UNIVERSITA’ COMMERCIALE “LUIGI BOCCONI”
PhD SCHOOL

PhD program in Statistics
Cycle: XXXV
Disciplinary Field: SECS-S/01

From post hoc explanations to Bayesian nonparametric models: unveiling hidden structures

Advisor: Omiros Papaspiliopoulos
Co-advisor: Daniele Durante

Candidate
Valentina Ghidini
ID number: 3112756

Year 2024

Acknowledgements

Without my family, I would not have completed this journey. I want to thank my mum, my dad and my brother for always supporting me, and for keeping things in perspective.

I also want to thank Marco, for the strenuous encouragement during the (few) highs and the (lot of) lows of these four years.

A special thank goes to Filippo: his technical and personal advices have been (and are) invaluable, and I am so glad we got to share the first years of study and research.

I would also like to acknowledge Michael Multerer, who took the time to review and correct this thesis. I am truly grateful for the opportunity to now work in his research group, and I look forward to continuing our collaboration.

Abstract

Over recent years, there has been a remarkable increase in the complexity of models and data structures. On the one side, as models become increasingly complex, there is a growing need to understand and explain the decision-making process of black boxes: this is important to enhance the trust of the users as well as to comply with legal requirements. On the other side, complex data structures, such as complex networks, require flexible models to effectively extract relevant information. The leitmotif of this thesis is the search for hidden structures in models and data. The first part is devoted to the topic of explainability. Namely, we introduce the Xi method, a comprehensive statistical framework to define post hoc explanations that possess theoretical guarantees. The rationale is to propose and evaluate a class of probabilistic sensitivity measures that quantifies the degree of association between covariates and generic model predictions. These explanations are designed to be applicable across different models and data types, regardless of their specific characteristics. The second part of this thesis focuses on Bayesian nonparametric models for community detection in complex networks. First, we define a stochastic block model for multiplex networks. Such a model identifies clusters specific to each layer, as well as a latent partition common to all the layers. A non-trivial computational scheme to perform posterior inference is also introduced. This framework has wide ranging applicability to a plethora of problems, including the analysis of latent structures in brain networks of different subjects. Secondly, we propose a stochastic block model specifically tailored for weighted networks with continuous and multidimensional node attributes. This model has the potential to effectively capture and utilize the information contained in these node features, while also being able to learn the optimal amount of information to incorporate from them. A real, motivating application is showcased, addressing the need to identify a meaningful latent partition within a transportation network.

Contents

Chapter 1

Introduction

Over the past decades, there has been a significant increase in the complexity of both data structures and models. This rise can be attributed to various factors, including advancements in technology, the widespread adoption of digital systems, and the availability of vast amounts of data. Machine learning and statistical models have evolved from simple linear regression and basic decision trees to more sophisticated structures such as deep neural networks or ensemble methods, just to cite two. These complex models possess a larger number of parameters, intricate architectures and learning mechanisms, enabling them to capture non-trivial patterns and relationships within the data. However, the trade-off for their enhanced predictive capabilities is often decreased interpretability [Murdoch et al., 2019].

In parallel, there has been a notable shift from simple, structured datasets to more intricate and heterogeneous ones. Traditional tabular datasets have expanded to encompass diverse formats, such as unstructured text and images. Alongside this expansion, complex data structures like graphs and networks have emerged. Analyzing and extracting insights from these systems within a rigorous mathematical framework requires sophisticated techniques capable of handling their inherent complexities. Bayesian nonparametrics is particularly suitable for this endeavor, since it offers endless potential flexibility in defining models for complex data.

The augmentation of complexity in both models and data structures has created challenges in obtaining, interpreting, and trusting the results produced by these systems. Comprehending the inner workings of complex models, spotting the factors driving their predictions, and quantifying the uncertainty associated with their outputs pose significant hurdles. Similarly, modeling complex data structures requires techniques that can handle high-dimensional, heterogeneous, dependent and interlinked data. The two problems outlined above can be reduced to the study of suitable hidden structures, either in already-trained models or in complex network systems. This present thesis explores these two parallel research directions.

The first part of the thesis is entirely devoted to explainability of black box models. Explainability (or interpretability) refers to the ability to explain the decision-making process and outputs of a machine learning model in a human-understandable manner [Murdoch et al., 2019, Rudin, 2019]. It encompasses the transparency, comprehensibility, and clarity of how a model arrives at its predictions or decisions. Interpretability is arguably essential in the modern era, where machine learning models are employed in every aspect of our daily lives. In fact, explainability helps building trust among users, by promoting accountability and ensuring that the model's behavior aligns with ethical and legal standards, which is fundamental especially in sensitive and critical areas, such as health or financial applications [Rudin, 2019]. It also helps to address ethical concerns related to bias, discrimination, and fairness in machine learning models [Miller, 2019]. By understanding how a model makes decisions, biases and discriminatory factors can be identified and mitigated, ensuring fair treatment and reducing potential harm to individuals or communities. Currently, there is an abundance of techniques available to comprehend the internal decision-making processes of models. Various taxonomies have been proposed to categorize these methods, but we can broadly distinguish between two approaches: transparent (white box) models and post hoc explanations of black boxes [Guidotti et al., 2018]. Transparent models, also known as white box models, are designed to be interpretable and understandable. These models have explicit rules or structures that allow us to directly interpret the decision-making process. Examples of white boxes include standard statistical models such as linear regression, decision trees, and rule-based systems [Hastie et al., 2009]. Recent works have also designed inherently black box models in a transparent way [Chen et al., 2019, 2020], in an effort to bridge the gap between the high predictive power of black boxes and the need for interpretability and understanding. On the other hand, post hoc explanations aim to shed light on any model by providing insights into how it arrives at its predictions. In fact, while black boxes may achieve high predictive performances, understanding their decision-making process is challenging due to their intricate internal mechanisms. Post hoc explanations analyze the model's behavior using techniques such as feature importance [Barber and Candés, 2015, Binder et al., 2016, Fisher et al., 2019], saliency maps [Springenberg et al., 2015, Lundberg and Lee, 2017, Petsiuk et al., 2019, Selvaraju et al., 2020], or surrogate models [Ribeiro et al., 2016], among others.

Chapter 2 contains an overview of the state-of-the-art techniques in the explainability literature, and introduces the Xi method. To the best of our knowledge, the Xi method is the first explainable technique that embeds the explanations in a statistically coherent framework. This allows the derivation of asymptotically consistent estimators for the explanations, as well as uncertainty quantification through standard statistical tools. In general, explanations of a model can be seen as the extraction of hidden knowledge in a post hoc manner: in the case of the Xi method, the explanations are veiled importance measures of the inputs with respect to the prediction task, obtained by using estimated distances between suitable probability distributions. The Xi method is also model agnostic and data agnostic, meaning that it can be used to interpret results of any model and on any data type (e.g. tabular, text or image).

The second part of the thesis centers around Bayesian nonparametric models for community detection in complex networks. Complex networks, often referred to as graphs, are a mathematical representation consisting of a set of nodes or vertices, and edges that describe the relationships or connections between those nodes. These networks are prevalent in various fields, including social networks, biological networks, transportation networks, and communication networks. The objective of this part of the thesis is to tackle the distinct challenges presented by this type of data and develop Bayesian nonparametric models specifically tailored to community detection. Community detection refers to the task of identifying groups or communities within a network [Fortunato and Newman, 2022]. These communities often represent subsets of nodes that exhibit stronger connections amongst themselves compared to nodes outside the community, or, alternatively, that experience similar connectivity patterns. By uncovering these latent partitions of the nodes, we can gain insights into the underlying structure and organization of the network. The starting point is the Stochastic Block Model (SBM) [Nowicki and Snijders, 2001, Schmidt and Morup, 2013] endowed with Bayesian nonparametric priors on the partitions [Legramanti et al., 2022]. This is a classical model-based technique used to find the best latent partition a posteriori. Using the standard version of the SBM, or most of its generalizations in the literature, it is possible to sample from the posterior distribution of the partitions employing a standard Gibbs sampler. This thesis extends the SBM in two, non-trivial directions.

Chapter 3 introduces a Bayesian hierarchical version of the SBM defined for multiplex networks (also known as edge-colored networks). A multiplex network is a multi-layer structure, with each layer representing a different type of relationship among the same entities [Kivelä et al., 2014]. A classical example for networks of this type are social networks, where a layer could represent friendships between individuals, another layer could represent professional collaborations, and yet another layer could represent family relationships. The model introduced in Chapter 3 is the multiplex Extended Stochastic Block Model (mESBM): the goal of the mESBM is to find two types of partitions within a multiplex network. On the one side, a set of layer-specific clusters within each layer of the edge-colored network, where each one groups nodes within the same layer; on the other side a general (or common) partition of the entities in the edge-colored graph. In the mESBM, the layer-specific partitions are not independent, since the model allows borrowing of information across the clusters in different layers through the dependence induced by the common grouping. Such dependence across layers is desirable, since the nodes represent the same entities, and we believe that it is important to acknowledge and exploit such information for the inference of all the partitions. It is worth noticing that even though the mESBM is a generalization of the extendend stochastic block model [Legramanti et al., 2022], its inference is not trivial: to sample from the posterior distribution of the clusters a combination of nested Monte Carlo algorithms needs to be employed. The model is subsequently applied to a multiplex network derived from human brain scans obtained from a group of patients, where each layer of the edge-colored network contains a functional brain map of a single subject: in this case, the partitions provided by the mESBM have a two-fold meaning. The layer-specific clusters refer to the groupings of brain areas that are specific to each individual patient: we argue that these subject-specific groups are related to possible mental illnesses and diagnosis. On the other hand, the common clustering provides an anatomical division of the human brain, shared by all the subjects involved in the study. The identification of this common partition provides valuable insights into the fundamental organization of the human brain, and suggests the presence of a shared functional structure.

Chapter 4 introduces the Poisson extended stochastic block model (pESBM), which generalizes the extended stochastic block model [Legramanti et al., 2022] to weighted networks with continuous and multidimensional node attributes. The specific objective of this generalization is to explore the latent structure of a transportation network provided by a local public company, which is embedded within a geographical space. The primary aim is to obtain spatially coherent clusters that reflect the inherent organization of the network. From an applicative viewpoint, understanding the geographical structure of the transportation network and obtaining spatially coherent clusters can have practical implications for transportation planning and optimization. It can inform decisions related to route planning, resource allocation, and infrastructure development, ultimately enhancing the efficiency and effectiveness of the transportation system in serving the needs of the different municipalities. In this transportation network, each node represents a municipality, and these municipalities are interconnected by public transport lines. To achieve spatially coherent clusters, the analysis takes into consideration the geographic information associated with each town. This information typically includes longitude and latitude coordinates that represent the geographic location of each node. The desired outcome is to encourage the model to generate radial clusters, which are commonly observed in transportation networks and useful for different practical scopes. Radial clusters refer to groupings of nodes centered around a hub, with nodes radiating outwards from that central point. In transportation networks, radial clusters can often be observed due to the organization of routes and transportation means. To achieve the objective of identifying such clusters in the transportation network, the analysis adopts a two-step approach. First, a suitable node attribute is selected, and then its supervision is incorporated into the estimation of the partition using a product partition model structure [Muller et al., 2011, Page and Quintana, 2015, 2018]. The selection of a suitable node attribute is an essential step in the analysis: we decided to use the distance from the main hub of the transportation network as a covariate. By incorporating such a covariate, we aim to induce the formation of radial clusters in the partitioning process: in fact, by utilizing the distance as node attribute, we encourage the partitioning algorithm to assign nodes at similar distances from the main hub into the same group, promoting the formation of radial clusters. Once the attribute is chosen, it can be used to guide the estimation of the partition. A product-partition-model structure is employed, which is a probabilistic model used to supervise the partitioning of the network with node attributes. This model allows for the incorporation of supervision through a so-called similarity function, which measures the cohesion of each cluster with respect to node attributes and encourages clusters with high similarity. The choice of the similarity function is crucial for the successful application of the model, since it directly affects the assessment of cohesion within clusters and plays a significant role in determining the quality of the partitioning results. Besides the similarity function, the pESBM also includes a smoothing parameter, which in the case of interest tunes the amount of spatial smoothing in the partitions provided by the geographical positions of the nodes, or, alternatively stated, the amount of information provided by the node attributes for the estimation of the partitions. Such a parameter can either be user-defined, or it can be inferred a posteriori in a mathematically coherent way. As for the latter case, inference of parameters in product partition models can be challenging due to the unavailability of their normalizing constant. However, a workaround to address this issue is to define an ad hoc joint prior distribution for both the smoothing parameter and the partitions, enabling the sampling from their posterior distributions.

Finally, I would like to acknowledge the different people I have worked with in every chapter of this thesis:

- Chapter 2 is a joint work with Emanuele Borgonovo and Elmar Plischke. It resulted in the following publication:
  - E. Borgonovo, **V. Ghidini**[1], R. Hahn, E. Plischke (2023) *Explaining classifiers with measures of statistical association*. Computational Statistics and Data Analysis, (182)107701.
- Chapter 3 is a joint work with Daniele Durante and Omiros Papaspiliopoulos.
- Chapter 4 is a joint work with Sirio Legramanti and Raffaele Argiento. Some works related to this chapter are available in the following two papers:
  - **V. Ghidini**, S. Legramanti and R. Argiento (2023) *Extended stochastic block model with spatial covariates for weighted brain networks*. BAYSM2022 (to appear);
  - **V. Ghidini**, S. Legramanti and R. Argiento (2023) *Binomial extended stochastic block model for brain networks*. Book of short paper SIS 2023 (to appear).

[1] Authors in alphabetical order

Part I

Explainability

Chapter 2

Explaining black boxes with measures of statistical association

2.1 Introduction

We have access to increasingly advanced and precise statistical models. However, they often come with a limitation: the difficulty of interpreting their parameters or predictions. Until a few decades ago, statistical models had simple structures and thus there was no urge to explain them: Machine Learning (ML) itself mostly relied on classical statistical models such as linear regression, generalized linear models, generalized additive models, which all come with a precise interpretation of the relationship between the input and the output. But the explosive growth and availability of data and the remarkable advancements in hardware technologies allowed the deployment of new, complex models, dramatically improving model accuracy and performances. The main and most innovative idea is to use multi-layer learning models (such as neural networks) [Goodfellow et al., 2016] to explore complex relationships among the data. This has lead to incredible applications in a huge amount of fields: speech and audio recognition, natural language processing or computer vision, just to cite some.

However, the opaque nature of ML models raises some ethical considerations: in spite of their high performance in many domains, it often proves extremely difficult to explain their decisions in a humanly understandable way. This is known as the *black box problem* [Molnar, 2018, Rudin, 2019], where researchers are unable to fully comprehend the factors contributing to the output of certain models. Alternatively stated, it is possible to observe the realizations of the input-output mapping of the model, but its internal operations do not make any sense from a human perspective. One of the main reasons causing this issue is the fact that most ML models are indeed sub-symbolic systems [Huang, 2010], trained with a data-driven approach, meaning that all the system's learning process is completely automatic, without any additional human-based knowledge. The only input for the training process is raw data. In this way, the model can construct its own representation of the entities and perform its own feature engineering, which often does not follow any logic from a human viewpoint. The growing discrepancy between human cogitation and sub-symbolic representations makes it more and more difficult to interpret the decisions of a ML model, since in principle the latter can extrapolate and exploit irrelevant information from a human perspective and ignore other fundamental features, making the decision process chaotic and unreadable from a standard point of view.

There are also other reasons at the root of the need for explanation of models' decisions: first of all, the variety of tasks that systems are required to attend in sensitive fields like healthcare or credit scoring demands an explanation for any decisions they are taking. This can also be linked to the trustfulness of the domain expert: a person is more likely to trust (or correct and improve) the decisions taken by an algorithm if the decision process can be explained, especially if the predictions are in contrast with his or her belief. Secondly, from a legal viewpoint [Goodman and Flaxman, 2017, Wachter et al., 2018], ML systems in Europe are required by law to comply with the General Data Protection Regulation (GDPR) since May 2018, which regulates algorithmic decision making. The most pertinent contribution of GDPR motivating explainability is the statement that the decisions *"which produces legal effects concerning him or her or of similar importance shall not be based on the data revealing racial or ethnic origins, political opinions, philosophical beliefs, health situations, gender or sexual orientation"*. Since it is usually impossible to maintain the same level of accuracy (and usefulness) of statistical models excluding completely these sensitive information from the training process, the need of explanations for a decision of an automated model arises, in order to exclude that the output is influenced by the personal details mentioned above. Giving an explanation is also potentially useful to spot some bias in the training data collection [Molnar, 2018]: for example, the model can have different levels of confidence for its predictions about different groups of people, due to one or some of the groups being underrepresented in the data, or it can take the right decision for the wrong reasons exploiting spurious correlations in the dataset.

The combination of these factors gave rise to the development of eXplainable Artificial Intelligence (XAI), also known as Interpretability or Explainability. This burgeoning field of research is dedicated to discovering novel techniques for explaining the underlying logic behind a model's output. While XAI originated in computer science, much of the literature has yet to analyze the explanations provided by various algorithms from a statistical perspective. In this chapter, we try to overcome this limitation by defining post hoc explanations computed using statistically consistent estimators.

2.1.1 Terminology

A formal and unique definition of interpretability does not exist in literature (at the time of writing). Quoting Murdoch et al. [2019], *"interpretability is a broad, poorly defined concept"*. However, Murdoch et al. [2019] introduce a definition of interpretable machine learning as *"the extraction of relevant knowledge from a model concerning relationships either contained in data* *or learned by the model"*. Miller [2019] writes extensively about all the semantics behind the terminology concerning explanations. In particular, he addresses the distinction between *interpretability*, *explainability*, *justification*, and *explanation*, which are recurrent terms in this literature. We report a mini-glossary below, to clarify the terms used in the reminder of this chapter.

- **Interpretability**: the degree to which an observer can understand the causes of a decision in a model from the model itself.
- **Explainability**: the degree to which an observer can understand the causes of a decision, possibly in a post hoc fashion and with the application of an additional tool.
- **Explanation**: a tool to obtain understanding of a prediction process, through the explanation of the decision of a model after it has been designed (and sometimes trained).
- **Justification**: it explains why a decision is good, but it does not necessarily aim to give an explanation of the actual decision-making process.

In ML, interpreting a model means giving an explanation to the decisions of a certain architecture, which must be at the same time an accurate proxy of its decision making process and understandable to humans. Notice that interpretability and explainability are used interchangeably in most of the literature, even though it can be argued that they have different meanings: interpretability concerns transparent ML models (that is, we are directly interpreting the model), while explainability may also regard post hoc explanations (that is, post hoc interpretations of black boxes). Basically, interpretability can be seen as a specific instance of explainability, which can be used only when we are dealing with an interpretable model. However, in the remainder, the two terms will be often used interchangeably in accordance with the literature.

2.2 Explainability methods

In this section, we present the main features characterizing explainability techniques.

2.2.1 Features of explainability methods

The explanations discussed in Section 2.1.1 are usually obtained through the application of explainability methods [Robnik-Šikonja and Bohanec, 2018]. To begin, we outline the main distinguishing features of each technique [Guidotti et al., 2018, Molnar, 2018]:

- *Expressive power*: it refers to the degree to which the form or structure of the explanation generated by the method contributes to its ease of understanding. For instance, an explanation may take the form of IF-THEN rules, natural language sentences, an interpretable model, or other formats, each with varying degrees of expressive power.

- *Translucency*: it refers to the extent to which an explanation method reveals the inner workings of the model. For instance, the explanation of a linear regression model is highly translucent since the coefficients have direct meaning. On the other hand, model-agnostic methods do not rely on the model's parameters and are therefore less translucent.

- *Portability*: it describes the range of models which the explanation method can be applied to.

2.2.2 Taxonomy of explainability methods

The literature provides different criterions to classify explainability methods. Before delving into our original work in Section 2.3, it is important to clarify some terminology that we have introduced to aid the understanding.

The first distinction [Guidotti et al., 2018] is between *model-agnostic* versus *model-aware* techniques: the former can be used with any kind of model, and usually rely on perturbing the input and studying how the output varies. The latter exploit some intrinsic feature of the model to provide the explanations (e.g. gradients in neural networks as in Binder et al. [2016], Selvaraju et al. [2020]). Another discrimination is given by *data-agnostic* versus *data-aware* algorithms. Data-agnostic algorithms can be applied to any kind of data structure, such as images, texts, or other formats. In contrast, data-aware techniques are designed to be applied to a specific data type. For example, RISE [Petsiuk et al., 2019] is an algorithm that can only be applied to images, while Partial Dependence Plots [Apley and Zhu, 2020] are designed for use with tabular data. Another important distinction is given by the following [Murdoch et al., 2019]:

- *White-box interpretability*: this set of techniques involves specifying an interpretable model a priori, without sacrificing accuracy. In general, this requires a huge effort in statistical modeling and feature engineering, and the result is a completely interpretable model, which does not need any additional approximation to provide an explanation.

- *Black-box explainability*: this set of algorithms is used to explain a black box model in a post hoc manner. It usually involves the interpretation of a simpler approximation of the model of interest [Ribeiro et al., 2016] or a perturbation of the input and a measure of how the output changes in function of that [Breiman, 2001, Petsiuk et al., 2019]. In this case, we then have two levels of approximations (which can be sources of errors): first, we estimate the true data generating process through a black box. Then, we use a proxy of the model as explanation. The advantage, however, is that this type of explanations can be obtained for any kind of model, without any additional field knowledge or feature selection.

Let us also distinguish two further types of explanatory techniques [Guidotti et al., 2018, Murdoch et al., 2019]:

- *Instance or Local methods* measure the impact of input features in a given model for a single data instance;
- *Model or Global methods* provide an interpretation of the behavior of the features of interest for the entire dataset, for the analyzed model. They can be the result of an aggregation of instance explanations over many training data points.

The two mentioned characteristics are not mutually exclusive: often, a global technique yields information at an instance level. Finally, we would like to add a novel distinction among XAI outputs:

- *Model-dependent explanations* provide a measure of the impact of input features in a given, specific model (e.g. Ribeiro et al. [2016], Petsiuk et al. [2019] and many others).
- *Model-independent explanations* provide a measure of feature importance for the prediction task, independently of a single model. Some examples may be found in Fisher et al. [2019], Dong and Rudin [2020], where this problem is addressed considering the influence on the Rashomon set of models. The pre hoc explanations introduced in Section 2.3 also belong to such a class.

We devote the next subsection to *model-agnostic* techniques, a class of methods that includes the Xi method proposed in Section 2.3.

Model-agnostic techniques

An explainable technique is defined as *model-agnostic* if it can be applied to any model, as it does not exploit any feature of its structure to obtain explanations. This category includes some families of variable importance indices, such as the ones obtained using sensitivity measures and perturbation-based techniques. We introduce such concepts in the next two paragraphs.

Sensitivity measures & feature importance Sensitivity Analysis (SA) [Saltelli, 2008] involves the quantification of how the uncertainty in a target variable, whether observed or predicted by a model, can be attributed to different sources in a given set of input variables. Although SA and XAI developed independently, there is a significant overlap in their goals and techniques. A class of indices proposed in SA are measures of statistical dependence [Borgonovo et al., 2016], which quantify the strength of dependence between the output Y and an input of interest X. We will generalize (and recall) such measures in Section 2.3.

There also exist other procedures estimating indices of feature importance using SA: some examples of this type of indices are knockoffs [Barber and Candés, 2015] and model-X knockoffs [Candès et al., 2018]. In both techniques, the aim is to construct knockoff versions of the original features which allow to find the smallest subset $\mathscr{S}$ of variables $\{X_j\}_{j=1}^p$ such that the response Y is independent of all the other variables conditionally on $\{X_{j\in\mathscr{S}}\}$. This is a feature selection procedure, and the most recent version allows its application to high-dimensional settings (with $p \geq n$). An alternative technique is the Sure Independence Screening (SIS) [Fan and Lv, 2008]: SIS is also a feature selection procedure which picks out a subset of covariates according to the coefficients $w = X^T Y$ (componentwise regression). A peculiar approach is provided by Fisher et al. [2019] and Dong and Rudin [2020]: the authors try to combine the concept of having a model-independent measure of feature importance with the goal of providing explanations for the general prediction task of a specific variable. They avoid the bias of model-dependent explanations by defining feature importance measures over a set of optimal models (i.e. the Rashomon set), emphasizing the consistency of the explanations across such set.

Perturbation-based techniques Perturbation-based techniques employ the corruption of (one or more) covariates of interest, and the study of the consequent effect on the predictions of the model. A perturbation is defined as noise introduced in the data (e.g. through permutation, random modification of the design matrix, noise addition). The most famous index of this type is the Permutation Variable Importance introduced by Breiman [2001] for Random Forests.

Definition 2.2.1. *The* ***Permutation Variable Importance Measure*** *for a covariate in a Random Forest is the average difference in the performance obtained after the permutation of the feature of interest, evaluated on the Out-Of-Bag observations (i.e. the ones excluded by the training of the single tree).*

Even if it was originally specified for Random Forests, this measure can be generalized to any model [Pavlov, 2019].

In Section 2.3, we focus on data-agnostic, black-box explainability, i.e. techniques able to provide a post hoc explanation for predictions by any model or on any kind of data. We propose the Xi method, a data-agnostic and model-agnostic technique, which is also global and can be model-independent or model-dependent according to each user's needs.

2.3 Explaining black-box classifiers with measures of statistical association

Understanding statistical dependence is a challenging task, especially when target variables live on a support without an intrinsic order. In this section (which comprises the original part of this chapter), we propose a new class of probabilistic sensitivity measures that quantifies the degree of association between covariates and generic targets used in classification. We show that the class possesses the zero-independence property and we introduce corresponding estimators, prove asymptotic consistency and use bootstrap to quantify uncertainty in the estimates. We illustrate the use of the new dependence measures in a ML context, providing post hoc explanations. The resulting approach, called Xi method, is demonstrated through applications involving different data formats: tabular, visual and textual.

2.3.1 Framework and motivation

The growing size and complexity of data structures and the simultaneous need of accurate predictions force analysts to employ black-box rather than transparent ML models in an increasing number of applications. While the success of such models extends the range of ML applications, it also increases the need of methods that aid explainability [Dunson, 2018, Rudin, 2019, Murdoch et al., 2019]. Determining feature importance is essential for model simplification, dimensionality reduction, and for understanding whether predictions are at risk of unfair discrimination [Fisher et al., 2019]. As underlined in Murdoch et al. [2019], there is a variety of techniques for calculating feature importance. Methods such as the Local Interpretable Model-agnostic Explanations (LIME) [Ribeiro et al., 2016], the Layerwise Relevance Propagation (LRP) [Binder et al., 2016], and the SHapley Additive exPlanations (SHAP) [Lundberg and Lee, 2017] yield feature importance measures at the individual prediction level. Techniques such as permutation or removal [Breiman, 2001], Shapley values [Lundberg and Lee, 2017, Lundberg et al., 2019], knockoffs [Barber and Candés, 2015, Candès et al., 2018] and Sure Independence Screening [Fan and Song, 2010] provide an indication of importance at the dataset level.

Measures of statistical association are an alternative way for assessing feature importance: they provide dataset scores and are model-agnostic [Murdoch et al., 2019]. While they have been widely studied beginning with works such as Pearson [1905], Hotelling [1936] and Renyi [1959], the size and complexity of modern datasets have generated new interest in their construction, as the recent works of Pan et al. [2020] and Chatterjee [2021] highlight. We note that several measures of statistical association rely on the assumption that the target is a real number (or vector). However, in several ML applications, targets and/or features may be images or objects, for which a mathematical order relation may not be appropriate. To illustrate, consider an image recognition task in which the target consists of three different types of pictures representing, say, cats, tigers and rabbits. Then the alphabetical ordering "cat", "rabbit", "tiger" is a possible ordering, but it is as valid as the ordering "cat", "tiger", "rabbit" that is based on the number of the characters in each word, and none of them qualifies as a natural ranking. The issue is underlined in recent works (e.g. Da Veiga [2015]) and the definition of probabilistic sensitivity measures for non-ordered outputs is a topical research subject.

To bridge this gap we propose a family of measures of statistical association that is well-defined also for non-ordered data. Our intuition is to rely on separation measurements between probability mass functions. Here, by separation measurement we mean any distance or divergence between probability mass functions that is positive, and that is null if and only if the probability mass functions coincide. Then, we show that the new class of sensitivity indices complies with Renyi's postulate D of measures of statistical dependence [Renyi, 1959]. This postulate, called zero-independence property in the following, requires that a measure of association is null if and only if the two random variables are statistically independent. We address the estimation of this new class of indicators for generic samples, and discuss their asymptotic convergence. We then use these probabilistic sensitivity measures in the context of explainability. A relevant aspect related to measures of statistical association is that they can be computed directly on the original dataset without the need of actually fitting a machine learning model. Thus, not only are they model-agnostic in explaining the behaviour of a black box, but they can also provide both model-dependent and model-independent explanations. Our aim is then to compare explanations provided by measures of statistical association first calculated on the original data (pre hoc explanations) and then on the forecasts of the ML model fitted to the data (post hoc explanations). This comparison provides an indication on whether the ML model predictions capture the statistical dependence originally present in the data. We call the resulting approach Xi method.

We proceed as follows: first, we discuss the methodological framework, with a focus on the choice of the separation measurement. Then, we address estimation, with a focus on a partition-based approach that allows us to obtain the explanations from a given dataset. We perform experiments to investigate the asymptotic convergence of the estimates and highlight limitations of the approach with particular reference to the curse of dimensionality. We then test the approach by performing experiments on datasets of alternative types, comprising tabular, image and textual data. For the first set of experiments, in which the ML model is a random forest, we compare the pre hoc and post hoc explanations provided by measures of statistical dependence with the post hoc explanations delivered by variable importance measures based on split and count [Breiman, 1984] and permutation [Breiman, 2001]. The rationale of this comparison is to test the agreement between results provided by measures of statistical association with one representative of variable importance measures that are post hoc and model-dependent (split and count) and one representative of measures that are post hoc but model-agnostic (permutation feature importance). Measures of statistical dependence produce additional and complementary insights with respect to alternatives currently in use, with the advantage of being model-agnostic and computationally convenient.

The remainder of the Section is organized as follows. Subsection 2.3.2 sets up the relevant framework and reviews the literature. Subsection 2.3.3 introduces the new dependence measures for classification. Subsection 2.3.4 presents the Xi method. Subsection 2.4 illustrates results for alternative datasets. Subsection 2.5 offers conclusions.

2.3.2 Feature importance and measures of association

Let $(\Omega, \mathscr{B}(\Omega), \mathbb{P})$ be a reference probability space (where Ω is called sample space, $\mathscr{B}(\Omega)$ is a Borel σ-algebra, and $\mathbb{P} : \mathscr{B}(\Omega) \to [0,1]$ is the reference probability measure) and let $\mathbf{X} = (X_1, X_2, \ldots, X_{n_\mathbf{X}})$ and L be random variables on this reference space with supports $\mathscr{X}$ and $\mathscr{L}$, respectively. Let $\mathbf{X}$ have the meaning of vector of features and let L represent target labels. For example, in image classification, we might be dealing with a set of images of a given resolution, say n_{pixels}. Frequently, pixels themselves are selected as features, yielding $\mathbf{X} = (X_1, X_2, \ldots, X_{n_{\text{pixels}}})$. The set of labels is then the list of objects depicted in the corresponding images. The support of each pixel ($\mathscr{X}_i, i = 1, \ldots, n_{\text{pixels}}$) is its range of values (for instance $\mathscr{X}_i = [0,1]$ in case of grayscale images), and the overall support $\mathscr{X}$ is the Cartesian product of the ranges of all pixels. The support of the target $\mathscr{L}$ is the list of all the possible labels associated to each image in the dataset, that is the set of objects represented in the data collection (e.g., cats, rabbits, tigers as we were mentioning in the previous section). We are interested in associating realizations of $\mathbf{X}$ to realizations of L through the input-output mapping $g : \mathscr{X} \times \Theta \to \mathscr{L}$ [Hastie et al., 2009, Zhao and Hastie, 2019]

$$\Lambda = g(\mathbf{X}, \theta), \tag{2.1}$$

where Λ denotes a model forecast and $\theta \in \Theta$ is a vector of parameters (or rules). In supervised learning, a dataset $(\mathbf{x}^{(n)}, L^{(n)})$, $n = 1, 2, ..., N$, of realizations of $\mathbf{X}$ and L is usually available (henceforth, N denotes the sample size). Splitting the sample into training and testing subsamples of sizes N^{Tr} and N^{Te} respectively (on the meaning and rationale for training and testing in supervised learning see classical references such as Hastie et al. [2009]), the parameters of the input-output mapping are determined via the solution of an optimization problem of the form

$$\theta^* = \operatorname*{argmin}_{\theta} \left\{ \frac{1}{N^{Tr}} \sum_{n=1}^{N^{Tr}} \mathscr{C}\left(L^{(n)}; g(\mathbf{x}^{(n)}, \theta)\right) \right\}, \tag{2.2}$$

where $\mathscr{C} : \mathscr{L} \times \mathscr{L} \to \mathbb{R}$ is a suitably defined objective (loss) function. In the remainder, we shall use the shorter $\widehat{g}(\mathbf{X})$ for $g(\mathbf{X}, \theta^*)$. Then, let $\mathbf{X}_{\pi(i)}$ denote the design matrix in which we have randomly permuted the realizations of the i^{th} feature X_i. Let also $\mathscr{C}(\mathbf{X}; \theta^*) = \frac{1}{N^{Tr}} \sum_{n=1}^{N^{Tr}} \mathscr{C}(L^{(n)}; g(\mathbf{x}^{(n)}, \theta^*))$ denote the value of the optimized loss function and $\mathscr{C}(\mathbf{X}_{\pi(i)}; \theta^*) = \frac{1}{N^{Tr}} \sum_{n=1}^{N^{Tr}} \mathscr{C}(L^{(n)}, g(\mathbf{x}_{\pi(i)}^{(n)}, \theta^*))$ the expected loss registered when feature i is permuted, both computed on the training dataset. Then, the permutation importance of feature X_i is given by

$$\text{PI}_i = \mathscr{C}(\mathbf{X}_{\pi(i)}; \theta^*) / \mathscr{C}(\mathbf{X}; \theta^*). \tag{2.3}$$

This indicator measures the deterioration in the ML model performance caused by disrupting the dependence between Y and X_i.

Measures of statistical dependence instead quantify importance from a different perspective, evaluating the degree of association between the target and one or more features. The problem of measuring statistical association roots back to works such as Pearson [1895, 1905] and Hotelling [1936] and in the statistical literature we find several measures of association and tests for independence. Recently, renewed interest has been generated by the type and size of modern datasets (see Chatterjee [2021] for a review). In particular, among measures of statistical dependence recently studied, we find the Hilbert–Schmidt independence criterion [Da Veiga, 2015], distance correlation [Székely et al., 2007, Székely and Rizzo, 2009, Chaudhuri and Hu, 2019] as well as a new correlation coefficient [Chatterjee, 2021]. To provide the background needed for the remainder of our investigation, we recall the following definition.

Definition 2.3.1 (Separation Measurement [Glick, 1975])**.** *Let $\mathscr{P}$ be the set of all probability measures on $(\Omega, \mathscr{B}(\Omega))$. A separation measurement between $\mathbb{P}, \mathbb{Q} \in \mathscr{P}$ is given by a function $\zeta : \mathscr{P} \times \mathscr{P} \to \mathbb{R}$ such that a) $\zeta(\mathbb{P}, \mathbb{Q}) \geq 0$, and b) $\zeta(\mathbb{P}, \mathbb{Q}) = 0$ if and only if $\mathbb{P} = \mathbb{Q}$.*

Let now Y and $X \subseteq \mathbf{X}$ (i.e. X is a subset of one or more covariates of interest contained in the original design matrix $\mathbf{X}$) be random variables on $(\Omega, \mathscr{B}(\Omega), \mathbb{P})$, and denote by $\mathbb{P}_Y$ and $\mathbb{P}_X$ their marginal laws, and by $\mathbb{P}_{Y|X}$ the conditional law of Y given X. Without loss of generality, we will assume X to be a univariate random variable (if not specified otherwise).

Definition 2.3.2 (Probabilistic Sensitivity Measure)**.** *We define*

$$\xi_X = \mathbb{E}_X[\zeta(\mathbb{P}_Y, \mathbb{P}_{Y|X})] \tag{2.4}$$

as the probabilistic sensitivity measure of X with respect to Y based on the separation measurement $\zeta(\cdot, \cdot)$.

In Equation (2.4), the expectation is taken with respect to the law of X, that is,

$$\xi_X = \int_{\mathscr{X}} \zeta(\mathbb{P}_Y, \mathbb{P}_{Y|X=x}) dF_X(x), \tag{2.5}$$

where $F_X(x)$ is the cumulative distribution function of X and the integral is interpreted in a Riemann-Stieltjes sense. To illustrate, if X is continuous with density $f_X(x)$ then Equation (2.5) becomes

$$\xi_X = \int_{\mathscr{X}} \zeta(\mathbb{P}_Y, \mathbb{P}_{Y|X=x}) f_X(x) dx. \tag{2.6}$$

If X is discrete with realizations $x_1, x_2, \ldots, x_n$, then Equation (2.5) becomes

$$\xi_X = \sum_{i=1}^{n} \zeta(\mathbb{P}_Y, \mathbb{P}_{Y|X=x_i}) p(X = x_i). \tag{2.7}$$

Also notice that if Y and X are independent, then $\mathbb{P}_Y = \mathbb{P}_{Y|X}$, implying $\zeta(\mathbb{P}_Y, \mathbb{P}_{Y|X}) = 0$ and consequently $\xi_X = 0$. This means that the index ξ_X complies with the zero-independence property [Renyi, 1959].

The choice of the separation measurement in Equation (2.4) determines the properties of the dependence measure ξ_X. For instance, if $\zeta(\cdot,\cdot)$ is invariant for monotonic transformation of Y then ξ_X is also monotonic transformation invariant [Borgonovo et al., 2014]. If Y is absolutely continuous then selecting the Kullback-Leibler divergence as separation measurement has a corresponding probabilistic sensitivity measure equal to the mutual information between X and Y [Soofi, 1994]. Recently, Taverniers et al. [2021] have proposed this importance measure in the context of neural network interpretability. They have developed a deep neural network to emulate the behavior of a complex system in a forecasting task, and then the mutual information is used to quantify feature importance with respect to the neural network predictions. Alternatively, if one selects the separation measurement as the Cramér-von Mises distance as in Gamboa et al. [2018], one obtains

$$\xi_X^{\text{CvM}} = \mathbb{E}_X \left[\int_{\mathbb{R}} \left(F_Y(y) - F_{Y|X}(y)\right)^2 dF_Y(y) \right]. \tag{2.8}$$

Notice that ξ_X^{CvM} is the limiting value of Chatterjee's new correlation coefficient [Chatterjee, 2021]. Both the mutual information and ξ_X^{CvM} are then transformation invariant and possess the zero-independence property.

The estimation of probabilistic sensitivity measures has been extensively studied, with computational breakthroughs obtained in works such as Chan et al. [2000], Saltelli [2002], Strong et al. [2012], and Gamboa et al. [2016], among others. Relevant to our work is a given-data estimation approach that enables direct estimation from a dataset. The key intuition dates back to Pearson [1905], and is extensively studied in works such as Strong et al. [2012], Strong and Oakley [2013] for the calculation of variance-based sensitivity measures, and Plischke and Borgonovo [2014] for distribution-based sensitivity measures. Recently, Gamboa et al. [2022] have shown that the newly introduced Chatterjee's rank-based correlation coefficient [Chatterjee, 2021] can be used as a given-data estimator for ξ_X^{CvM}. The connection leads to an elegant and advantageous approach for the calculation of a global sensitivity measure in the form of Equation (2.4) from a given dataset. However, the advantage is lost in a classification setting, because the targets are not necessarily elements of an ordered space and cannot be ranked univocally. This then opens the question of defining probabilistic sensitivity measures for classification tasks. We address this issue next.

2.3.3 Probabilistic sensitivity measures for supervised classification

Let $\mathscr{L} = \{\ell_1, \ell_2, \ldots, \ell_{n_L}\}$ denote the support of L, i.e. the set of labels in the target variable of interest. Without loss of generality, we assume that for all $\ell \in \mathscr{L}$, $\mathbb{P}(\{L = \ell\}) > 0$. The support $\mathscr{L}$ represents in general a list of objects and consequently the realizations of L may not be ordeable. Let $\mathbf{p}_L$ be the probability mass function (pmf) of L and $\mathbf{p}_{L|X}$ the conditional pmf given a feature (or a feature group) X. Without loss of generality, in the remainder we will consider X to be a single feature, if not explicited otherwise. We recall that $\mathbf{p}_L = \{\mathbb{P}(\{L = \ell_1\}), \ldots, \mathbb{P}(\{L = \ell_{n_L}\})\} = \{p_1, p_2, \ldots, p_{n_L}\}$ is a probability mass function if $p_l \geq 0$ for all $l = 1,\dots,n_L$ and $\sum_{l=1}^{n_L} p_l = 1$. Let $\mathscr{P}^{\text{mf}}$ denote the set of all probability mass functions on $(\mathscr{L},\mathscr{P}(\mathscr{L}))$ and let $\zeta(\cdot,\cdot)$ denote a separation measurement between probability mass functions, $\zeta : \mathscr{P}^{\text{mf}} \times \mathscr{P}^{\text{mf}} \to \mathbb{R}$. By Definition 2.3.1, ζ is such that given $\mathbf{p},\mathbf{q} \in \mathscr{P}^{\text{mf}}$ it holds $\zeta(\mathbf{p},\mathbf{q}) \geq 0$ and the equality holds if and only if $\mathbf{p} = \mathbf{q}$. Then, we propose the following definition.

Definition 2.3.3. *We call*

$$\xi_X^L = \mathbb{E}_X\left[\zeta(\mathbf{p}_L,\mathbf{p}_{L|X})\right] \tag{2.9}$$

the probabilistic sensitivity measure of X with respect to L based on $\zeta(\cdot,\cdot)$.

Formally, ξ_X^L in Equation (2.9) is a particular case of ξ_X in Equation (2.4). However, Equation (2.9) explicitly considers the marginal and conditional probability mass functions of the labels. The rationale is that probability mass functions are defined without ambiguity also when labels are non-ordered data. In fact, the corresponding cumulative distribution function would require an additional convention, that is, we need to order the labels and then stick to such lexicographic order. If an alternative order is used, we get an alternative cumulative distribution function. Relying on probability mass functions avoids the additional step of introducing an order relation. In Table 2.1, we report three potential choices for the separation measurement between two probability mass functions $\mathbf{p} = \{p_1,\dots,p_{n_L}\}$ and $\mathbf{q} = \{q_1,\dots,q_{n_L}\}$ defined on the same sample space. Specifically, $\zeta^1(\mathbf{p},\mathbf{q})$ is a separation measurement based on the 1-norm (absolute value of the differences), $\zeta^2(\mathbf{p},\mathbf{q})$ is a separation based on the 2-norm (square value of the differences) and $\zeta^{\text{KU}}(\mathbf{p},\mathbf{q})$ is based on an extension to the discrete case of the Kuiper distance [Kuiper, 1960] (which is, in turn, an extension of the Kolmogorov-Smirnov distance). In $\zeta^{\text{KU}}(\mathbf{p},\mathbf{q})$, we have $\text{range}(\mathbf{p}-\mathbf{q}) = \max_{l=1}^{n_L}(p_l - q_l) - \min_{l=1}^{n_L}(p_l - q_l)$.

Table 2.1: Three possible separation measurements based on the 1-norm, 2-norm and Kuiper distance between probability mass functions.

1-norm	**2-norm**	**Kuiper distance**
$\zeta^1(\mathbf{p},\mathbf{q}) = \sum_{l=1}^{n_L} \lvert p_l - q_l \rvert$	$\zeta^2(\mathbf{p},\mathbf{q}) = \sum_{l=1}^{n_L}(p_l - q_l)^2$	$\zeta^{\text{KU}}(\mathbf{p},\mathbf{q}) = \text{range}(\mathbf{p}-\mathbf{q})$

Example 2.3.4. *To illustrate the separation measurements in Table 2.1, let L be a categorical variable with support $\mathscr{L} = \{A,B,C\}$ and consider the probability mass functions given by $\mathbf{p} = [\frac{1}{3},\frac{1}{3},\frac{1}{3}]$ and $\mathbf{q} = [\frac{1}{4},\frac{1}{2},\frac{1}{4}]$. Then, we have:*

$$\begin{aligned}
\zeta^1(\mathbf{p},\mathbf{q}) &= \left|\frac{1}{3}-\frac{1}{4}\right| + \left|\frac{1}{3}-\frac{1}{2}\right| + \left|\frac{1}{3}-\frac{1}{4}\right| = \frac{1}{3},\\
\zeta^2(\mathbf{p},\mathbf{q}) &= \left(\frac{1}{3}-\frac{1}{4}\right)^2 + \left(\frac{1}{3}-\frac{1}{2}\right)^2 + \left(\frac{1}{3}-\frac{1}{4}\right)^2 = \frac{1}{24},\\
\zeta^{KU}(\mathbf{p},\mathbf{q}) &= \text{range}\left(\frac{1}{3}-\frac{1}{4},\frac{1}{3}-\frac{1}{2},\frac{1}{3}-\frac{1}{4}\right) = \frac{1}{12} - \left(-\frac{1}{6}\right) = \frac{1}{4}.
\end{aligned}$$

Analysts can select separation measurements between probability mass functions that go beyond the ones listed in Table 2.1, such as the Kullback-Leibler divergence, or the Hellinger distance, or any other particular choice that best suites her/his needs for the application at hand. The next results state that as long as the separation measurement follows the requirements in Definition 2.3.1, members of the ξ_X^L family in Equation (2.9) possess the zero-independence property. Indeed, recall that for all the measurements $\zeta(\cdot,\cdot)$ complying with Definition 2.3.1, $\zeta(\mathbf{p}_L,\mathbf{p}_{L|X}) = 0$ when $\mathbf{p}_L = \mathbf{p}_{L|X}$, that includes the case when L is independent of X.

Proposition 2.3.5. *Given the above setup, if $\zeta(\cdot,\cdot)$ is a separation measurement between probability mass functions then $\xi_X^L \geq 0$, and $\xi_X^L = 0$ if and only if L is independent of X.*

Let us now turn to the estimation of the probabilistic sensitivity measures of $X_i \in \mathbf{X}$ with respect to L, for $i \in \{1,\dots,n_X\}$. Let $\mathcal{X}_i$ denote the support of X_i. Let also $\mathcal{K}_i = \{\mathcal{X}_i^1,\mathcal{X}_i^2,\dots,\mathcal{X}_i^K\}$ denote a partition of $\mathcal{X}_i$, i.e., a finite or countable collection of K subsets of $\mathcal{X}_i$ such that $\mathcal{X}_i = \bigcup_{k=1}^K \mathcal{X}_i^k$ and $\mathcal{X}_i^m \cap \mathcal{X}_i^l = \emptyset$, for all $m \neq l$. A given-data estimator of $\xi_i^L := \xi_{X_i}^L$ in (2.9) is given by:

$$\xi_i(\mathcal{K}_i) = \sum_{k=1}^{K} p\big(X_i \in \mathcal{X}_i^k\big)\zeta\big(\mathbf{p}_L,\mathbf{p}_{L|X_i\in\mathcal{X}_i^k}\big), \tag{2.10}$$

where $\mathbf{p}_{L|X_i\in\mathcal{X}_i^k} = \big\{p_r^L(\mathcal{X}_i^k) = p(L=\ell_r|X_i \in \mathcal{X}_i^k); r = 1,2,\dots,n_L\big\}$ denotes the conditional distribution of L, for $\mathcal{X}_i^k \in \mathcal{K}_i$. Now, consider a fixed partition $\mathcal{K}_i = \{\mathcal{X}_i^1,\mathcal{X}_i^2,\dots,\mathcal{X}_i^K\}$ with cardinality K and a dataset of features and target realizations. Let N be the sample size, and, for $r = 1,\dots,n_L$, let N_r the number of the observations labeled with ℓ_r. For $\mathcal{X}_i^k \in \mathcal{K}_i$, let $N(\mathcal{X}_i^k)$ denote the number of input observations in $\mathcal{X}_i^k$, and $N_r^L(\mathcal{X}_i^k)$ the corresponding number of target observations with label ℓ_r. Then, using the plug-in principle, we obtain an estimate of $\xi_i(\mathcal{K}_i)$ in (2.10) from

$$\widehat{\xi}_i(K,N) = \sum_{k=1}^{K} \widehat{p}\big(\mathcal{X}_i^k\big)\zeta\big(\widehat{\mathbf{p}}_L,\widehat{\mathbf{p}}_{L|X_i\in\mathcal{X}_i^k}\big), \tag{2.11}$$

where $\widehat{p}(\mathcal{X}_i^k)$, $\widehat{\mathbf{p}}_L$ and $\widehat{\mathbf{p}}_{L|X_i\in\mathcal{X}_i^k}$ are plug-in estimates of, respectively, $p(X_i \in \mathcal{X}_i^k)$, $\mathbf{p}_L$ and $\mathbf{p}_{L|X_i\in\mathcal{X}_i^k}$. Observing that, for $i = 1,2,\dots,n_X$ and $r = 1,2,\dots,n_L$, $\widehat{p}(X_i \in \mathcal{X}_i^k) = N(\mathcal{X}_i^k)/N$, $\widehat{p}(L=\ell_r) = N_r/N$ and $\widehat{\mathbf{p}}_L = \{\widehat{p}(L=\ell_1),\dots,\widehat{p}(L=\ell_{n_L})\}$, as well as $\widehat{p}_r^L(\mathcal{X}_i^k) = N_r^L(\mathcal{X}_i^k)/N(\mathcal{X}_i^k)$ and $\widehat{\mathbf{p}}_{L|X_i\in\mathcal{X}_i^k} = \{\widehat{p}_1^L(\mathcal{X}_i^k),\dots,\widehat{p}_{n_L}^L(\mathcal{X}_i^k)\}$, we can rewrite (2.11) as

$$\widehat{\xi}_i(K,N) = \sum_{k=1}^{K} \frac{N(\mathcal{X}_i^k)}{N}\zeta\big(\widehat{\mathbf{p}}_L,\widehat{\mathbf{p}}_{L|X_i\in\mathcal{X}_i^k}\big). \tag{2.12}$$

Proposition 2.3.6. *Let $\zeta(\cdot,\cdot)$ be a separation measure between probability mass functions, and let $\zeta(\cdot,\cdot)$ be continuous and bounded almost everywhere as X_i varies in $\mathscr{X}_i$. Let $\widehat{\xi}_i(K,N)$ be defined by* (2.12). *Then, if X_i is discrete,*

$$\lim_{N\to\infty}\widehat{\xi}_i(K,N)=\xi_i^L.$$

If X_i is continuous,

$$\lim_{K\to\infty}\lim_{N\to\infty}\widehat{\xi}_i(K,N)=\xi_i^L. \tag{2.13}$$

Proposition 2.3.6 reassures us of the consistency of the estimator of ξ_i^L. From an implementation viewpoint, we need to distinguish the case in which X_i is discrete or categorical from the case in which it is continuous. If X_i is discrete then the partition $\mathscr{K}_i$ is fixed and immediately given by the support of X_i. If X_i is continuous then the two limits with respect to the sample size and the partition cardinality are nested. Theoretically, first one lets the sample size N tend to infinity and then one refines the partitions sending K to infinity — as evidenced already in Pearson [1905]. In practice, care must be taken in selecting the partition size at any finite sample N. Proposition 2.3.6 also implies that for bounded metrics $\zeta(\cdot,\cdot)$, the variance of $\zeta(\widehat{\mathbf{p}}_L,\widehat{\mathbf{p}}_{L|X_i\in\mathscr{X}_i^k})$ is finite for every $k=1,2,\dots,K$. As an immediate consequence, the following holds for the separation measurements in Table 2.1 the following holds.

Corollary 2.3.7. *The estimators of $\zeta^1(\mathbf{p}_L,\mathbf{p}_{L|X_i\in\mathscr{X}_i^k})$, $\zeta^2(\mathbf{p}_L,\mathbf{p}_{L|X_i\in\mathscr{X}_i^k})$, and $\zeta^{KU}(\mathbf{p}_L,\mathbf{p}_{L|X_i\in\mathscr{X}_i^k})$ are asymptotically consistent.*

Consistency of the estimators has an immediate advantage also with regard to uncertainty quantification. In particular, a natural way of quantifying uncertainty in a context as the one we are dealing with is to make use of the bootstrap method. We consider the bootstrap bias-reducing estimator in the version of Efron and Gong [1983], $\widehat{\widehat{\xi}}_i(K,N)$. More precisely, consider a sample of size N and a partition $\mathscr{K}_i$ of K elements. Then, we have

$$\widehat{\widehat{\xi}}_i(K,N)=2\bar{\xi}_i(K,N)-\widehat{\xi}_i(K,N), \tag{2.14}$$

where $\bar{\xi}_i(K,N)$ is the estimate of $\xi_i(K,N)$ produced by taking the mean over the bootstrap replicates with fixed partition $\mathscr{K}_i$ and $\widehat{\xi}_i(K,N)$ is the corresponding point estimate. By the theory of the bootstrap method, the asymptotic consistency of $\widehat{\xi}_i(K,N)$ implies the asymptotic consistency of $\widehat{\widehat{\xi}}_i(K,N)$. In our case, the consistency of $\widehat{\xi}_i(K,N)$ is ensured by Proposition 2.3.6.

