# Applications of Bayesian mixed effects models

**Vincent Chin**

Supervisors: Prof. Scott A. Sisson and Prof. Robert Kohn

A thesis in fulfilment of the requirements for the degree of Doctor of Philosophy

School of Mathematics and Statistics
Faculty of Science

October 2019

# Thesis/Dissertation Sheet

| | | |
|---|---|---|
| Surname/Family Name | : | **Chin** |
| Given Name/s | : | **Vincent** |
| Abbreviation for degree as give in the University calendar | : | **PhD** |
| Faculty | : | **Faculty of Science** |
| School | : | **School of Mathematics and Statistics** |
| Thesis Title | : | **Applications of Bayesian mixed effects models** |


**Abstract 350 words maximum: (PLEASE TYPE)**

Longitudinal study is an experimental design which takes repeated measurements of some variables from a study cohort over a specified time period. Collected data is most often modelled using a mixed effects model, which permits heterogeneity analysis of the variables over time. In this thesis, we apply the linear mixed effects models to applications that cover different domains of research. First, we consider the problem of estimating a multivariate probit model in a longitudinal data setting with emphasis on sampling a high-dimensional correlation matrix, and improving the overall efficiency of the posterior sampling approach via a dynamic variance reduction technique. The proposed method is used to analyse stated preference of female contraceptive products by Australian general practitioners, and hence provide insights to their behaviour in decision-making. Additionally, we introduced a multiclass classification model for growth trajectory that flexibly extends a piecewise linear model popular in the literature by allowing the number of classes to be data driven. Individual-specific random change points are introduced to model heterogeneity in growth phases realistically. The model is then applied on a birth cohort from the Healthy Birth, Growth and Development knowledge integration (HBGDki) project funded by the Bill and Melinda Gates Foundation. Finally, we investigate the evolution of unobserved executive functions of male soccer players representing a professional German Bundesliga club using a latent variable model, where cognitive outcomes from a test battery of neuropsychological assessments undergone by the players are manifestation of some underlying curves representing executive functions. This is the first study of its kind in soccer research that permits a longitudinal analysis of domain-generic and domain-specific executive functions.

**ORIGINALITY STATEMENT**

'I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at UNSW or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by others, with whom I have worked at UNSW or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.'

Signed ……………………………………………….............

Date ……………………………………………….............

# INCLUSION OF PUBLICATIONS STATEMENT

UNSW is supportive of candidates publishing their research results during their candidature as detailed in the UNSW Thesis Examination Procedure.

**Publications can be used in their thesis in lieu of a Chapter if:**

- The candidate contributed greater than 50% of the content in the publication and is the "primary author", ie. the candidate was responsible primarily for the planning, execution and preparation of the work for publication
- The candidate has approval to include the publication in their thesis in lieu of a Chapter from their supervisor and Postgraduate Coordinator.
- The publication is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in the thesis

Please indicate whether this thesis contains published material or not:

☐ This thesis contains no publications, either published or submitted for publication

☐ Some of the work described in this thesis has been published and it has been documented in the relevant Chapters with acknowledgement

☒ This thesis has publications (either published or submitted for publication) incorporated into it in lieu of a chapter and the details are presented below

**CANDIDATE'S DECLARATION**

I declare that:

- I have complied with the UNSW Thesis Examination Procedure
- where I have used a publication in lieu of a Chapter, the listed publication(s) below meet(s) the requirements to be included in the thesis.

| **Candidate's Name** | **Signature** | **Date (dd/mm/yy)** |
| --- | --- | --- |
| | | |

| **POSTGRADUATE COORDINATOR'S DECLARATION**<br>I declare that:<br>• the information below is accurate<br>• where listed publication(s) have been used in lieu of Chapter(s), their use complies with the UNSW Thesis Examination Procedure<br>• the minimum requirements for the format of the thesis have been met. | | |
|---|---|---|
| **PGC's Name** | **PGC's Signature** | **Date (dd/mm/yy)** |

**For each publication incorporated into the thesis in lieu of a Chapter, provide all of the requested details and signatures required**

| **Details of publication #1:**<br>*Full title:* Efficient data augmentation for multivariate probit models with panel data: An application to general practitioner decision making about contraceptives<br>*Authors:* Vincent Chin, David Gunawan, Denzil G. Fiebig, Robert Kohn and Scott A. Sisson<br>*Journal or book name:* Journal of the Royal Statistical Society: Series C (Applied Statistics)<br>*Volume/page numbers:* Volume 69, pp. 277-300 | | | | | | |
|---|---|---|---|---|---|---|
| **Status** | *Published* | **X** | *Accepted and In press* | | *In progress (submitted)* | |
| **The Candidate's Contribution to the Work**<br>Formulating concept of the paper, performing data analysis, writing the paper and addressing reviewers' comments | | | | | | |
| **Location of the work in the thesis and/or how the work is incorporated in the thesis:**<br>Chapter 3 | | | | | | |
| **PRIMARY SUPERVISOR'S DECLARATION**<br>I declare that:<br>• the information above is accurate<br>• this has been discussed with the PGC and it is agreed that this publication can be included in this thesis in lieu of a Chapter<br>• All of the co-authors of the publication have reviewed the above information and have agreed to its veracity by signing a 'Co-Author Authorisation' form. | | | | | | |
| **Primary Supervisor's name** | | **Primary Supervisor's signature** | | | **Date (dd/mm/yy)** | |

**COPYRIGHT STATEMENT**

'I hereby grant the University of New South Wales or its agents a non-exclusive licence to archive and to make available (including to members of the public) my thesis or dissertation in whole or part in the University libraries in all forms of media, now or here after known. I acknowledge that I retain all intellectual property rights which subsist in my thesis or dissertation, such as copyright and patent rights, subject to applicable law. I also retain the right to use all or part of my thesis or dissertation in future works (such as articles or books).'

'For any substantial portions of copyright material used in this thesis, written permission for use has been obtained, or the copyright material is removed from the final public version of the thesis.'

Signed ……………………………………………..........................

Date ……………………………………………..............................

**AUTHENTICITY STATEMENT**

'I certify that the Library deposit digital copy is a direct equivalent of the final officially approved version of my thesis.'

Signed ……………………………………………..........................

Date ……………………………………………..............................

# Acknowledgements

First of all, I would like to express my very profound gratitude to my supervisors, Professor Scott Sisson and Professor Robert Kohn, for their continuous guidance and support throughout my PhD study, and for putting up with my frequent ineptitude. The doors to their offices were always open whenever I ran into an issue, either personal or research related. Without their great insights and immense knowledge of Bayesian computation, I could not have completed this research work.

I am thankful to have had constructive collaborations with great researchers along my PhD journey. In particular, I am indebted to Professor Louise Ryan for the privilege to contribute to a project funded by the Bill and Melinda Gates Foundation in Chapter 4. Also, special thanks to Professor Denzil Fiebig, Dr. Job Fransen and Adam Beavan for sharing the data for Chapters 3 and 5, as well as providing subject matter expertise in the analyses.

I would like to acknowledge the Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS) for providing generous financial funding that supported my research. My sincere gratitude also goes to my fellow colleagues – Dr. Jarod Lee, Dr. David Gunawan, Dr. Boris Beranger and Yu Yang for their encouragement and valuable suggestions.

Last but not the least, I am grateful to my family for the unfailing support throughout this challenging journey. My accomplishment would have meant nothing without the unconditional love and care from them. To my friends (in no particular order!) – Ritchie, Chee Han, Aron and Firdaus, thank you for keeping me sane!

*To mum & dad.*

# Contents

# Nomenclature

**Parameter**

| | |
|---|---|
| $\boldsymbol{\theta}$ | Multidimensional parameter |
| $\boldsymbol{\theta}^{[i]}$ | $i$-th iterate of $\boldsymbol{\theta}$ in a Markov chain |
| $\boldsymbol{\theta}_i$ | $i$-th sample of $\boldsymbol{\theta}$ |
| $\Theta$ | Parameter space of $\boldsymbol{\theta}$ |
| $\theta_i$ | $i$-th margin of $\boldsymbol{\theta}$ |

**Probability and Distribution Functions**

| | |
|---|---|
| Cov | Covariance |
| $\mathbb{E}$ | Expected value |
| $\mathcal{HIW}$ | Hierarchical inverse-Wishart distribution |
| $\mathcal{IG}$ | Inverse-Gamma distribution |
| $\mathcal{IW}$ | Inverse-Wishart distribution |
| $\mathcal{TN}$ | Truncated normal distribution |
| $\mathcal{N}$ | Normal distribution |
| $\mathbb{P}$ | Probability |
| $\mathbb{V}$ | Variance |

