Università Commerciale "Luigi Bocconi"
PhD School

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

Advances in Bayesian modelling of array structured data

Advisor: Daniele Durante

PhD Thesis by
Federico Pavone
ID number: 3111700

Year: 2024

Abstract

Data organized in array structures arise in various domains. Each entry of the array serves as a statistical unit, while the dimensions correspond to indexing attributes. The inherent dependence among statistical units along the indexing attributes makes the array representation more suitable than the usual tabular format. Models for this type of data typically employ probabilistic low–rank factorizations, where the latent factors attempt to capture patterns within the indexing attributes responsible for the values of the outcome. It is of primary importance to correctly model the dependence within the latent factors eliciting structural information available from data. Our contribution consists of novel structured Bayesian factorization models for array data, with applications to mortality forecasts and network analysis.

We first address the problem of accurately forecasting future death–rate patterns for different age groups and time horizons for a country of interest. This type of data exhibits smooth structures of different natures across ages and years, which we flexibly account for in our model. We propose a novel B–spline process with locally–adaptive dynamic coefficients that outperforms state–of–the–art forecasting strategies by explicitly incorporating the core structures of period mortality trajectories within an interpretable formulation.

Next, we consider the problem of learning the underlying structure responsible for the connectivity patterns in the human brain. We analyze a population of networks representing the connections between brain regions for a set of subjects. These networks are characterized by a hierarchical or multiresolution organization of the nodes responsible for the connectivity. We propose a phylogenetic latent position model that effectively learns the multiresolution structure. The model reveals a tree organization of the brain regions coherent with known hemisphere and lobe partitions. Such a result uncovers interesting new possible clusterings of the brain regions at different levels of resolution. Finally, we explore the potential to incorporate additional covariates to inform the tree structure of the model responsible for the latent positions.

We have considered two settings of array data that exhibit distinct structural properties. Through Bayesian modelling, we have been able to leverage this information in the form of prior specification. Our results highlight the importance of incorporating these structures appropriately, leading to improved outcomes in both inferential and forecasting problems.

Acknowledgements

I would like to express my gratitude to my supervisor, Daniele Durante. His expertise and encouragement have been instrumental in shaping this work. Furthermore, I am grateful to him for introducing me to various research opportunities and creating a joyful and collaborative environment.

I would like to extend my heartfelt appreciation to all my colleagues of the doctorate, great companions of this journey. A special thanks goes to Filippo, Valentina, and Veronica, with whom I have shared all the difficulties and the joys from the first day.

Thanks to my BJJ fellows, for all the incredible early mornings we shared. You have taught me that, with the right pressure, you can overcome any difficulty.

Special thanks to Masha, who is a constant source of encouragement and inspiration.

To my family, I owe a debt of gratitude for their support throughout these years. I am particularly grateful to my brother, an exceptional guide, who has been always the first one to pave the way through new life decisions.

Contents

Introduction

Technological advancements have facilitated the collection and storage of vast quantities of data, resulting in the emergence of big data. With *big data*, it is not only the sample size to be large, but also the number of features. Using common statistical language, in big data both n and p are large, with any of the two dominating the other. The *big* attribute introduces unique challenges when it comes to statistical modeling of such data. For example, a key characteristic of high–dimensional spaces is that the distances between data points increase rapidly as the dimensionality of the space expands. This phenomenon gives rise to a range of adverse implications, collectively referred to as the *curse of dimensionality*, which are not encountered in low–dimensional spaces. Moreover, big data present additional challenges stemming from the complexity of the data sources. As George Box famously stated in 1976, *"all models are wrong, but some are useful"* (Box, 1976). This aphorism serves as a reminder that the natural world is invariably more intricate than the mathematical descriptions we incorporate into statistical models. When constructing a model, we must select which aspects of the data complexity to disregard, while striving to provide a *useful* approximation of reality. The large sample size and dimensionality of big data magnify the significance of non-negligible internal dependence structures that necessitate careful consideration in statistical modeling.

This thesis presents methodological advancements in the field of statistical modeling of multidimensional array data. This particular type of data is characterized by an array structure, where each entry within the array serves as a statistical unit, while the dimensions of the array correspond to indexing attributes. Although it is possible to organize such data in traditional tabular format, with rows representing statistical units and columns representing covariates, the inherent dependence among statistical units along the indexing attributes makes the array representation more suitable.

Array structured data can be found in various domains, such as factorial designs and contingency tables. In demography, life tables are employed to collect life indicators for a population, stratified by age and observed over time. Image data is represented by arrays, with two dimensions capturing the pixel position and potentially additional dimensions accounting for colour channels. In the context of network analysis, the adjacency matrix

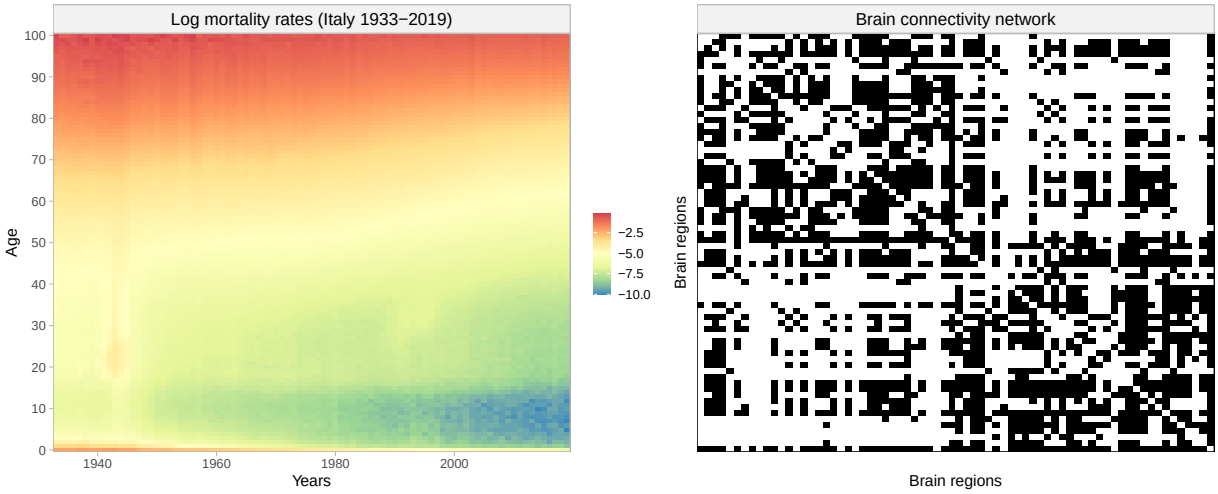


Figure 1: Examples of array data from the applications of the thesis. On the left, the age–specific log mortality rates of men in Italy from 1933 to 2019. On the right, one example of the connectivity network between brain regions.

is a two–dimensional array, encoding the edges of the network.

Figure 1 shows two examples of array data, from the applications discussed in the thesis. On the left, data are the age–specific log mortality rates for men in Italy from 1933 to 2020. Each entry is the log mortality rate for people in the population of a given age and during a given year. Both the time and age components create dependence in the mortality rates, as visible in the smooth patterns in the picture. On the right of Figure 1, it is shown the adjacency matrix representing the non–directional connectivity network of the brain of a subject. The nodes of the network represent the brain regions, which index both axes of the array. The entries represent presence or absence of a connection, respectively in black and white. Connections on the same row or column are dependent since they relate to the same brain region.

A successful model has to carefully balance the trade–off between a low–dimensional stochastic representation of the data generation mechanism, and proper modelling the dependence structures, in order to achieve good inference and out–of–sample predictive ability. Models for array data are typically based on probabilistic low–rank factorizations, where factors are possibly combined in a non–linear way. Considering the simplified case of a $n_1 \times n_2$ array $\mathbf{Y}$ and two factor matrices $\mathbf{U} \in \mathbb{R}^{K_1 \times n_1}$ and $\mathbf{V} \in \mathbb{R}^{K_2 \times n_2}$, we model the array entries as follows, for $i = 1, \dots, n_1$ and $j = 1, \dots, n_2$:

$$y_{ij} \overset{\text{ind}}{\sim} F(k(\mathbf{u}_i, \mathbf{v}_j), \xi_{ij}), \tag{1}$$

where ξ_{ij} collects extra terms, such as covariates, $k(\cdot, \cdot)$ is a possible non–linear function

combining the latent factors, $\mathbf{u}_i$ and $\mathbf{v}_j$ are the vectors corresponding to the i-th and j-th columns of $\mathbf{U}$ and $\mathbf{V}$, and F is the distribution for the observed y_{ij}. For instance, model (1) with $K_1 = K_2$ and a bilinear kernel $k(\mathbf{u}_i, \mathbf{v}_j) = \mathbf{u}_i^\intercal \mathbf{v}_j$ encompasses a series of probabilistic factorization models (Tipping and Bishop, 1999; Mnih and Salakhutdinov, 2007; Salakhutdinov and Mnih, 2008).

Occasionally, available data include additional covariates that can aid in predicting the response. However, often, the only data at hand is the array itself. In these cases, it becomes even more critical to include the correct structural information to properly model the response through the latent factors. Our contribution consists of novel structured Bayesian factorization models for array data, with applications to mortality forecast and network analysis. The Bayesian approach provides a natural way to flexibly incorporate structural information in the form of prior specification.

We now provide a brief overview of the current methodologies employed in mortality forecasting and network analysis, highlights their limitations in the context of our applications. Finally, we conclude summarizing the key contributions of the thesis.

Mortality forecasting

Several statistical models for mortality data are based on the structure of (1). The central mortality rates of a country are obtained computing the ratio between the age–specific period death counts and the average number of individuals at risk at the central time of the period, i.e. half of the year for yearly data. Let $\mathbf{Y} \in \mathbb{R}^{X \times T}$ be the array of age–specific log mortality rates for $n_1 = X$ ages and over $n_2 = T$ years.

The Lee–Carter model (Lee and Carter, 1992), a seminal work for statistical modelling of mortality data, factorizes the array of mortality rates with a bilinear product of age–specific $\mathbf{U} \in \mathbb{R}^{2 \times X}$ and period–specific factors $\mathbf{V} \in \mathbb{R}^{2 \times T}$, as follows

$$y_{ij} = \mathbf{u}_i^\intercal \mathbf{v}_j + \epsilon_{ij}, \tag{2}$$

where $\mathbf{u}_i = (u_{1i}, u_{2i})^\intercal$ and $\mathbf{v}_j = (1, v_{2j})^\intercal$, and ϵ_{ij} is a Gaussian noise error term. Under suitable constraints, u_{1i} captures the average mortality at age i, over all years. The core of the model is given by the choice of the bilinear kernel $k(\mathbf{u}_i, \mathbf{v}_j) = \mathbf{u}_i^\intercal \mathbf{v}_j$, which combines the contribution of the two latent factors. Similarly, Renshaw and Haberman (2006) extend the factorization structure of the original formulation, including also a cohort regression term $\xi_{ij} = \xi_{j-i}$. Forecasts are obtained by projecting the time–series of the estimated period–specific factors, under certain model assumptions.

Both models suffer from an over–simplistic dependence structure. Specifically, the latent factors $\mathbf{U}$ and $\mathbf{V}$ do not encode anyhow the known smooth variations of the mortality

rates both across ages and years. This leads to undesirable wiggly estimates of $\mathbf{u}_i$ and $\mathbf{v}_j$, once models are fitted to data. Under suitable choices, the forecast strategy introduces temporal–smoothness in the predictions. However, this mitigates only partially the problem of lack of proper smoothness in the mortality rates, leaving space for improvements in the accuracy of the forecasts.

Subsequent proposals attempt at including more structure. For instance, Cairns et al. (2009) impose a smoother dependence across ages via a linear basis expansion, as follows:

$$y_{ij} = v_{1j} + (i - \overline{X})v_{2j} + \epsilon_{ij}, \tag{3}$$

corresponding to a bilinear combination of the following factors:

$$\mathbf{u}_i = \begin{pmatrix} 1 \\ i - \overline{X} \end{pmatrix} \quad \text{and} \quad \mathbf{v}_j = \begin{pmatrix} v_{1j} \\ v_{2j} \end{pmatrix}, \tag{4}$$

where $\overline{X}$ is the average of the considered ages. The age–specific factor $\mathbf{u}_i$ are deterministic and induce linear smoothness across ages through the second component, while the period–specific factors $\mathbf{v}_j$ have no constraints. As before, after the fit of the model, predictions for future periods are obtained by projecting $\mathbf{v}_j$ for $j \geq T$, under a time–series model assumption. Similarly, Plat (2009) specify a piece–wise linear basis expansion over ages with an additional cohort term. These attempts at including smoothness show an increase interest at better eliciting prior information on the array structure in the model. However, the linear structure and the combination of few bases is generally not sufficient to flexibly characterize the broad spectrum of global and local changes in age–specific mortality rates across years, thereby affecting forecasting performances (e.g., Camarda, 2019).

An effective option for addressing this problem is to rely on a more structured and interpretable basis expansion that incorporates possible heterogeneity in mortality patterns for different age groups. Currie et al. (2004) explore this direction via two–dimensional penalized B–splines to jointly model age–period patterns. Hyndman and Ullah (2007), instead, consider a functional principal component decomposition (Ramsay and Silverman, 2005) of a nonparametric smoothing estimate $f(x)_j \,|_{x=i} = f_{ij}$ of the mortality, as a function of age at each period,

$$y_{ij} = f_{ij} + \epsilon_{ij}. \tag{5}$$

The nonparametric smoothing function is composed of orthogonal basis functions $\phi_j(x)$

and a location $\mu(x)$ as follows:

$$f_j(x) = \mu(x) + \sum_{k=1}^{K} \beta_{jk}\phi_k(x) + e_j(x), \tag{6}$$

where $e_j(x)$ is a Gaussian noise term. The smoothness of the mortality trajectories over time is obtained by modelling the dependence of the set of coefficients $\{\beta_{jk}\}_{k=1}^{K}$ over periods $j \in \{1, \dots, T\}$ under a suitable stochastic process. The basis functions $\{\phi_k(x) \mid_{x=i}\}_{k=1}^{K}$ evaluated at each age i have the role of the age factors $\mathbf{u}_i = (u_{1i}, \dots, u_{Ki})^\intercal$ in our notation, while the latent coefficients $\{\beta_{jk}\}_{k=1}^{K}$ stand for the period latent factors $\mathbf{v}_j = (v_{1j}, \dots, v_{1K})^\intercal$. Such a model specification allows for smoothness both across ages and periods, while preserving a flexible specification through the expanded functional basis expansion in (6). Camarda (2019), instead, further extend the framework of Currie et al. (2004) eliciting additional prior knowledge in form of spline constraints, obtaining improved forecasting performance.

State–of–the–art predictive models for mortality rates (Hyndman and Ullah, 2007; Camarda, 2019) prove that the effort of flexibly eliciting available information on the data structure in the model effectively improves the goodness–of–fit and the forecasting accuracy. However, these proposals still retain crucial limiting drawbacks. The flexible functional decomposition of Hyndman and Ullah (2007) has limited interpretability of the basis expansion, and consequently does not allow us to directly express known age patterns. The interpretable B–spline construction of Camarda (2019) forces a constant smoothing both across ages and periods, and prevents from treating the period component as a time–indexed stochastic process on which to impose a suitable dynamic model for principled inference and forecasting.

In the thesis, we provide a novel model for forecasting mortality rates that combines the interpretability of a suitable basis expansion across ages with an extension of the locally–adaptive Gaussian process of Zhu and Dunson (2013) capturing trajectories over time periods. Our proposal overcomes the major limitations of the current approaches that we mentioned above. Crucially, the model is able to directly learn the rate of changes of the mortality rates across periods allowing us to build a robust forecasting procedure, which has shown improved performances compared to state–of–the–art alternatives.

Network analysis

Networks, or *graphs*, are collections of edges between pairs of nodes. They are widely used to represent and model relational data. In this context, edges denote observed relations

between pairs of agents. A network of n nodes can be represented by a $n \times n$ adjacency matrix $\mathbf{Y}$, with each entry $[\mathbf{Y}]_{ij} = y_{ij}$ representing the connection between nodes i and j. If the relations are dichotomous, e.g. absence or presence, and unidirectional, then $\mathbf{Y}$ is symmetric with binary entries. The two–dimensional data array $\mathbf{Y}$ is expanded with additional dimensions in case of multilayer networks.

Let us consider here the case of a single undirected binary network. Statistical modelling aims at representing the stochastic process responsible for the observed connectivity patterns in the data, preferably with a low–dimensional structure. Edges are drawn independently from a Bernoulli distribution given the edge probability θ_{ij},

$$y_{ij} \mid \theta_{ij} \overset{\text{ind}}{\sim} \text{BERN}(\theta_{ij}). \tag{7}$$

Models differ in terms of the probabilistic construction that defines the pairwise connection probabilities θ_{ij}. Both dimensions of the array $\mathbf{Y}$ are indexed by the same attribute, which is the node index. In this context, model (1) represents the wide–class of latent variable models. Popular statistical models, such as the stochastic block model (Nowicki and Snijders, 2001), the latent position model (Hoff et al., 2002) and the eigenmodel (Hoff, 2007), belong to this class.

For instance, in the stochastic block model each node belongs to one of K latent blocks. The probability of a connection between two nodes only depends on the block membership. Considering for each node the K–dimensional vector of block assignments, with entries equal zero except for the correspondent block equal to 1, the factor matrices $\mathbf{Z} := \mathbf{U} = \mathbf{V} \in \mathbb{R}^{n \times K}$ correspond to the collection of these vectors. If we denote with $\mathbf{\Theta}_{\text{B}} \in [0,1]^{K \times K}$ the matrix of pairwise connection probabilities between blocks, the stochastic block model assumes

$$\theta_{ij} = k(\mathbf{z}_i, \mathbf{z}_j) = \mathbf{z}_i^{\intercal} \mathbf{\Theta}_{\text{B}} \mathbf{z}_j. \tag{8}$$

In the latent position model, instead, nodes are embed in a K–dimensional latent space. In this case, the factor matrices $\mathbf{U} = \mathbf{V} \in \mathbb{R}^{K \times n}$ represents the coordinates of the node latent positions. Denoting with $\mathbf{Z} := \mathbf{U} = \mathbf{V}$, the edge probabilities depend on the pairwise distances in the latent space as follows:

$$\theta_{ij} = k(\mathbf{z}_i, \mathbf{z}_j) = \text{expit}\left(a - \| \mathbf{z}_i - \mathbf{z}_j \|\right), \tag{9}$$

where expit$(\cdot)$ is the inverse of the logit function. The original model of Hoff et al. (2002) assumes independent Gaussian priors on the latent position. In many contexts, such as in social networks, it is common to observe group of nodes characterized by high intra–

connectivity, forming what is typically named a community. In order to better capture these patterns, Handcock et al. (2007) propose the latent cluster model, where the latent positions are drawn from a mixture of Gaussian distribution. This prior introduce dependence among the nodes and the mixture component allows us to identify community structures.

In the thesis, we consider multiple networks representing the structural brain connectivity over a set of subjects. These types of networks have peculiar characteristics. Edges correspond to white matter fibers connecting pairs of brain regions. The formation of these connections is expensive in terms of material and energy costs, therefore regions highly connected tends to be located closeby (Bullmore and Sporns, 2009). Moreover, the connectivity between brain regions show a multiresolution or hierarchical structure, shown also in common partitions such as the hemispheres and lobes ones. Therefore, the prior specifications of the latent factors should attempt at eliciting this structural information of the data.

The stochastic block model has been extended to allow for nested partitioning of nodes with several proposals, see e.g. Roy et al. (2006); Clauset et al. (2008); Schmidt and Morup (2013). However, the limited ability of this class of models to recover certain connectivity patterns, such as homophily and triangles, motivates the attempt to model brain networks with a different approach. Fosdick et al. (2019) overcome this issue effectively combining the stochastic block model with the latent positions model, but their approach is limited to a single network.

We instead propose a new model belonging to the latent position class, the phylogenetic latent position model. This allows us to capture important connectivity patterns, while at the same time organizing the nodes in a multiresolution structure shared across multiple networks. This is achieved by imposing a tree–structured dependence on the latent positions, which effectively generalises previous latent positions model such as Hoff et al. (2002) and Handcock et al. (2007). The proposed model has demonstrated the ability to infer meaningful multiresolution organizations of the nodes, both in the context of simulated data and in the analysis of brain networks.

Summary of the specific contributions

The applications that we consider in the thesis have peculiar array structures that we carefully leverage in our models. Figure 2 shows the two different types of structural information included in the marginal factorizations of our proposed models of Chapter 1 and Chapter 2. One the left side, smoothness over consecutive ages and years, whereas on the right side hierarchical structure of the brain regions.

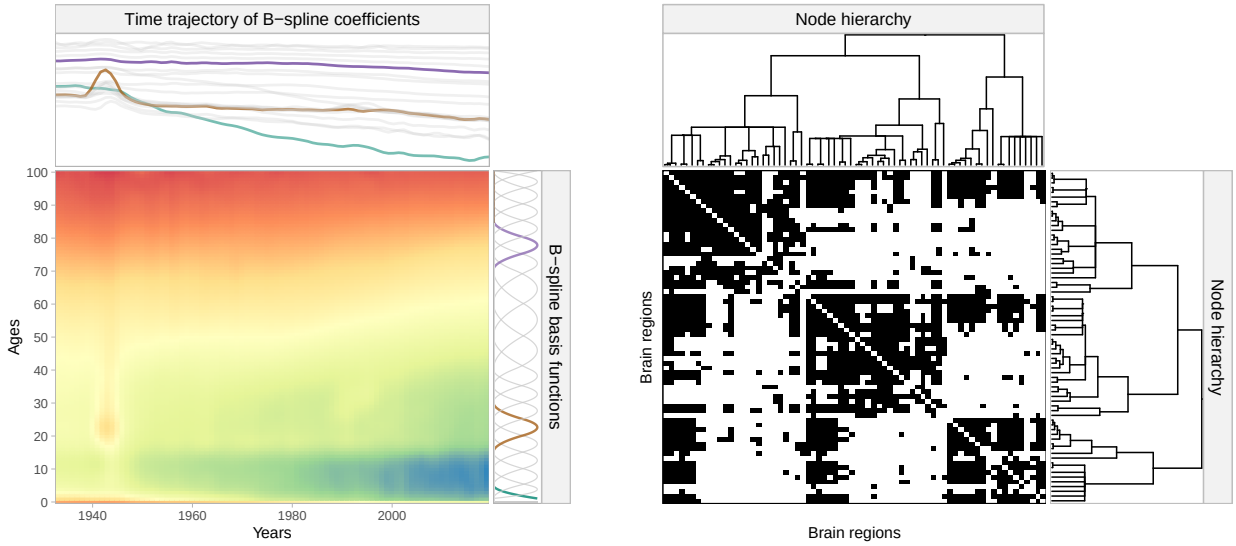


Figure 2: Same examples of Figure 1 under the model proposed in this thesis. On the left, the smoothed log–mortality rates after model fit. Smooth dependence across ages and years is respectively captured via B–spline basis expansion and Gaussian process dynamics. On the right, the adjacency matrix of brain connections after reordering the brain regions according to the estimated hierarchical structure.

Life tables, such as the mortality rates, have an explicit local dependence and smoothness, across ages and years. Available methods in the literature fail to properly express such nature. In Chapter 1, we propose an interpretable and flexible B–splines basis expansion to capture patterns across ages together with an extension of the locally adaptive Gaussian process of Zhu and Dunson (2013) that guarantees local adaptivity of the time component. The latter allows us to properly account for possible shocks in the mortality rates due to exogenous events, such as wars and epidemics. The Gaussian process directly learns the rates of change of the mortality rates through time, which are used to provide more accurate forecasts compared to state–of–the–art alternatives.

The type of dependence in the brain connectivity networks is instead of a different nature. Brain regions with similar functionality tends to be more connected. The classification in hemisphere and lobes reflects the hierarchical organization of the brain. The model should encode this information in the low–dimensional factorization. In Chapter 2, we specify a new latent position model where the node latent positions are the realization of Brownian motions over a phylogenetic tree. The latter allows us to directly infer the multiresolution organization of the brain regions. This model fills a gap in the latent position model literature for networks with hierarchical node structures. Finally, Chapter 3 explores the possibility to leverage additional covariates to inform the tree structure of the model.

The content of Chapter 1 is based on the following publication:

- Pavone, F., Legramanti, S., and Durante, D. (2022). Learning and forecasting of age-specific period mortality via B-spline processes with locally-adaptive dynamic coefficients. *arXiv preprint arXiv*:2209.12047.

A first version of the model has been used in:

- Pavone, F., and Legramanti, S. (2022). Bayesian analysis of mortality in Iceland via locally adaptive splines. *In Book of the short papers SIS 2022*. 520-525.

Chapter 1

Learning and Forecasting of Age–Specific Period Mortality via B–Spline Processes with Locally–Adaptive Dynamic Coefficients

JOINT WORK WITH SIRIO LEGRAMANTI AND DANIELE DURANTE

1.1 Introduction

Since the seminal contribution by Lee and Carter (1992) on stochastic modeling and forecasting of human mortality patterns, several efforts have been devoted towards the development of effective strategies characterized by increasing accuracy in predicting the future evolution of death rates for different age groups and countries (e.g., Booth and Tickle, 2008; Currie, 2016; Hunt and Blake, 2021). Due to its direct impact in guiding social, economic, environmental and health–care policies, such an endeavor is of paramount interest in a variety of fields, including demography (e.g., Lee and Miller, 2001; Li and Lee, 2005; Raftery et al., 2013; Li et al., 2013; Camarda, 2019), actuarial sciences (e.g., Renshaw and Haberman, 2003, 2006; Cairns et al., 2006; Plat, 2009; Currie, 2016) and statistics (e.g., Lee and Carter, 1992; Dellaportas et al., 2001; Hyndman and Ullah, 2007; Alexopoulos et al., 2019; Aliverti et al., 2022), among others. Nonetheless, despite this collective effort, there is still a lack of consensus on a superior solution. In fact, several peculiar characteristics of age–specific period mortality trajectories keep motivating active and ongoing innovations in stochastic modeling and forecasting of death rates via increasingly flexible

representations.

As illustrated in Figure 1.1, the age–specific period mortality surfaces exhibit a combination of global and local variations. When expressed as a function of age, these mortality trajectories display similar and generally–smooth shapes, whereas the overall dynamic evolution of these trajectories across periods exhibits a progressive downward shift, whose rate of change varies locally with both age classes and years. Although the inclusion of these core structures is expected to enhance both inference and forecasting performance, current literature still lacks a statistical model that can effectively address such goals within a single formulation. In fact, while successful extensions of the age–period bilinear formulation by Lee and Carter (1992) and of the additive age–period–cohort representation in, e.g., Holford (1983) improve flexibility via more general basis expansions of age effects with time–varying coefficients, the selected bases are either simple parametric functions often active in the whole age range (Brouhns et al., 2002; Czado et al., 2005; Cairns et al., 2006; Delwarde et al., 2007; Plat, 2009; Cairns et al., 2009; Haberman and Renshaw, 2011; O'Hare and Li, 2012; Wong et al., 2018) or are inferred via functional principal components analysis (Hyndman and Ullah, 2007; Hyndman et al., 2013). This implies that the induced death–rate forecasts are mainly based on a combination of global trends in mortality across ages which do not explicitly account for local heterogeneities in mortality levels and the corresponding rates of change for specific age classes. Recalling Figure 1.1, the mortality patterns exhibit both global and local variations across years and ages, thereby suggesting that a suitable representation capable of including these two behaviors would yield improved forecasts with respect to those obtained under a mainly–global perspective.

An effective option for addressing the aforementioned goal is to rely on a more structured and interpretable basis expansion that incorporates possible heterogeneity in mortality patterns for different age groups. Within this framework, the contribution by Heligman and Pollard (1980) provides a first effective answer which expresses the age pattern of mortality via a combination of three basis functions corresponding to infant mortality, accident hump and elderly–age mortality; see also Dellaportas et al. (2001), Mazzuco et al. (2018) and Alexopoulos et al. (2019) for subsequent extensions. While these formulations yield interpretable inference, the combination of only three bases is generally not sufficient to flexibly characterize the broad spectrum of global and local changes in age–specific mortality rates across years, thereby affecting forecasting performance (e.g., Camarda, 2019). To overcome this issue, a possible solution consists in specifying a richer set of basis functions, each active — i.e., non–zero — only in a subset of the ages, with these subsets varying across bases to cover the whole age range. Expressing the age pattern of mortality through a linear combination of these basis func-

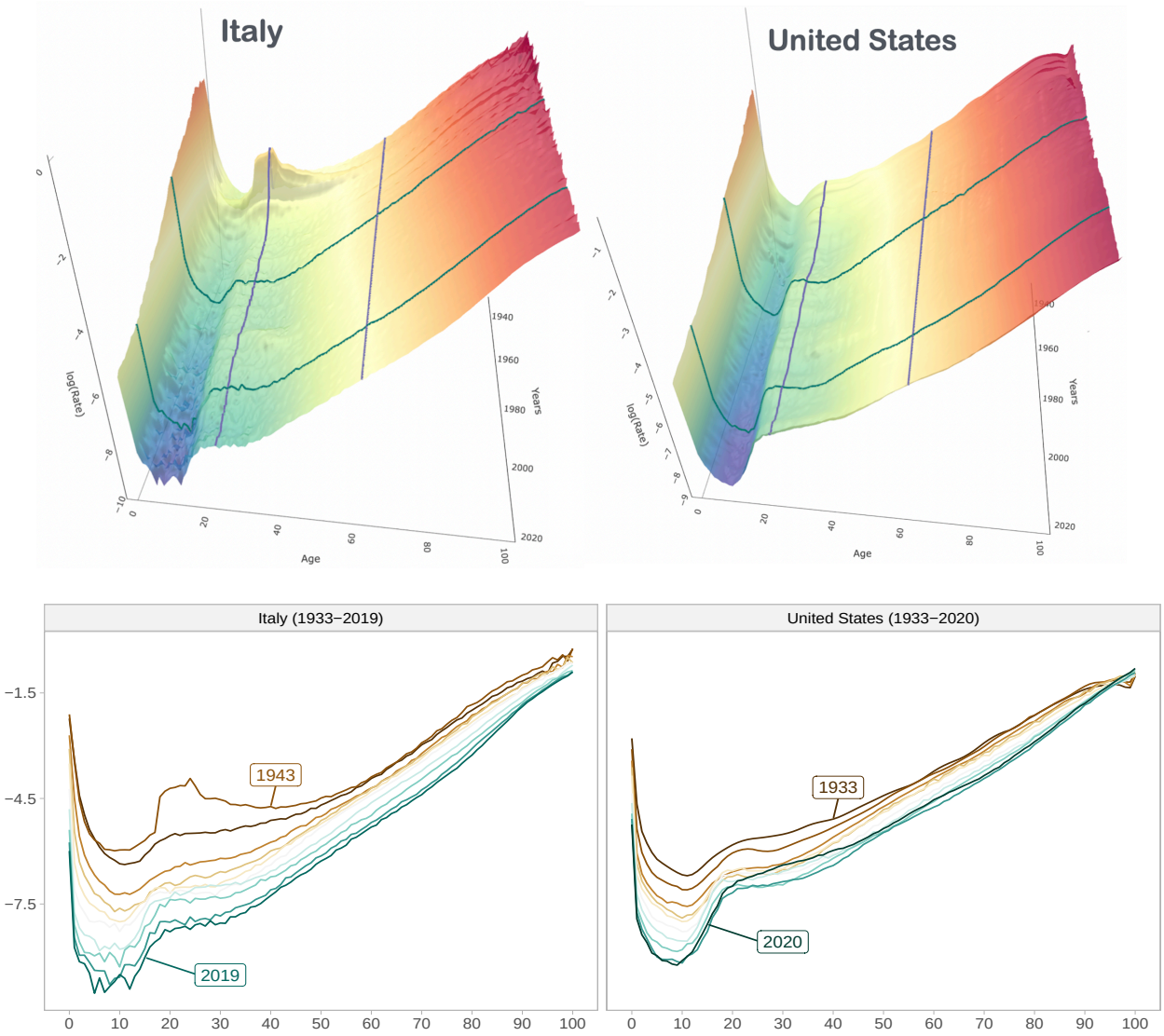


Figure 1.1: Graphical representations of the observed age–period log–mortality rates from 1933 until 2019 for Italy, and from 1933 to 2020 for United States. Top panels provide 3D visualizations of the age–period log–mortality rate surface, whereas bottom panels comprise a 2D illustration of the age–specific trajectories for each period. Data are retrieved from the HUMAN MORTALITY DATABASE (https://www.mortality.org/).

tions yields a globally–smooth, yet flexible, representation which additionally accounts for possible local heterogeneities in specific age classes via the control on the coefficients for the bases active in those classes. Such a direction has been partially explored in Currie et al. (2004) via two–dimensional penalized B–splines to jointly characterize age–period patterns of mortality; see also Camarda (2019) for a recent effective extension of this approach which incorporates suitable constraints and prior knowledge to improve forecasting performance. Although both formulations provide a sensible representation of age–specific period mortality surfaces, the two–dimensional B–splines perspective enforces a constant smoothing both across ages and periods, and prevents from treating the period component as a time–indexed stochastic process on which to impose a suitable dynamic model for principled inference and forecasting. As shown in Figure 1.1 with a focus on two of the countries analyzed in our application, while the age pattern of mortality often exhibits a smooth trajectory, the time changes in such a trajectory fluctuate between periods of rapid and slow variations, affecting the age classes with different magnitudes. These peculiar characteristics necessarily require a careful statistical model which can effectively combine interpretable basis expansions for the age patterns of mortality with a flexible stochastic process having locally–varying smoothness for the dynamic evolution of such patterns across periods. While the aforementioned contributions include some of these structures, there is still the lack of a unique representation that can effectively incorporate all these characteristics within a single formulation.