The selection of the partition cardinality K is a crucial step in implementing given-data estimators. The partition cardinality is, indeed, the sole hyperparameter of the design. It is well-known that this choice is associated with a bias-variance trade-off. On the one hand, the higher the value of K, the fewer the available realizations in each partition. We then have higher bias and lower variance for $\widehat{\mathbf{p}}_L$ and $\widehat{\mathbf{p}}_{L|X_i\in\mathscr{X}_i^k}$ for $k=1,\dots,K$. On the other hand, the smaller K is, the lower the bias but the higher the variance of the estimators.

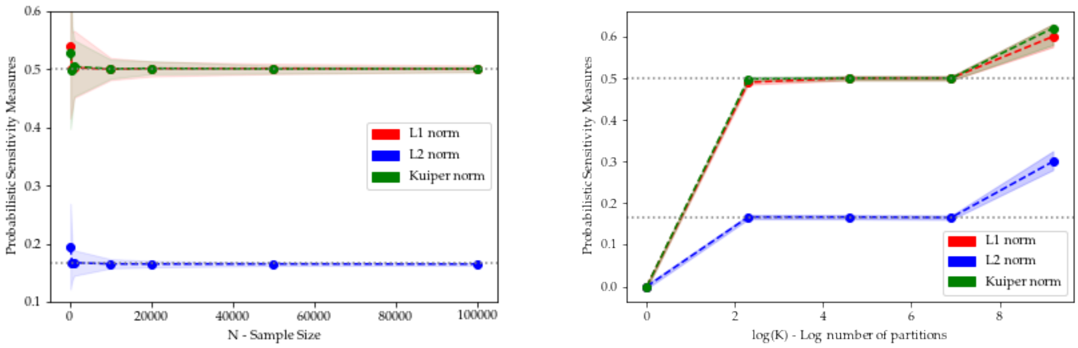
The choice of K in relationship to the sample size has been studied in depth in Strong and Oakley [2013]. The analysis in Strong and Oakley [2013] evidences a plateau effect: for large enough samples, the choice of K in a certain range does not impact the value of $\widehat{\xi}_i(K,N)$. To illustrate, consider Figure 1 in Strong and Oakley [2013, p. 759], that shows results of experiments at alternative values of K: fixing a sample of size $N = 10'000$, choosing K in a range between $K = 10$ and $K = 1'000$ yields similar values of $\widehat{\xi}_i(K,N)$. For larger (smaller) values of K (at the same sample of size $N = 10'000$), Strong and Oakley [2013] showcase an upward (downward) bias. The plateau effect can then be used to obtain guidance on the choice of K: Strong and Oakley [2013] recommend a value of K in the range that causes the plateau as a natural choice. To illustrate this aspect in our case, we report results of numerical experiments involving an analytical test case.

Example 2.3.8. *Consider the following fictitious binary classification problem, with a target random variable L with a binary support, i.e. $\mathscr{L} = \{0,1\}$. We simulate L using the following data generating process:*

$$L = \begin{cases} 0 & \text{if } 0 \le Y < 1, \\ 1 & \text{if } 1 \le Y \le 2, \end{cases}$$

where $Y = X_1 + X_2$, and X_1, X_2 are two independent and uniformly distributed random variables in [0,1], i.e. $X_1, X_2 \sim \mathrm{Unif}(0,1)$, here playing the role of covariates. Computing explicitly the true probability mass function for L on its support $\mathscr{L} = \{0,1\}$, we obtain $p(L=1) = 1 - p(L=0) = \frac{1}{2}$. Applying the definition in Equation (2.9), the analytical values of the probabilistic sensitivity measures based on the $1-$norm, $2-$norm and the Kuiper distance are respectively obtained computing the integrals $\xi_i^1 = \int_0^1 |1-2x_i| dx_i = \frac{1}{2}$, $\xi_i^2 = \int_0^1 (2x_i^2 - 2x_i + \frac{1}{2}) dx_i = \frac{1}{6}$, and $\xi_i^{KU} = \int_0^1 |1-2x_i| dx_i = \frac{1}{2}$, for $i = 1,2$. In Figure 2.1a, we report results at increasing sample sizes. Using a Sobol′ quasi-random sequence generator we produce a sequence of datasets with realizations of (X_1, X_2) and L, for the sample size ranging from $N = 100$ to $N = 100'000$. Using such a sequence allows easy reproducibility of this experiment. The shadows represent bootstrap confidence intervals. The estimates converge towards the analytical values as N increases, and the width of the bootstrap confidence intervals shrinks, thus confirming the asymptotic consistency of the estimates. Figure 2.1b presents results for a fixed sample size ($N = 10'000$) but selecting alternative partitions of increasing cardinality. In particular, we vary K from $K = 1$ to $K = 10'000$. Figure 2.1b shows that we register a plateau in the values of the estimates $\widehat{\xi}_i(K,N)$ for $K \in [20, 900]$. For values of K smaller than 20, we have a downward bias. Conversely, for values of K exceeding 900, we register an upward bias.

The results in Figure 2.1b are in agreement with the findings in Strong and Oakley [2013]. Moreover, as discussed also in Borgonovo et al. [2016], the downward bias is empirically explained by the fact that as K gets smaller the conditional and unconditional distributions tend to coincide. In fact, in the limiting case $K = 1$ the conditional and unconditional distributions are the same and the value of any measures of statistical association is null. Conversely, choosing $K = N$, we obtain exactly one point per partition. Then, the numerical calculation is between the marginal distribution of Y and a Dirac-δ mass cen-



(a) Bootstrap results as the sample size increases from $N = 10^2$ to $N = 10^5$.

(b) Bootstrap results as the partition size increases from $K = 1$ to $K = N$.

Figure 2.1: Bootstrap results varying N and K for $\widehat{\xi}_1^{\mathrm{KU}}(K,N)$, $\widehat{\xi}_1^{1}(K,N)$ and $\widehat{\xi}_1^{2}(K,N)$. The shaded areas represent 95%−bootstrap confidence intervals. In both Figures 2.1a and 2.1b, the y-axis represents the boostrap mean values of $\widehat{\xi}_1^{L}(K,N)$ based on the 1-norm (red), 2-norm (blue) and Kuiper norm (blue). In Figure 2.1a the x-axis reports the values of the sample size N, in Figure 2.1b the x-axis reports the values of the partition cardinality K on a logarithmic scale.

tered at the sole realization in the partition. Such distance is maximal or infinite for several metrics and therefore we register an upward bias.

Finally, any given-data estimator in Equation (2.11) is exposed to the curse of dimensionality when $X_i \in \mathbf{X}$ is multidimensional, that is, we are determining the joint importance of two or more features. In general, then, $X_i = (X_1, X_2 \ldots, X_s)$, $s \leq n_X$ requires us to condition on s features. We are then dealing with s-dimensional partition sets (to fix ideas, in one dimension we are dealing with intervals, in two dimension with rectangles, in s dimension with hyper-rectangles). These partition sets contain a number of realizations (data) that decreases exponentially with s. To illustrate, start with $s = 1$ and consider that we have available N realizations and a partition with cardinality K. Then, we can count on $J = N/K$ realizations per partition set (assuming equipopulated partition sets). For instance if $N = 10{,}000$ and $K = 25$, we have 400 data points per partition set. If we consider bi-dimensional partitions to find the joint importance of, say, features X_i and X_j $(i \neq j)$, then we could count on $J = N/(K_i K_j)$, where K_i and K_j are, respectively, the cardinalities of the partitions of the supports of X_i and X_j. At $N = 10{,}000$ selecting $K_i = K_j = 25$, we would have 16 data points per partition set. Compared to the previously available 400 points per partition set, this number of realizations is drastically lower and may prohibit an accurate estimation of the requested statistical quantities. However, this sparsity effect becomes noticeable when we increase the number of features (s) in the group, and is not related to the overall number of features n_X.

2.3.4 Getting explanations: the Xi method

In this section, we introduce a framework for obtaining pre hoc and post hoc explanations for ML models through measures of statistical dependence. The idea is to understand how close the values of explanations computed from the data are to those computed from the forecasts. Performing this analysis at the overall dataset level as well as at the level of each individual class leads to several indications about the features that are statistically important in the problem at hand. Starting with the data collection $T_L = \left\{(\mathbf{x}^{(n)}, L^{(n)}); n = 1,2,\ldots,N\right\}$, *dataset explanations* are defined as the collection of probabilistic sensitivity measures estimates:

$$\widehat{\xi}_X^L = \left\{\widehat{\xi}_i^L, i = 1,2,\ldots,n_X\right\}, \tag{2.15}$$

estimated from T_L by applying (2.11) (the superscript L denotes that the estimates are obtained using the data, i.e., $\widehat{\xi}_i^L$ is an estimate of $\xi_i^L = \mathbb{E}_X[\zeta(\mathbf{p}_L, \mathbf{p}_{L|X_i})]$). Consider now the dataset $T_\Lambda = \{(\mathbf{x}^{(n)}, \Lambda^{(n)}); n = 1,2,\ldots,N\}$, with $\Lambda^{(n)} = \widehat{g}(\mathbf{x}^{(n)})$, obtained from T_L by replacing the true labels with the corresponding ML model predictions. *Prediction explanations* are defined as the collection of probabilistic sensitivity measures

$$\widehat{\xi}_X^\Lambda = \{\widehat{\xi}_i^\Lambda, i = 1,2,\ldots,n_X\}, \tag{2.16}$$

with $\widehat{\xi}_i^\Lambda$ defined in Equation (2.9) and estimated by applying Equation (2.11) to T_Λ. Since they look at all the target classes simultaneously, $\widehat{\xi}_X^L$ and $\widehat{\xi}_X^\Lambda$ yield an understanding of the most important covariates for the overall dataset or prediction task. However, we can make the analysis more granular by analyzing what are the statistically most important features to predict a given label. To do this, we resort to the so-called one-hot encoding technique for the response variable [Hastie et al., 2009]. In particular, for n_L different labels $\ell_1, \ldots, \ell_{n_L}$, we codify the data by $L_1 = \mathbb{1}_{\{L=\ell_1\}}, \ldots, L_{n_L} = \mathbb{1}_{\{L=\ell_{n_L}\}}$. Then, we call the *$\ell_r$-dataset explanations* the quantities $\widehat{\xi}_X^{L_r} = \left\{\widehat{\xi}_i^{L_r}, i = 1,2,\ldots,n_X\right\}$, where $\widehat{\xi}_i^{L_r}$ is the estimator of $\xi_i^{L_r} = \mathbb{E}\left[\zeta(\mathbf{p}_{L_r}, \mathbf{p}_{L_r|X_i})\right], r = 1,\ldots,n_L$. Similarly, to answer the question of *what are the statistically important features for the ML model when predicting label ℓ_r*, we define the *ℓ_r-prediction explanations* as $\widehat{\xi}_X^{\Lambda_r} = \left\{\widehat{\xi}_i^{\Lambda_r}, i = 1,2,\ldots,n_X\right\}$ with $\xi_i^{\Lambda_r} = \mathbb{E}\left[\zeta(\mathbf{p}_{\Lambda_r}, \mathbf{p}_{\Lambda_r|X_i})\right]$ estimated by $\widehat{\xi}_i^{\Lambda_r}$, and Λ_r defined in a similar fashion as L_r.

The next step is to compare $\widehat{\xi}_X^\Lambda$ and $\widehat{\xi}_X^L$ (or their prediction counterparts $\widehat{\xi}_i^{L_r}, \widehat{\xi}_i^{\Lambda_r}$). This comparison can be done by a simple graphical visualization of the values of $\widehat{\xi}_X^\Lambda$ and $\widehat{\xi}_X^L$ or of the ranking they induce. For this task, one can use any discrepancy measure between vectors of real numbers. A possible choice is the Minkowski distance:

$$D^p\left(\widehat{\xi}_X^\Lambda, \widehat{\xi}_X^L\right) = \left\|\widehat{\xi}_X^\Lambda - \widehat{\xi}_X^L\right\|_p = \sqrt[p]{\frac{1}{n_X}\sum_{i=1}^{n_X} |\widehat{\xi}_i^\Lambda - \widehat{\xi}_i^L|^p}. \tag{2.17}$$

For $p = 1$, $D^p(\widehat{\xi}^{\Lambda}_X, \widehat{\xi}^{L}_X)$ is the Mean Absolute Deviation (MAD), while for $p = 2$ it is the Mean Squared Error (MSE). Information delivered by $D^p\left(\widehat{\xi}^{\Lambda}_X, \widehat{\xi}^{L}_X\right)$ can be used in alternative ways. For instance, the analyst may consider two sets of explanations $\widehat{\xi}^{\Lambda}_X$ and $\widehat{\xi}^{L}_X$ far apart if $D^p\left(\widehat{\xi}^{\Lambda}_X, \widehat{\xi}^{L}_X\right) > \delta$, for some threshold $\delta > 0$. If that is the case, and the model is fitting well, then the model is making good predictions but it is not picking up the same statistical relationships present in the data. Conversely, if $D^p\left(\widehat{\xi}^{\Lambda}_X, \widehat{\xi}^{L}_X\right) < \delta$, the model is predicting accurately and its forecasts recreate well the statistical dependence of the original dataset. In this respect, it is interesting to conduct the comparison for the training as well as the test datasets, to see if differences emerge between the two subsamples. Regarding the value of the threshold δ, the choice depends on the application of interest and on the separation measure $\zeta(\cdot,\cdot)$; we therefore refrain from providing a general rule. Also, often the interest is on the ordinal ranking induced by $\widehat{\xi}^{\Lambda}_X$ and $\widehat{\xi}^{L}_X$. In this case, a quantitative comparison is carried out using well-known statistical techniques such as the Spearman rank correlation coefficient (ρ^{Spear}, henceforth) [Spearman, 1904]. This very same procedure can be employed in a very similar fashion at the individual class level, i.e. when the sensitivity measures are calculated for each target class.

2.4 Test cases: tabular, visual and textual data

In this section, we illustrate the method and apply it to well-known datasets in different formats: in Subsection 2.4.1 we analyze a tabular dataset, in Subsection 2.4.2 an image dataset and in Subsection 2.4.3 a text dataset. In all the experiments, we train a ML model on a subset of the available data (the training set), and then we compute the explanations ξ^{Λ}_X, ξ^{L}_X on the test set, using the estimators defined in Equation (2.10) and fixing the number of partitions at $K = 10$. To fit the models, we employ the well-known `scikit-learn` package in Python, using the default loss functions. We report the ML model test accuracy as the percentage of correctly classified instances.

2.4.1 Tabular data: the wine dataset

In our first application, we use the well-known Wine dataset publicly available at `https://archive.ics.uci.edu/ml`. This dataset has been widely used in association with the task of predicting a wine quality based on its chemical properties [Dua and Graff, 2017]. The dataset collects quantitative data on eleven features (ranging from alcohol content to the pH of the wine), and the output is the corresponding wine quality measured on a scale from 1 to 10. The sample size is $N = 4898$, split into 60% for training and 40% for testing. The only preprocessing has been to replace the 39 missing entries in the design matrix $\mathbf{X}$ with either the mean or the median of the corresponding variable (we also performed our analysis omitting these entries, with unchanged results). The classes in the final dataset are unevenly populated, with 1457, 2198 and 880 entries, respectively, for Qualities 5, 6 and 7, with 163 and 175 entries for Qualities 4 and 8, with only 20 entries for Quality 3, and zero entries for Quality 1 and 2. We trained alternative ML models and registered a very similar accuracy for all. In the remainder, we shall focus on results obtained using a Random Forest (RF), which registers a test accuracy of 66% (1-off accuracy ~ 90%). Using the test set and the trained model, for all the variables $i = 1, \ldots, n_X$, we estimate the feature importance measures provided by Permutation Importance (PI_i) and Split and Count (SC_i), as well as the probabilistic sensitivity measures $\widehat{\xi}_i^L$, $\widehat{\xi}_i^\Lambda$ based on the 1-norm separation (for the sake of space we shall focus only on this norm in the reminder). Results are displayed in Figure 2.2. Panels (a) and (b) in Figure 2.2 compare the values of the dataset versus prediction explanations (both computed on the test set). A visual inspection suggests that the ranking induced by $\widehat{\xi}_X^L$ and $\widehat{\xi}_X^\Lambda$ are similar: alcohol stands out as the most important variable, the group volatile acidity, density and chlorides follows, and is followed in turn by a group comprising citric acid, total sulfur dioxide, free sulfur dioxide and sulphates , while the group pH, fixed acidity and residual sugar contains the three least relevant features. Table 2.2 reports results of the quantitative comparisons between $\widehat{\xi}_X^L, \widehat{\xi}_X^\Lambda$ with MAD and MSE in the second and third columns, respectively, and ρ^{Spear} in the last column. The small

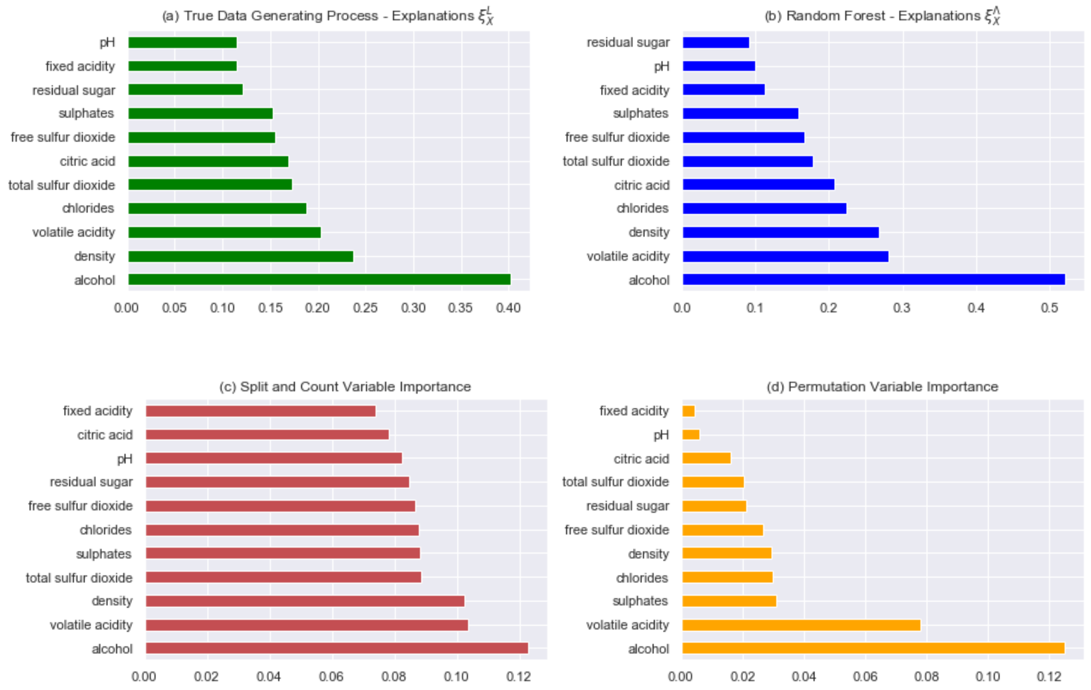


Figure 2.2: Panel 2.2a: dataset explanations $\widehat{\xi}_i^L$ based on the 1-norm separation. Panel 2.2b: prediction explanations $\widehat{\xi}_i^\Lambda$ based on the 1-norm separation. Panel 2.2c: Split and Count measure $\widehat{\mathrm{SC}}_i$, Panel 2.2d: Permutation Importance measure $\widehat{\mathrm{PI}}_i$, for $i = 1, \ldots, n_X$ — tabular data.

values of MAD and MSE and the simultaneously high value of ρ^{Spear} confirm the visual impression of Figure 2.2 concerning the overall agreement. Thus, the ML model forecasts actually reproduce well the original statistical dependence in the data and the covariates that are statistically important for the true data generating process are also important for the ML model predictions. Panels (c) and (d) in Figure 2.2 compare these results with indications provided by the Split and Count measure (SC) and the Permutation Importance measure (PI). The values of ρ^{Spear} between the ranking induced by the alternative importance measures amount at around 0.7, indicating a lower ranking agreement than in the previous case (Table 2.3). Indeed, while all importance measures agree on alcohol as the most important feature, the values of SC indicate a rather homogeneous influence of the remaining features (the values of SC are similar), while $\widehat{\xi}_X^{\Lambda}$ and $\widehat{\xi}_X^{L}$ display greater variability in their values, with alcohol, density and volatile acidity being statistically more important than pH, fixed acidity and residual sugar. One also notes that the ranking of the three most important variables is consistent for $\widehat{\xi}_X^{L}$, $\widehat{\xi}_X^{\Lambda}$ and SC, while PI assigns density a much lower importance. All importance measures suggest alcohol as the most important variable, so that this feature is not only the feature on which the target depends most strongly, but also the one that drives the ML model performance the most. Let us now examine results at the individual class level. We apply one-hot encoding to the model forecasts for the test set, with $N^{Te} = 2573$. Estimates of $\xi_i^{\Lambda_r}$ for $i = 1, \ldots, n_X$ and $r = 1, \ldots, n_L$ based on $\zeta^1(\cdot,\cdot)$ are reported in Figure 2.3. Each group of bars displays estimates of $\xi_i^{\Lambda_r}$ for a given class, from Quality 4 to Quality 8, as the model never predicts Quality 3 or Quality 9 and thus the bars would be all null. Similarly, we register extremely low values of $\xi_i^{\Lambda_r}$ for Quality 4 and

Table 2.2: Quantitative similarity between dataset explanations $\widehat{\xi}_X^{L}$ and prediction explanations $\widehat{\xi}_X^{\Lambda}$ based on the 1-norm separation — tabular data.

	MAD	MSE	ρ^{Spear}
1-norm	0.034	0.048	0.955
2-norm	0.010	0.017	0.986
Kuiper	0.040	0.055	0.936

Table 2.3: Spearman correlation coefficient comparing the ranking of dataset explanations $\widehat{\xi}_X^{L}$, prediction explanations $\widehat{\xi}_X^{\Lambda}$ (both based on the 1-norm separation), Split and Count (SC) measure, and Permutation Importance (PI) — tabular data.

Spearman Correlation	$\rho^{\text{Spear}}(\text{SC}, \widehat{\xi}_X^{L})$	$\rho^{\text{Spear}}(\text{SC}, \widehat{\xi}_X^{\Lambda})$	$\rho^{\text{Spear}}(\text{PI}, \widehat{\xi}_X^{L})$	$\rho^{\text{Spear}}(\text{PI}, \widehat{\xi}_X^{\Lambda})$
1-norm	0.809	0.709	0.791	0.709
2-norm	0.745	0.736	0.718	0.736
Kuiper	0.745	0.709	0.727	0.709

Quality 8. This is a consequence of the few predictions of the corresponding labels, with Quality 4 and 8 predicted, respectively, only 12 and 82 times on over 2500 entries. As a result, the one-hot encoding vectors are, effectively, vector of zeros independently of the values of the features. This effect is then captured by the values of the probabilistic sensitivity measures that are close to zero. The remaining bars indicate that alcohol is, statistically, the most important feature for the model when making a prediction on classes Quality 5, Quality 6 and Quality 7, followed by sulphates; fixed acidity raises in importance when the target is Quality 5 or Quality 6, while total sulfur dioxide becomes the third most important variable for predicting Quality 7.

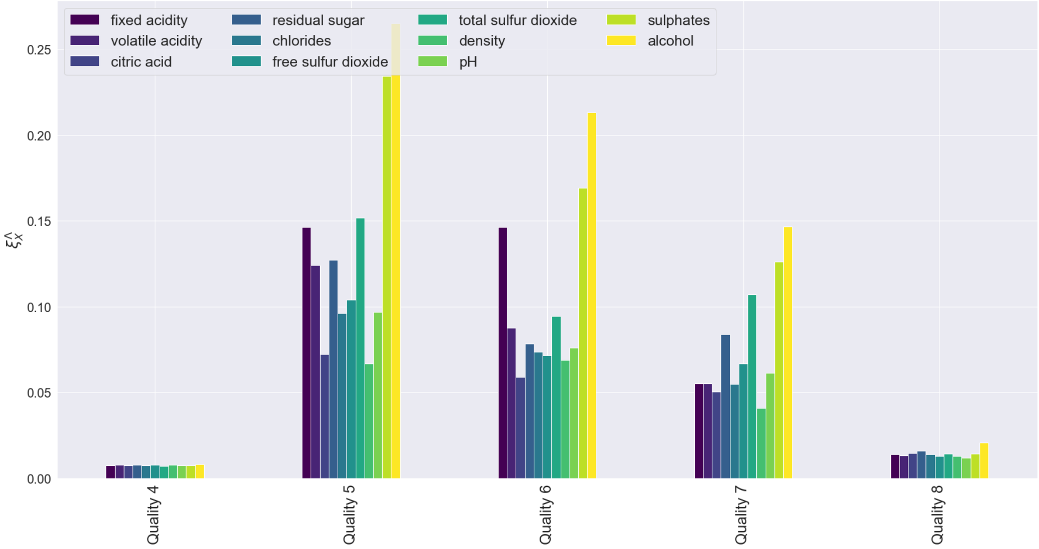


Figure 2.3: ℓ_r-prediction explanations $\widehat{\xi}^{\Lambda_r}_X$ based on the 1-norm separation, $r = 4, \ldots 8$ — tabular data.

In the previous discussion, we have referred to point estimates. However, we also performed corresponding uncertainty quantification to understand whether such observations are robust to variability in the estimates. We report a first set of results obtained by 2'500 bootstrap replicates of the test data and predictions, and for each sample we recompute $\widehat{\xi}^{L}_X$ as well as $\widehat{\xi}^{\Lambda}_X$. Figure 2.4 reports the corresponding bootstrap distributions as boxplots. The bootstrap confidence intervals in Figure 2.2a and 2.2b are narrow enough around the point estimates to let us state that the results obtained with point estimates are actually reliable, with a few (and negligible) outliers in the boxplot. Thus, in this case, uncertainty quantification does not alter the previous considerations.

2.4.2 Image data: fashion MNIST

In this section, we report results of experiments conducted on the well-known Fashion MNIST dataset [Xiao et al., 2017], a database by Zalando research containing images of

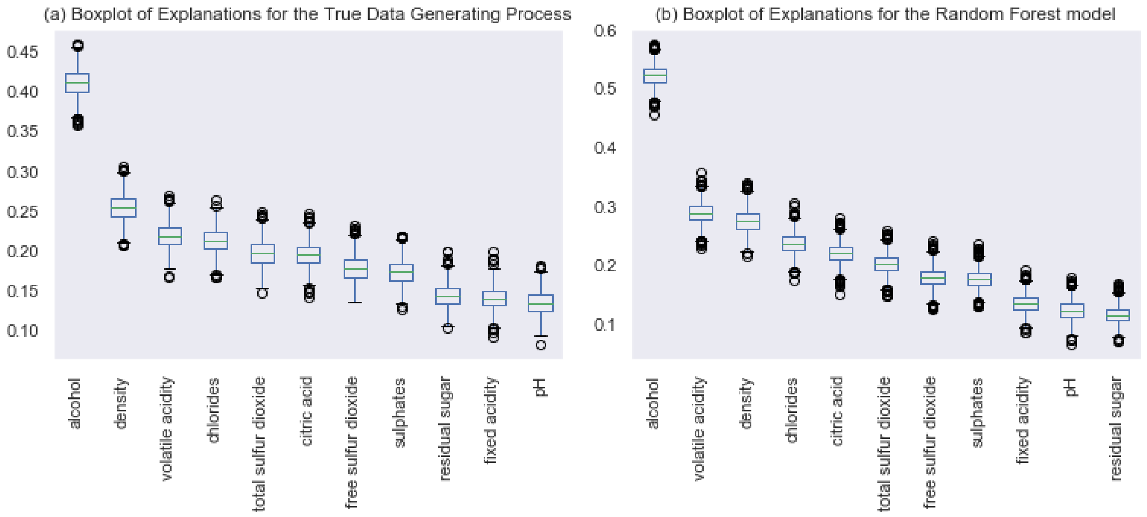


Figure 2.4: Panel 2a: bootstrap distributions of dataset explanations $\widehat{\xi}^{L}_{X}$. Panel 2b: bootstrap distributions of prediction explanations $\widehat{\xi}^{\Lambda}_{X}$ (both based on the 1-norm separation) — tabular data.

clothing articles, with N=70’000, of which 60’000 images are used for training and 10’000 for testing. The images are made of 28×28 grayscale pixels; each instance is associated with a label from ten classes: T-shirt/Top, Trouser, Pullover, Dress, Coat, Sandal, Shirt, Sneaker, Bag, Ankle boot. The ML model in this application is a pre-trained convolutional neural network with the LeNet architecture [LeCun et al., 1998]. LeNet is a 7-layer neural network consisting of the input layer, 2 convolutional layers each followed by a pooling layer and another convolutional layer followed by the output layer. The overall accuracy is 94%.

In this experiment, we focus on the explanations provided by the Xi method using the 1-norm separation measure. Figure 2.5 reports results of the investigation at the all-classes level. The first heatmap displays values of $\widehat{\xi}^{L}_{X}$, with lighter pixels being more important: to estimate the explanations, we use the entire dataset of 70’000 images. The second heatmap refers to dataset explanations $\widehat{\xi}^{L}_{X}$ on the test set, and the third one displays prediction explanations $\widehat{\xi}^{\Lambda}_{X}$ computed on the test set. A visual inspection shows a great similarity between the regions that are statistically important for the LeNet model and for the true data generating process. Interestingly, the exact center of the image does not contain the most important pixels, which are the ones surrounding the object (images are centred). Table 2.4 reports results for the quantitative comparison between $\widehat{\xi}^{L}_{X}$ and $\widehat{\xi}^{\Lambda}_{X}$: MAD and MSE are close to 0, and ρ^{Spear} is almost 1, indicating a high agreement between $\widehat{\xi}^{L}_{X}$ and $\widehat{\xi}^{\Lambda}_{X}$. Thus, the statistical dependence in the true data generating process is maintained in the forecasts.

Consider now the analysis at the individual class level. Figure 2.6 reports the heatmaps generated by $\widehat{\xi}^{\Lambda_r}_{X}$ for each of the image classes $r = 1, \ldots, n_L$ using the test data (the heatmaps

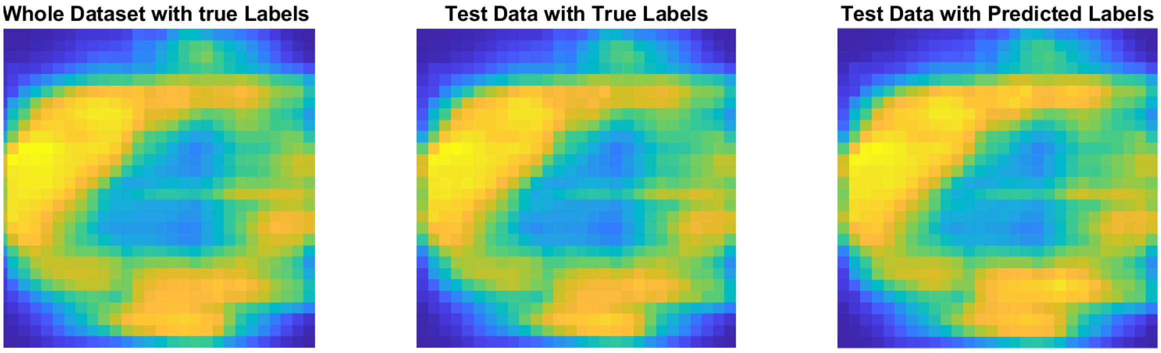


Figure 2.5: From left to right: dataset explanations $\widehat{\xi}_X^L$ (computed respectively using the whole dataset and the test set) and prediction explanations $\widehat{\xi}_X^\Lambda$, both based on the 1-norm separation separation — image data.

Table 2.4: Quantitative similarity between the dataset explanations $\widehat{\xi}_X^L$ and the prediction explanations $\widehat{\xi}_X^\Lambda$ based on the 1-norm separation — image data.

	MAD	MSE	ρ^{Spear}
1-norm	0.024	0.030	0.997
2-norm	0.004	0.005	0.994
Kuiper	0.009	0.011	0.993

of $\widehat{\xi}_X^{L_r}$ are similar and thus not displayed). The explanations at the individual-class level for the LeNet model show interesting insights: for example, the most important pixels for predicting the class T-shirt/Top highlight the lack of sleeve in such items; the most important ones for predicting trousers are the pixels that signal the empty space between the legs; for classes Dress or Shirt the focus is on the overall clothing piece (light blue pixels), with specific parts that are more important than others (yellow pixels). Also for this case study we performed an uncertainty quantification via the bootstrap. Results show stability in the estimates (details are not reported).

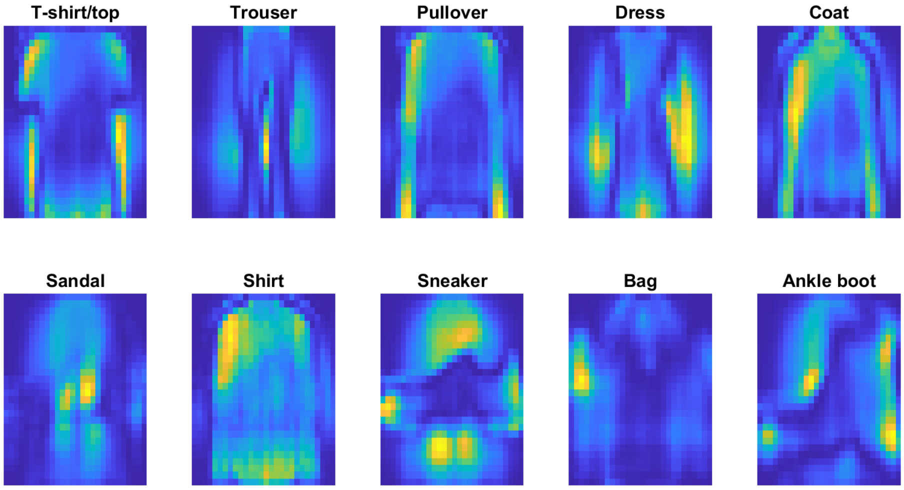
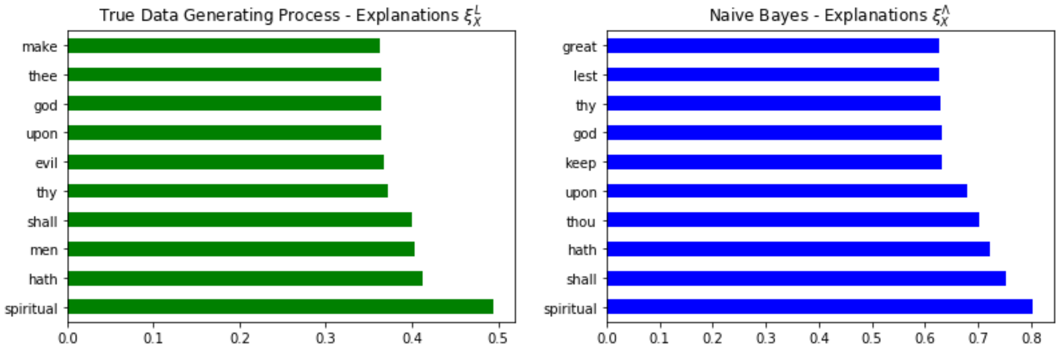


Figure 2.6: ℓ_r-prediction explanations $\widehat{\xi}_X^{\Lambda_r}$ based on the 1-norm separation, $r = 1,\ldots,n_L$ — image data.

2.4.3 Textual data: asian religious texts

The last application concerns the textual database available at `https://archive.ics.uci.edu/ml/datasets/A+study+of++Asian+Religious+and+Biblical+Texts`. In this case, the task is to predict the book of origin of an excerpt among eight sacred texts [Sah and Fokoué, 2019]: Book Of Ecclesiasticus, Book Of Ecclesiastes, Book Of Proverb, Book Of Wisdom (Christians' sacred books), Buddhism, Tao Te Ching (sacred text for Taoism), Upanishad and Yoga Sutra (sacred books for Hinduism), related to different religions in Asia (Hinduism, Buddhism, Taoism, Christianity). Features are the individual words in the Document Term Matrix (DTM) used to train the model, describing the frequency of terms that occur in the collection of documents. The classifier is a simple Naive Bayes, trained on 70% of the observations. The ML model test accuracy is 65%. As in the previous experiments, we estimate the probabilistic sensitivity measures based on the 1-norm using the test dataset. Figure 2.7 reports the ten most important words according to dataset explanations $\widehat{\xi}^L_X$ and prediction explanations $\widehat{\xi}^\Lambda_X$. It shows that while there is some difference in the actual values of the explanations, the rankings provided by both $\widehat{\xi}^L_X$ and $\widehat{\xi}^\Lambda_X$ are similar.



(a) (b)

Figure 2.7: Panel 2.7a: dataset explanations $\widehat{\xi}^L_X$ based on the 1-norm separation, Panel 2.7b: prediction explanations $\widehat{\xi}^\Lambda_X$ based on the 1-norm separation (ten most important features displayed) — text data.

Table 2.5: Quantitative similarity between the dataset explanations $\widehat{\xi}^L_X$ and the prediction explanations $\widehat{\xi}^\Lambda_X$ based on the 1-norm separation — text data.

	MAD	MSE	ρ^{Spear}
1-norm	0.030	0.049	0.923
2-norm	0.006	0.010	0.908
Kuiper	0.009	0.017	0.886

The quantitative comparison in Table 2.5 shows a value of ρ^{Spear} at about 90%; thus, we register a high overall ranking agreement, although the induced ranking are not perfectly coincident; at the same time values of the MAD and MSE between $\widehat{\xi}^L_X, \widehat{\xi}^\Lambda_X$ are close to 0. This last result shows that the values of the probabilistic sensitivity measures calculated on the data and on the forecasts are close on average; thus, small differences may result in different ranking, in spite of the features having a similar influence. At the individual class level, results are presented in Table 2.6. The columns in Table 2.6 list the ten most important words for each target class according to $\widehat{\xi}^{\Lambda_r}_X$, $r = 1, \ldots, n_L$. We note that the words that matter for the classification into a Hinduism sacred text (Upanishad, YogaSutra) mostly pertain to spiritualism and knowledge (*spiritual*, *wise*, *knows*, *teaching*) as well as with some specific

Table 2.6: Ten most important words provided by the ℓ_r-prediction explanations $\xi_i^{\Lambda_r}$ based on the 1-norm separation, $r = 1,\ldots,n_L$ — text data.

Book Of Ecclestasticus	Book Of Ecclesiastes	Book Of Proverbs	Book Of Wisdom	Buddhism	Tao Te Ching	Upanishad	Yoga Sutra
shall	mortal	shall	therefore	hath	tao	brahman	spiritual
s	maketh	hath	consciousness	thy	young	called	great
lord	souls	like	form	right	never	whole	wise
heart	flesh	life	psychic	qualities	together	last	aged
god	treasure	without	though	therefore	person	devas	forgotten
great	silver	loveth	life	bring	appeared	knows	declareth
thee	spirit	soul	perception	made	looked	enters	approved
thou	born	glory	vision	come	perhaps	definite	wherewith
truth	know	poor	another	make	knows	teaching	number
knowledge	tongue	keepeth	know	nature	root	things	rites

terms of Hindu philosophy (*Brahman, devas*). On the other hand, terms more related to *nature* appear in Buddhism. The most impactful word to predict the Tao Te Ching sacred text is, unsurprisingly, *Tao*. Finally, the predictions regarding the Christian sacred texts are more influenced by terms dealing with the contraposition of mortal (*flesh*,*mortal*, *born*) and eternal (*souls, spirit, form*) life as well as hints to *god* and *lord*. We performed an uncertainty quantification via bootstrap. Estimates remain stable (results not shown). For this test case, we also computed Breiman's Permutation Variable Importance (PI), however obtaining a null value for all the features. We believe that this result can be explained by two main reasons: first, the high dimensionality of the DTM, which contains roughly 8′300 columns. Permuting one feature out of so many does not result in a significant change in the loss function. Second, the DTM is highly sparse, with feature entries consisting mostly of zeros. Then, the permutation leads to a basically identical vector of realizations, and does not yield any particular change in the loss functions.

2.5 Discussion and future research directions

This work has introduced probabilistic sensitivity measures that are well-posed on unordered data and has applied them to ML classification tasks. The proposed measures of association are based on the separation between probability mass functions and can be estimated from a given dataset, without the need of fitting a ML model. They possess the zero-independence property. Also, we have proven that, provided that the indices are based on a bounded and continuous separation measurement between probability mass functions, the corresponding estimators are asymptotically consistent.

We have then proposed a framework to explain relationships of statistical dependence in a classification context. A key part of the method is the comparison of measures of statistical association calculated first on the original data and then on the ML model forecasts. The first set of estimates uncovers the target-features dependence of the data generating process, the second the target-feature dependence that emerges when the ML model forecasts replace the original targets. The framework, here called Xi method, offers several advantages: the importance measures are not computationally expensive, the explanations can be obtained for any kind of data (images, texts, tabular) and they come with theoretical guarantees. Also, when images are concerned, the sensitivity measures avoid data manipulations such as obscuring or removing pixels. This is a major advantage because it is well-known that results may change according to, e.g., the color used to mask portions of the image. Moreover, even if this work has focused on classification, the Xi method is applicable in a regression framework as well. In this case, measures of statistical association such as the one in Equation (2.8) (i.e., the Chatterjee correlation coefficient) and several others are available to be selected as statistical indicators (the measures introduced in this work are aimed at classification problems). A further advantage of the approach is that it allows a straightforward and computationally cheap uncertainty quantification, in line with the recommendation of, among others, Dunson [2018]. As for any method based on measures of statistical association, a first limitation is that if two features X_i and X_j are highly correlated, measures of pairwise dependence between target and features will assign similar importance to X_i and X_j, even if Y is a function of X_i only. In principle, this issue can be overcome by conditioning on one or more of the correlated features and estimating conditional measures of pairwise dependence. The estimation procedure may also suffer from the curse of dimensionality or of lack of data in the case of small sample sizes. Investigations aimed at studying these probabilistic sensitivity measures in a conditional setting as well as the extension of the method to multivariate forecasting problems are future research avenues.

Appendix

Proofs

Proof of Proposition 2.3.6. **1) Discrete input case**. Consider X_i a discrete variable with support $\mathcal{X}_i = \{x_i^1, x_i^2, \dots, x_i^K\}$. In this case, the natural partition choice is $\mathcal{K}_i = \{\mathcal{X}_i^1, \mathcal{X}_i^2, \dots, \mathcal{X}_i^K\}$, with $\mathcal{X}_i^k = \{X_i : X_i = x_i^k\}$, for $k = 1, 2, \dots, K$. Note that $\xi_i^L = \xi_i(\mathcal{X}_i) = \xi_i(\mathcal{K}_i)$ becomes

$$\xi_i^L = \xi_i(\mathcal{K}_i) = \sum_{k=1}^{K} p(X_i = x_i^k)\zeta(\mathbf{p}_L, \mathbf{p}_{L|X_i=x_i^k}). \tag{18}$$

Then, a given-data estimator for $\xi_i(\mathcal{K}_i)$ in Equation (18) can be written as

$$\widehat{\xi}_i(K,N) = \sum_{k=1}^{K} \widehat{p}\big(X_i = x_i^k\big)\zeta\big(\widehat{\mathbf{p}}_L, \widehat{\mathbf{p}}_{L|X_i=x_i^k}\big). \tag{19}$$

Then, consider a dataset $(\mathbf{X},L)$. Let N_i^k denote the number of realizations of X_i such that $X_i \in \mathscr{X}_i^k$ — i.e. $X_i = x_i^k$, for $k = 1,2,\dots,K$. Then, $\widehat{p}(X_i = x_i^k) = N^{-1}N_k$ is a consistent estimator of $p(X_i = x_i^k)$, by the law of large numbers. Similarly, letting N_r^L be the number of labels in category ℓ_r, $r = 1,2,\dots,n_L$, $\widehat{p}(L = \ell_r) = N^{-1}N_r^L$ is a consistent estimator of $p(L = \ell_r)$. Analogously, let $\widehat{p}(L = \ell_r|X_i = x_i^k) = (N_i^k)^{-1}N_r^L(\mathscr{X}_i^k)$, where $N_r^L(\mathscr{X}_i^k)$ counts the realizations of L equal to ℓ_r, when $X_i \in \mathscr{X}_i^k$. Then, this estimator is also consistent by the law of large numbers. Therefore $\widehat{\mathbf{p}}_L \to \mathbf{p}_L$ and $\widehat{\mathbf{p}}_{L|X_i=x_i^k} \to \mathbf{p}_{L|X_i=x_i^k}$ as $N \to \infty$. Now $\zeta(\widehat{\mathbf{p}}_L, \widehat{\mathbf{p}}_{L|X_i=x_i^k})$ is continuous, and, therefore, we have

$$\zeta\big(\widehat{\mathbf{p}}_L, \widehat{\mathbf{p}}_{L|X_i=x_i^k}\big) \underset{N\to\infty}{\longrightarrow} \zeta(\mathbf{p}_L, \mathbf{p}_{L|X_i=x_i^k}),$$

so that

$$\sum_{k=1}^{K} \widehat{p}\big(X_i = x_i^k\big)\zeta\big(\widehat{\mathbf{p}}_L, \widehat{\mathbf{p}}_{L|X_i=x_i^k}\big) \underset{N\to\infty}{\longrightarrow} \sum_{k=1}^{K} p\big(X_i = x_i^k\big)\zeta(\mathbf{p}_L, \mathbf{p}_{L|X_i=x_i^k}), \tag{20}$$

which implies $\widehat{\xi}_i(K,N) \to \xi_i^L$ as $N \to \infty$.

2) Absolutely continuous input case. If X_i is absolutely continuous, let $f_{X_i}(x_i)$ be the density of X_i. Consider now $\xi_i^L = \xi_i(\mathscr{X}_i)$ written as

$$\xi_i(\mathscr{X}_i) = \int\limits_{\mathscr{X}_i} \zeta\big(\mathbf{p}_L, \mathbf{p}_{L|X_i=x_i}\big) f_{X_i}(x_i)dx_i.$$

First, we note that the integral above is the limit, if it exists, of the following Riemann-Stieltjes sum:

$$\xi_i(\mathscr{X}_i) = \lim_{\delta\to 0} \xi_i(\mathscr{K}_i^\delta) = \lim_{\delta\to 0} \sum_{k=1}^{K(\delta)} \zeta\big(\mathbf{p}_L, \mathbf{p}_{L|X_i\in\mathscr{X}_i^k(\delta)}\big) p\big(X_i \in \mathscr{X}_i^k(\delta)\big),$$

where the set $\mathscr{X}_i^k$ is a member of a partition $\mathscr{K}_i^\delta$ of $\mathscr{X}_i$, and where K and δ denote the cardinality and norm of the partition, respectively. Consider now a dataset $(\mathbf{X},L)$. Fixing a partition $\mathscr{K}_i(K) = \{\mathscr{X}_i^1,\dots,\mathscr{X}_i^K\}$, the given data estimator $\widehat{\xi}_i(K,N)$ is written as

$$\widehat{\xi}_i(K,N) = \sum_{k=1}^{K} \zeta\big(\widehat{\mathbf{p}}_L, \widehat{\mathbf{p}}_{L|X_i\in\mathscr{X}_i^k}\big)\widehat{p}\big(X_i \in \mathscr{X}_i^k\big).$$

Now, let $N \to \infty$. If $\zeta(\cdot,\cdot)$ is continuous, then

$$\begin{aligned}\lim_{N\to\infty} \widehat{\xi}_i(K,N) &= \lim_{N\to\infty} \sum_{k=1}^{K} \zeta\big(\widehat{\mathbf{p}}_L, \widehat{\mathbf{p}}_{L|X_i\in\mathscr{X}_i^k}\big)\widehat{p}\big(X_i \in \mathscr{X}_i^k\big) \\ &= \sum_{k=1}^{K} \zeta\big(\mathbf{p}_L, \mathbf{p}_{L|X_i\in\mathscr{X}_i^k}\big)p\big(X_i \in \mathscr{X}_i^k\big) =: \xi_i(K)\end{aligned}$$

by the same argument holding for the consistency of the discrete case. Then, we consider a sequence of refining partitions of $\mathscr{X}_i$ such that $\mathscr{K}_i(K+1)$ is finer than $\mathscr{K}_i(K)$ and such that $\lim_{K\to\infty}\mathscr{K}_i(K) = \mathscr{X}_i$ (that is equivalent to $\lim_{\delta\to 0}\mathscr{K}_i^\delta = \mathscr{X}_i$ with the notation above). Then, by Rohlin's disintegration theorem, we have that $\widehat{\mathbf{p}}_{L|X_i\in\mathscr{X}_i^k} \to \widehat{\mathbf{p}}_{L|X_i=x_i^k}$ for almost every $x_i^k \in \mathscr{X}_i$. Then, by the continuity and boundedness of $\zeta(\cdot,\cdot)$ and the definition of Riemann-Stieltjes integral, we have that

$$\lim_{K\to\infty}\xi_i(K) = \xi_i^L.$$

□

Proof of Corollary 2.3.7. To prove the assertion, we first need to show that the three metrics are bounded. We start with the Kuiper metric. We have that

$$\begin{aligned}\zeta^{\mathrm{KU}}(\mathbf{p}_L,\mathbf{p}_{L|X_i\in\mathscr{X}_i^k}) = &\sup_{r=1,2,\ldots,n_L}\left\{p(L=\ell_r) - p(L=\ell_r|X_i\in\mathscr{X}_i^k)\right\}\\ &+\sup_{r=1,2,\ldots,n_L}\left\{p(L=\ell_r|X_i\in\mathscr{X}_i^k) - p(L=\ell_r)\right\}.\end{aligned}$$

Noting that the maximum value that $|p(L=\ell_r) - p(L=\ell_r|X_i\in\mathscr{X}_i^k)|$ can assume is 1, we have $\zeta^{\mathrm{KU}}(\mathbf{p}_L,\mathbf{p}_{L|X_i\in\mathscr{X}_i^k}) \le 2$. Similarly, for the 1-norm, we have

$$\begin{aligned}\zeta^1(\mathbf{p}_L,\mathbf{p}_{L|X_i\in\mathscr{X}_i^k}) &= \sum_{r=1}^{n_L}\left|p(L=\ell_r) - p(L=\ell_r|X_i\in\mathscr{X}_i^k)\right|\\ &\le \sum_{r=1}^{n_L}p(L=\ell_r) + \sum_{r=1}^{n_L}p(L=\ell_r|X_i\in\mathscr{X}_i^k) = 2.\end{aligned}$$

For the 2-norm, we have

$$\begin{aligned}\zeta^2(\mathbf{p}_L,\mathbf{p}_{L|X_i\in\mathscr{X}_i^k}) &= \sum_{r=1}^{n_L}\left[p(L=\ell_r) - \mathbf{p}(L=\ell_r|X_i\in\mathscr{X}_i^k)\right]^2\\ &= \sum_{r=1}^{n_L}p(L=\ell_r)^2 - 2\sum_{r=1}^{n_L}p(L=\ell_r)p(L=\ell_r|X_i\in\mathscr{X}_i^k) + \sum_{r=1}^{n_L}p(L=\ell_r|X_i\in\mathscr{X}_i^k)^2.\end{aligned}$$

Hence, because $p(L=\ell_r)^2 \le p(L=\ell_r)$ and $p(L=\ell_r|X_i\in\mathscr{X}_i^k)^2 \le p(L=\ell_r|X_i\in\mathscr{X}_i^k)$, then $\sum_{r=1}^{n_L}p(L=\ell_r)^2 \le \sum_{r=1}^{n_L}p(L=\ell_r) \le 1$, and similarly $\sum_{r=1}^{n_L}p(L=\ell_r|X_i\in\mathscr{X}_i^k)^2 \le 1$. This then leads to

$$\zeta^2(\mathbf{p}_L,\mathbf{p}_{L|X_i\in\mathscr{X}_i^k}) \le 2 - 2\sum_{r=1}^{n_L}p(L=\ell_r)p(L=\ell_r|X_i\in\mathscr{X}_i^k).$$

Because $\sum_{s=1}^{n_L}p(L=\ell_r)p(L=\ell_s|X_i=x_i)$ is positive, we have $\zeta^2(\mathbf{p}_L,\mathbf{p}_{L|X_i\in\mathscr{X}_i^k}) \le 2$. Consistency of the estimators then follows from Proposition 2.3.6. □

Part II

Bayesian nonparametric modeling for complex networks

Chapter 3

Bayesian nonparametric hierarchical models for multiplex networks

3.1 Introduction

Network data has become ubiquitous in modern applications, leading to renewed interest in finding block structures that may be hidden in graphs [Fortunato and Hric, 2016, Fortunato and Newman, 2022]. The objective of this work is to investigate various types of block structures within a multiplex network, which is a layer-organized graph representing different types of relationships among the same instances. Specifically, we focus on studying networks that exhibit connectivity patterns among brain areas detected for different subjects. In this setting, it is of considerable interest to infer both subject-specific clusters of cerebral regions, which are groupings based on personal brain interconnections, as well as a common clustering, which is a partition of brain areas common to all the patients involved in the study. We expect to find some physical similarities displayed by all the brain maps, as well as individual differences due to possible illnesses. One common method for obtaining brain network data involves constructing a functional connectivity network using the inverse covariance matrix of fMRI scans. In this approach, low values of the precision matrix indicate that pairs of brain areas are conditionally independent (e.g. Smith et al. [2011], Simpson et al. [2013]), meaning that the activity in one area is not strongly influenced by the activity in the other area. This allows researchers to construct a network that captures the patterns of functional connectivity among different regions of the brain. By analyzing the resulting network, researchers can gain insight into the relationships between functional brain regions and how they contribute to various cognitive processes or behaviors. While functional connectivity networks remain of fundamental interest, recent advances in Diffusion Tensor Imaging (DTI) technologies [Craddock et al., 2013] have led to an increased focus on structural brain network data that measures anatomical connections made by axonal pathways. As discussed by Sporns [2013], there are notable differences between functional and structural connectivity networks, and the latter is particularly useful in understanding the underlying anatomical basis of brain functions. In this work, we specifically focus on structural anatomical graphs that are provided by DTI scans. These graphs allow us to study the structural connections between different regions of the brain and understand how these connections contribute to various cognitive processes or behaviors. By analyzing the structural network, we can gain insights into the physical basis of the brain function and how it is altered in presence of diseases or conditions. Network science has emerged as the most effective tool for analyzing brain connectivity patterns and understanding how different regions of the brain are functionally and structurally connected. Through the use of network analysis techniques, researchers can identify subnetworks of brain regions that exhibit similar connectivity patterns, providing insights into the organization of the brain and the neural circuits that underlie different cognitive processes. Additionally, network science allows for the exploration of how alterations in brain connectivity patterns are associated with various neurological or psychiatric disorders. The contributions to the literature on brain networks span a wide range of analyses, from classical descriptive approaches [Bassett and Bullmore, 2007, Bullmore and Sporns, 2009, Rubinov and Sporns, 2010, van den Heuvel and Sporns, 2013, Bassett and Bullmore, 2017] to more complex inferential frameworks [Bullmore and Sporns, 2012, Baggio et al., 2018, Demir et al., 2020]. An in-depth survey of brain graphs can be found in Sporns [2010]. In addition to studying the properties of individual brain regions, some works have also investigated the presence of complex group structures involving multiple brain areas [Tononi et al., 1998, Eickhoff et al., 2018]. For instance, Zemanová et al. [2006] and Crofts et al. [2016] have identified functional group structures, suggesting the presence of subnetworks within the brain that may be involved in specialized processes. These findings provide further evidence for the modular organization of the brain and the importance of studying clusters in addition to individual brain regions. While the contributions to the literature on brain networks have provided valuable insights into the functioning of the human brain, most of these studies have focused on simpler data structures, such as standard graphs, and have used classical community detection algorithms [Newman, 2004, Blondel et al., 2008, Karrer and Newman, 2011, Barabási, 2013] to detect meaningful patterns among the nodes [Bassett and Bullmore, 2007, 2017, Faskowitz et al., 2018]. However, these techniques do not easily extend to multiplex networks without loss of information, as thoroughly discussed in Section 3.5. For example, in a multiplex network where each layer represents a brain map of a different patient, adaptations of standard techniques do not allow for information sharing across subjects who are part of the same study and undergoing the same scans. However, since the brain anatomy is largely similar across most humans, we would expect that the connections displayed by a single subject should contain information that is relevant to the inference of latent structures in the brain maps of all the other patients. Therefore, it is important to develop techniques that can effectively analyze multiplex brain networks and take into account both subject-specific connectivity patterns and shared structural properties. Thus, we believe that the search for partition patterns of brain regions of different patients (encoded in a multiplex network) should be two-fold: on the one side, we would like to infer patient-specific groupings of brain areas, which may be influenced by possible personal diagnoses. On the other hand, an anatomical clustering should be provided, highlighting common patterns to all the patients, estimated using all the available information in the multiplex network.

