UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Scienze Statistiche
Corso di Dottorato di Ricerca in Scienze Statistiche
Ciclo XXXII

Bayesian modelling of complex dependence structures

Coordinatore del Corso: Prof. Massimiliano Caporin

Supervisore: Prof. Bruno Scarpa

Co-supervisore: Prof. David B. Dunson

Dottorando: Emanuele Aliverti

29 November 2019

Abstract

Complex dependence structures characterising modern data are routinely encountered in a large variety of research fields. Medicine, biology, psychology and social sciences are enriched by intricate architectures such as networks, tensors and more generally high-dimensional dependent data. Rich dependence structures stimulate challenging research questions and open wide methodological avenues in different areas of statistical research, providing an exciting atmosphere to develop innovative tools. A primary interest in statistical modelling of complex data is on adequately extracting information to conduct meaningful inference, providing reliable results in terms of uncertainty quantification and generalisability into future samples. These aims require ad-hoc statistical methodologies to appropriately characterize the dependence structures defining complex data as such, further improving the understanding of the mechanisms underlying the observed configurations.

The focus of the thesis is on Bayesian modelling of complex dependence structures via latent variable constructs. This strategy characterises the dependence structure in an unobservable latent space, specifying the observed quantities as conditionally independent given a set of latent attributes, facilitating tractable posterior inference and an eloquent interpretation. The thesis is organized into three main parts, illustrating case studies from different fields of application and focused on studying modern challenges in neuroscience, psychology and criminal justice. Bayesian modelling of the complex data arising in these domains via latent features effectively provides valuable insights on different aspects of such structures, addressing the questions of interest and contributing to the scientific understanding.

Sommario

Strutture di dipendenza complesse sono molto diffuse in diverse applicazioni. Medicina, biologia, psicologia e scienze sociali sono arricchite da architetture complicate quali reti, tensori e più generalmente dati dipendenti ed ad alta dimensionalità. Strutture di dipendenza articolate stimolano complesse domande di ricerca ed aprono ampi spazi metodologici in diversi ambiti di ricerca statistica, creando una frizzante atmosfera nella quale sviluppare strumenti innovativi. Un obiettivo cruciale nella modellazione statistica di dati complessi consiste nell'estrazione di informazione per condurre inferenza coerente e ottenere risultati affidabili in termini di quantificazione dell'incertezza e di validità per dati futuri. Questi obiettivi necessitano di metodologie statistiche *ad-hoc* per caratterizzare un modo appropriato le strutture di dipendenza che definiscono dati complessi in quanto tali, migliorando ulteriormente la conoscenza dei meccanismi sottostanti tali strutture.

Questa tesi si concentra sulla modellazione Bayesiana di strutture di dipendenza complessa tramite costrutti a variabili latenti. Tale strategia caratterizza la struttura di dipendenza in uno spazio latente, specificando le quantità osservate come condizionatamente indipendenti dato un insieme di attributi latenti, i quali semplificano l'inferenza a posteriori e permettono un'eloquente interpretazione. La tesi è organizzata in tre parti principali, le quali illustrano diverse applicazioni in neuroscienze, psicologia e giustizia criminale. Una modellazione Bayesiana tramite variabili latenti dei dati complessi che nascono in questi ambiti fornisce interessanti intuizioni su diversi aspetti di tali strutture, rispondendo a diverse domande di ricerca e contribuendo alla conoscenza scientifica in materia.

Le parole sono importanti

Acknowledgements

Thanks to Bruno, the best mentor I could ever ask for. Directly and indirectly, you gave me a huge collection of advices, most of which are crucial for being a good statistician and a fruitful researcher. Your scientific supervision and personal support were essential for improving many aspects of my work and, most importantly, myself. The real question is whether I *really* learned something, but let's say for the moment that I took many notes.

I am also really grateful to David, the most brilliant professor and powerful idea-machine I've ever met. You've always motivated my creativity and enthusiasm in research, granting me the freedom to follow my interests and intuitions, and providing useful comments and smart feedbacks on any path I was trying to follow. You really showed me the right attitude, tools and rules to play this game.

I was lucky enough to meet smart colleagues that became good friends during these years. Max and Sally in Padova, Michael, Victor and Felipe at Duke, whose friendship, VIM advices and cheap booze helped me in many moments. A warm thanks go to my friends (and office mates) Ilaria, Alessandro, Mohammed, Federico, Anastasiia, Huiting, Moin and Andrea. I'm also thankful to Tony and Daniele for motivating me with a challenging trade-off between seducing sloppiness and obsessive-compulsive consistency in notation.

If I am writing this today, then I owe my parents for most everything. Their immense love, perpetual support and sincere pride were the few guarantees I've always had during all the choices I took in my life.

My last thanks is for Arianna, for supporting and "sopporting" me everyday.

Contents

List of Figures

List of Tables

Introduction

Overview

In a large variety of fields of application, structured high-dimensional data are commonly collected for different purposes. Some notable example include networks, tensors, functions, texts and images, routinely arising in different scientific fields such as medicine, biology, social sciences and demography, among many others. With complex data comes challenging research questions, opening new avenues to improve current scientific knowledge and ideally advise policy. Practitioners have often focused on preprocessing such complex data to force them into simpler structures such as matrices, vectors and scalars, where standard techniques work without efforts. However, such an approach inevitably destroys the intrinsic dependence underlying structured data, which is a precious source of information to investigate patterns, regularities and the relations among structured data and other factors of scientific interest.

For example, there has been growing interest in understanding the connectivity structure within human brains, investigating how regions are connected and how such structures relate with subject-specific traits, such as intelligence or diseases. A standard approach would attempt to reduce such rich structure into a set of predictors $\mathbf{x}_i$ and a response y_i amenable to standard regression or classification methods. Unfortunately, it has been observed that such an approach often leads to unreliable results, poorly interpretable inference and over-confident conclusions which do not generalize to future samples (Dunson, 2018). Therefore, formal statistical methodologies are required to understand data with complex structures, providing meaningful and interpretable results with solid assessment of uncertainty of estimation.

This thesis focuses on methods leveraging conditionally independence specifications, given a set of latent features which characterise the dependence structure in an unobservable space. Such an approach leads to substantial advantages in interpretation, providing a natural way to make sense out of complex dependence structures with a

moderate number of parameters and use such simple representations to address the research questions of interest. Compact latent representations provide also gains in computation, allowing to conduct posterior inference efficiently via data-augmentation algorithms leveraging Monte Carlo simulation or optimisation techniques.

The concrete advantages of latent variables approaches to inference are illustrated in different applications ranging from criminal justice, psychology and neuroscience. Such fields of research provide complex data structures and stimulate challenging research aims, with the focus being on modelling and explaining the dependence and association patterns characterising such complex data. Using latent structures, these aims can be successfully addressed with computational efficiency. Clearly, the methods proposed in the following chapters should not be regarded as exclusively limited to the specific illustrative applications, but could also be directly adapted to different case studies involving dataset similar in the structure, but arising in completely different field of research. This aspect is crucial in the development of modern statistical methodologies, which should always have an exhaustive view that is broad enough to embrace different fields and allow the statistician to "get to play in everyone's backyard".

Main contributions of the thesis

Latent space models for network data

Chapter 1 focuses on latent structures modelling for the analysis of network data, which are commonly collected in a large variety of fields of application. Some notable examples include social sciences (McPherson *et al.*, 2001), biology (Jonsson *et al.*, 2006), economics (Jackson, 2014) and neuroscience (Bullmore and Sporns, 2009), where the analysis of interconnected units can provide valuable insights on the functionality of the entire complex system. In practice, it is of interest to investigate different aspects of network data, ranging from simple descriptive statistics to complete specification of the network generating process and its dependence structure.

The benefits of network modelling leveraging latent structures are illustrated with two case studies motivated by the recent abundance of brain scan data, measuring the physical connections among pre-specified sets of brain regions in live humans. These data are increasingly available for multiple individuals along with additional information on the regions anatomy; for example, the 3–dimensional anatomical coordinates of the regions, and their membership to hemispheres and lobes. Although recent studies

have explored the spatial effects underlying brain networks, there is still a lack of statistical analyses on the net connectivity structure which is not explained by the physical proximity of the brain regions.

In answering the above questions, popular approaches in neuroscience focus on network summary statistics, such as the number of connections, the average path length and the clustering coefficient, among others (Rubinov and Sporns, 2010). The goal of these analyses is to evaluate whether the brain connectivity structure is characterized by small–world or scale–free properties and community structures, possibly varying with region–specific predictors (e.g. Bullmore and Sporns, 2009, 2012; Sporns, 2013; Stam, 2014). Although these descriptive analyses offer valuable insights, statistical modeling of brain network data is fundamental to provide scientific inference on heterogeneous structures. For example, brain connectivity could vary systematically in relation to predictors, or endogenously due to underlying dependence structures among the edges. The above considerations have motivated an increasing interest in more sophisticated statistical modeling of networks; for example, Exponentially Random Graphs Models (ERGM) provide a popular tools for the analysis of brain network data (e.g. Simpson *et al.*, 2011), but often lead to poor characterisations of the network structure and face computational bottlenecks (Hunter *et al.*, 2012).

In Chapter 1, latent structure models for network data will be extended to include region–specific covariates, while allowing joint modelling of multiple networks associated with different individuals. Explicitly accounting for these covariates in latent space models provides key insights on how network architectures relate to brain anatomy and which patterns departs from physical proximity. Indeed, the first method developed in Section 1.4 generalises latent space models with nodes clustering allowing finer grouping of brain regions via a mixed membership clustering. Such a specification allows to detect clusters of brain regions which are similar with respect to a subset of latent features, providing a more detailed explanation of the brain architecture underlying connectivity. The approach illustrated in Section 1.5, instead, provides an important computational contribution in the development of Bayesian models for network data. Specifically, a Mean-Field Variational Bayes algorithm to conduct approximate inference for the latent factor model for networks is developed and extended to include additional covariates. This contribution allows to conduct approximate Bayesian inference for high-quality network scans including hundreds of brain regions, and provide meaningful inference on the connectivity patterns and on its anatomical determinants.

Latent structures models for multivariate categorical data

Chapter 2 focuses on latent structure modelling for multivariate categorical data, which are crucial in a large number of fields of research. From medical studies to social sciences, there is an immense variety of applications in which the analysis of observations on categorical scales is a routine problem (e.g. Agresti, 2003). Categorical data are collected whenever individuals are asked to report *opinions* or feelings about something; for example, how close do they feel to a political party or their emotions during particular situations. The development of methods to analyse categorical data began well back in the 19th century, and has constantly received attention remaining a very active area of research (e.g. Fienberg and Rinaldo, 2007).

Several research questions can be addressed characterising the dependence structure underlying the categorical variables and their low-dimensional functionals, such as the marginal bivariate or conditional distributions (Agresti, 2003). Therefore, it is of substantial interest to develop parsimonious, yet flexible, statistical methodologies for categorical data, in order to characterize such complex structures and conduct meaningful and reliable statistical inference. This problem is particularly challenging with multivariate categorical data, since the resulting contingency is tremendously large, with a number of cells that grows exponentially with the number of features. Latent variable models provide concrete benefits to reduce the number of free parameters and provide efficient estimates by assuming that the dependence structure is modelled via latent features, with categorical variables given the latent structure having a simple specification.

Section 2.3 focuses on latent structure modelling for high-dimensional multivariate categorical data, introducing a novel methodology to combine two popular approaches for categorical data, log-linear modelling and latent class analysis. Specifically, the focus is on obtaining a model for categorical data which allows a simple interpretation of its parameters in terms of association among categorical variables, relying on latent structures to improve the flexibility of the approach in characterising higher order dependences. Compared to standard log-linear models, our approach can accommodate a large number of categorical variable with modest computational power, while compared with latent class models and tensor decompositions, our procedure is able to characterize complex dependent categorical data with a limited number of mixture components.

The advantages of the proposed approach are illustrated on a case study involving psychological questionnaires administered to suicide attempts patients at the hospital of Padova, carefully described in Section 2.2. Specifically, the focus of the application is on investigating which psycho pathological symptoms are associated in suicide attempt

survivors, and whether is there some associations structure between the psychological symptoms and the empathic profiles of suicide attempts. Although there has been some preliminary evidence on the relation and intensity among these two aspects in healthy individuals, a statistical assessment on a sample of suicide attempt is still lacking, and can provide important insights to understand dynamics underlying such a tragic act.

Latent structures models for removing dependence

Traditionally, statistical modelling through latent variables is motivated as a strategy to characterize intricate dependence structures in an unobservable latent spaces, specifying the observed quantities as conditional independent given such latent features. This approach is also the leading strategy of Chapter 1 and Chapter 2, where the benefits in terms of computational efficiency and interpretation are highlighted and illustrated through different case studies. Chapter 3, instead, focuses on a different setting in which latent variables are introduced as an effective tool to manipulate complex dependence structures at multiple levels.

In high-stakes decision processes, there has been growing interest on providing machine learning tools to guarantee that algorithms will be neutral with respect to "protected" information; for example, race or gender of the candidates during job interviews. It has been argued that decisions taken by algorithms would automatically be fair since personal prejudices carried by humans cannot influence automated algorithms, which only rely on computers. However, there is growing recognition that algorithms will reproduce the same biases observed in the training data. For example, if a company has been affected by gender-gap in the salaries, then automated algorithms will "learn" this aspect and reproduce it in predictions, therefore propagating the issues in future samples instead of reducing it.

An illustrative case study from criminal justice is described in Section 3.2, where interest is on predicting two years recidivism for defendants as a function of several demographic information. Specifically, such predictive tools are used in American courtrooms to advise judges on the likelihood that a defendant will be rearrested in the future, and such predictions are used to inform the court; for example, to decide the amount of the bail. Clearly, there is strong ethical interest in making such predictions independent of protected information – such as race membership – in order to guarantee that individuals will be treated the same regardless their racial group. Algorithms to predict recidivism are generally trained on arrest data, and there is strong evidence that such source of data is drastically biased with respect to race, since police patrols can choose

which neighborhood should be patrolled and who looks more suspicious. Without adjustment, predictions of algorithms trained on such data would automatically reproduce the same racial biases observed in the training sample, therefore providing predictions for minority groups which are systematically higher.

In order to solve these issues, Section 3.3 focuses on latent structure modelling for fair decision processes, introducing a Gaussian latent factor model which allows to characterise the dependence structure of the data via a set of low-dimensional latent variables, and facilitates to constrain the dependence among such a compact representation and the protected attributes during estimation. The proposed algorithm is fast and efficient, and allows to release adjusted dataset where practitioner can apply any algorithm with guarantees in terms of independence among resulting predictions and protected information.

Chapter 1

Latent space models for network data

1.1 Network data

A challenging field of research in which latent variable modelling leads to tractable and interpretable specification of complex dependence structures is *network science*. In general terms, a network can be defined as a collection of interconnected units, while from a mathematical perspective it is more rigorous to represent a network as a *graph* $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, which is a collection of two sets denoting the interconnected units and their connections, respectively. More specifically, $\mathcal{V} = \{1, \dots, n\}$ denotes the set of $n = |\mathcal{V}|$ interconnected *nodes* while $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ the *edges* interconnecting them; see Newman (2018) and references therein for an introduction to the topic. From a modelling perspective, it is more useful to represent a network as an $n \times n$ *adjacency* matrix $\mathbf{A}$ with elements a_{ij} characterising the edge from node i to node j. Such a representation is conceptually equivalent to a graph, but provides a simpler way to focus on the random nature of the edges and to properly model them as elements of an interconnected structure. In this chapter the focus will be on *undirected binary networks*, which are represented by symmetric binary matrices indicating the presence or absence of a connection between nodes. Since such adjacency matrices are symmetric, it is sufficient to characterise the lower-triangular part of $\mathbf{A}$, thereby letting $a_{ij} = a_{ji} = 1$ if there is and edge between the pair (i, j) with $i = 2, \dots, n$ and $j = 1, \dots, i-1$ and 0 otherwise.

From a statistical perspective, the main challenges in network modelling arise from the characterisation of the intricate dependence underlying such data. Since such connectivity structure defines and characterise the network as such, forcing network data

into standard structures or assuming oversimplified specifications leads to poor characterisations and tremendously biased estimates (e.g. Hoff *et al.*, 2002). On the other hand, it is necessary to reduce the complexity of those structures and characterise the main properties of the network with a parsimonious representation, comprising a moderate number of parameters performing a significant dimensionality reduction. Latent structures modelling, adapted to the context of network data, successfully addresses these two issues. Such methods assume *conditional* independence among the edges, given a latent structure which characterises the dependence patterns. Several specifications are available for both the latent variables and the conditional model, covering a large variety of methods. For example, discrete latent variables induce stochastic block-models (Nowicki and Snijders, 2001), while continuous variables define latent space models (Hoff *et al.*, 2002). More recent generalisations provide additional layers of complexity, such as mixed-membership stochastic block-model (Airoldi *et al.*, 2008) and latent space models with nodes clustering (Handcock *et al.*, 2007). See also Durante *et al.* (2017); Gollini and Murphy (2016) for recent extensions and generalisation to multiple networks. Models based on conditional independence assumptions are particularly appealing from a computational point of view, solving several estimation issues encountered in other popular models such as Exponential Random Graph Models (ERGM) (Hunter *et al.*, 2012). Beside this important aspect, latent variables offer important benefits from interpretative point of view; for example, it is quite natural to interpret discrete latent variables as clusters of nodes, while continuous latent variables are interpreted as positions in a low-dimensional latent space amenable for data visualisation.

1.2 Data description and motivation

As outlined in the Introduction chapter, neuroscience provides fascinating network data stimulating challenging research questions and novel statistical methodologies to address them. A particularly active area of research in network science is motivated by recent developments in neuroscience, where modern advances in neuro–imaging have made it possible to measure brain connectivity non–invasively in live humans (e.g. Craddock *et al.*, 2013; Smith *et al.*, 2011). These network data commonly denote functional synchronization in brain activity, measured via fMRI (e.g. Smith *et al.*, 2011), or structural interconnections among brain regions made by white matter fibers reconstructed from DTI (e.g. Craddock *et al.*, 2013). Refer also to Bullmore and Sporns (2009, 2012); Sporns (2013); Stam (2014) for an overview on state–of–the–art network data in neuroscience and the associated methods of analysis. Recently, there has been an increasing interest

on structural connectivity. Indeed, these data measure the physical connections among brain regions which provide the fundamental axonal pathways for brain signals (e.g. Craddock *et al.*, 2013). In particular, nodes of structural brain networks correspond to a specific brain parcellation providing a set of regions of interest, whereas the edges represent physical interconnections made by white matter fibers (e.g. Craddock *et al.*, 2013).

In this chapter, we analyze two problems involving structural brain networks data from two different studies. The fist application is drawn from the neuro–imaging study KKI-21, comprising data for $m = 21$ individuals with no history of neurological disease (Landman *et al.*, 2011). Here the focus is on the structural brain networks obtained by pre–processing the raw imaging data under the MIGRAINE pipeline (Roncal *et al.*, 2013), focusing on the $n = 68$ brain regions characterizing the Desikan atlas (Desikan *et al.*, 2006) The second application is drawn form the study described in Hagmann *et al.* (2008) available at HCP (2019). Specifically, the study focuses on extremely high-resolution scans with $n = 998$ brain regions obtained dividing the anatomical cortex into sections of about 1.5 cm^2 width for $m = 5$ healthy subjects. See Hagmann *et al.* (2008) for detailed information on the preprocessing, the pipelines and the construction of the structural brain network.

In both studies, the brain network of each individual k is available via a $n \times n$ symmetric adjacency matrix $\mathbf{A}^{(k)}$ having elements $a_{ij}^{(k)} = a_{ji}^{(k)} = 1$ if at least one white matter fiber has been observed between regions $i = 2, \ldots, n$ and $j = 1, \ldots, i-1$ in subject $k = 1, \ldots, m$, and $a_{ij}^{(k)} = a_{ji}^{(k)} = 0$ otherwise; see Figure 1.1 for an illustration in the first case study. Additionally, for each region i anatomical covariates are available. These predictors comprise the 3-dimensional anatomical centroid positions, denoted as x_i, y_i and z_i, for $i = 1, \ldots, n$, membership to the left or right hemisphere, and in which

FIGURE 1.1: Adjacency matrices characterizing the brain networks of two subjects. Black refers to an edge; white to a non-edge.

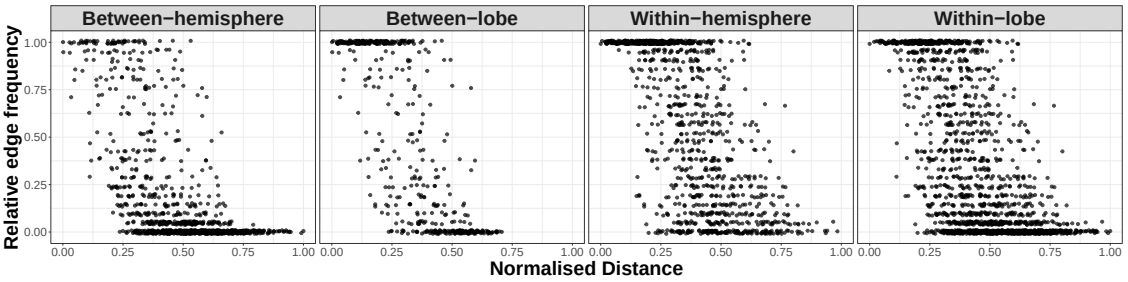

FIGURE 1.2: Plot of the relative edge frequencies for every pair of brain regions (i, j), with $i = 2, \ldots, n$ and $j = 1, \ldots, i-1$, versus their anatomical normalized Euclidean distance.

anatomical lobe (in the first application) or cortical cortex (in the second) the region is located.

The aim of both applications is to learn shared anatomical effects and latent structures underlying the replicated brain networks $\mathbf{A}^{(1)}, \ldots \mathbf{A}^{(k)}$. In this regard, Figure 1.2 focuses on the first application and suggests that brain regions tend to connect with others that are spatially closer, and belonging to the same hemisphere and lobe. However, these patterns are not sufficient to explain the whole variability in regions connectivity, thereby motivating studies on the net wiring structures which are not related to the observed anatomical covariates. A first step towards addressing this question is taken via latent space models for replicated brain network data, which characterize the edge probabilities as a function of observed and latent effects specific to every brain region. The former allows inference on how brain anatomy—measured via physical proximity, along with hemispheres and lobe membership—relates to brain connectivity. The latter characterizes, instead, patterns not captured by exogenous predictors, thus stimulating future studies to explain these departures via alternative determinants, such as regions' morphology or other biological processes. In the first case study, we develop a latent space model with local clustering, which allows to address the aims discussed above while providing further insights on the brain architecture not explained by anatomical constraints partitioning the brain regions in the latent space. The model is illustrated and quantitatively evaluated in Section 1.4. The second case study is analysed in Section 1.5, and motivates more computationally tractable algorithms leveraging latent factor models for networks, which effectively address the main aims of the applications for high-quality scans.

1.3 Latent space models for network data

Before going into the illustration of the contributions for this Chapter, it is worth recalling some details on the state-of-the-art modelling for network data via latent structures.

TABLE 1.1: Latent variable specification for network data under Equation (1.1).

	$\alpha(\mathbf{w}_i, \mathbf{w}_j)$	Additional parameters
Stochastic block-model	$\mathbf{w}_i^\intercal \boldsymbol{\Theta} \mathbf{w}_j$	$\boldsymbol{\Theta}$ symmetric
Latent distance model	$\|\|\mathbf{w}_i - \mathbf{w}_j\|\|_2$	-
Latent factor model	$\mathbf{w}_i^\intercal \boldsymbol{\Lambda} \mathbf{w}_j$	$\boldsymbol{\Lambda}$ diagonal

As outlined in Section 1.1, there is a rich literature on latent structure modelling for networks, with stochastic block models (Nowicki and Snijders, 2001) and latent distance models (Hoff *et al.*, 2002) being popular building blocks for many complex approaches. Focus for the moment, and without loss of generality, on a single binary network with associated adjacency matrix $\mathbf{A}$, and let $\text{pr}(a_{ij} = 1 \mid \pi_{ij}) = \pi_{ij} \in (0, 1)$, denote the population probability of an edge between node i and j, for each $i = 2, \ldots, n$, $j = 1, \ldots, i-1$. Following Hoff (2019), latent structure models for binary undirected networks can be defined with a common specification as follows.

$$\begin{aligned} (a_{ij} \mid \pi_{ij}) &\sim \text{Bern}(\pi_{ij}) \\ g(\pi_{ij}) &= \alpha(\mathbf{w}_i, \mathbf{w}_j) \end{aligned} \tag{1.1}$$

where α is a function of the node-specific latent features $\mathbf{w}_i \in \mathbb{R}^r$, $i = 1, \ldots, n$ and g is a link function which guarantees that $\pi_{ij} \in (0, 1)$, as in a classical Generalized Linear Model (GLM) specification. Generalisation to directed networks and weighted edges are straightforward, directly adapting Equation (1.1) to the undirected case or modifying the stochastic component accordingly.

Table 1.1 illustrates how different latent variable models can be represented using Equation (1.1). Specifically, stochastic block models correspond to discrete latent variables $\mathbf{w}_i \in \{0, 1\}^r$ indicating block-membership, with an $r \times r$ symmetric matrix $\boldsymbol{\Theta}$ specifying in-block specific probabilities over the main diagonal and between-block probabilities outside. Latent distance models are instead recovered by introducing latent positions $\mathbf{w}_i \in \mathbb{R}^r$ and computing the Euclidean distance between pairs of nodes in the latent space. Such an approach will be the building block of the next Section 1.4, where we generalize it in order to model multiple networks, include covariate information and perform clustering over the latent space. Finally, the latent factor model can be interpreted as an eigenvalue decomposition of the symmetric adjacency matrix, where $\mathbf{w}_i \in \mathbb{R}^r$ denotes its eigenvectors and $\boldsymbol{\Lambda}$ the diagonal $r \times r$ matrix of eigenvalues. We focus on such a flexible representation in Section 1.5, providing a computationally efficient algorithm to perform approximate posterior inference with high-resolution scan

data. Bayesian inference is generally performed via Markov Chain Monte Carlo (MCMC, adapting the data augmentation strategy of Albert and Chib (1993) in case of a probit link function g or using a Metropolis-Hasting routine. More recent developments focus instead on approximate inference for network models, leveraging Variational Bayes or approximate MCMC (Gollini and Murphy, 2016; Airoldi *et al.*, 2008; Lachal *et al.*, 2016).