Motivated by the above discussion and by the mortality data discussed in Section 1.1.1, we cover such a gap in Section 1.2 by defining a Poisson log–normal model for the age–specific death counts whose rate is parameterized via a novel B–spline process with locally–adaptive dynamic coefficients which extends the nested Gaussian process by Zhu and Dunson (2013) in a number of directions. Our novel process characterizes the age pattern of mortality via a suitable combination of interpretable B–spline bases — each active in different age intervals — and incorporates flexible dynamic changes in such patterns by allowing the splines coefficients to evolve in time via a system of stochastic differential equations that account for locally–varying smoothness in time trajectories, and facilitate borrowing of information across coefficients of contiguous splines. This representation is conceptually and practically more suitable than the bivariate B–splines approach in Currie et al. (2004) and Camarda (2019) since it allows to properly treat the period component as a dynamic locally–adaptive stochastic process rather than just a function of time with constant smoothness. In addition, it yields a more flexible characterization of age–period mortality patterns relative to classical parametric extensions of the Lee and Carter (1992) model in, e.g., Brouhns et al. (2002); Czado et al. (2005); Cairns et al. (2006); Delwarde et al. (2007); Plat (2009); Cairns et al. (2009); Haberman and Renshaw (2011);

O'Hare and Li (2012) and Wong et al. (2018), while preserving interpretability via the use of B–spline bases instead of those inferred from, e.g., functional principal components (Hyndman and Ullah, 2007; Hyndman et al., 2013).

As clarified in Sections 1.2–1.4, these advancements yield a model which is both flexible and interpretable, thereby improving accuracy in point forecasts, calibration of predictive intervals, and inference potentials relative to state–of–the–art formulations, at no additional cost in computational tractability. In fact, in Section 1.3 we derive a provably accurate Gaussian state–space approximation of the proposed model that allows the implementation of closed–form Kalman filter updates for smoothing, filtering and forecasting of both the trends and the first derivatives for the trajectories of the spline coefficients. This computational tractability is in contrast with recent flexible representations that require MCMC methods to benefit from a fully–Bayesian approach which further allows the choice of priors for the structural model parameters (e.g., Wong et al., 2018; Alexopoulos et al., 2019). Moreover, unlike for state–of–the–art extensions of the Lee and Carter (1992) model which generally employ an ARIMA formulation for the dynamic parameters (Brouhns et al., 2002; Czado et al., 2005; Cairns et al., 2006; Hyndman and Ullah, 2007; Plat, 2009; Cairns et al., 2009; Haberman and Renshaw, 2011; O'Hare and Li, 2012; Hyndman et al., 2013), our proposal explicitly incorporates and flexibly learns not only mortality trends but also the corresponding time–varying rates of change. Although the importance of accounting for dynamic rates of change in mortality forecasting has been recently illustrated in Camarda (2019), this concept has received limited attention to date and there is a lack of models which explicitly include and learn these higher–level patterns within a single formulation. The empirical performance illustrated in Section 1.4 for our proposed model clarifies that this additional structure is not only beneficial in delivering improved point forecasts and predictive intervals than state–of–the–art competing methods, but also allows to quantify and compare relevant mortality accelerations experienced both in past and recent years across different countries and age groups. For example, our model reveals substantially different patterns in age–specific mortality across countries during the last two decades and in the recent COVID–19 pandemic. Concluding remarks and future research directions are provided in Section 1.5, whereas codes and tutorial implementations are available at https://github.com/fpavone/BSP-mortality.

1.1.1 Motivating Application

The novel B–spline process (BSP) developed in Sections 1.2–1.3 is meant to provide a general modeling and forecasting solution that can be applied to any country. To this end, the motivating application we consider in Section 1.4 aims at illustrating the practical advantages of BSP in learning and forecasting several mortality patterns characterized by

a broad range of different evolutions across years and ages, ranging from smooth trajectories to more rapid shocks, over a broad time horizon. This motivates our focus on four illustrative countries, namely Italy, Sweden, United Kingdom and United States, whose gender–specific age–period log–mortality rates are available from the Human Mortality Database (https://www.mortality.org/) for a wide time range, that spans from 1933 until either 2019 or 2020 and displays different country–specific evolutions of mortality over periods and age classes, along with fluctuations of varying magnitude due to shocks.

Although the Human Mortality Database comprises data also for other countries, the corresponding time window is generally much shorter than those considered in Section 1.4. In addition, Italy, Sweden, United Kingdom and United States exhibit a number of peculiar characteristics which make these countries of particular interest not only in forecasting, but also for inference. More specifically, as we will illustrate in Figures 1.4 and 1.5, Italy and the United Kingdom provide interesting examples to quantify the ability of the proposed BSP in flexibly learning different magnitudes of the mortality shock, and the corresponding rates of change, associated with the World War II. Recalling, e.g., Vaupel and Lundstrom (1994), Sweden is historically characterized by low mortality rates, but the recent evidence of slower rates of increment in the life expectancy relative to other countries (Drefahl et al., 2014) and the less stringent policy adopted during COVID–19 (Wang et al., 2022a; Juul et al., 2022) make Sweden an interesting case study. The United States have also experienced a slower increment in life expectancy in recent years, which culminated in a decreasing pattern over the past decade (Woolf and Schoomaker, 2019). This specific behavior has motivated several explanatory studies mostly focused on peculiar mortality patterns and vulnerabilities associated with young and adult age classes (e.g., Remund et al., 2018; Glei, 2022), which could also explain particular differences in the age–specific excess mortality in the United States during COVID–19, relative to the patterns observed for the other countries (e.g., Katzmarzyk et al., 2020; Wiemers et al., 2020; Goldstein and Lee, 2020). The BSP formulation developed in Sections 1.2–1.3 is carefully designed to flexibly incorporate all these multifaceted patterns and, hence, the analysis of these four countries provides a comprehensive setting to obtain empirical evidence of improved performance in inference and forecasting relative to state–of–the–art alternatives.

As highlighted in Section 1.4, the proposed BSP yields improved forecasts also when applied to a different subgroup of countries from the Human Mortality Database, such as, for example, Czech Republic, Denmark and France. A discussion on future studies of BSP performance for low– and middle–income countries, whose data are currently not available in the Human Mortality Database, can be found in Section 1.5.

1.2 Model Formulation

Let d_{xt} and E_{xt} be the total death counts and the average number of individuals at risk (also known as *central exposed to risk*) at age x in period t, within a given population. Following the overarching focus in the literature on mortality modeling (e.g., Booth and Tickle, 2008; Hunt and Blake, 2021) our aim is to improve inference and forecasting of the observed central mortality rates defined as $m_{xt} = d_{xt}/\text{E}_{xt}$.

To this end, let $\overline{m}_{xt} = \mathbb{E}(m_{xt} \mid \overline{m}_{xt})$ denote the underlying expected mortality rate at age x within period t, we introduce in Section 1.2.1 a Poisson log–normal model for d_{xt}, whose rate parameter $\text{E}_{xt}\overline{m}_{xt}$ is allowed to flexibly vary across both ages x and periods t via a novel B–spline process for $\log \overline{m}_{xt}$. As clarified in Section 1.2.2, such a model admits a provably–accurate Gaussian state–space approximation which expresses the observed log–mortality rates $\log m_{xt} = \log(d_{xt}/\text{E}_{xt})$ as a linear combination of B–spline bases whose dynamic coefficients and the associated derivatives vary in time via a system of Gaussian state equations. This allows closed–form filtering, smoothing and forecasting of the coefficients trajectories and, as a consequence, of the induced patterns in the log–mortality rates $\log m_{xt}$ via a direct application of standard Kalman filter updates (Kalman, 1960); see Section 1.3.

1.2.1 B–Spline Process with Locally–Adaptive Dynamic Coefficients

Recalling the above discussion, we model the death counts d_{xt}, at each age $x \in \mathcal{X} \subset \mathbb{R}^+$ and period $t \in \mathcal{T} \subset \mathbb{R}^+$ via the Poisson log–normal distribution

$$(d_{xt} \mid \overline{m}_{xt}) \overset{\text{ind}}{\sim} \text{Poisson}(\text{E}_{xt}\overline{m}_{xt}), \quad \text{with} \quad (\log \overline{m}_{xt} \mid f_t(x)) \overset{\text{ind}}{\sim} \text{N}(f_t(x), \sigma^2_{\overline{m}}), \tag{1.1}$$

for every $x \in \mathcal{X}$ and $t \in \mathcal{T}$, where $f_t(x)$ denotes a flexible function of age x whose shape is allowed to vary with time t, whereas $\sigma^2_{\overline{m}}$ encodes the global amount of over–dispersion within the observed death counts. The Poisson log–normal assumption in (1.1) has been considered in Wong et al. (2018) to account for extra variability in the Poisson Lee–Carter model proposed by Brouhns et al. (2002) and Czado et al. (2005). Although this is a sensible modification which allows to formally incorporate age–specific heterogeneity in period mortality — possibly arising from differences in cohort effects — Wong et al. (2018) still rely on the classical Lee and Carter (1992) parametric bilinear form for $f_t(x)$. As illustrated in Table 1.1 (see column LC), such a form yields an overly–restrictive characterization of age–period mortality patterns which affects both inference and forecasting performance; see also Delwarde et al. (2007) for an additional example of a Poisson log–bilinear formulation which employs the classical Lee and Carter (1992) construction.

To address the above issues and incorporate the core patterns of mortality discussed within Section 1.1 and illustrated in Figure 1.1, we combine the statistical model in (1.1) with a flexible, yet interpretable, representation for $f_t(x)$ based on a novel B–spline process with locally–adaptive dynamic coefficients. This formulation defines $f_t(x)$ via a linear combination of p pre–selected B–spline basis functions of age, $g_1(x), \dots, g_p(x)$, whose associated coefficients $\beta_1(t), \dots, \beta_p(t)$ jointly evolve in time via a system of stochastic differential equations that induce locally–varying smoothness and borrowing of information across contiguous bases. In particular, let $\mathbf{b}(t) = [\beta_1(t), \partial\beta_1(t)/\partial t, a_1(t), \dots, \beta_p(t), \partial\beta_p(t)/\partial t, a_p(t)]^\intercal$ be the $(3p \times 1)$–dimensional process comprising the p B–splines coefficients $\beta_1(t), \dots, \beta_p(t)$, the corresponding first derivatives $\partial\beta_1(t)/\partial t, \dots, \partial\beta_p(t)/\partial t$, and the associated local instantaneous mean functions $a_1(t), \dots, a_p(t)$ which induce time–varying smoothness by controlling the expected value of the second derivatives at time t, namely $a_j(t) = \mathbb{E}[\partial^2\beta_j(t)/\partial^2 t \mid a_j(t)]$, for $j = 1, \dots, p$. Moreover, denote with $\boldsymbol{\varepsilon}_t = [\varepsilon_{\beta_1}(t), \varepsilon_{a_1}(t), \dots, \varepsilon_{\beta_p}(t), \varepsilon_{a_p}(t)]^\intercal$ a $(2p\times1)$–dimensional vector encoding independent Gaussian white noise processes. Then, leveraging these quantities and letting $\tau = t/\lambda$ be a reference time scale, the proposed BSP assumes

$$f_t(x) = \sum_{j=1}^{p} \beta_j(t) g_j(x), \qquad \text{for any } x \in \mathcal{X} \text{ and } t \in \mathcal{T}, \tag{1.2}$$

$$\partial\mathbf{b}(\lambda\tau)/\partial\tau = \lambda(\mathbf{I}_p \otimes \mathbf{C})\mathbf{b}(\lambda\tau) + (\mathbf{I}_p \otimes \mathbf{D})(\boldsymbol{\Omega}^{1/2}\boldsymbol{\varepsilon}_\tau), \qquad \text{for any } \lambda\tau = t \in \mathcal{T}, \tag{1.3}$$

where $\mathbf{I}_p$ is the $p \times p$ identity matrix, $\otimes$ denotes the Kronecker product, $\lambda > 0$ corresponds to a length–scale parameter which allows to preserve time unit invariance, $\boldsymbol{\Omega}$ is a suitably–specified $2p \times 2p$ correlation matrix which induces local borrowing of information across contiguous splines coefficients — via the correlation among the corresponding derivatives and local instantaneous means — whereas $\mathbf{C}$ and $\mathbf{D}$ are known system matrices defined as

$$\mathbf{C} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \qquad \mathbf{D} = \begin{bmatrix} 0 & 0 \\ \sigma_\beta & 0 \\ 0 & \sigma_a \end{bmatrix}, \tag{1.4}$$

with $\sigma_\beta > 0$ and $\sigma_a > 0$ denoting two scaling parameters. As clarified in (1.4), these matrices are pre–specified to induce the desired system of stochastic differential equations; see Zhu and Dunson (2013, A.6) for a related definition of $\mathbf{C}$ and $\mathbf{D}$ in the univariate case.

As illustrated in Figure 1.2, the B–spline process construction in equations (1.2)–(1.4) provides an effective formulation which treats $f_t(x)$ as a function of age x, via a linear combination of interpretable B–spline bases directly associated to specific age classes, and as a stochastic process of time $t = \lambda\tau$, leveraging a flexible system of stochastic dif-

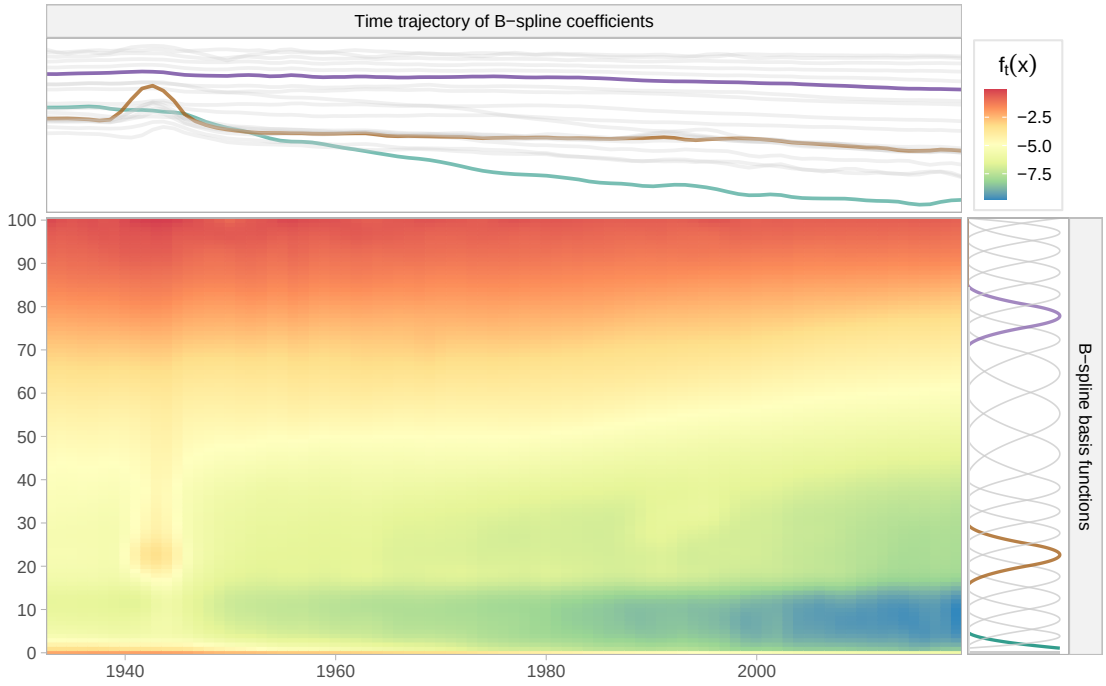


Figure 1.2: Illustrative example of a heatmap for $f_t(x)$, as defined in (1.2) via a linear combination of pre–specified B–spline bases (right–side panel) with coefficients varying across periods according to (1.3) (top–side panel). For illustrative purposes, three B–spline bases and the corresponding coefficients trajectories are highlighted with different colors; see Section 1.4 for details on the choice of the number and location of the B–spline bases.

ferential equations that jointly characterize the time trajectory of each spline coefficient $\beta_j(t)$, for $j = 1, \dots, p$, by explicitly modeling its smoothness across periods. In (1.3), such a smoothness is measured by the second–order derivative $\partial^2\beta_j(t)/\partial^2 t$ which is in turn centered on a higher level time–varying instantaneous mean function $a_j(t)$ that allows local adaptivity. Combining (1.2)–(1.4) with model (1.1) yields a unique representation for age–period mortality patterns that (i) accounts for age–specific heterogeneity in death counts via the log–normal assumption in (1.1), (ii) enforces a generally smooth trajectory for the age pattern of mortality through the linear combination of B–splines in (1.2), and (iii) explicitly allows these patterns to evolve in time between periods of rapid and slow variations, affecting age classes with different magnitudes, via the system of stochastic differential equations in (1.3)–(1.4) for the splines coefficients.

To further clarify representation (1.3)–(1.4), it shall be emphasized that such a construction extends the nested Gaussian process of Zhu and Dunson (2013) in a number of directions inherently motivated by our focus on modeling and forecasting of mortality rates. In fact, the original formulation by Zhu and Dunson (2013) does not consider (1.2), and provides a simpler version of (1.3) with a focus on inducing locally–varying

smoothness in a single trajectory via a Gaussian process (e.g., Rasmussen and Williams, 2006) for a derivative of selected order, which is in turn centered on a higher–level Gaussian process characterizing the local instantaneous mean function. Although considering separate nested Gaussian processes for the trajectories $f_t(x)$ of each age $x \in \mathcal{X}$ is a viable option, this representation is not invariant with respect to the choice of the time unit and, more crucially, it fails to borrow information among mortality patterns for contiguous ages. Recalling, e.g., Currie et al. (2004) and Camarda (2019) the latter property would be conceptually and practically useful since it is reasonable to expect that the mortality patterns of close age classes display a natural dependence. To this end, equations (1.2)–(1.3) extends Zhu and Dunson (2013) to a structured multivariate formulation which induces such a borrowing of information through the B–spline representation of $f_t(x)$ in (1.2) and by inducing dependence among the B–spline coefficients in (1.3) via the introduction of correlation between the white noises through the matrix $\boldsymbol{\Omega}$. This matrix has unit diagonal, and off–diagonal elements that are non–zero only for the entries $\boldsymbol{\Omega}_{j,l}$ whose indexes (j, l) are either both even or both odd, so as to induce correlation among the noises associated with the derivatives and local instantaneous means, respectively. As clarified in Section 1.3, by defining these non–zero correlations via suitable covariance functions (e.g., Rasmussen and Williams, 2006) allows to enforce a local borrowing of information which decays as the distance between age classes grows. The introduction of the length–scale parameter λ allows, instead, to preserve time unit invariance, so that, if the time scale is changed — e.g., from ∂t to $c \cdot \partial t$ — it is still possible to retrieve the same model specification with a suitable specification of the parameters λ, σ_β and σ_a. In fact, the scale of t is often arbitrary in practice and, hence, it is desirable to define the process in (1.2) with respect to the reference time $\tau = t/\lambda$. This modification is in line with similar operations considered in the Gaussian process literature when including a length–scale parameter in popular covariance functions (e.g., Rasmussen and Williams, 2006). The empirical results in Section 1.4 confirm that these extensions yield substantial gains in mortality forecasts relative to those obtained via a direct application of the original nested Gaussian process by Zhu and Dunson (2013) to each trajectory $f_t(x)$, $t \in \mathcal{T}$, separately for every age x.

Besides including the core age–period structures of mortality, model (1.1)–(1.4) crucially admits a provably accurate Gaussian state–space approximation, as described in Section 1.2.2 below. This representation further clarifies the proposed model and allows efficient computation via standard Kalman filter updates; see also Section 1.3 and refer to Section 1.4 for details on the choice of the number and location of the B–spline bases.

1.2.2 Gaussian State–Space Approximation

As a first step towards the derivation of an accurate and computationally tractable Gaussian state–space representation of the proposed formulation in (1.1)–(1.4), Proposition 1 proves that model (1.1) induces a distribution on the observed log–mortality rates $\log m_{xt} = \log(d_{xt}/\mathrm{E}_{xt})$ which can be closely approximated, for E_{xt} large enough, by the $\mathrm{N}(f_t(x), \sigma^2_{\overline{m}})$ assumed in (1.1) for the expected log–mortality.

Proposition 1. *Under model* (1.1), $\log m_{xt} = \log(d_{xt}/\mathrm{E}_{xt}) \to \mathrm{N}(f_t(x), \sigma^2_{\overline{m}})$ *in distribution, as* $\mathrm{E}_{xt} \to \infty$, *for any* $x \in \mathcal{X}$ *and* $t \in \mathcal{T}$.

Proposition 1 motivates direct focus on the observed log–mortality rates $\log m_{xt}$, which are of overarching interest in state–of–the–art studies (e.g., Land, 1986; Booth and Tickle, 2008; Currie, 2016; Hunt and Blake, 2021). In addition, it justifies the adoption of the Gaussian regression model $\log m_{xt} = f_t(x) + \nu_{xt}$, with $\nu_{xt} \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma^2_{\overline{m}})$ for any $x \in \mathcal{X}$, $t \in \mathcal{T}$, and $f_t(x)$ as in (1.2), which is arguably more tractable than the Poisson log–normal representation for the death counts in (1.1). Recalling Proposition 1, this approximation is provably accurate in settings with large enough E_{xt}, a common situation in mortality studies by country, where E_{xt} is typically in the order of tens–to–hundreds of thousands.

Although Proposition 1 yields a simpler construction, to obtain a fully tractable formulation it is also necessary to derive an alternative representation for the stochastic differential equations in (1.3)–(1.4) which is amenable to efficient computation, direct forecasting, interpretable inference and effective uncertainty quantification. Proposition 2 proves that, when observed at a finite collection of times $t_1, \dots, t_n$, as in the mortality–data context, equations (1.3)–(1.4) admit a tractable representation via a linear system of Gaussian state equations.

Proposition 2. *Let* $\mathbf{b}_{t_s}$ *denote the realization at a generic time* t_s *of the process* $\mathbf{b}(t)$ *defined in Section 1.2.1, i.e.,* $\mathbf{b}_{t_s} = [\beta_1(t), \partial\beta_1(t)/\partial t, a_1(t), \dots, \beta_p(t), \partial\beta_p(t)/\partial t, a_p(t)]^\intercal_{|t=t_s}$. *Then, for each finite grid of times* $t_s = t_1, \dots, t_n$, *with* $t_1 < \cdots < t_n$, *the system of stochastic differential equations in* (1.3)–(1.4) *admits the Gaussian state–equation representation*

$$\mathbf{b}_{t_{s+1}} = \mathbf{T}_{t_s}\mathbf{b}_{t_s} + \boldsymbol{\eta}_{t_s}, \quad \boldsymbol{\eta}_{t_s} \overset{\text{ind}}{\sim} \mathrm{N}_{3p}(\mathbf{0}, \mathbf{Q}_{t_s}), \qquad \text{for } t_s = t_1, \dots, t_n, \tag{1.5}$$

where $\mathbf{T}_{t_s}$ *denotes a* $3p \times 3p$ *block–diagonal transition matrix defined as*

$$\mathbf{T}_{t_s} = \mathbf{I}_p \otimes \begin{bmatrix} 1 & \lambda\delta_s & \lambda^2(\delta_s^2/2) \\ 0 & 1 & \lambda\delta_s \\ 0 & 0 & 1 \end{bmatrix}, \tag{1.6}$$

with $\delta_s = (t_{s+1} - t_s)/\lambda$, *while* $\mathbf{Q}_{t_s}$ *is a* $3p \times 3p$ *covariance matrix having generic block*

$$\mathbf{Q}_{t_s[j,l]} = \rho_{\beta[j,l]} \begin{bmatrix} (\delta_s^3/3)\lambda^2\sigma_\beta^2 & (\delta_s^2/2)\lambda\sigma_\beta^2 & 0 \\ (\delta_s^2/2)\lambda\sigma_\beta^2 & \delta_s\sigma_\beta^2 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \rho_{a[j,l]} \begin{bmatrix} (\delta_s^5/20)\lambda^4\sigma_a^2 & (\delta_s^4/8)\lambda^3\sigma_a^2 & (\delta_s^3/6)\lambda^2\sigma_a^2 \\ (\delta_s^4/8)\lambda^3\sigma_a^2 & (\delta_s^3/3)\lambda^2\sigma_a^2 & (\delta_s^2/2)\lambda\sigma_a^2 \\ (\delta_s^3/6)\lambda^2\sigma_a^2 & (\delta_s^2/2)\lambda\sigma_a^2 & \delta_s\sigma_a^2 \end{bmatrix} \tag{1.7}$$

for each $j = 1, \dots, p$ *and* $l = 1, \dots, p$, *with* $\rho_{\beta[j,l]}$ *and* $\rho_{a[j,l]}$ *denoting those entries of* $\boldsymbol{\Omega}$ *in* (1.3) *that measure the correlation among the derivatives of the coefficients associated with splines* j *and* l, *and between the corresponding local instantaneous means, respectively.*

Proposition 2 yields a simple state–space representation expressing the value of $\beta_j(t_{s+1})$ at time t_{s+1} via a second–order stochastic Taylor expansion of the trajectory $\beta_j(t)$ around the previous time point t_s, for each $j = 1, \dots, p$; see the form of $\mathbf{T}_{t_s}$ in (1.6). This allows to explicitly model and forecast not only class–specific mortality trends encoded in $\beta_1(t), \dots, \beta_p(t)$, but also the associated rates of change measured by $\partial\beta_1(t)/\partial t, \dots, \partial\beta_p(t)/\partial t$ and the corresponding instantaneous means $a_1(t), \dots, a_p(t)$. Recalling Section 1.2.1, the choice of the time scale of t is often arbitrary in practice. In fact, the actual values of $t_s = t_1, \dots, t_n$ are not necessary in equations (1.5)–(1.7), which only require to pre–specify the time lags δ_s among consecutive observations under the reference scale t/λ. In our equally–spaced context we set lags to 1, and then learn the appropriate scaling λ via maximum likelihood under (1.5)–(1.7).

The above representation is both flexible and interpretable, and further allows to borrow information across coefficients of different B–splines via the covariance matrix $\mathbf{Q}_{t_s}$ of the noise $\boldsymbol{\eta}_{t_s}$; see (1.7). The core parameters regulating the strength of this dependence are $\rho_{\beta[j,l]}$ and $\rho_{a[j,l]}$, for $j = 1, \dots, p$ and $l = 1, \dots, p$. Letting $\rho_{\beta[j,l]} = \mathbb{1}(j = l)$ and $\rho_{a[j,l]} = \mathbb{1}(j = l)$ yields no borrowing of information and, as a consequence, separate state–equations for each B–spline coefficient, whereas, whenever $\rho_{\beta[j,l]} \in (0, 1]$ and $\rho_{a[j,l]} \in (0, 1]$, the j–th and l–th splines display a dependence in the trajectories of the associated coefficients. More specifically, large values of $\rho_{\beta[j,l]}$ and $\rho_{a[j,l]}$ imply high correlation between the first derivatives and local instantaneous means functions, respectively, of the coefficient trajectories for splines j and l. This allows to borrow information in terms of both the overall trend and smoothness, while inducing dependence among the actual trajectories $\beta_j(t)$ and $\beta_l(t)$ under (1.5)–(1.7).

Recalling the above discussion and extending related ideas from P–splines representations (Eliers and Marx, 1996; Lang and Brezger, 2004), we define $\rho_{\beta[j,l]}$ and $\rho_{a[j,l]}$ to induce a local borrowing of information whose strength decays with a suitable distance between the j–th and l–th splines. More specifically, let $\bar{x}_j$ and $\bar{x}_l$ denote the ages at

which the B–spline functions $g_j(x)$ and $g_l(x)$ are maximized, respectively, we define

$$\rho_{\beta[j,l]} = \mathcal{K}(\bar{x}_j, \bar{x}_l; \boldsymbol{\gamma}_\beta), \qquad \text{and} \qquad \rho_{a[j,l]} = \mathcal{K}(\bar{x}_j, \bar{x}_l; \boldsymbol{\gamma}_a), \tag{1.8}$$

for every spline $j = 1, \dots, p$ and $l = 1, \dots, p$, where $\mathcal{K}(\bar{x}_j, \bar{x}_l; \boldsymbol{\gamma}_\beta)$ and $\mathcal{K}(\bar{x}_j, \bar{x}_l; \boldsymbol{\gamma}_a)$ denote user–selected covariance functions (e.g., Rasmussen and Williams, 2006, Ch. 4), which decay to zero as $|\bar{x}_j - \bar{x}_l|$ grows, and are defined such that $\mathcal{K}(\bar{x}_j, \bar{x}_j; \boldsymbol{\gamma}_\beta) = \mathcal{K}(\bar{x}_j, \bar{x}_j; \boldsymbol{\gamma}_a) = 1$ for $j = 1, \dots, p$. As a consequence, the time patterns of mortality are allowed to effectively share local information across contiguous age classes, and the strength of this dependence progressively decreases for distant ages with a pattern that depends on the selected covariance functions and on the associated parameters $\boldsymbol{\gamma}_\beta$ and $\boldsymbol{\gamma}_a$. Routinely–implemented examples of covariance functions are the squared exponential and the Matérn, among others (e.g., Rasmussen and Williams, 2006, Ch. 4); see Section 1.4 for details on suitable specifications of these covariance functions and the corresponding parameters in the mortality–data context.

Combining Propositions 1 and 2 yields the tractable Gaussian state–space model for the observed log–mortality rates

$$\log \mathbf{m}_{t_s} = \mathbf{Z}_{t_s} \mathbf{b}_{t_s} + \boldsymbol{\nu}_{t_s}, \qquad \boldsymbol{\nu}_{t_s} \overset{\text{ind}}{\sim} \mathrm{N}_k(\mathbf{0}, \mathbf{H}_{t_s}), \tag{1.9}$$

$$\mathbf{b}_{t_{s+1}} = \mathbf{T}_{t_s} \mathbf{b}_{t_s} + \boldsymbol{\eta}_{t_s}, \qquad \boldsymbol{\eta}_{t_s} \overset{\text{ind}}{\sim} \mathrm{N}_{3p}(\mathbf{0}, \mathbf{Q}_{t_s}), \tag{1.10}$$

for every time $t_s = t_1, \dots, t_n$, where $\log \mathbf{m}_{t_s} = (\log m_{x_1,t_s}, \dots, \log m_{x_k,t_s})^\intercal$ is the $(k \times 1)$–dimensional vector of the log–mortality rates observed for ages $x_1, \dots, x_k$ at time t_s, $\mathbf{Z}_{t_s} = [\mathbf{g}_1, \mathbf{0}, \mathbf{0}, \mathbf{g}_2, \mathbf{0}, \mathbf{0}, \dots, \mathbf{g}_p, \mathbf{0}, \mathbf{0}]$ denotes the $(k \times 3p)$–dimensional design matrix with non–zero columns $\mathbf{g}_j = [g_j(x_1), \dots, g_j(x_k)]^\intercal$, $j = 1, \dots, p$ comprising the values of the pre–selected B–splines bases at the observed ages $x_1, \dots, x_k$, $\mathbf{H}_{t_s} = \sigma^2_{\bar{m}} \mathbf{I}_k$, whereas $\mathbf{b}_{t_s}$, $\mathbf{T}_{t_s}$ and $\mathbf{Q}_{t_s}$ are defined as in Propositions 2.

As clarified in Section 1.3, the above Gaussian state–space formulation allows closed–form filtering, smoothing and forecasting via simple recursive equations obtained from a direct application of classical Kalman filter updates (Kalman, 1960; Koopman and Durbin, 2000); see also Durbin and Koopman (2012) and Chopin and Papaspiliopoulos (2020) for a general treatment of the Kalman filter and smoother in linear–Gaussian state–space models, and refer to the R package KFAS (Helske, 2017) for an effective implementation. This tractability is in contrast with the recently–proposed flexible mortality models which require MCMC routines (e.g., Wong et al., 2018; Alexopoulos et al., 2019) and, unlike for state–of–the–art formulations discussed in Section 1.1, model (1.9)–(1.10) holds not only for equally–spaced time grids $t_1, \dots, t_n$ but also for unequally–spaced ones. Such a generality is conceptually and practically useful in allowing inference and forecasting for

different time horizons, which is of interest, for example, during periods of mortality shocks to rapidly revise forecasts in the short term, e.g., within months, trimesters or semesters.

1.3 Filtering, Smoothing and Forecasting

In Section 1.3.1 we leverage model (1.9)–(1.10) to derive tractable strategies for probabilistic inference and prediction of the coefficients vector $\mathbf{b}_{t_s}$ defined in Proposition 2 and, as a direct consequence, of the $(k \times 1)$–dimensional log–mortality rates mean vector $\mathbf{f}_{t_s}$ defined as $\mathbf{f}_{t_s} = \mathbf{Z}_{t_s}\mathbf{b}_{t_s} = [f_{t_s}(x_1), \dots, f_{t_s}(x_k)]^\intercal$, with each $f_t(x)$ as in equation (1.2). To this end, we employ the classical Kalman filter (Kalman, 1960) under the model in (1.9)–(1.10) to obtain simple and closed–form recursive formulas for the filtering $p(\mathbf{b}_{t_s} \mid \log \mathbf{m}_{t_1}, \dots, \log \mathbf{m}_{t_s})$, predictive $p(\mathbf{b}_{t_{s+1}} \mid \log \mathbf{m}_{t_1}, \dots, \log \mathbf{m}_{t_s})$, and smoothing $p(\mathbf{b}_{t_s} \mid \log \mathbf{m}_{t_1}, \dots, \log \mathbf{m}_{t_n})$ distributions of $\mathbf{b}_{t_s}$, for $t_s = t_1, \dots, t_n$. Since $\mathbf{f}_{t_s} = \mathbf{Z}_{t_s}\mathbf{b}_{t_s}$, with $\mathbf{Z}_{t_s}$ known, the filtering, predictive and smoothing distributions for $\mathbf{f}_{t_s}$ can be directly derived from those of $\mathbf{b}_{t_s}$, for each $t_s = t_1, \dots, t_n$. Leveraging these results, we further develop in Section 1.3.2 a modern version of the celebrated Lee and Carter (1992) approach. Our proposed strategy provides future probabilistic projections of the B–splines coefficients via a simple random walk plus drift model where the drift component exploits the possibility of our formulation to explicitly learn not only mortality levels but also the corresponding rates of change. As illustrated in Section 1.4, this solution yields improved probabilistic forecasts of log–mortality rates relative to state–of–the–art alternatives.

The above formulas require the knowledge of the parameters $\sigma^2_{\overline{m}}$, σ^2_β, σ^2_a and λ. Due to the Gaussian form of model (1.9)–(1.10), these quantities can be estimated via maximization of the marginal likelihood for the Gaussian vectors $(\log \mathbf{m}_{t_1}, \dots, \log \mathbf{m}_{t_n})$, which is available in closed form, thereby allowing direct estimation; see Durbin and Koopman (2012, Ch. 7) and Chopin and Papaspiliopoulos (2020, Ch. 7) for further details on maximum likelihood estimation of system parameters in Gaussian state–space models, and refer to the R package KFAS (Helske, 2017) for an effective implementation. As discussed in the tutorial code at https://github.com/fpavone/BSP-mortality, in practice it is often recommended to add suitable penalizations and initialize the estimation procedure at different starting points, selecting as final estimate the one that yields the highest marginal likelihood, thereby avoiding possible issues associated with local modes. The covariance–function parameters $\boldsymbol{\gamma}_\beta$ and $\boldsymbol{\gamma}_a$ in (1.8) are instead fixed at default values which allow to induce suitable borrowing of information across spline coefficients; see Section 1.4 for details.