3.1.1 Relevant literature

The importance of learning block structures in network data has motivated a collective effort by various disciplines towards the development of methods for detecting node groups, ranging from algorithmic strategies [Newman, 2004, Von Luxburg, 2007, Blondel et al., 2008, Karrer and Newman, 2011] to model-based solutions, among which we focus on the Stochastic Block Model — SBM henceforth — [Nowicki and Snijders, 2001, Schmidt and Morup, 2013] and its generalizations, such as the mixed-membership SBM [Airoldi et al., 2008], the degree-corrected SBM [Karrer and Newman, 2011, Côme et al., 2021], the bipartite SBM [Larremore et al., 2014] and the Extended SBM (ESBM) [Legramanti et al., 2022]. The SBM is a popular generative model for binary network data that assumes the network is partitioned into clusters or blocks. In the SBM, the probability of an edge between two nodes only depends on their cluster memberships, rather than on the individual nodes themselves. This property of the SBM makes it particularly useful for inferring the underlying block structure of a network, as the probability of observing a given network can be written as a function of the block assignments and the block probabilities. Bayesian inference for SBMs involves estimating both the posterior partition of nodes into blocks and the posterior block probabilities, given the observed network data. Overall, it is a useful framework for studying block structures in networks, as it provides a flexible and efficient way to model and infer their latent composition. See Fortunato and Hric [2016] and Fortunato and Newman [2022] for an overview. Another notable class of clustering models for networks are latent space models, which have also been adapted to multiplex graphs by Gollini and Murphy [2016]. However, comparing the results of stochastic block models and latent space models poses a significant challenge, which arises because stochastic block models yield latent partitions, while latent space models provide not only partitions but also spatial embeddings for nodes. In other words, although both models yield clusters, latent space models go a step further by offering additional information through the embedding of nodes into a continuous space. Therefore, comparing these two approaches is difficult due to the distinct types of information they provide. For this reason, such a comparison is out of scope of this thesis and deferred to future work.

Additionally, some effort has been put into generalizing existing frameworks to deal with multiplex networks: among others, it is the case of spectral clustering [DeFord and Pauls, 2019], topological clustering [Yuvaraj et al., 2021] and also SBM [Barbillon et al., 2015, Vallè s-Català et al., 2016]. However, few of the approaches currently presented in the literature fully exploit the potential of the multiplex structure: usually, they either do not consider the division of the nodes in layers (allowing inter-layer clustering, often uninterpretable, or providing one common partition for the entire edge-colored network) or they do not acknowledge the identity of the nodes, which is the same across layers. Another common technique is the application of standard community-detection algorithms to a collapsed version of a multiplex network [Vallè s-Català et al., 2016]: this usually requires a previous transformation of the data. For example, a naive approach could be to fit a SBM on the supra-adjacency matrix of the multiplex network [Kivelä et al., 2014]: this requires collapsing the multiplex network a priori, with a non-quantifiable loss of information. On the other hand, it is possible to fit a standard SBM on each layer of the multiplex network, independently: this procedure ignores the shared identity of the nodes across the layers, treating them as independent (see Section 3.5 for additional details). To mitigate such a loss of information, we propose a hierarchical version of the SBM for multiplex networks, the multiplex ESBM (mESBM). We use a Bayesian nonparametric approach to define a flexible model: indeed, in nonparametric modeling, the structure of the model is not fixed and thus the model flexibility can adapt as needed according to the complexity of the data. Among other numerous advantages [Nowicki and Snijders, 2001], some Bayesian nonparametric priors on the partition do not require to set the number of final clusters a priori, which is an evident limit of standard parametric approaches. The mESBM generalizes the ESBM by Legramanti et al. [2022] by defining a hierarchical Bayesian nonparametric structure on the latent partitions. The mESBM estimates two types of partitions of the nodes: layer-specific clusters, i.e. clusters within each layer depending on the corresponding edges, and a common grouping that partitions the nodes according to the information provided by the entire multiplex network. In this way, the mESBM is able to exploit the information provided by the division in layers, i.e. how the nodes differently interact within each layer, as well as the common identity of the nodes through layers. The mESBM induces a borrowing of information mechanism across layers through the use of informed Gibbs-type priors [Gnedin and Pitman, 2004, De Blasi et al., 2015], and also provides a hyperparameter to tune the amount of information shared. Moreover, the mESBM is a partially exchangeable model [De Finetti, 1980, Diaconis, 1988]: the multiplex network is such that the data (i.e. edges/nodes) are exchangeable within each layer. Even though the mESBM was initially conceived as a generalization of the ESBM by Legramanti et al. [2022], this extension requires a completely different computational approach for the estimation of the latent partitions. In fact, due to the complex nature of the posterior distributions in the problem at hand, a standard Gibbs sampler, as described in Legramanti et al. [2022], cannot be directly applied to sample from these distributions. As a result, more advanced and non-trivial combinations of Monte Carlo algorithms are required to obtain approximate posterior samples.

The rest of the Chapter is organized as follows: Section 3.2 introduces the multiplex network data structure in detail, while Section 3.3 defines the mESBM. Section 3.4 provides the computational framework to estimate all the posterior partitions in the mESBM, while

Sections 3.6, 3.5 and 3.7 show the application of the mESBM and two competitors to a simulated multiplex network and to cerebral maps generated by DTI scans.

3.2 Multiplex networks

Network theory [Kivelä et al., 2014] is a fundamental part of data science, which describes and analyzes complex systems. Originally, this field dealt with simple (but not trivial) graph structures, namely unweighted and undirected networks [Barabási, 2013]. However, due to the ever growing complexity and availability of data, researchers have soon moved to more sophisticated frameworks. As complex systems evolved, models and data structures needed to adapt: directed and weighted networks became more popular, as well as networks displaying multiple type of connections. In order to recall the concepts this work builds on, a graph (or network) can be defined as follows [Kivelä et al., 2014]:

Definition 3.2.1. *A **graph** is a tuple $G = (W, E)$ where W is a set of $V = |W|$ nodes and $E \subseteq W \times W$ is the set of edges, connecting the nodes.*

Definition 3.2.1 includes weighted, unweighted, directed, and undirected networks, but other possible generalizations exist. Among them, Kivelä et al. [2014] formalize the idea of *multilayer network*, which includes as specific cases most of the types of graphs currently used in the literature (such as multiplex networks, i.e. the data structures object of this chapter). The authors establish the concepts of *aspects* $a = 1, \ldots, d$ (e.g. time or space) and *layers* $\mathbb{L} = \{L_a\}_{a=1}^{d}$: this allows the multilayer network to have multiple layers for different aspects (e.g. to represent different types of relationships among the nodes, possibly evolving over time).

Definition 3.2.2. *A **multilayer network** [Kivelä et al., 2014] is a tuple $\mathcal{M} = (G_M, W, \mathbb{L})$ where $G_M = (W_M, E_M)$ is a graph for each combination M of layer-aspect, W is the entire set of nodes and $\mathbb{L}$ as above.*

In this chapter, we consider multilayer networks with binary, undirected edges in each layer. Thus, if not specified otherwise, the edges will be considered undirected henceforth, and taking values in $\{0, 1\}$. Moreover, this chapter deals with a specific type of multilayer network, the so-called *multiplex* or *edge-colored* network.

Definition 3.2.3. *A **multiplex** (or **edge-colored**) **network** [Kivelä et al., 2014] can be defined as a sequence of graphs $G_{multiplex} = \{W, E, C\}$, where $E \subseteq W \times W \times C$ is the edge set, W is the set of $V = |W|$ nodes and C is the color set (i.e. labels of layers).*

Basically, edge-colored networks are multilayer networks with one aspect ($d = 1$), whose layer set $\mathbb{L}$ is defined by the color set C (the terms layer and color will be used interchangeably from now on); all the layers $c \in C$ contain the same set of nodes W, but they differ by the edge set $E_c \subseteq W \times W$, $c \in C$. Multiplex networks typically represent sets of interactions among the same individuals: in some instances they may present inter-layer edges linking nodes standing for the same entity in different layers. However, in this chapter we consider edge-colored networks with intra-layer edges only, i.e. where the links are between nodes in the same layer. Moreover, since the goal is clustering, self-loops are not considered, as customary in the literature.

Notation

In the remainder, let $Y = [Y_1, \dots, Y_K]$ be an edge-colored network with V nodes and K colors/layers, with $Y_1, \dots, Y_K$ denoting the $V \times V$ binary adjacency matrices defining the graph in each layer. Thus, each element $Y_{k,uv} \in \{0,1\}$ of the adjacency matrix Y_k denotes the presence or absence of a link between node u and node v in the k^{th} layer (with $u, v = 1, \dots, V$ and $k = 1, \dots, K$). Let $\mathbf{z}_k = (z_{k1}, \dots, z_{kV}) \in \{1, \dots, H_k\}^V$ for $k = 0, \dots, K$ be the node membership vector associated to a generic node partition into H_k groups, so that $z_{kv} = h$ if and only if v belongs to cluster h in partition $\mathbf{z}_k$, for $v = 1, \dots, V$ and $h = 1, \dots, H_k$. Moreover, let $\mathbf{z} = [\mathbf{z}_1, \dots, \mathbf{z}_K]'$ denote a $V \times K$ matrix with, as columns, the layer-specific partitions, i.e. clusters of the V nodes within each layer of the multiplex network; let the additional partition $\mathbf{z}_0$ be a general clustering of the V nodes, which does not refer to any specific layer. Let then n_{h_k} denote the size of cluster h_k in partition $\mathbf{z}_k$, for $k = 0, \dots, K$ and $h_k = 1, \dots, H_k$. Henceforth, the apex $^{-v}$ denotes a quantity computed excluding node v, for $v = 1, \dots, V$.

3.3 Multiplex extended stochastic block models (mESBM)

The goal of this work is to define a hierarchical Bayesian nonparametric model to cluster the nodes of a multiplex network in two different ways. Recalling the notation introduced in Section 3.2, the aim is to find two types of partitions: a) the layer-specific clusters $\mathbf{z} = [\mathbf{z}_1, \dots, \mathbf{z}_K]'$ for each layer of the multiplex network Y, and b) the general (or common) partition $\mathbf{z}_0$. The partitions $\mathbf{z}_1, \dots, \mathbf{z}_K$ group nodes within the same layer of the network (inter-layer clusters are not admitted), but the model allows borrowing of information across the partitions in different layers through the dependence induced by a common grouping $\mathbf{z}_0$. Such dependence across layers is desirable, since the nodes represent the same entities, and we believe that their identity is an important information to exploit for the inference of the layer-specific partitions. To pursue the goal, we define the multiplex Extended Stochastic Block Model (mESBM) for a binary, edge-colored network:

$$\begin{aligned}
Y_{k,uv}|z_{ku}=h,z_{kv}=l &\overset{\text{ind.}}{\sim} \text{Bernoulli}(\Phi_{k,hl}) && \text{for } 1\le u<v\le V;\, k=1,\dots,K,\\
\Phi_{k,hl} &\overset{\text{i.i.d.}}{\sim} \text{Beta}(a,b) && \text{for } 1\le h\le k\le H_k,\\
\mathbf{z}_1,\dots,\mathbf{z}_K|\mathbf{z}_0 &\overset{\text{i.i.d.}}{\sim} \text{informed Gibbs-type distribution}(\gamma;\quad \mathbf{z}_0,\alpha), && \qquad (3.1)\\
\mathbf{z}_0 &\sim \text{Gibbs-type distribution}(\beta).
\end{aligned}$$

In this setting, the probability of observing a link between node u and node v in the k^{th} layer (i.e. $Y_{k,uv}=1$) just depends on the layer-specific labels of such nodes z_{ku}, z_{kv}. Moreover, the block probabilities follow a conjugate Beta distribution. The first three levels of the hierarchical model in (3.1) define an Extended Stochastic Block Model (ESBM) Legramanti et al. [2022]. The novel part is the addition of a mechanism to borrow information across $\mathbf{z}_1,\dots,\mathbf{z}_K$ through $\mathbf{z}_0$. It is worth noticing that $\mathbf{z}_0,\mathbf{z}_1,\dots,\mathbf{z}_K$ identify labeled clusters: hence, technically, a vector $\mathbf{z}_k$ and $\mathbf{z}'_k$, $k=0,\dots,K$, may contain different labels (and thus be mismatched objects), even though they identify the same partition. Henceforth, we use the following convention to identify a unique labeling for each given partition, setting $z^*_{k1}\to z_{k1}=1$ and sequentially relabeling each membership value in $\mathbf{z}^*_k=(z^*_{k1},\dots,z^*_{kV})$ as

$$z^*_{ki}\to z_{ki}=\begin{cases}\max\{z_{k1},\dots,z_{ki-1}\}+1 & \text{if } z^*_{ki}\notin(z^*_{k1},\dots,z^*_{ki-1}),\\ z_{kj} & \text{for } j: z^*_{kj}=z^*_{ki},\end{cases} \qquad (3.2)$$

for $i=2,\dots,V$ and $k=0,\dots,K$. This operation effectively defines an equivalence class, so that all the labelled vectors attaining the same partition will be reduced to the same object. This procedure is equivalent to the canonical projection in Peng and Carvalho [2013]. In the next two sections we present and analyze the Gibbs-type distributions used as priors for the partitions included in the mESBM.

First however, since the mESBM described in Section 3.3 encompasses a family of models, we specify the multiplex Dirichlet Process Dirichlet Process (mDPDP) model in Equation (3.3) and we use it in the applied parts of this chapter (namely, Section 3.7).

$$\begin{aligned}
Y_{k,uv}|z_{ku}=h,z_{kv}=l &\overset{\text{ind.}}{\sim} \text{Bernoulli}(\Phi_{k,hl}) && \text{for } 1\le u<v\le V \text{ and } k=1,\dots,K, && (3.3)\\
\Phi_{k,hl}|\mathbf{z}_k &\overset{\text{i.i.d.}}{\sim} \text{Beta}(a,b) && \text{for } 1\le h\le l\le H_k, && (3.4)\\
\mathbf{z}_1,...,\mathbf{z}_K|\mathbf{z}_0 &\overset{\text{i.i.d.}}{\sim} \text{informed Dirichlet} && \text{process}(\gamma;\mathbf{z}_0,\alpha), && (3.5)\\
\mathbf{z}_0 &\sim \text{Dirichlet process}(\beta). && && (3.6)
\end{aligned}$$

Notice that the model in Equation (3.3) is part of the family defined by the mESBM, where the prior on $\mathbf{z}_0$ is a Dirichlet Process (DP), and the distribution of $\mathbf{z}_k|\mathbf{z}_0$ for $k=1,\dots,K$ is an informed DP. In the remainder, we fix the following hyperparameters: $\beta=\gamma=1, a=$

$1, b = 1$.

3.3.1 Gibbs-type distributions

Several partition models have been considered in the literature, and arguably the most notable class of such distributions are Gibbs-type distributions [Gnedin and Pitman, 2004, De Blasi et al., 2015]. Gibbs-type priors are defined on the space of partitions of V observations. More specifically, we define a probability mass function $p(\mathbf{z})$ (where, with a slight abuse of notation, $\mathbf{z}$ is any partition of the V nodes into H clusters) as Gibbs-type if and only if it has the form

$$p(\mathbf{z}) = \mathscr{W}_{V,H} \prod_{h=1}^{H} (1-\sigma)_{n_h - 1}, \tag{3.7}$$

where n_h is the cardinality of cluster $h = 1, \dots, H$, $\sigma < 1$ is the discount parameter and $\{\mathscr{W}_{V,H} : 1 \leq H \leq V\}$ is a collection of non-negative weights such that $\mathscr{W}_{V,H} = (V - H\sigma)\mathscr{W}_{V+1,H} + \mathscr{W}_{V+1,H+1}$ and $\mathscr{W}_{1,1} = 1$. A convenient feature of Gibbs-type distributions is the availability of closed-form predictive urn schemes [De Blasi et al., 2015]:

$$p(z_{V+1} = l|\mathbf{z}) \propto \begin{cases} \mathscr{W}_{V+1,H}(n_l - \sigma) & \text{for } l = 1, \dots, H, \\ \mathscr{W}_{V+1,H+1} & \text{for } l = H+1. \end{cases} \tag{3.8}$$

The urn scheme in (3.8) allows the derivation of simple algorithms to obtain samples from Gibbs-type distributions. In fact, the urn scheme in Equation (3.8) is coherent across sample sizes, that is:

$$\sum_{h=1}^{H+1} p(z_{V+1} = h, \mathbf{z}) = p(\mathbf{z}).$$

This property ensures that Equation 3.8 can be used to sequentially generate observations from a Gibbs-type distribution.

3.3.2 Informed Gibbs-type distributions

In this chapter, we call *informed Gibbs-type* distributions the family referred to as *supervised Gibbs-type* in Legramanti et al. [2022]. The intuition is that in the mESBM there are no covariates supervising the partition process, but instead the mESBM introduces a mechanism for sharing information across the layer-specific partitions $\mathbf{z}_1, \dots, \mathbf{z}_K$ inducing a dependency structure through a common clustering $\mathbf{z}_0$ (both to be learnt). More in general, informed Gibbs-type distributions are part of the family of Product Partition Models (PPMs) [Muller et al., 2011, Page and Quintana, 2015, 2018]. Let us then define an informed Gibbs-type distribution for a generic partition $\mathbf{z}$, given $\mathbf{z}_0$, as:

$$p(\mathbf{z}|\mathbf{z}_0) = \frac{q(\mathbf{z}|\mathbf{z}_0)}{c(\mathbf{z}_0)} = \frac{1}{c(\mathbf{z}_0)} \mathscr{W}_{V,H} \prod_{h=1}^{H} g(\mathbf{z}_{0h})(1-\sigma)_{n_h - 1}, \tag{3.9}$$

where $\mathbf{z}_{0h} = \{z_{0i} : z_i = h\}$, while $\mathbf{z}$, the discount parameter $\sigma < 1$ and the non-negative weights $\mathscr{W}_{V,H}$ are as in Section 3.3.1. The normalising constant $c(\mathbf{z}_0)$ involves a sum over the set of all the possible partitions $\mathcal{Z}$, whose cardinality is equal to the Bell number of the second kind, which grows exponentially with V:

$$c(\mathbf{z}_0) = \sum_{\mathbf{z}\in\mathcal{Z}} q(\mathbf{z}|\mathbf{z}_0).$$

In general (heuristically, for $V > 4$), the computation of $c(\mathbf{z_0})$ is unfeasible. The quantity $g(\mathbf{z}_{0h})$ is the *similarity function*, measuring the similarity of nodes included in cluster h in $\mathbf{z}$ with respect to their cluster assignment in $\mathbf{z}_0$, for $h = 1,\dots,H$. Since $\mathbf{z}_0$ is a categorical variable, we follow the recommended practice in the literature [Muller et al., 2011], relying on a fictitious model for $\mathbf{z}_0$, obtaining as similarity a Dirichlet-multinomial distribution:

$$g(\mathbf{z}_{0h}) = \frac{\Gamma(H_0\alpha)}{\prod_{h_0=1}^{H_0}\Gamma(\alpha)} \frac{1}{\Gamma(n_h + H_0\alpha)} \prod_{h_0=1}^{H_0} \Gamma(n_{hh_0} + \alpha). \tag{3.10}$$

Here, α is an important hyperparameter to be set a priori, and n_{hh_0} is the cardinality of the intersection between cluster h in $\mathbf{z}$ and cluster h_0 in $\mathbf{z}_0$. In general, the higher the overlap between $\mathbf{z}_0$ and $\mathbf{z}$, the higher the similarity function. Finally, notice that we include the normalising constant $\Gamma(H_0\alpha)/\prod_{h_0=1}^{H_0}\Gamma(\alpha)$ of the Dirichlet-multinomial distribution in the definition of the similarity function (differently from Legramanti et al. [2022]): this choice is well-motivated below. The informed Gibbs-type distribution also yields closed-form full conditionals:

$$p(z_v = h|\mathbf{z}_0, \mathbf{z}^{-v}) \propto \begin{cases} \dfrac{n^{-v}_{hz_0^v} + \alpha}{n_h^{-v} + H_0\alpha} \mathscr{W}_{V,H^{-v}}(n_h^{-v} - \sigma) & \text{for } h = 1,\dots,H^{-v}, \\ \dfrac{1}{H_0}\mathscr{W}_{V,H^{-v}+1} & \text{for } h = H^{-v} + 1. \end{cases} \tag{3.11}$$

Differently from the standard Gibbs-type family in Section 3.3.1, the urn scheme in Equation 3.11 is **not** coherent across sample sizes according to the standard definition, i.e.

$$\sum_{h=1}^{H+1} p(z_{V+1} = h, \mathbf{z}|\mathbf{z}_{0V+1}, \mathbf{z}_0) \neq p(\mathbf{z}|\mathbf{z_0}).$$

This implies that at each sample size $v = 1,\dots,V$ the underlying informed model $p(z_{1:v}|z_{01:v})$ is effectively changing for different values of v, making it impossible to simulate observations from an informed Gibbs-type distribution sequentially. However, Muller et al. [2011] introduce a definition of coherency for Product Partition Models (which include the general class of informed Gibbs-type distributions):

$$p_V(\mathbf{z}|\mathbf{z_0}) = \sum_{h=1}^{H+1}\sum_{h_0=1}^{H_0} p_{V+1}(\mathbf{z}, z_{V+1} = h|\mathbf{z_0}, z_{0V+1} = h_0)p(z_{0V+1} = h_0|\mathbf{z}_0).$$

Intuitively, this means that an informed Gibbs-type distribution is coherent across sample sizes only integrating out z_{0V+1}, i.e. when z_{0V+1} does not provide any information for the clustering of the $V+1$ observation in $\mathbf{z}$. This is opposed to the scope of the mESBM, that is to inform the layer-specific partitions $\mathbf{z}$ with a common clustering $\mathbf{z}_0$. Moreover, this definition of coherency does not entail a proper scheme for sequential simulation from the informed Gibbs-type family.

Borrowing of information: the role of the hyperparameter α

In the previous section, we highlighted the presence of the normalizing constant $\Gamma(H_0\alpha)/\prod_{h_0=1}^{H_0}\Gamma(\alpha)$ in the similarity function of the informed Gibbs-type distribution (reported below for convenience):

$$g(\mathbf{z}_{0h}) = \frac{\Gamma(H_0\alpha)}{\prod_{h_0=1}^{H_0}\Gamma(\alpha)}\frac{1}{\Gamma(n_h + H_0\alpha)}\prod_{h_0=1}^{H_0}\Gamma(n_{hh_0} + \alpha).$$

In the following, we argue that including such a normalising constant in the similarity function is important to provide a meaningful interpretation to $\alpha > 0$ as the parameter tuning the borrowing of information mechanism. Specifically, the lower α is, the higher the amount of borrowed information across $\mathbf{z}, \mathbf{z}_0$, which results in layer-specific partitions $\mathbf{z}$ more adherent to the common clustering $\mathbf{z}_0$. This mechanism becomes clear analyzing the limits of $p(\mathbf{z}|\mathbf{z}_0)$ with respect to the extreme values of α (for detailed computations, see the Appendix).

For $\alpha \to 0$,

$$\lim_{\alpha\to 0} g(\mathbf{z}_{0h}) = \frac{1}{\Gamma(n_h)}\prod_{h_0=1}^{H_0}\Gamma(n_{hh_0}),$$

yielding the maximum borrowing of information between $\mathbf{z}$ and $\mathbf{z}_0$. Thus, for low values of α, the influence of $\mathbf{z}_0$ on $\mathbf{z}$ through $p(\mathbf{z}|\mathbf{z}_0)$ only depends on the size of the intersections between clusters in $\mathbf{z}$ and clusters in $\mathbf{z}_0$, normalised, encouraging an overlap of $\mathbf{z}_1, \ldots, \mathbf{z}_K$ to $\mathbf{z}_0$.

On the other hand, for $\alpha \to \infty$,

$$\lim_{\alpha\to\infty} g(\mathbf{z}_{0h}) = c,$$

with $c \neq 0$ constant, and as a consequence $\lim_{\alpha\to\infty} p(\mathbf{z}|\mathbf{z}_0) = p(\mathbf{z})$ (where $p(\mathbf{z})$ is a standard Gibbs-type distribution of Section 3.3.1). Thus, for $\alpha \to \infty$, the informed model converges to a standard (not informed) Gibbs-type distribution, removing completely the influence of $\mathbf{z}_0$ on $\mathbf{z}$, basically modelling $\mathbf{z}_0, \mathbf{z}_1, \ldots, \mathbf{z}_K$ as independent draws from the same law. Hence, high values of α make the borrowing of information disappear.

3.4 Posterior computation and inference

In Section 3.3.1, we have highlighted two major features of informed Gibbs-type distributions, namely the unfeasibility of computing their normalising constant $c(\mathbf{z}_0)$ and their lack of projectivity, which hinders the sequential sampling from such a law. These features are not in general a problem, since the sampling of $\mathbf{z}|\mathbf{z}_0 \sim p(\mathbf{z}|\mathbf{z}_0)$ from an informed Gibbs-type distribution can be performed with a standard Gibbs sampler. However, they prevent the straightforward posterior sampling of $\mathbf{z}_0|\mathbf{z} \sim p(\mathbf{z}_0|\mathbf{z})$, as throughly explained next. Indeed, while the mESBM is defined in a straightforward way, the posterior estimation of the latent variables $\mathbf{z}, \mathbf{z}_0$ is not as simple. A first, naive approach is to use a standard Monte Carlo algorithm to provide (approximate) samples from the posterior distribution $p(\mathbf{z}, \mathbf{z}_0|Y)$, updating the partitions $\mathbf{z}_0, \mathbf{z}_1, \dots, \mathbf{z}_K$ one node at a time (for a comprehensive review of Monte Carlo methods, please refer to Chopin and Papaspiliopoulos [2020]). Such an algorithm would entail two steps:

1. Sample $\mathbf{z}|\mathbf{z}_0, Y \sim p(\mathbf{z}|\mathbf{z}_0, Y)$.
   For $k = 1, \dots, K$, sample $\mathbf{z}_k|\mathbf{z}_0, Y_k \sim p(\mathbf{z}_k|\mathbf{z}_0)p(Y_k|\mathbf{z}_k)$ using the likelihood and the informed urn scheme in Equation (3.1) to update the label of one node $v = 1, \dots, V$ at a time, according to the probabilites

$$p(z_{kv} = h_k|\mathbf{z}_0, \mathbf{z}_k^{-v}) \propto \begin{cases} \dfrac{n^{-v}_{h_k z_0^v} + \alpha}{n^{-v}_{h_k} + H_0\alpha} \mathscr{W}_{V,H_k^{-v}}(n^{-v}_{h_k} - \sigma) & \text{for } h_k = 1, \dots, H_k^{-v}, \\ \dfrac{1}{H_0} \mathscr{W}_{V,H_k^{-v}+1} & \text{for } h_k = H_k^{-v} + 1. \end{cases}$$

2. Sample $\mathbf{z}_0|\mathbf{z} \sim p(\mathbf{z_0}|\mathbf{z}) \propto \prod_{k=1}^K p(\mathbf{z}_k|\mathbf{z_0})p(\mathbf{z_0})$.
   The conditional distribution of $\mathbf{z}_0|\mathbf{z}$ is independent of the data Y, due to the hierarchical structure of the mESBM. However, $p(\mathbf{z}_k|\mathbf{z}_0) = \frac{q(\mathbf{z}_k|\mathbf{z}_0)}{c(\mathbf{z}_0)}$ and $c(\mathbf{z}_0) = \sum_{\mathbf{z}_k \in \mathcal{Z}} q(\mathbf{z}_k|\mathbf{z}_0)$ is a normalising constant that a) can not be discarded (since it depends on $\mathbf{z}_0$) and b) is unavailable in closed form and uncomputable for non-trivial cases. As a consequence, to sample from $p(\mathbf{z}_0|\mathbf{z})$, it would be necessary to compute the full conditional probabilities

$$\begin{aligned} p(z_{0v} = h_0|\mathbf{z}_1, \dots, \mathbf{z}_K, \mathbf{z}_0^{-v}) &\propto p(\mathbf{z}_1, \dots, \mathbf{z}_K|z_{0v} = h_0, \mathbf{z}_0^{-v})p(z_{0v} = h_0, \mathbf{z}_0^{-v}) \\ &= \prod_{k=1}^K p(\mathbf{z}_k|z_{0v} = h_0, \mathbf{z}_0^{-v})p(z_{0v} = h_0, \mathbf{z}_0^{-v}) \\ &= \prod_{k=1}^K \frac{q(\mathbf{z}_k|z_{0v} = h_0, \mathbf{z}_0^{-v})}{c(z_{0v} = h_0, \mathbf{z}_0^{-v})} p(z_{0v} = h_0, \mathbf{z}_0^{-v}), \end{aligned}$$

   where $c(\mathbf{z}_0)$ is impossible to obtain.

While a Gibbs sampler to sample from $p(\mathbf{z}|\mathbf{z}_0, Y)$ remains a valid choice, the same can not be said to sample from $p(\mathbf{z_0}|\mathbf{z})$ due to the unavailability of the normalising constant of the informed Gibbs-type distribution. Thus, it is not possible to carry out Step 2 of the previous algorithm to perform posterior inference of the common partition parameter $\mathbf{z}_0$. A first, alternative option could be to learn $\mathbf{z}_0|\mathbf{z}$ by replacing Step 2 of the Gibbs sampler with a Metropolis-Hastings iteration, that is:

2a. Propose $\mathbf{z}_0^* \sim h(\mathbf{z}_0^*|\mathbf{z}_0)$

2b. Accept $\mathbf{z}_0^*$ with probability

$$a = \min\left\{1, \frac{p(\mathbf{z}_0^*)h(\mathbf{z}_0|\mathbf{z}_0^*)p(\mathbf{z}|\mathbf{z}_0^*)}{p(\mathbf{z}_0)h(\mathbf{z}_0^*|\mathbf{z}_0)p(\mathbf{z}|\mathbf{z}_0)}\right\} = \min\left\{1, \frac{p(\mathbf{z}_0^*)h(\mathbf{z}_0|\mathbf{z}_0^*)q(\mathbf{z}|\mathbf{z}_0^*)}{p(\mathbf{z}_0)h(\mathbf{z}_0^*|\mathbf{z}_0)q(\mathbf{z}|\mathbf{z}_0)}\frac{c(\mathbf{z}_0)}{c(\mathbf{z}_0^*)}\right\}.$$

Once again, the acceptance probability a is not computable due to the presence of $c(\mathbf{z}_0^*), c(\mathbf{z}_0)$. In conclusion, the first major concern is that a standard Monte Carlo approach is not feasible because of the unavailability of the normalising constant of the informed Gibbs-type distribution $p(\mathbf{z}|\mathbf{z}_0)$.

Another major computational issue which prevents a straightforward posterior sampling of the partitions provided by the mESBM is the non-coherency of the informed full conditional probabilities in Equation (3.11). As already noted in Section 3.3.2, this prevents the sequential simulation from an informed Gibbs-type distribution. Being able to simulate from $p(\mathbf{z}|\mathbf{z}_0)$ is essential for the implementation of some Monte Carlo algorithms based on augmentation, among which a notable example is the Exchange algorithm [Murray et al., 2006, Caimo and Friel, 2011]. More specifically, consider the sequential generation of a sample from an informed Gibbs-type distribution $\mathbf{z} = (z_1,\dots,z_V)|\mathbf{z_0} = (z_{01},\dots,z_{0V})$: to this end, we would like to generate first $z_1|z_{01}$, then $z_2|z_1,z_{01},z_{02}$ and so forth. To this aim, we would need the urn scheme to be coherent across sample sizes in the following manner:

$$p_n(z_{1:n}|z_{0,1:n}) = \sum_{h=1}^{H_n+1} p_{n+1}(z_{1:n}, z_{n+1} = h|z_{0,1:n+1}) \quad \forall n = 1,\dots,V,$$

where $p_n(\cdot)$ is the distribution of $z_1,\dots,z_n$ given $z_{01},\dots,z_{0n}$ for $n = 1,\dots,V$ and H_n the number of clusters in $z_1,\dots,z_n$. As explained in Section 3.3.2, this condition is not satisfied by informed Gibbs-type distributions (and, in general, by product partition models), and consequently it is not possible to sequentially generate samples from such a law using the urn scheme in Equation (3.11). In the next two sections, we explore solutions to the two issues underlined above.

3.4.1 Exchange algorithm

The exchange algorithm [Murray et al., 2006, Caimo and Friel, 2011] is used in the literature with the goal of sampling from the posterior distribution of Exponential random graph models, among other things. Exponential random graphs have a computational issue similar to the mESBM: in particular, their likelihood can not be easily obtained because of an uncomputable normalising constant, yielded by a sum over all the possible graph configurations given by V nodes. The framework is then quite similar to the mESBM estimation, impossible to carry out using standard Monte Carlo methods. Adapting the algorithm in Caimo and Friel [2011] to sample from the posterior of the mESBM, we obtain a Monte Carlo algorithm which can replace Step 2 in the Gibbs sampler of the previous section, whose invariant distribution is the posterior law $p(\mathbf{z}_0|\mathbf{z})$:

2a. Propose $\mathbf{z}_0^* \sim h(\mathbf{z}_0^*|\mathbf{z}_0)$;

2b. Sample the auxiliary variables $\tilde{\mathbf{z}} \sim p(\mathbf{z}|\mathbf{z}_0^*) = q(\mathbf{z}|\mathbf{z}_0^*)/c(\mathbf{z}_0^*)$;

2c. Accept $\mathbf{z}_0^*$ with probability

$$a = \min\left\{1, \frac{p(\mathbf{z}_0^*)h(\mathbf{z}_0|\mathbf{z}_0^*)q(\mathbf{z}|\mathbf{z}_0^*)q(\tilde{\mathbf{z}}|\mathbf{z}_0)}{p(\mathbf{z}_0)h(\mathbf{z}_0^*|\mathbf{z}_0)q(\mathbf{z}|\mathbf{z}_0)q(\tilde{\mathbf{z}}|\mathbf{z}_0^*)}\right\}.$$

As a first step, the exchange algorithm requires to propose a new value $\mathbf{z}_0^* \sim h(\mathbf{z}_0^*|\mathbf{z}_0)$. To this aim, we present two different proposal distributions: first, the *global proposal distribution* propose the entire $V-$dimensional vector $\mathbf{z}_0^*$ from the prior distribution, using the prior urn scheme in Equation (3.8), i.e. $\mathbf{z}_0^* \sim p(\mathbf{z}_0)$, with $h(\mathbf{z}_0^*|\mathbf{z}_0) = p(\mathbf{z}_0)$. Second, the *local proposal distribution* chooses uniformly a random node, and updates its assignment according to the urn scheme in Equation (3.8). In this case, the resulting proposal distribution is

$$h(\mathbf{z}_0^*|\mathbf{z}_0) = \frac{1}{V}\sum_{i=1}^{V} p(z_{0i}^*|\mathbf{z}_0^{-i})\mathbb{1}_{\{v=i\}}.$$

In the remainder, we use the local proposal distribution for the Exchange algorithm. A statistical interpretation is available for the quantity $q(\tilde{\mathbf{z}}|\mathbf{z}_0)/q(\tilde{\mathbf{z}}|\mathbf{z}_0^*)$ in the acceptance probability. In fact, the exchange algorithm estimates the uncomputable ratio of normalising constants $c(\mathbf{z}_0)/c(\mathbf{z}_0^*)$ with a one-sample, unbiased, importance estimator:

$$\frac{c(\mathbf{z}_0)}{c(\mathbf{z}_0^*)} \approx \frac{q(\tilde{\mathbf{z}}|\mathbf{z}_0)}{q(\tilde{\mathbf{z}}|\mathbf{z}_0^*)}.$$

Looking at the steps above, it is clear that the exchange algorithm is suitable to simulate from a posterior distribution in the fashion of $p(\mathbf{z}_0|\mathbf{z}) \propto p(\mathbf{z}|\mathbf{z}_0)p(\mathbf{z}_0)$, where $p(\mathbf{z}|\mathbf{z}_0)$ has an uncomputable (but unavoidable) normalising constant. The only requirement is *being able to simulate the auxiliary variables* $\tilde{\mathbf{z}} \sim p(\mathbf{z}|\mathbf{z}_0)$ from the uncomputable distribution. This condition is easily satisfied by distributions coherent across sample sizes, since it is possible to sequentially simulate samples. Unfortunately, this is not the case for informed Gibbs-type distributions as thoroughly explained in Section 3.4. Caimo and Friel [2011] actually incur in the same issue, as it is not possible to sample exactly from exponential random graph laws: the authors use a Gibbs sampler within the Exchange algorithm to obtain approximate samples from such a distribution. Nevertheless, this necessity opens the question: how can we simulate from the informed Gibbs-type distribution $p(\mathbf{z}|\mathbf{z}_0)$, in order to generate the auxiliary variables $\tilde{\mathbf{z}}$ necessary to use the exchange algorithm, with the lightest possible computational burden? The next section will provide the answer.

3.4.2 Simulating from informed Gibbs-type distributions

In this subsection, we explore two ways of simulating the auxiliary variables from an informed Gibbs-type distribution $\tilde{\mathbf{z}}|\mathbf{z}_0 \sim p(\mathbf{z}|\mathbf{z}_0)$. Notice that, by definition of the mESBM in (3.1), $p(\mathbf{z}|\mathbf{z}_0) = \prod_{k=1}^{K} p(\mathbf{z}_k|\mathbf{z}_0)$: thus, we need to generate K independent and identically distributed auxiliary partitions $\tilde{\mathbf{z}}_k \sim p(\cdot|\mathbf{z}_0)$, $k = 1,\dots,K$, each from a univariate informed Gibbs-type law, one for each layer of the multiplex network Y. This procedure ensures $\tilde{\mathbf{z}} = \mathbf{z}$ in distribution as required by the Exchange algorithm.

Gibbs Sampler

A straightforward approach to simulate samples from an informed Gibbs-type distribution is to exploit the standard inferential framework (e.g. see Legramanti et al. [2022]), that is to employ a Gibbs sampler to simulate $\tilde{\mathbf{z}}|\mathbf{z}_0 \sim p(\mathbf{z}|\mathbf{z}_0)$. In this case, we sweep through each V-dimensional vector $\tilde{\mathbf{z}}_1,\dots,\tilde{\mathbf{z}}_K$ various times, updating one node at a time (in a random order) using Equation (3.11). This procedure is not impacted by the non-coherency of such an urn scheme, since we are updating vectors $\tilde{\mathbf{z}}_k$, $k = 1,\dots,K$, of fixed dimension V. Hence, at each step of the exchange algorithm we can generate the auxiliary variable $\tilde{\mathbf{z}} = [\tilde{\mathbf{z}}_1,\dots,\tilde{\mathbf{z}}_K]'$ as follows:

2a. Initialize $\tilde{\mathbf{z}} = [\tilde{\mathbf{z}}_1,\dots,\tilde{\mathbf{z}}_K]'$ (in the Exchange algorithm, one can retain the values of the auxiliary variables of the previous iteration);

For each network $k = 1,\dots,K$, and for a certain number of iterations do:

2b. Select a random order of update of the nodes;

2c. For each node v, update $\tilde{z}_{kv}|\tilde{\mathbf{z}}_k^{-v},\mathbf{z}_0 \sim p(\tilde{z}_{kv}|\tilde{\mathbf{z}}_k^{-v},\mathbf{z}_0)$ using Equation (3.11) and the order chosen in 2b.

Caimo and Friel [2011] follow an equivalent procedure to generate the auxiliary random variables identically distributed to exponential random graphs within each iteration of the Exchange algorithm. However, there is an obvious computational disadvantage, with respect to both the number of nodes (V) and the number of layers of the multiplex network (K). In fact, K, V-dimensional independent Markov Chains are estimated at each iteration through Gibbs sampling. To mitigate the computational complexity with respect to K, we propose the Sampling Importance Resampling in the next section.

Sampling Importance Resampling (SIR)

A second approach is to sample the auxiliary variables from the informed Gibbs-type distribution using the Sampling Importance Resampling (SIR) algorithm [Chopin and Papaspiliopoulos, 2020]. The main underlying idea is to propose $N \gg K$ independent values using as proposal the non-coherent urn scheme in Equation (3.11), and then to sample the actual $K \leq N$ independent, informed Gibbs-type auxiliary variables among those using importance weights. The steps are formalized below.

2a. Sample $N \gg K$ partitions of the V nodes using the non-coherent urn scheme in Equation (3.11), i.e. sample $\bar{\mathbf{z}}_i = (\bar{z}_{i1}, \dots, \bar{z}_{iV})$ for $i = 1, \dots, N$ as follows:

$$\begin{aligned}\bar{z}_{i1} | z_{01} &= 1, \\ \bar{z}_{i2} | \bar{z}_{i1}, z_{01}, z_{02} &\sim p_2(\bar{z}_{i2} | \bar{z}_{i1}, z_{01}, z_{02}) \text{ according to Equation (3.11)}, \\ &\vdots \\ \bar{z}_{iV} | \bar{z}_{i1}, \dots, \bar{z}_{iV-1}, \mathbf{z}_0 &\sim p_V(\bar{z}_{iV} | \bar{z}_{i1}, \dots, \bar{z}_{iV-1}, \mathbf{z}_0) \text{ according to Equation (3.11)}.\end{aligned}$$

Each final partition will have a joint distribution given by:

$$\bar{\mathbf{z}}_i \sim f(\bar{\mathbf{z}}_i | \mathbf{z}_0) = p_1(\bar{z}_{i1} | z_{01}) p_2(\bar{z}_{i2} | \bar{z}_{i1}, z_{01}, z_{02}) \dots p_V(\bar{z}_{iV} | \bar{z}_{i1}, \dots, \bar{z}_{iV-1}, \mathbf{z}_0)$$

with $p_v(\cdot)$, $v = 1, \dots, V$, as in Equation (3.11).

2b. To each partition $\bar{\mathbf{z}}_i$ for $i = 1, \dots, N$ assign a weight defined by:

$$w_i = \frac{q(\bar{\mathbf{z}}_i | \mathbf{z}_0)}{f(\bar{\mathbf{z}}_i | \mathbf{z}_0)}, \tag{3.12}$$

where $q(\bar{\mathbf{z}}_i | \mathbf{z}_0)$ is the (unnormalised) informed Gibbs-type distribution.

2c. Normalize the N weights obtained above:

$$\widehat{w}_i = \frac{w_i}{\sum_{n=1}^{N} w_n}.$$

2d. Finally, sample the K independent dimensions of $\tilde{\mathbf{z}} = [\tilde{\mathbf{z}}_1, \dots, \tilde{\mathbf{z}}_K]'$. For $k = 1, \dots, K$, sample $\tilde{\mathbf{z}}_k$ as:

$$\tilde{\mathbf{z}}_k = \begin{cases} \bar{\mathbf{z}}_1 \text{ with probability } \hat{w}_1, \\ \vdots \\ \bar{\mathbf{z}}_N \text{ with probability } \hat{w}_N. \end{cases}$$

Observe that we need to perform one SIR for each iteration of the Exchange algorithm, regardless of the number of layers in the multiplex network. However, we still need to choose the number of SIR samples N, which needs to increase with V and/or K. Gelman et al. [2004] suggests to carry out Step 2d sampling $\tilde{\mathbf{z}}$ from:

$$\tilde{\mathbf{z}}_k = \begin{cases} \bar{\mathbf{z}}_1 \text{ with probability } \hat{w}_1, \\ \vdots \\ \bar{\mathbf{z}}_N \text{ with probability } \hat{w}_N, \end{cases}$$

without replacement. In our case, sampling $\tilde{\mathbf{z}}$ with or without replacement does not seem to have a tangible impact on the final result, at least empirically, and we decided to stick with the standard SIR algorithm.

3.4.3 Posterior computation

Finally, we can define the algorithm to sample observations from the joint posterior distribution $p(\mathbf{z}, \mathbf{z}_0|Y)$. We combine the Gibbs sampler at the beginning of Section 3.4, with a step performed by SIR within the Exchange algorithm of Section 3.4.1. The final scheme is reported in Algorithm 1, which outputs a multidimensional Markov Chain with invariant distribution $p(\mathbf{z}, \mathbf{z}_0|Y)$.

3.4.4 Point estimation

While algorithmic techniques return a single estimated partition, the mESBM provides the empirical posterior distribution over the space of node partitions. Specifically, Algorithm 1 provides samples from $p(\mathbf{z}, \mathbf{z}_0|Y)$. To perform inference on the space of partitions, we adopt the approach of Wade and Ghahramani [2018]. In particular, we provide a posterior point estimate leveraging the Variation of Information (VI) metric [Meilă, 2007], that quantifies distances between two clusterings and ranges from 0 to $\log_2 V$. Intuitively, the lower the VI metric, the higher the overlap between two partitions; see Wade and Ghahramani [2018] for a detailed discussion. Under this scheme, a formal Bayesian point estimate for $\mathbf{z}, \mathbf{z}_0$ are the partitions minimizing the VI distance in the posterior sample, i.e.

$$\hat{\mathbf{z}}, \hat{\mathbf{z}}_0 = \underset{\mathbf{z}^*, \mathbf{z}_0^*}{\operatorname{argmin}} \mathbb{E}_{\mathbf{z}, \mathbf{z}_0}[\mathrm{VI}(\mathbf{z}, \mathbf{z}_0; \mathbf{z}^*, \mathbf{z}_0^*)] \tag{3.13}$$

Algorithm 1: Posterior sampling of latent partitions $\mathbf{z}_0, \mathbf{z}_1, \ldots, \mathbf{z}_K$

```
Input: $V \times V$ adjacency matrices $Y = [Y_1, \ldots, Y_K]$, number of auxiliary partitions $N$,
       initialization of $\mathbf{z}_0, \mathbf{z}_1, \ldots, \mathbf{z}_K$, hyperparameters.
Output: Posterior samples from $p(\mathbf{z}_0, \mathbf{z} | Y)$.
for each iteration do
    Sample $\mathbf{z}|\mathbf{z}_0, Y$ using a Gibbs sampler.
    for $k = 1, \ldots, K$ do
        for $v = 1, \ldots, V$ do
            Update $z_{kv} \sim p(z_{kv}|\mathbf{z}_k^{-v}, \mathbf{z}_0, Y_k) \propto p(z_{kv}|\mathbf{z}_k^{-v}, \mathbf{z}_0) p(Y_k|z_{kv}, \mathbf{z}_k^{-v})$ using
             Equations (3.9) and (3.11);
        end
    end
    Sample $\mathbf{z}_0|\mathbf{z}_1, \ldots, \mathbf{z}_K$ using SIR within Exchange algorithm.
    (1) Exchange algorithm: propose a new value $\mathbf{z}_0^*$.
    Sample $\mathbf{z}_0^* \sim h(\mathbf{z}_0^*|\mathbf{z}_0)$;
    (2) Exchange algorithm: sample the auxiliary variables
     using SIR, conditionally on $\mathbf{z}_0^*$.
    (2a) SIR: sample the proposed partitions.
    for $n = 1, \ldots, N$ do
        Sample $\bar{\mathbf{z}}_n|\mathbf{z}_0^*$ using the non-coherent urn scheme in Equation (3.11);
    end
    (2b) SIR: compute and normalize the weights.
    for $n = 1, \ldots, N$ do
        Compute the weights $w_n$ in Equation (3.12);
    end
    Normalize the weights $\hat{w}_n = \frac{w_n}{\sum_{j=1}^N w_j}$ for $n = 1, \ldots, N$;
    (2c) SIR: sample the auxiliary partitions $\tilde{\mathbf{z}} = [\tilde{\mathbf{z}}_1, \ldots, \tilde{\mathbf{z}}_K]'$.
    for $k = 1, \ldots, K$ do
        Sample $i \sim \text{Multinomial}(\hat{w}_1, \ldots, \hat{w}_N)$;
        Set the auxiliary partition $\tilde{\mathbf{z}}_k = \bar{\mathbf{z}}_i$;
    end
    (3) Exchange algorithm: sample $\mathbf{z}_0$.
    Update $\mathbf{z}_0 = \mathbf{z}_0^*$ with probability $a = \min\left\{1, \frac{p(\mathbf{z}_0^*) h(\mathbf{z}_0|\mathbf{z}_0^*) q(\tilde{\mathbf{z}}|\mathbf{z}_0) q(\mathbf{z}|\mathbf{z}_0^*)}{p(\mathbf{z}_0) h(\mathbf{z}_0^*|\mathbf{z}_0) q(\tilde{\mathbf{z}}|\mathbf{z}_0^*) q(\mathbf{z}|\mathbf{z}_0)}\right\}$.
end
```

where the expectation is taken with respect to the posterior distribution of $\mathbf{z}, \mathbf{z}_0$, i.e. $p(\mathbf{z}, \mathbf{z}_0|Y)$. The optimization in Equation (3.13) is typically carried out through Monte Carlo estimation or, alternatively, a greedy algorithm. From an practical viewpoint, we employ the `R` package `mcclust.ext` [Wade and Ghahramani, 2018].

3.5 Competitor models

To the best of our knowledge, a model to obtain both layer-specific and common partitions of the nodes of an edge-colored network has not been proposed in the literature yet. For this reason, we assess the two-fold output of the mESBM versus the partitions provided by two standard models adapted to the multiplex network framework. Such models are illustrated in the next subsections.