**Matrices**

$\boldsymbol{I}$ Identity matrix

$\boldsymbol{K}^{-1}$ Inverse of matrix $\boldsymbol{K}$

$\boldsymbol{R}$ Correlation matrix

$\mathrm{diag}(\boldsymbol{K})$ Diagonal entries of matrix $\boldsymbol{K}$

$|\boldsymbol{K}|$ Determinant of matrix $\boldsymbol{K}$

$\boldsymbol{\Sigma}$ Covariance matrix

**Dirichlet Process**

$\mathcal{DP}$ Dirichlet process

$\mathcal{G}$ Realisation from a Dirichlet process

$\mathcal{G}_0$ Base distribution of a Dirichlet process

**Hamiltonian Monte Carlo**

$\boldsymbol{M}$ Mass matrix

$\boldsymbol{u}$ Momentum

$\mathcal{H}$ Hamiltonian

$\mathfrak{L}$ Number of leapfrog updates

$\varepsilon$ Stepsize

**Other Symbols**

$\boldsymbol{\mathcal{M}}$ Space of all statistical models

$\boldsymbol{\mathcal{R}}$ Space of all correlation matrices

$\mathbb{R}$ Real numbers

| | |
|---|---|
| $\mathcal{M}$ | Statistical model |
| $\nu$ | Degrees of freedom |
| $D$ | Dimension of an observation |
| $f$ | A scalar function |
| $G$ | Number of mixture components |
| $N$ | Number of individuals |
| $n$ | Number of simulated samples or iterates |
| $P$ | Dimension of $\boldsymbol{\theta}$ |
| $T$ | Number of repeated measurements |

# List of Figures

# List of Tables

# Chapter 1

# Introduction

A longitudinal study is an experimental design which takes repeated measurements of some variables from a study cohort over a specified time period. This induces a correlation structure between observations collected on the same measurement subject, and hence special care is required when performing statistical analyses on the data. One of the most dominant methods used in longitudinal analysis of continuous outcomes is the linear mixed effects model proposed by Laird and Ware (1982).

Let $\boldsymbol{y}_{it} = (y_{1,it}, \ldots, y_{D,it})^\top$ be a vector of $D$ correlated continuous outcomes collected from individual $i = 1, \ldots, N$ at time period $t = 1, \ldots, T_i$. The linear mixed effects model is formulated as

$$\boldsymbol{y}_{it} = \boldsymbol{B}\boldsymbol{x}_{it} + \boldsymbol{A}_i\boldsymbol{z}_{it} + \boldsymbol{\epsilon}_{it}, \tag{1.1}$$

where $\boldsymbol{x}_{it} = (1, x_{1,it}, \ldots, x_{K_f-1,it})^\top$ and $\boldsymbol{z}_{it} = (1, z_{1,it}, \ldots, z_{K_r-1,it})^\top$ are both a set of exogenous variables assumed to be the same for all margins of $\boldsymbol{y}_{it}$, $\boldsymbol{B}$ is a $D \times K_f$ matrix of fixed-effects regression coefficients, $\boldsymbol{A}_i$ is a $D \times K_r$ matrix of random effects and $\boldsymbol{\epsilon}_{it} = (\epsilon_{1,it}, \ldots, \epsilon_{D,it})^\top$ is a $D$-vector of $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$ distributed correlated error term which models the dependence structure between the outcomes $\boldsymbol{y}_{it}$. The variable $\boldsymbol{x}_{it}$ is assumed to be uncorrelated with both $\boldsymbol{A}_i$ and $\boldsymbol{\epsilon}_{it}$. The fixed effects are constant across all subjects, whereas the random effects $\boldsymbol{\alpha} = \text{vec}(\boldsymbol{A})$ which are distributed as $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\alpha}})$ account for heterogeneity between subjects, thereby permitting an investigation of the evolution of individual-specific processes over time. Using the linear mixed effects model as a basis for model formulation, we analyse longitudinal data from three different domains of research – health economics, epidemiology and sports science, and introduce novel statistical methodologies in this thesis. Chapter 2 reviews the modelling approaches for these applications and provides an outline of the Bayesian framework used throughout the thesis for model estimation. The rest of the thesis is organised as follows.

Chapter 3 models the longitudinal stated preference survey described in Fiebig et al. (2017), whereby a study is designed to mimic the choice problem faced by general practitioners in a consultation where they need to match alternative female contraceptive products with a particular patient whose socio-economic and clinical characteristics are varied as part of the experimental design to cover a range of different life cycle and fertility stages. An analysis of the decision-making of these general practitioners can be performed using a multivariate probit model with mixed effects by extending the formulation in (1.1) to accommodate for binary responses. This is done by introducing normally distributed latent variables $\boldsymbol{y}_{it}^*$ following the data augmentation approach given in Chib and Greenberg (1998) such that the value of each margin in the observed binary outcomes $\boldsymbol{y}_{it}$ is determined by the sign of the corresponding margin in $\boldsymbol{y}_{it}^*$. Additionally, the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ of these latent normal random variates must be restricted to a correlation matrix $\boldsymbol{R}_{\boldsymbol{\epsilon}}$ so that the model parameters are uniquely identified (Chib and Greenberg, 1998). In a Bayesian context, Markov chain Monte Carlo (MCMC) sampling from the posterior distribution of $\boldsymbol{R}_{\boldsymbol{\epsilon}}$ is challenging as a result of the restrictions on the diagonal entries and the positive definiteness property of the matrix (Chib and Greenberg, 1998; Edwards and Allenby, 2003; Smith, 2013). Common sampling approaches for $\boldsymbol{R}_{\boldsymbol{\epsilon}}$ include the random walk Metropolis-Hastings algorithm (Chib and Greenberg, 1998; Gunawan et al., 2017) and the Griddy-Gibbs sampler (Barnard et al., 2000), which suffer from poor exploration of the parameter space (Sherlock et al., 2010) in addition to being computationally expensive when the dimension of $\boldsymbol{R}_\epsilon$ is large. To overcome this, we reparameterise $\boldsymbol{R}_\epsilon$ in a principled way and then carry out efficient Bayesian inference using Hamiltonian Monte Carlo (Duane et al., 1987; Neal, 2011) due to its ability to generate credible but distant candidate parameters for the Metropolis-Hastings algorithm, thereby reducing autocorrelation in the posterior samples. Motivated by variance reduction techniques, we also propose a novel method which integrates an antithetic variable (Hammersley and Morton, 1956) dynamically within the MCMC sampling algorithm to improve the mixing properties of the Markov chain associated with the regression parameter $\boldsymbol{B}$ and the random effects $\boldsymbol{\alpha}$, thereby increasing their effective sample size. Our analysis result of the motivating discrete choice experiment suggests that the joint probability of discussing combinations of contraceptive products with a patient shows medical practice variation (Wennberg et al., 1982; Scott and Shiell, 1997; Davis et al., 2000) among the general practitioners.

While it is well established that having access to professional reproductive health advice improves the general well-being of women (Darroch et al., 2011; Sundstrom et al., 2019), it is also equally important to address nutritional problems among young children which is prevalent in low to medium income countries (Onofiok and Nnanyelugo, 1998; Martorell, 1999; Lartey, 2008; Kirby and Danner, 2009; Keino et al., 2014). In Chapter 4, we propose a multiclass classification model for growth curves so that children with similar growth structures can be identified and appropriate targeted treatments or interventions can be designed and administered. Current approaches in the literature can be categorised as functional models (Abraham et al., 2003; James and Sugar, 2003; Ramsay and Silverman, 2005; Heard et al., 2006; Tokushige et al., 2007) or growth mixture models (Muthén and Shedden, 1999; Nagin, 1999; Muthén and Muthén, 2000; Li et al., 2001; Muthén, 2008). Functional approaches intrinsically assume that the data are infinite dimensional and defined over a continuum of time, which makes them a less attractive option in applications with sparse observations, especially so when analysing health data from low to medium income countries. On the other hand, growth mixture models are popular regression models in the epidemiological literature which extend naturally from the general formulation of the linear mixed effects models in (1.1) by replacing the exogenous variable $\boldsymbol{z}_{it}$ with basis function of time. Greater flexibility is also achieved by relaxing the distribution assumption on $\boldsymbol{\alpha}$ from a normal distribution to a more structured normal mixture distribution with $G$ components. This allows each mixture component to characterise subgroup-specific growth trajectories. However, choosing a suitable value of $G$ in a mixture distribution is a non-trivial problem. Most methods require the need to fit multiple models with different values of $G$, and then select the "best" model by performing a likelihood ratio test (Titterington et al., 1985), or considering a goodness-of-fit test (Verbeke and Lesaffre, 1996) or information criterion (Dasgupta and Raftery, 1998), among others. In order to circumvent this kind of model selection procedure, we model the mixture distribution non-parametrically using a Dirichlet process prior (Ferguson, 1973), which avoids the need to specify $G$ by allowing its value to be driven by the complexity of the data (Teh, 2011). Because children have individual differences in the onset of growth stages, we introduce individual-specific random knot change points whose prior distribution follow the even-numbered order statistics distribution in Green (1995) to probabilistically encourage consecutive change points to be uniformly spaced. Simulation results show that the random change point model outperforms the fixed change point model because it has fewer restrictions on knot locations. We apply the proposed model to analyse a longitudinal birth cohort from the Healthy Birth, Growth and Development knowledge integration project. Our result suggests that child growth may be influenced by gender and maternal education levels, and that children who experience severe faltering during their first year of life have lower IQ scores compared to their peers.