Stochastic block models have received much attention for at least the last decade. This popularity is mainly due to the ease of interpretation of such an approach, which allows to cluster nodes into groups as a byproduct of model estimation. Indeed, it has been observed that many real networks exhibits important clustering behaviour, with such a grouping driven by endogenous information or unobservable attributes (e.g. Wasserman and Faust, 1994). Motivated by the above consideration, Handcock *et al.* (2007) extended the latent distance model of Hoff *et al.* (2002) introducing a mixture of Gaussian prior over the latent coordinates $\mathbf{w}_i$, $i = 1, \dots, n$, in order to induce a model-based clustering of observations in the latent space. Specifically, Handcock *et al.* (2007) rely on the following specification.

$$\begin{aligned} (a_{ij} \mid \pi_{ij}) &\sim \text{Bern}(\pi_{ij}) \\ \text{logit}(\pi_{ij}) &= ||\mathbf{w}_i - \mathbf{w}_j||_2 \\ \mathbf{w}_i &\sim \sum_{h=1}^{H} \nu_h \text{N}_r(\boldsymbol{\mu}_h, \sigma_h \mathbf{I}_r) \end{aligned} \tag{1.2}$$

independently for each $i = 2, \dots, n$, $j = 1, \dots, i-1$ and $k = 1, \dots, m$. The model defined in Equation (1.2) allows to cluster nodes in the latent space into H spherical groups according to their position, and improves estimation of the latent structure by facilitating borrowing of information across nodes (Handcock *et al.*, 2007). In the next Section, we extend such an approach to deal with multiple networks, include covariate information and improve the clustering scheme induced by the prior distribution.

1.4 Latent space model with local clustering

1.4.1 Model specification

Let $\text{pr}(a_{ij}^{(k)} = 1 \mid \pi_{ij}) = \pi_{ij} \in (0, 1)$, denote the population probability of an edge between brain regions i and j, for each $i = 2, \dots, n$, $j = 1, \dots, i-1$ and $k = 1, \dots, m$. The focus of this chapter is on a flexible representation for π_{ij} which can learn anatomical effects and latent patterns in the observed brain network data. Adapting latent variable models for networks, the edges $a_{ij}^{(k)} = a_{ji}^{(k)}$, $k = 1, \dots, m$ are assumed as conditionally

independent Bernoulli variables given π_{ij}, thus obtaining

$$(a_{ij}^{(k)} \mid \pi_{ij}) \sim \text{Bern}(\pi_{ij}), \tag{1.3}$$

independently for each $i = 2, \ldots, n$, $j = 1, \ldots, i-1$ and $k = 1, \ldots, m$. To flexibly characterize variations in π_{ij} across pairs of nodes, while learning anatomical and latent patterns in the edge–specific probabilities, we introduce the logistic model

$$\text{logit}(\pi_{ij}) = \beta_0 + \beta_1 \texttt{hem}_{ij} + \beta_2 \texttt{lobe}_{ij} + \beta_3 d_{ij} - \bar{d}_{ij}, \tag{1.4}$$

for each pair $i = 2, \ldots, n$ and $j = 1, \ldots, i-1$, where $\texttt{hem}_{ij}$ and $\texttt{lobe}_{ij}$ are binary predictors indicating shared membership to the same hemisphere and lobe, respectively, whereas d_{ij} and $\bar{d}_{ij}$ denote the anatomical and *latent* Euclidean distances between brain regions i and j, respectively.

Based on Equation (1.4), the edge probability between regions i and j is allowed to change with their shared hemisphere and lobe membership, in addition to their anatomical distance d_{ij}. The parameter $\bar{d}_{ij}$ adds instead a further layer of flexibility, which allows modeling of edge probabilities that are not properly explained by brain anatomy. To address this issue, $\bar{d}_{ij}$ is expressed as a function of the regions' coordinates $\bar{x}_i$, $\bar{y}_i$ and $\bar{z}_i$, for $i = 1, \ldots, n$ in a latent space, via $\bar{d}_{ij} = \sqrt{(\bar{x}_i - \bar{x}_j)^2 + (\bar{y}_i - \bar{y}_j)^2 + (\bar{z}_i - \bar{z}_j)^2}$, thus obtaining a more parsimonious and interpretable formulation which borrows information via the shared dependence on a common set of latent coordinates. Indeed, according to Equation (1.4), the closer two brain regions are within this latent space, the more likely it is to observe a connection among them, after controlling for the anatomical structure. Hence, by providing inference on the latent positions $\bar{x}_i$, $\bar{y}_i$ and $\bar{z}_i$ for each brain region $i = 1, \ldots, n$, a deeper understanding of brain connectivity is allowed.

The statistical model in Equations (1.3) and (1.4) is in the same spirit of the flexible latent space model proposed by Hoff *et al.* (2002), and therefore is characterized by similar properties and theoretical support. However, differently from Hoff *et al.* (2002), the focus in this chapter is on joint modeling of multiple adjacency matrices $\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(m)}$, instead of just one. Although this difference may apparently require novel computational methods and inference procedures, note that, under the model in Equation (1.3), a sufficient statistic a_{ij} in the joint likelihood for the data $a_{ij}^{(k)}$, $k = 1, \ldots, m$, is $a_{ij} = \sum_{k=1}^{m} a_{ij}^{(k)} \sim \text{Binom}(m, \pi_{ij})$, for every pair of regions $i = 2, \ldots, n$ and $j = 1, \ldots, i-1$. Hence, joint modeling of multiple adjacency matrices $\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(m)}$ under the model in Equation (1.3)–(1.4), coincides with providing inference on the weighted network $\mathbf{A} = \sum_{k=1}^{m} \mathbf{A}^{(k)}$ with binomial edges having probabilities factorized

as in Equation (1.4).

1.4.2 Bayesian inference

A Bayesian approach to estimation and inference is followed. Consistent with this choice, the specification of a prior distributions for the coefficients in $\boldsymbol{\beta}$ and for the latent space coordinates $\bar{x}_i$, $\bar{y}_i$, $\bar{z}_i$ for $i = 1, \ldots, n$ in Equation (1.4) is seek to obtain a flexible, computationally tractable and interpretable characterization of the edge probabilities in the model defined in Equation (1.3). Indeed, besides studying the embedding structure of the regions in the latent space, a recent focus in neuroscience is on estimating groups of brain regions which are devoted to locally specialized processes (Bullmore and Sporns, 2009). Within the proposed Bayesian approach to inference, this aim can be addressed by considering a finite mixture of distributions as prior for the latent space coordinates, thus allowing regions with similar positions to cluster together. This choice also improves borrowing of information in inference on the latent coordinates by exploiting the grouping structure.

One possibility to accomplish the above goal is to rely on the latent space model with nodes clustering proposed by Handcock *et al.* (2007) and illustrated in Equation (1.2). In the model considered in this Chapter, this choice would imply that the 3–dimensional latent space can be partitioned into spherical groups, with regions in the same cluster having similar vectors of latent coordinates. Although this strategy improves flexibility and offers insights on community structures, clustering regions with respect to the join vector of latent coordinates $(\bar{x}_i, \bar{y}_i, \bar{z}_i)$, $i = 1, \ldots, n$, might provide an oversimplified characterization of complex connectivity patterns via a single joint partition. Indeed, joint clustering of the entire vector of latent coordinates might fail to detect groups of brain regions which are similar with respect to a subset of the latent traits, but are significantly different relatively to the others. This issue is particularly important for brain networks, since it is well known in the literature that regions co-operate in a large variety of different tasks. Incorporating this structure under a single joint clustering process would, in fact, lead to an inefficient allocation into too many clusters, or to an oversimplified representation via fewer groups with high within–cluster variability.

The above issue is addressed by performing separate clustering for each latent space dimension via mixtures of univariate Gaussian priors for every $\bar{x}_i$, $i = 1, \ldots, n$, $\bar{y}_i$, $i = 1, \ldots, n$, and $\bar{z}_i$, $i = 1, \ldots, n$, thus allowing the brain regions to belong to different clusters depending on the latent coordinate considered. This structure provides a more parsimonious, yet flexible, local borrowing of information, which relies on a separate

partition structure for each latent coordinate. Consistent with this assumption, we let

$$\begin{aligned}
\bar{x}_i &\sim P_x, \quad P_x = \sum_{h=1}^{H_{\mathbf{x}}} \nu_{\mathbf{x}_h} \mathrm{N}(\mu_{\mathbf{x}_h}, \sigma^2_{\mathbf{x}_h}), \quad i = 1, \ldots, n, \\
\bar{y}_i &\sim P_y, \quad P_y = \sum_{h=1}^{H_{\mathbf{y}}} \nu_{\mathbf{y}_h} \mathrm{N}(\mu_{\mathbf{y}_h}, \sigma^2_{\mathbf{y}_h}), \quad i = 1, \ldots, n, \\
\bar{z}_i &\sim P_z, \quad P_z = \sum_{h=1}^{H_{\mathbf{z}}} \nu_{\mathbf{z}_h} \mathrm{N}(\mu_{\mathbf{z}_h}, \sigma^2_{\mathbf{z}_h}), \quad i = 1, \ldots, n,
\end{aligned} \tag{1.5}$$

where $\nu_{\mathbf{x}_h} \in (0, 1)$ is the probability that the latent $\bar{x}$–coordinate of a generic brain region belongs to cluster h, with $\sum_{h=1}^{H_{\mathbf{x}}} \nu_{\mathbf{x}_h} = 1$. The parameters $(\mu_{\mathbf{x}_h}, \sigma^2_{\mathbf{x}_h})$ characterize, instead, the mean and variance of this coordinate in cluster h. A similar interpretation holds for $\nu_{\mathbf{y}_h}$, $(\mu_{\mathbf{y}_h}, \sigma^2_{\mathbf{y}_h})$ and $\nu_{\mathbf{z}_h}$, $(\mu_{\mathbf{z}_h}, \sigma^2_{\mathbf{z}_h})$, with respect to the latent $\bar{y}$–coordinate and $\bar{z}$–coordinate, respectively.

Note that priors in Equation (1.5) and the model in Equations (1.3) and (1.4) effectively generalize Hoff *et al.* (2002) and Handcock *et al.* (2007). Indeed, the standard latent space model of Hoff *et al.* (2002) is obtained as a degenerate case when the number of components $H_{\mathbf{x}} = H_{\mathbf{y}} = H_{\mathbf{z}}$ is set to 1, whereas the mixture model specification of Handcock *et al.* (2007) described in Equation (1.2) can be regarded as a particular case of Equation (1.5) in which the cluster membership is identical across the latent coordinates. It is worth emphasizing that such a single joint partition can be obtained, when required, by summarizing the marginal partitions into a single global clustering index. In fact, the 3 marginal assignments can be combined to create a joint similarity matrix $\mathbf{C}$ with elements $c_{ij} \in \{0, 1, 2, 3\}$ denoting the number of marginal partitions in which regions i and j share the same cluster. Based on $\mathbf{C}$, a single global grouping structure can be then obtained by applying classical clustering methods.

Following Rousseau and Mengersen (2011), a conservative upper bound H for $H_{\mathbf{x}}$, $H_{\mathbf{y}}$, $H_{\mathbf{z}}$ is specified, with adaptive deletion of redundant components favored via a sparse Dirichlet prior for the probabilities $\boldsymbol{\nu}_{\mathbf{x}} = (\nu_{\mathbf{x}_1}, \ldots, \nu_{\mathbf{x}_H})$, $\boldsymbol{\nu}_{\mathbf{y}} = (\nu_{\mathbf{y}_1}, \ldots, \nu_{\mathbf{y}_H})$, $\boldsymbol{\nu}_{\mathbf{z}} = (\nu_{\mathbf{z}_1}, \ldots, \nu_{\mathbf{z}_H})$. This choice provides

$$\begin{aligned}
\boldsymbol{\nu}_{\mathbf{x}} &\sim \text{Dirichlet}\left(\frac{1}{H}, \ldots, \frac{1}{H}\right), \\
\boldsymbol{\nu}_{\mathbf{y}} &\sim \text{Dirichlet}\left(\frac{1}{H}, \ldots, \frac{1}{H}\right), \\
\boldsymbol{\nu}_{\mathbf{z}} &\sim \text{Dirichlet}\left(\frac{1}{H}, \ldots, \frac{1}{H}\right).
\end{aligned} \tag{1.6}$$

The prior for the means and variances of the Gaussian kernels in Equation (1.5) is instead specified to favor simple posterior computation. Due to this, Normal–Inverse Gamma priors with common hyperparameters are considered. Focusing on the $\bar{x}$–coordinate, this choice implies

$$(\mu_{\mathbf{x}_h} \mid \sigma^2_{\mathbf{x}_h}) \sim \mathrm{N}(\mu_0, \sigma^2_{\mathbf{x}_h}/\kappa_0), \quad \sigma^{-2}_{\mathbf{x}_h} \sim \mathrm{Gamma}(\eta_0/2, \eta_0\xi_0/2), \tag{1.7}$$

for every component $h = 1, \ldots, H$, with μ_0 denoting the mean of the prior for the kernels' locations, κ_0 controlling the precision and (η_0, ξ_0) denoting the hyperparameters for the prior of the kernel variances. The priors for $(\mu_{\mathbf{y}_h}, \sigma^2_{\mathbf{y}_h})$ and $(\mu_{\mathbf{z}_h}, \sigma^2_{\mathbf{z}_h})$, are defined similarly to (1.7), for each $h = 1, \ldots, H$.

To conclude prior specification, the prior for the coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^\intercal$ is specified as multivariate Gaussian, obtaining

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^\intercal \sim \mathrm{N}_4(0, \boldsymbol{\Lambda}_0), \quad \boldsymbol{\Lambda}_0 = \mathrm{diag}(\lambda_0, \ldots, \lambda_3). \tag{1.8}$$

Due to the absence of a closed–form posterior, Bayesian inference for model defined in Equations (1.3) and (1.4) with priors Equations (1.5) and (1.8), proceeds via a (MCMC) strategy, whose key steps are summarized in Appendix A.1. The implementation benefits from the recently developed Pólya–Gamma data augmentation strategy (Polson *et al.*, 2013; Choi and Hobert, 2013) which allows recasting the logistic model defined in Equations (1.3) and (1.4) into a classical Bayesian regression having transformed Gaussian responses. This representation induces full–conditional conjugacy for all the model parameters except the latent coordinates, due to the Euclidean distance in Equation (1.4) which requires a Metropolis–Hastings step.

It is worth noticing that the latent positions are only identifiable up to translation, rotation and reflection. Handcock *et al.* (2007) address this issue via unity–norm constraints. Alternatively, it is possible to post–process the unconstrained samples via a Procrustean transform with the maximum likelihood estimate of the latent positions as reference vector (e.g. Krivitsky and Handcock, 2008). In fact, consistent with Krivitsky and Handcock (2008), also the empirical findings provided in this chapter suggest that unconstrained sampling improves computational tractability and performance. Moreover, since the Procrustean rotation is unique, the transformed samples coincide with draws from a constrained posterior (Hoff *et al.*, 2002). Finally, it is well known in the literature that label–switching issues of mixture models can affect posterior inference on the clustering structure. This issue is induced by the invariance of the likelihood function under different labelling of the mixture components (Stephens, 2000). The

presence of such an issue can be easily detected via visual inspection of the trace-plots of the MCMC chains; standard relabelling procedures can be applied when this issue arises (e.g. Stephens, 2000).

1.4.3 Simulation study

To asses the empirical performance of the proposed methods, a simulation study is considered. In particular, $m = 21$ replicated network data with $n = 68$ nodes are simulated from model defined in Equations (1.3) and (1.4), also including one dummy covariate mimicking the hemisphere membership; its associated β effect is set to 2. The latent space structure is instead assumed of growing complexity across the scenarios, in order to evaluate whether the proposed model and its priors can flexibly detect varying latent space architectures.

In the first scenario, the latent $\bar{x}$–coordinates are sampled from a mixture of two Gaussians, while the $\bar{y}$–coordinates and the $\bar{z}$–coordinates are generated from standard Gaussians. The second scenario considers instead a mixture of two Gaussians for the latent $\bar{x}$–coordinates and the latent $\bar{y}$–coordinates, and samples the $\bar{z}$–coordinates from a standard Gaussian. Finally, in the last scenario, all the three latent space coordinates are generated from a mixture of two Gaussians. The main challenge in these scenarios is due to different local complexities and varying cluster membership across latent dimensions. Indeed, even when the number of clusters is equal across the latent dimensions, the underlying group partitions can be different across the latent coordinates. Indeed, the true number of mixture components $(H^0_{\mathbf{x}}, H^0_{\mathbf{y}}, H^0_{\mathbf{z}})$ is equal to $(2, 1, 1)$ in the first scenario, $(2, 2, 1)$ in the second, and $(2, 2, 2)$ in the third.

Posterior inference is conducted by relying on default hyperparameters $\mathbf{\Lambda}_0 = \text{diag}(2, 2)$, $\kappa_0 = 2$, $\mu_0 = 0$ $\eta_0 = 30$, $\xi_0 = 1$, and set the upper bound H to 5. Moderate variations of the hyperparameters in sensitivity analyses did not lead to substantially different conclusions. This robustness to hyperparameters is possibly due to the borrowing of information provided by the latent space embedding and the clustering of the coordinates.

Posterior computation relies on 5000 iterations with a burn–in of 2500 and a tuned step size for Metropolis–Hastings steps to obtain an acceptance ratio close to 0.2 (e.g., Gelman *et al.*, 2014, 1996). Relabelling, when required, is performed via the R package `label.switching` (Papastamoulis, 2016; Stephens, 2000). Convergence and mixing are assessed via the trace–plots and effective sample sizes.

Figure 1.3 compares the true latent positions with their posterior means estimated from the MCMC samples. The mean square error among the true latent positions and

FIGURE 1.3: Bivariate plots for the latent positions and cluster membership in the simulation study. True coordinates are reported as light gray crosses. The estimated latent coordinates are illustrated as full points, with colors and shapes denoting the estimated cluster membership. First, second and third column display the estimated cluster membership induced by the $\bar{x}$–coordinate, $\bar{y}$–coordinate and $\bar{z}$–coordinate, respectively. First, second and third row refer to the first, second and third scenario, respectively.

their posterior estimates is 0.03 in the first scenario, 0.16 in the second, and 0.18 in the third, respectively. As expected, when the complexity of the latent space increases, the precision in recovering the underlying structure deteriorates. However, in each scenario, the posterior means of the latent positions still provide satisfactory estimates of the true latent structure. Estimates were satisfactory also for the number of clusters and the parameter β; see Appendix A.1.2 and Aliverti and Durante (2019) for additional details.

1.4.4 Application to the KKI-21 dataset

The methods described in Sections 1.4.1 and 1.4.2 are applied to the KKI-21 dataset outlined in Section 1.2. Posterior inference is performed with the same hyperparameters as in the simulation study, with a more conservative upper bound $H = 10$ and relying on 5000 MCMC samples with a burn–in of 2500, obtaining satisfactory convergence and mixing, measured via effective sample sizes. Posterior computation is based on a simple R implementation and requires approximately 2 minutes per 1000 iterations on a laptop with an INTEL(R) CORE(TM) I7-7700HQ @ 2.8 GHZ processor and 16GB of RAM running Linux.

Figure 1.4 highlights the gains in model flexibility and interpretability that can be obtained by characterizing the net connectivity structure via a latent space representation. This is done by comparing the performance of model defined in Equations (1.3) and (1.4) with a purely anatomical specification holding out in Equation (1.4) the latent distances $\bar{d}_{ij}$, for each $i = 2, \ldots, n$ and $j = 1, \ldots, i-1$. According to the left matrix in Figure 1.4, anatomical information have an effect in modeling edge probabilities, but are not sufficient to capture specific wiring mechanisms. In fact, the model based on purely anatomical predictors overestimates the intrahemispheric connectivity patterns associated with regions in the temporal lobe, while underestimating interhemisperic wiring mechanisms in the parietal and occipital lobes—among others. As shown in the right

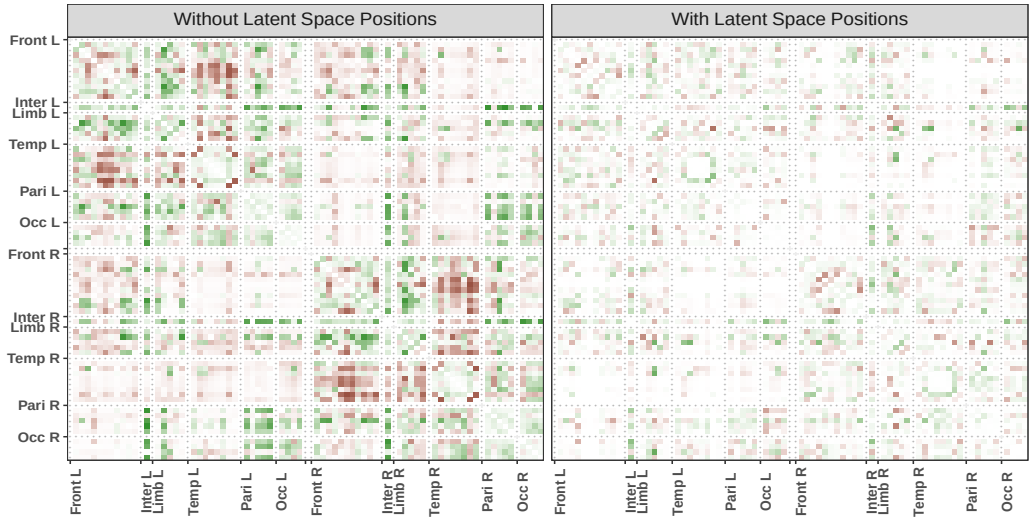

FIGURE 1.4: Graphical representation of the difference between the observed edge frequencies $\sum_{k=1}^{21} a_{ij}^{(k)}/21$ and the posterior mean $\mathrm{E}(\pi_{ij} \mid \mathbf{A})$ of the corresponding edge probabilities under model (1.3)–(1.4) with (right matrix) and without (left matrix) the latent space effects in (1.4). Brain regions are grouped by combinations of lobe and hemisphere membership. Colors range from dark red to dark green as the differences go from -1 to $+1$.

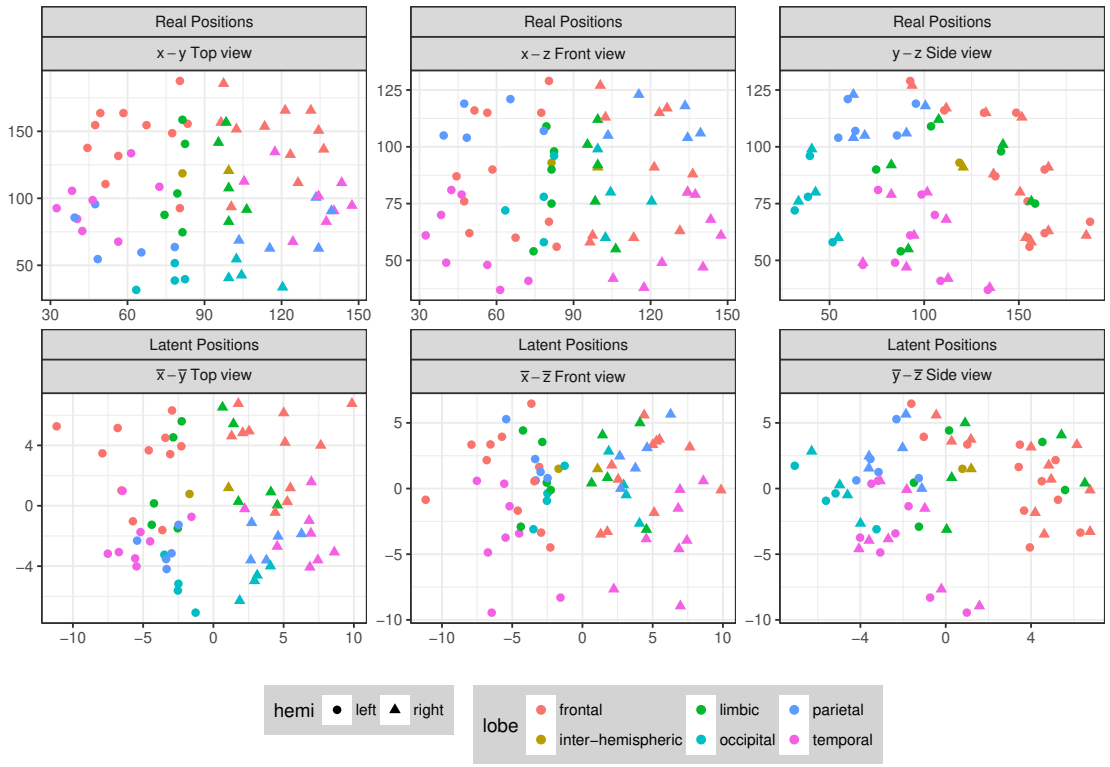

FIGURE 1.5: Graphical representation of the anatomical coordinates (upper panel) and the posterior means of the latent positions (lower panel). Shapes and colors refer to hemisphere and lobe membership.

matrix of Figure 1.4, incorporating a latent space component substantially increases performance in modeling edge probabilities, with the latent coordinates flexibly capturing net connectivity patterns not subject to anatomical constraints. These findings, which are specific to the proposed model construction, can provide relevant insights for neuroscientists, thus stimulating novel studies to explain these departures via alternative anatomical or morphological determinants.