3.5.1 Binomial extended stochastic block model (bESBM)

The binomial ESBM (bESBM) [Ghidini et al., 2023b] is an extended stochastic block model [Legramanti et al., 2022] defined for weighted, bounded, integer-valued edges. The bESBM introduce a binomial likelihood to fully exploit the information in a weighted network, instead of dichotomizing its integer edge weights to apply the standard ESBM, as it is customary in the literature. However, the bESBM is still defined for single-layered, weighted networks and can not be applied directly on a multiplex graph. Thus, in this application, we exploit the so-called supra-adjacency matrix [Kivelä et al., 2014] of the undirected, binary, edge-colored graph of interest, defined as:

$$Y^* = \sum_{k=1}^{K} Y_k,$$

where the Y_k for $k = 1, \dots, K$ are the binary, adjacency matrices of each layer in the original multiplex network. In the remainder, we will call Y^* the *collapsed adjacency matrix* of the *collapsed network*. In this way, we are intuitively summarising the information of K different undirected, binary networks into a single adjacency matrix, defining an undirected weighted graph where each edge counts the number of links between the two nodes in the original multiplex network. Given such a collapsed matrix obtained from a K-layer edge-colored network, it is immediate to see that $Y^*_{uv} \in \{0, 1, \dots, K\}$ for $u, v = 1, \dots, V$. Thus, we can define the bESBM as follows:

$$Y^*_{uv}|z_u = h, z_v = k, \Phi_{hk} \overset{\text{i.i.d.}}{\sim} \text{Binomial}(K, \Phi_{hk}) \qquad \text{for } 1 \le u < v \le V, \tag{3.14}$$

$$\Phi_{hk}|\mathbf{z} \overset{\text{ind.}}{\sim} \text{Beta}(a, b) \qquad \text{for } 1 \le h \le k \le H, \tag{3.15}$$

$$\mathbf{z} \sim \text{Gibbs-type}(\beta). \tag{3.16}$$

The likelihood and the full-conditional distributions of the bESBM are available, and it is possible to set up a standard Gibbs sampler to learn the posterior distribution $p(\mathbf{z}|Y^*)$. First of all, the cluster-specific probabilities Φ_{hk}, which are not of direct interest here, are marginalized out from Equation (3.14), yielding

$$\begin{aligned} p(Y^*|\mathbf{z}) = \prod_{h=1}^{H}\prod_{k=1}^{h-1}\Big\{\prod_{u,v:z_u=h;z_v=k}\binom{K}{y_{uv}}\Big\}\frac{B(m_{hk}+a, Kn_hn_k-m_{hk}+b)}{B(a,b)}\cdot \\ \cdot\Big\{\prod_{u<v:z_u=z_v=h}\binom{K}{y_{uv}}\Big\}\frac{B(m_{hh}+a,(K/2)(n_h-1)n_h-m_{hh}+b)}{B(a,b)}, \end{aligned} \tag{3.17}$$

where $B(\cdot,\cdot)$ is the Beta function, m_{hk} is the sum of the weights of the edges connecting nodes in cluster h and nodes in cluster k, while m_{hh} is the sum of edge weights within cluster h. Using Equation (3.17), we derive a collapsed Gibbs sampler that, at every iteration, updates the cluster membership of each node according to its full-conditional distribution

$$p(z_v = h|\mathbf{z}^{-v}, X, Y^*) \propto p(z_v = h|\mathbf{z}^{-v}, X)\frac{p(Y^*|\mathbf{z}^{-v}, z_v = h)}{p(Y^{*,-v}|\mathbf{z}^{-v})}, \tag{3.18}$$

where, as usual, the superscript $-v$ denotes quantities computed excluding node v. The final ratio in Formula (3.18) is computed using (3.17), yielding

$$\begin{aligned} &\frac{p(Y^*|z_v = l, \mathbf{z}^{-v})}{p(Y^{*,-v}|\mathbf{z}^{-v})} = \\ &= \prod_{j\neq v}\binom{K}{y_{vj}}\prod_{\substack{k=1\\k\neq l}}^{H^{-v}}\frac{B\big(m_{kl}^{-v}+\sum_{i<v,i:\mathbf{z}_i=k}y_{vj}^*+a, K(n_l^{-v}+1)n_k^{-v}-m_{kh}^{-v}-\sum_{i<v,i:\mathbf{z}_i=k}y_{vj}^*+b\big)}{B\big(m_{kl}^{-v}+a, K(n_l^{-v})n_k^{-v}-m_{kh}^{-v}+b\big)}\cdot \\ &\qquad\cdot\frac{B\big(m_{ll}^{-v}+\sum_{i\neq v,i:\mathbf{z}_i=l}y_{vj}^*+a, K\frac{n_l^{-v}(n_l^{-v}+1)}{2}-m_{ll}^{-v}-\sum_{i\neq v,i:\mathbf{z}_i=l}y_{vj}^*+b\big)}{B\big(m_{ll}^{-v}+a, K\frac{n_l^{-v}(n_l^{-v}-1)}{2}-m_{ll}^{-v}+b\big)}. \end{aligned}$$

Moreover, recall that the urn scheme associated to a Gibbs-type distribution is:

$$p(z_v = h|\mathbf{z}^{-v}, X) \propto \begin{cases} \mathscr{W}_{V,H^{-v}}(n_h^{-v}-\sigma) & \text{for } h = 1,\dots,H^{-v}, \\ \mathscr{W}_{V,H^{-v}+1} & \text{for } h = H^{-v}+1, \end{cases}$$

where $\mathscr{W}_{V,H}$ and σ are determined by the Gibbs-type prior of choice; see, e.g., Legramanti et al. [2022]. Once again, given the posterior samples of the partition $\mathbf{z}$, the point posterior estimate is found through the minimization of the Variation of Information measure (see Section 3.4.4):

$$\hat{\mathbf{z}}_0^* = \operatorname*{argmin}_{\mathbf{z}^*}\mathbb{E}_{\mathbf{z}}[\mathrm{VI}(\mathbf{z},\mathbf{z}^*)].$$

Then, the partition $\hat{\mathbf{z}}_0^*$ estimated using the bESBM is compared to the common partition $\hat{\mathbf{z}}_0$ output by the mESBM. The rationale is the comparison between two partitions that are obtained by exploiting information across all layers $Y_1,\ldots,Y_K$ of the multiplex network in two completely different ways: for the bESBM, we combine the layer-specific data through the collapsed graph, yielding an unquantifiable loss of information a priori, before the estimation of the model. On the contrary, the mESBM is able to borrow information across layers to estimate the common clustering of the nodes, thanks to the dependency induced by its hierarchical definition, maintaining the original data structure.

3.5.2 Layer-wise extended stochastic block models (ESBMs)

To assess the layer-wise specific clustering, we compare the posterior estimates of $\mathbf{z}_1,\ldots,\mathbf{z}_K$ to the clusters obtained by separately fitting a standard, independent ESBM for each layer. We follow the approach in Legramanti et al. [2022], estimating K independent partitions as:

$$\begin{aligned} Y_{k,uv} | z_{ku} = h, z_{kv} = l &\overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\Phi_{k,hl}), \\ \Phi_{k,hl} | \mathbf{z}_k &\overset{\text{i.i.d.}}{\sim} \text{Beta}(a,b), \\ \mathbf{z}_k &\sim \text{Gibbs-type distribution}(\beta), \end{aligned}$$

for $1 \leq h \leq l \leq H_k$ and $k = 1,\ldots,K$, where Φ_k is a $H_k \times H_k$ matrix containing the block probabilities. For the posterior estimation, we run the Gibbs sampler in Legramanti et al. [2022], and to find the point posterior partition we use the same approach of Section 3.4.4:

$$\hat{\mathbf{z}}_k^* = \underset{\mathbf{z}_k^*}{\operatorname{argmin}} \mathbb{E}_{\mathbf{z}_k} \left[\text{VI}(\mathbf{z}_k, \mathbf{z}_k^*) \right], \quad k = 1,\ldots,K.$$

As already mentioned, we compare the partitions estimated by independent ESBMs with the layer-specific clusters of the mESBM: the rationale is to see what is the impact of borrowing of information across layers for the estimation of the corresponding clustering, since we expect that exploiting the identity of the nodes in a multiplex network through an induced dependency on the partitions should provide more meaningful groups.

3.6 Simulation studies

To assess the performance of the mESBM with respect to the competitors described in Section 3.5, we consider simulated multiplex networks with $V = 60$ nodes and $K = 10$ layers. We perform two simulation studies: first, we consider the fully-informed case (Scenario 1), where all the layer partitions are actually centered on the same $\mathbf{z}_0$. Second, we introduce some noise in a half-informed case (Scenario 2), considering $K/2$ networks centered on $\mathbf{z}_0$, and the rest showing random edges. The rationale behind the data generation process is to produce adjacency matrices resembling the data of interest as much as possible (see Section 3.7), with two different levels of noise.

Data generation process The data considered in this simulation study are generated according to the scheme below:

1. Fix a (true) value for the common partition $\mathbf{z}_0$ of the V nodes. In this simulation, $\mathbf{z}_0$ splits the V nodes into four clusters of equal size. Figure 3.1 represents the true simulated coclustering structure.

2. Generate the layer-specific partitions $\mathbf{z}_1, \dots, \mathbf{z}_K$ as follows: for $k = 1, \dots, K$, first set $\mathbf{z}_k = \mathbf{z}_0$ and then change independently the label of $M = 10$ randomly chosen nodes. The initial, coclustering structure of the simulated partitions $\mathbf{z} = [\mathbf{z}_1, \dots, \mathbf{z}_K]'$ is displayed in Figure 3.3. Each layer-specific partition contains $H_k = 4$ different clusters, for $k = 1, \dots, K$.

3. Use a SBM [Nowicki and Snijders, 2001] to generate the layer-specific adjacency matrices $Y = [Y_1, \dots, Y_K]$ or $Y = [Y_1, \dots, Y_{K/2}]$, respectively for Scenario 1 and Scenario 2. First, fix the matrix $\Phi_{4\times 4}$, containing the block probabilities — see Figure 3.2:

$$\Phi = \begin{bmatrix} 0.8 & 0.7 & 0.4 & 0.1 \\ 0.7 & 0.8 & 0.1 & 0.4 \\ 0.4 & 0.1 & 0.8 & 0.7 \\ 0.1 & 0.4 & 0.7 & 0.8 \end{bmatrix}.$$

   Then, for Scenario 1, generate the $V \times V$ adjacency matrices $Y_1, \dots, Y_K$ using a SBM:

$$Y_{k,uv} | \mathbf{z}_k \sim \text{Bernoulli}(\Phi_{z_{ku}, z_{kv}}) \quad \text{for } k = 1, \dots, K; \quad u, v = 1, \dots, V.$$

   For Scenario 2, generate the first $K/2$ networks $Y_1, \dots, Y_{K/2}$ as above, and then simulate the remaining networks $Y_{K/2+1}, \dots, Y_K$ using a random Bernoulli distribution, with parameter 1/2:

$$Y_{k,uv} | \mathbf{z}_k \sim \text{Bernoulli}(1/2) \quad \text{for } k = K/2 + 1, \dots, K; \quad u, v = 1, \dots, V.$$

   The simulated multiplex networks for Scenario 1 and Scenario 2 are shown in Figures 3.4 and 3.5, respectively. It is clear that Scenario 1 is a fully-informed multiplex network: all the layers are centered on the same partition. Scenario 2 is noisier: half of the layers actually provide information on $\mathbf{z}_0$, while the other half is totally random.

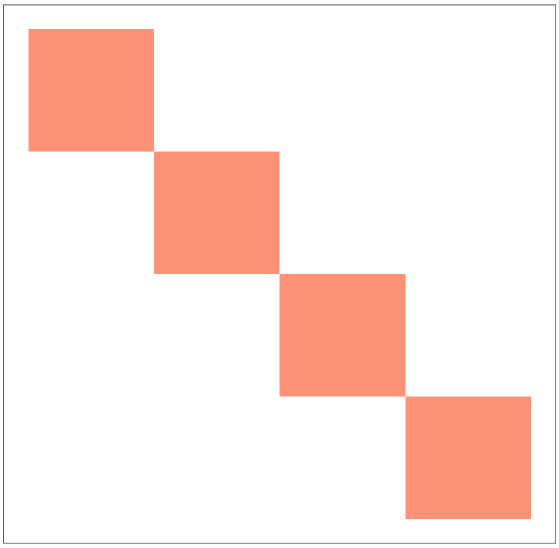


Figure 3.1: True coclustering structure of $\mathbf{z}_0$.

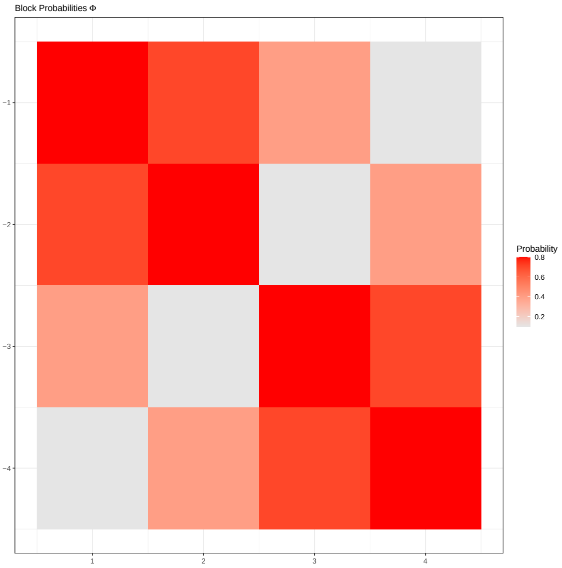


Figure 3.2: Block probabilities Φ.

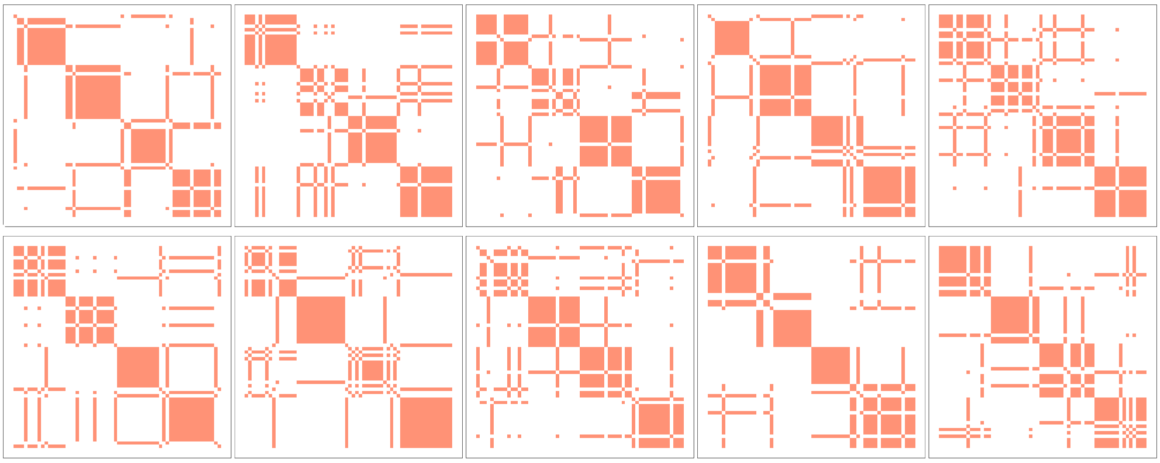


Figure 3.3: True layer-specific coclustering structures of $\mathbf{z}_1, \ldots, \mathbf{z}_{10}$.

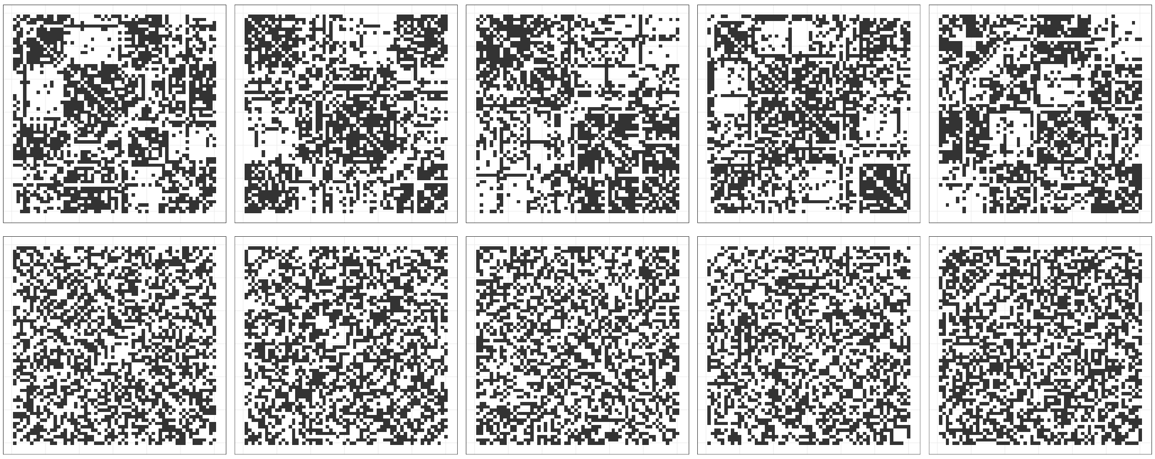


Figure 3.4: Simulated multiplex network $Y = [Y_1, \ldots, Y_{10}]$ under the first fully-informed scenario described in Section 3.6 — Scenario 1.

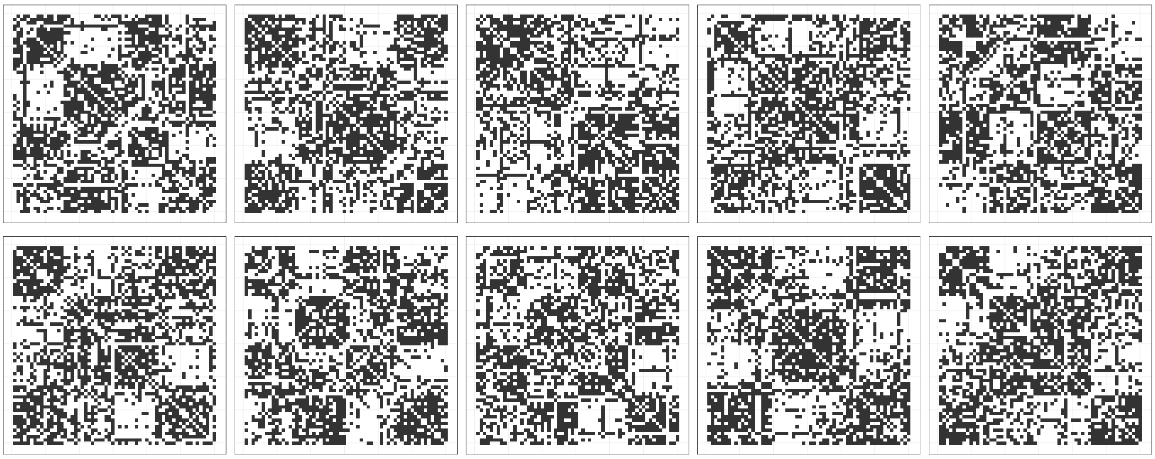


Figure 3.5: Simulated multiplex network $Y = [Y_1, \ldots, Y_{10}]$ under the second half-informed scenario described in Section 3.6 — Scenario 2.

Posterior estimation The mDPDP model in Equation (3.3) is estimated on the artificial multiplex networks using Algorithm 1. We run the algorithm for $10'000$ iterations, with $1'000$ iterations as burn-in, and with $N = 500$. Figures 3.6 and 3.7 show the posterior coclustering matrix of the approximate samples from $p(\mathbf{z}_0|\mathbf{z})$ in the two cases. The posterior samples clearly distinguish the four different clusters underlying the data generation process in Figure 3.1, with a higher, but still acceptable, level of uncertainty in the noisier case. Hence, in both Scenarios 1 and 2, the true underlying common partition is fully recovered.

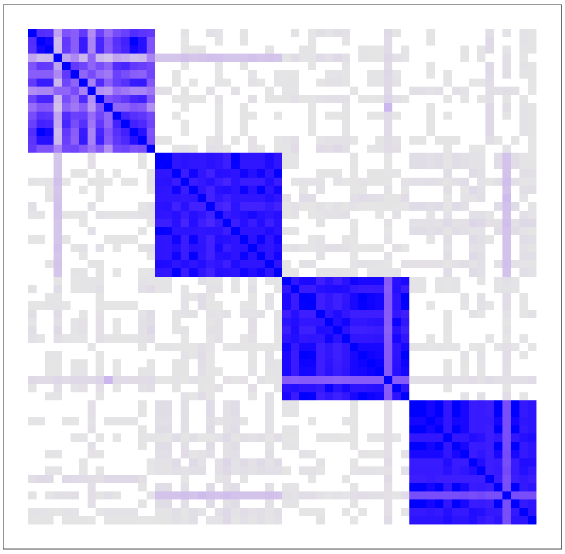


Figure 3.6: Posterior coclustering matrix of $\mathbf{z}_0$ under the multiplex extended stochastic block model — Scenario 1.

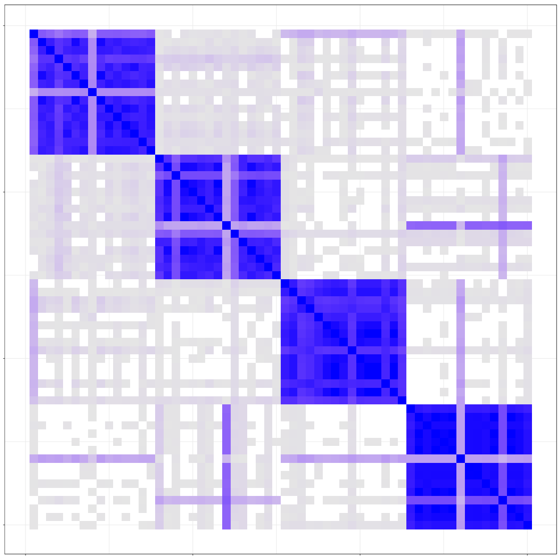


Figure 3.7: Posterior coclustering matrix of $\mathbf{z}_0$ under the multiplex extended stochastic block model — Scenario 2.

Figures 3.8 and 3.9 display the empirical coclustering matrices of the (approximate) posterior samples from $p(\mathbf{z}_k|Y,\mathbf{z}_0)$, for $k = 1,\ldots,K$ in the two scenarios of interest. Starting with Scenario 1, and comparing the posterior results to the true value of $\mathbf{z}$ in Figure 3.3, one can see that they are mostly similar, minus some noise (indicating exploration of the partition space during the estimation procedure). For example, in layers 5 and 7, we have some uncertainty in the off-diagonal blocks, probably due to the corresponding edge structure in the adjacency matrices. Nevertheless, the posterior estimation of $\mathbf{z},\mathbf{z}_0$ well reproduces all the true partitions underlying the data generating process. The same holds for Scenario 2: the true clustering structures are well recovered for the first $K/2$ networks, the ones where the edges are informed by the layer-specific partitions. On the other hand, the last $K/2$ partitions are non-existent, since the edges are random and not influenced by any specific clustering of the nodes. Therefore, the mDPDP model successfully captures all the existing layer-specific latent partitions, whether in scenarios with or without noise.

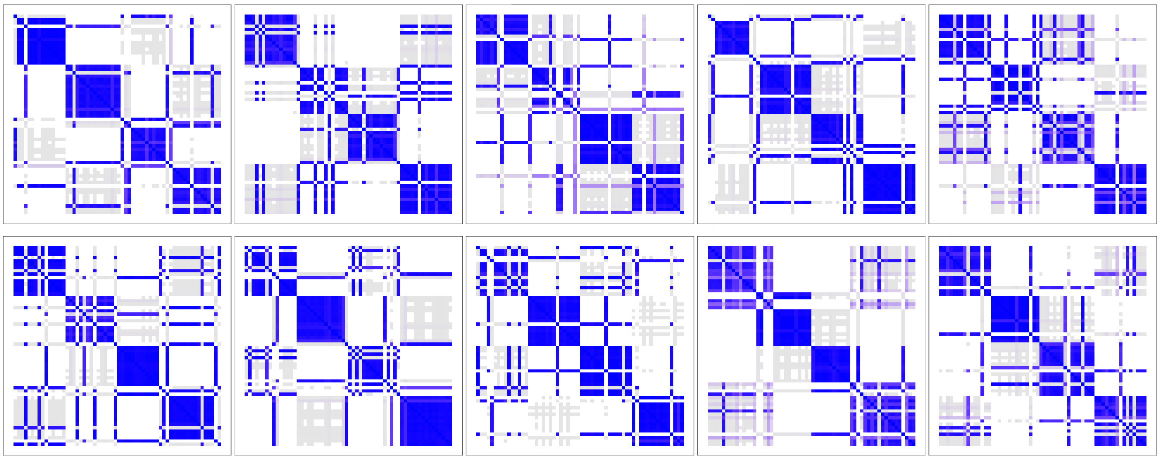


Figure 3.8: Posterior coclustering matrices of $\mathbf{z}_1, \ldots, \mathbf{z}_K$ under the multiplex extended stochastic block model — Scenario 1.

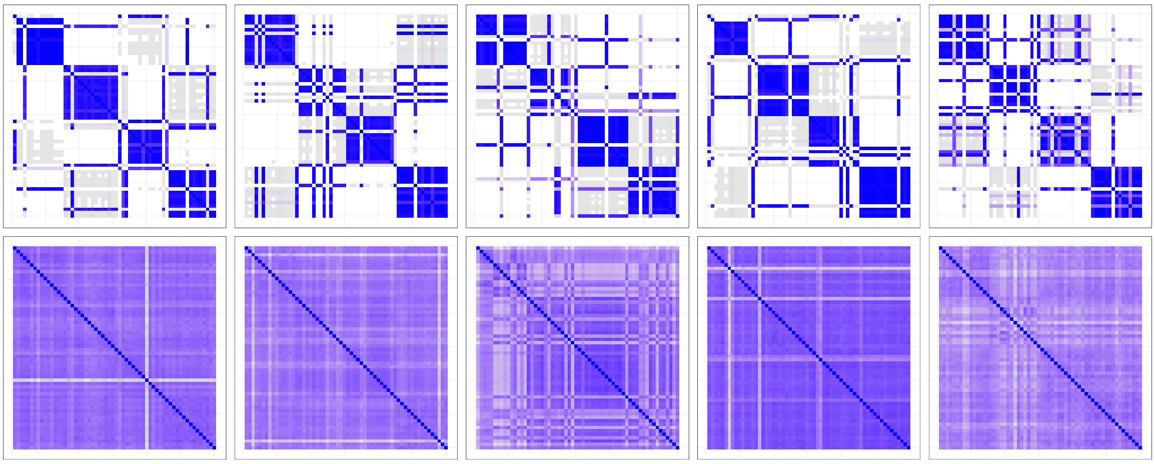


Figure 3.9: Posterior coclustering matrices of $\mathbf{z}_1, \ldots, \mathbf{z}_K$ under the multiplex extended stochastic block model — Scenario 2.

Competitors As already mentioned in the previous sections, we also compare the accuracy of the mESBM in two simulated scenarios to the performances of the two competitors mentioned in Section 3.5. In particular, recall that the posterior estimation of $\mathbf{z}_0$ is compared with the partition provided by the bESBM on the collapsed network (Figures 3.10, 3.11), and the posterior layer-specific partitions $\mathbf{z}_1, \ldots, \mathbf{z}_K$ are matched to the clustering estimated by independent ESBMs on each layer of the simulated multiplex network (Figures 3.14, 3.15). Both the competitor models are endowed with a Dirichlet process prior on the partitions, with a concentration parameter equal to 1. The corresponding posterior distributions are estimated using 10′000 iterations of the suitable Gibbs samplers, with 1′000 iterations as burn-in. The posterior partitions estimated by competitor models are displayed in Figures 3.12, 3.13 and 3.14, 3.15 respectively.

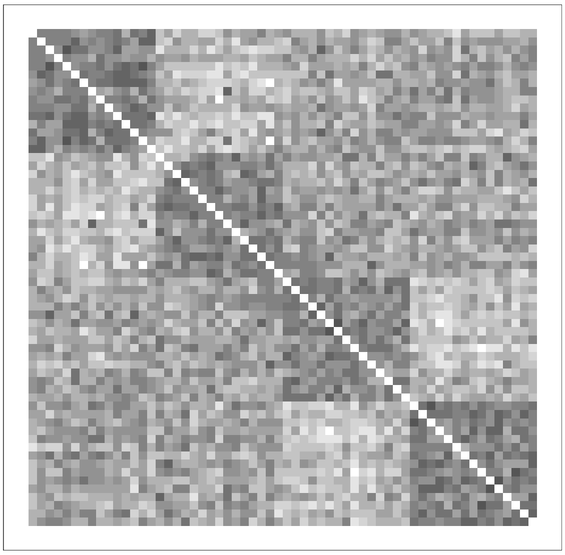


Figure 3.10: Collapsed adjacency matrix of the simulated network — Scenario 1.

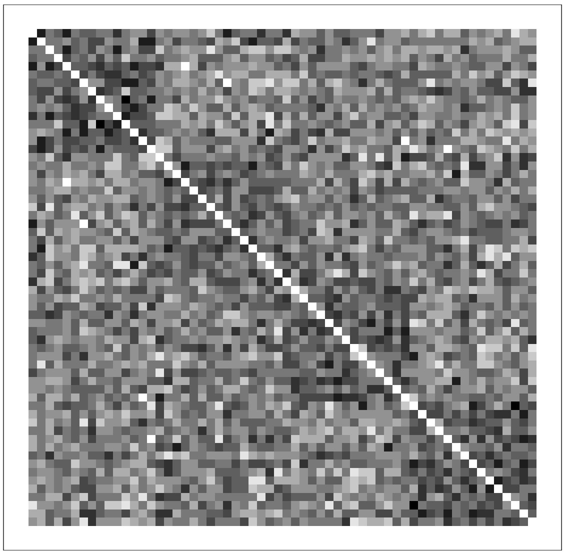


Figure 3.11: Collapsed adjacency matrix of the simulated network — Scenario 2.

Starting from Scenario 1, the posterior coclustering matrix of $\mathbf{z}_0^*$ provided by the bESBM in Figure 3.12 is overconfident: this could be a sign of a limited exploration of the partition space. On the contrary, in Figure 3.14, it is clear that the posterior estimations of the layer-specific partitions are way more variable (and less accurate) than the ones provided by the mESBM. Thus, introducing dependence among the layers of the multiplex network through $\mathbf{z}_0$ seems to both ameliorate the posterior estimations of the layer-specific partitions, making them more robust, and improve the exploration of the support during the Monte Carlo estimation of the common clustering. As for Scenario 2, Figure 3.13 shows that the bESBM is not able to recover the underlying common partition $\mathbf{z}_0$: hence, in presence of moderate noise, the bESBM can not retrieve the correct clustering. Moreover, Figure 3.15 shows that also the layer-specific partitions are not perfectly retrieved. In particular, they

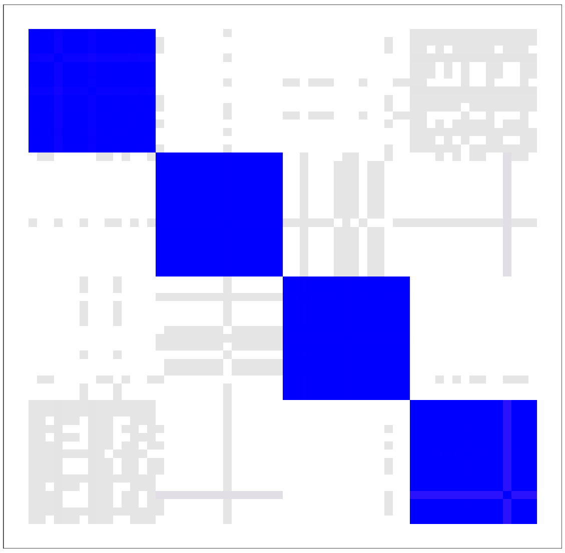


Figure 3.12: Posterior coclustering matrix of $\mathbf{z}_0^*$ under the binomial extended stochastic block model — Scenario 1.

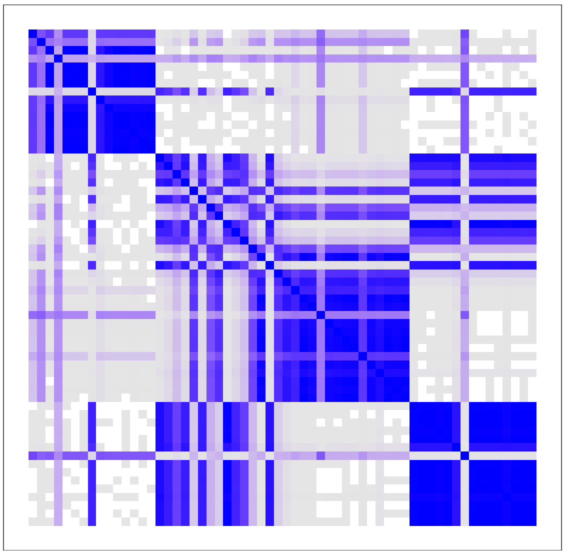


Figure 3.13: Posterior coclustering matrix of $\mathbf{z}_0^*$ under the binomial extended stochastic block model — Scenario 2.

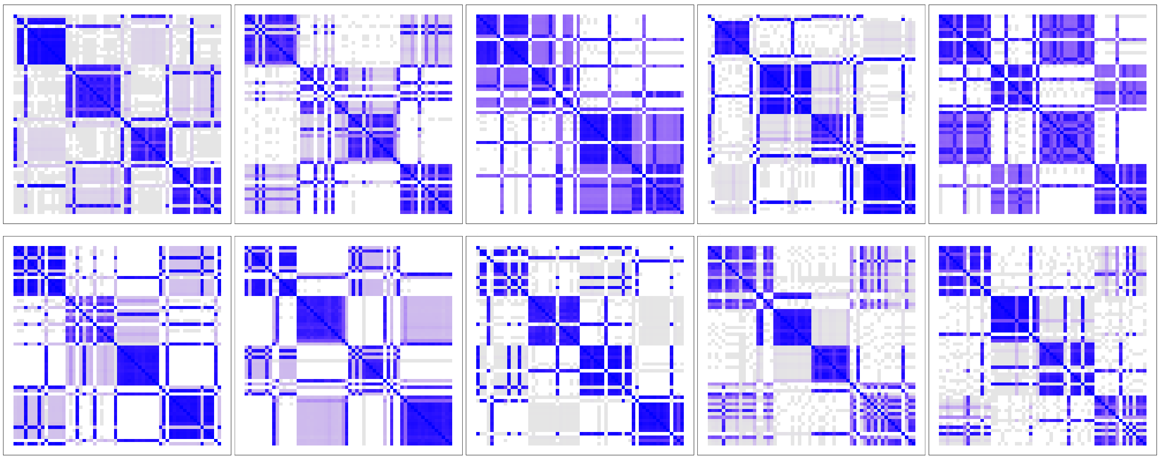


Figure 3.14: Posterior coclustering matrices of $\mathbf{z}_1^*,\dots,\mathbf{z}_K^*$ under independent, layer-wise extended stochastic block models — Scenario 1.

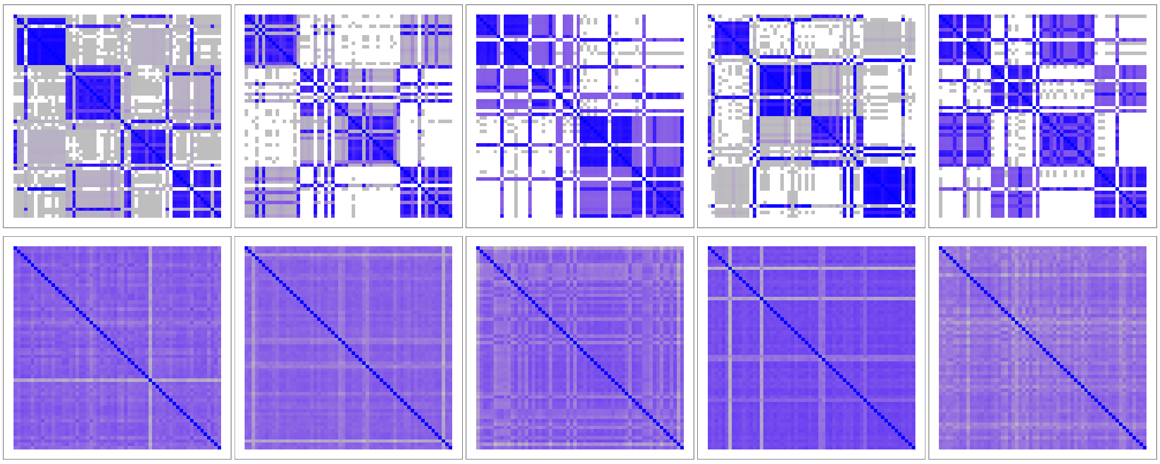


Figure 3.15: Posterior coclustering matrices of $\mathbf{z}_1^*, \ldots, \mathbf{z}_K^*$ under independent, layer-wise extended stochastic block models — Scenario 2.

show much more uncertainty and a lower accuracy with respect to the one obtained using the mESBM.

3.7 Application: human brain networks

In this section, we apply the mDPDP model described in Section 3.3 to a multiplex network representing the brain structural connections of different patients, collected using Diffusion Tensor Imaging.

3.7.1 Diffusion Tensor Imaging data

The data comes from a pilot study of the Enhanced Nathan Kline Institute-Rockland Sample project, whose description can be found at `http://fcon_1000.projects.nitrc.org/indi/enhanced/`. Specifically, we are interested in the pilot `NKI1` study, which comprises multimodal imaging data and subject-specific covariates for 24 subjects. Detailed information about the study can be found at `http://fcon_1000.projects.nitrc.org/indi/CoRR/html/nki_1.html`. The goal of this work is to study the structural networks of the brain of 24 different subjects, making use of measures of anatomical interconnection provided by white matter fibers among 68 cerebral regions of interest. In this dataset, the parcellation of the human brain follows the Desikan atlas nomenclature [Thomas Yeo et al., 2008]. By applying the model proposed in Section 3.3, the goal is to estimate a) the layer-specific partitions $\mathbf{z}$ to cluster areas of the brain for each subject and b) a common clustering $\mathbf{z}_0$ of the human brain, common to all the patients involved in the study. We conjecture the existence of a common partition $\mathbf{z}_0$ providing a physical division of the brain shared by all humans, as well as of different patient-specific partitions of the cerebral regions $\mathbf{z}_1, \ldots, \mathbf{z}_K$, influenced by the personal diagnoses of each subject. The data are collected using Diffusion Tensor Imaging (DTI) scans, with the aim of studying underlying structural connectivity patterns within the brain of each subject. DTI maps the diffusion of water molecules across the biological brain tissues, thereby allowing reconstruction of the white matter fibers which act as highways for the directional diffusion of water within the brain. The dataset of interest provides two different scans per patient, but we consider just the first one. Along the DTI results, the data contain some additional information, such as the lobe membership and 3D coordinates of each brain area or some subject-specific features. There are two regions (for the left and for the right hemisphere) marked as `unknown`, which are not taken into consideration in the subsequent analyses. Also, the DTI scans for four subjects are completely missing, and thus discarded from the beginning. The final multiplex network $Y = [Y_1, \ldots, Y_{20}]$ consists of $K = 20$ layers (one per each subject with non-missing information), each one encompassing a graph with 68 nodes (i.e. the brain areas of interest) and a variable number of edges representing the white matter fiber interconnections shown by the corresponding patient. Each graph is represented by its adjacency matrix $Y_1, \ldots, Y_K$. Originally, each element $Y_{k,uv}$ of Y_k denotes the number of white matter fibers connecting the corresponding pair of brain regions in subject k, for $k = 1, \ldots, 20$ and $u, v = 1, \ldots, 68$. Thus, each Y_k is technically a weighted adjacency matrix; however, we observe that the structural networks are sparse, with a lot of fiber counts being zero, and the others having a wide range of variability (we can have from 1 up to $\sim 30'000$ fibers connecting two regions). We then focus on a binary structural network, which simply detects the presence or absence of white fiber matters linking two brain areas. The final binary multiplex network is displayed in Figure 3.16.

3.7.2 Posterior estimation and inference

To analyze the data presented in Section 3.7.1, we use the mDPDP model in Equation (3.3). The following hyperparameters are fixed: $a = 1, b = 1$. The resulting setting is an uniform prior distribution on the block connections, which allows for the search of various cluster connectivity patterns and for easier generalization. In fact, in most applications, it is not clear a priori what type of clusters one should look for. In particular, notice that the mESBM is not finding communities as standardly defined in graphs, i.e. clusters of nodes densely connected within but not between. The mESBM aims at finding clusters with similar connectivity patterns, a general framework that encompasses both assortative and disassortative community structures, among others. Moreover, the concentration parameters of the DPs are set to $\gamma = \beta = 1$: different experiments have been performed varying such values, but the resulting partitions are pretty robust, possibly due to the fact that network data contain a lot of information, resulting in the likelihood value overwhelming the prior in the posterior computation.

As for the parameter α of the similarity function, which is regulating the borrowing of information mechanism (see Section 3.3.2), we experimented with different values: we

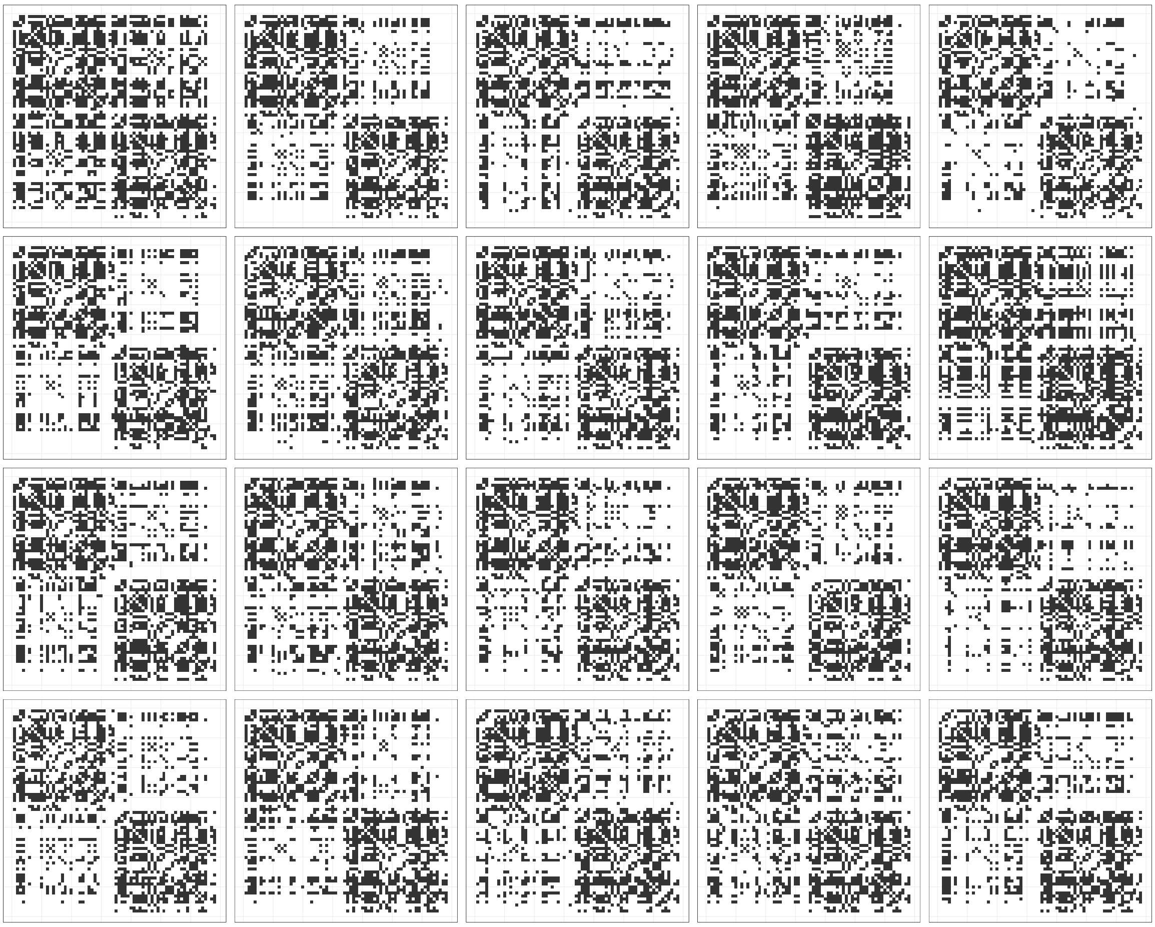


Figure 3.16: Multiplex network from Diffusion Tensor Imaging data represented through layer-specific adjacency matrices Y_k for each subject $k = 1, \ldots, 20$.

used $\alpha = 1$, $\alpha = 3$, $\alpha = 5$ and the extreme case $\alpha = 30$. For sensible values of α (i.e. $\alpha = 1,3,5$) the results are coherent and reliable, both in terms of the likelihood $p(Y|\mathbf{z})$ and of posterior coclustering matrices. Hence we discuss the results for $\alpha = 1$ in the remainder. To estimate the mDPDP model, we run Algorithm 1 for 50′000 iterations, with the parameter of SIR set to $N = 500$. The initialization of $\mathbf{z}_0$ is a random partition of the nodes into 10 clusters, while the initialization for $\mathbf{z}_k$, $k = 1,\ldots,K$ is provided by V singletons. Figure 3.17 displays the traceplot of the (unnormalised) log-likelihood $\log q(Y|\mathbf{z}) = \sum_{k=1}^{K} \log q(Y_k|\mathbf{z}_k)$, with 10′000 iterations of burn-in. The traceplot does not point towards issues of non-convergence.

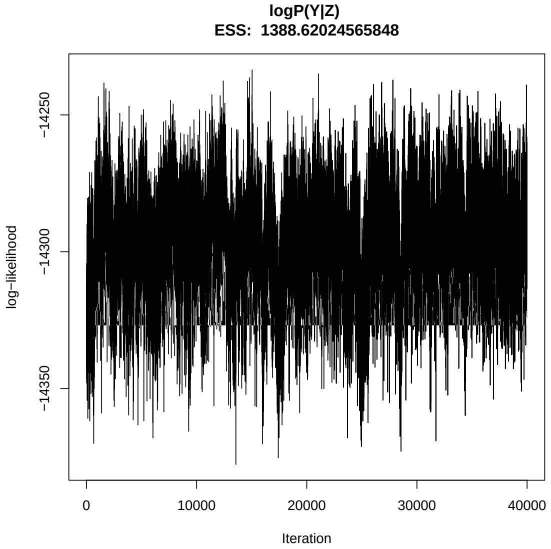


Figure 3.17: Traceplot of the log-likelihood $\log q(Y|\mathbf{z})$, across 50′000 iterations and with 10′000 iterations of burn-in.

Inference on the layer-specific partitions z Figure 3.18 displays the posterior coclustering matrices of the (approximate) samples $\mathbf{z}_1,\ldots,\mathbf{z}_K$ from $p(\mathbf{z}_1,\ldots,\mathbf{z}_K|Y,\mathbf{z}_0)$, obtained through Algorithm 1. Most of the subjects show two diagonal blocks, denoting similar structural connections among areas of the same hemisphere. Nevertheless, some patients display layer-specific partitions clustering together areas from opposite hemispheres (e.g. patients 1, 12, 16, 19), while other subjects (e.g. patients 6, 15, 17, 18) experience a partition with most of the clusters being included in the same half of the brain. Besides DTI scans for each patient, the dataset also contains some information about the subjects involved in the study: in particular, patients 6 and 17 experience alcohol and cannabis abuse, subject 7 is diagnosed with a major depressive disorder (MDD) as well as an eating disorder and cannabis dependence. Subject 13 also suffers from MDD. We argue that the areas for which the subject-specific partitions $\mathbf{z}_1,\ldots,\mathbf{z}_K$ vary the most across different patients, are also the ones commonly impacted by substance abuse and/or depressive disorders,

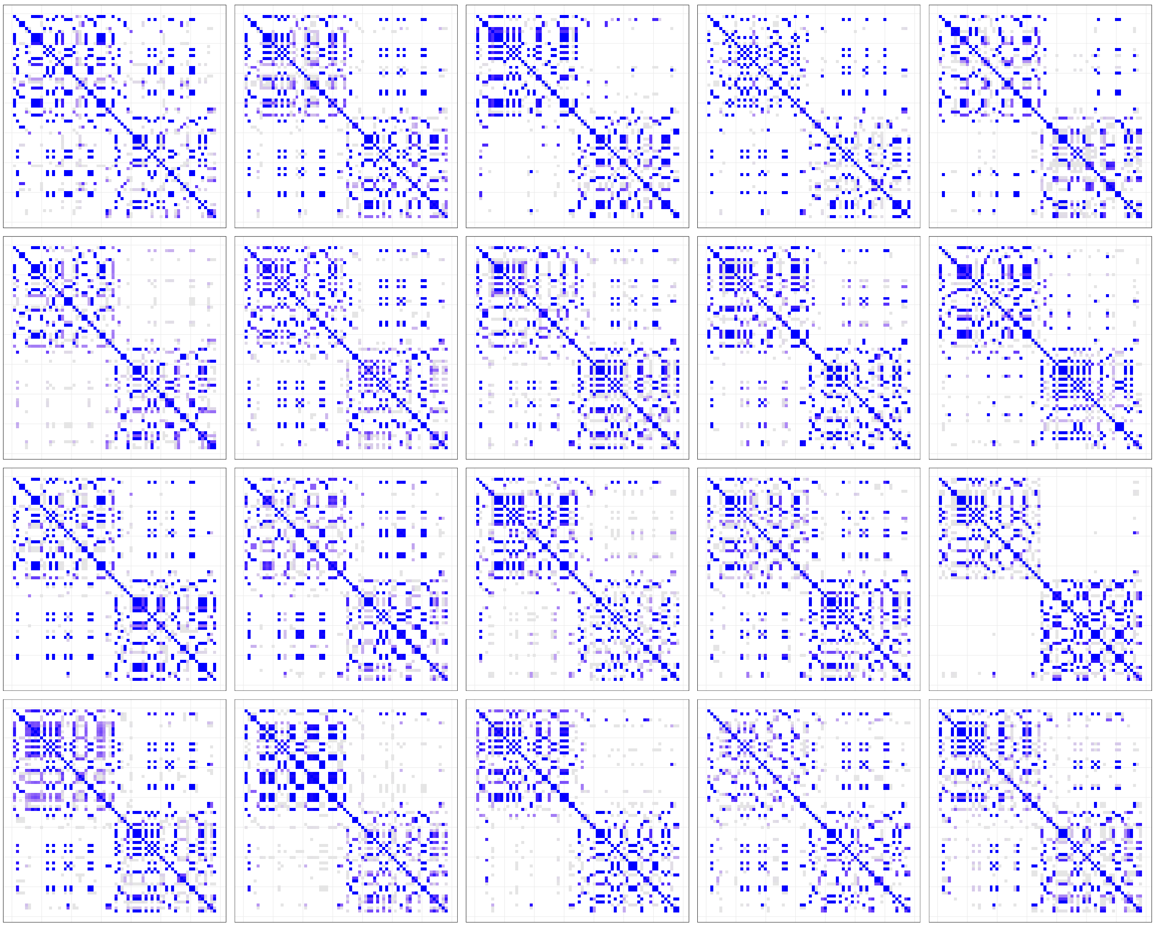


Figure 3.18: Posterior coclustering matrices of the samples $\mathbf{z}_1,\ldots,\mathbf{z}_K$ under the multiplex extended stochastic block model.

according to the literature. In particular, analysing the regions with the most uncertain cluster labelling (i.e. the areas where the mean posterior pairwise coclustering proportion across $\mathbf{z}$ is $\in (0.45, 0.55)$ with at least 10 nodes), one can notice that the brain zones that exhibit the most dubious clustering patterns across subjects belong to the *temporal* and *limbic* system. The temporal and limbic systems (containing the hippocampus and the amygdala, performing a primary role in decision making and emotional responses) are the parts of the brain which are most affected by both drug addiction [Franklin et al., 2002, Unterrainer et al., 2019] and MDD [O'Shea et al., 2018, Kim and Park, 2021]. In particular, it is known that patients with depressive disorders show increased activation in the limbic system, specifically in the amygdala, insula, and hippocampus, compared with healthy controls [Bellani et al., 2010, Zamoscik et al., 2014, Lemke et al., 2022]. MDD has also been linked to structural changes in temporal brain regions [Caetano et al., 2007, Garcia, 2012, Ramezani et al., 2014, Barbosa et al., 2021]. For a more in-depth analysis of the impact of MDD on brain areas, see Zeng et al. [2012], Helm et al. [2018] and Zhang et al. [2018].

Besides the coclustering matrices of the posterior samples of the layer-specific partitions $\mathbf{z}$ given $Y, \mathbf{z}_0$, we can analyze the point posterior estimates $\hat{\mathbf{z}}$ obtained through the minimization of the Variation of Information measure, as explained in Section 3.4.4. Figure 3.19 shows the brain map of each subject, where the nodes of the graph are placed in the correct bidimensional coordinates of the corresponding cerebral regions, while the edges represent subject-specific interconnections. Node colors correspond to the posterior point estimate of the layer-specific partitions $\hat{\mathbf{z}} = [\hat{\mathbf{z}}_1, \dots, \hat{\mathbf{z}}_K]'$. Table 3.1 and Figures 3.20, 3.21 show a preliminary analysis of the Variation of Information distance between each layer-specific partition $\mathbf{z}_1, \dots, \mathbf{z}_K$ and the common partition $\mathbf{z}_0$. In particular, Table 3.1 shows the summary statistics of the quantity $\text{VI}(\hat{\mathbf{z}}_0, \hat{\mathbf{z}}_k)$ for $k = 1, \dots, K$. Notice that the maximum possible value for such quantity is equal to $\log_2 V = 4.322$. Moreover, Figure 3.20 shows the values and the boxplot of $\text{VI}(\hat{\mathbf{z}}_0, \hat{\mathbf{z}}_k)$ for $k = 1, \dots, 20$. Some layer-specific partitions are closer to the common clustering estimated by the mESBM (e.g. $\mathbf{z}_7, \mathbf{z}_9, \mathbf{z}_{16}$), while other ones are more different (e.g. $\mathbf{z}_1, \mathbf{z}_5, \mathbf{z}_6, \mathbf{z}_{18}$). Finally, Figure 3.21 displays the empirical distributions of the Variation of Information between the common clustering $\hat{\mathbf{z}}_0$ estimated by the mESBM and each sampled posterior partitions $\mathbf{z}_k^t$, for $k = 1, \dots, K$ and $t = 1, \dots,$ #iterations. In this case, the distance of some layer-specific posterior samples from the common cluster $\hat{\mathbf{z}}_0$ show a higher variability of (e.g. for subject 2, 7, 12, 19) with respect to others (e.g. subject 6, 15).