Unlike anthropometric measurements such as height, weight and body mass index which can be collected physically, executive functions, which are higher order cognitive functioning underpinning other cognitive processes such as problem solving, planning and reasoning (Diamond, 2013), are usually examined using a test battery of neuropsychological assessments. Chapter 5 presents one of the first population-specific studies in the field of sports science research by analysing the developmental changes in executive functions of elite German soccer players aged between 10 and 21 years old participating in a longitudinal study. Previous research investigating this problem in an athlete population are based on cross-sectional data (see e.g. Verburgh et al., 2014; Huijgen et al., 2015; Sakamoto et al., 2018), which ignores potential heterogeneity between players that are due to unobservables such as the number of training hours and familiarity with the test. Furthermore, existing results from longitudinal studies are based on a general population (Zelazo et al., 2004; Huizinga and Smidts, 2010; Zelazo and Carlson, 2012) and the generalisation of these results to an athlete population is limited given that active participation in sports has shown to improve executive functioning (Jacobson and Matthaeus, 2014). The common factor theories (Birren and Fisher, 1995; Salthouse, 1996; Baltes and Lindenberger, 1997; Lindenberger and Baltes, 1997) argue that the evolutionary changes in cognitive functioning are shared among various types of cognitive variables (Salthouse et al., 1998). Therefore, it is instructive to assume that measurements of these cognitive variables reflect properties of a common latent cognitive process. To model the unobserved latent cognitive process, we consider a latent growth curve model (Meredith and Tisak, 1990; Dunson, 2000; Muthén, 2002; Proust et al., 2006) which introduces a two-level hierarchical structure to the formulation in (1.1) such that the first level given by a measurement model links the measured cognitive outcomes to the latent variable representing the executive functions in a linear fashion. The executive functions are, in turn, modelled by a structural model with random effects in the second level of the hierarchical model so that each individual has its own rate of growth centered around a population mean. Estimation results show that executive functions of these players which are responsible for excellence in soccer performance demonstrate a sharp increase from late childhood (10–12 years old) until early adolescence (12–15 years old), and then their increase remains very minimal. This developmental pattern implies that executive functions do not correlate with good performance in soccer, as claimed in the literature (Verburgh et al., 2014; Vestberg et al., 2012; Sakamoto et al., 2018). Finally, we provide some concluding remarks and discussion of potential future research work in Chapter 6.

# Chapter 2

# Literature review

Bayesian estimation methods for any general statistical model $\mathcal{M}$ requires computing the posterior distribution $\pi(\boldsymbol{\theta})$ (or $\pi(\boldsymbol{\theta}|\mathcal{M}, \boldsymbol{y})$ in a more precise notation) of model parameter $\boldsymbol{\theta}$ upon observing the data $\boldsymbol{y} = \{\boldsymbol{y}_{it}; i = 1, \ldots, N, t = 1, \ldots, T_i\}$ according to the Bayes' theorem,

$$\pi(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}|\mathcal{M}, \boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\theta}, \mathcal{M})p(\boldsymbol{\theta}|\mathcal{M})}{p(\boldsymbol{y}|\mathcal{M})}, \tag{2.1}$$

where $p(\boldsymbol{y}|\boldsymbol{\theta}, \mathcal{M})$ is the likelihood function of the model $\mathcal{M}$ and $p(\boldsymbol{\theta}|\mathcal{M})$ is the prior distribution on the parameter $\boldsymbol{\theta}$. The marginal likelihood $p(\boldsymbol{y}|\mathcal{M})$, which can be expressed as:

$$p(\boldsymbol{y}|\mathcal{M}) = \int_{\Theta} p(\boldsymbol{y}|\boldsymbol{\theta}, \mathcal{M})p(\boldsymbol{\theta}|\mathcal{M})d\boldsymbol{\theta}, \tag{2.2}$$

provides a measure of the average fit of the model to the data, and is therefore used extensively for Bayesian model selection.

A fundamental problem in statistical computing is estimating the expectation $\mathbb{E}_{\pi}[f(\boldsymbol{\theta})]$ of a scalar function $f$ of $\boldsymbol{\theta} \in \Theta$ with respect to its posterior distribution $\pi(\boldsymbol{\theta})$ in (2.1), i.e. evaluating the integral

$$\int_{\Theta} f(\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}. \tag{2.3}$$

A tractable closed form expression for (2.3) rarely exist for a high dimensional $\boldsymbol{\theta}$, but it can be approximated if samples from $\pi(\boldsymbol{\theta})$ can be drawn easily.

This chapter begins by introducing three different models for the analysis of longitudinal data that will be used in subsequent chapters, i.e. multivariate probit models, growth mixture models and latent growth curve models. We then review variance reduction techniques for Monte Carlo estimation of (2.3), as well as Markov chain Monte Carlo algorithms when $\pi(\boldsymbol{\theta})$ is challenging to sample directly. Finally, we address how model selection procedures can be circumvented, and model uncertainty can be accounted for when using Bayesian non-parametric models.

## 2.1 Longitudinal data analysis

### 2.1.1 Multivariate probit models

A multivariate probit model (MVP; Ashford and Sowden, 1970) is used commonly in situations where multiple correlated binary responses are observed. For example, the outcome $\boldsymbol{y}_{it}$ may represent an individual's preference for each of the $D$ different products. To accommodate for binary outcomes in (1.1), the MVP model can be written as:

$$\boldsymbol{y}^*_{it} = \boldsymbol{B}\boldsymbol{x}_{it} + \boldsymbol{A}_i\boldsymbol{z}_{it} + \boldsymbol{\epsilon}_{it},$$

using the latent variable formulation introduced in Chib and Greenberg (1998) so that each margin of $\boldsymbol{y}_{it}$ takes the value of 0 or 1 depending on the sign of the corresponding margin of the latent variable $\boldsymbol{y}^*_{it}$:

$$y_{d,it} = \mathbb{1}(y^*_{d,it} > 0), \quad d = 1, \ldots, D,$$

where $\mathbb{1}(\boldsymbol{E})$ denotes an indicator function of the event $\boldsymbol{E}$. Additionally, the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ must be restricted to a correlation matrix $\boldsymbol{R}_{\boldsymbol{\epsilon}}$ so that all model parameters are identifiable (Chib and Greenberg, 1998).

We now discuss related work on priors for $\boldsymbol{R}_{\boldsymbol{\epsilon}}$. Let $\boldsymbol{\mathcal{R}}^D$ be the space of all valid correlation matrices. Barnard et al. (2000) suggest a uniform prior over all correlation matrices in $\boldsymbol{\mathcal{R}}^D$, which is equivalent to the LKJ prior (Lewandowski et al., 2009) with unit shape, as suggested by the Stan Development Team (2017). The LKJ prior with unit shape is a regularising prior (McElreath, 2020) since each off-diagonal elements $r_{ij}, i \neq j$ of $\boldsymbol{R}_{\boldsymbol{\epsilon}}$ is marginally distributed according to a Beta$(\frac{D}{2}, \frac{D}{2})$ distribution over $(-1, 1)$ with both shape parameters $\frac{D}{2}$, which is informative in high dimensions because the Beta density increasingly concentrates around zero. Chib and Greenberg (1998) propose using a multivariate normal prior on the $r_{ij}$, with the support of the prior restricted to values of $r_{ij}$ which give a correlation matrix in $\boldsymbol{\mathcal{R}}^D$, while Liechty et al. (2004) introduce a mixture of normal distributions prior on $r_{ij}$ to express *a priori* knowledge of blocked structure in $\boldsymbol{R}_{\boldsymbol{\epsilon}}$. However, these choices of normal priors do not imply that all marginal densities of the $r_{ij}$ are the same due to the constraints imposed on the $r_{ij}$ for the resulting $\boldsymbol{R}_{\boldsymbol{\epsilon}}$ to be in $\boldsymbol{\mathcal{R}}^D$.

