UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Scienze Statistiche
SCUOLA DI DOTTORATO DI RICERCA IN SCIENZE STATISTICHE
CICLO XXVIII

BAYESIAN NONPARAMETRIC MODELING OF NETWORK DATA

Direttore della Scuola: Prof.ssa Monica Chiogna

Supervisore: Prof. Bruno Scarpa

Co-supervisore: Prof. David B. Dunson

31 Gennaio 2016

Dottorando: Daniele Durante

To Davide, Gianluca and Paolo

my sufficient statistics

Acknowledgements

Ph.D. is already stressful as it is. You start with no likelihood. Just a vague – almost improper – prior idea of what to expect. Someone – at first – might find these initial states quite annoying. Trust me, if you start updating your prior using a misspecified model and relying on complex noisy data, the worse has yet to come. I didn't make this mistake just by my own. So I would like to blame someone.

Let me start with Bruno, my highest Cook's distance data point. You've thrown me beyond the path since we met. This is already a crime for someone who expect to walk on a flat trail, but you kept leveraging the slope anytime I managed to handle your previous one. Unluckily, I found this tug-of-war quite attractive. What I've done so far and my excessive enthusiasm is – mostly – your fault.

David, you are not blameless. A smoothly increasing slope is – somehow – manageable. Once you joined this crime, you made the path purposely over-parameterized, hierarchical and nonparametric. Apparently, the worst likelihood ever. A great excuse to quit and do something else. Unfortunately, you are a brilliant professor and things mixed way better than expected. Working with you is like dreaming up methods that fully support free thinking and creativity in a non-standard research space. There is no more neighborhood I can hide from you now. Hard life.

There are many other professors accomplices of previous crimes. You know who you are, and I owe you. Just an honorable mention to Tony, my ergodicity theorem. Once you have a neat proof that a Ph.D. sooner or later reaches convergence, then it comes natural to keep on running your chain. Tony, I'm still running my naive and unscalable implementation waiting for convergence. This is your fault.

Let me blame also that huge network community of friends who – apparently – relieved my pains, starting with a good beer and nice talks after work, but ending with some crazy experience. Unfortunately I have been trained by my lovely flatmates of Via Makallè to run codes in parallel – i.e. work hard and play harder. Hence, I blame you guys for the difficult mornings spent coding or attending classes after a mad night. The list is long and includes my Padua family, my Duke one as well as the wonderful people I have shared part of my life with. I owe you all.

My L_1 neighbors Giulio, Roby, Ronnie, Giovi, Ismail and Haftu, are not sinless. I blame you guys, we built such a great compact subspace that it was natural to spend most of my last three years in the office. And – of course – once you are in the office there is nothing better to do, than working ...

I owe my family. My hyperparameters, the most skeptical reviewers but also the proudest mentors. I also owe you Veronica, my conjugate prior.

The only people to whom I apologize, are those who will be asked to carefully read this thesis. It is quite long and may suffer from label-switching. If – while you are reading – you want to blame someone, blame me. This work is the resulting posterior after updating my improper beliefs with the experiences I collected during these years. As George Box says: "Statisticians have the bad habit of falling in love with their models". In fact, even if it seems I'm complaining, I'm indeed the real guilty, as I felt in love with what the people I blamed before taught me during this wonderful Ph.D.

Padova, Daniele Durante

27 January 2016

Abstract

Network data representing relationship structures among a set of nodes are available in many fields of applications covering social science, neuroscience, business intelligence and broader relational settings. Although early probability models for networks date back almost sixty years, this field of research is still an object of intense and dynamic interest. A primary reason for the recent growth of statistical methodologies in modeling of networks is that the routine collection of such data is a recent development. Online social networks, novel neuroimaging technologies, improved business intelligence analyses and sophisticated computer algorithms monitoring world news media, currently provide increasingly complex network data sets along with novel motivating applications and new methodological questions. A challenging issue in such settings is that data are available via multiple network observations and hence the rich literature in modeling of a single network falls far short of the goal of providing flexible inference in this scenario.

Statistical modeling of replicated network data is still on its infancy and several questions remain about coherence of inference, flexibility, computational tractability and other key issues. Motivated by complex applications from different domains, this thesis aims to take a sizable step towards addressing these issues via Bayesian nonparametric modeling. The thesis is organized in two main frameworks, further divided in different topics. The first thread develops flexible and computationally tractable stochastic processes for modeling dynamic networks, which incorporate temporal dependence and exploit latent network structures. The second focuses on defining a provably flexible representation for the probabilistic generative mechanism underlying a network-valued random variable, which is able to provide valuable insights both on shared and subject – or phenotype – specific sources of variability in the network structure.

Sommario

I dati di rete misurano connessioni tra un insieme di nodi e ricorrono in molti campi di studio, tra cui le scienze sociali, le neuroscienze, il marketing ed altre discipline. Sebbene i primi modelli probabilistici per dati di rete risalgano a circa sessant'anni fa, questo campo di ricerca è tuttora oggetto di vivace ed intenso interesse. La principale motivazione per la recente crescita di metodologie statistiche per la modellazione di reti è legata alla sempre più massiccia accessibilità a dati di questo tipo. Le reti sociali online, i recenti sviluppi tecnologici nel monitoraggio di reti cerebrali e la disponibilità di algoritmi sofisticati per catalogare informazioni dai mezzi di comunicazione, forniscono dati di rete caratterizzati da una progressiva complessità e contribuiscono a nuovi interrogativi applicativi e metodologici. Un aspetto comune a queste nuove basi di dati è legato alla disponibilità di misure ripetute di reti, anzichè di una sola rete. Di conseguenza, l'ampia letteratura nello studio di una singola rete richiede generalizzazioni sostanziali per fornire adeguati strumenti inferenziali in questi nuovi scenari.

Le tecniche statistiche di modellazione per misure ripetute di reti sono ancora agli albori e diversi interrogativi rimangono ancora irrisolti in merito alla coerenza dei metodi inferenziali, alla maneggevolezza degli strumenti computazionali ed altre importanti questioni. Questa tesi è motivata da applicazioni complesse in diversi ambiti di studio e si pone l'obiettivo di compiere un passo considerevole nella risponda alle precedenti tematiche attraverso modelli Bayesiani non parametrici. Il lavoro è organizzato in due macro aree, a loro volta suddivise in diverse tematiche. La prima si pone l'obiettivo di sviluppare processi stocastici flessibili per la modellazione di reti dinamiche, capaci di incorporare sia la dipendenza temporale che quella di rete. La seconda macro area cerca invece di definire tecniche di rappresentazione flessibili per definire meccanismi probabilistici associati a variabili aleatorie di rete, con il fine di fornire informazioni chiave su strutture comuni di connessione e comprendere se e come queste si modifichino in funzione di altre variabili.

Contents

List of Figures

List of Tables

Introduction

Overview

Network data have attracted a considerable interest from several scientific communities in the recent years. A main motivation behind the popularity of this field relies on the unique focus of network science on relationship patterns among entities and their implication in several environments and phenomena. Early analyses in different fields of application have suggested that the network perspective provides an appealing direction in answering challenging scientific questions via formal inference on patterns and regularities among interacting units. The importance of this endeavor is well understood – for example – in the neuroscience community and has motivated recent developments in modeling of network data consisting of interconnection structures among anatomical regions in the human brain. Citing a recent review of the American Scientist

> *Networks of the Brain*'s (Sporns, 2010) most important contribution lies in connecting neuroscience with the science of networks [...] This is where we should be looking for solutions to the great mysteries of life and the mind.

This is clearly a simple example of a wider and intense interdisciplinary research embracing several disciplines such as social science, business intelligence, political science, biology and finance – among others.

The analysis of networks and the development of statistical methodologies for formal and robust inference on these data is a challenging task. Networks represent a type of object data – a concept encompassing a broad class of non-standard data types, ranging from functions to images and trees; refer to Wang and Marron (2007) and the references cited therein for an overview. Such data require adaptations of classical modeling frameworks to non-standard spaces. This is particularly true for inference on network data in which the set of methodologies and concepts required to learn underlying connectivity structures from the observed data is necessarily distinct from standard data analysis strategies.

Formally, a network can be represented by a graph $\mathbb{G} = (\mathbb{V}, \mathbb{E})$ where $\mathbb{V} = \{1, \dots, V\}$ denotes the set of nodes, while $\mathbb{E} \subset \mathbb{V} \times \mathbb{V}$ defines the set of connected pairs of nodes. A graph $\mathbb{G}$, naturally induces an adjacency matrix representation through the $V \times V$ matrix A having elements $A_{[vu]}$ informing on the specific relationship from node v to node u. Such relations can be binary if only the absence or presence of a connection is recorded, or can assume discrete or continuous values when a strength in the relationship is also available. Additionally, connections can be directed or undirected. In the former case a relationship from v to u can be different from the one connecting u to v, while in undirected networks $A_{[vu]} = A_{[uv]}$ for every $v = 2, \dots, V$ and $u = 1, \dots, v-1$. This thesis focuses on modeling of undirected binary networks with no self-relations. Such data are common in many applied fields and the methodologies developed for this scenario represent an important building block for generalizations dealing with more complex network structures.

Similarly to other types of object data, networks are characterized by specific topological properties. Since early contribution of Milgram (1967) focusing on small-world structural properties of networks – suggesting that most of the pairs can be joined by a relatively short path across interconnected nodes – several studies have been designed to learn recurring topological structures in real networks. Watts and Strogatz (1998) improve initial findings of Milgram (1967) on small-world networks by joining the analysis of short paths with the study of the nodes propensity to create transitive relations in a network. This topological characteristic is highly related with the concept of community structure, denoting the tendency of nodes to cluster in communities characterized by an high number of edges connecting nodes in the same community and comparatively few edges between nodes in different communities (Girvan and Newman, 2002). When these communities contains nodes that are similar with respect to other features such as language, race and age – among others – the network is said to have assortative mixing structures (Newman, 2003). Another seminal contribution of Barabási and Albert (1999) introduces the concept of scale free networks, in which few nodes – called hubs – have substantially more connections than others and the distribution of the number of edges connecting to a node follows a power-law.

Previous topological structures recur in networks from several domains, covering neuroscience (Bassett and Bullmore, 2006), social science (McPherson et al., 2001), bioinformatics (Jonsson et al., 2006) and finance (Górski et al., 2008) and have been shown to affect the functioning of networked systems in a fundamental way. This has motivated an intense initial focus on descriptive analyses of networks aimed at extracting summary statistics informative of specific topological properties. Such measures include the average length of all the shortest paths between pairs of nodes (average path length), the number of connections that each node has in the network (degree of a node), the relative frequency of the observed edges in the network with respect to the total number of possible edges (network density) and the propensity of nodes to form tight-knit groups in the network (transitivity) – among others;

refer to Börner et al. (2007) for additional summary statistics. These measures are typically further integrated with graphical visualizations – carefully tailored for network data – and statistical algorithms learning community structures, in order to provide a comprehensible description of the entire system. As highlighted in Tamassia (2007) and Fortunato (2010), respectively, these topics are still object of interest.

Although descriptive analyses provide valuable insights and represent a key initial step at the basis of deeper studies, statistical modeling of networks is currently an area of major research. In fact, networks are highly complex objects, characterized by several layers of global and local heterogeneous structures which can be dramatically altered by small amounts of random perturbations (Watts, 1999). As a results, explicitly accounting for variability in network structures via carefully tailored statistical models can lead to improved estimates of connectivity patterns and properties, while providing methodologies for formal inference in the network framework, including estimation techniques, hypothesis testing, uncertainty quantification and predictive methods.

Since the seminal random graph model of Erdös and Rényi (1959) – which consider edges as independent Bernoulli random variables with a common edge probability – several alternative specifications have been considered to induce suitable dependence structures among edges and model specific topological properties of interest. Chatterjee et al. (2011) replace the common edge probability assumption with a more flexible representation considering also node-specific propensities to form ties in the network. This specification represents the counterpart in the undirected case of the p_1 model proposed by Holland and Leinhardt (1981), which replaces the edge independence assumption with dyadic independence. Although these contributions allow tractable inference and are characterized by simple generative mechanisms, edge and dyadic independence assumptions have shown to be unrealistic in many empirical studies (Robins et al., 2007a).

Frank and Strauss (1986), generalize previous contributions to account for more realistic network structures in which two edges can be conditionally dependent – given the others – if they have a node in common. Their p^* model allows a more flexible characterization of transitivity patterns and falls within the more general class of exponential random graph models (ERGM). This popular family of statistical models defines the probability of a given network configuration A under an exponential family representation, with sufficient statistics representing suitably chosen network measures, such as number of edges, node degree, number of triangles or k-stars and others; refer to Wasserman and Pattison (1996) for a detailed overview and Robins et al. (2007a,b) for recent developments including covariates effects and more flexible characterizations. Although exponential random graphs can induce suitable dependence structures between edges and model some topological properties of interest,

these procedures are characterized by a number of drawbacks. Estimation relies on pseudo-likelihood (Strauss and Ikeda, 1990) and approximate Markov chain Monte Carlo methods (Snijders, 2002), due to the computational intractability of a full likelihood approach. Some specifications are prone to degeneracy (Handcock, 2003) and questions remain about coherence, inflexibility and other key issues (Chatterjee and Diaconis, 2013). Moreover, when the aim of statistical modeling is in developing a flexible characterization of the probabilistic generative mechanism underlying observed network data, exponential random graphs may lack flexibility in assigning the same probability to configurations having equal sufficient statistics, even when such configurations are very different.

Previous issues have motivated an intense research aimed at finding alternative specifications to exponential random graph models. An increasingly popular class of procedures including – among others – stochastic block models (Nowicki and Snijders, 2001), mixed membership stochastic block models (Airoldi et al., 2008) and latent space models (Hoff et al., 2002), assume edges as conditionally independent Bernoulli random variables given their corresponding edge probabilities, with these probabilities further characterized as a function of node-specific latent variables. As highlighted in Hunter et al. (2012), building on conditional independence rules out degeneracy issues and provides computational benefits in facilitating implementation of standard MCMC methods. Moreover, the shared dependence on a common set of node-specific latent coordinates can accurately characterize a broad variety of topological structures and dependencies within the network.

Stochastic block models (Nowicki and Snijders, 2001) and their generalizations (Kemp et al., 2006) can characterize block structures by defining edge probabilities as a function of nodes membership to latent communities and block probabilities between these communities. These formulations can recover different block patterns including assortative and disassortative structures, but have limited flexibility in relying on stochastic equivalence within the blocks. Mixed membership stochastic block models (Airoldi et al., 2008) and latent space models (Hoff et al., 2002) improve flexibility by not restricting nodes to belong to a single community. Airoldi et al. (2008) introduce mixed membership block structures which allow nodes to participate in multiple communities with node-specific degrees of affiliation. Hoff et al. (2002) define instead edge probabilities as a function of pairwise Euclidean distances between nodes in a latent space. This characterization can provably accommodate community behaviors, transitive relations, k-star structures along with predictors effects (Hoff et al., 2002) and has been recently generalized to capture additional network properties (Krivitsky et al., 2009) and account for different types of distance (Hoff, 2008). Refer to Hunter et al. (2012) for an overview of the computational methods associated with this class of models.

Main contributions of the thesis

Previous contributions cover a wide set of methodologies for statistical analysis of a single network observation, but fall far short of the goal of providing flexible inference in complex network data problems which are increasingly common in many fields. Online social networks, novel neuroimaging technologies, improved business intelligence analyses and sophisticated computer algorithms monitoring world news media, currently provide increasingly complex network data sets along with novel motivating applications and new methodological questions. Examples include dynamic networks, where data are available via time-varying adjacency matrices $A_{t_1}, \ldots, A_{t_n}$; multi-layer networks in which the data set is characterized by multiple network views A_k, with each layer $k = 1, \ldots, K$ measuring a different type of relationship on the same set of nodes; and population of networks when the replicated network data $A_1, \ldots, A_n$ consist of measurements of the same type of network on different individuals.

A common issue in such settings is that data are available via multiple network observations and hence the rich literature in modeling of a single network requires generalizations to carefully accommodate these data sets. Although a number of proposals is available, current methodologies for provably flexible and tractable inference in these scenarios are still on their infancy. Motivated by complex applications from different domains, this thesis aims to take a sizable step towards addressing the main flexibility, computational tractability and theoretical issues associated with available contributions.

After reviewing available statistical contributions in modeling of complex network data and providing a careful description of the motivating applications in Chapter 1, we focus on two main frameworks, further divided in different topics. Chapter 2 develops methodologies for modeling dynamic networks, while Chapter 3 aims to provide flexible inference procedures in analyzing populations of networks data. Concluding remarks and further directions of research are outlined in a final discussion.

Nonparametric Bayes modeling of dynamic networks

Motivated by applications to international relationships data and human interaction networks, Chapter 2 focuses on dynamically evolving binary relational matrices $A_{t_1}, \ldots, A_{t_n}$, with interest being on inference on the time-varying relationship structure and prediction. Previous proposals lack computational tractability and few theoretical results on the flexibility of the models are available. In Section 2.1 we aim to define a Bayesian nonparametric dynamic model which is provably general, reduces dimensionality and favors simple computation. Section 2.2 focuses instead on generalizing previous proposal in order to improve

the flexibility of the underlying stochastic process, while providing scalable algorithms for inference, forecasting and prediction of future networks.

Specifically, in Section 2.1 we propose a nonparametric Bayesian dynamic model, which reduces dimensionality in characterizing the collection of time-varying adjacency matrices through a lower-dimensional latent space representation, with the latent coordinates of the nodes evolving in continuous time via Gaussian processes (Rasmussen and Williams, 2006). Using a logistic mapping from the edge probability space to the latent relational space, we obtain a provably general formulation which can accommodate across-node heterogeneity in dynamic connectivity patterns. Posterior computation is available via a simple Gibbs sampler which leverages the recently developed Pólya-gamma data augmentation for Bayesian logistic regression (Polson et al., 2013). We provide theoretical results and illustrate performance via simulations. The model is applied to study dynamic international relationships.

Although providing a good methodological basis, previous approach faces the usual computational bottlenecks of Gaussian processes (GP) in scaling to large time windows, and the dynamic network inherits the stationary dependence structure of the latent GPs. Motivated by the importance of realistically modeling and forecasting dynamic networks of face-to-face human interactions, we generalize previous contribution by proposing a novel methodology for Locally Adaptive DYnamic (LADY) network inference in Section 2.2. Our LADY network model replaces GP with a dynamic latent space representation in which each subject's position evolves over time via a stochastic differential equation characterized by a simple state space formulation (Durbin and Koopman, 2012). This approach improves computational tractability utilizing results from Kalman filter (Durbin and Koopman, 2002) and allows locally varying smoothness in edge probability trajectories.

Nonparametric Bayes modeling of populations of networks

Methodologies developed in Chapter 3 are mostly motivated by neuroscience applications measuring a network of binary structural interconnections among brain regions for each subject along with a categorical variable such as creativity group.

When multiple network observations $A_1, \ldots, A_n$ are collected, current literature provides inference on the scale of the networks summary statistics which typically discards important information about the whole network structure and leads to different results depending on the summary measures considered. In Section 3.1 we develop fully general and provably flexible methods to nonparametrically estimate the probability mass function (pmf) for network-valued random variables, while favoring dimensionality reduction. Motivated by the interest in assessing evidence of differences in brain connectivity between low and high creativity subjects, previous methodology is further generalized to allow flexible changes in

the pmf across groups, facilitating robust inference and hypothesis testing on global and local associations between networks and categorical outcomes. While previous contribution is specifically motivated by neuroscience data, Section 3.2 takes the lead from the developed methods to define targeted cross-selling marketing campaigns exploiting mixed domain data from customers marginal preferences and co-subscription networks among products.

According to previous discussion, Section 3.1 proposes a fully generative Bayesian nonparametric approach for modeling the pmf of network-valued data. In particular, the probability mass function for the network-valued random variable is assigned a mixture model, allocating individuals to latent classes in terms of their network structure. Within a class, the edge probabilities are related to latent similarity measures via a logistic mapping. The similarity matrix is then factorized as the sum of a common component and a class-specific deviation that arises from embedding the nodes in a lower-dimensional latent space that takes into account the network structure. This mixture of low-rank factorizations is provably flexible and provides a valuable building block for formal inference on group differences in the network structure, allowing global and local hypothesis testing adjusting for multiplicity and robust to model misspecification. This is accomplished by generalizing the mixture of low-rank factorizations to a dependent mixture of low-rank factorizations which allows the pmf for network-valued data to shift nonparametrically between groups. An efficient Gibbs sampler is defined for posterior computation. We provide theoretical results on the flexibility of the model and assess testing performance in simulations. The approach is applied to provide novel results showing relationships between brain networks and creativity. An additional application is considered to learn how the Alzheimer's compromises the brain network.

This mixture of low-rank factorizations provides a general building block also in other applied settings. In Section 3.2 we exploit this representation for hierarchical joint modeling of customer preferences for specific products, along with co-subscription networks among such products encoding multi-buying behavior, across different insurance agencies. This formulation allows efficient targeting at both agency and customer levels, while providing key information on mono- and multi-product buying behaviors within clusters, informing cross-selling marketing campaigns.

Chapter 1

Modeling of complex networks

1.1 Motivations underlying dynamic networks

Multiple network observations can arise from dynamic monitoring of time-varying relational data. Real networks are often associated with a dynamic component and the development of statistical methodologies to learn how connectivity patterns are wired across time is a fundamental goal in many fields of application. The accurate characterization of these processes allows deeper insights in many complex phenomena, while providing inference and prediction strategies in different dynamical systems, covering information diffusion (Leskovec et al., 2007), disease contagion (Keeling and Eames, 2005), computer anomaly transmission (Idé and Kashima, 2004) and riots propagation (Berestycki et al., 2015) – among others; refer also to Holme and Saramäki (2012) for further examples.

Despite the importance of this endeavor, statistical modeling of dynamic networks is a recent field of research compared to the most popular literature on static network data. As highlighted in Goldenberg et al. (2009) the reasons for this delay are mostly related to the initial unavailability of dynamic network data along with the increased complexity in modeling their underlying structures. Since earliest Sampson (1969) monastery data set – encoding dynamic relationships between eighteen monks at three time points – the access to dynamic network data has registered an increasing growth till reaching substantially complex data structures in the last years. World Wide Web architectures (Papadimitriou et al., 2010), telecommunication infrastructures (Liu et al., 2011), recommendation systems (Sarkar et al., 2014) and novel tracking devises for face-to-face human contact (Stehlé et al., 2011), currently provide – among others – a rich variety of complex dynamic networks along with novel applied questions. In fact, beside modeling and forecasting of global network structures, there is also an increasing focus on studying edge-specific dynamic patterns. These

finer scale analyses are a key to detect anomalous behaviors and predict how information propagate from a node to the others in future scenarios.

In accomplishing this goal it is important to develop statistical methodologies which can flexibly accommodate and forecast complex temporal patterns of heterogeneity among nodes. Although the number of available contributions in statistical modeling of dynamic networks has registered an exponential growth in recent years, current proposals still raise open questions about inference, flexibility and computational tractability. Motivated by time-varying networks of geo-political relationships among international countries and dynamic human interaction data, we aim to take a further step towards improving the current state of art in this field of analysis. The following Sections 1.1.1–1.1.2 describe the data sets motivating the proposed methodologies along with a careful review of the available literature in modeling of dynamic networks.

1.1.1 Matrix-valued stochastic processes for international relationships

The last two decades have abounded with key financial and conflict events strongly affecting the world geo-political system. Notable examples include the 1997–2000 dot-com bubble (Taylor, 2009), the 2004-2007 United States housing bubble (Bernanke, 2007) and the subsequent 2007–2009 global financial crisis (Brunnermeier, 2009) along with the 2008–2012 global recession and the 2010–2012 European sovereign-debt crisis (Belkin et al., 2012). Beside perturbing financial events, recent years have been additionally characterized by several conflicts including the 2008–2009 Russian-Ukrainian gas crisis (Tsygankov, 2015) and the 2003–2011 Iraq war (Hinnebush, 2009) – among others. These events potentially have a major impact on the dynamic evolution of international relationships among pairs of countries and predicting future patterns is a key to anticipate possible crises. Motivated by the importance of this endeavor and by the current availability of sophisticated algorithms monitoring world news media – which allow access to wide catalogs of societal-scale dynamic behaviors – we aim to exploit media reports to learn dynamic geo-political patterns in the last two decades.

There is growing interest in mining massive daily news media and web data to learn and predict social and political patterns; refer to Michel et al. (2010), Leetaru (2011), Zaman et al. (2014) and the references cited therein for examples. Although media reports do not necessarily provide an unbiased view on world events, they provide useful data regarding the overall tone of public opinions (Wanta et al., 2004), including on relationships between countries. We focus on dynamic relationship networks among international countries based on the Global Database of Events, Language and Tone (GDELT) project.

GDELT is an open access database containing a comprehensive and high resolution catalog of geo-referenced sociopolitical events from 1979 to the present. Combining Conflict and

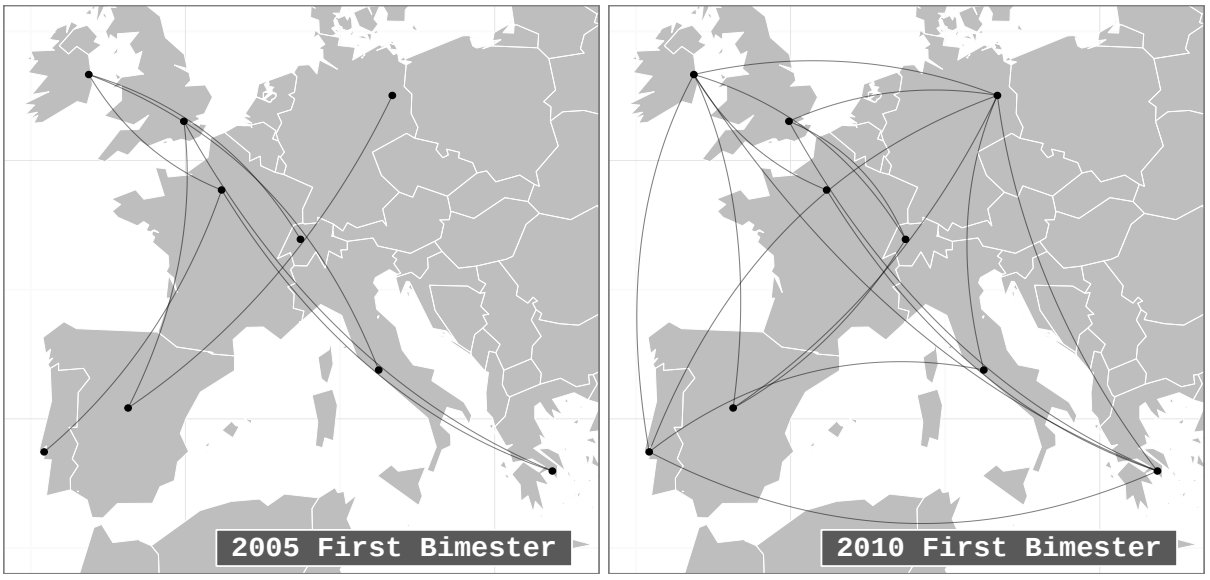


FIGURE 1.1: For some European countries in our data, dynamic relationships for selected times.

Mediation Event Observations (CAMEO) taxonomy for political events and actors (Schrodt, 2012) and Textual Analysis by Augmented Replacement Instructions (TABARI) open software for the machine coding of text data (Schrodt, 2014), GDELT provides a platform that daily monitors the world's news media reports and translates them into relational data. Specifically each row in the data set corresponds to a specific event record for which a variety of spatio-temporal and contextual information are available including – among others – the two agents interacting, their country of affiliation, the type of relationship recorded and the calendar date in which it was first reported; see Leetaru and Schrodt (2013) for a more detailed overview. This project is attracting increasing interest in the machine learning community (Schein et al., 2013; Hoff, 2014; Schein et al., 2014) and has been successfully utilized in several applied settings, covering domestic protests (Keneshloo et al., 2014), international conflicts (Brandt et al., 2013), political instabilities (Gao et al., 2013) and global disasters (Kwak and An, 2014). We are specifically interested in modeling of dynamic variations in the overall relationships among countries across the last twenty years as reported by the news media.

Our data consist of a sequence of $V \times V$ dynamic symmetric adjacency matrices $A_{t_1}, \ldots, A_{t_n}$ having entries $A_{t_i[vu]} = A_{t_i[uv]} = 1$ if there is a positive overall cooperative relationship among countries $v = 2, \ldots, V$ and $u = 1, \ldots, v-1$ at time t_i and $A_{t_i[vu]} = A_{t_i[uv]} = 0$, otherwise. For interpretability, we consider bimonthly relationships among the $V = 25$ countries heavily involved in financial crises and international conflicts in the last twenty years. Refer to Figure 1.1 for an illustration. Dynamic networks $A_{t_1}, \ldots, A_{t_{127}}$ are constructed exploiting variable `QuadClass` in the GDELT data set to first obtain matrices $A^{\text{hel}}_{t_1}, \ldots, A^{\text{hel}}_{t_{127}}$ and $A^{\text{conf}}_{t_1}, \ldots, A^{\text{conf}}_{t_{127}}$. These dynamic matrices have entries $A^{\text{hel}}_{t_i[vu]} = A^{\text{hel}}_{t_i[uv]}$ and $A^{\text{conf}}_{t_i[vu]} = A^{\text{conf}}_{t_i[uv]}$ encoding the total number of unique events among pairs of agents affiliated with countries

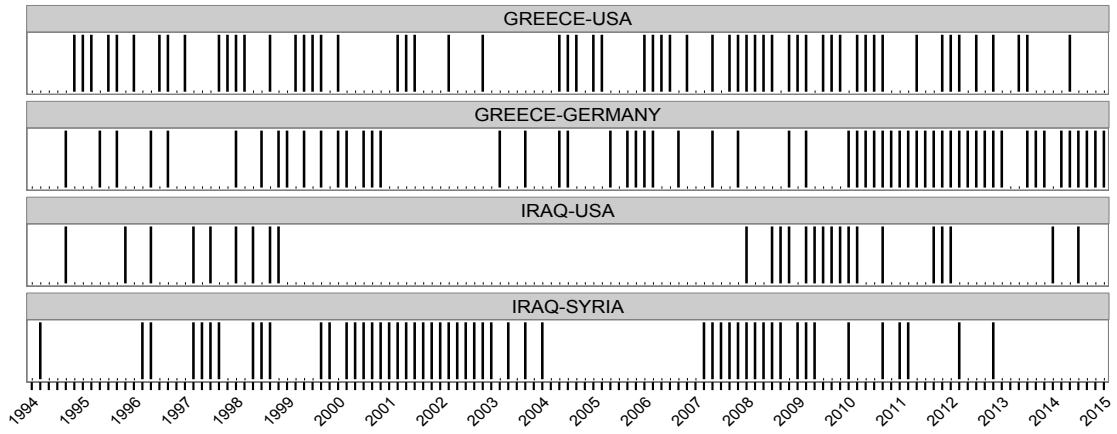


FIGURE 1.2: For select pairs of countries, barcode plot of their edges across time. Bar at t_i means $A_{t_i[vu]} = 1$.

$v = 2, \ldots, V$ and $u = 1, \ldots, v-1$, respectively, at time interval t_i. Matrix $A^{\text{hel}}_{t_i}$ counts material help events, while $A^{\text{conf}}_{t_i}$ counts material conflict events. The difference $\Delta_{t_i} = A^{\text{hel}}_{t_i} - A^{\text{conf}}_{t_i}$ provides an aggregated measure of the strength of positive association between each pair of countries, with $A_{t_i[vu]} = A_{t_i[uv]} = 1(\Delta_{t_i[vu]} = \Delta_{t_i[uv]} \geq 0)$ indicating an overall positive cooperative relationship between countries v and u at time t_i. In the previous notation $1(\cdot)$ represents the indicator function. Examples of positive events include sharing intelligence and economic aid, while examples of negative events include imposing embargo and stopping military assistance. In addition to ease in interpretation, we avoid joint modeling the dynamic count matrices $(A^{\text{hel}}_{t_i}, A^{\text{conf}}_{t_i})$ directly, to improve robustness, limiting sensitivity to missed and duplicate events. The latter are further controlled by a one-day filter which collapses event records having the same date, pairs of agents and relationship type.

As shown in Figure 1.2, the edge trajectories cycle with varying patterns of duration and inter-dependence across time. Capturing such behavior is important in assessing how the dynamic inter-relationships relate to key conflict and financial events. Occurrence times of such events can be included as time-varying predictors of the dynamic network. However, for simplicity and robustness, we instead focus on developing a dynamic network model, which is sufficiently general to account for dynamic variations in the network structure without requiring known events to be driving this variation.

Relevant literature in modeling of dynamic networks

There is a growing literature in statistical modeling of dynamic networks. A subset of contributions focus on the case in which the exact time of each edge event is observed; see for

example Butts (2008) and DuBois et al. (2013). We instead consider the case in which snapshots of a specific network are collected at multiple time points.

A popular class of procedures generalizes static exponential random graph models via discrete time Markov models (Robins and Pattison, 2001; Hanneke et al., 2010). These contributions define the transition probability from a network configuration at time t to a configuration at time $t+1$ under an exponential family representation with sufficient statistics covering network measures at time $t+1$ and suitable interaction terms with observed configuration at t; see also Krivitsky and Handcock (2014) for recent developments. Although discrete time exponential random graphs (TERGM) can leverage the several techniques available for the static case and explicitly account for topological properties as well as suitable dependence structures among edges, these procedures inherit the drawbacks of the static models they seek to generalize and are not tailored to accommodate irregular time grids.

The seminal contribution of Holland and Leinhardt (1977) and the subsequent improvements of Snijders (2001, 2005) and Snijders et al. (2010a), provide an alternative specification via continuous time Markov models in which changes in edge variables are conditionally independent given the current network observation and arise from nodes choices aimed at maximizing their utility based on the current network topology. Estimation is available via method of moments (Snijders, 2001), maximum likelihood (Snijders et al., 2010b) or Bayesian techniques (Koskinen and Snijders, 2007). Stochastic actor-oriented models provide a valuable methodology when the interest is on homogenous and time-constant effects of specific structures – covering for example, transitivity, degree popularity and exogenous variables – on network evolution, however the underlying homogeneity assumptions may fail to accommodate specific heterogenous connectivity patterns. Flexible modeling of dynamic network structures, while accommodating heterogenous behaviors is a key to improve prediction.

The importance of this endeavor has motivated increasing efforts in generalizing static latent variables models for network data to dynamic scenarios. Although these procedures do not explicitly parameterize interdependence between relations, the shared dependence on a common set of node-specific latent variables can induce rich dependence structures and allow for across-node heterogeneity in time-varying connectivity patterns. Early versions of dynamic stochastic block models (Yang et al., 2009, 2010) focus on time-varying nodes membership to blocks, while relying on time-constant block probability matrices. These initial versions have been recently generalized to more general scenarios in which both block-probabilities and nodes membership to blocks can change across possibly unequally spaced times. Xu and Hero (2014) and its recent generalization (Xu, 2015) take advantage from state space formulations, but require sufficient numbers of observations in each block to meet Gaussian assumptions for the sample mean, and rely on extended Kalman filter to linearize

the observation equation. Beside possible computational issues, time-dependent stochastic block models are specifically tailored for learning dynamic changes in block structures, and hence may fail in accurately characterizing time-varying network patterns different than those arising from block structures. Dynamic relational feature models (Foulds et al., 2011) partially improve flexibility by replacing the single block membership variable with vectors describing presence or absence of given features for each node. Although this representation accommodates more general network structures, the time-constant assumption for the feature-interaction matrix may restrict dynamic variations.

We dynamically model adjacency matrices by embedding the nodes in a low-dimensional latent Euclidean space, with their coordinates evolving in continuous time via Gaussian processes and edge probabilities constructed via a logistic mapping function. Hence, our work is most closely related to the literature on more general classes covering mixed membership stochastic block models (Airoldi et al., 2008) and latent space models (Hoff et al., 2002; Hoff, 2008). Dynamic mixed membership stochastic block models (Xing et al., 2010) and latent space models (Sarkar and Moore, 2005; Sewell and Chen, 2015) propagate information across time via state space models, Markov processes and random walk trajectories, respectively. Inference relies on several layers of approximation – such as extended Kalman filter and variational Bayes – without theory available to justify accuracy. In contrast, we provide a simple Gibbs sampling algorithm, which converges to the exact posterior and adaptively shrinks towards lower-dimensional structures.

1.1.2 Scalable and adaptive inference for face-to-face interaction data

The increasing availability of new sensing devices and wearable sensors to trace human interaction behaviors, allows growing access to these type of dynamic networks, while opening new avenues for studying underlying patterns in social interactions and how these processes relate to associated dynamic systems such as epidemic spreading. Recent studies have investigated dynamic face-to-face human interactions in several environments. Isella et al. (2011) focus on contact dynamics among individuals in two different scenarios, covering a scientific conference and a long-running museum exhibition, respectively. Vanhems et al. (2013) study interactions among staff members and patients in a hospital. Stehlé et al. (2011), Gemmetto et al. (2014) and Fournet and Barrat (2014), Mastrandrea et al. (2015) investigate face-to-face contact dynamics among students in primary and high schools, respectively; refer also to Barrat and Cattuto (2013) for a review.

Previous studies mostly focus on aggregate and time-varying descriptive analyses in order to provide a summarized overview of the topological structures underlying the observed networks and how these measures relate to environmental conditions and other variables.

Although these procedures provide valuable insights, holistic statistical models of how the human interaction networks dynamically evolve would provide improved ability to jointly infer how different network structures vary, while accounting for uncertainty. In addition, such models would be highly useful in terms of prediction and forecasting of interactions, which is of key interest in epidemiology – for example.

We are specifically interested in studying face-to-face dynamic interactions among individuals in a primary school in Lyon, France – see Stehlé et al. (2011) and Gemmetto et al. (2014) for additional details. Understanding key aspects of these interaction networks and prediction of future contacts is interesting sociologically and important in infectious disease epidemiology. Raw contact data are available at `http://www.sociopatterns.org` for 232 children between 6 and 12 years of age and 10 teachers, during two consecutive school days running from $\approx 08{:}40$ to $\approx 17{:}10$. The primary school is characterized by 5 grades, each divided in two classes comprising on average 24 children. Face-to-face contacts are monitored via wearable radio frequency identification devices (RFID), exchanging low-power radio packets when two individuals face each other at a distance of $\approx 1 - 1.5$ meters. This proximity range is chosen to represent a reasonable proxy of close social contact, while indicating a potential occasion of disease transmission (Stehlé et al., 2011). Raw data are available for consecutive windows of 20 seconds and encode which pairs of individuals established a face-to-face proximity contact during each of these time intervals; refer to Cattuto et al. (2010) for a description of RFID proximity-sensing infrastructures.

Initial descriptive analyses of these data highlight a very sparse and noisy structure with only 24 contacts – among the $29{,}161$ possible – monitored on average for every window of 20 seconds. This time scale might be too narrow to highlight recurring patterns in the dynamic evolution of underlying network topological structures. Hence, we aggregate the data in consecutive time windows of 10 minutes so that the resulting networks encode which pairs of individuals established at least one face-to-face proximity contact during each of these subsequent 10 minute time intervals. Focusing on binary connections instead of the cumulative number of contacts in the 10 minute time windows provides a simpler starting point. Moreover, under an epidemiological perspective, at least one proximity contact of 20 seconds may be sufficient for disease transmission. Although we loose short scale dynamics, these windows are sufficiently wide to highlight longer range patterns in the network topology, but maintain enough granularity to capture sharp changes which may occur in correspondence of breaks, lunch times and school hours. We found these underlying structures quite robust to moderate changes in the length of the time intervals, including 5, 15 and 20 minutes. Stehlé et al. (2011) consider a similar aggregation strategy to investigate dynamic changes in the averaged degree.

In analyzing these data, we seek inference and prediction procedures which are sufficiently

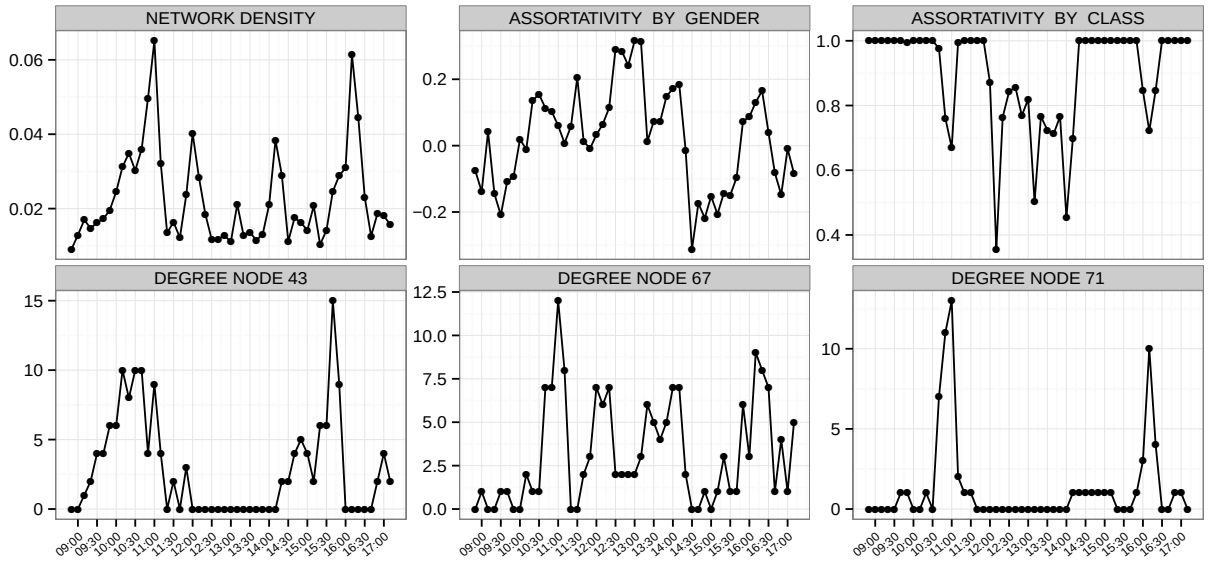


FIGURE 1.3: Time-varying observed network summary statistics for the first day of school. Upper panels: global measures. Lower panels: degree of selected nodes.

flexible to capture different types of dynamic changes in the network data. Dynamic changes in connectivity patterns may be influenced by underlying endogenous architectures as well as exogenous factors, such as changing spatial environments and class or gender homophily. Information on class membership and gender are available for all the individuals – except teachers – while approximate changes in spatio-temporal locations are provided for 5 classes out of 10 in Figure 10 of Stehlé et al. (2011). We focus on the students and teachers in these 5 classrooms, leading to a total of $V = 120$ nodes. Data from the first day $A_{t_1}, \ldots, A_{t_n}$ are the focus on inference, while contact networks $A^*_{t_1}, \ldots, A^*_{t_n}$ in the second day are considered to evaluate out-of-sample predictive performance. Each 120×120 adjacency matrix A_{t_i} has entries $A_{t_i[vu]} = A_{t_i[uv]} = 1$ if a face-to-face contact has been recorded between individuals $v = 2, \ldots, V$ and $u = 1, \ldots, v-1$ at time t_i, $i = 1, \ldots, n$, and $A_{t_i[vu]} = A_{t_i[uv]} = 0$, if no contact is observed.

As shown in Figure 1.3, the trajectories of global and node-specific summary measures cycle irregularly between phases characterized by slower and more rapid variations. Flexibly capturing such behavior is important to improve prediction and investigate how dynamic face-to-face interactions relate to specific events, such as school hours, breaks, lunch time and changing environments. Instead of directly including covariate information on gender, class membership and spatial locations in the model, we use these variables to assess the extent to which our model can learn known structure in the data. Current models for dynamic networks typically rely on homogeneity and stationarity assumptions, and hence have difficulties in modeling variation over time in the rate of change in the network. This can have a strong effect on the quality of inferences and predictions, with under-smoothing

during periods of stable contacts and over-smoothing across times of rapid variations. Motivated by face-to-face contact network data and by the need for flexible methods enforcing time-varying smoothness in dynamic network data, we develop a Locally Adaptive DYnamic (LADY) network model that characterizes the time-varying edge probabilities via latent processes, which have time-varying smoothness.

Relevant literature in scalable and flexible inference for dynamic processes

Methodologies for dynamic network inference outlined in Section 1.1.1 have two main drawbacks motivating further modifications to deal with face-to-face dynamic interaction data. Firstly, most of the proposed stochastic processes for network dynamics are insufficiently flexible to accommodate connectivity patterns cycling irregularly between periods of rapid and slow change. Secondly, although available models reduce dimensionality within the network by collapsing higher-order dependencies into lower-dimensional spaces, the proposed computational methods are typically not appropriate for fast forecasting and predictions.

Inappropriately restricting the smoothness of edge probability trajectories to be constant can have a major impact on the quality of inferences and predictions, with over-smoothing during times of rapid change and under-smoothing in correspondence of stable windows. To realistically characterize the face-to-face human interaction data, it is necessary to accommodate time-varying smoothness. Motivated by our application, we additionally look for fast online updating and forecasting procedures. Efficient strategies of this type remain partially unexplored, but are a key to timely prediction of future interactions and appropriate design of policies, such as disease surveillance and outbreak prevention.

There is a wide literature in modeling time-varying trajectories, covering Kalman filter (Kalman, 1960), Gaussian processes (Rasmussen and Williams, 2006), smoothing spline (Hastie and Tibshirani, 1990) and kernel smoothing methods (Silverman, 1984) – among others. Such approaches perform well for slowly-changing patterns with constant bandwidth parameters regulating implicitly or explicitly global smoothness; however, our interest is allowing smoothness to vary locally in continuous time. Possible extensions for local adaptivity include free knot splines (Friedman, 1991), which perform well in simulations but the different strategies proposed to select the number and the locations of knots via stepwise knot selection (Friedman, 1991), Bayesian knot selection (Smith and Kohn, 1996) or MCMC methods (George and McCulloch, 1993), prove to be computationally intractable for moderately large data sets. Zhu and Dunson (2013) recently address previous scalability issues by inducing local adaptivity via nested Gaussian processes (nGP). These processes explicitly model the trajectories' mth order derivatives via GP priors, which are in turn centered on a higher level

GP instantaneous mean that favors time-varying smoothness. Beside providing flexible models for locally adaptive inference in trajectories patterns, nGPs can be reformulated as a state space model (Durbin and Koopman, 2012) which allows implementation of scalable algorithms (Durbin and Koopman, 2002) substantially reducing the computational complexity.

Although nested Gaussian processes have been successfully generalized to characterize complex mean-covariance stochastic processes (Durante et al., 2014), similar proposals are lacking in the dynamic network field. To our knowledge only Snijders (2005) includes a notion of local adaptivity by considering a time-varying rate parameter in his actor-oriented model formulation. However, available algorithms for estimation and inference (Koskinen and Snijders, 2007; Snijders et al., 2010b) apparently face substantial issues in scaling to large time windows and approximate procedures via method of moments (Snijders, 2001) raise questions about accuracy. It is additionally worth noticing how their applications are substantially different than our face-to-face interaction networks in considering time-varying relational data generally observed for less than ten time points, instead of dynamic networks collected at a much finer time scale. In these wider time windows several homogeneity assumptions underlying their formulation may be unrealistic.

Motivated by these issues, we generalize the methodologies developed for data in Section 1.1.1, to realistically analyze dynamic face-to-face human interactions. The proposed procedure aims at enhancing flexibility in modeling of time-varying edge trajectories, while substantially improving scalability of inference and forecasting strategies. This is accomplished by considering a latent space formulation with nGPs to induce variability over time in the rate of change in the network structure. By considering a state space representation of the latent stochastic processes, we reduce the computational burden, while also developing simple procedures for fast forecasting and prediction as well as novel online updating strategies appropriate to streaming networks.

1.2 Motivations underlying population of networks

Current networks are not only dynamic, but also inherently multidimensional. Social actors can interact on different online networking platforms such as Twitter, Facebook and Linkedin – among others – or within the same online social network according to different types of relationship covering friendships, comments, likes, tags (Mankad and Michailidis, 2015). Other notable examples include trade networks among countries with respect to distinct products (De Domenico et al., 2015), protein–protein interactions of different types (De Domenico et al., 2015) and transportation networks arising from various services (De Domenico et al., 2014). From a statistical perspective, these multiple types of relationships induce a multi-layer network representation characterized by multiple adjacency matrices – called layers – which share the same set of nodes, but differ in their edges.

Although the last two decades have been characterized by increasing efforts in developing joint statistical models for multi-layer networks to improve understanding of complex interacting systems (Kivela et al., 2014; De Domenico et al., 2015), the frontier of network science has shifted again towards new multidimensional data. In fact, the pervasiveness of novel technologies currently allow the collection of replicated network data. For example, brain connectivity networks for a group of individuals in a study can be measured via accurate imaging techniques (Craddock et al., 2013); networks of passes among players are routinely collected for each team via tracking systems (Grund, 2012) and air transportations networks are constantly monitored for each airline company (Cardillo et al., 2013). These populations of networks data are substantially different than multi-layer networks in consisting of multiple observations of the same type of network on different statistical units, instead of measurements of different types of relationships on the same set of nodes. Hence, while multi-layer networks require statistical methodologies for joint modeling of multivariate edges, population of networks can be seen as realizations from a common network-valued random variable whose underlying probabilistic generative mechanism represents the focus of inference.

Flexible modeling of network-valued random variables requires substantial generalizations of current methodologies along with novel inference procedures. Motivated by applications to neuroscience and business intelligence data sets, we propose a fundamentally new approach based on defining a generative probabilistic model for replicated network data, while developing novel inference procedures to estimate and test for changes in the network architecture across groups. In Sections 1.2.1–1.2.2 we describe the data sets motivating the proposed methodologies and provide a careful review of the available contributions.

1.2.1 Flexible statistical modeling and inference for connectome data

There has been an increasing focus on using neuroimaging technologies to better understand the neural pathways underlying human behavior, abilities and neuropsychiatric diseases. The primary emphasis has been on relating the level of activity in specific brain regions to phenotypes. Activity measures are available via electroencephalography (EEG) and functional magnetic resonance imaging (fMRI) – among others – and the aim is to produce a spatial map of the locations in the brain across which activity levels display evidence of change with the phenotype.

Most statistical analyses are based on the massively univariate approach (Luo and Nichols, 2003), by which separate tests are performed to detect local variations for each brain region activity variable across phenotypes. These approaches do not consider dependence in activation structures, and face issues with low power when multiple testing corrections – such as the Benjamini and Hochberg (1995) false discovery rate (FDR) control – are employed. Refer to Genovese et al. (2002) for an application of the Benjamini and Hochberg (1995) procedure within the neuroscience field and Leek and Storey (2008), Clarke and Hall (2009) for a discussion of possible drawbacks in high-dimensional data sets with dependent variables. Graphical models for multivariate activity data represent a possible solution which gains power in multiple testing by accounting for specific dependence structures in the brain regions' activity variables. This is typically accomplished by incorporating information on the regions' spatial proximity in the brain (Worsley, 2003; Bowman et al., 2008; Tansey et al., 2014).

Although previous procedures are still object of interest, more recently there has been a paradigm shift in neuroscience away from the above modular approach and towards studying brain connectivity networks and their relationship with phenotypes (Fuster, 2000, 2006). It has been increasingly realized that it is naive to study region-specific activity in isolation, and the overall circuit structure across the brain is a more important predictor of phenotypes (Bressler and Menon, 2010). Brain connectivity data are now available to facilitate this task, with non-invasive imaging technologies providing accurate brain network data at increasing spatial resolution; see Stirling and Elliott (2008), Craddock et al. (2013) and Wang et al. (2014) for an overview and recent developments on brain scanning technologies. A common approach for constructing brain network data is based on the covariance in activity across brain regions estimated from fMRI data. For example, one can define a functional connectivity network from the inverse covariance matrix, with low values of the precision matrix suggesting evidence of conditional independence between pairs of brain regions (Ramsey et al., 2010; Smith et al., 2011; Simpson et al., 2013).

Although functional synchronizations matrices are popular data in the neuroscience field, such networks do not measure anatomical connections made by axonal pathways and hence

caution is required in interpreting results (Bressler and Menon, 2010). This has motivated recent developments in extracting brain structural networks from various MRI technologies, including structural and diffusion tensor imaging (Craddock et al., 2013). These brain imaging techniques map the diffusion of water molecules across biological tissues, rather than collecting brain activity measures specific to regions, thereby providing better candidates to estimate axonal pathways. As directional diffusion of water within the brain tends to occur along white matter tracts, current connectome pre-processing pipelines (Craddock et al., 2013; Roncal et al., 2013) can produce an adjacency matrix A_i for each individual $i = 1, \ldots, n$, with elements $A_{i[vu]} = A_{i[uv]} = 1$ if there is at least one white matter fiber connecting brain regions $v = 2, \ldots, V$ and $u = 1, \ldots, v-1$ in individual i and $A_{i[vu]} = A_{i[uv]} = 0$ otherwise. In our applications $V = 68$ and each node in the network characterizes a specific anatomical brain region according to the Desikan atlas (Desikan et al., 2006), with the first 34 in the left hemisphere and the remaining 34 in the right; see Figure 1.4 for an illustration. Hence, instead of focusing on multivariate activity data – under a modular paradigm – we aim to develop methodologies for network-valued data and take a further step towards improving the current state of art in the cognitive network field. Refer also to Sporns (2013) for a discussion on functional and structural connectivity networks.

Recent studies measure brain networks along with a categorical predictor, typically denoting each subject's membership to one of two possible groups. Examples include presence or absence of a neuropsychiatric disease, rest-stimulus states and evidence of an high or low level of creative cognition. In such studies, there is a need for methods assessing how the brain connectivity structure varies across groups.

The methods developed in this sections are directly motivated by ongoing studies of the neural pathways underlying creative cognition and Alzheimer's disease. In particular, there is focus on obtaining a greater understanding of how the connection structure in the brain varies between low and high creativity individuals as well as learning how the brain architecture is compromised by the Alzheimer's disease. Specifically, in the first data set connectomes A_i, $i = 1, \ldots, n$ are available for $n = 36$ subjects along with a creativity group indicator y_i, with $y_i = 1$ or $y_i = 2$ if subject i has low or high creativity, respectively. The first group comprises 19 subjects and the second 17, with creativity groups defined by the composite creativity index (CCI) (Jung et al., 2010). Alzheimer's data set focuses instead on brain structural connectivity networks A_i for $n = 92$ individuals, with 42 individual in the Alzheimer's disease group $y_i = 2$, and 50 subjects characterizing age-matched cognitively healthy individuals $y_i = 1$.

Statistical methods for analyzing these data sets have lagged far behind the increasingly routine collection of networks in neuroscience studies. Current practice focuses on overly restrictive procedures which fail in flexibly characterizing the richness of the brain network

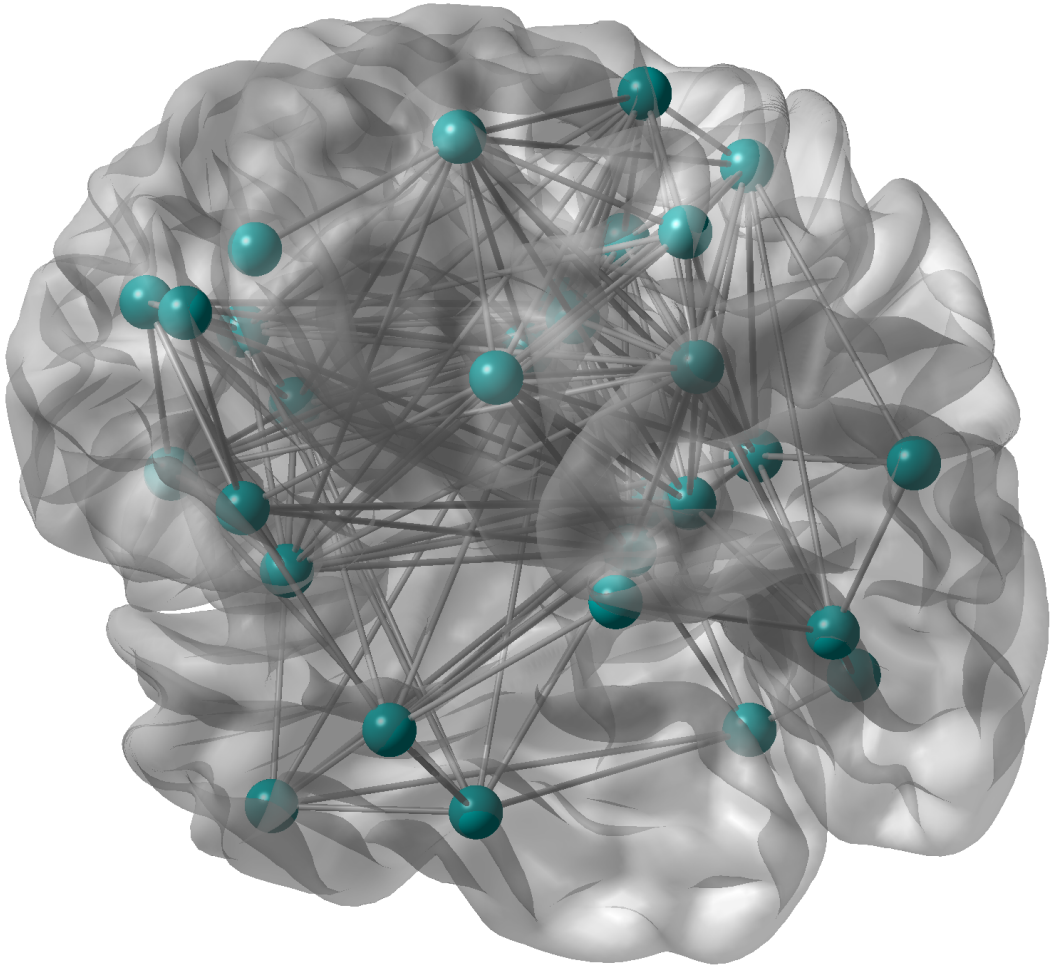


FIGURE 1.4: For a selected subject, graphical representation of his undirected structural brain network for selected brain regions. Node positions are given by their spatial coordinates in the brain.

structure and hence are prone to issues arising from model misspecification. See Arden et al. (2010) for a review of inconsistencies when relating brain networks to creative reasoning. Motivated by these issues and novel applications, we develop a probabilistic generative mechanism to draw tractable and efficient inference directly on the probability mass function associated to a network-valued random variable, rather than on network summary measures or multivariate activity data. In allowing the brain network data to be appropriately analyzed as network-valued, these methods enable substantial improvements in accurately detecting group differences, isolating specific aspects of the network that vary across neurological disorders or behavioral traits, and enhancing performance of predictive models.

Relevant literature in modeling of replicated network data

Much of today's literature focuses on analytic methods for understanding localized brain activity data, yet methodologies for analyzing brain network data A_i is still in its infancy.

Our main aim is to develop techniques to assess whether and how a network-valued random variable generating structural brain networks A_i, $i = 1, \ldots, n$ varies across two groups. In particular, it is of interest to test for global variation in the overall brain network structure across groups, while identifying specific local variations to understand if and which brain connections are changing.

There has been some emphasis in the literature on developing methods for addressing these goals; see Bullmore and Sporns (2009), Stam (2014) and the references cited therein for an overview. The main focus is on reducing each network A_i, $i = 1, \ldots, n$ to a vector of summary statistics $\theta_i = (\theta_{i1}, \ldots, \theta_{ip})^{\mathrm{T}}$ and then applying standard procedures such as the multivariate analysis of variance (MANOVA) to test for variations of these vectors across groups. Summary statistics are commonly chosen to represent global network characteristics of interest, such as the number of connections, average path length and clustering coefficient (Rubinov and Sporns, 2010). Similar procedures have been recently employed in exploring the relation between the brain network and neuropsychiatric diseases, such as Parkinson's (Olde Dubbelink et al., 2014) and Alzheimer's (Daianu et al., 2013), but analyses are sensitive to the chosen network topological measures, with substantially different results obtained for different types of summary statistics. Simpson et al. (2011) and Simpson et al. (2012) improve choice of network summary statistics via a data driven procedure which exploits exponential random graph models (Holland and Leinhardt, 1981; Frank and Strauss, 1986; Wasserman and Pattison, 1996; Robins et al., 2007a) and related validation procedures (Hunter et al., 2008a,b) to detect the topological measures that better characterize the observed networks. Although this is a valuable procedure, inference is still available only on the scale of the network summary statistics, which typically discards important information about the brain connectivity architecture that may crucially explain differences among groups.

An alternative approach is to avoid discarding information by separately testing for differences between groups in each edge probability while adjusting the significance threshold for multiple testing via FDR control. As there are $V(V-1)/2$ pairs of brain regions under study – with $V = 68$ using the Desikan atlas (Desikan et al., 2006) – the number of tests is substantial. Such massively univariate approaches do not exploit network information, leading to low power (Fornito et al., 2013), and underestimating the variations of the brain connections across groups. Recent proposals try to gain power by replacing the common Benjamini and Hochberg (1995) approach, with thresholding procedures that account for the network structure in the data (Zalesky et al., 2010). However, such approaches require careful interpretation, while being highly computationally intensive, requiring permutation testing and choice of suprathreshold links. Instead of controlling FDR thresholds, Scott et al. (2014) gain power in multiple testing by explicitly using auxiliary data – such as spatial proximity – to inform the posterior probability that specific pairs of nodes interact differently across groups or with respect to a baseline. Ginestet et al. (2014) focus instead on assessing evidence of global

changes in the brain structure by testing for group differences in the expected Laplacians.

Scott et al. (2014) and Ginestet et al. (2014) substantially improve the state of art in local and global hypothesis testing for network data, respectively, but are characterized by a similar key issue, motivating our methodology. Specifically, previous procedures test for changes across groups in marginal (Scott et al., 2014) or expected (Ginestet et al., 2014) structures associated to a much complex network-valued random variable, and hence cannot detect variations in the probabilistic generative mechanism that go beyond their focus. Similarly to much simpler settings, substantially different probability mass functions for a network-valued random variable can have equal expectation or induce the same marginal distributions – characterized by the edge probabilities. Hence, previous procedures are expected to fail in scenarios where the changes in the network-valued random variable are related to more complex functionals. Model misspecification can have a major effect on the quality of inference (Deegan, 1976; Begg and Lagakos, 1990; DiRienzo and Lagakos, 2001), providing biased and inaccurate conclusions.

In order to avoid the previous issues it is fundamental to define a statistical model which is sufficiently flexible to accurately approximate any probabilistic generative mechanism underlying the observed data. We address this goal by developing a fully generative Bayesian joint modeling approach for the data (y_i, A_i), $i = 1, \ldots, n$, which explicitly models the networks instead of reducing data to summary measures prior to statistical analysis, while avoiding misspecification issues in testing on changes in the brain network across groups.

Current practice for inference on populations of networks either conducts separate analyses for each A_i to extract local and global measures (Hagmann et al., 2008) or applies standard network analyses – outlined in the Introduction – after averaging $A_1, \ldots, A_n$ (Scheinerman and Tucker, 2010). These approaches fall short of addressing our interest in estimating the population distribution of adjacency matrices to efficiently infer common structures and individual – or group – differences in the architecture of interconnections in the brain; in fact, such differences are poorly understood but are thought to provide important drivers of variability in cognitive traits and disorders (Mueller et al., 2013). As the replicated network data arise from measurements of the same type of network on different individuals, it is appealing to develop a probabilistic generative mechanism, which efficiently borrows information across observed networks, while allowing networks with similar connectivity patterns to cluster close together. Individuals within the same class in terms of their brain network architecture may also have similar cognitive abilities or disorders; see e.g, Stam (2014).

The literature on multi-layer networks considers the case in which replicated network data arise from measurements of different types of relationships on the same set of nodes. Gollini and Murphy (2013) generalize Hoff et al. (2002) allowing the network density parameter to be layer-specific, while forcing the latent space to be shared across layers. Salter-Townshend

and McCormick (2013) consider a layer-specific latent space representation, while estimating dependence across observed networks. The resulting model is heavily parametric, inducing strong constraints on the individual network structure as well as the dependence between different layers; refer to Kivela et al. (2014) for a broader overview of statistical methodologies for multi-layer networks. Additionally – as previously discussed – multi-layer network data are fundamentally different from replicated network data in consisting of measurements of different types of relationships on the same set of nodes instead of measurements of the same type of network on different statistical units. In fact, our goal is not on joint modeling of multivariate edges, but on flexibly characterizing the probability mass function for a network-valued random variable.

An alternative to define a nonparametric model for the distribution of the random variable $\mathcal{A}$ generating networks A_i, $i = 1, \ldots, n$ is to rely on nonparametric models for multivariate binary data. In particular, a model for the probability mass function of the multivariate Bernoulli vector of edges between pairs of nodes, such as those in Dunson and Xing (2009) and Zhou et al. (2014), automatically induces a model for the distribution of $\mathcal{A}$. However, as the length of this binary vector is quadratic in the number of nodes, models that do not exploit the network structure for dimensionality reduction are expected to have poor performance when the number of nodes is moderate to large.

We instead explicitly consider network information in our model formulation, allowing testing on the association between the connectivity architecture and the categorical predictor, while borrowing information across subjects in learning the network structure. This is accomplished by factorizing the joint pmf for the random variable generating data (y_i, A_i), $i = 1, \ldots, n$ as the product of the marginal pmf of the categorical predictor and the conditional pmf for the network-valued random variable given the group membership defined by the categorical predictor. By modeling the collection of group-dependent pmfs for the network-valued random variable via a flexible mixture of low-rank factorizations with group-specific mixing probabilities, we develop a simple test for global variations in the entire distribution of the network-valued random variable rather than focusing only on given functionals. Differently from Ginestet et al. (2014), our procedure additionally incorporates simple local testing for changes in edge probabilities across groups, in line with Scott et al. (2014) methods – which in turn do not consider global tests. By explicitly borrowing strength within the network via matrix factorization representations we intrinsically control for multiplicity in our local multiple tests and substantially improve power compared to standard FDR control procedures.

Although being specifically motivated by neuroscience applications, previous methodologies apply to broader relational settings characterized by replicated network data and focused on flexible modeling of changes in network structures across categorical variables. This is the

case of the data set outlined in the next Section 1.2.2, considering a complex business intelligence problems associated to hierarchical cross-selling strategies.

1.2.2 Joint modeling of mixed domain data for cross-selling of products

Increasing business competition and market saturation have led companies to progressively shift the focus of their marketing strategies from the acquisition of new customers to an increased penetration of their customer base. Targeting existing customers via cross-selling services instead of attracting new ones, provides a more effective strategy for the growth of the company and additionally enhances customer retention by increasing the switching costs (Kamakura et al., 1991). As a consequence, mono-product customers buying a single product from a company, represent a key segment of the customer base and companies are naturally interested in expanding these customers buying behavior to additional products.

Cross-sell and up-sell strategies have been widely studied in marketing and business statistics; see e.g. Azzalini and Scarpa (2012). Common practice focuses on identifying shared acquisition patterns of products by customers, based on their ownership data. A first effort in addressing this aim can be found in Kamakura et al. (1991), where a latent trait model is presented for the probability that a customer would buy a particular product, based on its ownership of other products. Kamakura et al. (2003) combines instead data from a customer database with information from a survey to make probabilistic predictions of ownerships of products. Another approach is given by Verhoef and Donkers (2001) who define a multivariate probit model to predict the potential value of a current customer, and propose a two-by-two segmentation to create a better basis for customer specific strategies. Instead, Thuring (2012) develops a multivariate credibility method to identify an expected profitable set of customers for cross-selling, by estimating a customer specific latent risk profile, using claims as additional information; refer also to Thuring et al. (2012) and Kaishev et al. (2013) for recently developed cross-selling strategies and for a general overview on available methodologies.

Previous proposals exploit different sources, including customer demographics and survey data, to estimate co-subscription probabilities among pairs of products for each customer in a single agency. Differently from this setting, we do not observe customer demographic data for a single agency, but monitor mono-product customer preferences along with co-subscription networks among $V = 15$ products for $n = 130$ agencies operating in the Italian insurance market. Customer relationship management is becoming increasingly important to effectively operate in the insurance market. This sector is mostly stable in developed countries, and rising customer expectations, along with tight competition among top corporations and low growth potentials, force companies to efficiently exploit their database to create, manage and maintain their portfolio of profitable customers (Matiş and Ilieş, 2014).

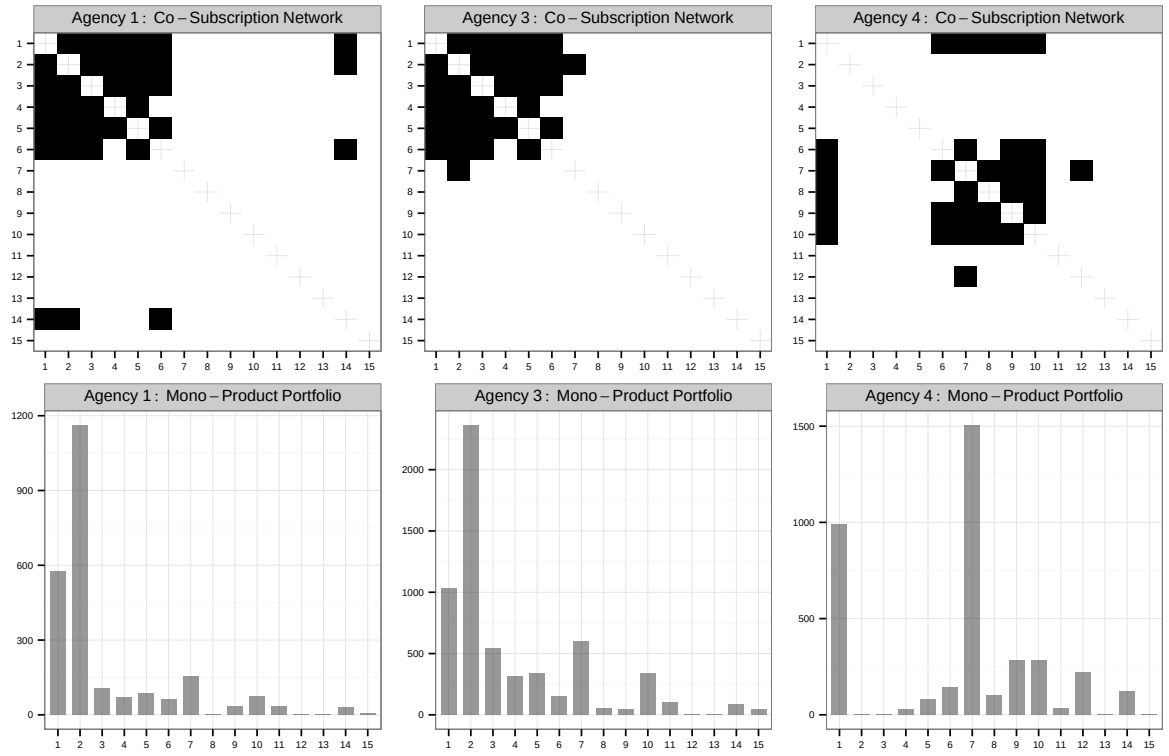


FIGURE 1.5: For selected agencies. Upper panel: observed co-subscription networks A_i. Black refers to an edge, white to a non-edge. Lower panel: total number of mono-product customers for each product $v = 1, \ldots, 15$ based on data d_{ij}, $j = 1, \ldots, n_i$.

