ambrosio</i>	357
A statistical model for software releases complexity prediction <i>Marco Ortu, Giuseppe Destefanis, Roberto Tonelli</i>	361
Comparison of serious diseases mortality in regions of V4 <i>Viera Pacáková, Lucie Kopecká</i>	365
Price and product design strategies for manufacturers of electric vehicle batteries: inferences from latent class analysis <i>Friederike Paetz</i>	369
A Mahalanobis-like distance for cylindrical data <i>Lucio Palazzo, Giovanni C. Porzio, Giuseppe Pandolfo</i>	373
Archetypes, prototypes and other types <i>Francesco Palumbo, Giancarlo Ragozini, Domenico Vistocco</i>	377
Generalizing the skew-t model using copulas <i>Antonio Parisi, Brunero Liseo</i>	381
Contamination and manipulation of trade data: the two faces of customs fraud <i>Domenico Perrotta, Andrea Cerasa, Lucio Barabesi, Mario Menegatti, Andrea Cerioli</i>	385
Bayesian clustering using non-negative matrix factorization <i>Michael Porter, Ketong Wang</i>	389

Exploring gender gap in international mobility flows through a network analysis approach <i>Ilaria Primerano, Marialuisa Restaino</i>	393
Clustering two-mode binary network data with overlapping mixture model and covariates information <i>Saverio Ranciati, Veronica Vinciotti, Ernst C. Wit, Giuliano Galimberti</i>	395
A stochastic blockmodel for network interaction lengths over continuous time <i>Riccardo Rastelli, Michael Fop</i>	399
Computationally efficient inference for latent position network models <i>Riccardo Rastelli, Florian Maire, Nial Friel</i>	403
Clustering of complex data stream based on barycentric coordinates <i>Parisa Rastin, Basarab Matei, Guénaél Cabanes</i>	407
An INDSCAL based mixture model to cluster mixed-type of data <i>Roberto Rocci, Monia Ranalli</i>	411
Topological stochastic neighbor embedding <i>Nicoleta Rogovschi, Nistor Grozavu, Basarab Matei, Younès Bennani, Seiichi Ozawa</i>	415
Functional data analysis for spatial aggregated point patterns in seismic science <i>Elvira Romano, Jonatan González Monsalve, Francisco Javier Rodríguez Cortés, Jorge Mateu</i>	419
ROC curves with binary multivariate data <i>Lidia Sacchetto, Mauro Gasparini</i>	420
Silhouette-based method for portfolio selection <i>Marco Scaglione, Carmela Iorio, Antonio D'Ambrosio</i>	424
Item weighted Kemeny distance for preference data <i>Mariangela Sciandra, Simona Buscemi, Antonella Plaia</i>	428
A fast and efficient modal EM algorithm for Gaussian mixtures <i>Luca Scrucca</i>	432
Probabilistic archetypal analysis <i>Sohan Seth</i>	436
Multilinear tests of association between networks <i>Daniel K. Sewell</i>	438

Use of multi-state models to maximise information in pressure ulcer prevention trials <i>Linda Sharples, Isabelle Smith, Jane Nixon</i>	442
Partial least squares for compositional canonical correlation <i>Violetta Simonacci Massimo Guarino, Michele Gallo</i>	445
Dynamic modelling of price expectations <i>Rosaria Simone, Domenico Piccolo, Marcella Corduas</i>	449
Towards axioms for hierarchical clustering of measures <i>Philipp Thomann, Ingo Steinwart, Nico Schmid</i>	453
Influence of outliers on cluster correspondence analysis <i>Michel Van de Velden, Alfonso Iodice D'Enza, Lisa Schut</i>	454
Earthquake clustering and centrality measures <i>Elisa Varini, Antonella Peresan, Jiancang Zhuang</i>	458
Co-clustering high dimensional temporal sequences summarized by histograms <i>Rosanna Verde, Antonio Irpino, Antonio Balzanella</i>	462
Statistical analysis of item pre-knowledge in educational tests: latent variable modelling and optimal statistical decision <i>Chen Yunxiao, Lu Yan, Irini Moustaki</i>	466
Evaluation of the web usability of the University of Cagliari portal: an eye tracking study <i>Gianpaolo Zammarchi, Francesco Mola</i>	468
Application of survival analysis to critical illness insurance data <i>David Zapletal, Lucie Kopecka</i>	472

DYNAMIC CLUSTERING OF NETWORK DATA: A HYBRID MAXIMUM LIKELIHOOD APPROACH

Maria Francesca Marino¹ and Silvia Pandolfi²

¹ Department of Statistics, Computer Science, Applications “G. Parenti”, University of Florence, (e-mail: mariafrancesca.marino@unifi.it)

² Department of Economics, University of Perugia, (e-mail: silvia.pandolfi@unipg.it)

ABSTRACT: Dynamic Stochastic Block Models (SBMs) represent an attractive field of research as they provide a flexible modeling tool for the dynamic clustering of network data. However, full maximum likelihood (ML) for this class of models is not a viable estimating strategy due to the intractability of the likelihood function: variational inference represents quite a standard alternative in the frequentist framework. However, despite its simplicity, it may lead to non-optimal estimators and may suffer from local maxima solutions. We extend the hybrid ML approach developed in the context of static SBMs to deal with dynamic networks also considering both weighted and unweighted relations as well as nodal attributes which may potentially affect the block structure.