Barnard et al. (2000) also propose decomposing a covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ as $\boldsymbol{S}\boldsymbol{R}_{\boldsymbol{\epsilon}}\boldsymbol{S}$, where $\boldsymbol{S}$ is a diagonal matrix of standard deviations and $\boldsymbol{R}_{\boldsymbol{\epsilon}}$ is a correlation matrix. They show that if $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \sim \mathcal{IW}(\nu, \boldsymbol{I})$, i.e. an inverse-Wishart distribution with degrees of freedom $\nu$ and the $D \times D$ identity matrix $\boldsymbol{I}$ as scale matrix, then the density of $\boldsymbol{R}_{\boldsymbol{\epsilon}}$ is

$$p(\boldsymbol{R}_{\boldsymbol{\epsilon}}) \propto |\boldsymbol{R}_{\boldsymbol{\epsilon}}|^{\frac{1}{2}(\nu-1)(D-1)-1} \left( \prod_{d=1}^{D} |\boldsymbol{R}_{\boldsymbol{\epsilon}}(-d; -d)| \right)^{-\frac{\nu}{2}}, \tag{2.4}$$

where $\boldsymbol{R}_{\boldsymbol{\epsilon}}(-d; -d)$ denotes the $d$-th principal submatrix of $\boldsymbol{R}_{\boldsymbol{\epsilon}}$, that is $\boldsymbol{R}_{\boldsymbol{\epsilon}}$ with its $d$-th row and column removed. The prior distribution in (2.4) induces a modified Beta distribution on each $r_{ij}$. In particular, the marginal densities of the $r_{ij}$ are uniform on $(-1, 1)$ when $\nu = D + 1$, which means that posterior inference is invariant to the ordering of the binary outcomes $\boldsymbol{y}$. Furthermore, recent results in Wang et al. (2018) establish that for such a choice of $\nu$, the corresponding matrix of partial correlations $\rho_{kl}$ has the LKJ distribution with unit shape parameter. This means that the prior weights on all $\rho_{kl}$ are greater around zero as the dimension $D$ increases, using the property of the LKJ distribution mentioned earlier. The informativity of

$\rho_{kl}$ is useful in practical applications, where more often than not a sparse structure on the partial correlation matrix is desirable to suggest conditional independence.

The dependence structures imposed by the marginally uniform prior on $r_{ij}$ when $\nu = D+1$ in (2.4) are less studied in the literature. Since analytical results for these properties are limited (Tokuda et al., 2011), we briefly illustrate these graphically instead. The results obtained are based on correlation matrices of dimension $D = 4$ but they can be generalised to higher dimensions. We generate $10^7$ samples from (2.4) with $\nu = D+1$ by normalising the covariance matrices drawn from an $\mathcal{IW}(D+1, \boldsymbol{I})$ distribution. Figure 2.1 illustrates the pairwise dependence structures among the correlations $r_{ij}$ and the partial correlations $\rho_{kl}$ when the pairs share (top panels) or do not share (bottom panels) common indices. When there is a shared index, the density on $(r_{12}, r_{13})$ tends to support similar values in absolute terms (the visible cross pattern), which is less apparent when there is no common index in $(r_{12}, r_{34})$. However, both distributions have most of their density on the vertices corresponding to $|r_{ij}| \approx 1$. This means that inference for all pairs of $r_{ij}$ is skewed towards jointly extreme values *a priori* (the univariate margin for each $r_{ij}$ is still uniform on $(-1, 1)$), although this effect diminishes with an increase in the number of observations. In contrast, pairs of partial correlations $\rho_{kl}$ exhibit no dependence structure regardless

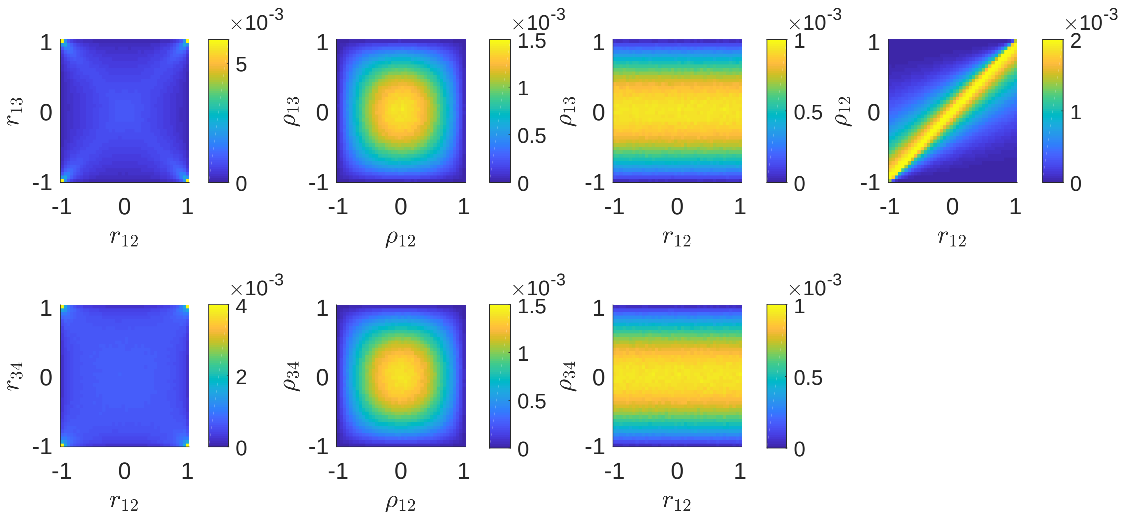



Figure 2.1: Bivariate density plots showing the dependence structures associated with the marginally uniform prior (2.4) on $\boldsymbol{R}_\epsilon$ with $\nu = D+1$, for pairs of parameters sharing common indices (top panels) and without a common index (bottom panels).

of whether or not there is a common index. Independence is also observed between $r_{ij}$ and $\rho_{kl}$, except when both indices of parameters are the same $(r_{12}, \rho_{12})$ in which case they are strongly positively correlated; see Figure 2.1, top row, rightmost.

### 2.1.2 Growth mixture models

The linear mixed effects model in (1.1) can also be used to study how growth curves of measurement subjects are shaped and change over time (Ghisletta et al., 2015). A univariate formulation of (1.1) that forms the basis of more complex growth models is the random intercept and slopes model given by:

$$y_{it} = \alpha_i + \beta_i \omega_{it} + \epsilon_{it}, \tag{2.5}$$

$$\begin{bmatrix} \alpha_i \\ \beta_i \end{bmatrix} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \text{with } \boldsymbol{\mu} = \begin{bmatrix} \mu_\alpha \\ \mu_\beta \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_\alpha^2 & \sigma_{\alpha\beta} \\ \sigma_{\alpha\beta} & \sigma_\beta^2 \end{bmatrix}, \tag{2.6}$$

$$\epsilon_{it} \sim \mathcal{N}(0, \sigma_\epsilon^2), \tag{2.7}$$

where $\alpha_i$ and $\beta_i$ are the normally distributed random intercept and slope whose variance-covariance is $(\sigma_\alpha^2, \sigma_\beta^2, \sigma_{\alpha\beta})$; $\omega_{it}$ is the age of individual $i$ on the $t$-th measurement occasion; $\epsilon_{it}$ is the random error term assumed to be uncorrelated with $\alpha_i$ and has variance $\sigma_\epsilon^2$. The growth model described in (2.5)–(2.7) allows the growth trajectory of each individual to have its own intercept and rate of growth. Acknowledging the potential of non-linear trends in growth structures, some applications consider more flexible construction of the growth trajectory using latent basis coefficients (McArdle and Epstein, 1987; Meredith and Tisak, 1990; Ram and Grimm, 2009; Grimm et al., 2011), linear splines (Pan and Goldstein, 1998; Tilling et al., 2014; Crozier et al., 2019) and fractional polynomials (Long and Ryoo, 2010; Tan et al., 2011), among others.

Equation (2.6) assumes that the study population is homogeneous such that individual growth profiles follow closely the trend of a population trajectory whose intercept-slope parameter is given by $\boldsymbol{\mu} = (\mu_\alpha, \mu_\beta)^\top$. However, this is rarely the case in most practical applications where multiple subgroups are hypothesised to present in a population. This restriction motivates the utilisation of finite mixture models (Gelman et al., 2013, Chapter 22) in the context of growth mixture models (Muthén and Shedden, 1999; Nagin, 1999; Muthén and Muthén, 2000; Li et al., 2001; Muthén, 2008), which allow greater flexibility by permitting different sets of intercept-slope parameter to capture group-specific growth trajectories. Formally, the distributional assumption on $(\alpha_i, \beta_i)^\top$ is relaxed to a normal mixture distribution:

$$\begin{bmatrix} \alpha_i \\ \beta_i \end{bmatrix} \sim \sum_{g=1}^{G} w_g \mathcal{N}(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g), \tag{2.8}$$

with positive weights $w_g > 0$ such that $\sum_{g=1}^{G} w_g = 1$. Each component $g$ in the mixture distribution therefore represents a particular class of growth trajectory as characterised by $\boldsymbol{\mu}_g$, and each child belongs to one of these $G$ subgroups probabilistically. The heterogeneous model assumed in (2.8) is used extensively for cluster analysis. For example, Verbeke and Lesaffre (1996) divided a schoolgirl population into "slow" growers and "fast" growers, Lin et al. (2002) identified different trajectories of prostate-specific antigen for the onset of prostate cancer, and Muthén (2004) studied mathematics achievement of students in U.S. public schools.