We observe preference data $d_{ij} \in \{1, \ldots, V\}$ denoting the product subscribed to by mono-product customer $j = 1, \ldots, n_i$ within agency $i = 1, \ldots, n$. Multi-product customer data are available via a $V \times V$ symmetric adjacency matrix A_i, with $A_{i[vu]} = A_{i[uv]} = 1$ if more than 5% of customers of agency i subscribe to both products $v = 2, \ldots, V$ and $u = 1, \ldots, v-1$ and $A_{i[vu]} = A_{i[uv]} = 0$ otherwise. The 5% threshold is used to focus on pairs of products involving sufficient number of customers; in our application, agencies have $1{,}700$ multi-product clients on average, so that products with edges between them have at least 85 customers subscribing to both products. Refer to Figure 1.5 for an illustrative example.

Each agency can define appropriate cross-selling strategies by exploiting its co-subscription network A_i to estimate the propensity of a customer who subscribed to product $v = 1, \ldots, V$ to additionally buy $u \neq v$. This leads to V different cross-selling strategies $q_{i1}, \ldots, q_{iV}$, with q_{iv} defining which additional product $u \neq v$ is the best offer to currently mono-product customers subscribed to v in agency i, with $u = \text{argmax}_u\{\text{pr}(\mathcal{A}_{i[vu]} = 1) : u \neq v\}$. Efficiently targeting advertising by offering customers the product mostly complementary to their current choice can substantially improve performance relative to untargeted advertising, while increasing satisfaction and reducing churn effects due to frequent and pointless cross-selling attempts (Kamakura et al., 2003). Satisfied customers are a key to enhancing positive word-of-mouth communication and are less sensitive to competing brands and price (Matiş and Ilieş, 2014).

The effectiveness at increasing the number of multi-product customers in agency i depends not only on the tendency for customers with v to also subscribe to u, quantified by $\text{pr}(\mathcal{A}_{i[vu]} = 1)$, but also the proportion of mono-product customers with v, defined by $p_i(v) = \text{pr}(d_{ij} = v)$. If $p_i(v)$ is low, then strategy q_{iv} targets a small portion of the customer base of agency i, and hence has a low ceiling on effectiveness. To take into account the role of $p_i(v)$, we associate each strategy q_{iv} with a performance indicator $e_{iv} = p_i(v)\text{max}\{\text{pr}(\mathcal{A}_{i[vu]} = 1) : u \neq v\}$, for each $v = 1, \ldots, V$ and $i = 1, \ldots, n$. Strategies with a high e_{iv} will target a sizable proportion of the available customers for that agency with advertising for a new product likely to be appealing to them.

In defining and evaluating cross-sell strategies, there are two important issues to take into consideration. Firstly, we are faced with statistical error in estimating the components underlying strategies q_{iv} and indicators e_{iv}, for each $v = 1, \ldots, V$ and $i = 1, \ldots, n$; this is a particular problem in estimating $\text{pr}(\mathcal{A}_{i[vu]} = 1)$ due to data sparsity. The second issue is that it is important to take into account the fact that administrative overhead can be reduced by using the same strategy for different agencies within the same company. For groups of agencies having sufficiently similar customer bases, an identical strategy can be used to reduce administrative cost without decreasing effectiveness. Motivated by this notion, we propose to address both the statistical error and administrative overhead issue through clustering of agencies according to the parameters characterizing their customer bases, and then administrating the same strategy to all agencies within a cluster.

As suggested by Figure 1.5, it is reasonable to expect agencies offering the same service to exhibit clusters, corresponding to common patterns in the composition of their mono-product portfolio and co-subscription behavior. Efficient detection of such clusters allows adaptive reduction of the total number of strategies to be devised from $q_{i1}, \ldots, q_{iV}$, $i = 1, \ldots, n$ to $q_{y1}, \ldots, q_{yV}$, $y = 1, \ldots, K < n$, with each group-specific strategy maintaining its effectiveness in targeting similar agencies. This higher level targeting and profiling represents a key to balance the need of the company to reduce costs and the importance of providing agencies with effective strategies that account for their specific structure. Providing agencies with sets of strategies suitably related with their structure is further important to increase their trust in the company and improve synergy.

We address this goal by developing a Bayesian hierarchical model, which adaptively associates shared strategies $q_{y1}, \ldots, q_{yV}$ and performance indicators $e_{y1}, \ldots, e_{yV}$ to groups of agencies characterized by a similar mono-product portfolio and co-subscription behavior. Each group-specific set of cross-sell strategies $q_{y1}, \ldots, q_{yV}$ is devised by learning group-specific propensity patterns among pairs of products from co-subscription networks of multi-product customers. Joining this information with the estimated group-specific distribution of mono-product customers across products, performance indicators $e_{y1}, \ldots, e_{yV}$ are constructed. To

our knowledge, this is the first approach in the literature that considers a two-level cross-sell segmentation of the customer base, which clusters agencies with similar client portfolio and profiles mono- and multi-product buying behavior within each group to define cross-sell strategies and related performance indicators.

Relevant literature in joint modeling of mixed domain data

There is an increasing statistical literature on joint modeling and co-clustering of mixed domain data. Most available procedures focus on learning the dependence between a univariate response variable and an object predictor, typically characterized by a function. Bigelow and Dunson (2009) favor clustering among predictor trajectories, with each cluster associated to a specific offset in a generalized linear model for the response variable. Although providing an appealing procedure for the sake of interpretability and inference, their model may lack flexibility in constraining predictor and response groups to be the same. This may require the introduction of many clusters to appropriately characterize the joint distribution of the mixed data, reducing the performance in estimating cluster-specific components and providing a biased overview of the underlying grouping structure.

Dunson et al. (2008) address the previous issue by modeling the conditional distribution of the response within each functional cluster via a cluster-dependent mixture representation, rather than considering only a cluster-specific offset in the conditional expectation. In improving flexibility via dependent mixture modeling, they can estimate more reliable clusters which better identify the underlying grouping structure, rather then characterizing the lack of fit of the model formulation; see also Banerjee et al. (2013) for a recent overview of this topic and additional methods.

Although we are similar to previous methods in looking for flexible and accurate joint modeling and co-clustering procedures for mixed domain data, our motivating data set is substantially different in considering categorical mono-product customer choices and network-valued co-subscription data. Flexible modeling of the conditional distribution of a network-valued random variable is still an ongoing issue, which requires careful representations in order to borrow information across edges, reduce the dimensionality and maintain flexibility in characterizing its conditional distribution.

We address these issues by exploiting previous methodologies for brain networks to propose a cluster-dependent mixture of low-rank factorizations, which allows the distribution of the co-subscription networks to flexibly change across clusters via cluster-specific mixing probabilities, while borrowing information across agencies in learning the shared mixture components. Considering cluster dependence only in the mixing probabilities allows further

dimensionality reduction, while providing simple and efficient computational methods. Differently from Dunson et al. (2008), we additionally avoid fixing the total number of clusters, but instead learn this key quantity from our data via a Chinese restaurant process prior for the cluster assignments.

Chapter 2

Dynamic networks

2.1 Nonparametric Bayes modeling of dynamic networks

Motivated by the applied problems outlined in Section 1.1.1, we dynamically model binary relational matrices by embedding the nodes in a low-dimensional latent Euclidean space, with their coordinates evolving in continuous time via Gaussian processes and edge probabilities constructed via a logistic mapping function. Posterior computation is available via a simple Gibbs sampler leveraging the recently developed Pólya-gamma data augmentation. We provide theoretical results on model flexibility, and illustrate its performance via simulation experiments and application to international relationships data.

2.1.1 Dynamic latent space model

Let A_{t_i} denote the adjacency matrix characterizing the undirected network with no self-relations, observed at the generic time $t_i \in \Re^+$. As self-relationships are not of interest and A_{t_i} is symmetric, we model $A_{t_1}, \ldots, A_{t_n}$ by defining a stochastic process for $\mathcal{L}(A_{t_1}), \ldots, \mathcal{L}(A_{t_n})$, with $\mathcal{L}(A_{t_i}) = (A_{t_i[21]}, A_{t_i[31]}, \ldots, A_{t_i[V1]}, A_{t_i[32]}, \ldots, A_{t_i[V2]}, \ldots, A_{t_i[V(V-1)]})^{\mathrm{T}}$ the vector encoding the lower triangular elements of A_{t_i}, which uniquely characterize the network as $A_{t_i[vu]} = A_{t_i[uv]}$ for every $v = 2, \ldots, V$, $u = 1, \ldots, v-1$ and $t_i = t_1, \ldots, t_n$. As a result, $\mathcal{L}(A_{t_i})$ is a vector of binary elements $\mathcal{L}(A_{t_i})_l \in \{0, 1\}$, encoding the presence or absence of an edge among the lth pair of nodes at time t_i for each $l = 1, \ldots, V(V-1)/2$.

Based on previous notation, developing a probabilistic representation for a sequence of time-varying undirected networks, translates into statistical modeling of a multivariate time series $\mathcal{L}(A_{t_1}), \ldots, \mathcal{L}(A_{t_n})$ arising from dynamic monitoring of $V(V-1)/2$ binary variables for n times. However, in accomplishing this goal it is important to explicitly account for the special structure of our data. Specifically, the key difference between a general unstructured

multivariate time series and our dynamic vectors of edges is that the observed networks are potentially characterized by specific underlying patterns – such as transitive relations, community structures and k-stars – which induce dependence among edges at each time t_i. As a result, by carefully accommodating the network structure in dynamic modeling of $\mathcal{L}(A_{t_1}), \ldots, \mathcal{L}(A_{t_n})$, one might efficiently borrow information within each $\mathcal{L}(A_{t_i})$ and across time, while reducing dimensionality and infer specific network properties along with their dynamic changes.

Consistently with previous discussion, we assume observed data $\mathcal{L}(A_{t_1}), \ldots, \mathcal{L}(A_{t_n})$ as n snapshots of a continuous latent process $\{\mathcal{L}(\mathcal{A}_t) : t \in \mathbb{T} \in \Re^+\}$ over a possibly unequally spaced time grid $t_1, \ldots, t_n$. Letting

$$\mathcal{L}(\mathcal{A}_t)_l \mid \pi_l(t) \sim \text{Bern}\left\{\pi_l(t)\right\}, \tag{2.1}$$

independently for each pair of nodes $l = 1, \ldots, V(V-1)/2$ and $t \in \mathbb{T}$, we aim to define a prior Π_π for the collection of dynamic edge probability vectors $[\pi(t) = \{\pi_1(t), \ldots, \pi_{V(V-1)/2}(t)\}^{\mathrm{T}} : t \in \mathbb{T}]$ with the goals being to obtain a provably flexible specification, maintain simple computations, perform dimensionality reduction to scale to moderately large V, allow missing values, accommodate observations over unequally spaced time grids and allow predictions including a measure of predictive uncertainty.

We construct each $\pi_l(t) \in (0,1)$ via a monotonic increasing link function $g(\cdot) : \Re \to (0,1)$ mapping a latent similarity measure among the lth pair of nodes at time t, $S_l(t) \in \Re$, into the probability space. We choose $g(\cdot)$ to be the logistic distribution function, obtaining

$$\text{E}\left\{\mathcal{L}(\mathcal{A}_t)_l \mid \pi_l(t)\right\} = \pi_l(t) = \left\{1 + e^{-S_l(t)}\right\}^{-1} \quad l = 1, \ldots, V(V-1)/2, \;\; t \in \mathbb{T}. \tag{2.2}$$

Without further assumptions on $S_l(t)$, one needs to model $V(V-1)/2$ stochastic processes separately – one for each time-varying similarity measure $S_l(t)$, for $l = 1, \ldots, V(V-1)/2$. In order to reduce dimensionality and account for the network structure among the nodes for every t, we express the similarity measures $S(t) = \{S_1(t), \ldots, S_{V(V-1)/2}(t)\}^{\mathrm{T}}$ as a quadratic combination of a set of node-specific coordinates in a latent space. Specifically, focusing on the lth pair, corresponding to nodes v and u, $v > u$, we let

$$S_l(t) = \mu(t) + X_v(t)^{\mathrm{T}} X_u(t) = \mu(t) + \sum_{r=1}^{R} X_{vr}(t) X_{ur}(t), \tag{2.3}$$

for every $l = 1, \ldots, V(V-1)/2$ and $t \in \mathbb{T}$, where $X_v(t) = \{X_{v1}(t), \ldots, X_{vR}(t)\}^{\mathrm{T}} \in \Re^R$ and $X_u(t) = \{X_{u1}(t), \ldots, X_{uR}(t)\}^{\mathrm{T}} \in \Re^R$ are the vectors of latent coordinates for nodes v and u at time t, respectively, while $\mu(t) \in \Re$ is a baseline trajectory centering the latent similarity process. According to (2.1)–(2.3), nodes with latent coordinates in the same direction will be

more similar and hence will have an higher edge probability. Recalling our motivating application on international relationships data, this construction has an appealing interpretation. In particular, each country is assigned a multifaceted latent position, which can be seen as representing its view on different debated topics or international policies. Countries having similar positions in the different attributes, both positive or negative, will be more likely to cooperate than countries with opposite positions. The similarity – or dissimilarity – will be higher the stronger the positions in the same direction – or opposite direction. Moreover factorization (2.3) reduces dimensionality from a $V(V-1)/2$ stochastic processes on the edge probabilities to $V \times R$ latent trajectories – typically $R \ll V$ – and one baseline process. In matrix form, equation (2.3) can be rewritten as

$$S(t) = \mu(t)1_{V(V-1)/2} + \mathcal{L}(X(t)X(t)^{\mathrm{T}}), \quad t \in \mathbb{T}, \tag{2.4}$$

where $1_{V(V-1)/2} = (1, \ldots, 1)^{\mathrm{T}}$, $X(t) \in \Re^{V \times R}$ defines the matrix of the R latent coordinates for the V nodes at time t and $\mathcal{L}(\cdot)$ is again the operator vectorizing the lower triangular elements of $X(t)X(t)^{\mathrm{T}}$, so that $\mathcal{L}(X(t)X(t)^{\mathrm{T}}) = \{X_2(t)^{\mathrm{T}}X_1(t), X_3(t)^{\mathrm{T}}X_1(t), \ldots, X_V(t)^{\mathrm{T}}X_{V-1}(t)\}^{\mathrm{T}}$. The factorization (2.4) is not unique. For example, if $\mu(t) = 0$ and letting $\tilde{X}(t) = X(t)Q$, with Q an $R \times R$ orthogonal matrix, then $\tilde{X}(t)\tilde{X}(t)^{\mathrm{T}} = X(t)QQ^{\mathrm{T}}X(t)^{\mathrm{T}} = X(t)X(t)^{\mathrm{T}}$. If one is interested in inference on the latent coordinates matrix $X(t)$, identifiability can be ensured via restrictions (Bollen, 1989) or Procrustean transformations (Hoff et al., 2002). However, since we instead focus our inferences on the trajectories of the latent similarities $S(t)$ and the edge probability vectors $\pi(t)$, we follow Ghosh and Dunson (2009) in avoiding identifiability constraints, as they are not necessary to ensure identifiability of $S(t)$ and $\pi(t)$.

Before considering prior specification, it is important to characterize the class of edge probability vectors $\pi(t)$ which can be represented as in (2.2) with latent similarities factorized as in (2.3). In fact, our model considers a latent space approach to network analysis assuming edges as conditionally independent given the corresponding edge probabilities and aims at accommodating and learning network structures by careful modeling of $\pi(t)$ via (2.2)–(2.3). As discussed in the Introduction and in Section 1.1.1, the shared dependence on a common set of node-specific latent coordinates can induce rich dependence structures and accommodate recurring network properties; see for example Hoff et al. (2002); Hoff (2008); Krivitsky et al. (2009) and Hunter et al. (2012). Although previous results are promising, it is important to develop formal theory in order to assess at which extent our formulation is enough general to accomplish previous goals. Theorem 2.1 and Corollary 2.2 state that for R sufficiently large, any edge probability vector $\pi(t)$ has representation (2.2)–(2.3).

Theorem 2.1. *For any $S(t) \in \Re^{V(V-1)/2}$ and $t \in \mathbb{T}$ there exist $\{X(t), \mu(t)\} \in \Re^{V \times R} \times \Re$ such that $S(t) = \mu(t)1_{V(V-1)/2} + \mathcal{L}(X(t)X(t)^{\mathrm{T}})$ for some R.*

Proof. Assume without loss of generality $\mu(t) = 0$. As there exist infinitely many $V \times V$ positive semidefinite matrices having lower triangular elements $S(t)$, let $\Xi(t)$ be one of these matrices such that $\mathcal{L}(\Xi(t)) = S(t)$. Letting $\tilde{R}$ be the rank of $\Xi(t) = \tilde{X}(t)\tilde{\Lambda}(t)\tilde{X}(t)^{\mathrm{T}}$, with $\tilde{\Lambda}(t)$ the diagonal matrix with the $\tilde{R}$ positive eigenvalues of $\Xi(t)$ and $\tilde{X}(t) \in \Re^{V \times \tilde{R}}$ the matrix with the corresponding eigenvectors, Theorem 2.1 holds after defining $X(t) = \{\tilde{X}(t)\tilde{\Lambda}(t)^{1/2}\ 0_{V\times(R-\tilde{R})}\}$. □

Corollary 2.2. *Any edge probability vector $\pi(t) \in (0,1)^{V(V-1)/2}$ admits representation* (2.2) *for every $t \in \mathbb{T}$, with latent similarities factorized as in* (2.3) *for some R.*

Proof. The proof follows immediately from Theorem 2.1 and from the fact that the element-wise mapping from $S_l(t)$ to $\pi_l(t)$ is one-to-one continuous for each $l = 1, \ldots, V(V-1)/2$. □

This ensures that our specification is sufficiently flexible to characterize any true generating process, and hence can be viewed as nonparametric given sufficiently flexible priors for the components.

2.1.2 Prior specification and theoretical properties

We specify independent prior distributions Π_X and Π_μ for $X_{\mathbb{T}} = \{X(t) : t \in \mathbb{T}\}$ and $\mu_{\mathbb{T}} = \{\mu(t) : t \in \mathbb{T}\}$ to induce a prior Π_π for $\pi_{\mathbb{T}} = \{\pi(t) : t \in \mathbb{T}\}$ through (2.2) and (2.3). This prior is defined to have large support, favor simple and efficient computation, allow missing values, induce a continuous time specification, and allow adaptive shrinkage towards lower-dimensional representations. Bhattacharya and Dunson (2011) proposed an approach for Bayesian shrinkage of the number of latent factors in a model for a single covariance matrix, and we extend their approach from independent Gaussian latent factors to Gaussian process latent factors. In particular, we let

$$X_{vr}(\cdot) \sim \mathrm{GP}(0, \lambda_r c_X), \tag{2.5}$$

independently for $v = 1, \ldots, V$ and $r = 1, \ldots, R$, with $c_X(t_i, t_j) = \exp\{-\kappa_X(t_i - t_j)^2\}$. We focus on the squared exponential correlation function in our applications to enforce smoothness in analyzing cooperation relationship data, but more elaborate choices can be made to allow cyclic trends, non-stationarity and other features. Recalling Rasmussen and Williams (2006), assumption (2.5) implies the following joint prior for the node-specific latent coordinates at the time grid $t_1, \ldots, t_n$ on which networks $\mathcal{L}(A_{t_1}), \ldots, \mathcal{L}(A_{t_n})$ are observed

$$\{X_{vr}(t_1), \ldots, X_{vr}(t_n)\}^{\mathrm{T}} \sim \mathrm{N}_n(0, \lambda_r K_X),$$

independently for $v = 1, \ldots, V$ and $r = 1, \ldots, R$, where the covariance matrix K_X has elements $K_{X[ij]} = \exp\{-\kappa_X(t_i - t_j)^2\}$ and λ_r represents a further scaling effect so that when $\lambda_r \approx 0$ the latent coordinates trajectories for dimension r collapse around the zero mean function. Hence to favor adaptive shrinkage we look for an hyperprior Π_λ for the vector of scaling parameters $\lambda = (\lambda_1, \ldots, \lambda_R)^{\mathrm{T}}$ that adaptively deletes redundant latent space dimensions which are not required to characterize the dynamic edge probability vectors according to the observed data. To accomplish this goal we adapt Bhattacharya and Dunson (2011) proposal to our setting by letting $\lambda \sim \text{MIG}(a_1, a_2)$, with $\text{MIG}(a_1, a_2)$ denoting the multiplicative inverse gamma distribution

$$\lambda_r = \prod_{m=1}^{r} \frac{1}{\vartheta_m}, \quad \vartheta_1 \sim \text{Ga}(a_1, 1), \ \vartheta_{m>1} \sim \text{Ga}(a_2, 1), \quad r = 1, \ldots, R, \tag{2.6}$$

where $\text{Ga}(a, b)$ denote the gamma distribution with mean a/b and variance a/b^2. Prior (2.6) adaptively penalizes overparameterized representations favoring elements λ_r to be increasingly concentrated towards 0 as r increases for appropriate choice of a_2. The parameter a_1 controls instead the overall level and variability of the entries in λ; see Bhattacharya and Dunson (2011) for further discussion and theoretical properties. To conclude the prior specification, we choose $\mu(\cdot) \sim \text{GP}(0, c_\mu)$, with $c_\mu(t_i, t_j) = \exp\{-\kappa_\mu(t_i - t_j)^2\}$.

To provide insight on the induced prior distribution for the edge probability process, we derive prior moments of the log-odds process, $S_l(t) = \log[\pi_l(t)/\{1 - \pi_l(t)\}]$, conditioning on the shrinkage parameters λ_r to highlight their effect on the prior. Focusing on the log-odds, prior moments have simple form and can be easily derived via straightforward calculations, obtaining

$$\text{E}\{S_l(t) \mid \lambda\} = 0, \quad \text{var}\{S_l(t) \mid \lambda\} = 1 + \sum_{r=1}^{R} \lambda_r^2, \quad \text{cov}\{S_l(t), S_{l^*}(t) \mid \lambda\} = 1,$$

for each fixed time $t \in \mathbb{T}$ and indexes $l = 1, \ldots, V(V-1)/2$ and $l^* = 1, \ldots, V(V-1)/2$, $l^* \neq l$, with covariances across time given by

$$\begin{aligned} \text{cov}\{S_l(t_i), S_l(t_j) \mid \lambda\} &= \exp\{-\kappa_\mu(t_i - t_j)^2\} + \sum_{r=1}^{R} \lambda_r^2 \exp\{-2\kappa_X(t_i - t_j)^2\}, \\ \text{cov}\{S_l(t_i), S_{l^*}(t_j) \mid \lambda\} &= \exp\{-\kappa_\mu(t_i - t_j)^2\} \quad (t_i, t_j \in \mathbb{T}). \end{aligned}$$

A priori the log-odds of an edge has mean zero and the variance increases with the sum of shrinkage parameters, while the covariance between the log-odds for different edges at the same time is fixed at one. When the λ_rs are all close to zero, the correlation between the log-odds for different edges at the same time is close to one and the covariance over time is controlled primarily by κ_μ. As the λ_rs increase, κ_X plays more of a role in controlling the

dependence in the log-odds of a given edge at different times.

An important issue is the support of the induced prior Π_π. Specifically, we are interested in whether there is a positive probability of generating a $\{\pi(t) : t \in \mathbb{T}\}$ that is arbitrarily close to any true $\{\pi^0(t) : t \in \mathbb{T}\}$. Theorem 2.3 states the large support property for Π_S, while Corollary 2.4 provides the same property for Π_π exploiting the continuity of the logistic mapping.

Theorem 2.3. *Let Π_S denote the induced prior on $\{S(t) : t \in \mathbb{T}\}$ based on the specified prior $\Pi_X \times \Pi_\mu$. If $\mathbb{T}$ is compact, then for all element-wise continuous $S^0(t)$ and for every $\epsilon > 0$,*

$$\text{pr}\left\{\sup_{t\in\mathbb{T}} ||S(t) - S^0(t)||_1 < \epsilon\right\} > 0.$$

Proof. Let $B_{\epsilon_0}(t_0) = \{t : |t - t_0| < \epsilon_0\}$ denote an ϵ_0-neighborhood around t_0, with $t_0 \in \mathbb{T}$ and $\epsilon_0 > 0$. Exploiting the compactness of $\mathbb{T}$, for any open cover of ϵ_0-neighborhoods, we can always define a finite subcover such that $\mathbb{T} \subset \cup_{t_0\in\mathbb{T}_0} B_{\epsilon_0}(t_0)$, with $|\mathbb{T}_0| = n$. Hence:

$$\text{pr}\left\{\sup_{t\in\mathbb{T}} ||S(t) - S^0(t)||_1 < \epsilon\right\} = \text{pr}\left\{\max_{t_0\in\mathbb{T}_0} \sup_{t\in B_{\epsilon_0}(t_0)} ||S(t) - S^0(t)||_1 < \epsilon\right\}.$$

Since $\text{pr}\{\max_{t_0\in\mathbb{T}_0} \sup_{t\in B_{\epsilon_0}(t_0)} ||S(t) - S^0(t)||_1 < \epsilon\} > 0$, if and only if $\text{pr}\{\sup_{t\in B_{\epsilon_0}(t_0)} ||S(t) - S^0(t)||_1 < \epsilon\} > 0$, for every $t_0 \in \mathbb{T}_0$, we only need to prove $\text{pr}\{\sup_{t\in B_{\epsilon_0}(t_0)} ||S(t) - S^0(t)||_1 < \epsilon\} > 0$ for each ϵ_0-neighborhood, independently. Using the triangle inequality, a lower bound for this probability is

$$\text{pr}\left\{\sup_{t\in B_{\epsilon_0}(t_0)} ||S^0(t_0) - S^0(t)||_1 + \sup_{t\in B_{\epsilon_0}(t_0)} ||S(t_0) - S(t)||_1 + ||S(t_0) - S^0(t_0)||_1 < \epsilon\right\}, \quad (2.7)$$

and provided that the first term in (2.7) states the continuity property for a deterministic component, which is independent from the second and third events, we can further lower bound the previous probability by

$$\text{pr}\left\{\sup_{t\in B_{\epsilon_0}(t_0)} ||S^0(t_0) - S^0(t)||_1 < \frac{\epsilon}{3}\right\} \text{pr}\left\{\sup_{t\in B_{\epsilon_0}(t_0)} ||S(t_0) - S(t)||_1 < \frac{\epsilon}{3} \mid ||S(t_0) - S^0(t_0)||_1 < \frac{\epsilon}{3}\right\}$$
$$\times \text{ pr}\left\{||S(t_0) - S^0(t_0)||_1 < \frac{\epsilon}{3}\right\}. \quad (2.8)$$

We prove each of the terms in (2.8) in turn.

As every function $S^0_l(\cdot)$ is continuous, following Theorem 4.10 in Rudin (1976), this implies that also $S^0(\cdot) = \{S^0_1(\cdot), \ldots, S^0_{V(V-1)/2}(\cdot)\}^{\text{T}}$ is continuous. As a results, for every $\epsilon/3 > 0$, there exists an $\epsilon_{0,1} > 0$ such that:

$$||S^0(t_0) - S^0(t)||_1 < \frac{\epsilon}{3}, \quad |t - t_0| < \epsilon_{0,1}.$$

Hence, $\text{pr}\left\{\sup_{t\in B_{\epsilon_{0,1}}(t_0)} ||S^0(t_0) - S^0(t)||_1 < \epsilon/3\right\} = 1$.

The second term states the continuity property of $S(t)$ in a neighborhood of t_0, with the conditional event restricting the analysis to the subset of all the realizations of $S(t)$, with $S(t_0)$ lying in a $\epsilon/3$ neighborhood of $S^0(t_0)$. We first prove the continuity property of $S(t)$ in its unrestricted sample space. The continuity in a subset will follow as a consequence.

Given the Gaussian process prior on the elements of $X(\cdot)$, the equation

$$\{X(t)X(t)^{\mathrm{T}}\}_{vu} = \sum_{r=1}^{R} X_{vr}(t)X_{ur}(t), \quad t \in \mathbb{T},$$

represents a finite sum over pairwise products of almost surely continuous functions, implying that also elements in $X(t)X(t)^{\mathrm{T}}$ are almost surely continuous on $\mathbb{T}$. Therefore $S(t) = \mu(t)1_{V(V-1)/2} + \mathcal{L}(X(t)X(t)^{\mathrm{T}})$ is almost surely element-wise continuous on $\mathbb{T}$ since the baseline $\mu(\cdot)$ is itself almost surely continuous given the Gaussian process prior assumption. Therefore, similarly as before, for every $\epsilon/3 > 0$, there exists an $\epsilon^*_{0,2} > 0$ such that

$$\text{pr}\left\{\sup_{t\in B_{\epsilon^*_{0,2}}(t_0)} ||S(t_0) - S(t)||_1 < \frac{\epsilon}{3}\right\} = 1.$$

Since we proved that all realizations from $S(t)$ are continuous in a neighborhood of t_0, the same will be true for the subset of the sample space induced by the condition $||S(t_0) - S^0(t_0)||_1 < \epsilon/3$. Hence for every $\epsilon/3 > 0$, we can always identify an $\epsilon_{0,2} > 0$, such that

$$\text{pr}\left\{\sup_{t\in B_{\epsilon_{0,2}}(t_0)} ||S(t_0) - S(t)||_1 < \frac{\epsilon}{3} \;\middle|\; ||S(t_0) - S^0(t_0)||_1 < \frac{\epsilon}{3}\right\} = 1.$$

To prove the last term, by Theorem 2.1, $\text{pr}\left\{||S(t_0) - S^0(t_0)||_2 < \epsilon/3\right\}$ can be always factorized as

$$\text{pr}\left\{||\mu(t_0) \times 1_{V(V-1)/2} + \mathcal{L}(X(t_0)X(t_0)^{\mathrm{T}}) - \mu^0(t_0) \times 1_{V(V-1)/2} - \mathcal{L}(X^0(t_0)X^0(t_0)^{\mathrm{T}})||_1 < \frac{\epsilon}{3}\right\}, \quad (2.9)$$

with $\{X^0(t_0), \mu^0(t_0)\} \in \Re^{V\times R} \times \Re$ such that $S^0(t_0) = \mu^0(t_0)1_{V(V-1)/2} + \mathcal{L}(X^0(t_0)X^0(t_0)^{\mathrm{T}})$. Using the triangle inequality, a lower bound for (2.9) is

$$\text{pr}\left\{||\mathcal{L}(X(t_0)X(t_0)^{\mathrm{T}}) - \mathcal{L}(X^0(t_0)X^0(t_0)^{\mathrm{T}})||_1 < \frac{\epsilon}{6}\right\} \text{pr}\left[||1_{V(V-1)/2}\left\{\mu(t_0) - \mu^0(t_0)\right\}||_1 < \frac{\epsilon}{6}\right].$$

Based on the support of the Gaussian prior,

$$\text{pr}\left[||1_{V(V-1)/2}\left\{\mu(t_0) - \mu^0(t_0)\right\}||_1 < \frac{\epsilon}{6}\right] = \text{pr}\left\{|\mu(t_0) - \mu^0(t_0)| < \frac{\epsilon}{6V(V-1)/2}\right\} > 0.$$

For studying the first term of the previous decomposition, note that we need to show the full

support of the prior on the space of the vectorized lower triangular elements of a symmetric matrix. Hence it suffices to show that the induced prior on $X(t_0)X(t_0)^{\mathrm{T}}$ assigns positive probability to a neighborhood of every possible $V \times V$ positive semidefinite matrix. Note that, because self-relationships are not of interest, there is no loss of generality in focusing on the space of positive semidefinite matrices, since for every configuration of latent similarities there exist infinitely many positive semidefinite matrices having these quantities as off-diagonal elements. To prove this property, write $X(t_0)X(t_0)^{\mathrm{T}} = \sum_{r=1}^{R} X_r(t_0)X_r(t_0)^{\mathrm{T}}$, where $X_r(t_0) = \{X_{1r}(t_0), \ldots, X_{Vr}(t_0)\}^{\mathrm{T}}$ is distributed, according to our prior specification, as $\mathrm{N}_V(0, \lambda_r I_V)$, implying that $X_r(t_0)X_r(t_0)^{\mathrm{T}} \mid \lambda_r \sim \mathrm{W}_V(\lambda_r I_V, 1)$ independently for all $r = 1, \ldots, R$, where $\mathrm{W}_V(\cdot, \cdot)$ denotes the Wishart distribution. Using the triangle inequality

$$\mathrm{pr}\left\{||X(t_0)X(t_0)^{\mathrm{T}} - X^0(t_0)X^0(t_0)^{\mathrm{T}}||_1 < \frac{\epsilon}{6}\right\} \geq \prod_{r=1}^{R} \mathrm{pr}\left\{||X_r(t_0)X_r(t_0)^{\mathrm{T}} - X_r^0(t_0)X_r^0(t_0)^{\mathrm{T}}||_1 < \frac{\epsilon}{6R}\right\}.$$

Since $X_r^0(t_0)X_h^0(t_0)^{\mathrm{T}}$ is an arbitrary positive semidefinite rank-1 symmetric matrix in $\Re^{V \times V}$, and based on the support of the Wishart distribution

$$\mathrm{pr}\left\{||X_r(t_0)X_r(t_0)^{\mathrm{T}} - X_r^0(t_0)X_r^0(t_0)^{\mathrm{T}}||_1 < \frac{\epsilon}{6R}\right\} > 0 \quad r = 1, \ldots, R.$$

Thus $\mathrm{pr}\left\{||X(t_0)X(t_0)^{\mathrm{T}} - X^0(t_0)X^0(t_0)^{\mathrm{T}}||_1 < \epsilon/6\right\} > 0$ and combining it with the large support property previously proved for the prior on the baseline $\mu(\cdot)$, we obtain

$$\mathrm{pr}\left\{||S(t_0) - S^0(t_0)||_1 < \frac{\epsilon}{3}\right\} > 0.$$

Letting $\epsilon_0 = \min(\epsilon_{0,1}, \epsilon_{0,2})$, with $\epsilon_{0,1}$ and $\epsilon_{0,2}$ defined as above, the proof follows from the positivity of the three probabilities in (2.8), for every $S^0(\cdot)$ and $\epsilon > 0$. □

Corollary 2.4. *Let Π_π the induced prior on $\{\pi(t) : t \in \mathbb{T}\}$ based on the specified prior $\Pi_X \times \Pi_\mu$. If $\mathbb{T}$ is compact, then for all element-wise continuous $\pi^0(t)$ and for every $\delta > 0$,*

$$\mathrm{pr}\left\{\sup_{t \in \mathbb{T}} ||\pi(t) - \pi^0(t)||_1 < \delta\right\} > 0.$$

Proof. Since the elements of $\pi(t)$ are defined as a one-to-one continuous mapping of the elements of $S(t)$ through the function $g(\cdot)$, by definition of continuity we have that for every $\delta > 0$ there exists an $\epsilon > 0$ such that

$$\sup_{t \in \mathbb{T}} ||g\{S(t)\} - g\{S^0(t)\}||_1 = \sup_{t \in \mathbb{T}} ||\pi(t) - \pi^0(t)||_1 < \delta,$$

for all $S(t)$ such that $\sup_{t \in \mathbb{T}} ||S(t) - S^0(t)||_1 < \epsilon$, where $g\{S(t)\}$ means that the function $g(\cdot)$ is applied to every element of $S(t)$. Finally, since by Theorem 2.3 the event $\sup_{t \in \mathbb{T}} ||S(t) - S^0(t)||_1 < \epsilon$ has positive probability, the same holds for $\sup_{t \in \mathbb{T}} ||\pi(t) - \pi^0(t)||_1 < \delta$. □

Theorem 2.3 and Corollary 2.4 provide key results to ensure good performance in our application because without prior support about the true data generating process, the posterior cannot possibly concentrate around the truth.

2.1.3 Posterior computation

Posterior computation is performed by adapting a Pólya-gamma data augmentation scheme for Bayesian logistic regression (Polson et al., 2013); see Choi and Hobert (2013) for results on uniform ergodicity of the algorithm. Letting $y_i \sim \text{Bern}(\pi_i)$, independently with $\pi_i = (1 + e^{-x_i^{\mathrm{T}}\beta})^{-1}$, Polson et al. (2013) show that conditionally on Pòlya-gamma augmented data $\omega_i \mid - \sim \text{PG}(1, x_i^{\mathrm{T}}\beta)$, the contribution to the likelihood for the ith observation is

$$\propto \exp\left[-\frac{\omega_i}{2}\left\{(y_i - 0.5)/\omega_i - x_i^{\mathrm{T}}\beta\right\}^2\right], \quad i = 1, \ldots, n. \tag{2.10}$$

Equation (2.10) is the kernel of a Gaussian distribution for data $(y_i - 0.5)/\omega_i$, with mean $x_i^{\mathrm{T}}\beta$ and variance $1/\omega_i$. Hence, letting $\beta \sim \text{N}_p(b, B)$ be the prior for the coefficients, given Pòlya-gamma augmented data, the Bayesian logistic regression on y_i can be recast in terms of Bayesian linear regression with Gaussian response $(y_i - 0.5)/\omega_i$. This allows a Gibbs algorithm, which alternates between $\omega_i \mid - \sim \text{PG}(1, x_i^{\mathrm{T}}\beta)$ and $\beta \mid - \sim \text{N}_p(\mu_\beta, \Sigma_\beta)$, where $\Sigma_\beta = (X^{\mathrm{T}}\Omega X + B^{-1})^{-1}$, $\mu_\beta = \Sigma_\beta(X^{\mathrm{T}}z + B^{-1}b)$, $z = (y_1 - 1/2, \ldots, y_n - 1/2)^{\mathrm{T}}$ and $\Omega = \text{diag}(\omega_1, \ldots, \omega_n)$.

Recalling model (2.1), with probabilities defined as in (2.2) and latent similarities from (2.3), we develop an efficient Gibbs sampler, which uses Pólya-gamma augmented data and converges to the exact posterior, while avoiding accuracy issues arising from analytic approximations, such as Laplace or variational Bayes. Detailed steps are outlined in Algorithm 1.

Algorithm 1 Gibbs sampler for the dynamic latent space model

---

[1] Sample Pólya-gamma augmented data

for each $l = 1, \ldots, V(V-1)/2$ and $t_i = t_1, \ldots, t_n$ **do**

Update each augmented data $\omega_l(t_i)$ from the full conditional Pólya-gamma

$$\omega_l(t_i) \mid - \sim \text{PG}\left\{1, \mu(t_i) + \mathcal{L}(X(t_i)X(t_i)^{\mathrm{T}})_l\right\}.$$

end for

---

[2] Sample the baseline trajectory $\mu = \{\mu(t_1), \dots, \mu(t_n)\}^{\mathrm{T}}$ from

$$\mu \mid - \sim \mathrm{N}_n \left(\Sigma_\mu \begin{bmatrix} \sum_{l=1}^{V(V-1)/2} \{\mathcal{L}(A_{t_1})_l - 1/2 - \omega_l(t_1)\mathcal{L}(X(t_1)X(t_1)^{\mathrm{T}})_l\} \\ \vdots \\ \sum_{l=1}^{V(V-1)/2} \{\mathcal{L}(A_{t_n})_l - 1/2 - \omega_l(t_n)\mathcal{L}(X(t_n)X(t_n)^{\mathrm{T}})_l\} \end{bmatrix}, \Sigma_\mu \right),$$

with $\Sigma_\mu = \left\{ \mathrm{diag}\left(\sum_{l=1}^{V(V-1)/2} \omega_l(t_1), \dots, \sum_{l=1}^{V(V-1)/2} \omega_l(t_n) \right) + K_\mu^{-1} \right\}^{-1}$, and K_μ the Gaussian process covariance matrix with entries $K_{\mu[ij]} = \exp\{-\kappa_\mu (t_i - t_j)^2\}$.

---

[3] Sample the matrix of coordinates trajectories $X(t_1), \dots, X(t_n)$
for $v = 1, \dots, V$ **do**

Block-sample $\{X_v(t_1), \dots, X_v(t_n)\}$ given $X_{(-v)} = \{X_u(t_i) : u \neq v, t_i = t_1, \dots t_n\}$.

1. Define $X_{(v)} = \{X_{v1}(t_1), \dots, X_{v1}(t_n), \dots, X_{vR}(t_1), \dots, X_{vR}(t_n)\}^{\mathrm{T}}$
2. Define a Bayesian logistic regression with $X_{(v)}$ acting as coefficient vector and having prior, according to GP, $X_{(v)} \sim \mathrm{N}_{n \times R}\{0, \mathrm{diag}(\lambda_1, \dots, \lambda_R) \otimes K_x\}$
3. The Bayesian logistic regression to update $X_{(v)}$ is a follow

$$\mathcal{L}(A)_{(v)} \sim \mathrm{Bern}(\pi_{(v)}) \quad \mathrm{logit}(\pi_{(v)}) = 1_{V-1} \otimes \mu + \tilde{X}_{(-v)} X_{(v)},$$

   with $\mathcal{L}(A)_{(v)}$ obtained by stacking vectors $\{\mathcal{L}(A_{t_1})_l, \dots, \mathcal{L}(A_{t_n})_l\}^{\mathrm{T}}$ for all pairs l having v as a one of the two nodes, and $\pi_{(v)}$ are the corresponding vector of edge probabilities. Finally, $\tilde{X}_{(-v)}$ is the matrix of regressors with entries suitably chosen from $X_{(-v)}$, to reproduce (2.4) for the sub-sample considered.
4. Hence exploiting previous formulation, the Pólya-gamma sampling provides

$$X_{(v)} \mid - \sim \mathrm{N}_{n \times R}\left(\mu_{x_{(v)}}, \Sigma_{x_{(v)}} \right),$$

   with $\Sigma_{x^{(v)}} = \left\{ \tilde{X}_{(-v)}^{\mathrm{T}} \Omega_{(v)} \tilde{X}_{(-v)} + \mathrm{diag}(\lambda_1^{-1}, \dots, \lambda_R^{-1}) \otimes K_x^{-1} \right\}^{-1}$, $\Omega_{(v)}$ a diagonal matrix with the corresponding Pólya-gamma augmented data and mean vector given by $\mu_{x_{(v)}} = \Sigma_{x_{(v)}} \left[\tilde{X}_{(-v)}^{\mathrm{T}} \left\{ \mathcal{L}(A)_{(v)} - \ 1_{V-1} \otimes 1_N 0.5 - \Omega_{(v)}(1_{V-1} \otimes \mu) \right\} \right]$.

end for

---

[4] Sample the gamma quantities defining the shrinkage parameters $\lambda_1, \dots, \lambda_R$

$$\vartheta_1 \mid - \ \sim \ \mathrm{Ga}\left\{ a_1 + \frac{V \times n \times R}{2}, 1 + \frac{1}{2} \sum_{m=1}^{R} \theta_m^{(-1)} \sum_{v=1}^{V} X_{vm}^{\mathrm{T}} K_x^{-1} X_{vm} \right\},$$

$$\vartheta_r \mid - \ \sim \ \mathrm{Ga}\left\{ a_2 + \frac{V \times n \times (R - r + 1)}{2}, 1 + \frac{1}{2} \sum_{m=r}^{R} \theta_m^{(-r)} \sum_{v=1}^{V} X_{vm}^{\mathrm{T}} K_x^{-1} X_{vm} \right\},$$

where $\theta_m^{(-r)} = \prod_{t=1,t\neq r}^{m} \vartheta_t$ for $r = 1, \ldots, R$ and $X_{vm} = \{X_{vm}(t_1), \ldots, X_{vm}(t_n)\}^{\mathrm{T}}$.

In performing posterior computation, we fix R at a conservative upper bound, allowing unnecessary extra dimensions to be effectively removed through posterior distributions for λ_r that are concentrated near zero. The results are not sensitive to R unless R is chosen to be too small, in which case λ_R not concentrated near zero provides evidence that R should be increased.

Given MCMC chains for $\mu(t_1), \ldots, \mu(t_n)$ and $X(t_1), \ldots, X(t_n)$, posterior samples for time-varying latent similarities $S(t_1), \ldots, S(t_n)$ and edge probability vectors $\pi(t_1), \ldots, \pi(t_n)$ can be easily derived by applying equations (2.3) and (2.2), respectively. Our algorithm can also easily handle missing values by adding a further step imputing the unobserved binary edges from their conditional distribution in (2.1) given the current state of the chain.

Previous strategy provides also a useful procedure for forecasting new networks $\mathcal{L}(A_{t_{n+1}})$. Under our Bayesian paradigm, a strategy to obtain one-step-ahead forecasts is to rely on the expectation of the forecasted predictive distribution $\mathrm{E}\{\mathcal{L}(\mathcal{A}_{t_{n+1}}) \mid \mathcal{L}(A_{t_1}), \ldots, \mathcal{L}(A_{t_n})\}$, having elements $\mathrm{E}\{\mathcal{L}(\mathcal{A}_{t_{n+1}}) \mid \mathcal{L}(A_{t_1}), \ldots, \mathcal{L}(A_{t_n})\}_l = \mathrm{E}\{\mathcal{L}(\mathcal{A}_{t_{n+1}})_l \mid \mathcal{L}(A_{t_1}), \ldots, \mathcal{L}(A_{t_n})\}$ defined as

$$\begin{aligned} \mathrm{E}\{\mathcal{L}(\mathcal{A}_{t_{n+1}})_l \mid \mathcal{L}(A_{t_1}), \ldots, \mathcal{L}(A_{t_n})\} &= \mathrm{E}_{\pi_l(t_{n+1})}[\mathrm{E}\{\mathcal{L}(\mathcal{A}_{t_{n+1}})_l \mid \pi_l(t_{n+1})\} \mid \mathcal{L}(A_{t_1}), \ldots, \mathcal{L}(A_{t_n})] \\ &= \mathrm{E}\{\pi_l(t_{n+1}) \mid \mathcal{L}(A_{t_1}), \ldots, \mathcal{L}(A_{t_n})\}, \end{aligned} \tag{2.11}$$

for each $l = 1, \ldots, V(V-1)/2$. Recall also that we use standard font $\mathcal{L}(A)$ to define the observed vectorized adjacency matrix and italics notation $\mathcal{L}(\mathcal{A})$ to denote its associated random variable. Equation (2.11) simply requires the posterior mean of the edge probabilities at time t_{n+1}. Under our model, the posterior distribution of future $\pi(t_{n+1})$ with $t_{n+1} > t_n$ given the observed networks $\mathcal{L}(A_{t_1}), \ldots, \mathcal{L}(A_{t_n})$, can be obtained by simply performing the previous posterior computations adding to the observed dataset $\mathcal{L}(A_{t_1}), \ldots, \mathcal{L}(A_{t_n})$ a new vector $\mathcal{L}(A_{t_{n+1}})$ of missing values and make inference on the posterior distribution for $\pi(t_{n+1})$.

2.1.4 A note on the multiplicative inverse gamma prior

Before assessing model performance in simulations and applications, it is worth considering an in-depth analysis on the properties of the multiplicative inverse gamma prior in (2.6) introduced by Bhattacharya and Dunson (2011). This prior provides a useful building block in the following analyses, but some of the authors statements need clarifications and a careful study of prior properties is required to ensure appropriate use in routine applications.

Bhattacharya and Dunson (2011) are motivated by increasingly high-dimensional problems requiring statistical methodologies which adaptively induce sparsity and automatically

delete redundant components not required to characterize the data. Although shrinkage is implicit in some Bayesian inference procedures (Morris, 1983), it is increasingly common to further enhance adaptive deletion of redundant components via carefully defined priors which are designed to stochastically penalize over-parameterized representations. Other key examples include stick-breaking representation (Sethuraman, 1994) and spike-and-slab priors (Ishwaran and Rao, 2005). These procedures facilitate adaptive shrinkage by defining priors which concentrate increasing mass towards values deleting the effect of parameters associated to growing dimensions of the statistical model. For example, the stick-breaking prior for the weights in a mixture model assigns growing mass around zero for the weights associated to increasing mixture components. Hence, as the number of components grows, their weights tends to concentrate around zero a priori, meaning that increasing components have a decreasing importance in defining the density.

Although these procedures provide key strategies to deal with high-dimensional data, their performance may be sensitive to several choices including model definition, hyperparameters settings and prior specification for other quantities not directly related to shrinkage; refer to Roos and Held (2011) and the references cited therein for a discussion. Indeed, the theoretical analysis of these shrinkage priors is currently object of intense interest. Key questions include, among others, assessing how prior properties and hyperparameter settings guarantee improved theoretical performance of the posterior distribution and whether specific shrinkage priors can accurately recover the true dimensions of the parametric space a posteriori. Early results are available in simple models (Rousseau and Mengersen, 2011), and it is an active area of research to extend these asymptotic results to more general settings.

Although previous questions are of fundamental interest, in making these contributions standard practice, it is first important to provide the researcher with strategies to check whether the prior actually achieves the shrinkage behavior he seeks when using these methods. This property doesn't holds for all the hyperparameters settings, with poor choices leading to completely opposite behaviors which vanish the motivations for the use of a shrinkage prior and do not justify higher computational complexity in performing posterior computation under these methods. Focusing on the shrinkage prior (2.6) we use, its cumulative shrinkage property is clearly not maintained for all the values of the hyperparameters a_1 and a_2. Although Bhattacharya and Dunson (2011) consider hyperpriors on a_1 and a_2 to learn these key quantities from the data, a wide set of statistical models building on their contribution fix such hyperparameters following some of the authors statements on the behavior of (2.6) in relation to a_1 and a_2. In particular Bhattacharya and Dunson (2011) claim that the quantities $1/\lambda_r$ – acting as precision parameters in their statistical model – are stochastically increasing under the restriction $a_2 > 1$, meaning that the cumulative shrinkage behavior of the prior is guaranteed for choices of a_2 greater than 1.

	$1/\lambda_r$			λ_r		
	Q_1	Q_2	Q_3	Q_1	Q_2	Q_3
$a_2 = 1.1$						
$r = 1$	0.29	0.69	1.39	0.72	1.44	3.50
$r = 2$	0.13	0.45	1.28	0.78	2.20	7.50
$r = 3$	0.07	0.30	1.08	0.92	3.35	14.86
$r = 4$	0.04	0.19	0.88	1.13	5.12	28.16

TABLE 2.1: Stochastic behavior of the priors for the model parameters. For $a_2 = 1.1 > 1$, first (Q_1), second (Q_2) and third (Q_3) quartiles for $1/\lambda_r$ and λ_r at increasing dimensions $r = 1, \dots, 4$, when $\lambda = (\lambda_1, \dots, \lambda_4)^{\mathrm{T}} \sim \text{MIG}(1, 1.1)$. Quantities are obtained by 1,000,000 simulated data from prior (2.6).

Although it is true that $\mathrm{E}(1/\lambda_r) = \mathrm{E}(\prod_{m=1}^{r} \vartheta_m) = \prod_{m=1}^{r} \mathrm{E}(\vartheta_m) = a_1 a_2^{r-1}$ increases with r when $a_2 > 1$, such property doesn't necessarily imply stochastic ordering and cumulative shrinkage. Indeed, as shown in Table 2.1, a value of $a_2 = 1.1 > 1$ induces priors on parameters $1/\lambda_r$ which seem stochastically decreasing as r increases. This leads to stochastically increasing distributions on λ_r. Hence, a researcher choosing $a_2 = 1.1$ will obtain a prior with an opposite behavior with respect to the one he seeks when using a multiplicative inverse gamma prior. Beside this, even increasing expectation in $1/\lambda_r$, doesn't necessarily implies decreasing expectation in λ_r and therefore growing shrinkage on average. In fact, $\mathrm{E}(\lambda_r) = \mathrm{E}(\prod_{m=1}^{r} 1/\vartheta_m) = \prod_{m=1}^{r} \mathrm{E}(1/\vartheta_m) = 1/\{(a_1 - 1)(a_2 - 1)^{r-1}\}$. Hence for values $1 < a_2 < 2$ both $1/\lambda_r$ and λ_r increase in expectation with growing dimension. Motivated by these misleading results, we aim to improve the characterization of the multiplicative inverse gamma process and add further insights compared to those available in Bhattacharya and Dunson (2011), in terms of prior properties. This is a key to avoid undesired behaviors similar to those for $a_2 = 1.1$.

Stochastic ordering and shrinkage in the multiplicative inverse gamma prior

Consistently with the previous discussion let us focus on studying the stochastic behavior of the sequence $\lambda_1, \dots, \lambda_R$, with each λ_r, $r = 1, \dots, R$, defined as the cumulative product of r independent inverse gamma random variables $1/\vartheta_1, \dots, 1/\vartheta_r$. Stochastic ordering $\lambda_1 \succeq \dots \succeq \lambda_R$ is an appealing property for the multiplicative inverse gamma prior in facilitating increasing shrinkage as the dimension index r grows. This requires showing that $\mathrm{pr}(\lambda_{r+1} \leq \xi) \geq \mathrm{pr}(\lambda_r \leq \xi)$ for every r and $\xi > 0$, or equivalently that $\mathrm{E}\{g(\lambda_{r+1})\} \leq \mathrm{E}\{g(\lambda_r)\}$ for all the increasing functions $g(\cdot)$, for which expectation exists; refer to page 4 in Shaked and Shanthikumar (2007) .

Unfortunately, as stated in Lemma 2.5, for every $a_2 > 1$ it is always possible to find an increasing function $g^*(\cdot)$ for which $\mathrm{E}\{g^*(\lambda_{r+1})\} > \mathrm{E}\{g^*(\lambda_r)\}$, for every $r = 1, \dots, R$, meaning that stochastic ordering is never met under the multiplicative inverse gamma prior. In proving $\lambda_1 \not\succeq \dots \not\succeq \lambda_R$, let us first derive the quantity $\mathrm{E}(\lambda_r^c)$, for $r = 1, \dots, R$ and $c > 0$. According to (2.6), this requires first $\mathrm{E}(1/\vartheta^c)$, where ϑ is a generic gamma random variable $\vartheta \sim \text{Ga}(a, 1)$.

	$a_2 = 1$			$a_2 = 1.5$			$a_2 = 2$		
	0.1	1	5	0.1	1	5	0.1	1	5
$r = 1$	4×10^{-5}	0.37	0.82	4×10^{-5}	0.37	0.82	4×10^{-5}	0.37	0.82
$r = 2$	6×10^{-3}	0.28	0.65	0.02	0.41	0.77	0.02	0.51	0.84
$r = 3$	0.01	0.22	0.52	0.05	0.43	0.74	0.10	0.60	0.87
$r = 4$	0.02	0.18	0.42	0.08	0.44	0.73	0.19	0.67	0.89

TABLE 2.2: Behavior of the cumulative distribution functions. For selected $a_2 \in \{1, 1.5, 2\}$, values of $\text{pr}(\lambda_r \leq \xi)$, evaluated at selected $\xi \in \{0.1, 1, 5\}$ for increasing $r = 1, \ldots, R$. Without loss of generality a_1 is fixed at 1. When $r = 1$ the $\text{pr}(\lambda_r \leq \xi)$ is analytically available as $F_{\lambda_1}(\xi) = 1 - \gamma(a_1, 1/\xi)/\Gamma(a_1)$, with $\gamma(\cdot, \cdot)$ the incomplete gamma function. When instead $r = 2, \ldots, R$, the quantity $\text{pr}(\lambda_r \leq \xi)$ is evaluated numerically as $\sum_{q=1}^{N}\{1 - \gamma(a_2, \lambda_{r-1}^{(q)}/\xi)/\Gamma(a_2)\}/N$, with $N = 1{,}000{,}000$ and $\lambda_{r-1}^{(q)}$ sampled from (2.6) for each $q = 1, \ldots, N$. This approximation follows from the Markovian structure of the multiplicative inverse gamma which guarantees that $F_{\lambda_r}(\xi) = \text{E}_{\lambda_{r-1}}\{F_{\lambda_r|\lambda_{r-1}}(\xi)\}$ with $F_{\lambda_r|\lambda_{r-1}}(\xi)$ the cdf of $\lambda_r \mid \lambda_{r-1}$ which is still an inverse gamma with shape a_2 and scale λ_{r-1}.

Hence

$$\begin{aligned} \text{E}(1/\vartheta^c) &= \int_0^{+\infty} \vartheta^{-c} \frac{1}{\Gamma(a)} \vartheta^{a-1} e^{-\vartheta} \text{d}\vartheta = \frac{1}{\Gamma(a)} \int_0^{+\infty} \vartheta^{(a-c)-1} e^{-\vartheta} \text{d}\vartheta \\ &= \frac{\Gamma(a-c)}{\Gamma(a)} \int_0^{+\infty} \frac{1}{\Gamma(a-c)} \vartheta^{(a-c)-1} e^{-\vartheta} \text{d}\vartheta = \frac{\Gamma(a-c)}{\Gamma(a)}, \quad (a > c). \end{aligned} \tag{2.12}$$

Exploiting (2.12), $\text{E}(\lambda_r^c) = \text{E}\{(\prod_{m=1}^{r} 1/\vartheta_m)^c\} = \text{E}(\prod_{m=1}^{r} 1/\vartheta_m^c) = \{\Gamma(a_1 - c)/\Gamma(a_1)\}\{\Gamma(a_2 - c)^{r-1}/\Gamma(a_2)^{r-1}\}$, $a_1 > c$, $a_2 > c$. Exploiting this result, Lemma 2.5 proves $\lambda_1 \not\geq \ldots \not\geq \lambda_R$.

Lemma 2.5. *For every $a_2 > 1$, there always exists an increasing function $g^*(\cdot)$ for which $E\{g^*(\lambda_r)\}$ exists and such that $E\{g^*(\lambda_{r+1})\} > E\{g^*(\lambda_r)\}$, for every $r = 1, \ldots, R$.*

Proof. Without loss of generality let $a_1 = a_2$ and $0 < c_{a_2} < a_2$, and consider $g^*(\lambda_r) = \lambda_r^{c_{a_2}}$. As c_{a_2} is positive and since $\lambda_r \in (0, +\infty)$ for every $r = 1, \ldots, R$, the function $g^*(\lambda_r)$ is increasing in the parametric space of λ_r. Moreover, as $c_{a_2} < a_2$, $\text{E}\{g^*(\lambda_r)\}$ exists for every $r = 1, \ldots, R$. Exploiting results in (2.12),

$$\begin{aligned} \text{E}(\lambda_{r+1}^{c_{a_2}}) - \text{E}(\lambda_r^{c_{a_2}}) &= \frac{\Gamma(a_1 - c_{a_2})}{\Gamma(a_1)} \frac{\Gamma(a_2 - c_{a_2})^r}{\Gamma(a_2)^r} - \frac{\Gamma(a_1 - c_{a_2})}{\Gamma(a_1)} \frac{\Gamma(a_2 - c_{a_2})^{r-1}}{\Gamma(a_2)^{r-1}} \\ &= \frac{\Gamma(a_1 - c_{a_2})}{\Gamma(a_1)} \frac{\Gamma(a_2 - c_{a_2})^{r-1}}{\Gamma(a_2)^{r-1}} \left\{ \frac{\Gamma(a_2 - c_{a_2})}{\Gamma(a_2)} - 1 \right\}. \end{aligned} \tag{2.13}$$

Hence showing $\text{E}(\lambda_{r+1}^{c_{a_2}}) - \text{E}(\lambda_r^{c_{a_2}}) > 0$, requires finding a value $0 < c_{a_2} < a_2$, such that $\Gamma(a_2 - c_{a_2}) > \Gamma(a_2)$. According to the well known properties and functional form of the gamma function $\Gamma(\cdot)$, it is always possible to find a value c_{a_2} less than a_2 but sufficiently close to a_2, such that their difference $a_2 - c_{a_2}$ is close to zero enough to obtain $\Gamma(a_2 - c_{a_2}) > \Gamma(a_2)$. This proves absence of stochastic ordering. □

Although absence of stochastic ordering is an undesired property, it doesn't necessarily affect the shrinkage behavior for which the prior has been developed. According to the Markov

inequality $\text{pr}(\lambda_r \geq \xi) \leq \text{E}(\lambda_r)/\xi = 1/\{\xi(a_1-1)(a_2-1)^{r-1}\}$. Hence, a value $a_2 > 2$, increasingly concentrates the upper bound towards zero, meaning that the prior achieves shrinkage for growing r, when $a_2 > 2$. However this property may not hold for $1 < a_2 \leq 2$, representing a subset of the possible hyperparameters settings suggested by Bhattacharya and Dunson (2011). Moreover, decreasing upper bound as r grows, do not univocally characterize the behavior of $\text{pr}(\lambda_{r+1} \leq \xi)$ compared to $\text{pr}(\lambda_r \leq \xi)$. Although stochastic ordering doesn't hold for all $\xi > 0$, analyzing the cumulative distribution function of each λ_r is still important to understand if this property is valid on subsets of $(0, +\infty)$ of interest.

Indeed, as shown in Table 2.2, for every $a_2 \in \{1, 1.5, 2\}$, including $a_2 = 1$, the prior assigns increasing mass to small intervals of zero, such as $(0, \xi = 0.1)$, as r grows, with this mass increasingly higher for growing a_2. This facilitates shrinkage. However as ξ grows, stochastic order no longer holds for all values of a_2, with increasing a_2 apparently enlarging the subset of the parametric space in which $\text{pr}(\lambda_{r+1} \leq \xi) \geq \text{pr}(\lambda_r \leq \xi)$. For example when $a_2 = 1.5$, the property $\text{pr}(\lambda_{r+1} \leq \xi) \geq \text{pr}(\lambda_r \leq \xi)$ is true when $\xi = 0.1$ and $\xi = 1$, but not for $\xi = 5$, while $\text{pr}(\lambda_{r+1} \leq \xi) \geq \text{pr}(\lambda_r \leq \xi)$ holds for all $\xi \in \{0.1, 1, 5\}$ when $a_2 = 2$. Lemma 2.6 proves that stochastic order $\text{pr}(\lambda_{r+1} \leq \xi) \geq \text{pr}(\lambda_r \leq \xi)$ holds as $\xi \to 0^+$ for every $r = 2, \ldots, R$ and $a_2 > 0$.

Lemma 2.6. *For every* $a_2 > 0$, $\lim_{\xi \to 0^+}\{\text{pr}(\lambda_{r+1} \leq \xi)/\text{pr}(\lambda_r \leq \xi)\} \geq 1$.

Proof. To prove Lemma 2.6 let us first study likelihood ratio order $\lambda_{r+1} \leq_{\text{lr}} \lambda_r$ as $\xi \to 0^+$. Adapting Shaked and Shanthikumar (2007) page 42 to our case, this requires showing $\lim_{\xi \to 0^+} \lim_{\Delta \to 0^+}\{f_{\lambda_{r+1}}(\xi)/f_{\lambda_r}(\xi) - f_{\lambda_{r+1}}(\xi+\Delta)/f_{\lambda_r}(\xi+\Delta)\} \geq 0$, where $f_{\lambda_{r+1}}(\xi)$ and $f_{\lambda_r}(\xi)$ are the probability density functions of λ_{r+1} and λ_r, respectively. As the probability density function for a product of independent gammas is available via sophisticated Meijer G-functions (Springer and Thompson, 1970), let us first focus on $f_{\lambda_{r+1}|\lambda_{r-1}}(\xi)$ and $f_{\lambda_r|\lambda_{r-1}}(\xi)$ representing the conditional density function of λ_{r+1} and λ_r, given $\lambda_{r-1} > 0$, respectively. As $\lambda_r = \lambda_{r-1}/\vartheta_r$, with $1/\vartheta_r \sim \text{Inv-Ga}(a_2, 1)$, from the standard properties of the inverse gamma random variable, $f_{\lambda_r|\lambda_{r-1}}(\xi)$ is easily available as the probability density function for the random variable $\lambda_r \mid \lambda_{r-1} \sim \text{Inv-Ga}(a_2, \lambda_{r-1})$. To compute $f_{\lambda_{r+1}|\lambda_{r-1}}(\xi)$ note instead that $\lambda_{r+1} = \lambda_h/\vartheta_{r+1}$, with $1/\vartheta_{r+1} \sim \text{Inv-Ga}(a_2, 1)$. Hence $\lambda_{r+1} \mid \lambda_r \sim \text{Inv-Ga}(a_2, \lambda_r)$ and $\lambda_r \mid \lambda_{r-1} \sim \text{Inv-Ga}(a_2, \lambda_{r-1})$. Exploiting this Markov property

$$\begin{aligned} f_{\lambda_{r+1}|\lambda_{r-1}}(\xi) &= \int_0^{+\infty} f_{\lambda_{r+1}|\lambda_r=x}(\xi) f_{\lambda_r|\lambda_{r-1}}(x)\mathrm{d}x = \frac{\lambda_{r-1}^{a_2}}{\Gamma(a_2)^2}\xi^{-a_2-1}\int_0^{+\infty} x^{-1}e^{-\frac{\lambda_{r-1}}{x}-\frac{x}{\xi}}\mathrm{d}x \\ &= \frac{\lambda_{r-1}^{a_2}}{\Gamma(a_2)^2}\xi^{-a_2-1}\int_0^{+\infty} y^{-1}e^{-y-\frac{\lambda_{r-1}/\xi}{y}}\mathrm{d}y, \end{aligned} \tag{2.14}$$

where the last equality in (2.14) follows after the change of variable $x/\xi = y$. To evaluate (2.14), note that from the theory of Bessel functions $K_\nu(z) = 0.5(0.5z)^\nu \int_0^{+\infty} t^{-\nu-1}\exp(-t -$

$z^2/4t)\mathrm{d}t$, where $K_\nu(\cdot)$ is the modified Bessel function of the second kind with parameter ν; refer to Watson (1966) page 183. Hence, after changing λ_{r-1}/ξ with $\{2(\lambda_{r-1}/\xi)^{1/2}\}^2/4$ and rewriting (2.14) to highlight the Bessel component, $f_{\lambda_{r+1}|\lambda_{r-1}}(\xi)$ is

$$\begin{aligned} f_{\lambda_{r+1}|\lambda_{r-1}}(\xi) &= \frac{\lambda_{r-1}^{a_2}}{\Gamma(a_2)^2}\xi^{-a_2-1}\frac{2}{2}\{2(\lambda_{r-1}/\xi)^{\frac{1}{2}}\}^0\int_0^{+\infty} y^{-1}e^{-y-\frac{\{2(\lambda_{r-1}/\xi)^{1/2}\}^2}{4y}}\mathrm{d}y \\ &= \frac{2\lambda_{r-1}^{a_2}}{\Gamma(a_2)^2}\xi^{-a_2-1}K_0\{2(\lambda_{r-1}/\xi)^{\frac{1}{2}}\}. \end{aligned} \tag{2.15}$$

Once $f_{\lambda_{r+1}|\lambda_{r-1}}(\xi)$ is available as in (2.15) let us evaluate the likelihood ratio order as $\xi \to 0^+$. According to previous discussion this requires showing that $\lim_{\xi\to 0^+}\lim_{\Delta\to 0^+}\{g(\xi) - g(\xi+\Delta)\} \geq 0$, where $g(\xi) = f_{\lambda_{r+1}|\lambda_{r-1}}(\xi)/f_{\lambda_r|\lambda_{r-1}}(\xi)$ is defined as

$$g(\xi) = \frac{2\lambda_{r-1}^{a_2}}{\Gamma(a_2)^2}\xi^{-a_2-1}K_0\{2(\lambda_{r-1}/\xi)^{\frac{1}{2}}\}\frac{\Gamma(a_2)}{\lambda_{r-1}^{a_2}}\xi^{a_2+1}e^{\frac{\lambda_{r-1}}{\xi}} = \frac{2}{\Gamma(a_2)}K_0\{2(\lambda_{r-1}/\xi)^{\frac{1}{2}}\}e^{\frac{\lambda_{r-1}}{\xi}}. \tag{2.16}$$

In proving Lemma 2.6 note that the limit $\lim_{\xi\to 0^+}\lim_{\Delta\to 0^+}\{g(\xi) - g(\xi+\Delta)\} \geq 0$ if and only if $\lim_{\xi\to 0^+}\lim_{\Delta\to 0^+}\{g(\xi) - g(\xi+\Delta)\}/\Delta \geq 0$, provided that $\Delta > 0$. By the standard definition of first derivative $\mathrm{d}g(\xi)/\mathrm{d}\xi = \lim_{\Delta\to 0^+}\{g(\xi+\Delta) - g(\xi)\}/\Delta$, previous inequality reduces to prove $\lim_{\xi\to 0^+}\mathrm{d}g(\xi)/\mathrm{d}\xi \leq 0$. Let us compute $\mathrm{d}g(\xi)/\mathrm{d}\xi$, with $g(\xi)$ from (2.16).

$$\frac{\mathrm{d}g(\xi)}{\mathrm{d}\xi} = \frac{2}{\Gamma(a_2)}e^{\frac{\lambda_{r-1}}{\xi}}\left\{K_1\{2(\lambda_{r-1}/\xi)^{\frac{1}{2}}\}\lambda_{r-1}^{\frac{1}{2}}\xi^{-\frac{3}{2}} - K_0\{2(\lambda_{r-1}/\xi)^{\frac{1}{2}}\}\lambda_{r-1}\xi^{-2}\right\}.$$

Adapting results in page 378 of Abramowitz and Stegun (1964) to our case, as $\xi \to 0^+$, $K_0\{2(\lambda_{r-1}/\xi)^{1/2}\} \approx K_1\{2(\lambda_{r-1}/\xi)^{1/2}\} \approx 0.5\pi^{1/2}\lambda_{r-1}^{-1/4}\xi^{1/4}e^{-2(\lambda_{r-1}/\xi)^{1/2}}$. Hence, as $\xi \to 0^+$

$$\begin{aligned} \frac{\mathrm{d}g(\xi)}{\mathrm{d}\xi} &\approx \frac{\pi^{1/2}}{\Gamma(a_2)}e^{\frac{\lambda_{r-1}}{\xi}(1-2\lambda_{r-1}^{3/2}\xi^{1/2})}\lambda_{r-1}^{-1/4}\xi^{1/4}\left\{\lambda_{r-1}^{1/2}\xi^{-3/2} - \lambda_{r-1}\xi^{-2}\right\} \\ &= \frac{\pi^{1/2}\lambda_{r-1}^{3/4}}{\Gamma(a_2)}e^{\frac{\lambda_{r-1}}{\xi}(1-2\lambda_{h-1}^{3/2}\xi^{1/2})}\xi^{-7/4}(\lambda_{r-1}^{-1/2}\xi^{1/2} - 1). \end{aligned}$$

Since $\lim_{\xi\to 0^+}(\lambda_{r-1}^{-1/2}\xi^{1/2} - 1) = -1$ and $\lim_{\xi\to 0^+} e^{\lambda_{r-1}(1-2\lambda_{r-1}^{3/2}\xi^{1/2})/\xi}\xi^{-7/4} = +\infty$, it follows that $\lim_{\xi\to 0^+}\mathrm{d}g(\xi)/\mathrm{d}\xi \leq 0$ for every $a_2 > 0$ and $\lambda_{r-1} > 0$. This proves $\lambda_{r+1} \mid \lambda_{r-1} \leq_{\mathrm{lr}} \lambda_r \mid \lambda_{r-1}$ as $\xi \to 0^+$. As order in likelihood ratio implies stochastic order (Shaked and Shanthikumar, 2007; page 43), previous results guarantees $\lambda_{r+1} \mid \lambda_{r-1} \preceq \lambda_r \mid \lambda_{r-1}$ for every $a_2 > 0$ when $\xi \to 0^+$. Finally, since stochastic order is closed under mixtures (Shaked and Shanthikumar, 2007; page 6) and provided that $\lambda_{r+1} \mid \lambda_{r-1} \preceq \lambda_r \mid \lambda_{r-1}$ holds for every λ_{r-1} when $\xi \to 0^+$, it follows that $\lim_{\xi\to 0^+}\{\mathrm{pr}(\lambda_{r+1} \leq \xi)/\mathrm{pr}(\lambda_r \leq \xi)\} \geq 1$ for every $a_2 > 0$, proving Lemma 2.6. □

Lemma 2.6 is appealing in guaranteeing that the prior assigns increasing mass to a small neighborhood of zero for all $a_2 > 0$ as r grows, facilitating shrinkage. However this interval may be substantially small. Hence, from an applied perspective, it is worth assessing

	$a_2 = 1$	$a_2 = 1.5$	$a_2 = 2$	$a_2 = 2.5$	$a_2 = 3$
$a_1 = 1$					
$r = 1 \to r = 2$	$(0, \xi = 0.52)$	$(0, \xi = 1.52)$	$(0, \xi > 100)$	$(0, \xi > 100)$	$(0, \xi > 100)$
$r = 2 \to r = 3$	$(0, \xi = 0.33)$	$(0, \xi = 1.61)$	$(0, \xi > 100)$	$(0, \xi > 100)$	$(0, \xi > 100)$
$r = 3 \to r = 4$	$(0, \xi = 0.22)$	$(0, \xi = 1.72)$	$(0, \xi > 100)$	$(0, \xi > 100)$	$(0, \xi > 100)$
$r = 4 \to r = 5$	$(0, \xi = 0.14)$	$(0, \xi = 1.88)$	$(0, \xi > 100)$	$(0, \xi > 100)$	$(0, \xi > 100)$
$a_1 = 2$					
$r = 1 \to r = 2$	$(0, \xi = 0.33)$	$(0, \xi = 0.65)$	$(0, \xi = 1.79)$	$(0, \xi = 25.94)$	$(0, \xi > 100)$
$r = 2 \to r = 3$	$(0, \xi = 0.21)$	$(0, \xi = 0.66)$	$(0, \xi = 3.18)$	$(0, \xi > 100)$	$(0, \xi > 100)$
$r = 3 \to r = 4$	$(0, \xi = 0.13)$	$(0, \xi = 0.68)$	$(0, \xi = 5.67)$	$(0, \xi > 100)$	$(0, \xi > 100)$
$r = 4 \to r = 5$	$(0, \xi = 0.09)$	$(0, \xi = 0.69)$	$(0, \xi = 9.50)$	$(0, \xi > 100)$	$(0, \xi > 100)$

TABLE 2.3: Solutions of (2.17) for $r = 1, \dots 4$, based on different combinations of a_1 and a_2. In evaluating (2.17), $N = 1{,}000{,}000$.

whether this property holds for a larger subset of the parametric space. This requires finding for which values $\xi > 0$ the inequality $F_{\lambda_{r+1}}(\xi) - F_{\lambda_r}(\xi) = \text{pr}(\lambda_{r+1} \leq \xi) - \text{pr}(\lambda_r \leq \xi) \geq 0$, holds. As previously discussed, derivation of the cumulative distribution for the product of independent inverse gammas is a cumbersome task. Few results are obtained for the product of two gammas (Withers and Nadarajah, 2013). However also in these simpler settings analytical forms are available only for specific values of a_2 via sophisticated combinations of modified Bessel and Struve functions.