1.3.1 Filtering, Prediction and Smoothing

Due to the Gaussian form of model (1.9)–(1.10), the filtering, predictive and smoothing distributions for $\mathbf{b}_{t_s}$ are still multivariate normals $\mathrm{N}_{3p}(\boldsymbol{\mu}_{t_s|t_{1:s}}, \Sigma_{t_s|t_{1:s}})$, $\mathrm{N}_{3p}(\boldsymbol{\mu}_{t_{s+1}|t_{1:s}}, \Sigma_{t_{s+1}|t_{1:s}})$ and $\mathrm{N}_{3p}(\boldsymbol{\mu}_{t_s|t_{1:n}}, \Sigma_{t_s|t_{1:n}})$, respectively, with the mean vectors and covariance matrices that can be derived sequentially in time via recursive equations (Kalman, 1960). More specifically, let $\boldsymbol{\mu}_{t_s|t_{1:s-1}}$ and $\Sigma_{t_s|t_{1:s-1}}$ be the predictive mean vector and covariance matrix for $\mathbf{b}_{t_s}$ given the log–mortality rates observed until time t_{s-1}. Then, recalling, e.g., Durbin and Koopman (2012, Ch. 4), the filtering distribution for $\mathbf{b}_{t_s}$ is a $3p$–variate Gaussian with mean vector $\boldsymbol{\mu}_{t_s|t_{1:s}}$ and covariance matrix $\Sigma_{t_s|t_{1:s}}$ equal to

$$\begin{aligned}\boldsymbol{\mu}_{t_s|t_{1:s}} &= \boldsymbol{\mu}_{t_s|t_{1:s-1}} + \Sigma_{t_s|t_{1:s-1}}\mathbf{Z}_{t_s}^{\intercal}(\mathbf{Z}_{t_s}\Sigma_{t_s|t_{1:s-1}}\mathbf{Z}_{t_s}^{\intercal} + \mathbf{H}_{t_s})^{-1}(\log \mathbf{m}_{t_s} - \mathbf{Z}_{t_s}\boldsymbol{\mu}_{t_s|t_{1:s-1}}),\\ \Sigma_{t_s|t_{1:s}} &= \Sigma_{t_s|t_{1:s-1}} - \Sigma_{t_s|t_{1:s-1}}\mathbf{Z}_{t_s}^{\intercal}(\mathbf{Z}_{t_s}\Sigma_{t_s|t_{1:s-1}}\mathbf{Z}_{t_s}^{\intercal} + \mathbf{H}_{t_s})^{-1}\mathbf{Z}_{t_s}\Sigma_{t_s|t_{1:s-1}}.\end{aligned} \tag{1.11}$$

The above results are a direct consequence of the closure under conditioning of multivariate Gaussians and, when combined with (1.10), directly yield the mean vector $\boldsymbol{\mu}_{t_{s+1}|t_{1:s}}$ and covariance matrix $\Sigma_{t_{s+1}|t_{1:s}}$ for the predictive Gaussian distribution of $\mathbf{b}_{t_{s+1}}$. More specifically, leveraging the closure under linear combinations of multivariate Gaussians, we obtain

$$\begin{aligned}\boldsymbol{\mu}_{t_{s+1}|t_{1:s}} &= \mathbf{T}_{t_s}\boldsymbol{\mu}_{t_s|t_{1:s}},\\ \Sigma_{t_{s+1}|t_{1:s}} &= \mathbf{T}_{t_s}\Sigma_{t_s|t_{1:s}}\mathbf{T}_{t_s}^{\intercal} + \mathbf{Q}_{t_s}.\end{aligned} \tag{1.12}$$

Equations (1.11)–(1.12) provide simple closed–form formulas that allow to filter and forecast $\mathbf{b}_{t_s}$ recursively from time t_1 until time t_n by iterating among filtering and prediction steps. Recalling, e.g., Durbin and Koopman (2012, Ch. 4), such a recursion is initialized at t_1 from a $\mathrm{N}_{3p}(\boldsymbol{\mu}_{t_1|t_0}, \Sigma_{t_1|t_0})$. Although several starting strategies can be considered (e.g., Durbin and Koopman, 2012), we rely on a data–driven approach and fix $\boldsymbol{\mu}_{t_1|t_0}$ at a frequentist estimate based on a simple spline regression, while $\Sigma_{t_1|t_0}$ is set to $10\mathbf{I}_{3p}$ to induce a relatively diffuse initialization.

The forward recursions within (1.11)–(1.12) can be also combined with backward iterations to obtain the mean vector $\boldsymbol{\mu}_{t_s|t_{1:n}}$ and covariance matrix $\Sigma_{t_s|t_{1:n}}$ of the Gaussian smoothing distribution for each t_s, by iterating backward in time from t_n to t_1 via the expressions

$$\begin{aligned}\boldsymbol{\mu}_{t_s|t_{1:n}} &= \boldsymbol{\mu}_{t_s|t_{1:s-1}} + \Sigma_{t_s|t_{1:s-1}}\mathbf{r}_{t_{s-1}},\\ \Sigma_{t_s|t_{1:n}} &= \Sigma_{t_s|t_{1:s-1}} - \Sigma_{t_s|t_{1:s-1}}\mathbf{V}_{t_{s-1}}\Sigma_{t_s|t_{1:s-1}},\end{aligned} \tag{1.13}$$

where $\mathbf{r}_{t_{s-1}}$ and $\mathbf{V}_{t_{s-1}}$ are obtained from the backward equations $\mathbf{r}_{t_{s-1}} = \mathbf{Z}_{t_s}^{\intercal}(\mathbf{Z}_{t_s}\Sigma_{t_s|t_{1:s-1}}\mathbf{Z}_{t_s}^{\intercal} +$

$\mathbf{H}_{t_s})^{-1}(\log \mathbf{m}_{t_s} - \mathbf{Z}_{t_s}\boldsymbol{\mu}_{t_s|t_{1:s-1}}) + \mathbf{L}_{t_s}^\intercal \mathbf{r}_{t_s}$ and $\mathbf{V}_{t_{s-1}} = \mathbf{Z}_{t_s}^\intercal(\mathbf{Z}_{t_s}\Sigma_{t_s|t_{1:s-1}}\mathbf{Z}_{t_s}^\intercal + \mathbf{H}_{t_s})^{-1}\mathbf{Z}_{t_s} + \mathbf{L}_{t_s}^\intercal \mathbf{V}_{t_s}\mathbf{L}_{t_s}$, with $\mathbf{L}_{t_s} = \mathbf{T}_{t_s} - \mathbf{T}_{t_s}\Sigma_{t_s|t_{1:s-1}}\mathbf{Z}_{t_s}^\intercal(\mathbf{Z}_{t_s}\Sigma_{t_s|t_{1:s-1}}\mathbf{Z}_{t_s}^\intercal + \mathbf{H}_{t_s})^{-1}\mathbf{Z}_{t_s}$, and initialization $\mathbf{r}_{t_n} = \mathbf{0}$ and $\mathbf{V}_{t_n} = \mathbf{0}_{3p\times 3p}$; see Durbin and Koopman (2012, Ch. 4.4) for a detailed derivation of (1.13) leveraging again standard properties of multivariate Gaussian distributions.

Since $\mathbf{f}_{t_s} = \mathbf{Z}_{t_s}\mathbf{b}_{t_s}$, it immediately follows that the filtering, predictive and smoothing distributions for the log–mortality rates mean function $f_{t_s}(x)$, $x = x_1, \dots, x_k$, can be directly obtained from those of $\mathbf{b}_{t_s}$ in equations (1.11), (1.12) and (1.13), respectively. This implies

$$\begin{aligned}
(\mathbf{f}_{t_s} \mid \log \mathbf{m}_{t_1}, \dots, \log \mathbf{m}_{t_s}) &\sim \mathrm{N}_k(\mathbf{Z}_{t_s}\boldsymbol{\mu}_{t_s|t_{1:s}}, \mathbf{Z}_{t_s}\Sigma_{t_s|t_{1:s}}\mathbf{Z}_{t_s}^\intercal), \\
(\mathbf{f}_{t_{s+1}} \mid \log \mathbf{m}_{t_1}, \dots, \log \mathbf{m}_{t_s}) &\sim \mathrm{N}_k(\mathbf{Z}_{t_{s+1}}\boldsymbol{\mu}_{t_{s+1}|t_{1:s}}, \mathbf{Z}_{t_{s+1}}\Sigma_{t_{s+1}|t_{1:s}}\mathbf{Z}_{t_{s+1}}^\intercal), \\
(\mathbf{f}_{t_s} \mid \log \mathbf{m}_{t_1}, \dots, \log \mathbf{m}_{t_n}) &\sim \mathrm{N}_k(\mathbf{Z}_{t_s}\boldsymbol{\mu}_{t_s|t_{1:n}}, \mathbf{Z}_{t_s}\Sigma_{t_s|t_{1:n}}\mathbf{Z}_{t_s}^\intercal),
\end{aligned} \tag{1.14}$$

for each $t_s = t_1, \dots, t_n$. These results yield closed–form Gaussian distributions that facilitate probabilistic inference on mortality levels and the corresponding rates of change across ages and periods, beyond currently–available analyses. Moreover, as clarified in Section 1.3.2, this perspective allows to improve point forecasts and predictive intervals of future log–mortality rates at different times. Crucially, these quantities can be readily computed via user–friendly and optimized R packages for state–space models. For example, the filtering, predictive and smoothing distributions in (1.11)–(1.14) can be obtained via the KFS function from the KFAS package (Helske, 2017), after specifying model (1.9)–(1.10) via the function SSMODEL.

1.3.2 Forecasting

As is clear from (1.9), the results in Section 1.3.1 are useful not only for inference, but also to obtain probabilistic forecasts for the vector of future log–mortality rates $\log \mathbf{m}_{t_{s^*}} = (\log m_{x_1,t_{s^*}}, \dots, \log m_{x_k,t_{s^*}})^\intercal$ for $t_{s^*} = t_{n+1}, \dots, t_{n+n^*}$, from the predictive distribution of $\mathbf{f}_{t_s}$. While such an approach is expected to yield accurate results for short–term forecasts, the quantitative studies in Section 1.4 suggest that the inherent local adaptivity of the model developed in Section 1.2 might yield less stable and shock–robust mortality projections for those medium–to–large time horizons that are of interest in demography.

To address this aspect and deliver accurate probabilistic forecasts at different time horizons, we derive a simple strategy that combines the proposed B–spline construction in Section 1.2 with the celebrated random walk plus drift projections by Lee and Carter (1992), in order to obtain an improvement in log–mortality rates forecasts relative to state–of–the–art competitors. In fact, despite its simplicity, the random walk plus drift construction is empirically supported by the globally–linear trend of log–mortality

rates in medium–to–large time horizons, which make Lee and Carter (1992) projections still competitive. Nonetheless, as mentioned in e.g., Hyndman and Ullah (2007), these forecasts still rely on an age–period bilinear formulation that fails to account for heterogeneity in age–specific mortality dynamics, and, in addition, the estimation of the drift component is not robust to mortality shocks. The model we propose in Section 1.2 accounts for both effects, thus suggesting that incorporating the random walk plus drift forecasting strategy within the proposed B–spline process with locally–adaptive dynamic coefficients would yield improvements over the original Lee and Carter (1992) strategy and, as clarified in Section 1.4, also with respect to other state–of–the–art methods, both in terms of point forecasts and quality of the predictive intervals.

Consistent with the above discussion, and recalling equations (1.2) and (1.9), we obtain point forecasts for the future log–mortality rates $\log \mathbf{m}_{t_{s^*}}$, via

$$\hat{\mathbf{f}}_{t_{s^*}} = [\mathbf{g}_1, \dots, \mathbf{g}_p]\hat{\boldsymbol{\beta}}_{t_{s^*}} \quad \text{for } t_{s^*} = t_{n+1}, \dots, t_{n+n^*}, \tag{1.15}$$

where $[\mathbf{g}_1, \dots, \mathbf{g}_p]$ denotes the $(k \times p)$–dimensional B–splines matrix having columns $\mathbf{g}_j = [g_j(x_1), \dots, g_j(x_k)]^\intercal$, $j = 1, \dots, p$, whereas $\hat{\boldsymbol{\beta}}_{t_{s^*}} = [\hat{\beta}_1(t_{s^*}), \dots, \hat{\beta}_p(t_{s^*})]^\intercal$ is the $(p \times 1)$–dimensional vector comprising the forecasts for the B–splines coefficients at $t_{s^*} > t_n$ from the p–variate random walk plus drift model

$$\begin{aligned} \boldsymbol{\beta}_{t_{s^*+1}} &= \boldsymbol{\beta}_{t_{s^*}} + \hat{\lambda}\delta_{s^*}\boldsymbol{\Delta}_{t_{s^*}} + \boldsymbol{\omega}_{t_{s^*}}, & \boldsymbol{\omega}_{t_{s^*}} &\overset{\text{i.i.d}}{\sim} \mathrm{N}_p(\mathbf{0}, \mathbf{W}), \\ \boldsymbol{\Delta}_{t_{s^*+1}} &= \boldsymbol{\Delta}_{t_{s^*}} + \boldsymbol{\epsilon}_{t_{s^*}} & \boldsymbol{\epsilon}_{t_{s^*}} &\overset{\text{i.i.d}}{\sim} \mathrm{N}_p(\mathbf{0}, \sigma^2_\Delta \mathbf{I}_p). \end{aligned} \tag{1.16}$$

In (1.16), each $\mathbf{W}_{[j,l]}$ is set equal to $\sigma^2_\omega \rho_{\beta[j,l]}$, for $j = 1, \dots, p$ and $l = 1, \dots, p$, with $\rho_{\beta[j,l]}$ as in (1.8), $\hat{\lambda}$ is the maximum marginal likelihood estimate of λ discussed in Section 1.3, whereas the starting values $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\Delta}}$ for $\boldsymbol{\beta}_{t_n}$ and $\boldsymbol{\Delta}_{t_n} = [\Delta_{1,t_n}, \dots, \Delta_{p,t_n}]^\intercal$, respectively, are defined in order to ensure flexible, yet shock–robust, point forecasts at each $t_{s^*} = t_{n+1}, \dots, t_{n+n^*}$. More specifically, $\hat{\boldsymbol{\beta}}$ corresponds to the mean of the smoothing distribution for $\boldsymbol{\beta}_{t_n}$, whereas $\hat{\boldsymbol{\Delta}}$ is defined as the median of the estimates of $\partial\beta_j(t)/\partial t$, over the last 25 years $t_n, \dots, t_{n-24}$ computed under the smoothing distribution in (1.13), for every $j = 1, \dots, p$. Since the smoothing distribution in (1.13) is Gaussian, these estimates coincide with the elements having position $2 + 3(j-1)$ in $\boldsymbol{\mu}_{t_s|t_{1:n}}$, for each $j = 1, \dots, p$ and $t_s = t_n, \dots, t_{n-24}$. Rather than projecting forward in time a single global dynamic component, as in Lee and Carter (1992), strategy (1.15)–(1.16) gains accuracy by extrapolating multiple time dynamics corresponding to different age classes, while relying on a natural initialization for the drift terms which leverages the ability of the proposed model to explicitly learn dynamics also in the first order derivatives quantifying rates of change in mortality levels. To ensure robustness to shocks while adapting to the most recent

globally–linear trend dynamics, the starting drift term $\hat{\boldsymbol{\Delta}}$ is set to the median, rather than the mean, of these estimates over the last 25 years, which was found to be a robust default time horizon in the application to multiple countries in Section 1.4. This is in line with similar results in Lee and Carter (1992) on the time horizon to condition on.

Although the above strategy yields improved point forecasts, the derivation of effective predictive intervals requires accurate estimates of σ^2_ω and σ^2_Δ in (1.16), along with the variance parameter σ^2_ψ in the observation equation which yields the forecasted $\hat{\mathbf{f}}_{t_{s^*}}$ in (1.15), namely

$$\log \mathbf{m}_{t_{s^*}} = [\mathbf{g}_1, \dots, \mathbf{g}_p]\boldsymbol{\beta}_{t_{s^*}} + \boldsymbol{\nu}_{t_{s^*}}, \qquad \boldsymbol{\nu}_{t_{s^*}} \sim \mathrm{N}_k(\mathbf{0}, \sigma^2_\psi \mathbf{I}_k). \tag{1.17}$$

Consistent with the strategy adopted for deriving the point forecasts, these three variances are obtained via maximum marginal likelihood under model (1.16)–(1.17) applied to data from t_{n-24} until t_n. To suitably connect model (1.9)–(1.10) with the simpler formulation in (1.16)–(1.17), $\boldsymbol{\Delta}_{t_{n-24}}$ is initialized at $\mathrm{N}_p(\hat{\boldsymbol{\mu}}_{\boldsymbol{\Delta}}, \mathrm{diag}(\hat{\boldsymbol{\sigma}}^2_{\Delta}))$, where $\hat{\boldsymbol{\mu}}_{\boldsymbol{\Delta}}$ is the median of the estimated $\partial\beta_j(t)/\partial t$, for each $j = 1, \dots, p$, under the smoothing distribution in (1.13), over the 25 years preceding t_{n-24}, while the generic entry $\hat{\sigma}^2_{\Delta_j}$ in $\hat{\boldsymbol{\sigma}}^2_{\Delta}$ corresponds to the sample variance of the median estimate $\hat{\mu}_{\Delta_j}$, for $j = 1, \dots, p$, computed via a set of simulations from the smoothing distribution. The initial vector $\boldsymbol{\beta}_{t_{n-24}}$ is instead assumed to follow the p–variate Gaussian distribution where the mean vector is obtained from the one–step–ahed projection under (1.16) of the smoothing estimate for $\boldsymbol{\beta}_{t_{n-25}}$ provided by model (1.9)–(1.10), whereas the covariance matrix coincides with that of the predictive distribution at t_{n-25} under model (1.9)–(1.10).

From a practical perspective, the above strategies can be still implemented via standard R libraries for time series analysis, and, as previously–discussed, are reminiscent of the two–step approach by Lee and Carter (1992), which relies on an in–sample estimate of the global time–specific effect and then fits, for future projections, a random walk plus drift model on the subset of these time effects corresponding to a suitably–defined most recent window.

1.4 Learning and Forecasting of Mortality Across Countries

In order to quantify the improvements provided by the novel BSP developed in Sections 1.2–1.3, we consider extensive analyses and performance comparisons with a main focus on the gender–specific age–period log–mortality rates for the countries discussed in Section 1.1.1, across a wide time horizon that spans from 1933 until either 2019 or

2020 depending on data availability in the Human Mortality Database at https://www.mortality.org/.

In modeling the age–period log–mortality rates for the four countries analyzed we follow common practice in demographic studies (e.g., Haberman and Renshaw, 2011; Currie, 2016; Camarda, 2019) and consider a separate analysis for each combination of gender–country, leading to a total of eight different implementations of the BSP model in Sections 1.2–1.3 and of selected state–of–the–art competitors. More specifically, for each combination of gender and country we study the age–period log–mortality rates via the Gaussian state–space approximation in (1.9)–(1.10) of the BSP model defined in Section 1.2.1, employing the $p = 20$ B–spline bases $g_1(x), \dots, g_{20}(x)$ illustrated in Figure 1.3, and considering Matérn covariance functions for $\mathcal{K}(\bar{x}_j, \bar{x}_l; \boldsymbol{\gamma}_\beta)$ and $\mathcal{K}(\bar{x}_j, \bar{x}_l; \boldsymbol{\gamma}_a)$ in (1.8) (e.g., Rasmussen and Williams, 2006, Ch. 4). The specification of a total of $p = 20$ bases over the observed age range $x_1 = 0, \dots, x_{101} = 100$ is motivated by a similar choice for the age dimension in the bivariate B–splines construction of Camarda (2019). Consistent with the graphical evidence in Figure 1.1, these B–splines are more dense at early and late ages to achieve increased flexibility in capturing local dynamic variations for such classes. For the same reasons, the first B–spline $g_1(x)$ is the only one that is active at $x_1 = 0$ since mortality at age 0 is known to display peculiar patterns relative to those from age 1 onward, thereby requiring increased flexibility relative to the other classes (Camarda, 2019). The Matérn covariance functions parameters $\boldsymbol{\gamma}_\beta$ are instead set at $(0.5, 1)$ to induce local borrowing of information only across close ages, whereas no correlation is enforced on the instantaneous means to increase the flexibility in modeling shocks affecting only specific age classes. Although these covariance parameters could be estimated, together with $\sigma^2_{\overline{m}}$, σ^2_β, σ^2_a and λ, via maximum marginal likelihood under model (1.9)–(1.10), as clarified in Table 1.1 and in Figures 1.4–1.5, the suggested settings provide robust default choices to accurately learn and forecast several mortality patterns across different countries. In fact, moderate changes in the Matérn covariance parameters as well as in the number and location of the B–spline bases did not change the final conclusions. Notice also that, although the squared exponential covariance function provides another routinely–implemented alternative, such a function can be recovered as a special case of the Matérn one which, therefore, yields a more general class with a higher degree of flexibility.

As a first assessment, we evaluate in Table 1.1 the performance in point forecasting of the BSP formulation proposed in Sections 1.2–1.3, and quantify the improvements relative to the state–of–the–art competitors discussed in Section 1.1. These include the classical Lee and Carter (1992) (LC) and age–period–cohort (APC) (Holford, 1983; Osmond, 1985) models, along with the subsequent developments and extensions in Renshaw and

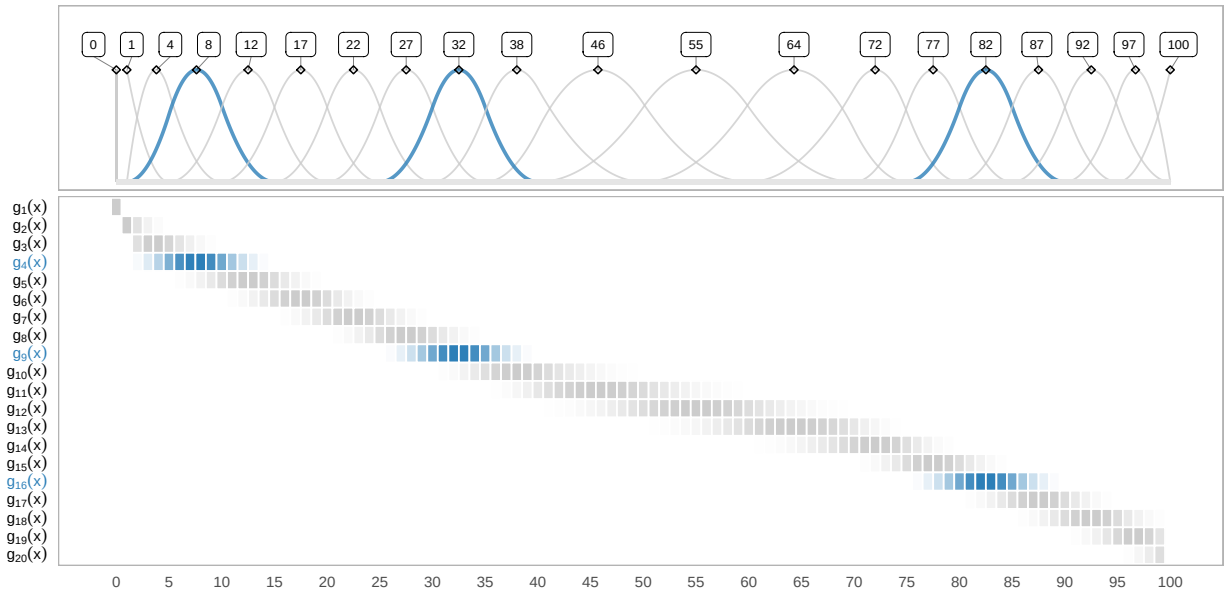


Figure 1.3: Graphical representation of the $p = 20$ selected B–spline bases $g_1(x), \dots, g_{20}(x)$ along with a heatmap clarifying the intensity of each spline in the corresponding age range. The number $\bar{x}_j$ associated to each spline $g_j(x)$ in the top panel denotes the age at which such a spline takes its maximum value, for each $j = 1, \dots, 20$. For illustrative purposes, three bases and the corresponding active range of ages are highlighted in blue.

Haberman (2006) (RH), Cairns et al. (2006) (CBD), Hyndman and Ullah (2007) (HU), Plat (2009) (PLAT) and Camarda (2019) (CP), which currently represent the leading methods in mortality forecasting. As illustrated in the code available at https://github.com/fpavone/BSP-mortality, these models can be readily implemented via standard R functions in the packages StMoMo (Villegas et al., 2018), demography (Hyndman et al., 2014) and in the R code within the supplementary materials of Camarda (2019). Recalling Section 1.3, parameter estimation, inference and forecasting under the Gaussian state–space formulation of the proposed BSP approach can be instead effectively implemented via the R package KFAS (Helske, 2017).

Table 1.1 summarizes the performance in point forecasting of the above strategies over different time horizons, ranging from 1–step–ahead to 10–step–ahead. For the eight gender–country combinations, these forecasts are obtained by sequentially fitting each model on the observed age–period log–mortality rates from 1933 up until a last year ranging from 1990 to 2010, and then predicting, for each of these final years from 1990 to 2010, the age–period log–mortality rates in the subsequent ten years. This produces, under each model and step–ahead predictive horizon, a total $2 \times 101 \times 21$ forecasts per country — except for Italy whose data are available only until 2019 — which correspond to the different combinations of gender, ages and last observation time, thereby providing a large sample of predictions to accurately compare the different methods. Table 1.1

displays the overall median of the absolute differences between these forecasts and the actual observed log–mortality rates, across countries, gender, ages and the last observation times. Results provide empirical evidence for the improved forecasting accuracy of the proposed BSP, which outperforms all the state–of–the–art alternatives for every predictive horizon. As expected, CP (Camarda, 2019) and HU (Hyndman and Ullah, 2007) are the most competitive alternatives. Recalling Section 1.1, also these strategies rely on a flexible basis expansion, but are not as effective as the proposed BSP in incorporating all the core structures of age–period mortality surfaces, thereby allowing our procedure to further improve forecasting accuracy. Notice that, although the magnitude of these improvements is not always remarkable, the BSP approach remains systematically more accurate than all the methods considered, both in terms of medians and the two quartiles. Such a finding was further confirmed in additional studies of other countries in the HUMAN MORTALITY DATABASE (i.e., Czech Republic, Denmark and France) and when comparing the forecasting performance of the different methods via the mean squared error, rather than the median of the absolute error. The latter measure is preferred in Table 1.1 since it provides a more robust and direct measure of the actual distance between the forecasted and observed mortality rates. We shall also emphasize that, in these contexts, even a small reduction in the predictive errors for the log–mortality rates can have a major impact in population forecasts since, as is clear from equation (1.1), such rates are multiplied by the central exposed to risk E_{xt} when modeling the total death counts d_{xt}, with E_{xt} in the order of tens–to–hundreds of thousands in common population analyses at the country level. This reasoning applies also to other demographic measures derived as a function of the mortality rates, such as, for example, the life expectancy at birth whose forecasts can be directly obtained via the R package DEMOGRAPHY from those produced for m_{xt}. Also in this case, BSP was still found to outperform all the state–of–the–art competitors and almost halved the 10–step–ahead predictive errors of both CP and HU. Finally, we want to highlight that our model has shown consistent forecasting accuracy across countries when computing the country–specific median absolute errors. The results, which we do not report here, are in accordance with the values shown in Table 1.1.

From a computational perspective, all the methods analyzed in Table 1.1, including the proposed BSP strategy, facilitate tractable and scalable implementations which yield runtimes for estimation, inference and forecasting always below one minute. This is several orders of magnitude lower than the yearly time scale at which mortality data are typically analyzed, thereby providing effective solutions for rapid updating of inferences and forecasts.

To further clarify the major advantages of the BSP construction, we also considered

Step ahead	BSP	CP	HU	PLAT	RH	APC	LC	CBD
1	**0.032** [0.01, 0.07]	0.033 [0.01, 0.07]	0.038 [0.02, 0.09]	0.097	0.067	0.126	0.107	0.141
2	**0.037** [0.02, 0.08]	0.040 [0.02, 0.08]	0.047 [0.02, 0.09]	0.104	0.075	0.137	0.114	0.149
3	**0.044** [0.02, 0.09]	0.048 [0.02, 0.10]	0.056 [0.03, 0.11]	0.113	0.085	0.150	0.123	0.159
4	**0.050** [0.02, 0.10]	0.057 [0.03, 0.11]	0.064 [0.03, 0.12]	0.121	0.094	0.161	0.129	0.171
5	**0.056** [0.03, 0.11]	0.066 [0.03, 0.12]	0.072 [0.03, 0.14]	0.131	0.103	0.175	0.138	0.182
6	**0.063** [0.03, 0.12]	0.073 [0.03, 0.13]	0.080 [0.04, 0.15]	0.140	0.113	0.190	0.145	0.193
7	**0.070** [0.03, 0.13]	0.081 [0.04, 0.15]	0.088 [0.04, 0.16]	0.150	0.127	0.206	0.154	0.204
8	**0.076** [0.03, 0.14]	0.088 [0.04, 0.16]	0.094 [0.04, 0.18]	0.160	0.141	0.224	0.162	0.218
9	**0.083** [0.04, 0.16]	0.095 [0.04, 0.17]	0.102 [0.05, 0.19]	0.168	0.155	0.237	0.172	0.241
10	**0.093** [0.04, 0.17]	0.105 [0.05, 0.19]	0.110 [0.05, 0.21]	0.179	0.174	0.260	0.180	0.253

Table 1.1: For the eight methods under analysis and ten predictive horizons, overall median of the absolute difference between the forecasted and observed log–mortality rates computed from all the country–gender–age–year combinations. Bold values denote the best performance for each predictive horizon, whereas the gray column corresponds to the proposed B–spline process with locally–adaptive dynamic coefficients. For the three top performing methods, the first and third quartiles of the absolute differences are also reported within brackets.

predictive comparisons against direct implementations of the simpler building–blocks underlying the proposed formulation and forecasting approach. More specifically, instead of relying on the strategy outlined in Section 1.3.2, we considered forecasts obtained either under the predictive distribution (1.12) of the original BSP formulation in (1.9)–(1.10), or from the direct use of separate nested Gaussian processes for the trajectories $f_t(x)$ of every age $x \in \mathcal{X}$ rather than employing the more structured formulation proposed within equations (1.2)–(1.4). Focusing again on a time horizon ranging form 1–step–ahead to 10–step–ahead forecasts, the overall medians of the absolute differences between the forecasted and observed log–mortality rates were $[0.035, 0.045, 0.057, 0.067, 0.078, 0.091, 0.103, 0.115, 0.128, 0.144]$ for the first alternative strategy and $[0.058, 0.109, 0.179, 0.264, 0.370, 0.490, 0.630, 0.785, 0.958, 1.149]$ for the second. Comparing these results with those in the first column of Table 1.1 clarifies the key advantages of the proposed BSP construction that achieves improved predictive accuracy by carefully borrowing information across ages via a structured B–spline representation with dependence across the dynamic coefficients, which is subsequently leveraged to develop the parsimonious, yet effective, forecasting strategy outlined in Section 1.3.2.

Let us conclude the analysis of forecasting performance by assessing the calibration of the predictive intervals under the different methods considered in Table 1.1. To this end, Table 1.2 displays the relative proportion, computed from all the different combinations of country–gender–age–year, of the 95% predictive intervals which contain the observed log–mortality rates. Also in this setting, the BSP intervals computed under the methods illustrated in Section 1.3.2 achieve improved overall calibration relative to those obtained

Step ahead	BSP	CP	HU	PLAT	RH	APC	LC	CBD
1	**0.955**	0.497	0.980	0.428	0.377	0.309	0.230	0.328
2	**0.955**	0.466	0.982	0.493	0.441	0.312	0.295	0.395
3	**0.953**	0.436	0.982	0.513	0.457	0.370	0.330	0.469
4	**0.952**	0.415	0.984	0.521	0.464	0.406	0.364	0.498
5	**0.950**	0.388	0.984	0.532	0.464	0.421	0.384	0.466
6	**0.948**	0.357	0.982	0.511	0.468	0.403	0.396	0.513
7	**0.948**	0.331	0.979	0.522	0.453	0.424	0.406	0.505
8	**0.946**	0.311	0.977	0.551	0.436	0.413	0.413	0.506
9	**0.945**	0.300	0.976	0.565	0.421	0.430	0.429	0.503
10	**0.937**	0.286	0.972	0.570	0.404	0.422	0.429	0.512

Table 1.2: For the eight methods under analysis and ten predictive horizons, relative proportion, computed from all the country–gender–age–year combinations, of the 95% predictive intervals containing the observed log–mortality rates. Bold values denote the best performance for each predictive horizon, whereas the gray column corresponds to the proposed B–spline process with locally–adaptive dynamic coefficients.

under the competing strategies. Comparing results in Table 1.2 with those in Table 1.1, the poor performance of PLAT, RH, APC, LC and CBD is mainly attributable to the bias in the point forecasts at which such intervals are centered, whereas CP suffers from an underestimation of the predictive variance, possibly due to challenges in the implementation of the employed bootstrap strategy within a time–series context. As illustrated in Table 1.2, HU is the only competitive strategy, although it exhibits an over–coverage tendency with intervals having a similar length to those obtained under the proposed BSP construction. We shall emphasize that, when stratifying by age, the calibration of the BSP intervals is generally less accurate at younger ages than older ones, thus motivating additional future refinements.