Equation (2.8) requires specifying the "correct" number of subgroups $G$, which is non-trivial. Under a classical statistical approach, a likelihood ratio test (Titterington et al., 1985) is performed, but the asymptotic distribution of the test statistic under the null hypothesis is unknown (Ghosh and Sen, 1985), as opposed to the conventional $\chi^2$ distribution. Verbeke and Lesaffre (1996) consider a goodness-of-fit test by comparing the probability distribution of random variables derived from linear combinations of the observations against a uniform distribution using the Kolmogorov-Smirnov test. From the Bayesian perspective, Richardson and Green (1997) adapt the reversible jump MCMC (Green and Han, 1992), which permits dimension-changing moves between the parameter subspaces corresponding to different values of $G$ in the Metropolis-Hastings algorithm. Dasgupta and Raftery (1998) use the Bayesian information criterion (BIC) approximation to the Bayes factor as a basis for the selection of $G$, from which there is strong evidence to prefer the model with a larger value of $G$ if the BIC value increases by more than 10 upon an increase of one additional mixture component. Sugar and James (2003) propose computing the average Mahalanobis distance between the observations and their respective subgroup means for a range of values $G$. They show theoretically that the "true" value of $G$ contributes to the largest drop in the distance. More recently, Fúquene et al. (2019) develop a family of repulsive prior distributions to penalise recurring components so that each subgroup is well distinguished. An extensive review of other trans-dimensional MCMC methods and likelihood-based approaches for finite mixture models under model specification uncertainty of $G$ is described in Frühwirth-Schnatter (2006).

### 2.1.3 Latent growth curve models

An extension of the linear mixed effects model in (1.1) to latent variable models can be formalised in a latent growth curve model (McArdle, 1986; Meredith and Tisak, 1990; Muthén, 1991; Duncan et al., 1994; Stoolmiller, 1995; Bollen and Curran, 2006; Duncan et al., 2013), which can be described by a two-level hierarchical structure:

$$\text{Measurement model:} \quad y_{d,it} \quad = \eta_{d,i} + \boldsymbol{x}_{it}^{\top}\boldsymbol{\gamma}_d + c_d\zeta_i(\omega_{it}) + \epsilon_{d,it}, \tag{2.9}$$

$$\text{Structural model:} \quad \zeta_i(\omega_{it}) = \alpha_i + \beta_i\omega_{it} + e_{it}. \tag{2.10}$$

The first level of the hierarchical structure in (2.9) is a measurement model that relates the observed outcome $y_{d,it}$ to the latent construct $\zeta_i$ scaled by $c_d$; $\eta_{d,i} \sim \mathcal{N}(0, \sigma^2_{\eta d})$ is the random effects and $\boldsymbol{\gamma}_d$ is a $K_f$-vector of contrasts for the $d$-th outcome margin associated with the exogenous variables $\boldsymbol{x}_{it}$. The second level of the structure shown in (2.10) is a structural model having the same formulation as the random intercept and slopes model in (2.5) to describe the evolutionary process of the latent construct $\zeta_i$ over time $\omega_{it}$. Sammel and Ryan (1996) proposed a structural model in which fixed effect covariates are allowed to affecting $\zeta_i$ directly. In order for all model parameters to be identifiable, the random error $e_{it}$ is assumed to have a $\mathcal{N}(0,1)$ distribution and $c_1$ is a constant taking a value of 1.

An important feature of the latent growth curve model is that the time variable $\omega_{it}$ acts on the latent construct $\zeta_i$ directly, which in turn influences the $D$-dimensional observation $\boldsymbol{y}_{it} = (y_{1,it}, \ldots, y_{D,it})^\top$ such that the cross-sectional correlation structure in $\boldsymbol{y}_{it}$ is due to $\zeta_i$ (Roy and Lin, 2000). In other words, the model estimates a common growth trajectory shared between the marginal outcomes to characterise their observed variability across time (Rolfe, 2010; Wolf, 2016). Other possible extensions of latent growth curve modelling include accommodating a mixture of binary, ordinal, count and continuous data (Dunson, 2003), relaxing linear relationship between the observed outcomes and the latent construct (Proust et al., 2006; Proust-Lima et al., 2009), allowing individually varying measurement occasions (Sterba, 2014) and using a semi-parametric smooth function (Jacqmin-Gadda et al., 2010) or a finite mixture model in the structural model (Berlin et al., 2014; Lai et al., 2016).

## 2.2 Monte Carlo integration

Monte Carlo methods, which date back to the work of Metropolis and Ulam (1949), use stochastic simulation to approximate (2.3) via

$$\hat{f}_n^{MC} = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{\theta}_i), \quad \boldsymbol{\theta}_i \overset{iid}{\sim} \pi(\boldsymbol{\theta}), \tag{2.11}$$

with $\boldsymbol{\theta}_i$ being independent and identically distributed samples drawn from $\pi(\boldsymbol{\theta})$. The strong law of large numbers (Loève, 1977, Chapter 17) guarantees that the unbiased Monte Carlo estimate $\hat{f}_n^{MC}$ converges almost surely to the expectation $\mathbb{E}_\pi[f(\boldsymbol{\theta})]$ as $n \to \infty$. Moreover, the variance (and the mean squared error) of $\hat{f}_n^{MC}$ is given by

$$\mathbb{V}(\hat{f}_n^{MC}) = \frac{1}{n}\mathbb{V}(f(\boldsymbol{\theta})),$$

for $\mathbb{E}_\pi[f^2(\boldsymbol{\theta})] < \infty$. It is possible in certain cases to construct an estimator that is more efficient than $\hat{f}_n^{MC}$. To produce estimates of (2.3) with a lower variance for the same amount of simulation effort, or equivalently, achieve the same level of variability as $\hat{f}_n^{MC}$, but use fewer than $n$ samples, we discuss a few variance reduction techniques (Robert and Casella, 2004; Kroese et al., 2013; Rubinstein and Kroese, 2017) in the following sections.

### 2.2.1 Rao-Blackwellisation

A Rao-Blackwellised estimator (Gelfand and Smith, 1990; Casella and Robert, 1996) is based on the principle that analytical computation should be carried out as much as possible (Liu, 2001). Suppose that $\boldsymbol{\theta} = (\theta_1, \theta_2)^\top$ has two margins for simplicity and the conditional expectation $\mathbb{E}_\pi[f(\boldsymbol{\theta})|\theta_2]$ is analytically tractable, then given random samples $\{\theta_{2,1}, \ldots, \theta_{2,n}\}$ of $\theta_2$,

$$\hat{f}_n^{RB} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_\pi[f(\boldsymbol{\theta})|\theta_2 = \theta_{2,i}],$$

is an unbiased estimator of $\mathbb{E}_\pi[f(\boldsymbol{\theta})]$ by the law of total expectations, and additionally by the law of total variance,

$$\mathbb{V}(\hat{f}_n^{RB}) = \frac{1}{n}\mathbb{V}(\mathbb{E}_\pi[f(\boldsymbol{\theta})|\theta_2]) \le \mathbb{V}(\hat{f}_n^{MC}).$$

Variance reduction is achieved by replacing the random samples of $f(\boldsymbol{\theta})$ in (2.11) with exact values of its conditional expectation $\mathbb{E}_\pi[f(\boldsymbol{\theta})|\theta_2]$. Integrating out $\theta_1$ eliminates some of the randomness in the naive estimator and in turn, this marginalisation procedure means lower computational costs since only the simulation of $\theta_2$ is required for the approximation of $\mathbb{E}_\pi[f(\boldsymbol{\theta})]$ in (2.3). The Rao-Blackwellisation method has been used to estimate, for example, posterior model probabilities in Bayesian variable selection for regression models (Ghosh and Clyde, 2011) and to compute smoothed expectations in a general state space model (Olsson and Ryden, 2011).