To overcome previous issues let us exploit the Markovian structure of the multiplicative inverse gamma process which guarantees that $\lambda_r \mid \lambda_{r-1} \perp \lambda_{r-2}, \dots, \lambda_1 \sim \text{Inv-Ga}(a_2, \lambda_{r-1})$, $r = 2, \dots, R$. Exploiting this property we can write the previous inequality between the cumulative distribution functions as $\text{E}_{\lambda_r}\{F_{\lambda_{r+1}|\lambda_r}(\xi)\} - \text{E}_{\lambda_{r-1}}\{F_{\lambda_r|\lambda_{r-1}}(\xi)\} = \text{E}_{\lambda_r}\{\Gamma(a_2, \lambda_r/\xi)/\Gamma(a_2)\} - \text{E}_{\lambda_{r-1}}\{\Gamma(a_2, \lambda_{r-1}/\xi)/\Gamma(a_2)\} \geq 0$. As the previous expectations require the probability density functions for a product of inverse gamma, and provided that these quantities are available via sophisticated Meijer G-functions, the quantities $\text{E}_{\lambda_r}\{\Gamma(a_2, \lambda_r/\xi)/\Gamma(a_2)\}$ are not analytically available. Hence, to address our goal let us focus on finding the solutions for the numerical approximation of $\text{E}_{\lambda_r}\{F_{\lambda_{r+1}|\lambda_r}(\xi)\} - \text{E}_{\lambda_{r-1}}\{F_{\lambda_r|\lambda_{r-1}}(\xi)\} \geq 0$, which can be easily obtained as

$$\frac{1}{N}\sum_{q=1}^{N}\frac{\Gamma(a_2, \lambda_r^{(q)}/\xi)}{\Gamma(a_2)} - \frac{1}{N}\sum_{q=1}^{N}\frac{\Gamma(a_2, \lambda_{r-1}^{(q)}/\xi)}{\Gamma(a_2)} \geq 0, \quad r = 2, \dots, R, \tag{2.17}$$

where samples $\lambda_r^{(q)}$ and $\lambda_{r-1}^{(q)}$, $q = 1, \dots, N$ are easily available as cumulative products of r and $r-1$ independent inverse gammas from (2.6), respectively. Note that when $r = 1$, inequality (2.17), reduces to $\sum_{q=1}^{N}\{\Gamma(a_2, \lambda_1^{(q)}/\xi)/\Gamma(a_2)\}/N - \Gamma(a_1, 1/\xi)/\Gamma(a_1) \geq 0$.

In order to provide guidelines for possible behaviors of the multiplicative gamma process prior, Table 2.3 reports solutions of (2.17) for different combinations of a_1 and a_2, at increasing $r = 1, \dots, 4$. Note how, consistently with Lemma 2.6, stochastic order $\text{pr}(\lambda_{r+1} \leq \xi) \geq \text{pr}(\lambda_r \leq \xi)$ holds in an interval of zero for every a_2 and r in Table 2.3, with this interval becoming

increasingly larger as a_2 grows for every combination of r and a_1. Also a_1 plays a role in defining the dimension of such interval. According to Table 2.3, the higher is a_1 the smaller is the interval where stochastic order holds for every combination of a_2 and r. Note also how, the dimensions of such interval generally grows as r increases for the combinations of a_1 and a_2 considered, with exception of $a_2 = 1$. Finally it is worth noticing how the subset of $(0, +\infty)$ where stochastic order holds become substantially wide when a_2 is moderately higher than a_1. Although Table 2.3 focuses on few standard cases, this study provides the researchers with the basic guidelines and tools to evaluate the properties of the multiplicative inverse gamma prior at all possible combinations of a_1, a_2 and dimensions r. This is a key to check desirable behaviors in our practical applications.

2.1.5 Simulation study

We conduct a simulation study to evaluate the performance of the proposed approach in accommodating dynamic heterogenous connectivity patterns. We focus on estimating the dynamic edge probabilities and on out-of-sample forecasting. We additionally compare our results with an approach that uses only temporal information. We generate a set of 15×15 time-varying matrices A_{t_i} with $t_i \in \mathbb{T}^{\text{sim}} = \{1, \ldots, 40\}$. Each edge $\mathcal{L}(A_{t_i})_l$, $l = 1, \ldots, 15(15-1)/2$, $i = 1, \ldots, 40$, is simulated according to (2.1) with probabilities obtained from (2.2)–(2.3), generating $\{\mu(1), \ldots, \mu(40)\}^{\text{T}}$ from a GP$(0, c_\mu)$ with length scale $\kappa_\mu = 0.01$ and choosing two time-varying latent coordinates $\{X_{v1}(1), \ldots, X_{v1}(40)\}^{\text{T}}$, $\{X_{v2}(1), \ldots, X_{v2}(40)\}^{\text{T}}$, from Gaussian processes with length scale $\kappa_x = 0.01$, independently for each node $v = 1, \ldots, 15$. To evaluate out-of-sample predictive performance, we perform posterior inference taking $\mathcal{L}(A_{t_{40}})$ to be a vector missing edges, and then compare our predictions with the simulated data $\mathcal{L}(A_{t_{40}})$.

For inference we choose $R = 10$ and length scales $\kappa_\mu = \kappa_x = 0.05$. Recalling discussion in Section 2.1.4 we set $a_1 = 2.5$ and $a_2 = 3.5$ for the shrinkage parameters. According to Bhattacharya and Dunson (2011), this choice additionally ensures the existence of the first two moments for the induced priors on elements $S_l(t)$ at every $t \in \mathbb{T}$. We consider $5{,}000$ Gibbs iterations, and discard the first $1{,}000$. Mixing has been assessed via effective sample sizes for the quantities of interest, represented by $\pi_l(t_i)$, for $l = 1, \ldots, V(V-1)/2$ and $t_i \in \mathbb{T}^{\text{sim}}$ after burn-in. Most of these values were around $1{,}700$ out of $4{,}000$, suggesting good mixing. We additionally assess convergence by investigating the Gelman and Rubin (1992) potential scale reduction factors (PSRF). These are computed by dividing each chain in four consecutive sub-chains of length $1{,}000$ after burn-in, and comparing within and between sub-chains variance. The median of the PSRFs for the chains of the edge probabilities at every time, is 1.01, with the 99% of these PSRFs being less than 1.2, providing evidence that convergence has been reached.

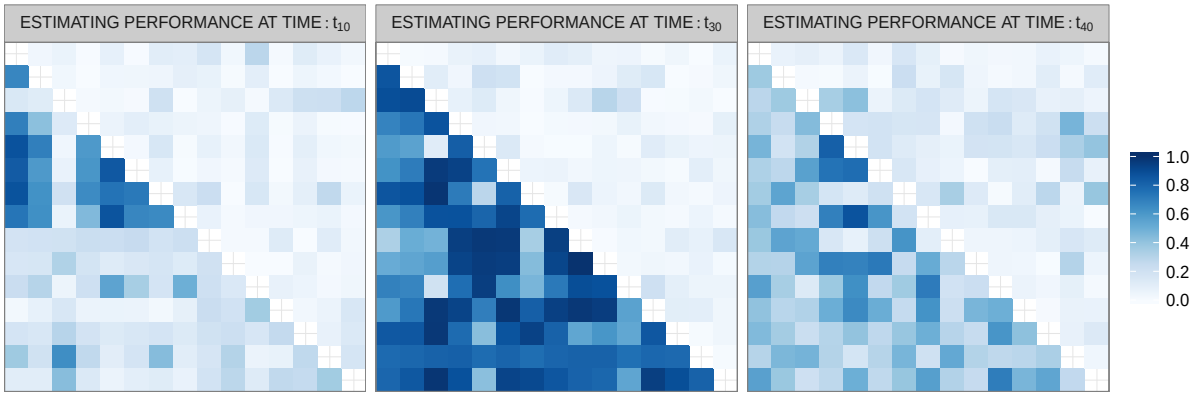


FIGURE 2.1: For selected times t_i, plot of the posterior mean for edge probabilities $\hat{\pi}_l(t_i)$, $l = 1, \ldots, V(V-1)/2$ – rearranged in matrix form – (lower triangular), and absolute value of the difference $|\hat{\pi}_l(t_i) - \pi_l^0(t_i)|$ (upper triangular).

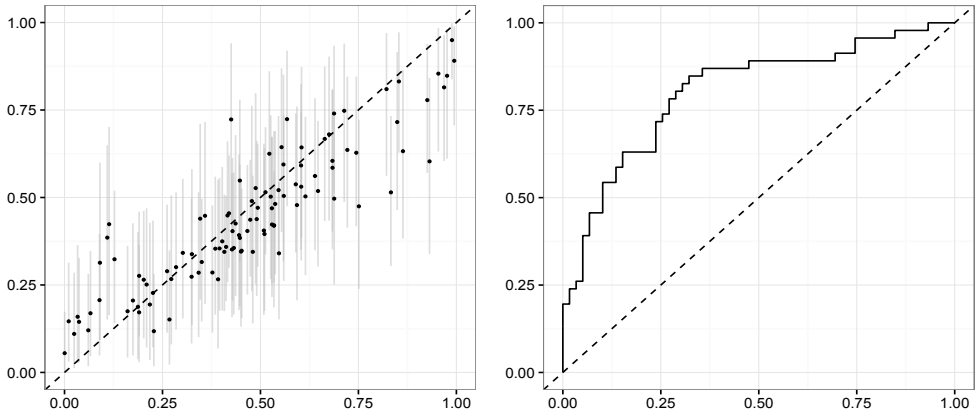


FIGURE 2.2: Left plot: plots of the true edge probabilities at time t_{40}, $\pi_l^0(t_{40})$ (x-axis) versus their posterior mean $\hat{\pi}_l(t_{40})$ (y-axis), $l = 1, \ldots, V(V-1)/2$. Segments denote the 0.95 highest posterior density intervals. Right plot: forecasting performance assessed via the receiver operating characteristic curve (ROC) generated using $\hat{\pi}_l(t_{40})$ and the observed edges $\mathcal{L}(A_{t_{40}})_l$, $l = 1, \ldots, V(V-1)/2$.

The graphical representation in Figure 2.1 of the estimated edge probabilities – rearranged in matrix form – and their difference with the corresponding true values for some selected times t_i highlights the good performance of our approach in estimation and forecasting. The latter can be noticed by focusing on the third matrix assessing performance at t_{40}, recalling that data at t_{40} were held out in deriving the posterior distribution. Accurate forecasting performance is further highlighted in Figure 2.2, displaying the true $\pi_l^0(t_{40})$ against the corresponding estimates $\hat{\pi}_l(t_{40})$, along with the ROC curve when predicting $\mathcal{L}(A_{t_{40}})$ with the expectation of its forecasted predictive distribution according to equation (2.11).

Figure 2.3 compares the performance of our model with respect to selected edge probability trajectories $\pi_l(t_1), \ldots, \pi_l(t_n)$ with the inferential results when each of these edge probability processes $\pi_l(t_1), \ldots, \pi_l(t_n)$ is estimated with the same setting of our model but using only the time series of the corresponding edges $\mathcal{L}(A_{t_1})_l, \ldots, \mathcal{L}(A_{t_n})_l$ without borrowing information across the network. The sub-optimality of the independent approach is apparent in terms

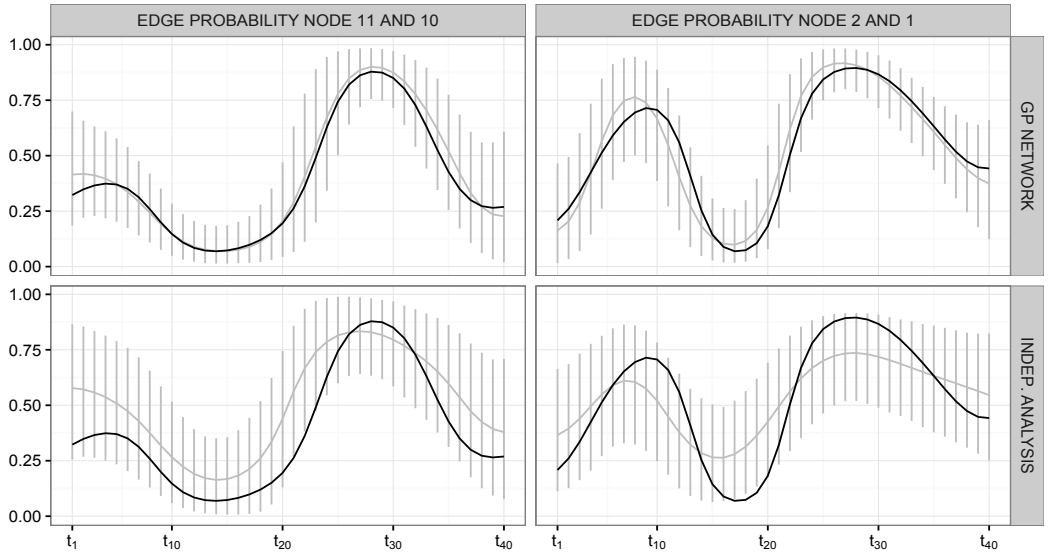


FIGURE 2.3: Model comparison. Upper panels: for selected pairs of nodes, plots of the true edge probability trajectories (black lines), pointwise posterior means (gray lines) and 0.95 highest posterior density intervals (gray segments) for our model. Lower panels: same quantities estimated using only temporal information without exploring network structure. Specifically, we estimate each edge probability trajectory using only the time series of the edges observed for the corresponding pair of nodes, instead of considering the entire network information.

of over-smoothed trajectories and wider highest posterior density intervals. When network structure is taken into account, the model provides accurate estimates, with posterior distributions better concentrating around the true parameters, while adaptively deleting latent space dimensions not required to characterize the observed data. In particular, we find that the posterior mean for λ_r drop to small values for $r = 3, \ldots, 10$. This implies that these later dimensions trajectories are flat and have limited influence.

Borrowing information across the network over time has the additional advantage of reducing sensitivity to the choice of hyperparameters, in particular with respect to the length scales in the Gaussian process priors. Our approach can be easily modified to learn the length scales from the data as in Murray and Adams (2010). However, since we obtain similar results when instead letting $\kappa_\mu = \kappa_x = 0.03$, $\kappa_\mu = \kappa_x = 0.1$ and $\kappa_\mu = \kappa_x = 0.5$ in sensitivity analyses, we preferred to simply elicit the length scales to favor smooth trajectories a priori.

2.1.6 Application to international cooperation relationships networks

We apply our dynamic network model outlined in Sections 2.1.1–2.1.2 to GDELT relationships data $A_{t_1}, \ldots, A_{t_{127}}$ described in Section 1.1.1, considering the same settings as the in simulation study. Also mixing via effective sample sizes and convergence based on Gelman and Rubin (1992) potential scale reduction factors, are on similar values.

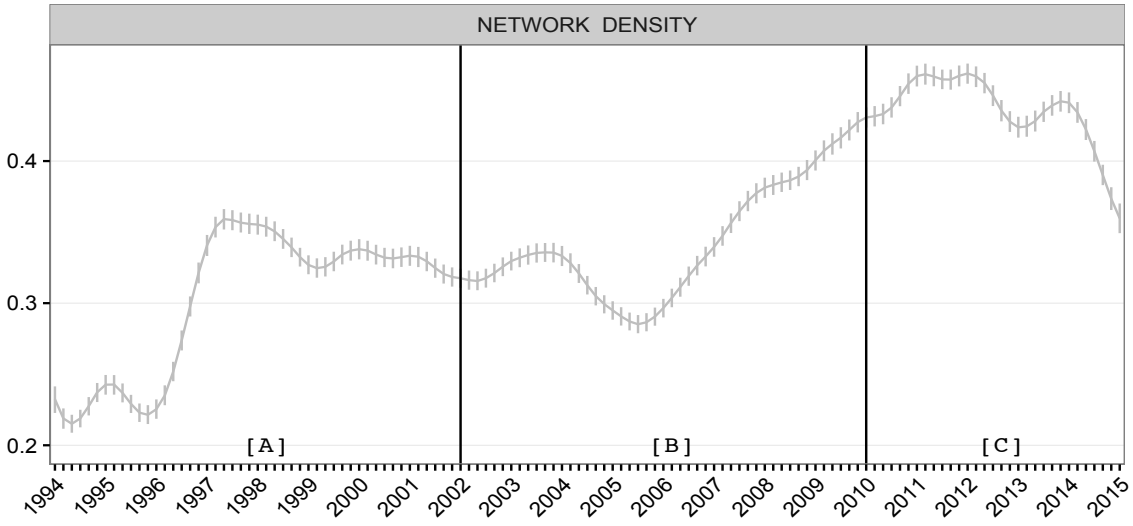


FIGURE 2.4: Trajectory of the posterior mean for the expected network density (gray line) and pointwise posterior interquartile range (gray segments). [A] Mexican economic crisis (≈1995), Asian financial crisis (≈1997–1998), Russian financial crisis (≈1998), Turkish financial crisis (≈2000–2001), Argentinian financial crisis (≈1999–2001), raise and burst of Dot-com bubble (≈1997–2001); [B] Raise and burst of housing bubble (≈2002–2007) and global financial crisis (≈2007–2009); [C] European debt crisis (≈2010–2013), Russian financial crisis (≈2014).

The trajectory of the posterior mean for the expected network density in Figure 2.4 provides an appealing overview of the overall dynamic connectivity behavior in relation to key financial and economic international events. Note that the posterior distribution of this quantity – at every time t_i – can be easily derived as a function of the posterior samples for the edge probabilities, as $\mathrm{E}[\sum_{l=1}^{V(V-1)/2} \mathcal{L}(\mathcal{A}_{t_i})_l/\{V(V-1)/2\}] = \sum_{l=1}^{V(V-1)/2} \mathrm{E}\{\mathcal{L}(\mathcal{A}_{t_i})_l\}/\{V(V-1)/2\} = \sum_{l=1}^{V(V-1)/2} \pi_l(t_i)/\{V(V-1)/2\}$. It is first interesting to notice how the posterior mean for the time-varying expected network density evolves on a range between 0.2 and 0.5, meaning that there is not an overall strong tendency towards material cooperations compared to material conflicts in the time window considered. This can be partially explained by the fact that 7 nodes on a total of 25 represent Arab countries, which traversed long conflictual periods with the other nations, but also may reflect an overall general tendency of mass media towards negative news reports, such as material conflicts, rather than positive ones; see for example Thapthiang (2013) and the references cited therein.

Although evolving on low values, the overall dynamic connectivity behavior is characterized by a positive trend, which traverses several changes and cyclical periods interestingly related to the key financial events occurring in the time window considered. We observe a rapid change in the expected network density at the burst of the Asian financial crisis in 1997, which then remains on similar high levels in the subsequent years, while displaying bumps in correspondence of the main crises occurred in [A] – i.e. the Mexican peso crisis in 1995, the 1997–1998 Asian financial crisis, the Russian flu in 1998, the 2000–2001 Turkish economic crisis and the Argentina great depression of 1999–2001. Refer to Eun and Resnick (2010) and the references cited therein. Previous crises are generally accompanied by rescue

packages and increased material cooperation relationships among international countries to organize bailout investments and avoid facing spread of contagion in case of financial collapse of the countries affected (Eun and Resnick, 2010). Our estimated increments for the overall propensity towards material cooperations in correspondence of previous crises confirm this behavior, with the persistent high levels after 1997 potentially related to the growth of the 1997–2001 Dot-com bubble, which facilitated worldwide investments.

The estimated expected network density remains approximately on the same level in later years until further increasing from 2006, with the burst of the United States housing bubble (Taylor, 2009) and the subsequent 2007–2009 Global financial crisis (Brunnermeier, 2009) [B]. This behavior is interestingly consistent with our previous conclusions for time window [A] and key analyses from Bernanke (2007), Brunnermeier (2009) and Taylor (2009). In particular, similarly to the Dot-com bubble, the behavior prior to 2006 can be related to the growth of the United States housing bubble, which was stimulated by the unusually low interest rates decision of the Federal Reserve to mitigate the effects of the Dot-com bubble (Taylor, 2009) and facilitated a wide network of material investments among countries under a "global saving glut" scenario (Bernanke, 2007). The increase of the expected network density in later years is instead reasonably associated with need for international material cooperation to provide the many bailouts and bank rescue packages in order to avoid bankruptcy or spread of financial collapses. Refer to Brunnermeier (2009) for an overview of the interventions required on key financial institutions covering Northern Rock, Bear Stearns, JP Morgan, Lehman Brother and others. A similar scenario applies during the subsequent European debt crisis, which required important bailout investments by the European Stability Mechanism and the International Monetary Fund to face the most acute phases of the crises for Greece, Ireland, Portugal in 2010–2011 and Spain in 2012 (Belkin et al., 2012). Our dynamic network model captures also these events with high levels in 2010 – 2012, which are followed by a last increment in correspondence with the recent 2014 Russia ruble crisis.

We provide further insights to specific events by focusing on the estimated dynamic cooperation probabilities between selected pairs of countries, outlined in Figure 2.5. Results in the top panels confirm previous discussion on the European debt crisis with a specific focus to Greece. Consistently with the leading role of Germany in guaranteeing financial stability of the Eurozone, the estimated cooperation probability between Greece and Germany rapidly increases exactly at the burst of the Greek debt crisis and later stabilizes at very high levels. Conversely, the relationships between United States and Greece are instead characterized by a decreasing trend starting in 2010. This may be a result of efforts to reduce inter-connection with a country in crisis.

The last two panels provide insights on the effect of recent conflicts on the dynamics of the estimated cooperation probabilities. In the three time windows of the middle panels we

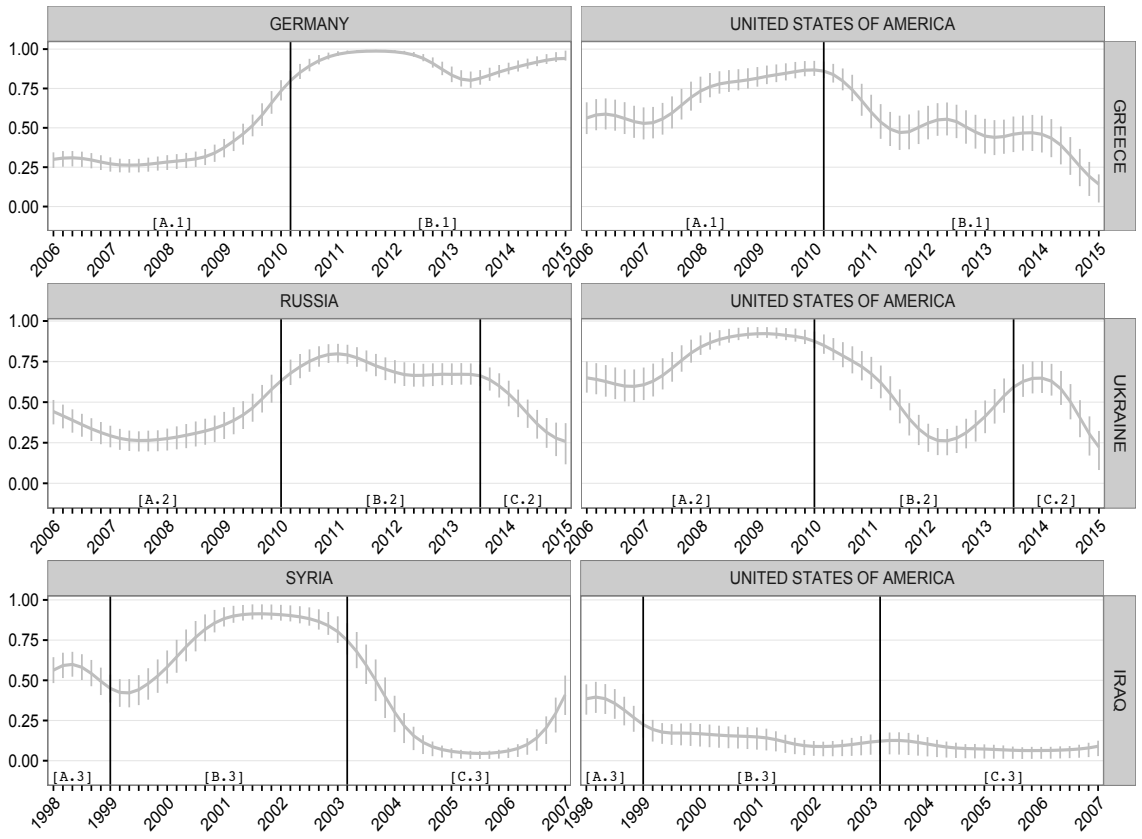


FIGURE 2.5: For selected time windows. Upper panels: posterior mean (gray lines) and pointwise posterior interquartile range (gray segments) for the dynamic cooperation probabilities between Greece–USA and Greece–Germany. Middle panels: same quantities with respect to Ukraine–USA and Ukraine–Russia. Lower panels: same quantities with respect to Iraq–USA and Iraq–Syria.

learn opposite behavior of United States and Russia in their cooperation relationships with Ukraine. In particular, window [A.2] refers to Viktor Yushchenko president (2005–2010) and Yulia Tymoshenko prime minister (2007–2010) period who deepened relations with United States after the Orange Revolution in 2004 and supported NATO membership for Ukraine while progressively increasing conflicts occasions with Russia, which culminated in the 2008–2009 Russian-Ukrainian gas crisis (Tsygankov, 2015). Consistently with the previous political background, we learn evident lower cooperation relationships between Russia and Ukraine with respect to United States and Ukraine, with the latter evolving on very high levels after a bump in 2007 when the prime minister Viktor Yanukovych was succeeded by Yulia Tymoshenko. Differently from Yushchenko and Tymoshenko, Yanukovych improved relations with Russia since he was elected president in 2010, renouncing any aspirations to join NATO and allowing Russia's Black Sea Fleet to stay in the Crimean port of Sevastopol (Tsygankov, 2015). This change of regime is evident in the two trajectories of the estimated edge probabilities in [B.2] which cross in correspondence of Yanukovych election to reach higher and lower levels for Russia and United States, respectively. As expected, a further increment is evident in [C.2] during the 2013–2015 Ukrainian–Russia crisis and the related ousting of Yanukovych, when the estimated relationships between Ukraine and United States returns to high levels

while those with Russia sharply drop.

The lower panels focus instead on the Iraq war. As expected, relationships between United States and Iraq evolve on very low values, with an estimated decrement at the end of 1998 in correspondence of the 1998 Iraq Liberation Act and the subsequent operation Desert Fox in December 1998. Cooperation relationships between Syria and Iraq register instead an increase around the 2000 with the raise of Bashar al-Assad as president of Syria [B.3], and remains on high levels until the 2003–2011 Iraq war [C.3]. These results appear to be consistent with improved economic relations between Iraq and Syria under the Bashar al-Assad regime, mostly related to Iraq oil exports at subsidized prices, which was later shut down by the United States invasion of Iraq in 2003 (Hinnebush, 2009).

We conclude our analysis by evaluating in-sample prediction and out-of-sample forecasting performance in the GDELT data application. This is accomplished by performing estimation until 2014 – using networks from t_1 to t_{126} – and forecasting edges at t_{127} – coinciding with the first bimester of 2015 – according to the procedure outlined in equation (2.11). We obtain an area ROC the curves of 0.79, and 0.71 for in-sample prediction and out-of-sample forecasting, respectively, providing good results given the complexity of GDELT data.

2.2 Locally adaptive dynamic network inference

Motivated by dynamic face-to-face human interaction data described in Section 1.1.2, we generalize methodologies developed in Section 2.1 to improve flexibility and computational tractability. Our Locally Adaptive DYnamic (LADY) network model, relies on the same latent space formulation previously defined in equations (2.1)–(2.3), but substantially modifies prior specification by considering nested Gaussian process (nGP) priors on the latent coordinates to flexibly accommodate time-varying smoothness patterns. Using matrix factorization procedures, our LADY network model can accommodate moderately large V, and considering a state space representation of the nGP, we improve scalability of inference and provide novel procedures for fast forecasting of future networks. Adapting the Pólya-gamma data augmentation strategy to our specific setting, we develop a novel and efficient Gibbs sampler for posterior computations, which utilizes standard results of Kalman filter (Durbin and Koopman, 2002) for transformed Gaussian data.

2.2.1 From Gaussian process to nested Gaussian process priors

Although the dynamic latent space model developed in Section 2.1 provides a continuous time and highly general methodology that accommodates missing data, accounts for across-node heterogeneity and scales to moderately large V, there are two issues which may arise when focusing on data sets similar to those described in Section 1.1.2. Firstly, the proposed coordinates processes assume a stationary dependence structure, and hence tends to under-smooth during periods of stability and over-smooth during periods of sharp changes. Secondly, the well known computational problems with usual GP regression are inherited, leading to difficulties in developing strategies for fast forecasting of future networks. If we define stationary processes, which assume that the correlation between the realizations at times t_i and t_j only depend on the time separation $(t_i - t_j)^2$, it is straightforward to show that the resulting network-valued stochastic process will inherit this stationarity. To realistically characterize the face-to-face human interaction data, it is necessary to accommodate non-stationarity. However, this needs to be done in a careful way to avoid needing to estimate many parameters related to non-stationarity and face computational intractability. Although there is a rich literature on incorporating non-stationarity in GPs, such models tend to be highly challenging to implement even in simpler settings in which data consist of direct error-prone measurements of a single function.

With these issues in mind, we maintain the same model (2.1)–(2.3) previously described in Section 2.1.1, but rely on nested GPs (nGPs) (Zhu and Dunson, 2013) rather than GPs to induce highly flexible stochastic processes on $\{\mu(t) : t \in \mathbb{T}\}$ and $\{X_{vr}(t) : t \in \mathbb{T}\}$ for every

$v = 1, \ldots, V$ and $r = 1, \ldots, R$. nGPs explicitly model time-varying smoothness by defining stochastic differential equations for the function's derivatives. Focusing on the trajectory $\{X_{vr}(t) : t \in \mathbb{T}\}$, the stochastic differential equation representation for the nGP can be accurately characterized by the following state equations for $\{X_{vr}(t) : t \in \mathbb{T}\}$, it's first order derivative $\{X'_{vr}(t) : t \in \mathbb{T}\}$ and the local instantaneous mean $\{U_{vr}(t) : t \in \mathbb{T}\}$ – where $U_{vr}(t) = \mathrm{E}\{X'_{vr}(t) \mid U_{vr}(t)\}$.

$$\begin{aligned}\begin{bmatrix} X_{vr}(t_{i+1}) \\ X'_{vr}(t_{i+1}) \\ U_{vr}(t_{i+1}) \end{bmatrix} &= \begin{bmatrix} 1 & \delta_i & 0 \\ 0 & 1 & \delta_i \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X_{vr}(t_i) \\ X'_{vr}(t_i) \\ U_{vr}(t_i) \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \eta_{i_{X_{vr}}} \\ \eta_{i_{U_{vr}}} \end{bmatrix}, \\ &= T_i \begin{bmatrix} X_{vr}(t_i) \\ X'_{vr}(t_i) \\ U_{vr}(t_i) \end{bmatrix} + Q_i \begin{bmatrix} \eta_{i_{X_{vr}}} \\ \eta_{i_{U_{vr}}} \end{bmatrix},\end{aligned} \tag{2.18}$$

independently for $v = 1, \ldots, V$ and $r = 1, \ldots, R$, with $(\eta_{i_{X_{vr}}}, \eta_{i_{U_{vr}}})^{\mathrm{T}} \sim \mathrm{N}_2(0, \Sigma_{vr})$, $\Sigma_{vr} = \mathrm{diag}(\sigma^2_{X_{vr}}\delta_i, \sigma^2_{U_{vr}}\delta_i)$ and $\delta_i = t_{i+1} - t_i$ sufficiently small. Similarly, the state equations implied for $\{\mu(t) : t \in \mathbb{T}\}$ are

$$\begin{aligned}\begin{bmatrix} \mu(t_{i+1}) \\ \mu'(t_{i+1}) \\ M(t_{i+1}) \end{bmatrix} &= \begin{bmatrix} 1 & \delta_i & 0 \\ 0 & 1 & \delta_i \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu(t_i) \\ \mu'(t_i) \\ M(t_i) \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \eta_{i_\mu} \\ \eta_{i_M} \end{bmatrix}, \\ &= T_i \begin{bmatrix} \mu(t_i) \\ \mu'(t_i) \\ M(t_i) \end{bmatrix} + Q_i \begin{bmatrix} \eta_{i_\mu} \\ \eta_{i_M} \end{bmatrix},\end{aligned} \tag{2.19}$$

where $(\eta_{i_\mu}, \eta_{i_M})^{\mathrm{T}} \sim \mathrm{N}_2(0, \Sigma_\mu)$, with $\Sigma_\mu = \mathrm{diag}(\sigma^2_\mu\delta_i, \sigma^2_M\delta_i)$.

Although there exists other possible methods for accommodating local adaptivity in the latent trajectories, state equations (2.18)–(2.19) along with observation equations (2.1)–(2.3) form an appealing nonlinear logistic state space model for adaptive dynamic network inference, which characterizes the latent positions at time t_{i+1} as a first-order stochastic Taylor expansion of the same quantities at t_i. This choice is further appealing in improving scalability of the inference procedures, while facilitating implementation of tractable online updating and prediction strategies by adapting available techniques associated to state space models. Although previous state equations can be easily extended to model higher order derivatives for the latent coordinates' trajectories and their local instantaneous means, equations (2.18)–(2.19) prove to be sufficiently flexible in inducing adaptive patterns according to our results.

Maintaining formulation (2.1)–(2.3) is appealing in our motivating application. This construction recalls Hoff (2008) static eigenmodel, providing a flexible class of latent variables models for social networks which allows for across-node heterogeneity while accommodating several topological properties. According to Hoff (2008), model (2.1)–(2.3) generalizes stochastic block models (Nowicki and Snijders, 2001) and latent distance models (Hoff et al.,

2002), and hence can suitably accommodate block structures, homophily behaviors as well as transitive contact patterns. These properties are – potentially – key factors underlying our face-to-face interaction data. For instance, during school hours or lunch times the contact networks are expected to exhibit block structures due to shared environments by students belonging to the same class or groups of classes. Breaks are instead potentially associated with transitive patterns arising from friendship among students in different classes or homophily by gender.

2.2.2 Posterior computation

Current methodologies leveraging state space formulations in dynamic network analysis require several layers of approximation to perform statistical inference, without theory available to justify accuracy. The reason behind these approximate methods is that the observation equation in (2.1)–(2.3) is neither Gaussian, nor linear and hence efficient algorithms (Durbin and Koopman, 2002) for inference in state space models (Durbin and Koopman, 2012) can't be directly applied. Differently from available contributions, we develop a novel and efficient Gibbs sampler to obtain samples from the exact posterior of $\{\pi(t) : t \in \mathbb{T}\}$ based on the statistical model (2.1)–(2.3) and priors (2.18)–(2.19). This is accomplished by efficiently exploiting Pòlya-gamma augmented data (Polson et al., 2013) to obtain a Gaussian state space model for transformed data. By block-sampling in turn the latent coordinate processes for each node v conditionally on the latent positions of the others $u = 1, \dots, V,\ u \neq v$, we further obtain a linear observation equation, which allows us to apply standard results from Kalman filtering (Durbin and Koopman, 2012). Specifically, the Gibbs sampler for Bayesian inference in our LADY network model alternates between steps outlined in Algorithm 2.

Algorithm 2 Gibbs sampler for the LADY network model

---

[1] Sample Pólya-gamma augmented data

for each $l = 1, \dots, V(V-1)/2$ and $t_i = t_1, \dots, t_n$ **do**

Update each augmented data $\omega_l(t_i)$ from the full conditional Pólya-gamma

$$\omega_l(t_i) \mid - \sim \text{PG}\left\{1, \mu(t_i) + \mathcal{L}(X(t_i)X(t_i)^{\text{T}})_l\right\}.$$

end for

---

[2] Update $\mu = \{\mu(t_1), \dots \mu(t_n)\}^{\text{T}}$, $\mu' = \{\mu'(t_1), \dots \mu'(t_n)\}^{\text{T}}$, $M = \{M(t_1), \dots M(t_n)\}^{\text{T}}$.

Adapting representation (2.10) to our model, the likelihood for $\mu = \{\mu(t_1), \dots \mu(t_n)\}^{\mathrm{T}}$ given the Pólya-gamma augmented data and the latent coordinate processes is

$$\begin{aligned}
&\propto \prod_{i=1}^{n} \exp\left[-\sum_{l=1}^{V(V-1)/2} \frac{\omega_l(t_i)}{2} \left\{ (\mathcal{L}(A_{t_i})_l - 0.5)/\omega_l(t_i) - \mu(t_i) - \mathcal{L}(X(t_i)X(t_i)^{\mathrm{T}})_l \right\}^2 \right], \\
&\propto \prod_{i=1}^{n} \exp\left[-\frac{\sum_{l=1}^{V(V-1)/2} \omega_l(t_i)}{2} \left\{ \mu(t_i)^2 - 2\mu(t_i) \frac{\sum_{l=1}^{V(V-1)/2} \psi_l(t_i)}{\sum_{l=1}^{V(V-1)/2} \omega_l(t_i)} \right\} \right], \\
&\propto \prod_{i=1}^{n} \exp\left[-\frac{\sum_{l=1}^{V(V-1)/2} \omega_l(t_i)}{2} \left\{ \frac{\sum_{l=1}^{V(V-1)/2} \psi_l(t_i)}{\sum_{l=1}^{V(V-1)/2} \omega_l(t_i)} - \mu(t_i) \right\}^2 \right], \qquad (2.20)
\end{aligned}$$

with $\psi_l(t_i) = \mathcal{L}(A_{t_i})_l - 0.5 - \omega_l(t_i)\mathcal{L}(X(t_i)X(t_i)^{\mathrm{T}})_l$. Hence, letting $\omega_{\mu(t_i)} = \sum_{l=1}^{V(V-1)/2} \omega_l(t_i)$ and $\mathcal{L}(A)_{\mu(t_i)} = \sum_{l=1}^{V(V-1)/2} \psi_l(t_i) / \sum_{l=1}^{V(V-1)/2} \omega_l(t_i)$ for each $i = 1, \dots, n$, it is easy to notice that (2.20) is the likelihood for the baseline vector μ arising from the model

$$\mathcal{L}(A)_{\mu(t_i)} = \mu(t_i) + \epsilon_{\mu(t_i)}, \quad i = 1, \dots, n, \qquad (2.21)$$

where $\epsilon_{\mu(t_i)} \sim \mathrm{N}(0, 1/\omega_{\mu(t_i)})$ independently for $i = 1, \dots, n$. Hence, combining the observation equation (2.21) for transformed data with state equations (2.19), we obtain a linear Gaussian state space model, which allows simple updating for $\mu = \{\mu(t_1), \dots \mu(t_n)\}^{\mathrm{T}}$, $\mu' = \{\mu'(t_1), \dots \mu'(t_n)\}^{\mathrm{T}}$ and $M = \{M(t_1), \dots M(t_n)\}^{\mathrm{T}}$ via the simulation smoother of Durbin and Koopman (2002) . This has a computational complexity of $O(n)$ and diffuse initialization at t_1, $\{\mu(t_1), \mu'(t_1), M(t_1)\}^{\mathrm{T}} \sim \mathrm{N}_3(0, 100 \cdot \mathrm{I}_3)$.

---

[3] Sample states matrices $X(t_1), \dots, X(t_n)$**,** $X'(t_1), \dots, X'(t_n)$ **and** $U(t_1), \dots, U(t_n)$
for $v = 1, \dots, V$ **do**

Sample $X_v(t_i)$, $X'_v(t_i)$, $U_v(t_i)$, $t_i = t_1, \dots, t_n$ given $X_{(-v)} = \{X_u(t_i) : u \neq v, t_i = t_1, \dots t_n\}$, considering a similar derivation to the one in step **[2]**.

1. Let $X_{(-v)}(t_i)$ the $(V-1) \times R$ coordinate matrix at t_i with the vth row held out

2. Define the $(V-1) \times 1$ vector of transformed Gaussian data for the observation equation $\mathcal{L}(A)_{X_v(t_i)} = \mathrm{diag}\{\Omega_{(-v)}(t_i)\}^{-1}\{A_{t_i[(-v)v]} - 0.5 \cdot 1_{V-1} - \mu(t_i)\Omega_{(-v)}(t_i)\}$, where $A_{t_i[(-v)v]}$ denotes the vth column of A_{t_i} after discarding the vth row and $\Omega_{(-v)}(t_i)$ the $(V-1) \times 1$ vector of corresponding Pòlya-gamma augmented data

3. Update states $\{X_{vr}(t_i), X'_{vr}(t_i), U_{vr}(t_i) : r = 1, \dots, R,\ t_i = t_1, \dots, t_n\}$ by applying the simulation smoother of Durbin and Koopman (2002) to the state space model having state equation (2.18) and observation equation

$$\mathcal{L}(A)_{X_v(t_i)} = X_{(-v)}(t_i)X_v(t_i) + \epsilon_{X_v(t_i)}, \quad i = 1, \dots, n,$$

with $\epsilon_{X_v(t_i)} \sim \mathrm{N}_{V-1}(0, \mathrm{diag}\{\Omega_{(-v)}(t_i)\}^{-1})$ independently for $i = 1, \ldots, n$. The simulation smoother is initialized with diffuse states $\{X_{vr}(t_1), X'_{vr}(t_1), U_{vr}(t_1)\}^{\mathrm{T}} \sim \mathrm{N}_3(0, 100 \cdot \mathrm{I}_3)$ for each $r = 1, \ldots, R$.

end for

---

[4] Update the hyperprior for the noise variances σ^2_μ and σ^2_M

Letting $\sigma^2_\mu \sim$ Inv-Ga(a_μ, b_μ) and $\sigma^2_M \sim$ Inv-Ga(a_M, b_M) the hyperpriors for the noise variances in the states equation (2.19), their full conditional distribution is

$$\sigma^2_\mu \mid - \;\sim\; \text{Inv-Ga}\left[a_\mu + \frac{n-1}{2}, b_\mu + \frac{1}{2}\sum_{i=1}^{n-1} \frac{\{\mu'(t_{i+1}) - \mu'(t_i) - M(t_i)\delta_i\}^2}{\delta_i}\right],$$
$$\sigma^2_M \mid - \;\sim\; \text{Inv-Ga}\left[a_M + \frac{n-1}{2}, b_M + \frac{1}{2}\sum_{i=1}^{n-1} \frac{\{M(t_{i+1}) - M(t_i)\}^2}{\delta_i}\right].$$

---

[5] Update the noise variances $\sigma^2_{X_{vr}}$ and $\sigma^2_{U_{vr}}$, $v = 1, \ldots, V$ and $r = 1, \ldots, R$

for $v = 1, \ldots, V$ and $r = 1, \ldots, R$ **do**

Letting $\sigma^2_{X_{vr}} \sim$ Inv-Ga(a_X, b_X), $\sigma^2_{U_{vr}} \sim$ Inv-Ga(a_U, b_U), the hyperpriors for the noise variances in the states equation (2.18), their full conditional distribution is

$$\sigma^2_{X_{vr}} \mid - \;\sim\; \text{Inv-Ga}\left[a_X + \frac{n-1}{2}, b_X + \frac{1}{2}\sum_{i=1}^{n-1} \frac{\{X'_{vr}(t_{i+1}) - X'_{vr}(t_i) - U_{vr}(t_i)\delta_i\}^2}{\delta_i}\right],$$
$$\sigma^2_{U_{vr}} \mid - \;\sim\; \text{Inv-Ga}\left[a_U + \frac{n-1}{2}, b_U + \frac{1}{2}\sum_{i=1}^{n-1} \frac{\{U_{vr}(t_{i+1}) - U_{vr}(t_i)\}^2}{\delta_i}\right].$$

end for

Given MCMC chains for $\mu(t_1), \ldots, \mu(t_n)$ and $X(t_1), \ldots, X(t_n)$, posterior samples for latent similarities $S(t_1), \ldots, S(t_n)$ and edge probability vectors $\pi(t_1), \ldots, \pi(t_n)$ can be easily derived by applying equations (2.3) and (2.2), respectively. To estimate R, we repeat the above algorithm for increasing R, stopping when there is no substantial improvement in in-sample edge prediction based on area under the ROC curve. As in-sample prediction strategies may suffer from over-fitting issues, we additionally assess our choice of R by exploring out-of-sample prediction and forecasting performance.

The proposed Gibbs sampler allows substantial improvements compared to the procedures outlined in Section 2.1. Replacing GP with nGPs reduces the computational burden from $O(n^3)$ to $O(n)$, with n denoting the length of the time series, while also allowing flexible locally varying smoothness. Moreover, the state space representation further allows efficient dynamic updating and forecasting procedures exploiting results from Kalman filter.

2.2.3 Forecasting, predicting and online updating

Forecasting and predicting

Forecasting a future network based on past data is particularly appealing in our motivating application as it allows to timely design specific policies, such as outbreak prevention. For example, if a subject contract a disease at time t_n, forecasts at time t_{n+1} are a key to understand which children are at risk of contagion as a result of face-to-face proximity interaction.

Recalling strategies outlined at the end of Section 2.1.3, one-step-ahead forecasts for a future network $\mathcal{L}(A_{t_{n+1}})$ can be obtained from the expectation of the forecasted predictive distribution as in equation (2.11). In this respect, an appealing feature of our LADY network model – compared to procedures developed in Section 2.1 – is that the entire posterior distribution for $\pi_l(t_{n+1})$, can be easily obtained by applying the equation

$$\pi_l(t_{n+1}) = \left[1 + e^{-\{\mu(t_n)+\delta_n\mu'(t_n)\}-\{X_v(t_n)+\delta_n X'_v(t_n)\}^{\mathrm{T}}\{X_u(t_n)+\delta_n X'_u(t_n)\}}\right]^{-1}, \tag{2.22}$$

to the posterior samples of the latent states at time t_n, where v and u are the nodes corresponding to pair l, for each $l = 1, \ldots, V(V-1)/2$. This procedure is substantially faster than the one proposed in Section 2.1.3 which requires re-running posterior computations adding to the observed dataset $\mathcal{L}(A_{t_1}), \ldots, \mathcal{L}(A_{t_n})$ a new vector $\mathcal{L}(A_{t_{n+1}})$ of missing values.

Recalling our data set structure, beside forecasting contacts at the next time within the first day, it is additionally of interest to predict the whole network dynamics in the second day, based on estimates from the previous day. In particular, letting $\mathcal{L}(A^*_{t_i})$ the random vector denoting presence or absence of contacts among pairs of nodes at time t_i in the second day, we predict edges $\mathcal{L}(A^*_{t_i})_l$, $l = 1, \ldots, V(V-1)/2$ by focusing on the expected value of the posterior predictive distribution

$$\begin{aligned} \mathrm{E}\{\mathcal{L}(\mathcal{A}_{t_i})_l \mid \mathcal{L}(A_{t_1}), \ldots, \mathcal{L}(A_{t_n})\} &= \mathrm{E}_{\pi_l(t_i)}[\mathrm{E}\{\mathcal{L}(\mathcal{A}_{t_i})_l \mid \pi_l(t_i)\} \mid \mathcal{L}(A_{t_1}), \ldots, \mathcal{L}(A_{t_n})] \\ &= \mathrm{E}\{\pi_l(t_i) \mid \mathcal{L}(A_{t_1}), \ldots, \mathcal{L}(A_{t_n})\}, \end{aligned} \tag{2.23}$$

for each $l = 1, \ldots, V(V-1)/2$ and time t_i, where the expectation in (2.23) simply coincides with the posterior mean of the edge probability trajectories. Clearly equation (2.23) relies on the assumption that dynamic contact networks at the second day are governed by the same statistical model underlying data at the first day. Although this assumption is not necessarily valid in other analyses of real world dynamic networks, it provides a reasonable choice in our motivating application. In fact, as the overall schedule of a school remains in general substantially unchanged across subsequent days, it is reasonable to expect that the contact network at a given time in the second day may be governed by similar underlying patterns to those occurring at the same time in the first day.

Online updating

Online updating is particularly appealing in several real world dynamic networks. Recalling our motivating application, once the model has been estimated on data $A_{t_1}, \ldots, A_{t_n}$, new contact networks $A_{t_{n+1}}, \ldots, A_{t_{n+n^*}}$ can stream in. Hence, in order to timely update policies, such as disease surveillance, it is important to have a fast online updating algorithm for the posterior of the edge probability vectors $\pi(t_{n+1}), \ldots, \pi(t_{n+n^*})$, including data from new networks $A_{t_{n+1}}, \ldots, A_{t_{n+n^*}}$, without the need to rerun posterior computation for the whole data from t_1 to t_{n+n^*}.

Current procedures for dynamic network inference are insufficiently flexible to accommodate online updating strategies. Our LADY network model is instead amenable to such fast dynamic updating due to the latent Kalman filter formulation. Conditionally on the posterior means and covariances of the latent states at time n and the estimated noise variances in the state equation, our online updating algorithm efficiently cycles between steps **[1]**, **[2]** and **[3]** only for new data $A_{t_{n+1}}, \ldots A_{t_{n+n^*}}$, with the simulation smoother in **[2]** and **[3]** initialized at t_{n+1} using the one step ahed predictive distribution from the Kalman filter. Specifically we initialize states $\{\mu(t_{n+1}), \mu'(t_{n+1}), M(t_{n+1})\}^{\mathrm{T}}$ at t_{n+1} in **[2]** by assuming $\{\mu(t_{n+1}), \mu'(t_{n+1}), M(t_{n+1})\}^{\mathrm{T}}$ are distributed according to

$$\mathrm{N}_3(T_n[\hat{\mathrm{E}}\{\mu(t_n)\}, \hat{\mathrm{E}}\{\mu'(t_n)\}, \hat{\mathrm{E}}\{M(t_n)\}]^{\mathrm{T}}, T_n\hat{\Gamma}_{\mu,n}T_n^{\mathrm{T}} + Q_n\mathrm{diag}(\hat{\sigma}^2_{\mu}\delta_n, \hat{\sigma}^2_{M}\delta_n)Q_n^{\mathrm{T}}),$$

where $[\hat{\mathrm{E}}\{\mu(t_n)\}, \hat{\mathrm{E}}\{\mu'(t_n)\}, \hat{\mathrm{E}}\{M(t_n)\}]^{\mathrm{T}}$ is the vector of posterior means for the states at time n, $\hat{\Gamma}_{\mu,n}$ is their 3×3 posterior covariance matrix and $\hat{\sigma}^2_{\mu}$, $\hat{\sigma}^2_{M}$ are the estimated state noise variances using the initial dataset from t_1 to t_n. A similar initialization is considered in **[3]** for $\{X_{vr}(t_{n+1}), X'_{vr}(t_{n+1}), U_{vr}(t_{n+1})\}^{\mathrm{T}}$ obtaining

$$\mathrm{N}_3(T_n[\hat{\mathrm{E}}\{X_{vr}(t_n)\}, \hat{\mathrm{E}}\{X'_{vr}(t_n)\}, \hat{\mathrm{E}}\{U_{vr}(t_n)\}]^{\mathrm{T}}, T_n\hat{\Gamma}_{X_{vr},n}T_n^{\mathrm{T}} + Q_n\mathrm{diag}(\hat{\sigma}^2_{X_{vr}}\delta_n, \hat{\sigma}^2_{U_{vr}}\delta_n)Q_n^{\mathrm{T}}),$$

for $v = 1, \ldots, V$ and $r = 1, \ldots, R$. Although the algorithm fixes the hyperparameters corresponding to the noise variances in the state equations at their posterior means, these quantities are time-constant and hence can be accurately estimated by borrowing information across the whole time window. It is however straightforward to modify the algorithm to update the posterior distribution also for these quantities given the latent states stored in the initial sampling from t_1 to t_n and the updated ones from t_{n+1} to t_{n+n^*}. This strategy may be useful when n is small. We found few differences between the two procedures in our simulations and hence prefer the first strategy.

It is also worth noticing that our procedure does not update $\pi(t_1), \ldots, \pi(t_n)$, given new data $A_{t_{n+1}}, \ldots A_{t_{n+n^*}}$, but focuses only on the posterior of $\pi(t_{n+1}), \ldots, \pi(t_{n+n^*})$. This may

affect the ability of our procedures to properly propagate uncertainty and reduce performance in updating $\pi(t_{n+1}), \ldots, \pi(t_{n+n^*})$. To mitigate this issue, while maintaing computational scalability, we run online updating for data $A_{t_{n-j}}, \ldots, A_{t_n}, A_{t_{n+1}}, \ldots A_{t_{n+n^*}}$ instead of only $A_{t_{n+1}}, \ldots A_{t_{n+n^*}}$. We found this correction to improve performance even when a small j number of past networks is included along with new data.

2.2.4 Model checking

Before moving to simulations and application, it is worth developing procedures for model evaluation. Assessing the performance of a statistical model in characterizing the observed and future data is fundamental to guarantee robust inference; refer to Chapter 6 in Gelman et al. (2014) for common procedures in model checking within the Bayesian paradigm. This practice is even more important in the network framework, providing complex data structures and comprising a wide set of possible statistical models.

Our methods fall within the class of latent variable modeling of dynamic networks. Although these procedures are appealing in accommodating heterogenous structures and facilitate tractable inference strategies, the types of higher-order dependencies included may be limited by the conditional independence assumption and the characterization of the latent variables. Exponential random graph models overcome this issue by explicitly parameterizing interdependence among edges, but typically rely on restrictive homogeneity assumptions.

Although conditional independence may at first appear overly-restrictive, multivariate categorical data – such as a vectorized adjacency matrix – can be expressed as conditionally independent given a sufficient number of latent factors without imposing any assumptions on the joint distribution; see for example Dunson and Xing (2009) for recent theoretical results. Investigating previous property requires analysis of the posterior predictive distribution $\mathrm{p}\{\mathcal{L}(\mathcal{A}_{t_1}), \ldots, \mathcal{L}(\mathcal{A}_{t_n}) \mid \mathcal{L}(A_{t_1}), \ldots, \mathcal{L}(A_{t_n})\}$ defined as

$$\int \prod_{i=1}^{n} \prod_{l=1}^{V(V-1)/2} \mathrm{p}\{\mathcal{L}(\mathcal{A}_{t_i})_l \mid \pi_l(t_i)\} \mathrm{d}\Pi\{\pi(t_1), \ldots, \pi(t_n) \mid \mathcal{L}(A_{t_1}), \ldots, \mathcal{L}(A_{t_n})\},$$

where $\mathrm{p}\{\mathcal{L}(\mathcal{A}_{t_i})_l \mid \pi_l(t_i)\}$ is the Bernoulli probability mass function in (2.1) for the univariate random variable $\mathcal{L}(\mathcal{A}_{t_i})_l$ measuring presence or absence of a contact among the lth pair of nodes at time t_i. $\Pi\{\pi(t_1), \ldots, \pi(t_n) \mid \mathcal{L}(A_{t_1}), \ldots, \mathcal{L}(A_{t_n})\}$ is instead the joint posterior distribution for the edge probability trajectories given observed data $\mathcal{L}(A_{t_1}), \ldots, \mathcal{L}(A_{t_n})$.

Although the posterior predictive distribution is not analytically available, it is straightforward to simulate from $\mathrm{p}\{\mathcal{L}(\mathcal{A}_{t_1}), \ldots, \mathcal{L}(\mathcal{A}_{t_n}) \mid \mathcal{L}(A_{t_1}), \ldots, \mathcal{L}(A_{t_n})\}$ exploiting equation (2.1) along with posterior samples for $\pi_l(t_i)$, $l = 1, \ldots, V(V-1)/2$ and $i = 1, \ldots, n$. Specifically,

for each MCMC sample of $\pi_l(t_i)$, $l = 1, \ldots, V(V-1)/2$ and $i = 1, \ldots, n$, we simulate contacts among pairs of nodes from conditionally independent Bernoulli random variables given their corresponding $\pi_l(t_i)$, obtaining samples from $\mathrm{p}\{\mathcal{L}(\mathcal{A}_{t_1}), \ldots, \mathcal{L}(\mathcal{A}_{t_n}) \mid \mathcal{L}(A_{t_1}), \ldots, \mathcal{L}(A_{t_n})\}$.

Exploiting samples from $\mathrm{p}\{\mathcal{L}(\mathcal{A}_{t_1}), \ldots, \mathcal{L}(\mathcal{A}_{t_n}) \mid \mathcal{L}(A_{t_1}), \ldots, \mathcal{L}(A_{t_n})\}$, we evaluate the performance of our model in accommodating specific dynamic topological structures characterizing observed data by investigating their density arising from the posterior predictive distribution. Recalling our motivating application, we focus on dynamic network density and node degree, along with time-varying homophily by class and gender. The last two quantities are measured by the assortativity coefficient; see Newman (2003), equation 2.

When the interest is on disease surveillance and outbreak prevention, time-varying network density is a key quantity in summarizing the total number of contacts including those leading to potential contagion. Node degrees are instead appealing in providing a measure of the number of subjects at risk of contagion if a child contract a disease at a certain time. Evolution of homophily structures across time and environmental conditions are instead of interest from a social science perspective; see for example Stehlé et al. (2013) for a study on gender homophily in face-to-face contact networks from an aggregated perspective.

Although we compare quantities from the posterior predictive distribution to those arising from the same data obtained to estimate such distribution, there is no guarantee we will obtain a good matching. Even the best fit, may lead to substantially biased inference if the statistical model is insufficiently flexible in accommodating specific topological structures characterizing the observed networks. Our goal is assessing to what extent the LADY network model can accommodate such properties. We additionally perform out-of-sample model checking by evaluating forecasting and predictive performance.

2.2.5 Simulation study

We implement a simulation study to assess the performance of our LADY network model in correctly estimating varying smoothness patterns, accommodating streaming data and predicting future networks. We consider a dynamic networks with $V = 15$ nodes monitored for $n = 50$ equally spaced times from $t_1 = 0$ to $t_{50} = 15$. The time varying edges $\mathcal{L}(A_{t_i})_l$, $l = 1, \ldots, V(V-1)/2$, are simulated from model (2.1) with edge probabilities evolving in time across five regimes mimicking – in a simple version – possible scenarios associated to our face-to-face children interactions; refer to Figure 2.6 for a description of the true generative process underlying edge probabilities.

Specifically we consider three classes comprising five students each and define also a gender variable. There are 8 males and 7 females almost equally divided in the different classes.

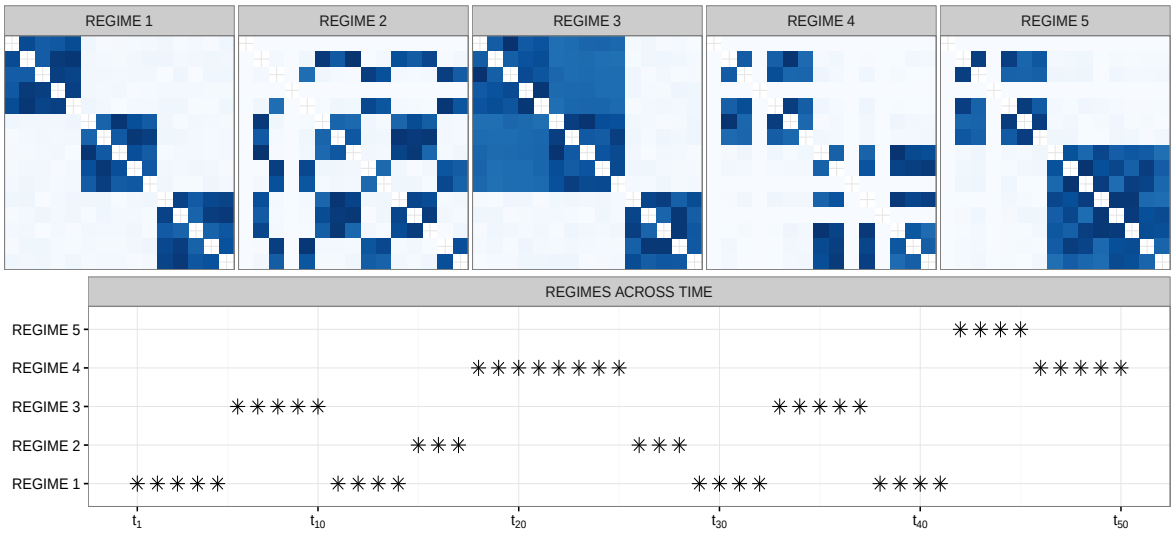


FIGURE 2.6: Upper panels: true edge probabilities – arranged in matrix form – for the regimes in the simulation; colors go from white to dark blue as the probability goes from 0 to 1. Lower panels: graphical representation showing for every time which regime – i.e. edge probabilities – is considered to simulate the data.

The first regime represents school hours and is characterized by high probability of contact between students in the same class, and low chance of face-to-face interaction among children in different classes. The second regime encodes high gender homophily which may arise during breaks in which all children can interact and reveal friendship structures; see also descriptive analyses in Stehlé et al. (2013). The third regime is characterized by the first two classes sharing the same room – for school hours or during breaks – and hence, beside high within class probabilities of contact, we observe also moderately high chance of contact between students in the first two classes. Regime four represents a possible scenario we have observed in our data during lunch times and confirmed in Figure 10 of Stehlé et al. (2011). Specifically students in the second class are almost equally divided in two groups with one attending lunch with children in the first class and the other with those in the third class. Hence we observe two block structures, with an additional subset of the students having no contacts with the others in leaving the school for lunch times. Regimes five and four may also reasonably characterize contact networks during the end of the school day, with groups of students gathering in the same room and progressively leaving the school.

Although this generative mechanism represents a substantially simplified version of our complex data set, the basic underlying structures and the rapid changes in specific topological patterns are in line with those we expect in our application. Moreover considering edge probabilities obtained under scenarios different than (2.2)–(2.3) and evolving in time across a regime-switching process instead of the state equations (2.18)–(2.19) has the additional benefit of providing a more fair validation of our LADY network methodology, as the true edge probability processes are not generated from our model.

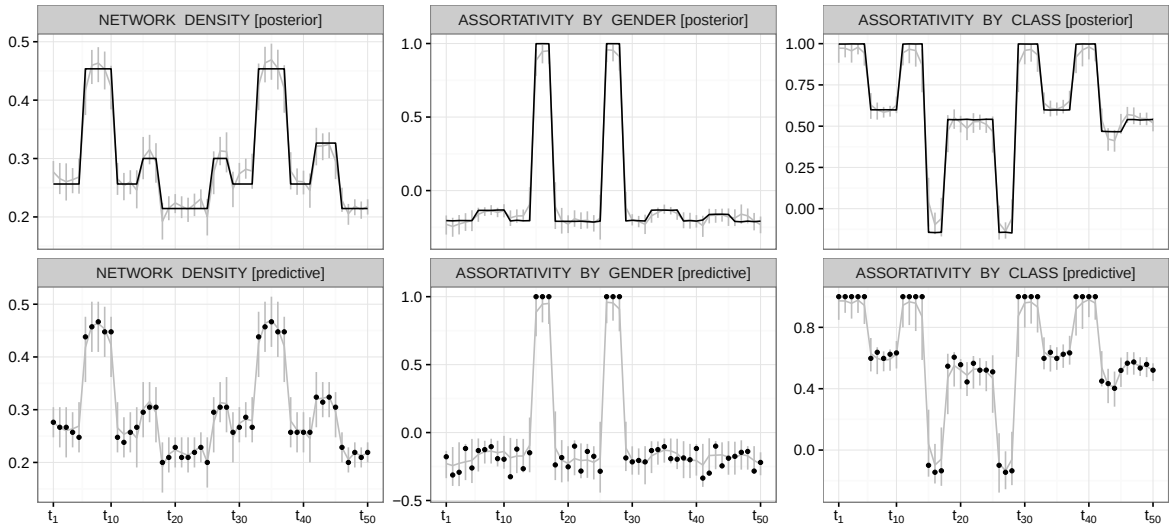


FIGURE 2.7: Upper panels: time-varying posterior mean (gray line) and pointwise 0.95 highest posterior density intervals (gray segments) for expected network summary statistics covering network density, assortativity by gender and by class; true values are represented by the black line. Lower panels: for the same summary statistics, time-varying mean (gray line) and pointwise 0.95 predictive intervals (gray segments) obtained from the posterior predictive distribution; black dots represents the corresponding time-varying network measures computed from the simulated data.