The improvements in forecasting performance of BSP motivate additional analyses and country comparisons of age–period mortality surfaces, which are further facilitated by the interpretable construction of the proposed model in Sections 1.2–1.3. In fact, as illustrated in Figures 1.4–1.5, BSP allows to formally study and compare changes in mortality patterns across years and specific ages via inference on the location and variability of the smoothing distribution for the coefficients of the splines active in those age classes. These selected trajectories are displayed in Figure 1.4, along with the corresponding first derivatives, and highlight interesting differences across countries in the dynamic evolution of age–specific mortality rates. For instance, in the first row of Figure 1.4, BSP learns a peculiar trajectory for infant mortality in Italy, characterized by a structural break soon after the World War II which leads to a faster decay in infant mortality with respect to other countries. This remarkable change is aligned with the so–called *Italian miracle*, a phase of rapid economic growth and improved life conditions after the World War II (e.g., Ginsborg, 1990), progressively bringing infant mortality in Italy to even lower levels than

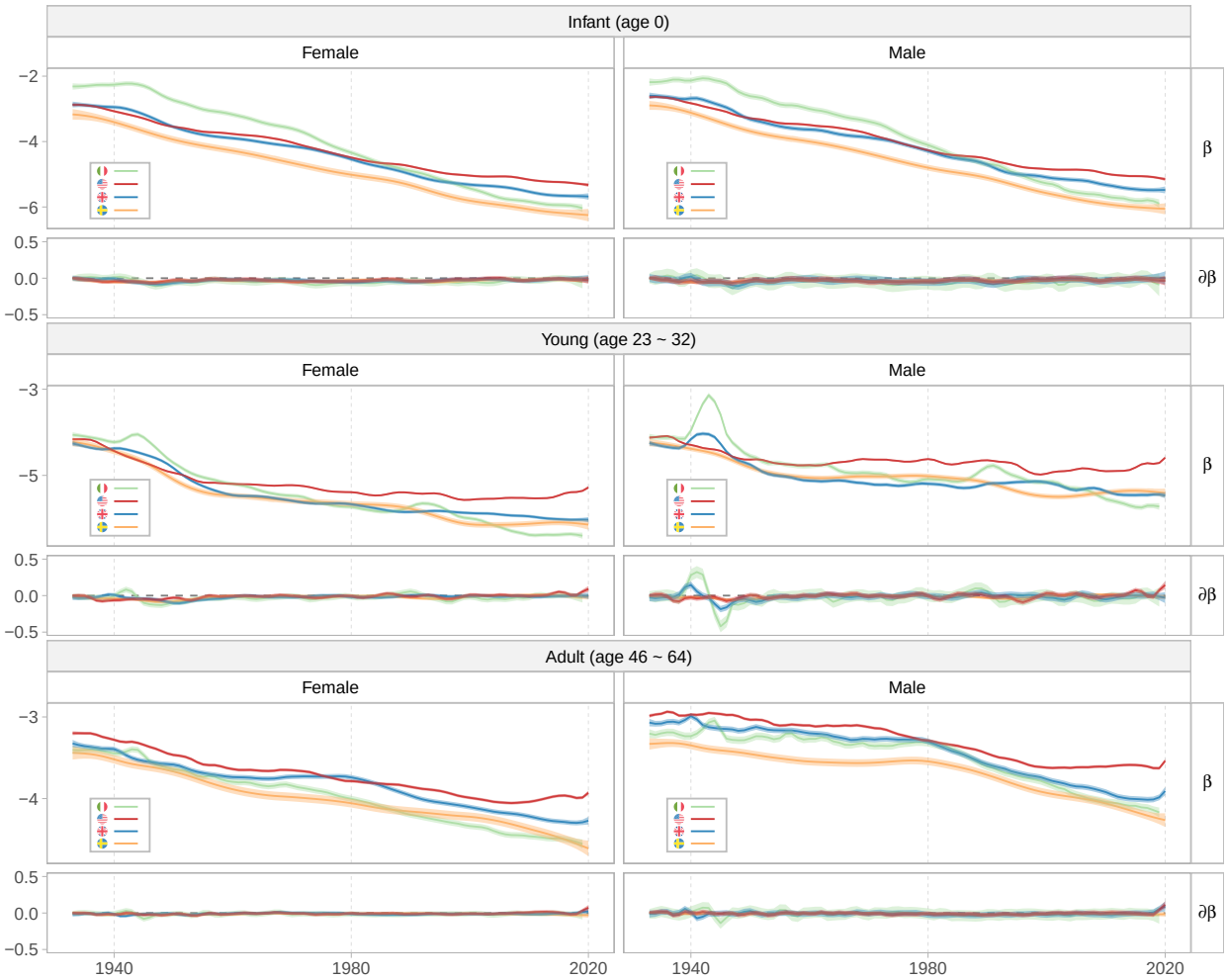


Figure 1.4: For females (left) and males (right), and the four countries under analysis, time trajectories of three representative B–splines coefficients along with the corresponding first derivatives, as obtained from the smoothing distribution under the proposed BSP model. The straight lines correspond to the trajectories of the means, whereas shaded areas denote the pointwise 95% credible intervals.

those registered in countries such as the United Kingdom and United States. The local adaptivity of BSP can be instead appreciated in the second row of Figure 1.4, where the proposed model learns the rapid mortality increment corresponding to the World War II, which, as expected, is mainly evident for Italian males, albeit visibile also for males in the United Kingdom. Notice that, since most of the United States military deaths happened abroad, these counts do not contribute to US mortality as recorded in the HUMAN MORTALITY DATABASE. In the second row of Figure 1.4, BSP also learns an evident excess–mortality peak in Italy for both young males and females in the late '80s and early '90s. This provides quantitative evidence for the rapid and severe combined effect of AIDS, car accidents and overdoses in Italy during those years for young age classes (e.g., Conti et al., 1994, 1997). Despite this shock, Italy displays a generally decreasing trend in the splines coefficients associated to young age classes, which interestingly departs from the general stagnation, or even increasing trend, that BSP learns for the other three countries, especially in the last two decades. This is particularly remarkable in the United States which display peculiar mortality patterns characterized by slower mortality decrements or even increments, mainly evident since the '80s, for all the three age classes analyzed in Figure 1.4. These quantitive findings further support a number of studies on the recent US mortality crisis as a consequence of specific disparities and vulnerabilities associated with young and adult age classes (Ho and Preston, 2010; Woolf and Schoomaker, 2019; Glei, 2022; Preston and Vierboom, 2021; Case and Deaton, 2021). As is clear from Figure 1.4, all the smoothing distributions analyzed are characterized by limited uncertainty, thus supporting the reliability of these findings.

Notably, the aforementioned patterns are also associated with differences in the rates of change of mortality levels during COVID–19, as inferred from the analysis of the first derivatives of the three splines coefficients in Figure 1.4 for year 2020. The ability of BSP to explicitly model and quantify uncertainty also in these rates of change in mortality trends crucially allows to learn a mortality shock during COVID–19 in young age classes only for the United States and not for the other countries under analysis; see the panels $\partial\beta$ in the second row of Figure 1.4. These findings are further expanded in Figure 1.5 where the focus is on the smoothing distribution of the differences between each spline coefficient in year 2020 and its average in the previous five years, for Sweden, United Kingdom and United States; data for Italy in year 2020 are not yet available in the HUMAN MORTALITY DATABASE at https://www.mortality.org/. Consistent with the discussion of Figure 1.4, BSP infers a noticeably–high excess mortality in the United States, for both females and males, which is surprisingly visibile from very young age classes onward, and whose magnitude is much higher than in the United Kingdom and Sweden. This key finding further corroborates recent studies on the association between COVID–19 effects

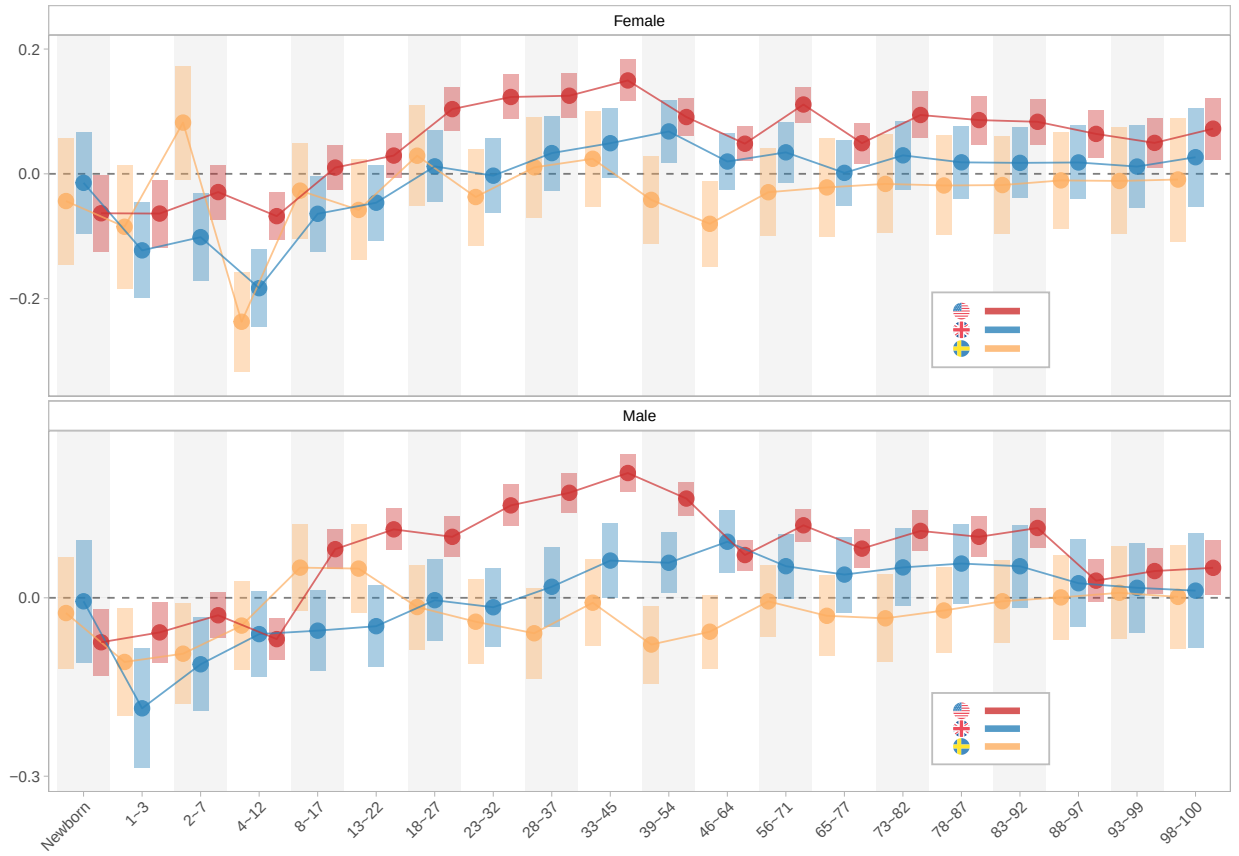


Figure 1.5: For females and males, means (colored dots) and 95% credible intervals (colored boxes) of the smoothing distribution for the difference between the spline coefficients in year 2020 and the corresponding average over 2014–2019, for United States, United Kingdom and Sweden. Data for Italy in 2020 are not yet available in the Human Mortality Database. This representation provides a summary of excess mortality in 2020.

and the peculiar pre–existing us disparities and vulnerabilities, especially in relation to risk factors (e.g., Wiemers et al., 2020). Despite the less stringent policies adopted in Sweden, the covid–19 mortality shock for such a country is less evident than the one registered in the United Kingdom and United States. It shall be emphasized that also Sweden experienced an excess mortality during the first and second wave of the covid–19 pandemic (e.g., Juul et al., 2022). However, when aggregating all–causes mortality at a yearly scale, such increments become less visibile and systematic, pointing toward a possible mortality displacement effect (Juul et al., 2022), also known as *harvesting*; namely a phase of excess deaths followed by a mortality deficit that has a balancing effect when aggregating at a larger time scale.

1.5 Conclusion and Future Research Directions

We propose a novel B–spline process with locally–adaptive dynamic coefficients for accurate learning and forecasting of mortality patterns across ages and periods. Such a process decomposes the age–period mortality surface as a flexible, yet interpretable, function of age, and crucially treats the dynamics of this function across periods via a suitable stochastic process of time that explicitly incorporates the core structures of mortality evolution through a set of stochastic differential equations. This allows to (i) incorporate and learn differences in the time patterns of mortality across age classes, while borrowing information between close ages, (ii) explicitly infer and project not only age–specific mortality trends, but also the corresponding rates of change, (iii) characterize dynamics that fluctuate between periods of rapid and slow variation, (iv) devise simple and accurate forecasting strategies for log–mortality rates which are both flexible and shock–robust, (v) develop computationally–efficient methods for filtering, smoothing and prediction of mortality patterns via closed–form Kalman filter recursions.

To the best of our knowledge, none of the solutions currently available in the literature accounts for all the above properties within a single formulation. In fact, as illustrated in the application in Section 1.4, the proposed model generally improves forecasting performance and crucially expands the set of findings which can be obtained from the analysis of age–period mortality data. This can open new avenues to formally compare differences in mortality patterns across ages, countries and years, while quantifying possible heterogeneities in the rate of change of mortality and in the impact of shocks, such as the recent COVID–19 pandemic for which we infer notable differences across countries.

Besides providing an important contribution to the literature on mortality modeling, the proposed formulation also motivates several future advancements. A relevant direction is to extend the B–spline process in Section 1.2 for joint modeling of multiple populations, possibly from high, middle and low income countries. Although the HUMAN MORTALITY DATABASE has data only for the first group, such an extension can be accomplished within our formulation by considering a mixture of B–spline processes that would further allow to cluster countries characterized by similar age–period mortality patterns. This facilitates borrowing of information for countries with low population size or studies at a local level, and incorporates improved coherency in mortality forecasts, an important aspect in recent multi–population studies (e.g., Li and Lee, 2005; Wen et al., 2021; Wang et al., 2022b). Alternatively, it would be of interest to specify country–specific B–spline processes with locally–adaptive dynamic coefficients and then induce dependence among such processes via a suitable graphical model (Lauritzen, 1996), thus allowing inference on conditional independence structures in age–period mortality dynamics among different countries, while borrowing information to improve inference and forecasting. This

can be particularly useful also in joint modeling of male and female mortality.

The above directions are also of interest when the focus is on joint modeling of mortality patterns for different causes–of–death, rather than countries (e.g., Kjærgaard et al., 2019).

1.6 Proofs of Propositions

Proof of Proposition 1. The proof adapts similar derivations considered by Liang et al. (2014) in the context of binomial logistic–normals. More specifically, under the model in (1.1), it holds that $d_{xt} = \sum_{i=1}^{\mathrm{E}_{xt}} y_{i,xt}$, where $y_{i,xt}$, $i = 1, \ldots, \mathrm{E}_{xt}$, denote auxiliary variables such that $(y_{i,xt} \mid \overline{m}_{xt}) \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\overline{m}_{xt})$. Hence, for fixed $\overline{m}_{xt}$, by the weak law of large numbers we have that $d_{xt}/\mathrm{E}_{xt} \to \overline{m}_{xt}$ in probability, as $\mathrm{E}_{xt} \to \infty$. This also implies convergence in distribution, i.e., $\lim_{\mathrm{E}_{xt}\to\infty} \text{pr}(d_{xt}/\mathrm{E}_{xt} \leq u \mid \overline{m}_{xt}) = \mathbb{1}(\overline{m}_{xt} \leq u)$. Leveraging this result and applying the dominated convergence theorem, it follows that

$$\begin{aligned}
\lim_{\mathrm{E}_{xt}\to\infty} \text{pr}(d_{xt}/\mathrm{E}_{xt} \leq u) &= \lim_{\mathrm{E}_{xt}\to\infty} \int_0^\infty \text{pr}(d_{xt}/\mathrm{E}_{xt} \leq u \mid \overline{m}_{xt}) p(\overline{m}_{xt}) \partial \overline{m}_{xt} \\
&= \int_0^\infty \lim_{\mathrm{E}_{xt}\to\infty} \text{pr}(d_{xt}/\mathrm{E}_{xt} \leq u \mid \overline{m}_{xt}) p(\overline{m}_{xt}) \partial \overline{m}_{xt} \\
&= \int_0^\infty \mathbb{1}(\overline{m}_{xt} \leq u) p(\overline{m}_{xt}) \partial \overline{m}_{xt} = \text{pr}(\overline{m}_{xt} \leq u).
\end{aligned}$$

Hence, d_{xt}/E_{xt} converges in distribution to the assumed log–normal for $\overline{m}_{xt}$ in equation (1.1) and, as a direct consequence of the continuous mapping theorem, it follows that $\log m_{xt} = \log(d_{xt}/\mathrm{E}_{xt}) \to \text{N}(f_t(x), \sigma^2_{\overline{m}})$ in distribution, as $\mathrm{E}_{xt} \to \infty$, for any $x \in \mathcal{X}$ and $t \in \mathcal{T}$. □

Proof of Proposition 2. The proof follows directly from the results in Appendix A.6 of Zhu and Dunson (2013) after replacing $\mathbf{C}$ with $\lambda(\mathbf{I}_p \otimes \mathbf{C})$ and $\mathbf{D}$ with $(\mathbf{I}_p \otimes \mathbf{D})\boldsymbol{\Omega}^{1/2}$. □

Chapter 2

Phylogenetic Latent Position Models for Populations of Networks

JOINT WORK WITH DANIELE DURANTE AND ROBIN RYDER

2.1 Introduction

Networks are typically used in statistics to represent relational data. These consist of a set V of n individuals or units, corresponding to the nodes of the network, and observations of relations between them, represented by the edges. The set $E = \{y_{ij} \in \{0, 1\}$, with $i, j = 1, \dots, n\}$ collects the pairwise relationships, where 0 and 1 respectively represent the absence or presence of a connection.

There has been a rising interest in analysing this type of data in order to infer connectivity patterns and topological properties of the network. The goal of statistical modelling of networks is to describe the connectivity structure by a relatively low–dimensional probabilistic model. This idea is enforced by the fact that real–world networks frequently exhibit low–dimensional structures which are responsible for the overall connectivity, such as homophily, core–periphery, and block connectivity. A variety of methods have been proposed for the analysis of a single network, such as the *exponential random graph* (Frank and Strauss, 1986), the *stochastic block model* (Nowicki and Snijders, 2001), and the *latent position* or *latent space model* (Hoff et al., 2002). These have been followed by several generalisations and extensions (e.g. Handcock et al., 2007; Hoff, 2007; Airoldi et al., 2008; Krivitsky et al., 2009; Fosdick et al., 2019; Schweinberger et al., 2020; Legramanti et al., 2022; Ricci et al., 2022).

In many domains, there is an increasing availability of replicated observations of relational data. A separate analysis of each network would ignore the inherent dependence between observations, given by the underlying common structure. For instance, in *dy-*

namical networks the same network is observed at different time shots. Models attempt at capturing the dynamics responsible for the change of the connectivity over time (Xu and Hero, 2013; Xu, 2015; Durante and Dunson, 2014, 2016; Miscouridou et al., 2018). In *multilayer or multiplex networks*, a set of nodes is observed under multiple contexts corresponding to different observed set of edges, which share similar connectivity structures (Kivelä et al., 2014; Gollini and Murphy, 2016; Young et al., 2022). We focus here on a particular case of multiplex networks, which we call *populations of networks* (see, e.g. Scheinerman and Tucker, 2010; Durante et al., 2017; Arroyo et al., 2021; Lunagómez et al., 2021). For a fixed set of nodes V, we assume to observe multiple sets of edges $\{E_m\}_{m=1}^N$ given by independent realisations of the underlying random graph representing the edge generation mechanism. All observed networks $\{(V, E_1), \dots, (V, E_N)\}$ share the same nodes V. The multiple observations allow us to infer quantities of interest about the underlying random graph.

In particular, we consider a frequent setting in biology and neuroscience, in which the network of the brain connectivity has been recorded for a group of subjects (see, e.g. Van Essen et al., 2012; Zuo et al., 2014; Zhang et al., 2018). We examine the brain anatomical connectivity measured via diffusion tensor imaging (DTI) for a set of $N = 20$ individuals. DTI records how water molecules diffuse across brain tissues. The white matter fibres of the brain facilitate the diffusion of water, and thus DTI allows for recovering the *structural* brain network given by the white matter. See Craddock et al. (2013); Stam (2014); Sporns (2022) for a technical discussion of the retrieval of the brain connectivity via DTI and other techniques.

The dataset we analyse comes from a pilot study of the Enhanced Nathan Kline Institute Rockland Sample project[1], see Figure 2.1 for an example. Networks obtained by post–processing DTI scans are subject to natural variability across individuals and inherent measurement errors. This double stochastic nature of the observations motivates the need of statistical modelling for such data.

Several models for populations of networks have been proposed in order to infer the underlying structure responsible for the observed connectivity patterns in the brain (Durante et al., 2017; Wang et al., 2019; Aliverti and Durante, 2019; Schweinberger et al., 2020), or to build comparison between heterogeneous sets of subjects (Durante and Dunson, 2018; Carboni et al., 2021, 2023). For instance, Durante et al. (2017) leverage replicated observations to learn the connectivity patterns with a latent space based mixture model, in which the mixture components are shared across networks. Wang et al. (2019) extends the eigenmodel (Hoff, 2007) to populations of networks allowing for a common term capturing the baseline connectivity shared across networks, together with a subject–specific

[1] http://fcon_1000.projects.nitrc.org/indi/enhanced/

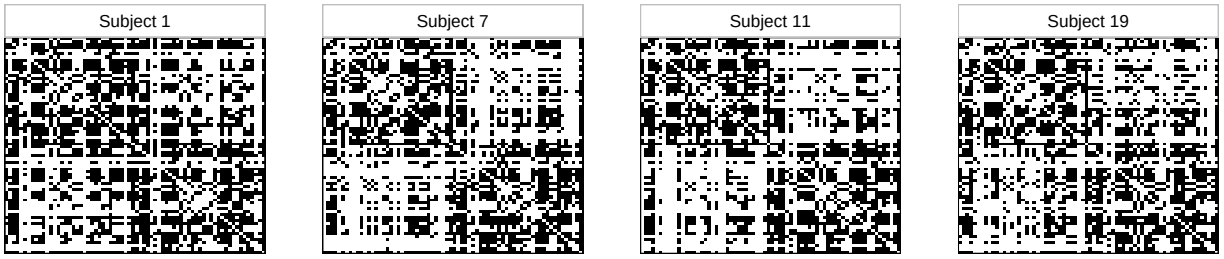


Figure 2.1: Example of four adjacency matrices of the brain connectivity networks considered in Section 2.5. Network nodes corresponds to the $n = 70$ brain regions defined by the Desikan atlas (Desikan et al., 2006).

factor.

However, the existence of different widely used partitions of the brain at different level of granularity – such as the Desikan atlas (Desikan et al., 2006), the division in lobes and hemispheres – motivates a multiresolution view of the structure of the brain, eventually affecting the connectivity (Hilgetag et al., 2000; Sporns et al., 2005; Moreno-Dominguez et al., 2014; Betzel and Bassett, 2017; Urchs et al., 2019; Li et al., 2023). Previous attempts at statistical modelling brain networks in the latent position model class (e.g. Durante et al., 2017; Wang et al., 2019; Aliverti and Durante, 2019) fail at allowing for such structure in the model specification.

We present a novel latent position model for populations of networks. Our proposal is designed to capture multiresolution connectivity structures that are shared across multiple networks. This is accomplished by assuming that the latent positions of each network are sampled from a branching Brownian motion. Crucially, all the Brownian motions associated with the networks share the same branching structure, which is effectively represented by a phylogenetic tree. As a consequence, the latent positions within each network exhibit a common dependence structure. Leveraging the phylogenetic tree component enables us to infer the multiresolution organization of the nodes, thereby facilitating the identification of nested clusters that capture community–like connectivity patterns within the networks.

Phylogenetic trees have been largely developed and extensively employed in the field of evolutionary biology (see Felsenstein (2004) for an introduction). However, their utility extends beyond this field and finds applications in various other domains, such as demography (Opgen-Rhein et al., 2005; Drummond et al., 2005), and linguistics (Ryder and Nicholls, 2011; Sagart et al., 2019). The evolutionary interpretation of the phylogenetic tree is attractive and it opens new interesting research directions in network modelling. However, we recognize that inference regarding the potential evolutionary processes of networks necessitates domain–specific calibrations and assumptions. These considera-

tions go beyond the scope of our current work. As a result, further investigations, along with external validation by domain experts, must be conducted before employing our proposal for inferring evolutionary processes. In our model, we utilise phylogenetic trees as mathematical constructs that enable us to achieve a hierarchical representation of the network nodes and effectively model their interdependence.

The objective of learning the multiresolution structure of networks arise in many domains and it has been previously addressed in the literature. The main works in this direction has been done primarily in the stochastic block model framework (Nowicki and Snijders, 2001), and limited to single network modelling. Typically, network nodes are placed at the leaves of a binary–splitting tree. Each internal node of the tree has a parameter representing the probability of a connection between pairs of nodes that have that tree node as the most recent common ancestor, e.g. see (Roy et al., 2006; Clauset et al., 2008; Roy and Teh, 2008; Herlau et al., 2012). Schmidt and Morup (2013) generalise binary splits to multifurcating trees via Gibbs fragmentation processes (McCullagh et al., 2008). Schweinberger and Snijders (2003), instead, embeds the nodes of the network in a latent ultrametric space. The ultrametric property of the space encodes the tree organization of the nodes. The probability of a connection between two nodes is treated as a parameter depending only on the ultrametric distance between the nodes. The authors assign uniform priors on both the distance and the probability parameters. Even though the model embeds the nodes in a latent space, the structure is again one of a tree–based stochastic block model, where the ultrametric distance prior is a distribution on the class of partitioning trees, with probability parameters assigned to each internal node of the tree.

It is worth noticing that in these works, while the tree parameter allows for inferring the hierarchical clustering of the nodes, the connection probabilities are assumed to be independent of the tree. This is a restrictive assumption, as one may expect that connections between similar pairs of blocks, characterized by most common recent ancestors being close in the tree, have similar probabilities. Additionally, while the stochastic block model effectively captures community structures and assortative mixing, it falls short in accurately modelling crucial local connectivity patterns that are characteristic of the brain structural connectivity (Bullmore and Sporns, 2009). For instance, brain regions located nearby have higher chance to be connected as the material and energy costs to form white matter fibres are high, thereby likely showing in the network patterns such as homophily and triangles. Latent position models, instead, are able to represent well these types of structures through the latent space representation (Kaur et al., 2023). Fosdick et al. (2019) overcome this problem combining the stochastic block model with the latent position model, respectively for the between–blocks and within–blocks connection

probabilities, obtaining two levels of resolution of the network. While such an approach has shown interesting results applied to social sciences (Ng et al., 2021), the methodology is still limited to a single network and only allows for two levels of resolution.

Our contribution stands out from the previously mentioned approaches with the introduction of the phylogenetic latent position model. This innovative framework combines the benefits of the latent position model class with the ability to capture multiresolution structures and community–like connectivity patterns, shared across a population of networks. Moreover, the connection probabilities are inherently dependent on the tree component of the model responsible for the hierarchical organization of the nodes. In our proposal, each network is associated with its own set of latent positions, differently from Gollini and Murphy (2016), where the latent positions are shared across all networks. We believe that this assumption is restrictive and can be relaxed by only sharing the latent tree structure. By allowing network–specific latent positions, our model accommodates a wider range of variations and captures the inherent diversity within the population of networks.

The remaining of the Chapter is organized as follows. In Section 2.2, we introduce phylogenetic trees. In Section 2.3, we define our model and discuss how to sample from the posterior distribution. In Section 2.4 and 2.5, we respectively apply our model to a set of simulated examples and to the brain networks data. Finally, we conclude with a discussion in Section 2.6.

2.2 Phylogenetic Trees

We briefly introduce the topic of phylogenetic trees in order to facilitate the reading of the remaining of the Chapter, for those who are not familiar with the subject.

Phylogenies are the natural object for thinking about evolution. The major contributions in developing theory and methods for phylogenetic trees come from the evolutionary biology and population genetics communities. A phylogenetic tree g is a tree endowed with branch lengths. From a mathematical perspective, there are different ways of considering them. They can be defined as random trees with random branch lengths, or equivalently as point processes in the product space of time and tree–node indexes, or as branching processes (Aldous, 2001).

For what matters our work, we consider trees conditioned on having n leaves corresponding to the network nodes, at the time of the observations. Let us consider an example of the construction of a phylogenetic tree seen as a branching process, in the simplest case where only bifurcations are allowed. It is a usual convention to consider time to increase from the leaves to the root of the tree, placing the leaves at time $\tau = 0$

and the root at time $\tau = \tau_0 > 0$. In this case, time takes the meaning of *age* or *depth* of a node, with the root of the tree being the oldest node at depth τ_0 and the leaves the youngest at depth 0. Such convention is common in many mathematical derivations and software implementations. However, when reporting results in applications it is natural to use the opposite time orientation, which follows the *evolutionary time* t. In this case, the time measure is also called the *height* of a node. The root is set to be at height $t = 0$ and leaves at height $t = t_0 = \tau_0$. The reader has to get used to both conventions.

The process starts from a single node, the root. At τ_0 the root gives birth to two offspring nodes. The branch of each offspring grows independently from the other, for some amount of time after which, in turn, it gives birth to two new offspring. The process goes on recursively, like a growing tree. At time $\tau = 0$, a snapshot of the tree shows the n current offspring nodes, which we refer to as *leaves*, *tips*, or *individuals*. Notice that when the tree node j gives birth to l and m, we do not count j anymore among the current leaves or individuals of the tree.

If we assume that the amount of time between the generation of a node and the moment it gives birth to offspring is independent of the other nodes and follows an exponential distribution with rate $b > 0$, then the process described before is the *Yule process* (Yule, 1925). The Yule process is a special case of the *birth and death process* (see, e.g. Harris, 1963; Ross, 2014), where the birth rate is b and the death rate is 0. For general birth and death rates $(b, d > 0)$, the process is changed by assuming that each node undergoes two competing events: giving birth to two offspring with rate b, or dying with rate d. If death happens before giving birth, then the branch of the tree related to the node is removed from the tree.

Depending on the assumptions, there are different ways to compute the density of a given phylogenetic tree under the birth and death process of rates (b, d). If we condition the process on having n leaves at the present, on the root giving birth to two sub–branching processes at time τ_0, and on both of them surviving with any descendant to the present $\tau = 0$, then we can write the density of a given tree g with branching times $\tau_0 > \tau_1 > \cdots > \tau_{n-1} > 0$ as follows:

$$f(g) = (n-1)! \left(\frac{p_1(\tau_0)}{1 - p_0(\tau_0)} \right)^2 \prod_{i=1}^{n-1} b \, p_1(\tau_i), \tag{2.1}$$

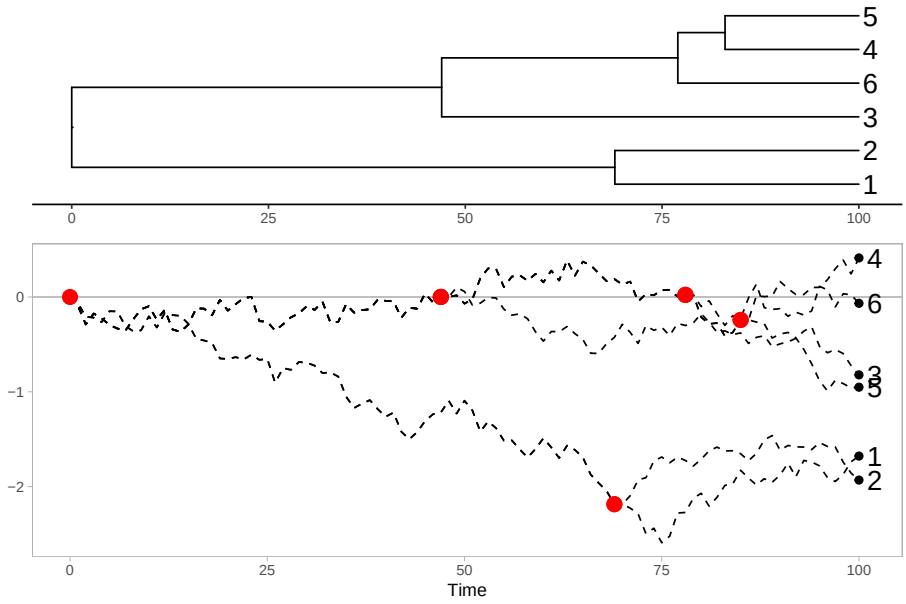


Figure 2.2: Example of the realization of a branching Brownian motion with the tree representing the branching structure.

where the auxiliary functions $p_1(\tau)$ and $p_0(\tau)$ are defined as:

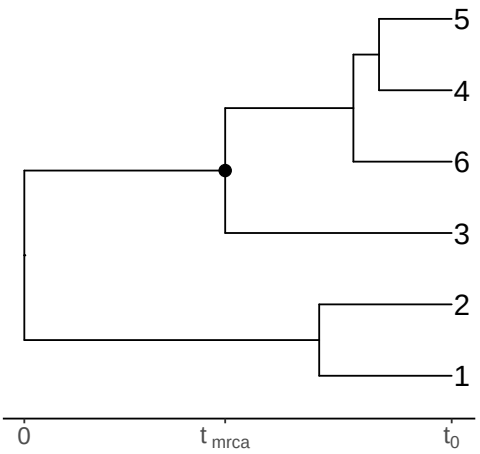
$$\begin{aligned} p_0(\tau) &= 1 - \frac{b-d}{b - d\, e^{-(b-d)\tau}} \\ p_1(\tau) &= \frac{(b-d)^2\, e^{-(b-d)\tau}}{(b - d\, e^{-(b-d)\tau})^2}. \end{aligned} \tag{2.2}$$

See Stadler (2010, 2013) for the derivation of different density functions under different assumptions, and Kendall (1949) for a detailed study of the birth and death process.

In many applications, the phylogenetic tree is the skeleton for some Markov process which is observed at the leaves. For example, this can represent genetic variations of nucleotides of a set of observed DNA sequences (see, e.g. Felsenstein, 1981). In our framework, instead, we consider branching Brownian motions (BBM) starting at the origin at $t = 0$, see Figure 2.3 for an example.

The phylogenetic tree represents the branching structure of the Brownian motion. On each branch, the process evolves as an independent Brownian motion with a given rate σ^2. It follows that the random vector collecting the values of the BBM at the tips of the tree is distributed as a multivariate normal distribution. The marginal variance of each component is given by the product $t_0\sigma^2$, between the total height of the tree t_0 and the rate of the Brownian motion σ^2. The branching structure affects the covariance between

the tips values. By simple computations, one finds out that the covariance between two tips is equal to the variance of the Brownian motion particle corresponding to the most recent common ancestors of the tips in the tree. For instance, consider the following tree example and let $(x_1, x_2, x_3, x_4, x_5, x_6)^\intercal$ be the random vector representing the Brownian motion at the tips of the tree. It follows that the variance–covariance of, e.g., x_3 and x_5, is given by:



$$\begin{aligned} \mathrm{Var}(x_3) &= t_0\sigma^2 \\ \mathrm{Var}(x_5) &= t_0\sigma^2 \\ \mathrm{Cov}(x_3, x_5) &= t_{\mathrm{mrca}}\sigma^2. \end{aligned} \tag{2.3}$$

In absence of historical observations, i.e. between time 0 and t_0, the rate of the Brownian motion is identifiable only for a fixed time scale, and vice versa. Therefore, we will consider in our model a fixed and arbitrary age of the root t_0. We discuss in Section 2.6 possible extensions of our model to, e.g., non–constant rate Brownian motions.