Minimum	First quartile	Mean	Median	Third quartile	Maximum
0.958	1.327	1.447	1.435	1.678	1.766

Table 3.1: Summary statistics of the posterior Variation of Information distance $\text{VI}(\hat{\mathbf{z}}_0, \hat{\mathbf{z}}_k)$ between the layer-specific partitions $\hat{\mathbf{z}}, \dots, \hat{\mathbf{z}}_K$ and the common partition $\hat{\mathbf{z}}_0$.

Inference on the common partition $\mathbf{z}_0$ The most innovative part of the mESBM is the possibility to infer the common partition $\mathbf{z}_0$. Figure 3.22 displays the tering matrix of the

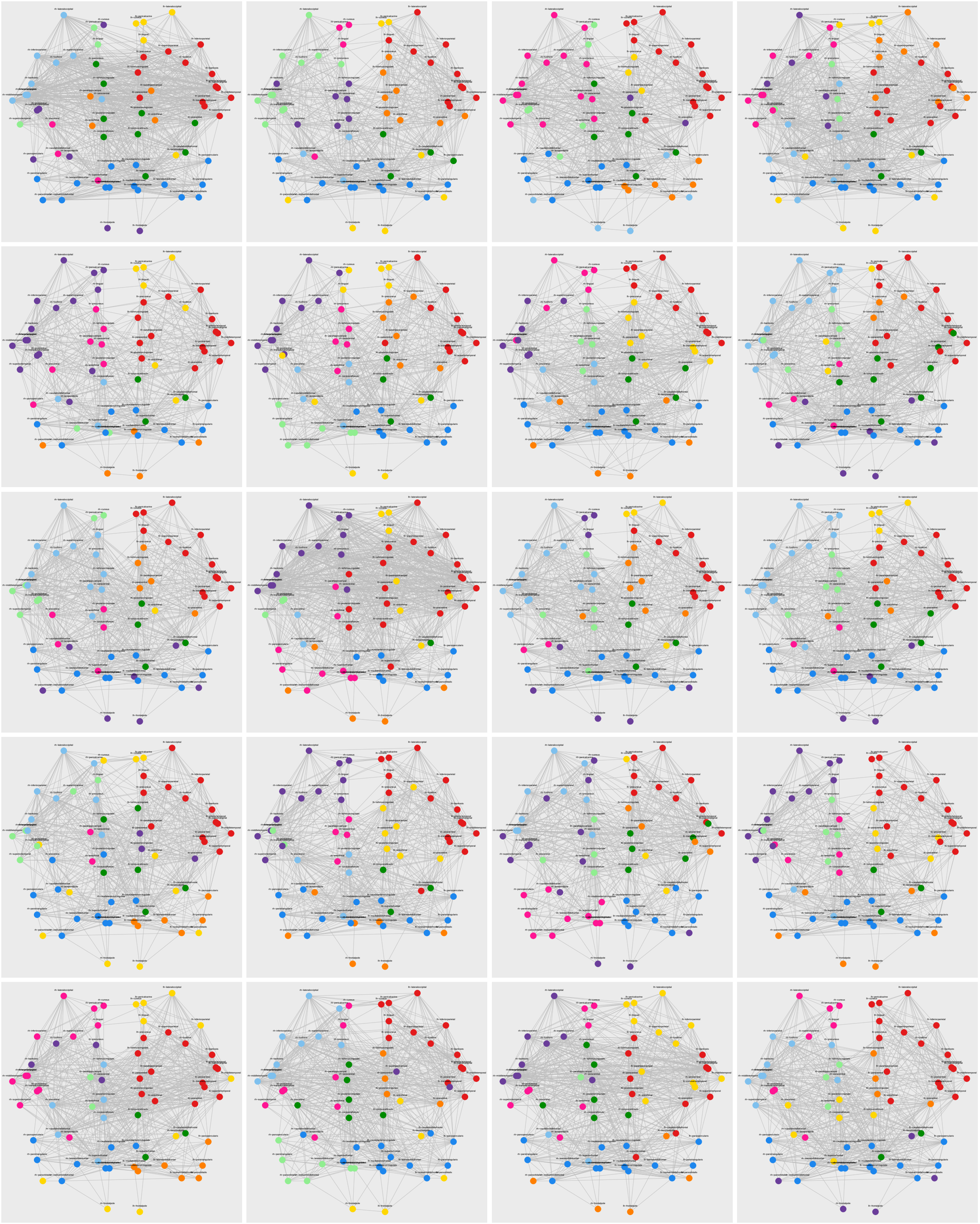


Figure 3.19: Graphical representation of the subject-specific brain networks: node colors correspond to the posterior point estimate of the layer-specific partition $\hat{\mathbf{z}}$ under the multiplex extended stochastic block model.

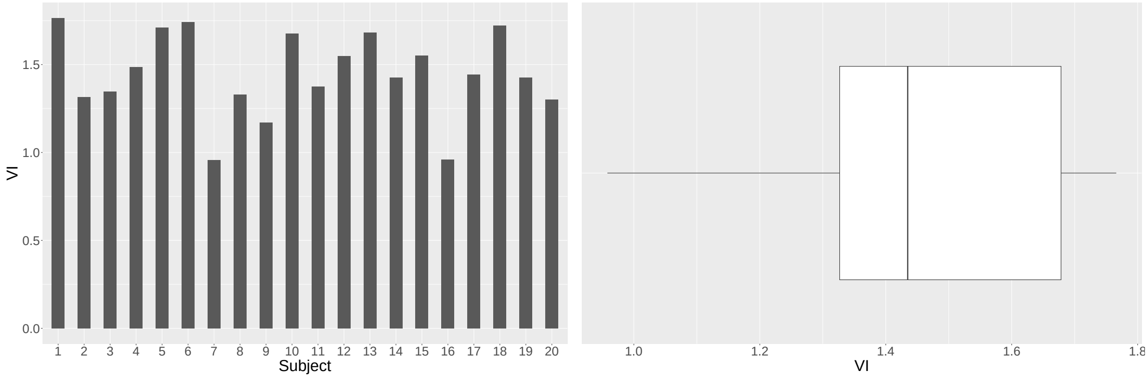


Figure 3.20: Values and empirical distribution of the posterior Variation of Information distance VI($\hat{\mathbf{z}}_0, \hat{\mathbf{z}}_k$) between the layer-specific partitions $\hat{\mathbf{z}}, \dots, \hat{\mathbf{z}}_K$ and the common partition $\hat{\mathbf{z}}_0$.

approximate samples $\mathbf{z}_0 \sim p(\mathbf{z}_0 | \mathbf{z}, Y)$, obtained using Algorithm 1. The x-axis and y-axis display the names of the cerebral regions, in the fashion of the Desikan atlas. The acronyms `lh` and `rh` stand for `left hemisphere` and `right hemisphere`, respectively. Looking at the reordered posterior coclustering matrix in Figure 3.22, the diagonal blocks show that most of the clusters include areas within the same hemisphere, on both the left and the right side. However, some areas of the brain are grouped together with their counterparts in the opposite hemisphere for most of the iterations: in particular, the `rostral anterior cingulate` and `rostral middle frontal` in the right hemisphere cluster with their counterparts in left hemisphere, and the same happens with the `caudal anterior cingulate, medial orbito-frontal, parsopercularis, parstriangularis`. Such a partition is evident looking at the point estimate of the posterior partition $\hat{\mathbf{z}}_0$, obtained once again through the minimization of the Variation of Information measure (see Section 3.4.4). Figure 3.23 shows a brain network (in this case, of subject 1), where each node represents a brain area and it is plotted in the actual bidimensional coordinates of its position in the human anatomy. The edges denote the presence of white fiber connections in the patient. The color of each node corresponds to the cluster labelling according to the point posterior estimate $\hat{\mathbf{z}}_0$. Analyzing the brain map in Figure 3.23, the posterior partition $\hat{\mathbf{z}}_0$ seems to provide a physical division of the brain: the turquoise and the black clusters are symmetric (with the only exception of the `rh-isthumuscingulate`), grouping mirrored areas in the right and the left hemisphere, respectively. On the contrary, the red and the blue groups are spread across hemispheres: in particular, they combine the frontal part of the brain and they are symmetrical, in the sense that they include exactly the same areas from both the left and the right hemisphere. The blue and red clusters make up the majority of the temporal and frontal lobes, and thus the entire frontal part of the brain. The purple and green clusters are also symmetrical, with the same exception of the `rh-isthumuscingulate` area. Hence, the mESBM learns a non-trivial anatomical segmentation of the human brain: the learnt structure is a lobe/hemisphere segmentation,

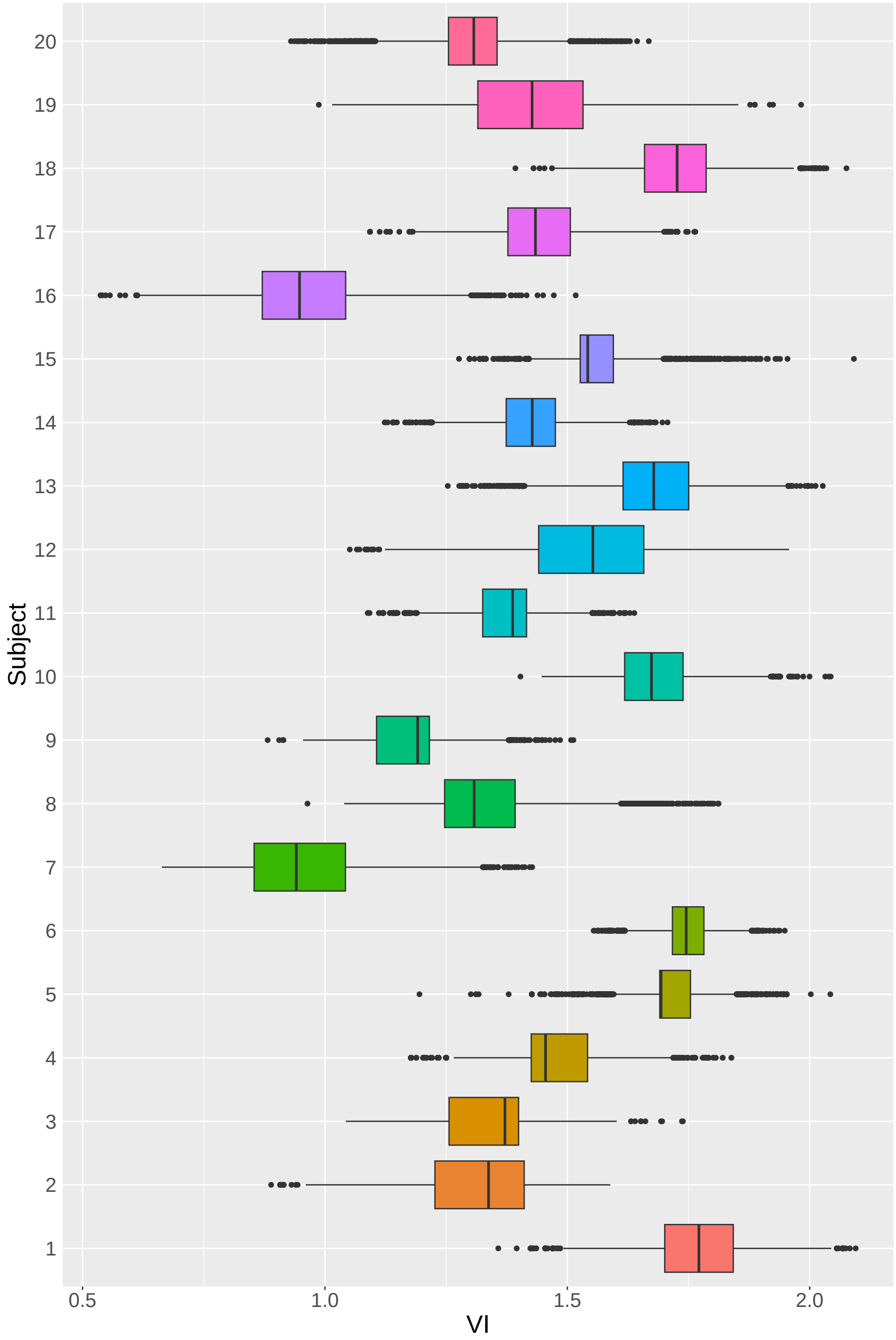


Figure 3.21: Empirical distribution of the posterior Variation of Information distance $\mathrm{VI}(\mathbf{z}_0, \mathbf{z}_k^t)$ per each subject, across iterations.

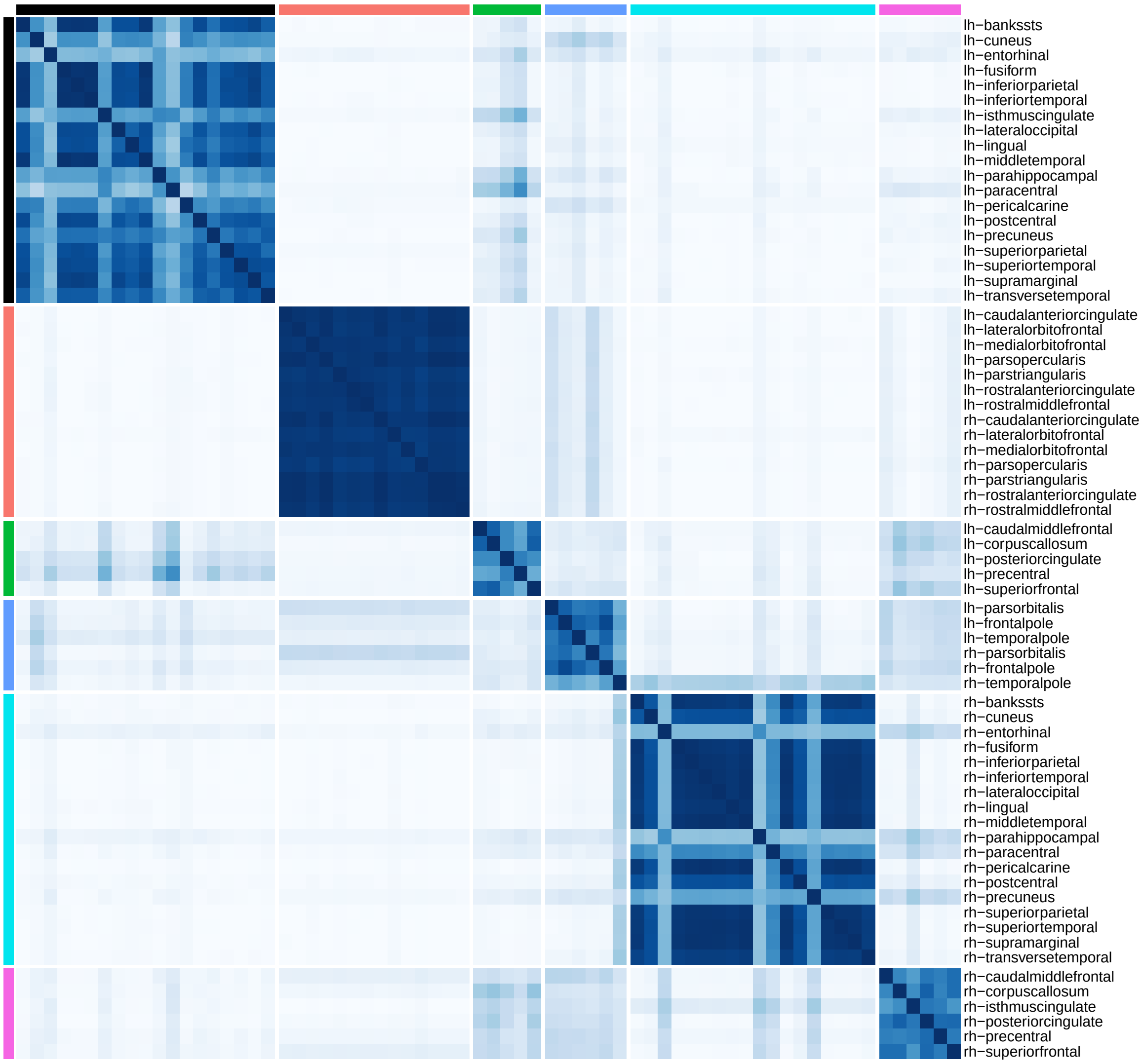


Figure 3.22: Posterior coclustering matrix of $\mathbf{z}_0$ under the multiplex extended stochastic block model.

which is more complex than a standard hemisphere or lobe division. In particular, the model discerns both the hemisphere and lobe division in the posterior part of the brain (where the hemisphere information is important, according to the literature), while it only provides a lobe division in the frontal part, where the lobes show similar connectivity patterns regardless of whether they are located on the left or right part.

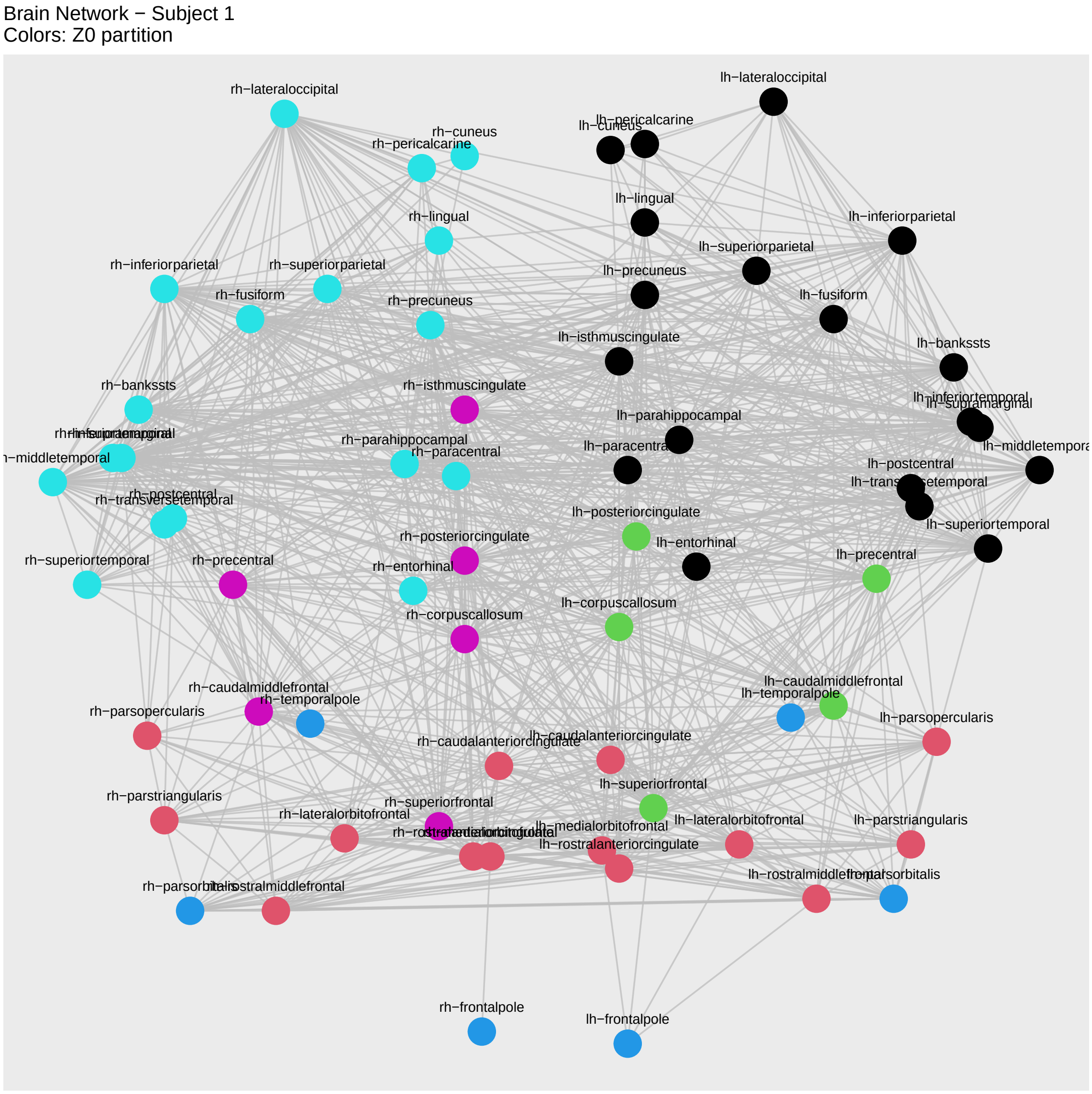


Figure 3.23: Brain map of patient 1: the colors denote the clusters provided by the posterior point estimation of $\mathbf{z}_0$ under the multiplex extended stochastic block model.

3.7.3 Competitor models

In this section, we assess the performances of the mESBM versus the two competitors illustrated in Section 3.5. More precisely, we compare the estimation of the common partition $\mathbf{z}_0$ to the clustering $\mathbf{z}_0^*$ provided by the bESBM on the collapsed network, and the layer-specific groups $\mathbf{z}$ to the partitions $\mathbf{z}^*$ obtained by independent ESBM on each layer of the multiplex network. Both the competitor models are defined with a DP prior on the partitions, with a concentration parameter equal to 1. The corresponding posterior distributions are estimated using $10'000$ iterations of Gibbs sampler, with $1'000$ iterations of burn-in. Figures 3.24 and 3.25 show the collapsed graph obtained from the supra-adjacency matrix of the edge-colored graph $Y^* = \sum_{k=1}^{K} Y_k$. Figure 3.24 shows the resulting weighted adjacency matrix: the division in two distinct blocks (i.e. the two hemispheres) is evident. Figure 3.25 shows the resulting weighted graph, where the edge color is proportional to the corresponding weight: the vast majority of the darkest edges are connecting areas within the same hemisphere, meaning that the strongest connections are between regions in the same half of the brain. The result of the posterior estimation provided by the bESBM are shown in Figures 3.26 and 3.27. Figure 3.26 shows the coclustering matrix of the samples obtained from the posterior distribution $p(\mathbf{z}_0^*|Y)$ of the bESBM. There seems to be almost no uncertainty, denoting an unjustified overconfidence in the estimation of the latent partition. Figure 3.27 shows the collapsed graph, where the node colors are with respect to the

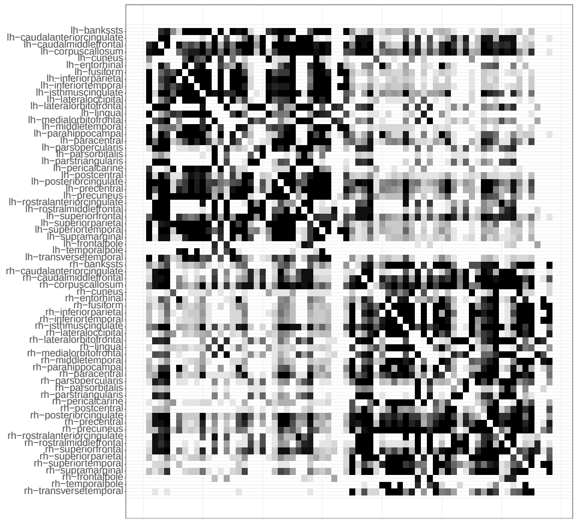


Figure 3.24: Collapsed adjacency matrix $Y^* = \sum_{k=1}^{K} Y_k$. The y-axis show the name of the brain areas, according to the Desikan parcellation. The color of each element is proportional to the edge weight (the darker, the higher).

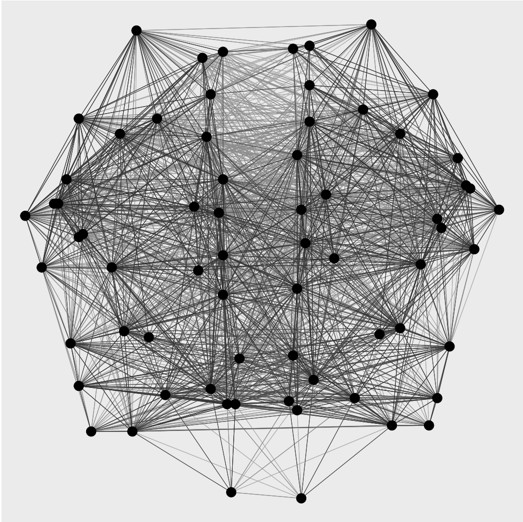


Figure 3.25: Graphical representation of the collapsed brain network $Y^* = \sum_{k=1}^{K} Y_k$: each node, representing a brain area, is placed in the correct coordinates as in the human anatomy. The edge colors are proportional to their weight (the darker, the higher).

point posterior estimation of the bESBM partition $\hat{\mathbf{z}}_0^*$ (obtained by minimising the Variation of Information measure). The grouping of the frontal part of the brain (orange clusters) is similar to the one provided by the mESBM in Figure 3.6 and reported on the right side for the sake of convenience. However, we lose all the distinction between right and left hemisphere we observed in the posterior part of the brain partition $\hat{\mathbf{z}}_0$ estimated by the mESBM: in this case, the black cluster is scattered through hemispheres, and the blue and light blue groups do not have a symmetric counterpart in the left hemisphere. In general, we do not have a physical division of the brain areas anymore: groups do not capture nodes in close proximity, but are spread across opposite sides of the human brain. Moreover, the estimated partition $\hat{\mathbf{z}}_0^*$ does not show any symmetry (contrary to the expectations, given the symmetrical nature of the human brain).

A further comparison is between the layer-specific partitions provided by the mESBM, and the clustering obtained by fitting K independent ESBMs on each layer of the multiplex network. Figure 3.28 shows the K coclustering matrices of the posterior samples provided by independent, layer-wise ESBMs: they show more uncertainty about the clustering process (gray areas), making the point estimation more difficult. Thus, it seems that adding a dependence among the partitions $\mathbf{z}_1,\ldots,\mathbf{z}_K$ through a common clustering $\mathbf{z}_0$ also improves the final estimation of layer-specific groups.

3.8 Discussion and future research directions

This work proposes the first —- to our knowledge —- general model in the literature to perform inference on both layer-specific and common partitions of multiplex networks. It also presents a non-trivial algorithm to estimate the posterior clusters. The application on brain maps obtained through DTI scans illustrates how our approach could find valuable patterns in the data: the estimated layer-specific partitions seems related to possible mental illnesses of the subjects involved in the study, while the common clustering provides a physical and anatomical division of the brain areas, shared by all humans. Differently from the competitors tested in this work, the mESBM does not require any preprocessing of the data (such as collapsing all the layers of the multiplex network into one - which may yield an important loss of information), and exploits all the features contained in such a data structure (for example, the identity of the nodes in all the layers). Although we specifically focus on brain networks, our method can be applied in many other settings. For example, it may be used to study the evolution of patterns of connections among individuals through time: in this case, we could apply the mESBM to a temporal network, where each layer contains the interactions among the same V nodes at a specific time point. Moreover, it would interesting to generalize the mESBM in several directions: first, its counterpart in the case of weighted, edge-colored networks could be defined. This extension is pretty straightforward, changing only the block prior and the likelihood distribution of the mESBM. Secondly, a supervised framework is also worth of exploration; for example,

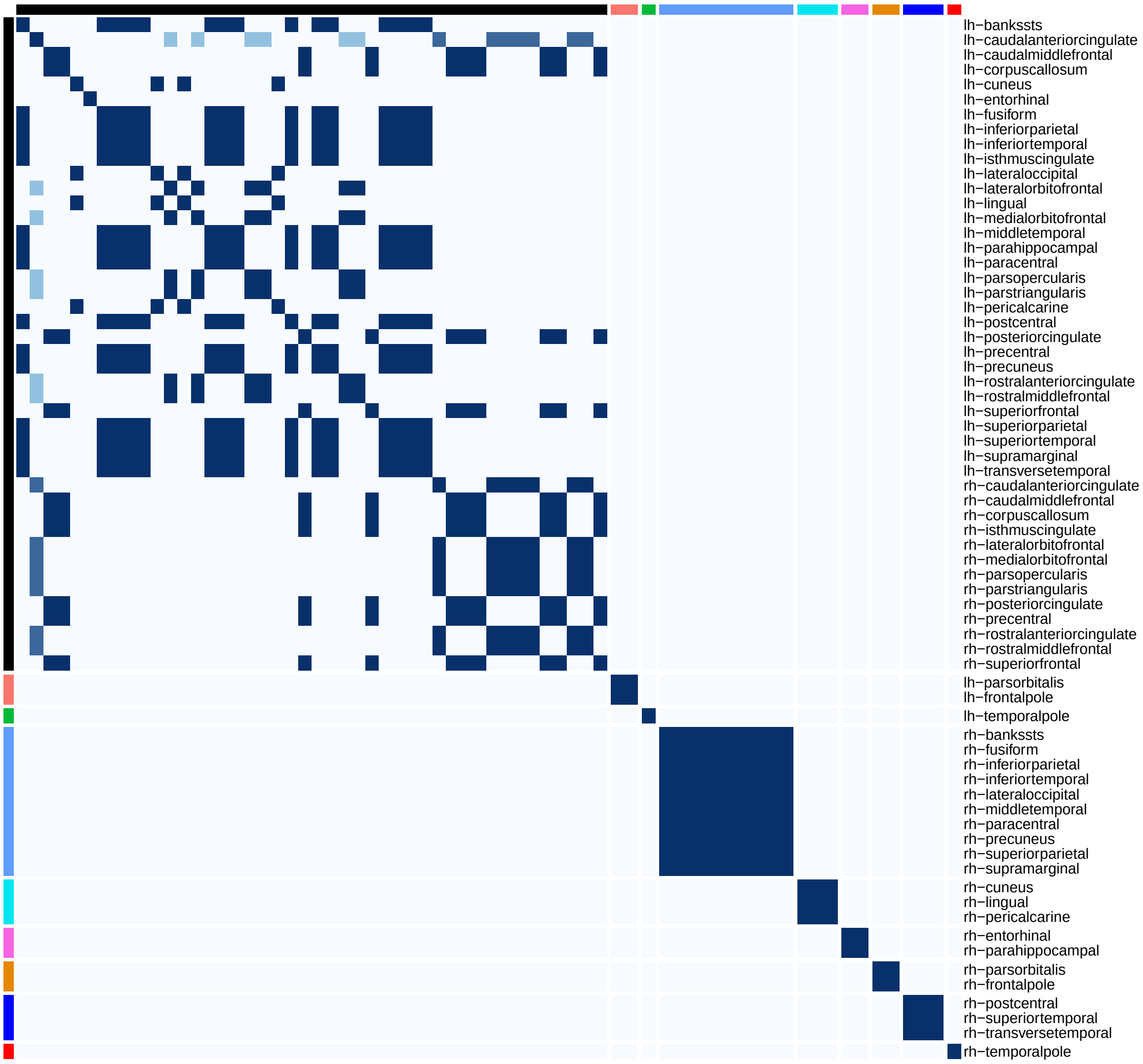


Figure 3.26: Coclustering matrix of the posterior samples estimated by the binomial extended stochastic block model.

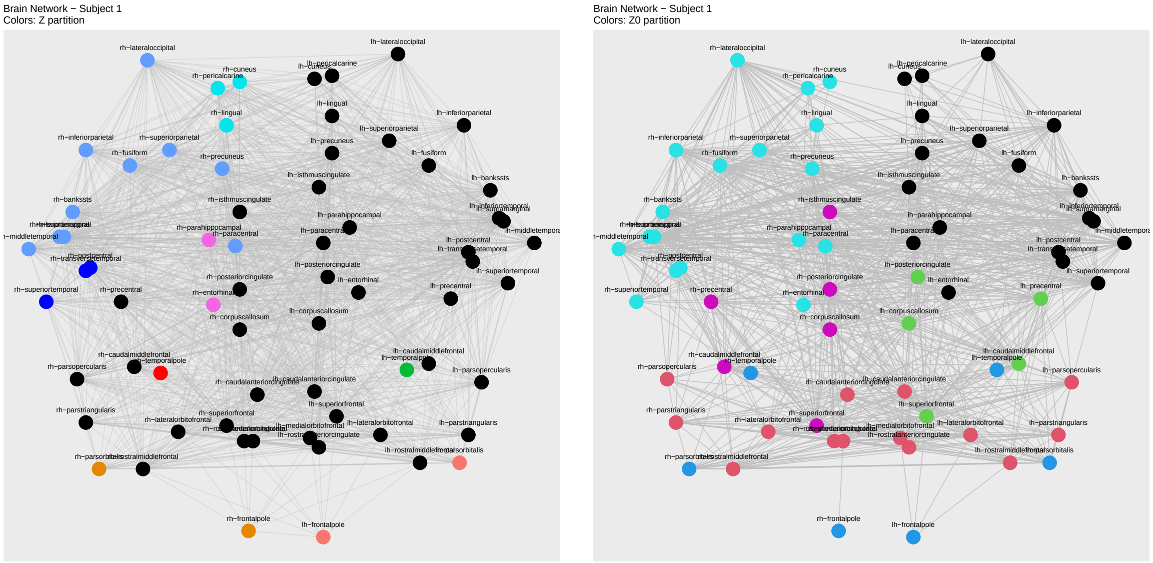


Figure 3.27: Graphical representation of a brain network: each node, representing a brain area, is placed in the correct coordinates as in the human anatomy. The node colors are with respect to the point posterior estimation of $\hat{\mathbf{z}}_0^*$ under the binomial extended stochastic block model (left) and under the multiplex extended stochastic block model estimated common partition (right).

in the brain application, the layer-specific partitions could be informed by features of each subject, and the common clustering could be supervised by brain structural characteristics. Another possible modelling extension is to inject knowledge about the common clustering by centering its prior distribution on a partition of interest, using for example the centered partition process [Paganin et al., 2021], or any law encompassing such a desired structure. The proposed inferential framework remains valid, regardless of the prior distribution defined on the common partition.

Appendix

Borrowing of information

A fundamental parameter of the mESBM defined in Section 3.3 is the similarity parameter α, which regulates the amount of information borrowed between $\mathbf{z}_0$ and $\mathbf{z} = [\mathbf{z}_1, \dots, \mathbf{z}_K]'$. In fact, the similarity parameter α plays a key role not only in updating the network partitions $\mathbf{z}$ given the common grouping $\mathbf{z}_0$, but also in proposing the auxiliary variables $\tilde{\mathbf{z}} \sim p(\mathbf{z}|\mathbf{z}_0)$ and in accepting $\mathbf{z}_0$ in the exchange step (see Algorithm 1). In this section, we study how the similarity function $g(\mathbf{z}_{0h})$ and the informed Gibbs-type distribution $p(\mathbf{z}|\mathbf{z}_0)$ behave when $\alpha \to 0$ or $\alpha \to \infty$. Without loss of generality, we consider the case $K = 1$ (and $\mathbf{z} = \mathbf{z}_1$).

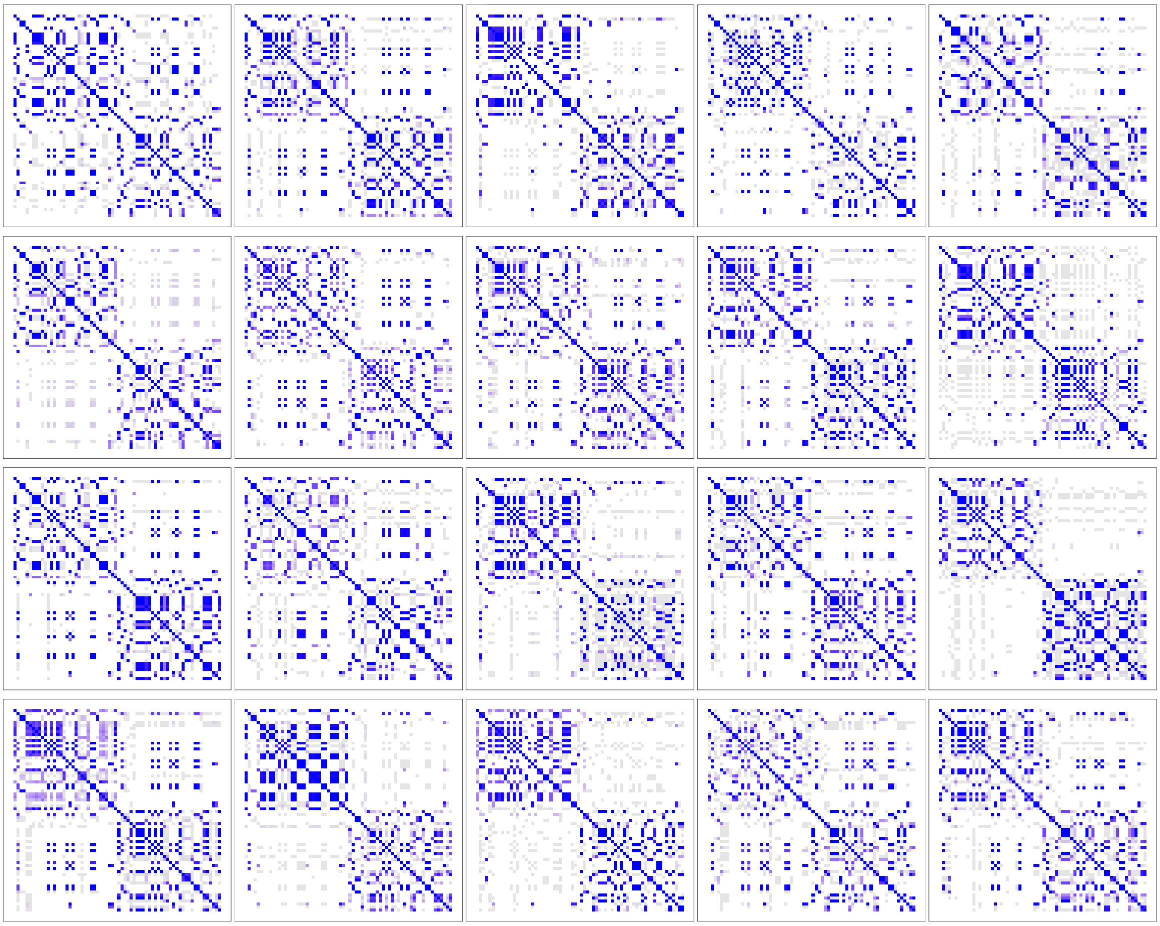


Figure 3.28: Coclustering matrix of the posterior samples estimated by the K independent extended stochastic block models on each layer.

We will show the results both for the model using the unnormalised and the normalised similarity function, justifying our final choice of retaining the normalisation constant in the similarity function of the mESBM.

Similarity Function

Following the notation of Section 3.3.2, we consider the following similarity function of the informed Gibbs-type distribution $p(\mathbf{z}|\mathbf{z}_0)$, given that the information is provided by a categorical variable ($\mathbf{z}_0$):

$$g(\mathbf{z}_{0h}) = \frac{\Gamma(H_0\alpha)}{\prod_{h_0=1}^{H_0}\Gamma(\alpha)} \frac{1}{\Gamma(n_h + H_0\alpha)} \prod_{h_0=1}^{H_0} \Gamma(n_{hh_0} + \alpha) = \frac{\Gamma(H_0\alpha)}{\prod_{h_0=1}^{H_0}\Gamma(\alpha)} k(\mathbf{z}_{0h}) \propto k(\mathbf{z}_{0h})$$

for $h = 1,\ldots,H$. Thus, $g(\mathbf{z}_{0h})$ is the normalised Dirichlet-multinomial distribution (where the normalising constant depends on α, with respect to which we are studying the limits) and $k(\mathbf{z}_{0h})$ is its unnormalised counterpart. Potentially, they both can define a similarity function for the informed Gibbs-type distribution.

Unnormalised similarity function $k(\mathbf{z}_{0h})$ Let us first study the limit of the unnormalised similarity function $k(\mathbf{z}_{0h})$ as $\alpha \to 0$ and $\alpha \to \infty$. For $\alpha \to 0$:

$$\lim_{\alpha\to 0} k(\mathbf{z}_{0h}) = \lim_{\alpha\to 0} \frac{1}{\Gamma(n_h + H_0\alpha)} \prod_{h_0=1}^{H_0} \Gamma(n_{hh_0} + \alpha) = \frac{1}{\Gamma(n_h)} \prod_{h_0=1}^{H_0} \Gamma(n_{hh_0}).$$

In this case we have maximum borrowing of information between $\mathbf{z}$ and $\mathbf{z}_0$. The informed Gibbs-type distribution $p(\mathbf{z}|\mathbf{z}_0)$ is only influenced by the size of the intersections between clusters in $\mathbf{z}$ and clusters in $\mathbf{z}_0$, normalised. For $\alpha \to \infty$:

$$\lim_{\alpha\to\infty} k(\mathbf{z_{0h}}) = \lim_{\alpha\to\infty} \frac{1}{\Gamma(n_h + H_0\alpha)} \prod_{h_0=1}^{H_0} \Gamma(n_{hh_0} + \alpha) < \lim_{\alpha\to\infty} \frac{\Gamma(\alpha + V)^{H_0}}{\Gamma(1 + H_0\alpha)} = 0,$$

where the last equality holds for $H_0 > 1$. If $H_0 = 1$, $k(\mathbf{z}_{0h}) = 1$ and $\lim_{\alpha\to\infty} k(\mathbf{z}_{0h}) = 1$.

Normalised similarity function $g(\mathbf{z}_{0h})$ Let us now consider the limits of the normalised similarity function $g(\mathbf{z}_{0h})$. For $\alpha \to 0$, the result is the same as before, with $k(\mathbf{z}_{0h})$:

$$\lim_{\alpha\to 0} g(\mathbf{z}_{0h}) = \lim_{\alpha\to 0} k(\mathbf{z}_{0h}) = \frac{1}{\Gamma(n_h)} \prod_{h_0=1}^{H_0} \Gamma(n_{hh_0})$$

For $\alpha \to \infty$, the result is different. The normalised similarity function does not converge to zero anymore, but to a positive constant:

$$\begin{aligned}\lim_{\alpha\to\infty} g(\mathbf{z}_{0h}) &= \lim_{\alpha\to\infty} \frac{\Gamma(H_0\alpha)}{\prod_{h_0=1}^{H_0}\Gamma(\alpha)} \frac{1}{\Gamma(n_h+H_0\alpha)} \prod_{h_0=1}^{H_0}\Gamma(n_{hh_0}+\alpha)\\ &= \lim_{\alpha\to 0}\int_{\Delta_{H_0-1}} \prod_{h_0=1}^{H_0} p_{h_0}^{n_{hh_0}} \frac{\Gamma(H_0\alpha)}{\prod_{h_0=1}^{H_0}\Gamma(\alpha)} \prod_{h_0=1}^{H_0} p_{h_0}^{\alpha-1}\,\mathrm{d}\underline{p}\\ &= \prod_{h_0=1}^{H_0}\left(\frac{1}{H_0}\right)^{n_{hh_0}} = \left(\frac{1}{H_0}\right)^{\sum_{h_0=1}^{H_0} n_{hh_0}} = \left(\frac{1}{H_0}\right)^{n_h} > 0,\end{aligned}$$

with $\underline{p} = (p_1,\dots,p_{H_0})$ and Δ_{H_0-1} the support of the Dirichlet distribution. In the latter, we exploit the property of the Dirichlet distribution with a growing concentration parameter: in the limit, its variance converges to zero and the process converges to its mean. Hence, for $\alpha \to \infty$, $g(\mathbf{z}_{0h})$ converges to a constant value, does not depend on the structure of $\mathbf{z}_0$ anymore and fails in sharing information between $\mathbf{z}$ and $\mathbf{z}_0$ through $p(\mathbf{z}|\mathbf{z}_0)$.

Informed Gibbs-type distributions

Using the results of the previous section, we can study how informed Gibbs-type distributions vary according to limits with respect to α, using as similarity function either $g(\mathbf{z}_{0h})$ or $k(\mathbf{z_{0h}})$. For $\alpha \to 0$, we get the same limit for $p(\mathbf{z}|\mathbf{z}_0) = \frac{q(\mathbf{z}|\mathbf{z}_0)}{c(\mathbf{z}_0)}$ in both cases:

$$\begin{aligned}\lim_{\alpha\to 0} q(\mathbf{z}|\mathbf{z}_0) &= \lim_{\alpha\to 0} \mathscr{W}_{V,H} \prod_{h=1}^{H} g(\mathbf{z}_{0h})(1-\sigma)_{n_h-1} = \lim_{\alpha\to 0} \mathscr{W}_{V,H} \prod_{h=1}^{H} k(\mathbf{z}_{0h})(1-\sigma)_{n_h-1}\\ &= \mathscr{W}_{V,H} \prod_{h=1}^{H} \frac{1}{\Gamma(n_h)} \prod_{h_0=1}^{H_0} \Gamma(n_{hh_0})(1-\sigma)_{n_h-1}.\end{aligned}$$

On the other hand, the result for $\alpha \to \infty$ is different for the two similarity functions. However, in both cases, for $\alpha \to \infty$, the informed Gibbs-type distribution $p(\mathbf{z}|\mathbf{z}_0)$ converges to a law independent of $\mathbf{z}_0$, basically nullifying the influence of $\mathbf{z}_0$ on $\mathbf{z}$ in different ways.

Informed Gibbs-type distributions $p(\mathbf{z}|\mathbf{z}_0)$ with unnormalised similarity function $k(\mathbf{z}_{0h})$

$$\lim_{\alpha\to\infty} q(\mathbf{z}|\mathbf{z}_0) = \lim_{\alpha\to\infty} \mathscr{W}_{V,H} \prod_{h=1}^{H} k(\mathbf{z}_{0h})(1-\sigma)_{n_h-1} = \delta_{\{\mathbf{z}=(1,1,\dots,1)\}},$$

since $k(\mathbf{z}_{0h})$ always converges to zero, but for the partition $\mathbf{z} = (1,1,\dots,1)$ it converges at a slower pace. Hence, for $\alpha \to \infty$, the informed Gibbs-type distribution converges to a degenerate Dirac measure on the configuration with all the nodes in one cluster, making all the layer-specific partitions almost surely identical. This is not clearly desirable, and also the reason why we have introduced the normalising constant in the similarity function in the mESBM.

Informed Gibbs-type distribution $p(\mathbf{z}|\mathbf{z}_0)$ with normalised similarity function $g(\mathbf{z}_{0h})$

$$\begin{aligned}
\lim_{\alpha\to\infty} q(\mathbf{z}|\mathbf{z_0}) &= \lim_{\alpha\to\infty} \mathscr{W}_{V,H} \prod_{h=1}^{H} g(\mathbf{z}_{0h})(1-\sigma)_{n_h-1} = \mathscr{W}_{V,H} \prod_{h=1}^{H} \Big(\frac{1}{H_0}\Big)^{n_h} (1-\sigma)_{n_h-1} \\
&= \mathscr{W}_{V,H} \Big[\prod_{h=1}^{H} \Big(\frac{1}{H_0}\Big)^{n_h}\Big] \prod_{h=1}^{H} (1-\sigma)_{n_h-1} \\
&= \Big(\frac{1}{H_0}\Big)^{V} \mathscr{W}_{V,H} \prod_{h=1}^{H} (1-\sigma)_{n_h-1} \propto p(\mathbf{z}),
\end{aligned}$$

that is, the informed Gibbs-type distribution $p(\mathbf{z}|\mathbf{z_0})$ is converging to its (uninformed) Gibbs-type counterpart for $\alpha \to \infty$. Thus, in this case, the mESBM reduces to the standard ESBM, without any supervision or borrowing of information, and with a non-degenerate prior on the layer-specific partitions.

Chapter 4

Spatially-informed Bayesian clustering for weighted networks

4.1 Introduction

Complex networks can represent transportation connections between an origin and a destination, such as roads, railways, and airline routes: these transportation networks are intricate and dynamic, comprising numerous interconnected nodes and links. Analyzing these networks as a whole can be challenging due to their size and intricate relationships. Clustering comes into play to simplify and understand their underlying structure by grouping similar entities (embedded in the same geographical space) together. These subsets, known as clusters or communities, represent groups of entities that are closely related to each other compared to nodes in other clusters. The clustering process can be guided by various factors, such as geographical proximity, traffic flow patterns, connectivity, or other relevant attributes, depending on the specific application and context.

In network analysis, clustering of network nodes is a fundamental task that has been studied extensively; see, e.g., see Fortunato and Hric [2016] and Fortunato and Newman [2022] for a review. A first taxonomy of clustering is that into *soft* and *hard* clustering. The former allows nodes to belong to multiple clusters with different probabilities, while in the latter each node is assigned to a single community. In this chapter we focus on hard clustering. Given a clustering aim — either soft or hard — a further subdivision of clustering techniques is that into algorithmic and model-based strategies.

Algorithmic clustering is typically based on the optimization of some measure of goodness of the resulting partition or, alternatively, on automatic detection of groups. Arguably, the most famous algorithmic technique is greedy clustering [Clauset et al., 2004, Newman, 2004], which outputs the partition with the highest (local) modularity. The Louvain algorithm [Blondel et al., 2008] provides a hierarchical clustering, that recursively merges communities until reaching a single one and executes a greedy clustering on the condensed graph. Finally, spectral clustering [Von Luxburg, 2007] is based on the spectral decomposition of some similarity matrix of the nodes. Possible choices for such a matrix are the adjacency or the Laplacian matrix of a graph, as well as some of their regularized versions [Amini et al., 2013]. All the aforementioned methods are well-defined for both weighted and unweighted networks.

On the other hand, model-based techniques assume an underlying probabilistic model for the data, opening the way for statistical analysis. One class of such models are latent space models [Gollini and Murphy, 2016], but their analysis and comparison is out of the scope of this thesis and deferred to future work. We focus instead on the well-known stochastic block model (SBM) [Nowicki and Snijders, 2001, Schmidt and Morup, 2013], and its generalizations, such as the mixed-membership SBM [Airoldi et al., 2008], the degree corrected SBM [Karrer and Newman, 2011, Côme et al., 2021], the bipartite SBM [Larremore et al., 2014] and the extended stochastic block model (ESBM) [Legramanti et al., 2022]. The SBM was originally defined on binary networks, but several adaptations to weighted graphs have then been formulated [Aicher et al., 2013, Peixoto, 2018, Xu et al., 2020, Ng and Murphy, 2021]. Peng and Carvalho [2013] introduce a Bayesian degree-corrected SBM with priors on the node-specific parameters. However, they develop a model for binary networks which requires setting the number of desired clusters a priori. Herlau et al. [2014] define the degree-corrected infinite relational model, which overcomes the issue of choosing the number of communities a priori. However, most of these models do not include covariate supervision.

In fact, network data often come with node attributes, which can inform the node partition. Some clustering techniques, both model-based and algorithmic, were then designed to leverage this additional information. In the Bayesian literature, a widely used strategy to include covariates information in the partition process is through product partition models [Hartigan, 1990, Barry and Hartigan, 1992, Dahl, 2008, Park and Dunson, 2010, Muller et al., 2011]. Legramanti et al. [2022] apply this general strategy — which is not restricted to network data — in the context of SBMs. This is made possible by the fact that their extended SBM relies on Gibbs-type priors, which have a product-partition-model structure. In the context of soft clustering, a model-based solution for weighted networks with node attributes is provided by Yang et al. [2013]. Algorithmic alternatives for attribute-assisted clustering include Combe et al. [2015] and Binkiewicz et al. [2017], Mu et al. [2022], who respectively adapt the Louvain algorithm and spectral clustering to include node attributes. Zhang et al. [2016] propose to maximize a joint community detection criterion, which takes into account both edges information and node attributes. For a more comprehensive review about clustering algorithms for node-attributed networks, see Chunaev [2020].

As explained in the next sections, the goal of this work is to present a model-based method for clustering municipalities. These municipalities are represented as nodes within a transportation network, and our aim is to group them based on their flow patterns, taking into account their geographical location. From a methodological and applicative viewpoint, the work of Egidi et al. [2023] is extremely related to ours: the authors introduce a Bayesian latent mixture model aimed at designing administrative structures on the basis of commuting flows between municipalities, represented as nodes. They propose to augment the geographical node attributes of a transport network, including the longitude and latitude of each municipality, with a third, latent variable accounting for hidden factors impacting the commuting structure, such as geographical barriers or socio-economics differences between the units.

We apply the methodology developed in this chapter to the data provided by Azienda dei Trasporti di Bergamo (ATB), one of the public transportation companies operating in Bergamo and surroundings. As carefully explained in the next sections, the proposed model produces a clustering of municipalities, which is determined by considering both their connectivity patterns and geographical positions. Clusters of this type can be useful to a public transportation company to, for example, establish new pricing zones in order to decide where to change the price of the tickets or monthly subscriptions, still remaining within budget constraints. In fact, clusters of municipalities allow for a better understanding of travel patterns and passenger demand. By analyzing flows within and between clusters, transportation companies can identify high-demand routes and adjust ticket prices accordingly. For example, routes connecting busy urban clusters may have higher demand and, therefore, warrant higher ticket returns compared to routes serving less-populated areas. Moreover, some clusters of municipalities often have multiple transportation providers operating within the same region, and the presence of competition can also influence ticket pricing strategies. Transportation clusters may also provide opportunities for pricing incentives and discounts. ATB could introduce promotional fares, loyalty programs, or group discounts targeting specific clusters or travel corridors, to attract more passengers, encourage loyalty, and stimulate demand within the cluster. Considering as primary goal the definition of new zones for ticket pricing, we can notice that pricing zones usually have radial shapes. Radial clusters are relatively common in transportation networks, especially in urban areas. They refer to a network structure where transportation routes radiate outward from a central hub to surrounding regions. This configuration is often observed in areas with a central main town and where transportation routes extend outward to connect the surrounding municipalities. Our modeling approach focuses on developing a Bayesian clustering model for networks that specifically incorporates the concept of radial clusters. To achieve this, we utilize a relevant covariate for each node, namely the distance of each municipality from the main hub of the network. By incorporating the distance as node attribute, our Bayesian clustering model will effectively encourage the formation of radial clusters within the network. This means that municipalities with the same distance from the main hub of the network will have a higher tendency to be grouped together, forming clusters that radiate outward from the central hub.