In performing posterior inference under our LADY network model, we choose diffuse priors for the noise variances in the state equations by letting $a_\mu = a_M = a_X = a_U = b_\mu = b_M = b_X = b_M = 0.01$, and run $5{,}000$ Gibbs iterations discarding the first $1{,}000$. To learn R we consider our selection procedure by performing posterior computation for increasing $R = 1, 2, \ldots$ and provide posterior inference for the model having R total latent coordinates such that $\text{AUC}_{R+1} - \text{AUC}_R < 0.01$. The AUC for the model with only the baseline process is 0.59, while those for formulations with $R = 1$ and $R = 2$ are 0.97 and 0.99, respectively. Increasing the coordinates from $R = 2$ to $R = 3$ we found no substantial improvement with an AUC of 0.992. Hence, consistently with our procedure we provide inference with $R = 2$.

Mixing via effective sample sizes for the quantities of interest is on similar values to those obtained in Section 2.1.5. For the same quantities we assess convergence by investigating the Gelman and Rubin (1992) potential scale reduction factors (PSRF) – computed as in the simulation in Section 2.1.5. The median of the PSRFs for the chains of the edge probabilities $\pi_l(t_i)$, $l = 1, \ldots, V(V-1)/2$, and $i = 1, \ldots, n$, is 1.05, with the 99% of these PSRFs being less than 1.3, providing evidence that convergence has been reached. Similar results are obtained for the PSRF of selected network measures of interest including time-varying expected homophily by gender and class, expected density $\text{E}[\sum_{l=1}^{V(V-1)/2} \mathcal{L}(\mathcal{A}_{t_i})_l / \{V(V-1)/2\}] = \sum_{l=1}^{V(V-1)/2} \pi_l(t_i)/\{V(V-1)/2\}$, and expected node degree $\text{E}\{\sum_{l \in L_v} \mathcal{L}(\mathcal{A}_{t_i})_l\} = \sum_{l \in L_v} \pi_l(t_i)$ for each $v = 1, \ldots, V$ and $i = 1, \ldots, n$, where L_v is the set of pairs of nodes $\{(v, u) : u \in \mathbb{V}, u \neq v\}$. As the expectation of the assortativity coefficient is not analytically available as a function of the edge probabilities, we derive posterior samples for the assortativity coefficients via Monte Carlo methods. Specifically for each posterior sample of $\pi_l(t_i)$, $l = 1, \ldots, V(V-1)/2$

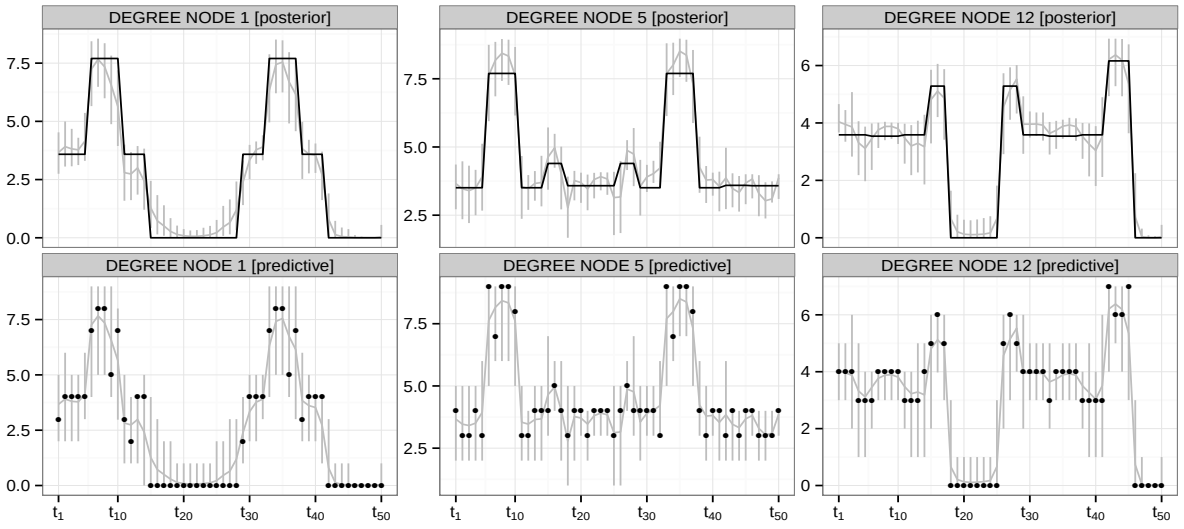


FIGURE 2.8: Upper panels: time-varying posterior mean (gray line) and pointwise 0.95 highest posterior density intervals (gray segments) for the expected degree of selected nodes; true values are represented by the black line. Lower panels: for the same summary statistics, time-varying mean (gray line) and pointwise 0.95 predictive intervals (gray segments) obtained from the posterior predictive distribution; black dots represents the corresponding time-varying node degrees computed from the simulated data.

and $i = 1, \ldots, n$, we simulate 100 networks from (2.1) and obtain approximated samples from the posterior distribution of the time-varying expected gender and class assortativity by computing these coefficients for the 100 simulated networks and averaging them.

As shown in the upper panels of Figures 2.7 and 2.8, enforcing local adaptivity in the time-varying trajectories of the edge probabilities while accommodating across-node heterogeneity, allow us to accurately characterize rapid changes in the true expected measures of interest, including time-varying network density, homophily structures and node degrees. Moreover, although we rely on a latent variable representation which does not explicitly parameterize dependencies among edges, our LADY network model can accurately accommodate topological structures of interest characterizing the observed dynamic networks. This is highlighted in the lower panels of Figures 2.7 and 2.8, comparing summary statistics form the observed data, with their distribution arising from the posterior predictive distribution, consistently with the model checking procedures outlined in Section 2.2.4. All observed quantities are inside the 0.95 posterior predictive intervals, suggesting good fit.

Table 2.4 compares forecasting and predictive performance of our model to those associated with two selected competitors, for times from t_{45} to t_{50}. Specifically we compare out-of-sample edge prediction of our LADY network model, with results obtained under the Gaussian process dynamic network from Section 2.1 and Hanneke et al. (2010) temporal ERGM (TERGM). Our procedure in Section 2.1 relies on the same model formulation (2.1)–(2.3) but does not allow varying smoothness over time. Hanneke et al. (2010) TERGM is instead a substantially different model which explicitly account for the effect of topological structures

in model formulation, rather than considering latent variables.

In performing posterior computation under the Gaussian process dynamic network model, we consider the same hyperparameters settings of the simulation study 2.1.5, fixing $R = 2$ – as in the LADY network model for this simulation – and increasing the GP length scales κ_μ and κ_x from 0.05 to 0.1 to improve performance in capturing sudden changes. We considered several different choices of length scales and selected the one with the best performance. The TERGM is instead estimated via bootstrapped pseudolikelihood procedures (Desmarais and Cranmer, 2012) exploiting `R` packages `btergm` and `xergm`. In defining the linear predictor under the TERGM representation we consider a p^* ERGM specification with alternating k-starts (Robins et al., 2007a) and triangle effects to account for transitivity patterns and include gender and class variables both in terms of main and homophily effects – using functions `nodefactor()` and `nodematch()`, respectively. Finally we account for temporal dependence by including a stability term which measures the tendency of an edge – or non-edge – at time t_i to be also observed – or not observed – at the next time t_{i+1}. The main effects of node covariates as well as homophily by gender were not significant, hence we drop these variables in assessing forecasting and predictive performance. It is also worth noticing how considering time constant homophily effects prevent the TERGM from capturing the strong gender homophily in the two time windows shown in Figure 2.7. We also attempted an actor-oriented model using `R` package `RSiena` but found computational issues in terms of convergence for the time-specific parameters in the rate function.

For each time t_i, $i = 44, \ldots, 49$ forecasting performance is assessed by estimating the three different models using data from t_1 to t_i, and forecasting edges at time t_{i+1}. Forecasts under the GP dynamic network follow procedures outlined at the end of Section 2.1.3. Under the TERGM, forecasting of future networks proceed via simulation methods using the `gof()` function in the `R` package `ergm`; see also Hunter et al. (2008b). Finally for our LADY network model we consider a potentially more challenging strategy which proceeds by first online updating the posterior distribution of the edge probabilities at t_i using estimates from t_1 to t_{i-1} according to procedures in Section 2.2.3 – with $j = 5$ – and then forecasts edges at time t_{i+1} by applying the forecasting methods outlined in Section 2.2.3 to the posterior distribution of the edge probabilities from the online updating. Joining online updating and forecasting is appealing in providing a fast strategy which avoids re-running posterior computation for the whole data set when a one-step-ahead forecast in required.

In evaluating predictive performance we instead simulate new networks $A^*_{t_{45}}, \ldots, A^*_{t_{50}}$ from the same mechanism considered to generate training data – see Figure 2.6 – and compare the AUC based on the estimates from the three competing methods – exploiting training data $A_{t_1}, \ldots, A_{t_{50}}$ – and the new simulated networks $A^*_{t_{45}}, \ldots, A^*_{t_{50}}$. Edge prediction under our

	AREA UNDER THE ROC CURVE					
	t_{45}	t_{46}	t_{47}	t_{48}	t_{49}	t_{50}
LADY Network: forecasting performance	0.94	0.82	0.99	0.97	0.99	0.98
LADY Network: predictive performance	0.97	0.98	0.99	0.99	0.98	0.99
Dynamic GP Network: forecasting performance	0.91	0.81	0.91	0.83	0.97	0.97
Dynamic GP Network: predictive performance	0.97	0.98	0.97	0.97	0.98	0.98
TERGM: forecasting performance	0.92	0.78	0.95	0.91	0.97	0.92
TERGM: predictive performance	0.98	0.83	0.97	0.95	0.97	0.96

TABLE 2.4: For our model and selected competitors, forecasting and predictive performance for data from t_{45} to t_{50}.

LADY network model and the GP dynamic network model in Section 2.1 use equation (2.23). For TERGM we exploit again simulation procedures from the `gof()` function.

As shown in Table 2.4 our procedure is characterized by improved forecasting and predictive performance compared to GP dynamic network model and TERGM. The dynamic GP network in Section 2.1 accommodates heterogenous structures but assumes time-constant smoothness. Hanneke et al. (2010) explicitly account for several higher-order dependencies but force the model parameters to be shared among nodes and typically constant across time. These assumptions lead to reduced performance compared to our procedure which incorporates both across-node heterogeneity and time-varying smoothness. These results additionally highlight the good performance of our online updating procedures.

As expected forecasting performance decreases at t_{46} since the models have no experience of sudden regime changes. However it is interesting to notice how accommodating locally adaptive processes provides rapid adjustments of the estimates to new regimes once they are observed, improving subsequent forecasts. Dynamic GP network model requires more times to adapt to new regimes due to the time-constant smoothness assumption. Reduced performance at t_{46} is not an issue when predicting new networks generated under the same mechanism, as the whole training data set $A_{t_1}, \ldots, A_{t_{50}}$ already inform on regime changes. Clearly in the out-of-sample prediction exercise, performance depends on the flexibility of the model in accommodating rapid regime changes along with their associated network structures.

Inference under our LADY network model takes ≈ 30 minutes for posterior computation, ≈ 6 minutes for online updating and ≈ 1 second for forecasting. Dynamic GP network model is substantially slower in performing posterior computation – ≈ 95 minutes – due to the computational bottlenecks of the Gaussian processes. Estimation under TERGM is instead faster than previous procedures, but simulations methods for forecasting and predictions require more time. It is additionally important to underline that our algorithms are based on

a naive R (version 3.1.1) implementation in a machine with one Intel Core i5 2.3GHz processor and 4GB of RAM.

2.2.6 Application to fate-to-face human interaction data

We apply our LADY network model outlined in Section 2.2.1 to the face-to-face contact data $A_{t_1}, \ldots, A_{t_{51}}$ described in Section 1.1.2, under the same settings of the simulation study, with $R = 4$. We select $R = 4$ as adding a further dimension increases the area under the ROC curve by less than 0.01, while $\text{AUC}_4 - \text{AUC}_3 > 0.01$. In performing posterior inference we consider $5{,}000$ Gibbs samples with a burn-in of $1{,}000$. Convergence and mixing are assessed via Gelman and Rubin (1992) potential scale reduction factors and effective sample sizes, respectively, for the quantities of interest, obtaining comparable results to those in the simulation study.

Considering four coordinates provides an area under the ROC curve for in-sample prediction of $\text{AUC}_4 = 0.978$. This is already an interesting results in suggesting that the 120×120 time-varying adjacency matrices can be adequately characterized by collapsing information into a substantially lower-dimensional space. This insight is further confirmed by results in Figure 2.9 highlighting accurate performance not only in edge prediction but also in modeling time-varying network structures of interest.

The trajectory of the posterior mean for the expected network density in upper left plot of Figure 2.9 provides an interesting overview of the overall dynamic contact behavior, consistent with school schedule and changing environments summarized in Figure 10 of Stehlé et al. (2011). It is first interesting to notice how the expected network density evolves on low values suggesting a sparse network, with our adaptive procedure additionally capturing rapid increase in contacts occurring in correspondence of school breaks and the beginning or the end of lunch times for groups of students. According to the left plot in the lower panel of Figure 2.9, the posterior predictive distribution arising from our formulation is sufficiently flexible in accommodating the evolution of this summary statistics.

In studying dynamic homophily patterns, we investigate the posterior distribution of the time-varying expected assortativity coefficients by class and gender, computed for the 115 students. We hold out teachers in homophily studies as we don't have gender information for these nodes and we are interested in social interactions among children – consistently with Stehlé et al. (2013). In investigating gender homophily, Stehlé et al. (2013) focus on a single network obtained aggregating contacts that are observed in pre-selected nonconsecutive time windows when proximity occasions are expected to have less environmental restrictions – i.e. break and lunch times. Although this is a reasonable procedure, information on spatial environments or events are not always available and the choice of aggregation intervals is

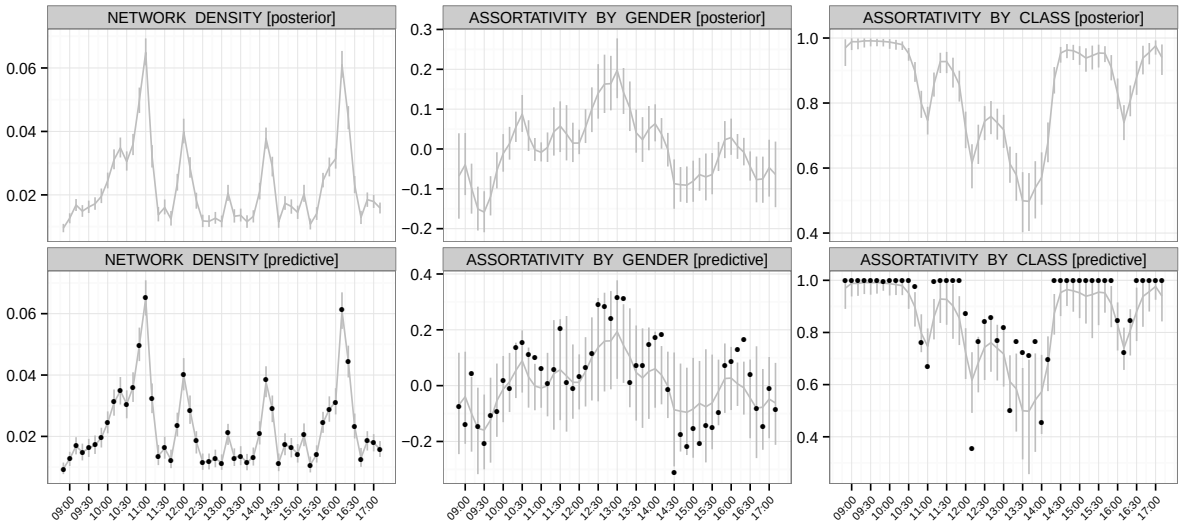


FIGURE 2.9: Upper panels: time-varying posterior mean (gray line) and pointwise 0.95 highest posterior density intervals (gray segments) for expected network summary statistics covering density, assortativity by gender and assortativity by class. Lower panels: for the same summary statistics, time-varying mean (gray line) and pointwise 0.95 predictive intervals (gray segments) obtained from the posterior predictive distribution; black dots represents the corresponding time-varying network measures computed from the observed data.

not necessary unique. Moreover, investigating gender homophily for a single aggregated network provides only an averaged overview of a dynamic system. We instead study homophily structures as they evolve in time, and allow these quantities to be different in nonconsecutive time windows. Our results in the upper middle plot of Figure 2.9 partially confirm findings in Stehlé et al. (2013), with the posterior distributions of the dynamic expected assortativity coefficients concentrated on positive values during break and lunch times. However expected assortativity is higher during lunch compared to breaks, with the posterior for these quantities including the value 0 during the last break. Hence Stehlé et al. (2013) may over-estimate gender homophily in correspondence of break times and under-estimate this property during lunches.

Expected assortativity by class is always positive, with the posterior distributions concentrating on substantially high values during school hours, when contacts are restricted by the spatial environments displayed in Figure 10 of Stehlé et al. (2011); refer to the upper right plot of Figure 2.9. Model checking in the lower middle and right plots of Figure 2.9 highlights an overall good performance of our procedures in characterizing also these higher-order homophily structures. These are key results, provided that we embed a 120×120 dynamic network into a substantially lower-dimensional space made by four latent coordinates, without any further information on the dynamic effect of exogenous variables. Few issues are found in accommodating rapid changes in assortativity by class. A reason behind this slight lack of fit is that $R = 4$ latent coordinates may not be sufficient to characterize class homophily in specific time windows. It is still an active area of research to accommodate latent space dimensions which adaptively change as a function of time. Similarly to our procedure,

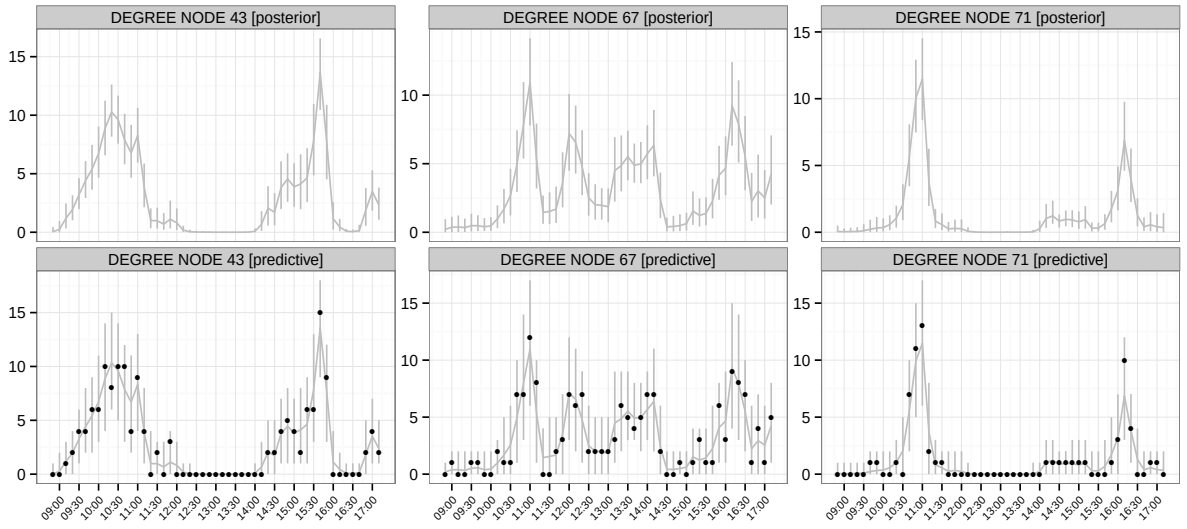


FIGURE 2.10: Upper panels: time-varying posterior mean (gray line) and pointwise 0.95 highest posterior density intervals (gray segments) for the expected degree of selected nodes. Lower panels: for the same summary statistics, time-varying mean (gray line) and pointwise 0.95 predictive intervals (gray segments) obtained from the posterior predictive distribution; black dots represents the corresponding time-varying node degrees computed from the observed data.

most of available contributions rely on time-constant space dimensions which can adequately characterize the whole dynamic network structure. Although a subset of the observed class assortativity coefficients are not within the 0.95 predictive intervals, most of these values are contained in the 0.99 predictive intervals. Hence we maintain $R = 4$ to avoid over-fitting.

Beside accommodating global network structures our procedure can flexibly characterize node-specific activity measures of interest. According to the upper panels of Figure 2.10, incorporating node heterogeneity and time-varying smoothness, allow us to flexibly account for substantially different patterns and dynamic changes in expected node degrees. As shown in the lower panels of Figure 2.10, the posterior predictive distributions for the dynamic node degrees arising from our estimates are characterized by a very accurate performance in accommodating these time-varying observed quantities.

Beside representing a key for robust inference, previous results are fundamental to guarantee accurate performance in forecasting of future network structures. Recalling our motivating application, once the model has been estimated on data from t_1 to t_{i-1}, a new contact network A_{t_i} can stream in along with the information that a subject – or a subset of them – has contracted a specific disease at t_i. Hence, for the sake of outbreak prevention it is fundamental to fast update estimates at time t_i and forecast the contact network structures at the next time t_{i+1}. Our LADY network model can suitably accomplish this task by online updating the posterior distribution for the edge probabilities at t_i exploiting strategies in Section 2.2.3 – with $j = 5$ – and then forecast the posterior distribution of the same quantities at the next time t_{i+1} by applying equation (2.22) to the MCMC samples from the online updating. Once

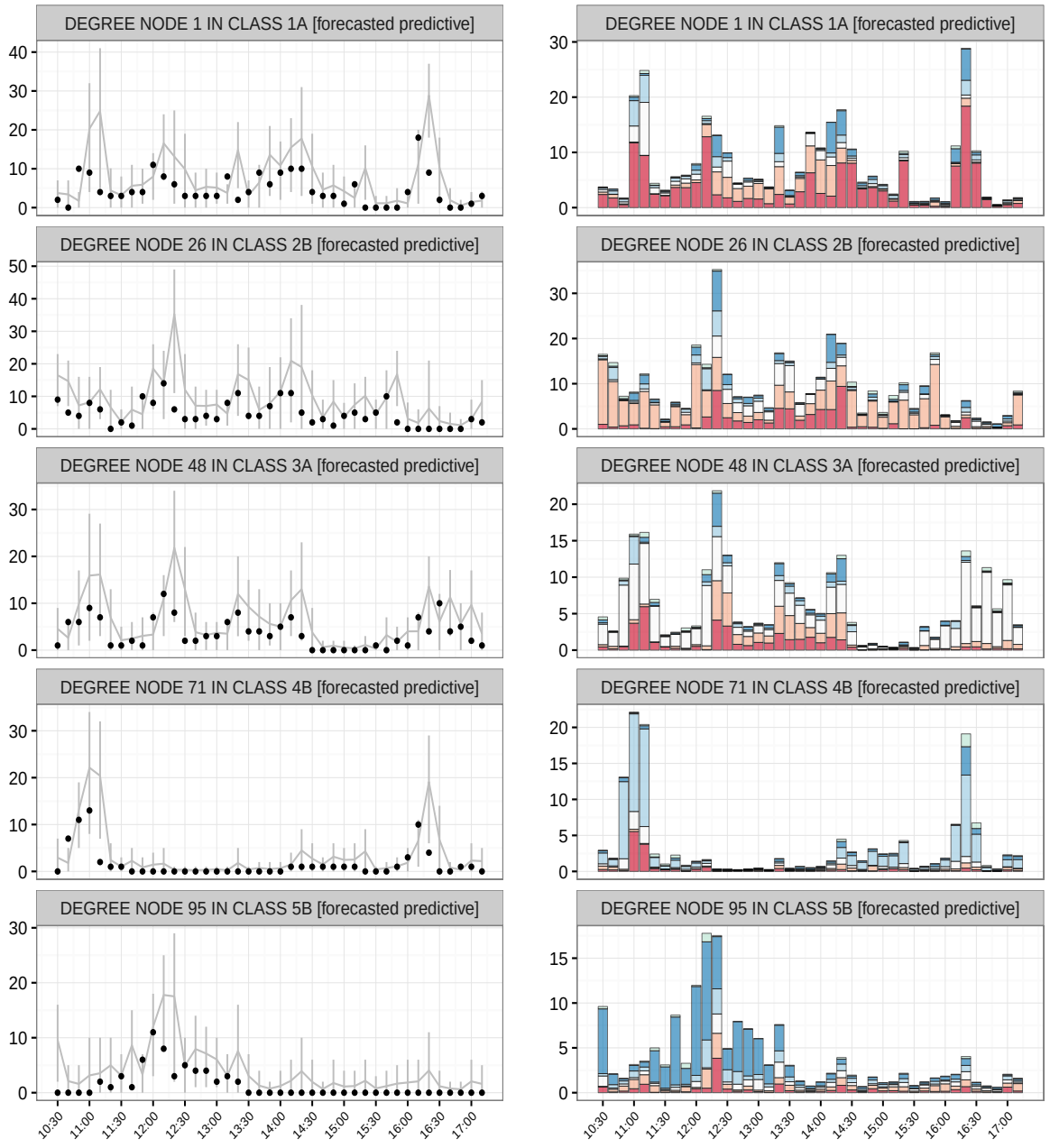


FIGURE 2.11: Left panels: for selected subjects in the five different classes, time-varying mean (gray line) and pointwise 0.95 predictive intervals (gray segments) of their degree obtained from the one-step-ahead forecasted predictive distribution from t_{11} to t_{51}; black dots represents the corresponding time-varying node degrees computed from the observed data. Right panels: for the same subjects barplots representing the time-varying mean of their degree obtained from the one-step-ahead forecasted predictive distribution from t_{11} to t_{51}. Colors in the bars represent the proportion of the forecasted degree due to connections with each class. Dark red (first class), light red (second class), white (third class), light blue (fourth class), dark blue (fifth class), green (teachers).

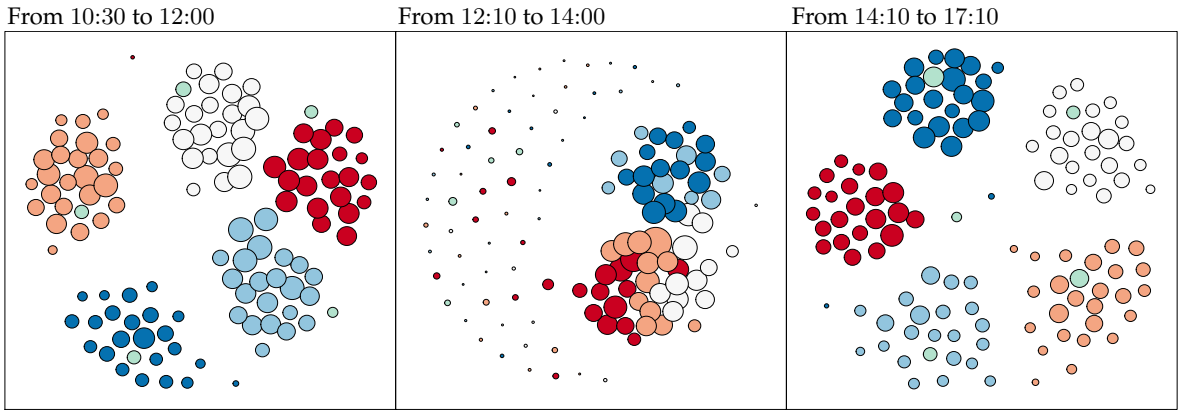


FIGURE 2.12: Weighted network visualization with weights obtained averaging the mean of the one-step-ahead forecasted predictive distributions over three time windows. Edges are not displayed to facilitate graphical analysis. Nodes positions are obtained applying the Fruchterman and Reingold (1991) force-directed placement algorithm. Nodes dimensions are proportional to their forecasted degree averaged over each time window and colors indicate class membership. Dark red (first class), light red (second class), white (third class), light blue (fourth class), dark blue (fifth class), green (teachers).

these quantities are available it is straightforward to derive the approximate forecasted predictive distribution at t_{i+1} along with related quantities of interest such as the expected value for forecasting edges and the predictive distribution of future topological structures. Figures 2.11, 2.12 and the upper left plot of Figure 2.13 evaluate the performance of our joint online updating and forecasting procedure for times from t_{11} to t_{51}, under different perspectives.

Left panels of Figure 2.11 compare observed node degrees for selected subjects in the five different classes, with their mean and quantiles arising from the forecasted predictive distribution. Dynamic node degrees are a key for disease surveillance and accurate forecasts for these quantities are fundamental to measure the infectivity for each individual at future times. According to left panels of Figure 2.11 our strategies provide in general a good performance in forecasting dynamic degrees. We observe, however, a slight tendency towards over-estimating these quantities. Although previous bias is of course undesired, it is worth noticing that for the sake of outbreak prevention, slightly over-estimating node degrees suggests conservative policies which are preferable to biased mild actions understating chance of contagion.

Right panels of Figure 2.11 add further insights by highlighting the proportion of the forecasted degree due to connections with students in the different classes. This provides an higher-level measure of which groups of nodes are at risk of contagion at t_{i+1} if a given individual contracts a disease at t_i, for each $i = 10, \ldots, 50$. Results further confirm our good performance in forecasting heterogenous activity patterns and dynamic changes in node degrees. Consistently with previous findings on homophily structures, contacts with individuals from the same class represent an high proportion of the forecasted dynamic degrees. This is more evident during school hours, than breaks or lunch times where we forecast more mixed patterns including increased across classes contacts as well as students apparently

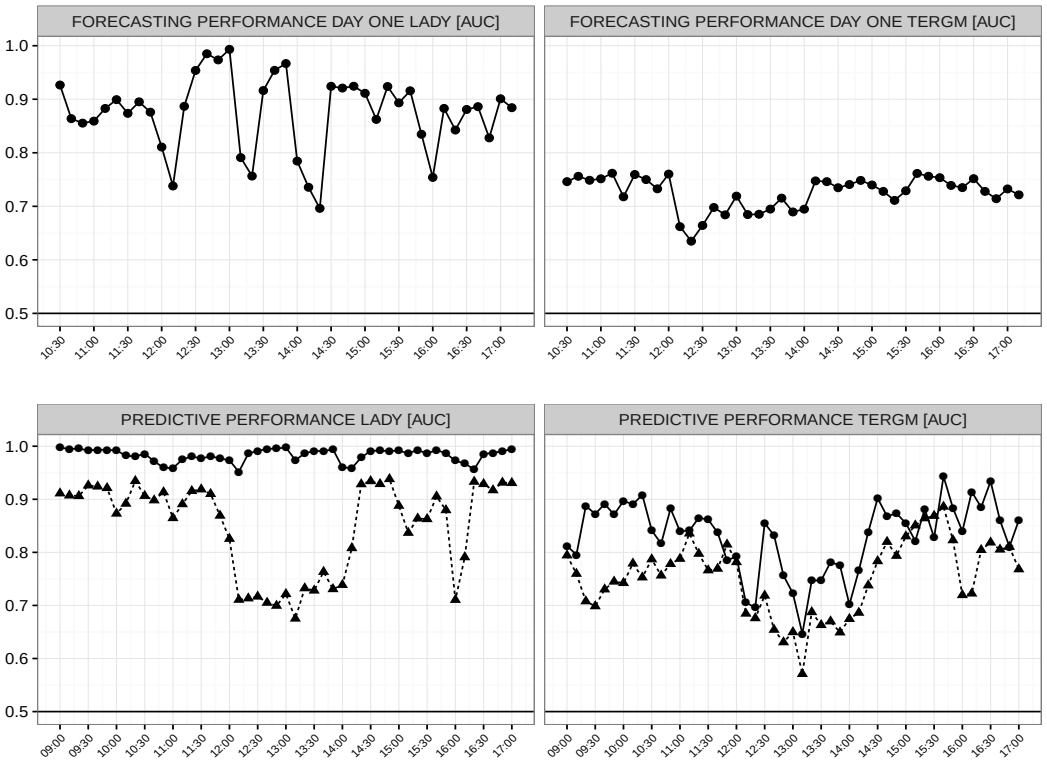


FIGURE 2.13: Upper panels: for times from t_{11} to t_{51} in day one, comparison of the forecasting performance between our LADY network model and TERGM. Performance is assessed via the area under the ROC curve generated using the mean of the one-step-ahead forecasted predictive distribution and observed networks. Lower panels: in-sample (day one: solid line) and out-of-sample (day two: dashed line) predictive performance of LADY network model and TERGM. Performance is assessed via the area under the ROC curve generated using the mean of the predictive distribution and observed networks for every time in day one and two, respectively.

leaving the school – such as for example node 71.

These findings are confirmed in Figure 2.12 providing a graphical representation of future networks with nodes positions depending on the forecasted edges – according to procedures (2.11) – averaged over three time windows of interest. Although we do explicitly include environmental information, as shown in Figure 2.12 our procedure is sufficiently flexible to account for these structures from an unsupervised perspective. Consistently with Figure 10 in Stehlé et al. (2011) we forecast evident community structures induced by class membership during the morning hours, with students in classes 1A, 3A and 4B being spatially closer than those in the remaining classes. This is consistent with classes 1A, 3A and 4B sharing the playground during the morning break according to Figure 10 in Stehlé et al. (2011). Lunch times are characterized by a sparse structure with two communities and a wide set of students having essentially no face-to-face contacts. The first community comprises students in classes 1A, 2B and part of those in class 3A. The second includes children from classes 4B, 5B and the remaining students from class 3A. Also these forecasts are consistent with the approximate school schedule presented in Stehlé et al. (2011), with a subset of the students leaving the school during lunch and the remaining children sharing the canteen in two different groups

at consecutive times. As expected results in the afternoon hours are similar to those in the morning ones, with a slightly more sparse structure due the fact that children increasingly leave the school towards the end of the day.

To further evaluate our forecasts, the upper left plot in Figure 2.13 assesses forecasting performance by showing for each time from t_{11} to t_{51} the AUC based on the expected value of the forecasted predictive distribution – according to our online updating and forecasting procedure – and observed data $A_{t_{11}}, \dots, A_{t_{51}}$. The AUC evolves on high values, suggesting overall good performance in forecasting of future edges, with more evident decrements in correspondence of the beginning, mid and end of the lunch time windows. These times are characterized by rapid variations in contact behavior due to children rapidly changing environments; refer to Figure 10 in Stehlé et al. (2011). Hence – recalling also insights in the simulation study – this decreased forecasting performance is reasonably related to the fact that the model has no experience of sudden regime changes. Although we face reduced forecasting performance in specific times, our procedure almost always improves forecasts of a TERGM estimated maintaining the same linear predictor of the simulation study. Refer to the upper right plot in Figure 2.13.

We conclude our analysis by evaluating in-sample and out-of-sample predictive performance. In the former case we compare the mean of the predictive distribution from equation (2.23) to observed edges in the first day via AUC. Out-of-sample predictive performance is instead assessed using the same procedure but comparing predicted edges – training the model with data from day one – to observed networks from the second day. As time t_{51} is not available in the second day, we assess in-sample and out-of-sample predictive performance using data and estimates from t_1 to t_{50}. Results are displayed in the lower plots of Figure 2.13.

As expected in-sample edge prediction is very accurate under our LADY network model. We also obtain a general good performance when predicting edges at the second day, based on estimates from day one. More evident differences compared to in-sample performance are found in correspondence of lunch times and the afternoon break. This may suggest that the dynamic contact networks at the second day are governed by slightly different underlying patterns than those associated with the first day, for these time windows. Also in this case we almost always improve results from the TERGM in both prediction tasks. These results further confirm the need of procedures accounting for heterogenous and dynamic dependence patterns in such frameworks.

Chapter 3

Populations of networks

3.1 Nonparametric modeling of populations of networks

In neuroscience there is increasing interest in relating the structural connection network defined by white matter tracts in the human brain and cognitive traits or neuropsychiatric disorders. There is evidence that the structural network is a more important driver of variability in cognitive traits and disorders than measures of human brain activity – extracted from fMRI. Recent connectomics pipelines can obtain the brain network based on diffusion tensor imaging and structural MRI. This produces a network-valued random variable for each individual in a study. Motivated by data outlined in Section 1.2.1 we develop novel nonparametric Bayes methods for analyzing network-valued data, and for performing inference on the relationship between brain networks and cognitive traits or neurological disorders. These methods allow the probability mass function of the network-valued data to shift nonparametrically between groups, via a dependent mixture of low-rank factorizations, facilitating global and local hypothesis testing adjusting for multiplicity and robust against model misspecification. An efficient Gibbs sampler is defined for posterior computation. We provide theoretical results on the flexibility of the model and show dramatic improvements relative to current approaches in studying creative reasoning and Alzheimer's disease data.

3.1.1 Notation and motivation

Let (y_i, A_i) represent the group membership and the undirected network observation, respectively, for subject $i = 1, \ldots, n$, with $y_i \in \mathbb{Y} = \{1, 2\}$ and A_i the $V \times V$ adjacency matrix characterizing the connections among the anatomical regions in his brain.

As a step towards our goal of defining a joint model and testing procedures for data (y_i, A_i), $i = 1, \ldots, n$, we first develop a probabilistic generative mechanism for the random variable

generating replicated network data $A_1, \ldots, A_n$. In addressing this goal we look for a statistical representation which can provably characterize a wide class of probabilistic generative processes, while maintaining tractable computations via efficient dimensionality reduction. Accomplishing this aim is fundamental to develop accurate testing procedures to assess evidence of changes in the brain network across groups, limiting concerns about lack of robustness to model misspecification.

Consistently with discussion in Section 2.1.1 – as the brain networks are available via undirected edges and self-relationships are not of interest – we model the observed adjacency matrices $A_1, \ldots, A_n$ by focusing on the random variable $\mathcal{L}(\mathcal{A})$ generating data $\mathcal{L}(A_1), \ldots, \mathcal{L}(A_n)$ with $\mathcal{L}(A_i) = (A_{i[21]}, A_{i[31]}, \ldots, A_{i[V1]}, A_{i[32]}, \ldots, A_{i[V2]}, \ldots, A_{i[V(V-1)]})^{\mathrm{T}} \in \mathbb{A}_V = \{0,1\}^{V(V-1)/2}$ the vector encoding the lower triangular elements of A_i, which uniquely define the network as $A_{i[vu]} = A_{i[uv]}$ for every $v = 2, \ldots, V$, $u = 1, \ldots, v-1$ and $i = 1, \ldots, n$.

Data $\mathcal{L}(A_1), \ldots, \mathcal{L}(A_n)$ are realizations from a multivariate Bernoulli random variable $\mathcal{L}(\mathcal{A})$. Since there are finitely many network configurations, $\mathcal{L}(\mathcal{A})$ can be seen as a categorical random variable with each category representing one of the possible network configurations $\mathcal{L}(A) = a \in \mathbb{A}_V = \{0,1\}^{V(V-1)/2}$. Considering for example $V = 3$, the network-valued random variable $\mathcal{L}(\mathcal{A})$ has $2^{V(V-1)/2} = 8$ possible categories $\{(0,0,0); (1,0,0); \ldots; (1,1,1)\}$ and $2^{V(V-1)/2} - 1 = 7$ parameters are required to fully characterize the pmf $p_{\mathcal{L}(\mathcal{A})}(a) = \text{pr}\{\mathcal{L}(\mathcal{A}) = a\}$, $a \in \mathbb{A}_V$ under the restriction $\sum_{a \in \mathbb{A}_V} p_{\mathcal{L}(\mathcal{A})}(a) = 1$; see Dai et al. (2013) for properties and recent results on the multivariate Bernoulli random variable.

The number of parameters is intractable and massively larger than the sample size n even in small V settings. In the motivating neuroscience study, brain images have been processed to obtain adjacency matrices for each subject considering $V = 68$ anatomical brain regions. This implies that, in the absence of constraints, there are $2^{68(68-1)/2} - 1 = 2^{2278} - 1$ free parameters to estimate characterizing $p_{\mathcal{L}(\mathcal{A})}$. Clearly no studies will ever have this many subjects, and hence it is necessary to substantially reduce dimensionality to make the problem tractable. However, in reducing dimension, it is important to avoid making overly restrictive assumptions that lead to inadequate characterization of the observed network data.

To solve this problem, our goal is to develop a provably flexible and tractable factorization for $p_{\mathcal{L}(\mathcal{A})}$, which reduces dimensionality of the parameter space, while retaining flexibility in characterizing $p_{\mathcal{L}(\mathcal{A})}$ incorporating network structure. In fact, the key difference between a network-valued random variable and an unstructured categorical random vector is that the network configurations share a common underlying structure which informs edge probabilities. As a result, by carefully combining mixture representations with matrix factorization procedures in constructing the edge probabilities, one might efficiently borrow information across units and within each network, while characterizing individual variability. By placing priors on the components within this factorization, we induce a prior Π for $p_{\mathcal{L}(\mathcal{A})}$, with

full support over the $2^{V(V-1)/2}$-dimensional probability simplex $\mathcal{P}_{2^{V(V-1)/2}}$, while obtaining appealing asymptotic properties.

3.1.2 Low-rank factorization mechanism

In developing a flexible representation for the probabilistic generative mechanism underlying data $A_1, \ldots, A_n$, it is fundamental to account for the special structure of the random variable $\mathcal{L}(\mathcal{A})$. In particular, $\mathcal{L}(\mathcal{A})$ is a multivariate Bernoulli random variable characterized by a network structure underlying its entries $\mathcal{L}(\mathcal{A})_l$, $l = 1, \ldots, V(V-1)/2$, with the structure potentially having small-world, scale free, transitive, community or hub behaviors.

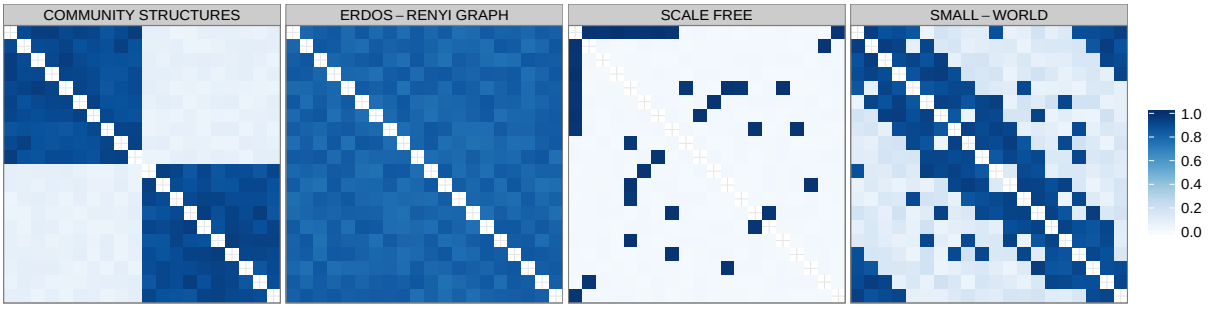
As discussed in the Introduction there is a rich literature on borrowing network information across edges and modeling of network data and their topological characteristics. Classical approaches in modeling of a single network observation, such as Erdös and Rényi (1959), p_1-models (Holland and Leinhardt, 1981) and p^* models (Frank and Strauss, 1986), define the probability $p_{\mathcal{L}(\mathcal{A})}(a)$ of a given network configuration $a \in \mathbb{A}_V$ under an exponential family representation, with sufficient statistics representing suitably chosen network measures. Although exponential random graphs can induce suitable dependence structures between edges and model some topological properties of interest, these procedures are characterized by drawbacks in terms of estimation (Strauss and Ikeda, 1990), possible degeneracy issues and inflexibility in assigning the same probability to configurations having equal network measures, even when such configurations are very different (Chatterjee and Diaconis, 2013).

Based on these possible issues and to provide tractable computation, we borrow information across edges by considering a latent variable approach to network analysis. Recalling our Introduction, latent variable modeling of networks is accomplished by assuming edges $\mathcal{L}(\mathcal{A})_l$, $l = 1, \ldots, V(V-1)/2$ as conditionally independent Bernoulli random variables given their corresponding edge probabilities $\pi_l \in (0,1)$, $l = 1, \ldots, V(V-1)/2$. This leads to the following representation for $p_{\mathcal{L}(\mathcal{A})}$

$$p_{\mathcal{L}(\mathcal{A})}(a) = \prod_{l=1}^{V(V-1)/2} \pi_l^{a_l}(1-\pi_l)^{1-a_l}, \quad a \in \mathbb{A}_V. \tag{3.1}$$

As shown in Figure 3.1, under suitable choices of $\pi = (\pi_1, \ldots, \pi_{V(V-1)/2})^{\mathrm{T}} \in (0,1)^{V(V-1)/2}$ equation (3.1) can assign high probability to network configurations having specific properties, such as community structure, scale free, small-world and hub behaviors.

Stochastic block models (Nowicki and Snijders, 2001) and their generalizations can characterize modularity structures by defining π as a function of node memberships to communities and block probabilities between these communities. Although estimation of block structures



COMMUNITY STRUCTURES	ERDOS – RENYI GRAPH	SCALE FREE	SMALL – WORLD
Deg. mean: 8.5 (0.4)	**Deg. mean:** 15 (0.5)	**Deg. mean:** 2.1 (0.1)	Deg. mean: 8.5 (0.4)
Deg. var.: 1.4 (0.5)	**Deg. var.:** 2.8 (0.9)	**Deg. var.:** 3.1 (0.4)	Deg. var.: 2.2 (0.7)
Trans.: 0.7 (0.1)	Trans.: 0.8 (0.1)	Trans.: 0.1 (0.1)	**Trans.:** 0.5 (0.1)
Av. p. leng.: 1.6 (0.1)	Av. p. leng.: 1.2 (0.1)	Av. p. leng.: 2.7 (0.2)	**Av. p. leng.:** 1.4 (0.1)
Assort.: 0.8 (0.1)	Assort.: −0.1 (0.1)	Assort.: 0.1 (0.1)	Assort.: 0.2 (0.1)

FIGURE 3.1: Example of possible edge probability matrices generating networks with given topological properties under conditional independence assumption of the edges. For each edge probability matrix, we report mean and standard deviation of key topological measures calculated on 1,000 networks whose edges are simulated from conditionally independent Bernoulli random variables given their edge probabilities defined in the four matrices.

is often of interest, such models have limited flexibility. Mixed membership stochastic block models (Airoldi et al., 2008) and latent space models (Hoff et al., 2002) improve flexibility by not restricting nodes to belong to a single community. Hoff et al. (2002) define π as a function of pairwise Euclidean distances between nodes in a latent space. This characterization can provably capture community behaviors, transitive relations, and k-star structures (Hoff et al., 2002) and has been generalized to accommodate additional network properties (Krivitsky et al., 2009; Hoff, 2008).

In line with the factorizations considered for dynamic network inference in Sections 2.1 and 2.2, our probabilistic low-rank factorization characterizes the edge probability vectors as

$$\pi = [1 + \exp\{-\mathcal{L}(X\Lambda X^{\mathrm{T}})\}]^{-1}, \quad \pi \in (0,1)^{V(V-1)/2}, \tag{3.2}$$

with the logistic mapping from $\Re$ to $(0,1)$ applied element-wise. Equation (3.2) defines edge probabilities through a low-rank factorization of their log-odds $S = (S_1, \ldots, S_{V(V-1)/2})^{\mathrm{T}} = \mathcal{L}(X\Lambda X^{\mathrm{T}}) \in \Re^{V(V-1)/2}$, with $X \in \Re^{V\times R}$ the matrix of the R latent coordinates for the V nodes, and Λ a diagonal weight matrix with $\mathrm{diag}(\Lambda) = (\lambda_1, \ldots, \lambda_R)^{\mathrm{T}} = \lambda \in \Re^R_{\geq 0}$. Notation $\Re^R_{\geq 0}$ refers to the space of vectors with R real non negative elements.

The low-rank factorization allows dimensionality reduction from $V(V-1)/2$ edge probabilities to $V \times R$ latent coordinates and R weights – typically $R \ll V$ – while facilitating adaptive collapsing on lower-dimensional models by appropriately shrinking the weights λ_r towards 0 as r increases. Moreover characterizing the log-odds via the weighted dot product of the nodes latent coordinates has an appealing interpretation. Recalling our motivating neuroscience application, the coordinate $X_{vr} \in \Re$ may measure the activity of brain region v within pathway r. According to the dot product construction, regions with activities in the same

direction – both positive or negative – will be more similar. The similarity – or dissimilarity – will be higher the stronger the activity is in the same – or opposite – direction.

Although the mechanism generates networks from conditionally independent edges given π, the shared dependence on a common set of node-specific latent coordinates induced by the dot product representation of S facilitates borrowing of information across observed edges in estimating π_l for $l = 1, \ldots, V(V-1)/2$ and can accurately characterize a broad variety of network structures. In particular, equation (3.2) can arbitrarily represent every possible edge probability vector $\pi \in (0,1)^{V(V-1)/2}$ by exploiting the one-to-one continuity of the logistic mapping and the fact that there exist infinitely many positive semidefinite matrices having S as lower triangular element vector. This allows our low-rank factorization mechanism to capture specific network properties by adaptively modeling the edge probability vector.

3.1.3 Nonparametric mixture of low-rank factorizations

The low-rank factorization described above has two main drawbacks motivating further modifications. Firstly, a single factorization of the edge probability vector does not characterize variability across different networks. Secondly, a common edge probability vector is not sufficiently flexible to characterize every possible probabilistic mechanism for generating network data. For example, it is easy to show that equation (3.1) cannot represent the pmf for the network-valued random variable $\mathcal{L}(\mathcal{A})$ that generates either disconnected or fully connected networks with equal probability $p_{\mathcal{L}(\mathcal{A})}\{(0,0,\ldots,0)\} = p_{\mathcal{L}(\mathcal{A})}\{(1,1,\ldots,1)\} = 0.5$ and assigns $p_{\mathcal{L}(\mathcal{A})}(a) = 0$ for all configurations a different from $(0,0,\ldots,0)$ and $(1,1,\ldots,1)$.

We improve the flexibility by considering a hierarchical representation that characterizes individual variability through the introduction of a low-rank factorization mechanism for each network A_i. To characterize variability across networks, the unit-specific edge probability vectors π_i, $i = 1, \ldots, n$ are treated as random effects and are assigned a common discrete probability measure P. Specifically we let

$$\mathcal{L}(\mathcal{A}_i)_l \mid \pi_{il} \overset{\text{indep}}{\sim} \text{Bern}(\pi_{il}), \quad l = 1, \ldots, V(V-1)/2, \quad i = 1, \ldots, n, \tag{3.3}$$

$$\pi_i \mid P \overset{\text{iid}}{\sim} P = \sum_{h=1}^{H} \nu_h \delta_{\pi^{(h)}}, \quad \pi^{(h)} = \left[1 + \exp\{-Z - \mathcal{L}(X^{(h)}\Lambda^{(h)}X^{(h)\text{T}})\}\right]^{-1}, \tag{3.4}$$

where $\delta_{\pi^{(h)}}$ denotes a mass concentrated at $\pi^{(h)}$ and ν_h the probability that a randomly selected network is allocated to class h. This choice allows clustering of networks into H latent classes, with networks in the same class h having identical edge probability vector $\pi^{(h)}$. Each $\pi^{(h)}$ is in turn factorized to allow inference on shared versus class-specific components of variability in the networks connectivity behavior. Specifically, according to (3.4) each $\pi^{(h)}$ is defined as a function of a similarity vector $Z \in \Re^{V(V-1)/2}$ shared across all networks and a

class-specific deviation $D^{(h)} \in \Re^{V(V-1)/2}$. The shared vector Z is modeled as unstructured. By borrowing information across all networks in all classes, we can accurately infer Z without additional structural constraints in our experience. There is much less information in the data about the class-specific deviations, and we rely on a low-rank matrix factorization as in equation (3.2) obtaining $D^{(h)} = \mathcal{L}(X^{(h)}\Lambda^{(h)}X^{(h)\mathrm{T}})$ for every $h = 1, \ldots, H$. See Figure 3.2 for a graphical representation of the probabilistic generative mechanism associated with the mixture of low-rank factorizations characterizing $p_{\mathcal{L}(\mathcal{A})}$.

Allowing a separate factorization for each $D^{(h)}$ induces highly flexible deviations in connectivity behavior with h. Network properties and topological structures can vary substantially, with some classes having small-world behaviors, while others indicate strong community patterns. Considering a common edge probability vector as in Nowicki and Snijders (2001), Airoldi et al. (2008) and Hoff et al. (2002) has the major disadvantage of reducing such variability in forcing $p_{\mathcal{L}(\mathcal{A})}$ to concentrate its mass on a subset of configurations characterized by a specific network property via (3.1), while ruling out others. Model (3.3)–(3.4) instead adaptively assigns probability to different subsets of configurations, each one potentially characterized by a different network property.

By marginalizing out the unit-specific edge probability vectors π_i in (3.3)–(3.4), we obtain the following representation for the pmf $p_{\mathcal{L}(\mathcal{A})}$ associated with the network-valued random variable $\mathcal{L}(\mathcal{A})$ generating networks $\mathcal{L}(A_1), \ldots, \mathcal{L}(A_n)$:

$$p_{\mathcal{L}(\mathcal{A})}(a) = \sum_{h=1}^{H} \nu_h \prod_{l=1}^{V(V-1)/2} \left\{\pi_l^{(h)}\right\}^{a_l} \left\{1 - \pi_l^{(h)}\right\}^{1-a_l}, \tag{3.5}$$

for every $a \in \mathbb{A}_V$, with each $\pi^{(h)}$ factorized as

$$\pi^{(h)} = \left[1 + \exp\{-Z - D^{(h)}\}\right]^{-1}, \quad D^{(h)} = \mathcal{L}(X^{(h)}\Lambda^{(h)}X^{(h)\mathrm{T}}), \quad h = 1, \ldots, H. \tag{3.6}$$

Beside considerably reducing the dimensionality from $2^{V(V-1)/2} - 1$ to $H\{1 + R(V + 1)\} + V(V - 1)/2 - 1$ parameters, as formalized in Lemma 3.1, our mixture of low-factorizations can represent any possible pmf $p_{\mathcal{L}(\mathcal{A})} \in \mathcal{P}_{2^{V(V-1)/2}}$ defined on a network-valued sample space. This confirms the full flexibility of our construction, which can be viewed as nonparametric given appropriately chosen priors for the components.

Lemma 3.1. *Any $p_{\mathcal{L}(\mathcal{A})} \in \mathcal{P}_{2^{V(V-1)/2}}$ admits representation (3.5) for some H with ν_h probability weights such that $\sum_{h=1}^{H} \nu_h = 1$ and each $\pi^{(h)} \in (0, 1)^{V(V-1)/2}$ factorized as in* (3.6) *for some R.*

Proof. To prove the full generality of (3.5), note that $p_{\mathcal{L}(\mathcal{A})}$ is the probability mass function over the cells in a contingency table with the lth variable denoting presence or absence of an edge between the lth pair of nodes. Hence Lemma 3.1 follows immediately from Theorem 1

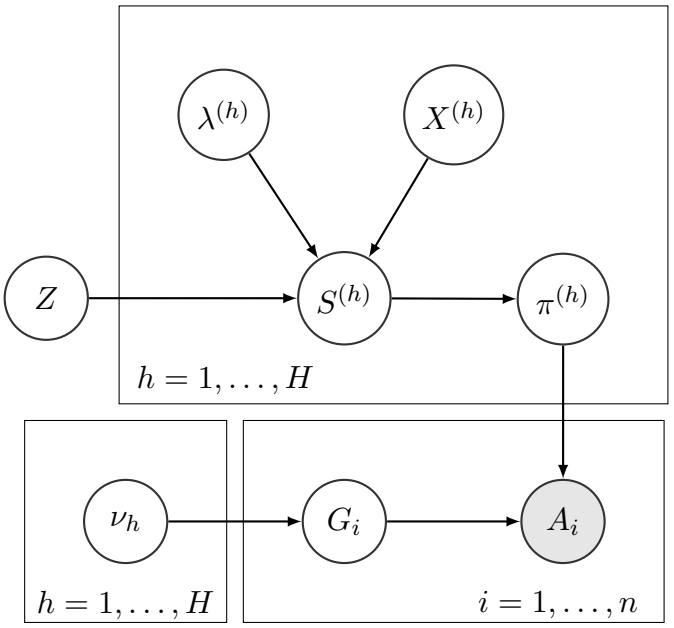


FIGURE 3.2: Graphical representation of the probabilistic mechanism generating networks A_i ($i = 1, \ldots, n$) under the mixture of low-rank factorizations representation in (3.3)–(3.4). In particular, for each i choose one low-rank factorization mechanism by sampling the latent class indicator $G_i \in \{1, \ldots, H\}$ from p_G with $p_G(h) = \text{pr}(G_i = h) = \nu_h$. Given $G_i = h$ and the corresponding edge probability vector $\pi^{(h)}$ arising from the low-rank representation, generate the network A_i by sampling its edges $\mathcal{L}(A_i)_l$, $l = 1, \ldots, V(V-1)/2$ from conditionally independent Bernoulli variables.

of Dunson and Xing (2009) with $\psi_h^{(l)} = (\pi_l^{(h)}, 1 - \pi_l^{(h)})^{\text{T}}$ for $l = 1, \ldots, V(V-1)/2$, as long as any $\pi^{(h)} \in (0,1)^{V(V-1)/2}$ can be represented via (3.6) for $h = 1, \ldots, H$. Assume without loss of generality $Z = 0_{V(V-1)/2}$. Since the logistic mapping is one-to-one and continuous it suffices to show that any $D^{(h)} \in \Re^{V(V-1)/2}$ can be expressed as $D^{(h)} = \mathcal{L}(X^{(h)}\Lambda^{(h)}X^{(h)\text{T}})$, with $X^{(h)} \in \Re^{V \times R}$ and $\Lambda^{(h)}$ a $R \times R$ diagonal matrix with non-negative entries for each $h = 1, \ldots, H$. As there exist infinitely many positive semidefinite matrices having lower triangular elements $D^{(h)}$, let $\Xi^{(h)}$ be one of these matrices such that $\mathcal{L}(\Xi^{(h)}) = D^{(h)}$. Letting $R^{0(h)}$ denote the rank of $\Xi^{(h)} = \tilde{X}^{(h)}\tilde{\Lambda}^{(h)}\tilde{X}^{(h)\text{T}}$, with $\tilde{\Lambda}^{(h)}$ the diagonal matrix with the $R^{0(h)}$ positive eigenvalues of $\Xi^{(h)}$ and $\tilde{X}^{(h)} \in \Re^{V \times R^{0(h)}}$ the matrix with the corresponding eigenvectors, Lemma holds after defining $X^{(h)} = (\tilde{X}^{(h)} \;\; 0_{V \times (R - R^{0(h)})})$ and $\tilde{\Lambda}^{(h)}$ diagonal, with $\Lambda_{rr}^{(h)} = \tilde{\Lambda}_{rr}^{(h)}$ for $r \leq R^{0(h)}$ and 0 otherwise. □

Factorization (3.6) is not unique. For example, letting $\tilde{Z} = Z + Q$ and $\tilde{D}^{(h)} = D^{(h)} - Q$, $h = 1, \ldots, H$ then $\tilde{Z} + \tilde{D}^{(h)} = Z + Q + D^{(h)} - Q = Z + D^{(h)}$. This further affects the uniqueness of the factorization $D^{(h)} = \mathcal{L}(X^{(h)}\Lambda^{(h)}X^{(h)\text{T}})$. Moreover, there exist infinitely many diagonalizable positive semidefinite matrices having $D^{(h)}$ as lower triangular elements. Although these issues do not affect the identifiability of each class-specific edge probability vector $\pi^{(h)}$, for $h = 1, \ldots, H$ required to characterize $p_{\mathcal{L}(\mathcal{A})}$ via (3.5), they may lead to misleading conclusions when studying common network properties and class-specific connectivity patterns.

Similar identifiability issues arise routinely in Bayesian factorizations and nonparametric models, which tend to be purposely over-parameterized. Such over-parameterization often has a beneficial effect on computational efficiency and does not lead to problems when inference focuses on identifiable functionals of the parameters; see for example Ghosh and Dunson (2009) and Bhattacharya and Dunson (2011). In our specific setting, the low-rank factorization is appealing in reducing dimensionality and accommodating network information. We propose an approach for inference on identified quantities of interest. To study shared network patterns, we focus on the expected value $\bar{\pi} = \mathrm{E}\{\mathcal{L}(\mathcal{A})\} = \sum_{a \in \mathbb{A}_V} a p_{\mathcal{L}(\mathcal{A})}(a)$. According to Proposition 3.2, this quantity can be easily computed under our model as the weighted sum of the edge probability vectors $\pi^{(h)}$, with weights given by the mixing probabilities ν_h.

Proposition 3.2. *Under representation* (3.5) *for* $p_{\mathcal{L}(l\mathcal{A})}$*, the expected value for the network-valued random variable* $\mathcal{L}(\mathcal{A})$ *is given by* $\bar{\pi} = E\{\mathcal{L}(\mathcal{A})\} = \sum_{a \in \mathbb{A}_V} a p_{\mathcal{L}(\mathcal{A})}(a) = \sum_{h=1}^{H} \nu_h \pi^{(h)}$.

Proof. Focusing on the general element l with $l = 1, \ldots, V(V-1)/2$, we need show that $\bar{\pi}_l = \sum_{a \in \mathbb{A}_V} a_l p_{\mathcal{L}(\mathcal{A})}(a) = \sum_{h=1}^{H} \nu_h \pi_l^{(h)}$ for Proposition 3.2 to hold. Under representation (3.5) for $p_{\mathcal{L}(A)}$ and letting $\mathbb{A}_V^{-l}$ denote the set containing all the possible network configurations for the node pairs except the lth, we can write $\bar{\pi}_l$ as

$$
\begin{aligned}
& 1 \cdot \{ \sum_{\mathbb{A}_V^{-l}} \sum_{h=1}^{H} \nu_h \pi_l^{(h)} \prod_{l^* \neq l} (\pi_{l^*}^{(h)})^{a_{l^*}} (1 - \pi_{l^*}^{(h)})^{1 - a_{l^*}} \} + 0 \cdot \{ \sum_{\mathbb{A}_V^{-l}} \sum_{h=1}^{H} \nu_h (1 - \pi_l^{(h)}) \prod_{l^* \neq l} (\pi_{l^*}^{(h)})^{a_{l^*}} (1 - \pi_{l^*}^{(h)})^{1 - a_{l^*}} \} \\
= & \sum_{\mathbb{A}_V^{-l}} \sum_{h=1}^{H} \nu_h \pi_l^{(h)} \prod_{l^* \neq l} (\pi_{l^*}^{(h)})^{a_{l^*}} (1 - \pi_{l^*}^{(h)})^{1 - a_{l^*}} = \sum_{h=1}^{H} \nu_h \pi_l^{(h)} \sum_{\mathbb{A}_V^{-l}} \prod_{l^* \neq l} (\pi_{l^*}^{(h)})^{a_{l^*}} (1 - \pi_{l^*}^{(h)})^{1 - a_{l^*}}.
\end{aligned}
$$

Proposition 3.2 follows after noticing that $\prod_{l^* \neq l} (\pi_{l^*}^{(h)})^{a_{l^*}} (1 - \pi_{l^*}^{(h)})^{1 - a_{l^*}}$ is the joint pmf of $V(V-1)/2 - 1$ independent Bernoulli random variables having joint sample space $\mathbb{A}_V^{-l}$ and hence the summation over $\mathbb{A}_V^{-l} = \{0,1\}^{V(V-1)/2-1}$, provides $\sum_{\mathbb{A}_V^{-l}} \prod_{l^* \neq l} (\pi_{l^*}^{(h)})^{a_l^*} (1 - \pi_{l^*}^{(h)})^{1 - a_{l^*}} = 1$. □

To study class-specific connectivity patterns, we rely on $\pi^{(h)}$ and the differences $\bar{\pi}^{(h)} = \pi^{(h)} - \bar{\pi}$ for each $h = 1, \ldots, H$. As this type of inference is class-specific, it is important to check for label switching issues (Stephens, 2000). Although it is not the case in our specific simulations and application, when trace-plots suggest label switching issues are encountered, one possibility is to relabel the classes at each MCMC iteration via post-processing algorithms, such as Stephens (2000).

3.1.4 Prior specification and properties

Results in Section 3.1.3 ensure that any true probability mass function for a population of networks $p^0_{\mathcal{L}(\mathcal{A})} \in \mathcal{P}_{2^{V(V-1)/2}}$ admits representation (3.5), with class-specific edge probability vectors $\pi^{(h)}$ factorized as in (3.6). Although this is a key result, it is not guaranteed that the same flexibility is maintained after choosing independent priors $Z \sim \Pi_Z$, $\nu = (\nu_1, \ldots, \nu_H) \sim \Pi_\nu$, $X^{(h)} \sim \Pi_X$ and $\lambda^{(h)} \sim \Pi_\lambda$, for $h = 1, \ldots, H$.

Letting $\mathbb{B}_\epsilon\{p^0_{\mathcal{L}(\mathcal{A})}\} = \{p_{\mathcal{L}(\mathcal{A})} : \sum_{a \in \mathbb{A}_V} |p_{\mathcal{L}(\mathcal{A})}(a) - p^0_{\mathcal{L}(\mathcal{A})}(a)| < \epsilon\}$ denote an L_1 neighborhood around any $p^0_{\mathcal{L}(\mathcal{A})} \in \mathcal{P}_{2^{V(V-1)/2}}$, we place simple and very general conditions on Π_Z, Π_ν, Π_X and Π_λ, so that the prior Π on $p_{\mathcal{L}(\mathcal{A})}$ induced through (3.5)–(3.6) has full support on $\mathcal{P}_{2^{V(V-1)/2}}$, meaning that $\Pi[\mathbb{B}_\epsilon\{p^0_{\mathcal{L}(\mathcal{A})}\}] > 0$ for any $p^0_{\mathcal{L}(\mathcal{A})} \in \mathcal{P}_{2^{V(V-1)/2}}$ and $\epsilon > 0$. Theorem 3.3 provides sufficient conditions on Π_ν and the prior for the class-specific edge probability vectors Π_π under which the prior Π for $p_{\mathcal{L}(\mathcal{A})}$, induced through representation (3.5), has full support on $\mathcal{P}_{2^{V(V-1)/2}}$. Lemma 3.4 provides sufficient conditions on Π_Z, Π_X, and Π_λ to ensure that the induced prior Π_π through (3.6) meets condition (ii) in Theorem 3.3.

Theorem 3.3. *Let Π be the prior induced on the probability mass function $p_{\mathcal{L}(\mathcal{A})}$ through (3.5) and H^0 be the number of components required to represent $p^0_{\mathcal{L}(\mathcal{A})}$ as in* (3.5)*. Then for any $p^0_{\mathcal{L}(\mathcal{A})} \in \mathcal{P}_{2^{V(V-1)/2}}$, $\Pi[\mathbb{B}_\epsilon\{p^0_{\mathcal{L}(\mathcal{A})}\}] > 0$ for all $\epsilon > 0$ under the following conditions:*

(i) $H \geq H^0$ so that H is an upper bound on H^0;

(ii) $\Pi_\pi\{\pi^{(1)}, \ldots, \pi^{(H)} : \sum_{h=1}^{H} \sum_{l=1}^{V(V-1)/2} |\pi_l^{(h)} - \pi_l^{0(h)}| < \epsilon_\pi\} > 0$, for any collection of edge probability vectors $\{\pi^{0(1)}, \ldots, \pi^{0(H)} : \pi^{0(h)} \in (0,1)^{V(V-1)/2}, h = 1, \ldots, H\}$ and $\epsilon_\pi > 0$;

(iii) $\Pi_\nu\{\mathbb{B}_{\epsilon_\nu}(\nu^0)\} > 0$, for any ν^0 in the probability simplex $\mathcal{P}_H$ and $\epsilon_\nu > 0$.

Proof. As it is always possible to factorize $p^0_{\mathcal{L}(\mathcal{A})}$ according to (3.5), we can express the L_1 distance $\sum_{a \in \mathbb{A}_V} | p_{\mathcal{L}(\mathcal{A})}(a) - p^0_{\mathcal{L}(\mathcal{A})}(a) |$ between $p_{\mathcal{L}(\mathcal{A})}$ and $p^0_{\mathcal{L}(\mathcal{A})}$ as

$$\sum_{a \in \mathbb{A}_V} \left| \sum_{h=1}^{H} \nu_h \prod_{l=1}^{V(V-1)/2} \left\{\pi_l^{(h)}\right\}^{a_l} \left\{1 - \pi_l^{(h)}\right\}^{1-a_l} - \sum_{h=1}^{H} \nu_h^0 \prod_{l=1}^{V(V-1)/2} \left\{\pi_l^{0(h)}\right\}^{a_l} \left\{1 - \pi_l^{0(h)}\right\}^{1-a_l} \right|,$$

with vector $\nu^0 = (\nu_1^0, \ldots, \nu_{H^0}^0, 0_{H-H^0}) \in \mathcal{P}_H$, and H^0 the rank of the tensor $p^0_{\mathcal{L}(\mathcal{A})}$. Hence

$$\Pi[\mathbb{B}_\epsilon\{p^0_{\mathcal{L}(\mathcal{A})}\}] = \int 1 \left(\sum_{a \in \mathbb{A}_V} | \, p_{\mathcal{L}(\mathcal{A})}(a) - p^0_{\mathcal{L}(\mathcal{A})}(a) \, | < \epsilon \right) d\Pi_\nu(\nu) d\Pi_\pi(\pi^{(1)}, \ldots, \pi^{(H)}).$$

Following Dunson and Xing (2009) and recalling the independence between Π_ν and Π_π, a sufficient condition for the latter to be strictly positive is that Π_ν has full support on the probability simplex $\mathcal{P}_H$, and $\Pi_\pi\{\mathbb{B}_{\epsilon_\pi}(\pi^{0(1)}, \ldots, \pi^{0(H)})\} = \Pi_\pi\{\pi^{(1)}, \ldots, \pi^{(H)} : \sum_{h=1}^{H} \sum_{l=1}^{V(V-1)/2} |\pi_l^{(h)} -$

$\pi_l^{0(h)}| < \epsilon_\pi\} > 0$, for any collection $\{\pi^{0(1)}, \ldots, \pi^{0(H)} : \pi^{0(h)} \in (0,1)^{V(V-1)/2}, h = 1, \ldots, H\}$ and $\epsilon_{\pi_h} > 0$, which follow from conditions (iii) and (ii) in Theorem 3.3, proving the result. □

Lemma 3.4. *Let Π_π be the prior for the class-specific edge probability vectors induced by Π_Z, Π_X and Π_λ through (3.6), and denote with R^0 a value of R for which Lemma 3.1 holds, when $p^0_{\mathcal{L}(\mathcal{A})}$ is factorized as in* (3.5) *with H^0 components. Then, the following sufficient conditions imply (ii) in Theorem 3.3:*

(i) $R \geq R^0$ so that R is an upper bound on R^0;

(ii) Π_Z has full L_1 support on $\Re^{V(V-1)/2}$;

(iii) Π_X has full L_1 support on the space of $V \times R$ real matrices $\Re^{V \times R}$;

(iv) Π_λ has full L_1 support on $\Re^R_{\geq 0}$.