These results motivate inference on the posterior mean $\hat{\bar{x}}_i$, $\hat{\bar{y}}_i$ and $\hat{\bar{z}}_i$, $i = 1, \ldots, n$ of the 3–dimensional latent coordinates, which are compared with the corresponding anatomical ones in Figure 1.5. According to Figure 1.5, the 3–dimensional anatomical and latent positions are interestingly related. However, as expected, the coordinates of the regions in the latent space do not coincide with the anatomical positions, and characterize those net connectivity patterns after controlling for anatomical constraints. For example, consistent with results in Figure 1.4, the brain regions in the temporal lobe have a more peripheral position in the latent space, whereas regions in the parietal and occipital lobes become more central. These findings, provided by the latent space embedding, suggest that the connectivity structures which are not explained by the observed anatomical covariates might still have a relation with the physical proximity

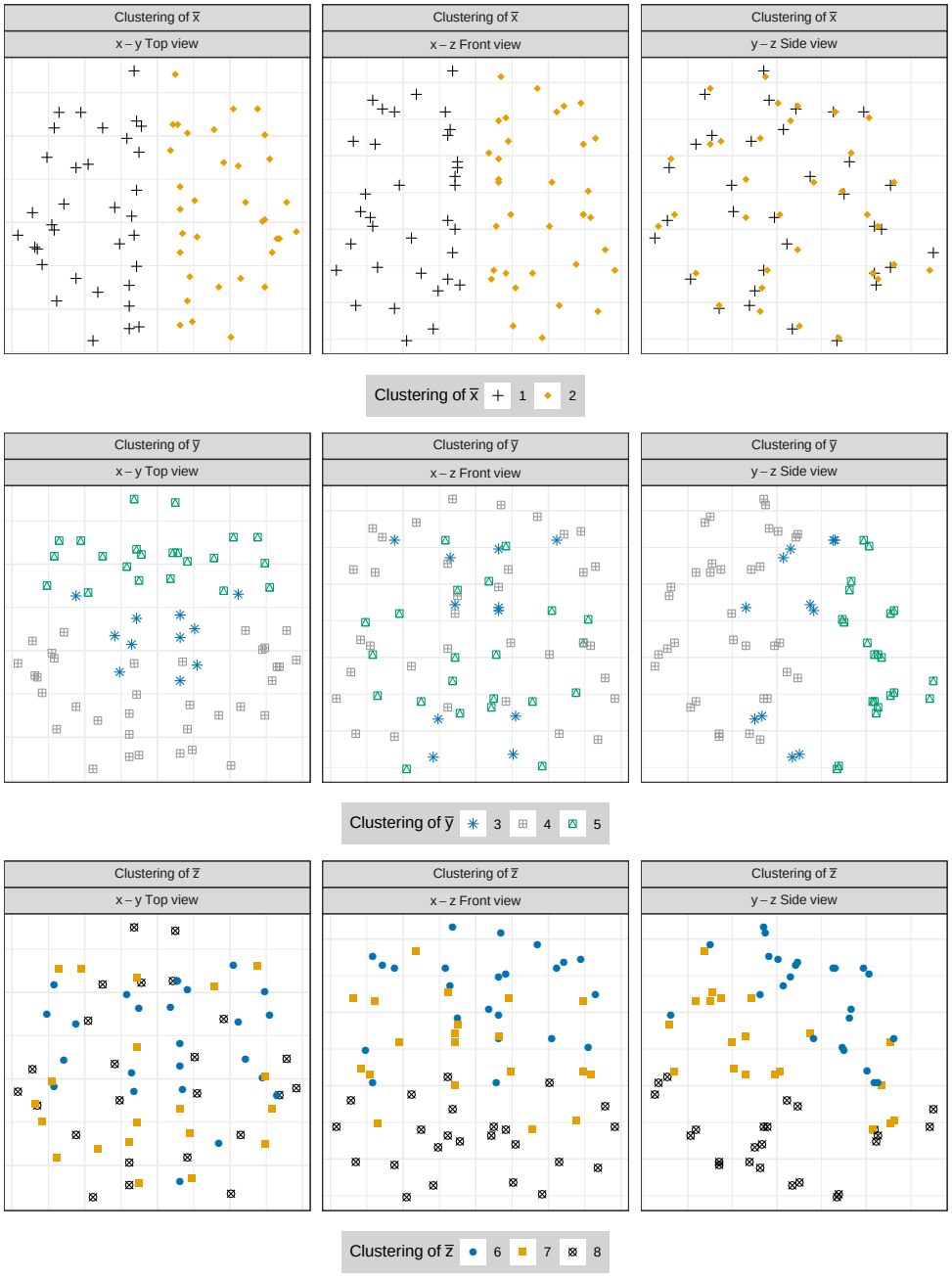

FIGURE 1.6: Graphical representation of the estimated cluster partitions and anatomical coordinates. Colors and shapes are defined by the estimated cluster membership for the $\bar{x}$–coordinate (first column), $\bar{y}$–coordinate (second column) and $\bar{z}$–coordinate (third column). The locations of the brain regions are given by their anatomical centroids. Each row corresponds to a different bivariate view of the 3–dimensional brain locations.

among the regions. This might be due, for instance, to local neighboring structures, but additional determinants should be considered to fully explain structural brain connectivity, such as the shape and morphology of the brain regions.

To conclude the analysis, Figure 1.6 compares the results from the clustering scheme with the anatomical coordinates of the brain regions. The first column of Figure 1.6

TABLE 1.2: Summaries of the posterior distribution for the parameters in $\boldsymbol{\beta}$.

	Mean	Median	Std. Dev.	Cred. Int.$_{95\%}$
`Intercept`	7.27	7.27	0.18	(6.94, 7.60)
`hemisphere`	0.60	0.61	0.18	(0.29, 0.92)
`lobes`	0.24	0.24	0.06	(0.13,0.35)
`distance`	-0.35	-0.35	0.06	(-0.47,-0.23)

represents the partition induced by the latent $\bar{x}$–coordinate, and clearly encodes the brain segmentation into the two hemispheres. This result is coherent with Figure 1.5, where the distinction between hemispheres still persists in the latent space and is mainly captured by the latent $\bar{x}$–coordinate. The grouping structures characterizing the latent $\bar{y}$–coordinate and $\bar{z}$–coordinate, in the second and third column, mainly incorporate spatial brain segmentation from back–to–front and top–to–bottom, respectively, with the clustering on the $\bar{y}$–coordinate partially related also to lobe segmentation, although not exactly overlapping. See also Appendix A.1.3 for additional results.

The posterior summaries for the parameters controlling the effects of the anatomical covariates in Equation (1.4) are reported in Table 1.2. The insights in Figure 1.2 are confirmed in Table 1.2, thereby highlighting a general preference for the brain regions to connect with others that are spatially closer and within the same hemisphere (e.g. Bullmore and Sporns, 2012; Sporns, 2013). Variations in the hyper-parameters' specification did not affect the main empirical conclusion drawn in this Chapter.

1.5 Latent factor model

1.5.1 Motivation and model specification

The methodologies introduced and described in the previous sections are based on MCMC estimation, and face severe computational bottlenecks when the number of nodes exceeds few hundreds (Hoff, 2019). This drawback clearly limits the analysis on high-quality brain scans, which are becoming increasingly popular in neuroscience and motivates more challenging research questions (e.g. Stanley *et al.*, 2013). Therefore, it is crucial to develop methodologies which can scale to such dimensions while accounting for the dependence structures and the topology of the brain network.

This section focuses on the latent factor model (LFM) for networks, a latent space model which relies on a set of multiplicative random effects to characterise network dependences (Hoff, 2008). The development of scalable computational tools for such a method is particularly important for the network literature, since it has been shown that

the LFM generalises popular latent variable methods for network data. For example, the stochastic block models and the latent distance models are specific cases of the LFM (Hoff, 2008).

Adapting the notation introduced in Section 1.4.1, the LFM is specified as follows.

$$\begin{aligned}(a_{ij}^{(k)} \mid \pi_{ij}) &\sim \text{Bern}(\pi_{ij}) \\ \text{logit}(\pi_{ij}) &= \beta_0 + \beta_1 \texttt{hem}_{ij} + \beta_2 \texttt{cortex}_{ij} + \beta_3 d_{ij} + \tilde{d}_{ij},\end{aligned} \tag{1.9}$$

for each pair $i = 2, \ldots, n$ and $j = 1, \ldots, i-1$, with $\tilde{d}_{ij} = \psi_x \tilde{x}_i \tilde{x}_j + \psi_y \tilde{y}_i \tilde{y}_j + \psi_z \tilde{z}_i \tilde{z}_j$ and with $\texttt{cortex}_{ij}$ denoting membership to the same cerebral cortex, used instead of lobe membership since such an information was not provided in this dataset. In analogy with the latent space model described in Section 1.4.1, the vector $(\tilde{x}_i, \tilde{y}_i, \tilde{z}_i) \in \mathbb{R}^3$ corresponds to the position of the brain region i in the 3-dimensional latent space. The additional parameters $(\psi_x, \psi_y, \psi_z) \in \mathbb{R}^3$ measure the overall importance and the direction of the first, second and third latent dimension on the brain network connectivity pattern. For example, a value $\psi_x > 0$ implies that brain regions are more "similar" if they have the same sign in the first latent dimension $\tilde{x}$, while the absolute value $|\psi_x|$ determines how relevant the first coordinate is in characterising the observed connectivity pattern. Therefore, the quantity $\tilde{d}_{ij} \in \mathbb{R}$ can be interpreted as a weighted similarity among region i and j in the latent space, with more similar regions having larger similarities resulting in greater probability of being connected. See Hoff (2019) for a recent overview and a comparison among the LFM and other latent variable methods for networks.

We seek simple specification of the prior distributions, leading to efficient computational algorithms. With this motivation in mind, the prior distributions for the coefficients and latent structures are specified as follows.

$$\begin{aligned}\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^\intercal \sim \text{N}_4(0, \boldsymbol{\Sigma}_0), \qquad \boldsymbol{\Sigma}_0 = \text{diag}(\sigma_0, \ldots, \sigma_3), \\ \tilde{x}_i \sim \text{N}(0,1), \quad \tilde{y}_i \sim \text{N}(0,1), \quad \tilde{z}_i \sim \text{N}(0,1) \quad i = 1, \ldots, n, \\ (\psi_x, \psi_y, \psi_x) \sim \text{N}_3(0, \gamma_{\psi_0} I_3)\end{aligned} \tag{1.10}$$

Note that, compared with the model of Equation (1.3)-(1.4) and the priors specified in Equation (1.5), the proposed LFM for networks is much simpler and does not focus on estimating clusters of brain regions in the latent space. Also, the multiplicative similarities of the LFM are less interpretable than the Euclidean distance used in Equation (1.4), since the analogy with the anatomical counterpart is lost. There is a clear trade-off between ease of interpretation and fast computation, and the approaches

described in this Chapter lie at the extremes of this continuum. The selection of a particular approach should be driven by practical consideration on the problem under study, discussing with the practitioners what type of inference is more appropriate and the practical implications of each scenario.

1.5.2 Approximate Bayesian inference via variational methods

In this Section, we provide a coincise review of approximate Bayesian inference, focusing on Variational Bayes (VB). The term "variational approximations" comes from a specific mathematical topic developed in the 18-th century known as *variational calculus*, while its application in statistics is more recent and dates back at the early years of 2000 (e.g. Jordan *et al.*, 1999). Recent applications involving massive amount of data have stimulated additional interest on such a body of techniques, which have been successfully used to conduct inference for complex models in extremely high-dimensional settings.

In order to provide a general notation, let us denote as $\mathbf{y}$ the set of observed data and as $\boldsymbol{\vartheta}$ the set of model's parameters, including "standard" parameters and latent variables, when present. The focus of VB is on finding an *approximate* posterior distribution $q(\boldsymbol{\vartheta})$ in a suitable family $\mathcal{Q}$ of densities which provides the best approximation of the *true* posterior distribution $p(\boldsymbol{\vartheta} \mid \mathbf{y})$, proportional to the conditional likelihood $p(\mathbf{y} \mid \boldsymbol{\vartheta})$ times a joint prior $p(\boldsymbol{\vartheta})$. A popular choice in VB is to use *optimisation* to minimize the Kullback-Lielber (KL) divergence among the approximate posterior and the truth, thereby obtaining

$$q^{\star}(\boldsymbol{\vartheta}) = \arg\min_{q \in \mathcal{Q}} \mathrm{KL}\left\{q(\boldsymbol{\vartheta}) \,||\, p(\boldsymbol{\vartheta} \mid \mathbf{y})\right\}. \tag{1.11}$$

Since the posterior distribution is analytically intractable, direct minimisation of Equation (1.11) is not directly computable. In practice, VB optimizes an alternative objective called the evidence lower bound (ELBO), formally equal to

$$\mathrm{ELBO}(q) = \mathbb{E}_{q(\boldsymbol{\vartheta})}\left[\log\left\{\frac{q(\boldsymbol{\vartheta})}{p(\boldsymbol{\vartheta} \mid \mathbf{y})}\right\}\right], \tag{1.12}$$

with expectation taken with respect to the variational distribution $q(\boldsymbol{\vartheta})$. It is easy to show the ELBO in Equation (1.12) corresponds to the negative KL divergence in Equation (1.11) plus the marginal likelihood (e.g. Blei *et al.*, 2017). Therefore, maximising the ELBO in Equation (1.12) is equivalent to minimizing the KL divergence of Equation (1.11); see Blei *et al.* (2017); Bishop (2006) for a proof. The family $\mathcal{Q}$ needs to be explicitly specified in order to complete the definition of the optimisation problem. There is a clear trade-off between the accuracy of the approximation and computational

convenience of the VB routine, with more complex families leading to less efficient algorithms but more accurate approximations. For example, $\mathcal{Q}$ can be specified as the space of Gaussian distributions with appropriate size, and the optimisation of Equation (1.11) reduces at finding the best Gaussian approximation of the posterior in KL sense. A particularly important family is induced by the Mean Field (MF) approximation, which specifies posterior independence among *blocks* of parameters (Blei *et al.*, 2017, e.g.). The use of MF with conditionally conjugate exponential families is motivated both from a practical and a theoretical perspective. Indeed, the factorisation of the approximate posterior density is the only assumption which is required to obtain closed form expressions for the optimal distributions, further facilitating analytical derivations and computations (e.g. Ormerod and Wand, 2010).

1.5.3 Approximate Bayesian inference for the LFM

From an algebraic standpoint, the main advantage of the LFM lies in the linearity of the parameters with respect to the log-odds of the probability of observing an edge. This feature, combined with the Pólya–Gamma data augmentation of Polson *et al.* (2013), allows to recast the model in terms of a conditionally conjugate exponential family. Indeed, the LFM has served as a building block for more complicated models; for example, Durante *et al.* (2017); Sewell and Chen (2017). Beside allowing to implement a Gibbs Sampler, the availability of closed form expressions for the full-conditional distributions facilitates also approximate Bayesian inference based on VB (Bishop, 2006; Blei *et al.*, 2017).

In order to use a more compact notation, denote as $\mathbf{W}$ the $(n \times 3)$ matrix of latent coordinates with generic row $\mathbf{w}_i^\intercal = (\tilde{x}_i, \tilde{y}_i, \tilde{z}_i)$, $i = 1, \ldots, n$, and denote as $\boldsymbol{\psi} = (\psi_x, \psi_y, \psi_x)$. In particular, the following product restriction will be assumed for the variational family of distributions.

$$q(\boldsymbol{\beta}, \mathbf{W}, \boldsymbol{\omega}, \boldsymbol{\psi}) = q(\boldsymbol{\beta}) q(\mathbf{W}) q(\boldsymbol{\omega}) q(\boldsymbol{\psi}). \tag{1.13}$$

Under the MF factorisation in Equation (1.13), it is possible to show (e.g. Blei *et al.*, 2017, Sec 2.4) that the optimal distributions $q^\star$ have closed form expressions proportional to

$$\begin{aligned}
q^\star(\boldsymbol{\beta}) &\propto \exp\left\{\mathbb{E}_{q(\mathbf{W}, \boldsymbol{\psi}, \boldsymbol{\omega}}\left[\log p(\boldsymbol{\beta} \mid -)\right]\right\}, \\
q^\star(\boldsymbol{\psi}) &\propto \exp\left\{\mathbb{E}_{q(\boldsymbol{\beta}, \mathbf{W}, \boldsymbol{\omega})}\left[\log p(\boldsymbol{\psi} \mid -)\right]\right\}, \\
q^\star(\mathbf{w}_i) &\propto \exp\left\{\mathbb{E}_{q(\boldsymbol{\beta}, \{\mathbf{w}_j\}_{j \neq i}, \boldsymbol{\psi}, \boldsymbol{\Omega}_i)}\left[\log p(\mathbf{w}_i \mid -)\right]\right\} \quad i = 1, \ldots, n, \\
q^\star(\omega_{ij}) &\propto \exp\left\{\mathbb{E}_{q(\boldsymbol{\beta}, \mathbf{w}_i, \mathbf{w}_j, \boldsymbol{\psi})}\left[\log p(\omega_{ij} \mid -)\right]\right\} \quad i = 2, \ldots, n,\ j = 1, \ldots, i,
\end{aligned} \tag{1.14}$$

with $\boldsymbol{\Omega}_i$ denoting the i-th row of the $n \times n$ matrix of Pòlya–Gamma augmented variables such that $[\boldsymbol{\Omega}]_{ij} = \omega_{ij}$, and where each expectation is taken with respect to the optimal distributions of the parameters indicated in the subscripts. Note that $p(\boldsymbol{\beta} \mid -)$ denotes the density of the full-conditional distribution for $\boldsymbol{\beta}$, and similarly for the other parameters.

Since the LFM falls within the class of conditionally conjugate exponential families, each full conditional distribution — available in closed form — is in the exponential family, being either multivariate Gaussian or Pòlya–Gamma. Therefore, the optimal distribution for each factor is in the same parametric (exponential) family of the corresponding full conditional distribution, with natural parameters replaced with variational expectations (Hoffman *et al.*, 2013) which can be easily computed for both the Gaussian and the Pòlya–Gamma distributions. Since each expectation in Equation (1.14) is a functional of different parameters, the optimal solution can be found with iterative methods; for example, iteratively maximising each variational distribution on the basis on the current values of the remaining parameters, until convergence. Such a technique is known as Coordinate Ascent Variational Inference (CAVI), and guarantees a monotonic sequence of the ELBO ensuring converge to a local maxima (Blei *et al.*, 2017). Pseudo code illustrating the analytical form of the updates in Equation (1.14) is reported in the Appendix A.2

Note that VB for the latent distance model is also available (Gollini and Murphy, 2016; Salter-Townshend and Murphy, 2013). However, the lack of conjugacy requires to introduce further approximations of the expected log-likelihood via Taylor expansions (Salter-Townshend and Murphy, 2013) or via Jensen's inequality (Gollini and Murphy, 2016). Instead, exploiting the conditionally conjugacy induced trough the Pòlya–Gamma data augmentation, it is possible to leverage on a pure MF factorisation, thus allowing further improvements via recent computational advances in the variational inference literature (e.g. Blei *et al.*, 2017); see also Durante and Rigon (2019) for related arguments.

1.5.4 Simulation study

To evaluate the empirical performance of the VB algorithm, a simulation study is conducted focusing on different scenarios. The focus of the simulations will be on determining whether approximate Bayesian inference for the LFM provides reasonable estimates for artificial networks simulated under different data generating processes.

In the first setting, artificial networks are simulated form a LFM with $H = 2$ latent factors, $\boldsymbol{\psi} = (2, -2)$ and latent factors generated from standard multivariate Gaussians.

Networks in the second scenario are generated from from a latent distance model with $H = 2$ latent coordinates randomly sampled from multivariate Gaussians. The third and last setting focuses on stochastic block models (Nowicki and Snijders, 2001). Specifically, 2 latent blocks with equal membership probabilities are considered and, conditionally on group allocation, the probability of a connection within the same group is 0.6 in the first block and 0.8 in the second, while the probability of connection across groups is equal 0.2. For each setting described above, networks are generated with a number of nodes $n = \{100, 500, 1000, 5000\}$.

The approach of Salter-Townshend and Murphy (2013), which performs VB for the latent distance model, is considered as a competitor, relying on the R implementation available through the package VBLPCM and using default configuration. Prior parameters for the LFM are specified as $\sigma_0^2 = 5$, $\gamma_{\psi_0} = 5$, and the algorithm is run until changes in the ELBO are lower than 10^{-5}. Performance is assessed in terms of the Area Under the Roc curve (AUC) and Accuracy (ACC) of the adjacency matrices predicted via posterior means under both methods.

Table 1.3 reports the results of the simulations. First, second and third block of rows correspond to the first, second and third scenario, while columns are associated with different network sizes. The competitor is denoted as LDM. The empirical findings of the simulations suggest that in the first scenario, the LFM has an overall better performance than the competitor, and the gap increases in particular when the number of nodes is large. In the second scenario, the performance of the two approaches is similar, with the competitor providing better results with $n = 500$ and $n = 1000$. In the third and last scenario, the competitor approach has poor performance, while the LFM provides satisfactory results, better than the competitor in all the settings considered.

It is worth highlighting that since both algorithms perform *approximate* Bayesian inference, in general it is not guaranteed that the approximate posterior — which is

TABLE 1.3: Results for the simulation study. LFM indicates the method proposed in the Chapter, while the competitor approach of (Salter-Townshend and Murphy, 2013).

		$n = 100$		$n = 500$		$n = 1000$		$n = 5000$	
		LFM	LDM	LFM	LDM	LFM	LDM	LFM	LDM
Scenario 1	AUC	0.861	0.743	0.849	0.793	0.849	0.799	0.848	0.609
	ACC	0.685	0.622	0.663	0.688	0.662	0.787	0.658	0.591
Scenario 2	AUC	0.730	0.693	0.659	0.688	0.650	0.677	0.648	0.558
	ACC	0.565	0.651	0.547	0.593	0.597	0.673	0.673	0.459
Scenario 3	AUC	0.787	0.586	0.736	0.706	0.727	0.696	0.713	0.593
	ACC	0.705	0.591	0.694	0.631	0.698	0.661	0.700	0.541

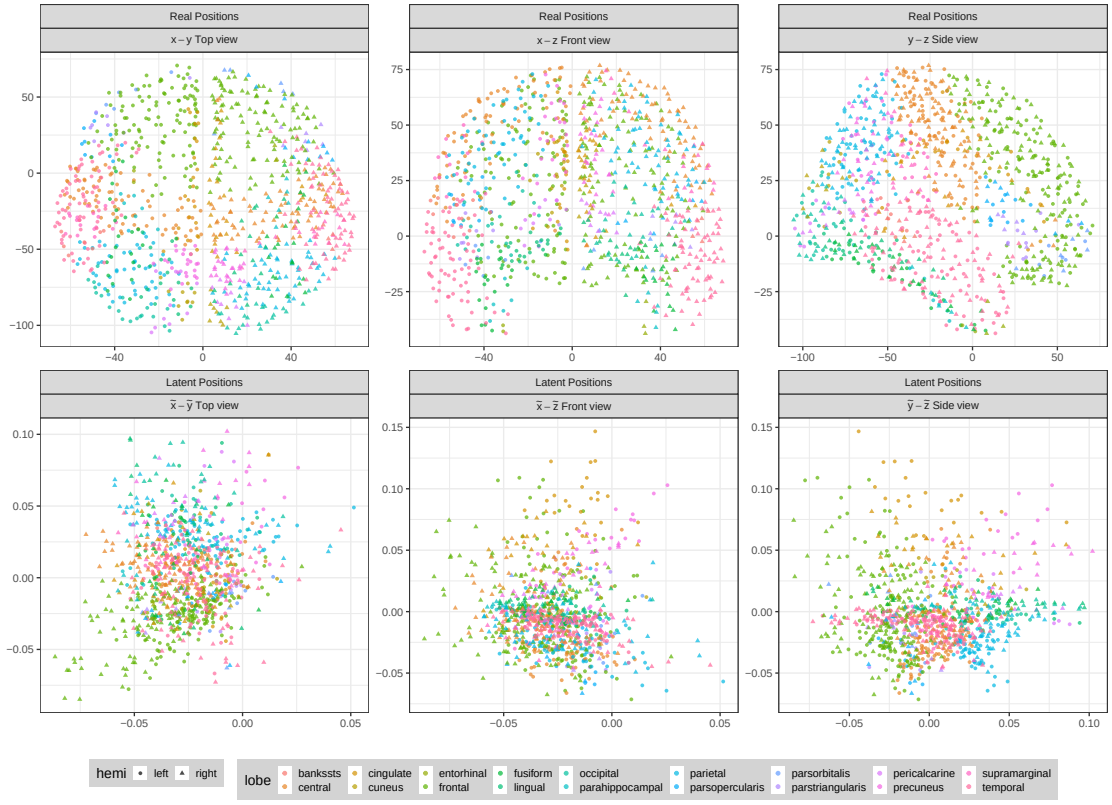

FIGURE 1.7: Graphical representation of the anatomical coordinates (upper panel) and the posterior means of the latent positions (lower panel). Shapes and colors refer to hemisphere and lobe membership.

the best member in the approximating family — provides a good approximation of the true posterior distribution. Indeed, if the approximating family is poorly chosen or if the approximation of the marginal likelihood is inaccurate, the optimised lower bound might not be tight enough to provide a reasonably good approximation. The empirical results of the simulation, instead, suggest that under both approaches the resulting posterior is a reasonable approximation, providing accurate predictions from moderate to large network settings.

1.5.5 Application to high-quality brain imaging

The simulations results motivate inference on the mean of the approximate posterior distributions $q^\star(\tilde{x}_i)$, $q^\star(\tilde{y}_i)$, $q^\star(\tilde{z}_i)$, $i = 2, \ldots, n$ of the 3–dimensional latent coordinates, which are compared with the corresponding anatomical ones in Figure 1.7. Coherently with the results of Section 1.4.4, also the coordinates of the regions in the latent space do not coincide with the anatomical positions measured for high-scan imaging. For example, the brain regions in the cingulate and entorhinal areas have a more peripheral position in the latent space, whereas regions in the precuneus and supramarginal areas

become more central. Differently from the results of the results of Section 1.4.4, the effect of hemisphere segmentation is not apparent in the latent coordinates, and is entirely captured by the covariates information. This effect might be due to the different structure imposed by the multiplicative similarity of the LFM, which characterise more details of the unobserved structure. These findings are coherent with the results on the KKI-21 dataset, and suggest that the connectivity which are not explained by the observed anatomical covariates might still have a relation with the physical properties of the brain. Also the effect of the different covariates is similar to the findings of Section 1.4.4, highlighting a general preference for brain regions to connect with regions that are closer (-0.536), belonging to the same areas (0.796) and hemisphere (1.211). This findings were expected and coherent with the main empirical findings in the brain network literature (e.g. Bullmore and Sporns, 2012; Sporns, 2013). However, as already discussed, the contributions proposed in this chapter complete and refine these findings by explicitly modelling also the connectivity architectures not captured by the anatomical covariates.

Chapter 2

Latent structures models for multivariate categorical data

2.1 Categorical data

Latent structure modelling provides concrete benefits also in the analysis of multivariate categorical data, which are routinely collected in many application areas. The challenging complexity of such data relies in the intricate interaction structure across the different categorical variables, which often provides precious insights on many research questions but whose estimation is very difficult when the number of variables is moderate. Indeed, categorical data can be organized as multiway contingency tables, where individuals are cross classified according to their values for the different variables. As the number of cells in the table grows exponentially with the number of variables, many or even most cells will contain zero observations; for example, 16 categorical variables with 4 categories each define a contingency table with a total number of cells larger than 1-billion, and clearly no study will ever collect so many individuals. This severe sparsity motivates appropriate statistical methodologies that effectively reduce the number of free parameters, with latent structure analysis being a successful option (e.g. Lazarsfeld, 1950; Dunson and Xing, 2009; Bhattacharya and Dunson, 2012; Zhou *et al.*, 2015; Russo *et al.*, 2019).

Latent variable models for categorical data are specified in terms of one or more latent features, with observed variables modelled as conditionally independent given the latent features. Marginalising over the latent structures, complex dependence patterns across the categorical variables are induced (e.g. Andersen, 1982). Some representative examples include latent class analysis (Lazarsfeld, 1950) and the normal ogive model

(Lawley, 1943), where an univariate latent variable with discrete or continuous support, respectively, captures the dependence structure among the observed categorical variables. Leveraging data-augmentation schemes, estimation of latent variable models is feasible in high-dimensional applications via MCMC, Expectation-Maximisation (EM) (Dempster *et al.*, 1977) and combinations of the two (e.g. Fruhwirth-Schnatter *et al.*, 2019). Beside providing tractable computational methods, in several applications the heterogeneity of the population can be studied making inference on the latent structures; for example, estimating groups of individuals with similar response patterns, or measuring item difficulties (Andersen, 1982).