The rest of this chapter is organized as follows: Section 4.2 defines a family of models for Bayesian network clustering and also discusses the prior settings, including a mechanism to infer the amount of node attribute information in the model in a mathematically rigorous way. Section 4.3 provides the computational framework to sample from the posterior distribution of such models. Section 4.4 shows results from alternative algorithmic approaches to cluster the transport network of interest and displays the performances of the proposed model on transport data and Section 4.5 offers a conclusive discussion, along with some new, desirable research directions.

4.2 Poisson extended stochastic block models (pESBM)

Consider an undirected integer-weighted network with V nodes, and its $V \times V$ symmetric adjacency matrix Y, whose elements $Y_{uv} = Y_{vu} \in \{0,1,2,\ldots\}$ contain the integer weight of the edge connecting nodes u and v, for $u,v = 1,\ldots,V$. Self-loops are not informative for our clustering purposes, and thus we set $Y_{vv} = 0$ for each $v = 1,\ldots,V$. Denote with $\mathbf{x}_v = (x_{v1},\ldots,x_{vp})$ the p-dimensional row vector of covariates associated to node v, and with X the $V \times p$ matrix obtained by stacking all the $\mathbf{x}_v$'s. Finally, let $\mathbf{z} = (z_1,\ldots,z_V) \in \{1,\ldots,H\}^V$ be the vector of node memberships associated to a partition of the V nodes into H groups, so that $z_v = h$ if and only if node v belongs to cluster h. To model the integer edge weights, we define the following Poisson extended stochastic block model (pESBM):

$$\begin{aligned}
Y_{uv}|z_u = h, z_v = k, \lambda_{hk} &\overset{\text{ind.}}{\sim} \text{Poisson}(\lambda_{hk}), && \text{for } 1 \le u < v \le V, && (4.1)\\
\lambda_{hk}|\mathbf{z} &\overset{\text{i.i.d.}}{\sim} \text{Gamma}(a,b), && \text{for } 1 \le h \le k \le H, && (4.2)\\
p(\mathbf{z},\alpha;X) &\propto p(\mathbf{z}|\alpha;X)\cdot\pi(\alpha) && && (4.3)\\
&\propto \mathscr{W}_{V,H}\prod_{h=1}^{H}(1-\sigma)_{n_h-1}\, g(X_h^*)^{\alpha}\cdot\pi(\alpha),
\end{aligned}$$

where X_h^* is the $n_h \times p$ matrix obtained by stacking the row vectors of the covariates of nodes in cluster h. Each element Y_{uv} below the diagonal of Y is modeled with a conditionally independent Poisson distribution whose rate depends solely on the cluster memberships of the involved nodes, u and v. As a result, there are as many distinct Poisson rates λ_{hk} as the possible unordered pairs of clusters, including identical ones. To exploit conjugacy, each distinct Poisson rate is given an independent Gamma(a,b) prior. In Equation (4.3) the conditional distribution $p(\mathbf{z}|\alpha;X)$ is a Gibbs-type prior, modified by the similarity function $g(X_h^*)^{\alpha}$. The former depends on the cardinality n_h of each cluster and is parametrized by the discount parameter $\sigma < 1$, while $\{\mathscr{W}_{V,H} : 1 \le H \le V\}$ is a collection of non-negative weights such that $\mathscr{W}_{V,H} = (V - H\sigma)\mathscr{W}_{V+1,H} + \mathscr{W}_{V+1,H+1}$ and $\mathscr{W}_{1,1} = 1$. Instead, the similarity $g(X_h^*)$ is a function of the covariates of nodes in cluster h, and is raised to power $\alpha \ge 0$ thus allowing to adjust the relative importance of node covariates. The law $\pi(\alpha)$ is an auxiliary distribution, needed to properly define the joint prior for $\alpha, \mathbf{z}$ in a computationally convenient way. According to the auxiliary distribution of choice, we can get different prior distributions. We analyze two scenarios in the next subsections.

It is worth noticing that $\mathbf{z}$ identifies labeled clusters: hence, technically, a vector $\mathbf{z}$ and $\mathbf{z}'$

may contain different labels (and thus be mismatched objects), even though they identify the same partition. Henceforth, we use the following convention to identify a unique labeling for each given partition, setting $z_1^* \rightarrow z_1 = 1$ and sequentially relabeling each membership value in $\mathbf{z}^* = (z_1^*, \dots, z_V^*)$ as

$$z_i^* \rightarrow z_i = \begin{cases} \max\{z_1, \dots, z_{i-1}\} + 1 & \text{if } z_i^* \notin (z_1^*, \dots, z_{i-1}^*), \\ z_j & \text{for } j : z_j^* = z_i^*, \end{cases} \tag{4.4}$$

for $i = 2, \dots, V$. This operation effectively defines an equivalence class, so that labeled vector attaining the same partition will be reduced to the same object. This procedure is equivalent to the canonical projection in Peng and Carvalho [2013]. In the next section, we motivate the prior choice of the pESBM, and elaborate on the role of the smoothing parameter α.

4.2.1 Prior specification

The prior specification for the clusters $\mathbf{z}$ and the smoothing parameter α provided by the pESBM is not trivial: in this section, we first establish our choice for the similarity function used throughout this chapter. Then, we present the class of Product Partition Models (to which Gibbs-type distributions belong to), used to define the prior in the pESBM. Moreover, according to the choice of the auxiliary distribution $\pi(\alpha)$, we can either allow the user to specify the smoothing parameter, or we can learn it from the data. We present these two scenarios in the next subsections.

A crucial choice: the similarity function

The choice of the similarity function $g(\cdot)$ is of paramount importance in applications, and obviously depends on the type of node covariates. The chosen $g(\cdot)$ reflects the assumed agreement of each cluster with respect to the node attributes, and should increase as the covariates of nodes in cluster h get more similar. Among the possible ways to define the similarity function [e.g. Dahl, 2008, Muller et al., 2011], we follow Muller et al. [2011], Page and Quintana [2015, 2018] and derive its core as the marginal distribution of a conjugate Bayesian model:

$$\tilde{g}(X_h^*) = \int \prod_{v : z_v = h} p(\mathbf{x}_v | \boldsymbol{\xi}_h) p(\boldsymbol{\xi}_h) \mathrm{d}\boldsymbol{\xi}_h.$$

Even though covariates are not random, this procedure offers modeling advantages including predictive coherence; see Muller et al. [2011], Page and Quintana [2015] for more details. In the motivating transportation network application, each node $v = 1, \dots, V$, which represents a municipality, is equipped with its latitude and longitude, and consequently with its distance from the major hub of the network (Bergamo, the main city). As explained in Section 4.1, our objective is to induce radial clustering, particularly relevant for transport networks and useful to, for example, define new pricing zones for public transport. To achieve this, municipalities with similar distances from Bergamo can be encouraged to join the same cluster, thereby increasing the likelihood of partitioning municipalities into rings surrounding the main city. Hence, in the application, the node attribute supervising the partition process is the distance of each municipality from Bergamo. We standardize such a distance (in order to provide a more general framework, which can be useful in other applications), and we denote the final covariate of interest $\mathbf{x}_v \in \mathbb{R}^p$, with $p = 1$ in our case. Hence, we choose to derive the similarity using a conjugate Gaussian-Inverse Wishart model. Namely, we assume that the similarity function for the general case of $p-$dimensional covariates is defined as:

$$\mathbf{x}_v | \boldsymbol{\mu}_h, \Sigma_h \sim \mathcal{N}_p(\boldsymbol{\mu}_h, \Sigma_h)$$

for each node v in cluster h, and we set

$$\boldsymbol{\mu}_h | \Sigma_h \sim \mathcal{N}_p(\boldsymbol{\mu}_0, k_0^{-1}\Sigma_h), \quad \Sigma_h \sim \text{Inverse-Wishart}(\Sigma_0, \nu_0)$$

for $h = 1, \dots, H$, with $k_0 > 0$ and $\nu_0 > p+1$. By conjugacy, marginalizing out $\boldsymbol{\mu}_h$ and Σ_h, we obtain a multivariate Student's t kernel, which is higher when the covariates of the nodes in cluster h are more concentrated around $\bar{\mathbf{x}}_h$, which denotes centroid of cluster h:

$$\begin{aligned}\tilde{\tilde{g}}(X_h^*) = &\frac{2^{p(\nu_0+n_h)/2}}{\sqrt{k_0+n_h}}\Gamma_p\left(\frac{\nu_0+n_h}{2}\right)\cdot \\ &\cdot\left|\Sigma_0^{-1} + \frac{k_0 n_h}{k_0+n_h}(\bar{\mathbf{x}}_h - \boldsymbol{\mu}_0)^T(\bar{\mathbf{x}}_h - \boldsymbol{\mu}_0) + \sum_{v:z_v=h}(\mathbf{x}_v - \bar{\mathbf{x}}_h)^T(\mathbf{x}_v - \bar{\mathbf{x}}_h)\right|^{-\frac{\nu_0+n_h}{2}}, \end{aligned} \tag{4.5}$$

where $\Gamma_p(x) = \pi^{(p-1)/2}\Gamma(x)\Gamma_{p-1}(x-1/2)$, $\Gamma_1(\cdot) = \Gamma(\cdot)$ is the Gamma function, and $|\cdot|$ is the matrix determinant; recall that $\bar{\mathbf{x}}_h$, $\boldsymbol{\mu}_0$ and $\mathbf{x}_v$ are row vectors in $p > 1$ dimensions. Detailed computations can be found in the Appendix. From the analytic form of the similarity, it is clear that attributes of the nodes in cluster h with more variability yield lower values of $\tilde{\tilde{g}}(X_h^*)$, thus discouraging municipalities with too-diverse distances from Bergamo from being clustered together.

We also perform two additional transformations, with the scope of rescaling the similarity function. First, we normalize the value of each similarity measure $\tilde{\tilde{g}}(X_h^*)$, inspired by the calibrated similarity function of Page and Quintana [2018]:

$$\tilde{g}(X_h^*) \overset{\text{def}}{=} \frac{\tilde{\tilde{g}}(X_h^*)}{\sum_{h=1}^{H}\tilde{\tilde{g}}(X_h^*)}.$$

This operation rescales the original support of the similarity function from $[0, \infty)$ to $[0, 1]$. Moreover, we also apply an additional linear transformation:

$$g(X_h^*) \overset{\text{def}}{=} \frac{e^2-1}{e}\tilde{g}(X_h^*) + \frac{1}{e}.$$

This last transformation has the effect of changing again the support of $g(X_h^*)$ to $[1/e, e]$ causing, more importantly, the support of $\log g(X_h^*)$ to be $[-1,1]$. This is necessary to control the mean a posteriori in case of non-trivial inference on α, as carefully explained below.

Supervised product partition models (PPMx)

Product partition models (PPMs) [Hartigan, 1990] have gained significant popularity in the domains of machine learning and probabilistic modeling due to their ability to capture dependencies among variables. These models provide a framework for representing complex relationships and interactions between variables in a structured manner, with their main application being probabilistic clustering. PPMx [Muller et al., 2011, Page and Quintana, 2015, 2018], an extension of PPMs, introduces additional flexibility by allowing instances to be partitioned into multiple groups estimated by leveraging covariate information. Concerning our work, we notice that the conditional distribution $p(\mathbf{z}|\alpha;X)$ in Equation (4.3) is an instance of a PPMx, since it can be written as

$$p(\mathbf{z}|\alpha;X) \propto \prod_{h=1}^{H} \mathscr{W}_{V,H}(1-\sigma)_{n_h-1} g(X_h^*)^{\alpha} = \prod_{h=1}^{H} c(S_h) g(X_h^*)^{\alpha}, \tag{4.6}$$

where S_h is the cluster with label h defined in the partition $\mathbf{z}$, and $c(\cdot)$ is called cohesion function. In the pESBM, the cohesion function is chosen according to the Gibbs-type distribution of interest. Moreover, the definition of the similarity function $g(\cdot)$ is crucial, as well motivated below. It is important to note that PPMx models usually lack a closed-form normalizing constant, which hinders their computation and posterior inference. In fact,

$$p(\mathbf{z}|\alpha;X) = \frac{1}{k(\alpha)} \prod_{h=1}^{H} c(S_h) g(X_h^*)^{\alpha}, \tag{4.7}$$

where $k(\alpha) = \sum_{\mathbf{z}\in\mathcal{Z}} \prod_{h=1}^{H} c(S_h) g(X_h^*)^{\alpha}$. This means that the normalising constant $k(\alpha)$ can only be obtained through a sum over the space of all the possible partitions of the V nodes, which is computationally unattainable even for moderate values of V. In general, this issue prevents the straightforward calculation of the posterior distribution of any parameter of the PPMx model, including α. In the next sections we use two auxiliary distributions $\pi(\alpha)$ to define ad hoc prior distributions for α and $\mathbf{z}$ trying to circumvent this problem.

Poisson extended stochastic block model, with fixed smoothing parameter

The simplest choice for the auxiliary distribution $\pi(\alpha)$ used in the prior of Equation (4.3) is a point mass on a user-defined value for the smoothing parameter

$$\pi(\alpha') = \delta_{\alpha}(\alpha').$$

This specification yields the following joint distribution, which basically amounts to a PPMx with a fixed smoothing parameter:

$$p(\mathbf{z},\alpha;X) = p(\mathbf{z}|\alpha;X) \propto \prod_{h=1}^{H} c(S_h) g(X_h^*)^\alpha = \mathscr{W}_{V,H} \prod_{h=1}^{H} (1-\sigma)_{n_h-1} g(X_h^*)^\alpha. \tag{4.8}$$

With this model, there is no (non-trivial) posterior inference on α to perform; moreover, α acts as a user-defined smoothing parameter. In fact, it is clear from Equation (4.6) that α regulates the amount of information provided by the node attributes in the distribution of the partitions $\mathbf{z}$. This is particularly clear taking the logarithm of the PPMx:

$$\log p(\mathbf{z}|X,\alpha) = \sum_{h=1}^{H} \log c(S_h) + \alpha \sum_{h=1}^{H} \log g(X_h^*) - \log k(\alpha).$$

Notice that $\alpha = 0$ corresponds to the case of a PPM prior with no supervision from the node attributes. We furtherly validate our intuition regarding the influence of α by sampling from the prior distribution for various values of α, specifically $\alpha = 0$ (no supervision), $\alpha = 1$ (weak supervision), $\alpha = 5$ (middle supervision) and $\alpha = 500$ (strong supervision). This allows us to see how the induced prior partitions change according to the smoothing parameter. Prior sampling is possible through the specification of a standard Gibbs sampler, which is feasible since PPMx endowed with Gibbs-type distributions have closed-form urn schemes:

$$p(z_v = h|\mathbf{z}_{-v},\alpha;X) \propto \begin{cases} \mathscr{W}_{V,H_{-v}}(n_h-\sigma)\frac{g(X_{h,-v}^*\cup\{\mathbf{x}_v\})^\alpha}{g(X_{h,-v}^*)^\alpha} & \text{for } h = 1,\dots,H_{-v}, \\ \mathscr{W}_{V,H_{-v}+1} g(\mathbf{x}_v)^\alpha & \text{for } h = H_{-v}+1. \end{cases} \tag{4.9}$$

In the subsequent experiments, the cohesion function $c(\cdot)$ we chose to model the partitions a priori is defined by the Gnedin process, which yields $\sigma = -1$, $\mathscr{W}_{V,H} = (\gamma)_{V-H}\prod_{h=1}^{H-1}(h^2 - \gamma h)/\prod_{v=1}^{V-1}(v^2+\gamma v)$ for some $\gamma \in (0,1)$. For a Gnedin process with parameter $\gamma \in (0,1)$, the abovementioned prior urn scheme becomes

$$p(z_v = h|\mathbf{z}_{-v},\alpha;X) \propto \begin{cases} (n_{h,-v}+1)(V-H_{-v}+\gamma)\frac{g(X_{h,-v}^*\cup\{\mathbf{x}_v\})^\alpha}{g(X_{h,-v}^*)^\alpha} & \text{for } h = 1,\dots,H_{-v}, \\ (H_{-v}^2 - \gamma H_{-v}) g(\mathbf{x}_v)^\alpha & \text{for } h = H_{-v}+1. \end{cases}$$

The next paragraphs contain the random partitions induced by the prior distribution, with different degree of information provided by the node attributes.

Sampling from the prior, with no spatial information We report the result of 5'000 samples from the law on the partitions induced by the Gnedin process [Gnedin and Pitman, 2004, Gnedin, 2010], with $\gamma = 0.3$ (corresponding to a conservative framework, where the prior expected number of clusters approximately equal to ~ 35). A burn-in of 500 iterations has been applied. Precisely, we sample from

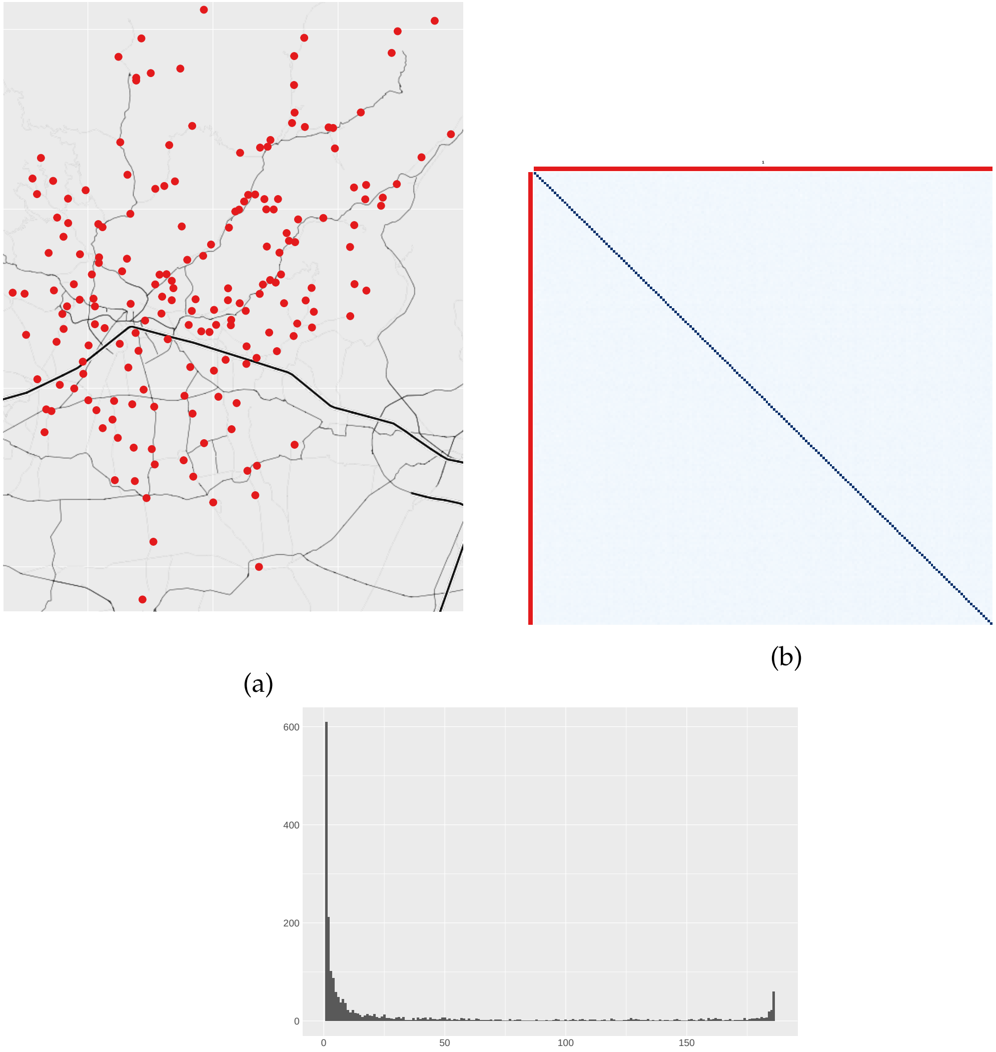
$$p(\mathbf{z}) = (\gamma)_{V-H} \frac{\prod_{h=1}^{H-1}(h^2 - \gamma h)}{\prod_{v=1}^{V-1}(v^2 + \gamma v)} \prod_{h=1}^{H} (2)_{n_h - 1}. \tag{4.10}$$

The Gnedin process is particularly suitable to model the clusters in the application of interest, since it has a finite prior expected number of clusters which can be learned from the process at hand. Among other advantages, the Gnedin process has a simple analytical expression for the urn scheme (Equation (4.2.1)) and for the distribution of the number of clusters H for $V \to \infty$:

$$p(H = h) = \frac{\gamma(1-\gamma)_{h-1}}{h!}.$$

Such a law has a mode for $H = 1$, that is $p(\mathbf{z})$ has a mode in the configuration with all the nodes in the same cluster, heavy tails and infinite expectation in the number of clusters, for infinite data [Gnedin, 2010]. Hence, the associated partition law favors parsimonious representations of the block structure of transportation patterns among municipalities, potentially more interpretable and usable from an operational perspective, but still preserves a positive prior mass for a high number of clusters. Figure 4.1 shows the partition induced by the prior distribution for $V = 186$ nodes of interest (obtained by minimizing the Variation of Information measure — see Wade and Ghahramani [2018] and Sections 3.4.4, 4.3), the prior coclustering matrix and the prior distribution of H. In the scatterplot, the points are placed in the geographical coordinates of the municipalities they represent. As expected, given a substantial number of nodes, the prior mode is for $H = 1$ and thus the prior partition estimate is a cluster containing all the instances. However, the off-diagonal values of the coclustering matrix are not equal to 0, and the distribution of H shows a heavy tail, giving a prior positive mass to every possible $H = 1, \ldots, 186$.

Sampling from the prior: α fixed The aim of this paragraph is to show that different values of the smoothing parameter α actually induce dissimilar prior partitions of the nodes of interest. The setting and procedure are identical to the previous paragraph. Figure 4.2 contains the partitions induced by the described prior PPMx: the data points are displayed according to latitude and longitude, with colors according to the prior induced partition, for different values of α: (a) $\alpha = 1$, (b) $\alpha = 5$, (c) $\alpha = 500$. Such values are on different scales: the reason for that is throughly explained in Section 4.4. The relationship between increasing α and the emergence of more radial clusters is evident. However, if α is too low (e.g. $\alpha = 1$), the information provided by the covariates is basically null and the estimated partition is indistinguishable from the unsupervised case. As a result, we can state that, a priori, we are effectively promoting the formation of radial clusters by changing the smoothing parameter α, as desired. Figure 4.3 shows the prior coclustering matrices, induced by the very same PPMx, for different values of α.

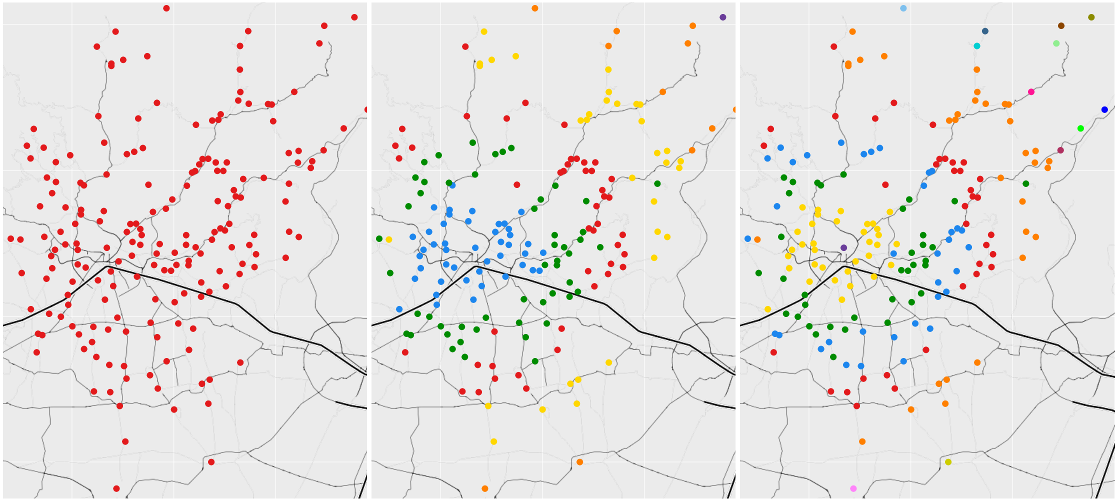


(a)

(b)

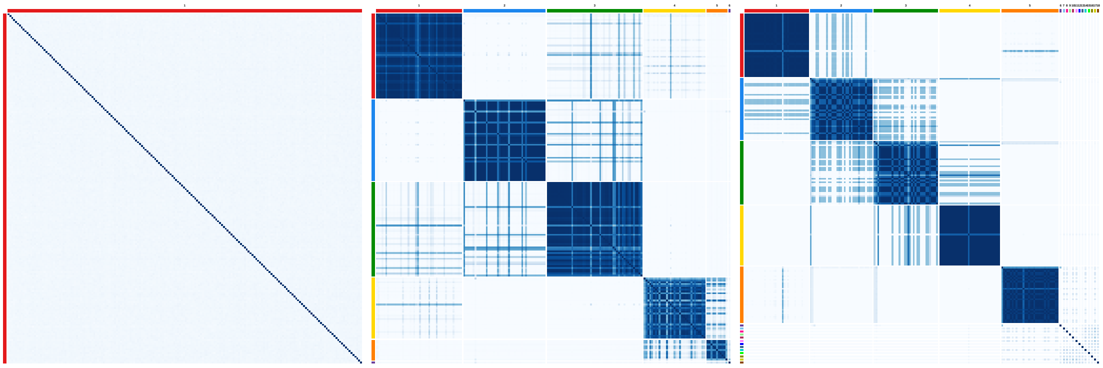
(c)

Figure 4.1: Partition induced by the (unsupervised) product partition model, i.e. a standard Gnedin process: (a) data points in the dataset of interest, displayed according to latitude and longitude and colored with respect to the prior induced partition; (b) coclustering matrix of the data points, with side colors according to the prior induced partition; (c) prior distribution of the number of clusters H.



(a) (b) (c)

Figure 4.2: Partitions induced by a supervised product partition model, endowed with Gnedin process and a Student's kernel as similarity function. The supervision is provided by the distance from Bergamo. The data points are displayed according to latitude and longitude, with colors with respect to the prior induced partition, for different values of α: (a) $\alpha = 1$, (b) $\alpha = 5$, (c) $\alpha = 500$.



(a) (b) (c)

Figure 4.3: Coclustering matrices induced by a supervised product partition model, endowed with Gnedin process and a Student's kernel as similarity function. The supervision is provided by the distance from Bergamo. Rows and columns are reordered according to the prior induced partition, for different values of α: (a) $\alpha = 1$, (b) $\alpha = 5$, (c) $\alpha = 500$.

Poisson extended stochastic block model, with inferred smoothing parameter

Even though we can visually explore the partitions induced by the prior distribution on the covariate space for different values of α (at least in our application), it is crucial to have the capability to learn the smoothing parameter through posterior inference. This ability is highly desirable in order to enhance the modeling process. However, PPMx presents intrinsic difficulties in estimating their parameters a posteriori (see also Chapter 3, Section 3.3.1). In fact, given a general, non-trivial prior $\alpha \sim p(\alpha)$, and considering a PPMx as prior model for $p(\mathbf{z}|\alpha;X)$, in order to be able to learn α from such a model sampling from its posterior through a Gibbs sampler, we would need the following distribution:

$$p(\alpha|\mathbf{z};X) \propto p(\mathbf{z}|\alpha;X)p(\alpha) = \frac{1}{k(\alpha)} \prod_{h=1}^{H} c(S_h)g(X_h^*)^{\alpha} p(\alpha). \tag{4.11}$$

Once again, we face the normalising constant $k(\alpha)$ which cannot be discarded (since it depends on the value of the smoothing parameter), and cannot be computed in closed form. To circumvent this issue, the rationale of the pESBM is to define a new, joint prior distribution on the couple $(\mathbf{z},\alpha)$ such that a) $p(\mathbf{z}|\alpha;X)$ is the PPMx in Equation (4.6) and b) $p(\alpha|\mathbf{z};X)$ is easy to sample from. This will allow us to set up a Gibbs sampler to infer both $\alpha,\mathbf{z}$. Such a joint distribution can be defined by means of an auxiliary law $\pi(\alpha)$ on α which provides easy integration to obtain $p(\alpha|\mathbf{z};X)$. In our case, we define $\pi(\alpha)$ to be:

$$\pi(\alpha) \sim \mathrm{TN}_{[0,\infty)}(\mu,\sigma^2), \tag{4.12}$$

that is we use as auxiliary distribution a Normal with mean μ and variance σ^2, truncated in $[0,\infty)$ (i.e. a truncated normal). In this way, we obtain the joint distribution of $(\mathbf{z},\alpha)$ as

$$\begin{aligned} p(\mathbf{z},\alpha;X) &\overset{\text{def}}{=} p(\mathbf{z}|\alpha;X)\cdot\pi(\alpha) \qquad (4.13)\\ &= \frac{1}{k^*}\left\{\prod_{h=1}^{H} c(S_h)\right\}\exp\left\{-\frac{\alpha^2}{2\sigma^2} + \frac{2\alpha}{2\sigma^2}\left(\mu + \sigma^2 \sum_{h=1}^{H} \log g(X_h^*)\right)\right\}\mathbb{1}_{\{\alpha\geq 0\}}, \end{aligned}$$

where k^* is a (new) normalising constant defined as

$$k^* = \sum_{\mathbf{z}\in\mathcal{Z}} \int_0^{\infty} \left\{\prod_{h=1}^{H} c(S_h)\right\}\exp\left\{-\frac{\alpha^2}{2\sigma^2} + \frac{2\alpha}{2\sigma^2}\left(\mu + \sigma^2 \sum_{h=1}^{H} \log g(X_h^*)\right)\right\}\mathrm{d}\alpha.$$

Notice that the auxiliary function $\pi(\alpha)$ is not the prior distribution of the smoothing parameter. However, the joint prior of Equation (4.13) allows the derivation of the actual, marginal prior of α:

$$\begin{aligned} p(\alpha;X) &= \sum_{\mathbf{z}\in\mathcal{Z}} p(\mathbf{z},\alpha;X) \qquad (4.14) \\ &= \sum_{\mathbf{z}\in\mathcal{Z}} \frac{k_{TN}}{k^*}\left\{\prod_{h=1}^{H} c(S_h)\right\}\exp\left\{-\frac{1}{2\sigma^2}\left(\mu+\sum_{h=1}^{H}\log g(X_h^*)\right)^2\right\}\cdot \mathrm{TN}_{[0,\infty)}\left(\mu+\sigma^2\sum_{h=1}^{H}\log g(X_h^*),\sigma^2\right) \\ &= \sum_{\mathbf{z}\in\mathcal{Z}} w_{\mathbf{z}}\mathrm{TN}_{[0,\infty)}\left(\mu+\sigma^2\sum_{h=1}^{H}\log g(X_h^*),\sigma^2\right), \end{aligned}$$

that is a mixture of truncated normals with weights $w_{\mathbf{z}}$ summing to 1. The quantity k_{TN} is the normalising constant of the truncated normal in $[0,\infty)$ with mean $\mu+\sigma^2\sum_{h=1}^{H}\log g(X_h^*)$ and variance σ^2. The specification of the joint prior in Equation (4.13) also enables a straightforward computation of the conditional distribution $p(\alpha|\mathbf{z};X)$, which yields

$$\alpha|\mathbf{z};X \sim \mathrm{TN}_{[0,\infty)}\left(\mu+\sigma^2\sum_{h=1}^{H}\log g(X_h^*),\sigma^2\right). \qquad (4.15)$$

The detailed computations are reported in the Appendix. This suggests that we can enhance the sampling process using a standard Gibbs sampler to also sample from the posterior distribution of α, just adding a computationally simple step. Finally, it is important to exercise caution when selecting the two hyperparameters, μ and σ^2. It is worth noting that, due to the definition of the similarity function described in Section 4.2.1, we observe that $g(X_h^*)\in[1/e,e]$ for all $h=1,\dots,H$. Consequently, $\log g(X_h^*)\in[-1,1]$ and $\sum_{h=1}^{H}\log g(X_h^*)\in[-H,H]$. This implies that the mean of the truncated normal distribution is constrained and cannot approach negative infinity, as it would be the case if we used the untransformed similarity. Additionally, this constraint prevents the posterior sampling from being concentrated solely in the tails of a normal distribution truncated in the positive interval for small values of μ (e.g. $\mu=0$). However, the choice of μ is crucial (as evident in the prior samples displayed in Figure 4.4): the higher it is, the higher the (spatial) smoothing induced by the model.

Sampling from the prior: α random Figures 4.4 and 4.5 report respectively the partition and the coclustering structure induced by the prior distribution $p(\mathbf{z},\alpha;X)$ of Equation (4.13), for different values of μ and σ^2. The prior samples are obtained through a standard Gibbs sampler, which also updates α according to the distribution in (4.15). The setting and the choice of the hyperparameters is the same of the previous experiments. Figure 4.4 shows the data points of interest, representing municipalities, in their geographical coordinates (latitude and longitude): the colors are with respect to the induced prior estimation obtained by minimising the Variation of Information measure [Wade and Ghahramani, 2018]. Once again, increasing μ (and thus shifting to the right the posterior distribution of α) results in a higher amount of information provided by the node attributes in X (that is, the distance of each municipality from Bergamo) and, in this application, in more radial clusters. The clusters also appear to have a smoother radial shape. Figure 4.5 shows the corresponding coclustering matrices, with rows and columns reordered according to the induced partition, and side colors denoting such a clustering. More interestingly, Figure 4.6 reports the estimated prior marginal distribution of α according to different values of μ, σ^2, with the red line corresponding to the prior mean of α, respectively equal to $\bar{\alpha} = 0.030$ for $\mu = 0, \sigma^2 = 0.1$; $\bar{\alpha} = 5.196$ for $\mu = 5, \sigma^2 = 0.25$ and $\bar{\alpha} = 219.936$ for $\mu = 500, \sigma^2 = 25$. As expected, the higher μ, the higher the induced α.

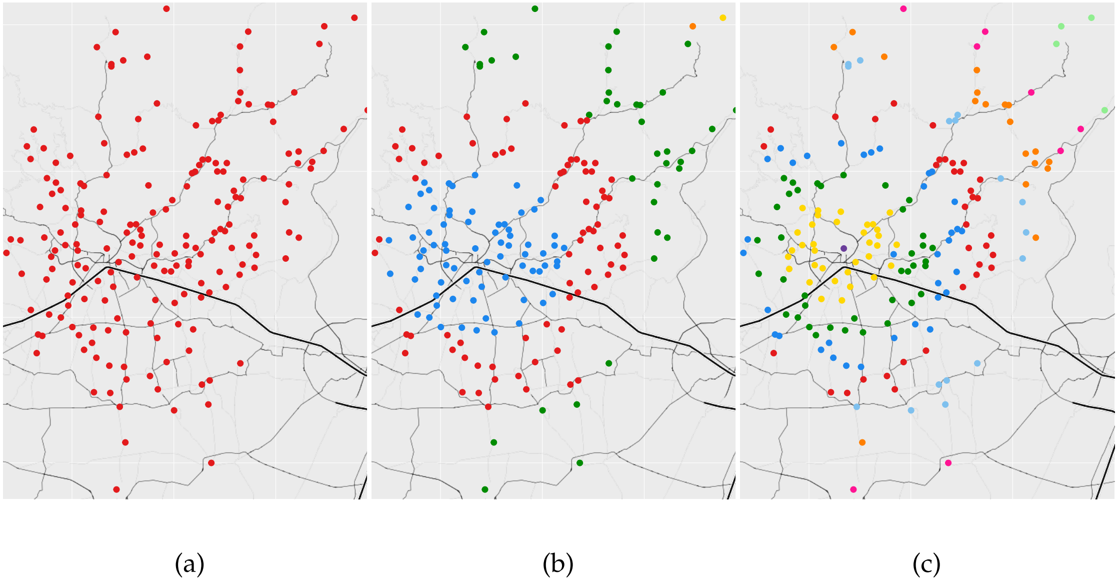
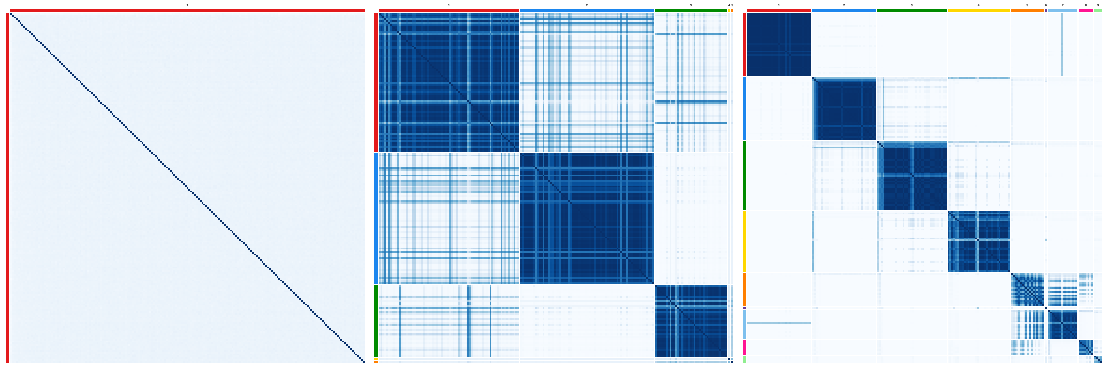


Figure 4.4: Partitions induced by the joint prior distribution of the Poisson extended stochastic block model, endowed with Gnedin process and a Student's kernel as similarity function. The supervision is provided by the distance from Bergamo. The data points are displayed according to latitude and longitude, with colors with respect to the prior induced partition, for different values of μ, σ^2: (a) $\mu = 0, \sigma^2 = 0.1$, (b) $\mu = 5, \sigma^2 = 0.25$, (c) $\mu = 500, \sigma^2 = 25$.

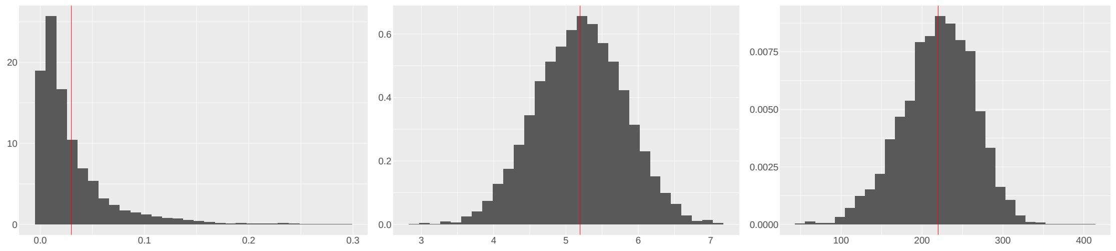
4.3 Posterior computation and inference

In this section, we outline the inferential framework used to sample from the posterior distribution of the pESBM discussed in Section 4.2. Specifically, we describe the sampling procedure for obtaining the posterior distribution of $\mathbf{z}$ in the pESBM in the two scenarios described in Section 4.2.1, namely the pESBM with a fixed and an inferred smoothing parameter. Additionally, we provide the corresponding point estimates.



(a) (b) (c)

Figure 4.5: Coclustering matrices induced by the joint prior distribution of the Poisson extended stochastic block model, endowed with Gnedin process and a Student's kernel as similarity function. The supervision is provided by the distance from Bergamo. Rows and columns are reordered according to the prior induced partition, for different values of μ, σ^2: (a) $\mu = 0, \sigma^2 = 0.1$, (b) $\mu = 5, \sigma^2 = 0.25$, (c) $\mu = 500, \sigma^2 = 25$.



(a) (b) (c)

Figure 4.6: Marginal distribution of α induced by the joint prior distribution of the Poisson extended stochastic block model, endowed with Gnedin process and a Student's kernel as similarity function, for different values of μ, σ^2: (a) $\mu = 0, \sigma^2 = 0.1$, (b) $\mu = 5, \sigma^2 = 0.25$, (c) $\mu = 500, \sigma^2 = 25$. The supervision is provided by the distance from Bergamo. The red line is the posterior mean.

4.3.1 Posterior computation

In this section, we present the procedure used to sample from the posterior of the pESBM discussed in Section 4.2. Specifically, we focus on the Gibbs sampler for the pESBM with a random smoothing parameter, with the prior defined in Equation 4.13. It is important to note that if we want to estimate the pESBM with a fixed value for α (as described in Equation 4.8), we can utilize the same algorithm, omitting the sampling step for the smoothing parameter. Posterior inference for the partition provided by the pESBM in (4.1)–(4.3) is carried out via a collapsed Gibbs sampler. In particular, since the cluster-specific connection rates λ_{hk} in (4.1)–(4.2) are not of direct interest, we follow the common practice of treating them as nuisance parameters and marginalizing them out, thus obtaining

$$\begin{aligned} p(Y|\mathbf{z}) = \left(\frac{b^a}{\Gamma(a)}\right)^{\frac{H^2}{2}} \prod_{h=1}^{H}\prod_{k=1}^{h-1} \frac{1}{\prod_{u,v:z_u=h,z_v=k} y_{uv}!} \frac{\Gamma(m_{kh}+a)}{(n_h n_k+b)^{m_{kh}+a}} \cdot \\ \cdot \frac{1}{\prod_{u<v:z_u=z_v=h} y_{uv}!} \frac{\Gamma(m_{hh}+a)}{[n_h(n_h-1)/2+b]^{m_{hh}+a}}, \end{aligned} \tag{4.16}$$

where m_{hk} is the sum of edge weights between clusters h and k, while m_{hh} is the sum of edge weights within cluster h. We then derive a collapsed Gibbs sampler that, at each iteration, updates the cluster membership of each node v according to its full conditional

$$p(z_v = h|\mathbf{z}_{-v}, \alpha, Y, X) \propto p(z_v = h|\mathbf{z}_{-v}, \alpha, X) \cdot \frac{p(Y|z_v = h, \mathbf{z}_{-v})}{p(Y_{-v}|\mathbf{z}_{-v})}, \tag{4.17}$$

with subscript $-v$ denoting objects and quantities obtained by removing node v. The last term in (4.17) can be computed from Equation (4.16), yielding

$$\begin{aligned} &\frac{p(Y|z_v = h, \mathbf{z}_{-v})}{p(Y_{-v}|\mathbf{z}_{-v})} = \\ &= \frac{1}{\prod_{u\neq v} y_{uv}!} \prod_{\substack{k=1\\k\neq h}}^{H_{-v}} \frac{\Gamma(m_{hk}^{-v} + \sum_{u:z_u=k} y_{vu} + a)}{\Gamma(m_{hk}^{-v}+a)} \frac{[n_h^{-v} n_k^{-v} + b]^{m_{hk}^{-v}+a}}{[(n_h^{-v}+1)n_k^{-v}+b]^{m_{hk}^{-v}+\sum_{u:z_u=k} y_{uv}+a}} \cdot \\ &\qquad \cdot \frac{\Gamma(m_{hh}^{-v} + \sum_{u:z_u=h} y_{uv} + a)}{\Gamma(m_{hh}^{-v}+a)} \frac{[\frac{n_h^{-v}(n_h^{-v}-1)}{2} + b]^{m_{hh}^{-v}+a}}{[\frac{n_h^{-v}(n_h^{-v}+1)}{2} + b]^{m_{hh}^{-v}+\sum_{u:z_u=h} y_{uv}+a}}, \end{aligned}$$

while $p(z_v = h|\mathbf{z}_{-v}, \alpha, X)$ is the urn scheme of the chosen supervised Gibbs-type distribution. As already mentioned, for the transportation network application in Section 4.4, among Gibbs-type priors, we opt for a Gnedin process [Gnedin and Pitman, 2004], under which, as $V \to \infty$, the number of clusters is random but not infinite like under,for example, the Dirichlet process.

It is possible to conduct posterior inference for the spatial smoothing parameter α as well. Due to the joint prior specification in Equation (4.13), the smoothing parameter possesses a closed-form full-conditional posterior distribution, independent of the data Y given the partition $\mathbf{z}$ (detailed computations are included in the Appendix)

$$\alpha|\mathbf{z};X \sim \mathrm{TN}_{[0,\infty)}\Big(\mu+\sigma^2\sum_{h=1}^{H}\log g(X_h^*),\sigma^2\Big).$$

Therefore, given the partition $\mathbf{z}$ sampled from the posterior in the previous iteration, updating α is computationally efficient, since it requires only one sample from a Truncated Normal distribution, with the mean being dependent on the partition $\mathbf{z}$. Algorithm 2 summarizes the collapsed Gibbs sampler for the pESBM in (4.13) with inference on the spatial smoothing parameter.

Algorithm 2: Posterior sampling of the joint distribution of $\alpha,\mathbf{z}$ yielded by the pESBM.

```
Input: Adjacency matrix $Y_{V\times V}$, design matrix $X_{V\times p}$, similarity function,
       hyperparameters.
Output: Posterior samples from $p(\alpha,\mathbf{z}|Y,X)$.
for each iteration do
    Sample $\mathbf{z}|\alpha;Y,X$ using a Gibbs sampler.
    for $v = 1,\dots,V$ do
        Update $z_v \sim p(z_v|\mathbf{z}^{-v};Y,X)$ using Equations (4.9), (4.16) and (4.17);
    end
    Sample $\alpha|\mathbf{z};X$ .
    Update $\alpha|\mathbf{z};X \sim \mathrm{TN}_{[0,\infty)}\Big(\mu+\sigma^2\sum_{h=1}^{H}\log g(X_h^*),\sigma^2\Big)$.
end
```

4.3.2 Point estimation

The Gibbs sampler described above (as well as the one employed to generate prior samples discussed in Section 4.2.1) produces joint samples of $\mathbf{z}$ and α. However, we may want to also obtain a point estimate from these samples. Estimating α is relatively straightforward, as we can utilize any reasonable numerical summary of its posterior distribution, such as Maximum A Posteriori (MAP) estimation, mean, or median. On the other hand, obtaining a representative partition from a sample of different partitions is more challenging. To tackle this, we have opted to follow the approach proposed by Wade and Ghahramani [2018], which provides a point estimate from partition samples by minimizing the Variation of Information. The approach is described in detail in Section 3.4.4.

4.4 Application: transport networks

In this section, we illustrate the application of the two versions of the pESBM (discussed in Section 4.2) on real-world data from a public transport company. We specifically examine the take of these two approaches on the final partitioning of the data, fixing at first the smoothing parameter and then inferring it. Through these visual illustrations, our objective is to highlight the distinctions and effectiveness of the pESBM in capturing the latent structures within the transportation network. By analyzing the inferred partitions, we aim to provide a clear understanding of how each approach finds different clusters in the data. This analysis enables us to assess the strengths and limitations of the two models in capturing the complex relationships and dependencies in the transportation network.

4.4.1 Data

The dataset used in this study was collected by Azienda dei Trasporti di Bergamo (ATB), the public transport company operating in Bergamo and its surrounding areas. Each row of the original dataset contains information pertaining to a user's account, identified by their unique card number. This information includes details such as the purchase of monthly subscriptions, gender, year of birth, as well as the origin and destination of the subscription. The dataset specifically covers the year 2019, which was chosen as the last complete year without the disruptive effects of the COVID-19 pandemic. As the pandemic significantly impacted the transportation network by reducing the number of unique users and municipalities involved, our focus was on analyzing 2019 as a representative *regular* year. Any examination of differences in the transportation network during the subsequent years will be considered as part of future work. Our objective is to cluster municipalities based on both their connectivity patterns and their geographical position. To accomplish this, we aggregate the relevant data into a weighted network where each node represents a municipality. The edges in the network indicate the presence of public transport subscriptions between two municipalities, with the weight of each edge representing the cumulative duration of such subscriptions. Self-loops, which indicate subscriptions within the same municipality, are disregarded for our clustering analysis. We treat the network as undirected, since subscriptions enable users to freely travel between two municipalities. As node attributes, we employ the distance of each municipality from Bergamo. This information is obtained by extracting the municipality longitude and latitude from `simplemaps.com/data/it-cities`. The inclusion of such node attribute encourages municipalities at the same distance from Bergamo to cluster together, promoting the formation of radial clusters.

Figure 4.7 displays the transportation network in two equivalent ways: on the top left, the resulting graph with nodes (representing $V = 186$ municipalities) located on the map of Bergamo and surroundings, with a zommed-in detail in the top right plot. On the bottom, the corresponding adjacency matrix. In the adjacency matrix, each element Y_{uv} represents the number of monthly subscriptions between municipality u and municipality v. To visualize this information in the graph, the color of each edge is proportional to the corresponding subscription count. The two figures clearly indicate that the network of interest exhibits a star-shaped structure, with Bergamo as the center (and the main hub) and the only node connected to all the others. Furthermore, the periphery of the network predominantly shows connections only to the center of the graph. On the other hand, municipalities that are closer to Bergamo exhibit some interconnections among themselves.

Figure 4.8 includes three exploratory plots of the network of interest. Figure 4.8a represents the degree distribution across the entire network: it is positively skewed, with a single observation (corresponding to the city of Bergamo) on the far-right tail. This was highly predictable, since Bergamo is by far the most populated city of its province, with about 120′000 inhabitants (the second most populated town of the province is Treviglio, with about 30′000 residents). Moreover, the public transport system is mainly radial, with the vast majority of bus lines radiating from Bergamo. Figure 4.8b contains a scatterplot of the log-degree of each municipality versus its number of residents. The overall trend is vaguely increasing, but not showing a clear linear relationship between the number of residents and the degree of the nodes (the correlation index is $R^2 = 0.02$). Thus, there seems to be highly populated municipalities where private or alternative transportation means are preferred over public transportation provided by ATB. This suggests that these nodes may be underserved by ATB, or it is possible that alternative transportation companies or services cater to the transportation needs of these municipalities. Finally, Figure 4.8c displays the node log-degree versus the distance of each municipality from Bergamo: generally speaking, the data do not show specific functional relationship, and the linear correlation is weak ($R^2 = 0.0003$). Once again, it is surprising to observe that the dataset includes municipalities in close proximity to Bergamo with a relatively low number of monthly subscriptions. This discrepancy suggests that despite their geographical proximity to Bergamo, these municipalities have a lower usage of public transport services provided by ATB: this could be again due to the presence of transport means supplied by other companies.

The next subsections contain the results obtained using three algorithmic competitors and the pESBM. As for the pESBM, we run Algorithm 2 on the data described in Section 4.4.1, fixing the smoothing parameter α at first, and then inferring it. The algorithm outputs samples from the posterior distribution $p(\mathbf{z}, \alpha | Y, X)$ of the pESBM. The resulting point estimate $\hat{\mathbf{z}}$ is obtained by minimizing the Variation of Information [Wade and Ghahramani, 2018] (see Section 4.3 for more details). We run the algorithm for 10′000 iterations, with 1′000 iterations as burn in, initializing it with a random initial partition. The pESBM is endowed with a Gnedin marginal prior cohesion function, with parameter $\gamma = 0.3$, which corresponds to an unsupervised prior expectation of 35 clusters (a conservative estimate). As for the gamma prior parameters a and b, we set them to $a = \bar{Y}^2$, $b = \bar{Y}$, in order to have $\mathbb{E}\big[\mathbb{E}[Y|\lambda_{hk}]\big] = \mathbb{E}[\lambda_{hk}] = \bar{Y}$ and $\text{var}\big[\mathbb{E}[Y|\lambda_{hk}]\big] = \text{var}[\lambda_{hk}] = 1$, following a moment-matching criterion but setting the variance a priori to 1. Finally, we set the hyperparameters of the

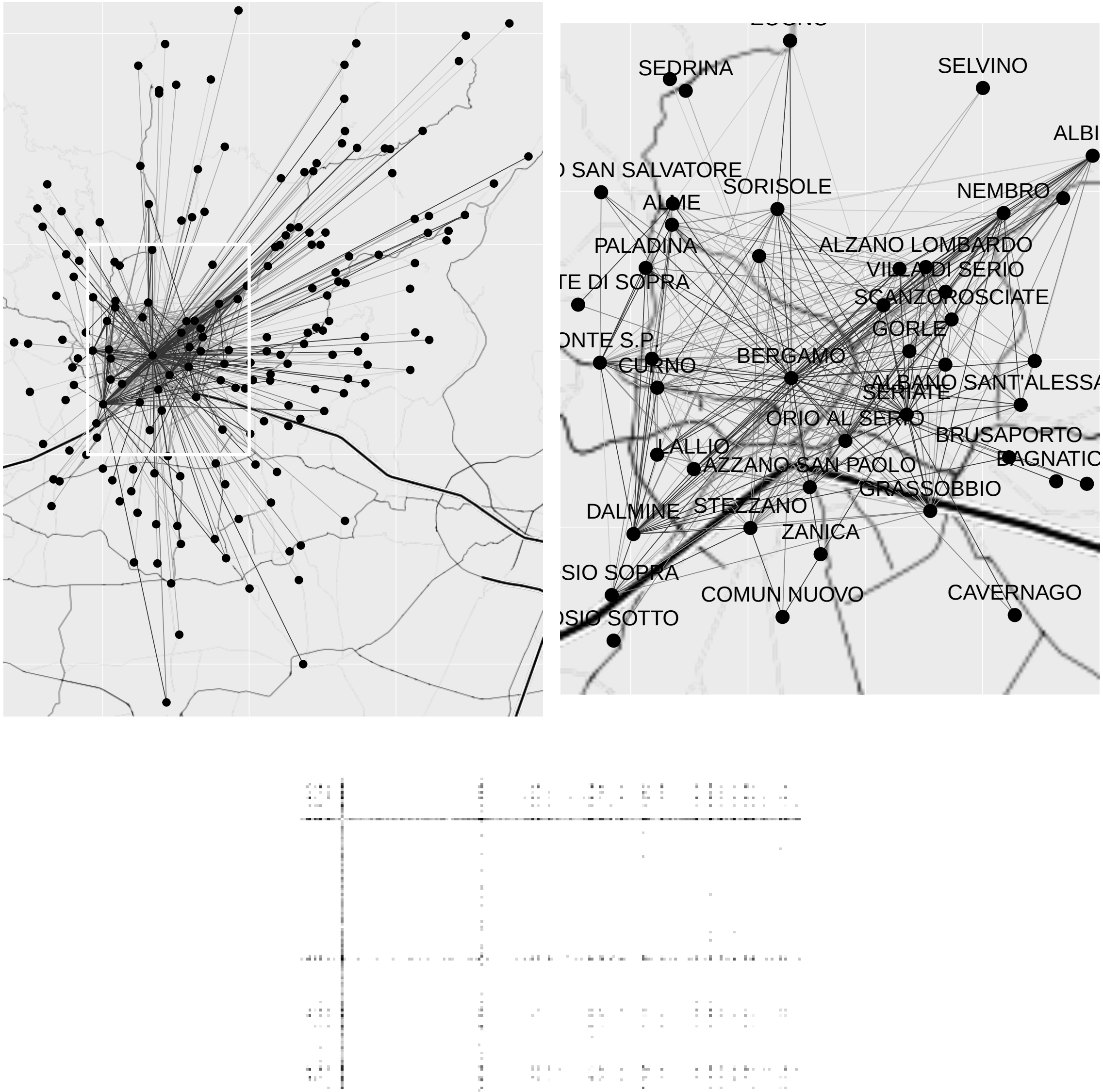
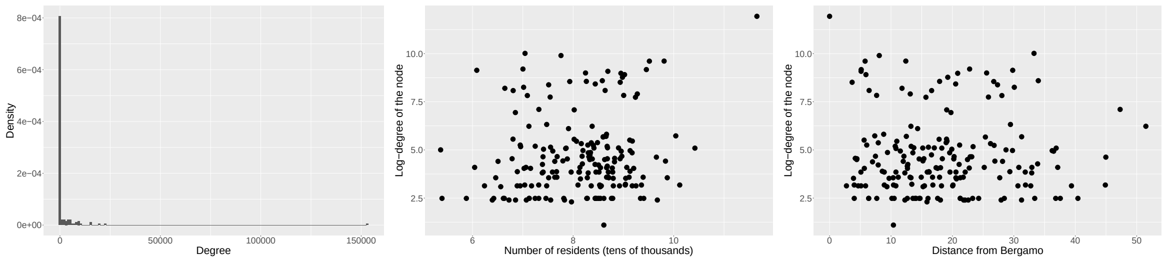


Figure 4.7: Graphical representation (top left) with a zoomed-in detail (top right), and adjacency matrix (bottom) of the considered transportation network. Nodes are placed according to the longitude and latitude of the corresponding municipality. Edge colors are proportional to edge weights (the darker, the higher), which in turn are given by the total number of subscription months among the involved municipalities.