KEYWORDS: stochastic blockmodels, latent Markov models, approximate inference.

1 Introduction

Stochastic Block Models (SBMs; e.g. Snijders & Nowicki, 1997) are widely employed in the social network literature when the focus is on clustering nodes with respect to their social behavior. Recently, a growing interest has focused on the evolution of networks over time. Dynamic SBMs (e.g., Yang *et al.*, 2011, Matias & Miele, 2017, Bartolucci *et al.*, 2018) assume that the probability of observing a connection between two nodes depends on the corresponding block membership only. The evolution of the latter over time is represented by a homogeneous, discrete, latent Markov chain. Nodes belonging to a given block (state) at a given occasion share similar social behaviors.

Besides the easiness of interpretation of the model, Maximum Likelihood (ML) inference remains problematic due to the intractability of the likelihood function. Among approximate solutions available in the statistical literature, the variational approach represents a quite standard choice in the frequentist framework (e.g. Matias & Miele, 2017).

Here, we extend the hybrid ML approach recently introduced by Bartolucci *et al.*, 2017 in the framework of static SBM to deal with its dynamic counterpart, also relaxing the homogeneity assumption underlying the latent structure of the model. Moreover, to account for both binary and valued relations, we further modify the estimation algorithm to accommodate different types of (conditional) response distributions.

2 Dynamic stochastic block models

For a network of n nodes observed at T occasions, let $\mathbf{Y}^{(t)}$ denote the $n \times n$ adjacency matrix observed at occasion t , whose generic element $Y_{ij}^{(t)}, i, j = 1, \dots, n, j \neq i$, summarizes the relation existing between node i and j at occasion t . In the case of binary relations, $Y_{ij}^{(t)}$ will be a binary variable taking only zero/one values; in the case of valued relations, $Y_{ij}^{(t)}$ will be a count or a continuous variable. Without loss of generality, we focus on undirected relations without self-loops, so that $\mathbf{Y}^{(t)}$ denotes a symmetric matrix with missing values on the main diagonal. Finally, let $\mathcal{Y} = \{\mathbf{Y}^{(1)}, \dots, \mathbf{Y}^{(T)}\}$.

Dynamic SBMs (e.g., Yang *et al.*, 2011, Matias & Miele, 2017, Bartolucci *et al.*, 2018) assume that each node in the network belongs to one of k distinct blocks identified by individual- and time-specific, discrete, latent variables $U_i^{(t)} \in \{1, \dots, k\}$. These evolve over time according to a homogeneous latent Markov chain with initial probabilities $\lambda_u, u = 1, \dots, k$ and transition probabilities $\pi_{u|v}, u, v = 1, \dots, k$. Also, the dynamic SBMs postulate a *local independence assumption* between nodes: conditional on $U_i^{(t)} = u_1$ and $U_j^{(t)} = u_2$, responses $Y_{ij}^{(t)}$ are independent and identically distributed with distribution function depending on u_1 and u_2 only; that is,

$$\left[Y_{ij}^{(t)} \mid U_i^{(t)} = u_1, U_j^{(t)} = u_2 \right] \sim p \left(y_{ij}^{(t)} \mid \boldsymbol{\Psi}_{u_1 u_2} \right),$$

with $\boldsymbol{\Psi}_{u_1 u_2}$ being a suitable vector of parameters. Therefore, at each occasion, the response variable distribution only depends on the block membership of nodes i and j at that occasion.

2.1 Assessing the impact of nodal attributes on block structure

In some cases, the homogeneity assumption underlying the latent structure of the model may be too restrictive. For instance, in a friendship network, individual features, such as gender or age, may play a role in determining

the evolution of the block structure over time. To enhance the flexibility of the model, initial and transition probabilities defining the latent Markov chain may be parametrically specified as follows:

$$\begin{aligned}\log \frac{\lambda_{iu}}{\lambda_{i1}} &= \tau_{u,\lambda} + \mathbf{x}'_{i,\lambda} \boldsymbol{\phi}_{u,\lambda}, \quad u = 2, \dots, k, \\ \log \frac{\pi_{iu|v}}{\pi_{iu|u}} &= \tau_{uv,\pi} + \mathbf{x}'_{i,\pi} \boldsymbol{\phi}_{uv,\pi}, \quad u, v = 1, \dots, k, v \neq u.\end{aligned}$$

Here, $\tau_{u,\lambda}$ and $\tau_{uv,\pi}$ denote state specific intercepts, while $\boldsymbol{\phi}_{u,\lambda}$ and $\boldsymbol{\phi}_{uv,\pi}$ denote the parameters measuring the impact of nodal attributes in $\mathbf{x}_{i,\lambda}$ and $\mathbf{x}_{i,\pi}$ on the initial and the transition probabilities of the latent Markov chain, respectively.