The success of latent structure modelling and the recent computational developments have motivated several extensions in different areas, with methods based on more complicated multivariate latent structures becoming increasingly popular. Some example include Grade of Membership models (Erosheva, 2005) and Mixed Membership models (Airoldi *et al.*, 2014). Specific latent variable models for multivariate categorical admits natural Bayesian nonparametric specifications allowing an infinite number of components (Dunson and Xing, 2009; Bhattacharya and Dunson, 2012; Zhou *et al.*, 2015). See also Carota *et al.* (2015).

An alternative class of models for categorical data consists of log-linear models, which represent the logarithms of cell probabilities as linear terms of parameters related to each cell index, and with coefficients that can be interpreted as interactions among the categorical variables (Agresti, 2003). The relationship between Multinomial and Poisson log-likelihoods allows one to obtain Maximum Likelihood estimates (MLE) for log-linear models leveraging standard Generalized Linear Model (GLM) algorithms (e.g., Fisher-Scoring), with the vectorized table of cell counts used as a response variable. As already discussed, the number of cells of the contingency table grows exponentially with the number of variables, leading to infinite MLE (Fienberg and Rinaldo, 2007). To overcome this issue and obtain unique estimates, it is assumed that a large set is coefficients is zero, and estimation is performed via penalised likelihood (Nardi and Rinaldo, 2012; Tibshirani *et al.*, 2015; Wainwright and Jordan, 2008; Ravikumar *et al.*, 2010). However, the computation and storage of the joint cells counts — required to fit the approaches mentioned above — becomes unfeasible even for moderate values of the number of variables p.

Bayesian approaches for inference in log-linear models restrict consideration to specific nested model subclasses; for example, hierarchical, graphical or decomposable log-linear models (Lauritzen, 1996). Conjugate priors on the model coefficients are available

(Massam *et al.*, 2009), but exact Bayesian inference is still complicated since the resulting posterior distribution is not particularly useful, lacking of closed form expressions for important functionals – such as credible intervals – and sampling algorithms to perform inference via Monte Carlo integration. As an alternative, the posterior distribution can be analytically approximated with a Gaussian distribution, if the number of cells is not excessively large (Johndrow and Bhattacharya, 2018). When the focus is on selecting log-linear models with high posterior evidence, stochastic search algorithms evaluating the exact or approximate marginal likelihood are available (Dobra and Massam, 2010). Unfortunately, the size of the model space is enormous and these algorithms scale poorly with the number of variables, being essentially unfeasible in applications with more than 15 binary variables (Johndrow and Bhattacharya, 2018). While hierarchical log-linear model can be justified from a practical perspective, the same does not hold for other nested subclasses, which are justified only from an algebraic point of view. For example, hierarchical log-linear models including all two-factor interactions are very popular in practice, but do not belong to the graphical subclass and are therefore excluded from consideration by these methods.

2.2 Data description and motivation

We consider a psychiatric study on suicide attempt, a dramatic phenomenon which has motivated a huge variety of scientific studies over the past decades (e.g. Nock *et al.*, 2008; De Leo *et al.*, 2004). Studies on suicide attempt survivals are crucial for the development of novel intervention treatments based on the early identification of psychological symptoms (e.g. Hawton and Fagg, 1988), and also for accurate descriptions of the psycho–pathological profiles more likely to attempt suicide acts. For example, depression and hostility symptoms are often associated in suicide attempts (Ben-Ya'acov and Amir, 2004), while some recent work have also suggested that empathy could be an important risk factor associated with specific psychiatric disease and the suicidal act (e.g. Lachal *et al.*, 2016).

In particular, it is of interest to analyse the psychopathology of attempt suicide patients, their empathic profile and the possible interactions across these two psychological aspects. This case study has been motivated by a collaboration with doctor Paolo Scocco from Padova Hospital, which is kindly acknowledged for providing the data and for the stimulating discussions on the definition of the problem and on the interpretation of the

results from a clinical perspective.[1] Individuals analysed in the study correspond to a sample of 58 inpatients hospitalized after an attempted suicide at the psychiatric ward of Padova Hospital (Italy) between January 2017 and December 2018. For the purposes of the study, an "attempted suicide" was defined as a person who deliberately harmed their body, and spontaneously declared that the act was intended to end their life. When the person was not sure about the reasons for their act, attempted suicide was diagnosed when the self-harm caused medically serious consequences requiring hospitalization. See also Goodfellow *et al.* (2019); De Leo *et al.* (2004) for additional comments. Data were collected by self administered questionnaires aimed at evaluating different psychological aspects of attempted suicidal, with the Symptom Check List (SCL-90) (Derogatis *et al.*, 1973) and the Interpersonal Reactivity Index (IRI) (Davis, 1980) being reliable instruments for these purposes.

Specifically, the SCL-90 is commonly used to describe psychiatric symptoms, using 90 items scored on a five-point Likert scale; additionally, scores can be grouped into nine subscales (somatization, obsessive-compulsive, interpersonal sensitivity, depression, anxiety, hostility, phobic anxiety, paranoid ideation, psychoticism) corresponding to well-defined psychiatric profiles (Derogatis *et al.*, 1973). As suggested by the clinician, it is of particular interest to focus on 4 subscales of the questionnaire: obsessive-compulsive (OC), depression (DEP), anxiety (ANX) and hostility (HOS), encompassing a total of 28 items. See Appendix B.2 for an illustration of the items under investigation. The reliability of the Italian version of the instrument has been assessed in previous studies (e.g. Prunas *et al.*, 2012).

The IRI is a 28-item instrument scored on five-point Liker scale that measures the emotional and cognitive components of a person's empathy, into four subscales. Indeed, the IRI measures the cognitive capacity to see things from the point of view of others (Perspective Taking, PT), the tendency to experience reactions of sympathy, concern and compassion for other people undergoing negative experiences (Empathic Concern, EC), the tendency to experience distress and discomfort in witnessing other people's negative experiences (Personal Distress, PD) and the capacity to strongly identify oneself with fictitious characters in movies, books, and plays (Fantasy, FS). See also Davis (1983) for further comments and again Appendix B.2 for a more detailed description of the dataset.

[1] I would also like to acknowledge Prof. Giovanna Capizzi for introducing me to Dr. Scocco and for her active contribution in the discussions and the definition of the problem.

In the psychological literature, investigation of the relationship among different empathic profiles and psycho-pathological symptoms has been a challenging research objective in the last years. Generally, variations in empathy are also associated with depression (Cusi *et al.*, 2011; Schreiter *et al.*, 2013), obsessive compulsive disorders (Fontenelle *et al.*, 2009), anxiety (Perrone-McGovern *et al.*, 2014) and hostility (Guttman and Laporte, 2002). For example, a frequent symptom of depression is the inability to perceive our own emotions, which is also realistically associated with the inability to comprehend other individuals' ones (e.g. Cusi *et al.*, 2011). Another example include anxiety symptoms, which are likely to be associated with personal distress and hostility (Guttman and Laporte, 2002). However, the relationship among psycho-pathological symptoms and empathic profiles in attempt suicidal is still not completely understood. Indeed, individuals who attempted suicide might exhibit unexpected association patterns across the psycho-pathological diseases and empathic profiles. For instance, a depressed individual with a strongly empathic profile would be in the state of having inconsistent thoughts, unable to describe its own feeling and overwhelmed by other people's emotions, including their worse suffering. Such a state, sometimes referred to as "cognitive dissonance" in psychology, has been identified as a potential relevant cause of attempt suicide (Zhang and Lester, 2008). At the same time, it is reasonable to assume that several associations are similar between attempted suicidal people and the healthy population, and it is of interest to include such information in a statistical model.

As outlined in Section 2.1, latent structure models are efficiently estimated in high-dimensional settings via data-augmentation algorithms. Therefore, such approaches are effective at estimating very high-dimensional models in reasonable time with solid assessment of uncertainty when a Bayesian approach to inference is followed. However, interest often relies on low-dimensional functionals of such massively high-dimensional structures — such as some bivariate measures of association — but it is not clear what type of functional form is imposed over such quantities of interest. Indeed, when interest is on such functionals, it is customary to post-process the MCMC output and conduct Monte Carlo inference on such quantities (e.g. Bhattacharya and Dunson, 2012; Dunson and Xing, 2009). Therefore, the inclusion of simple prior information on the association structure is unpractical within this class of model. On the other hand, log-linear models directly parametrise the interactions among the categorical variables (Agresti, 2003) and the lower-dimensional marginal distributions (Bergsma and Rudas, 2002; Roverato *et al.*, 2013), but estimation is generally unfeasible when the number of variables is moderate to high, due to the huge computational bottlenecks and the massively large model space. Sparse log-linear models and latent class structures are deeply related in

the way in which sparsity is induced (Johndrow *et al.*, 2017), but a formal methodology mixing the benefits of the two families is still lacking.

Motivated by the above considerations, Section 2.3 introduces a novel class of Bayesian tools for multivariate categorical data, which we refer to as MILLS. The focus of this methodology is to combine the two large classes of methods for categorical data — log-linear models and latent structure analysis — in order to reduce the reciprocal disadvantages illustrated above. We propose to model the multivariate categorical data as a composite mixture of log-linear models with first order interactions, thereby allowing characterisation of the bivariate distributions with simple and robust models while accounting for dependence beyond first order via mixing different local models. Such a specification allows one to model the categorical data with a simple, yet flexible, specification which can take into account complex dependencies with a relatively small number of components. In addition, inference on low-dimensional marginal distributions – and induced measures of association — is obtained efficiently.

The idea of mixing simple low-dimensional models for modelling complex data to reduce the growth of the number of parameters has a long history in statistics. A notable example of mixing simple models accounting for first order dependencies is the mixture of transition matrix of Raftery (1985), originally developed to model higher-order Markov chains. See also Fruhwirth-Schnatter *et al.* (2019) for related ideas.

2.3 Composite mixture of log-linear models for categorical data

2.3.1 Log-linear models

The notation of Lauritzen (1996) is adopted. Let $V = \{1, \ldots, p\}$ index a set of p categorical variables. Let $(Y_j, j \in V)$ denote the categorical variable taking values in the finite set $\mathcal{I}_j$ with dimension $|\mathcal{I}_j| = d_j$. For simplicity in exposition, and without loss of generality, we can assume $\mathcal{I}_j = \{1, \ldots, d_j\}$. Categorical data are generally collected as an $n \times p$ data matrix with elements $y_{ij} \in \mathcal{I}_j$, $i = 1, \ldots, n$, $j = 1, \ldots, p$, and can also be represented as a contingency table. Let $\mathcal{I}_V = \bigtimes_{j \in V} \mathcal{I}_j$ denote the set with generic element $\boldsymbol{i} = (i_1, \ldots, i_p)$. The elements $\boldsymbol{i}$ of $\mathcal{I}_v$ are referred to as the *cells* of the contingency table $\mathcal{I}_V$, which has size $|\mathcal{I}_V| = \prod_{j=1}^{p} d_j$. Given a sample of size n, the number of observations falling in the generic cell $\boldsymbol{i}$ is denoted as $y(\boldsymbol{i})$, with $\sum_{\boldsymbol{i} \in \mathcal{I}_V} y(\boldsymbol{i}) = n$.

A log-linear model is a generalised linear model for the resulting multinomial likelihood, which represents the logarithms of cell probabilities with additive terms. Let $\mathbf{p} = (p(\boldsymbol{i}), \boldsymbol{i} \in \mathcal{I}_v)$ denote the vectorised cell probabilities and let $\boldsymbol{\vartheta}$ denote the set of log-linear coefficients. Following Letac and Massam (2012); Johndrow and Bhattacharya (2018), it is possible to relate cell probabilities and log-linear coefficients as follows:

$$\log \mathbf{p} = \mathbf{X}\boldsymbol{\vartheta}, \tag{2.1}$$

where $\mathbf{X}$ is a full rank $|\mathcal{I}_V| \times |\mathcal{I}_V|$ matrix if the transformation is invertible; for example, when $\mathbf{X}$ is the identity matrix, the so-called identity parametrisation is obtained. Identifiability is imposed through careful specification of the matrix $\mathbf{X}$, which determines the model parametrisation and, consequently, constraints on the parameters (Agresti, 2003). Equation (2.1) can be extended to embrace a larger class of invertible and non-invertible log-linear parametrisations; for example, marginal parametrisations (e.g. Bergsma and Rudas, 2002; Roverato *et al.*, 2013; Lupparelli *et al.*, 2009). In general, it is desirable to specify a sparse set of k coefficients with $k \ll |\mathcal{I}_v|$, corresponding to some notion of interactions among the categorical variables; for example, representing conditional or marginal independence (Agresti, 2003). When a sparse parameterisation is employed, it is common to remove in Equation (2.1) the columns of $\mathbf{X}$ associated with zero coefficients, thereby obtaining a more parsimonious design matrix with dimension $|\mathcal{I}_V| \times k$. In this article we focus on the corner parameterisation, which is particularly popular in the literature for categorical data (Agresti, 2003; Massam *et al.*, 2009; Letac and Massam, 2012), and is generally the default choice in statistical software. The columns of $\mathbf{X}$ under the corner parameterisation can be formally expressed in terms of Moebius inversion (e.g. Letac and Massam, 2012, Preposition 2.1); see also Massam *et al.* (2009, Lemma 2.2). For simplicity in exposition, we prefer to use matrix notation.

Let $\mathbf{y} = (y(\boldsymbol{i}), i \in \mathcal{I}_v)$ denote the vectorised cell counts. The likelihood function associated with the multinomial sampling and log-linear parameters can be expressed, in matrix form, as follows.

$$\prod_{\boldsymbol{i} \in \mathcal{I}_V} p(\boldsymbol{i})^{y(\boldsymbol{i})} = \exp\left\{\mathbf{y}^\intercal \mathbf{X}\boldsymbol{\vartheta} - n\kappa(\boldsymbol{\vartheta})\right\} = \exp\left\{\tilde{\mathbf{y}}^\intercal \boldsymbol{\vartheta} - n\kappa(\boldsymbol{\vartheta})\right\}, \tag{2.2}$$

with $\kappa(\boldsymbol{\vartheta}) = \log\left[\mathbf{1}^\intercal \exp(\mathbf{X}\boldsymbol{\vartheta})\right]$. Such a parametrisation yields a very compact data reduction, since the canonical statistics $\mathbf{y}^\intercal \mathbf{X} = \tilde{\mathbf{y}}^\intercal$ correspond to the marginal cell counts relative to the highest interaction term included in the model (Massam *et al.*, 2009; Agresti, 2003). In particular, we will consider hierarchical log-linear models which include all the main effects and all the first-order interactions; under such a specification,

the canonical statistics $\tilde{\mathbf{y}}$ correspond to the marginal bivariate and univariate tables (e.g Agresti, 2003).

2.3.2 Composite likelihood

The log-partition function in Equation (2.2) involves a sum of $|\mathcal{I}_V|$ terms, the total number of cells. Clearly, this operation becomes infeasible even for very modest values of p, thereby preventing the likelihood function from being evaluated and numerically maximised. Approximations of intractable likelihoods have been proposed in the literature, with Monte Carlo Maximum Likelihood (Snijders, 2002; Geyer and Thompson, 1992) being one valid option. Composite likelihoods provide a computationally tractable alternative to the joint likelihood, replacing the original density with a product of carefully chosen marginal or conditional distributions; see Varin *et al.* (2011) for an extensive overview. Extending the work of Meng *et al.* (2013), Massam and Wang (2018) focused on Composite Maximum Likelihood Estimation for log-linear models, with a careful choice of the conditional and marginal distributions based on the conditional dependence graph. In practice, the dependence graph is unknown and its estimation can be very demanding and affected by large uncertainty (Dobra and Massam, 2010); hence, it is more desirable to rely on a model specification which is coherent with any unknown underlying structure, without relying on a pre-selected graphical structure.

We propose to replace the joint likelihood with a more simple and robust alternative. Denote as $\mathcal{P}_2$ the set of subsets of V with cardinality 2. For each $E_2 \in \mathcal{P}_2$ let $\boldsymbol{y}_{E_2}$ denote the vectorised E_2-marginal bivariate table of counts. We define, for each $\boldsymbol{y}_{E_2}$, a saturated log-linear model with corner parametrisation as follows.

$$\mathbf{p}(\boldsymbol{y}_{E_2};\, \boldsymbol{\vartheta}_{E_2}) = \exp\left\{\boldsymbol{y}_{E_2}^{\intercal}\mathbf{X}_2\boldsymbol{\vartheta}_{E_2} - n\kappa_2(\boldsymbol{\vartheta}_{E_2})\right\} = \exp\left\{\tilde{\boldsymbol{y}}_{E_2}^{\intercal}\boldsymbol{\vartheta}_{E_2} - n\kappa_2(\boldsymbol{\vartheta}_{E_2})\right\}, \tag{2.3}$$

where $\kappa_2(\boldsymbol{\vartheta}_{E_2}) = \log\left[\mathbf{1}^{\intercal}\exp(\mathbf{X}_2\boldsymbol{\vartheta}_{E_2})\right]$ and with $\dim\boldsymbol{\vartheta}_{E_2} = \dim\tilde{\boldsymbol{y}}_{E_2} = |\mathcal{I}_{E_2}| = \prod_{j\in E_2} d_j$ and $\boldsymbol{\vartheta}_{E_2} \in \mathbb{R}^{|\mathcal{I}_{E_2}|}$. There is an important difference between $\boldsymbol{y}_{E_2}$ and $\tilde{\boldsymbol{y}}_{E_2}$. The former refers to the E_2-marginal bivariate table, while the latter refers to the sufficient statistics of the log-linear model with corner parametrisation, which are indeed elements of the bivariate and univariate E_2-marginal table; see, for example, Agresti (2003).

We define a surrogate likelihood function combining the distributions defined in Equation (2.3) as follows:

$$\exp\left\{\sum_{E_2\in\mathcal{P}_2}\log\mathbf{p}(\boldsymbol{y}_{E_2};\boldsymbol{\vartheta}_{E_2})\right\} = \exp\left\{\sum_{E_2\in\mathcal{P}_2}\tilde{\boldsymbol{y}}_{E_2}^{\intercal}\boldsymbol{\vartheta}_{E_2} - n\sum_{E_2\in\mathcal{P}_2}\kappa_2(\boldsymbol{\vartheta}_{E_2})\right\} \tag{2.4}$$

Equation (2.4) is constructed with the same motivation of composing simplified likelihood from marginal densities in composite likelihood estimation; see, for example, Cox and Reid (2004); Varin *et al.* (2011). Differently from Massam and Wang (2018), we include contribution for all the bivariate distributions in Equation (2.4) since the underlying graphical structure is not known a priori, and it is not possible to decide which marginal densities should be included accordingly.

The use of Equation (2.4) can also be justified and motivated from an inferential point of view. When interest is on making inference on low-dimensional marginal distributions, such as univariates and bivariates, estimates based on the pseudo likelihood in Equation (2.4) and the original likelihood Equation (2.2) are equivalent, since the joint model is a closed exponential family which includes only first order interactions in the sufficient statistics (Mardia *et al.*, 2009, Theorem 2). With respect to this consideration, it is also worth highlighting that the sufficient statistics $\tilde{\boldsymbol{y}}_{E_2}$ of the simplified model Equation (2.3) is actually a subset of the sufficient statistics of the joint model $\tilde{\mathbf{y}}$ in Equation (2.2) and that $\bigcup_{E_2 \in \mathcal{P}_2} \tilde{\boldsymbol{y}}_{E_2} = \tilde{\mathbf{y}}$. When, instead, inference focuses on higher-order distributions, parameters of Equation (2.4) can be combined effectively to provide estimates on such quantities.

We have motivated the use of the simplified likelihood in terms of focusing inference on low-dimensional margins, which is often the objective in a variety of applications. Even when interest is on such low-dimensional margins, a model which includes only first order interactions might be oversimplified, since the presence of dependencies beyond first order might actually lead to misleading results and a very poor representation. We propose to address the above issues by mixing different composite likelihoods. Denote with $\boldsymbol{i}_{E_2}$ the elements of $\mathcal{I}_{E_2}$, cells of the E_2-marginal bivariate table. The contribution to the composite likelihood for a single observation $y_i = (y_{i1}, \ldots, y_{ip})$ can be expressed as

$$\tilde{\mathbf{p}}(y_i; \boldsymbol{\vartheta}) = \exp\left\{ \sum_{E_2 \in \mathcal{P}_2} \sum_{\boldsymbol{i}_{E_2} \in \mathcal{I}_{E_2}} \mathbb{I}(y_i, \boldsymbol{i}_{E_2}) \mathbf{X}_2 \boldsymbol{\vartheta}_{E_2} - \sum_{E_2 \in \mathcal{P}_2} \kappa_2(\boldsymbol{\vartheta}_{E_2}) \right\}, \tag{2.5}$$

where $\mathbb{I}(y_i, \boldsymbol{i}_{E_2})$ denotes a vectorial indicator function over the vectorised E_2-marginal table, corresponding to a vector of length $|\mathcal{I}_{E_2}|$ with all elements equal to 0 and 1 in the position corresponding to the cell in which the E_2 component of y_i falls. The product over all the data points is clearly equivalent to the full likelihood in Equation (2.4)

In order to mix different models, we rely on a latent structure construction. We suppose that the population can be divided into H latent group, each characterised by a group-specific composite likelihood which takes into account the marginal bivariate

association structure in the group. In doing so, we introduce for each multivariate observations y_i a latent group indicator z_i with $\text{pr}[z_i = h] = \nu_h$, $\nu_h > 0$ and $\sum_{h=1}^{H} \nu_h = 1$. Conditionally on group membership, we regard Equation (2.5) as the likelihood contribution for y_i conditionally on z_i. Considering only observations y_i such that $z_i = h$ and denoting with n_h the number of observations in group h, we can interpret Equation (2.4) as a model for the contingency table conditional on cluster membership, as

$$\tilde{\mathbf{p}}(\boldsymbol{y}^h; \boldsymbol{\vartheta}^h, \mathbf{z}) \;= \exp\left\{ \sum_{E_2 \in \mathcal{P}_2} \tilde{\boldsymbol{y}}_{E_2}^{h\intercal} \boldsymbol{\vartheta}_{E_2}^h - n_h \sum_{E_2 \in \mathcal{P}_2} \kappa_2(\boldsymbol{\vartheta}_{E_2}^h) \right\}. \tag{2.6}$$

We can interpret Equation (2.6) as local models to characterise the association structures of the categorical variable for the subjects in the h-th group. To make inference at the population level, the local models needs to be combined in order to yield a realistic global one, and clearly such an operation should be carefully addressed. Marginalising over the latent feature and considering the contribution for all the data points, we obtain a joint model with likelihood function equal to

$$\tilde{\mathbf{p}}(\boldsymbol{y}; \boldsymbol{\vartheta}, \boldsymbol{\nu}) = \prod_{i=1}^{n} \sum_{h=1}^{H} \nu_h \, \tilde{\mathbf{p}}(y_i; \boldsymbol{\vartheta}^h), \tag{2.7}$$

with $\boldsymbol{\vartheta}^h = \{\boldsymbol{\vartheta}_{E_2}^h\}_{E_2 \in \mathcal{P}_2}$, $\boldsymbol{\vartheta} = \{\boldsymbol{\vartheta}^h\}_{h=1}^H$ and $\boldsymbol{\nu} = \{\nu_h\}_{h=1}^H$.

Equation (2.7) corresponds to a finite mixture of composite first order log-linear models, an will be refer to as MILLS in the sequel. Inference on bivariate distributions is simple under the proposed MILLS specification. For example, a natural estimator for the E_2 bivariate probabilities is given by

$$p(\boldsymbol{i}_{E_2}) = \sum_{h=1}^{H} \nu_h \, \frac{\exp\left(\boldsymbol{\vartheta}_{E_2}^h\right)}{1 + \mathbf{1}^\intercal \exp(\mathbf{X}_2 \boldsymbol{\vartheta}_{E_2}^h)}, \tag{2.8}$$

which corresponds to a weighted average of the local model, with weights given by the mixture proportions.

2.3.3 Bayesian inference

We proceed with a Bayesian approach to inference, and specify prior distributions for the parameters $\boldsymbol{\nu}$ and $\boldsymbol{\vartheta}_{E_2}^h$. For computational convenience, we rely on spare Dirichlet priors for the mixture weights $\boldsymbol{\nu}$, thereby letting

$$\boldsymbol{\nu} \sim \text{Dirichlet}\left(\frac{1}{H}, \ldots, \frac{1}{H}\right). \tag{2.9}$$

Choosing a conservative upper bound H, sparse Dirichlet weights favor the deletion of redundant components (Rousseau and Mengersen, 2011). The prior distributions for the log-odds of the cell probabilities are specified as standard multivariate Gaussian distributions. This choice allows computational tractability adapting the Pólya–Gamma data augmentation to multinomial regression (Polson *et al.*, 2013), and to include simple prior information on the marginal bivariate association structure by choosing appropriate prior mean values. This choice implies

$$\boldsymbol{\vartheta}^h_{E_2} \sim \mathrm{N}_{|\mathcal{I}_{E_2}|}(\mu_{E_2}, \sigma^2 I), \quad E_2 \in \mathcal{P}_2, \quad h = 1, \ldots, H. \tag{2.10}$$

The statistical model defined in Equation (2.7) does not correspond to a genuine likelihood, since it's not a distribution function of the observed data given a parameter value. Therefore, some additional attention is required before proceeding with standard Bayesian inference. There is a rich literature on the use of alternative likelihoods for Bayesian inference; for example, approximate likelihood (Efron, 1993), partial likelihood (Raftery *et al.*, 1995), empirical likelihood (Lazar, 2003) and adjusted profile likelihood (Chang and Mukerjee, 2006), among many others. See also Greco *et al.* (2008) for related arguments. The use of composite likelihoods as components of Bayesian inference has been investigated more recently (Ribatet *et al.*, 2012; Pauli *et al.*, 2011).

In particular, we will conduct inference using the composite posterior distribution

$$\tilde{p}(\boldsymbol{\vartheta}, \boldsymbol{\nu} \mid \boldsymbol{y}) = \frac{p(\boldsymbol{\vartheta})p(\boldsymbol{\nu}) \prod_{i=1}^{n} \sum_{h=1}^{H} \nu_h \, \tilde{\mathbf{p}}(y_i; \boldsymbol{\vartheta}^h)}{\int\int p(\boldsymbol{\vartheta})p(\boldsymbol{\nu}) \prod_{i=1}^{n} \sum_{h=1}^{H} \nu_h \, \tilde{\mathbf{p}}(y_i; \boldsymbol{\vartheta}^h) \mathrm{d}\boldsymbol{\vartheta} \mathrm{d}\boldsymbol{\nu}}, \tag{2.11}$$

which is guaranteed to be proper by the following Lemma.