(a) Node degree distribution (b) Node log-degree vs Residents (c) Node degree distribution

Figure 4.8: (a) Node degree distribution; (b) node log-degree versus number of residents in the corresponding municipality; (c) node log-degree versus distance of the corresponding municipality from Bergamo.

similarity function to $\mu_0 = 0, \Sigma_0 = 1, \nu_0 = 2, k_0 = 0.1$. Subsections 4.4.3, 4.4.4 and 4.4.5 contain respectively the results for the unsupervised and the supervised model, with and without inference on the smoothing parameter. Recall that the supervision is carried out by means of the node attributes, containing the standardized distance of each municipality from Bergamo. The convergence of the algorithms used in the study is assessed through traceplots, which are provided in the Appendix. These checks ensure that the algorithms have reached a stable state and that the results obtained are reliable.

4.4.2 Algorithmic competitors

We consider three competitor algorithmic approaches for weighted networks. First, we apply the Louvain algorithm [Blondel et al., 2008] and spectral clustering [Von Luxburg, 2007], where the latter has been applied to the Laplacian matrix of the weighted graph. At this stage, we are not leveraging the information of the covariate yet. Then, we use the covariate-assisted spectral clustering [Binkiewicz et al., 2017], where the node attributes are provided by the distance of each municipality from Bergamo. Figure 4.9 displays the partition of the nodes according to these algorithmic clustering procedures for the weighted network of interest. The number of clusters for each algorithm has been chosen with the elbow method, looking at the scree plot of the eigenvalues of the input matrix. These algorithmic techniques output a clustering either too or not enough coarsened that is arguably of little use for guiding policy decisions, by either the transport company or public administrations. The covariate-assisted spectral clustering discerns a periphery (in blue), far from Bergamo, but the clusters are not radial at all, despite the carefully chosen attribute node supervising the partition estimation. None of algorithms acknowledges the unique nature of Bergamo in the network topology.

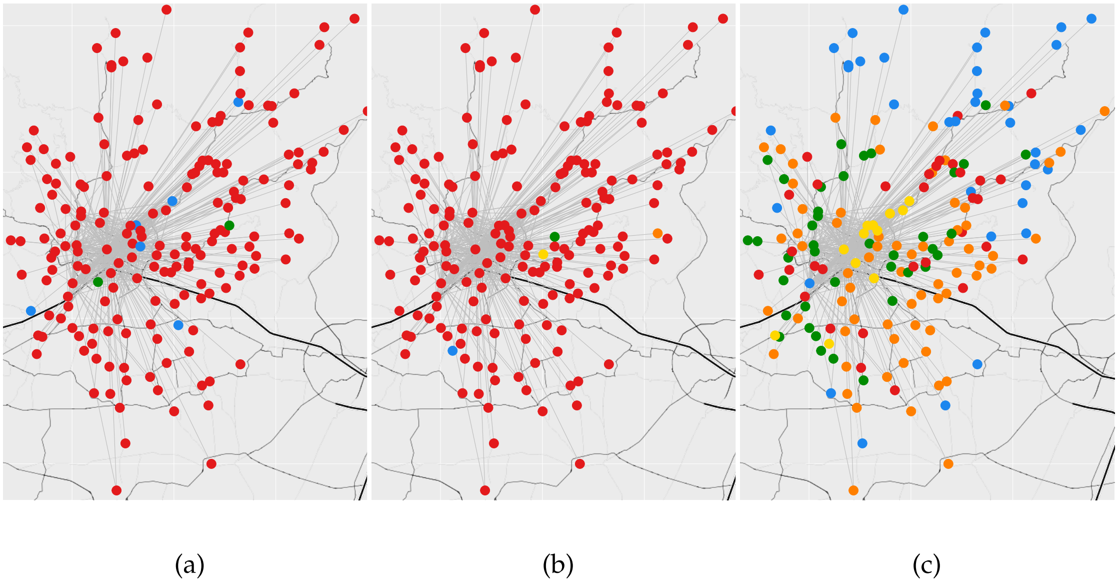


Figure 4.9: Clustering provided by (a) Louvain algorithm, (b) spectral algorithm, (c) covariate-assisted spectral clustering on the transportation network of interest.

4.4.3 Unsupervised clustering

Figures 4.10 and 4.11 report respectively the graph and the reordered adjacency matrix, where the colors denote the posterior point partition provided by the pESBM trained with no information from the node attributes (i.e. for $\alpha = 0$). Figure 4.10 shows an interesting geographical displacement of the clusters: the red and yellow clusters group municipalities on the external belt of the province of Bergamo, mainly in the surrounding valleys. Moreover, the two groups display the same connectivity patterns, since they link exclusively to Bergamo, with essentially no connectivity with the other clusters. However, there is a notable disparity in the number of monthly subscriptions generated by the municipalities in each cluster, with an average of approximately 27 subscriptions in the red cluster and around 156 subscriptions in the yellow cluster. This discrepancy can be attributed to a crucial difference between the two groups, namely the average population size of the municipalities. The red cluster comprises municipalities with an average population of approximately 3′500 residents, whereas the yellow cluster consists of municipalities with an average population of around 6′000 residents. This observation highlights that the pESBM model can effectively distinguish more populated municipalities without an explicit training, solely relying on a proxy variable, such as the number of subscriptions. We consider this to be a valuable information not to be overlooked in the partition process and consequently do not incorporate any degree correction in our model, as done in previous work. On that note, we argue that the degree correction [Karrer and Newman, 2011, Herlau et al., 2014] is not suitable for the application of interest: indeed, we want to cluster together municipalities with the same connectivity patterns, where the similarity also takes into account the weights of the edges and the degree of the nodes, without correcting or disregarding such an important source of information. Lastly, we notice some unexpected municipalities in the red group located in the west part of the Bergamo province, such as Presezzo, Bonate Sopra, Bonate Sotto, and Mapello. These are towns with a high number of residents and located close to Bergamo. Normally, one would expect these municipalities to have a high number of subscriptions to the main city and be clustered with the municipalities in the green and blue clusters, which represent areas with higher connectivity and transportation services. However, in this case, the municipalities west of Bergamo are primarily served by another transport company instead of ATB. As a result, the number of monthly subscriptions for the bus lines provided by ATB is relatively low. This low subscription rate makes them comparable to (and clustered with) small municipalities in the valleys, which may have limited public transportation options and lower population densities.

The green cluster encompasses the municipalities in the first belt surrounding the main city of Bergamo. This cluster is distinguished by a significant level of internal movement, as depicted in Figure 4.11. The municipalities included in the green cluster are notable because they are exclusively served by ATB, without the presence of other public transport companies. This observation indicates that the residents of these municipalities heavily rely on ATB as their primary mode of public transportation. The blue cluster is similar, in the sense that it includes municipalities on the second radial belt from Bergamo. The blue cluster is characterized by a low (almost non-existent) internal movement, but also by strong connections with the green municipalities. It also shows low connections with the yellow cluster, but not with the red one. Lastly, we have the singleton cluster consisting of Bergamo, the main hub and center of the network. It is desirable for Bergamo to be clustered by itself due to its unique position in the network topology. Being the only node connected to all the other municipalities, clustering Bergamo separately acknowledges its distinct role and connectivity within the transportation network as the origin and destination of the vast majority of bus lines. Figure 4.12 displays the network representation of the inferred clusters: each node represents one group and edges are weighted using a plug-in estimate of the between-clusters rate matrix Λ. Node sizes are proportional to group cardinalities, while edge colors represent their weights (the darker the higher). From Figure 4.12, it is clear that the Bergamo orange singleton plays a hub role. Similarly, the green cluster is connected to all the other groups, and also shows internal movements, represented by the self-loops in the metagraph. On the contrary, the largest clusters (red, yellow) are also the ones with less and more moderate connections. Moreover, the red and yellow clusters show similar connectivity patterns, but, as noticed before, the yellow one groups municipalities with double the average number of residents with respect to the red cluster, a fact that obviously has a strong influence on the resulting number of subscriptions and therefore on the strength of the connections. Table 4.1 shows the sample estimate of the average block-connection, which are an estimate of expected monthly subscriptions between two different municipalities according to their cluster labels. Notice that the estimated within rate $\hat{\lambda}_{\text{orange, orange}}$ representing the number of subscriptions within the city of Bergamo is 0. However, in the original dataset the number of monthly subscriptions within Bergamo is 610. This is due to the fact that, by definition, the pESBM do not model self-loops and thus this information has not been used to train the model. Nevertheless, the estimated between-clusters rate Λ can potentially be useful for administrative decisions and optimization processes.

4.4.4 Supervised clustering, with fixed smoothing parameter

In this section, we report the results for the pESBM described in Section 4.2 with the spatial-smoothing parameter α fixed (that is, for $\pi(\alpha) = \delta_{\alpha}(\alpha)$). In particular, we consider three scenarios: no supervision ($\alpha = 0$, already shown in Section 4.4.3 but also reported below for convenience), moderate supervision ($\alpha = 5$) and strong supervision ($\alpha = 500$). Such values have been chosen according to the study of the induced partition a priori (see Section 4.2.1). A similar study for the partitions induced a posteriori on a different set of data is contained in Ghidini et al. [2023a,b]. Notice that the chosen values for α are on wildly different scales: from the moderate-supervision scenario to the highly-supervised one, we increase such parameter by a factor of 100. This is due to the fact that the (unnormalised) log-cohesion and the log-similarity function in Equation 4.7 are on different scales: in particular, the log-cohesion is 100 times bigger than the log-similarity. Thus, to value more the information coming from the node attributes with respect to the connections, we have to increase the smoothing parameter accordingly.

Figures 4.13 and 4.14 show the transportation graph and the reordered adjacency matrix, where the colors denote the posterior point partition provided by the pESBM trained with different levels of information specified by the distance of each municipality from Bergamo. It is clear that increasing α, we also increase the spatial smoothing provided by the node attribute, inducing more radial clusters. The pESBM with a moderate supervision ($\alpha = 5$) still distinguishes the two groups of municipalities in the surrounding valleys of Bergamo with the same connectivity pattern but different number of subscriptions (red and green nodes in Figure 4.13b). Additionally, the singleton Bergamo is maintained as a distinct cluster. However, in this case, the municipalities in the first and second belt are combined into a single cluster, denoted by the blue group. Notably, the blue cluster exhibits significant internal connections, distinguishing it from the other clusters. With a strong supervision ($\alpha = 500$), the clusters are clearly radial: the municipalities in the valleys are now clustered according to their distance from Bergamo (which is overwhelming the information provided by the strength of their connections). However, the closest municipalities to Bergamo are again partition into a first and second belt, similarly to the pESBM for $\alpha = 0$. Bergamo is still clustered as a singleton.

Figure 4.15 displays the posterior coclustering matrices obtained for the estimation of

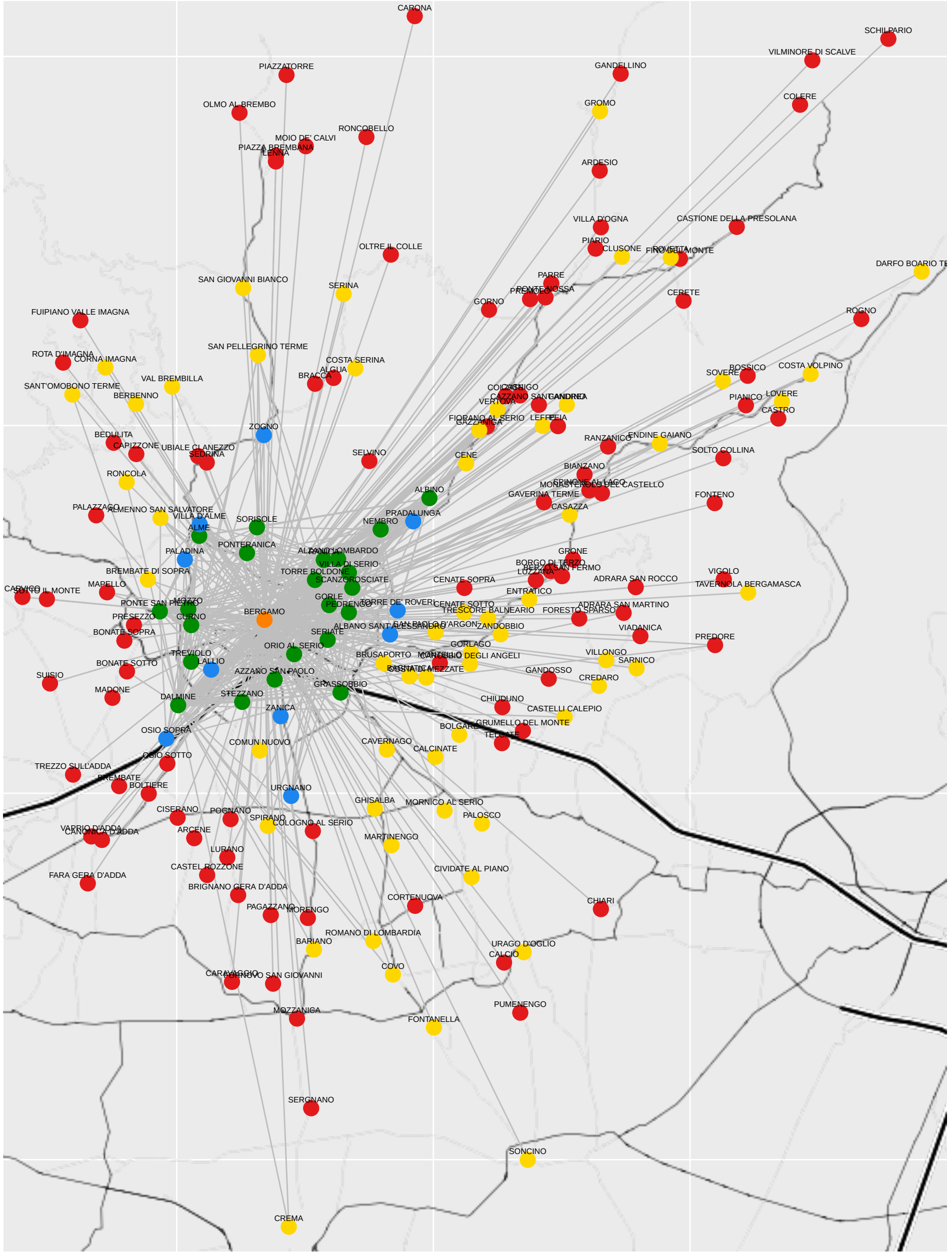


Figure 4.10: Graphical representation of the considered transportation network: node colors correspond to the posterior point estimate of the partition under the Poisson extended stochastic block model endowed with a Gnedin prior.

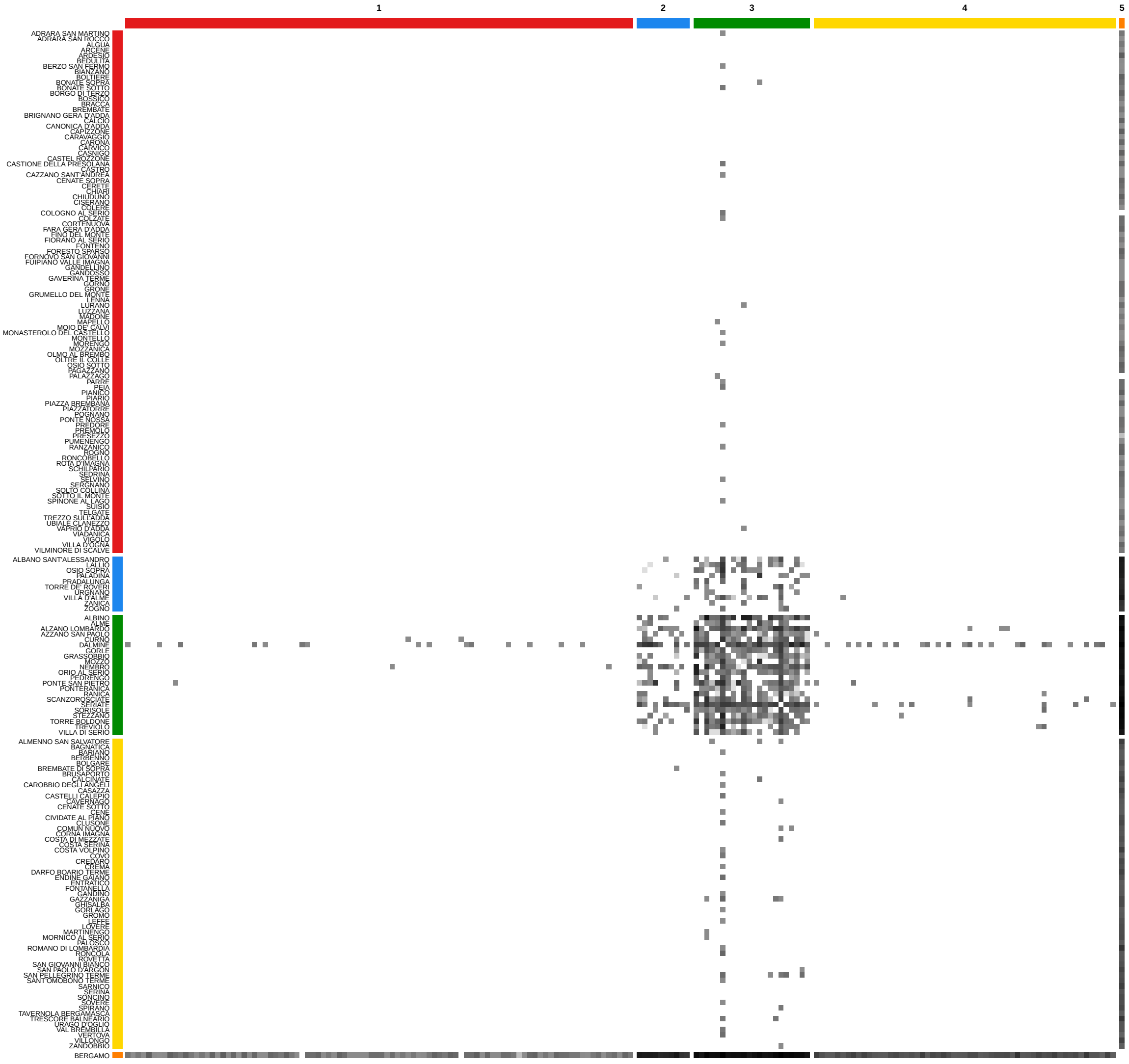


Figure 4.11: Adjacency matrix of the considered transportation network: the side colors correspond to the posterior point estimate of the partition under the Poisson extended stochastic block model, endowed with a Gnedin prior. The left column contains the names of the municipalities, the top row the labels of the clusters.

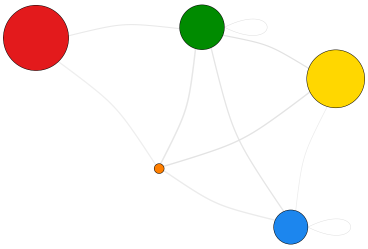
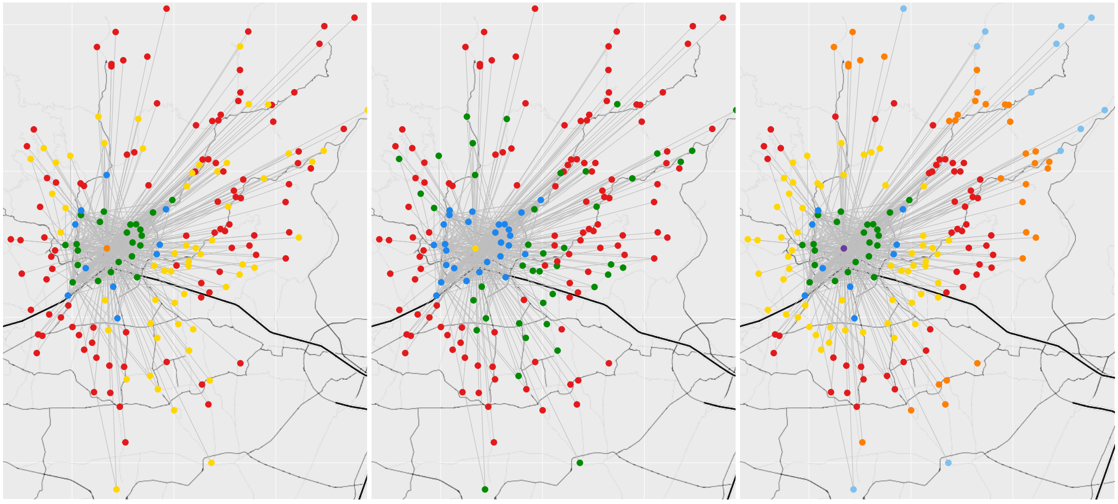


Figure 4.12: Network representation of the inferred clusters in the transportation network. Each node denotes one group and edges are weighted by the estimated block rates Λ. Node sizes are proportional to cluster cardinalities, edge colors to the corresponding weights (the darker, the higher).

	🟥	🟦	🟩	🟨	🟧
🟥	0.00	0.00	0.13	0.00	27.62
🟦	0.00	0.49	15.71	0.02	1329.80
🟩	0.13	15.71	102.65	0.66	5797.64
🟨	0.00	0.02	0.66	0.00	156.60
🟧	27.62	1329.80	5797.64	156.60	0

Table 4.1: Empirical estimates for Λ given by the sample average weight of edges between clusters.



(a)

(b)

(c)

Figure 4.13: Graphical representation of the considered transportation network: node colors correspond to the posterior point estimate of the partition under the Poisson extended stochastic block model, endowed with a Gnedin prior and supervised with the distance of each municipality from Bergamo. Three scenarios are presented, according to the fixed value of the smoothing parameter: (a) no supervision ($\alpha = 0$), (b) moderate supervision ($\alpha = 5$), (c) strong supervision ($\alpha = 500$).

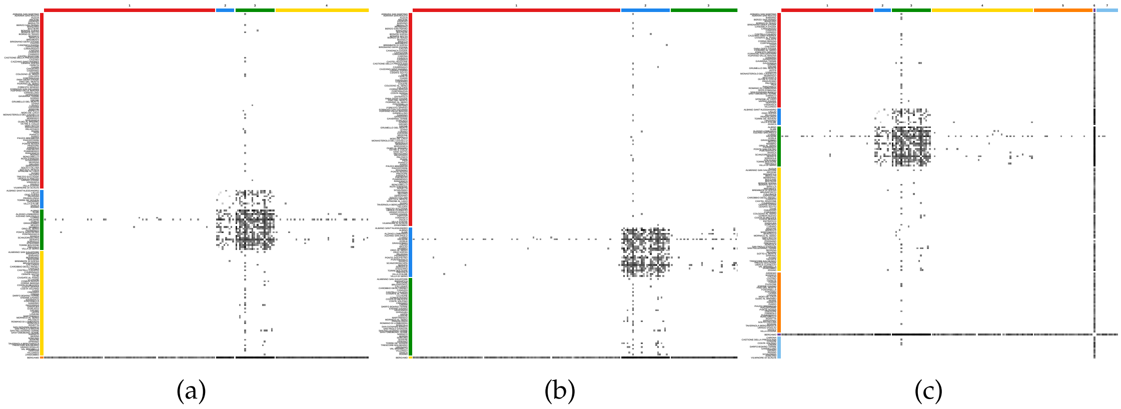

(a) (b) (c)

Figure 4.14: Adjacency matrix of the considered transportation network: the side colors correspond to the posterior point estimate of the partition under the Poisson extended stochastic block model, endowed with a Gnedin prior and supervised with the distance of each municipality from Bergamo. Three scenarios are presented, according to the fixed value of the smoothing parameter: (a) no supervision ($\alpha = 0$), (b) moderate supervision ($\alpha = 5$), (c) strong supervision ($\alpha = 500$).

the pESBM with a fixed smoothing parameter, specifically for $\alpha = 0, 5, 500$. These matrices provide insights into the level of uncertainty associated with the posterior estimation of the clusters. For $\alpha = 0$ and $\alpha = 5$, the coclustering matrices display less uncertainty, leading to relatively certain point estimates provided by the pESBM. However, when more information is incorporated from the node attributes (i.e., $\alpha = 500$), the posterior estimation exhibits increased uncertainty, particularly for nearby groups of municipalities that share similar connectivity patterns (e.g. light blue-orange clusters, blue-green clusters). The presence of uncertainty in these cases highlights the complexity and overlapping nature of the underlying structures in the data, emphasizing the need to carefully consider the influence of node attributes in the clustering process. One possible reason for the uncertainty could be that the two sources of information, represented by X and Y, are not perfectly aligned. For example, certain nodes may exhibit different characteristics or behaviors in terms of their connections (Y) even though their their attributes (X) are similar, causing uncertainty in the clustering process.

4.4.5 Supervised clustering, with inferred smoothing parameter

In this section, we report the results for the pESBM described in Section 4.2 with additional inference on the spatial-smoothing parameter α. In particular, we consider again three scenarios, induced by the prior distribution on α: low supervision ($\mu = 0, \sigma^2 = 0.1$), moderate supervision ($\mu = 5, \sigma^2 = 0.25$) and strong supervision ($\mu = 500, \sigma^2 = 25$). Such values have been chosen according to the study of the induced partition a priori (see Section 4.2.1). In these experimental setting, we vary the hyperparameter μ, σ^2 jointly and accordingly,

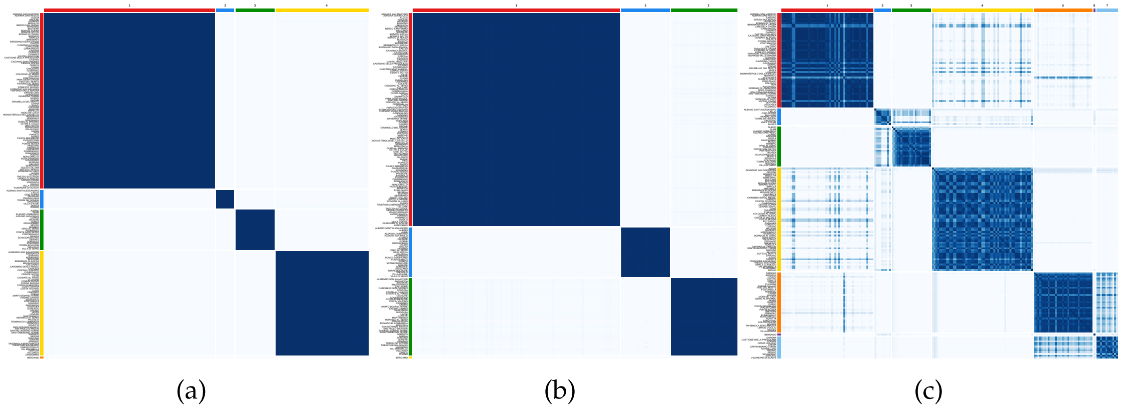


Figure 4.15: Posterior coclustering matrix of the considered transportation network: the side colors correspond to the posterior point estimate of the partition under the Poisson extended stochastic block model, endowed with a Gnedin prior and supervised with the distance of each municipality from Bergamo. Three scenarios are presented, according to the fixed value of the smoothing parameter: (a) no supervision ($\alpha = 0$), (b) moderate supervision ($\alpha = 5$), (c) strong supervision ($\alpha = 500$).

to define a sensible auxiliary distribution $\pi(\alpha)$. However, we argue that while the value of μ is clearly influencing the distribution of α a posteriori (in particular, the higher it is, the more shifted the distribution to the right), the hyperparameter σ^2 is not impacting the posterior that much. The reason is clear from the shape of Equation (4.15). There, we can see that the higher σ^2 is, the higher the impact of $\log g(X_h^*)$ mean but also the higher the marginal, posterior variance of α. Thus, even if increasing σ^2 changes the mean, it also changes the variance resulting in a bigger support around the posterior marginal mean. To support such conjecture, several experiments with μ fixed and varying σ^2 are reported in the Appendix.

Figures 4.16 and 4.17 show the transportation graph and the reordered adjacency matrix, where the colors denote the posterior point partition provided by the pESBM trained with the three scenarios above, and Figure 4.18 the corresponding posterior coclustering matrix. The results obtained with the pESBM with an inferred α are consistent with those obtained using the pESBM with a fixed smoothing parameter. However, with this second model, the partitions estimated with low and moderate supervision remain the same, indicating that the clustering structure is preserved for reasonable supervision levels. However, the level of uncertainty associated with the partitions is generally higher. When more information from the node attributes is incorporated, as in the case of moderate supervision, the posterior estimation exhibits higher uncertainty, as depicted in Figure 4.18. This suggests that incorporating additional information from the node attributes increases the complexity of the clustering problem and introduces more uncertainty in the estimation. Furthermore, as previously observed, the pESBM with high supervision induces radial clusters and also leads to even more increased uncertainty in the posterior estimation.

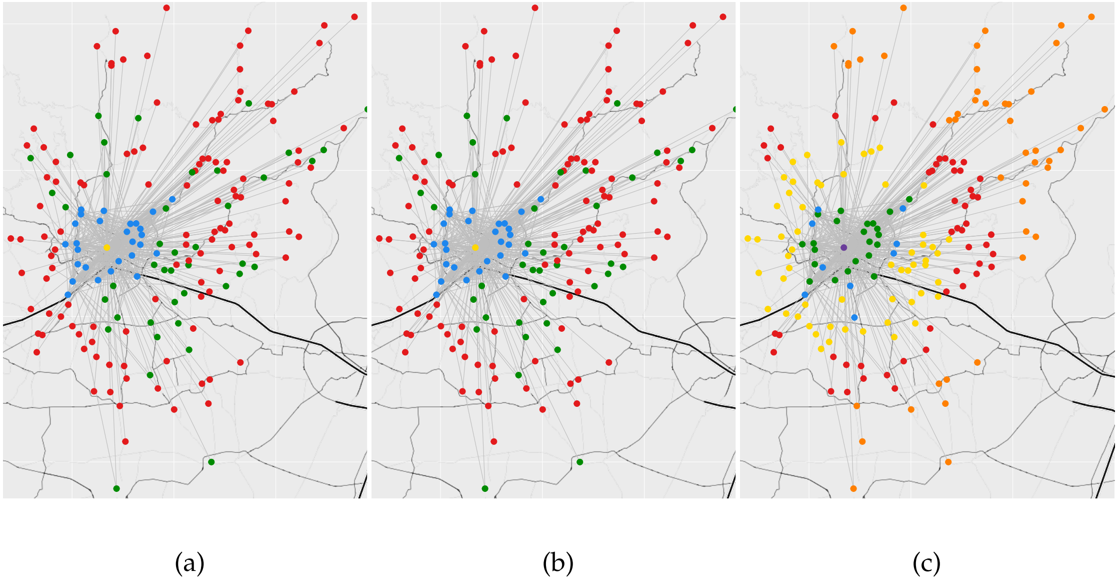


Figure 4.16: Graphical representation of the considered transportation network: node colors correspond to the posterior point estimate of the partition under the Poisson extended stochastic block model, endowed with a Gnedin prior and supervised with the distance of each municipality from Bergamo. Three scenarios are presented, according to the prior hyperparameters of the smoothing parameter: (a) low supervision ($\mu = 0, \sigma^2 = 0.1$), (b) moderate supervision ($\mu = 5, \sigma^2 = 0.25$), (c) strong supervision ($\mu = 500, \sigma^2 = 25$).

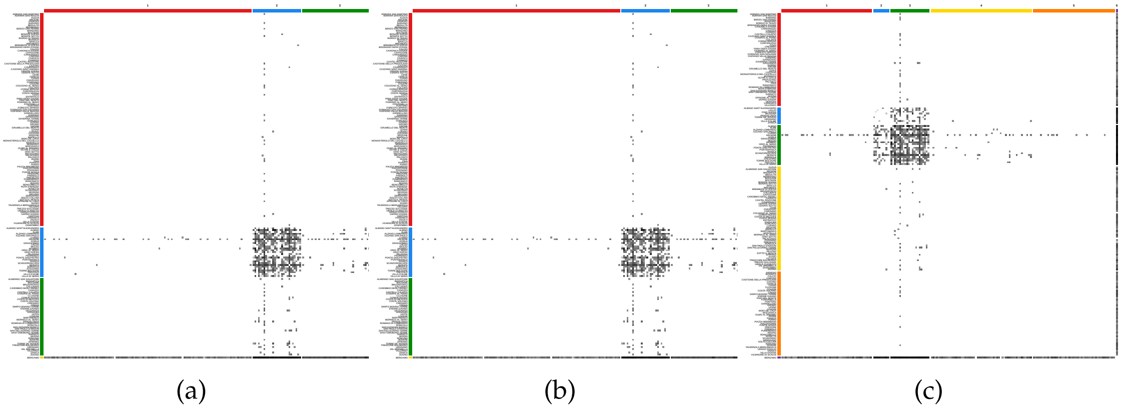
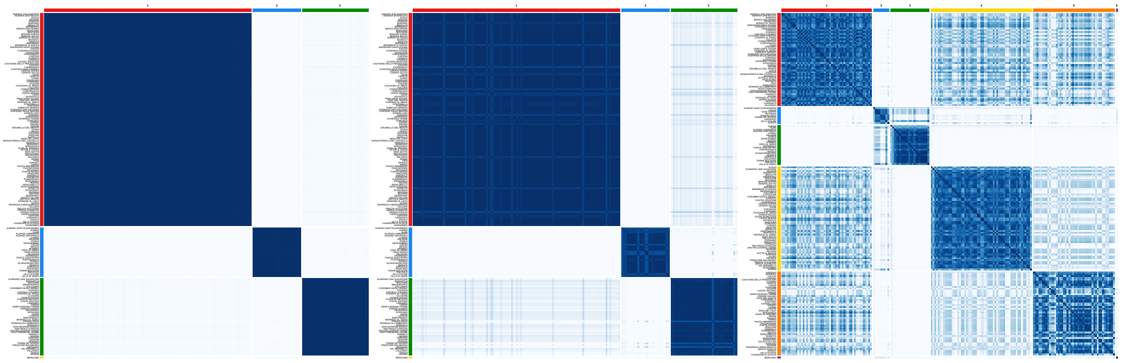


Figure 4.17: Adjacency matrix of the considered transportation network: the side colors correspond to the posterior point estimate of the partition under the Poisson extended stochastic block model, endowed with a Gnedin prior and supervised with the distance of each municipality from Bergamo. Three scenarios are presented, according to the prior hyperparameters of the smoothing parameter: (a) low supervision ($\mu = 0, \sigma^2 = 0.1$), (b) moderate supervision ($\mu = 5, \sigma^2 = 0.25$), (c) strong supervision ($\mu = 500, \sigma^2 = 25$).



(a) (b) (c)

Figure 4.18: Posterior coclustering matrix of the considered transportation network: the side colors correspond to the posterior point estimate of the partition under the Poisson extended stochastic block model, endowed with a Gnedin prior and supervised with the distance of each municipality from Bergamo. Three scenarios are presented, according to the prior hyperparameters of the smoothing parameter: (a) low supervision ($\mu = 0, \sigma^2 = 0.1$), (b) moderate supervision ($\mu = 5, \sigma^2 = 0.25$), (c) strong supervision ($\mu = 500, \sigma^2 = 25$).

4.4.6 Quantitative evaluation

In this subsection, we report a first, quantitative evaluation of the partitions obtained. Such evaluation can be carried out using standard goodness-of-fit indices for clustering, such as the Rand index or the Silhouette. Table 4.2 displays the Silhouette indices for the clusters obtained in Subsections 4.4.4 and 4.4.5. In general, as expected, models with an increasing level of supervision perform better when the Silhouette is computed taking into account not only the similarity with respect to the edge weights, but also with respect to the node attributes. Moreover, models with random α seems to perform at least as well as the models with fixed α, with a clear advantage in the scenario of low or high supervision.

	Low/no supervision	Moderate supervision	High supervision
α fixed - Sil(Y)	0.500	**0.545**	-0.127
α fixed - Sil(X,Y)	0.500	**0.545**	-0.126
α fixed - Sil(X)	-0.087	0.02	0.375
α random - Sil(Y)	**0.545**	**0.545**	-0.053
α random - Sil(X,Y)	**0.545**	**0.545**	-0.052
α random - Sil(X)	0.02	0.02	**0.403**

Table 4.2: Silhouette index computed for all the posterior point estimates obtained in Subsections 4.4.4 and 4.4.5.

4.5 Discussion and future research directions

In conclusion, this chapter has demonstrated the effectiveness of the Poisson extended stochastic block model (pESBM) in uncovering the underlying structures and patterns of a public transportation network. By incorporating both connectivity patterns and node attributes, such as the distance from the main hub, the pESBM enables the identification of meaningful clusters of municipalities which can be useful for policy decision as, e.g., the definition of new pricing zones. The results have provided valuable insights into the considered network, such as the distinction of underserved municipalities.

There are several directions for future and current work, as the development of this project is still in progress at the time of writing. Firstly, defining classes of similarity functions that consider administrative needs, such as measuring similarity based on polar coordinates or bus line alignment, can provide more tailored and relevant clustering results. Additionally, extending the model to multiplex networks, possibly incorporating different means of transportation, would enable a comprehensive analysis of transportation patterns and facilitate strategic decision-making for public transport companies. Integration of data from multiple public transport companies and the expansion of the analysis to a regional level can also provide a broader perspective on transportation networks, uncovering core-periphery structures and supporting administrative divisions. Multi-level clustering techniques can be employed to capture hierarchical organization within transportation networks, enabling a finer-grained understanding of network dynamics. Also, latent space models could be employed to study latent embeddings of the network. Their usage could be two-fold: on the one hand, geographical distance between two municipalities could be used as an edge covariate in order to find the latent embedding which is not explained by the geographical information. On the other hand, the prior on the latent node positions can be centered on the true geographical coordinates of the municipalities.

Methodologically, there are opportunities for further improvements in the estimation and computational aspects of the pESBM. One area of focus is the development of better inference methods for estimating the smoothing parameter α, with particular attention to reducing the reliance of the posterior density on the prior parameters. Furthermore, addressing computational complexity issues is crucial for improving the scalability and applicability of Bayesian nonparametric models like the pESBM. Efficient algorithms and techniques, such as approximate inference methods or parallel computing, could be explored to reduce the computational burden and facilitate the analysis of larger networks.

Lastly, considering the temporal variation and the impact of external factors on transportation networks, such as the COVID-19 pandemic, would provide valuable insights into evolving traffic flows and congestion patterns. Comparative analyses of pre-pandemic and post-pandemic years can shed light on the resilience and adaptability of transportation systems. Overall, this study has showcased the capabilities of the pESBM in capturing the complexity of transportation networks, thus potentially facilitating transportation planning, network optimization, and policy-making.

Appendix

We are interested in augmenting our pESBM with node attributes, incorporating such information using cohesion functions. Using the standard approach proposed by Muller et al. [2011], we can define the probability of a partition as follows

$$p(\mathbf{z};\alpha,X)\propto \mathscr{W}_{V,H}\prod_{h=1}^{H}(1-\sigma)_{n_h-1}g(X_h^*)^\alpha$$

where

- $\mathscr{W}_{V,H},\sigma$ are defined by the choice of the Gibbs-type prior,
- $g(X_h^*)$ is the cohesion function of cluster h.

According to the recipe proposed by Muller et al. [2011], a standard way to obtain a cohesion function as a marginal distribution is:

$$g(X_h^*)=\int\prod_{i:z_i=h}q(\mathbf{x}_i|\boldsymbol{\xi}_h)q(\boldsymbol{\xi}_h)\mathrm{d}\boldsymbol{\xi}_h.$$

We consider the case in which the node attributes are p-dimensional vectors $\mathbf{x}_i, i=1,\ldots,V$.

$$\mathbf{x}_i|\boldsymbol{\mu}_h,\Sigma_h\sim\mathscr{N}_p(\boldsymbol{\mu}_h,\Sigma_h),\qquad\text{for } i:z_i=h$$

which yields

$$q(\mathbf{x}_i|\boldsymbol{\xi}_h=(\boldsymbol{\mu}_h,\Sigma_h))\propto|\Sigma_h|^{-\frac{1}{2}}\exp\Big\{-\frac{p}{2}(\mathbf{x}_i-\boldsymbol{\mu}_h)^T\Sigma_h^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_h)\Big\}$$

The conjugate distribution of a multivariate gaussian random variable is the Normal-Inverse-χ^2 distribution, i.e.

$$\Sigma_h\sim\text{Inv-Wishart}_{\nu_0}(\Lambda_0^{-1})$$
$$\boldsymbol{\mu}_h|\Sigma_h\sim\mathscr{N}_p(\boldsymbol{\mu}_0,\frac{\Sigma_h}{k_0})$$

and the joint prior distribution is

$$p(\boldsymbol{\mu}_h,\Sigma_h)\propto|\Sigma_h|^{-\left(\frac{\nu_0+p}{2}+1\right)}\exp\Big\{-\frac{1}{2}\mathrm{tr}(\Lambda_0\Sigma_h^{-1})-\frac{k_0}{2}(\boldsymbol{\mu}_h-\boldsymbol{\mu}_0)^T\Sigma_h^{-1}(\boldsymbol{\mu}_h-\boldsymbol{\mu}_0)\Big\}.$$

Since it is a conjugate model, the posterior distribution is again a Normal-Inverse-χ^2 distribution, available at page 73 in Gelman et al. [2004] (recall that $X_h^* = \{x_i : z_i = h\}$):

$$p(\boldsymbol{\mu}_h, \Sigma_h | X_h^*) \propto |\Sigma_h|^{-\left(\frac{\nu_n+p}{2}+1\right)} \exp\Big\{ -\frac{1}{2}\text{tr}(\Lambda_n \Sigma_h^{-1}) - \frac{k_n}{2}(\boldsymbol{\mu}_h - \boldsymbol{\mu}_n)^T \Sigma_h^{-1} (\boldsymbol{\mu}_h - \boldsymbol{\mu}_n) \Big\}.$$

where

- $\boldsymbol{\mu}_n = \frac{k_0}{k_0+n_h}\boldsymbol{\mu}_0 + \frac{n_h}{k_0+n_h}\bar{\mathbf{x}}_h$;
- $k_n = k_0 + n_h$;
- $\nu_n = \nu_0 + n_h$;
- $\Lambda_n = \Lambda_0 + S + \frac{k_0 n_h}{k_0+n_h}(\bar{\mathbf{x}}_h - \boldsymbol{\mu}_0)(\bar{\mathbf{x}}_h - \boldsymbol{\mu}_0)^T$, $S = \sum_{i:z_i=h}(\mathbf{x}_i - \bar{\mathbf{x}}_h)(\mathbf{x}_i - \bar{\mathbf{x}}_h)^T$.

Thus, the cohesion function becomes (keeping in mind that $\xi_h = (\boldsymbol{\mu}_h, \Sigma_h)$):

$$g(X_h^*) = \int \prod_{i:z_i=h} q(\mathbf{x}_i|\xi_h) q(\boldsymbol{\xi}_h) \mathrm{d}\boldsymbol{\xi}_h \propto \int q(\boldsymbol{\xi}_h|X_h^*) \mathrm{d}\boldsymbol{\xi}_h$$

and using the posterior distribution above:

$$= \int |\Sigma_h|^{-\left(\frac{\nu_n+p}{2}+1\right)} \exp\Big\{ -\frac{1}{2}\text{tr}(\Lambda_n \Sigma_h^{-1}) - \frac{k_n}{2}(\boldsymbol{\mu}_h - \boldsymbol{\mu}_n)^T \Sigma_h^{-1} (\boldsymbol{\mu}_h - \boldsymbol{\mu}_n) \Big\} \mathrm{d}\boldsymbol{\mu}_h \, \mathrm{d}\Sigma_h$$

Here, we can exploit the normalising constant of the multivariate gaussian and of the Inverse-Wishart distribution getting:

$$= (2\pi k_n)^{-\frac{1}{2}} \int |\Sigma_h|^{-\left(\frac{\nu_n+p+1}{2}\right)} \exp\{-\frac{1}{2}\text{tr}(\Lambda_n \Sigma_h^{-1})\} \, \mathrm{d}\Sigma_h$$

$$\propto k_n^{-\frac{1}{2}} 2^{\frac{p}{2}\nu_n} \Gamma_p(\frac{\nu_n}{2}) |\Lambda_n|^{-\frac{\nu_n}{2}}$$

$$= \frac{2^{p(\nu_0+n_h)/2}}{\sqrt{k_0+n_h}} \Gamma_p\left(\frac{\nu_0+n_h}{2}\right) \cdot \left| \Sigma_0^{-1} + \frac{k_0 n_h}{k_0+n_h}(\bar{\mathbf{x}}_h - \boldsymbol{\mu}_0)^T(\bar{\mathbf{x}}_h - \boldsymbol{\mu}_0) + \sum_{v:z_v=h}(\mathbf{x}_v - \bar{\mathbf{x}}_h)^T(\mathbf{x}_v - \bar{\mathbf{x}}_h) \right|^{-\frac{\nu_0+n_h}{2}}$$

Supervised full conditional distribution As before, we compute the posterior probabilities of the labels according to the formula:

$$p(z^v = h | \mathbf{z}^{-v}, X, Y) \propto p(z^v = h | \mathbf{z}^{-v}, X) \frac{p(Y | z^v = h, \mathbf{z}^{-v})}{p(Y^{-v} | \mathbf{z}^{-v})}$$

where

$$p(z^v = h | \mathbf{z}^{-v}, X) \propto \frac{p(z^v = h, \mathbf{z}^{-v} | X)}{p(\mathbf{z}^{-v} | X^{-v})} = \frac{p(z^v = h, \mathbf{z}^{-v})}{p(\mathbf{z}^{-v})} \frac{g({X^*}_h^v)^\alpha}{g({X^*}_h^{-v})^\alpha} = p(z^v = h | \mathbf{z}^{-v}) \frac{g({X^*}_h{}^v)^\alpha}{g({X^*}_h{}^{-v})^\alpha}$$

Notice that $p(z^v = h|\mathbf{z}^{-v})$ is provided by the urn scheme of the Gibbs-type distribution of choice. Thus, we are just left with the computation of the ratio $\frac{g(X^{*v}_{h})}{g(X^{*-v}_{h})}$, where X^{*v}_{h} is the set of covariates in cluster h adding the v node to such a cluster, and X^{*-v}_{h} is the set of covariates in cluster h, discarding node v. Taking the log posterior distribution, we get:

$$\begin{aligned}
&\log p(z^v = h|\mathbf{z}^{-v}, X, Y)\\
&\propto \log p(z^v = h|\mathbf{z}^{-v}, X) + \log p(Y|z^v = h, \mathbf{z}^{-v}) - \log p(Y^{-v}|\mathbf{z}^{-v})\\
&= \log p(z^v = h|\mathbf{z}^{-v}) + \alpha[\log g(X^{*v}_{h}) - \log g(X^{*-v}_{h})] + \log p(Y|z^v = h, \mathbf{z}^{-v}, X) - \log p(Y^{-v}|\mathbf{z}^{-v}, X^{-v})
\end{aligned}$$

and we can compute $\log g(X^{*v}_{h}) - \log g(X^{*-v}_{h})$ as follows:

$$\begin{aligned}
\log g(X^{*}{}_{h}) = &-\frac{1}{2} log(k_0 + n_h) + (\nu_0 + n_h)\log 2 + \frac{1}{2}\log\pi + \log\Gamma\left(\frac{\nu_0 + n_h}{2}\right) + \log\Gamma\left(\frac{\nu_0 + n_h - 1}{2}\right)\\
&- \frac{\nu_0 + n_h}{2}\log|\Lambda_0 + \sum_{i:z_i = h}(\mathbf{x}_i - \bar{\mathbf{x}}_h)(\mathbf{x}_i - \bar{\mathbf{x}}_h)^T + \frac{k_0 n_h}{k_0 + n_h}(\bar{\mathbf{x}}_h - \boldsymbol{\mu}_0)(\bar{\mathbf{x}}_h - \boldsymbol{\mu}_0)^T|
\end{aligned}$$

As a consequence:

$$\begin{aligned}
&\log g(X^{*v}_{h}) - \log g(X^{*-v}_{h}) =\\
&= -\frac{1}{2}\log(k_0 + n_h^{-v} + 1) + \frac{1}{2}\log(k_0 + n_h^{-v}) + \log\Gamma\left(\frac{\nu_0 + n_h^{-v} + 1}{2}\right) - \log\Gamma\left(\frac{\nu_0 + n_h^{-v} - 1}{2}\right)\\
&- \frac{\nu_0 + n_h^{-v} + 1}{2}\log|\Lambda_0 + \sum_{i:z_i = h \text{ or } i = v}(x_i - \bar{\mathbf{x}}^v_h)(x_i - \bar{\mathbf{x}}^v_h)^T + \frac{k_0(n_h^{-v} + 1)}{k_0 + n_h^{-v} + 1}(\bar{\mathbf{x}}^v_h - \boldsymbol{\mu}_0)(\bar{\mathbf{x}}^v_h - \boldsymbol{\mu}_0)^T|\\
&+ \frac{\nu_0 + n_h^{-v}}{2}\log|\Lambda_0 + \sum_{i:z_i = h \text{ and } i \neq v}(\mathbf{x}_i - \bar{\mathbf{x}}_h^{-v})(\mathbf{x}_i - \bar{\mathbf{x}}_h^{-v})^T + \frac{k_0 n_h^{-v}}{k_0 + n_h^{-v}}(\bar{\mathbf{x}}_h^{-v} - \boldsymbol{\mu}_0)(\bar{\mathbf{x}}_h^{-v} - \boldsymbol{\mu}_0)^T|
\end{aligned}$$

for $h = 1, \ldots, H^{-v}$ and

$$\begin{aligned}
&\log g(X^{*v}_{h}) - \log g(X^{*-v}_{h}) =\\
&= -\frac{1}{2}\log(k_0 + 1) + \frac{1}{2}\log(k_0) + \log\Gamma\left(\frac{\nu_0 + 1}{2}\right) - \log\Gamma\left(\frac{\nu_0 - 1}{2}\right)\\
&- \frac{\nu_0 + 1}{2}\log|\Lambda_0 + \frac{k_0}{k_0 + 1}(\mathbf{x}_v - \boldsymbol{\mu}_0)(\mathbf{x}_v - \boldsymbol{\mu}_0)^T| + \frac{\nu_0}{2}\log|\Lambda_0|
\end{aligned}$$

for $h = H^{-v} + 1$.

Inference on the smoothing parameter α

In this section, we report the inferential framework for the smoothing parameter α. The trick is to exploit an auxiliary law $\pi(\cdot)$ on α which provides easy integration to obtain $p(\alpha|\mathbf{z})$.