### 2.2.2 Control variates

The construction of a control variate estimator relies on the availability of an exact solution to the expectation $\mathbb{E}_\pi[h(\boldsymbol{\theta})]$ for some proxy function $h$. A common choice of $h$ is the Taylor expansion of $f$, which is used extensively in a class of scalable Bayesian computational methods (Giles et al., 2016; Bardenet et al., 2017; Bierkens et al., 2019). A control variate estimator of $\mathbb{E}_\pi[f(\boldsymbol{\theta})]$ is then defined by

$$\hat{f}_n^{CV} = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{\theta}_i) - w(h(\boldsymbol{\theta}_i) - \mathbb{E}_\pi[h(\boldsymbol{\theta})]),$$

where $w$ is some weighting constant. Denote the correlation coefficient between $f(\boldsymbol{\theta})$ and $h(\boldsymbol{\theta})$ by $\rho$ and the covariance between the pair by $\mathrm{Cov}(f(\boldsymbol{\theta}), h(\boldsymbol{\theta}))$, it is straightforward to show that

$$\mathbb{V}(\hat{f}_n^{CV}) = \frac{1}{n}\Big(\mathbb{V}(f(\boldsymbol{\theta})) - 2w\,\mathrm{Cov}(f(\boldsymbol{\theta}), h(\boldsymbol{\theta})) + w^2\mathbb{V}(h(\boldsymbol{\theta}))\Big)$$

attains its minimum value of $(1-\rho^2)\mathbb{V}(\hat{f}_n^{MC})$ at the optimal weighing constant

$$w^* = \frac{\mathrm{Cov}(f(\boldsymbol{\theta}), h(\boldsymbol{\theta}))}{\mathbb{V}(h(\boldsymbol{\theta}))}, \tag{2.12}$$

hence $\hat{f}_n^{CV}$ has lower variability the more highly correlated the samples of $f(\boldsymbol{\theta})$ and $h(\boldsymbol{\theta})$ are. However, achieving the optimal condition in (2.12) is infeasible in practice because computing $w^*$ requires knowing $\mathbb{E}_\pi[f(\boldsymbol{\theta})]$. One possible solution is estimating $w^*$ using the sample moments of $f(\boldsymbol{\theta})$ and $h(\boldsymbol{\theta})$, see e.g. Glynn and Szechtman (2002). Alternatively, a more sophisticated approach based on the score function in Brooks and Gelman (1998) and Philippe and Robert (2001) can be implemented.

### 2.2.3 Antithetic variables

An estimator constructed from independent samples may not always be desirable. The formulation of an antithetic variable estimator is in fact similar to that of $\hat{f}_n^{MC}$, but with potentially correlated samples of $f(\boldsymbol{\theta})$. Suppose that both $\boldsymbol{\theta}$ and $\tilde{\boldsymbol{\theta}}$ are distributed according to $\pi(\boldsymbol{\theta})$, then the antithetic variable estimator of $\mathbb{E}_\pi[f(\boldsymbol{\theta})]$ is

$$\hat{f}_n^{AV} = \frac{1}{n}\sum_{i=1}^{n/2}(f(\boldsymbol{\theta}_i) + f(\tilde{\boldsymbol{\theta}}_i)),$$

with its variance given by

$$\mathbb{V}(\hat{f}_n^{AV}) = \frac{1}{n}\Big(\mathbb{V}(f(\boldsymbol{\theta})) + \mathrm{Cov}(f(\boldsymbol{\theta}), f(\tilde{\boldsymbol{\theta}}))\Big).$$

Therefore, $\hat{f}_n^{AV}$ will have a lower variance compared to $\hat{f}_n^{MC}$ provided that the joint distribution for $(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}})$ is chosen so that $\mathrm{Cov}(f(\boldsymbol{\theta}), f(\tilde{\boldsymbol{\theta}})) < 0$. In some cases, the construction of the dependence structure is simple, for example letting $\tilde{\boldsymbol{\theta}}$ be the corresponding values of $\boldsymbol{\theta}$ reflected about the mean of a symmetric $\pi(\boldsymbol{\theta})$ (Geweke, 1988), while others require careful designs (see Robert and Casella (2004) for a discussion). The intuition behind this style of variance reduction is that if one of a pair in $(f(\boldsymbol{\theta}_i), f(\tilde{\boldsymbol{\theta}}_i))$ overestimates $\mathbb{E}_\pi[f(\boldsymbol{\theta})]$ then the other provides a natural correction and vice versa.

## 2.3 Markov chain simulation

So far, our discussion assumes that it is trivial to simulate from $\pi(\boldsymbol{\theta})$. However, this is usually not the case as $\pi(\boldsymbol{\theta})$ is often a non-standard probability density function even for a simple model. This section describes a Bayesian posterior sampling approach known as Markov chain Monte Carlo (MCMC) methods, which produce approximate samples from $\pi(\boldsymbol{\theta})$ without having to sample directly from $\pi(\boldsymbol{\theta})$. Technically, MCMC methods generate a Markov chain whose invariant distribution is $\pi(\boldsymbol{\theta})$.

### 2.3.1 Gibbs sampling

The Gibbs sampler was formalised by Geman and Geman (1984) as an MCMC tool for simulating from high-dimensional distributions arising in image restoration, and subsequently developed further in the statistics literature by Gelfand and Smith (1990) to compute estimates of marginal probability distributions. Suppose that it is impossible to sample directly from the joint posterior distribution $\pi(\boldsymbol{\theta})$ where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_P)^\top$, but that sampling from the full conditional distribution of each margin $\pi_p(\theta_p|\theta_1, \ldots, \theta_{p-1}, \theta_{p+1}, \ldots, \theta_P)$ for $p = 1, \ldots, P$ is straightforward. The Gibbs sampler proceeds by updating the margins systematically, one at a time, conditional on the current values of the other margins (see Algorithm 1). Under relatively general conditions, it can be shown that the generated Markov chain $\{\boldsymbol{\theta}^{[1]}, \ldots, \boldsymbol{\theta}^{[n]}\}$, where $\boldsymbol{\theta}^{[i]} = (\theta_1^{[i]}, \ldots, \theta_P^{[i]})^\top$ is the sample generated in the $i$-th iteration, has invariant distribution $\pi(\boldsymbol{\theta})$ (Robert and Casella, 2004). However, the Gibbs sampler is limited since it requires the ability to sample from all $P$ full conditional distributions.

**Algorithm 1** Gibbs sampling

Initialise $\boldsymbol{\theta}^{[1]}$ with a random value such that $\pi(\boldsymbol{\theta}^{[1]}) > 0$. For $i = 2, \ldots, n$,

1. Sample $\theta_1^{[i]} \sim \pi_1(\theta_1|\theta_2^{[i-1]}, \ldots, \theta_P^{[i-1]})$.

   $\vdots$

$p$. Sample $\theta_p^{[i]} \sim \pi_p(\theta_p|\theta_1^{[i]}, \ldots, \theta_{p-1}^{[i]}, \theta_{p+1}^{[i-1]} \ldots, \theta_P^{[i-1]})$.

   $\vdots$

$P$. Sample $\theta_P^{[i]} \sim \pi_P(\theta_P|\theta_1^{[i]}, \ldots, \theta_{P-1}^{[i]})$.

### 2.3.2 Metropolis-Hastings algorithm

The Metropolis-Hastings (MH) algorithm, which is due to the work of Metropolis et al. (1953) and Hastings (1970), is a powerful Bayesian inference tool for generating samples of model parameters $\boldsymbol{\theta}$ from their posterior distribution $\pi(\boldsymbol{\theta})$ for more general problems compared to the Gibbs sampler as it only requires the likelihood function of the model of interest to be analytically tractable. Each iteration of the scheme involves sampling a proposed state $\boldsymbol{\theta}'$ from an arbitrary proposal distribution $q(\boldsymbol{\theta}'|\boldsymbol{\theta})$. The proposed value $\boldsymbol{\theta}'$ is then accepted or rejected with a certain probability according to the MH acceptance ratio to reflect how likely it is to move from the current state of $\boldsymbol{\theta}^{[i-1]}$ to the new state $\boldsymbol{\theta}'$ under the target distribution $\pi(\boldsymbol{\theta})$ (see

Algorithm 2). The posterior density $\pi(\boldsymbol{\theta})$ only needs to be known up to a normalising constant (marginal likelihood) because the normalising constant is cancelled out in the acceptance ratio. A rejection of $\boldsymbol{\theta}'$ implies that there is no change in the state between successive iterations. Note that the Gibbs sampler is a special instance of the MH algorithm with the full conditional distribution as the proposal distribution, for which the acceptance probability is always 1. Furthermore, the Markov chain generated will also yield the invariant distribution $\pi(\boldsymbol{\theta})$ when some of the Gibbs steps in the Gibbs sampler are replaced by the equivalent MH step updates (Johnson et al., 2013). Despite being a universal method for sampling difficult $\pi(\boldsymbol{\theta})$, the MH sampler is known to be very inefficient for high dimensional $\boldsymbol{\theta}$ as it performs local updates which then generate highly correlated samples (Sherlock et al., 2010; Neal, 2011).

**Algorithm 2** Metropolis-Hastings algorithm

Initialise $\boldsymbol{\theta}^{[1]}$ with a random value such that $\pi(\boldsymbol{\theta}^{[1]}) > 0$. For $i = 2, \ldots, n$,

1. Sample $\boldsymbol{\theta}' \sim q(\cdot|\boldsymbol{\theta}^{[i-1]})$.
2. Calculate the MH acceptance ratio given by

$$a^{[i-1]} = a(\boldsymbol{\theta}'|\boldsymbol{\theta}^{[i-1]}) = \min\left\{1, \frac{\pi(\boldsymbol{\theta}')q(\boldsymbol{\theta}^{[i-1]}|\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta}^{[i-1]})q(\boldsymbol{\theta}'|\boldsymbol{\theta}^{[i-1]})}\right\}.$$

3. Sample $\mathfrak{u} \sim \text{Uniform}(0, 1)$.
4. Set $\boldsymbol{\theta}^{[i]} \leftarrow \boldsymbol{\theta}'$ if $a^{[i-1]} > \mathfrak{u}$, otherwise set $\boldsymbol{\theta}^{[i]} \leftarrow \boldsymbol{\theta}^{[i-1]}$.