2.2 Model inference

Let $\mathcal{U} = \{\mathbf{U}_i, i = 1, \dots, n\}$ denote the overall set of latent variables in the model, with $\mathbf{U}_i = (U_i^{(1)}, \dots, U_i^{(T)})'$; the observed network distribution (i.e., the likelihood) is obtained by marginalizing out all these latent variables from the joint distribution of $\mathcal{Y}$ and $\mathcal{U}$. That is,

$$p(\mathcal{Y}) = \sum_{\mathcal{U}} \left\{ \prod_{i=1}^n \left[\lambda_{iu_i^{(1)}} \prod_{t=2}^T \pi_{iu_i^{(t)}|u_i^{(t-1)}} \prod_{t=1}^T \prod_{j>i} p(y_{ij}^{(t)} | U_i^{(t)} = u_i^{(t)}, U_j^{(t)} = u_j^{(t)}) \right] \right\}, \quad (1)$$

where $\sum_{\mathcal{U}}$ denotes the sum over the support of $\mathcal{U}$. Due to such a summation, computing the likelihood in eq. (1) becomes prohibitive even for small n and T . To overcome the issue, an alternative inferential procedure based on a variational approximation to the likelihood function is frequently considered in the literature (e.g., Matias & Miele, 2017). Although this method is well principled and computationally fast, it may lead to non-optimal estimators, and it may suffer from local maxima solutions.

3 Hybrid ML inference for dynamic SBMs

In this work, we extend the hybrid ML approach introduced by Bartolucci *et al.*, 2017 in the framework of static SBMs to make inference on model parameters. This lies in between a full ML and a classification likelihood inference. In this latter, the realization of the discrete latent variables in the models are considered as fixed parameters to be estimated. In this respect, the

following hybrid log-likelihood function is defined:

$$\ell_{hyb} = \sum_{i=1}^n \left\{ \log \sum_{u_i^{(1)} \dots u_i^{(T)}} \lambda_{iu_i^{(1)}} \prod_{t=2}^T \pi_{iu_i^{(t)} | u_i^{(t-1)}} \prod_{t=1}^T p(\mathbf{y}_i^{(t)} | U_i^{(t)} = u_i^{(t)}; \mathbf{U}_{(-i)}^{(t)} = \tilde{\mathbf{u}}_{(-i)}^{(t)}) \right\},$$

where $\mathbf{y}_i^{(t)}$ is the vector of all observed responses $y_{ij}^{(t)}$, $j \neq i$, for unit i at occasion t , $\mathbf{U}_{(-i)}^{(t)}$ denotes the vector of latent variables referring to time t and associated to all nodes in the network, but for the i -th one, and $\tilde{\mathbf{u}}^{(t)}$ denotes the corresponding realization. These latter are considered as fixed discrete parameters in the model taking values in the set $\{1, \dots, k\}$. Last, $p(\mathbf{y}_i^{(t)} | U_i^{(t)} = u_i^{(t)}, \mathbf{U}_{(-i)}^{(t)} = \tilde{\mathbf{u}}_{(-i)}^{(t)}) = \prod_{j \neq i} p(y_{ij}^{(t)} | U_i^{(t)} = u_i^{(t)}; \mathbf{U}_{(-i)}^{(t)} = \tilde{\mathbf{u}}_{(-i)}^{(t)})$.

Let $\boldsymbol{\theta}$ denote the vector of all model parameters; this can be estimated by alternating the following steps until convergence: (i) *Classification step* – ℓ_{hyb} is maximized wrt $\tilde{\mathbf{u}}^{(t)}$, $t = 1, \dots, T$, by identifying the configuration in the set $\{1, \dots, k\}$ providing the highest log-likelihood value; (ii) *Expectation step* – for each i , the expected value of the individual complete-data (hybrid) log-likelihood is computed, given the current parameter estimates $\boldsymbol{\theta}$ and $\tilde{\mathbf{u}}^{(t)}$; (iii) *Maximization step* – the expected value of the complete-data (hybrid) log-likelihood is maximized wrt $\boldsymbol{\theta}$.

References

- BARTOLUCCI, F., MARINO, M.F., & PANDOLFI, S. 2017. Stochastic block models for social network data: inferential developments. *In: Proc. of the 32nd International Workshop on Statistical Modelling, Vol. I*.
- BARTOLUCCI, F., MARINO, M.F., & PANDOLFI, S. 2018. Dealing with reciprocity in dynamic stochastic block models. *Computational Statistics & Data Analysis*, **123**, 86–100.
- MATIAS, C., & MIELE, V. 2017. Statistical clustering of temporal networks through a dynamic stochastic block model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **79**, 1119–1141.
- SNIJDERS, T.A., & NOWICKI, K. 1997. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of classification*, **14**(1), 75–100.
- YANG, T., CHI, Y., ZHU, S., GONG, Y., & JIN, R. 2011. Detecting communities and their evolutions in dynamic social networks - a Bayesian approach. *Machine Learning*, **82**, 157–189.