Lemma 2.1. *The quantity in Equation* (2.11) *is a proper probability distribution.*

Proof. In order to show that Equation (2.11) is a probability distribution, it is necessary to show that the normalising constant exists and is finite. Hence, it is necessary to show that

$$\int\int \pi(\boldsymbol{\vartheta})\pi(\boldsymbol{\nu}) \prod_{i=1}^{n} \sum_{h=1}^{H} \nu_h \, \tilde{\mathbf{p}}(y_i \mid \boldsymbol{\vartheta}^h) \mathrm{d}\boldsymbol{\vartheta} \mathrm{d}\boldsymbol{\nu} < \infty. \tag{2.12}$$

Since the priors specified in Equations (2.9) and (2.10) are proper distributions, it is sufficient to show that

$$\sup_{\boldsymbol{\vartheta}, \boldsymbol{\nu}} \prod_{i=1}^{n} \sum_{h=1}^{H} \nu_h \, \tilde{\mathbf{p}}(y_i \mid \boldsymbol{\vartheta}^h) < \infty. \tag{2.13}$$

Boundedness holds under the MILLS specification since

$$\tilde{\mathbf{p}}(y_i \mid \boldsymbol{\vartheta}^h) = \prod_{E_2 \in \mathcal{P}_2} \exp\left\{ \sum_{\boldsymbol{i}_{E_2} \in \mathcal{I}_{E_2}} \mathbb{I}(y_i, \boldsymbol{i}_{E_2}) \mathbf{X}_2 \boldsymbol{\vartheta}^h_{E_2} - \kappa_2(\boldsymbol{\vartheta}^h_{E_2}) \right\}, \tag{2.14}$$

is always bounded, being a product of probabilities. □

Since the quantity in Equation (2.11) is a proper distribution, we can conduct inference via MCMC, simulating from the pseudo posterior and conducting inference via Monte-Carlo integration. Key steps of the Gibbs Sampler are illustrate in Appendix B. An alternative procedure for fast inference on the Maximum-A-Posteriori (MAP) can be developed through an EM algorithm with nested Expectation step, adapting the strategy of Durante *et al.* (2019) to the composite MILLS kernel.

2.4 Simulation study

In order to evaluate the model performance, we considered a simulation study over three different settings. In each scenario, we focus on a challenging setting with $p = 15$ variables with $d_1 = \ldots = d_{15} = 4$, dividing the categorical variables in groups characterised by different dependence structure. In the first scenario, variables in the first group $\mathcal{J} = (5, 10, 12, 15)$ are generated from a latent class model with 5 latent classes with equal membership probabilities and within-group probabilities generated from a symmetric Dirichlet distribution with unit parameters. The remaining variables are generated from independent Dirichlet-Multinomial distributions with hyper-parameter $\boldsymbol{\alpha} = (3, 3, 3, 3)^\intercal$. This setting generates a subset of variables with a relevant dependence structure, while leaves the other unstructured. In the second scenario, variables $\mathcal{J} = (1, 2, 3, 4, 5)$ are generated from a dense log-linear model with first order interactions. Main effects are randomly sampled from standard Gaussians, while coefficients associated to the same interaction term are sampled from standard Gaussians and constrained to have the same sign. Since the dependence structure is explicitly parametrised, this settings allows complete control on its intensity and direction, while in the first setting this operation was not possible. The remaining variables are generated from independent Dirichlet-Multinomial distributions with hyper-parameter $\boldsymbol{\alpha} = (3, 3, 3, 3)^\intercal$. Note that simulation from a log-linear model is doable since we are considering a small subset of categorical variables, with $|\mathcal{J}| = 5$. Simulation from a log-linear model encompassing a substantially larger groups of variables is unfeasible, since it would lead to the same issues described in Section 2.3.1. The third and last scenario

builds on the second one, but focus on imposing additional structure in the subset of variables $\mathcal{J}' = (5, 6, 7, 8, 9, 10)$, generated from a dense hierarchical log-linear model with second order interactions. Main effects are sampled from standard Gaussians, while coefficients associated to the same interaction term are sampled from standard Gaussians and constrained to have the same sign.

The focus of these settings is on inducing challenging data generating processes, characterised by complex dependencies over subsets of specific categorical variables. Posterior inference in each scenario for the MILLS approach relies on 3000 iterations collected after a burn-in period of 2000, setting a conservative upper bound $H = 5$ and $\sigma_2 = 3$. The priors for the log-linear coefficients were centered on 0 in each mixture kernel. As a competitor approach, we considered a finite approximation of the PARAFAC tensor decomposition proposed in Dunson and Xing (2009), approximating the Dirichlet Process with a finite Dirichlet distribution with 10 components and sparse hyper-parameter. Estimation is performed via Hamiltonian Monte-Carlo using the R package `rstan` (Team, 2018) and simulating 3000 iterations after a burn-in period of 2000.

Posterior inference focuses on marginal bivariate associations, measured via Cramer-V. In particular, we consider the posterior median of the Cramer-V under both methods and the posterior probability that the Cramer-V between variable j and variable $j\prime$ – denoted as $\rho_{j\,j\prime}$ – is above a specified significance threshold. Such an estimator can also be used to conduct a formal test to asses the presence of a significant bivariate relationship among pairs of categorical variables. Coherently with studies in social sciences, we choose a threshold of 0.2 for the Cramer-V, and focus only on those $\rho_{j\,j\prime} > 0.2$ (e.g. King *et al.*, 2008; Russo *et al.*, 2018).

Results are reported in Figure 2.1, 2.2 and 2.3, indicating a satisfactory performance of both methods in detecting the underlying dependence patterns. Indeed, first and second panel of Figure 2.1, 2.2 and 2.3, report the posterior median of the Cramer-V under the MILLS and the PARAFAC decomposition, respectively. Bivariate distributions with true Cramer-V greater than 0 are denoted are red crosses. Results suggest a good performance of both methods in estimating the Cramer-V under different data generating processes. However, the PARAFAC underestimates several associations patterns when the generating process is complex, such as the third scenario illustrated in Figure 2.3. This behaviour is confirmed in the third and fourth panels of Figure 2.1, 2.2 and 2.3, which illustrate the estimated $\hat{\text{pr}}(\rho_{j\,j\prime} > 0.2)$ under the MILLS and the PARAFAC decomposition, respectively, and confirming a tendency of the PARAFAC to assign very low

TABLE 2.1: Posterior estimates for the number of the active components in the simulation studies. The upper bound for MILLS is fixed at $H = 5$. Modal frequencies are denoted in boldface.

		1	2	3	4	5	6	7	8
Scenario 1	PARAFAC	0.000	0.000	0.000	0.000	**0.561**	0.463	0.165	0.011
	MILLS	0.166	**0.803**	0.030	0.001	0.000	·	·	·
Scenario 2	PARAFAC	0.000	0.000	0.003	**0.997**	0.000	0.000	0.000	0.000
	MILLS	**0.935**	0.064	0.001	0.000	0.000	·	·	·
Scenario 3	PARAFAC	0.000	0.000	0.000	0.000	**0.467**	0.440	0.093	0.000
	MILLS	**0.821**	0.174	0.005	0.000	0.000	·	·	·

evidence to some significant associations patterns, in particular under complex data generating process. This behaviour is not surprising, since the latent class model requires several components to adequately characterise complex dependence patters, while the more structured MILLS kernels mitigates such problem and provides satisfactory results with substantially fewer components. However, results from Figure 2.3 also suggest that MILLS might instead *overestimate* some associations more severely than a PARAFAC decomposition, providing an higher false discovery rate. This issue should be investigated with more attention, and it might be due to the inclusion of some redundancy in the composite-likelihood specification in Equation (2.7). Since such an issue is not particularly severe, and affects only a small number of settings and of association terms, we can arguably trust the results of posterior inference from the current specification.

To further confirm this intuition, Table 2.1 shows the posterior estimates for the number of active components under both approaches. This estimates are obtained computing, at each step of the MCMC, the total number of non-empty mixture components. For example, the first scenario provides evidence of 5 mixture components for the PARAFAC versus 2 components for MILLS. Similar results are achieved also in the second and third scenario, in which MILLS focuses on one mixture components while PARAFAC requires 4 or 5 components. These results are highly expected, and confirm the ability of the MILLS composite kernel to induce dependence patterns within each component, therefore allowing to model categorical data with fewer components than a PARAFAC decomposition, which leverage a conditional independence assumption and often requires a prohibitive number of mixture components (Johndrow *et al.*, 2017).

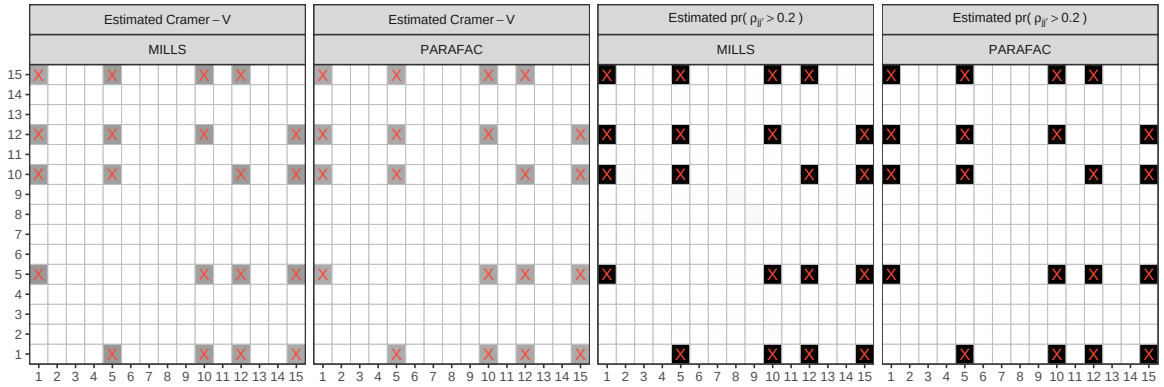

FIGURE 2.1: First scenario. Posterior median for the Cramer-V and $\text{pr}(\rho_{j\,j\prime} > 0.2)$ estimated under the MILLS approach and a competitor latent class model. Colors range from white to black as values range from 0 to 1. Red crosses indicate the bivariate distribution with significant association under the data generating process.

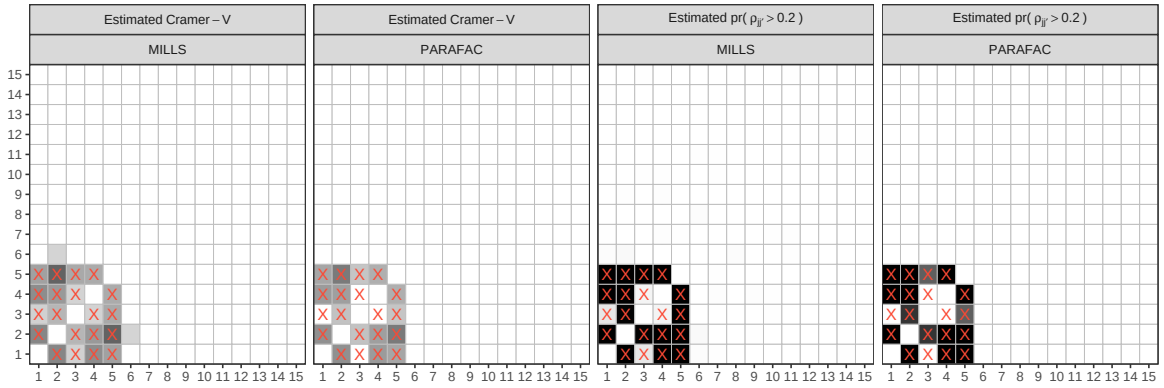

FIGURE 2.2: Second scenario. Posterior median for the Cramer-V and $\text{pr}(\rho_{j\,j\prime} > 0.2)$ estimated under the MILLS approach and a competitor latent class model. Colors range from white to black as values range from 0 to 1. Red crosses indicate the bivariate distribution with significant association under the data generating process.

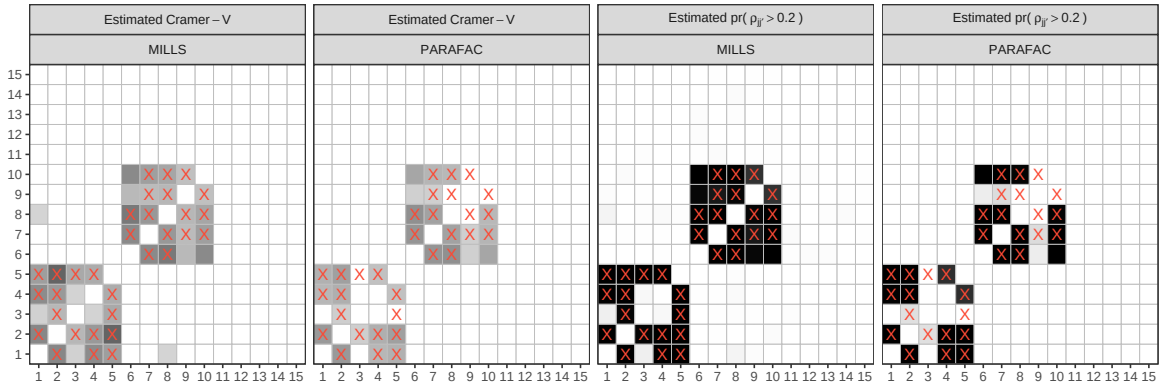

FIGURE 2.3: Third scenario. Posterior median for the Cramer-V and $\text{pr}(\rho_{j\,j\prime} > 0.2)$ estimated under the MILLS approach and a competitor latent class model. Colors range from white to black as values range from 0 to 1. Red crosses indicate the bivariate distribution with significant association under the data generating process.

2.5 Application

The method described in Section 2.3.2 is applied over the dataset illustrated in Section 2.2. Posterior inference is performed with a more conservative upper bound $H = 10$ and relying on 5000 MCMC samples with a burn–in of 2500, obtaining satisfactory convergence and mixing, measured via effective sample sizes and analysis of the trace plots. Posterior computation is based on a mixed R and C++ implementation and requires approximately 1 minutes per 1000 iterations on a laptop with an INTEL(R) CORE(TM) I7-7700HQ @ 2.8 GHZ processor and 16GB of RAM running Linux.

Differently from the simulations, we have opted for a different tool to visualise results from posterior inference. Indeed, it is useful to visualise the estimated association structure as a *network*, with nodes representing categorical variables and edges the presence or the degree of a specific pairwise measure of association. It is worth highlighting that the tool "network" is used here in the sense of the graphical modelling literature , where the focus is on assessing the presence of marginal or conditional dependence across a set of variables, and eventually measuring its intensity and direction (e.g. Lauritzen, 1996). Therefore, the sets of graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ correspond here to the index set of categorical

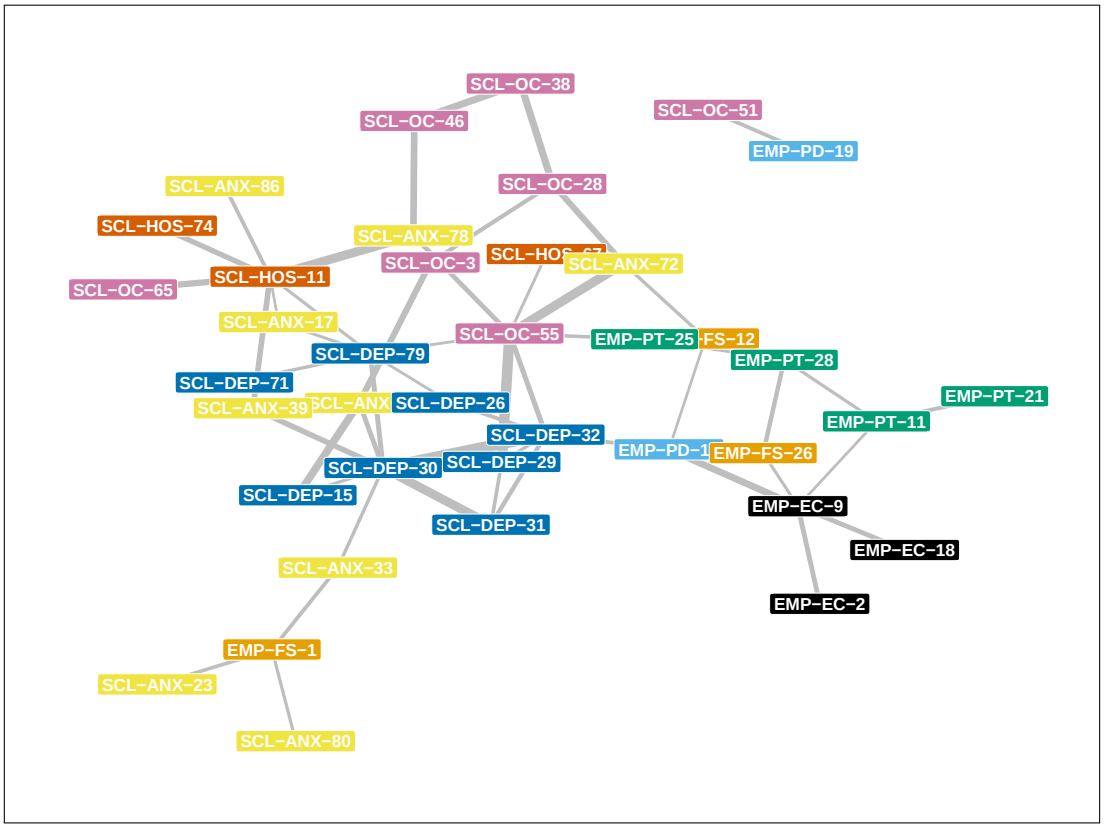

FIGURE 2.4: Association structure of the items. Color of the nodes varies with subscales, while edges width varies with the value of the posterior mean of the pairwise Cramer-V. Edges with values lower than 0.4 are not reported.

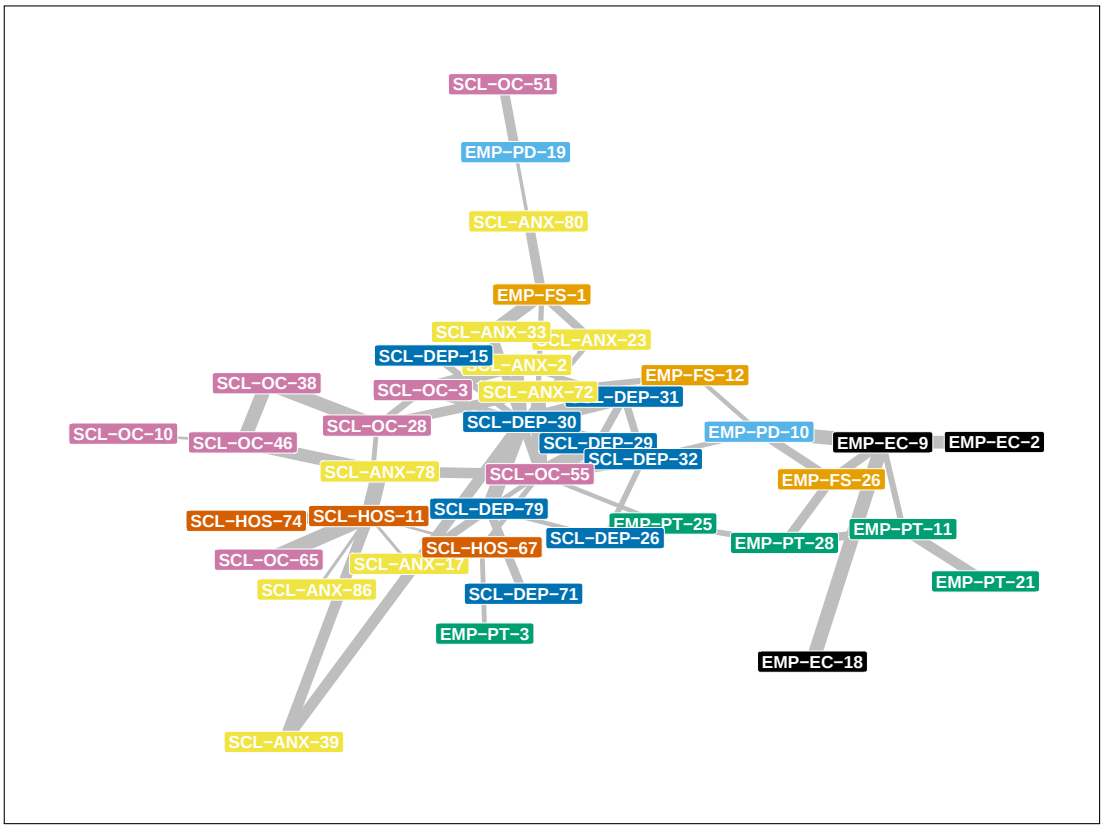

FIGURE 2.5: Association structure of the items. Color of the nodes varies with subscales, while edges width varies with the posterior probabilities $\hat{\text{pr}}(\rho_{j\,j\prime} > 0.2)$. Edges with values lower than 0.9 are not reported.

variables $V = \{1, \ldots, p\}$ and to the posterior estimates of the marginal association, respectively. Such an interpretation is totally different from the "data" interpretation of Chapter 1, and indeed highlights how network science can be a powerful ally in many field of application.

Figure 2.4 illustrates the posterior median of the estimated pairwise Cramer-V. Edges size varies according with the posterior median of the Cramer-V, with thicker edges corresponding to stronger associations. In order to further improve the graphical visualisation of the results, we have focused only on pairs with an estimated Cramer-V above 0.4, excluding from the visualisation nodes without connections. Results suggest a general tendency of items to create links with others that are in the same sub-scales, with items in the SCL-DEP being strongly associated and suggesting a strong associations among depressive profiles in suicide attempts patients. This result is expected and confirms the validity of the tools to measure psyco-pathological symptoms and empathic profiles. More interesting associations involve items in different subscales. For example, it is worth highlight the presence of an association between the items SCL-ANX-23 (*"Suddenly scared for no reason"*) and SCL-ANX-80 (*"Feeling that familiar things are*

strange or unreal") with the item from the IRI questionnaire EMP-1 (*"I daydream and fantasize, with some regularity, about things that might happen to me"*). Another interesting association involves the items SCL-51 (*"Your mind going blank"*) and IRI-19 (*"I am usually not effective in dealing with emergencies."*), likely to define individuals with low-capacity to handle panic situations.

Figure 2.5 further refines this findings providing additionally insights on the association structure among the items. Specifically, the posterior probabilities $\hat{\text{pr}}(\rho_{j\,j'} > 0.2)$ are computed for each pair of variable post-processing the MCMC output. Figure reports only edges associated with estimated posterior probabilities greater than 0.5. The main empirical findings are similar with Figure 2.4, suggesting a general preference to observe strong associations within the same subscales, and suggesting the presence of connections between depressive states and obsessive compulsive profiles.

Chapter 3

Latent structures models for removing dependence

3.1 Biased data

In the previous chapters, the use of latent variable modelling was motivated as a strategy to characterize complex dependence structures in an unobservable latent space, therefore modelling the observed quantities as conditionally independent given the latent variables. This modelling approach allows one to take into account intricate structures balancing flexibility with a drastic reduction in the number of parameters, and conduct posterior inference efficiently in involved case studies. The complexity of the dependence structures of this chapter arise instead from different considerations on the data generating process and the research objectives. Specifically, it is of interest to characterize such structures in order to restrain some specific aspects and achieve conditions motivated by peculiar case studies. Indeed, we will show that modification of the main strategy used so far can be successfully introduced to solve these more broader issues.

Recently, there has been growing interest in developing algorithms and statistical tools to assist, and eventually replace, humans in high-stakes decision processes. Some examples include credit scoring, hiring and sentencing, among many others (e.g. Dunson, 2018). The desire for introducing algorithms in such decisions settings comes from different considerations, ranging from attempting to reduce unethical differences to improving the efficiency and costs of the recruitment process (e.g. Friedler *et al.*, 2019). Indeed, it has been often argued that decisions relying on automated systems would be automatically *fair*, efficient and objective, since computers are not supposed to encode prejudices or to make decision driven by subjective arguments. Fairness and objectivity are crucial in high-stake decision processes, since these outcomes have significantly

impact on the entire society as a whole and on individuals.

The relevance of this problem has motivated different research threads on evaluating fairness of automated procedures. Indeed, the centrality of the data in the development of statistical tools becomes even more relevant in the context of fair decisions making; see, for example, the report of Munoz *et al.* (2016). Modern statistical applications often rely on large datasets collected within observational studies for a convenience sample of individuals, where interest is on detecting relationships between features and outcome variables; for example, diseases or behaviors. Such processes create datasets in which the sampling mechanism is complex, unknown, and often subject to some form of systematic bias which might affect fairness and objectivity of decisions based on automated tools (Dunson, 2018). Indeed, when selection bias exists in the sampling mechanism, the data often encode spurious associations, and there is growing recognition that machine learning algorithms will reproduce and often amplify bias in the data upon which they were trained (e.g. Angwin *et al.*, 2016; Zech *et al.*, 2018). For example, recruiters in high-tech companies commonly short-list potential candidates on the basis of their curriculum-vitae, leveraging algorithms trained on past interviews (e.g. Hoffman *et al.*, 2017). If a systematic gender-gap has been observed in the past, this tendency will be propagated into the estimated algorithms and into future predictions, therefore amplifying such phenomena instead of mitigating it.

More recently, great attention has been devoted to the *algorithmic* aspect of the problem of fair high-stake decision making (Corbett-Davies *et al.*, 2017). This line of research has focused on the development of algorithms to predict an outcome of interest as a function of different covariates, excluding from the analysis variables measuring sensitive attributes; for example, predicting the salary during a job interview on the basis of the information contained in the curriculum-vitae, excluding gender of the candidate to avoid gender-gap effects. However, strong associations are often observed among sensitive attributes and other demographic features, and the naive exclusion of sensitive information is not sufficient to mask those information, which can still be retained on the basis of other covariates. More recently, researchers have proposed to include specific penalisation into the algorithms, optimizing models with respect to *ad-hoc* loss functions; for example, equating the proportions, across ethnic groups, of individuals incorrectly classified (Dwork *et al.*, 2012). Such an approach requires to include very specific constraints, which are often impossible to be satisfied simultaneously (Friedler *et al.*, 2016) and are limited to the specific model which included them.