### 2.3.3 Hamiltonian Monte Carlo

Hamiltonian Monte Carlo (HMC; Neal, 2011; Betancourt, 2017) was initially introduced in the physics literature by Duane et al. (1987) and first used for statistical applications in Neal (1996). The widespread use of HMC is attributed to its ability to sample credible but distant proposal parameters for the MH algorithm, relying on the gradient information of the posterior density $\pi(\boldsymbol{\theta})$, and often its Hessian as well.

Instead of generating a Markov chain whose invariant distribution is $\pi(\boldsymbol{\theta})$, HMC introduces an auxiliary momentum variable $\boldsymbol{u}$ and targets the augmented distribution

$$\pi(\boldsymbol{\theta}, \boldsymbol{u}) \propto \exp(-\mathcal{H}(\boldsymbol{\theta}, \boldsymbol{u})),$$

where

$$\mathcal{H}(\boldsymbol{\theta}, \boldsymbol{u}) = -\log \pi(\boldsymbol{\theta}) + \frac{1}{2}\boldsymbol{u}^\top \boldsymbol{M}^{-1}\boldsymbol{u},$$

is known as the Hamiltonian which is made up of potential energy and kinetic energy components. The potential energy is defined as minus the log density of $\boldsymbol{\theta}$ under the target distribution $\pi(\boldsymbol{\theta})$, while the kinetic energy is due to the movement of the momentum variable $\boldsymbol{u}$ which is assumed to follow a $\mathcal{N}(\mathbf{0}, \boldsymbol{M})$ pseudo-prior with mass matrix $\boldsymbol{M}$. The Hamiltonian system is used to describe the evolution of $\boldsymbol{\theta}$ and $\boldsymbol{u}$ over time $t$ via the differential equations

$$\frac{d\boldsymbol{\theta}}{dt} = \frac{\partial \mathcal{H}}{\partial \boldsymbol{u}} \quad \text{and} \quad \frac{d\boldsymbol{u}}{dt} = -\frac{\partial \mathcal{H}}{\partial \boldsymbol{\theta}}. \tag{2.13}$$

Essentially, the loss in the kinetic energy of $\boldsymbol{u}$ is used to drive $\boldsymbol{\theta}$ to a region of higher density, and vice versa. An inherent property of the Hamiltonian is that it conserves energy, i.e. $d\mathcal{H}/dt = 0$ so that the proposed state $(\boldsymbol{\theta}', \boldsymbol{u}')$ obtained by solving (2.13) is always accepted. However, solution to the dynamics in (2.13) is typically intractable for complex $\pi(\boldsymbol{\theta})$ and requires implementing the leapfrog integrator (Neal, 2011), which discretises continuous time by a stepsize $\varepsilon$ so that

$$\begin{aligned}
\boldsymbol{u}(t+\varepsilon/2) &= \boldsymbol{u}(t) - (\varepsilon/2)\frac{\partial \mathcal{H}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}(t)) \\
\boldsymbol{\theta}(t+\varepsilon) &= \boldsymbol{\theta}(t) + \varepsilon\frac{\partial \mathcal{H}}{\partial \boldsymbol{u}}(\boldsymbol{u}(t+\varepsilon/2)) \\
\boldsymbol{u}(t+\varepsilon) &= \boldsymbol{u}(t+\varepsilon/2) - (\varepsilon/2)\frac{\partial \mathcal{H}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}(t+\varepsilon)).
\end{aligned} \tag{2.14}$$

Proposed values $\boldsymbol{\theta}'$ and $\boldsymbol{u}'$ obtained after a trajectory of length $\mathcal{T} = \mathfrak{L}\varepsilon$ by iterating the procedures in (2.14) $\mathfrak{L}$ times are then accepted with probability

$$\min\{1, \exp(\mathcal{H}(\boldsymbol{\theta}, \boldsymbol{u}) - \mathcal{H}(\boldsymbol{\theta}', \boldsymbol{u}'))\}. \tag{2.15}$$

Since the leapfrog integrator introduces approximation error, the MH accept-reject step in (2.15) is necessary in order to preserve the invariant distribution of the Markov chain generated as $\pi(\boldsymbol{\theta}, \boldsymbol{u})$. Samples from $\pi(\boldsymbol{\theta})$ are obtained by retaining the margins associated with $\boldsymbol{\theta}$. A summary description of HMC is provided in Algorithm 3. Despite providing an approximate solution to the differential equations in (2.13), the leapfrog integrator still largely conserves the energy in the Hamiltonian (Neal, 2011). This translates to a high sampler acceptance probability even when the proposal is distant, thereby suppressing the random walk behaviour observed in typical MH algorithms. Note that the updates performed in the leapfrog integrator are entirely deterministic, and the stochasticity of HMC is due to the random sampling of the momentum variable $\boldsymbol{u}$.

There are three tuning parameters which affect the performance of HMC sampling: the choice of the covariance matrix $\boldsymbol{M}$ of the momentum $\boldsymbol{u}$, the number of leapfrog updates $\mathfrak{L}$ and the stepsize $\varepsilon$. Betancourt (2017) suggest setting $\boldsymbol{M}$ to be the precision matrix of the variable $\boldsymbol{\theta}$ in order to achieve uniform energy level sets which lead to efficient exploration of the scaled target space. For a small value of $\mathfrak{L}$, exploration of the parameter space is reduced to a random walk. On the other hand, the simulated trajectory reverses direction and the proposed value $\boldsymbol{\theta}'$ is close to the initial value $\boldsymbol{\theta}$ if $\mathfrak{L}$ is chosen to be too large. To automate the tuning of this parameter, Hoffman

**Algorithm 3** Hamiltonian Monte Carlo

Initialise $\boldsymbol{\theta}^{[1]}$ with a random value such that $\pi(\boldsymbol{\theta}^{[1]}) > 0$. For $i = 2, \ldots, n$,

1. Sample $\boldsymbol{u}^{[i-1]} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{M})$.
2. Sample $(\boldsymbol{\theta}', \boldsymbol{u}')$ using the leapfrog integrator in (2.14).
3. Set $\boldsymbol{u}' \leftarrow -\boldsymbol{u}'$.
4. Calculate the MH acceptance ratio given by
$$a^{[i-1]} = a((\boldsymbol{\theta}', \boldsymbol{u}')|(\boldsymbol{\theta}^{[i-1]}, \boldsymbol{u}^{[i-1]})) = \min\{1, \exp(\mathcal{H}(\boldsymbol{\theta}^{[i-1]}, \boldsymbol{u}^{[i-1]}) - \mathcal{H}(\boldsymbol{\theta}', \boldsymbol{u}'))\}.$$
5. Sample $\mathfrak{u} \sim \text{Uniform}(0, 1)$.
6. Set $(\boldsymbol{\theta}^{[i]}, \boldsymbol{u}^{[i]}) \leftarrow (\boldsymbol{\theta}', \boldsymbol{u}')$ if $a^{[i-1]} > \mathfrak{u}$, otherwise set $(\boldsymbol{\theta}^{[i]}, \boldsymbol{u}^{[i]}) \leftarrow (\boldsymbol{\theta}^{[i-1]}, \boldsymbol{u}^{[i-1]})$.

and Gelman (2014) develop the No-U-Turn Sampler (NUTS) which determines an optimal value of $\mathfrak{L}$ using a tree building algorithm. The tree building algorithm can be briefly described as follows: a single leapfrog update is performed from the current state of $\boldsymbol{\theta}$, followed by two more updates and then four more, with the doubling process being terminated when the simulated trajectory of $\boldsymbol{\theta}$ first retraces its steps back to its initialised value. The number of doubling procedure undertaken is known as the tree depth. In the same paper, the authors also provide a heuristic for setting $\varepsilon$ based on the dual averaging method of Nesterov (2009), which is used predominantly for non-smooth and stochastic convex optimisation, to target a desired level of acceptance probability.