In the Bayesian framework, phylogenetic trees are random entities. Inference on trees is based on a set of tree samples from the posterior distribution of the model given data. The uncertainty in the posterior can be visualised by plotting the trees under a common leaf ordering (see, e.g. the bottom plot in Figure 2.6). However, this approach is limited by the fact that, in some cases, the chosen ordering has a significant influence on the visual impact of the graphical representation of the set of trees. Softwares such as DENSITREE (Bouckaert, 2010) and GGTREE (Yu et al., 2017) implement different algorithms to choose the optimal leaf ordering for better visualization.

Alternatively, one can build estimates or summaries from a set of trees. Defining summary statistics for complex objects like trees is challenging. However, there is a very common set of tools to summarize a set of trees with a single one, which are called the *consensus trees* (see, e.g. Felsenstein, 2004, Ch. 30). We consider in particular the *majority–rule consensus tree* of level p. The topology of the majority–rule consensus tree merges two nodes in a subtree, if the branching occurs with a frequency of at least p in the set of trees. Given the consensus topology, branch lengths can be computed in several ways,

e.g. as the mean or the least square length of the lengths of the branches in the set of trees. As alternative summaries, it is also possible to consider the maximum a posteriori topology or the maximum clade credibility tree.

Regarding implementations, there are several softwares for Bayesian phylogenetic models like BEAST or RevBayes (Bouckaert et al., 2014; Höhna et al., 2016), together with a wide universe of R packages (R Core Team, 2023) to explore, manipulate, and fit phylogenetic trees (see, e.g. Gearty et al., 2023).

2.3 Phylogenetic Latent Position Models

Consider a single network (V, E) represented by the $n \times n$ adjacency matrix $\mathbf{Y}$, with $[\mathbf{Y}]_{ij} = y_{ij} = 1$ if nodes i and j are connected and $y_{ij} = 0$ otherwise, for $i, j \in \{1, \dots, n\}$. We focus on undirected networks with no self–loops, corresponding to symmetric $\mathbf{Y}$ with null diagonal entries $y_{ii} = 0$.

We follow the latent position model approach (Hoff et al., 2002). Edges are drawn independently from a Bernoulli distribution given the edge probability θ_{ij},

$$y_{ij} \mid \theta_{ij} \overset{\text{ind}}{\sim} \text{BERN}(\theta_{ij}). \tag{2.4}$$

Network nodes are embedded in a K dimensional latent space, where each node i is represented by a vector of latent coordinates $\mathbf{z}_i \in \mathbb{R}^K$. The latent space representation is used to model the connection probabilities. The closer are two nodes in the latent space, the higher is the probability of an edge between them. This is obtain by modelling the edge probabilities through a logistic regression depending on the Euclidean distance between the latent positions, as follows:

$$\text{logit}\, \theta_{ij} = a - \| \mathbf{z}_i - \mathbf{z}_j \|, \tag{2.5}$$

where a captures the overall edge density in the network. In case covariates are available, these are included with an additional regression term to the right–hand side of (2.5).

Let us consider now the setting of populations of networks and introduce the phylogenetic latent position model. We have a set of networks $\{\mathbf{Y}^{(1)}, \dots, \mathbf{Y}^{(N)}\}$, represented by $n \times n$ adjacency matrices with binary entries $[\mathbf{Y}^{(m)}]_{ij} = y_{ij}^{(m)}$. We assume that conditioning on the latent positions, the edges of each network are independent from the other networks. The model likelihood (2.4) for a single network is directly extended to

populations of networks as follows: independently for each $m \in \{1, \dots, N\}$,

$$\begin{aligned} y_{ij}^{(m)} \mid \mathbf{Z}^{(m)}, a &\overset{\text{ind}}{\sim} \text{BERN}(\theta_{ij}^{(m)}) \qquad i, j = 1, \dots, n \\ \text{logit}\, \theta_{ij}^{(m)} &= a - \| \mathbf{z}_i^{(m)} - \mathbf{z}_j^{(m)} \|, \end{aligned} \tag{2.6}$$

where $\mathbf{Z}^{(m)} = [\mathbf{z}_1^{(m)}, \dots, \mathbf{z}_n^{(m)}] \in \mathbb{R}^{K \times n}$ collects the latent positions of network m. We assume that all networks are embedded in the same latent space $\mathbb{R}^K$. For each network $m \in \{1, \dots, N\}$, the latent positions $\mathbf{Z}^{(m)}$ are sampled from a K–dimensional branching Brownian motion (BBM). All Brownian motions associated with the networks share a common branching structure, which is represented by the phylogenetic tree g. This implies the following prior for the latent positions: for $m \in \{1, \dots, N\}$,

$$\begin{aligned} \mathbf{Z}^{(m)} \mid g, \sigma^2 &\overset{\text{iid}}{\sim} \text{BBM}_K(\sigma^2, g) \\ g \mid b &\sim \text{BDT}_n(b, 0) \end{aligned} \tag{2.7}$$

where σ^2 is the diffusion parameter of the Brownian motion, and BDT_n is the birth and death process prior on a tree with n leaves with birth and death rates b and $d = 0$, which is equivalent to the Yule process. This prior choice has shown to provide enough flexibility for the purpose of our applications and it reduces the number of parameters of the model.

The phylogenetic tree g is responsible for capturing the correlation structure among the nodes in each network and for each dimension of the latent space $\mathbb{R}^K$. In particular, the tree induces a $n \times n$ covariance matrix Σ_g, see example (2.3). The vector of the k-th components of the latent positions $\mathbf{Z}^{(m)}$ of each network m, i.e. $([\mathbf{z}_1^{(m)}]_k, \dots, [\mathbf{z}_n^{(m)}]_k)^\intercal$, follows a n–dimensional normal distribution centred in zero and with covariance $\sigma^2 \Sigma_g$, with independence across components $k \in \{1, \dots, K\}$. For a fixed m, this can be written as follows:

$$\mathbf{Z}^{(m)} \mid g, \sigma^2 \sim \text{BBM}_K(\sigma^2, g) \quad \Longleftrightarrow \quad \begin{pmatrix} [\mathbf{z}_1^{(m)}]_k \\ \vdots \\ [\mathbf{z}_n^{(m)}]_k \end{pmatrix} \mid g, \sigma^2 \overset{\text{iid}}{\sim} \mathcal{N}_n(0, \sigma^2 \Sigma_g), \tag{2.8}$$

where the right–hand side is independently and identically distributed over the latent space dimensions $k \in \{1, \dots, K\}$. The normal distribution is the result of observing the evolution of the branching Brownian motion at the tips of the tree g. Figure 2.3 shows an example of the latent positions for three networks, together with the shared phylogenetic tree.

We complete the model specification by setting independent priors for the remaining

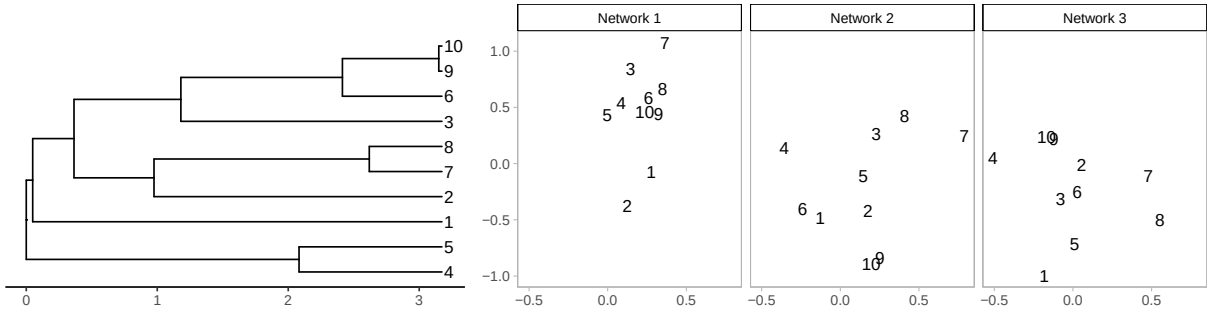


Figure 2.3: Example of the realization of a branching Brownian motion with $K = 2$ over a tree with $n = 10$ leaves. On the left the underlying phylogenetic tree. On the right the n network nodes in the latent space for $N = 3$ networks.

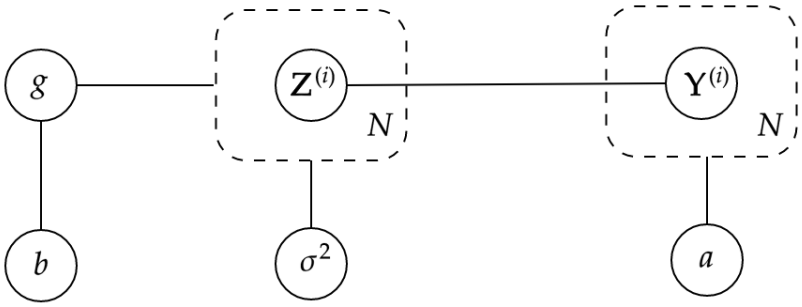


Figure 2.4: Graphical model representation of the phylogenetic latent position model.

model parameters as follows:

$$\begin{aligned} \sigma^2 &\sim \text{INV-GAMMA}(\alpha_\sigma, \beta_\sigma) \\ b &\sim \text{INV-GAMMA}(\alpha_b, \beta_b) \\ a &\sim \mathcal{N}(0, \sigma_a^2). \end{aligned} \tag{2.9}$$

Figure 2.4 shows the graphical model of the phylogenetic latent position model. Dashed boxes represent sets of replicated random variables. Edges between circles indicates one–to–one dependence, whereas edges between circle and boxes indicates that the random variable in the circle is a shared parameter for all random variables in the box.

For the remaining of the Chapter, we set the latent space dimension to $K = 3$. We have preliminary explored different values for K with the latent position model of Hoff et al. (2002) through the LATENTNET software (Krivitsky and Handcock, 2008) in the context of

our experiments. The value $K = 3$ has shown to achieve a good representation of the networks, while allowing for a low–dimensional model structure and reasonable computational costs. Finally, we fix the prior hyperparameters as follows: $\alpha_b = \alpha_\sigma = \beta_b = \beta_\sigma = 1$, and $\sigma_a = 2$. In our experience, the model has proved to be robust to small deviations from these default values.

2.3.1 Signal and Model Identifiability

From an inferential perspective, the main quantity of interest in our model is the latent phylogenetic tree g, responsible for the multiresolution structure of the network nodes. The size n of each network and the number of observed networks N play two opposite roles for what matters the identifiability of g.

Considering a single network, data are indirectly placed at the leaves of g due to the latent position representation of the network. Conceptually, the flow of the information from the data to the latent layers of the model has the following path. The observed edges inform the probabilities of connection between nodes. These affect the latent positions, since the pairwise distances define the connection probabilities. The latent positions are instances of the branching Brownian motion at the leaves of g. In this flow, the topology of the tree is one layer more latent than the latent positions $\mathbf{Z}^{(m)}$. The amount of information contained in a single network of n nodes is low to infer a tree with n leaves. The problem gets harder as n increases, as the size of the tree increases as well. This is somehow equivalent to the problem of estimating a $n \times n$ covariance matrix Σ of a vector of n components observing K replicates, with $K \ll n$, but in our case with constraints on the structure of the covariance matrix (i.e., $\Sigma = \Sigma_g$). Observing N networks, instead, corresponds to observing KN replicates of the same covariance structure Σ_g. Therefore, as N grows there is increasing signal in the data to better infer the latent hierarchical organization of the nodes given by the phylogenetic tree g.

The connection probabilities depend on the latent positions only through their pairwise distance, see equation (2.5). In the classical latent space model, where the latent positions follow a K-dimensional isotropic normal distribution, the latent positions are identifiable up to rotations, translations, and reflections (this class of transformations forms an equivalence class for the latent positions). By post–processing the MCMC draws via a suitable Procrustean transformation (Hoff et al., 2002; Handcock et al., 2007), it is possible to obtain a set of posterior samples of the latent positions all belonging to the same equivalence class.

In the prior (2.7), instead, the conditional distribution of the latent positions is not isotropic, because the covariance Σ_g is not diagonal. This is expected to improve the identifiability of the latent positions, but it is likely not enough to have complete identifi-

ability. If they are of direct interest, we suggest to rely on the Procrustean transformation as in (Hoff et al., 2002). In our experience, we have observed that this condition is not harmful for what matters the inference of the phylogenetic tree g. In general, the weak identifiability of the latent positions could potentially lead to a multimodal posterior of g. The existence of different configurations of the latent positions with similar likelihood and prior values may correspond to different tree topologies associated to each configuration, with similar posterior probabilities. However, we have not experienced this in practice. In our experiments, we have observed good concentration of the posterior and no signs of this behaviour in the convergence diagnostics of the posterior sampling. Nevertheless, we believe that it is of primary importance to better study this aspect of the model in the near future, ideally providing theoretical guarantees for the identifiability of the tree.

2.3.2 Posterior Computations via Gibbs-Sampling

We design a Gibbs sampling algorithm targeting the joint posterior of the model (2.6)–(2.7)–(2.9), given the data. We give here a detailed description of each sampling step of the full conditional distributions of the algorithm, which leverages – where possible – parallelization and conjugate updates. When conjugacy is not available, a Metropolis–Hasting step is adopted.

Except for tree moves, all Metropolis–Hasting based moves use symmetric Gaussian proposals. We target the ideal acceptance rate of $\bar{\alpha} = 0.23$ through the adaptation algorithm of Andrieu and Thoms (2008). Such value should guarantee a good balance between sampling nearby the current parameter value of the chain – in a region of high posterior probability – and exploring the remaining of the parameter space. In practice, at the end of iteration s we adapt the standard deviation η_s of the proposal distribution for a given parameter as follows:

$$\log \eta^s = \log \eta^{s-1} + s^{-0.8}(\alpha^s - \bar{\alpha}), \tag{2.10}$$

where α_s denotes the acceptance probability computed at iteration s.

In the following, we provide formulas for a generic iteration s of the sampler. We use superscripts s and $s+1$ for the parameter involved in the sampling, respectively indicating the current value and the proposed value at the given iteration. In order to simplify the notation, we do not explicitly indicate the iteration for the parameters not involved in the sampling. Their values are assumed to be the most recently computed ones.

We denote the generic laws for observations and parameters with $\mathcal{L}(\cdot)$ and $\pi(\cdot)$. When necessary, we explicitly denote the density function computed in a given value, separating distribution parameters with semicolons (e.g., $\mathcal{N}(x; \mu, \sigma^2)$ is the density of the normal

distribution computed in x, with mean μ and variance σ^2). Finally, the horizontal bar in the writing $\pi(\cdot \mid -)$ denotes conditioning on all the other parameters and observations.

Parameter a Due to conditional independence (see Figure 2.4), we factorize the full conditional of a as follows:

$$\pi(a \mid -) \propto \pi(a) \prod_{m=1}^{N} \mathcal{L}(\mathbf{Y}^{(m)} \mid a, \mathbf{Z}^{(m)}). \tag{2.11}$$

At iteration s, the acceptance probability for the Metropolis–Hasting step is given by:

$$\alpha_a^s = \min\left\{1, \frac{\mathcal{N}(a^{s+1}; 0, \sigma_a^2) \prod_{m=1}^{N} \mathcal{L}(\mathbf{Y}^{(m)} \mid a^{s+1}, \mathbf{Z}^{(m)})}{\mathcal{N}(a^{s}; 0, \sigma_a^2) \prod_{m=1}^{N} \mathcal{L}(\mathbf{Y}^{(m)} \mid a^{s}, \mathbf{Z}^{(m)})}\right\}, \tag{2.12}$$

where the likelihood terms can be computed as follows:

$$\begin{aligned} \mathcal{L}(\mathbf{Y}^{(m)} \mid a, \mathbf{Z}^{(m)}) &= \prod_{i,j=1\,:\,i<j}^{n} (\theta_{ij}^{(m)})^{y_{ij}^{(m)}} (1-\theta_{ij}^{(m)})^{1-y_{ij}^{(m)}} \\ \operatorname{logit} \theta_{ij}^{(m)} &= a - \| z_i^{(m)} - z_j^{(m)} \|. \end{aligned} \tag{2.13}$$

Alternatevely to Metropolis–Hasting with Gaussian proposal, it is possible to consider the Polya–Gamma augmentation (Polson et al., 2013).

Birth rate b The birth rate b is a positive parameter, therefore we consider symmetric Gaussian proposals on the log scale, as follows:

$$\log b^{s+1} = \log b^s + \epsilon_b^s, \tag{2.14}$$

where $\epsilon_b^s \overset{\text{ind}}{\sim} \mathcal{N}(0, \eta_s^2)$. Due to conditional independence, the full conditional of b only depends on the tree g. It follows that we can write the acceptance probability as follows:

$$\alpha_b^s = \min\left\{1, \frac{\gamma^{-1}(b^{s+1}; \alpha_b, \beta_b)\, \pi_{\text{BDT}}(g; b^{s+1}, 0)}{\gamma^{-1}(b^{s}; \alpha_b, \beta_b)\, \pi_{\text{BDT}}(g; b^{s}, 0)}\right\}, \tag{2.15}$$

where $\gamma^{-1}(\cdot; \alpha, \beta)$ denotes the density of the Inverse–Gamma distribution, $\pi_{\text{BDT}}(\cdot; b, 0)$ is the density of the birth and death process given by (2.1).

Tree g In order to sample from the full conditional of the phylogenetic tree g, we implement four different symmetric moves, which ensure ergodicity of the chains in the tree space. In particular, we implement the followings, under the condition that they do not violate the ultrametric constraint of the tree (i.e., tree leaves are all at the same height):

1. **Tips interchange**: it consists in randomly selecting two leaves of the tree and swapping them;
2. **Subtree exchange**: it consists in randomly selecting two subtrees and swapping them;
3. **Tree-node age move**: it consists in randomly selecting an internal node and shifting his age. This corresponds to expand or contract the length of the branch connecting the selected node and its parent in the tree. Accordingly to this, the child branches of the selected nodes are respectively contracted or expanded in order to keep unchanged the total height of tree.
4. **Subtree pruning and regrafting**: it consists in randomly selecting a subtree, pruning it and re–attaching it to another suitable position in the tree. Any branch which will not violate the time order of the parent, selected node, and the children, represents a suitable position.

The full conditional of g is proportional to the prior on g times the conditional distribution of the latent positions (see Figure 2.4). It follows that we can write the acceptance probability for any of the above moves as follows:

$$\alpha_g^s = \min\left\{1, \frac{\pi_{\text{BDT}}(g^{s+1}; b, 0)\ \prod_{m=1}^{N} \pi(\mathbf{Z}^{(m)} \mid \sigma^2, g^{s+1})}{\pi_{\text{BDT}}(g^{s}; b, 0)\ \prod_{m=1}^{N} \pi(\mathbf{Z}^{(m)} \mid \sigma^2, g^{s})}\right\}. \tag{2.16}$$

Denoting with $[\mathbf{Z}^{(m)}]_k = ([\mathbf{z}_1^{(m)}]_k, \dots, [\mathbf{z}_n^{(m)}]_k)^\intercal \in \mathbb{R}^n$ the vector collecting the k-th coordinates of the latent positions of the nodes in network m, we can write the conditional density of the latent positions $\mathbf{Z}^{(m)}$ in the following way:

$$\pi(\mathbf{Z}^{(m)} \mid \sigma^2, g) = \prod_{k=1}^{K} \mathcal{N}([\mathbf{Z}^{(m)}]_k; \mathbf{0}, \sigma^2 \Sigma_g), \tag{2.17}$$

where Σ_g is the covariance matrix induced by the tree g.

The prior density (2.1) is invariant with respect to the *tips interchange* move, since it does not depend on the labelling order of the leaves. Therefore, the ratio between $\pi_{\text{BDT}}(g^{s+1}; b, 0)$ and $\pi_{\text{BDT}}(g^{s}; b, 0)$ cancels out in the acceptance probability of this move.

Brownian motion rate σ^2 The prior on σ^2 is conjugate with the conditional distribution of the latent positions. We can write the full conditional of σ^2 as follows:

$$\pi(\sigma^2 \mid -) \propto \pi(\sigma^2) \prod_{m=1}^{N} \pi(\mathbf{Z}^{(m)} \mid \sigma^2, g), \tag{2.18}$$

and leverage the conjugacy to sample from such distribution. The collection of the latent coordinates related to a given dimension, for a given network, is a multivariate normal with covariance $\sigma^2\Sigma_g$. If we denote with $\tilde{\mathbf{z}}$ the vector piling all the latent coordinates of all the networks,

$$\tilde{\mathbf{z}} = \left([\mathbf{Z}^{(1)}]_1^\intercal, \dots, [\mathbf{Z}^{(1)}]_K^\intercal, \dots, [\mathbf{Z}^{(N)}]_1^\intercal, \dots, [\mathbf{Z}^{(N)}]_K^\intercal\right)^\intercal \in \mathbb{R}^{nKN}, \tag{2.19}$$

we can rewrite (2.18) as follows:

$$\pi(\sigma^2 \mid -) \propto \gamma^{-1}(\sigma^2; \alpha_\sigma, \beta_\sigma)\, \mathcal{N}(\tilde{\mathbf{z}}; 0, \sigma^2 \mathbf{I}_{KN} \otimes \Sigma_g), \tag{2.20}$$

where $\mathbf{I}_{KN}$ denotes the KN–dimensional identity matrix and the operator $\otimes$ is the Kronecker product. Equation 2.20 shows the conjugate Normal–Inverse–Gamma model. Therefore, the Gibbs sampler step for σ^2 requires sampling from an INV-GAMMA$(\tilde{\alpha}, \tilde{\beta})$, with:

$$\begin{aligned} \tilde{\alpha} &= \alpha_\sigma + \frac{1}{2} nKN \\ \tilde{\beta} &= \beta_\sigma + \frac{1}{2} \tilde{\mathbf{z}}^\intercal \left(\mathbf{I}_{KN} \otimes \Sigma_g^{-1}\right) \tilde{\mathbf{z}}. \end{aligned} \tag{2.21}$$

Latent positions Z The full conditional of the latent positions $\mathbf{Z}^{(m)}$ depends only on network $\mathbf{Y}^{(m)}$, g and σ^2 (see Figure 2.4). The conditional independence of $\mathbf{Z}^{(m)}$ from $\mathbf{Y}^{(l)}$, for $l \neq m$, allows us to implement parallel updates of the latent positions of each network. In each parallel update, there are $n \times K$ components to sample, which are the K latent coordinates of the n nodes of network m. In our implementation, we jointly update the latent coordinates of each node, cycling over all nodes. For every proposal relative to network m, the acceptance probability is given by:

$$\alpha_z^s = \min\left\{1, \frac{\pi(\mathbf{Z}^{(m)\,s+1} \mid g, \sigma^2)\, \mathcal{L}(\mathbf{Y}^{(m)} \mid a, \mathbf{Z}^{(m)\,s+1})}{\pi(\mathbf{Z}^{(m)\,s} \mid g, \sigma^2)\, \mathcal{L}(\mathbf{Y}^{(m)} \mid a, \mathbf{Z}^{(m)\,s})}\right\}, \tag{2.22}$$

where the quantities $\mathcal{L}(\mathbf{Y}^{(m)} \mid -)$ and $\pi(\mathbf{Z}^{(m)} \mid -)$ can be computed from (2.13) and (2.17). Notice that the possibility to parallelize over each network m is highlighted from the fact

that only index m appears in equation (2.22).

2.3.3 Diagnosing Convergence of Markov Chain Monte Carlo for Phylogenetic Trees

The reliability of Bayesian posterior inference based on Markov chain Monte Carlo algorithms is based on the assumption that the Markov chains are correctly targeting the posterior distribution, they have reached stationarity after the burnin period, and the autocorrelation between samples is low. The first assumption is ensured by the proper implementation of the Gibbs sampler. However, careful inspection of the posterior samples is required to ensure the validity of the other ones.

Assessing the convergence of the chains to the stationary distribution is an important step in the modelling workflow. While plenty of tools have been designed to check the convergence of numerical quantities (see, e.g. Brooks and Gelman, 1998; Brooks et al., 2011; Vehtari et al., 2021), diagnosing convergence for trees is further more challenging. A first inspection consists in checking the mixing of numerical summaries of trees, such as branch lengths, total tree height (if random), and tree likelihood, using standard tools for numerical MCMC.

Dedicated softwares, such as the R–package RWTY (Warren et al., 2017), implements more sophisticated diagnostics, such as topology trace plots for trees (Lanfear et al., 2016). Given a distance in the tree space (see, e.g. Robinson and Foulds, 1981; Critchlow et al., 1996), topology trace plots are based on the trace plot of the distances between the tree samples and a fixed tree, named the focal tree. The latter can be, for instance, a random sample from the tree prior, or one of the posterior samples. In a similar fashion, it is possible to defined autocorrelation for trees. For instance, the topology autocorrelation (Lanfear et al., 2016) is based on the pairwise tree distance between trees at given lags in the chains. As the lag increases, the distances are expected to stabilise on a fixed average value. Additionally, it is possible to monitor other quantities, such as the clade splitting frequencies, and to inspect the projection of the tree samples on a low–dimensional space via multidimensional scaling (Hillis et al., 2005). Nonetheless, assessing proper convergence of MCMC remains still an important open problem of active research. See, e.g., Kelly et al. (2023) for a recent advancement and a broader discussion on the topic.

In all our applications, we check the goodness of the sampling relying on the MCMC diagnostics available in the R–package RWTY (Warren et al., 2017).

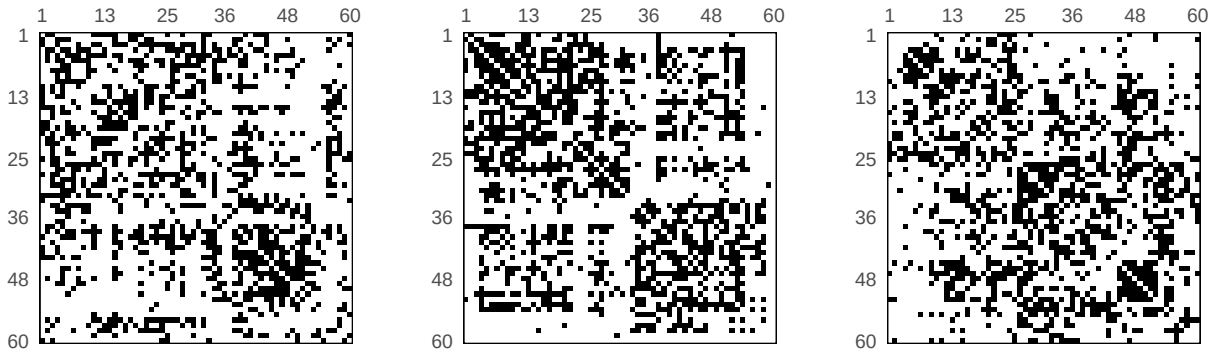


Figure 2.5: Prior simulation - Example of networks simulated under the model generating process.

2.4 Simulated Networks

We apply the proposed model to various simulated scenarios for evaluation. Firstly, we simulate data from the prior to assess the ability of the model to correctly identify the parameters. Then, we examine two different types of simulated networks that exhibit distinct structures: one lacks hierarchical organization among nodes, while the other demonstrates a tree structure. The data generating processes for the these scenarios follow Legramanti et al. (2022).

For all simulations, we run four chains initialised at random samples of the parameters, each for 20000 iterations and with a thinning frequency of 20. We consider a burnin window of 6000 iterations before thinning, based on the results of the MCMC diagnostic analysis.

Prior simulation We simulate data from the process (2.6)–(2.7) for a total of $N = 20$ networks, each with $n = 60$ nodes. We fix $\sigma^2 = 0.6$, and g to a sample from $\text{BDT}_n(0.7, 0)$. Based on these values, we sample the latent positions for each network according to (2.7), choosing $K = 3$. We fix $a = 2.6$ and sample the networks edges according to the observation equation of model (2.6). Figure 2.5 shows three of the simulated networks.

We fit the model on the sampled networks, initializing the chains at a random tree scaled to have the same length of the true tree. All other parameters are randomly initialized at values sampled from their priors. Table 2.1 shows a summary of the posterior draws for parameters a, b, and σ^2. Overall, there is good recovery of the true parameters of the data generating process. Parameter a showed slower mixing and a posterior distribution centered on a slightly lower value than the true one.

Figure 2.6 and 2.7 summarise the posterior of the phylogenetic tree g. In particular, the

	True value	Mean	Median	Sd	q5	q95
b	0.7	0.67	0.66	0.10	0.52	0.82
σ^2	0.6	0.67	0.66	0.06	0.57	0.78
a	2.6	2.52	2.52	0.05	2.43	2.60

Table 2.1: Prior simulation - Summary of posterior samples.

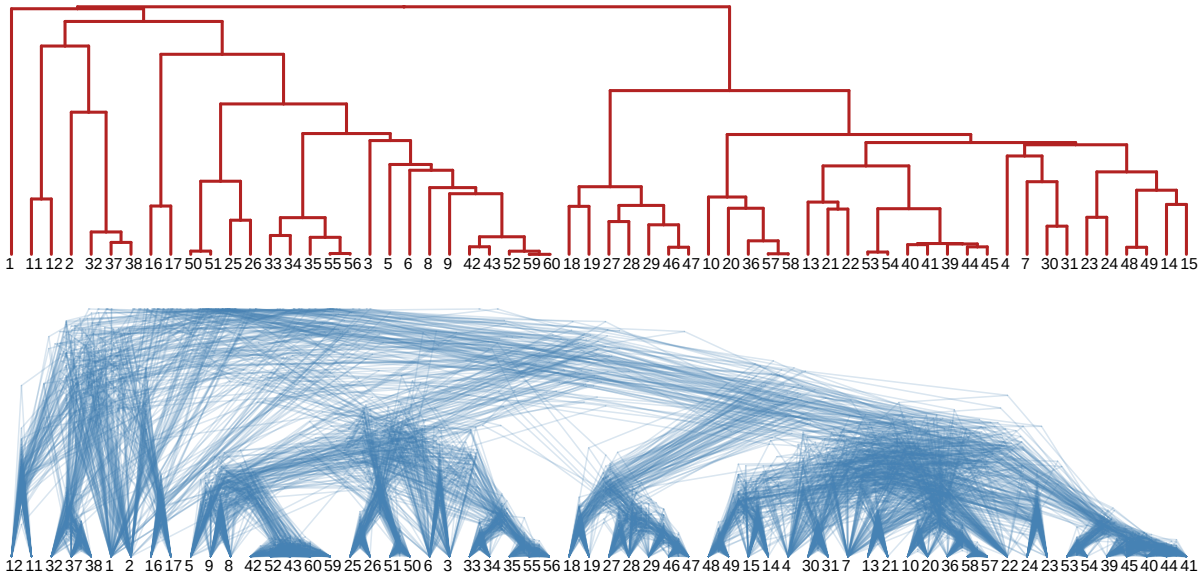


Figure 2.6: Prior simulation - True tree and posterior samples (subset of 100 samples).

first figure reports the true tree and a subset of 100 posterior samples of g. It is interesting to notice two things: firstly, the posterior samples recover part of the structure of the true tree; secondly, the uncertainty in the posterior is higher closer to the root of the tree. This is not surprising, as this part of the tree is the most difficult to identify since it relates to pairs of nodes almost uncorrelated.

Given the difficulties of comparing two large trees as the ones in Figure 2.6, we show in Figure 2.7 the consensus tree between the true tree of the data generating process and the posterior consensus tree of level 0.6. Such a tree has sub–tree splittings if they are present in both the true tree and the posterior consensus tree. The presence of many sub–structures in the consensus tree shows the ability to recover most of the true tree.

Independent groups and hierarchical structures We consider now two types of generating processes for networks with $n = 80$ nodes, presenting different block structures, which we refer to as (a) the independent groups, and (b) the dependent groups structure. In both cases, nodes are partitioned in 5 groups and the probability of an edge

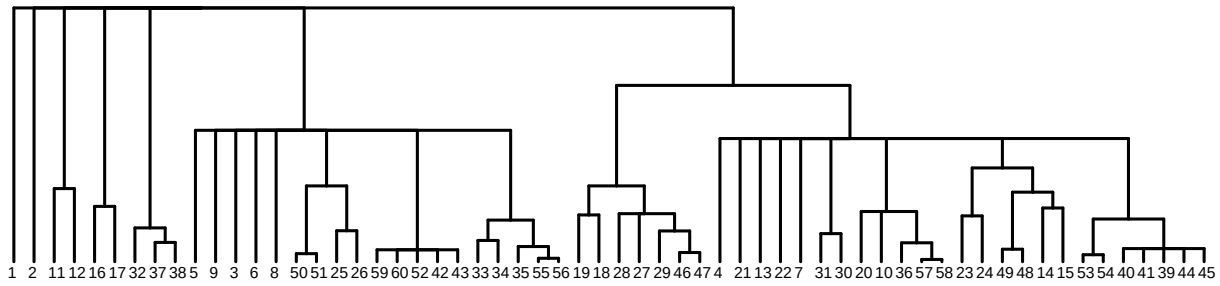


Figure 2.7: Prior simulation - consensus between the true tree and the posterior consensus tree of level 0.6.

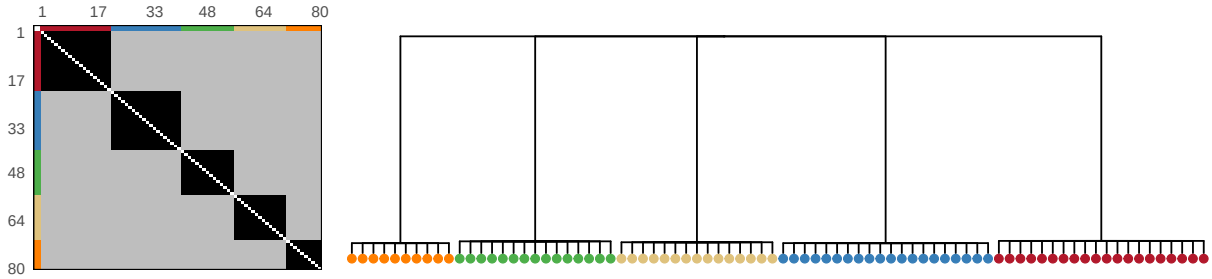
between two nodes depend on the group memberships.