Proof. Letting Π_S be the prior on the class-specific latent similarity vectors induced by Π_Z, Π_X and Π_λ through factorization $S^{(h)} = Z + \mathcal{L}(X^{(h)}\Lambda^{(h)}X^{(h)\mathrm{T}})$, $h = 1, \ldots, H$, we first show that for any collection $\{S^{0(1)}, \ldots, S^{0(H)} : S^{0(h)} \in \Re^{V(V-1)/2}, h = 1, \ldots, H\}$ and $\epsilon_s > 0$, $\Pi_S\{\mathbb{B}_{\epsilon_s}(S^{0(1)}, \ldots, S^{0(H)})\} = \Pi_S\{S^{(1)}, \ldots, S^{(H)} : \sum_{h=1}^H \sum_{l=1}^{V(V-1)/2} | S_l^{(h)} - S_l^{0(h)} | < \epsilon_s\} > 0$. Let R be chosen so as to satisfy condition (i), then according to the proof of Lemma 3.1, we can factorize the previous probability as

$$\mathrm{pr}\left\{\sum_{h=1}^{H}\sum_{l=1}^{V(V-1)/2} |Z_l - Z_l^0 + \mathcal{L}(X^{(h)}\Lambda^{(h)}X^{(h)\mathrm{T}})_l - \mathcal{L}(X^{0(h)}\Lambda^{0(h)}X^{0(h)\mathrm{T}})_l| < \epsilon_s\right\}, \tag{3.7}$$

with $\mathrm{diag}(\Lambda^{0(h)}) = \lambda^{0(h)} = (\lambda_1^{0(h)}, \ldots, \lambda_{R^{0(h)}}^{0(h)}, 0_{R-R^{0(h)}})^{\mathrm{T}}$. Under the independence of Π_Z, Π_X and Π_λ, and exploiting the triangle inequality, a lower bound for the previous quantity is

$$\mathrm{pr}\left\{\sum_{l=1}^{V(V-1)/2} | Z_l - Z_l^0 | < \frac{\epsilon_s}{2H}\right\} \prod_{h=1}^{H} \mathrm{pr}\left\{\sum_{l=1}^{V(V-1)/2} | \mathcal{L}(X^{(h)}\Lambda^{(h)}X^{(h)\mathrm{T}})_l - \mathcal{L}(X^{0(h)}\Lambda^{0(h)}X^{0(h)\mathrm{T}})_l | < \frac{\epsilon_s}{2H}\right\}.$$

Hence, (3.7) is positive if both terms are positive. The positivity of the first term follows from (ii) of the Lemma. To prove the positivity of the second term, proof of Lemma 3.1 ensures that for any $\epsilon_s/(2H)$ there exist infinitely many radii $\epsilon_{X^{(h)}}, \epsilon_{\lambda^{(h)}}$, for which $\sum_{v=1}^V \sum_{r=1}^R |X_{vr}^{(h)} - X_{vr}^{0(h)}| < \epsilon_{X^{(h)}}$ and $\sum_{r=1}^R |\lambda_r^{(h)} - \lambda_r^{0(h)}| < \epsilon_{\lambda^{(h)}}$ imply that $\sum_{l=1}^{V(V-1)/2} | \mathcal{L}(X^{(h)}\Lambda^{(h)}X^{(h)\mathrm{T}})_l - \mathcal{L}(X^{0(h)}\Lambda^{0(h)}X^{0(h)\mathrm{T}})_l | < \epsilon_s/(2H)$ for every $h = 1, \ldots, H$. Thus to prove the positivity of the second term and recalling the independence between Π_X and Π_λ, it is sufficient to show that for every $h = 1, \ldots, H$ we have $\Pi_X\{\mathbb{B}_{\epsilon_{X^{(h)}}}(X^{0(h)})\} > 0$, for any $X^{0(h)} \in \Re^{V \times R}$ and $\epsilon_{X^{(h)}} > 0$ and $\Pi_\lambda\{\mathbb{B}_{\epsilon_{\lambda^{(h)}}}(\lambda^{0(h)})\} > 0$, for any $\lambda^{0(h)} \in \Re^R_{\geq 0}$ and $\epsilon_{\lambda^{(h)}} > 0$, representing conditions (iii) and (iv) of the Lemma, respectively.

Let $\pi_l^{0(h)} = 1/\{1+\exp(-S_l^{0(h)})\}$, $l = 1, \ldots, V(V-1)/2$, $h = 1, \ldots, H$, with $S^{0(h)} \in \Re^{V(V-1)/2}$ factorized as before, and denote with Π_π the prior on the class-specific edge probability vectors, induced by Π_S through the one-to-one continuous logistic mapping applied element-wise. To conclude the proof we need to show that $\Pi_\pi\{\mathbb{B}_{\epsilon_\pi}(\pi^{0(1)}, \ldots, \pi^{0(H)})\} > 0$ given that $\Pi_S\{\mathbb{B}_{\epsilon_s}(S^{0(1)}, \ldots, S^{0(H)})\} > 0$ is true. Since the logistic mapping is one-to-one element-wise continuous, by the general definition of continuity, for any $\epsilon_\pi > 0$, there exists an $\epsilon_s > 0$, such that

$$\sum_{h=1}^{H}\sum_{l=1}^{V(V-1)/2} \mid 1/\{1+\exp(-S_l^{(h)})\} - 1/\{1+\exp(-S_l^{0(h)})\} \mid = \sum_{h=1}^{H}\sum_{l=1}^{V(V-1)/2} \mid \pi_l^{(h)} - \pi_l^{0(h)} \mid < \epsilon_\pi,$$

for all collections $\{S^{(1)}, \ldots, S^{(H)} : S^{(h)} \in \Re^{V(V-1)/2}, h = 1, \ldots, H\}$ satisfying condition $\sum_{h=1}^{H}\sum_{l=1}^{V(V-1)/2} |S_l^{(h)} - S_l^{0(h)}| < \epsilon_s$. Since we proved that the event $\sum_{h=1}^{H}\sum_{l=1}^{V(V-1)/2} |S_l^{(h)} - S_l^{0(h)}| < \epsilon_s$ has non-null probability for any $\{S^{0(1)}, \ldots, S^{0(H)} : S^{0(h)} \in \Re^{V(V-1)/2}, h = 1, \ldots, H\}$, by the continuity of the mapping the same holds for $\sum_{h=1}^{H}\sum_{l=1}^{V(V-1)/2} |\pi_l^{(h)} - \pi_l^{0(h)}| < \epsilon_\pi$ for any collection $\{\pi^{0(1)}, \ldots, \pi^{0(H)} : \pi^{0(h)} \in (0,1)^{V(V-1)/2}, h = 1, \ldots, H\}$, concluding the proof. □

Results in Theorem 3.3 and Lemma 3.4 provide simple sufficient conditions on the priors for the components in our factorization under which the induced prior for $p_{\mathcal{L}(\mathcal{A})}$ has full L_1 support. It is additionally important to notice that Lemmas 3.1 and 3.4 hold for more general functions $g(\cdot) : \Re \to (0,1)$ mapping from the latent similarity space to the edge probability space, as long as $g(\cdot)$ is one-to-one continuous.

Large prior support is a key condition of a Bayesian nonparametric model, which also relates to asymptotic behavior of the posterior distribution of $p_{\mathcal{L}(\mathcal{A})}$. The usual asymptotic focus in the network literature is on the case in which the number of nodes $V \to \infty$ in a single network having a particular structure; see Sussman et al. (2012), Tang et al. (2013), Sussman et al. (2014). Our asymptotic theory is unique in the literature in focusing on consistent estimation of the entire population distribution for a network-valued random variable, when the number of network realizations $n \to \infty$. For tractability we focus on the fixed V case, though it is interesting to study the behavior allowing V to increase with n. This is related to a small but growing literature on Bayesian asymptotics in high-dimensional models, but most of the focus has been on substantially simpler models, such as linear regression; see e.g., Ghosal (2000), Ghosal and Belitser (2003), Armagan et al. (2013).

As the pmf for $\mathcal{L}(\mathcal{A})$ is characterized by finitely many parameters $p_{\mathcal{L}(\mathcal{A})}(a), a \in \mathbb{A}_V$, which are all identifiable, full L_1 support is sufficient to guarantee that the posterior distribution assigns probability one to any L_1 neighborhood of the true data-generating probability mass function as the number of networks $n \to \infty$. In particular, we have the strong posterior

consistency property,

$$\lim_{n\to\infty} \Pi[\mathbb{B}_\epsilon\{p^0_{\mathcal{L}(\mathcal{A})}\} \mid \mathcal{L}(A_1), \ldots, \mathcal{L}(A_n)] = 1, \quad \text{for every } \epsilon > 0,$$

with probability one when $p^0_{\mathcal{L}(\mathcal{A})}$ is the true probability mass function.

For Theorem 3.3 and Lemma 3.4 to hold, we need to choose H and R as upper bounds on H^0 and R^0, respectively. Then, the priors for the different components in our factorization are chosen to favor collapsing out the redundant dimensions, so that the posterior will concentrate on $\nu_h \approx 0$ for $h > H^0$ and $\lambda_r^{(h)} \approx 0$ for $r > R^{0(h)}$, with $R^{0(h)}$ denoting the sufficient number of coordinates required to represent the true edge probability vector $\pi^{0(h)}$ via the low-rank factorization in (3.6), for each $h = 1, \ldots, H$. This is achieved by a double shrinkage prior.

The first layer of shrinkage corresponds to deleting extra clusters that are not needed to characterize the data. We take the lead from Rousseau and Mengersen (2011) by letting

$$(\nu_1, \ldots, \nu_H) \sim \text{Dirichlet}\left(\frac{1}{H}, \ldots, \frac{1}{H}\right). \tag{3.8}$$

In a simpler case involving Gaussian mixtures, Rousseau and Mengersen (2011) showed that prior (3.8) will induce effective deletion of the extra mixture components, with the posterior concentrating on the true number of components H^0. It is an active area of research to extend these asymptotic results on over-fitted mixtures to more general settings, but our empirical results suggest that such efficient deletion of extra components also occurs in our case. It is straightforward to verify that condition (iii) in Theorem 3.3 is met under this prior.

The second layer of shrinkage induces collapsing on lower rank structures within each class. As there are infinitely many positive semidefinite matrices having $D^{(h)}$ as lower triangular elements, we are not specifically interested in consistently recovering a true rank for each class-specific deviation vector, but instead look for a prior Π_λ that adaptively deletes redundant latent space dimensions which are not required to characterize $\pi^{(h)}$, $h = 1, \ldots, H$ via (3.6) according to the data. We looked for a similar behavior when defining priors for the scaling parameters in the GP latent coordinates in Section 2.1.2, obtaining good results in simulations and applications. Hence, adapting previous choice to this framework we let $\lambda^{(h)} \sim \text{MIG}(a_1, a_2)$, independently for $h = 1, \ldots, H$, obtaining

$$\lambda_r^{(h)} = \prod_{m=1}^{r} \frac{1}{\vartheta_m^{(h)}}, \quad \vartheta_1^{(h)} \sim \text{Ga}(a_1, 1),\ \vartheta_{m>1}^{(h)} \sim \text{Ga}(a_2, 1), \quad r = 1, \ldots, R, \tag{3.9}$$

independently for each $h = 1, \ldots, H$. Prior (3.9) adaptively penalizes overparameterized representations for each $\pi^{(h)}$, $h = 1, \ldots, H$ by favoring elements $\lambda_r^{(h)}$ to be stochastically

decreasing towards 0 as r increases for appropriate values of a_2; see our discussion in Section 2.1.4 for further details in this prior. Additionally Π_λ has a Markovian structure with $\lambda_r^{(h)} \mid \lambda_{r-1}^{(h)} \sim$ Inv-Ga$(a_2, \lambda_{r-1}^{(h)})$, allowing the joint distribution of $\lambda^{(h)}$ to be factorized as the product of inverse gamma distributions. This property facilitates proving Lemma 3.5, ensuring condition (iv) in Lemma 3.4 is met under this prior choice.

Lemma 3.5. *Let Π_λ correspond to the MIG(a_1, a_2), then Π_λ has full L_1 support on $\Re^R_{\geq 0}$.*

Proof. Let λ^0 be a general vector with R positive elements $\lambda_r^0 \in \Re_{>0}$, $r = 1, \ldots, R$. We first show that $\Pi_\lambda\{\lambda : \sum_{r=1}^R |\lambda_r - \lambda_r^0| < \epsilon_\lambda\} > 0$, when Π_λ coincides with the MIG(a_1, a_2). Letting $\mathbb{B}_{\epsilon_\lambda}(\lambda^0) = \{\lambda : |\lambda_r - \lambda_r^0| < \epsilon_\lambda / R, r = 1, \ldots, R\}$ a lower bound for the previous probability is $\Pi_\lambda\{\mathbb{B}_{\epsilon_\lambda}(\lambda^0)\}$, and exploiting the Markovian property of the MIG(a_1, a_2) we can factorize this probability as $\int_{\mathbb{B}_{\epsilon_\lambda}(\lambda^0)} f(\lambda_1) \prod_{r=2}^R f(\lambda_r \mid \lambda_{r-1}) d\lambda$, where $f(\lambda_r \mid \lambda_{r-1})$ is the conditional density of λ_r given λ_{r-1}.

Hence, the joint MIG(a_1, a_2) prior for λ can be factorized as the product of conditional densities with $\lambda_1 \sim$ Inv-Ga$(a_1, 1)$ and $\lambda_r \mid \lambda_{r-1} \sim$ Inv-Ga(a_2, λ_{r-1}) for each $r = 2, \ldots, R$. Therefore, since the Inv-Ga(a, b) has full support over $\Re_{>0}$ for any $a > 0, b > 0$ and provided that by definition $\lambda_{r-1} > 0$ for every $r = 2, \ldots, R$, it follows that $\Pi_\lambda\{\mathbb{B}_{\epsilon_\lambda}(\lambda^0)\} > 0$. This proof holds also for vectors $\lambda^0 = (\lambda_1^0, \ldots, \lambda_{R^0}^0, 0_{R-R^0})^{\mathrm{T}} \in \Re_{\geq 0}$ with non negative elements as every neighborhood of λ^0 contains a subset of $\Re^R_{>0}$ for which prior support has been shown. This concludes the proof. □

Finally, priors Π_Z and Π_X are chosen to meet conditions (ii) and (iii), respectively, in Lemma 3.4, while favoring simple posterior computation. Consistently with these aims we assume

$$Z \sim \mathrm{N}_{V(V-1)/2}(\mu, \Sigma), \quad \mu \in \Re^{V(V-1)/2}, \; \Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma^2_{V(V-1)/2}). \tag{3.10}$$

Prior Π_X is defined by assigning independent standard Gaussians

$$X_{vr}^{(h)} \sim \mathrm{N}(0, 1), \quad v = 1, \ldots, V, \;\; r = 1, \ldots, R, \;\; h = 1, \ldots, H. \tag{3.11}$$

Beside meeting full prior support conditions and leading to efficient posterior computation, the previous choices allow simple derivations for the prior moments of the class-specific log-odds $S_l^{(h)} = Z_l + \sum_{r=1}^R \lambda_r^{(h)} X_{vr}^{(h)} X_{ur}^{(h)}$ for each $h = 1, \ldots, H$ and $l = 1, \ldots, V(V-1)/2$, with (v, u) denoting the pair of nodes indexed by l. Specifically, based on priors (3.10)-(3.11) and conditioning on $\lambda^{(h)}$ to highlight their effect in the prior, it is straightforward to show that

$$\mathrm{E}\{S_l^{(h)} \mid \lambda^{(h)}\} = \mu_l, \quad \mathrm{var}\{S_l^{(h)} \mid \lambda^{(h)}\} = \sigma_l^2 + \sum_{r=1}^R \{\lambda_r^{(h)}\}^2, \quad \mathrm{cov}\{S_l^{(h)}, S_{l^*}^{(h)} \mid \lambda^{(h)}\} = 0, \tag{3.12}$$

for each $h = 1, \ldots, H$, $l = 1, \ldots, V(V-1)/2$ and $l^* = 1, \ldots, V(V-1)/2$ with $l^* \neq l$. The covariance between log-odds in classes $h = 1, \ldots, H$ and $h^* = 1, \ldots, H$ with $h^* \neq h$ is instead

$$\begin{aligned} \text{cov}\{S_l^{(h)}, S_l^{(h^*)} \mid \lambda^{(h)}, \lambda^{(h^*)}\} &= \sigma_l^2, \quad l = 1, \ldots, V(V-1)/2, \\ \text{cov}\{S_l^{(h)}, S_{l^*}^{(h^*)} \mid \lambda^{(h)}, \lambda^{(h^*)}\} &= 0, \quad l^* \neq l. \end{aligned}$$

A priori the log-odd of a given edge has the same mean μ_l in all classes with σ_l^2 controlling the edge-specific component of variability shared across groups as well as the covariance between the log-odds of the same edge across different classes. Parameters $\lambda^{(h)}$ add instead a class-specific component of variability in the log-odds of each edge. When the $\lambda^{(h)}$ are all close to zero, the correlation between the log-odds for the same edge across different groups is close to one collapsing model (3.5)–(3.6) to (3.1). The prior covariance between the log-odds of different edges is instead 0.

3.1.5 Posterior computation

Given priors defined as in equations (3.8)–(3.11), posterior computation for the statistical model having likelihood (3.5) with $\pi^{(h)}$ from (3.6) is available in a simple form adapting Polson et al. (2013) Pólya-gamma data augmentation for Bayesian logistic regression.

Specifically, the proposed Gibbs sampler exploits the graphical representation of our hierarchical construction (3.3)–(3.4) outlined in Figure 3.2 to first allocate each observation $\mathcal{L}(A_i)$, $i = 1, \ldots, n$, into one of the classes and then updates Z, $X^{(h)}$, $\lambda^{(h)}$, for $h = 1, \ldots, H$, via Bayesian logistic regression within each class. Detailed steps for the Gibbs sampler are outlined in Algorithm 3.

Algorithm 3 Gibbs sampler for the mixture of low-rank factorizations model

---

[1] Allocate each network observation to one of the classes

for $i = 1, \ldots, n$ **do**

Sample the class indicator G_i from the full conditional discrete distribution with

$$\text{pr}(G_i = h \mid -) = \frac{\nu_h \prod_{l=1}^{V(V-1)/2} \{\pi_l^{(h)}\}^{\mathcal{L}(A_i)_l} \{1 - \pi_l^{(h)}\}^{1-\mathcal{L}(A_i)_l}}{\sum_{m=1}^{H} \nu_m \prod_{l=1}^{V(V-1)/2} \{\pi_l^{(m)}\}^{\mathcal{L}(A_i)_l} \{1 - \pi_l^{(m)}\}^{1-\mathcal{L}(A_i)_l}},$$

for each $h = 1, \ldots, H$, with $\pi^{(h)}$ defined in (3.6)

end for

---

[2] Sample the class probabilities $\nu_1, \ldots, \nu_H$ from the full conditional Dirichlet

$$(\nu_1, \ldots, \nu_H) \mid - \sim \text{Dirichlet}\left\{\frac{1}{H} + \sum_{i=1}^{n} 1(G_i = 1), \ldots, \frac{1}{H} + \sum_{i=1}^{n} 1(G_i = H)\right\}.$$

Comment: Recalling representation (3.3)–(3.4), networks in the same class are independent and identically distributed conditionally on the class-specific edge probability vectors $\pi^{(h)}$, $h = 1, \ldots, H$. Hence, to update Z, $X^{(h)}$ and $\lambda^{(h)}$, $h = 1, \ldots, H$ at each step, it is sufficient to adapt Polson et al. (2013) Pólya-gamma data augmentation for aggregated networks $\mathcal{L}(A^{(1)}), \ldots, \mathcal{L}(A^{(H)})$, with $\mathcal{L}(A^{(h)}) = \sum_{G_i=h} \mathcal{L}(A_i)$, for $h = 1, \ldots, H$ and, according to our model formulation,

$$\mathcal{L}(A^{(h)})_l \mid Z, X^{(h)}, \lambda^{(h)} \quad \sim \text{Binom}[n_h, 1/\{1 + \exp(-Z_l - \mathcal{L}(X^{(h)}\Lambda^{(h)}X^{(h)\text{T}})_l)\}],$$

independently for $l = 1, \ldots, V(V-1)/2$ and $h = 1, \ldots, H$, with n_h the number of networks in class h at a given iteration. This provides also a key result in reducing the computational complexity, as at each step the number of augmented Pólya-gamma variables to be sampled depends on the number of classes instead of the sample size n. Hence, after the grouping steps, the MCMC proceeds as follows

---

[3] Sample Pólya-gamma augmented data
for each $l = 1, \ldots, V(V-1)/2$ and $h = 1, \ldots, H$ **do**
Update each augmented data $\omega_l^{(h)}$ from the full conditional Pólya-gamma

$$\omega_l^{(h)} \mid - \sim \text{PG}\left\{n_h, Z_l + \mathcal{L}(X^{(h)}\Lambda^{(h)}X^{(h)\text{T}})_l\right\},$$

for every $h = 1, \ldots, H$ and $l = 1, \ldots, V(V-1)/2$, with $\text{PG}(b, c)$ denoting the Pólya-gamma distribution with parameters $c \in \Re$ and $b > 0$.
end for

---

[4] Sample the shared similarity vector Z from its Gaussian full conditional

$$Z \mid - \sim \text{N}_{V(V-1)/2}\{\mu_Z, \text{diag}(\sigma^2_{Z_1}, \ldots, \sigma^2_{Z_{V(V-1)/2}})\},$$

with μ_Z having elements $\mu_{Z_l} = \sigma^2_{Z_l}[\sigma_l^{-2}\mu_l + \sum_{h=1}^{H}\{\mathcal{L}(A^{(h)})_l - n_h/2 - \omega_l^{(h)}\mathcal{L}(X^{(h)}\Lambda^{(h)}X^{(h)\text{T}})_l\}]$, where $\sigma^2_{Z_l} = 1/(\sigma_l^{-2} + \sum_{h=1}^{H}\omega_l^{(h)})$, for each $l = 1, \ldots, V(V-1)/2$.

Comment: To maintain conjugacy in sampling the class-specific parameters defining $D^{(h)}$, $h = 1, \ldots, H$, we reparameterize the model to update quantities $\bar{X}^{(h)} = X^{(h)}\Lambda^{(h)1/2}$ and

$\Lambda^{(h)}$, $h = 1, \ldots, H$. Hence, we can rewrite $D^{(h)} = \mathcal{L}(\bar{X}^{(h)}\bar{X}^{(h)\mathrm{T}})$, and according to our prior specification $\bar{X}^{(h)}_{vr}|\lambda^{(h)}_r \sim \mathrm{N}(0, \lambda^{(h)}_r)$ independently for $v = 1, \ldots, V$, $r = 1, \ldots, R$ and $h = 1, \ldots, H$, with independent MIG(a_1, a_2) priors on $\lambda^{(h)}$. Hence, the Gibbs sampler proceeds af follows

---

[5] Sample the class-specific weighted matrices $\bar{X}^{(1)}, \ldots, \bar{X}^{(H)}$
for $h = 1, \ldots, H$ and $v = 1, \ldots, V$ **do**
Block-sample the vth row of $\bar{X}^{(h)}$.

1. Define $\bar{X}^{(h)}_v = (\bar{X}^{(h)}_{v1}, \ldots, \bar{X}^{(h)}_{vR})^{\mathrm{T}}$ and let $\bar{X}^{(h)}_{(-v)}$ denote the $(V-1) \times R$ matrix obtained by removing the vth row in $\bar{X}^{(h)}$. Consider the logistic regression

$$\mathcal{L}(A^{(h)})_{(v)} \sim \mathrm{Binom}(n_h, \pi^{(h)}_{(v)}), \quad \mathrm{logit}(\pi^{(h)}_{(v)}) = Z_{(v)} + \bar{X}^{(h)}_{(-v)}\bar{X}^{(h)}_v,$$

with $\mathcal{L}(A^{(h)})_{(v)}$ and $Z_{(v)}$ obtained by stacking elements $\mathcal{L}(A^{(h)})_l$ and Z_l, respectively, for all l corresponding to pairs having v as a one of the two nodes, and ordered consistently with the linear predictor.

2. Exploiting previous formulation, and letting $\Omega^{(h)}_{(v)}$ be the diagonal matrix with the corresponding Pólya-gamma augmented data, the full conditional is

$$\bar{X}^{(h)}_v \mid - \sim \mathrm{N}_R\left\{\left(\bar{X}^{(h)\mathrm{T}}_{(-v)}\Omega^{(h)}_{(v)}\bar{X}^{(h)}_{(-v)} + \Lambda^{(h)^{-1}}\right)^{-1}\eta^{(h)}_v, \left(\bar{X}^{(h)\mathrm{T}}_{(-v)}\Omega^{(h)}_{(v)}\bar{X}^{(h)}_{(-v)} + \Lambda^{(h)^{-1}}\right)^{-1}\right\},$$

with $\eta^{(h)}_v = \bar{X}^{(h)\mathrm{T}}_{(-v)}\{\mathcal{L}(A^{(h)})_{(v)} - 1_{V-1}n_h/2 - \Omega^{(h)}_{(v)}Z_{(v)}\}$

end for

---

[6] Sample the gamma quantities defining the shrinkage weights $\lambda^{(1)}, \ldots, \lambda^{(H)}$
for $h = 1, \ldots, H$ **do**

$$\begin{aligned}
\vartheta^{(h)}_1 \mid - &\sim \mathrm{Ga}\left\{a_1 + \frac{VR}{2}, 1 + \frac{1}{2}\sum_{m=1}^{R}\theta^{(-1)}_m\sum_{v=1}^{V}(\bar{X}^{(h)}_{vm})^2\right\},\\
\vartheta^{(h)}_{r>1} \mid - &\sim \mathrm{Ga}\left\{a_2 + \frac{V \times (R-r+1)}{2}, 1 + \frac{1}{2}\sum_{m=r}^{R}\theta^{(-r)}_m\sum_{v=1}^{V}(\bar{X}^{(h)}_{vm})^2\right\},
\end{aligned}$$

where $\theta^{(-r)}_m = \prod_{t=1, t\neq r}^{m}\vartheta^{(h)}_t$ for $r = 1, \ldots, R$.
end for

The above steps are all straightforward and mixing is efficient in our experience. Moreover, given MCMC chains for the previous quantities the class-specific edge probability vectors $\pi^{(h)}$ are easily available as $\pi^{(h)} = [1 + \exp\{-Z - D^{(h)}\}]^{-1}$ with $D^{(h)} = \mathcal{L}(\bar{X}^{(h)}\bar{X}^{(h)\mathrm{T}})$, for each

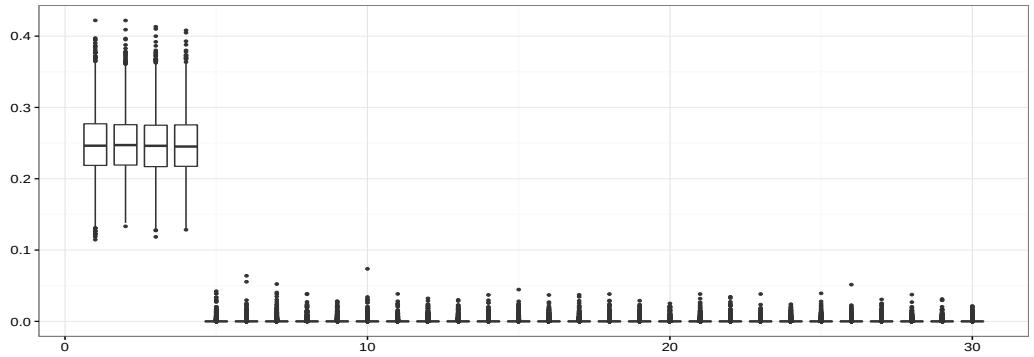


FIGURE 3.3: Boxplots summarizing the posterior distribution of the mixing probabilities. Latent classes are reordered to be in decreasing order of the posterior mean of ν_h.

$h = 1, \ldots, H$. Finally to obtain a posterior distribution for $p_{\mathcal{L}(\mathcal{A})}$ it sufficient to apply equation (3.5) to the posterior samples of $\pi^{(h)}$ and ν_h, $h = 1, \ldots, H$.

3.1.6 Simulation study

We conduct simulation studies to evaluate the performance of our approach in accurately estimating the population distribution of network data, accounting for broad differences in network properties across classes of networks. Of particular interest are community structures (Faust and Wasserman, 1992), scale freeness (Barabási and Albert, 1999), small-worldness (Watts and Strogatz, 1998) and classical random graph behaviors (Erdös and Rényi, 1959). Although the latter is overly-restrictive and rarely met in applications, it provides a null model in many network analyses.

We consider four latent classes and simulate 25 networks for each class by sampling their edges from conditionally independent Bernoulli random variables given their corresponding class-specific edge probabilities. We focus on networks having $V = 20$ nodes to facilitate graphical presentation. Each class-specific edge probability vector is carefully constructed to assign high probability to a subset of network configurations characterized by a specific property via (3.1). In particular, one class is associated with simulated networks characterized by two latent communities. Networks generated under a second class have a behavior similar to Erdös and Rényi (1959) random graphs. Another class assigns high probability to scale free networks generated under the Barabási and Albert (1999) model. Finally networks in the remaining class display small-world properties according to the Watts and Strogatz (1998) generative model. Figure 3.1 provides a graphical representation of these true edge probability vectors rearranged in matrix form.

The goal in defining this challenging simulation scenario is to assess whether our approach can accurately characterize a collection of networks having such broad and widely different

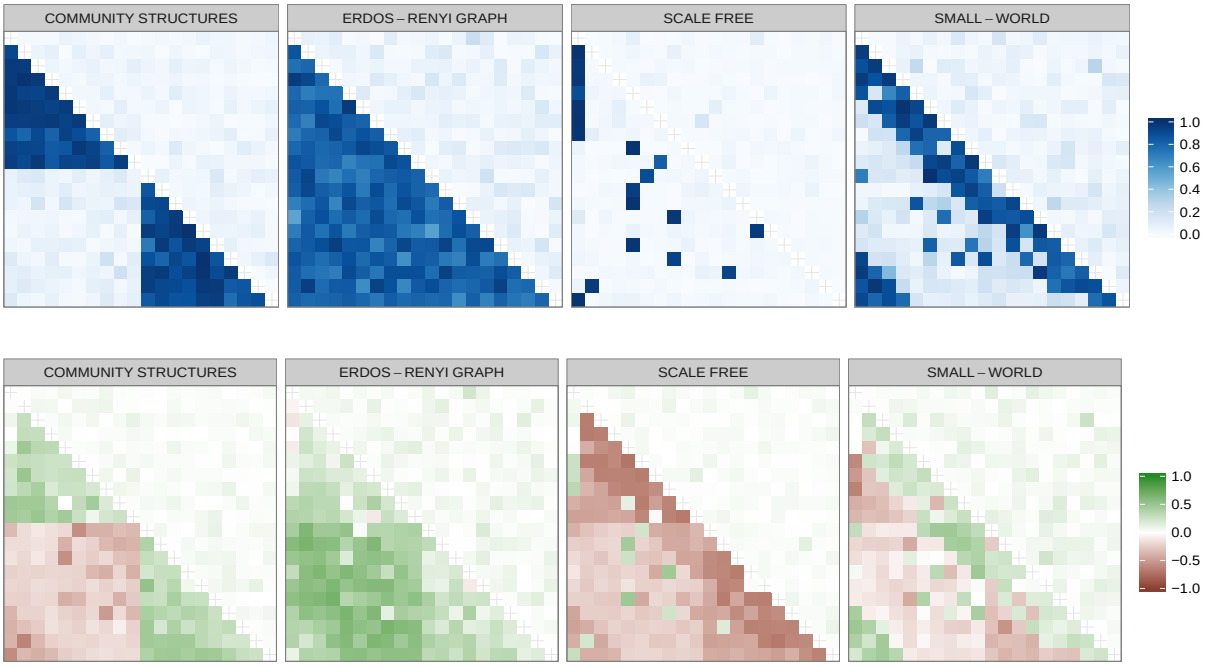


FIGURE 3.4: Upper panel: for each non-empty latent class, posterior mean $\hat{\pi}^{(h)}$ of $\pi^{(h)}$ (lower triangular) and absolute value of the difference $|\hat{\pi}^{(h)} - \pi^{0(h)}|$ between estimated and true values (upper triangular). Lower panel: for the same classes posterior mean $\hat{\pi}^{(h)} - \hat{\bar{\pi}}$ of $\pi^{(h)} - \bar{\pi}$ (lower triangular) and absolute value of the difference $|\hat{\pi}^{(h)} - \hat{\bar{\pi}} - \pi^{0(h)} + \bar{\pi}^0|$ between estimated and true values (upper triangular).

properties. We analyze the simulated data under model (3.5)–(3.6) with priors (3.8)–(3.11). Exploiting results in (3.12), we consider $\mu_1 = \ldots = \mu_{V(V-1)/2} = 0$ to obtain priors for each $\pi^{(h)}$ centered on the Erdös and Rényi (1959) random graph, and let $\sigma_1^2 = \ldots = \sigma_{V(V-1)/2}^2 = 10$ to represent uncertainty in this shared structure. To favor deletion of unnecessary dimensions in each class, we consider $a_1 = 2.5$ and $a_2 = 3.5$ in the MIG(a_1, a_2) prior. This enforces adaptive shrinkage for growing r – as outlined in Section 2.1.4 – allows class-specific variability in the prior for each $\pi^{(h)}$ according to (3.12), and ensures the existence of the first two moments for the induced priors on elements $D_l^{(h)}$.

Our approach can be easily modified to learn the hyperparameters from the data via hyperpriors on quantities a_1, a_2 and μ_l, σ_l^2 for each $l = 1, \ldots, V(V-1)/2$. However, we obtained similar results when instead considering other hyperparameter settings, such as $\mu_l \in (-1, -0.5, 0.5, 1)$, $\sigma_l^2 \in (1, 100, 200)$ for each $l = 1, \ldots, V(V-1)/2$ and $a_1 \in (5, 10)$, $a_2 \in (5, 10)$. Higher values for a_1 and a_2 are not recommended in inducing priors on $\lambda^{(h)}$, $h = 1, \ldots, H$ strongly concentrated at 0, forcing $D^{(h)} \approx 0$. As a result, the edge probability vectors are forced to be equal across the different classes a priori, collapsing model (3.5)–(3.6) to (3.1).

We generate $5{,}000$ Gibbs iterations, with upper bounds $H = 30$ and $R = 10$, and set a burn-in of $1{,}000$. Trace-plots and Gelman and Rubin (1992) potential scale reduction factors for the quantities investigated in Figures 3.3–3.4 suggest this burn-in is sufficient for convergence

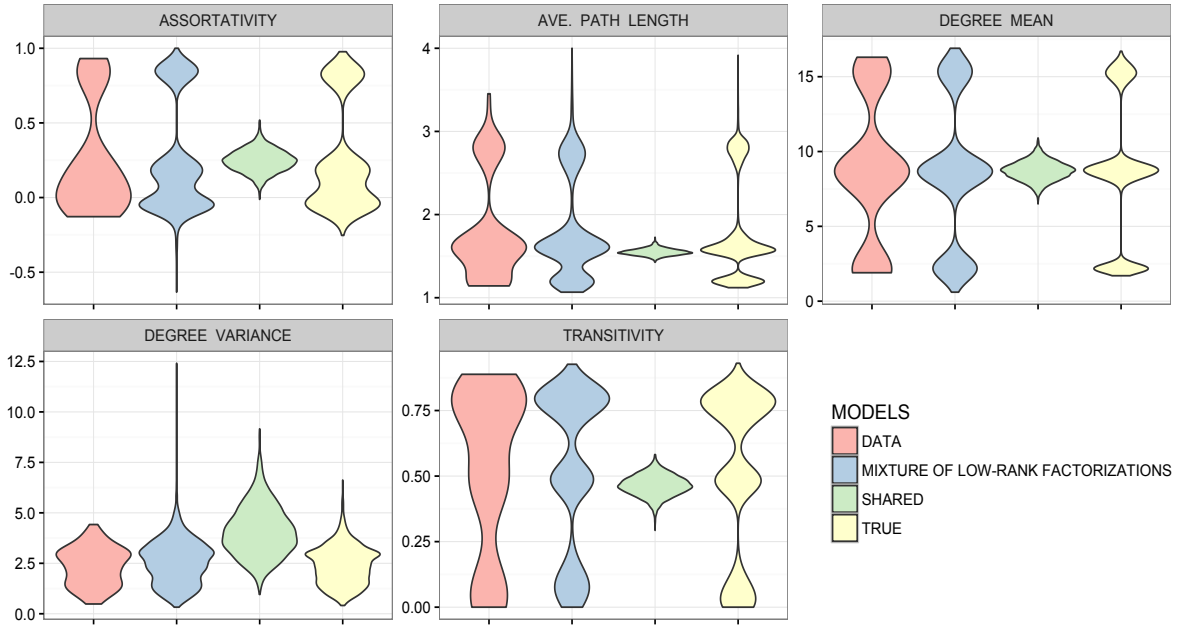


FIGURE 3.5: Violin plots showing the distribution of key topological features computed from networks simulated under different scenarios. DATA: computed from our 100 simulated networks. MIXTURE OF LOW-RANK FACTORIZATIONS: computed on 4,000 networks simulated from the posterior predictive distribution of our model. SHARED: computed on 4,000 networks whose edges are simulated from conditionally independent Bernoulli with edge probabilities given by $\sum_{i=1}^{n} \mathcal{L}(A_i)_l/n$, $l = 1, \ldots, V(V-1)/2$. TRUE: computed on 4,000 networks simulated from our model as in Figure 3.2 with $\pi^{(h)}$ and ν_h, $h = 1, \ldots, H$ set at true values.

and show no evidence of label switching issues. We additionally monitor mixing via effective sample sizes for the same quantities, with most of these values $\approx 1{,}200$ out of $4{,}000$, providing a good mixing result. The algorithm required 43 minutes to perform posterior computation based on a naive R (version 3.1.1) implementation in a machine with one Intel Core i5 2.3GHz processor and 4GB of RAM. As the posterior for $p_{\mathcal{L}(A)}$ is a complex object to visualize, we evaluate inference performance by focusing on posteriors for quantities ν_h and $\pi^{(h)}$, $h = 1, \ldots, H$, which characterize $p_{\mathcal{L}(A)}$ under equation (3.5).

Figure 3.3 highlights the good performance of Π_ν in adaptively deleting redundant classes while properly estimating the true mixing probabilities. We obtain similar good results in recovering the true class-specific edge probability vectors and their deviations from the expected value of the network-valued random variable as shown in Figure 3.4. Borrowing of information across replicated observations and within each network provides very accurate inference for the class-specific edge probabilities under very different scenarios, with the 0.95 highest posterior density intervals containing the true class-specific edge probabilities in 90% of the replicates. Figure 3.4 confirms the ability of the model to accurately approximate and efficiently estimate a very broad range of true network structures. Inferences are robust to the choice of upper bounds $R = 10 < V = 20$ due to the carefully specified prior favoring adaptive deletion of extra dimensions.

Accurate estimation of the mixing probabilities ν_h and the class-specific edge probability vectors $\pi^{(h)}$, $h = 1, \ldots, H$ ensures accurate inference and estimation for the true pmf $p^0_{\mathcal{L}(\mathcal{A})}$ arising from (3.5), while providing efficient clustering behavior. We correctly group all the simulated networks in four latent classes via maximum a posteriori estimates (MAP) of their G_i using the MCMC samples. We obtained similarly good performance in very different simulations having more subtle differences between classes mimicking the brain network application data.

Figure 3.5 presents violin plots demonstrating the ability of our model to accurately characterize the distribution of key topological features with respect to those associated with the true pmf $p^0_{\mathcal{L}(\mathcal{A})}$. Modeling of $p_{\mathcal{L}(\mathcal{A})}$ via (3.1), as in the previous literature, significantly reduces performance even when considering a different parameter for each edge probability, with these parameters consistently estimated by exploiting all the $n = 100$ simulated network data via $\hat{\pi} = \sum_{i=1}^{n} \mathcal{L}(A_i)/n$. Such an approach averages across network behaviors and hence cannot capture multi-modality patterns indicative of subsets of data characterized by different network properties.

Note that – in line with the discussion in Section 2.2.4 – the distributions of the network summary statistics in Figure 3.5 for our procedure are based on the posterior predictive distribution associated to our model. Although the latter is not analytically available, it is straightforward to simulate from the posterior predictive distribution exploiting our constructive representation in Figure 3.2 and posterior samples for the quantities in (3.5)–(3.6). Specifically for each MCMC sample of the parameters in (3.5)–(3.6) – after convergence – we generate a network from our model exploiting the mechanism in Figure 3.2, to obtain the desired samples form the posterior predictive distribution.

3.1.7 Global and local testing for group differences in brain networks

The model described in Section 3.1.3 provides a tractable and provably general characterization for the pmf of a network-valued random variable. However, when the focus is on inference and testing on changes in this pmf across groups, previous methodologies need to be generalized.

Let $p_{\mathcal{Y},\mathcal{L}(\mathcal{A})}$ be the joint pmf for the random variable $\{\mathcal{Y}, \mathcal{L}(\mathcal{A})\}$ with $p_{\mathcal{Y},\mathcal{L}(\mathcal{A})}(y, a) = \text{pr}\{\mathcal{Y} = y, \mathcal{L}(\mathcal{A}) = a\}$, $y \in \mathbb{Y} = \{1, 2\}$ and $a \in \mathbb{A}_V$ a network configuration. Assessing evidence of global association between $\mathcal{Y}$ and $\mathcal{L}(\mathcal{A})$, formally requires testing the system of hypotheses

$$H_0 : p_{\mathcal{Y},\mathcal{L}(\mathcal{A})} = p_{\mathcal{Y}} p_{\mathcal{L}(\mathcal{A})} \quad \text{versus} \quad H_1 : p_{\mathcal{Y},\mathcal{L}(\mathcal{A})} \neq p_{\mathcal{Y}} p_{\mathcal{L}(\mathcal{A})}, \tag{3.13}$$

where $p_{\mathcal{Y}} \in \mathcal{P}_2$ is the marginal pmf of the grouping variable and $p_{\mathcal{L}(\mathcal{A})} \in \mathcal{P}_{|\mathbb{A}_V|}$ denotes the unconditional pmf for the network-valued random variable $p_{\mathcal{L}(\mathcal{A})}(a) = \text{pr}\{\mathcal{L}(\mathcal{A}) = a\}$, $a \in \mathbb{A}_V$. System (3.13) assesses evidence of global changes in the entire probability mass function, rather than on selected functionals or summary statistics, and hence is more general than Ginestet et al. (2014) and joint tests on network measures.

Recalling our neuroscience application, rejection of H_0 implies that there are differences in the brain architecture across creativity groups, but fails to provide insights on the reasons for this association. The global differences may be attributable to several underlying mechanisms, including variations in specific interconnection circuits. As discussed in Section 1.2.1, local testing on group changes in edge probabilities is of key interest in neuroscience applications in highlighting which brain connections variables $\mathcal{L}(\mathcal{A})_l \in \{0, 1\}$, $l = 1, \ldots, V(V-1)/2$ – characterizing the marginals of $\mathcal{L}(\mathcal{A})$ – are potentially responsible for the global association between $\mathcal{Y}$ and $\mathcal{L}(\mathcal{A})$. Hence, consistently with these interests, we incorporate in our analyses also the multiple local tests

$$H_{0l} : p_{\mathcal{Y},\mathcal{L}(\mathcal{A})_l} = p_{\mathcal{Y}} p_{\mathcal{L}(\mathcal{A})_l} \quad \text{versus} \quad H_{1l} : p_{\mathcal{Y},\mathcal{L}(\mathcal{A})_l} \neq p_{\mathcal{Y}} p_{\mathcal{L}(\mathcal{A})_l}, \tag{3.14}$$

for each $l = 1, \ldots, V(V-1)/2$, to assess whether each brain connection $\mathcal{L}(\mathcal{A})_l$ has no association with $\mathcal{Y}$, or differs across low and high creativity subjects, respectively. In the system (3.14), the quantity $p_{\mathcal{Y},\mathcal{L}(\mathcal{A})_l}(y, a_l)$ denotes $\text{pr}\{\mathcal{Y} = y, \mathcal{L}(\mathcal{A})_l = a_l\}$, while $p_{\mathcal{L}(\mathcal{A})_l}(a_l) = \text{pr}\{\mathcal{L}(\mathcal{A})_l = a_l\}$, $l = 1, \ldots, V(V-1)/2$, $a_l \in \{0, 1\}$.

In order to develop robust and tractable methodologies to test the global system (3.13) and the multiple locals in (3.14), it is fundamental to consider a representation for $p_{\mathcal{Y},\mathcal{L}(\mathcal{A})}$ which is provably flexible in approximating any joint probabilistic generative mechanism underlying data (y_i, A_i), $i = 1, \ldots, n$. As $\mathcal{L}(\mathcal{A})$ is an highly multidimensional random variable on a non-standard space we additionally require dimensionality reduction in characterizing $p_{\mathcal{Y},\mathcal{L}(\mathcal{A})}$, while looking for a representation which facilitates simple derivation of $p_{\mathcal{Y},\mathcal{L}(\mathcal{A})_l}(y, a_l)$ and $p_{\mathcal{L}(\mathcal{A})_l}(a_l)$ from $p_{\mathcal{Y},\mathcal{L}(\mathcal{A})}$. This is a key to ensure tractable methodologies for testing the multiple local systems in (3.14). According to these goals, we start by factorizing $p_{\mathcal{Y},\mathcal{L}(\mathcal{A})}$ as

$$p_{\mathcal{Y},\mathcal{L}(\mathcal{A})}(y, a) = p_{\mathcal{Y}}(y) p_{\mathcal{L}(\mathcal{A})|y}(a) = \text{pr}(\mathcal{Y} = y)\text{pr}\{\mathcal{L}(\mathcal{A}) = a \mid \mathcal{Y} = y\}, \tag{3.15}$$

for every $y \in \mathbb{Y}$ and $a \in \mathbb{A}_V$. It is always possible to characterize the joint pmf $p_{\mathcal{Y},\mathcal{L}(\mathcal{A})} \in \mathcal{P}_{2\times|\mathbb{A}_V|}$ as the product of the marginal $p_{\mathcal{Y}} \in \mathcal{P}_2$ for the grouping variable and the conditional pmfs $p_{\mathcal{L}(\mathcal{A})|y} \in \mathcal{P}_{|\mathbb{A}_V|}$ of the network-valued random variable given the group membership $y \in \mathbb{Y}$. This also favors inference on how the network structure varies across the two groups, with $p_{\mathcal{L}(\mathcal{A})|1}$ and $p_{\mathcal{L}(\mathcal{A})|2}$ fully characterizing such variations. Although we treat $\mathcal{Y}$ as a random variable through a prospective likelihood, the method we propose is valid also for studies that sample groups under a retrospective design.

Under factorization (3.15), the global test coincides with assessing whether the conditional pmf of the network-valued random variable remains equal or shifts across the two groups. Hence, under (3.15), the system (3.13), reduces to

$$H_0 : p_{\mathcal{L}(\mathcal{A})|1} = p_{\mathcal{L}(\mathcal{A})|2} \quad \text{versus} \quad H_1 : p_{\mathcal{L}(\mathcal{A})|1} \neq p_{\mathcal{L}(\mathcal{A})|2}. \tag{3.16}$$

In order to develop provably general and robust strategies to test (3.16) the key challenge relies in flexibly modeling the conditional pmfs $p_{\mathcal{L}(\mathcal{A})|1}$ and $p_{\mathcal{L}(\mathcal{A})|2}$ characterizing the distribution of the network-valued random variable in the first and second group, respectively. For every group $y \in \mathbb{Y}$, one needs a parameter $p_{\mathcal{L}(\mathcal{A})|y}(a)$ for each network configuration $a \in \mathbb{A}_V$ to uniquely characterize $p_{\mathcal{L}(\mathcal{A})|y}$, with the number of possible configurations being $|\mathbb{A}_V| = 2^{V(V-1)/2}$. Hence, a possible naive procedure to test the system (3.16) is to jointly assess evidence of $H_0 : p_{\mathcal{L}(\mathcal{A})|1}(a) = p_{\mathcal{L}(\mathcal{A})|2}(a)$ for every $a \in \mathbb{A}_V$, against the alternative $H_1 : p_{\mathcal{L}(\mathcal{A})|1}(a) \neq p_{\mathcal{L}(\mathcal{A})|2}(a)$ for some $a \in \mathbb{A}_V$. Although this strategy is fully general and robust against model misspecification, in our motivating application, $|\mathbb{A}_{68}| = 2^{68(68-1)/2} - 1 = 2^{2278} - 1$ parameters are required to uniquely define the pmf of the brain network in each group $y \in \mathbb{Y}$ under the usual restriction $\sum_{a \in \mathbb{A}_{68}} p_{\mathcal{L}(\mathcal{A})|y}(a) = 1$. Clearly this number of parameters to test is massively larger than the sample size available in neuroscience applications. Hence, to facilitate tractable testing procedures it is necessary to substantially reduce dimensionality. However, in reducing dimension, it is important to avoid making overly restrictive assumptions that lead to formulations sensitive to issues arising from model misspecification.

Methodologies developed in Sections 3.1.3–3.1.5 address this dimensionality issue in modeling of the network's pmf without a categorical response, via mixture of low-rank factorizations. We generalize this approach to characterize changes of $\mathcal{L}(\mathcal{A})$ across groups, while accommodating tractable procedures for global and local testing on group differences in the network structure. This is accomplished via a simple modification of equation (3.5), which replaces group-constant mixing probabilities $\nu = (\nu_1, \ldots, \nu_H) \in \mathcal{P}_H$ with group-specific quantities $\nu_y = (\nu_{1y}, \ldots, \nu_{Hy}) \in \mathcal{P}_H$ for each $y \in \{1, 2\}$, while maintaing the same low-rank factorization (3.6) for the class-specific edge probability vectors $\pi^{(h)}$, $h = 1, \ldots, H$. Replacing ν with ν_y in equation (3.5), leads to the dependent mixture representation

$$p_{\mathcal{L}(\mathcal{A})|y}(a) = \text{pr}\{\mathcal{L}(\mathcal{A}) = a \mid \mathcal{Y} = y\} = \sum_{h=1}^{H} \nu_{hy} \prod_{l=1}^{V(V-1)/2} \left\{\pi_l^{(h)}\right\}^{a_l} \left\{1 - \pi_l^{(h)}\right\}^{1-a_l}, \tag{3.17}$$

for each group $y \in \{1, 2\}$ and configuration $a \in \mathbb{A}_V$, where ν_{hy} is the probability that the brain network of a randomly selected subject within predictor group $y_i = y$ is allocated to class h, and $\pi_l^{(h)} \in (0, 1)$ – from factorization (3.6) – defines the probability of an edge among the lth pair of nodes in class h, for each $l = 1, \ldots, V(V-1)/2$ and $h = 1, \ldots, H$. Hence, representation (3.17) defines $p_{\mathcal{L}(\mathcal{A})|1}$ and $p_{\mathcal{L}(\mathcal{A})|2}$ via a flexible dependent mixture model, which

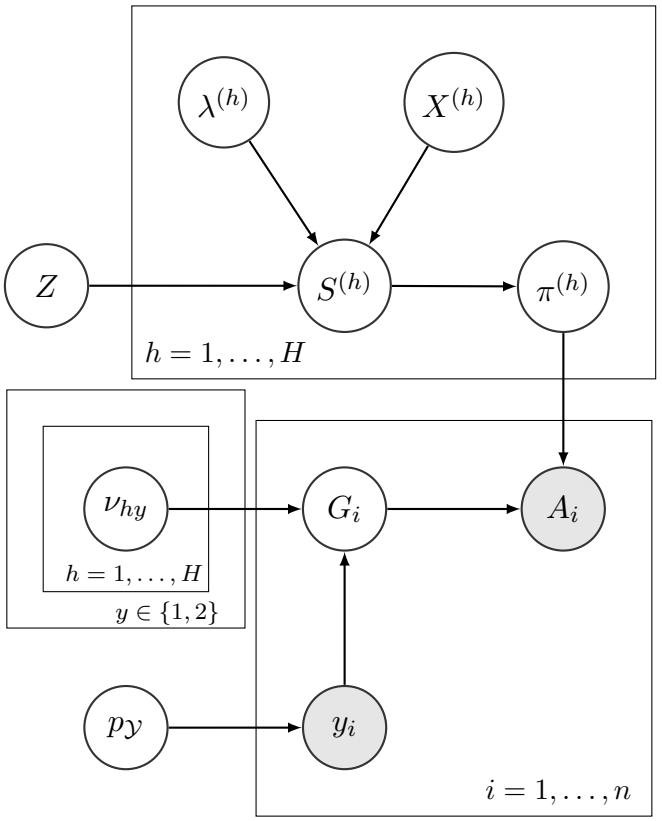


FIGURE 3.6: Graphical representation of the mechanism to generate data $\{y_i, \mathcal{L}(A_i)\}$, under representation (3.15) and (3.17) for the joint pmf $p_{\mathcal{Y},\mathcal{L}(\mathcal{A})}$, with class-specific edge probability vectors $\pi^{(h)} = (\pi_1^{(h)}, \ldots, \pi_{V(V-1)/2}^{(h)})^{\mathrm{T}}$ factorized as in (3.6).

borrows strength across the two groups in characterizing the shared mixture components, while allowing flexible modeling of the conditional pmfs $p_{\mathcal{L}(\mathcal{A})|y}$ via group-specific mixing probabilities $\nu_y = (\nu_{1y}, \ldots, \nu_{Hy}) \in \mathcal{P}_H$ for $y = 1$ and $y = 2$.

Figure 3.6 outlines the mechanism to generate data $\{y_i, \mathcal{L}(A_i)\}$ from the random variable $\{\mathcal{Y}, \mathcal{L}(\mathcal{A})\}$ with pmf factorized as in (3.15), (3.17) and class-specific edge probability vectors from (3.6). According to Figure 3.6 the group indicator y_i is sampled from the categorical random variable with two levels and pmf $p_{\mathcal{Y}}$. The network $\mathcal{L}(A_i)$ is instead generated conditioned on y_i under the mixture representation in (3.17). In particular, given $y_i = y$ we first choose a mixture component by sampling the latent class indicator $G_i \in \{1, \ldots, H\}$ from $p_{G|y}$ with $p_{G|y}(h) = \text{pr}(G_i = h \mid \mathcal{Y} = y) = \nu_{hy}$. Then, given $G_i = h$ and the corresponding edge probability vector $\pi^{(h)}$ – factorized as in (3.6) – the network $\mathcal{L}(A_i)$ is generated by sampling its edges $\mathcal{L}(A_i)_l$, $l = 1, \ldots, V(V-1)/2$ from conditionally independent Bernoulli variables. Hence, the dependence on the groups is introduced in the latent class assignment mechanism via group-specific mixing probabilities, so that brain networks in the same class h share a common edge probability vector $\pi^{(h)}$, with the probability assigned to each class changing across the two groups. This simple generative mechanism is appealing in facilitating tractable posterior computation and inference.

A key in the previous representation is that it allows substantial dimensionality reduction, while preserving flexibility. As stated in Corollary 3.6 – generalizing Lemma 3.1 – such a representation is sufficiently flexible to jointly characterize the collection of group-dependent pmfs $p_{\mathcal{L}(\mathcal{A})|1}$, $p_{\mathcal{L}(\mathcal{A})|2}$.

Corollary 3.6. *Any collection of group-dependent probability mass functions* $p_{\mathcal{L}(\mathcal{A})|y} \in \mathcal{P}_{|\mathbb{A}_V|}$, $y \in \{1,2\}$ *can be characterized as in* (3.17) *for some* H *with class-specific edge probability vectors* $\pi^{(h)}$, $h = 1, \ldots, H$ *factorized as in* (3.6) *for some* R.

Proof. According to Lemma 3.1 we can represent each $p_{\mathcal{L}(\mathcal{A})|y}(a)$, $y \in \{1,2\}$ as $p_{\mathcal{L}(\mathcal{A})|y}(a) = \sum_{h=1}^{H_y} \nu_{hy}^* \prod_{l=1}^{V(V-1)/2} \{\pi_l^{(hy)}\}^{a_l} \{1-\pi_l^{(hy)}\}^{1-a_l}$, with each $\pi_l^{(hy)}$ factorized as $\text{logit}\{\pi_l^{(hy)}\} = Z_l^{(y)} + \sum_{r=1}^{R_y} \lambda_r^{(hy)} X_{vr}^{(hy)} X_{ur}^{(hy)}$, $l = 1, \ldots, V(V-1)/2$ and $h = 1, \ldots, H_y$. Hence Corollary 3.6 follows after choosing $\pi^{(h)}, h = 1, \ldots, H$ as the sequence of unique class-specific edge probability vectors $\pi^{(hy)}$ appearing in the previous factorization for at least one group y, and letting the group-specific mixing weights in (3.17) be $\nu_{hy} = \nu_{hy}^*$ if $\pi^{(h)} = \pi^{(hy)}$ and $\nu_{hy} = 0$ otherwise. □

This additionally ensures that any joint probability mass function $p_{\mathcal{Y},\mathcal{L}(\mathcal{A})}$ for the random variable $\{\mathcal{Y}, \mathcal{L}(\mathcal{A})\}$ admits representation (3.15) and (3.17) with class-specific edge probability vectors from (3.6) and hence our formulation can be viewed as fully general and robust against model misspecification in testing (3.16), given sufficiently flexible priors for the components.

We could have considered more complicated scenarios with group dependence introduced also in the quantities characterizing the mixture components in (3.6). However, including group dependence only in the mixing probabilities favors borrowing of information across the groups in modeling $\pi^{(h)}$, $h = 1, \ldots, H$, while massively reducing the number of parameters to test in (3.16) from $2\{2^{V(V-1)/2} - 1\}$ to $2(H-1)$. Specifically, the characterization of $p_{\mathcal{L}(\mathcal{A})|y}$ in (3.17) further simplifies the system (3.16) to only testing the equality of the group-specific mixing probability vectors

$$H_0 : (\nu_{11}, \ldots, \nu_{H1}) = (\nu_{12}, \ldots, \nu_{H2}) \text{ versus } H_1 : (\nu_{11}, \ldots, \nu_{H1}) \neq (\nu_{12}, \ldots, \nu_{H2}). \quad (3.18)$$

Recalling Corollary 3.6, under our formulation, the system (3.18) uniquely characterizes $H_0 : p_{\mathcal{Y},\mathcal{L}(\mathcal{A})} = p_{\mathcal{Y}} p_{\mathcal{L}(\mathcal{A})}$ versus $H_1 : p_{\mathcal{Y},\mathcal{L}(\mathcal{A})} \neq p_{\mathcal{Y}} p_{\mathcal{L}(\mathcal{A})}$.

In developing methodologies for the multiple local tests in (3.14) under our model formulation, we measure the association between $\mathcal{L}(\mathcal{A})_l$ and $\mathcal{Y}$ exploiting the model-based version

of Cramer's V, proposed in Dunson and Xing (2009), obtaining

$$\begin{aligned}\rho_l^2 &= \frac{1}{\min\{2,2\}-1}\sum_{y=1}^{2}\sum_{a_l=0}^{1}\frac{\left\{p_{\mathcal{Y},\mathcal{L}(\mathcal{A})_l}(y,a_l)-p_{\mathcal{Y}}(y)p_{\mathcal{L}(\mathcal{A})_l}(a_l)\right\}^2}{p_{\mathcal{Y}}(y)p_{\mathcal{L}(\mathcal{A})_l}(a_l)}\\ &= \sum_{y=1}^{2}\sum_{a_l=0}^{1}\frac{\left\{p_{\mathcal{Y}}(y)p_{\mathcal{L}(\mathcal{A})_l|y}(a_l)-p_{\mathcal{Y}}(y)p_{\mathcal{L}(\mathcal{A})_l}(a_l)\right\}^2}{p_{\mathcal{Y}}(y)p_{\mathcal{L}(\mathcal{A})_l}(a_l)}\\ &= \sum_{y=1}^{2}p_{\mathcal{Y}}(y)\sum_{a_l=0}^{1}\frac{\left\{p_{\mathcal{L}(\mathcal{A})_l|y}(a_l)-p_{\mathcal{L}(\mathcal{A})_l}(a_l)\right\}^2}{p_{\mathcal{L}(\mathcal{A})_l}(a_l)}. \qquad (3.19)\end{aligned}$$

Measuring the local association with $\rho_l \in (0,1)$ provides an appealing choice in terms of interpretation, with $\rho_l = 0$ meaning that $p_{\mathcal{Y},\mathcal{L}(\mathcal{A})_l} = p_{\mathcal{Y}}p_{\mathcal{L}(\mathcal{A})_l}$, and hence the random variable $\mathcal{L}(\mathcal{A})_l$ modeling the presence or absence of an edge among the lth pair of nodes, has no differences across groups. Beside incorporating a fully general and tractable global test, our model formulation is particularly appealing also in addressing issues associated to local multiple testing in the network framework. First, as stated in Proposition 3.7, each ρ_l, $l = 1,\dots,V(V-1)/2$, can be easily computed from the quantities in our model.

Proposition 3.7. *Based on equations* (3.15) *and* (3.17) *,* $p_{\mathcal{L}(\mathcal{A})_l|y}(1) = 1-p_{\mathcal{L}(\mathcal{A})_l|y}(0) = \sum_{h=1}^{H}\nu_{hy}\pi_l^{(h)}$, *and* $p_{\mathcal{L}(\mathcal{A})_l}(1) = 1-p_{\mathcal{L}(\mathcal{A})_l}(0) = \sum_{y=1}^{2}p_{\mathcal{Y}}(y)\sum_{h=1}^{H}\nu_{hy}\pi_l^{(h)}$.

Proof. The steps recall those considered to prove Proposition 3.2. In particular, recalling factorization (3.17) and letting $\mathbb{A}_V^{-l}$ the set containing all the possible network configurations for the node pairs except the lth one, we have that $p_{\mathcal{L}(\mathcal{A})_l|y}(1)$ is equal to

$$\sum_{\mathbb{A}_V^{-l}}\sum_{h=1}^{H}\nu_{hy}\pi_l^{(h)}\prod_{l^*\neq l}\{\pi_{l^*}^{(h)}\}^{a_{l^*}}\{1-\pi_{l^*}^{(h)}\}^{1-a_{l^*}} = \sum_{h=1}^{H}\nu_{hy}\pi_l^{(h)}\sum_{\mathbb{A}_V^{-l}}\prod_{l^*\neq l}\{\pi_{l^*}^{(h)}\}^{a_{l^*}}\{1-\pi_{l^*}^{(h)}\}^{1-a_{l^*}}$$

Then Proposition 3.7 follows after noticing that $\prod_{l^*\neq l}\{\pi_{l^*}^{(h)}\}^{a_{l^*}}\{1-\pi_{l^*}^{(h)}\}^{1-a_{l^*}}$ is the joint pmf of independent Bernoulli random variables and hence the summation over the whole joint sample space $\mathbb{A}_V^{-l} = \{0,1\}^{V(V-1)/2-1}$, provides $\sum_{\mathbb{A}_V^{-l}}\prod_{l^*\neq l}\{\pi_{l^*}^{(h)}\}^{a_{l^*}}\{1-\pi_{l^*}^{(h)}\}^{1-a_{l^*}} = 1$. The proof for the marginal $p_{\mathcal{L}(\mathcal{A})_l}(1) = \sum_{y=1}^{2}p_{\mathcal{Y}}(y)\sum_{h=1}^{H}\nu_{hy}\pi_l^{(h)}$ follows directly from previous result after noticing that $p_{\mathcal{L}(\mathcal{A})_l}(1) = \sum_{y=1}^{2}p_{\mathcal{Y},\mathcal{L}(\mathcal{A})_l}(y,1) = \sum_{y=1}^{2}p_{\mathcal{Y}}(y)p_{\mathcal{L}(\mathcal{A})_l|y}(1)$. □

Joining results from Propositions 3.2 and 3.7 note that the vector $\bar{\pi}_y \in (0,1)^{V(V-1)/2}$, containing the group-specific edge probabilities $\bar{\pi}_{yl} = \text{pr}\{\mathcal{L}(\mathcal{A})_l = 1 \mid \mathcal{Y} = y\} = p_{\mathcal{L}(\mathcal{A})_l|y}(1) = \sum_{h=1}^{H}\nu_{hy}\pi_l^{(h)}$, coincides with the conditional expectation of the network-valued random variable $\text{E}\{\mathcal{L}(\mathcal{A}) \mid \mathcal{Y} = y\} = \sum_{a\in\mathbb{A}_V}ap_{\mathcal{L}(\mathcal{A})|y}(a) = \sum_{h=1}^{H}\nu_{hy}\pi^{(h)}$ given the group membership $y \in \{1,2\}$. These quantities are of key interest for inference in providing a summarized overview on how the network structure changes on average across groups.

A second key benefit for local testing provided by our model formulation, is that the shared dependence on a common set of node-specific latent coordinates characterizing the construction of the edge probability vector $\pi^{(h)}$ within each class $h = 1, \ldots, H$ in (3.6), explicitly accounts for specific dependence structures in brain connections. According to Hoff (2008), factorization (3.6) can accurately accommodate key topological properties including block structures, homophily behaviors and transitive edge patterns – among others. As a results – in line with Scott et al. (2014) – informing our local testing procedures about these structures, is expected to substantially improve power, compared to standard FDR control procedures.

3.1.8 Prior specification and posterior computation

We specify independent priors for the quantities $p_{\mathcal{Y}} \sim \Pi_y$, $Z = (Z_1, \ldots, Z_{V(V-1)/2})^{\mathrm{T}} \sim \Pi_Z$, $X^{(h)} = (X_1^{(h)}, \ldots, X_V^{(h)})^{\mathrm{T}} \sim \Pi_X$, $\lambda^{(h)} = (\lambda_1^{(h)}, \ldots, \lambda_R^{(h)})^{\mathrm{T}} \sim \Pi_\lambda$, $h = 1, \ldots, H$ and $\nu_y = (\nu_{1y}, \ldots, \nu_{Hy}) \sim \Pi_\nu$, $y \in \{1, 2\}$, to induce a prior Π on the joint pmf $p_{\mathcal{Y},\mathcal{L}(\mathcal{A})}$ with full support over the $2\times|\mathbb{A}_V|$ dimensional simplex $\mathcal{P}_{2\times|\mathbb{A}_V|}$, while obtaining desirable asymptotic behavior, simple posterior computation and allowance for testing. Prior support is a key to retain the flexibility associated to our statistical model and testing procedures, when performing posterior inference under a Bayesian paradigm.

As $p_{\mathcal{Y}}$ is the pmf of a categorical random variable on 2 levels, we let $1 - p_{\mathcal{Y}}(2) = p_{\mathcal{Y}}(1) \sim \text{Beta}(a, b)$, and consider the same prior specification discussed in Section 3.1.4 for the quantities in (3.6) by choosing Gaussian priors (3.10) for the entries in Z, standard Gaussians (3.11) for the elements in $X^{(h)}$ and multiplicative inverse gammas (3.9) for $\lambda^{(h)}$, $h = 1, \ldots, H$. A key of our prior specification is incorporation of global testing (3.18) in the definition of Π_ν. Specifically letting $\upsilon = (\upsilon_1, \ldots, \upsilon_H)$ and $\upsilon_y = (\upsilon_{1y}, \ldots, \upsilon_{Hy})$, we induce Π_ν through

$$\begin{aligned} \nu_y &= (1-T)\upsilon + T\upsilon_y, \quad y \in \{1,2\}, \\ \upsilon &\sim \text{Dir}(a_1, \ldots, a_H), \quad \upsilon_y \sim \text{Dir}(a_1, \ldots, a_H),\ y \in \{1,2\}, \\ T &\sim \text{Bern}\{\text{pr}(H_1)\}. \end{aligned} \tag{3.20}$$

In (3.20), T is a hypothesis indicator, with $T = 0$ for H_0 and $T = 1$ for H_1. Under H_1, we generate group-specific mixing weights independently, while under H_0 we have equal weight vectors. By choosing small values for the hyperparameters in the Dirichlet priors, we additionally favor automatic deletion of redundant components (Rousseau and Mengersen, 2011). In assessing evidence in favor of the alternative, we can rely on the posterior probability, $\text{pr}[H_1 \mid \{y, \mathcal{L}(A)\}] = 1 - \text{pr}[H_0 \mid \{y, \mathcal{L}(A)\}]$ which can be easily obtained from the output of the Gibbs sampler proposed below. Specifically, under prior (3.20) and exploiting the hierarchical structure of our dependent mixture model – summarized in Figure 3.6 – the

full conditional $\text{pr}(T = 1 \mid -) = \text{pr}(H_1 \mid -) = 1 - \text{pr}(H_0 \mid -)$ is simply

$$
\begin{aligned}
&= \frac{\text{pr}(H_1) \prod_{y=1}^{2} \int \{\prod_{i:y_i=y} \text{pr}(G_i \mid \upsilon_y, y_i)\} d\Pi_{\upsilon_y}}{\text{pr}(H_0) \int \{\prod_{i=1}^{n} \text{pr}(G_i \mid \upsilon)\} d\Pi_{\upsilon} + \text{pr}(H_1) \prod_{y=1}^{2} \int \{\prod_{i:y_i=y} \text{pr}(G_i \mid \upsilon_y, y_i)\} d\Pi_{\upsilon_y}}, \\
&= \frac{\text{pr}(H_1) \prod_{y=1}^{2} \int (\prod_{h=1}^{H} \upsilon_{hy}^{n_{hy}}) d\Pi_{\upsilon_y}}{\text{pr}(H_0) \int (\prod_{h=1}^{H} \upsilon_{h}^{n_h}) d\Pi_{\upsilon} + \text{pr}(H_1) \prod_{y=1}^{2} \int (\prod_{h=1}^{H} \upsilon_{hy}^{n_{hy}}) d\Pi_{\upsilon_y}}, \\
&= \frac{\text{pr}(H_1) \prod_{y=1}^{2} \text{B}(a + \bar{n}_y)/\text{B}(a)}{\text{pr}(H_0)\text{B}(a + \bar{n})/\text{B}(a) + \text{pr}(H_1) \prod_{y=1}^{2} \text{B}(a + \bar{n}_y)/\text{B}(a)},
\end{aligned}
\tag{3.21}
$$

with $n_h = \sum_{i=1}^{n} \text{I}(G_i = h)$, $n_{hy} = \sum_{i:y_i=y} \text{I}(G_i = h)$, $a = (a_1, \ldots, a_H)$, $\bar{n} = (n_1, \ldots, n_H)$, $\bar{n}_y = (n_{1y}, \ldots, n_{Hy})$ and B indicates the multivariate beta function $\text{B}(x) = \prod_{i=1}^{q} \Gamma(x_i)/\Gamma(\sum_{i=1}^{q} x_i)$ with $\Gamma(x_i)$ the gamma function. It is straightforward to derive the equalities $\int (\prod_{h=1}^{H} \upsilon_h^{n_h}) d\Pi_{\upsilon} = \text{B}(a + \bar{n})/\text{B}(a)$ and $\int (\prod_{h=1}^{H} \upsilon_{hy}^{n_{hy}}) d\Pi_{\upsilon_y} = \text{B}(a + \bar{n}_y)/\text{B}(a)$, $y \in \{1, 2\}$ using known results of the Dirichlet-multinomial conjugacy.