3.2 Criminal justice bias

An important area in which unwanted associations arise is in criminal justice data. There has been much recent attention on the use of criminal risk assessment models, many of which use demographic, criminal history, and other information to predict whether someone who has been arrested will be re-arrested in the future. These predictions then inform decisions on pre-trial detention, sentencing, and parole. In practice, the data used to train the models are based on police arrest records, which are well known to be subject to racial bias (Simoiu *et al.*, 2017; Rudovsky, 2001). For example, different studies have highlighted that records in police databases are not representative of the phenomena of interest (criminality), likely due to police patrols discretion in choosing which neighborhoods should be patrolled and who should be detained (Lum and Isaac, 2016). When risk assessment models are trained on such data, racial minority groups which are oversampled in the training data tend to be systematically assigned to higher risk categories on average (e.g. Lum and Isaac, 2016; Angwin *et al.*, 2016).

In this chapter we will focus on an arrest record database, which we refer to as COMPAS dataset (Angwin *et al.*, 2016). Data consists of police records for $n = 6180$ defendants in Broward County, Florida, collected during 2013 and 2014 and referred to detection of minor offenses; for example, marijuana possession. The database is publicly available at the link `github.com/propublica/compas-analysis`; see also Larson *et al.* (2016) for a detailed description of the data collection process. For each individuals, several demographic information are available, including age, sex, and information on prior offenses, along with their outcomes within 2 years of the decision and the race of the defendant. We will focus the analysis on the demographic information and all the interactions terms among them, for a total of $p = 63$ covariates.

To confirm our intuition on the presence of racial bias, Figure 3.1 reports the empirical distribution functions for the out-of-sample predictions of the probability of recidivism. Specifically, a Bayesian logistic regression estimated on the training data was used to provide estimates for the probability of recidivism over an independent test set. The empirical distribution of the predicted values is compared across racial groups in Figure 3.1. Predictions are reported for two different regressions, including race as a control and removing it from the analysis, respectively in the first and second panel. Results suggest that non-Caucasian individuals are systematically assigned to higher risk of recidivism, with the black dotted curves being shifted towards higher risks of recidivism. These results also suggest that removing racial information from the analysis, as suggested for a long time by practitioners, is basically ineffective in mitigating systematically different

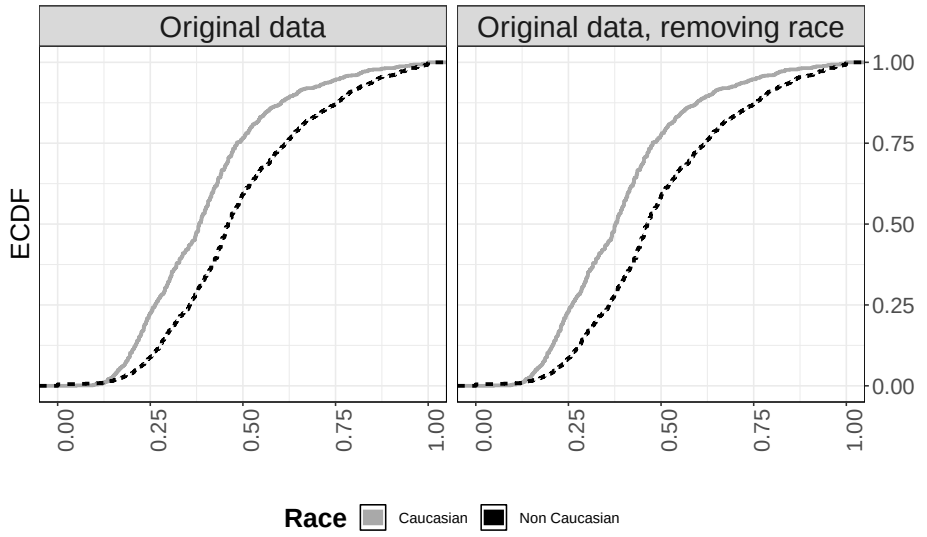

FIGURE 3.1: Out-of-sample predictions on the probability of recidivism for the COMPAS dataset. Gray continuous lines refer to Caucasian, black dotted to non-Caucasian.

predictions.

The focus of this chapter is on providing an adjusted dataset to guarantee that any algorithm can be estimated with strong guarantees in terms of fairness of predictions. Specifically, predictions are defined as "fair" (with respect to race) if they are independent from race (Johndrow and Lum, 2019). Leveraging latent structure modelling, we characterize the dependence among the observed data, and then we *constrain* such structure to remove aspects which depend on racial information, thereby guaranteeing fairness of predictions obtained from models estimated on such structures. This aim is addressed with a constrained Gaussian Latent Factor Model for continuous data, which decomposes the dependence structure using a compact representation which retains the main features of the data and impose independence among the estimated latent structure and the sensitive attributes via constrained optimisation.

3.3 Gaussian Latent Factor Model

3.3.1 Model specification

Let $\mathbf{X}$ denote an $n \times p$ data matrix of p features measured over n subjects, and let $\mathbf{Z}$ denote an additional group membership variable. Lastly, let $\mathbf{Y}$ define a response variable. In the recidivism example, these quantities correspond to the data matrix of demographic information, the race of the defendant and two-years recidivism, respectively. The focus of our approach is on providing a general procedure to obtain predictions for $\mathbf{Y}$, denoted as $\hat{\mathbf{Y}}$, such that $\hat{\mathbf{Y}} \perp\!\!\!\perp \mathbf{Z}$. Following Johndrow and Lum (2019), we develop

procedure to create an adjusted matrix $\widetilde{\mathbf{X}}$ with $\widetilde{\mathbf{X}} \perp\!\!\!\perp \mathbf{Z}$. This condition is sufficient to guarantee that predictions $\hat{\mathbf{Y}}$ based on the adjusted matrix are also independent of $\mathbf{Z}$ (Johndrow and Lum, 2019). Since such a procedure focuses directly on preprocessing the data, instead of on a specific algorithm to predict recidivism, it allows to estimate *any* algorithm on the adjusted matrix $\widetilde{\mathbf{X}}$ with guarantees of fairness of the predicted values. In order to address this aim, we propose an explicit statistical model for the data matrix $\mathbf{X}$, which allows us to characterise the dependence structure across the different features leveraging a low-dimensional latent variable representation, and to impose further constraints across such representation and the protected variable $\mathbf{Z}$ during estimation.

Let $\mathbf{x}_i^\intercal \in \mathbb{R}^p$ define a generic p-dimensional row of $\mathbf{X}$. We will suppose that $\mathbf{x}_i$ is generated from a Gaussian latent factor model, specified as follows.

$$\begin{aligned} (\mathbf{x}_i \mid \mathbf{\Lambda}, \mathbf{w}_i, \mathbf{\Sigma}) &\sim \mathrm{N}_p(\mathbf{\Lambda}\mathbf{w}_i, \mathbf{\Sigma}), \\ (\mathbf{w}_i) &\sim \mathrm{N}_k(0, I_k) \quad i = 1, \dots, n, \end{aligned} \tag{3.1}$$

with $\mathbf{\Sigma} = \mathrm{diag}(\sigma_1, \dots, \sigma_p)$, $\mathbf{\Lambda}$ $p \times k$ loading matrix and $\mathbf{w}_i$ k-dimensional latent factors. Using a matrix notation, the matrix $\mathbf{X}$ generated from a Gaussian latent factor model can be represented ad follows.

$$\mathbf{X} = \mathbf{W}\mathbf{\Lambda}^\intercal + \boldsymbol{\varepsilon}, \tag{3.2}$$

with $\mathbf{W}$ denoting the $n \times k$ latent factors matrix and $\boldsymbol{\varepsilon} = (\epsilon_1, \dots, \epsilon_n)^\intercal$ denoting the $n \times p$ error matrix, with $\epsilon_i^\intercal \sim \mathrm{N}_p(0, \mathbf{\Sigma})$ for $i = 1, \dots, n$. Popular choices for identifiability impose some additional structure in the loading matrix $\mathbf{\Lambda}$; for example, lower triangular constraints (Geweke and Zhou, 1996). This specification allows to remove the invariance of the latent space with respect to orthogonal transformations, and provides a direct way to interpret the latent factor by ordering them in terms of their relative importance. In our case study, interpretation of the latent factors is not required, as long as their joint structure captures the relevant properties of the data. Indeed, the choice of the coordinates used to describe such latent space is not relevant, and therefore we prefer to avoid constraints and rely on the over-parametrised representation of Equation (3.1).

Conditionally on $\mathbf{w}_i$, each $\mathbf{x}_i$ is assumed to be uncorrelated given the latent factors. Similarly to the latent structure methods of Chapter 2, dependence is explicitly obtained marginalising over the distribution of the latent factors, with standard Gaussian theory

showing that

$$(\mathbf{x}_i \mid \boldsymbol{\Lambda}, \boldsymbol{\Sigma}) \;\sim\; \mathrm{N}_p(\mathbf{0}, \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\intercal + \boldsymbol{\Sigma}), \quad i = 1, \ldots, n.$$

The Gaussian latent factor model, similarly to the approaches discussed in Chapter 1 and Chapter 2, provides a simple and effective way to model high-dimensional data with intricate dependence structures using a moderate number of parameters. Indeed, the matrix $\mathbf{W}$ provides a natural candidate for a low-dimensional representation of the data, in analogy with Probabilistic PCA (Tipping and Bishop, 1999) and other probabilistic dimensionality reduction techniques (e.g. Bishop, 2006, chapter 12). It is also worth stressing that the dependence structure among the columns $\mathbf{X}$ is entirely modelled through the latent factors $\mathbf{W}$, while from the loading matrix $\boldsymbol{\Lambda}$ specifies the location of each column of $\mathbf{X}$ as a linear combination of the columns of $\mathbf{W}$. Therefore, we focus on estimating a *constrained* version of the latent factors $\mathbf{W}$, which guarantees an accurate low-dimensional approximation of $\mathbf{X}$ and is such that $\mathbf{W} \perp\!\!\!\perp \mathbf{Z}$.

3.3.2 Constrained Bayesian Inference

A Bayesian approach to inference is followed. For a complete specification of the model defined in Equation (3.1), we specify prior distributions for the elements of the loading matrix $\boldsymbol{\Lambda}$ and of $\boldsymbol{\Sigma}$. For computational convenience, conjugate priors are specified, thereby letting

$$\boldsymbol{\lambda}_j \sim \mathrm{N}_k(0, \boldsymbol{\Psi}), \quad \sigma_j^2 \sim \mathrm{Gamma}(a_0, b_0) \quad j = 1, \ldots, p, \tag{3.3}$$

with $\boldsymbol{\Psi} = \mathrm{diag}(\psi, \ldots, \psi)$. This choice allows one to obtain closed form expressions for the full conditional distributions and, for example, implement a simple Gibbs Sampler algorithm to draw from the posterior distribution $p(\mathbf{W}, \boldsymbol{\lambda}, \boldsymbol{\Sigma} \mid \mathbf{X})$ or derive an highly-scalable MFVB routine, similarly to the approach followed in Section 1.5.3. This aim can be addressed recasting the Gaussian latent factor model expressed with matrix notation in Equation (3.1) into conditionally multivariate linear regressions, in the same spirit of the LFM for network data of Section 1.5; see also Lopes and West (2004) for related arguments. Although a complete characterisation of the posterior distribution is crucial in a large number of applications, the case study of this chapter justifies faster and scalable inferential procedures focusing only point estimates. Indeed, since interest is on providing a single low-dimensional representation of the data, a precise quantification of the uncertainty of the estimation process is not necessary for addressing such an aim.

Computational methods to estimate the posterior mode of high-dimensional distributions — sometimes referred to as Maximum-A-Posterior (MAP) estimation — are crucial in a large variety of applications. Indeed, several numerical problems can be recasted as optimisation procedure for MAP estimation, thereby allowing to introduce recent advances in Bayesian optimisation (e.g. Cockayne *et al.*, 2018). Among different methods for optimising high-dimensional functions, the EM algorithm (Dempster *et al.*, 1977) is certainly one of the most popular ones. Although it is more commonly used for likelihood maximisation, the EM algorithm naturally extends to MAP estimation with minor modifications (Dempster *et al.*, 1977).

A straightforward application of the Bayes theorem allows us to express the posterior distribution as follows.

$$p(\mathbf{W}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma} \mid \mathbf{X}) = \frac{p(\boldsymbol{\Lambda})p(\boldsymbol{\Sigma}) \prod_{i=1}^{n} p(\mathbf{x}_i \mid \boldsymbol{\Lambda}, \mathbf{w}_i, \boldsymbol{\Sigma})p(\mathbf{w}_i)}{\int\int\int p(\boldsymbol{\Lambda})p(\boldsymbol{\Sigma}) \prod_{i=1}^{n} p(\mathbf{x}_i \mid \boldsymbol{\Lambda}, \mathbf{w}_i, \boldsymbol{\Sigma})p(\mathbf{w}_i) \mathrm{d}\mathbf{W}\boldsymbol{\Lambda}\mathrm{d}\boldsymbol{\Sigma}}, \tag{3.4}$$

with $p(\mathbf{x}_i \mid \boldsymbol{\Lambda}, \mathbf{w}_i, \boldsymbol{\Sigma})$ and $p(\mathbf{w}_i)$ defined in Equation (3.1) and the prior distributions in Equation (3.3). Since the parameters are marginalised out in the normalising constants, it follows that

$$\arg\max_{\mathbf{W},\boldsymbol{\Lambda},\boldsymbol{\Sigma}} p(\mathbf{W}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma} \mid \mathbf{X}) = \arg\max_{\mathbf{W},\boldsymbol{\Lambda},\boldsymbol{\Sigma}} p(\boldsymbol{\Lambda})p(\boldsymbol{\Sigma}) \prod_{i=1}^{n} p(\mathbf{x}_i \mid \boldsymbol{\Lambda}, \mathbf{w}_i, \boldsymbol{\Sigma})p(\mathbf{w}_i), \tag{3.5}$$

and therefore maximising the posterior distribution is equivalent to maximising the likelihood function times the prior distribution.

This results notably simplifies MAP estimation via EM algorithm, since it allows to adapt a standard EM routine for MLE introducing the contribution of the prior only in the M step of the algorithm as a penalty term (McLachlan and Krishnan, 2007, Section 6.5). In case of conditionally conjugate exponential families with closed-form M-step, this extension is further simplified since it generally require minor algebraic adaptations.

We modify the EM algorithm for MAP estimation in the Gaussian latent factor model by adding an additional step to impose further constraints in the latent vectors $\mathbf{W}$. The Gaussian specification simplifies such a constraint since uncorrelation also implies independence under a multivariate Gaussian specification; see also Takai (2012) for related arguments on constrained EM algorithms and their properties.

Specifically, following Bishop (2006, Sec. 12.2.4), the E-step of the algorithm leads to

$$\mathbb{E}[\mathbf{w}_i] = \mathbf{G}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\mathbf{x}_i, \quad i = 1, \ldots, n \tag{3.6}$$

with $\mathbf{G} = (I_k + \boldsymbol{\Lambda}^\intercal\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})^{-1}$. Such expectations are adjusted with a projection step in which we compute the residuals of a multivariate regression of $\mathbb{E}[\mathbf{W}]$ over $\mathbf{Z}$, thereby letting

$$\hat{\mathbf{w}}_i = \mathbb{E}[\mathbf{w}_i] - \mathbf{Z}(\mathbf{Z}^\intercal\mathbf{Z})^{-1}\mathbf{Z}^\intercal\mathbb{E}[\mathbf{W}]. \tag{3.7}$$

The M-step is available in closed form, plugging the adjusted $\hat{\mathbf{w}}_i$. This choice leads to

$$\begin{aligned} \boldsymbol{\Lambda} &= \left[\sum_{i=1}^{n} \mathbf{x}_i\hat{\mathbf{w}}_i^\intercal\right]\left[n\mathbf{G} + \sum_{i=1}^{n}\hat{\mathbf{w}}_i\hat{\mathbf{w}}_i^\intercal + \boldsymbol{\Psi}^{-1}\right]^{-1} \\ \sigma_j &= \left[2b_0 + \sum_{i=1}^{n}(x_{ij} - \boldsymbol{\lambda}_j^\intercal\hat{\mathbf{w}}_i)^2\right][2a_0 + n + 2]^{-1}. \end{aligned} \tag{3.8}$$

The algorithm proceeds iteratively until variations in the parameters are sufficiently small. In our experience, speed of convergence can be greatly improved with appropriate initialisations; for example, initialising $\hat{\mathbf{w}}_i$ at the left singular vectors of $\mathbf{X}$. We refer to this procedure as Constrained Maximum a Posterior for the Gaussian latent factor model in the sequel; for brevity, CMAP.

3.4 Simulation Study

We conduct a simulation study to evaluate the empirical performance of the proposed algorithm. The focus of the simulations is on assessing the success in removing the influence of the group variable from predictions for future subjects and evaluation of

TABLE 3.1: Simulation studies. Out-of sample prediction of the response variable

		PSVA	COMBAT	CMAP
Scenario 1	RMSE	82.26	41.66	**31.08**
	MAE	91.94	30.49	**22.76**
	MDAE	32.88	21.69	**15.02**
Scenario 2	RMSE	19.29	17.54	**11.25**
	MAE	15.46	14.19	**9.15**
	MDAE	13.11	12.58	**8.1**
Scenario 3	RMSE	20.82	**12.77**	12.88
	MAE	16.38	**10.10**	10.12
	MDAE	13.65	10.26	**7.93**

the goodness of fit of predictions. We also compare our method with two competitor approaches developed in biostatistics. Specifically, we focus on the COMBAT method of Johnson *et al.* (2007) and the PSVA approach of Leek and Storey (2007). Indeed, there is a strong connection among the problem of fair predictions and batch effect removal in biostatistics, where high dimensional data are subject to selection biases due to the experimental design (Aliverti *et al.*, 2019, 2018).

The first scenario focuses on a setting with $n = 1000$, $p = 200$ and true low-rank structure for $\mathbf{X}$ with $k = 10$. Specifically, the data matrix $\mathbf{X}$ is constructed in two steps. Firstly, we simulate a loading matrix $\mathbf{S}$, with size (n, k), and a score matrix $\mathbf{U}$ with size (k, p), with entries sampled from independent Gaussian distributions. A group variable $\mathbf{Z}$ of length n is sampled from independent Bernoulli distributions with probability equal to 0.5. Each p-dimensional row of the $(n \times p)$ data matrix $\mathbf{X}$ is drawn from a p-variate standard Gaussian distribution with mean vector $\mu_i = (s_i - \lambda z_i)\mathbf{U}$, $i = 1, \dots, n$ and λ sampled from a k-variate Gaussian distribution. Lastly, a continuous response $\mathbf{Y}$ with elements y_i, $i = 1, \dots, n$ is sampled from independent Gaussians with mean $(s_i - \lambda z_i)\beta$ and elements of β sampled uniformly in $(-5, 5)$. We highlight that in this setting the data matrix $\mathbf{X}$ has a low-rank structure and the response variable $\mathbf{Y}$ is a function both of the group variable $\mathbf{Z}$ and on the low-dimensional embedding of $\mathbf{X}$.

In the second setting the generating process for data matrix $\mathbf{X}$ is identical to the first setting, while the response variable $\mathbf{Y}$ does not depend on $\mathbf{Z}$. Indeed, elements y_i of $\mathbf{Y}$ are sampled from standard Gaussians with mean vector $\mu_i = s_i\beta$, $i = 1, \dots, n$. Therefore, the response $\mathbf{Y}$ depends only on the low-dimensional embedding of $\mathbf{X}$. The third setting focuses on a "large p - small n" setting, in which the dimension of the data matrix $\mathbf{X}$ is $n = 100, p = 2000$ with $k = 10$. The construction of the matrix $\mathbf{X}$ follows the first setting, and dimensions of the score and loading matrix are changed accordingly.

In each setting, data are divided into a training set and a test set, with size equal to 3/4 and 1/4 of the observations, respectively. Therefore, the number of observations in the training and test set is equal to $(750, 250)$ in the first, second and fourth scenario, and equal to $(150, 50)$ in the third. Adjustment methods are applied separately on the train and test sets. Separate linear regressions are estimated on the adjusted training sets, and predictions $\hat{\mathbf{Y}}$ are provided for the adjusted test sets.

Table 3.1 reports the root mean square error (RMSE), mean absolute error (MAE) and median absolute error (MDAE) for the adjusted predictions from the competitors (COMBAT, PSVA) and the proposed method (CMAP). Results suggest that the performance of CMAP for the Gaussian latent factor model is similar to the competitors in

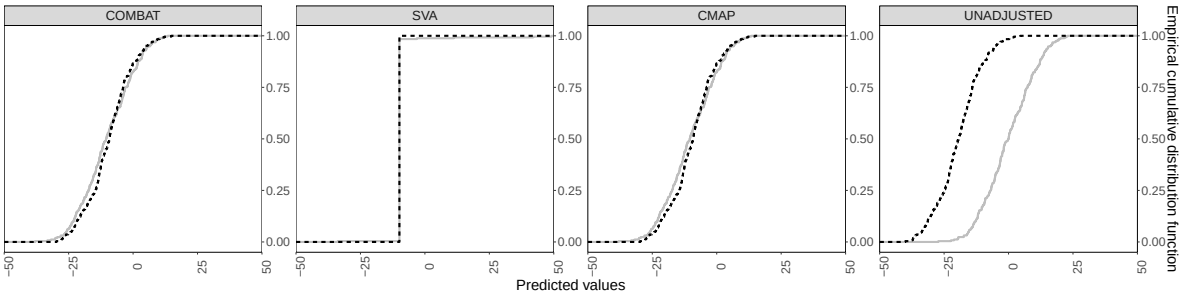

FIGURE 3.2: Empirical cumulative distribution functions for $\hat{\mathbf{Y}}$ in the first simulation scenario.

terms of predicting the response $\mathbf{Y}$. Specifically, the proposed method outperforms the competitors in the first and second scenario, with a better performance with respect to all the metrics considered. In the third scenario, we observe a slight better performance for the PSVA. However, it is worth mentioning that the performance of CMAP is very close with the best performing competitor.

Figure 3.2 illustrates the predictive performance of the three methods over the independent test set. Predictions are reported also for an unadjusted case (fourth panel) to illustrate that a simple model without adjustment actually leads to relevant difference across predictions. Figure 3.2 also justifies the substantially worse performance of SVA in the first setting. Although the predictive gap has been reduced, predictions are quite poor and the method is not able to adjust predictions without loosing in predictive power. Instead, COMBAT and CMAP report reasonable results, both in terms of predictive power and in term of reduction of predictive gap.

3.5 Application to the criminal justice dataset

The adjusted Gaussian latent factor model is applied over the dataset described in Section 3.2. Data was randomly divided into a training set and a test set, with 3/4 and 1/4 of observations, respectively; CMAP was independently applied over the two sets, and a logistic regression and a random forest estimated on the adjusted training set was used to provide predictions on the test set.

Figure 3.3 compares the out-of-sample predictive distribution for the predicted risk of recidivism under a random forest (left panel) and a logistic regression (right panel). Compared with the motivating Figure 3.1, the gap between the two curves is notably reduced, leading to predictions which are more similar across different racial groups for both methods. Specifically, under the proposed approach, predictions for different racial groups are very similar in terms of estimated probabilities, although the predictive gap is not completely removed.

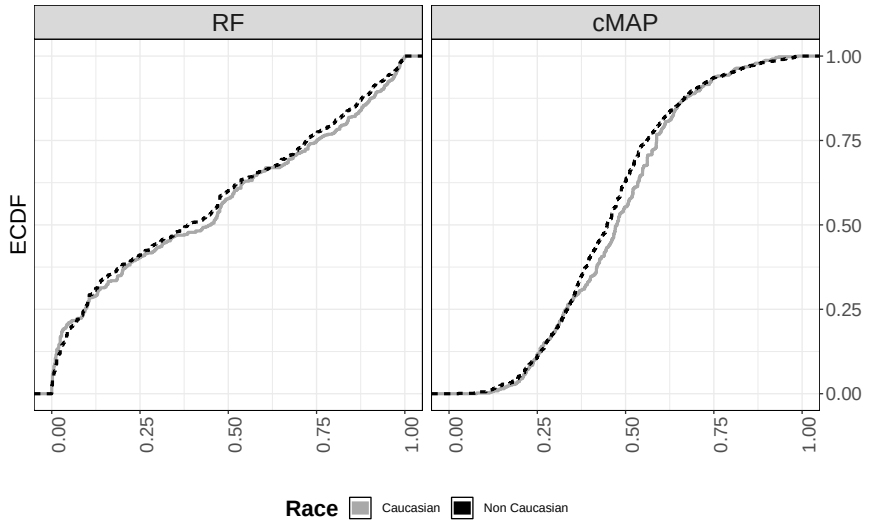

FIGURE 3.3: Empirical cumulative distribution functions for $\hat{\mathbf{Y}}$ under two adjusted approaches. Light solid gray refers to white ethnicity, dotted black to non-white.

Table 3.2 reports results for the model previously described and other competitor approaches. The first and second columns of Table 3.2 represent, respectively, results for a logistic regression using all the available original covariates and their interaction terms (LOGISTIC) and all the variables, interaction terms and race (LOGISTIC,Z). The column compares instead predictive performance for a random forest (RF), using all the unadjusted available covariates and their interactions. Predictive performance is measured via Accuracy (ACC), Area Under the Roc Curve (AUC), True Positive Rates (TPR), True Negative Rates (TNR), Positive Predicted Values (PPV) and Negative Predicted Values (NPV) for the out-of-sample predictions. Results are averaged over 50 splits over train and test set, and standard deviations across splits are reported in brackets. Results for the unadjusted procedures suggest a reasonably good performance in predicting the risk of recidivism, with the logistic regression which includes racial information being the overall best model. However, as highlighted in Figure 3.1, such predictions are systematically different by race, producing predictions for black individuals which are assigned to higher values on average.

TABLE 3.2: Predictive performance on the COMPAS dataset.

	LOGISTIC	LOGISTIC, Z	RANDOM FOREST	LOGISTIC,CMAP	RF, CMAP
ACC	0.669 (0.02)	**0.671** (0.01)	0.671 (0.01)	**0.654** (0.01)	0.606 (0.01)
AUC	0.712 (0.02)	**0.714** (0.02)	0.712 (0.01)	**0.708** (0.01)	0.642 (0.01)
TNR	0.719 (0.02)	0.716 (0.04)	**0.766** (0.02)	0.659 (0.02)	0.608 (0.01)
TPR	0.609 (0.05)	**0.617** (0.03)	0.558 (0.02)	**0.649** (0.01)	0.602 (0.02)
PPV	0.644 (0.02)	0.645 (0.02)	**0.665** (0.02)	0.613 (0.01)	0.562 (0.02)
NPV	0.689 (0.02)	**0.692** (0.01)	0.675 (0.01)	**0.692** (0.01)	0.647 (0.01)

The second part of Table 3.2 illustrates results for the adjusted procedures. Each adjusted method relies on datasets adjusted trough CMAP with $k = 20$. Specifically, CMAP adjustment was estimated independently over the train and test set, and we rely on a logistic regression model (LOGISTIC, CMAP) and a random forest (RF, CMAP) estimated over the adjusted training data to perform predictions over the adjusted test sets. Predictive performance after adjustment is highly comparable with the unadjusted procedures; for example, AUC for logistic regression drops from 0.712 to 0.708. The empirical distributions for the predicted values is illustrated in Figure 3.3, and suggests that adjustments is successful at removing dependence in predictions from both approaches. These results, combined, provide a compelling argument in favour of the proposed method, which allows to successfully remove dependence of predictions without significantly affecting predictive power.