The independent groups structure is given by 5 groups of nodes of size (20, 20, 15, 15, 10) with high intra–connectivity and low between–connectivity. In this setting, the connectivity of each group is independent from any other group. The dependent groups structure, instead, has 5 groups of size (25, 20, 15, 15, 5), each with high intra–connectivity. The first and the fourth groups have high between–connectivity, whereas the fifth has high between–connectivity with all the others, but the first one. This results in groups which are more similar than others.

In this experiment, we simulate $N = 10$ networks for each setting and fit the phylogenetic latent position model. Figure 2.8 shows, for each scenario, the matrix of the true connection probabilities and the estimated consensus tree obtained from the posterior samples. The matrices have entries of either high connection probability (set to 0.75 and corresponding to black areas), or low probability (set to 0.25 and corresponding to gray areas). On the margins of the matrices, the five colors (red, blue, green, brown, and orange) shows the affiliation of each node to one of the 5 groups. On the right, the consensus trees have leaves with colored tips showing the same affiliation. The results we show in this section do not change qualitatively when the same experiment is repeated increasing the number of simulated networks to $N = 20$. This is because the underlying tree structure is relatively simple, and $N = 10$ networks represents already a sufficiently large sample to infer the tree.

In the independent groups scenario (Figure 2.8a), the estimated consensus tree shows absence of hierarchy and correctly clusters nodes in their corresponding groups. In the second setting displayed in Figure 2.8b, instead, the model identifies a tree structure. Groups (1–4) and (2–3–5) are put on two different subtrees, coherently with their distinguishable connectivity patterns. Groups 1 (red) and 4 (brown) share similar edge probabilities, which only differ for the connectivity with group 5 (orange) of small size. While they are correctly put on the same subtree, the model however fails at separating the two

(a) Independent groups structure



(b) Dependent groups structure

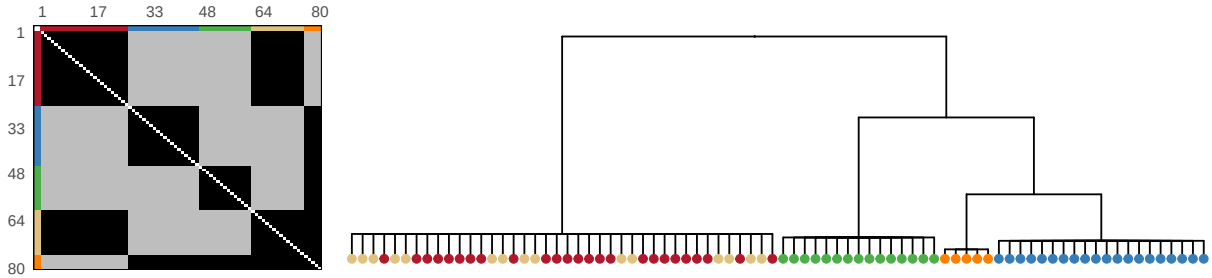


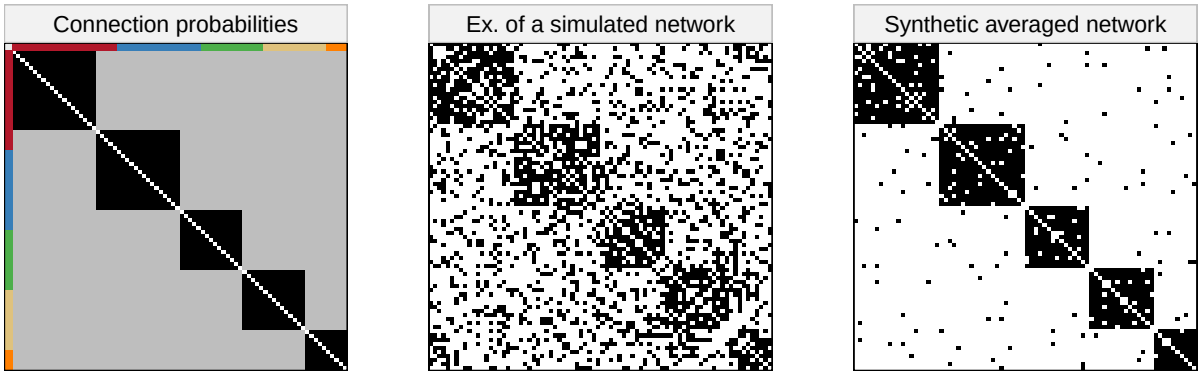
Figure 2.8: Matrices of true connection probabilities and estimate consensus trees of level 0.6. Black refers to high probability (0.75), whereas gray to low probability (0.25). The colours on the sides of the matrices and at the tips of the trees refer to the original group structure of the networks.

groups. Groups (2–3–5) have the same within and between connectivity structure and, in this case, the model is also able to cluster each group correctly.

In order to highlight the benefits of the proposed model, we consider a comparison with an alternative ad–hoc strategy for estimating the multiresolution structure of the nodes of a set of networks. To the best of our knowledge, there are not other latent position models that provide hierarchical clustering of the nodes. Therefore, we consider a modification of the latent cluster model of Handcock et al. (2007) to incorporate the information of multiple networks and infer the tree structure of the nodes. The latent cluster model is a generalization of the latent position model of Hoff et al. (2002), in which the latent positions are sampled from a mixture of Gaussian distributions. The mixture components enforce a clustering structure of the nodes in the latent space.

We adapt the latent cluster model in the following way. We create a synthetic network with an edge occurring when there are more than 5 edges between the two corresponding nodes in the simulated population of $N = 10$ networks. Note that since the true connection probabilities are either 0.75 or 0.25, the synthetic network is characterized on average by a less noisy edge block structure than the single networks (see Figure 2.9). We fit the

(a) Independent groups structure



(b) Dependent groups structure

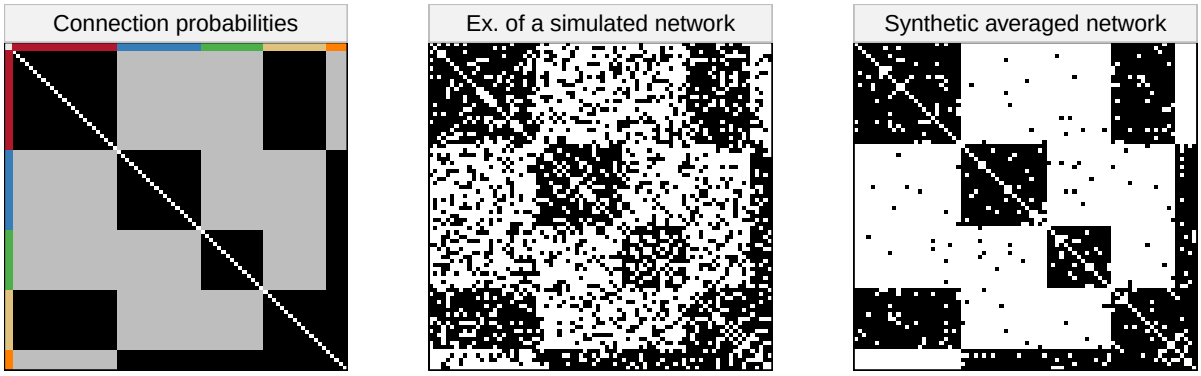


Figure 2.9: Matrices of true connection probabilities (0.75 in black, 0.25 in gray), examples of one of the simulated networks, and synthetic networks generated for the latent cluster model combined with hierarchical clustering.

latent cluster model to such network with $G = 5$ mixture components, corresponding to the exact number of clusters of nodes of the ground truth, using the default options in the LATENTNET R–package (Krivitsky and Handcock, 2008). We consider the maximum likelihood estimate of the latent positions and perform hierarchical clustering based on the Ward criterion (Ward Jr, 1963) with Euclidean distances. We have also considered – as an alternative – the latent cluster model (Handcock et al., 2007) with binomial likelihood applied to the network with weighted edges corresponding to the sum of the binary entries in the adjacency matrices of the $N = 10$ simulated networks. The binomial latent cluster model can be combined with the hierarchical clustering of the estimated latent positions with the same procedure described before. However, the result under the Binomial model provides much worse clustering of the nodes, thereby we do not consider it here.

The dendogram of the hierarchical clustering of the latent cluster model is represented with a dashed line in Figure 2.10. The consensus tree obtained with our model is instead displayed with a solid line. In general, the dendogram obtained from hierarchical clustering has the challenge of identifying which bifurcations are capturing a nested clustering structure, and which are instead only due to the binary–aggregating nature of the algorithm. One common rule–of–thumb is to look at how stable (i.e. long) are the branches in the dendogram. Conversely, the consensus tree for our model captures the uncertainty about the tree structure and, by construction, it presents only the branching structures which have a certain given posterior frequency.

In the simplest scenario of the independent groups structure (Figure 2.10a), both models correctly cluster nodes at the leaf level. The dendogram has the most stable branches corresponding to the 5 groups of nodes, similarly to the consensus tree. However, in the second scenario of Figure 2.10b, the dendogram captures only certain sub–clusters of nodes at the leaf level, while failing at reconstructing any meaningful tree structure. The estimated cluster assignments of the latent cluster model to the mixture components (not shown here) provides better grouping then the dendogram, but it still results in a worse clustering than the one inferred from the consensus tree from our model, and it lacks the desired multiresolution property.

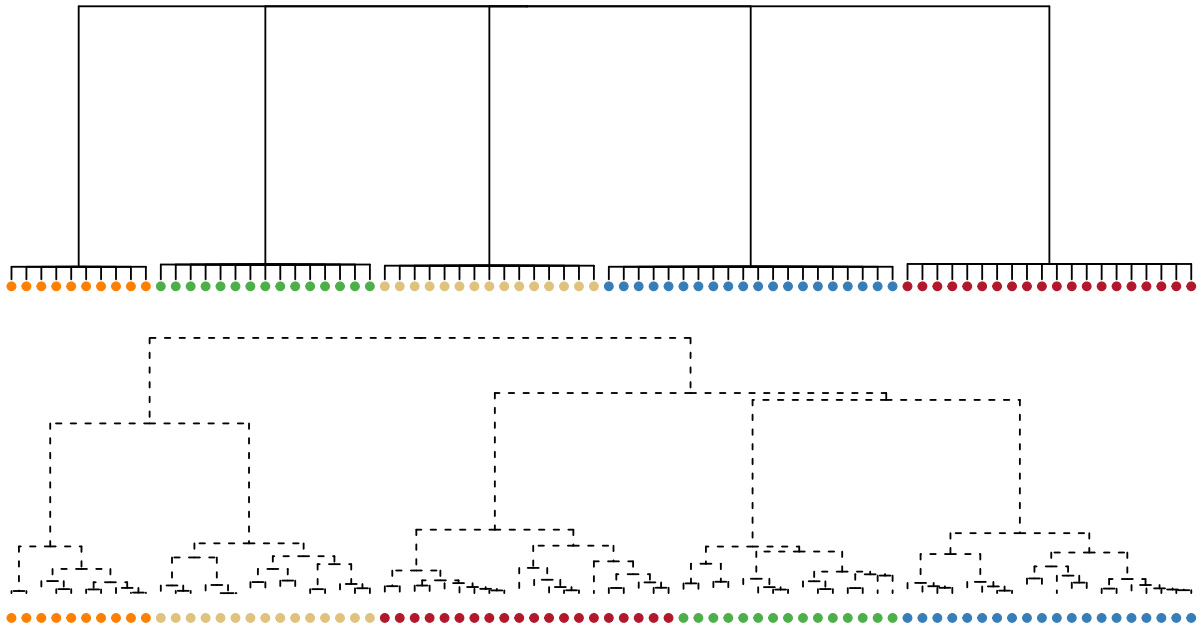
2.5 Brain Connectivity Networks

We analyze the brain connectivity networks from the Enhanced Nathan Kline Institute Rockland Sample project. The brain parcellation in $n = 70$ regions is based on the Desikan atlas (Desikan et al., 2006), which is considered a standard in neuroscience applications. The networks we analyse are obtained as post–processing of two DTI scans of $N = 20$ individuals. Each DTI scan measures the water diffusion between brain regions through white matter fibres. With such a procedure, for each individual it is obtained a weighted network with edges counting the number of white matter fibres connecting each pairs of brain regions, for each of the DTI scans. An illustrative representation of a post–processing procedure similar to our case is given in Figure 2.11, taken from Kim et al. (2016).

In our application, we consider the binary networks having an edge between two regions, if at least one white matter fibre is recorded in one of the two DTI scans, see Figure 2.12. The inherent possibility of measurement errors in the post–processing procedure together with the variability of the brain connectivity across subjects is suitably captured by the random nature of the edges in the statistical model.

For each brain region, i.e. network node, we have a collection of additional informa-

(a) Independent groups structure



(b) Dependent groups structure

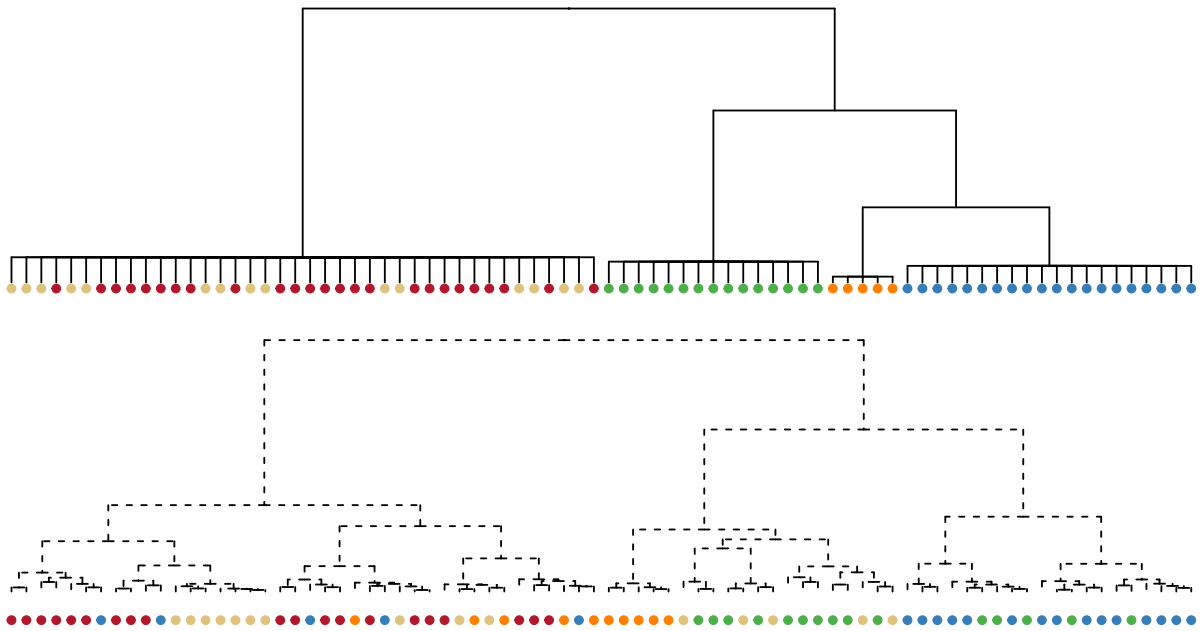


Figure 2.10: Comparison of the inferred hierarchy of the network nodes between our proposal (solid line) and the combination of the latent cluster model of Handcock et al. (2007) with hierarchical clustering (dashed line).

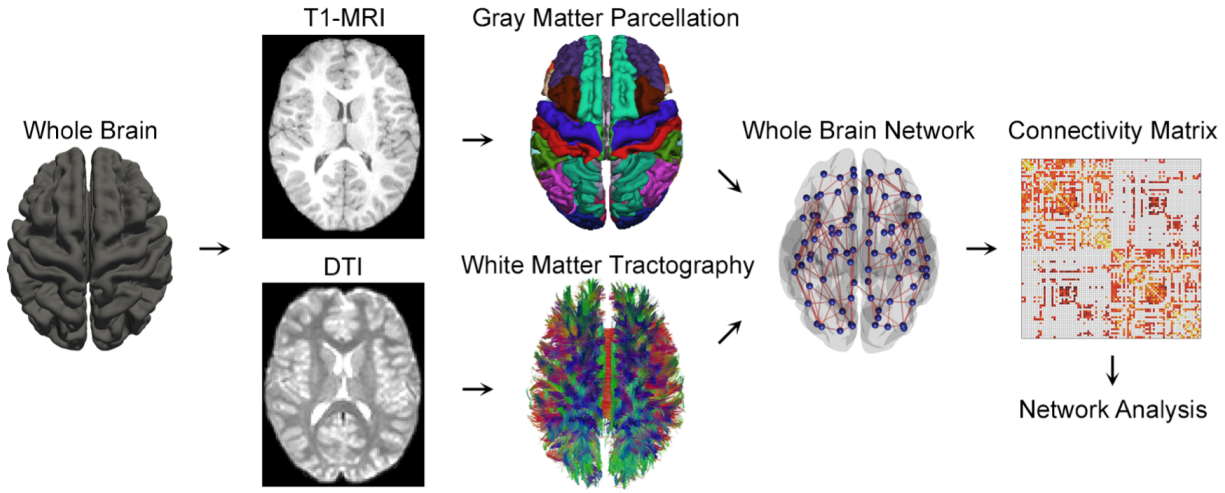


Figure 2.11: Example of a processing procedure to obtain structural brain connectivity networks from DTI scans. Image from Kim et al. (2016).

tion on 3–D spatial locations, hemisphere and lobe memberships. We do not include this information in our model, but instead we leverage them to assess the goodness of fit as they are expected to correlate with the underlying tree characterising the connectivity structure.

We fit the phylogenetic latent position model running four chains initialised at random samples of the parameters, each for 10^5 iterations and with a thinning frequency of 50. We consider a burnin window of 15000 iterations before thinning, based on the results of the MCMC diagnostic analysis. In this case, the sampling has shown slow mixing of the chains. We discuss in Section 2.6 possible future improvements on this side, based on alternative MCMC algorithms.

Figure 2.13 shows the consensus tree obtained from the posterior samples. The tree leaves report the indexing numbers and the names of the brain regions, together with the lobe memberships represented by the coloured boxes on the right. The pie charts located at few internal nodes of the consensus tree show the proportion of brain regions relative to the two hemispheres in the subtree rooted at the given tree node.

The split at the root reveals an uncommon macro–partiton of the brain regions in two groups, roughly corresponding to the frontal and the backward parts of the brain, in contrast with the canonical two–hemispheres division. The smaller subtree – at the bottom – collects mainly brain regions placed in the frontal lobe. The *Rostralanteriorcingulate* and the *Caudalanteriorcingulate*, both for the left and right hemispheres, are the only brain regions in the limbic lobe, specifically placed in the anterior part as the name indicates. The subsequent split generates two subtrees, which reflect the right and left

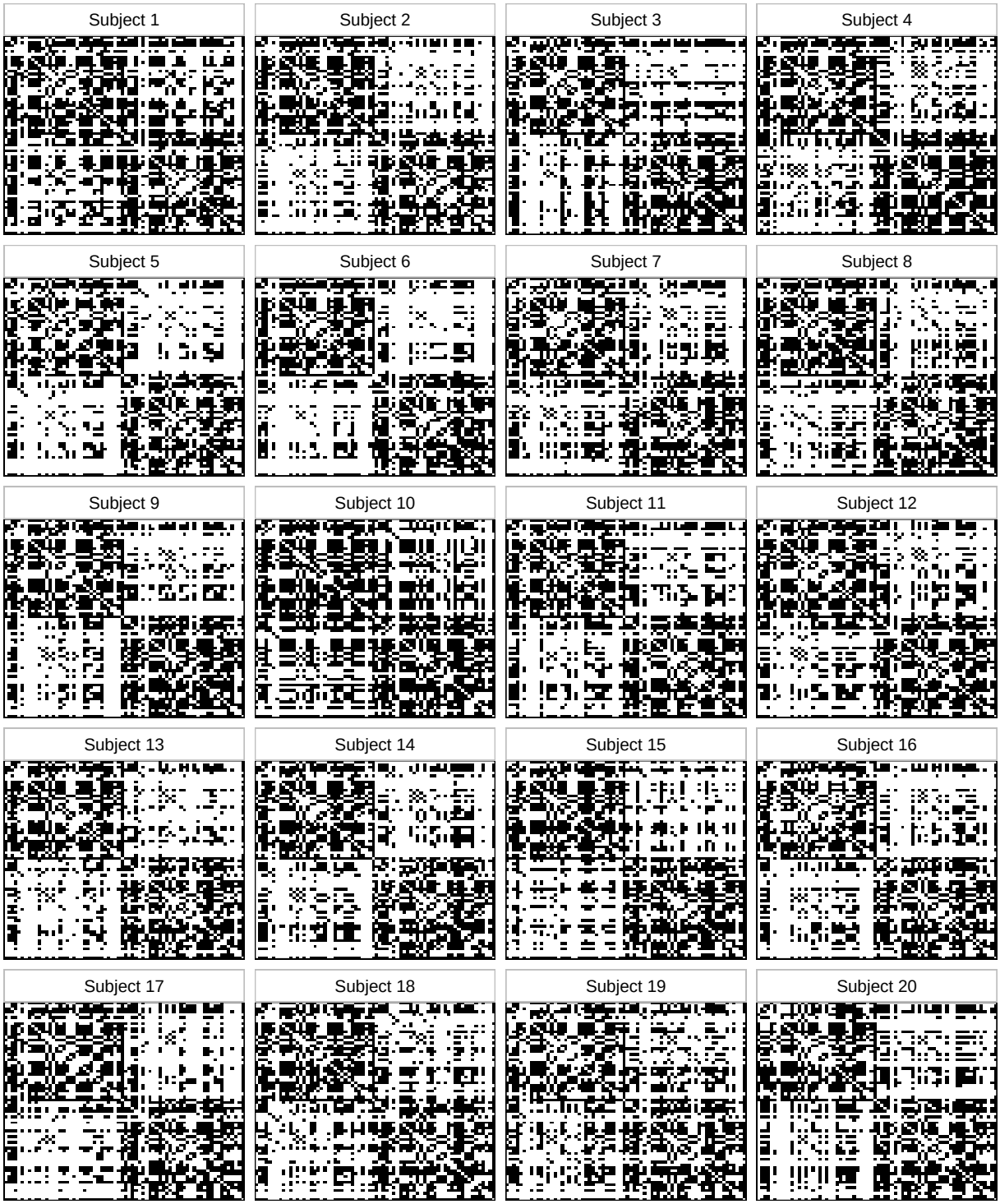


Figure 2.12: Brain connectivity networks for the $N = 20$ patients.

hemisphere structures. The larger group generated from the root split – at the top of the tree – collects regions belonging to all the other lobes, including few frontal ones. Likewise, it subsequently divides the regions in right and left hemispheres. At the leaf level, the lobe membership is partially a grouping factor.

The networks present also two unlabelled regions, corresponding to nodes 1 and 36. These two regions are placed in the inter–hemispheric lobe of the brain, next to the *Corpuscallosum*, in the right and left hemispheres (Desikan et al., 2006). These two regions are placed in the tree in accordance with other regions sharing the same hemisphere membership. Moreover, both of them form a subtree at the leaf level with the *Corpuscallosum*, showing that the proximity in the brain relates to the similar connectivity patterns.

The white matter fibre connectivity of the brain is partially explained by spatial closeness of the brain regions. Connecting regions far apart is more expensive in terms of material and energy costs, thereby highly connected neurones tends to be located closeby (Bullmore and Sporns, 2009). Therefore, we expect that the latent positions of our model partially reflect the true space distribution of the centroids of the brain regions. The left plot in Figure 2.14 compares the posterior average pairwise distances in the latent space against the centroid pairwise distances between brain regions available in the additional data. The high correlation between the two proves that our model is correctly capturing, through the latent space representation, one of the main reasons of the connectivity. The colours refer to the age of the most recent common ancestor of the two nodes involved in the pairwise distance. Age 0 corresponds to the leaves of the tree, and it is denoted with the darkest colour. On the right, Figure 2.14 shows the matrix of the ages of the most recent common ancestors. The rows and the columns corresponds to the brain regions reordered according to the leaf ordering in the consensus tree of Figure 2.13. The block–structures along the diagonal follow the nested grouping given by the tree representation. The closer two brain regions are on the tree, the younger is the most recent common ancestor of the two, and the higher is the correlation of their connectivity. As expected, the largest distances in the left plot correspond to regions that are divided in the tree closer to the root, which means they have an older most recent common ancestor.

2.6 Discussion

We propose a novel latent position model for populations of networks. As illustrated in the simulations and in the analysis of the brain connectivity networks, the model is able to infer meaningful tree structures underlying the connectivity patterns between nodes, and shared across networks. This is achieved by assuming that the node latent positions are sampled from branching Brownian motions, with the branching structure given by

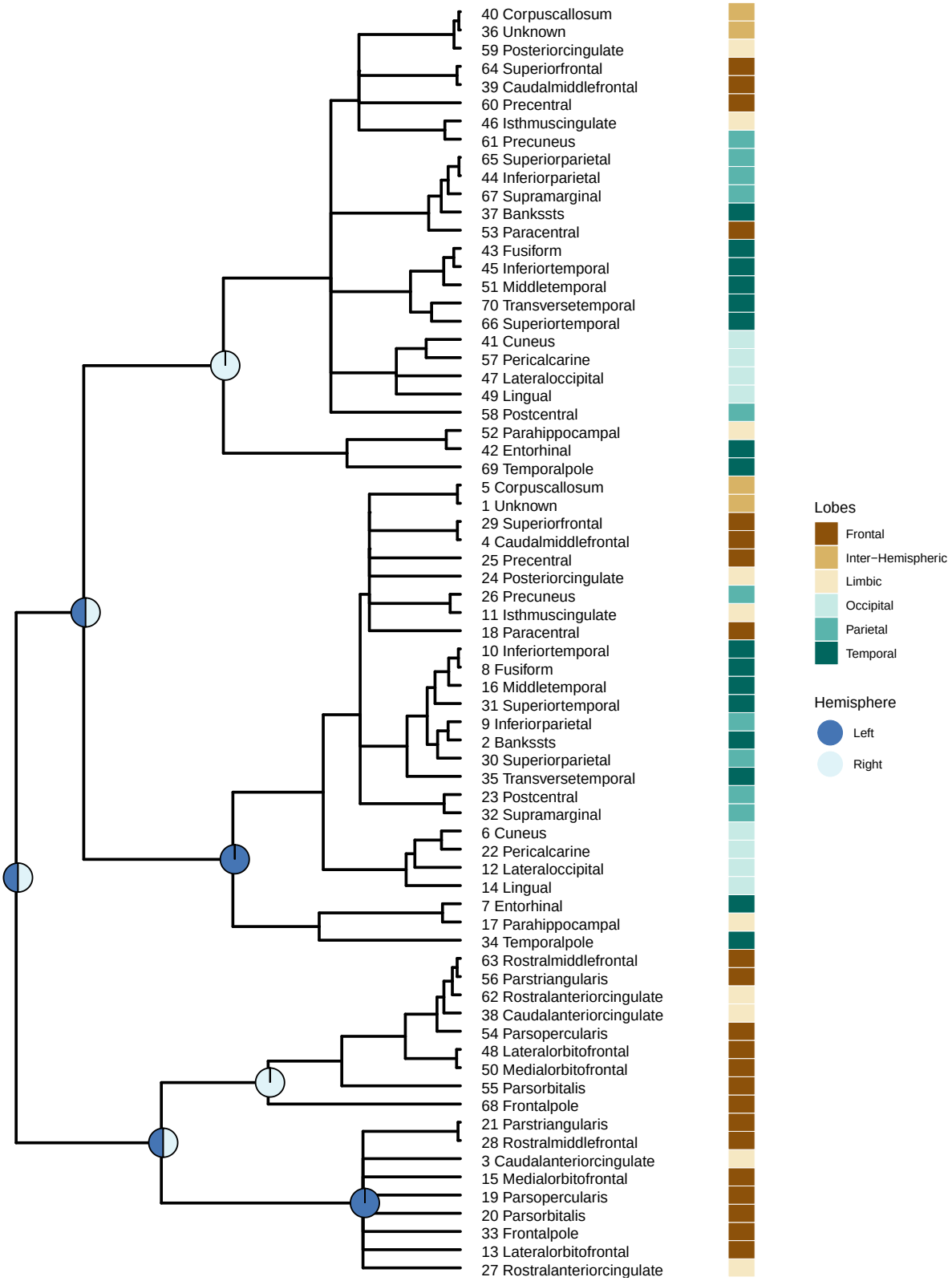


Figure 2.13: Consensus tree of level 0.6 for the brain connectivity networks. Coloured boxes at the tips show the lobe membership, whereas the pie charts at the internal node of the tree report the proportion of nodes in the right and left hemisphere in the respective subtrees.

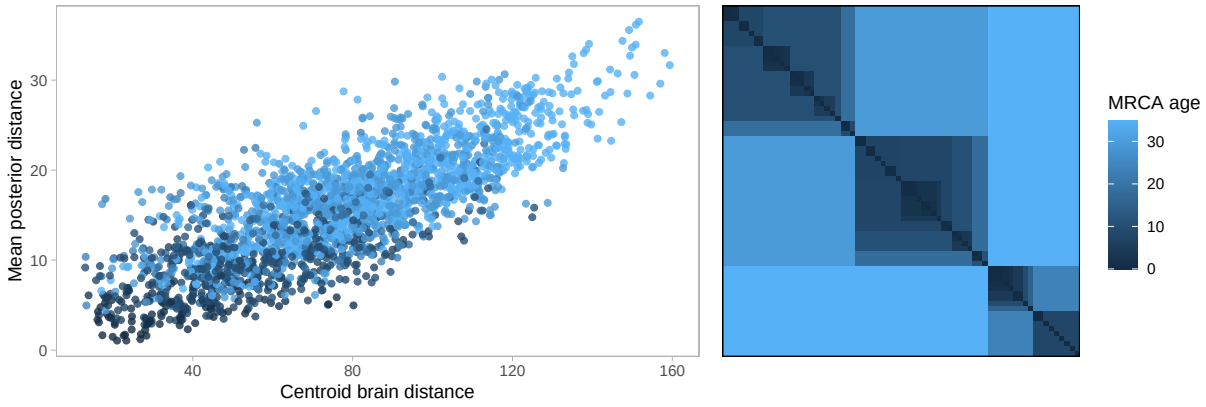


Figure 2.14: On the left, posterior average pairwise distances and centroid distances of the brain regions. On the right, the matrix of the ages of the most recent common ancestors from the consensus tree. Rows and columns of the matrix correspond to the network nodes reordered according to the consensus tree of Figure 2.13.

a phylogenetic tree common to all networks. The Bayesian formulation of the model allows us to obtain posterior samples of the phylogenetic tree. The posterior uncertainty on the tree helps us building summaries which only report branching structures with a posterior frequency higher than a chosen threshold, via the consensus tree. Incorporating the uncertainty, we can better distinguish cases showing absence of hierarchical organisation of the nodes, as shown in Section 2.4. To the best of our knowledge, this is the first work in the latent position model class having all the above features.

In the context of the brain application, the model has revealed an interesting multiresolution view of the brain regions of the Desikan atlas (Desikan et al., 2006). This opens the possibility to consider new different possible macro–partitions of the brain, coherently with the structural connectivity. It would be interesting to study in the future the possible relation between these and the functional connectivity of the brain (Babaeeghazvini et al., 2021). Moreover, several studies has shown that both the functional and structural connectivity of the brain change between healthy subjects and those affected by mental diseases (Stingo et al., 2013; Chekouo et al., 2016; Peterson et al., 2020). It would be of great interest to examine how the multiresolution organization inferred by our model may change in these cases, based on suitable comparisons between phylogenetic trees.

The phylogenetic tree component of the model makes it more challenging to efficiently sample from the posterior. In this work, we have presented a Gibbs sampler which leverages, where possible, parallelization of independent updates and conjugacy. However, there is room for improvements for what matter sampling. State–of–the–art softwares for phylogenetic inference implement advanced MCMC algorithms, such as par-

allel tempering and coupling, in order to improve the computational efficiency and obtain better mixing of the chains (Altekar et al., 2004; Müller and Bouckaert, 2020). These algorithms represent interesting options worth to consider in the future.

There are several directions in which is possible to extend our model. A possibility is to generalise the phylogenetic component of the model allowing for rate correlation across dimensions of the latent space, and rate changes across branches for the Brownian motion processes. Similarly, it would be of interest to study the effect of other prior choices for the phylogenetic tree. Both aspects lead to increased flexibility of the latent positions, allowing to better fit complex connectivity structures. The increased dimension of the parameter space constitutes the main challenge in these directions.

Additionally, it is appealing the idea of leveraging field–knowledge to provide an evolutionary interpretation of the model, whenever applicable. This could open up new ways of forecasting the future evolution of a network. The generating process of the model allows us to further grow the tree after $t = t_0$, the evolutionary time at which the observations are placed. Depending on the tree prior, the growing process can involve adding new nodes to the network, if a new bifurcation happens, and removing some of the observed nodes, if the death rate is positive $d > 0$. It would be interesting to study the properties of these forward projections of the networks and consider possible applications, also in other domains. For instance, in social sciences this may help forecasting how the connectivity structure of sub–communities of people evolves as the number of affiliates increases, represented by new bifurcations in the associated subtree. In ecology, mycorrhizal networks represent connections between trees and mushrooms that are known to change over time as new species join the network (Simard et al., 2012). Under suitable extensions, the model might learn these patterns, allowing us to build predictions through forward projections.

We have discussed in Section 2.3.1 the necessity of repeated measurements (i.e., multiple networks) in order to learn the tree organization of the nodes. Experimental results support the growing ability of the model to recover the underlying tree for increasing number of networks N. An important line of theoretical research consists in studying the asymptotic behaviour of the model in terms of consistency and posterior contraction toward the true tree for the number of networks N growing to infinity, assuming the existence of a tree representing the ground truth. In addition, we believe it is also important to investigate the finite sample regime, in order to provide further guarantees of reliable inference. For instance, it would be interesting to empirically check frequentist finite sample properties of the tree estimator. Moreover, as discussed in Section 2.3.1, we remark the necessity to further study the full identifiability of the tree related to the identifiability of the latent positions.

Similarly to other latent space models (e.g. Handcock et al., 2007), in our proposal we specify the dependence structure at the level of the latent positions. We believe that studying what type of dependence is induced at the level of the observable random variables constitutes an interesting change of perspective that might bring further theoretical understanding of the model. Such reasoning does not restrict to our proposal, but it applies to other latent space models as well.