Although providing a key choice for performing global testing, it is impractical to adopt formulation (3.20) for each local point null $H_{0l} : \rho_l = 0$ versus $H_{1l} : \rho_l \neq 0$, $l = 1, \ldots, V(V-1)/2$. Hence, we replace local point nulls with small interval nulls $H_{0l} : \rho_l \leq \epsilon$ versus $H_{1l} : \rho_l > \epsilon$. This choice allows $\text{pr}[H_{1l} \mid \{y, \mathcal{L}(A)\}] = 1 - \text{pr}[H_{0l} \mid \{y, \mathcal{L}(A)\}]$ to be easily estimated as the proportion of Gibbs samples in which $\rho_l > \epsilon$. Moreover – as noted in Berger and Sellke (1987) and Berger and Delampady (1987) – testing the small interval hypothesis $H_{0l} : \rho_l \leq \epsilon$ is in general more realistic and provides – under a Bayesian paradigm – essentially the same results than those obtained when assessing evidence of $H_{0l} : \rho_l = 0$.

Beside key computational properties, as stated in Corollary 3.8 – generalizing theoretical results in Section 3.1.4 on full prior support – our choices induce a prior Π for $p_{\mathcal{Y},\mathcal{L}(\mathcal{A})}$ with full L_1 support over $\mathcal{P}_{2\times|\mathbb{A}_V|}$, meaning that Π can generate a $p_{\mathcal{Y},\mathcal{L}(\mathcal{A})}$ within an arbitrarily small L_1 neighborhood of the true data-generating model $p^0_{\mathcal{Y},\mathcal{L}(\mathcal{A})}$, allowing the truth to fall in a wide class.

Corollary 3.8. *Based on the priors $\Pi_y, \Pi_Z, \Pi_X, \Pi_\lambda$, and Π_ν, and letting $B_\epsilon(p^0_{\mathcal{Y},\mathcal{L}(\mathcal{A})}) = \{p_{\mathcal{Y},\mathcal{L}(\mathcal{A})} : \sum_{y=1}^{2} \sum_{a\in\mathbb{A}_V} |p_{\mathcal{Y},\mathcal{L}(\mathcal{A})}(y, a) - p^0_{\mathcal{Y},\mathcal{L}(\mathcal{A})}(y, a)| < \epsilon\}$ denote the L_1 neighborhood around $p^0_{\mathcal{Y},\mathcal{L}(\mathcal{A})}$, then for any $p^0_{\mathcal{Y},\mathcal{L}(\mathcal{A})} \in \mathcal{P}_{2\times|\mathbb{A}_V|}$ and $\epsilon > 0$, $\Pi\{B_\epsilon(p^0_{\mathcal{Y},\mathcal{L}(\mathcal{A})})\} > 0$.*

Proof. Recalling Corollary 3.6 and factorization (3.15) we can always represent the L_1 distance $\sum_{y=1}^{2} \sum_{a\in\mathbb{A}_V} |p_{\mathcal{Y},\mathcal{L}(\mathcal{A})}(y, a) - p^0_{\mathcal{Y},\mathcal{L}(\mathcal{A})}(y, a)|$ between $p_{\mathcal{Y},\mathcal{L}(\mathcal{A})}$ and $p^0_{\mathcal{Y},\mathcal{L}(\mathcal{A})}$ as

$$
\sum_{y=1}^{2} \sum_{a\in\mathbb{A}_V} |p_{\mathcal{Y}}(y) \sum_{h=1}^{H} \nu_{hy} \prod_{l=1}^{V(V-1)/2} \{\pi_l^{(h)}\}^{a_l} \{1 - \pi_l^{(h)}\}^{1-a_l} - p^0_{\mathcal{Y}}(y) \sum_{h=1}^{H} \nu^0_{hy} \prod_{l=1}^{V(V-1)/2} \{\pi_l^{0(h)}\}^{a_l} \{1 - \pi_l^{0(h)}\}^{1-a_l}|,
$$

with $\nu^0_{hy} = \nu^{*0}_{hy}$ if $\pi^{0(h)} = \pi^{0(hy)}$ and $\nu^0_{hy} = 0$ otherwise. Hence $\Pi\{B_\epsilon(p^0_{\mathcal{Y},\mathcal{L}(\mathcal{A})})\}$ is

$$\int 1(\sum_{y=1}^{2} \sum_{a \in \mathbb{A}_V} |p_{\mathcal{Y},\mathcal{L}(\mathcal{A})}(y,a) - p^0_{\mathcal{Y},\mathcal{L}(\mathcal{A})}(y,a)| < \epsilon) d\Pi_y(p_{\mathcal{Y}}) d\Pi_\nu(\nu_1,\nu_2) d\Pi_\pi(\pi^{(1)},\dots,\pi^{(H)}).$$

Recalling results in Dunson and Xing (2009) a sufficient condition for the previous integral to be strictly positive is that $\Pi_y\{p_y : \sum_{y=1}^2 |p_{\mathcal{Y}}(y) - p^0_{\mathcal{Y}}(y)| < \epsilon_y\} > 0$, $\Pi_\pi\{\pi^{(h)}, h = 1,\dots,H : \sum_{h=1}^H \sum_{l=1}^{V(V-1)/2} |\pi^{(h)}_l - \pi^{0(h)}_l| < \epsilon_\pi\} > 0$ and $\Pi_\nu\{\nu_y, y \in \mathbb{Y} : \sum_{y=1}^2 \sum_{h=1}^H |\nu_{hy} - \nu^0_{hy}| < \epsilon_\nu\} > 0$ for every $\epsilon_\pi > 0$, $\epsilon_y > 0$ and $\epsilon_\nu > 0$. The large support for $p_{\mathcal{Y}}$ is directly guaranteed from the beta prior. Similarly, according to Theorem 3.3 and Lemma 3.4 the same hold for the joint prior over the sequence of class-specific edge probability vectors $\pi^{(h)}, h = 1,\dots,H$ induced by priors Π_Z, Π_X and Π_λ in factorization (3.6). Finally marginalizing out the testing indicator T and recalling our prior specification for the mixing probabilities in (3.20) a lower bound for $\Pi_\nu\{\nu_y, y \in \mathbb{Y} : \sum_{y=1}^2 \sum_{h=1}^H |\nu_{hy} - \nu^0_{hy}| < \epsilon_\nu\}$ is

$$\text{pr}(H_0)\Pi_\upsilon\{\upsilon : \sum_{y=1}^{2}\sum_{h=1}^{H} |\upsilon_h - \nu^0_{hy}| < \epsilon_\nu\} + \text{pr}(H_1) \prod_{y=1}^{2} \Pi_{\upsilon_y}\{\upsilon_y : \sum_{h=1}^{H} |\upsilon_{hy} - \nu^0_{hy}| < \epsilon_\nu/2\}.$$

If the true model is generated under no association, previous equation reduces to

$$\text{pr}(H_0)\Pi_\upsilon\{\upsilon : \sum_{h=1}^{H} |\upsilon_h - \nu^0_{h}| < \epsilon_\nu/2\} + \text{pr}(H_1) \prod_{y=1}^{2} \Pi_{\upsilon_y}\{\upsilon_y : \sum_{h=1}^{H} |\upsilon_{hy} - \nu^0_{h}| < \epsilon_\nu/2\},$$

with the Dirichlet priors for υ and $\upsilon_y, y \in \{1,2\}$ ensuring the positivity of both terms. When instead $\nu^0_{h1} \neq \nu^0_{h2}$ for some $h = 1,\dots,H$, the positivity of $\text{pr}(H_0)\Pi_\upsilon\{\upsilon : \sum_{y=1}^2 \sum_{h=1}^H |\upsilon_h - \nu^0_{hy}| < \epsilon_\nu\}$ is not guaranteed, but $\text{pr}(H_1)\prod_{y=1}^2 \Pi_{\upsilon_y}\{\upsilon_y : \sum_{h=1}^H |\upsilon_{hy} - \nu^0_{hy}| < \epsilon_\nu/2\}$ remains positive for every ϵ_ν under the independent Dirichlet priors for the quantities υ_y, $y \in \{1,2\}$, proving the Corollary. □

Full prior support is a key property to ensure good performance in posterior inference and testing, because without prior support about the true data-generating pmf, the posterior cannot possibly concentrate around the truth. Moreover, as $p_{\mathcal{Y},\mathcal{L}(\mathcal{A})}$ is characterized by finitely many parameters $p_{\mathcal{Y},\mathcal{L}(\mathcal{A})}(y,a)$, $y \in \mathbb{Y}$, $a \in \mathbb{A}_V$, Corollary 3.8 is sufficient to guarantee that the posterior assigns probability one to any arbitrarily small neighborhood of the true joint pmf as $n \to \infty$, meaning that $\Pi[B_\epsilon(p^0_{\mathcal{Y},\mathcal{L}(\mathcal{A})}) \mid \{y_1,\mathcal{L}(A_1)\},\dots,\{y_n,\mathcal{L}(A_n)\}]$ converges almost surely to 1, when the true joint pmf is $p^0_{\mathcal{Y},\mathcal{L}(\mathcal{A})}$.

Posterior computation is easily available adapting the Gibbs sampler 3 for $p_{\mathcal{L}(\mathcal{A})}$ factorized as in (3.5)–(3.6) to the new statistical model characterizing $p_{\mathcal{Y},\mathcal{L}(\mathcal{A})}$ via (3.15) and (3.17) with $\pi^{(h)}$ as in (3.6). Algorithm 4 provides detailed steps for the proposed Gibbs sampler.

Algorithm 4 Gibbs sampler for the dependent mixture of low-rank factorizations model

[1] Allocate each network observation to one of the classes

for $i = 1, \ldots, n$ **do**

Sample the class indicator G_i from the discrete distribution with probabilities

$$\text{pr}(G_i = h \mid -) = \frac{\nu_{hy_i} \prod_{l=1}^{V(V-1)/2} \{\pi_l^{(h)}\}^{\mathcal{L}(A_i)_l} \{1 - \pi_l^{(h)}\}^{1-\mathcal{L}(A_i)_l}}{\sum_{m=1}^{H} \nu_{my_i} \prod_{l=1}^{V(V-1)/2} \{\pi_l^{(m)}\}^{\mathcal{L}(A_i)_l} \{1 - \pi_l^{(m)}\}^{1-\mathcal{L}(A_i)_l}},$$

for each $h = 1, \ldots, H$, with $\pi^{(h)}$ defined in (3.6)

end for

---

[2] Sample the testing indicator T from the full conditional Bernoulli with probability (3.21).

---

[3] Update the group-specific mixing probabilities

If $T = 0$, let $\nu_y = \upsilon$, $y \in \{1, 2\}$ with υ updated from the full conditional Dirichlet

$$(\upsilon_1, \ldots, \upsilon_H) \mid - \sim \text{Dirichlet}(1/H + n_1, \ldots, 1/H + n_H).$$

Otherwise, if $T = 1$, update ν_y from

$$(\nu_{1y}, \ldots, \nu_{Hy}) \mid - \sim \text{Dirichlet}(1/H + n_{1y}, \ldots, 1/H + n_{Hy})$$

independently for each $y \in \{1, 2\}$.

---

[4] Update quantities Z, $X^{(h)}$ **and** $\lambda^{(h)}$, $h = 1, \ldots, H$

Given G_i, $i = 1, \ldots, n$, the updating for quantities Z, $X^{(h)}$ and $\lambda^{(h)}$, $h = 1, \ldots, H$ proceeds as in Algorithm 3 via Polyá-gamma data augmentation. Specifically first sample Polyá-gamma augmented data as in step **[3]** of Gibbs sampler 3 and then update Z, $X^{(h)}$ and $\lambda^{(h)}$, $h = 1, \ldots, H$ following steps **[4]**, **[5]** and **[6]**, respectively, of Algorithm 3.

---

[5] Update the marginal group probabilities $p_{\mathcal{Y}}(1) = 1 - p_{\mathcal{Y}}(2)$ from

$$p_{\mathcal{Y}}(1) \mid - \sim \text{Beta}(a + n_1, b + n_2),$$

with $n_y = \sum_{i=1}^{n} \text{I}(y_i = y)$.

Since the number of mixing components in (3.17) and the dimensions of the latent spaces in (3.6) are not known in practice, we perform posterior computation by fixing H and R at conservative upper bounds. The priors are chosen to allow adaptive emptying of the redundant

components, with the posteriors for parameters controlling unnecessary dimensions concentrated near zero. If all the classes h are occupied, then H should be increased. Similarly, if the posterior for $\lambda_R^{(h)}$ is not concentrated near zero for any h, then R should be increased.

3.1.9 Simulation study

We consider simulation studies to evaluate the performance of our method in accurately estimating the joint pmf for the pair $\{\mathcal{Y}, \mathcal{L}(\mathcal{A})\}$, in correctly assessing the global hypothesis of association among the network-valued random variable and the categorical predictor, and in identifying local variations in each edge probability across groups.

For comparison we also implement a MANOVA procedure – see e.g. Krzanowski (1988) – to test for global variations across groups of the random vector Θ of summary measures, with realization θ_i of Θ comprising the most commonly used network summary statistics – i.e. network density, transitivity, average path length and assortativity – computed for each network i. Refer to Kantarci and Labatut (2013) for an overview on these topological network measures and Bullmore and Sporns (2009), Rubinov and Sporns (2010), Bullmore and Sporns (2012) for a discussion on their importance in characterizing wiring mechanisms within brain networks. For local testing, we compare our procedure to the results obtained when testing on the association between $\mathcal{L}(\mathcal{A})_l$ and $\mathcal{Y}$ for each $l = 1, \ldots, V(V-1)/2$ via separate two-sided Fisher's exact tests – see e.g. Agresti (2002). We consider exact tests to avoid issues arising from the χ^2 approximations in sparse tables.

We simulate $n = 50$ pairs (y_i, A_i) from our model in (3.15) and (3.17), with y_i from a categorical variable having two equally likely groups $p_{\mathcal{Y}}^0 = (0.5, 0.5)$ and A_i a $V \times V$ network with $V = 20$ nodes. We consider $H = 2$ latent classes, with $\pi^{0(h)}$ defined as in (3.6). Brain networks are typically characterized by tighter intra-hemispheric than inter-hemispheric connections (Roncal et al., 2013). Hence, we consider two node blocks $\{1, \ldots, 10\}$ and $\{11, \ldots, 20\}$ characterizing left and right hemisphere, respectively, and generate entries in Z^0 to favor more likely connections between pairs in the same block than pairs in different blocks. To assess the local testing performance, we induce group differences only on a subset of nodes $V^* \subset V$. A possibility to favor this behavior is to consider $R = 1$, $\lambda^{0(1)} = \lambda^{0(2)} = 1$ and let $X_v^{0(h)} \neq 0$ only for nodes $v \in V^*$, while fixing the latent coordinates of the remaining nodes to 0. As a result, no variations in edge probabilities are displayed when mixing probabilities remain constant, while only local differences are highlighted when mixing probabilities shift across groups. Under the dependence scenario, data are simulated with group-specific mixing probabilities $\nu_1^0 = (0.8, 0.2)$, $\nu_2^0 = (0.2, 0.8)$. Instead, constant mixing probabilities $\nu_1^0 = \nu_2^0 = (0.5, 0.5)$ are considered under independence. Even if we focus only on 20 nodes to facilitate graphical analyses, our dependent mixture of low-rank factorizations scales to much higher V settings. Refer to discussion in Section 3.1.6.

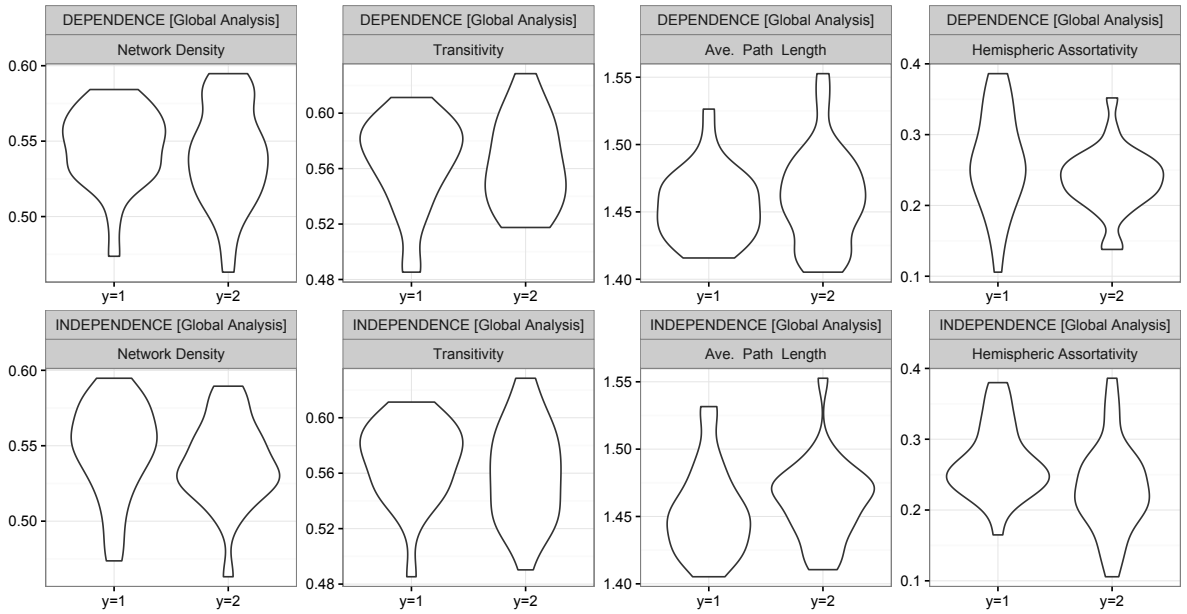


FIGURE 3.7: For the two scenarios, observed changes across the two groups of selected network summary statistics. These measures are computed for each simulated network under the two scenarios and summarized via violin plots.

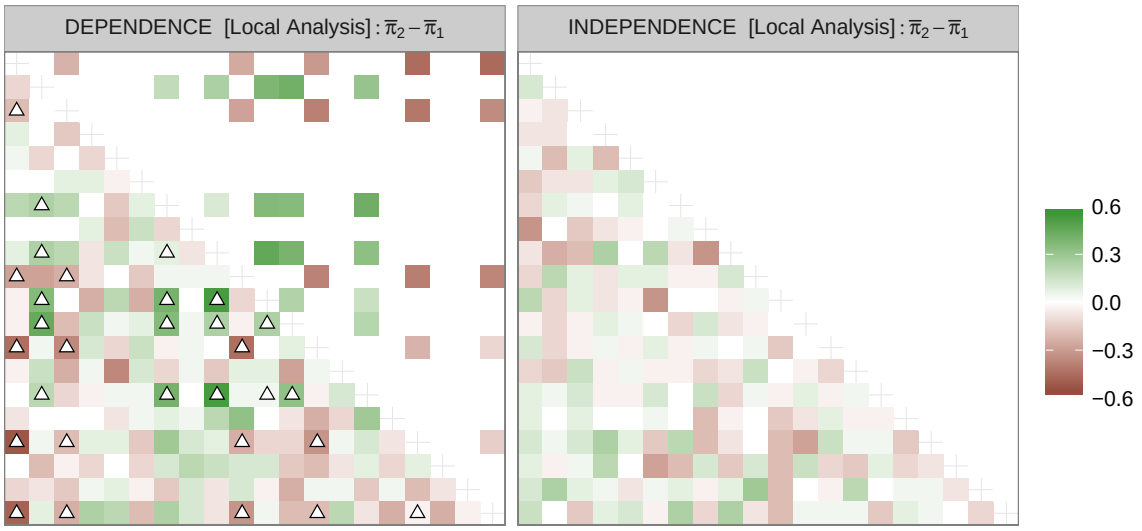


FIGURE 3.8: Lower triangular: group difference between the relative edge frequencies for each pair of nodes computed from the simulated data. Upper triangular: true group difference in edge probabilities arising from the generative processes considered in the simulations. Previous quantities are displayed for the dependence (left) and independence (right) simulation scenarios. Triangles highlight edges which truly differ across groups in the dependence simulation scenario.

As show in Figures 3.7–3.8, although our dependence simulation setting may appear – at first – simple, it provides a challenging scenario for procedures assessing evidence of global association by testing on variations in the network summary measures. In fact, we choose values $X_v^{0(h)}$ for the nodes $v \in V^*$ such that the resulting summary statistics for the simulated networks do no display changes across groups also in the dependence scenario. Hence a

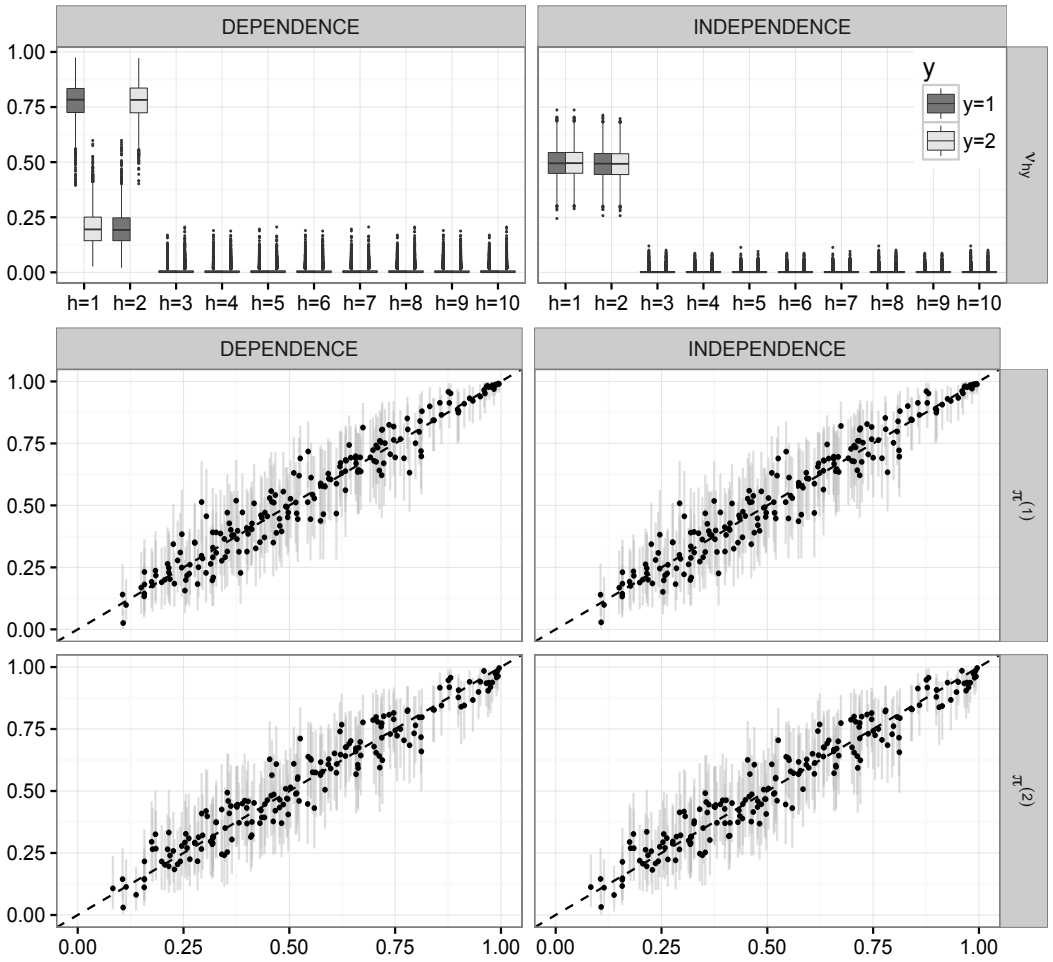


FIGURE 3.9: Upper panels: boxplots based on the posterior samples of the group-specific mixing probabilities under the dependence (left) and independence (right) scenarios. Lower panels: under the same scenarios and for the non-empty classes $h = 1$ and $h = 2$, plots of the true class-specific edge probabilities $\pi_l^{0(h)}$ (x-axis) versus their posterior mean $\hat{\pi}_l^{(h)}$ (y-axis), $l = 1, \ldots, V(V-1)/2$. Segments denote the corresponding 0.95 highest posterior density intervals.

global test relying on network summary measures is expected to fail in detecting association between $\mathcal{Y}$ and $\mathcal{L}(\mathcal{A})$, as variations in the network pmf are only local – i.e. in a subset of its marginals $\mathcal{L}(\mathcal{A})_l$. On the other hand, powerful local testing procedures are required to efficiently detect this small set of edge probabilities truly changing across the two groups.

In both scenarios, inference is accomplished by considering $H = R = 10$, $\text{pr}(H_1) = \text{pr}(H_0) = 0.5$ and letting $p_{\mathcal{Y}}(1) \sim \text{Beta}(1/2, 1/2)$. To favor deletion of unnecessary classes h, we fix the hyperparameter vector in the Dirichlet for υ and υ_y to $a = (a_1 = 1/H, \ldots, a_H = 1/H)$. As noted in Ishwaran and Zarepour (2002), this choice provides also a finite approximation to the Dirichlet process. For priors Π_Z, Π_X and Π_λ, we choose the same hyperparameter settings of the simulation study in Section 3.1.6. We collect $5{,}000$ Gibbs iterations, discarding the first $1{,}000$. In both scenarios converge and mixing are assessed via Gelman and Rubin (1992) potential scale reduction factors (PSRF) and effective sample sizes, respectively. Previous quantities are computed for the parameters of interest for inference in the

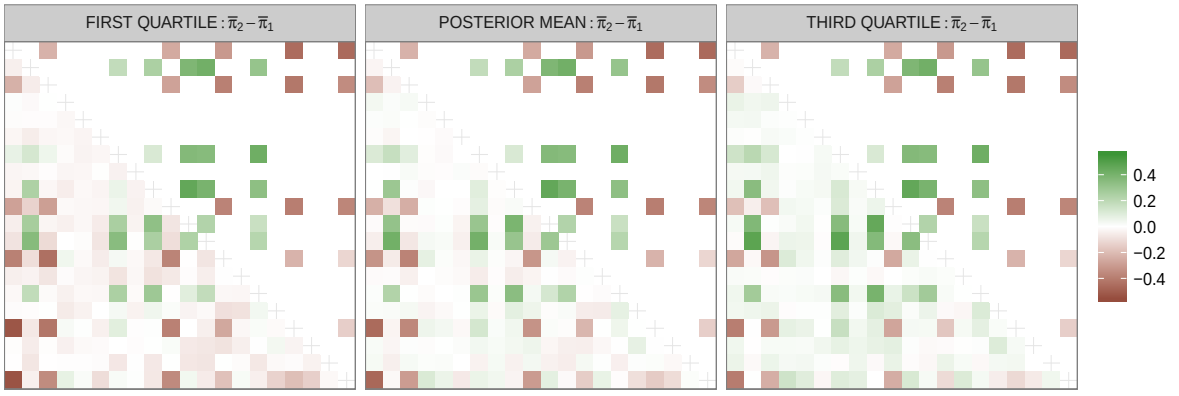


FIGURE 3.10: Lower triangular: for the dependence simulation scenario, mean and quartiles for the posterior distribution of the difference between the edge probabilities in the second group $\bar{\pi}_2$ and first group $\bar{\pi}_1$. Upper triangular: for the same scenario, true difference $\bar{\pi}_2^0 - \bar{\pi}_1^0$.

simulation, covering the group-specific mixing probabilities ν_{hy}, $h = 1, \ldots, H$, $y \in \{1, 2\}$, the class-specific edge probabilities comprising vectors $\pi^{(h)}$, the Cramer's V coefficients ρ_l, $l = 1, \ldots, V(V-1)/2$ for local testing and the group-specific edge probability vectors $\bar{\pi}_y$, with elements $\bar{\pi}_{yl} = p_{\mathcal{L}(\mathcal{A})_l|y}(1) = \text{pr}\{\mathcal{L}(\mathcal{A})_l = 1 \mid \mathcal{Y} = y\}$ defined in Proposition 3.7. As discussed in Section 3.1.7 this vector coincides with the group-specific mean network structure $\text{E}\{\mathcal{L}(\mathcal{A}) \mid \mathcal{Y} = y\} = \sum_{h=1}^{H} \nu_{hy}\pi^{(h)}$. In both scenarios, most of the effective samples sizes are around $2{,}000$ out of $4{,}000$ samples, demonstrating excellent mixing performance. Similarly, all the PSRFs are less than 1.1, providing evidence that convergence has been reached.

Our testing procedure allows accurate inference on the global association between $\mathcal{L}(\mathcal{A})$ and $\mathcal{Y}$. We obtain $\hat{\text{pr}}[H_1 \mid \{y, \mathcal{L}(A)\}] > 0.99$ for the association scenario and $\hat{\text{pr}}[H_1 \mid \{y, \mathcal{L}(A)\}] < 0.01$ when y_i and A_i, $i = 1, \ldots, n$ are generated independently. Instead, the MANOVA testing procedure on the summary statistics vector fails to reject the null hypothesis of no association in both scenarios at a level $\alpha = 0.1$ – as expected. This result further highlights how global network measures may fail in accurately characterizing the whole network architecture. We obtain similarly good performance in correctly recovering the true pmf for $\{\mathcal{Y}, \mathcal{L}(\mathcal{A})\}$ under both scenarios. This is highlighted in Figure 3.9. As the posteriors for $p_{\mathcal{L}(\mathcal{A})|1}$ and $p_{\mathcal{L}(\mathcal{A})|2}$ are complex objects to visualize, we evaluate inference performance in Figure 3.9 by focusing on posteriors for quantities ν_y, $y \in \{1, 2\}$ and $\pi^{(h)}$, $h = 1, \ldots, H$, which characterize $p_{\mathcal{L}(\mathcal{A})|1}$ and $p_{\mathcal{L}(\mathcal{A})|2}$ under equation (3.17). As Figures 3.9 provides inference on class-specific quantities, we additionally accounted for label switching via the Stephens (2000) relabeling algorithm. However, no relabeling was necessary in our simulations.

As expected we learn posterior distributions for the mixing probabilities which shift over the grouping variable or remain constant under dependence and independence, respectively, as shown in the upper panel of Figure 3.9. Note also how the sparse Dirichlet priors for quantities u and u_y allow us to efficiently remove redundant dimensions. Borrowing of information across the groups provides accurate estimates of the class-specific edge probability

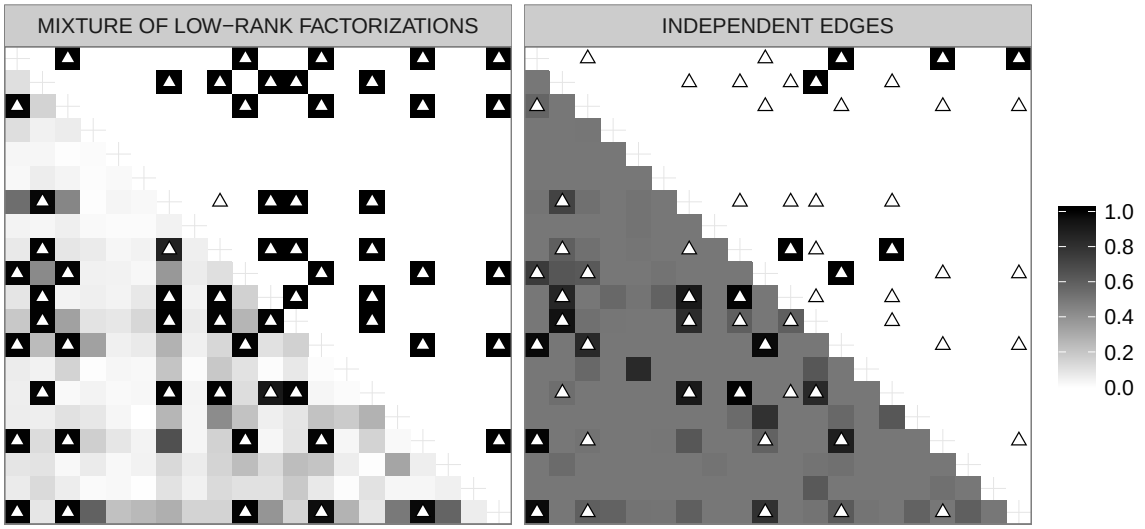


FIGURE 3.11: Lower triangular: $\hat{\text{pr}}[H_{1l} \mid \{y, \mathcal{L}(A)\}] = \text{pr}[\rho_l > 0.1 \mid \{y, \mathcal{L}(A)\}]$ (left) and calibrated Fisher's exact tests p-values $1/(1 - ep_l \log p_l)$ if $p_l < 1/e$, 0.5 otherwise (right), to allow comparison with $\hat{\text{pr}}[H_{1l} \mid \{y, \mathcal{L}(A)\}]$. Upper triangular: Accepted (white) and rejected (black) local null hypotheses. Triangles highlight edges which truly differ across groups.

vectors $\pi^{(h)}$, with posterior distribution concentrated around the true values, as confirmed in the lower panels of Figure 3.9. We obtain similar performance in estimating $p_{\mathcal{Y}}$, with the posterior concentrated around the true $p^0_{\mathcal{Y}}$. These results ensure accurate inference and estimation for the true joint pmfs $p^0_{\mathcal{Y},\mathcal{L}(\mathcal{A})}$ underlying simulated data in the dependence and independence scenarios.

Focusing on the dependence scenario, Figure 3.10 shows how accounting for sparsity and network information – via our dependent mixture of low-rank factorizations – provides accurate inference on local variations in edge probabilities, correctly highlighting pairs of nodes whose connectivity differs across groups in the true generating process and explicitly characterizing uncertainty through the posterior distribution. Conducting inference on each pair of nodes separately provides instead poor estimates – refer to left plot in Figure 3.8 – with the sub-optimality arising from inefficient borrowing of information across the edges. This lack of efficiency strongly affects also the local testing performance as shown in Figure 3.11, with our procedure having higher power than the one obtained via separate Fisher's exact tests. In Figure 3.11, each Fisher's exact test p-value is calibrated via $1/(1 - ep_l \log p_l)$ if $p_l < 1/e$ and 0.5 otherwise, to allow better comparison with $\hat{\text{pr}}[H_{1l} \mid \{y, \mathcal{L}(A)\}]$ (Sellke et al., 2001). Moreover, we adjust for multiplicity in the Fisher's exact tests by rejecting all local nulls having a p-value below p^*, with p^* the Benjamini and Hochberg (1995) threshold to maintain a false discovery rate FDR ≤ 0.1. Under our local Bayesian testing procedure we reject all H_{0l} such that $\hat{\text{pr}}[H_{1l} \mid \{y, \mathcal{L}(A)\}] > 0.9$, with $\epsilon = 0.1$. We do not explicitly control for FDR in order to assess whether our Bayesian procedures and the borrowing of information across local tests induced by factorization (3.6) contain the intrinsic adjustment for multiple testing, we expect. Results in Figure 3.11 confirm our expectations.

	Type I error	Type II error	FWER	FDR
	Global testing procedure			
Mixture of low-rank factorizations	0.01	0.01		
MANOVA on summary measures	0.09	0.90		
	Local testing procedure			
Mixture of low-rank factorizations	0.0004	0.0587	0.0600	0.0023
Separate Fisher's exact tests	0.0036	0.5983	0.4000	0.0387

TABLE 3.1: Comparison of error rates for our procedure against MANOVA on summary statistics for global testing and separate Fisher's exact tests for local hypotheses.

	Minimum	Mean	Median	Maximum
	Area under the ROC curve			
Mixture of low-rank factorizations	0.969	0.999	1.000	1.000
Separate Fisher's exact tests	0.810	0.921	0.923	0.989

TABLE 3.2: Summary of the AUCs computed for the 100 simulated datasets in the dependence scenario, to assess performance of local testing at varying thresholds. The ROC curves are constructed using the true hypotheses indicators – $\delta_l = 0$ if H_{0l} is true, $\delta_l = 1$ if H_{1l} is true, $l = 1, \ldots, V(V-1)/2$ – and the acceptance or rejection decisions based on our procedure and Fisher's exact tests at varying the thresholds on posterior probabilities or FDR, respectively.

To assess frequentist operating characteristics, we repeated the above simulation exercise for 100 simulated datasets under both dependence and independence scenarios. The MANOVA test is performed under a threshold $\alpha = 0.1$, while the decision rule in the local Fisher's exact tests is based on the Benjamini and Hochberg (1995) threshold to maintain a false discovery rate FDR ≤ 0.1. Under our Bayesian procedure we reject the global null if $\hat{\text{pr}}[H_1 \mid \{y, \mathcal{L}(A)\}] > 0.9$. As in our settings the prior odds $\text{pr}(H_1)/\text{pr}(H_0) = 1$, previous threshold implies rejecting the global null when the Bayes factor provides an evidence against H_0 which is substantially close or higher than strong (Kass and Raftery, 1995). According to sensitivity analyses, reasonable changes in the previous threshold do not affect the final conclusions. Consistently with our initial simulation we reject local nulls if $\hat{\text{pr}}[H_{1l} \mid \{y, \mathcal{L}(A)\}] > 0.9$. Also in this case results are not substantially affected by moderate changes in the threshold both in simulation and application, hence we maintain this choice to preserve coherence in our analyses.

Table 3.1 confirms the superior performance of our approach in maintaining all error rates close to zero, in both global and local testing, while intrinsically adjusting for multiplicity. The information reduction via summary measures for the global test and the lack of a network structure in the local Fisher's exact tests lead to procedures with substantially less power. Although Table 3.1 has been constructed using an FDR control of 0.1 in the Fisher's exact tests and a threshold of 0.9 under our local testing procedure, we maintain superior performance allowing the thresholds to vary, as shown in Table 3.2.

In considering sample size versus type I and type II error rates, it is interesting to assess the rate at which the posterior probability of the global alternative $\text{pr}[H_1 \mid \{y, \mathcal{L}(A)\}]$ converges to 0 and 1 under H_0 and H_1, respectively, as n increases. We evaluate this behavior by simulating 100 datasets as in the previous simulation for increasing sample sizes $n = 20$, $n = 40$ and

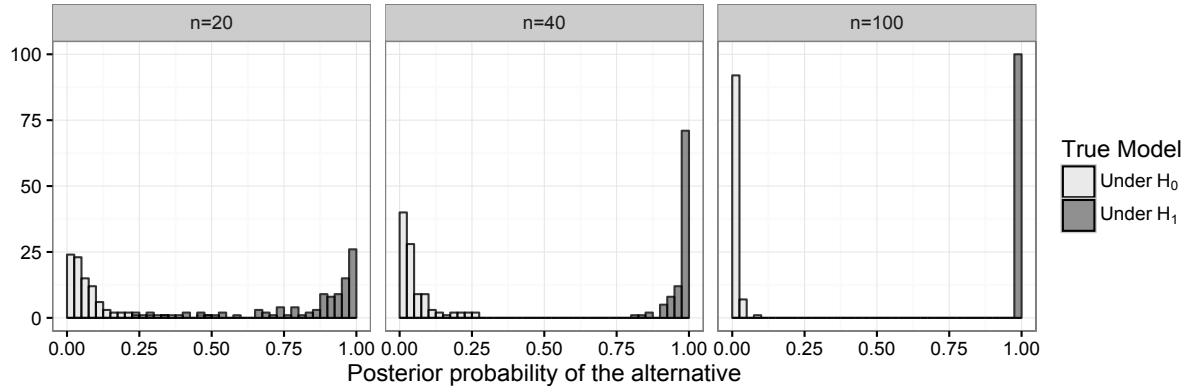


FIGURE 3.12: For increasing sample sizes n, histograms of the estimated posterior probabilities of the global alternative H_1 in each of the 100 simulations under association and no association.

$n = 100$ and for each scenario. Figure 3.12 provides the histograms showing the estimated posterior probabilities of H_1 for the 100 simulated datasets under the two scenarios and for increasing sample sizes. The separation between scenarios is evident for all sample sizes, with $\hat{\text{pr}}[H_1 \mid \{y, \mathcal{L}(A)\}]$ consistently concentrating around 0 and 1 under the no association and association scenario, respectively, as n increases. When $n = 20$ the test has lower power, with $32/100$ samples having $\hat{\text{pr}}[H_1 \mid \{y, \mathcal{L}(A)\}] < 0.9$ when H_1 is true. However, type I errors were rare, with $1/100$ samples having $\hat{\text{pr}}[H_1 \mid \{y, \mathcal{L}(A)\}] > 0.9$ when data are generated under H_0. These values are very close to 0 when the sample size is increased to $n = 40$ and $n = 100$, with the latter showing strongly concentrated estimates around 1 and 0, when H_1 is true and H_0 is true, respectively.

We conclude our simulation studies by considering a final scenario in which there is a strong association between $\mathcal{L}(\mathcal{A})$ and $\mathcal{Y}$, but this dependence arises from changes in more complex functionals of the probabilistic generative mechanism, instead of edge probabilities. Specifically, we simulate $n = 50$ pairs (y_i, A_i) from our model (3.15) and (3.17), with y_i from a categorical variable having $p^0_{\mathcal{Y}} = (0.5, 0.5)$ and A_i a $V \times V$ network with $V = 20$ nodes. In defining (3.17) we consider $H = 3$ components and again split the nodes in two blocks $V_1 = \{1, \ldots, 10\}$ and $V_2 = \{11, \ldots, 20\}$, characterizing – for example – the two different hemispheres. When $h = 1$, the vector $\pi^{0(1)}$ characterizes this block structure, with the probability of an edge between pairs of nodes in the same block set at 0.75, while nodes in different blocks have 0.5 probability to be connected. Vectors $\pi^{0(2)}$ and $\pi^{0(3)}$ maintain the same within block probability of 0.75 as in $\pi^{0(1)}$, but have different across block probability. In component $h = 2$ the latter increases by 0.3 – from 0.5 to 0.8 – while in component $h = 3$ this quantity decreases by the same value – from 0.5 to 0.2. As a result, when letting $\nu^0_1 = (1, 0, 0)$ and $\nu^0_2 = (0, 0.5, 0.5)$ it is easy to show that the group-specific edge probabilities – characterizing the distribution of each edge in the two groups – remain equal $\bar{\pi}^0_1 = \bar{\pi}^0_2$, even if the probability mass function jointly assigned to these edges changes across groups $p^0_{\mathcal{L}(\mathcal{A})|1} \neq p^0_{\mathcal{L}(\mathcal{A})|2}$. This provide a subtle scenario for the several procedures assessing evidence of changes in the brain across groups,

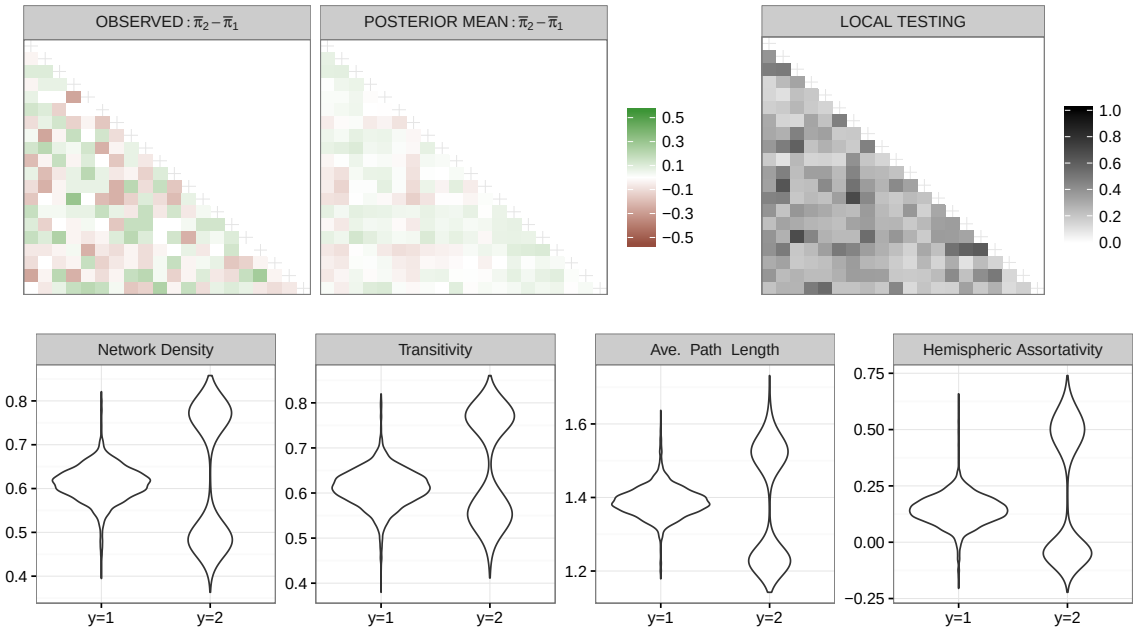


FIGURE 3.13: Model performance in the final simulation scenario. Upper-left adjacency matrix: group difference between the relative edge frequencies for each pair of nodes computed from the simulated data (lower triangular) versus true group difference in edge probabilities (upper triangular). Upper-middle adjacency matrix: posterior mean of the difference between the edge probabilities in the two groups (lower triangular) versus true group difference in edge probabilities (upper triangular). Upper-right adjacency matrix: $\hat{\text{pr}}[H_{1l} \mid \{y, \mathcal{L}(A)\}] = \text{pr}[\rho_l > 0.1 \mid \{y, \mathcal{L}(A)\}]$ (lower triangular) and accepted (white) or rejected (black) local null hypotheses (upper triangular). Lower panels: violin plots representing the density of selected network summary statistics in the two groups, arising from the posterior predictive distribution associated with our model.

by focusing on marginal or expected quantities. In such setting, these strategies should – correctly – find no difference in edge probabilities and hence may be – wrongly – prone to conclude that the brain network does not change across groups. Underestimating associations may be a dangerous fallacy in understating – for example – the effect of a neurological disorder that induces changes in more complex functionals of the brain network.

We apply our procedures to these simulated data under the same settings of our initial simulations, obtaining very similar effective sample sizes and PSRFs. As shown in the upper panels of Figure 3.13 the posterior probabilities for all the local alternatives are lower than 0.9 and hence our multiple testing procedure accepts H_{0l} for every $l = 1, \ldots, V(V-1)/2$. Beside correctly assessing the evidence of no changes in edge probabilities across the two groups, our global test is able to detect variations in more complex functionals of the brain network. In fact we obtain $\hat{\text{pr}}[H_1 \mid \{y, \mathcal{L}(A)\}] > 0.99$, meaning that although there is no evidence of changes in edge probabilities across the two groups, the model finds a strong association between $\mathcal{L}(\mathcal{A})$ and $\mathcal{Y}$.

The type of these variations can be observed in the lower panels of Figure 3.13 showing

the distribution of the selected network summary statistics arising from the posterior predictive distribution associated to our model. Although the latter is not analytically available, it is straightforward to simulate from the posterior predictive distribution exploiting our constructive representation in Figure 3.6 and posterior samples for the quantities in (3.15), (3.17) and (3.6), in line with the strategy outlined at the end of Section 3.1.6. According to the lower panels of Figure 3.13 there are substantial changes in the pmf of the network data across groups. In group one our model infers network summary measures having unimodal distributions, while in the second group we learn substantially different bimodal distributions. This behavior was expected based on our simulation, and hence these results further confirm the accuracy of our global test along with the good performance of our model in flexibly characterizing the distribution of a network-valued random variable and its variations across groups.

3.1.10 Application to brain network data and creativity

Although there is increasing interest in understanding how the structural interconnections in the brain play a critical role in creative cognition, as discussed in Section 1.2.1 and in Arden et al. (2010), current findings lack agreement due to the absence of a unifying approach to statistical inference in this field. The major effort of our procedures is in addressing these issues via provably general formulations which can flexibly characterize the richness of the network structure and avoid ad-hoc data reduction strategies prior to statistical modeling.

According to previous discussion, we evaluate our procedure on the creativity data set described in Section 1.2.1 using the same settings as in the simulation examples, but with upper bound H increased to $H = 15$. This choice proves to be sufficient with classes $h = 12, \ldots, 15$ having no observations and redundant dimensions of the latent spaces efficiently removed. The efficiency of the Gibbs sampler was very good, with effective sample sizes around $1{,}500$ out of $4{,}000$. Similarly the PSRFs provide evidence that convergence has been reached, as the highest of these quantities is 1.15. These checks on mixing and convergence are performed for the chains associated to quantities of interest for inference and testing. These include, the Cramer's V coefficients ρ_l, $l = 1, \ldots, V(V-1)/2$ for local testing, the group-specific edge probability vectors $\bar{\pi}_1$ and $\bar{\pi}_2$, the unconditional edge probability vector $\bar{\pi} = p_{\mathcal{Y}}(1)\bar{\pi}_1 + p_{\mathcal{Y}}(2)\bar{\pi}_2$ and the expectation of selected network summary statistics.

Our results provide interesting insights into the global relation between the brain network and creativity, with $\hat{\text{pr}}[H_1 \mid \{y, \mathcal{L}(A)\}] = 0.995$ strongly favoring the alternative hypothesis of association between brain region interconnections and level of creativity. In order to further assess the robustness of our global test we also performed posterior computation by randomly matching the observed group membership variables y_i with the brain networks

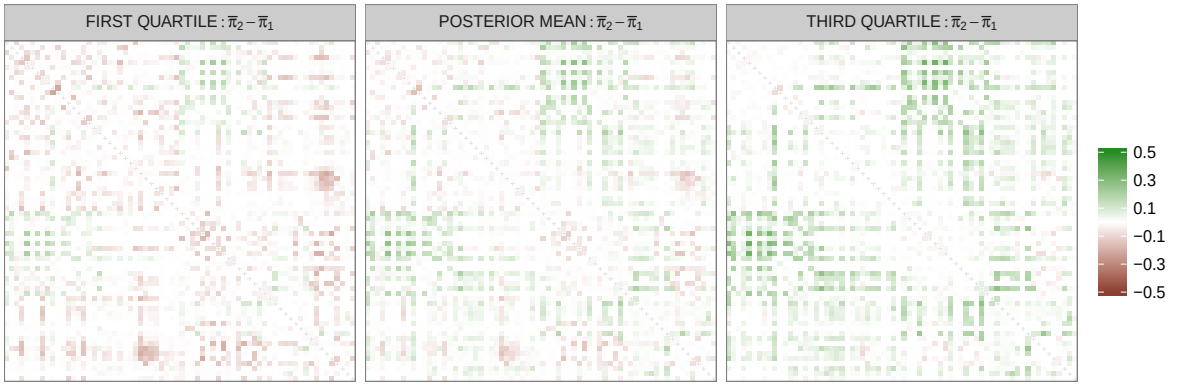


FIGURE 3.14: Mean and quartiles for the posterior distribution of the difference between the edge probabilities in high creativity subjects $\bar{\pi}_2$ and low creativity subjects $\bar{\pi}_1$.

$\mathcal{L}(A_i)$. In 10 of these trials we always obtained – as expected – low $\hat{\text{pr}}[H_1 \mid \{y, \mathcal{L}(A)\}] \leq 0.2$. This reasonably confirms the reliability of our conclusions.

We also attempted to apply the MANOVA test as implemented in the simulation experiments, with the same network statistics – i.e. network density, transitivity, average path length and assortativity by hemisphere. These are popular and key measures in neuroscience in informing on fundamental properties in brain network organization such as small-world, homophily patterns and scale-free behaviors (Bullmore and Sporns, 2009; Rubinov and Sporns, 2010; Bullmore and Sporns, 2012). In our dataset, the average path length was undefined for three subjects, as there were no paths between several pairs of their brain regions. Replacing these undefined shortest path lengths with the maximum path lengths, we obtain no significant difference in summary measures across creativity groups with a p-value of 0.111. When excluding this topological measure, we instead obtain a borderline p-value of 0.054. This sensitivity to the choice of summary statistics further motivates tests that avoid choosing topological measures, which is an inherently arbitrary exercise.

A key of our procedure is in providing efficient dimensionality reduction via mixture modeling and matrix factorization procedures, while preserving general flexibility in characterizing replicated network data. In fact, we obtain excellent performance – with an AUC $= 0.97$ – in edge prediction exploiting the posterior mean of the group-specific edge probabilities. Specifically we consider $\hat{\bar{\pi}}_1$ in predicting edges for brains in the low creativity group and $\hat{\bar{\pi}}_2$ for brains in the high creativity group. Beside providing a flexible approach in joint modeling of networks and categorical predictors, our methodology represents also a powerful tool to predict y_i given the subject's full brain network structure. In fact, under our framework, the probability that a subject i has high creativity, conditionally on his brain structural connectivity network A_i, is simply

$$\text{pr}\{\mathcal{Y}_i = 2 \mid \mathcal{L}(A_i)\} = 1 - \text{pr}\{\mathcal{Y}_i = 1 \mid \mathcal{L}(A_i)\} = \frac{p_{\mathcal{Y}}(2) p_{\mathcal{L}(\mathcal{A})|2}(a_i)}{p_{\mathcal{Y}}(2) p_{\mathcal{L}(\mathcal{A})|2}(a_i) + p_{\mathcal{Y}}(1) p_{\mathcal{L}(\mathcal{A})|1}(a_i)}, \quad (3.22)$$

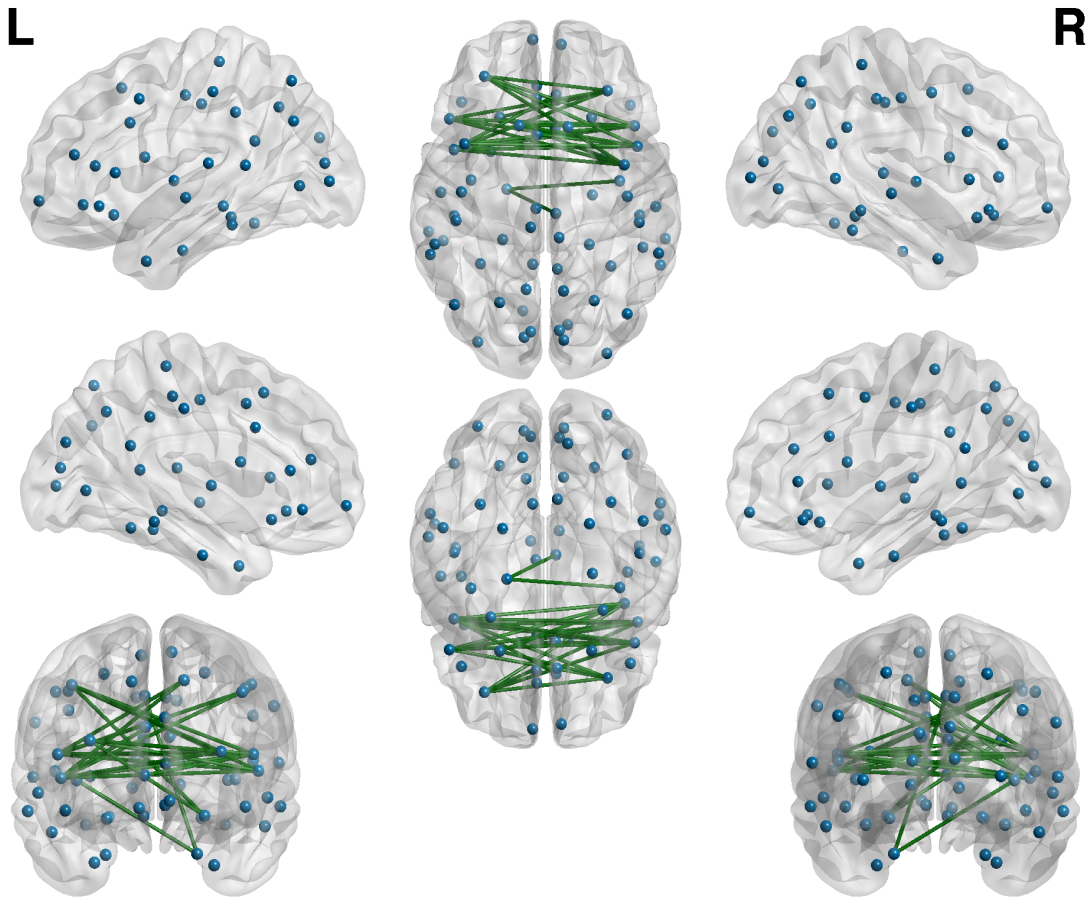


FIGURE 3.15: Brain network visualization exploiting results from our local testing procedure. We only display those connections which provide evidence of changes across high and low creativity subjects based on our procedure – i.e. $\hat{\text{pr}}[H_{1l} \mid \{y, \mathcal{L}(A)\}] > 0.9$. Edge color is green – or red – if its estimated probability in high creativity subjects is greater – or less – than low creativity ones. Regions positions are given by the spatial coordinates in the brain, with the same brain displayed from different views.

where $a_i = \mathcal{L}(A_i)$ is the network configuration of the ith subject and $p_{\mathcal{L}(\mathcal{A})|y}(a_i)$, $y \in \{1, 2\}$ can be easily computed from (3.17). We obtain an AUC $= 0.87$ in predicting the creativity group y_i using the posterior mean of $\text{pr}\{\mathcal{Y}_i = 2 \mid \mathcal{L}(A_i)\} = 1 - \text{pr}\{\mathcal{Y}_i = 1 \mid \mathcal{L}(A_i)\}$ for each $i = 1, \ldots, n$. Hence, allowing the conditional pmf of the network-valued random variable to shift across groups via group-specific mixing probabilities provides a good characterization of the dependence between brains and creativity, leading to accurate prediction of the creativity group.

Previous results highlight a good fit of our model to the data, motivating further analyses and interpretation of the results with respect to available literature. Figure 3.14 provides summaries of the posterior distribution for $\bar{\pi}_2 - \bar{\pi}_1$, with $\bar{\pi}_2 = \sum_{h=1}^{H} \nu_{h2} \pi^{(h)}$ and $\bar{\pi}_1 = \sum_{h=1}^{H} \nu_{h1} \pi^{(h)}$ encoding the edge probabilities in high and low creativity groups, respectively, as well as the conditional expectation of the corresponding network-valued random variable. Most of these connections have a similar probability in the two groups, with more evident local differences for connections among brain regions in different hemispheres. Highly creative

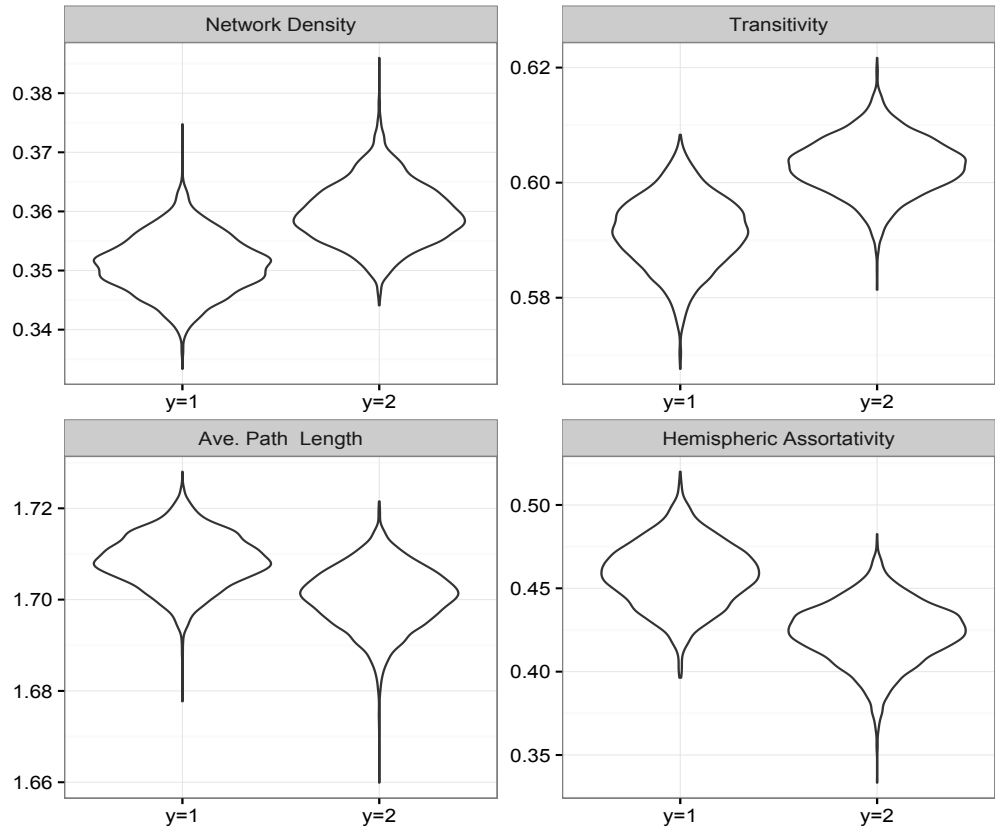


FIGURE 3.16: Violin plots representing the posterior distribution for the conditional expectation of selected network summary statistics in the two creativity groups.

individuals display a higher propensity to form inter-hemispheric connections. Differences in intra-hemispheric circuits are less evident. These findings are confirmed by Figure 3.15 including also results from our local testing procedure. As in the simulation we set $\epsilon = 0.1$ and the decision rule rejects the local nulls when $\hat{\text{pr}}[H_{1l} \mid \{y, \mathcal{L}(A)\}] > 0.9$. These choices provide reasonable settings based on simulations, and results are robust to moderate changes in the thresholds.

Early studies show that intra-hemispheric connections are more likely than inter-hemispheric connections for healthy individuals (Roncal et al., 2013). This is also evident in our dataset, with subjects having a proportion of intra-hemispheric edges of 0.55 over the total number of possible intra-hemispheric connections, against a proportion of about 0.21 for the inter-hemispheric ones. Our estimates in Figure 3.14 and local tests in Figure 3.15 highlight differences only in terms of inter-hemispheric connectivity, with high creative subjects having a stronger propensity to connect regions in different hemispheres. This is consistent with the idea that creative innovations arise from communication of brain regions that ordinarily are not connected (Heilman et al., 2003).

These findings contribute to the ongoing debate on the sources of creativity in the human

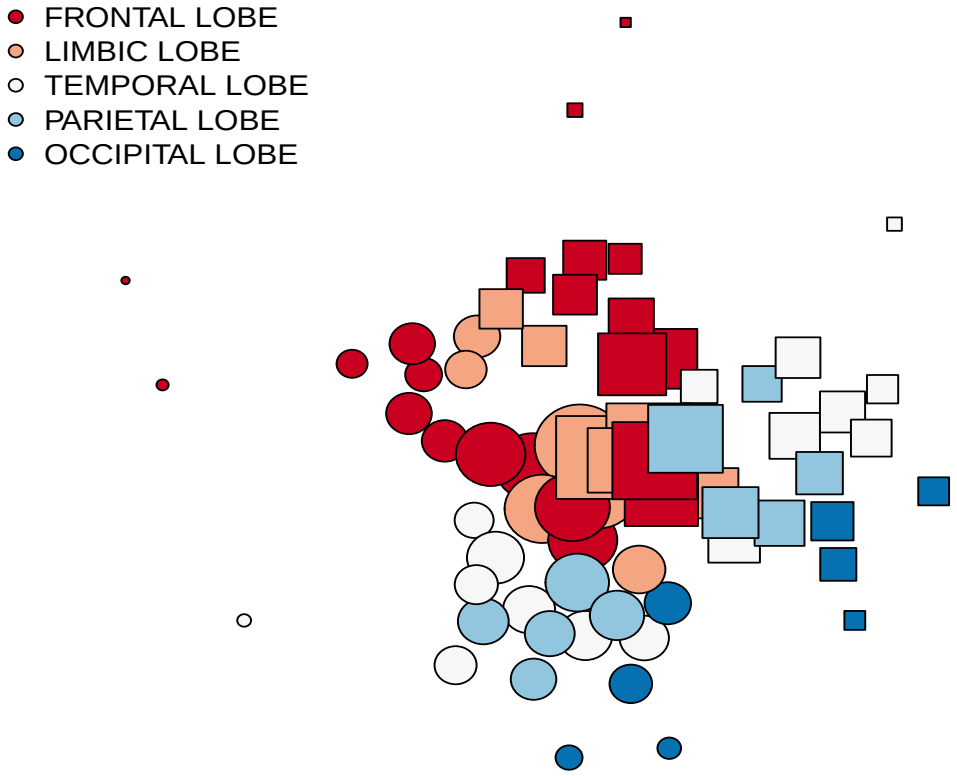


FIGURE 3.17: Weighted network representation with weights given by the posterior mean of the unconditional edge probabilities $\text{pr}\{\mathcal{L}(\mathcal{A})_l = 1\} = \bar{\pi}_l = p_{\mathcal{Y}}(1)\bar{\pi}_{1l} + p_{\mathcal{Y}}(2)\bar{\pi}_{2l}$, $l = 1, \ldots, V(V-1)/2$. Edges are not displayed to facilitate graphical analysis. Nodes positions are obtained by applying the Fruchterman and Reingold (1991) force-directed placement algorithm and sizes are proportional to their degree computed from the estimated $\bar{\pi}$. Circles and squares represent brain regions in the left and right hemispheres, respectively. Colors define anatomical lobe membership, according to Kang et al. (2012) classification of brain regions in anatomical lobes.

brain, with original theories considering the right-hemisphere as the seat of creative thinking, and more recent empirical analyses highlighting the importance of the level of communication between the two hemispheres of the brain; see Sawyer (2012), Shobe et al. (2009) and the references cited therein. Beside the different techniques in monitoring brain networks and measuring creativity, as stated in Arden et al. (2010), previous lack of agreement is likely due to the absence of a unifying approach to statistical inference in this field. Our method addresses this issue, while essentially supporting modern theories considering creativity as a result of cooperating hemispheres.

According to Figure 3.15 the differences in terms of inter-hemispheric connectivity are found mainly in the frontal lobe, where the co-activation circuits in the high creativity group are denser. This result is in line with recent findings highlighting the major role of the frontal lobe in creative cognition (Carlsson et al., 2000; Jung et al., 2010; Takeuchi et al., 2010). Previous analyses focus on variations in the activity of each region in isolation, with Carlsson et al. (2000) and Takeuchi et al. (2010) inferring an increase in cerebral blood flow and fractional anisotropy, respectively, for highly creative subjects, and Jung et al. (2010) showing a

negative association between creativity and cortical thickness in frontal regions. We instead provide inference on interconnections among these regions, with increased bilateral frontal connectivity for creative subjects, consistent with both the attempt to enhance frontal activity as suggested by Carlsson et al. (2000) and Takeuchi et al. (2010) or reduce it according to Jung et al. (2010).

Figure 3.16 shows the effect of the increased inter-hemispheric frontal connectivity – in high creativity subjects – on the posterior distribution of the key expected network summary statistics in the two groups. Although the expectation for most of these quantities cannot be analytically derived as a function of the parameters in (3.15) and (3.17), it is straightforward to obtain posterior samples for the previous measures via Monte Carlo methods exploiting the constructive representation in Figure 3.6. According to Figure 3.16 the brains in high creativity subjects are characterized by an improved architecture – compared to low creativity subjects – with increased connections, higher transitivity and shortest paths to connect pairs of nodes. As expected also hemispheric assortativity decreases. This is consistent with our local testing procedure providing evidence of increased inter-hemispheric activity and unchanged intra-hemispheric connectivity structures across the two groups. Previous results are also indicative of small-world structures in highlighting high transitivity and low average path length, with brains for high creativity subjects having a stronger small-world topology than subjects with low creativity. This property is a key in characterizing brain networks and hence our findings are in line with general results in neuroscience (Bullmore and Sporns, 2009).

We conclude our analysis by assessing the performance of our model formulation in characterizing also unconditional network structures. This is accomplished by providing a graphical network visualization based on the posterior mean of the unconditional expectation for the network-valued random variable arising from our model formulation. This quantity is easily available as $\bar{\pi} = p_{\mathcal{Y}}(1)\bar{\pi}_1 + p_{\mathcal{Y}}(2)\bar{\pi}_2$ and coincides also with the unconditional edge probability vector. Node positions in Figure 3.17 again highlight the two blocks induced by the hemispheres while additionally showing how regions in the same anatomical lobe are in general spatially closer. These results are consistent with neuroscience literature (Bullmore and Sporns, 2009), while being in line with the real spatial coordinates of the regions in the brain. This is a key insight on the performance of our model, as we learn previous structures only exploiting connectivity patterns without informing the model on spatial proximity of the nodes or their membership to hemispheres and lobes.

3.1.11 Application to brain network data and Alzheimer's

There is fundamental interest in understanding the relationship between the brain connectivity structure and neurodegenerative disorders, such as Alzheimer's, Parkinson's and other

dementias (Stam, 2014). These diseases are mainly found in aged populations and affect the normal functions of the central and peripheral nervous system causing, among others, muscle weakness, loss of coordination and cognitive impairment.