Conclusions

Discussion

This thesis has focused on Bayesian modelling for complex dependent structures. Through different case studies, it has been shown how latent variable specifications provide a powerful modelling strategy to characterise dependence structures with flexible specifications amenable to efficient computational algorithms. Modelling dependence in an unobservable space also allows to improve the interpretation of the results, focusing on a compact representation of the data which characterises the main property of the rich dependence structure.

The aim of Chapter 1 has been on illustrating the use of latent structure specifications with network data. Specifically, novel latent space models are developed for learning shared anatomical effects and latent structures underlying replicated brain networks, focusing on two applications involving structural brain scans. The main empirical findings are consistent across the two case studies, suggesting a general preference for the brain regions to connect with others that are spatially closer and within the same hemisphere. Moreover, both applications indicate an interesting relationship between the estimated positions of the brain regions in the latent space and their anatomical counterpart, confirming the presence of an intricate structure in the brain architecture which cannot be explained only in terms of anatomical determinants.

Chapter 2 has focused on latent structures for multivariate categorical data, providing a novel methodology to combine the benefits of latent class analysis with log-linear modelling. The findings on the psychiatric case study of interest shown interesting association patterns across different psychological profiles, suggesting important associations among patients with depressive states and obsessive compulsive disorders, and among the fantasy component of empathy and empathic concern. Moreover, there is evidence of significant associations across the fantasy component of empathy and some psychological symptoms. Our approach makes a first step toward understanding the

relation across important psychological dimension in suicide attempt patients, and the inferential results can contribute to the scientific knowledge on the topic.

The focus of Chapter 3 has been on illustrating the use of latent structures to remove dependence patterns in high-stakes decisions processes, focusing on a case study in criminal justice. The methods illustrated in Chapter 3 significantly reduce the predictive gap in predictions of recidivism across black and white defendants, providing a model-based pre-processing tool which allows to estimate any algorithm on the adjusted data. Moreover, accuracy of adjusted predictions is only slightly reduced after adjustment, providing an additional important motivation for introducing such an approach in courtrooms.

Future directions

Some potential future directions are certainly worth to be mentioned. The methods of Chapter 1 relate with an increasing interest in Bayesian approaches for multi-resolution medical imaging data (e.g. Peruzzi and Dunson, 2018). Clearly, high quality scans require more expensive capacities and efforts in terms of medical equipment and data storage, which do not automatically lead to more interesting inferential results. A desirable objective is to investigate the *optimal* resolution of such scans, providing the best trade-off between costs and benefits in terms of power. An alternative improvement involves the modelling of more detailed data, such as white fibers counts instead of their presence or absence. Some approaches which might be useful to develop novel tools in this direction comes from the ecological literature, where there has been much interest in modelling species counts via latent determinants (e.g. Gotelli and Ellison, 2004).

Some developments for Chapter 2 involve a more rigorous characterisation of the theoretical properties of the baseline MILLS model. There has been some interest in the literature on developing asymptotic arguments for Bayesian modelling with non-standard likelihood function; for example, Pauli *et al.* (2011). Also, the statistical efficiency of the algorithm could be improved by introducing additional latent variables characterising groups of categorical variables with similar associations. This direction is currently under investigation.

Chapter 3 motivates further extensions to deal with non-continuous data, relaxing the multivariate Gaussian assumption explicitly allowing for variables on discrete scales, such as counts and categorical variables. One possibility to address this objective is to introduce a link function similarly to a GLM specification, therefore considering the GLFM as a further latent model for the linear predictor of such specification.

Appendix A

Appendix for Chapter 1

A.1 Latent space model with local clustering

A.1.1 Computational Details

Let $\mathbf{S}$ denote the $n(n-1)/2$ vector comprising the lower–triangular elements of $\mathbf{A}$, and define $\bar{\mathbf{S}} = \mathbf{S} - (m/2) \cdot \mathbf{1}_{n(n-1)/2}$, with $\mathbf{1}_{n(n-1)/2}$ a $n(n-1)/2$ vector of ones. Moreover, let $\mathbf{X}$ be the $n(n-1)/2 \times 4$ matrix with a first column equal to $\mathbf{1}_{n(n-1)/2}$, and the remaining three comprising the vectorized version of the edge covariates `hem`, `lobe`, and d. Finally, define $\bar{\mathbf{d}}$ as the $n(n-1)/2$ vector corresponding to the vectorized version of $\bar{d}_{ij}$. Under these settings, and adapting the Pólya–Gamma data augmentation scheme for logistic regression Polson *et al.* (2013), the Metropolis–within–Gibbs routine to draw samples from the posterior distribution, iterates among the following steps.

- **Pólya–Gamma update**. Update the augmented data from the Pólya-Gamma full–conditional

$$(\omega_l \mid -) \sim \text{PG}(m,\, \mathbf{X}_l\boldsymbol{\beta} - \bar{\mathbf{d}}_l),$$

  for every $l = 1, \ldots, n(n-1)/2$.

- **Coefficients update**. Sample $\boldsymbol{\beta}$ from the full–conditional

$$(\boldsymbol{\beta} \mid -) \sim \text{N}_4[(\mathbf{X}^\intercal\boldsymbol{\Omega}\mathbf{X} + \boldsymbol{\Lambda}_0^{-1})^{-1}(\mathbf{X}^\intercal\bar{\mathbf{S}} + \mathbf{X}^\intercal\boldsymbol{\Omega}\bar{\boldsymbol{d}}), (\mathbf{X}^\intercal\boldsymbol{\Omega}\mathbf{X} + \boldsymbol{\Lambda}_0^{-1})^{-1}],$$

  with $\boldsymbol{\Omega}$ denoting the $n(n-1)/2 \times n(n-1)/2$ diagonal matrix with elements $(\omega_1, \ldots, \omega_{n(n-1)/2})$.

After the above steps, the algorithm proceeds separately for the $\bar{x}$, $\bar{y}$ and $\bar{z}$ latent dimensions. The detailed steps to update the latent $\bar{x}$–coordinates are reported below.

The steps associated with the latent $\bar{y}$–coordinates and $\bar{z}$–coordinates, proceed in a similar manner.

- **Cluster assignment update**. Let $c_{\mathbf{x}_i}$ denote the clustering indicator for the brain region i with respect to the latent coordinate $\bar{x}$. Assign each $i = 1, \ldots, n$ to one of the mixture components by sampling from the full–conditional categorical variable with probabilities

$$\text{pr}(c_{\mathbf{x}_i} = h \mid -) = \frac{\nu_{\mathbf{x}_h}\phi(\bar{x}_i; \mu_{\mathbf{x}_h}, \sigma^2_{\mathbf{x}_h})}{\sum_{q=1}^{H} \nu_{\mathbf{x}_q}\phi(\bar{x}_i; \mu_{\mathbf{x}_q}, \sigma^2_{\mathbf{x}_q})},$$

  for each $h = 1, \ldots, H$ and $i = 1, \ldots, n$, where $\phi(\bar{x}_i; \mu_{\mathbf{x}_h}, \sigma^2_{\mathbf{x}_h})$ indicates the density of the Gaussian with parameters $(\mu_{\mathbf{x}_h}, \sigma^2_{\mathbf{x}_h})$ evaluated at $\bar{x}_i$.

- **Latent coordinates update**. Let $\text{pr}(\mathbf{S} \mid \boldsymbol{\beta}, \bar{\mathbf{d}}, \mathbf{X})$ denote the joint distribution of the observed edges under model (1.3)–(1.4). Update the each latent coordinate $\bar{x}_i$, for $i = 1, \ldots, n$, from a random walk Metropolis–Hastings step with Gaussian proposal and full–conditional distribution $(\bar{x}_i \mid c_{\mathbf{x}_i} = h, -)$ proportional to $\phi(\bar{x}_i; \mu_{\mathbf{x}_h}, \sigma^2_{\mathbf{x}_h})\text{pr}(\mathbf{S} \mid \boldsymbol{\beta}, \bar{\mathbf{d}}, \mathbf{X})$.

- **Kernel parameters update**. Let

$$n_{\mathbf{x}_h} = \sum_{i=1}^{n} \mathbf{I}[c_{\mathbf{x}_i} = h],\ \bar{m}_{\mathbf{x}_h} = n^{-1}_{\mathbf{x}_h} \sum_{i=1}^{n} \bar{x}_i \mathbf{I}[c_{\mathbf{x}_i} = h],\ s^2_{\mathbf{x}_h} = \sum_{i=1}^{n} (\bar{x}_i - \bar{m}_{\mathbf{x}_h})^2 \mathbf{I}[c_{\mathbf{x}_i} = h]$$

  denote the cluster size, the intra–cluster mean and the intra–cluster sum of squares deviations from the mean. Sample $(\mu_{\mathbf{x}_h}, \sigma^2_{\mathbf{x}_h})$ from the Normal–Inverse Gamma

$$(\sigma^{-2}_{\mathbf{x}_h} \mid -) \sim \text{Gamma}(\eta_{\mathbf{x}_h}/2, \eta_{\mathbf{x}_h}\xi_{\mathbf{x}_h}/2)$$
$$(\mu_{\mathbf{x}_h} \mid \sigma^2_{\mathbf{x}_h}, -) \sim \text{N}\left(\frac{\kappa_0\mu_0 + n_{\mathbf{x}_h}\bar{m}_{\mathbf{x}_h}}{\kappa_0 + n_{\mathbf{x}_h}}, \frac{\sigma^2_{\mathbf{x}_h}}{\kappa_0 + n_{\mathbf{x}_h}}\right)$$

  with $\eta_{\mathbf{x}_h} = \eta_0 + n_{\mathbf{x}_h}$ and $\xi_{\mathbf{x}_h} = [\eta_0\xi_0 + s^2_{\mathbf{x}_h} + \kappa_0 n_{\mathbf{x}_h}(\bar{m}_{\mathbf{x}_h} - \mu_0)^2/(\kappa_0 + n_{\mathbf{x}_h})]/\eta_{\mathbf{x}_h}$ for each $h = 1, \ldots, H$.

- **Mixing probabilities update**. Sample the mixing probabilities of the prior for the $\bar{x}$–coordinate from the full conditional

$$(\boldsymbol{\nu}_{\mathbf{x}} \mid -) \sim \text{Dirichlet}\left(\frac{1}{H} + n_{\mathbf{x}_1}, \ldots, \frac{1}{H} + n_{\mathbf{x}_H}\right)$$

.

TABLE A.2: Summaries of the posterior distribution for the parameter β in the simulation studies.

	Mean	Median	Std. Dev.	Cred. Int.$_{95\%}$
Scenario 1	1.99	1.99	0.03	(1.92, 2.05)
Scenario 2	1.98	1.98	0.04	(1.90, 2.05)
Scenario 3	2.05	2.04	0.05	(1.96, 2.14)

A.1.2 Simulation study

Table A.1 provides additional details on the simulation study conducted in Section 1.4.3. In particular, Table A.1 illustrates the posterior distribution for the number of active components $\bar{H}_{\mathbf{x}}$, $\bar{H}_{\mathbf{y}}$ and $\bar{H}_{\mathbf{z}}$ characterizing the clustering structure induced by the latent coordinates $\bar{x}$, $\bar{y}$ and $\bar{z}$, respectively. These posterior distributions can be easily obtained by computing, for each step of the MCMC, the total number of non–empty mixture components having at least one region assigned in steps **[3**–x**]**, **[3**–y**]** and **[3**–z**]**, respectively. Table A.1 provides the relative frequency tables obtained from the posterior samples of these quantities and confirms the ability of our model to learn the correct number of components in each simulation scenario. This result is also confirmed in Figure 1.3. Lastly, Table A.2 shows how the posterior distribution of the coefficient β correctly concentrates around the truth.

TABLE A.1: In each simulation scenario, posterior distribution for the number of active components $\bar{H}_{\mathbf{x}}$, $\bar{H}_{\mathbf{y}}$ and $\bar{H}_{\mathbf{z}}$ characterizing the clustering structure induced by the $\bar{x}$, $\bar{y}$ and $\bar{z}$ latent coordinates, respectively.

		1	2	3	4	5
Scenario 1	$\bar{H}_{\mathbf{x}}$	0.00	0.88	0.11	0.01	0.00
	$\bar{H}_{\mathbf{y}}$	0.67	0.20	0.08	0.02	0.03
	$\bar{H}_{\mathbf{z}}$	0.63	0.26	0.10	0.01	0.00
Scenario 2	$\bar{H}_{\mathbf{x}}$	0.00	0.76	0.22	0.02	0.00
	$\bar{H}_{\mathbf{y}}$	0.00	0.87	0.12	0.01	0.00
	$\bar{H}_{\mathbf{z}}$	0.68	0.18	0.11	0.02	0.01
Scenario 3	$\bar{H}_{\mathbf{x}}$	0.00	0.61	0.32	0.05	0.02
	$\bar{H}_{\mathbf{y}}$	0.01	0.58	0.35	0.06	0.00
	$\bar{H}_{\mathbf{z}}$	0.00	0.55	0.33	0.11	0.01

A.1.3 Additional details on the application

Consistent with Figure 1.6 and recalling the analyses on the number of clusters discussed in the simulation study, Table A.3 provides evidence of two clusters for the latent $\bar{x}$–coordinate and three for the $\bar{y}$–coordinate and the $\bar{z}$–coordinate.

TABLE A.3: Posterior distribution for the number of active components $\bar{H}_{\mathbf{x}}$, $\bar{H}_{\mathbf{y}}$ and $\bar{H}_{\mathbf{z}}$ characterizing the clustering structure induced by the $\bar{x}$, $\bar{y}$ and $\bar{z}$ latent coordinates, respectively. Results are shown for a maximum of six non–empty clusters, since higher values were associated with 0 relative frequencies.

	1	2	3	4	5	6
$\bar{H}_x$	0.005	0.785	0.200	0.006	0.004	0.000
$\bar{H}_y$	0.000	0.183	0.655	0.157	0.005	0.000
$\bar{H}_z$	0.003	0.085	0.642	0.253	0.016	0.001

A.2 Latent factor model

A.2.1 Computational Details

Let $\boldsymbol{\Psi}$ denote the diagonal matrix with elements (ψ_x, ψ_y, ψ_z), and let $\boldsymbol{\Upsilon}$ the square matrix with elements $[\boldsymbol{\Upsilon}]_{ij} = \upsilon_{ij} = \beta_0 + \beta_1 \texttt{hem}_{ij} + \beta_2 \texttt{cortex}_{ij} + \beta_3 d_{ij}$. The conditionally conjugacy allows to express each optimal factor in the same exponential family form of its full conditional distribution, with natural parameters replaced with variational expectations. Therefore, in order to develop a CAVI algorithm, it is sufficient to express the analytical form of the natural parameters of the exponential family and illustrate each variational expectation.

- **Pólya–Gamma augmented variables**. The optimal distribution for ω_{ij} is Pólya–Gamma with natural parameter

$$\mathbb{E}[\boldsymbol{\eta}_{\omega_{ij}}] = -0.5 \left[\mathbb{E}(\mathbf{w}_i^\intercal \boldsymbol{\Psi} \mathbf{w}_j + \upsilon_{ij})^2\right]^{\frac{1}{2}}. \tag{A.1}$$

  Algebraic manipulations and the MF factorisation in Equation (1.13) shows that variational expectations of the natural parameter is equal to

$$\begin{aligned} -0.5 \Bigg[\mathbf{1}^\intercal \Big\{ &\mathbb{E}\left[\boldsymbol{\Psi}\boldsymbol{\Psi}^\intercal\right] \otimes \mathbb{E}\left[\mathbf{w}_i \mathbf{w}_i^\intercal\right] \otimes \mathbb{E}\left[\mathbf{w}_j \mathbf{w}_j^\intercal\right] + \mathbb{E}\left[(\upsilon_{ij})^2\right] + \\ &2\Big[\mathbb{E}[\mathbf{w}_i]\mathbb{E}[\boldsymbol{\Psi}]\mathbb{E}[\mathbf{w}_j] + \mathbb{E}\left[\upsilon_{ij}\right]\Big]\mathbf{1}\Big\}\Bigg]^{\frac{1}{2}} \end{aligned} \tag{A.2}$$

with expectations taken with respect to the other optimal factors in Equation (1.14), which correspond to multivariate Gaussian distributions with natural parameters that will described shortly.

- **Latent Factors**. Let $\mathbf{W}_{[-i]}$ denote the matrix $\mathbf{W}$ with the i-th row removed, let $\mathbf{\Omega}_{[-i]}$ denote the $n \times n$ matrix with element $[\mathbf{\Omega}]_{ij} = \omega_{ij}$ and similarly $\mathbf{\Omega}_{[-i]}$. The optimal distribution for $\mathbf{w}_i$ is Gaussian with natural parameters

$$\mathbb{E}\left[\boldsymbol{\eta}^{(1)}_{\mathbf{w}_i}\right] = \mathbb{E}\left[\mathbf{W}^{\intercal}_{[-i]}\mathbf{\Omega}_{[-i]}\mathbf{W}_{[-i]} + I\right], \quad \mathbb{E}\left[\boldsymbol{\eta}^{(2)}_{\mathbf{w}_i}\right] = \mathbb{E}\left[\mathbf{W}^{\intercal}_{[-i]}(\mathbf{A}_i - 0.5m - \mathbf{\Omega}_i\mathbf{\Upsilon}_i)\right], \tag{A.3}$$

which can be further decomposed as

$$\mathbb{E}[\boldsymbol{\eta}^{(1)}_{\mathbf{w}_i}] = \mathbb{E}[\mathbf{W}_{[-i]}\mathbf{W}^{\intercal}_{[-i]}]\mathbb{E}[\mathbf{\Omega}_{[-i]}]+I), \quad \mathbb{E}\left[\boldsymbol{\eta}^{(2)}_{\mathbf{w}_i}\right] = \mathbb{E}[\mathbf{W}^{\intercal}_{[-i]}](\mathbf{A}_i - 0.5m - \mathbb{E}[\mathbf{\Omega}_i]\mathbb{E}[\mathbf{\Upsilon}_i]). \tag{A.4}$$

It is worth recalling that if $X \sim \text{PG}(m.p)$, than $\mathbb{E}[X] = \frac{m}{2p}\tanh(p/2)$.

- **Loading matrix**. Let $\widetilde{\mathbf{W}}$ define the $n(n-1)/2 \times 3$ matrix

$$\Big[[\tilde{x}_2\tilde{x}_1, \tilde{x}_3\tilde{x}_1, \ldots, \tilde{x}_n\tilde{x}_{n-1}]^{\intercal}, [\tilde{y}_2\tilde{y}_1, \tilde{y}_3\tilde{y}_1, \ldots, \tilde{y}_n\tilde{y}_{n-1}]^{\intercal}, [\tilde{z}_2\tilde{z}_1, \tilde{z}_3\tilde{z}_1, \ldots, \tilde{z}_n\tilde{z}_{n-1}]^{\intercal}\Big],$$

and let $\widetilde{\mathbf{\Omega}}$ the $n(n-1)/2 \times n(n-1)/2$ diagonal matrix with elements $\boldsymbol{\omega} = (\omega_{12}, \omega_{13}, \ldots, \omega_{n-1n})$. The optimal distribution for $(\psi_x, \psi_y, \psi_z)^{\intercal}$ is Gaussian with natural parameters

$$\mathbb{E}\left[\boldsymbol{\eta}^{(1)}_{\boldsymbol{\psi}}\right] = \mathbb{E}\left[\widetilde{\mathbf{W}}^{\intercal}\widetilde{\mathbf{\Omega}}\widetilde{\mathbf{W}} + \gamma_{\psi_0}I\right], \quad \mathbb{E}\left[\boldsymbol{\eta}^{(2)}_{\boldsymbol{\psi}}\right] = \mathbb{E}\left[\widetilde{\mathbf{W}}^{\intercal}\Big(\bar{\mathbf{S}} - \mathbf{\Omega}\mathcal{L}(\mathbf{\Upsilon})\Big)\right]. \tag{A.5}$$

The first canonical parameter involves quite demanding expectations involving Gaussian cross products. Indeed, denote as $\mathbf{W}^{\dagger}$ the $3 \times 3 \times n(n-1)/2$ array obtained stacking matrices

$$\left[\Big[\mathbb{E}[\mathbf{w}_1\mathbf{w}_1^{\intercal}] \otimes \mathbb{E}[\mathbf{w}_2\mathbf{w}_2^{\intercal}]\Big], \Big[\mathbb{E}[\mathbf{w}_1\mathbf{w}_1^{\intercal}] \otimes \mathbb{E}[\mathbf{w}_3\mathbf{w}_3^{\intercal}]\Big], \ldots, \Big[\mathbb{E}[\mathbf{w}_{n-1}\mathbf{w}_{n-1}^{\intercal}] \otimes \mathbb{E}[\mathbf{w}_n\mathbf{w}_n^{\intercal}]\Big]\right]$$

in slice order. Then it can shown that

$$\mathbb{E}\left[\widetilde{\mathbf{W}}^{\intercal}\widetilde{\mathbf{\Omega}}\widetilde{\mathbf{W}}\right] = \mathbf{W}^{\dagger} \times_3 \mathbb{E}[\boldsymbol{\omega}] \tag{A.6}$$

- **Coefficients update** Similarly to the update of the loading matrix, the problem can be recasted into a conditionally Gaussian regression leading to an optimal

distribution for $\boldsymbol{\beta}$ which is Gaussian with

$$\mathbb{E}\left[\boldsymbol{\eta}_{\boldsymbol{\beta}}^{(1)}\right] = \mathbb{E}\left[\mathbf{X}^\intercal \boldsymbol{\Omega} \mathbf{X} + \boldsymbol{\Sigma}_0\right], \quad \mathbb{E}\left[\boldsymbol{\eta}_{\boldsymbol{\beta}}^{(2)}\right] = \mathbb{E}\left[\mathbf{X}^\intercal (\bar{\mathbf{S}} - \boldsymbol{\Omega}\mathcal{L}(\mathbf{W}\boldsymbol{\Psi}\mathbf{W}^\intercal)\right], \tag{A.7}$$

which is particularly simple since all the terms are linear in the variational expectations.

FIGURE A.1: Graphical representation of the difference between the observed edge frequencies $\sum_{k=1}^{21} a_{ij}^{(k)}/21$ and the posterior mean $\mathrm{E}(\pi_{ij} \mid \mathbf{A})$ of the corresponding edge probabilities under model (1.3)–(1.4) with (right matrix) and without (left matrix) the latent space effects in (1.4). Brain regions are grouped by combinations of lobe and hemisphere membership. Colors range from dark red to dark green as the differences go from -1 to $+1$.

Appendix B

Appendix for Chapter 2

B.1 Gibbs sampler for mills

The Gibbs sampler algorithm for the MILLS approach described in Section 2.3 iterates across the following steps.

- **Cluster Allocation**. For $i = 1, \ldots, n$, update each z_i sampling from its full conditional categorical distributions with

$$\text{pr}[z_i = h \mid -] \propto \nu_h \tilde{\mathbf{p}}(y_i \mid \boldsymbol{\vartheta}^h) \tag{B.1}$$

- **Mixture weights**. Update $\boldsymbol{\mu}$ from its full conditional Dirichlet distribution.

$$(\boldsymbol{\nu} \mid -) \sim \text{Dirichlet}\left(\frac{1}{H} + n_1, \frac{1}{H} + n_H\right), \tag{B.2}$$

  with $n_h = \sum_{i=1}^{n} \mathbf{I}[z_i = h]$

- **Kernel Update**. This step is performed separately for each bivariate distribution leveraging a nested Pòlya-Gamma data augmentation for multinomial likelihood and Gaussian prior for the log-odds (Polson *et al.*, 2013, Supplementary Material). Indeed, it is more convenient to sample directly the log-odds $\bar{\boldsymbol{\vartheta}}^h_{E_2} = \mathbf{X}^{-1}\boldsymbol{\vartheta}^h_{E_2}$ and then transform the sampled values after each iteration into the canonical parameters.

  Focusing on a specific marginal table $\bar{\boldsymbol{y}} = \boldsymbol{y}^h_{E_2} = (\bar{y}_1, \ldots, \bar{y}_{|E_2|})$, conditional on cross-classifying individuals i such that $z_i = h$, we focus on updating the associated $\bar{\boldsymbol{\vartheta}} = \bar{\boldsymbol{\vartheta}}^h_{E_2}$ with elements $\bar{\boldsymbol{\vartheta}} = \{0, \vartheta_2, \ldots, \vartheta_{|E_2|}\}$.

  Specifically, for $k = 2, \ldots, |E_2|$,

1. Define $\bar{\boldsymbol{y}}_k = \vartheta_k - c_k$, with $c_k = \log(1 + \sum_{j \neq k} \exp(\vartheta_j))$
2. Sample a nested augmented Pòlya-Gamma value from $(\omega \mid -) \sim \text{PG}(n_h, \psi)$
3. Sample the log odds from $(\vartheta_k \mid -) \sim \text{N}\left(\frac{\bar{y}_k - n_h/2 - \omega c_k}{\omega + 1/\sigma^2}, \frac{1}{\omega + 1/\sigma^2}\right)$

B.2 Additional data information

As outlined in Section 2.2, we focus on two different instruments measuring the empathic profile and the psychopathology of attempt suicidal patients. Table B.1 illustrates the questions of the IRI-28 tool, which asks to the subjects "The following statements inquire about your thoughts and feelings in a variety of situations. For each item, indicate how well it describes you". Subjects respond to the questions with letters (A,B,C,D,E), ranging from "This item does not describe me very well" (A) to "This item describes me very well" (E). See `this link` for an illustration. Some items are associated with positive behaviour, while some others with negative ones. However, the methodologies of Chapter 2 focus on unordered categorical data, and the ordering of the answers is not a concern. The subscales described in Section 2.2 are reported in brackets.

Table B.2 report the subscale of the SCL-90 questionnaire, focusing on the psycho pathological profiles of interest measured via the questions reported in Table B.2. The subjects are asked "'How much were you bothered by", and respond with numbers in $[0 - 4]$, respectively corresponding to "Not at all", "A little bit", "Moderately", "Quite a bit", "Extremely". See `this link` for an illustration of the entire SCL-90 questionnaire. The subscales described in Section 2.2 are reported in brackets.