Finally, there are few interesting research directions concerning the specification of the latent space itself. For instance, the optimal choice of the dimensionality K is a common open problem of latent position models (Kaur et al., 2023). Additionally, the geometry of the latent space has shown to have a strong impact on the type of networks representable by a given dimensionality of the latent space (Smith et al., 2019). In particular, we believe it would be interesting to study the possibility to specify the phylogenetic latent position model using a hyperbolic latent space, as this choice seems to better accommodate tree–like structures in the represented networks (Krioukov et al., 2010; Lubold et al., 2023).

Chapter 3

Incorporating Prior Knowledge of Phylogenetic Structure in Latent Position Models

JOINT WORK WITH DANIELE DURANTE AND ROBIN RYDER

3.1 Introduction

We consider again the brain connectivity networks from the Enhanced Nathan Kline Institute Rockland Sample project, which we analyzed in Chapter 2. The data consist of observations of the connectivity structure between $n = 70$ brain regions, for $N = 20$ patients. The brain parcellation is the same for all subjects. For each brain region, there is additional information on hemisphere and lobe memberships, along with 3–D spatial coordinates.

In this Chapter we aim at leveraging the available covariates in the form of a supervised prior for the phylogenetic tree. This is motivated by the fact that the additional data show accordance with the phylogenetic tree inferred in Chapter 2.5. We first begin by showing the latter in terms of phylogenetic signal. Later, we change perspective and introduce the phylogenetic tree supervised prior.

In Chapter 2, we infer the tree structure underlying the node connectivity patterns based on only observed edges, summarized in the consensus tree of Figure 3.1. The partition induced by the phylogenetic tree reveals clustering patterns that align with the physiological characteristics of the brain regions, such as their hemisphere and lobe memberships, as well as their spatial locations. This observation naturally raises an intriguing question: is there any evidence of phylogenetic signal in the node attributes that corresponds to the phylogenetic tree within our model? Or, equivalently, can we interpret

the node attributes as realizations of a process evolving on the same tree as the latent positions of the network?

We answer these questions by studying the presence of phylogenetic signal under the assumption of a given model specification for each feature. In this case, there are both continuous (x–y–z spatial coordinates) and discrete (hemisphere and lobe memberships) measurements. The simplest model choices for these two types of data, in the context of evolutionary trees, are the Brownian motion (Felsenstein, 1973) and the continuous time Markov chain (CTMC) with discrete state space (Lewis, 2001).

In order to study the presence or absence of phylogenetic signal in the node attributes with respect to the tree inferred from the observed networks, we rely on the widely used Pagel's λ (Pagel, 1999).

Pagel's λ model combines the process describing the evolution of the observed attributes (e.g., BBM or CTMC) with a tree transformation, in which branches are scaled according to $\lambda \in [0, 1]$. Depending on the value of λ, the transformation expands the length of the terminal branches (i.e., those connecting to a leaf) and compresses the internal branches, leaving the total length of the tree unchanged. Figure 3.2 shows an example of the transformation for a given tree, with $\lambda = 0, 0.5, 1$. Values of $\lambda = 1$ and $\lambda = 0$ are the limiting cases. The first one leaves the tree unchanged, while the second one compresses to 0 the lengths of the internal branches, leading to a degenerate tree of star–like shape.

The effect of the transformation is to modulate the dependence structure induced by the tree. As λ decreases to 0, the dependence between the components of the process at the leaves diminishes. Eventually, in the star–like tree case, the original tree becomes a nuisance parameter for the process and the components become fully independent.

This becomes clearer if we consider the example of the Pagel's λ model with a Brownian motion process. We use the same notation of Chapter 2, thereby σ^2 is the rate of the Brownian motion and t_0 is the total height of the tree. If $\lambda = 1$, i.e. no transformations, the components of the Brownian motion at the leaves of a phylogenetic tree g follows a multivariate normal distribution, with covariance matrix given by:

$$\sigma^2 \Sigma_g = \begin{pmatrix} t_0\sigma^2 & t_{12}\sigma^2 & \cdots & t_{1n}\sigma^2 \\ t_{12}\sigma^2 & t_0\sigma^2 & \cdots & t_{2n}\sigma^2 \\ \vdots & \vdots & \ddots & \vdots \\ t_{1n}\sigma^2 & t_{2n}\sigma^2 & \cdots & t_0\sigma^2 \end{pmatrix}, \tag{3.1}$$

where the generic entry t_{ij} denotes the height of the most recent common ancestor in g between leaves i and j. If we denote with $\tilde{g}_\lambda$ the tree obtained by transforming the original tree g with a given λ, the covariance matrix of the components of the Brownian

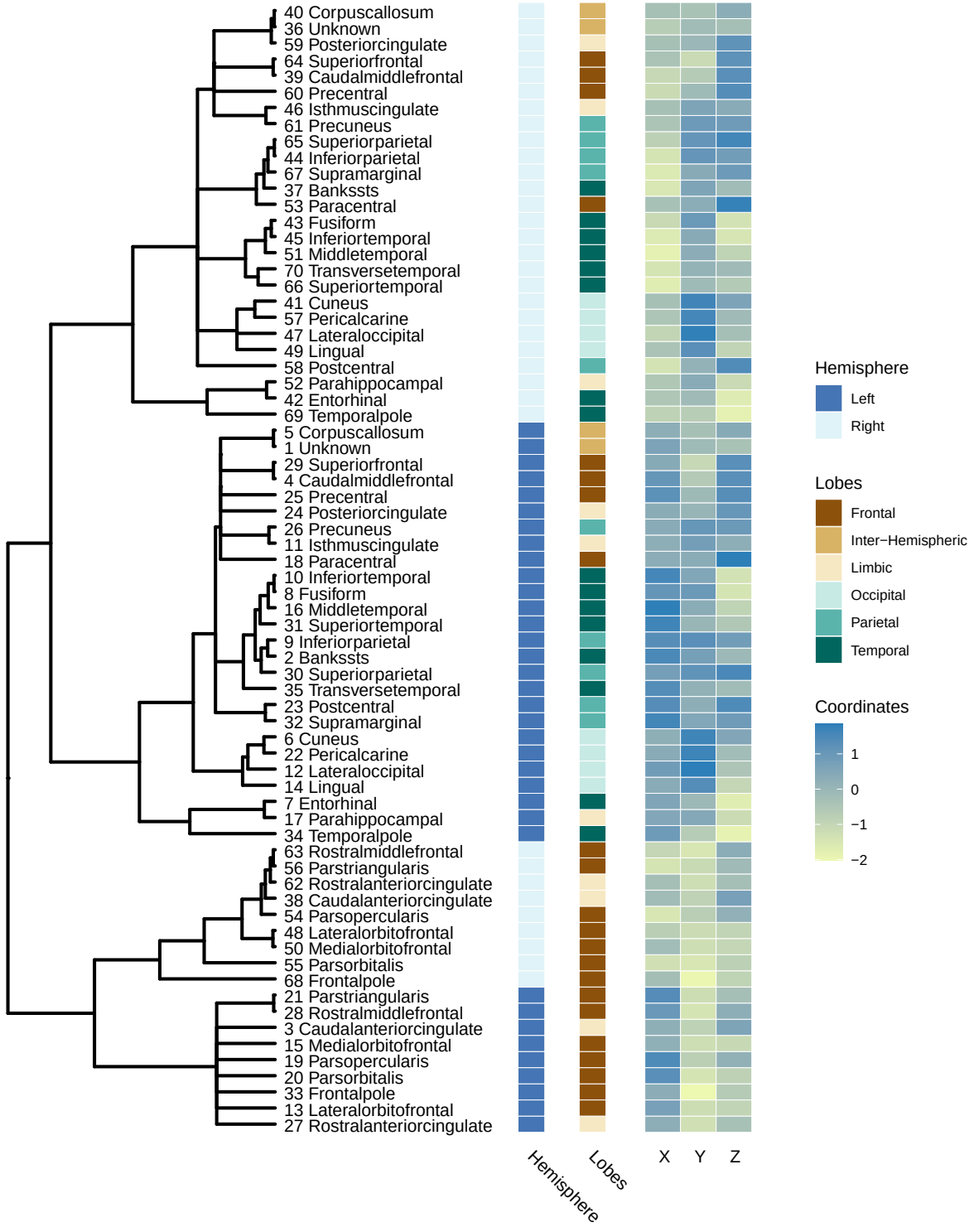


Figure 3.1: Posterior consensus tree at level 0.6 from the analysis of Chapter 2 and node attributes, for the brain networks.

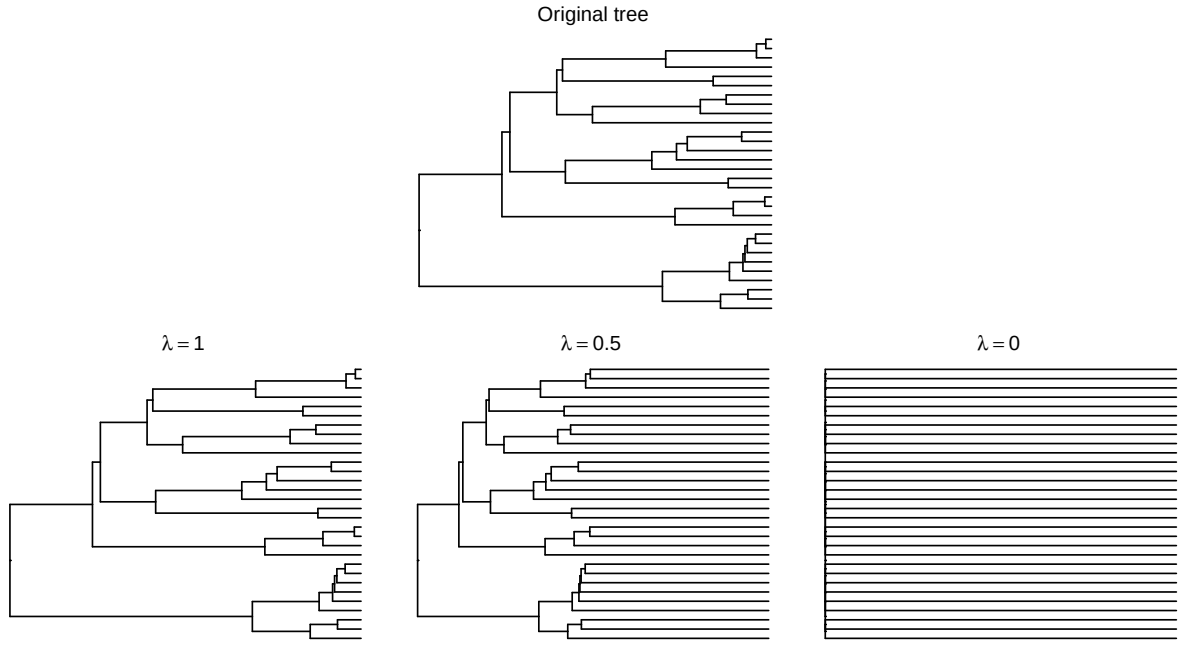


Figure 3.2: Example of Pagel's λ transformations.

motion at the leaves of $\tilde{g}_\lambda$ changes as follows:

$$\sigma^2 \Sigma_{\tilde{g}_\lambda} = \begin{pmatrix} t_0\sigma^2 & \lambda t_{12}\sigma^2 & \cdots & \lambda t_{1n}\sigma^2 \\ \lambda t_{12}\sigma^2 & t_0\sigma^2 & \cdots & \lambda t_{2n}\sigma^2 \\ \vdots & \vdots & \ddots & \vdots \\ \lambda t_{1n}\sigma^2 & \lambda t_{2n}\sigma^2 & \cdots & t_0\sigma^2 \end{pmatrix}. \tag{3.2}$$

The branch transformation is equivalent to changing the covariance matrix multiplying by λ the off diagonal elements. This reduces the covariances, while keeping the marginal variances fixed. For $\lambda = 0$, the covariance matrix becomes diagonal $\sigma^2 \Sigma_{\tilde{g}_0} = t_0 \sigma^2 \mathbf{I}$, corresponding to the case in which each component follows an independent Brownian motion process.

Given data and a fixed tree g, the parameter λ and the rate of the process describing the observations (e.g., σ^2 in the example of the branching Brownian motion) are typically estimated via maximum likelihood. The obtained value of λ can be seen as a measure of the intensity of the phylogenetic signal in the observations, with respect to the tree g. Clearly, the value of the phylogenetic signal is conditioned on the model assumptions for the branching process, thereby the strength of the signal varies under different models.

In our application, we leverage Pagel's λ as a measure of phylogenetic signal as fol-

lows. For each posterior sample of the phylogenic tree g obtained from the model fit in Chapter 2, we perform maximum likelihood estimation for the Pagel's λ model with Brownian motion and CTMC processes, respectively for continuous and discrete attributes. The procedure estimates both the optimal value of λ and the rate parameter of the processes. For the Brownian motions, we assume a constant rate σ^2, one for each of the standardised spatial coordinates x, y, and z. For the hemisphere and lobe membership covariates, we use two CTMC's with state space respectively given by {*Left, Right*} and {*Frontal, Inter–Hemispheric, Limbic, Occipital, Parietal, Temporal*}, and uniform transition rates between states, which are equivalent to M_k models (Lewis, 2001).

Figure 3.3 shows the estimated rates and λ's, over the tree posterior samples. All features show strong phylogenetic signal, corresponding to λ close to 1. Small deviations from 1 suggest that a slightly different tree topology or rate assumptions may better fit the data. The estimated rates are smaller for those features that show in Figure 3.3 a stable grouping structure throughout the tree, see e.g. hemisphere membership and x–locations. Intuitively, a feature that varies slowly from the root to the leaves will have homogeneous values at leaves that are grouped in subtrees close to the tips – i.e. the root of the subtree is a node closer to the leaves than to the tree root. Contrarily, heterogenous values require a fast changing evolution of the process in other to change states over short branches.

Pagel's λ is a useful tool to investigate the presence of phylogenetic signal. However, it is worth noticing that it does not represent a proper statistical testing procedure. Moreover, the branch transformation implied by λ is only one of the possible ways of modifying the tree topology to better fit the data, under model assumptions. We leave to future work the study of the results under different tree transformations (see, e.g. Gittleman and Kot, 1990; Blomberg et al., 2003; Münkemüller et al., 2012) and model specifications.

The division in hemispheres and lobes derives from the known macro–organization of the brain. The strong phylogenetic signal of these attributes shows accordance between the known partitions and the one learnt from observed networks. This proves that the phylogenetic latent position model is able to recover meaningful structures of the brain regions.

The above one is a motivating result that suggests an interesting change of perspective. We want to consider these features as an actual source of information available in the data. Can we include the node attributes in the model in order to inform the latent tree structure?

In the remaining of the Chapter, we provide a first step in this direction. In Section 3.2, we discuss how covariates are typically included in latent position models. In the context of our model, we propose a novel supervised phylogenetic tree prior in order to lever-

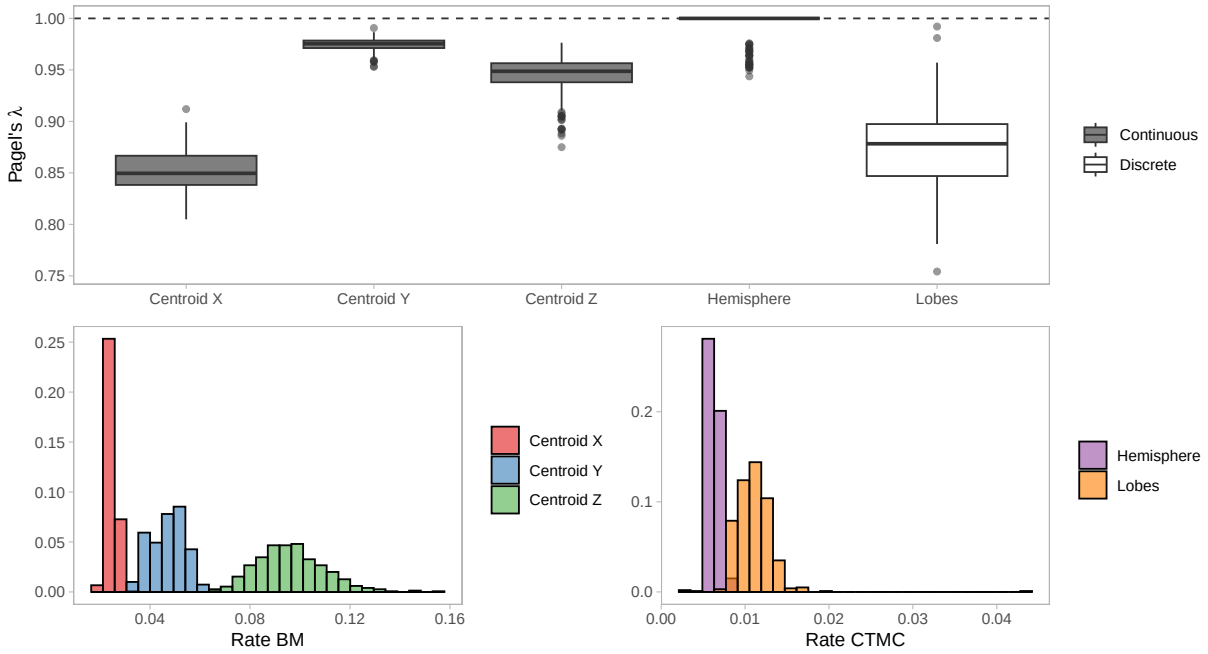


Figure 3.3: Estimated rates and λ's for testing phylogenetic signal in brain connectivity data.

age available features to inform the tree structure. This opens several new interesting research directions, that we discuss in Section 3.3.

3.2 Leveraging Exogenous Covariates in Phylogenetic Latent Position Models

When covariates are included in a statistical model, the underlying assumption is that they are responsible for the distribution of the outcome. The way they are integrated in the model should reflect the type of dependence between them and the response variable.

Let $\mathbf{X}$ be the $q \times n$ matrix of covariates associated with the n network nodes, where q denotes the number of features. In latent position models, covariates are added as extra regression terms affecting the edge probabilities (Hoff et al., 2002; Handcock et al., 2007). Denoting with $\mathbf{x}_i$ the i-th column of $\mathbf{X}$, collecting the q features of node i, the regression terms are typically constructed through a certain map $f(\mathbf{x}_i, \mathbf{x}_j) \in \mathbb{R}^p$ building the $p-$variate information for the edge connecting i and j. For instance, this map can measure the similarity between the covariates of the two nodes.

Using the same notation of Chapter 2, we can include observed features in the phylogenetic latent position model in the same fashion of other latent position models as

follows:

$$\begin{aligned} y_{ij}^{(m)} \mid \mathbf{Z}^{(m)}, a, \boldsymbol{\beta} &\overset{\text{ind}}{\sim} \text{BERN}(\theta_{ij}^{(m)}) \qquad i, j = 1, \dots, n \\ \text{logit}\, \theta_{ij}^{(m)} &= a + \boldsymbol{\beta}^\intercal f(\mathbf{x}_i, \mathbf{x}_j) - \| \mathbf{z}_i^{(m)} - \mathbf{z}_j^{(m)} \| \\ \mathbf{Z}^{(m)} \mid g, \sigma &\overset{\text{iid}}{\sim} \text{BBM}_K(\sigma^2, g) \\ g \mid b &\sim \text{BDT}_n(b, 0), \end{aligned} \tag{3.3}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ is the vector of regression coefficients for the edge attributes $f(\mathbf{x}_i, \mathbf{x}_j)$. The model is completed including suitable priors for $\boldsymbol{\beta}$. Gormley and Murphy (2010) extends this idea by considering additional possibilities to include covariates in the framework of the latent cluster model of Handcock et al. (2007), e.g. affecting the weights of the mixture representation.

The way covariates are included in (3.3) has consequences on the latent representation. In (3.3), the latent positions capture the *residual connection probability* that is not accounted for by the covariates. However, in some cases, it is likely that the covariates correlate with the latent space structure. Proceeding as in (3.3), the correlation is ignored leading to possible harmful consequences for the inference of the latent hierarchical structure of the nodes. For instance, consider the extreme case where the covariates fully describe the connection probabilities of a set of networks, whose nodes have a multiresolution organization. The latent positions would essentially capture random noise, showing absence of any tree–structure regardless of the truth.

A similar point is raised in the discussion of Handcock et al. (2007), in which the latent positions follow a mixture of normal distributions. Similarly to our case, the dependence between the covariates and the mixture assignments can lead to poor clustering of the network nodes. In the discussion, Gormley and Murphy (2007) suggest to allow the mixture probabilities to depend on covariates, while Sylvia and Alex (2007) propose to specify cluster–specific regression coefficients. Handcock et al. (2007) remark in their answer the importance of working in the direction of jointly specifying the dependence between the latent space structure and the covariates.

In the context of the phylogenetic latent position model, we argue that the interplay that matters between features and the latent space structure of the nodes translates into phylogenetic signal of the covariates, with respect to the underlying tree responsible for the latent positions. This is the case, for instance, in the brain connectivity networks, as showed in Section 3.1. Such type of dependence between the covariates and the latent space is not the only possible one. However, if the objective of the inference is the phylogenetic tree g, then this is the primary type of dependence which may negatively impact the tree inference.

One option to account for the dependence between the node attributes $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]$

and the latent positions in a principled way is to specify a *supervised prior* $\pi(g \mid \mathbf{X})$ on the phylogenetic tree. Different ways of defining $\pi(g \mid \mathbf{X})$ are possible. A natural option of doing this is to specify a probabilistic model for the covariates given the tree $\pi(\mathbf{X} \mid g)$, and to define the supervised prior as the posterior of this model combined with the unsupervised prior $\pi(g)$,

$$\pi(g \mid \mathbf{X}) \propto \pi(\mathbf{X} \mid g)\pi(g). \tag{3.4}$$

In general, the term $\pi(\mathbf{X} \mid g)$ can be any function expressing the coherence of the covariates $\mathbf{X}$ with respect to the tree g (also not representing a probabilistic model). It is possible to tune how strongly the covariates should affect the tree by tempering the covariate likelihood $\pi(\mathbf{X} \mid g)$ with a weight ω,

$$\pi(g \mid \mathbf{X}, \omega) \propto \pi(\mathbf{X} \mid g)^{\omega}\pi(g), \tag{3.5}$$

resulting, e.g., in a vanishing effect for $\omega \to 0$.

One advantage of decomposing the supervised prior as in (3.5) is that we only have to slightly change the Gibbs sampler for the unsupervised model in order to sample from the posterior of the supervised model. The only adjustment required is to modify the acceptance probability for the tree g by weighting the acceptance ratio by the ratio of the covariate likelihoods as follows:

$$\alpha_g^s = \min\left\{1, \frac{\prod_{m=1}^{N} \pi(\mathbf{Z}^{(m)} \mid \sigma^2, g^{s+1})\pi(g^{s+1})}{\prod_{m=1}^{N} \pi(\mathbf{Z}^{(m)} \mid \sigma^2, g^{s})\pi(g^{s})} \frac{\pi(\mathbf{X} \mid g^{s+1})^{\omega}}{\pi(\mathbf{X} \mid g^{s})^{\omega}}\right\}, \tag{3.6}$$

while all the other acceptance probabilities remain unchanged. As we have already mentioned in Section 3.1, simple default choices for the covariate likelihood can be Brownian motion processes for continuous features, and continuous time Markov chains for discrete features, with the state space corresponding to the set of possible values.

The supervised prior in (3.4–3.5) can be seen as a two–steps procedure of a fully Bayesian model for the joint observations $(\mathbf{Y}^{(1)}, \dots, \mathbf{Y}^{(N)}, \mathbf{X})$. Figure 3.4 shows the correspondent graphical model representation, in which $\boldsymbol{\phi}$ collects possible parameters for the covariate likelihood $\pi(\mathbf{X} \mid g) := \pi(\mathbf{X} \mid g, \boldsymbol{\phi})$, such as the diffusion parameter of the BM or the transition rates of the CTMC. Leveraging the fully Bayesian view, it is possible to specify priors for $\boldsymbol{\phi}$ and include an additional step in the Gibbs sampler. However, evaluations of likelihoods that depend on trees, such as $\pi(\mathbf{X} \mid g, \boldsymbol{\phi})$ or $\pi(\mathbf{Z}^{(m)} \mid g, \sigma^2)$, are computationally expensive. An additional sampling step for $\boldsymbol{\phi}$, which involves evaluating $\pi(\mathbf{X} \mid g, \boldsymbol{\phi})$, further increases the overall computational cost of the Gibbs sampler.

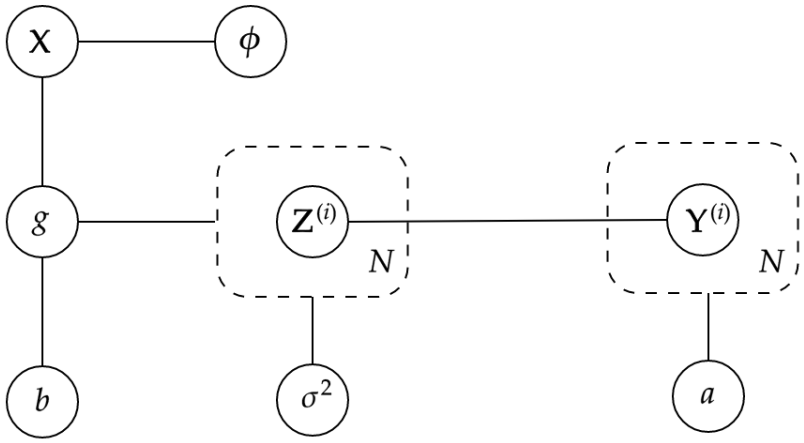


Figure 3.4: Graphical model representation of the equivalent fully Bayesian model.

Even though a complete Bayesian approach is preferable, there are few possibilities for setting $\boldsymbol{\phi}$ in advance and easing computations, at least for a first exploratory model fit. If possible, domain expertises can be leveraged to fix $\boldsymbol{\phi}$ to reasonable values. Otherwise, if the results of the unsupervised model are available, a reasonable heuristic is to set $\boldsymbol{\phi}$ to a value in line with the estimated ones obtained while checking for phylogenetic signal (e.g., the average value), for instance using Pagel's λ (Pagel, 1999). An interesting direction for future research regards the possibility to design procedures in order to estimate $\boldsymbol{\phi}$ in a principled way, and ease the computational costs of the posterior sampling.

It is worth remarking that while the *latent* nature of the nodes latent positions makes a simple process like a constant rate Brownian motion suitable, it is likely that in complex real–world applications *observed* node attributes require more sophisticated model choices than uniform rate CTMC or constant rate BM. Many generalizations are available in the evolutionary model literature (see, e.g. Yang, 2006), such as the early–burst model (Harmon et al., 2010) or the ACDC model (Blomberg et al., 2003), in which the rate of evolution increases or decreases exponentially from the root to the tips, or Pagel's λ–κ–δ, which all perform different types of branch–rate transformations (Pagel, 1999). The discussion on which better suits the covariates strictly depends on the context of the application. The challenges for the implementation of such procedures is left for future research.

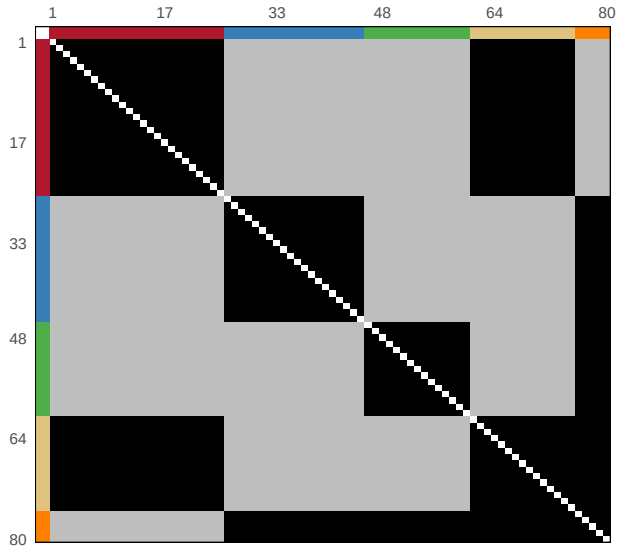


Figure 3.5: Matrix of connection probabilities of the data generating process. High probability of connection (0.75) in black, low probability (0.25) in gray. The group structure (1–5) is represented in the marginal colour bands.

3.2.1 Supervised Tree Prior: A Simulated Data Example

Let us consider the simulated networks with dependent groups structure of Section 2.4, composed of 5 groups of size (25, 20, 15, 15, 5) with a total of $n = 80$ nodes, for $N = 10$ generated networks. Figure 3.5 shows the matrix of connection probabilities of the data generating process, with marginal colours representing the groups. In Section 2.4, the model with unsupervised prior was able to capture most of the structure of the networks, but it failed at separating in different subtrees the first (red) and fourth (brown) groups (see the consensus tree in Figure 2.8b). The difficulty of separating these two groups in the latent space is in part due to the fact that they only differ for the connectivity with the fifty block (orange), composed by only 5 nodes.

We define two binary covariates which partially inform about the true group structure of the networks, in order to compare the posterior under the prior $\pi(g)$ and the supervised prior $\pi(g \mid \mathbf{X})$. In particular, we define the $2 \times n$ matrix $\mathbf{X}$ as follows:

$$[\mathbf{X}]_{ci} = X_c(i), \quad \text{for } i \in \{1, \dots, n\} \text{ and } c \in \{1, 2\}, \tag{3.7}$$

where,

$$X_1(i) = \begin{cases} 1 & \text{if node } i \text{ is in group 4 (beige) or 5 (orange),} \\ 0 & \text{otherwise,} \end{cases}$$
$$X_2(i) = \begin{cases} 1 & \text{if node } i \text{ is in group 1 (red) or 5 (orange),} \\ 0 & \text{otherwise.} \end{cases} \tag{3.8}$$

Both X_1 and X_2 are correlated with the structure explaining the connectivity of the networks. The way they are defined ensures that groups 1 and 4 are distinguishable, whereas there is no information regarding the difference between groups 2 and 3.

We set the supervised prior $\pi(g \mid \mathbf{X})$ using the weighted probabilistic formulation (3.5). Let $\mathbf{X}_c$, for $c \in \{1, 2\}$, be the vector corresponding to the c-th row of $\mathbf{X}$, which is the collection of values of feature c for all nodes. We assume independent models for the covariates as follows:

$$\pi(g \mid \mathbf{X}) \propto \pi(\mathbf{X}_1 \mid g, \phi_1)^{\omega} \pi(\mathbf{X}_2 \mid g, \phi_2)^{\omega} \pi(g), \tag{3.9}$$

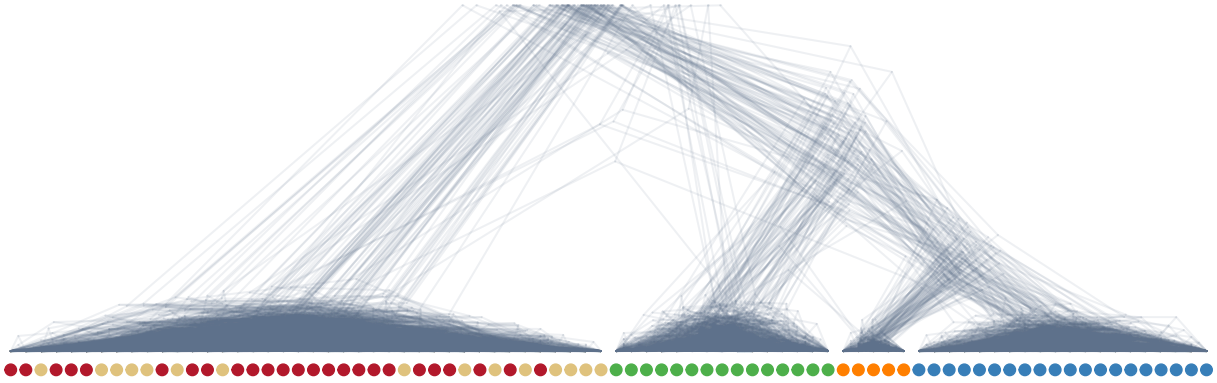
where $\pi(\mathbf{X}_c \mid g, \phi_c)$ stands for the likelihood of the M_2 model (Lewis, 2001), corresponding to a CTMC with homogenous transition rates ϕ_c. Following Chapter 2, we assume the Yule process as unsupervised prior $\pi(g)$. We set $\omega = 10$, in accordance with the fact that the same covariates are observed for all $N = 10$ networks.

As tree–based likelihood evaluations are computationally expensive, we fix the rates ϕ_1 and ϕ_2. In order to choose suitable values, we consider the artificial tree $\tilde{g}$ obtained by modifying the topology of the consensus tree estimated with the unsupervised model (Figure 2.8b) in such a way to separate nodes of blocks 1 and 4 in two different subtrees. We compute the maximum likelihood estimates $\hat{\phi}_1$ and $\hat{\phi}_2$ fitting two M_2 models on $\tilde{g}$ respectively using $\mathbf{X}_1$ and $\mathbf{X}_2$ as data. This strategy allows us to ease computations in the context of this preliminary simulation study.

We fit the model to the simulated networks with the supervised prior $\pi(g \mid \mathbf{X})$, setting $\phi_1 = \hat{\phi}_1$ and $\phi_2 = \hat{\phi}_2$. As the tree total height is not identified, when fitting the model we fix the tree height to the one of $\tilde{g}$. In such a way, we ensure that ϕ_1 and ϕ_2 are on a suitable scale.