Alarming prevalence projections of dementia cases by the World Health Organization in 2006, and the rapid development of brain imaging technologies in recent years, have stimulated intensive research aimed at understanding how the brain structure is compromised with specific neurological diseases. This is key to improving diagnosis as well as providing increasingly targeted therapies. According to the Centers for Disease Control and Prevention (CDC), Alzheimer's disease (AD) is the most common form of dementia and the sixth leading cause of death in the United States. Unlike cancer and heart disease death rates, which are expected to decline, the growth of elderly population in the age range most commonly affected by dementia is leading to an increase of the death rates due to AD (James et al., 2014). This has strongly motivated intensive research aimed at finding the sources of AD in the human brain to develop increasingly refined diagnosis and prognosis procedures as well as improved therapy.

Current understanding of variations in brain behavior across AD is mostly available via early neuropathological studies (Hopper and Vogel, 1976), and contributions analyzing joint or local changes in the activity of each region under the modular paradigm (Thompson et al., 2001). More recent proposals shift increasingly away from the above approach towards studying brain activity networks via changes of the covariance in activity across brain regions for AD and controls (Bokde et al., 2006). However, functional connectivity matrices estimated from fMRI data do not reflect the underlying axonal pathways that can give rise to changes in function, and often require caution in interpreting the results (Bressler and Menon, 2010). This has motivated an increasing interest in structural connectivity matrices estimated from diffusion scans. Early studies on these data proceed by assessing variations of global brain network measures or region-specific connectivity statistics across AD and controls (Daianu et al., 2013). As previously noted, these methods may fail in flexibly characterizing the richness of the brain network structure, leading to inconsistent results. To address these issues, we apply our methodology to brain networks and Alzheimer's disease data described in Section 1.2.1. Posterior analysis is performed with the same settings of the application to creativity in Section 3.1.10, obtaining comparable results in terms of mixing and convergence.

The global testing procedure in (3.18) strongly favors the hypothesis of association between brain structural connectivity and AD diagnosis with $\hat{\text{pr}}[H_1 \mid \{y, \mathcal{L}(A)\}] > 0.99$. This confirms findings in Daianu et al. (2013) highlighting significant variations in brain network summary measures when comparing AD patients with cognitively healthy controls.

As expected the estimated significant differences between the edge probabilities in AD group and control group in Figure 3.18 show an overall less connected brain network for the

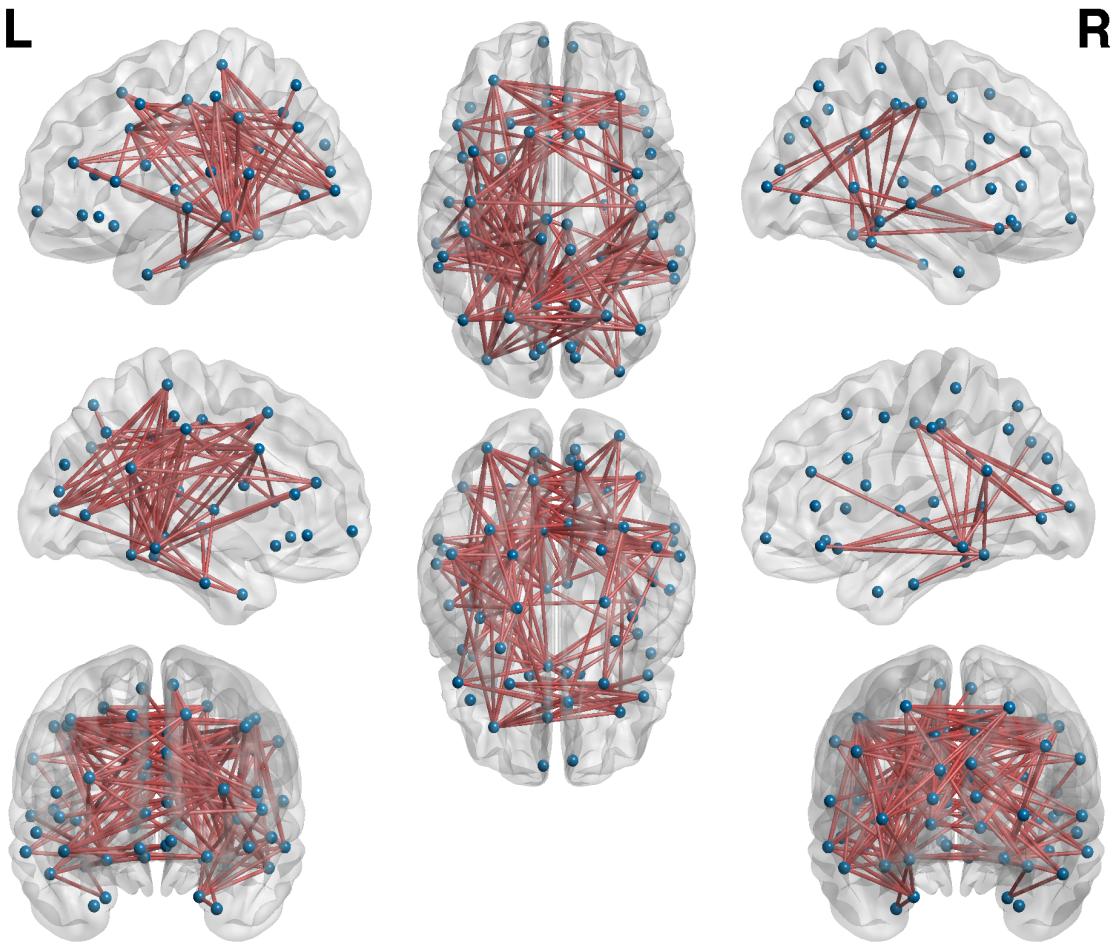


FIGURE 3.18: Brain network visualization exploiting results from our local testing procedure. We only display those connections which appear to be compromised by Alzheimer's based on our procedure – i.e. $\hat{\text{pr}}[H_{1l} \mid \{y, \mathcal{L}(A)\}] > 0.9$. Edge color is green – or red – if its estimated probability in Alzheimer's disease subjects is greater – or less – than healthy individuals. Regions positions are given by the spatial coordinates in the brain, with the same brain displayed from different views.

AD group compared to controls, in line with Daianu et al. (2013) and literature on AD. The main differences appear in terms of intra-hemispheric connections in the left hemisphere, while fewer local differences are found also in terms of inter-hemispheric connections and right intra-hemispheric. This major role of the left hemisphere agrees with Daianu et al. (2013) and Thompson et al. (2001). These findings are confirmed in Figure 3.19 summarizing the posterior distributions for elements in $\bar{\pi}_2 - \bar{\pi}_1$, with $\bar{\pi}_2 = \sum_{h=1}^{H} \nu_{h2}\pi^{(h)}$ and $\bar{\pi}_1 = \sum_{h=1}^{H} \nu_{h1}\pi^{(h)}$ encoding the edge probabilities in Alzheimer's disease and control groups, respectively. According to Figure 3.19 the entire posterior distributions – and not only posterior means – tend to concentrate on negative values for almost all connections. This further confirms the major effect of AD in compromising brain connectivity circuits.

The agreement with previous studies highlights the consistency of our methodology, which has the additional benefit of providing inference not only on the scale of the network summary measures but in terms of variations of the entire pmf for the brain network-valued

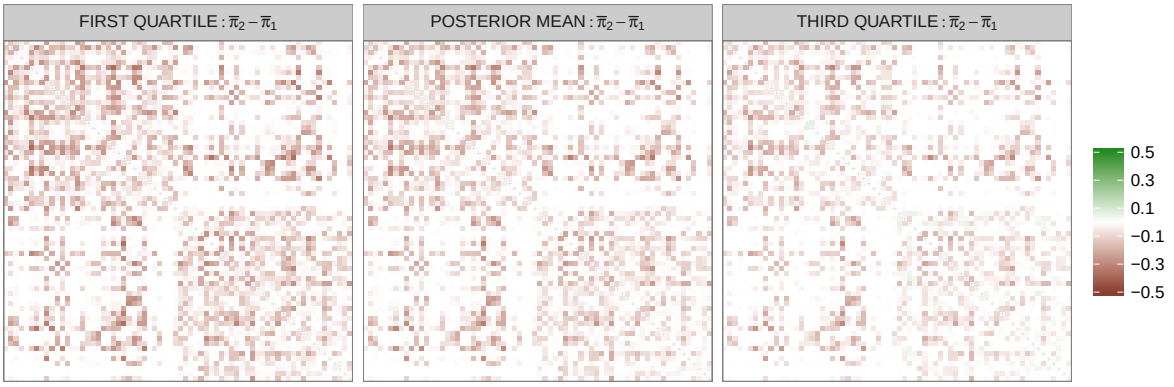


FIGURE 3.19: Mean and quartiles for the posterior distribution of the difference between the edge probabilities in Alzheimer's disease subjects $\bar{\pi}_2$ and and age-matched cognitively healthy controls $\bar{\pi}_1$.

random variable representing brain interconnections. This rules out the issue of conflicting conclusions when different network statistics are considered, while also avoiding ad-hoc choices when defining certain summary measures. Recalling for example Daianu et al. (2013) one may obtain different results when considering an order for the k-core different from 18. An additional benefit of our approach, as outlined in the simulation study, is that local testing intrinsically controls for multiplicity, while out-performing frequentist competitors controlling for FDR, in terms of power. Recalling the application to AD, this leads to a procedure which can more easily identify connections significantly varying between control and AD subjects. This is evident when comparing Figure 3.18 to results in Figure 1 in Daianu et al. (2013) learning less significant local differences. This result may be related to the use of a region-specific network statistic which displays low variations across case and controls as well as the choice of an overly conservative level for the FDR and the less power related to massively univariate local testing procedures.

Our approach doesn't rely on the choice of network summary measures and intrinsically controls for multiplicity, overcoming previous issues while strongly gaining power. As a result we learn more connections significantly varying between control and AD groups. This provides interesting new insights according to Figure 3.20, which displays for each region $v = 1, \dots, V$ the total number of connections among v and the remaining $V - 1$ regions significantly varying between controls and AD group under our local testing procedure (3.19) with $\epsilon = 0.1$. To highlight the roles of higher level brain systems, regions are grouped in anatomical lobes according to Kang et al. (2012) and in hemispheres. To facilitate comparison, we additionally maintain the same region's ID as in Table 3 of Daianu et al. (2013)

Results in Figure 3.20 highlight the connectivity breakdown for regions in the left hemisphere while providing new insights with respect to Daianu et al. (2013). In particular we learn the major role of regions in the left limbic lobe consistently with initial neuropathological studies (Hopper and Vogel, 1976; Blesa et al., 1995) and more recent empirical findings

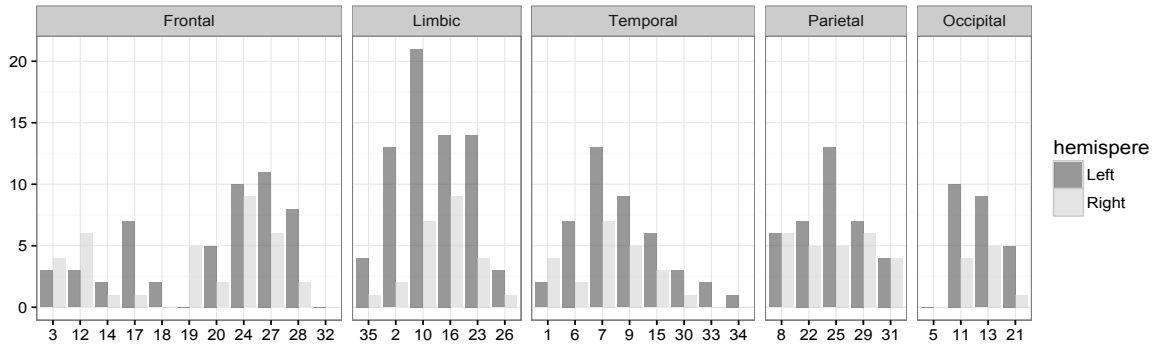


FIGURE 3.20: Test degree for each brain region – classified in left and right hemisphere and corresponding lobe. The test degree of region v is defined as the total number of connections among v and the remaining $V - 1$ regions significantly varying in Alzheimer's group. To facilitate comparison, we maintain the same region's ID as in Table 3 of Daianu et al. (2013)

via MRI (Deoni et al., 2011; Thiebaut de Schotten et al., 2014) highlighting the key role of the limbic system in memory, attention and executive functioning, while focusing on this lobe as one of the areas mainly affected by AD. Significant changes are also found in the connectivity of the other anatomical lobes such as temporal, parietal and occipital, consistent with Smith et al. (2001), Azari et al. (1992), Thompson et al. (2003) and Horwitz et al. (1987).

According to Figure 3.20 the regions mostly affected by AD in terms of connectivity behavior are the left isthmus of the cingulate (10L), left parahippocampal (16L), left posterior cingulate (23L), left fusiform (7L) and left precuneus (25L) – among others. These results provide a unifying answer to different insights arising from several studies, typically focusing on the activity of a subset of regions. Parahippocampal atrophy is found in Kesslak et al. (1991) and Thangavel et al. (2008); Zhou et al. (2008) highlights abnormal connectivity in hippocampus and posterior cingulate, while Kim et al. (2013) learn reduced functional activity in hippocampus and precuneus, with the latter showing atrophy also in Karas et al. (2007). Metabolic reduction in the posterior cingulate is studied in Minoshima et al. (1997) and Liang et al. (2008). Reduced functional connectivity in the fusiform is found in Golby et al. (2005) and Bokde et al. (2006) via fMRI. Fewer studies are available on the role of the isthmus of the cingulate with only a recent work of Zhu et al. (2013) trying a first attempt in this direction. We provide a unifying vision, consistent with previous literature, while highlighting the role of the isthmus. This region represents an anatomical bridge between the parahippocampal and the posterior cingulate, two critical regions extensively explored in the literature in terms of atrophy and metabolic reduction in AD subjects. Hence a reduced metabolic activity and increased atrophy of parahippocampal and the posterior cingulate, may be related to a disruption of the circuits from the left cingulate isthmus.

We conclude our application by evaluating the ability of our procedure in equation (3.22) to assess evidence of AD according to the subject's full brain network structure. Current

prediction procedures exploit either region activity vectors (Eskildsen et al., 2015) or network summary statistics vectors θ_i (Friedman et al., 2014; Prasad et al., 2015), rather than the whole brain network $\mathcal{L}(A_i)$, to predict y_i. We evaluate our procedure in (3.22) in terms of in-sample and out-of-sample predictive performance. In the first case, we compute (3.22) for each subject after considering all data in model estimation. Out-of-sample predictions is instead performed by training the model on 69 subjects and predicting the AD status via (3.22) on the remaining one fourth of the individuals, with the training and test samples randomly selected. Our methodology provides an overall good predictive performance, with an area under the ROC curve of 0.91 for in-sample prediction and 0.83 for out-of-sample. The accuracy is instead 87% in the former, and 75% in the latter. These results out perform Eskildsen et al. (2015), and Friedman et al. (2014) when summary statistics θ_i are extracted from undirected brain networks, while providing similar performance to Prasad et al. (2015). It is important to note that Prasad et al. (2015) utilizes substantially more information in considering both weighted and flow connectivity networks for a total of $298{,}600$ network summary measures, rather than only binary connections encoding presence or absence of fibers.

3.2 Bayesian modeling of mixed domain data

Methodologies developed in Section 3.1 provide a general procedure for answering applied questions also outside the neuroscience field. Motivated by a complex business intelligence problem for targeted advertising of cross-selling strategies in different agencies, we develop a flexible joint model for mixed domain data including mono-product customer preferences and co-subscription networks measuring multiple buying behavior across different agencies. Our procedure defines shared sets of cross-selling strategies by effectively clustering agencies characterized by common mono- and multi-product customer behavior. Each segment is carefully profiled by modeling customer preferences and co-subscription networks via the dependent mixture of low-rank factorization proposed in Section 3.1.7. Exploiting such estimates, we construct cluster-specific sets of cross-selling strategies informing for each product v which additional product $u \neq v$ should be offered to obtain the highest probability of a co-subscription by a mono-product customer subscribed to v. We evaluate the effectiveness of each strategy via performance indicators accounting also for mono-product customer preferences. We provide simple algorithms for posterior computation and assess performance in simulations and application to the data set described in Section 1.2.2.

3.2.1 Joint modeling of mono-product data and co-subscription networks

As a step towards our goal of designing efficient cross-sell strategies, we first develop joint models for the data $\{d_i = (d_{i1}, \ldots, d_{in_i}), A_i\}$, for each agency $i = 1, \ldots, n$, which characterize the distribution of the mono-product customer subscriptions along with the co-subscription network for multi-product customers. The models are chosen to be flexible while automatically clustering different agencies that have similar customer mono-product and co-subscription network profiles. This clustering is useful for borrowing information across agencies in efficiently and effectively learning the joint distribution of the mono- and co-subscription behavior of the customer bases. In addition, clustering provides a useful simplification in design of strategies.

Let $(y_1, \ldots, y_n)$ denote a vector of cluster assignments, with $y_i \in \{1, \ldots, K\}$ indicating the cluster membership of agency i. Agencies within the same cluster are characterized by a similar composition of their mono-product portfolio as well as a comparable co-subscription behavior. To complete a specification of the joint model, we need to define a cluster-specific probabilistic representation $p_{\mathcal{D}|y}$ of mono-product customer choice, as well as a cluster-specific probabilistic generative mechanism $p_{\mathcal{L}(\mathcal{A})|y}$ underlying co-subscription networks. The latter is a key to define the cross-selling strategies $q_{y1}, \ldots, q_{yV}$ in each cluster $y = 1, \ldots, K$, while the former provides the additional information to construct performance indicators $e_{y1}, \ldots, e_{yV}$, according to discussion in Section 1.2.2.

As a mono-product customer can be associated with only one subscription $v = 1, \ldots, V$, it is straightforward to define a probabilistic representation of the mono-product customer behavior within each cluster y. In particular, we introduce a cluster-specific vector $p_{\mathcal{D}|y} = \{p_{\mathcal{D}|y}(1), \ldots, p_{\mathcal{D}|y}(V)\}$, with element $p_{\mathcal{D}|y}(v)$ defining the probability that a mono-product customer in an agency within cluster y subscribes to product v. Assuming independence of customer choices, the joint probability for data d_i in agency i given its membership to cluster y is simply

$$p_{\mathcal{D}|y}(d_{i1})p_{\mathcal{D}|y}(d_{i2}) \cdots p_{\mathcal{D}|y}(d_{in_i}) = \prod_{v=1}^{V} p_{\mathcal{D}|y}(v)^{n_{iv}}, \quad i = 1, \ldots, n, \tag{3.23}$$

where n_{iv} is the total number of mono-product customers in agency i who subscribed to product v for each $v = 1, \ldots, V$.

Within each cluster y, co-subscription networks are realizations from a network-valued random variable with associated conditional probability mass function $p_{\mathcal{L}(\mathcal{A})|y}$ where $p_{\mathcal{L}(\mathcal{A})|y}(a)$ defines the probability of observing the network configuration $\mathcal{L}(A) = a \in \mathbb{A}_V$ in cluster y. Similarly to methodologies developed in previous Sections, we define a probabilistic generative mechanism for the adjacency matrices – characterizing co-subscription networks – by modeling their lower triangular elements. In fact, since networks are undirected and self-relationships are not of interest, the probability mass function on the entire symmetric adjacency matrix, coincides with the one on its lower triangular elements.

As there are $2^{V(V-1)/2} = |\mathbb{A}_V|$ distinct network configurations $a \in \mathbb{A}_V$, we cannot estimate $p_{\mathcal{L}(\mathcal{A})|y}$ nonparametrically without dimensionality reduction. Interestingly, these methodological issues coincide with those addressed in Section 3.1.7. In particular, the dependent mixture of low-rank factorizations in equation (3.17) has been specifically developed to reduce dimension while maintaining flexibility in characterizing changes in the distribution of brain networks across behavioral – or disease – groups. Replacing brain networks with co-subscription networks and behavioral – or disease – groups with agency-specific cluster indicators, we are then faced with a common underlying goal and hence we can exploit the same dependent mixture of low-rank factorizations to characterize $p_{\mathcal{L}(\mathcal{A})|y}$. Hence, generalizing equations (3.17) to the multiple group case, this leads to the following probability for the co-subscription network $\mathcal{L}(A_i) = a_i$ in agency i given its membership to cluster y:

$$p_{\mathcal{L}(\mathcal{A})|y}(a_i) = \sum_{h=1}^{H} \nu_{hy} \prod_{l=1}^{V(V-1)/2} \left\{\pi_l^{(h)}\right\}^{a_{il}} \left\{1 - \pi_l^{(h)}\right\}^{1-a_{il}}, \tag{3.24}$$

with each $\pi^{(h)} = (\pi_1^{(h)}, \ldots, \pi_{V(V-1)/2}^{(h)})^{\mathrm{T}}$ factorized as

$$\pi^{(h)} = \left[1 + \exp\{-Z - D^{(h)}\}\right]^{-1}, \quad D^{(h)} = \mathcal{L}(X^{(h)}\Lambda^{(h)}X^{(h)\mathrm{T}}), \quad h = 1, \ldots, H. \tag{3.25}$$

Equations (3.24)–(3.25) carefully incorporate cluster dependence in (3.24) via group-specific mixing probabilities $\nu_y = (\nu_{1y}, \ldots, \nu_{Hy})$, $y = 1, \ldots, K$ as well as network information by considering a different low-rank factorization representation (3.25) for the class-specific edge probability vectors $\pi^{(h)}$, $h = 1, \ldots, H$.

Recalling discussion in Sections 3.1.2, 3.1.3, 3.1.7, and focusing on class h, the low-rank factorization mechanism assumes the undirected edges are realizations from conditionally independent Bernoulli random variables given their specific edge probabilities $\pi_l^{(h)} \in (0, 1)$, and then borrows network dependence across these edge probabilities $\pi_l^{(h)}$ $l = 1, \ldots, V(V-1)/2$ via lower dimensional representations. In particular, according to (3.25) – and letting l corresponds to the pair of products v and u – each $\pi_l^{(h)}$ is constructed as a function of the pairwise similarity among products v and u in a latent space, with this similarity arising from the dot product of the products' latent coordinate vectors $X_v^{(h)} = \{X_{v1}^{(h)}, \ldots, X_{vR}^{(h)}\}^{\mathrm{T}} \in \Re^R$, $v = 1, \ldots, V$, with $X_v^{(h)\mathrm{T}}$ the vth row of $X^{(h)}$. Hence, products having coordinates in the same direction are more likely to be co-subscribed than products characterized by coordinates in opposite directions, with the $R \times R$ matrix $\Lambda^{(h)} = \mathrm{diag}(\lambda_1^{(h)}, \ldots, \lambda_R^{(h)}) = \lambda^{(h)}$ weighting the similarity in each dimension r by a non-negative parameter $\lambda_r^{(h)}$.

The low-rank factorization in each class h provides an appealing choice in reducing the dimensionality from $V(V-1)/2$ edge probabilities to $V \times R$ latent coordinates and R weights – typically $R \ll V$ – and has been shown to provide an highly flexible characterization of the connectivity patterns and network structures, according to simulations in Section 3.1.6. Moreover, according to Corollary 3.6, mixing together H low-rank factorization mechanisms as in equation (3.24) guarantees full flexibility in approximating the collection of cluster-specific probability mass functions $p_{\mathcal{L}(\mathcal{A})|y} \in \mathcal{P}_{|\mathbb{A}_V|}$, $y = 1, \ldots, K$ for the co-subscription networks.

Our focus is on using the resulting flexible and parsimonious joint model for the mono-product portfolio and multi-product network to develop targeted strategic marketing policies. Figure 3.21 provides an example of the output from our model for decision making in business intelligence when there are $n = 8$ agencies and $K = 3$ latent clusters. According to Figure 3.21 agencies 1, 4 and 5 have a similar composition of their mono-product portfolio and comparable co-subscription behavior as $y_1 = y_4 = y_5 = 1$. In Figure 3.21, mono-product preferences are simply available via $p_{\mathcal{D}|1}$, while co-subscription behavior is summarized by the expectation for the network-valued random variable in cluster $y = 1$, $\bar{\pi}_1 = \sum_{a \in \mathbb{A}_V} a p_{\mathcal{L}(\mathcal{A})|1}(a) = \sum_{h=1}^{H} \nu_{h1} \pi^{(h)}$ according to Proposition (3.2). As discussed in Section 3.1.7 these quantities coincide also with the co-subscription probabilities for pairs of products in each cluster and hence can be used to define the set of cross-selling marketing strategies $q_{11}, \ldots, q_{1V}$ in cluster $y = 1$. The same description holds for clusters $y = 2$ and $y = 3$.

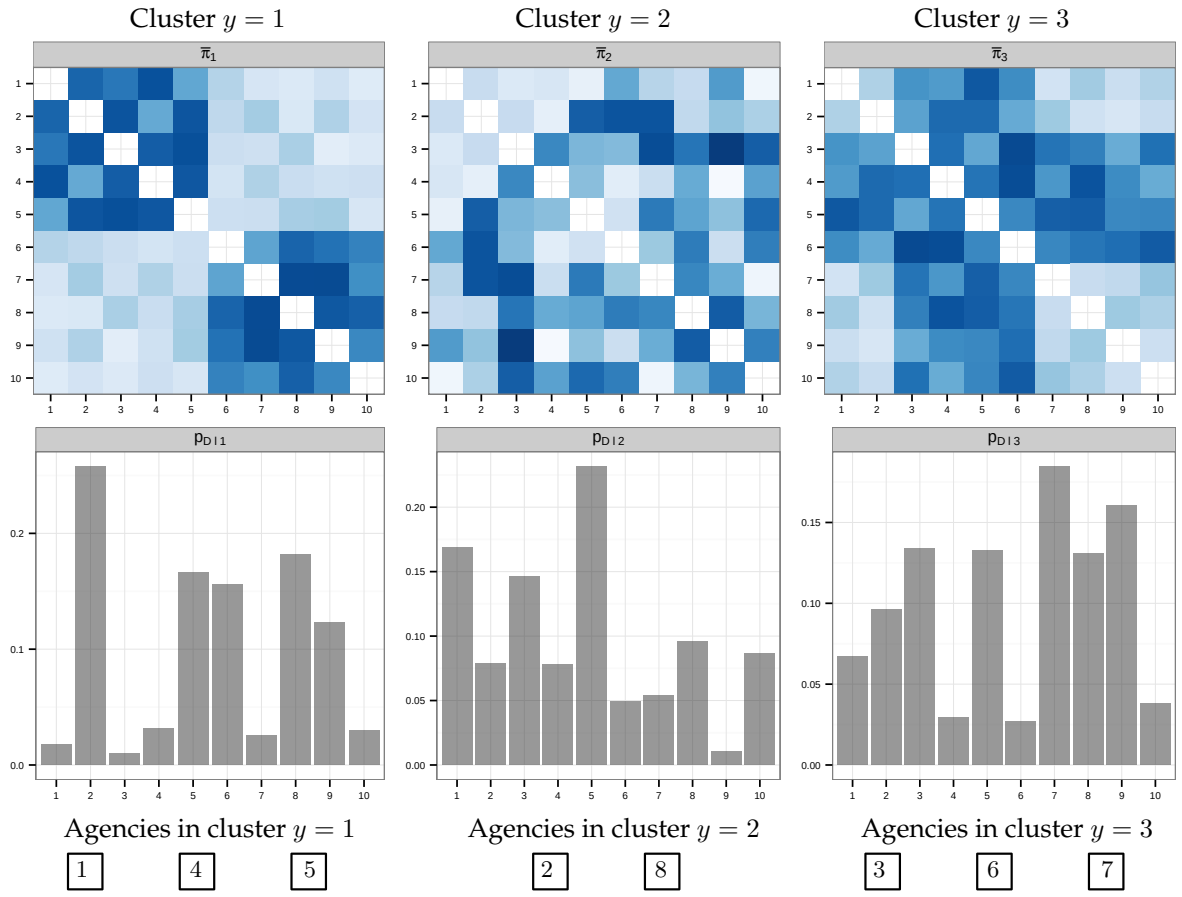


FIGURE 3.21: Example of a possible output from our model for decision making in business intelligence. The vectors of edge probabilities in each cluster $\bar{\pi}_1, \bar{\pi}_2$ and $\bar{\pi}_3$ are rearranged in adjacency matrix form. Color goes from white to dark blue as the probability goes from 0 to 1.

These quantities are estimated from our model and considered when defining the cluster-specific cross-selling marketing strategies $q_{y1}, \ldots, q_{yV}$ and computing their corresponding performance indicators $e_{y1}, \ldots, e_{yV}$. Adapting discussion in Section 1.2.2 to the output of our model, cross-selling strategy q_{yv}, offers to mono-product customers who subscribed to v in cluster y the additional product u, with $u = \text{argmax}_u\{\text{pr}(\mathcal{A}_{[vu]} = 1 \mid y) : u \neq v\}$ where the probability of a co-subscription of products v and u for agencies in cluster y, $\text{pr}(\mathcal{A}_{[vu]} = 1 \mid y)$, is easily available from our model as $\bar{\pi}_{yl}$, with l the index denoting the pair v and u in the vectorized representation of the adjacency matrix. The performance measure of q_{yv} is $e_{yv} = p_{\mathcal{D}|y}(v)\text{max}\{\text{pr}(\mathcal{A}_{[vu]} = 1 \mid y) : u \neq v\}$.

As motivated above, the main purpose of our analysis is to cluster agencies having similar customer bases in terms of mono-product and multi-product subscriptions, while providing accurate estimates of the performance measures for different marketing campaigns. In implementing computation, it is useful to rely on an equivalent hierarchical specification to factorization (3.24)–(3.25), which introduces an additional class index for each agency i,

$G_i \in \{1, \ldots, H\}$, as follows:

$$\begin{aligned} \mathcal{L}(\mathcal{A}_i)_l \mid \pi_{il} &\overset{\text{indep}}{\sim} \text{Bern}(\pi_{il}), \quad l = 1, \ldots, V(V-1)/2, \quad i = 1, \ldots, n, \\ \pi_i \mid G_i = h, \pi^{(h)} &= \pi^{(h)}, \quad \pi^{(h)} = \left[1 + \exp\{-Z - \mathcal{L}(X^{(h)}\Lambda^{(h)}X^{(h)\text{T}})\}\right]^{-1} \\ \text{pr}(G_i = h \mid y_i = y) &= \nu_{hy}, \quad h = 1, \ldots, H, \end{aligned} \tag{3.26}$$

independently for $i = 1, \ldots, n$. Recalling that y_i is the cluster index for agency i, representation (3.26) shows that a given cluster of agencies has a common set of weights over the components in the mixture model for the co-subscription network. The next section proposes a Bayesian approach to inference under the proposed model, which can be implemented via a simple Markov chain Monte Carlo algorithm. However – before proceeding to prior specification – it is worth noticing that although the hierarchical representation in (3.26) recalls the specification displayed in Figure 3.6, the statistical model developed in this Section has a key difference compared to the one outlined in Section 3.1.7. In particular, while in the previous neuroscience framework the predictor groups $y_1, \ldots, y_n$ represent exogenous observed data, in this business intelligence problem such quantities are model parameters – of key interest for inference – endogenously determined by mono-product choices and co-subscription behaviors shared across subsets of agencies.

3.2.2 Prior specification

As previously discussed, the assignment vector $(y_1, \ldots, y_n)$ of the agencies to K clusters is not observed from the data, but is a key unknown quantity in our analysis. This raises novel issues compared to methodologies developed in Section 3.1.7. A fundamental one is how to appropriately choose the total number of clusters K. Although K may be subject to budget restrictions and fixed a priori, this quantity is typically unknown in practical applications. Hence, in providing inference on such quantities, it is important to consider carefully tailored priors for the cluster assignments, which allow adaptive and automatic learning of the number of groups in our data.

There exists a considerable Bayesian nonparametric literature defining probabilistic generative mechanisms for clustering which allow the number of groups K to be random. A widely used prior for random partitions is the Chinese restaurant process (CRP) (Aldous, 1985), in which each cluster attracts new units in proportion to its size. In particular, letting $(y_1, \ldots, y_n) \sim \text{CRP}(\alpha_c)$, the prior distribution over clusters for the ith agency, conditioned on the membership of the others $y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n$ is

$$\text{pr}(y_i = y \mid y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n) = \begin{cases} \frac{n_{y,-i}}{n-1+\alpha_c} & \text{for } y = 1, \ldots, K_{-i}, \\ \frac{\alpha_c}{n-1+\alpha_c} & \text{for } y = K_{-i}+1, \end{cases} \tag{3.27}$$

where $n_{y,-i}$ is the total number of agencies associated to cluster y, excluding the ith one, and $\alpha_c > 0$ is a concentration parameter controlling the expected number of groups occupied by at least one of the n agencies $\mathrm{E}(K) = \sum_{i=1}^{n} \alpha_c/(i-1+\alpha_c) = O(\alpha_c \log n)$. High values of α_c favor a larger number of clusters a priori.

Equation (3.27) defines the conditional distribution of y_i given the other assignments, after rearranging indices so that $1, \ldots, K_{-i}$ clusters are nonempty after removing y_i. These operations are possible since the cluster labels are arbitrary and observations are exchangeable based on the CRP prior. In fact, the joint probability for any particular cluster assignment under the CRP representation is

$$\mathrm{pr}(y_1, \ldots, y_n) = \frac{\alpha_c^K \prod_{y=1}^{K} (n_y - 1)!}{\prod_{i=1}^{n} (i - 1 + \alpha_c)}. \tag{3.28}$$

Equation (3.28) depends only on the total number of agencies in each cluster n_y, $y = 1, \ldots, K$ and hence is invariant under permutation of elements or rearrangements of cluster indices; refer to Griffiths and Ghahramani (2011) and Gershman and Blei (2012) for an introductory overview. According to our aim, exchangeability is also a desirable property in characterizing absence of any particular knowledge about the type of agencies that would justify treating them differently from one another. Although we focus on the CRP, our model can easily accommodate other commonly used priors for random partitions, such as the stick-breaking construction for the Dirichlet process, the random partition induced by the Pitman-Yor process, and the Kingman paintbox; refer to Hjort et al. (2010) for a general overview.

Accurate clustering of agencies also relies on careful modeling of the sequence of cluster-specific mono-product portfolios $p_{\mathcal{D}|y}$, $y = 1, \ldots, K$ and the collection of cluster-specific probabilistic generative mechanism $p_{\mathcal{L}(\mathcal{A})|y}$, $y = 1, \ldots, K$ associated to the co-subscription networks. Efficient estimation of these quantities is also fundamental to develop accurate cross-selling strategies $q_{y1}, \ldots, q_{yV}$ in each cluster y and quantify their performance via $e_{y1}, \ldots, e_{yV}$. Hence we look for large support priors on these quantities which don't rule out a priori any generative mechanism while maintaining tractable computations.

As $p_{\mathcal{D}|y}$ is the probability mass function for a discrete random variable with V categories, we simply let

$$p_{\mathcal{D}|y} = \{p_{\mathcal{D}|y}(1), \ldots, p_{\mathcal{D}|y}(V)\} \sim \mathrm{Dirichlet}(\alpha_1, \ldots, \alpha_V), \tag{3.29}$$

independently for each cluster $y = 1, \ldots, K$.

The prior for the collection of co-subscription network probabilistic generative mechanisms $p_{\mathcal{L}(\mathcal{A})|y}$, $y = 1, \ldots, K$ is instead defined by choosing independent priors for the quantities in

factorization (3.24)–(3.25). To maintain computational tractability and recalling prior specification in Sections 3.1.4 and 3.1.8, we consider independent normal priors as in (3.10) for elements in Z and standard Gaussians for the latent coordinates $X^{(h)}$ for each $h = 1, \ldots, H$ according to equation (3.11). Adapting Rousseau and Mengersen (2011), we choose independent Dirichlet priors with small parameters for each cluster-specific mixing probability vector $\nu_y = (\nu_{1y}, \ldots, \nu_{Hy}) \sim \text{Dirichlet}(1/H, \ldots, 1/H)$ to favor deletion of redundant mixture components not required to characterize the co-subscription networks. Finally, as the dimension of latent spaces is unknown – consistently with discussion in Section 3.1.4 – we consider independent MIG(a_1, a_2) priors (Bhattacharya and Dunson, 2011) on the weights vector $\lambda^{(h)}$ for each $h = 1, \ldots, H$ to adaptively deletes redundant latent space dimensions not required to characterize the co-subscription probabilities; refer to equation (3.9) to recall the structure of the multiplicative inverse gamma prior.

Beside providing simple computational algorithms for posterior inference, a minor generalization of proof of Corollary 3.8 to the multiple group case, guarantees that the previous specifications induce a prior on the collection of probabilistic generative mechanisms $p_{\mathcal{L}(\mathcal{A})|y}$, $y = 1, \ldots, K$, for the co-subscription networks, with full support properties. Full prior support is a key property to ensure good performance in defining correct group-specific cross-selling strategies, because without prior support about the true data generating collection $p^0_{\mathcal{L}(\mathcal{A})|y}$, $y = 1, \ldots, K$, the posterior cannot possibly concentrate around the truth.

3.2.3 Posterior computation

Posterior computation is available via a simple Gibbs sampler outlined in Algorithms 5 and 6, which exploits results in Neal (2000) to allocate agencies to clusters under the CRP prior and steps in Algorithm 3 to update the quantities in equations (3.24)–(3.25) via Pólya-gamma data augmentation (Polson et al., 2013). Algorithm 5 proceeds as follows:

Algorithm 5 Part I of the Gibbs sampler for joint modeling of mixed domain data

---

Conditionally on cluster assignments $(y_1, \ldots, y_n)$ update priors for quantities in equations (3.23), (3.24) and (3.25), according to the following steps.

---

[1] Allocate each network observation to one of the classes

for $i = 1, \ldots, n$ **do**

Sample the class indicator G_i from the discrete distribution with probabilities

$$\text{pr}(G_i = h \mid -) = \frac{\nu_{hy_i} \prod_{l=1}^{V(V-1)/2} \{\pi_l^{(h)}\}^{\mathcal{L}(A_i)_l} \{1 - \pi_l^{(h)}\}^{1-\mathcal{L}(A_i)_l}}{\sum_{m=1}^{H} \nu_{my_i} \prod_{l=1}^{V(V-1)/2} \{\pi_l^{(m)}\}^{\mathcal{L}(A_i)_l} \{1 - \pi_l^{(m)}\}^{1-\mathcal{L}(A_i)_l}},$$

for each $h = 1, \ldots, H$, with $\pi^{(h)}$ defined in (3.25)

end for

---

[2] Update the cluster-specific mixing probabilities

for $y = 1, \dots, K$ **do**

Update each ν_y from

$$(\nu_{1y}, \dots, \nu_{Hy}) \mid - \sim \text{Dirichlet}(1/H + n_{1y}, \dots, 1/H + n_{Hy})$$

end for

---

[3] Update quantities Z, $X^{(h)}$ and $\lambda^{(h)}$, $h = 1, \dots, H$

Given G_i, $i = 1, \dots, n$, the updating for quantities Z, $X^{(h)}$ and $\lambda^{(h)}$, $h = 1, \dots, H$ proceeds as in Algorithm 3 via Polyá-gamma data augmentation. Specifically first sample Polyá-gamma augmented data as in step **[3]** of Gibbs sampler 3 and then update Z, $X^{(h)}$ and $\lambda^{(h)}$, $h = 1, \dots, H$ following steps **[4]**, **[5]** and **[6]**, respectively, of Algorithm 3.

---

[4] Update cluster-specific mono-product portfolio structures

for $y = 1, \dots, K$ **do**

Update each $\{p_{\mathcal{D}|y}(1), \dots, p_{\mathcal{D}|y}(V)\}$ from

$$\{p_{\mathcal{D}|y}(1), \dots, p_{\mathcal{D}|y}(V)\} \mid - \sim \text{Dirichlet}\left(\alpha_1 + \sum_{i:y_i=y} n_{i1}, \dots, \alpha_V + \sum_{i:y_i=y} n_{iV}\right)$$

end for

Algorithm 5 provides a detailed overview of the steps in our MCMC to update the cluster-specific probabilistic representation of the mono-product customer choice $p_{\mathcal{D}|y}$ and the cluster-specific probabilistic generative mechanism $p_{\mathcal{L}(\mathcal{A})|y}$ underlying co-subscription networks for each cluster $y = 1, \dots, K$, conditionally on cluster assignments $y_1, \dots, y_n$. All the steps are straightforward to compute, exploiting the data augmentation strategy described in (3.26) for updating $p_{\mathcal{L}(\mathcal{A})|y}$. The latter provides key computational benefits also when sampling from the full conditional of the cluster assignments described in Algorithm 6.

Algorithm 6 Part II of the Gibbs sampler for joint modeling of mixed domain data

---

Conditionally on samples for the quantities in equations (3.23), (3.24) and (3.25), update the cluster assignments $(y_1, \dots, y_n)$.

---

[5] Sample cluster assignments $(y_1, \dots, y_n)$ via sequential re-seating

for $i = 1, \dots, n$ **do**

Update each y_i conditionally on $y_{-i} = (y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_n)$

1. Remove agency i since we are going to sample its cluster membership y_i.

2. If no other agencies are in the same cluster of i, this cluster becomes empty and is removed along with its associated mono-product portfolio structure and co-subscription network probability mass function.

3. Re-order cluster indices so that $1, \ldots, K_{-i}$ are non-empty.

4. Update the cluster of i from the full conditional categorical variable with cluster probabilities

$$\text{pr}(y_i = y \mid -) \propto \begin{cases} \frac{n_{y,-i}}{n-1+\alpha_c} \text{pr}\left\{d_i, \mathcal{L}(A_i), G_i \mid y_i = y, p_{\mathcal{D}|y}, \nu_y, \pi^{(G_i)}\right\} & \text{for } y \leq K_{-i}, \\ \frac{\alpha_c}{n-1+\alpha_c} \text{pr}\left\{d_i, \mathcal{L}(A_i), G_i \mid y_i = K_{-i}+1, \pi^{(G_i)}\right\} & \text{for } y = K_{-i}+1, \end{cases} \tag{3.30}$$

5. If i is assigned a new cluster $K_{-i}+1$, add a new cluster and sample a new $\nu_{K_{-i}+1}$ and $p_{\mathcal{D}|K_{-i}+1}$ conditionally on G_i and d_i according to steps **[2]** and **[4]**, respectively.

end for

To perform steps in Algorithm 6 one needs to compute conditional probabilities in equation (3.30) at each MCMC iteration. Although this is apparently a cumbersome task, our model formulation (3.23)–(3.25) along with its hierarchical representation in (3.26) allows key simplifications, substantially improving the computational tractability of our procedures. Specifically, under our model, the conditional probability $\text{pr}\left\{d_i, \mathcal{L}(A_i), G_i \mid y_i = y, p_{\mathcal{D}|y}, \nu_y, \pi^{(G_i)}\right\}$ can be factorized as

$$\text{pr}(d_i \mid y_i = y, p_{\mathcal{D}|y})\text{pr}(G_i \mid y_i = y, \nu_y)\text{pr}\{\mathcal{L}(A_i) \mid G_i, \pi^{(G_i)}\}. \tag{3.31}$$

According to (3.31) inducing cluster-dependence only through the mixing probabilities ν_y, while considering cluster-independent mixing components in (3.24)–(3.25), has the key benefit of maintaining $\text{pr}\{\mathcal{L}(A_i) \mid G_i, \pi^{(G_i)}\}$ constant across the clusters assignments. As a result

$$\begin{aligned} \text{pr}\left\{d_i, \mathcal{L}(A_i), G_i \mid y_i = y, p_{\mathcal{D}|y}, \nu_y, \pi^{(G_i)}\right\} &\propto \text{pr}(d_i, G_i \mid y_i = y, p_{\mathcal{D}|y}, \nu_y) = \\ &= \text{pr}(d_i \mid y_i = y, p_{\mathcal{D}|y})\text{pr}(G_i \mid y_i = y, \nu_y), \end{aligned}$$

for every $y = 1, \ldots K$, where both terms are multinomial likelihoods with independent Dirichlet priors for their class probabilities $p_{\mathcal{D}|y}$ and ν_y, respectively. This allows simple computation of the posterior probabilities for clusters assignments in (3.30) of the algorithm,

obtaining

$$\text{pr}(y_i = y|-) \propto \begin{cases} \frac{n_{y,-i}}{n-1+\alpha_c} \prod_{v=1}^{V} p_{\mathcal{D}|y}(v)^{n_{iv}} \prod_{h=1}^{H} \nu_{hy}^{1_{(h)}(G_i)} & \text{for } y \leq K_{-i}, \\ \frac{\alpha_c}{n-1+\alpha_c} \text{pr}(d_i|y_i = K_{-i}+1)\text{pr}(G_i|y_i = K_{-i}+1) & \text{for } y = K_{-i}+1, \end{cases} \tag{3.32}$$

where the two marginal likelihoods corresponding to a newly occupied cluster are easily available exploiting the multinomial-Dirichlet conjugacy. In particular it is easy to show that

$$\text{pr}(d_i|y_i = K_{-i}+1) = \int \prod_{v=1}^{V} p_{\mathcal{D}|K_{-i}+1}(v)^{n_{iv}} d\Pi(p_{\mathcal{D}|K_{-i}+1}) = \frac{\Gamma(\sum_{v=1}^{V} \alpha_v)}{\prod_{v=1}^{V} \Gamma(\alpha_v)} \frac{\prod_{v=1}^{V} \Gamma(\alpha_v + n_{iv})}{\Gamma\{\sum_{v=1}^{V}(\alpha_v + n_{iv})\}},$$

for the mono-product portfolio, and

$$\text{pr}(G_i|y_i = K_{-i}+1) = \int \prod_{h=1}^{H} \nu_{h,K_{-i}+1}^{1_{(h)}(G_i)} d\Pi(\nu_{K_{-i}+1}) = \frac{\Gamma(\sum_{h=1}^{H} 1/H)}{\prod_{h=1}^{H} \Gamma(1/H)} \frac{\prod_{h=1}^{H} \Gamma\{1/H + 1_{(h)}(G_i)\}}{\Gamma[\sum_{h=1}^{H}\{1/H + 1_{(h)}(G_i)\}]},$$

for the augmented indicator variable in the cluster-dependent mixture of low-rank factorizations.

Hence, considering only cluster dependence in the mixture probabilities ν_y, $y = 1, \ldots, K$ and exploiting the augmented data G_i, $i = 1, \ldots, n$ in the mixture representation for $p_{\mathcal{L}(\mathcal{A})|y}$, $y = 1, \ldots, K$ allows a massive gain in computational tractability for step **[5]** in Algorithm 6. In fact, while $\text{pr}(d_i \mid y_i = K_{-i} + 1)$ and $\text{pr}(G_i \mid y_i = K_{-i} + 1)$ can be easily derived in closed form, the marginal likelihood of the multi-product networks with respect to the edge probability vectors arising from the low-rank factorization construction in (3.25) is not analytically available. According to factorization (3.31), this quantity can be avoided in step **[5]** as it doesn't change across clusters.

3.2.4 Simulation study

We consider a simulation study to evaluate the performance of our model in accurately recovering clusters of agencies and in efficiently estimating the key quantities required to define the set of cross-selling strategies for each group and their associated performance indicators. In simulating data, we look for a scenario possibly mimicking the structure of our application or related problems.

According to these aims we focus on $n = 200$ agencies equally divided in $K = 4$ latent clusters and consider a total number of $V = 15$ products as in our application. Graphical analyses of our data – highlighted in Figure 1.5 – show that mono-product customers typically concentrate on a small subset of the available products with high probability, while choosing the remaining set with very low frequency. We maintain this behavior in constructing $p^0_{\mathcal{D}|y}$,

$y = 1, \ldots, 4$, while additionally looking for a challenging scenario with small changes in $p^0_{\mathcal{D}|y}$ across clusters. This is obtained by letting $p^0_{\mathcal{D}|1}(v)$ and $p^0_{\mathcal{D}|2}(v)$ equal for all products except for permuting 1 and 9, so as $p^0_{\mathcal{D}|1}(1) = p^0_{\mathcal{D}|2}(9)$ and $p^0_{\mathcal{D}|1}(9) = p^0_{\mathcal{D}|2}(1)$. We adopt a similar strategy for clusters $y = 3$ and $y = 4$ by considering $p^0_{\mathcal{D}|3}(v) = p^0_{\mathcal{D}|4}(v)$ for all products v except 3 and 7, where we let $p^0_{\mathcal{D}|3}(3) = p^0_{\mathcal{D}|4}(7)$ and $p^0_{\mathcal{D}|3}(7) = p^0_{\mathcal{D}|4}(3)$. Based on the previous $p^0_{\mathcal{D}|y}$, $y = 1, \ldots, 4$ we simulate mono-product subscription data d_{ij}, $i = 1, \ldots, 200$ and $j = 1, \ldots, 500$ from the discrete random variable with probability mass function $p^0_{\mathcal{D}|y}$ where $y = 1$ for agencies $i = 1, \ldots, 50$, $y = 2$ for $i = 51, \ldots, 100$, $y = 3$ for $i = 101, \ldots, 150$ and $y = 4$ for $i = 151, \ldots, 200$. Although agencies in our application have at least $\approx 1{,}000$ mono-product customers, we consider a smaller number $n_i = 500$ for each agency $i = 1, \ldots, 200$ to evaluate the performance of the model when there is less information in the data.

Co-subscription networks are simulated exploiting the constructive representation (3.26) of the mixture model in (3.24). We consider $H = 3$ mixture components with each edge probability matrix $\pi^{0(h)}$ generated to mimic a possible co-subscription scenario. Specifically $\pi^{0(1)}$ is characterized by one dense community among 10 possibly highly related products, while assigning low probability to the remaining pairs of products. Matrix $\pi^{0(2)}$ represents the case of 4 hub products which occur with high probability in consumer multiple choices and fix the remaining co-subscription probabilities at low values. Finally, to reduce separation among mixture components and provide a more challenging scenario, matrix $\pi^{0(3)}$ is very similar to $\pi^{0(2)}$ with exception of product $v = 4$ which is held out from the hub products. In avoiding the low-rank factorization construction (3.25) in the definition of $\pi^{0(h)}$, $h = 1, \ldots, 3$, we additionally aim to evaluate the performance of factorization (3.25) in accurately characterizing the co-subscription probability matrices for each class h.

In simulating networks A_i, $i = 1, \ldots, 200$ from the hierarchical representation in (3.26), we consider cluster-dependent mixing probabilities $\nu^0_1 = \nu^0_2 = (0.9, 0.05, 0.05)$, $\nu^0_3 = (0.05, 0.9, 0.05)$ and $\nu^0_4 = (0.05, 0.05, 0.9)$. This choice allows the first co-subscription scenario defined by $\pi^{0(1)}$ to be very likely in agencies belonging to clusters $y = 1$ and $y = 2$. Scenarios characterized by $\pi^{0(2)}$ and $\pi^{0(3)}$ are instead more likely in clusters $y = 3$ and $y = 4$, respectively. Note that letting $\nu^0_1 = \nu^0_2$ further reduces separation among clusters $y = 1$ and $y = 2$. According to our simulation these two clusters have very similar mono-product choices according to $p^0_{\mathcal{D}|y}$ and equal generative process for the co-subscription networks $p^0_{\mathcal{L}(\mathcal{A})|1} = p^0_{\mathcal{L}(\mathcal{A})|2}$, providing an appealing scenario to evaluate the clustering performance of our model.

We analyze the simulated data under our model (3.23)–(3.25), considering the previously specified priors. As in Section 3.1.6 , we set $a_1 = 2.5$, $a_2 = 3.5$ and $\sigma^2_l = 10$, $l = 1, \ldots, V(V-1)/2$. Quantities μ_l, $l = 1, \ldots, V(V-1)/2$ are instead defined as $\mu_l = \text{logit}\{\sum_{i=1}^n \mathcal{L}(A_i)_l/n\}$ in order to center the mixture representation for $p_{\mathcal{L}(\mathcal{A})|y}$ around a co-subscription structure shared by all the simulated agencies in the company. We adopt a similar strategy for the

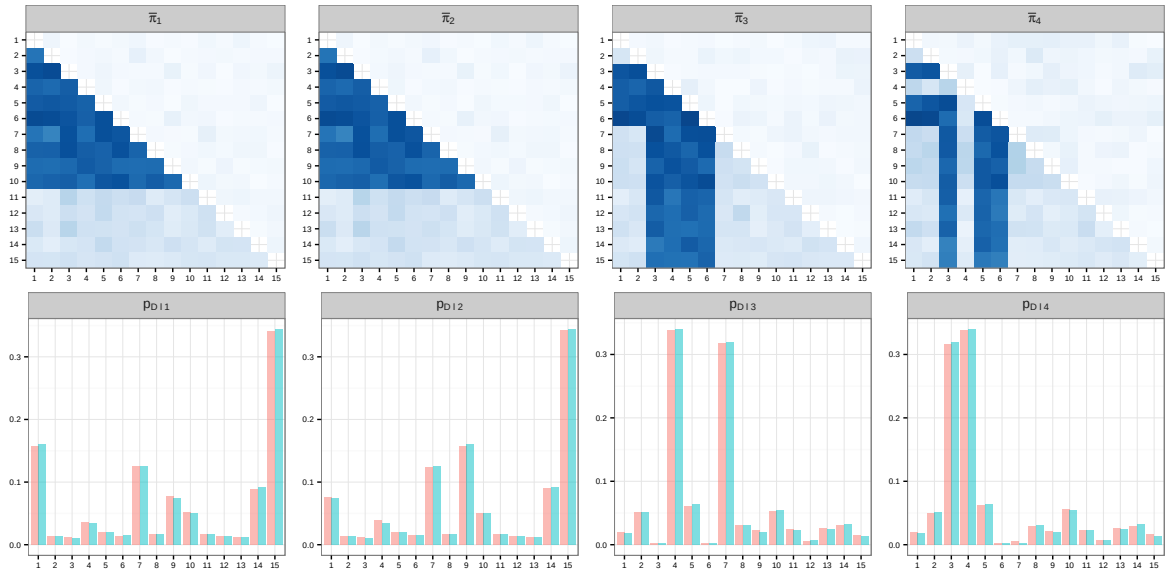


FIGURE 3.22: For the four non-empty clusters. Upper panel: posterior mean $\hat{\bar{\pi}}_y$ of the co-subscription probabilities among pairs of products in each cluster $y = 1, \dots, 4$ (lower triangular) and absolute value of the difference with the truth $|\hat{\bar{\pi}}_y - \bar{\pi}^0_y|$, $y = 1, \dots, 4$ (upper triangular). Color goes from white to dark blue as the probability goes from 0 to 1. Lower panels: posterior mean $\hat{p}_{\mathcal{D}|y}$ of the mono-product choices in each cluster $y = 1, \dots, 4$ (red) and true $p^0_{\mathcal{D}|y}$, $y = 1, \dots, 4$ (green).

hyperparameters of the Dirichlet prior (3.29) for the mono-product portfolio by setting $\alpha_v = \sum_{i=1}^{n} n_{iv}/n$ for each $v = 1, \dots, V$, in order to center (3.29) around an averaged portfolio for the entire company. This choice allows a more reasonable characterization of the marginal likelihood for mono-product data in (3.32), rather than considering a sparse and symmetric Dirichlet which may fail in adding further clusters, when required. Finally we set the concentration parameter $\alpha_c = 1$ in the CRP prior for the cluster assignments, according to standard practice. Although this parameter could be learned from our data as in Escobar and West (1995), we found results robust to moderate changes in α_c according to sensitivity analyses.

We perform posterior inference considering $5{,}000$ Gibbs iterations and set $H = 15$ and $R = 10$ as upper bounds for the number of mixture components and the dimension of the latent spaces, respectively. These upper bounds provide a good choice, with the sparse Dirichlet prior for the mixing probabilities ν_y, $y = 1, \dots, K$ and the multiplicative inverse gamma for the weights $\lambda^{(h)}$, $h = 1, \dots, H$ adaptively removing redundant components. Potential scale reduction factors for the quantities considered for inference suggest convergence is reached after a burn-in of $1{,}000$ and mixing is very good in our experience. As our inference focuses on cluster-specific structures it is additionally important to first check for label switching issues and relabel the clusters at each MCMC iteration using for example Stephens (2000) in case such issue is encountered. Traceplots suggest label switching isn't an issue in our simulation.

In providing inference on cluster assignments we initialize our algorithm by considering all agencies in a single group corresponding to a unique set of cross-selling strategies common to all agencies, and then assign agencies to clusters via maximum a posteriori estimates

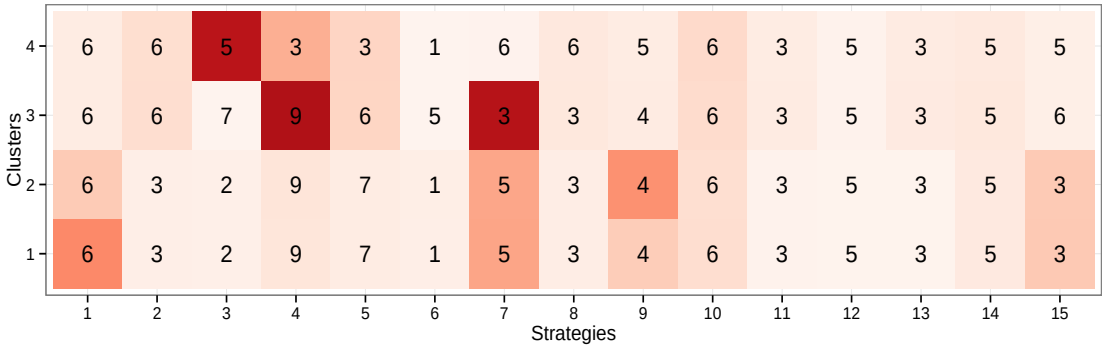


FIGURE 3.23: Plot of the estimated cluster-specific cross-selling strategies $\hat{q}_{y1}, \ldots, \hat{q}_{yV}$ along with their performance indicators $\hat{e}_{y1}, \ldots, \hat{e}_{yV}$. In particular each cell $[y, v]$ of the matrix defines the cross-selling strategy for mono-product customers subscribed to product v in agencies belonging to cluster y. The number in such cell corresponds to the best offer $u = \text{argmax}_u\{\hat{\text{pr}}(\mathcal{A}_{[vu]} = 1 \mid y) : u \neq v\}$ according to $\hat{q}_{yv}$, while the color is proportional to the corresponding estimated performance $\hat{e}_{yv} = \hat{p}_{\mathcal{D}|y}(v)\text{max}\{\hat{\text{pr}}(\mathcal{A}_{[vu]} = 1 \mid y) : u \neq v\}$. The estimated probability of a co-subscription for each pair of products v and u for agencies in cluster y, $\hat{\text{pr}}(\mathcal{A}_{[vu]} = 1 \mid y)$, is easily available from estimates via $\hat{\bar{\pi}}_{yl}$ for each $y = 1, \ldots, K$ and $l = 1, \ldots, V(V-1)/2$.

of $y_1, \ldots, y_n$ based on the MCMC samples. Flexible modeling of the cluster-specific mono-product portfolios and co-subscription networks along with the CRP prior for the cluster memberships, allow us to identify the $K = 4$ true clusters in our data and correctly group all the simulated agencies, including those in clusters $y = 1$ and $y = 2$. These groups are characterized by very subtle differences in their generating process.

Accurate clustering performance further allows for efficient estimation of the cluster-specific components required in defining the sets of cross-selling strategies. According to Figure 3.22 we correctly estimate the matrix of co-subscription probabilities among pairs of products $\bar{\pi}^0_y$ in each cluster $y = 1, \ldots, 4$ as well as the probability mass function $p^0_{\mathcal{D}|y}$ characterizing mono-product choices in each group $y = 1, \ldots, 4$. These results further highlight the flexibility of the low-rank factorizations in flexibly characterizing co-subscription probability matrices.

Previous quantities are a key to define the cluster-specific cross-selling strategies $q_{y1}, \ldots, q_{yV}$ and related performance indicators $e_{y1}, \ldots, e_{yV}$, as shown in Figure 3.23. Consistently with results in Figure 3.22, cross-selling strategies are the same in clusters 1 and 2 as $\hat{\bar{\pi}}_1 \approx \hat{\bar{\pi}}_2$, while performance indicators differ only for strategies targeting mono-product customers subscribed to $v = 1$ or $v = 9$; the first are more profitable in cluster $y = 1$ while the second in cluster $y = 2$. This is consistent with our estimates in Figure 3.22 highlighting $\hat{p}_{\mathcal{D}|1}(1) > \hat{p}_{\mathcal{D}|1}(9)$ and $\hat{p}_{\mathcal{D}|2}(1) < \hat{p}_{\mathcal{D}|2}(9)$. Mono-product customers subscribed to $v = 4$ and $v = 7$ are highly profitable in cluster $k = 3$ in being highly represented and having high co-subscription probability with at least one additional product. Customers subscribed to $v = 4$ are instead no more a segment worth targeting for cross-selling in cluster $y = 4$. Although $v = 4$ is highly populated, according to $\hat{\bar{\pi}}_4$ it is not possible to find any additional product u having high co-subscription probability with $v = 4$.

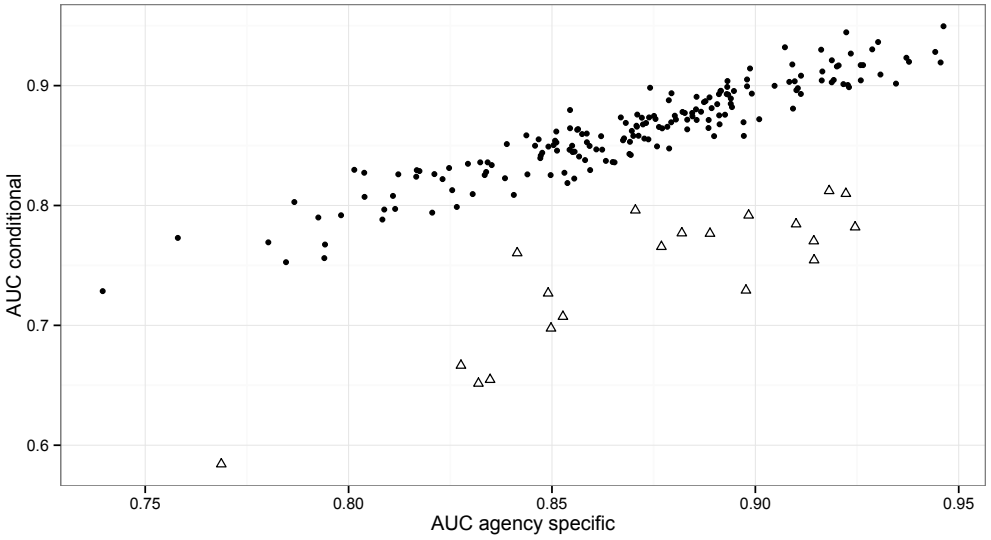


FIGURE 3.24: For each agency $i = 1, \ldots, n$ plot of the area under the ROC curve (AUC) when predicting its observed edges in the co-subscription network data $\mathcal{L}(A_i)_l$, $l = 1, \ldots, V(V-1)/2$ with the corresponding $\hat{\mathrm{E}}(\mathcal{L}(\mathcal{A}) \mid y_i = y)_l = \hat{\bar{\pi}}_{yl}$ versus the same quantity obtained replacing $\hat{\bar{\pi}}_{yl}$ with $\hat{\pi}_{il}$. Triangles represents agencies in which prediction via $\hat{\pi}_i$ provides an AUC which exceeds the one associated to $\hat{\bar{\pi}}_y$ by more than 0.05.

Figure 3.24 compares the prediction performance of the estimated cluster-specific vectors encoding co-subscription probabilities $\hat{\bar{\pi}}_y$, $y = 1, \ldots, 4$ with respect to the same quantities $\hat{\pi}_i$ specific to each agency, obtained from representation (3.26). According to results in Figure 3.24 most of the co-subscription networks are adequately characterized by the co-subscription probability vectors $\hat{\bar{\pi}}_y$, $y = 1, \ldots, 4$ specific to their clusters, with most of the AUC greater than 0.7 and the predictive performance not significantly improved when considering the more refined agency-specific co-subscriptions probabilities $\hat{\pi}_i$. More evident improvements are found for agencies represented by triangles. For such agencies, the company may devise ad-hoc cross-selling strategies based on their specific $\hat{\pi}_i$ rather than considering the same cross-selling advertising associated with the clusters they belong to.

To evaluate the fit with respect to the mono-product portfolios, we consider the standardized L_1 distance between observed and estimated product frequencies $\epsilon_{D_i} = \sum_{v=1}^{V} |n_{iv}/n_i - \hat{p}_{\mathcal{D}|y}(v)|/V$ with y denoting the cluster in which agency i is allocated. In our simulation the maximum of these quantities is $\max(\epsilon_{D_1}, \ldots, \epsilon_{D_n}) = 0.013$ meaning that mono-product data in each cluster $y = 1, \ldots, 4$ are adequately characterized by their cluster-specific estimate of $p_{\mathcal{D}|y}$.

3.2.5 Application to cross-selling marketing in an insurance company

We apply the model outlined in Section 3.2.1 to our motivating business intelligence dataset described in Section 1.2.2 which comprises mono-product choice data and co-subscription

networks for $n = 130$ agencies selling $V = 15$ different insurance products. Posterior computation is performed considering the same settings as in the simulation study. Also in this case we obtain convergence, good mixing performance and no issues of label switching according to traceplots and potential scale reduction factors for the quantities we consider in posterior analyses.

The posterior distribution for the cluster assignments suggests a total of $K = 18$ clusters in our data. This is already an appealing reduction of dimensionality in requiring the company to define $K = 18$ sets of cross-selling strategies $q_{y1}, \ldots, q_{yV}$, $y = 1, \ldots, 18$, rather than considering $n = 130$ different sets of campaigns $q_{i1}, \ldots, q_{iV}$, $i = 1, \ldots, 130$. According to posterior summaries in Figure 3.25 for the three mostly populated clusters which characterize the $\approx 50\%$ of the agencies, our procedures additionally allow a good joint representation of the different mono- and multi-product sources of variability in our data, while providing interesting insights on customer mono and multiple buying behavior with respect to insurance products. Posterior quartiles additionally highlight that the posterior distributions are efficiently concentrated around our estimates.

According to Figure 3.25, although mono- and multi-product customers of agencies in clusters $y = 2$ and $y = 3$ are characterized by a very similar behavior, our flexible procedure is able to capture subtle differences when comparing $\hat{p}_{\mathcal{D}|2}$ with $\hat{p}_{\mathcal{D}|3}$. Both groups have relatively high preferences for products $v = 1$ (house insurance) and $v = 2$ (car insurance), however the latter is slightly more populated in cluster $y = 3$ at the expense of house insurance policies. Correctly identifying differences in $p_{\mathcal{D}|y}$ is a key to evaluate and rank the cross-selling campaigns $q_{y1}, \ldots, q_{yV}$, $y = 1, \ldots, 18$ according to their performances indicators $e_{y1}, \ldots, e_{yV}$. Co-subscription probabilities are instead very similar in groups $y = 2$ and $y = 3$, which share an interesting community structure among products $v = 1, \ldots, 6$, while assigning low probability to the remaining pairs of products. Interestingly, these products refer to home insurance ($v = 1$), car insurance ($v = 2$), insurance on savings ($v = 3$), on investments ($v = 4$), retirement plans ($v = 5$) and insurance on injuries ($v = 6$), representing the polices mostly co-occurring in standard choices for families and individual consumers.

Cluster $y = 7$ is instead highly different than $y = 2$ and $y = 3$ in containing mono-product customers with high preferences for business activities insurance ($v = 7$) and multi-product customers characterized by a substantially different community structure in their co-subscription behavior. This community contains home insurance ($v = 1$), insurance on injuries ($v = 6$), business activities insurance ($v = 7$), payment protection insurance ($v = 8$), income protection insurance ($v = 9$) and liability insurance ($v = 10$). Hence, agencies in $y = 7$ are likely to deal mostly with business customers rather than families or individual consumers and require substantially different cross-selling strategies with respect to those associated with clusters $y = 2$ and $y = 3$.