Finally, for illustrative purposed, Table B.3 reports the descriptive statistics of the data.

TABLE B.1: IRI-28 questionnaire. Subjects answer with their level of agreement with letters ranging from A ("Does not describe me") to E ("Describes me very well").

ID		SUB
1.	I daydream and fantasize, with some regularity, about things that might happen to me.	(FS)
2.	I often have tender, concerned feelings for people less fortunate than me.	(EC)
3.	I sometimes find it difficult to see things from the "other guy's" point of view.	(PT)
4.	Sometimes I don't feel very sorry for other people when they are having problems.	(EC)
5.	I really get involved with the feelings of the characters in a novel.	(FS)
7.	I am usually objective when I watch a movie or play, and I don't often get completely caught up in it.	(FS)
8.	I try to look at everybody's side of a disagreement before I make a decision.	(PT)
9.	When I see someone being taken advantage of, I feel kind of protective towards them.	(EC)
10.	I sometimes feel helpless when I am in the middle of a very emotional situation.	(PD)
11.	I sometimes try to understand my friends better by imagining how things look from their perspective.	(PT)
12.	Becoming extremely involved in a good book or movie is somewhat rare for me.	(FS)
13.	When I see someone get hurt, I tend to remain calm.	(PD)
14.	Other people's misfortunes do not usually disturb me a great deal.	(EC)
15.	If I'm sure I'm right about something, I don't waste much time listening to other people's arguments.	(PT)
16.	After seeing a play or movie, I have felt as though I were one of the characters.	(FS)
17.	Being in a tense emotional situation scares me.	(PD)
18.	When I see someone being treated unfairly, I sometimes don't feel very much pity for them.	(EC)
19.	I am usually pretty effective in dealing with emergencies.	(PD)
21.	I believe that there are two sides to every question and try to look at them both.	(PT)
23.	When I watch a good movie, I can very easily put myself in the place of a leading character.	(FS)
25.	When I'm upset at someone, I usually try to "put myself in his shoes" for a while.	(PT)
26.	When I am reading an interesting story or novel, I imagine how I would feel if the events in the story were happening to me.	(FS)
28.	Before criticizing somebody, I try to imagine how I would feel if I were in their place.	(PT)

TABLE B.2: SCL-90 subscales. Subjects answer with their level of agreement with numbers ranging from 0 ("Not at all") to 4 ("Extremely").

ID		SUB	ID		SUB
2.	Nervousness or shakiness inside	(ANX)	39.	Heart pounding or racing	(ANX)
3.	Unwanted thoughts, words, or ideas that won't leave your mind	(OC)	45.	Having to check and double-check what you do	(OC)
5.	Loss of sexual interest or pleasure	(DEP)	46.	Difficulty making decisions frighten you	(OC)
9.	Trouble remembering things	(OC)	54.	Feeling hopeless about the future	(DEP)
10.	Worried about sloppiness or carelessness	(OC)	55.	Trouble concentrating	(OC)
11.	Feeling easily annoyed or irritated	(HOS)	57.	Feeling tense or keyed up	(ANX)
14.	Feeling low in energy or slowed down	(DEP)	63.	Having urges to beat, injure, or harm someone	(HOS)
15.	Thoughts of ending your life	(DEP)	65.	Having to repeat the same actions such as	(OC)
17.	Trembling	(ANX)	·	touching, counting, washing	·
20.	Crying easily	(DEP)	67.	Having urges to break or smash things	(HOS)
22.	Feeling of being trapped or caught	(DEP)	71.	Feeling everything is an effort	(DEP)
23.	Suddenly scared for no reason	(ANX)	72.	Spells of terror or panic	(ANX)
24.	Temper outbursts that you could not control	(HOS)	74.	Getting into frequent arguments	(HOS)
26.	Blaming yourself for things	(DEP)	78.	Feeling so restless you couldn't sit still	(ANX)
28.	Feeling blocked in getting things done	(OC)	79.	Feelings of worthlessness	(DEP)
29.	Feeling lonely	(DEP)	80.	Feeling that familiar things are strange or unreal	(ANX)
30.	Feeling blue	(HOS)	81.	Shouting or throwing things	(HOS)
31.	Worrying too much about things	(DEP)	86.	Feeling pushed to get things done	(ANX)
32.	Feeling no interest in things	(DEP)			
33.	Feeling fearful	(ANX)			
38.	Having to do things very slowly to insure correctness	(OC)			

TABLE B.3: Univariate descriptive statistics for the observed data.

ITEM	A	B	C	D	E
IRI-28, IT 1	10	10	22	9	5
IRI-28, IT 2	4	7	9	17	19
IRI-28, IT 3	9	19	12	14	2
IRI-28, IT 4	19	10	13	8	6
IRI-28, IT 5	8	12	12	14	10
IRI-28, IT 7	10	11	18	12	5
IRI-28, IT 9	3	6	7	14	26
IRI-28, IT 10	4	9	14	12	17
IRI-28, IT 11	5	8	17	12	14
IRI-28, IT 12	19	13	9	7	8
IRI-28, IT 13	13	12	14	9	8
IRI-28, IT 14	21	15	8	6	6
IRI-28, IT 15	10	9	13	14	10
IRI-28, IT 16	15	8	14	9	10
IRI-28, IT 17	11	10	12	11	12
IRI-28, IT 18	27	7	7	7	8
IRI-28, IT 19	11	12	7	10	16
IRI-28, IT 21	5	9	14	16	12
IRI-28, IT 23	8	12	15	4	17
IRI-28, IT 25	12	13	15	10	6
IRI-28, IT 26	12	11	8	14	11
IRI-28, IT 28	4	11	12	15	14

ITEM	0	1	2	3	4
SCL-90, IT 2	4	8	17	15	12
SCL-90, IT 3	10	8	11	13	14
SCL-90, IT 5	24	5	6	5	16
SCL-90, IT 9	13	22	8	6	7
SCL-90, IT 10	15	18	13	8	2
SCL-90, IT 11	12	22	8	8	6
SCL-90, IT 14	9	10	12	15	10
SCL-90, IT 15	17	14	10	3	12
SCL-90, IT 17	28	11	9	4	4
SCL-90, IT 20	21	14	6	11	4
SCL-90, IT 22	27	9	6	5	9
SCL-90, IT 23	29	10	10	4	3
SCL-90, IT 26	15	14	13	8	6
SCL-90, IT 28	10	20	11	6	9
SCL-90, IT 29	9	10	7	12	18
SCL-90, IT 30	3	9	14	14	16
SCL-90, IT 31	13	12	9	13	9
SCL-90, IT 32	13	11	5	14	13
SCL-90, IT 33	22	12	7	9	6
SCL-90, IT 38	13	19	9	8	7
SCL-90, IT 39	27	11	7	6	5
SCL-90, IT 45	23	14	7	9	3
SCL-90, IT 46	15	19	8	5	9
SCL-90, IT 51	24	13	8	5	6
SCL-90, IT 55	12	16	11	10	7
SCL-90, IT 63	40	6	6	2	2
SCL-90, IT 65	36	11	6	2	1
SCL-90, IT 67	41	2	7	4	2
SCL-90, IT 71	16	12	8	12	8
SCL-90, IT 72	30	6	9	7	4
SCL-90, IT 74	33	9	9	2	3
SCL-90, IT 78	29	8	7	6	6
SCL-90, IT 79	13	15	5	13	10
SCL-90, IT 80	31	7	8	5	5
SCL-90, IT 86	15	13	17	6	5

Bibliography

Agresti, A. (2003) *Categorical data analysis*. Volume 482. John Wiley & Sons.

Airoldi, E. M., Blei, D. M., Erosheva, E. A. and Fienberg, S. E. (2014) *Handbook of mixed membership models and their applications*. CRC press.

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.

Albert, J. H. and Chib, S. (1993) Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association* **88**(422), 669–679.

Aliverti, E. and Durante, D. (2019) Spatial modeling of brain connectivity data via latent distance models with nodes clustering. *Statistical Analysis and Data Mining: The ASA Data Science Journal* **12**(3), 185–196.

Aliverti, E., Lum, K., Johndrow, J. E. and Dunson, D. B. (2018) Removing the influence of a group variable in high-dimensional predictive modelling. *arXiv preprint arXiv:1810.08255* .

Aliverti, E., Tilson, J., Filer, D., Babcock, B., Colaneri, A., Ocasio, J., Gershon, T. R., Wilhelmsen, K. C. and Dunson, D. B. (2019) Batch correction of high-dimensional data. *arXiv preprint arXiv:1911.06708* .

Andersen, E. B. (1982) Latent structure analysis: a survey. *Scandinavian Journal of Statistics* pp. 1–12.

Angwin, J., Larson, J., Mattu, S. and Kirchner, L. (2016) Machine bias: there's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica* https://www.propublica.org/article/machine-bias-risk-assessments-in- criminal-sentencing.

Ben-Ya'acov, Y. and Amir, M. (2004) Posttraumatic symptoms and suicide risk. *Personality and Individual Differences* **36**(6), 1257–1264.

Bergsma, W. P. and Rudas, T. (2002) Marginal models for categorical data. *The Annals of Statistics* **30**(1), 140–159.

Bhattacharya, A. and Dunson, D. B. (2012) Simplex factor models for multivariate unordered categorical data. *Journal of the American Statistical Association* **107**(497), 362–377.

Bishop, C. M. (2006) *Pattern recognition and machine learning*. Springer, New York.

Blei, D. M., Kucukelbir, A. and McAuliffe, J. D. (2017) Variational inference: a review for statisticians. *Journal of the American Statistical Association* **112**(518), 859–877.

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* **13**, 336–349.

Carota, C., Filippone, M., Leombruni, R., Polettini, S. *et al.* (2015) Bayesian nonparametric disclosure risk estimation via mixed effects log-linear models. *The Annals of Applied Statistics* **9**(1), 525–546.

Chang, I. H. and Mukerjee, R. (2006) Probability matching property of adjusted likelihoods. *Statistics & Probability Letters* **76**(8), 838–842.

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.

Cockayne, J., Oates, C. J., Ipsen, I. C. and Girolami, M. (2018) A bayesian conjugate gradient method. *Bayesian Analysis* **14**(3), 937–1012.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S. and Huq, A. (2017) Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 797–806.

Cox, D. R. and Reid, N. (2004) A note on pseudolikelihood constructed from marginal densities. *Biometrika* **91**(3), 729–737.

Craddock, R., Jbabdi, S., Yan, C., Vogelstein, J., Castellanos, F., Di Martino, A., Kelly, C., Heberlein, K., Colcombe, S. and Milham, M. (2013) Imaging human connectomes at the macroscale. *Nature Methods* **10**, 524–539.

Cusi, A. M., MacQueen, G. M., Spreng, R. N. and McKinnon, M. C. (2011) Altered empathic responding in major depressive disorder: relation to symptom severity, illness burden, and psychosocial outcome. *Psychiatry Research* **188**(2), 231–236.

Davis, M. H. (1980) A muntidimensional approach to individual differences in empathy. *JSAS Catalogue of Selected Documents in Psychology* **10**.

Davis, M. H. (1983) Measuring individual differences in empathy: evidence for a multidimensional approach. *Journal of personality and social psychology* **44**(1), 113.

De Leo, D., Burgis, S., Bertolote, J. M., Kerkhof, A. and Bille-Brahe, U. (2004) Definitions of suicidal behaviour. *Suicidal behaviour: Theories and research findings* pp. 17–39.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* **39**(1), 1–22.

Derogatis, L., Lipman, R. and Covi, L. (1973) Scl-90: an outpatient psychiatric rating scale–preliminary report. *Psychopharmacology bulletin* **9**(1), 13.

Desikan, R., Ségonne, F., Fischl, B., Quinn, B., Dickerson, B., Blacker, D., Buckner, R., Dale, A., Maguire, R. and Hyman, B. (2006) An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage* **31**, 968–980.

Dobra, A. and Massam, H. (2010) The mode oriented stochastic search (moss) algorithm for log-linear models with conjugate priors. *Statistical methodology* **7**(3), 240–253.

Dunson, D. B. (2018) Statistics in the big data era: failures of the machine. *Statistics & Probability Letters* **136**, 4–9.

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., Canale, A. and Rigon, T. (2019) A nested expectation–maximization algorithm for latent class models with covariates. *Statistics & Probability Letters* **146**, 97–103.

Durante, D., Dunson, D. B. and Vogelstein, J. T. (2017) Nonparametric bayes modeling of populations of networks. *Journal of the American Statistical Association* **112**(520), 1516–1530.

Durante, D. and Rigon, T. (2019) Conditionally conjugate mean-field variational bayes for logistic models. *Statistical Science (in press)* .

Dwork, C., Hardt, M., Pitassi, T., Reingold, O. and Zemel, R. (2012) Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226.

Efron, B. (1993) Bayes and likelihood calculations from confidence intervals. *Biometrika* **80**(1), 3–26.

Erosheva, E. A. (2005) Comparing latent structures of the grade of membership, rasch, and latent class models. *Psychometrika* **70**(4), 619–628.

Fienberg, S. E. and Rinaldo, A. (2007) Three centuries of categorical data analysis: log-linear models and maximum likelihood estimation. *Journal of Statistical Planning and Inference* **137**(11), 3430–3445.

Fontenelle, L. F., Soares, I. D., Miele, F., Borges, M. C., Prazeres, A. M., Rangé, B. P. and Moll, J. (2009) Empathy and symptoms dimensions of patients with obsessive–compulsive disorder. *Journal of Psychiatric Research* **43**(4), 455–463.

Friedler, S. A., Scheidegger, C. and Venkatasubramanian, S. (2016) On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236* .

Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P. and Roth, D. (2019) A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 329–338.

Fruhwirth-Schnatter, S., Celeux, G. and Robert, C. P. (2019) *Handbook of mixture analysis*. Chapman and Hall/CRC.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2014) *Bayesian Data Analysis*. CRC press.

Gelman, A., Roberts, G. O. and Gilks, W. R. (1996) Efficient metropolis jumping rules. *Bayesian Statistics* **5**(599-608), 42.

Geweke, J. and Zhou, G. (1996) Measuring the pricing error of the arbitrage pricing theory. *The review of financial studies* **9**(2), 557–587.

Geyer, C. J. and Thompson, E. A. (1992) Constrained monte carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 657–699.

Gollini, I. and Murphy, T. B. (2016) Joint modeling of multiple network views. *Journal of Computational and Graphical Statistics* **25**(1), 246–265.

Goodfellow, B., Kõlves, K. and De Leo, D. (2019) Contemporary definitions of suicidal behavior: a systematic literature review. *Suicide and Life-Threatening Behavior* **49**(2), 488–504.

Gotelli, N. J. and Ellison, A. M. (2004) *Primer of ecological statistics*. Sinauer Associates Publishers.

Greco, L., Racugno, W. and Ventura, L. (2008) Robust likelihood functions in bayesian inference. *Journal of Statistical Planning and Inference* **138**(5), 1258–1270.

Guttman, H. and Laporte, L. (2002) Alexithymia, empathy, and psychological symptoms in a family context. *Comprehensive psychiatry* **43**(6), 448–455.

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. S., Raftery, A. E. and Tantrum, J. M. (2007) Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **170**(2), 301–354.

Hawton, K. and Fagg, J. (1988) Suicide, and other causes of death, following attempted suicide. *The British Journal of Psychiatry* **152**(3), 359–366.

HCP (2019) Human connectome project. Available at: `http://umcd.humanconnectomeproject.org/umcd/default/get_study_data/Hagmann_PLoSBiol_2008`, Accessed 08-August-2019.

Hoff, P. D. (2008) Modeling homophily and stochastic equivalence in symmetric relational data. In *Advances in Neural Information Processing Systems*, pp. 657–664.

Hoff, P. D. (2019) Additive and multiplicative effects network models. *Statistical Sciente (to appear)* .

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.

Hoffman, M., Kahn, L. B. and Li, D. (2017) Discretion in hiring. *The Quarterly Journal of Economics* **133**(2), 765–800.

Hoffman, M. D., Blei, D. M., Wang, C. and Paisley, J. (2013) Stochastic variational inference. *The Journal of Machine Learning Research* **14**(1), 1303–1347.

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.

Jackson, M. O. (2014) Networks in the understanding of economic behaviors. *Journal of Economic Perspectives* **28**(4), 3–22.

Johndrow, J. E. and Bhattacharya, A. (2018) Optimal gaussian approximations to the posterior for log-linear models with diaconis–ylvisaker priors. *Bayesian Analysis* **13**(1), 201–223.

Johndrow, J. E., Bhattacharya, A. and Dunson, D. B. (2017) Tensor decompositions and sparse log-linear models. *Annals of Statistics* **45**(1), 1–38.

Johndrow, J. E. and Lum, K. (2019) An algorithm for removing sensitive information: application to race- independent recidivism prediction. *The Annals of Applied Statistics* **13**(1), 189–220.

Johnson, W. E., Li, C. and Rabinovic, A. (2007) Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics* **8**(1), 118–127.

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**(1), 2.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S. and Saul, L. K. (1999) An introduction to variational methods for graphical models. *Machine learning* **37**(2), 183–233.

King, B. M., Minium, E. W. and Rosopa, P. J. (2008) *Statistical reasoning in the behavioral sciences*. John Wiley & Sons Hoboken, New Jersey.

Krivitsky, P. N. and Handcock, M. S. (2008) Fitting latent cluster models for networks with latentnet. *Journal of Statistical Software* **24**(5).

Lachal, J., Orri, M., Revah-Levy, A. and Moro, M. (2016) Empathy in adolescent suicidal behaviors: perspectives from the adolescents, their parents and their healthcare professionals. *European Psychiatry* **33**, S328.

Landman, B. A., Huang, A. J., Gifford, A., Vikram, D. S., Lim, I. A. L., Farrell, J. A., Bogovic, J. A., Hua, J., Chen, M., Jarso, S. *et al.* (2011) Multi-parametric neuroimaging reproducibility: a 3-t resource study. *Neuroimage* **54**(4), 2854–2866.

Larson, J., Mattu, S., Kirchner, L. and Angwin, J. (2016) How we analyzed the compas recidivism algorithm. *ProPublica* **9**. https://www.propublica.org/article/how-we-analyzed-the-compas- recidivism-algorithm.

Lauritzen, S. L. (1996) *Graphical models*. Volume 17. Clarendon Press.

Lawley, D. N. (1943) On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburgh Section A: Mathematics* **61**(3), 273–287.

Lazar, N. A. (2003) Bayesian empirical likelihood. *Biometrika* **90**(2), 319–326.

Lazarsfeld, P. F. (1950) The logical and mathematical foundation of latent structure analysis. *Studies in Social Psychology in World War II Vol. IV: Measurement and Prediction* pp. 362–412.

Leek, J. T. and Storey, J. D. (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics* **3**(9), e161.

Letac, G. and Massam, H. (2012) Bayes factors and the geometry of discrete hierarchical loglinear models. *The Annals of Statistics* **40**(2), 861–890.

Lopes, H. F. and West, M. (2004) Bayesian model assessment in factor analysis. *Statistica Sinica* **14**(1), 41–68.

Lum, K. and Isaac, W. (2016) To predict and serve? *Significance* **13**(5), 14–19.

Lupparelli, M., Marchetti, G. M. and Bergsma, W. P. (2009) Parameterization and fitting of discrete bi-directed graph models. *Scandinavian Journal of Statistics* **36**(3), 559–576.

Mardia, K. V., Kent, J. T., Hughes, G. and Taylor, C. C. (2009) Maximum likelihood estimation using composite likelihoods for closed exponential families. *Biometrika* **96**(4), 975–982.

Massam, H., Liu, J. and Dobra, A. (2009) A conjugate prior for discrete hierarchical log-linear models. *The Annals of Statistics* **37**(6A), 3431–3467.

Massam, H. and Wang, N. (2018) Local conditional and marginal approach to parameter estimation in discrete graphical models. *Journal of Multivariate Analysis* **164**, 1–21.

McLachlan, G. and Krishnan, T. (2007) *The* EM *algorithm and extensions*. Volume 382. John Wiley & Sons.

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.

Meng, Z., Wei, D., Wiesel, A. and Hero III, A. O. (2013) Distributed learning of gaussian graphical models via marginal likelihoods. In *Artificial Intelligence and Statistics*, pp. 39–47.

Munoz, C., Smith, M. and Patil, D. (2016) Big data: a report on algorithmic systems, opportunity, and civil rights. Technical report, Executive Office of the President. Available at: `https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf`.

Nardi, Y. and Rinaldo, A. (2012) The log-linear group-lasso estimator and its asymptotic properties. *Bernoulli* **18**(3), 945–974.

Newman, M. (2018) *Networks*. Oxford university press.

Nock, M. K., Borges, G., Bromet, E. J., Alonso, J., Angermeyer, M., Beautrais, A., Bruffaerts, R., Chiu, W. T., De Girolamo, G., Gluzman, S. *et al.* (2008) Cross-national prevalence and risk factors for suicidal ideation, plans and attempts. *The British Journal of Psychiatry* **192**(2), 98–105.

Nowicki, K. and Snijders, T. A. B. (2001) Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association* **96**(455), 1077–1087.

Ormerod, J. T. and Wand, M. P. (2010) Explaining variational approximations. *The American Statistician* **64**(2), 140–153.

Papastamoulis, P. (2016) Label.switching: an R package for dealing with the label switching problem in MCMC outputs. *Journal of Statistical Software* **69**(1).

Pauli, F., Racugno, W. and Ventura, L. (2011) Bayesian composite marginal likelihoods. *Statistica Sinica* **21**(1), 149–164.

Perrone-McGovern, K. M., Oliveira-Silva, P., Simon-Dack, S., Lefdahl-Davis, E., Adams, D., McConnell, J., Howell, D., Hess, R., Davis, A. and Gonçalves, Ó. F. (2014) Effects of empathy and conflict resolution strategies on psychophysiological arousal and satisfaction in romantic relationships. *Applied Psychophysiology and Biofeedback* **39**(1), 19–25.

Peruzzi, M. and Dunson, D. B. (2018) Bayesian modular and multiscale regression. *arXiv preprint arXiv:1809.05935* .

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.

Prunas, A., Sarno, I., Preti, E., Madeddu, F. and Perugini, M. (2012) Psychometric properties of the italian version of the scl-90-r: a study on a large community sample. *European psychiatry* **27**(8), 591–597.

Raftery, A., Madigan, D. and Volinsky, C. T. (1995) Accounting for model uncertainty in survival analysis improves predictive performance. *Bayesian Statistics* (5), 323–349.

Raftery, A. E. (1985) A model for high-order markov chains. *Journal of the Royal Statistical Society: Series B (Methodological)* **47**(3), 528–539.

Ravikumar, P., Wainwright, M. J., Lafferty, J. D. *et al.* (2010) High-dimensional ising model selection using $\ell 1$-regularized logistic regression. *The Annals of Statistics* **38**(3), 1287–1319.

Ribatet, M., Cooley, D. and Davison, A. C. (2012) Bayesian inference from composite likelihoods, with an application to spatial extremes. *Statistica Sinica* pp. 813–845.

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. *et al.* (2013) Migraine: Mri graph reliability analysis and inference for connectomics. In *Global Conference on Signal and Information Processing, IEEE*, pp. 313–316.

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.

Roverato, A., Lupparelli, M. and La Rocca, L. (2013) Log-mean linear models for binary data. *Biometrika* **100**(2), 485–494.

Rubinov, M. and Sporns, O. (2010) Complex network measures of brain connectivity: uses and interpretations. *Neuroimage* **52**(3), 1059–1069.

Rudovsky, D. (2001) Law enforcement by stereotypes and serendipity: racial profiling and stops and searches without cause. *U. Pa. J. Const. L.* **3**, 296.

Russo, M., Durante, D. and Scarpa, B. (2018) Bayesian inference on group differences in multivariate categorical data. *Computational Statistics & Data Analysis* **126**, 136–149.

Russo, M., Singer, B. H. and Dunson, D. B. (2019) Multivariate mixed membership modeling: inferring domain-specific risk profiles. *arXiv preprint arXiv:1901.05191* .

Salter-Townshend, M. and Murphy, T. B. (2013) Variational bayesian inference for the latent position cluster model for network data. *Computational Statistics & Data Analysis* **57**(1), 661–671.

Schreiter, S., Pijnenborg, G. and Aan Het Rot, M. (2013) Empathy in adults with clinical or subclinical depressive symptoms. *Journal of Affective Disorders* **150**(1), 1–16.

Sewell, D. K. and Chen, Y. (2017) Latent space approaches to community detection in dynamic networks. *Bayesian Analysis* **12**(2), 351–377.

Simoiu, C., Corbett-Davies, S. and Goel, S. (2017) The problem of infra-marginality in outcome tests for discrimination. *The Annals of Applied Statistics* **11**(3), 1193–1216.

Simpson, S. L., Hayasaka, S. and Laurienti, P. J. (2011) Exponential random graph modeling for complex brain networks. *PloS one* **6**(5), e20039.

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. (2002) Markov chain monte carlo estimation of exponential random graph models. *Journal of Social Structure* **3**(2), 1–40.

Sporns, O. (2013) Structure and function of complex brain networks. *Dialogues in Clinical Neuroscience* **15**, 247–262.

Stam, C. (2014) Modern network science of neurological disorders. *Nature Reviews Neuroscience* **15**, 683–695.

Stanley, M. L., Moussa, M. N., Paolini, B. M., Lyday, R. G., Burdette, J. H. and Laurienti, P. J. (2013) Defining nodes in complex brain networks. *Frontiers in Computational Neuroscience* **7**.

Stephens, M. (2000) Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **62**(4), 795–809.

Takai, K. (2012) Constrained em algorithm with projection method. *Computational statistics* **27**(4), 701–714.

Team, S. D. (2018) *Rstan: the R interface to stan. R package version 2.17.3.*

Tibshirani, R., Wainwright, M. J. and Hastie, T. (2015) *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC.

Tipping, M. E. and Bishop, C. M. (1999) Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **61**(3), 611–622.

Varin, C., Reid, N. and Firth, D. (2011) An overview of composite likelihood methods. *Statistica Sinica* pp. 5–42.

Wainwright, M. J. and Jordan, M. I. (2008) Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning* **1**(1–2), 1–305.

Wasserman, S. and Faust, K. (1994) *Social network analysis: Methods and applications*. Volume 8. Cambridge university press.

Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J. and Oermann, E. K. (2018) Confounding variables can degrade generalization performance of radiological deep learning models. *arXiv preprint arXiv:1807.00431* .

Zhang, J. and Lester, D. (2008) Psychological tensions found in suicide notes: a test for the strain theory of suicide. *Archives of Suicide Research* **12**(1), 67–73.

Zhou, J., Bhattacharya, A., Herring, A. H. and Dunson, D. B. (2015) Bayesian Factorizations of Big Sparse Tensors. *Journal of the American Statistical Association* **110**(512), 1562–1576.