Figure 3.6 shows a subset of posterior tree samples under $\pi(g)$ – Figure 3.6a – and under the supervised prior $\pi(g \mid \mathbf{X})$ – Figure 3.6b. Comparing the two plots, we can see the effect of the covariates on the posterior. Under the supervised prior, groups 1 (red) and 4 (brown) split in two different subtrees. This is something which is not required by the latent positions themselves – they do not split under $\pi(g)$ – but rather by the covariates. Under $\pi(g)$, group 3 (green) separates from groups 5 and 2 (orange and blue),

(a) Posterior under prior $\pi(g)$



(b) Posterior under supervised prior $\pi(g \mid \mathbf{X})$

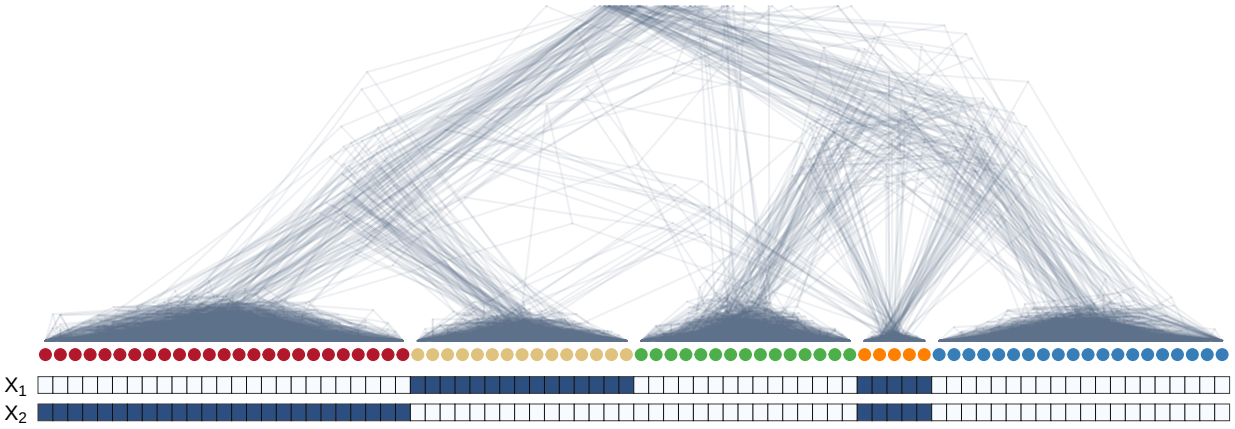


Figure 3.6: Posterior samples of the phylogenetic tree under the prior $\pi(g)$ and the supervised prior $\pi(g \mid \mathbf{X})$. At the bottom, the node–specific binary covariates $\mathbf{X}_1$, $\mathbf{X}_2$ (values 1 and 0 respectively in dark and light colours).

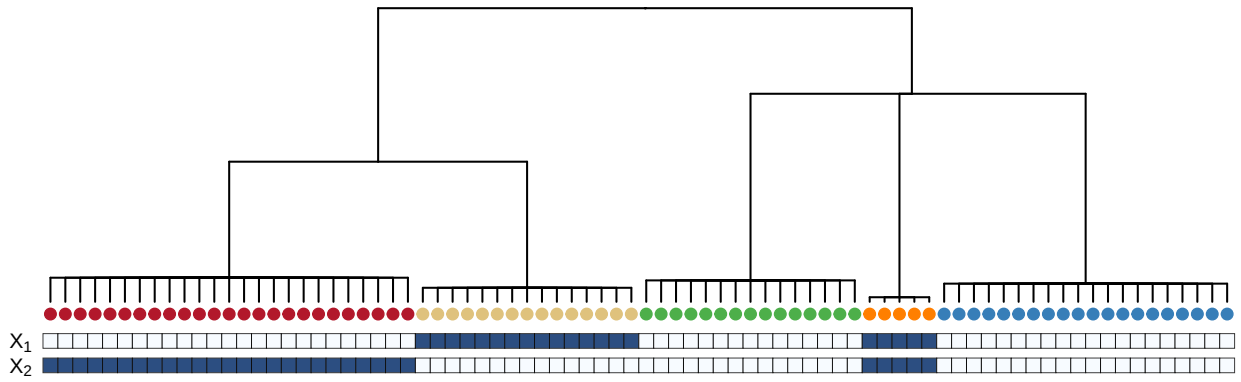


Figure 3.7: Consensus tree at level 0.6 under supervised prior $\pi(g \mid \mathbf{X})$. At the bottom, the node–specific binary covariates $\mathbf{X}_1$, $\mathbf{X}_2$ (values 1 and 0 respectively in dark and light colours).

and afterward groups 5 and 2 split. The topology is the result of a trade–off between the latent positions disposal and the prior on the tree. Indeed, while groups (2, 3, 5) clearly show different connectivity patterns from groups 1 and 4 (Figure 3.5), the ordering in which they split in the tree is mostly affected by the prior, rather than by the observed networks. Under the supervised prior, instead, the three groups split in any of the two ordering (2,(3,5)) or (3,(2,5)) with roughly the same frequencies, corresponding to a trifurcation in the majority–rule consensus tree (Figure 3.7). The covariates do not privilege any of the two orderings and assign the same value to groups 2 (blue) and 3 (green). It is worth noticing that, in this example, the values of the log–likelihood for the two models are roughly the same, indicating that both models similarly fit the data. In Figure 3.6b, we can also notice a small number of sampled trees connecting among them groups with same values in $\mathbf{X}_1$ and $\mathbf{X}_2$, but in contrast with the dominant topology learnt from the connectivity patterns through the latent positions.

These preliminary results demonstrate the potential of utilizing a supervised prior to leverage exogenous covariates in order to inform the tree structure responsible for the latent positions. Nevertheless, it is important to acknowledge the challenges that accompany this methodology, highlighting the need for further research, which we discuss in Section 3.3.

3.3 Discussion

Including covariates in the model as additional regression terms can be harmful for the tree inference, if they correlate with the structure of the latent positions. However, this is a difficult condition to check in advance. A possible practical way to decide whether

to use supervised or unsupervised prior on the tree consists in fitting the phylogenetic latent position model to only observed edges, and check a posteriori the strength of the phylogenetic signal, following some of the methods we mentioned in Section 3.1 (see, e.g. Münkemüller et al., 2012).

In Chapter 2, we discussed how the ability of the phylogenetic latent position model to identify the tree structure of the nodes increases with the number of observed networks. The inclusion of covariates through the supervised prior (3.5) is equivalent to add observations of the dependence structure induced by the tree, in addition to the node latent positions. The combination of these two sources of information might result both in more concentrated posterior distributions or more uncertainty, depending on the agreement between the tree structure of the latent positions and of the covariates.

The influence of the covariates can be tuned by weighting the covariate likelihood with ω. However, the choice of the weight is not trivial. Intuitively, N networks modelled with a K dimensional latent space contribute to the tree with NK sets of n latent coordinates, while q node–specific covariates with weight ω contribute with ωq sets of n values. Reasoning about the ratio $NK/\omega q$ gives a rule of thumb to balance the contribution of the latent positions and the covariates, for what matters the tree inference. Better understanding the effect of ω constitutes an important research topic, which hopefully can lead to design principled procedures to choose its value, possibly according to prior knowledge regarding the domain of the application. Additionally, future research is required to study the sensitivity of the model related to the choice of ω and the number of observed networks N.

Another significant challenge is the increased computational cost of the Gibbs sampler. The computations of likelihoods based on trees are expensive operations. Under the supervised tree priors, the calculation of the acceptance probability for the tree g necessitates q additional evaluations of tree–based likelihoods. Moreover, treating $\boldsymbol{\phi}$ as random requires an additional sampling step which further amplifies the computational burden. To mitigate these effects, one potential strategy, especially in cases where the model is being initially explored, is to pre–specify and fix the value of $\boldsymbol{\phi}$ in advance. In Section 3.2, we propose few heuristics tailored to specific scenarios. However, future research should delve into developing suggestions and principled solutions that can be applicable to general settings.

Lastly, it is important to note that utilizing a supervised tree prior does not preclude the inclusion of covariates within the logistic regression term. Instead, it opens up an intriguing avenue for future research to explore the optimal integration of these two components. For instance, inspired by Sylvia and Alex (2007), one interesting direction worth investigating involves specifying distinct sets of regression coefficients reflecting

the macro partitions defined by the tree structure close to the root. The underlying idea is that the effect of a given covariate may vary for nodes that are located at different distances within the tree. This proposal represents just one among numerous potential ideas that can guide an engaging path of research in this area.

Bibliography

Airoldi, E. M., Blei, D., Fienberg, S., and Xing, E. (2008). Mixed membership stochastic blockmodels. *Advances in Neural Information Processing Systems*, 21.

Aldous, D. J. (2001). Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. *Statistical Science*, pages 23–34.

Alexopoulos, A., Dellaportas, P., and Forster, J. J. (2019). Bayesian forecasting of mortality rates by using latent Gaussian models. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182(2):689–711.

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

Aliverti, E., Mazzuco, S., and Scarpa, B. (2022). Dynamic modeling of mortality via mixtures of skewed distribution functions. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 185(3):1030–1048.

Altekar, G., Dwarkadas, S., Huelsenbeck, J. P., and Ronquist, F. (2004). Parallel metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics*, 20(3):407–415.

Andrieu, C. and Thoms, J. (2008). A tutorial on adaptive MCMC. *Statistics and Computing*, 18:343–373.

Arroyo, J., Athreya, A., Cape, J., Chen, G., Priebe, C. E., and Vogelstein, J. T. (2021). Inference for multiple heterogeneous networks with a common invariant subspace. *The Journal of Machine Learning Research*, 22(1):6303–6351.

Babaeeghazvini, P., Rueda-Delgado, L. M., Gooijers, J., Swinnen, S. P., and Daffertshofer, A. (2021). Brain structural and functional connectivity: A review of combined works of diffusion magnetic resonance imaging and electro-encephalography. *Frontiers in Human Neuroscience*, 15:721206.

Betzel, R. F. and Bassett, D. S. (2017). Multi-scale brain networks. *NeuroImage*, 160:73–83.

Blomberg, S. P., Garland Jr, T., and Ives, A. R. (2003). Testing for phylogenetic signal in comparative data: Behavioral traits are more labile. *Evolution*, 57(4):717–745.

Booth, H. and Tickle, L. (2008). Mortality modelling and forecasting: A review of methods. *Annals of Actuarial Science*, 3(1-2):3–43.

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M. A., Rambaut, A., and Drummond, A. J. (2014). BEAST 2: A software platform for Bayesian evolutionary analysis. *PLoS Computational Biology*, 10(4):e1003537.

Bouckaert, R. R. (2010). DensiTree: Making sense of sets of phylogenetic trees. *Bioinformatics*, 26(10):1372–1373.

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799.

Brooks, S., Gelman, A., Jones, G., and Meng, X.-L. (2011). *Handbook of Markov chain Monte Carlo*. CRC Press.

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics*, 7(4):434–455.

Brouhns, N., Denuit, M., and Vermunt, J. K. (2002). A Poisson log-bilinear regression approach to the construction of projected lifetables. *Insurance: Mathematics and Economics*, 31(3):373–393.

Bullmore, E. and Sporns, O. (2009). Complex brain networks: Graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3):186–198.

Cairns, A. J., Blake, D., and Dowd, K. (2006). A two-factor model for stochastic mortality with parameter uncertainty: Theory and calibration. *Journal of Risk and Insurance*, 73(4):687–718.

Cairns, A. J., Blake, D., Dowd, K., Coughlan, G. D., Epstein, D., Ong, A., and Balevich, I. (2009). A quantitative comparison of stochastic mortality models using data from England and Wales and the United States. *North American Actuarial Journal*, 13(1):1–35.

Camarda, C. G. (2019). Smooth constrained mortality forecasting. *Demographic Research*, 41:1091–1130.

Carboni, L., Achard, S., and Dojat, M. (2021). Network embedding for brain connectivity. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1722–1725. IEEE.

Carboni, L., Dojat, M., and Achard, S. (2023). Nodal-statistics-based equivalence relation for graph collections. *Physical Review E*, 107(1):014302.

Case, A. and Deaton, A. (2021). *Deaths of despair and the future of capitalism*. Princeton University Press.

Chekouo, T., Stingo, F. C., Guindani, M., and Do, K.-A. (2016). A Bayesian predictive model for imaging genetics with application to schizophrenia. *The Annals of Applied Statistics*, 10(3):1547 – 1571.

Chopin, N. and Papaspiliopoulos, O. (2020). *An introduction to sequential Monte Carlo*. Springer.

Clauset, A., Moore, C., and Newman, M. E. (2008). Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191):98–101.

Conti, S., Farchi, G., and Prati, S. (1994). AIDS as a leading cause of death among young adults in Italy. *European Journal of Epidemiology*, 10(6):669–673.

Conti, S., Masocco, M., Farchi, G., Rezza, G., and Toccaceli, V. (1997). Premature mortality in Italy during the first decade of the AIDS epidemic: 1984-1993. *International Journal of Epidemiology*, 26(4):873–879.

Craddock, R. C., Jbabdi, S., Yan, C.-G., Vogelstein, J. T., Castellanos, F. X., Di Martino, A., Kelly, C., Heberlein, K., Colcombe, S., and Milham, M. P. (2013). Imaging human connectomes at the macroscale. *Nature Methods*, 10(6):524–539.

Critchlow, D. E., Pearl, D. K., and Qian, C. (1996). The triples distance for rooted bifurcating phylogenetic trees. *Systematic Biology*, 45(3):323–334.

Currie, I. D. (2016). On fitting generalized linear and non-linear models of mortality. *Scandinavian Actuarial Journal*, 2016(4):356–383.

Currie, I. D., Durban, M., and Eilers, P. H. (2004). Smoothing and forecasting mortality rates. *Statistical Modelling*, 4(4):279–298.

Czado, C., Delwarde, A., and Denuit, M. (2005). Bayesian Poisson log-bilinear mortality projections. *Insurance: Mathematics and Economics*, 36(3):260–284.

Dellaportas, P., Smith, A. F., and Stavropoulos, P. (2001). Bayesian analysis of mortality data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 164(2):275–291.

Delwarde, A., Denuit, M., and Eilers, P. (2007). Smoothing the Lee–Carter and Poisson log-bilinear models for mortality forecasting: A penalized log-likelihood approach. *Statistical Modelling*, 7(1):29–48.

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., et al. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, 31(3):968–980.

Drefahl, S., Ahlbom, A., and Modig, K. (2014). Losing ground-Swedish life expectancy in a comparative perspective. *PloS One*, 9(2):e88357.

Drummond, A. J., Rambaut, A., Shapiro, B., and Pybus, O. G. (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular Biology and Evolution*, 22(5):1185–1192.

Durante, D. and Dunson, D. B. (2014). Nonparametric Bayes dynamic modelling of relational data. *Biometrika*, 101(4):883–898.

Durante, D. and Dunson, D. B. (2016). Locally adaptive dynamic networks. *The Annals of Applied Statistics*, pages 2203–2232.

Durante, D. and Dunson, D. B. (2018). Bayesian inference and testing of group differences in brain networks. *Bayesian Analysis*, 13(1).

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

Durbin, J. and Koopman, S. J. (2012). *Time Series Analysis by State Space Methods*. Oxford University Press.

Eliers, P. and Marx, B. (1996). Flexible smoothing using B-splines and penalized likelihood (with comments and rejoinder). *Statistical Science*, 11(2):89–121.

Felsenstein, J. (1973). Maximum-likelihood estimation of evolutionary trees from continuous characters. *American Journal of Human Genetics*, 25(5):471.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17:368–376.

Felsenstein, J. (2004). *Inferring phylogenies*, volume 2. Sinauer Associates Sunderland, MA.

Fosdick, B. K., McCormick, T. H., Murphy, T. B., Ng, T. L. J., and Westling, T. (2019). Multiresolution network models. *Journal of Computational and Graphical Statistics*, 28(1):185–196.

Frank, O. and Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association*, 81(395):832–842.

Gearty, W., O'Meara, B., Berv, J., Ballen, G. A., Ferreira, D., Lapp, H., Schmitz, L., Smith, M. R., Upham, N. S., and Nations, J. A. (2023). CRAN task view: Phylogenetics. https://cran.r-project.org/web/views/Phylogenetics.html. Accessed: 04-05-2023.

Ginsborg, P. (1990). *A history of contemporary Italy: 1943-80*. Penguin UK.

Gittleman, J. L. and Kot, M. (1990). Adaptation: Statistics and a null model for estimating phylogenetic effects. *Systematic Zoology*, 39(3):227–241.

Glei, D. A. (2022). The US midlife mortality crisis continues: Excess cause-specific mortality during 2020. *American Journal of Epidemiology*, In press.

Goldstein, J. R. and Lee, R. D. (2020). Demographic perspectives on the mortality of COVID-19 and other epidemics. *Proceedings of the National Academy of Sciences*, 117(36):22035–22041.

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

Gormley, I. C. and Murphy, T. B. (2007). Discussion on the paper by Handcock, Raftery and Tantrum. *Journal of the Royal Statistical Society: Series A Statistics in Society*, 170(2):327.

Gormley, I. C. and Murphy, T. B. (2010). A mixture of experts latent position cluster model for social network data. *Statistical Methodology*, 7(3):385–405.

Haberman, S. and Renshaw, A. (2011). A comparative study of parametric mortality projection models. *Insurance: Mathematics and Economics*, 48(1):35–55.

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

Harmon, L. J., Losos, J. B., Jonathan Davies, T., Gillespie, R. G., Gittleman, J. L., Bryan Jennings, W., Kozak, K. H., McPeek, M. A., Moreno-Roark, F., Near, T. J., Purvis, A., Ricklefs, R. E., Schluter, D., Schulte II, J. A., Seehausen, O., Sidlauskas, B. L., Torres-Carvajal, O., Weir, J. T., and Mooers, A. O. (2010). Early bursts of body size and shape evolution are rare in comparative data. *Evolution*, 64(8):2385–2396.

Harris, T. E. (1963). *The theory of branching processes*, volume 6. Springer Berlin.

Heligman, L. and Pollard, J. H. (1980). The age pattern of mortality. *Journal of the Institute of Actuaries*, 107(1):49–80.

Helske, J. (2017). KFAS: exponential family state space models in R. *Journal of Statistical Software*, 78(10):1–39.

Herlau, T., Mørup, M., Schmidt, M. N., and Hansen, L. K. (2012). Detecting hierarchical structure in networks. In *2012 3rd International Workshop on Cognitive Information Processing (CIP)*, pages 1–6. IEEE.

Hilgetag, C.-C., O'Neill, M. A., and Young, M. P. (2000). Hierarchical organization of macaque and cat cortical sensory systems explored with a novel network processor. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 355(1393):71–89.

Hillis, D. M., Heath, T. A., and John, K. S. (2005). Analysis and visualization of tree space. *Systematic Biology*, 54(3):471–482.

Ho, J. Y. and Preston, S. H. (2010). US mortality in an international context: Age variations. *Population and Development Review*, 36(4):749–773.

Hoff, P. (2007). Modeling homophily and stochastic equivalence in symmetric relational data. *Advances in Neural Information Processing Systems*, 20.

Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098.

Höhna, S., Landis, M. J., Heath, T. A., Boussau, B., Lartillot, N., Moore, B. R., Huelsenbeck, J. P., and Ronquist, F. (2016). RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Systematic Biology*, 65(4):726–736.

Holford, T. R. (1983). The estimation of age, period and cohort effects for vital rates. *Biometrics*, pages 311–324.

Hunt, A. and Blake, D. (2021). On the structure and classification of mortality models. *North American Actuarial Journal*, 25(sup1):S215–S234.

Hyndman, R. J., Booth, H., Tickle, L., and Maindonald, J. (2014). Demography: Forecasting mortality, fertility, migration and population data. *R Package version 1.18, https://CRAN.R-project.org/package=demography*.

Hyndman, R. J., Booth, H., and Yasmeen, F. (2013). Coherent mortality forecasting: The product-ratio method with functional time series models. *Demography*, 50(1):261–283.

Hyndman, R. J. and Ullah, M. S. (2007). Robust forecasting of mortality and fertility rates: A functional data approach. *Computational Statistics & Data Analysis*, 51(10):4942–4956.

Juul, F. E., Jodal, H. C., Barua, I., Refsum, E., Olsvik, Ø., Helsingen, L. M., Løberg, M., Bretthauer, M., Kalager, M., and Emilsson, L. (2022). Mortality in Norway and Sweden during the COVID-19 pandemic. *Scandinavian Journal of Public Health*, 50(1):38–45.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engeneering*, 82(1):35–45.

Katzmarzyk, P. T., Salbaum, J. M., and Heymsfield, S. B. (2020). Obesity, noncommunicable diseases, and COVID-19: A perfect storm. *American Journal of Human Biology*, 32(5):e23484.

Kaur, H., Rastelli, R., Friel, N., and Raftery, A. E. (2023). Latent position network models. *arXiv preprint arXiv:2304.02979*.

Kelly, L. J., Ryder, R. J., and Clarté, G. (2023). Lagged couplings diagnose Markov chain Monte Carlo phylogenetic inference. *The Annals of Applied Statistics*, 17(2):1419–1443.

Kendall, D. G. (1949). Stochastic processes and population growth. *Journal of the Royal Statistical Society. Series B (Methodological)*, 11(2):230–282.

Kim, D.-J., Yu, J. H., Shin, M.-S., Shin, Y.-W., and Kim, M.-S. (2016). Hyperglycemia reduces efficiency of brain networks in subjects with type 2 diabetes. *PLoS One*, 11(6):e0157268.

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

Kjærgaard, S., Ergemen, Y. E., Kallestrup-Lamb, M., Oeppen, J., and Lindahl-Jacobsen, R. (2019). Forecasting causes of death by using compositional data analysis: The case of cancer deaths. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 68(5):1351–1370.

Koopman, S. J. and Durbin, J. (2000). Fast filtering and smoothing for multivariate state space models. *Journal of Time Series Analysis*, 21(3):281–296.

Krioukov, D., Papadopoulos, F., Kitsak, M., Vahdat, A., and Boguñá, M. (2010). Hyperbolic geometry of complex networks. *Physical Review E*, 82(3):036106.

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

Krivitsky, P. N., Handcock, M. S., Raftery, A. E., and Hoff, P. D. (2009). Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. *Social Networks*, 31(3):204–213.

Land, K. C. (1986). Methods for national population forecasts: A review. *Journal of the American Statistical Association*, 81(396):888–901.

Lanfear, R., Hua, X., and Warren, D. L. (2016). Estimating the effective sample size of tree topologies from Bayesian phylogenetic analyses. *Genome Biology and Evolution*, 8(8):2319–2332.

Lang, S. and Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics*, 13(1):183–212.

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

Lee, R. and Miller, T. (2001). Evaluating the performance of the Lee-Carter method for forecasting mortality. *Demography*, 38(4):537–549.

Lee, R. D. and Carter, L. R. (1992). Modeling and forecasting US mortality. *Journal of the American Statistical Association*, 87(419):659–671.

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

Lewis, P. O. (2001). A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic Biology*, 50(6):913–925.

Li, M., Xu, X., Cao, Z., Chen, R., Zhao, R., Zhao, Z., Dang, X., Oishi, K., and Wu, D. (2023). Multi-modal multi-resolution atlas of the human neonatal cerebral cortex based on microstructural similarity. *NeuroImage*, 272:120071.

Li, N. and Lee, R. (2005). Coherent mortality forecasts for a group of populations: An extension of the Lee-Carter method. *Demography*, 42(3):575–594.

Li, N., Lee, R., and Gerland, P. (2013). Extending the Lee-Carter method to model the rotation of age patterns of mortality decline for long-term projections. *Demography*, 50(6):2037–2051.

Liang, Y., Sun, D., He, C., and Schootman, M. (2014). Modeling bounded outcome scores using the binomial-logit-normal distribution. *Chilean Journal of Statistics*, 5:3–14.

Lubold, S., Chandrasekhar, A. G., and McCormick, T. H. (2023). Identifying the latent space geometry of network models through analysis of curvature. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(2):240–292.

Lunagómez, S., Olhede, S. C., and Wolfe, P. J. (2021). Modeling network populations via graph distances. *Journal of the American Statistical Association*, 116(536):2023–2040.

Mazzuco, S., Scarpa, B., and Zanotto, L. (2018). A mortality model based on a mixture distribution function. *Population Studies*, 72(2):191–200.

McCullagh, P., Pitman, J., and Winkel, M. (2008). Gibbs fragmentation trees. *Bernoulli*, 14(1):988–1002.

Miscouridou, X., Caron, F., and Teh, Y. W. (2018). Modelling sparsity, heterogeneity, reciprocity and community structure in temporal interaction data. *Advances in Neural Information Processing Systems*, 31.

Mnih, A. and Salakhutdinov, R. R. (2007). Probabilistic matrix factorization. *Advances in Neural Information Processing Systems*, 20.

Moreno-Dominguez, D., Anwander, A., and Knösche, T. R. (2014). A hierarchical method for whole-brain connectivity-based parcellation. *Human brain mapping*, 35(10):5000–5025.

Müller, N. F. and Bouckaert, R. R. (2020). Adaptive Metropolis-coupled MCMC for BEAST 2. *PeerJ*, 8:e9473.

Münkemüller, T., Lavergne, S., Bzeznik, B., Dray, S., Jombart, T., Schiffers, K., and Thuiller, W. (2012). How to measure and test phylogenetic signal. *Methods in Ecology and Evolution*, 3(4):743–756.

Ng, T. L. J., Murphy, T. B., Westling, T., McCormick, T. H., and Fosdick, B. (2021). Modeling the social media relationships of irish politicians using a generalized latent space stochastic blockmodel. *The Annals of Applied Statistics*, 15(4):1923–1944.

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

Opgen-Rhein, R., Fahrmeir, L., and Strimmer, K. (2005). Inference of demographic history from genealogical trees using reversible jump Markov chain Monte Carlo. *BMC Evolutionary Biology*, 5(1):1–13.

Osmond, C. (1985). Using age, period and cohort models to estimate future mortality rates. *International Journal of Epidemiology*, 14(1):124–129.

O'Hare, C. and Li, Y. (2012). Explaining young mortality. *Insurance: Mathematics and Economics*, 50(1):12–25.

Pagel, M. (1999). Inferring the historical patterns of biological evolution. *Nature*, 401(6756):877–884.

Peterson, C. B., Osborne, N., Stingo, F. C., Bourgeat, P., Doecke, J. D., and Vannucci, M. (2020). Bayesian modeling of multiple structural connectivity networks during the progression of Alzheimer's disease. *Biometrics*, 76(4):1120–1132.

Plat, R. (2009). On stochastic mortality modeling. *Insurance: Mathematics and Economics*, 45(3):393–404.

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American Statistical Association*, 108(504):1339–1349.

Preston, S. H. and Vierboom, Y. C. (2021). Excess mortality in the United States in the 21st century. *Proceedings of the National Academy of Sciences*, 118(16):e2024850118.

R Core Team (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raftery, A. E., Chunn, J. L., Gerland, P., and Ševčíková, H. (2013). Bayesian probabilistic projections of life expectancy for all countries. *Demography*, 50(3):777–801.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer New York.

Rasmussen, C. E. and Williams, C. K. (2006). *Gaussian processes for machine learning*. MIT Press Cambridge, MA.

Remund, A., Camarda, C. G., and Riffe, T. (2018). A cause-of-death decomposition of young adult excess mortality. *Demography*, 55(3):957–978.

Renshaw, A. E. and Haberman, S. (2003). Lee-Carter mortality forecasting with age-specific enhancement. *Insurance: Mathematics and Economics*, 33(2):255–272.

Renshaw, A. E. and Haberman, S. (2006). A cohort-based extension to the Lee-Carter model for mortality reduction factors. *Insurance: Mathematics and Economics*, 38(3):556–570.

Ricci, F. Z., Guindani, M., and Sudderth, E. (2022). Thinned random measures for sparse graphs with overlapping communities. *Advances in Neural Information Processing Systems*, 35:38162–38175.

Robinson, D. F. and Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1-2):131–147.

Ross, S. M. (2014). *Introduction to probability models*. Academic Press.

Roy, D. M., Kemp, C., Mansinghka, V., and Tenenbaum, J. (2006). Learning annotated hierarchies from relational data. *Advances in Neural Information Processing Systems*, 19.

Roy, D. M. and Teh, Y. (2008). The Mondrian process. *Advances in Neural Information Processing Systems*, 21.

Ryder, R. J. and Nicholls, G. K. (2011). Missing data in a stochastic Dollo model for binary trait data, and its application to the dating of Proto-Indo-European. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 60(1):71–92.

Sagart, L., Jacques, G., Lai, Y., Ryder, R. J., Thouzeau, V., Greenhill, S. J., and List, J.-M. (2019). Dated language phylogenies shed light on the ancestry of Sino-Tibetan. *Proceedings of the National Academy of Sciences*, 116(21):10317–10322.

Salakhutdinov, R. and Mnih, A. (2008). Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. In *Proceedings of the 25th International Conference on Machine Learning*, pages 880–887.

Scheinerman, E. R. and Tucker, K. (2010). Modeling graphs using dot product representations. *Computational Statistics*, 25:1–16.

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

Schweinberger, M., Krivitsky, P. N., Butts, C. T., and Stewart, J. R. (2020). Exponential-family models of random graphs: Inference in finite, super and infinite population scenarios. *Statistical Science*, 35(4):627–662.

Schweinberger, M. and Snijders, T. A. (2003). Settings in social networks: A measurement model. *Sociological Methodology*, 33(1):307–341.

Simard, S. W., Beiler, K. J., Bingham, M. A., Deslippe, J. R., Philip, L. J., and Teste, F. P. (2012). Mycorrhizal networks: Mechanisms, ecology and modelling. *Fungal Biology Reviews*, 26(1):39–60.

Smith, A. L., Asta, D. M., and Calder, C. A. (2019). The geometry of continuous latent space models for network data. *Statistical Science*, 34(3):428–453.

Sporns, O. (2022). Structure and function of complex brain networks. *Dialogues in Clinical Neuroscience*.

Sporns, O., Tononi, G., and Kötter, R. (2005). The human connectome: A structural description of the human brain. *PLoS Computational Biology*, 1(4):e42.

Stadler, T. (2010). Sampling-through-time in birth–death trees. *Journal of Theoretical Biology*, 267(3):396–404.

Stadler, T. (2013). How can we improve accuracy of macroevolutionary rate estimates? *Systematic Biology*, 62(2):321–329.

Stam, C. J. (2014). Modern network science of neurological disorders. *Nature Reviews Neuroscience*, 15(10):683–695.

Stingo, F. C., Guindani, M., Vannucci, M., and Calhoun, V. D. (2013). An integrative Bayesian modeling approach to imaging genetics. *Journal of the American Statistical Association*, 108(503):876–891.

Sylvia, R. and Alex, L. (2007). Discussion on the paper by Handcock, Raftery and Tantrum. *Journal of the Royal Statistical Society: Series A Statistics in Society*, 170(2):344.

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

Urchs, S., Armoza, J., Moreau, C., Benhajali, Y., St-Aubin, J., Orban, P., and Bellec, P. (2019). Mist: A multi-resolution parcellation of functional brain networks. *MNI Open Research*, 1:3.

Van Essen, D. C., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T. E., Bucholz, R., Chang, A., Chen, L., Corbetta, M., Curtiss, S. W., et al. (2012). The Human Connectome Project: A data acquisition perspective. *NeuroImage*, 62(4):2222–2231.

Vaupel, J. and Lundstrom, H. (1994). Longer life expectancy? Evidence from Sweden of reductions in mortality rates at advanced ages. *Studies in the Economics of Aging*, pages 79–102.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved Rˆ for assessing convergence of MCMC (with discussion). *Bayesian analysis*, 16(2):667–718.

Villegas, A. M., Kaishev, V. K., and Millossovich, P. (2018). StMoMo: An R package for stochastic mortality modelling. *Journal of Statistical Software*, 84(3):1–38.

Wang, H., Paulson, K. R., Pease, S. A., Watson, S., Comfort, H., Zheng, P., Aravkin, A. Y., Bisignano, C., Barber, R. M., Alam, T., et al. (2022a). Estimating excess mortality due to the COVID-19 pandemic: A systematic analysis of COVID-19-related mortality, 2020–21. *The Lancet*, 399(10334):1513–1536.

Wang, L., Zhang, Z., and Dunson, D. (2019). Common and individual structure of brain networks. *The Annals of Applied Statistics*, 13(1):85–112.

Wang, P., Pantelous, A. A., and Vahid, F. (2022b). Multi-population mortality projection: The augmented common factor model with structural breaks. *International Journal of Forecasting*, In Press.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244.

Warren, D. L., Geneva, A. J., and Lanfear, R. (2017). RWTY (R We There Yet): An R package for examining convergence of Bayesian phylogenetic analyses. *Molecular Biology and Evolution*, 34(4):1016–1020.

Wen, J., Cairns, A. J., and Kleinow, T. (2021). Fitting multi-population mortality models to socio-economic groups. *Annals of Actuarial Science*, 15(1):144–172.

Wiemers, E. E., Abrahams, S., AlFakhri, M., Hotz, V. J., Schoeni, R. F., and Seltzer, J. A. (2020). Disparities in vulnerability to complications from COVID-19 arising from disparities in preexisting conditions in the United States. *Research in Social Stratification and Mobility*, 69:100553.

Wong, J. S., Forster, J. J., and Smith, P. W. (2018). Bayesian mortality forecasting with overdispersion. *Insurance: Mathematics and Economics*, 83:206–221.

Woolf, S. H. and Schoomaker, H. (2019). Life expectancy and mortality rates in the United States, 1959-2017. *JAMA*, 322(20):1996–2016.

Xu, K. (2015). Stochastic block transition models for dynamic networks. In *Artificial Intelligence and Statistics*, pages 1079–1087. PMLR.

Xu, K. S. and Hero, A. O. (2013). Dynamic stochastic blockmodels: Statistical models for time-evolving networks. In *Social Computing, Behavioral-Cultural Modeling and Prediction: 6th International Conference, SBP 2013, Washington, DC, USA, April 2-5, 2013. Proceedings 6*, pages 201–210. Springer.

Yang, Z. (2006). *Computational molecular evolution*. Oxford University Press.

Young, J.-G., Kirkley, A., and Newman, M. (2022). Clustering of heterogeneous populations of networks. *Physical Review E*, 105(1):014312.

Yu, G., Smith, D. K., Zhu, H., Guan, Y., and Lam, T. T.-Y. (2017). ggtree: An R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, 8(1):28–36.

Yule, G. U. (1925). A mathematical theory of evolution, based on the conclusions of Dr. JC Willis, FRS. *Philosophical Transactions of the Royal Society of London Series B*, 213:21–87.

Zhang, Z., Descoteaux, M., Zhang, J., Girard, G., Chamberland, M., Dunson, D., Srivastava, A., and Zhu, H. (2018). Mapping population-based structural connectomes. *NeuroImage*, 172:130–145.

Zhu, B. and Dunson, D. B. (2013). Locally adaptive Bayes nonparametric regression via nested Gaussian processes. *Journal of the American Statistical Association*, 108(504):1445–1456.

Zuo, X.-N., Anderson, J. S., Bellec, P., Birn, R. M., Biswal, B. B., Blautzik, J., Breitner, J., Buckner, R. L., Calhoun, V. D., Castellanos, F. X., et al. (2014). An open science resource for establishing reliability and reproducibility in functional connectomics. *Scientific Data*, 1(1):1–13.