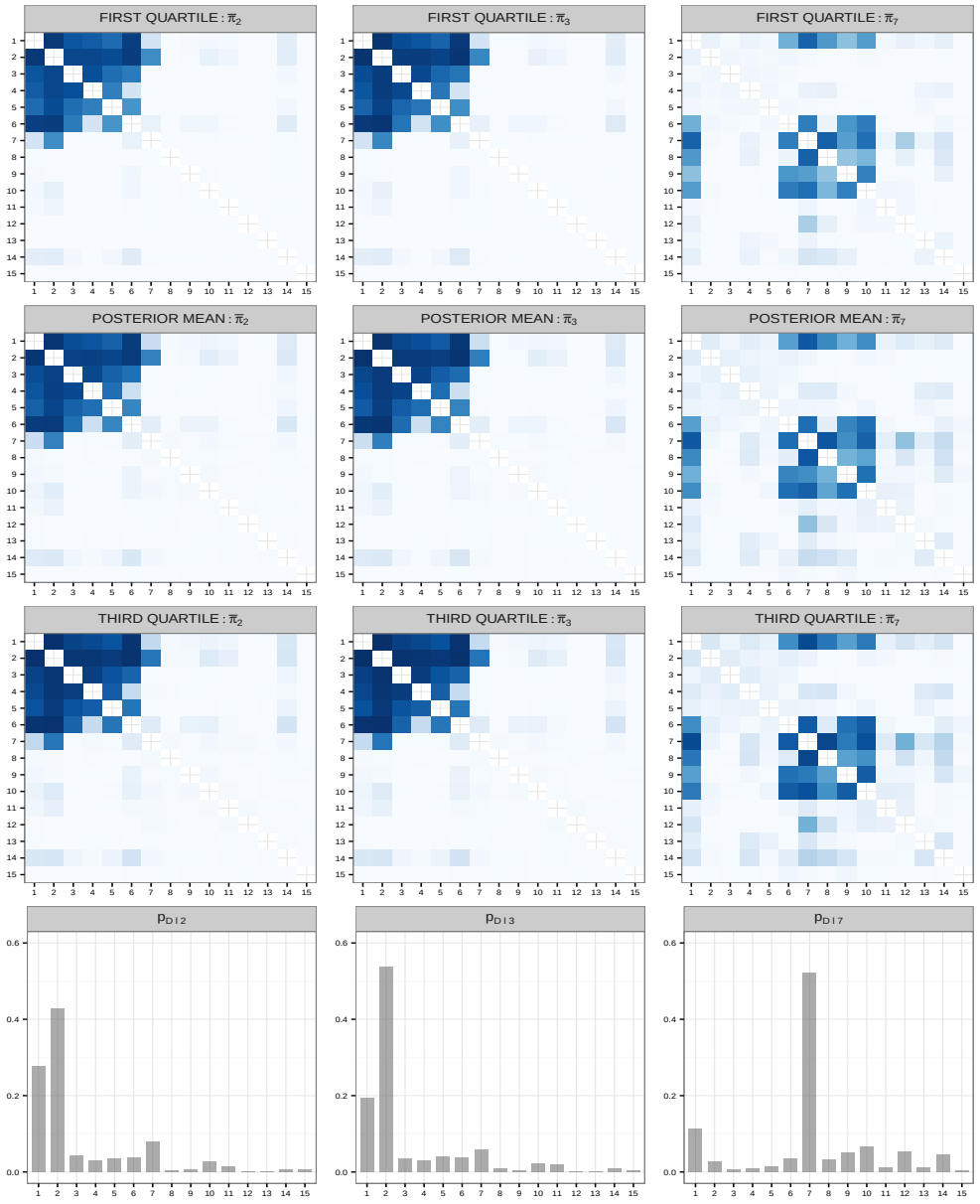


FIGURE 3.25: For the three mostly populated clusters. Summary of the posterior distribution of their co-subscription probabilities among pairs of products $\bar{\pi}_y$, rearranged in matrix form. Color goes from white to dark blue as the probability goes from 0 to 1. Lower panel: posterior mean $\hat{p}_{\mathcal{D}|y}$ of their mono-product choices.

The left matrix in Figure 3.26 provides a more general overview for the set of cross-selling strategies $\hat{q}_{y1}, \ldots, \hat{q}_{yV}$ in each cluster $y = 1, \ldots, 18$ estimated from $\hat{\bar{\pi}}_y$. Most of the clusters are characterized by similar cross-selling strategies closely related to the multiple buying behavior of families and individual consumers discussed for clusters $y = 2$ and $y = 3$ in Figure 3.25, with only slight differences in the co-subscription probabilities $\max\{\hat{\text{pr}}(\mathcal{A}_{[1u]} = 1 \mid y) : u \neq 1\}, \ldots, \max\{\hat{\text{pr}}(\mathcal{A}_{[Vu]} = 1 \mid y) : u \neq V\}$. Clusters $y = 18, 14, 12$ are instead more similar to $y = 7$ in having business-related insurance products, which co-occur more than family related ones, such as $v = 3, 4, 5$. Cluster $y = 9$ is finally characterized by a substantially different set of cross-selling strategies. In interpreting results for $y = 9$ it is worth noticing that

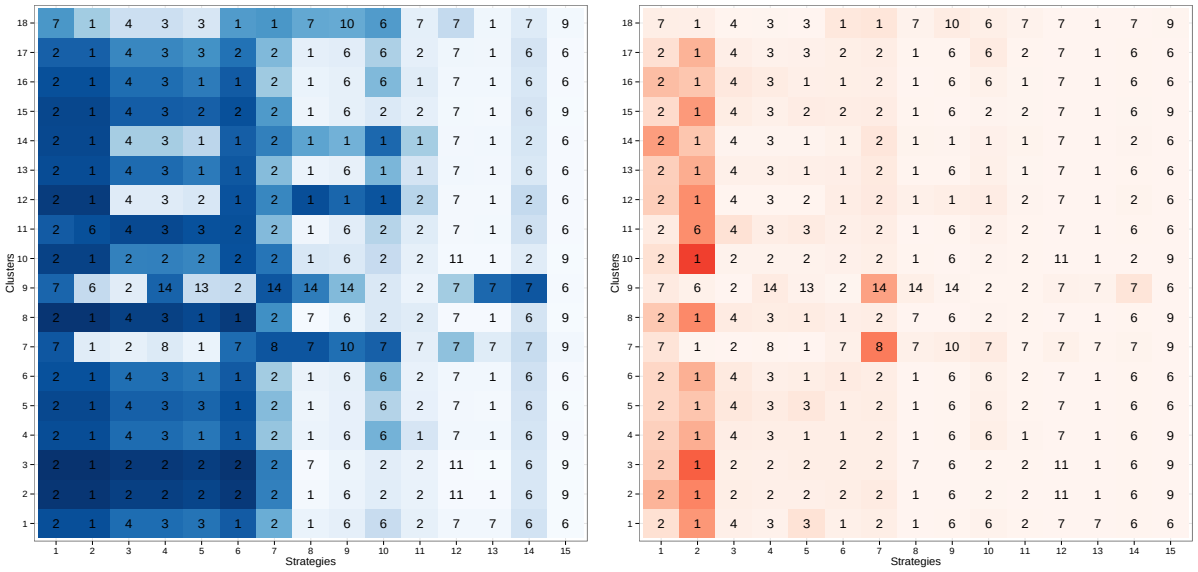


FIGURE 3.26: Plot of the estimated cluster-specific cross-selling strategies $\hat{q}_{y1}, \ldots, \hat{q}_{yV}$ along with their co-subscription probabilities $\max\{\hat{\text{pr}}(\mathcal{A}_{[1u]} = 1 \mid y) : u \neq 1\}, \ldots, \max\{\hat{\text{pr}}(\mathcal{A}_{[Vu]} = 1 \mid y) : u \neq V\}$ (left matrix) and performance indicators $\hat{e}_{y1}, \ldots, \hat{e}_{yV}$ (right matrix). In particular each cell $[y, v]$ of the matrix defines the cross-selling strategy for mono-product customers subscribed to product v in agencies belonging to cluster y. The number in such cell corresponds to the best offer $u = \text{argmax}_u\{\hat{\text{pr}}(\mathcal{A}_{[vu]} = 1 \mid y) : u \neq v\}$ according to $\hat{q}_{yv}$. The color of each cell in the left matrix is proportional to the corresponding estimated co-subscription probability $\max\{\hat{\text{pr}}(\mathcal{A}_{[vu]} = 1 \mid y) : u \neq v\}$. The color in the right matrix is proportional to the corresponding estimated performance $\hat{e}_{yv} = \hat{p}_{\mathcal{D}|y}(v)\max\{\hat{\text{pr}}(\mathcal{A}_{[vu]} = 1 \mid y) : u \neq v\}$. The estimated probability of a co-subscription for each pair of products v and u for agencies in cluster y, $\hat{\text{pr}}(\mathcal{A}_{[vu]} = 1 \mid y)$, is easily available from estimates via $\hat{\tilde{\pi}}_{yl}$ for each $y = 1, \ldots, K$ and $l = 1, \ldots, V(V-1)/2$.

basic coverage for medical expenses is guaranteed in Italy by public institutions, and health insurance policy ($v = 14$) provides further benefits in accessing health care. In fact $v = 14$ is rarely observed in the inferred cross-selling strategies with exception of cluster $y = 9$ where the health insurance policy ($v = 14$) and the business activities insurance ($v = 7$) co-occur with high probability in the co-subscription networks of agencies belonging to cluster $y = 9$. Hence, this group may refer to agencies dealing with high income business costumers which can afford additional expenses for an improved health care. This is further explicit after noticing that $v = 14$ co-occurs with high probability with insurance policies on investments ($v = 4$) in cluster $y = 9$.

When considering performance $\hat{e}_{y1} = \hat{p}_{\mathcal{D}|y}(1)\max\{\hat{\text{pr}}(\mathcal{A}_{[1u]} = 1 \mid y) : u \neq 1\}, \ldots, \hat{e}_{yV} = \hat{p}_{\mathcal{D}|y}(V)\max\{\hat{\text{pr}}(\mathcal{A}_{[Vu]} = 1 \mid y) : u \neq V\}$ in right matrix of Figure 3.26, we clearly notice how cross-selling strategies targeting mono-product customers with car insurance ($v = 2$) or home insurance ($v = 1$) are in general more effective in creating new multi-product customers. Beside being characterized by high co-occurence patterns with other polices, these products are also highly populated by mono-product costumers as home insurance ($v = 1$) represents a common policy for families and car insurance ($v = 2$) is compulsory in Italy.

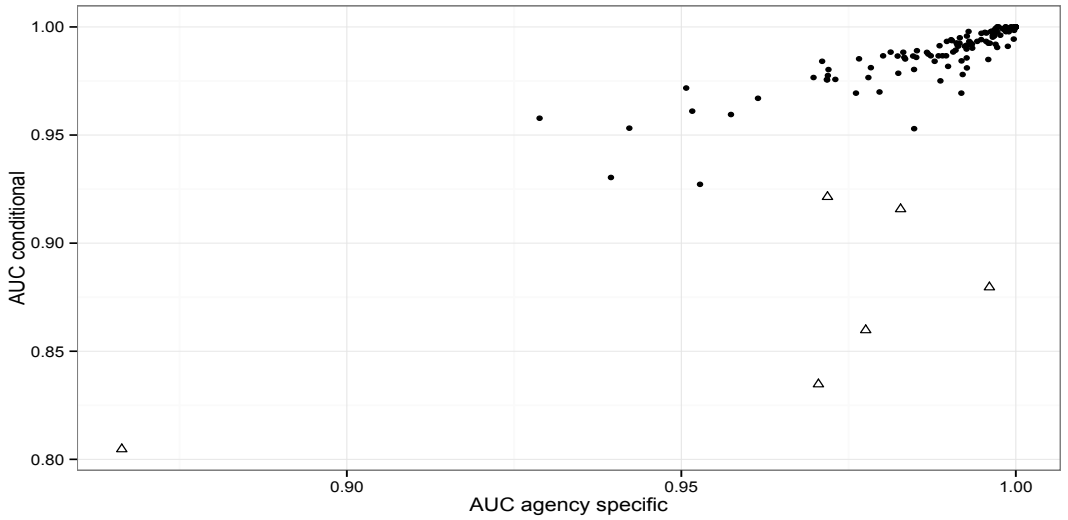


FIGURE 3.27: For each agency $i = 1, \dots, 130$ plot of the area under the ROC curve (AUC) when predicting its observed edges in the co-subscription network data $\mathcal{L}(A_i)_l$, $l = 1, \dots, V(V-1)/2$ with the corresponding $\hat{\mathrm{E}}(\mathcal{L}(\mathcal{A}) \mid y_i = y)_l = \hat{\bar{\pi}}_{yl}$ versus the same quantity obtained replacing $\hat{\bar{\pi}}_{yl}$ with $\hat{\pi}_{il}$. Triangles represents agencies in which prediction via $\hat{\pi}_i$ provides an AUC which exceeds the one associated to $\hat{\bar{\pi}}_y$ by more than 0.05.

Customers subscribed to business activities insurance ($v = 7$) are instead profitable for agencies in clusters $y = 7$ and $y = 9$. This segment of the customer base is in fact highly populated in $y = 7$ and $y = 9$, while having high co-subscription probability with $u = 8$ and $u = 14$, respectively.

Finally it is additionally interesting to notice how mono-product customers subscribed to $v = 3$ are associated with a cross-selling strategy $\hat{q}_{y3}$ offering $u = 4$ in almost all the clusters. The same is true in the reverse case. This is a consistent finding as $v = 3$ and $v = 4$ correspond to insurance on savings and insurance on investments, respectively, and hence such polices are reasonably related in customer multiple buying behavior.

According to results in Figure 3.27, we do a good job in characterizing the observed co-subscription networks $\mathcal{L}(A_i)$, $i = 1, \dots, 130$ considering the co-subscription probability vectors $\hat{\bar{\pi}}_y$, $y = 1, \dots, 18$ specific to their clusters. All the AUC are greater than 0.8 and the prediction performance is in general not significantly improved when considering the more refined agency-specific co-subscriptions probabilities $\hat{\pi}_i$, with exception of a few agencies represented by triangles. For such agencies, the company may devise cross-selling strategies based on their specific $\hat{\pi}_i$, but it is still reasonable to rely on strategies in Figure 3.26 based on the good performance associated to $\hat{\bar{\pi}}_y$, $y = 1, \dots, 18$.

Our estimates provide also a good fit with respect to the mono-product portfolios with the maximum of the standardized L_1 distances between observed and estimated product frequencies for each agency being $\max(\epsilon_{D_1}, \dots, \epsilon_{D_{130}}) = 0.041$.

Conclusion

Discussion

Network science is a stimulating field. It embraces several disciplines and provides increasingly complex data sets, new interlocutors as well as novel scientific questions. Within this framework, the main goal of statistical research is to constantly catch up with the ongoing changes and provide novel methods balancing the need for provably flexible formulations with the demand of tractable inference procedures.

This thesis starts from important statistical questions associated with new applied problems, to propose novel methods for Bayesian inference in complex network data. Taking inspiration from different methodologies – such as tensor decomposition, matrix factorization, functional data analysis and mixture modeling – we developed novel procedures to study dynamic networks and populations of networks, when edges are binary and undirected. A primary emphasis has been on carefully characterizing the statistical models and prior distributions to efficiently account for the different sources of information in our data, while providing tractable inference procedures and simple computational strategies guaranteeing ease of implementation.

Procedures developed in Chapters 2 and 3 incorporate network information by defining edge probabilities as a function of nodes coordinates in a latent space, with this shared dependence on a common set of latent positions, allowing characterization of a broad variety of network structures – as confirmed in simulations and highlighted in theoretical studies. This choice further facilitates scaling to moderately large V in requiring estimation of a smaller set of latent coordinates instead of direct modeling of $V(V-1)/2$ edge probabilities.

Methodologies developed in Chapter 2 further require incorporation of dynamic information on top of the network one. In Section 2.1 this is accomplished by allowing the latent coordinates to evolve in continuous time via Gaussian process priors, providing a general procedure for dynamic network inference with full support properties and simple strategies for posterior computation. Section 2.2 replaces Gaussian process priors with nested Gaussian processes to allow the smoothness level of the underlying trajectories to change across time

rather than being time-constant. This additionally allows scaling to larger time windows by leveraging state space models which reduce the Gaussian process computation burden from $O(n^3)$ to $O(n)$ and facilitates implementation of fast online updating and forecasting algorithms.

Although we evaluated the flexibility and the key benefits of our methods in different simulations and applications to time-varying international relationship data and face-to-face dynamic human interaction networks, the procedures developed in Chapter 2 have a broad range of possible applications – also outside the dynamic filed – when the interest is on learning changes of a network structure across a continuous variable – not necessarily time. Recalling the neuroscience applications outlined in Section 1.2.1, this is for example the case of brain networks collected for each subject along with continuous phenotypes such as intelligences scores.

Differently from procedures in Chapter 2, methodologies developed in Chapter 3 focus on replicated observations of the same network for different units, rather than time-varying relational structures. Hence – instead of incorporating dynamic information – the goal in Chapter 3 is to carefully borrow strength within each network and across different units to nonparametrically estimate the probability mass function of a network-valued random variable. In Section 3.1 this is accomplished by combining the latent space approach – to incorporate network information – with mixture models – to share information across different replicates. As highlighted in detailed simulations and theoretical studies, this model is unique in providing a flexible approximation to the population distribution of binary and undirected network-valued data, while representing a flexible and general building block to develop efficient testing methods for changes across the levels of categorical predictors. This is obtained by allowing the mixture probabilities to vary across groups, providing highly efficient and computationally tractable Bayesian global and local testing procedures that adjust intrinsically for multiple comparisons and are robust against issues arising from model misspecification. In allowing the network data to be appropriately analyzed as network-valued, these methods enable substantial improvements in accurately detecting group differences, isolating specific aspects of the network that vary across behavioral traits and neurological disorders, and enhancing performance of predictive models as outlined in the application to creative cognition data and Alzheimers disorders.

Contributions in Section 3.1 have great potentials beyond neuroscience. As highlighted in Section 3.2 the dependent mixture of low-rank factorizations provides a key to define joint models for flexible and computationally tractable statistical analyses of mixed domain data. Although Section 3.2 focuses on inference and co-clustering for business intelligence data, the procedures developed have a broad range of additional applications. Examples include efficient allocation of resources across health care or public services based on mono- and

multi-service citizens data in different cities or states. Similarly, our strategies can be useful in defining optimal task assignments based on mono and multi-task data from players in different teams.

Future directions of research

Although the procedures developed in this thesis take an initial step towards addressing important open questions associated with complex network data, there are several important areas for future research. The first – common to all the methods developed in this thesis – is to generalize our procedures to weighted network data. Presence or absence of a relationship among pairs of nodes is progressively replaced by a measure of strength in this relation, typically in the form of counts. Examples include event counts in international relationships data, number of contacts in face-to-face interaction data, fiber counts in connectomes and frequency of customers subscribed to pairs of products in business analysis.

Weighted networks contain potentially more information than binary ones, but statistical methods for analyzing these data are still on their infancy compared the most popular literature on binary networks. As discussed in this thesis, in the binary case, data consist of indicators of connections between each pair of nodes. Such data are essentially multivariate binary, with network-structured dependence. Incorporating information on weighted edges, data take the form of multivariate counts, again with network-structured dependence. There are subtleties involved in modeling of multivariate counts. It is common to incorporate latent variables in Poisson factor models (Dunson and Herring, 2004; Gopalan et al., 2014). However, as noted in Canale and Dunson (2011), there is a pitfall in such models due to the dual role of the latent variable component in controlling the degree of dependence and the magnitude of over-dispersion in the marginal distributions. Canale and Dunson (2011) address these issue via a rounded kernel method which improves flexibility in estimating the distribution of count variables, while allowing simple generalizations for dynamic inference (Canale and Dunson, 2013). We are currently adapting these procedures to define novel stochastic processes for dynamic networks of counts and develop the first nonparametric approach for estimating the population distribution of weighted networks of counts.

While modeling of weighted networks require new statistical models and theoretical justifications along with adapted testing procedures and new algorithms for posterior inference, it is substantially easier to generalize the methods for undirected networks developed in this thesis to directed ones. This can be accomplished by simply replacing the low-rank factorization mechanism based on eigen-decomposition procedures, with singular value factorizations of the latent similarities. Clearly in such directed cases it is necessarily to model the entire adjacency matrix, rather than only its lower triangular elements.

Beside future directions of research common to all the methods outlined, there are also several possible improvements specific to each topic addressed in this thesis. Dynamic networks are becoming increasingly multivariate. For example, using information from GDELT – described in Section 1.1.1 – one can potentially define a multi-layer dynamic network data set $A_{t_i}^{(k)}$, $t_i = t_1, \ldots, t_n$, $k = 1, \ldots, K$ with $A_{t_i[vu]}^{(k)} = A_{t_i[uv]}^{(k)}$ denoting presence or absence of a connection among countries v and u time t_i with respect to relationship type k. This leads to an highly complex $V \times V \times K \times n$ array for which flexible statistical methodologies still need to be developed. One possibility is to borrow information among different layers using the mixture models developed in Chapter 3, with the kernels defined by the stochastic processes developed in Chapter 2 to account for network and dynamic information.

Moving to connectome data considered in Section 3.1, it is important to develop statistical methods that explicitly take into account errors in constructing the structural brain connection network, including in alignment and in recovering fiber tracts, taking as input the raw data collected from the imaging machines. This represents a substantial computational hurdle, but may yield improvements in performance including better uncertainty quantification. An additional aspect is taking into account network structure other than a simple measure of the number of fibers connecting regions – for example, information on volume and relative spatial locations in the brain could also be incorporated. Additionally, scaling to massive networks is a key issue to deal with the high spatial resolution provided by modern imaging technologies. In the absence of careful modifications our computational algorithms fail in scaling to very large nodes sets V. Developing models that exploit sparsity in the network, or avoid sampling through efficient optimization algorithms, provide promising directions. In such settings, it is additionally of interest to develop further theoretical results to assess asymptotic properties of the posterior distribution for $p_{\mathcal{L}(\mathcal{A})}$ as the cardinality V of the node set increases with n. One possibility is to adapt recent Bayesian nonparametric asymptotic theory for multivariate categorical data with increasing number of variables (Zhou et al., 2014) to our specific setting.

Finally, focusing on methodologies for inference and co-clustering of mixed domain data in Section 3.2, our procedures develop and evaluate targeted cross-selling strategies aimed at stimulating customers multiple buying behavior in different agencies. It is worth considering further research to formally enter additional information – such as costs of the strategies and product prices – in a carefully defined loss function and define strategies within a Bayesian framework via minimization of expected posterior loss.

Bibliography

Abramowitz, M. and Stegun, I. A. (1964). *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*. Number 55. Courier Corporation.

Agresti, A. (2002). *Categorical data analysis*. Second edition. Wiley.

Airoldi, E. M., Blei, D. M., Fienberg, S. E. and Xing, E. P. (2008). Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9:1981–2014.

Aldous, D. (1985). Exchangeability and related topics. In Hennequin, P., editor, *École d'Été de Probabilités de Saint-Flour XIII 1983*, volume 1117 of *Lecture Notes in Mathematics*, pages 1–198. Springer Berlin Heidelberg.

Arden, R., Chavez, R. S., Grazioplene, R. and Jung, R. E. (2010). Neuroimaging creativity: A psychometric view. *Behavioural Brain Research*, 214(2):143–156.

Armagan, A., Dunson, D. B., Lee, J., Bajwa, W. U. and Strawn, N. (2013). Posterior consistency in linear models under shrinkage priors. *Biometrika*, 100(4):1011–1018.

Azari, N., Rapoport, S., Grady, C., Schapiro, M., Salerno, J., Gonzales-Aviles, A. and Horwitz, B. (1992). Patterns of interregional correlations of cerebral glucose metabolic rates in patients with dementia of the alzheimer type. *Neurodegeneration*, 1:101–111.

Azzalini, A. and Scarpa, B. (2012). *Data analysis and data mining: An introduction*. Oxford University Press.

Banerjee, A., Murray, J. and Dunson, D. (2013). Bayesian learning of joint distributions of objects. *Journal of Machine Learning, Workshops & Proceedings*, 31:1–9.

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509–512.

Barrat, A. and Cattuto, C. (2013). Temporal networks of face-to-face human interactions. In *Understanding Complex Systems*, pages 191–216. Springer Science Business Media.

Bassett, D. S. and Bullmore, E. (2006). Small-world brain networks. *The Neuroscientist*, 12(6):512–523.

Begg, M. D. and Lagakos, S. (1990). On the consequences of model misspecification in logistic regression. *Environmental Health Perspectives*, 87:69.

Belkin, P., Nelson, R.M. Mix, D. and Weiss, M. (2012). The eurozone crisis: Overview and issues for congress. Technical Report R42377, Congressional Research Service.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

Berestycki, H., Nadal, J. P. and Rodíguez, N. (2015). A model of riots dynamics: Shocks, diffusion and thresholds. *Networks and Heterogeneous Media*, 10(3):443–475.

Berger, J. O. and Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, 2(3):317–335.

Berger, J. O. and Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American Statistical Association*, 82(397):112–122.

Bernanke, B. S. (2007). Global imbalances: recent developments and prospects. Speech 317, Board of Governors of the Federal Reserve System, U.S.

Bhattacharya, A. and Dunson, D. B. (2011). Sparse Bayesian infinite factor models. *Biometrika*, 98(2):291–306.

Bigelow, J. L. and Dunson, D. B. (2009). Bayesian semiparametric joint models for functional predictors. *Journal of the American Statistical Association*, 104(485):26–36.

Blesa, R., Mohr, E., Miletich, R. and Chase, T. (1995). Limbic system dysfunction in alzheimer's disease. *Journal of neurology, neurosurgery, and psychiatry*, 59(4):450.

Bokde, A., Lopez-Bayo, P., Meindl, T., Pechler, S., Born, C., Faltraco, F., Teipel, S., Möller, H.-J. and Hampel, H. (2006). Functional connectivity of the fusiform gyrus during a face-matching task in subjects with mild cognitive impairment. *Brain*, 129(5):1113–1124.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. Wiley-Blackwell.

Börner, K., Sanyal, S. and Vespignani, A. (2007). Network science. *Annual Review of Information Science and Technology*, 41(1):537–607.

Bowman, F. D., Caffo, B., Bassett, S. S. and Kilts, C. (2008). A Bayesian hierarchical framework for spatial modeling of fmri data. *NeuroImage*, 39(1):146–156.

Brandt, P. T., Freeman, J. R., Lin, T. and Schrodt, P. A. (2013). Forecasting conflict in the cross-straits: Long term and short term predictions. In *Annual Meeting of the American Political Science Association*.

Bressler, S. L. and Menon, V. (2010). Large-scale brain networks in cognition: Emerging methods and principles. *Trends in Cognitive Sciences*, 14(6):277–290.

Brunnermeier, M. K. (2009). Deciphering the liquidity and credit crunch 2007–2008. *Journal of Economic Perspectives*, 23(1):77–100.

Bullmore, E. and Sporns, O. (2009). Complex brain networks: Graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3):186–198.

Bullmore, E. and Sporns, O. (2012). The economy of brain network organization. *Nature Reviews Neuroscience*, 77(3):586–595.

Butts, C. T. (2008). A relational event framework for social action. *Sociological Methodology*, 38(1):155–200.

Canale, A. and Dunson, D. B. (2011). Bayesian kernel mixtures for counts. *Journal of the American Statistical Association*, 106(496):1528–1539.

Canale, A. and Dunson, D. B. (2013). Nonparametric Bayes modelling of count processes. *Biometrika*, 100(4):801–816.

Cardillo, A., Gómez-Gardeñes, J., Zanin, M., Romance, M., Papo, D., del Pozo, F. and Boccaletti, S. (2013). Emergence of network features from multiplexity. *Scientific Reports*, 3.

Carlsson, I., Wendt, P. E. and Risberg, J. (2000). On the neurobiology of creativity. Differences in frontal activity between high and low creative subjects. *Neuropsychologia*, 38(6):873–885.

Cattuto, C., den Broeck, W. V., Barrat, A., Colizza, V., Pinton, J.-F. and Vespignani, A. (2010). Dynamics of person-to-person interactions from distributed RFID sensor networks. *PLoS ONE*, 5(7):e11596.

Chatterjee, S. and Diaconis, P. (2013). Estimating and understanding exponential random graph models. *The Annals of Statistics*, 41(5):2428–2461.

Chatterjee, S., Diaconis, P. and Sly, A. (2011). Random graphs with a given degree sequence. *Annals of Applied Probability*, 21(4):1400–1435.

Choi, H. M. and Hobert, J. P. (2013). The Pòlya-gamma Gibbs sampler for Bayesian logistic regression is uniformly ergodic. *Electronic Journal of Statistics*, 7:2054–2064.

Clarke, S. and Hall, P. (2009). Robustness of multiple testing procedures against dependence. *The Annals of Statistics*, 37(1):332–358.

Craddock, R. C., Jbabdi, S., Yan, C.-G., Vogelstein, J. T., Castellanos, F. X., Martino, A. D., Kelly, C., Heberlein, K., Colcombe, S. and Milham, M. P. (2013). Imaging human connectomes at the macroscale. *Nature Methods*, 10(6):524–539.

Dai, B., Ding, S. and Wahba, G. (2013). Multivariate bernoulli distribution. *Bernoulli*, 19(4):1465–1483.

Daianu, M., Jahanshad, N., Nir, T. M., Toga, A. W., Jack, C. R., Weiner, M. W. and Thompson, P. M. (2013). Breakdown of brain connectivity between normal aging and Alzheimer's disease: A structural k-core network analysis. *Brain Connectivity*, 3(4):407–422.

De Domenico, M., Nicosia, V., Arenas, A. and Latora, V. (2015). Structural reducibility of multilayer networks. *Nature Communications*, 6:6864.

De Domenico, M., Sole-Ribalta, A., Gomez, S. and Arenas, A. (2014). Navigability of interconnected networks under random failures. *Proceedings of the National Academy of Sciences*, 111(23):8351–8356.

Deegan, J. (1976). The consequences of model misspecification in regression analysis. *Multivariate Behavioral Research*, 11(2):237–248.

Deoni, S., Correia, S., Su, T., Man, J., Lehman, K., Malloy, P. and Salloway, S. (2011). Investigating limbic system myelin alteration in Alzheimer's disease. *Alzheimer's & Dementia*, 7(4):S57–S58.

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., Albert, M. S. and Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *NeuroImage*, 31(3):968–980.

Desmarais, B. A. and Cranmer, S. J. (2012). Statistical mechanics of networks: Estimation and uncertainty. *Physica A: Statistical Mechanics and its Applications*, 391(4):1865–1876.

DiRienzo, A. G. and Lagakos, S. W. (2001). Effects of model misspecification on tests of no randomized treatment effect arising from coxs proportional hazards model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(4):745–757.

DuBois, C., Butts, C. T., McFarland, D. and Smyth, P. (2013). Hierarchical models for relational event sequences. *Journal of Mathematical Psychology*, 57(6):297–309.

Dunson, D. B. and Herring, A. H. (2004). Bayesian latent variable models for mixed discrete outcomes. *Biostatistics*, 6(1):11–25.

Dunson, D. B., Herring, A. H. and Siega-Riz, A. M. (2008). Bayesian inference on changes in response densities over predictor clusters. *Journal of the American Statistical Association*, 103(484):1508–1517.

Dunson, D. B. and Xing, C. (2009). Nonparametric Bayes modeling of multivariate categorical data. *Journal of the American Statistical Association*, 104(487):1042–1051.

Durante, D., Scarpa, B. and Dunson, D. B. (2014). Locally adaptive factor processes for multivariate time series. *Journal of Machine Learning Research*, 15:1493–1522.

Durbin, J. and Koopman, S. J. (2002). A simple and efficient simulation smoother for state space time series analysis. *Biometrika*, 89(3):603–616.

Durbin, J. and Koopman, S. J. (2012). *Time Series Analysis by State Space Methods*. Oxford University Press.

Erdös, P. and Rényi, A. (1959). On random graphs. *Publicationes Mathematicae Debrecen*, (5):290–297.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588.

Eskildsen, S. F., Coupé, P., Fonov, V. S., Pruessner, J. C., Collins, D. L., Initiative, A. D. N., et al. (2015). Structural imaging biomarkers of Alzheimer's disease: Predicting disease progression. *Neurobiology of aging*, 36:S23–S31.

Eun, C. S. and Resnick, B. G. (2010). *International Financial Management*. Tata McGraw-Hill Education.

Faust, K. and Wasserman, S. (1992). Blockmodels: Interpretation and evaluation. *Social Networks*, 14(1):5–61.

Fornito, A., Zalesky, A. and Breakspear, M. (2013). Graph analysis of the human connectome: Promise, progress, and pitfalls. *NeuroImage*, 80:426–444.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3-5):75–174.

Foulds, J., Dubois, C., Asuncion, A. U., Butts, C. T. and Smyth, P. (2011). A dynamic relational infinite feature model for longitudinal social networks. *Journal of Machine Learning, Workshops & Proceedings*, 15:287–295.

Fournet, J. and Barrat, A. (2014). Contact patterns among high school students. *PLoS ONE*, 9(9):e107878.

Frank, O. and Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association*, 81(395):832–842.

Friedman, E. J., Young, K., Asif, D., Jutla, I., Liang, M., Wilson, S., Landsberg, A. S. and Schuff, N. (2014). Directed progression brain networks in alzheimer's disease: Properties and classification. *Brain connectivity*, 4(5):384–393.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–67.

Fruchterman, T. M. and Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and Experience*, 21(11):1129–1164.

Fuster, J. M. (2000). The module: crisis of a paradigm. *Neuron*, 26(1):51–53.

Fuster, J. M. (2006). The cognit: a network model of cortical representation. *International Journal of Psychophysiology*, 60(2):125–132.

Gao, J., Leetaru, K. H., Hu, J., Cioffi-Revilla, C. and Schrodt, P. (2013). Massive media event data analysis to assess world-wide political conflict and instability. In *Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 284–292.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Aki, V. and Rubin, D. B. (2014). *Bayesian data analysis*. Taylor & Francis.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472.

Gemmetto, V., Barrat, A. and Cattuto, C. (2014). Mitigation of infectious disease at school: targeted class closure vs school closure. *BMC Infectious Disease*, 14:695.

Genovese, C. R., Lazar, N. A. and Nichols, T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage*, 15(4):870–878.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

Gershman, S. J. and Blei, D. M. (2012). A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*, 56(1):1–12.

Ghosal, S. (2000). Asymptotic normality of posterior distributions for exponential families when the number of parameters tends to infinity. *Journal of Multivariate Analysis*, 74(1):49–68.

Ghosal, S. and Belitser, E. (2003). Adaptive Bayesian inference on the mean of an infinite-dimensional normal distribution. *The Annals of Statistics*, 31(2):536–559.

Ghosh, J. and Dunson, D. B. (2009). Default prior distributions and efficient posterior computation in Bayesian factor analysis. *Journal of Computational and Graphical Statistics*, 18(2):306–320.

Ginestet, C. E., Balanchandran, P., Rosenberg, S. and Kolaczyk, E. D. (2014). Hypothesis testing for network data in functional neuroimaging. *arXiv:1407.5525*.

Girvan, M. and Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826.

Golby, A., Silverberg, G., Race, E., Gabrieli, S., O'Shea, J., Knierim, K., Stebbins, G. and Gabrieli, J. (2005). Memory encoding in Alzheimer's disease: An fmri study of explicit and implicit memory. *Brain*, 128(4):773–787.

Goldenberg, A., Fienberg, S. E., Zheng, A. X. and Airoldi, E. M. (2009). A survey of statistical network models. *Foundations and Trends in Machine Learning*, 2(2):129–233.

Gollini, I. and Murphy, T. B. (2013). Joint modelling of multiple network views. *Journal of Computational and Graphical Statistics*, to appear.

Gopalan, P., Ruiz, F. J., Ranganath, R. and Blei, D. M. (2014). Bayesian nonparametric poisson factorization for recommendation systems. *Journal of Machine Learning, Workshops & Proceedings*, 33:275–283.

Górski, A. Z., Drożdż, S. and Kwapień, J. (2008). Scale free effects in world currency exchange network. *The European Physical Journal B*, 66(1):91–96.

Griffiths, T. L. and Ghahramani, Z. (2011). The indian buffet process: An introduction and review. *The Journal of Machine Learning Research*, 12:1185–1224.

Grund, T. U. (2012). Network structure and team performance: The case of english premier league soccer teams. *Social Networks*, 34(4):682–690.

Hagmann, P., Cammoun, L., Gigandet, X., Meuli, R., Honey, C. J., Wedeen, V. J. and Sporns, O. (2008). Mapping the structural core of human cerebral cortex. *PLoS Biology*, 6(7):e159.

Handcock, M. (2003). Assessing degeneracy in statistical models of social networks. Technical Report 39, Center for Statistics and the Social Sciences University of Washington.

Hanneke, S., Fu, W. and Xing, E. P. (2010). Discrete temporal models of social networks. *Electronic Journal of Statistics*, 4:585–605.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman & Hall.

Heilman, K. M., Nadeau, S. E. and Beversdorf, D. O. (2003). Creative innovation: Possible brain mechanisms. *Neurocase*, 9(5):369–379.

Hinnebush, R. (2009). Syrian foreign policy under bashar al-asad. *Ortadoğu Etütleri*, 1:7–26.

Hjort, N. L., Holmes, C., Müller, P. and Walker, S. G. (2010). *Bayesian nonparametrics*. Cambridge University Press.

Hoff, P. (2014). Multilinear tensor regression for longitudinal relational data. Technical Report 631, Department of Statistics, University of Washington.

Hoff, P. D. (2008). Modeling homophily and stochastic equivalence in symmetric relational data. In Platt, J., Koller, D., Singer, Y. and Roweis, S., editors, *Advances in Neural Information Processing Systems 20*, pages 657–664. Curran Associates, Inc.

Hoff, P. D., Raftery, A. E. and Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098.

Holland, P. W. and Leinhardt, S. (1977). A dynamic model for social networks. *The Journal of Mathematical Sociology*, 5(1):5–20.

Holland, P. W. and Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, 76(373):33–50.

Holme, P. and Saramäki, J. (2012). Temporal networks. *Physics Reports*, 519(3):97–125. Temporal Networks.

Hopper, M. and Vogel, F. (1976). The limbic system in Alzheimer's disease. A neuropathologic investigation. *The American Journal of Pathology*, 85:1–20.

Horwitz, B., Grady, C. L., Schlageter, N., Duara, R. and Rapoport, S. (1987). Intercorrelations of regional cerebral glucose metabolic rates in Alzheimer's disease. *Brain research*, 407(2):294–306.

Hunter, D. R., Goodreau, S. M. and Handcock, M. S. (2008a). Goodness of fit of social network models. *Journal of the American Statistical Association*, 103(481):248–258.

Hunter, D. R., Handcock, M. S., Butts, C. T., Goodreau, S. M. and Morris, M. (2008b). ergm: A package to fit, simulate and diagnose exponential-family models for networks. *Journal of statistical software*, 24(3):1–29.

Hunter, D. R., Krivitsky, P. N. and Schweinberger, M. (2012). Computational statistical methods for social network models. *Journal of Computational and Graphical Statistics*, 21(4):856–882.

Idé, T. and Kashima, H. (2004). Eigenspace-based anomaly detection in computer systems. In *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining, KDD'04*.

Isella, L., Stehlé, J., Barrat, A., Cattuto, C., Pinton, J.-F. and Van den Broeck, W. (2011). What's in a crowd? Analysis of face-to-face behavioral networks. *Journal of theoretical biology*, 271(1):166–180.

Ishwaran, H. and Rao, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *The Annals of Statistics*, 33(2):730–773.

Ishwaran, H. and Zarepour, M. (2002). Dirichlet prior sieves in finite normal mixtures. *Statistica Sinica*, 12:941–963.

James, B. D., Leurgans, S. E., Hebert, L. E., Scherr, P. A., Yaffe, K. and Bennett, D. A. (2014). Contribution of Alzheimer disease to mortality in the United States. *Neurology*, 82(12):1045–1050.

Jonsson, P. F., Cavanna, T., Zicha, D., and Bates, P. A. (2006). Cluster analysis of networks generated through homology: automatic identification of important protein communities involved in cancer metastasis. *BMC Bioinformatics*, 7:2.

Jung, R. E., Segall, J. M., Bockholt, H. J., Flores, R. A., Smith, S. M., Chavez, R. S. and Haier, R. J. (2010). Neuroanatomy of creativity. *Human Brain Mapping*, 31(3):398–409.

Kaishev, V. K., Nielsen, J. P. and Thuring, F. (2013). Optimal customer selection for cross-selling of financial services products. *Expert Systems with Applications*, 40(5):1748–1757.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45.

Kamakura, W. A., Ramaswami, S. N. and Srivastava, R. K. (1991). Applying latent trait analysis in the evaluation of prospects for cross-selling of financial services. *International Journal of Research in Marketing*, 8(4):329–349.

Kamakura, W. A., Wedel, M., de Rosa, F. and Mazzon, J. A. (2003). Cross-selling through database marketing: A mixed data factor analyzer for data augmentation and prediction. *International Journal of Research in Marketing*, 20(1):45–65.

Kang, X., Herron, T. J., Cate, A. D., Yund, E. W. and Woods, D. L. (2012). Hemispherically-unified surface maps of human cerebral cortex: Reliability and hemispheric asymmetries. *PLoS ONE*, 7(9):e45582.

Kantarci, B. and Labatut, V. (2013). Classification of complex networks based on topological properties. In *2013 International Conference on Cloud and Green Computing*. IEEE.

Karas, G., Scheltens, P., Rombouts, S., van Schijndel, R., Klein, M., Jones, B., van der Flier, W., Vrenken, H. and Barkhof, F. (2007). Precuneus atrophy in early-onset Alzheimers disease: A morphometric structural mri study. *Neuroradiology*, 49(12):967–976.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.

Keeling, M. J. and Eames, K. T. (2005). Networks and epidemic models. *Journal of The Royal Society Interface*, 2(4):295–307.

Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada, T. and Ueda, N. (2006). Learning systems of concepts with an infinite relational model. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1*, AAAI'06, pages 381–388. AAAI Press.

Keneshloo, Y., Cadena, J., Korkmaz, G. and Ramakrishnan, N. (2014). Detecting and forecasting domestic political crises. In *Proceedings of the 2014 ACM conference on Web science - WebSci'14*.

Kesslak, J. P., Nalcioglu, O. and Cotman, C. W. (1991). Quantification of magnetic resonance scans for hippocampal and parahippocampal atrophy in Alzheimer's disease. *Neurology*, 41(1):51–51.

Kim, J., Kim, Y.-H. and Lee, J.-H. (2013). Hippocampus–precuneus functional connectivity as an early sign of Alzheimer's disease: A preliminary study using structural and functional magnetic resonance imaging data. *Brain research*, 1495:18–29.

Kivela, M., Arenas, A., Barthelemy, M., Gleeson, J. P., Moreno, Y. and Porter, M. A. (2014). Multilayer networks. *Journal of Complex Networks*, 2(3):203–271.

Koskinen, J. H. and Snijders, T. A. (2007). Bayesian inference for dynamic social network data. *Journal of Statistical Planning and Inference*, 137(12):3930–3938.

Krivitsky, P. N. and Handcock, M. S. (2014). A separable model for dynamic networks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):29–46.

Krivitsky, P. N., Handcock, M. S., Raftery, A. E. and Hoff, P. D. (2009). Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. *Social Networks*, 31(3):204–213.

Krzanowski, W. (1988). *Principles of multivariate analysis: A user's perspective*. Oxford University Press.

Kwak, H. and An, J. (2014). A first look at global news coverage of disasters by using the GDELT dataset. In *Lecture Notes in Computer Science*, pages 300–308.

Leek, J. T. and Storey, J. D. (2008). A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences*, 105(48):18718–18723.

Leetaru, K. (2011). Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space. *First Monday*, 16(9).

Leetaru, K. and Schrodt, P. A. (2013). Gdelt: Global data on events, location and tone, 1979-2012. In *International Studies Association Annual Conference*.

Leskovec, J., McGlohon, M., Faloutsos, C., Glance, N. and Hurst, M. (2007). Patterns of cascading behavior in large blog graphs. In *Proceedings of the 2007 SIAM International Conference on Data Mining*, pages 551–556.

Liang, W. S., Reiman, E. M., Valla, J., Dunckley, T., Beach, T. G., Grover, A., Niedzielko, T. L., Schneider, L. E., Mastroeni, D., Caselli, R., et al. (2008). Alzheimer's disease is associated with reduced expression of energy metabolism genes in posterior cingulate neurons. *Proceedings of the National Academy of Sciences*, 105(11):4441–4446.

Liu, S., Li, L., Faloutsos, C. and Ni, L. M. (2011). Mobile phone graph evolution: Findings, model and interpretation. In *2011 IEEE 11th International Conference on Data Mining Workshops*. Institute of Electrical & Electronics Engineers (IEEE).

Luo, W.-L. and Nichols, T. E. (2003). Diagnosis and exploration of massively univariate neuroimaging models. *NeuroImage*, 19(3):1014–1032.

Mankad, S. and Michailidis, G. (2015). Analysis of multiview legislative networks with structured matrix factorization: Does twitter influence translate to the real world? Technical Report RHS 2529074, Robert H. Smith School Research.

Mastrandrea, R., Fournet, J. and Barrat, A. (2015). Contact patterns in a high school: A comparison between data collected using wearable sensors, contact diaries and friendship surveys. *PLOS ONE*, 10(9):e0136497.

Matiş, C. and Ilieş, L. (2014). Customer relationship management in the insurance industry. *Procedia Economics and Finance*, 15:1138–1145.

McPherson, M., Smith-Lovin, L. and Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A. and Aiden, E. L. (2010). Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182.

Milgram, S. (1967). The small world problem. *Psychology Today*, 2(1):60–67.

Minoshima, S., Giordani, B., Berent, S., Frey, K. A., Foster, N. L. and Kuhl, D. E. (1997). Metabolic reduction in the posterior cingulate cortex in very early Alzheimer's disease. *Annals of Neurology*, 42:85–94.

Morris, C. (1983). Parametric empirical Bayes inference: Theory and applications. *Journal of the American Statistical Association*, 78(381):47–55.

Mueller, S., Wang, D., Fox, M. D., Yeo, B. T., Sepulcre, J., Sabuncu, M. R., Shafee, R., Lu, J. and Liu, H. (2013). Individual variability in functional connectivity architecture of the human brain. *Neuron*, 77(3):586–595.

Murray, I. and Adams, R. P. (2010). Slice sampling covariance hyperparameters of latent Gaussian models. In *Advances in Neural Information Processing Systems*, pages 1732–1740.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265.

Newman, M. E. J. (2003). Mixing patterns in networks. *Physical Review E*, 67(2).

Nowicki, K. and Snijders, T. A. B. (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96(455):1077–1087.

Olde Dubbelink, K. T. E., Hillebrand, A., Stoffers, D., Deijen, J. B., Twisk, J. W. R., Stam, C. J. and Berendse, H. W. (2014). Disrupted brain network topology in Parkinson's disease: A longitudinal magnetoencephalography study. *Brain*, 137(1):197–207.

Papadimitriou, P., Dasdan, A. and Garcia-Molina, H. (2010). Web graph similarity for anomaly detection. *Journal of Internet Services and Applications*, 1(1):19–30.

Polson, N. G., Scott, J. G. and Windle, J. (2013). Bayesian inference for logistic models using Pólya–gamma latent variables. *Journal of the American Statistical Association*, 108(504):1339–1349.

Prasad, G., Joshi, S. H., Nir, T. M., Toga, A. W., Thompson, P. M., (ADNI, A. D. N. I., et al. (2015). Brain connectivity and novel network measures for Alzheimer's disease classification. *Neurobiology of aging*, 36:S121–S131.

Ramsey, J. D., Hanson, S. J., Hanson, C., Halchenko, Y. O., Poldrack, R. A. and Glymour, C. (2010). Six problems for causal inference from fmri. *Neuroimage*, 49(2):1545–1558.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. The MIT Press.

Robins, G. and Pattison, P. (2001). Random graph models for temporal processes in social networks. *The Journal of Mathematical Sociology*, 25(1):5–41.

Robins, G., Pattison, P., Kalish, Y. and Lusher, D. (2007a). An introduction to exponential random graph models for social networks. *Social Networks*, 29(2):173–191.

Robins, G., Snijders, T., Wang, P., Handcock, M. and Pattison, P. (2007b). Recent developments in exponential random graph models for social networks. *Social Networks*, 29(2):192–215.

Roncal, W. G., Koterba, Z. H., Mhembere, D., Kleissas, D. M., Vogelstein, J. T., Burns, R., Bowles, A. R., Donavos, D. K., Ryman, S., Jung, R. E., Wu, L., Calhoun, V. and Vogelstein, R. J. (2013). Migraine: Mri graph reliability analysis and inference for connectomics. In *IEEE Global Conference on Signal and Information Processing*. IEEE.

Roos, M. and Held, L. (2011). Sensitivity analysis in Bayesian generalized linear mixed models for binary data. *Bayesian Analisis*, 6(2):259–278.

Rousseau, J. and Mengersen, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):689–710.

Rubinov, M. and Sporns, O. (2010). Complex network measures of brain connectivity: Uses and interpretations. *NeuroImage*, 52(3):1059–1069.

Rudin, W. (1976). *Principles of Mathematical Analysis*. McGraw–Hill, 3rd edition.

Salter-Townshend, M. and McCormick, T. H. (2013). Latent space models for multiview network data. Technical Report 622, Department of Statistics, University of Washington.

Sampson, S. F. (1969). *Crisis in a cloister*. PhD thesis, Cornell University.

Sarkar, P., Chakrabarti, D. and Jordan, M. (2014). Nonparametric link prediction in large scale dynamic networks. *Electronic Journal of Statistics*, 8(2):2022–2065.

Sarkar, P. and Moore, A. W. (2005). Dynamic social network analysis using latent space models. *SIGKDD Explor. Newsl.*, 7(2):31–40.

Sawyer, K. R. (2012). *Explaining Creativity: The Science of Human Innovation*. Oxford University Press.

Schein, A., Moore, J. and Wallach, H. (2013). Inferring multilateral relations from dynamic pairwise interactions. In *NIPS 2013 Workshop on Networks*.

Schein, A., Paisley, J., Blei, D. and Wallach, H. (2014). Inferring polyadic events with Poisson tensor factorization. In *NIPS 2014 Workshop on Networks*.

Scheinerman, E. R. and Tucker, K. (2010). Modeling graphs using dot product representations. *Computational Statistics*, 25(1):1–16.

Schrodt, P. (2012). Cameo conflict and mediation event observations event and actor codebook. `http://data.gdeltproject.org/documentation/CAMEO.Manual.1.1b3.pdf`.

Schrodt, P. (2014). Tabari textual analysis by augmented replacement instructions. `http://eventdata.parusanalytics.com/tabari.dir/TABARI.0.8.4b3.manual.pdf`.

Scott, J. G., Kelly, R. C., Smith, M. A., Zhou, P. and Kass, R. E. (2014). False discovery rate regression: An application to neural synchrony detection in primary visual cortex. *Journal of the American Statistical Association*, to appear.

Sellke, T., Bayarri, M. J. and Berger, J. O. (2001). Calibration of p values for testing precise null hypotheses. *The American Statistician*, 55(1):62–71.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650.

Sewell, D. K. and Chen, Y. (2015). Latent space models for dynamic networks. *Journal of the American Statistical Association*, to appear.

Shaked, M. and Shanthikumar, J. G. (2007). *Stochastic orders*. Springer Science & Business Media.

Shobe, E. R., Ross, N. M. and Fleck, J. I. (2009). Influence of handedness and bilateral eye movements on creativity. *Brain and Cognition*, 71(3):204–214.

Silverman, B. W. (1984). Spline smoothing: The equivalent variable kernel method. *Annals of Statististics*, 12(3):898–916.

Simpson, S. L., Bowman, F. D. and Laurienti, P. J. (2013). Analyzing complex functional brain networks: Fusing statistics and network science to understand the brain. *Statistics surveys*, 7:1.

Simpson, S. L., Hayasaka, S. and Laurienti, P. J. (2011). Exponential random graph modeling for complex brain networks. *PLoS One*, 6(5):e20039.

Simpson, S. L., Moussa, M. N. and Laurienti, P. J. (2012). An exponential random graph modeling approach to creating group-based representative whole-brain connectivity networks. *NeuroImage*, 60(2):1117–1126.

Smith, M., Esiri, M., Barnetson, L., King, E. and Nagy, Z. (2001). Constructional apraxia in Alzheimers disease: Association with occipital lobe pathology and accelerated cognitive decline. *Dementia and geriatric cognitive disorders*, 12(4):281–288.

Smith, M. and Kohn, R. (1996). Nonparametric regression using Bayesian variable selection. *Journal of Econometrics*, 75(2):317–343.

Smith, S. M., Miller, K. L., Salimi-Khorshidi, G., Webster, M., Beckmann, C. F., Nichols, T. E., Ramsey, J. D. and Woolrich, M. W. (2011). Network modelling methods for fmri. *Neuroimage*, 54(2):875–891.

Snijders, T. A. B. (2001). The statistical evaluation of social network dynamics. *Sociological Methodology*, 31(1):361–395.

Snijders, T. A. B. (2002). Markov chain Monte Carlo estimation of exponential random graph models. *Journal of Social Structure*, 3(2):1–40.

Snijders, T. A. B. (2005). Models for longitudinal network data. In *Models and Methods in Social Network Analysis*, pages 215–247.

Snijders, T. A. B., van de Bunt, G. G. and Steglich, C. E. (2010a). Introduction to stochastic actor-based models for network dynamics. *Social Networks*, 32(1):44–60.

Snijders, T. A. B., Koskinen, J. and Schweinberger, M. (2010b). Maximum likelihood estimation for social network dynamics. *Annals of Applied Statistics*, 4(2):567–588.

Sporns, O. (2010). *Networks of the Brain*. The MIT Press, 1st edition.

Sporns, O. (2013). Structure and function of complex brain networks. *Dialogues in clinical neuroscience*, 15(3):247.

Springer, M. D. and Thompson, W. E. (1970). The distribution of products of beta, gamma and gaussian random variables. *SIAM J. Appl. Math.*, 18(4):721–737.

Stam, C. J. (2014). Modern network science of neurological disorders. *Nature Reviews Neuroscience*, 15(10):683–695.

Stehlé, J., Charbonnier, F., Picard, T., Cattuto, C. and Barrat, A. (2013). Gender homophily from spatial behavior in a primary school: A sociometric study. *Social Networks*, 35(4):604–613.

Stehlé, J., Voirin, N., Barrat, A., Cattuto, C., Isella, L., Pinton, J.-F., Quaggiotto, M., den Broeck, W. V., Régis, C., Lina, B. and Vanhems, P. (2011). High-resolution measurements of face-to-face contact patterns in a primary school. *PLoS ONE*, 6(8):e23176.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):795–809.

Stirling, J. and Elliott, R. (2008). *Introducing Neuropsychology*. Routledge.

Strauss, D. and Ikeda, M. (1990). Pseudolikelihood estimation for social networks. *Journal of the American Statistical Association*, 85(409):204–212.

Sussman, D. L., Tang, M., Fishkind, D. E. and Priebe, C. E. (2012). A consistent adjacency spectral embedding for stochastic blockmodel graphs. *Journal of the American Statistical Association*, 107(499):1119–1128.

Sussman, D. L., Tang, M. and Priebe, C. E. (2014). Consistent latent position estimation and vertex classification for random dot product graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(1):48–57.

Takeuchi, H., Taki, Y., Sassa, Y., Hashizume, H., Sekiguchi, A., Fukushima, A. and Kawashima, R. (2010). White matter structures associated with creativity: Evidence from diffusion tensor imaging. *NeuroImage*, 51(1):11–18.

Tamassia, R. (2007). *Handbook of Graph Drawing and Visualization*. Chapman & Hall.

Tang, M., Sussman, D. L. and Priebe, C. E. (2013). Universally consistent vertex classification for latent positions graphs. *The Annals of Statistics*, 41(3):1406–1430.

Tansey, W., Koyejo, O., Poldrack, R. A. and Scott, J. G. (2014). False discovery rate smoothing. *arXiv:1411.6144*.

Taylor, J. B. (2009). The financial crisis and the policy responses: An empirical analysis of what went wrong. NBER Working Papers 14631, National Bureau of Economic Research, Inc.

Thangavel, R., Van Hoesen, G. W. and Zaheer, A. (2008). Posterior parahippocampal gyrus pathology in Alzheimer's disease. *Neuroscience*, 154(2):667–676.

Thapthiang, N. (2013). An analysis of news reporting and its effects, using ibil model: Lee gardens plaza and cs pattani hotels cases. *Procedia-Social and Behavioral Sciences*, 91:411–420.

Thiebaut de Schotten, M., Bakardjian, H., Lista, S., Teipel, S., Dyrba, M., Filippi, M., Frisoni, G. B., Fellgiebel, A., Bokde, A., Klöppel, S., et al. (2014). Advanced diffusion weighting imaging (dwi) tractography of the limbic system: Novel biomarkers of neurodegenerative changes during progression/conversion from cognitive normality to AD dementia. *Alzheimer's & Dementia*, 10(4):P37.

Thompson, P. M., Hayashi, K. M., De Zubicaray, G., Janke, A. L., Rose, S. E., Semple, J., Herman, D., Hong, M. S., Dittmer, S. S., Doddrell, D. M., et al. (2003). Dynamics of gray matter loss in Alzheimer's disease. *The Journal of Neuroscience*, 23(3):994–1005.

Thompson, P. M., Mega, M. S., Woods, R. P., Zoumalan, C. I., Lindshield, C. J., Blanton, R. E., Moussai, J., Holmes, C. J., Cummings, J. L. and Toga, A. W. (2001). Cortical change in alzheimer's disease detected with a disease-specific population-based brain atlas. *Cerebral Cortex*, 11(1):1–16.

Thuring, F. (2012). A credibility method for profitable cross-selling of insurance products. *Annals of Actuarial Science*, 6(01):65–75.

Thuring, F., Nielsen, J., Guillén, M. and Bolancé, C. (2012). Selecting prospects for cross-selling financial products using multivariate credibility. *Expert Systems with Applications*, 39(10):8809–8816.

Tsygankov, A. (2015). Vladimir Putin's last stand: The sources of Russia's Ukraine policy. *Post-Soviet Affairs*, 31(4):279–303.

Vanhems, P., Barrat, A., Cattuto, C., Pinton, J.-F., Khanafer, N., Régis, C., a Kim, B., Comte, B. and Voirin, N. (2013). Estimating potential infection transmission routes in hospital wards using wearable proximity sensors. *PLoS ONE*, 8(9):e73970.

Verhoef, P. C. and Donkers, B. (2001). Predicting customer potential value an application in the insurance industry. *Decision Support Systems*, 32(2):189–199.

Wang, H. and Marron, J. S. (2007). Object oriented data analysis: Sets of trees. *The Annals of Statistics*, 35(5):1849–1873.

Wang, J., He, L., Zheng, H. and Lu, Z.-L. (2014). Optimizing the magnetization-prepared rapid gradient-echo (MP-RAGE) sequence. *PLoS ONE*, 9(5):1–12.

Wanta, W., Golan, G. and Lee, C. (2004). Agenda setting and international news: Media influence on public perceptions of foreign nations. *Journalism & Mass Communication Quarterly*, 81(2):364–377.

Wasserman, S. and Pattison, P. (1996). Logit models and logistic regressions for social networks: An introduction to markov graphs and p^*. *Psychometrika*, 61(3):401–425.

Watson, G. N. (1966). *A Treatise on the Theory of Bessel Functions*. Cambridge University Press, 2nd edition.

Watts, D. J. (1999). *Small Worlds: The Dynamics of Networks between Order and Randomness*. Princeton University Press, 1st edition.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of small-world networks. *Nature*, 393(6684):440–442.

Withers, C. S. and Nadarajah, S. (2013). On the product of gamma random variables. *Quality & Quantity*, 47(1):545–552.

Worsley, K. (2003). Detecting activation in fmri data. *Statistical Methods in Medical Research*, 12(5):401–418.

Xing, E. P., Fu, W. and Song, L. (2010). A state-space mixed membership blockmodel for dynamic network tomography. *Annal of Applied Statistics*, 4(2):535–566.

Xu, K. S. (2015). Stochastic block transition models for dynamic networks. *Journal of Machine Learning, Workshops & Proceedings*, 38:1079–1087.

Xu, K. S. and Hero, A. O. (2014). Dynamic stochastic blockmodels for time-evolving social networks. *IEEE Journal of Selected Topics in Signal Processing*, 8(4):552–562.

Yang, T., Chi, Y., Zhu, S., Gong, Y. and Jin, R. (2009). A Bayesian approach toward finding communities and their evolutions in dynamic social networks. In *Proceedings of the 2009 SIAM International Conference on Data Mining*, pages 990–1001. Society for Industrial & Applied Mathematics (SIAM).

Yang, T., Chi, Y., Zhu, S., Gong, Y. and Jin, R. (2010). Detecting communities and their evolutions in dynamic social networks—a Bayesian approach. *Machine Learning*, 82(2):157–189.

Zalesky, A., Fornito, A. and Bullmore, E. T. (2010). Network-based statistic: Identifying differences in brain networks. *NeuroImage*, 53(4):1197–1207.

Zaman, T., Fox, E. B. and Bradlow, E. T. (2014). A Bayesian approach for predicting the popularity of tweets. *Annal of Applied Statistics*, 8(3):1583–1611.

Zhou, J., Bhattacharya, A., Herring, A. H. and Dunson, D. B. (2014). Bayesian factorizations of big sparse tensors. *Journal of the American Statistical Association*, to appear.

Zhou, Y., Dougherty, J. H., Hubner, K. F., Bai, B., Cannon, R. L. and Hutson, R. K. (2008). Abnormal connectivity in the posterior cingulate and hippocampus in early Alzheimer's disease and mild cognitive impairment. *Alzheimer's & Dementia*, 4(4):265–270.

Zhu, B. and Dunson, D. B. (2013). Locally adaptive Bayes nonparametric regression via nested Gaussian processes. *Journal of the American Statistical Association*, 108(504):1445–1456.

Zhu, D. C., Majumdar, S., Korolev, I. O., Berger, K. L. and Bozoki, A. C. (2013). Alzheimer's disease and amnestic mild cognitive impairment weaken connections within the default-mode network: A multi-modal imaging study. *Journal of Alzheimer's Disease*, 34(4):969–984.

Daniele Durante

CONTACT INFORMATION

Dipartimento di Scienze Statistiche
Università degli Studi di Padova
Via Cesare Battisti, 241
35121 Padova

✉ durante@stat.unipd.it
web: fare-ricerca/durante-daniele
or: danieledurante2.wix.com/daniele-durante

RESEARCH INTERESTS

Network Data, Bayesian Nonparametrics, Tensor Factorization, Stochastic Processes and State Space Models, Models for Latent Variables, Machine Learning.

CURRENT POSITION

Università degli Studi di Padova, Department of Statistics, Padova, Italy

Ph.D., Statistics (XXVIII cycle). From 1 January 2013 to 31 December 2015.

- Thesis Topic: *Bayesian Nonparametric Modeling of Network Data*
- Advisors: Bruno Scarpa and David B. Dunson

VISITING PERIODS

Study visits during the Ph.D.

- Department of Statistical Science, Duke University, Durham, NC, USA (03/2014 – 07/2014)
- Department of Statistical Science, Duke University, Durham, NC, USA (09/2014 – 02/2015)

PAST POSITIONS

Academic

Research assistant. University of Padova. Research project: *Development of a model for mortality rates using Bayesian nonparametric methods*, headed by prof. Stefano Mazzuco. (from 10/2012 to 01/2013)

Tutor. University of Padova, Department of Statistical Sciences. Exercises and small lectures for undergraduate students. (from 10/2011 to 10/2012)

Research assistant. Ca' Foscari University. Research project: *Mothers smoke-free*, headed by prof. Stefano Campostrini. (from 05/2011 to 09/2011)

Business

Freelance Statistical Research Consultant. Collaboration with STEP. Research project: *Analysis of functional data analysis (EEG).* (from 01/2012 to 07/2012)

Freelance Statistical Business Consultant. Collaboration with QUAERIS. Research project: *Analysis of customer satisfaction.* (from 07/2011 to 09/2011)

EDUCATION

Università degli Studi di Padova, Padova, Italy

M.S., Statistical Sciences, Department of Statistics (110/110 cum laude). (from 2010 to 2012)

- Thesis Topic: *Locally Adaptive Factor Processes for Multivariate Time Series*
- Advisor: Bruno Scarpa

B.S., Statistics, Economy and Finance, Department of Statistics (110/110 cum laude). (from 2007 to 2010)

Awards

Academic

- Winner of the *David P. Byar Award* for the top paper among the student travel award winners. Biometrics Section of the American Statistical Association. (2015)
- Winner of the *ISBA Lifetime Members Junior Researcher Award*. International Society for Bayesian Analysis. (2014)
- Winner of the *Grand Data Challenge* of the 2014 International Conference on Social Computing, Behavioral-Cultural Modeling, & Prediction. (2014)
- Winner of the *Laplace Award* for the top paper among the student travel award winners. Section on Bayesian Statistical Sciences of the American Statistical Association. (2013)

Dissemination of Statistics

- *My Statistician Friend*. Honorable mention in the "video contest" sponsored for the International Year of Statistics 2013. (2013)
- *The Statistical Calendar*. Awarded as the best project of dissemination of Statistics in the multimedia competition: "La statistica e la professione di statistico: idee per la promozione e la diffuzione". ISTAT, Rome. (2012)

Publications

Peer–reviewed Journal Articles

- Durante, D. and Dunson, D. B., (2014). Nonparametric Bayes dynamic modelling of relational data. *Biometrika*, 101, 883–898.
- Durante, D. and Dunson, D. B., (2014). Bayesian dynamic financial networks with time-varying predictors. *Statistics & Probability Letters*, 93, 19–26.
- Durante, D., Scarpa, B. and Dunson, D. B., (2014). Locally adaptive factor processes for multivariate time series. *Journal of Machine Learning Research*, 15, 1493–1522.
- Durante, D., (2012). Qualitative latent variables: a comparison between SEM and LCA. *Quaderni di Statistica*, 14, 97–100.

Peer–reviewed Conference Proceedings

- Durante, D. and Dunson, D. B., (2015). Bayesian regression with network predictors. *Proceedings of the XLVIII conference of the Italian Statistical Society*, 1–7.
- Durante, D. and Dunson, D. B., (2014). Bayesian Logistic Gaussian Process Models for Dynamic Networks. *Artificial Intelligence and Statistics (AISTAT). Journal of Machine Learning Research-Workshop & Proceedings*, 33, 194–201.
- Durante, D., (2014). Analysis of Italian Financial Market via Bayesian Dynamic Covariance Models. In Lanzarone, E. and Ieva, F. *The contribution of Young Researchers to Bayesian Statistics*. Springer, 63, 171–177.
- Durante, D., Scarpa, B. and Dunson, D. B., (2013). Locally Adaptive Bayesian Multivariate Time Series. *Advances in Neural Information Processing Systems (NIPS)*, 26, 1664–1672.

Dissemination Articles

- Durante, D., Canale, A., Guidolin, M., Finos, L. and Scarpa, B., (2015). A night of Statistics, problem solving, and teamwork. *Proceedings of the XLVIII conference of the Italian Statistical Society*, 1–3.
- Durante, D., Vidotto, D. and Vettori, S. (2015). La bussola del ricercatore statistico. In Campostrini S., Ghellini, G. and Tuzzi, A. (eds.) *Con senso di misura, riflessi statistici da alcuni allievi di Lorenzo Bernardi.* Cleup, 25–36.
- Durante, D., (2013). My Statistician Friend. La Statistica vista con gli occhi degli studenti. *Induzioni*, 45, 103–116.

Manuscripts Under Review

- Durante, D., Paganin, S., Scarpa, B. and Dunson, D. B. (2015). Bayesian modeling of networks in complex business intelligence problems. *arXiv: 1510.00646*, submitted.
- Durante, D. and Dunson, D. B. (2016). Locally Adaptive Dynamic Networks. *arXiv: 1505.05668.* (Annals of Applied Statistics. Revision submitted).
- Durante, D. and Dunson, D. B. (2015). Bayesian inference on group differences in brain networks. *arXiv: 1411.6506.* (Bayesian Analysis. Revision requested).
- Durante, D., Dunson, D. B and Vogelstein J. T. (2015). Nonparametric Bayes modeling of Populations of Networks. *arXiv: 1406.7851.* (Journal of the American Statistical Association. Revision submitted).
- Durante, D., Shah, I. and Torelli, N., (2014). Bayesian nonparametric modeling of contraceptive use in India. *arXiv:1405.7555*, submitted.

Conferences presentation

- Bayesian Connectomics. *Department of Mathematics and Statistics, Lancaster University.* Lancaster, UK, October 7, 2015 (invited seminar)
- Bayesian Regression with Network Predictors. *SIS2015.* Treviso, Italy, September 10, 2015 (invited talk)
- Bayesian Inference on Group Differences in Brain Networks. *JSM2015.* Seattle, USA, August 13, 2015 (invited talk)
- Locally Adaptive DYnamic (LADY) networks. *StatMP.* Padua, Italy, May 28, 2015 (invited seminar)
- La Bussola del Ricercatore Statistico. *Una giornata in ricordo di Lorenzo Bernardi.* Padua, Italy, May 15, 2015 (invited talk)
- Bayesian Inference on network data. *ARS'15.* Capri, Italy, April 29, 2015 (invited talk)
- Modelli Bayesiani Non Parametrici per Popolazioni di Reti. *Modelli bayesiani non parametrici e applicazioni.* Padua, Italy, September 17, 2014 (invited talk)
- Nonparametric Bayes dynamic modeling of relational data. *ISBA2014.* Cancun, Mexico, July 15, 2014 (contributed talk)
- Bayesian Logistic Gaussian Process Models for Dynamic Networks. *AISTAT2014.* Reykjavik, Iceland, April 22, 2014 (poster)
- Friends in Joy and Sorrow: Analysis of the 2007-2012 Global Financial Crisis via Bayesian Nonparametric Dynamic Networks. *SBP 2014.* Washington DC, USA, April 4, 2014 (invited talk)

- Locally Adaptive Bayesian Multivariate Time Series. *JSM2013*. Montreal, Canada, August 6, 2013 (contributed talk)
- Analysis of Italian Financial Market via Bayesian Covariance Regression. *BAYSM 2013*. Milan, Italy, June 5, 2013 (contributed talk)
- Locally Adaptive Bayesian Covariance Regression. *Workshop on Bayesian nonparametric models for functional data: theory and application in biometry and marketing.* Padua, Italy, October 5, 2012 (invited talk)
- Qualitative latent variables: a comparison between SEM and LCA. *Methods and models for latent variables Final Conference of PRIN 2008*. Naples, Italy, May 1719, 2012 (contributed talk)
- The Statistical Calendar. *ISTAT giornata italiana della Statistica.* Rome, Italy, October 20, 2011 (invited talk)
- Analisi di Customer Satisfaction, l'applicabilità dellapproccio LISREL. *Statistica in Azienda, Statistici in Azienda*. Padua, Italy, June 15, 2010 (poster session)

EDITORIAL ACTIVITY

Referee for: Computational Statistics & Data Analysis; AISTAT; SBP Conference; Quaderni di Statistica.

ORGANIZATION OF SCIENTIFIC EVENTS

- Chair of the event *"Stats under the Stars"*. Padua, 8-9 September, 2015 (Satellite event of the SIS intermediate scientific meeting)
- Joint organiser of *"Statistica e Data Science per il Business"*. Padua, 8 September, 2015 (Satellite event of the SIS intermediate scientific meeting)

MEMBERSHIPS

Italian Statistical Society (SIS); American Statistical Association (ASA) & Biometrics Section.

TEACHING EXPERIENCE

Università degli Studi di Padova

School of Statistics

- Introduction to network analysis (specialist seminar during the class: Analisi dei Dati e Data Mining). June, 2015.