In our case, we define $\pi(\cdot)$ to be:

$$\pi(\alpha) \sim \mathrm{TN}_{[0,\infty)}(\mu, \sigma^2),$$

that is we use as auxiliary distribution for α a normal with mean μ and variance σ^2, truncated in $[0,\infty)$ (i.e. a truncated normal). In this way, we obtain the joint distribution of $(\mathbf{z}, \alpha)$

$$\begin{aligned}
p(\mathbf{z},\alpha) &= p(\mathbf{z}|\alpha)\pi(\alpha) \\
&\propto \Big[\prod_{h=1}^{H} c(S_h)\Big]\exp\Big\{\alpha \sum_{h=1}^{H} \log g(X_h^*)\Big\} \cdot \exp\Big\{-\frac{(\alpha-\mu)^2}{2\sigma^2}\Big\}\mathbb{1}_{\{\alpha\geq 0\}} \\
&\propto \frac{1}{k^*}\Big\{\prod_{h=1}^{H} c(S_h)\Big\}\exp\Big\{-\frac{\alpha^2}{2\sigma^2} + \frac{2\alpha}{2\sigma^2}\Big(\mu + \sigma^2 \sum_{h=1}^{H} \log g(X_h^*)\Big)\Big\}\mathbb{1}_{\{\alpha\geq 0\}},
\end{aligned}$$

where k^* is a (new) normalising constant defined as

$$k^* = \sum_{\mathbf{z}\in\mathcal{Z}} \int_0^\infty \Big\{\prod_{h=1}^{H} c(S_h)\Big\}\exp\Big\{-\frac{\alpha^2}{2\sigma^2} + \frac{2\alpha}{2\sigma^2}\Big(\mu + \sigma^2 \sum_{h=1}^{H} \log g(X_h^*)\Big)\Big\}\mathrm{d}\alpha.$$

From here, the posterior distribution of the smoothing parameter can be obtained, using

$$\begin{aligned}
p(\alpha|\mathbf{z}) \propto p(\mathbf{z},\alpha) &\propto \exp\Big\{-\frac{\alpha^2}{2\sigma^2} + \frac{2\alpha}{2\sigma^2}\Big(\mu + \sigma^2 \sum_{h=1}^{H} \log g(X_h^*)\Big)\Big\}\mathbb{1}_{\{\alpha\geq 0\}} \\
&\propto \exp\Big\{-\frac{1}{2\sigma^2}[\alpha - (\mu + \sigma^2 \sum_{h=1}^{H} \log g(X_h^*))]^2\Big\}\mathbb{1}_{\{\alpha\geq 0\}} \\
&\Rightarrow \alpha|\mathbf{z} \sim \mathrm{TN}_{[0,\infty)}\Big(\mu + \sigma^2 \sum_{h=1}^{H} \log g(X_h^*), \sigma^2\Big).
\end{aligned}$$

Finally, we can also obtain the actual prior distribution of α as:

$$\begin{aligned}
p(\alpha) &= \sum_{\mathbf{z}\in\mathcal{Z}} p(\alpha,\mathbf{z}) \\
&= \sum_{\mathbf{z}\in\mathcal{Z}} \frac{1}{k^*}\Big[\prod_{h=1}^{H} c(S_h)\Big]\exp\Big\{-\frac{1}{2\sigma^2}(\mu + \sum_{h=1}^{H} \log g(X_h^*))^2\Big\}\exp\Big\{-\frac{1}{2\sigma^2}[\alpha - (\mu + \sigma^2 \sum_{h=1}^{H} \log g(X_h^*))]^2\Big\}\mathbb{1}_{\{\alpha\geq 0\}}
\end{aligned}$$

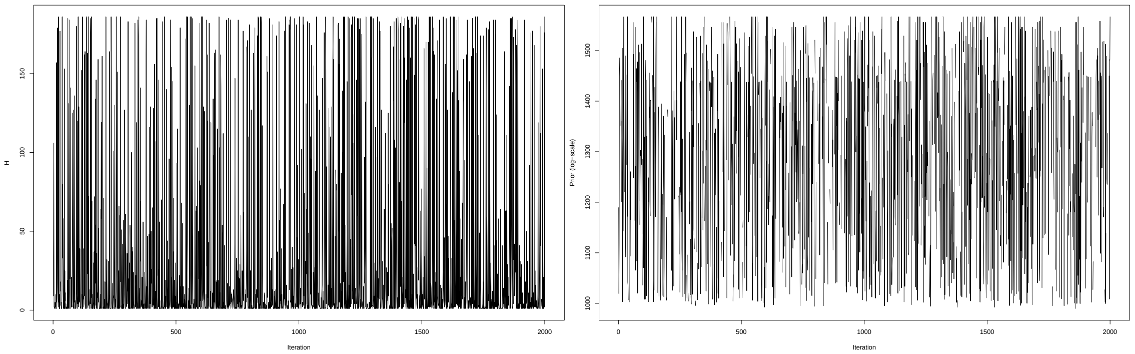
Denoting by k_{TN} the normalising constant of the truncated normal in $[0,\infty)$ with mean $\mu+\sigma^2\sum_{h=1}^{H}\log g(X_h^*)$ and variance σ^2, we get:

$$\begin{aligned}
p(\alpha) &= \sum_{\mathbf{z}\in\mathcal{Z}} p(\alpha,\mathbf{z}) \\
&= \sum_{\mathbf{z}\in\mathcal{Z}} \frac{k_{TN}}{k^*}\Big[\prod_{h=1}^{H} c(S_h)\Big]\exp\Big\{-\frac{1}{2\sigma^2}\big(\mu+\sum_{h=1}^{H}\log g(X_h^*)\big)^2\Big\}\cdot \mathrm{TN}_{[0,\infty)}\big(\mu+\sum_{h=1}^{H}\log g(X_h^*),\sigma^2\big) \\
&= \sum_{\mathbf{z}\in\mathcal{Z}} w_{\mathbf{z}}\cdot \mathrm{TN}_{[0,\infty)}\big(\mu+\sum_{h=1}^{H}\log g(X_h^*),\sigma^2\big),
\end{aligned}$$

that is a mixture of truncated normal distributions.

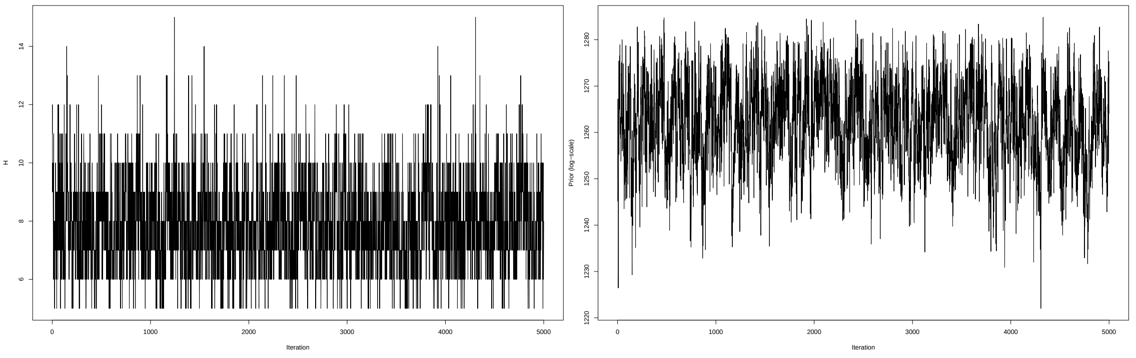
MCMC diagnostic

Prior sampling



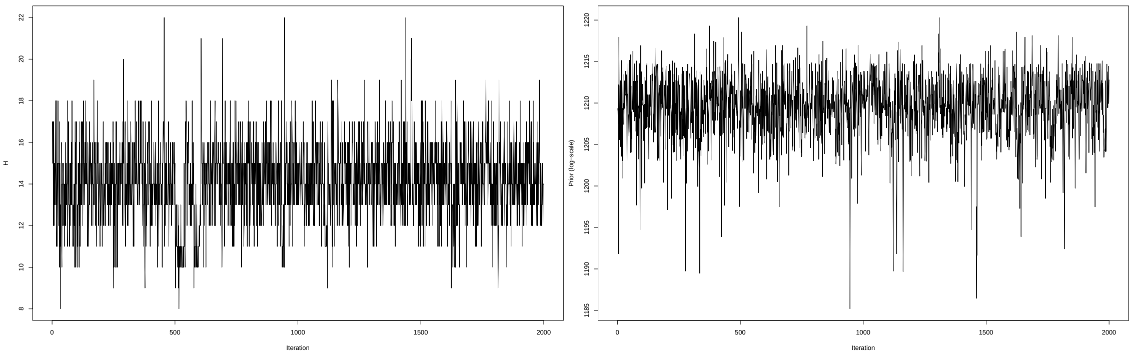
(a) (b)

Figure 19: Traceplots for the unsupervised prior sampling for: (a) the number of clusters H, (b) the log-prior $p(\mathbf{z})$.



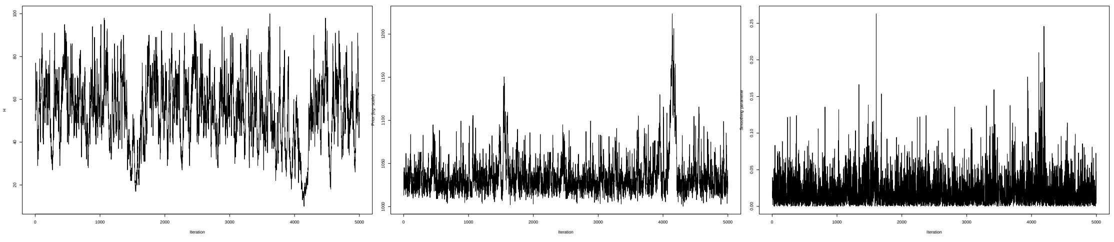
(a) (b)

Figure 20: Traceplots for prior sampling with smoothing parameter $\alpha = 5$ for: (a) the number of clusters H, (b) the log-prior $p(\mathbf{z}|X)$.



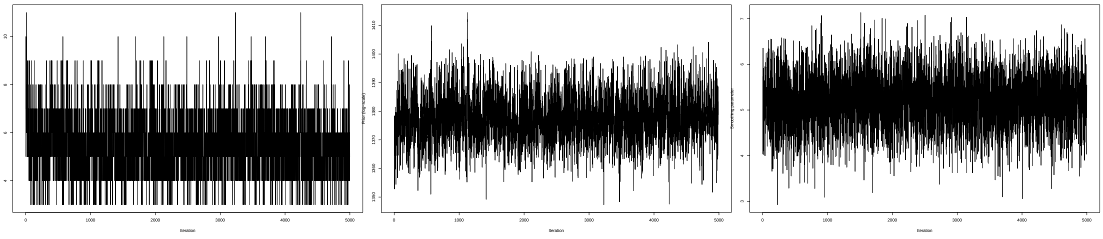
(a) (b)

Figure 21: Traceplots for prior sampling with smoothing parameter $\alpha = 500$ for: (a) the number of clusters H, (b) the log-prior $p(\mathbf{z}|X)$.



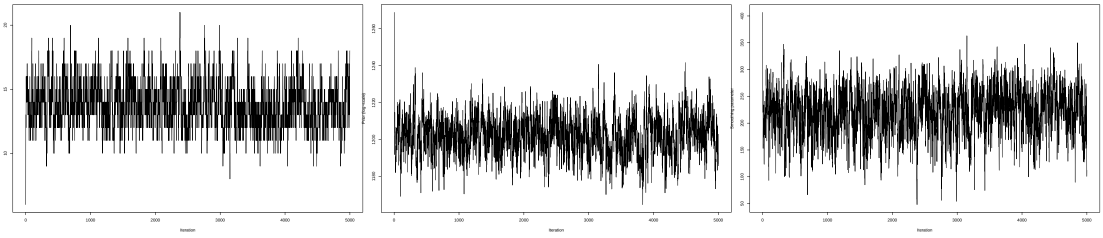
(a) (b) (c)

Figure 22: Traceplots for prior sampling with inferred smoothing parameter for $\mu = 0$ for: (a) the number of clusters H, (b) the log-prior $p(\mathbf{z}|X)$, (c) the smoothing parameter α.



(a) (b) (c)

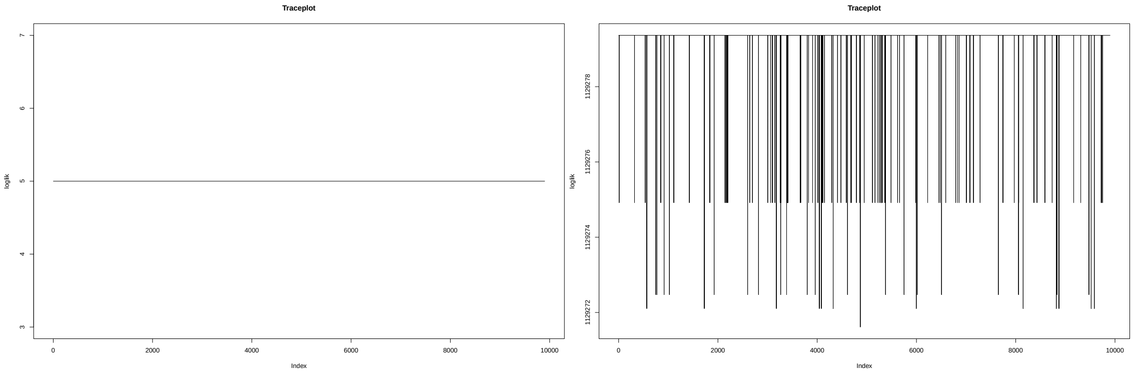
Figure 23: Traceplots for prior sampling with inferred smoothing parameter for $\mu = 5$ for: (a) the number of clusters H, (b) the log-prior $p(\mathbf{z}|X)$, (c) the smoothing parameter α.



(a) (b) (c)

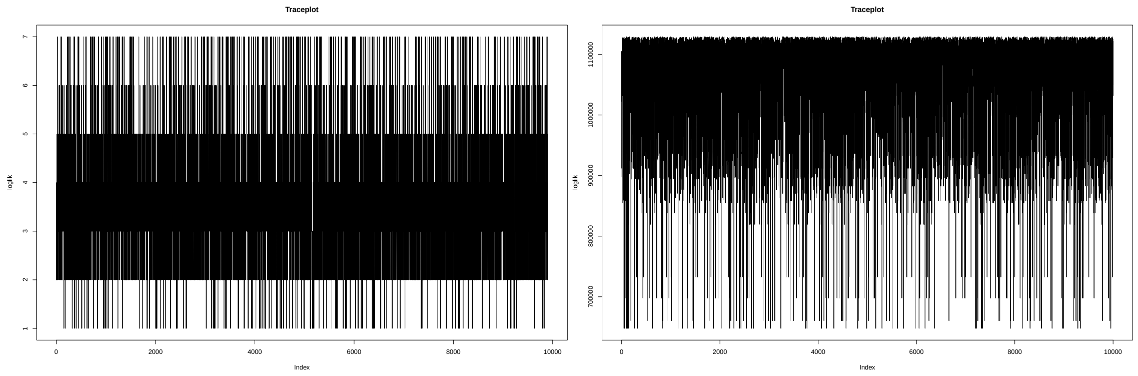
Figure 24: Traceplots for prior sampling with inferred smoothing parameter for $\mu = 500$ for: (a) the number of clusters H, (b) the log-prior $p(\mathbf{z}|X)$, (c) the smoothing parameter α.

Posterior sampling



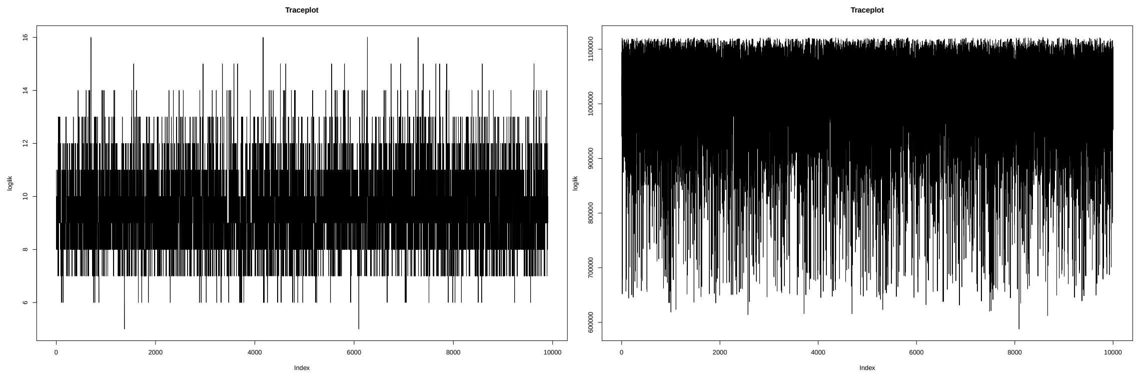
(a) (b)

Figure 25: Traceplots for unsupervised, posterior sampling for: (a) the number of clusters H, (b) the log-likelihood (unnormalised) $q(Y|\mathbf{z})$.



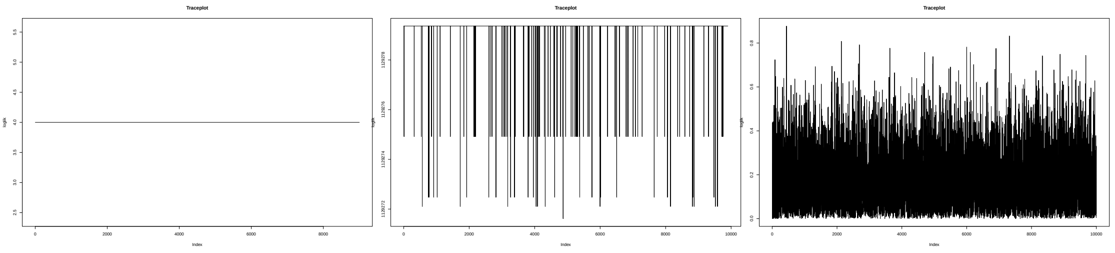
(a) (b)

Figure 26: Traceplots for supervised, posterior sampling for fixed smoothing parameter $\alpha = 5$: (a) the number of clusters H, (b) the log-likelihood (unnormalised) $q(Y|\mathbf{z})$.



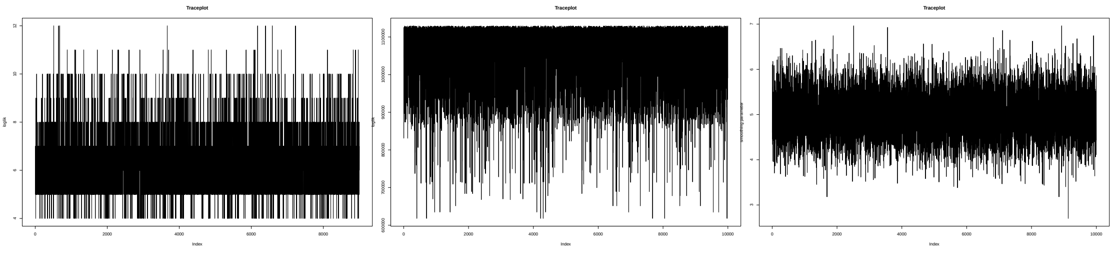
(a) (b)

Figure 27: Traceplots for supervised, posterior sampling for fixed smoothing parameter $\alpha = 500$: (a) the number of clusters H, (b) the log-likelihood (unnormalised) $q(Y|\mathbf{z})$.



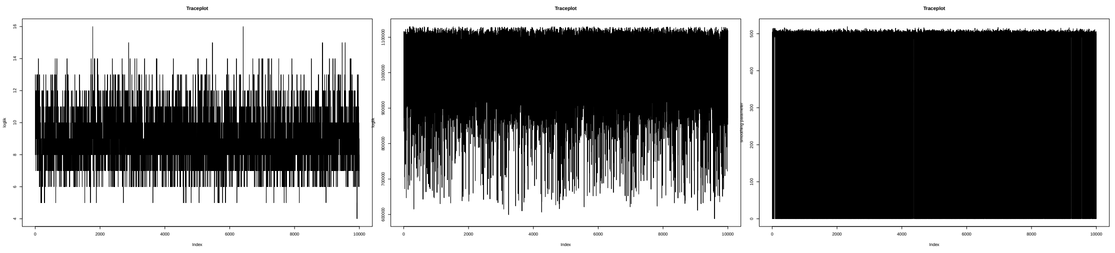
(a) (b) (c)

Figure 28: Traceplots for supervised, posterior sampling with inferred smoothing parameter for $\mu = 0, \sigma^2 = 0.1$ for: (a) the number of clusters H, (b) the log-likelihood (unnormalised) $q(Y|\mathbf{z})$, (c) the smoothing parameter α.



(a) (b) (c)

Figure 29: Traceplots for supervised, posterior sampling with inferred smoothing parameter for $\mu = 5, \sigma^2 = 0.25$ for: (a) the number of clusters H, (b) the log-likelihood (unnormalised) $q(Y|\mathbf{z})$, (c) the smoothing parameter α.

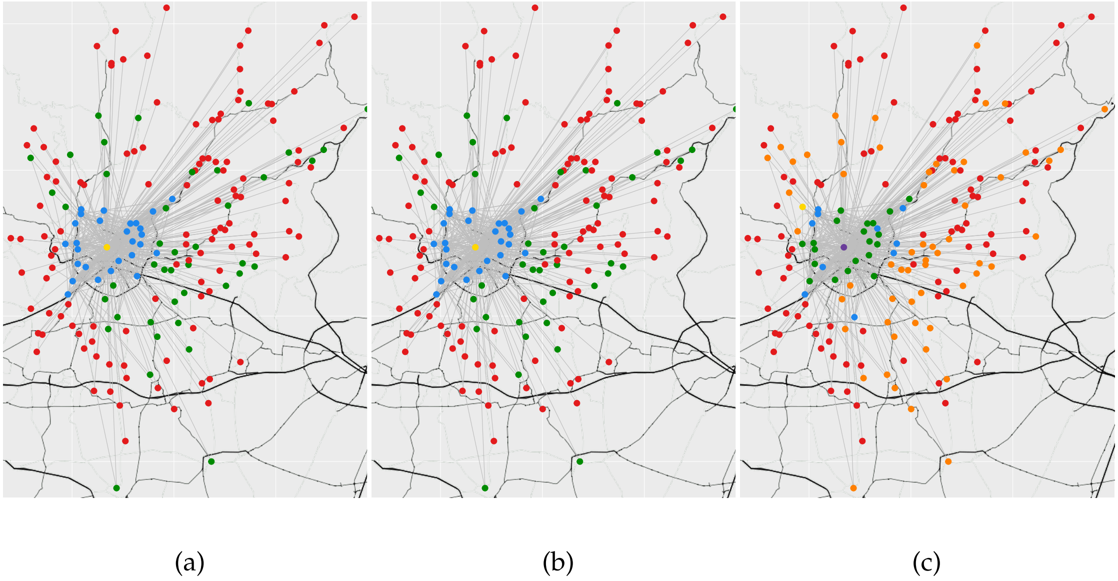


(a) (b) (c)

Figure 30: Traceplots for supervised, posterior sampling with inferred smoothing parameter for $\mu = 500, \sigma^2 = 25$ for: (a) the number of clusters H, (b) the log-likelihood (unnormalised) $q(Y|\mathbf{z})$, (c) the smoothing parameter α.

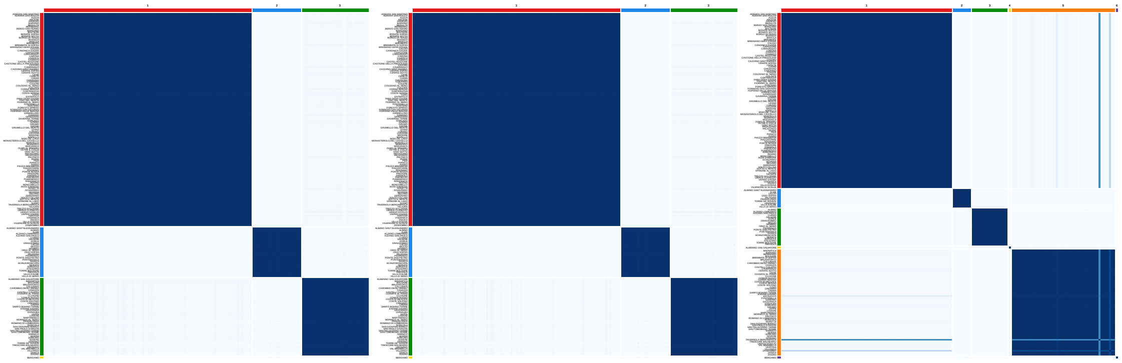
Impact of the hyperparameter σ^2 on the learned posterior

Here we report the results of the posterior inference of the model reported in Section 4.2.1. The aim of this experiment is to study the impact of the hyperparameter σ^2 on the posterior inference. Even though different experiments have been performed, we report here the study for $\mu = 5$, and $\sigma^2 = 5, 10, 25$. For sensible values of σ^2 (i.e. $\sigma^2 = 5, 10$) the inferred posterior partition does not change at all. For a high value of σ^2, the partition slightly changes, even though the macro areas (orange, red, green, yellow and blue) are preserved. However, this is an extreme situation: in general, one would not choose such an auxiliary distribution. The standard deviation is in fact equal to the variance, yielding a coefficient of variation equal to 1. Anyway, the results are reported for didactic purposes, and confirm the robustness of the model with respect to the hyperparameter σ^2.



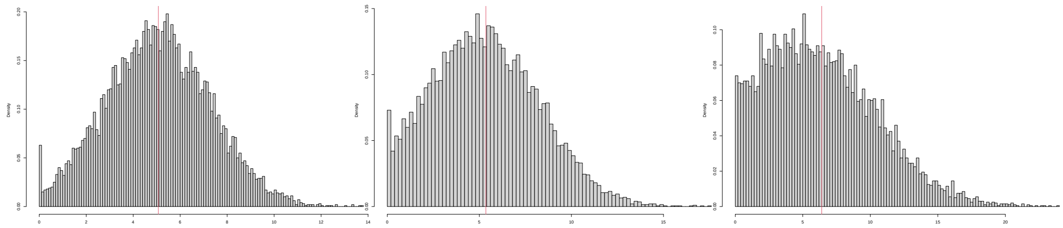
(a) (b) (c)

Figure 31: Graphical representation of the learned partitions a posteriori, keeping $\mu = 5$ fixed and varying σ^2: (a) $\sigma^2 = 5$, (b) $\sigma^2 = 10$, (c) $\sigma^2 = 25$.



(a) (b) (c)

Figure 32: Posterior coclustering matrix of the of the learned partitions a posteriori, keeping $\mu = 5$ fixed and varying σ^2: (a) $\sigma^2 = 5$, (b) $\sigma^2 = 10$, (c) $\sigma^2 = 25$.



(a) (b) (c)

Figure 33: Posterior distribution of the smoothing parameter α provided by the Poisson extended stochastic block model, keeping $\mu = 5$ fixed and varying σ^2: (a) $\sigma^2 = 5$, (b) $\sigma^2 = 10$ (c) $\sigma^2 = 25$.

Chapter 5

Discussion

The ideas and findings presented in this thesis could serve as a foundation for numerous promising directions for future research. In particular, Chapter 2 introduces a novel framework for defining explanations in a statistically coherent manner. This work could be a starting point for the definition and estimation of explanations with both theoretical guarantees and good practical heuristic results. By building upon this initial framework, future research can strive to enhance our understanding of explanations and their application in various domains, leading to more robust and reliable methods for explaining black-box models, with the possibility of performing uncertainty quantification. Furthermore, there is potential for advanced exploration of the proposed Xi method itself. While Chapter 2 presents the Xi method for classifiers, a straightforward generalization could be the definition of similar post hoc explanations for regression models. This extension would allow for a broader application of the Xi method and provide valuable insights. In addition, a significant direction to pursue is the improvement of the computational framework for the corresponding explanations. Firstly, it is crucial to evaluate and refine the proposed estimators for the probability distributions involved, as there may exist more optimal alternatives. Secondly, while the Xi method defines explanations that already consider the dependence structure of a set of random variables, estimating explanations in this case can be challenging due to the curse of dimensionality. Therefore, there is a clear necessity to explore and develop more effective methods that can incorporate the influence of dependence structures into explanations. By addressing these computational challenges, researchers can enhance the interpretability and practicality of the explanations, ultimately advancing the field of statistical explainability.

Chapter 3 introduces the multiplex extended stochastic block model (mESBM), a generalization of the stochastic block model for multiplex networks, graph-like structures representing different types of relationships among the same nodes. The mESBM goes beyond traditional models by providing both layer-specific groupings, that capture the latent partitions within each layer, and a common clustering of the nodes representing a general latent structure. A first, natural extension of the mESBM would involve adapting it to handle weighted multiplex networks, where the edge weights can be continuous or discrete. This extension can be defined by modifying the likelihood distribution and the prior on the block matrix while preserving the probabilistic structure of the partitions. Similar extensions could also be explored for bipartite or directed networks, changing the definition of the likelihood accordingly. A further direction to explore is the introduction of supervision in the partitioning process through suitable covariates. These covariates can either be node attributes or layer attributes, providing flexibility in the specification of the model. This approach enables the utilization of additional information to guide the partitioning process and enhance the quality and interpretability of the results. Regarding the application, it would be valuable to expand the analysis of the results presented in Chapter 3: namely, the relationship between the subject-specific partitions and the diagnosed mental illnesses for each patient could be furtherly unraveled. To this aim, collaborating with neuroscientists and physicians is of fundamental importance. Furthermore, applying the mESBM to other contexts may be compelling. For instance, examining multiplex transportation networks that encompass various graphs representing different means of transportation within the same cities can reveal the underlying structure, and could offer valuable insights for urban planning and transportation management. Also investigating multiplex social networks would present another relevant real-world scenario to explore: understanding the intricate relationships and dynamics across different layers of social interactions could provide interesting observations into various fields.

Chapter 4 introduces the Poisson extended stochastic block model (pESBM) for community detection in weighted networks with continuous and multidimensional node attributes. The starting point is the analysis of a real-world transportation network embedded in a geographical space, with the goal of obtaining latent radial clusters around a central hub. By focusing on the spatial arrangement of the transportation network and leveraging the formation of radial clusters, the aim is to uncover meaningful patterns and insights that can inform decision-making processes. However, it is important to acknowledge that this work is still ongoing and evolving. In particular, specific modifications need to be implemented in the inferential framework of the smoothing parameter α to mitigate the dependency of its posterior distribution on the parameters of the auxiliary distributions. This aspect becomes particularly significant when dealing with complex modeling tasks, where it is desirable to minimize the user intervention as much as possible. The goal of these necessary modifications is to enhance the autonomy and efficiency of the inference process, leading to more reliable, robust and automated results. Furthermore, it is important to consider that the proposed similarity function may not be the only or the optimal choice for the given context. Alternative approaches could be explored, including the incorporation of different geometric features of the network. For instance, leveraging polar coordinates could offer alternative means to define clusters with non-traditional shapes: incorporating the angle of each municipality with respect to the central hub may lead to angular clusters, i.e. groups of nodes that are encouraged to cluster based on their angular proximity. By exploiting additional geometrical aspects of the network, researchers can expand the possibilities for cluster formation and capture more nuanced patterns within the data.

Bibliography

C. Aicher, A. Z. Jacobs, and A. Clauset. Adapting the stochastic block model to edge-weighted networks. *arXiv preprint*, 1305.5782, 2013.

E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9(65):1981–2014, 2008.

A. A. Amini, A. Chen, P. J. Bickel, and E. Levina. Pseudo-likelihood methods for community detection in large sparse networks. *Annals of Statistics*, 41(4):2097–2122, 2013.

D. W. Apley and J. Zhu. Visualizing the effects of predictor variables in Black Box supervised learning models. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 82(4):1059–1086, 2020.

H. C. Baggio, A. Abos, B. Segura, A. Campabadal, A. Garcia-Diaz, C. Uribe, Y. Compta, M. J. Marti, F. Valldeoriola, and C. Junque. Statistical inference in brain graphs using threshold-free network-based statistics. *Human Brain Mapping*, 39(6):2289–2302, 2018.

A. L. Barabási. Network science. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1987):20120375, 2013.

R. F. Barber and E. J. Candés. Controlling the false discovery rate via knockoffs. *Annals of Statistics*, 43(5):2055–2085, 2015.

P. Barbillon, S. Donnet, E. Lazega, and A. Bar-Hen. Stochastic block models for multiplex networks: an application to networks of researchers. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180:295–314, 2015.

M. Barbosa, L. Pimentel-Silva, M. Nogueira, T. Rezende, C. Yasuda, and F. Cendes. Major depressive disorder and pharmacoresponse independently influence amygdalar t2 signal changes in temporal lobe epilepsy. *Journal of the Neurological Sciences*, 429:117871, 2021.

D. Barry and J. A. Hartigan. Product partition models for change point problems. *The Annals of Statistics*, 1(1):260–279, 1992.

D. Bassett and E. Bullmore. Small-world brain networks. *The Neuroscientist : a review journal bringing neurobiology, neurology and psychiatry*, 12(1):512–523, 2007.

D. S. Bassett and E. T. Bullmore. Small-world brain networks revisited. *The Neuroscientist*, 23(5):499–516, 2017.

M. Bellani, M. Baiano, and P. Brambilla. Brain anatomy of major depression I. Focus on hippocampus. *Epidemiology and Psychiatric Sciences*, 19(4):298–301, 2010.

A. Binder, S. Bach, G. Montavon, K. R. Müller, and W. Samek. Layer-wise relevance propagation for deep neural network architectures. *Lecture Notes in Electrical Engineering*, 376: 913–922, 2016.

N. Binkiewicz, J. T. Vogelstein, and K. Rohe. Covariate-assisted spectral clustering. *Biometrika*, 104(2):361–377, 2017.

V. D. Blondel, J. L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 10:1–12, 2008.

E. Borgonovo, S. Tarantola, E. Plischke, and M. D. Morris. Transformations and invariance in the sensitivity analysis of computer experiments. *Journal of the Royal Statistical Society, Series B*, 76(5):925–947, 2014.

E. Borgonovo, G. Hazen, and E. Plischke. A common rationale for global sensitivity measures and their estimation. *Risk Analysis*, 36(10):1871–1895, 2016.

L. Breiman. *Classification and regression trees*. Chapman&Hall, New York, 1984.

L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

E. Bullmore and O. Sporns. Complex brain networks: Graph theoretical analysis of structural and functional systems. *Nature reviews. Neuroscience*, 10:186–198, 2009.

E. Bullmore and O. Sporns. The economy of brain network organization. *Nature reviews, Neuroscience*, 13:336–349, 2012.

S. C. Caetano, M. Fonseca, J. P. Hatch, R. L. Olvera, M. Nicoletti, K. Hunterm, B. Lafer, S. R. Pliszka, and J. C. Soares. Medial temporal lobe abnormalities in pediatric unipolar depression. *Neuroscience Letters*, 427:142–147, 2007.

A. Caimo and N. Friel. Bayesian inference for exponential random graph models. *Social Networks*, 33(1):41–55, 2011.

E. Candès, Y. Fan, L. Janson, and J. Lv. Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 80(3):551–577, 2018.

K. Chan, A. Saltelli, and S. Tarantola. Winding stairs: a sampling tool to compute sensitivity indices. *Statistics and Computing*, 10(3):187–196, 2000.

S. Chatterjee. A new coefficient of correlation. *Journal of the American Statistical Association*, 116(536):2009–2022, 2021.

A. Chaudhuri and W. Hu. A fast algorithm for computing distance correlation. *Computational Statistics and Data Analysis*, 135:15–24, 2019.

C. Chen, O. Li, C. Tao, A. J. Barnett, J. Su, and C. Rudin. This looks like that: Deep learning for interpretable image recognition. *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 801:12–37, 2019.

Z. Chen, Y. Bei, and C. Rudin. Concept whitening for interpretable image recognition. *Nature Machine Intelligence*, 2(12):772–782, 2020.

N. Chopin and O. Papaspiliopoulos. *An Introduction to Sequential Monte Carlo*. Springer, 2020.

P. Chunaev. Community detection in node-attributed social networks: a survey. *Computer Science Review*, 37:1002–1086, 2020.

A. Clauset, M. E. Newman, and C. Moore. Finding community structure in very large networks. *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, 70(6):6–27, 2004.

D. Combe, C. Largeron, M. Géry, and E. Egyed-Zsigmond. I-Louvain: an attributed graph clustering method. *Advances in Intelligent Data Analysis XIV*, pages 181–192, 2015.

E. Côme, N. Jouvin, P. Latouche, and C. Bouveyron. Hierarchical clustering with discrete latent variable models and the integrated classification likelihood. *Advances in Data Analysis and Classification*, 15(4):957–986, 2021.

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

J. J. Crofts, M. Forrester, and R. D. O'Dea. Structure-function clustering in multiplex brain networks. *Europhysics Letters*, 116(1):180–203, 2016.

S. Da Veiga. Global sensitivity analysis with dependence measures. *Journal of Statistical Computation and Simulation*, 85(7):1283–1305, 2015.

D. B. Dahl. Distance-based probability distribution for set partitions with applications to bayesian nonparametrics. *JSM Proceedings. Section on Bayesian Statistical Science, American Statistical Association*, 2008.

P. De Blasi, S. Favaro, A. Lijoi, R. H. Mena, I. Prunster, and M. Ruggiero. Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):212–229, 2015.

B. De Finetti. On the condition of partial exchangeability. *Studies in inductive logic and probability*, 2:193–205, 1980.

D. R. DeFord and S. D. Pauls. Spectral clustering methods for multiplex networks. *Physica A: Statistical Mechanics and its Applications*, 533:1219–1249, 2019.

U. Demir, M. A. Gharsallaoui, and I. Rekik. Clustering-based deep brain multigraph integrator network for learning connectional brain templates. *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis*, pages 109–120, 2020.

P. Diaconis. Recent progress on de Finetti's notions of exchangeability. *Bayesian Statistics*, 3:111–125, 1988.

J. Dong and C. Rudin. Exploring the cloud of variable importance for the set of all good models. *Nature Machine Intelligence*, 2:810–824, 2020.

D. Dua and C. Graff. UCI Machine Learning Repository, 2017. URL `http://archive.ics.uci.edu/ml`.

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

B. Efron and G. Gong. A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, 37(1):36–48, 1983.

L. Egidi, F. Pauli, N. Torelli, and Z. Susanna. Clustering spatial networks through latent mixture models. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 186(1): 137–156, 2023.

S. Eickhoff, B. T. Yeo, and S. Genon. Imaging-based parcellations of the human brain. *Nature Reviews Neuroscience*, 19:672–686, 2018.

J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 70(5):849–911, 2008.

J. Fan and R. Song. Sure independence screening in generalized linear models with NP-dimensionality. *Annals of Statistics*, 38(6):3567–3604, 2010.

J. Faskowitz, X. Yan, X.-N. Zuo, and O. Sporns. Weighted stochastic block models of the human connectome across the life span. *Scientific Reports*, 8:12997, 2018.

A. Fisher, C. Rudin, and F. Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20:1–81, 2019.

S. Fortunato and D. Hric. Community detection in networks: a user guide. *Physics Reports*, 659:1–44, 2016.

S. Fortunato and M. E. J. Newman. 20 years of network community detection. *Nature Physics*, 18(8):848–850, 2022.

T. Franklin, P. Acton, J. Maldjian, J. Gray, J. Croft, C. Dackis, C. O'Brien, and A. Childress. Decreased gray matter concentration in the insular, orbitofrontal, cingulate, and temporal cortices of cocaine patients. *Biological psychiatry*, 51:134–142, 2002.

F. Gamboa, A. Janon, T. Klein, A. Lagnoux, and C. Prieur. Statistical inference for Sobol pick-freeze Monte Carlo method. *Statistics*, 50(4):881–902, 2016.

F. Gamboa, T. Klein, and A. Lagnoux. Sensitivity analysis based on Cramér von Mises distance. *SIAM/ASA J. Uncertainty Quantification*, 6(2):522–548, 2018.

F. Gamboa, P. Gremaud, T. Klein, and A. Lagnoux. Global sensitivity analysis: A novel generation of mighty estimators based on rank statistics. *Bernoulli*, 28(4):2345–2374, 2022.

C. Garcia. Depression in temporal lobe epilepsy: A review of prevalence, clinical features, and management considerations. *Epilepsy research and treatment*, 2012:809–843, 2012.

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. edition, 2004.

V. Ghidini, S. Legramanti, and R. Argiento. Extended stochastic block model with spatial covariates for weighted brain networks. *Bayesian Statistics, New Generations New Approaches (BAYSM2022)*, 2023a. To appear.

V. Ghidini, S. Legramanti, and R. Argiento. Binomial extended stochastic block model for brain networks. *Book of short papers SIS 2023*, 2023b. To appear.

N. Glick. Measurements of separation among probability densities or random variables. *Canadian Journal of Statistics*, 3(2):267–276, 1975.

A. Gnedin. Species sampling model with finitely many types. *Electronic Communications in Probability*, 15:79–88, 2010.

A. Gnedin and J. Pitman. Exchangeable Gibbs partitions and Stirling triangles. *Zapiski Nauchnykh Seminarov*, 325:83–102, 2004.

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

I. J. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, Cambridge, MA, USA, 2016.

B. Goodman and S. Flaxman. European union regulations on algorithmic decision making and a "right to explanation". *AI Magazine*, 38(3):50–57, 2017.

R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, D. Pedreschi, and F. Giannotti. A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5):1–42, 2018.

J. Hartigan. Partition models. *Communications in Statistics - Theory and Methods*, (19):2745–2756, 1990.

T. J. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning*. Springer-Verlag, second edition, 2009.

K. Helm, K. Viol, T. Weiger, P. Tass, C. Grefkes, D. Del Monte, and G. Schiepek. Neuronal connectivity in major depressive disorder: A systematic review. *Neuropsychiatric Disease and Treatment*, 14:2715–2737, 2018.

T. Herlau, M. N. Schmidt, and M. Mørup. Infinite-degree-corrected stochastic block model. *Physical Review E*, 90(3):032819, 2014.

H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3):321–377, 1936.

L. F. Huang. Artificial intelligence. *2010 The 2nd International Conference on Computer and Automation Engineering, ICCAE 2010*, 4:575–578, 2010.

B. Karrer and M. E. Newman. Stochastic blockmodels and community structure in networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 83(1):1–11, 2011.

I. B. Kim and S. C. Park. The entorhinal cortex and adult neurogenesis in major depression. *International Journal of Molecular Sciences*, 22:117–125, 2021.

M. Kivelä, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter. Multilayer networks. *Journal of complex networks*, 2(3):203–271, 2014.

N. H. Kuiper. Tests concerning random points on a circle. *Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen, Series A*, 63:38–47, 1960.

D. B. Larremore, A. Clauset, and A. Z. Jacobs. Efficiently inferring community structure in bipartite networks. *Physical Review E*, 90(1):012805, 2014.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

S. Legramanti, T. Rigon, D. Durante, and D. B. Dunson. Extended stochastic block models with application to criminal networks. *The Annals of Applied Statistics*, 16(4):2369 – 2395, 2022.

H. Lemke, S. Probst, A. Warneke, L. Waltemate, A. Winter, K. Thiel, S. Meinert, V. Enneking, F. Breuer, M. Klug, J. Goltermann, C. Hülsmann, D. Grotegerd, R. Redlich, K. Dohm, E. Leehr, J. Repple, N. Opel, K. Brosch, T. Meller, J. Pfarr, K. Ringwald, S. Schmitt, F. Stein, A. Krug, A. Jansen, I. Nenadic, T. Kircher, T. Hahn, and U. Dannlowski. The course of disease in major depressive disorder is associated with altered activity of the limbic system during negative emotion processing. *Biological Psychiatry Cognitive Neuroscience Neuroimaging*, 3(7):323–332, 2022.

S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 2017:4766–4775, 2017.

S. M. Lundberg, G. G. Erion, and S.-I. Lee. Consistent individualized feature attribution for tree ensembles. *ICML Workshop*, 1802:1–9, 2019.

M. Meilă. Comparing clusterings—an information based distance. *Journal of multivariate analysis*, 98(5):873–895, 2007.

T. Miller. Explanation in artificial intelligence: insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

C. Molnar. *Interpretable Machine Learning - a guide for making Black Box models explainable*. 2018.

C. Mu, A. Mele, L. Hao, J. Cape, A. Athreya, and C. E. Priebe. On spectral algorithms for community detection in stochastic blockmodel graphs with vertex covariates. *IEEE Transactions on Network Science and Engineering*, 9(5):3373–3384, 2022.

P. Muller, F. Quintana, and G. L. Rosner. A product partition model with regression on covariates. *Journal of Computational and Graphical Statistics*, 20(1):260–278, 2011.

W. J. Murdoch, C. Signh, K. Kumbier, R. Abbasi-Asl, and B. Yu. Definitions, methods and applications in interpretabile Machine Learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080, 2019.

I. Murray, Z. Ghahramani, and D. J. C. MacKay. MCMC for doubly-intractable distributions. *arXiv preprint*, 1206.6848, 2006.

M. E. Newman. Analysis of weighted networks. *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, 70(5):056131, 2004.

T. L. J. Ng and T. B. Murphy. Weighted stochastic block model. *Statistical Methods & Applications*, 30(5):1365–1398, 2021.

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

D. M. O'Shea, V. M. Dotson, A. J. Woods, E. C. Porges, J. B. Williamson, A. O'Shea, and R. Cohen. Depressive symptom dimensions and their association with hippocampal and entorhinal cortex volumes in community dwelling older adults. *Frontiers in Aging Neuroscience*, 10, 2018.

S. Paganin, A. H. Herring, A. F. Olshan, and D. B. Dunson. Centered partition processes: Informative priors for clustering (with discussion). *Bayesian Analysis*, 16(1), 2021.

G. Page and F. Quintana. Spatial product partition models. *Bayesian Analysis*, 4:265–298, 2015.

G. Page and F. Quintana. Calibrating covariate informed product partition models. *Statistics and Computing*, 28:1009–1031, 2018.

W. Pan, X. Wang, H. Zhang, H. Zhu, and J. Zhu. Ball covariance: a generic measure of dependence in Banach space. *Journal of the American Statistical Association*, 115(529):307–317, 2020.

J. H. Park and D. Dunson. Bayesian generalized product partition model. *Statistica Sinica*, 20(3):1203–1226, 2010.

Y. L. Pavlov. *Random forests*. VSP, 2019.

K. Pearson. Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242, 1895.

K. Pearson. On the general theory of skew correlation and non-linear regression. *Mathematical Contributions to the Theory of Evolution, Drapers' Company Research Memoirs*, 14, 1905.

T. P. Peixoto. Nonparametric weighted stochastic block models. *Physical Review E*, 97(1): 1–20, 2018.

L. Peng and L. Carvalho. Bayesian degree-corrected stochastic blockmodels for community detection. *arXiv preprint*, 1309.4796, 2013.

V. Petsiuk, A. Das, and K. Saenko. RisE: Randomized input sampling for explanation of black-box models. In *British Machine Vision Conference 2018, BMVC 2018*, 2019.

E. Plischke and E. Borgonovo. What about totals? alternative approaches to factor fixing. *Safety, Reliability and Risk Analysis: Beyond the Horizon - Proceedings of the European Safety and Reliability Conference, ESREL 2013*, pages 3339–3344, 2014.

M. Ramezani, I. Johnsrude, A. Rasoulian, R. Bosma, R. Tong, T. Hollenstein, K. Harkness, and P. Abolmaesumi. Temporal-lobe morphology differs between healthy adolescents and those with early-onset of depression. *NeuroImage: Clinical*, 6:145–155, 2014.

A. Renyi. On measures of statistical dependence. *Acta Mathematica Academiae Scientiarum Hungarica*, 10:441–451, 1959.

M. T. Ribeiro, S. Singh, and C. Guestrin. "Why should i trust you?" - explaining the predictions of any classifier. *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

M. Robnik-Šikonja and M. Bohanec. Perturbation-based explanations of prediction models. *Human and Machine Learning*, pages 159–175, 2018.

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

C. Rudin. Stop explaining black-box Machine Learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

P. Sah and E. Fokoué. What do Asian religions have in common? An unsupervised text analytics exploration. *arXiv preprint*, 1912.10847, 2019.

A. Saltelli. Making best use of model valuations to compute sensitivity indices. *Computer Physics Communications*, 145:280–297, 2002.

A. Saltelli. *Global sensitivity analysis: the primer*. 2008.

M. N. Schmidt and M. Morup. Nonparametric Bayesian modeling of complex networks: An introduction. *IEEE Signal Processing Magazine*, 30(3):110–128, 2013.

R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-CAM: visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, 2020.

S. Simpson, F. Bowman, and P. Laurienti. Analyzing complex functional brain networks: dusing statistics and network science to understand the brain. *Statistics Surveys*, 2:234–257, 2013.

S. M. Smith, K. L. Miller, G. S. Khorshidi, M. A. Webster, C. F. Beckmann, T. E. Nichols, J. Ramsey, and M. W. Woolrich. Network modelling methods for fmri. *NeuroImage*, 54: 875–891, 2011.

E. S. Soofi. Capturing the intangible concept of information. *Journal of the Americal Statistical Association*, 89(428):1243–1254, 1994.

C. Spearman. The proof and measurement of the association between two things. *American Journal of Psychology*, 15:72–101, 1904.

O. Sporns. *Networks of the Brain*. MIT Press, 2010.

O. Sporns. Structure and function of complex brain networks. *Dialogues in clinical neuroscience*, 15:247–262, 2013.

J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. *3rd International Conference on Learning Representations, ICLR 2015 - Workshop Track Proceedings*, pages 1–14, 2015.

M. Strong and J. E. Oakley. An efficient method for computing partial expected value of perfect information for correlated inputs. *Medical Decision-Making*, 33:755–766, 2013.

M. Strong, J. E. Oakley, and J. Chilcott. Managing structural uncertainty in health economic decision models: a discrepancy approach. *Journal of the Royal Statistical Society, Series C*, 61(1):25–45, 2012.

G. J. Székely and M. L. Rizzo. Brownian distance covariance. *Annals of Applied Statistics*, 3: 1236–1265, 2009.

G. J. Székely, M. L. Rizzo, and N. K. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, 2007.

S. Taverniers, E. J. Hall, M. A. Katsoulakis, and D. M. Tartakovsky. Mutual information for explainable deep learning of multiscale systems. *Journal of Computational Physics*, 444: 1105–1151, 2021.

B. Thomas Yeo, M. R. Sabuncu, R. Desikan, B. Fischl, and P. Golland. Effects of registration regularization and atlas sharpness on segmentation accuracy. *Medical Image Analysis*, 12 (5):603–615, 2008.

G. Tononi, A. R. McIntosh, D. Russell, and G. M. Edelman. Functional clustering: Identifying strongly interactive brain regions in neuroimaging data. *NeuroImage*, 7(2):133–149, 1998.

H. F. Unterrainer, M. Hiebler-Ragger, K. Koschutnig, J. Fuchshuber, K. Ragger, C. M. Perchtold, I. Papousek, E. M. Weiss, and A. Fink. Brain structure alterations in poly-drug use: Reduced cortical thickness and white matter impairments in regions associated with affective, cognitive, and motor functions. *Frontiers in Psychiatry*, 10, 2019.

T. Vallè s-Català, F. A. Massucci, R. Guimerà, and M. Sales-Pardo. Multilayer stochastic block models reveal the multilayer structure of complex networks. *Physical Review X*, 6 (1):11–36, 2016.

M. P. van den Heuvel and O. Sporns. Network hubs in the human brain. *Trends in Cognitive Sciences*, 17(12):683–696, 2013.

U. Von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

S. Wachter, B. Mittelstadt, and C. Russell. Counterfactual explanations without opening the black box: automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2):842–887, 2018.

S. Wade and Z. Ghahramani. Bayesian cluster analysis: point estimation and credible balls (with discussion). *Bayesian Analysis*, 13(2):559 – 626, 2018.

H. Xiao, K. Rasul, and R. Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint*, 1708.07747, 2017.

M. Xu, V. Jog, and P. L. Loh. Optimal rates for community estimation in the weighted stochastic block model. *Annals of Statistics*, 48(1):183–204, 2020.

J. Yang, J. McAuley, and J. Leskovec. Community detection in networks with node attributes. In *2013 IEEE 13th International Conference on Data Mining*, 2013.

M. Yuvaraj, A. Dey, V. Lyubchich, Y. Gel, and H. V. Poor. Topological clustering of multilayer networks. *Proceedings of the National Academy of Sciences*, 118:112–193, 2021.

V. Zamoscik, S. Huffziger, U. Ebner-Priemer, C. Kuehner, and P. Kirsch. Increased involvement of the parahippocampal gyri in a sad mood predicts future depressive symptoms. *Social cognitive and affective neuroscience*, 9(12):2034–2040, 2014.

L. Zemanová, C. Zhou, and J. Kurths. Structural and functional clusters of complex brain networks. *Physica D: Nonlinear Phenomena*, 224(1):202–212, 2006.

L.-L. Zeng, H. Shen, L. Liu, L. Wang, B. Li, P. Fang, Z. Zhou, Y. Li, and D. Hu. Identifying major depression using whole-brain functional connectivity: a multivariate pattern analysis. *Brain*, 135(5):1498–1507, 2012.

F.-F. Zhang, W. Peng, J. A. Sweeney, Z.-Y. Jia, and Q.-Y. Gong. Brain structure alterations in depression: Psychoradiological evidence. *CNS Neuroscience & Therapeutics*, 24(11):994–1003, 2018.

Y. Zhang, E. Levina, and J. Zhu. Community detection in networks with node features. *Electronic Journal of Statistics*, 10(2):3153–3178, 2016.

Q. Zhao and T. Hastie. Causal interpretations of black box models. *Journal of Business & Economic Statistics*, 0(0):1–10, 2019.