### 2.3.4 Assessing convergence

It is crucial to monitor the convergence of all parameters drawn using the algorithms described in Sections 2.3.1–2.3.3 in order to ensure that the samples are representative of the posterior distribution $\pi(\boldsymbol{\theta})$. For simplicity, suppose that we have a univariate parameter $\theta$ and we simulate $m$ parallel chains, each of length $n$ after discarding the burn in. Defining $\theta^{[i,j]}, i = 1, \ldots, n, j = 1, \ldots, m$ as the $i$-th iterate of $\theta$ in the $j$-th chain, the between- and within-chain variances are given, respectively, by

$$B = \frac{n}{m-1}\sum_{j=1}^{m}(\bar{\theta}^{[\cdot,j]} - \bar{\theta}^{[\cdot,\cdot]})^2, \quad \text{where } \bar{\theta}^{[\cdot,j]} = \frac{1}{n}\sum_{i=1}^{n}\theta^{[i,j]}, \quad \bar{\theta}^{[\cdot,\cdot]} = \frac{1}{m}\sum_{j=1}^{m}\bar{\theta}^{[\cdot,j]}$$

$$W = \frac{1}{m}\sum_{j=1}^{m}s_j^2, \quad \text{where } s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\theta^{[i,j]} - \bar{\theta}^{[\cdot,j]})^2.$$

Convergence can be assessed by computing the factor by which the scale of the current distribution for $\theta$ is reduced in the limit of $n \to \infty$ (Gelman and Rubin, 1992), which can be estimated from

$$\hat{R} = \sqrt{\frac{n-1}{n} + \frac{B}{nW}}, \tag{2.16}$$

which approaches 1 as $n \to \infty$.

### 2.3.5 Effective number of simulation draws

The inefficiency of an MCMC sampler in estimating $\mathbb{E}_\pi[f(\boldsymbol{\theta})]$ is usually measured by the integrated autocorrelation time (Roberts and Rosenthal, 2009), which is defined as

$$\text{IACT}_f = 1 + \sum_{j=1}^{\infty} \rho_{j,f},$$

where $\rho_{j,f}$ is the lag $j$ autocorrelation function of the MCMC iterates of $f(\boldsymbol{\theta})$ after convergence. Alternatively, one can measure the efficiency of the sampler by computing the effective sample size per MCMC iteration, which by definition is the reciprocal of the IACT (Kass et al., 1998; Robert and Casella, 2004), i.e.

$$\text{ESS} = \frac{1}{\text{IACT}_f}.$$

A small value of the IACT, under the assumptions that the invariant distribution of the generated Markov chain is $\pi(\boldsymbol{\theta})$, is desirable in practice as it indicates that the Markov chain mixes well (Pitt et al., 2012; Doucet et al., 2015).

## 2.4 Bayesian non-parametric methods

With advances in the computing technology, it has now become a common practice for practitioners to consider various models of differing complexity $\{\mathcal{M}_1, \ldots, \mathcal{M}_Q\}$, and obtain the predictive distribution $p(\boldsymbol{y}_\star|\boldsymbol{y})$ of a future observation $\boldsymbol{y}_\star$ from the same stochastic process that generated $\boldsymbol{y}$ via Bayesian model averaging (e.g. Draper, 1995; Hoeting et al., 1999; Clyde and George, 2004), i.e.

$$p(\boldsymbol{y}_\star|\boldsymbol{y}) = \sum_{q=1}^{Q} p(\boldsymbol{y}_\star|\mathcal{M}_q)\pi(\mathcal{M}_q|\boldsymbol{y}),$$

where $p(\boldsymbol{y}_\star|\mathcal{M}_q)$ is the predictive distribution of $\boldsymbol{y}_\star$ conditional on the model $\mathcal{M}_q$ whose parameter is $\boldsymbol{\theta}_{\mathcal{M}_q}$ and given prior model probabilities $p(\mathcal{M}_q)$, $\pi(\mathcal{M}_q|\boldsymbol{y})$ is the posterior model probability upon observing $\boldsymbol{y}$ computed from

$$\pi(\mathcal{M}_q|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\mathcal{M}_q)p(\mathcal{M}_q)}{\sum_{q=1}^{Q} p(\boldsymbol{y}|\mathcal{M}_q)p(\mathcal{M}_q)}. \tag{2.17}$$

An important quantity required for the analysis in (2.17) is the marginal likelihood of model $\mathcal{M}_q$ whose expression is given in (2.2). However, computation for (2.2) is notoriously challenging and $\{\mathcal{M}_1, \ldots, \mathcal{M}_Q\}$ only represents a finite subset of models in the space $\boldsymbol{\mathcal{M}}$ of all possible models (Ghahramani, 2013).

This section describes a special class of methods known as Bayesian non-parametric (BNP) methods (Hjort et al., 2010; Gershman and Blei, 2012) which adopt a model $\mathcal{M}_\infty \in \boldsymbol{\mathcal{M}}$ whose parameter $\boldsymbol{\theta}$ lies in an infinite-dimensional space $\Theta$. BNP models account for model uncertainty by using a finite number of parameters in $\Theta$ to describe the generating mechanism of $\boldsymbol{y}$ in such a way that the complexity of the model $\mathcal{M}_\infty$ is entirely data-driven (Orbanz and Teh, 2010), rather than comparing multiple models that vary in complexity (Gershman and Blei, 2012). Common examples of BNP methods used in statistical applications include Gaussian processes (Rasmussen and Williams, 2006) and Dirichlet processes (Ferguson, 1973).

### 2.4.1 Dirichlet processes

The Dirichlet process (DP) belongs to a family of stochastic processes which is used extensively in Bayesian mixture modelling (Rasmussen, 2000; Zhang et al., 2005; da Silva, 2007) when the number of components $G$ in a mixture distribution, i.e.

$$\sum_{g=1}^{G} w_g p_{\mathcal{K}}(\boldsymbol{y}|\boldsymbol{\theta}_g) \quad \text{with} \quad \sum_{g=1}^{G} w_g = 1, \tag{2.18}$$

where each component density $p_{\mathcal{K}}(\boldsymbol{y}|\boldsymbol{\theta}_g)$ comes from the same family of distribution $\mathcal{K}(\boldsymbol{\theta})$ weighted by $w_g \geq 0$, is unknown *a priori*. Let $\mathcal{G}_0$ be a probability measure on the measurable space $(\Theta, \mathcal{F})$, where $\Theta$ is a non-empty set and $\mathcal{F}$ is the $\sigma$-algebra on $\Theta$. A realisation, $\mathcal{G}$, from a Dirichlet process $\mathcal{DP}(\lambda, \mathcal{G}_0)$ with concentration parameter

$\lambda > 0$ and base distribution $\mathcal{G}_0$ is a discrete probability distribution, taking the form

$$\mathcal{G} = \sum_{g=1}^{\infty} w_g \delta_{\boldsymbol{\theta}_g}, \tag{2.19}$$

where $\{w_g\}_{g=1}^{\infty}$ is an infinite sequence of non-negative weights which sum to 1, $\{\boldsymbol{\theta}_g\}_{g=1}^{\infty}$ are independent random samples (also known as atoms) drawn from $\mathcal{G}_0$ and $\delta_{\boldsymbol{\theta}}$ is a point mass located at $\boldsymbol{\theta}$. The base distribution $\mathcal{G}_0$ determines the support of the (almost surely) discrete distribution $\mathcal{G}$ in (2.19), whereas the concentration parameter $\lambda$ controls the variability of $\mathcal{G}$ around $\mathcal{G}_0$.

While the sampling of the atoms is straightforward, the construction of $\{w_g\}_{g=1}^{\infty}$ is non-trivial. Sethuraman (1994) provides a constructive definition of the infinite sequence of weights in the DP using the stick-breaking process whereby metaphorically, a stick, initially of unit length, is repeatedly broken at a random lengths, as determined by a Beta random variable $\gamma$. In such a manner, the weights are constructed as

$$w_g = \gamma_g \prod_{h=1}^{g-1}(1 - \gamma_h), \quad \gamma_g \sim \text{Beta}(1, \lambda). \tag{2.20}$$

Note that the weight decreases stochastically so that $w_g$ is almost negligible for a large value of $G$ (which depends on $\lambda$).

We now discuss some theoretical properties of a DP. For any finite measurable partition $S_1, \ldots, S_J$ of $\Theta$ such that $S_j \cap S_{j'} = \varnothing$ for $j \neq j'$ and $\bigcup_{j=1}^{J} S_j = \Theta$, the marginal distribution of a DP is distributed according to a Dirichlet distribution by definition (Ferguson, 1973), so that

$$(\mathcal{G}(S_1), \ldots, \mathcal{G}(S_J)) \sim \text{Dirichlet}(\lambda\mathcal{G}_0(S_1), \ldots, \lambda\mathcal{G}_0(S_J)).$$

Using properties of the Dirichlet distribution, the mean and variance of a DP can be easily obtained as

$$\mathbb{E}[\mathcal{G}(S)] = \mathcal{G}_0(S), \quad \text{and} \quad \mathbb{V}(\mathcal{G}(S)) = \frac{\mathcal{G}_0(S)(1 - \mathcal{G}_0(S))}{1 + \lambda},$$

for any measurable set $S$ of $\Theta$. The concentration parameter $\lambda$ dictates the variability of $\mathcal{G}$ around $\mathcal{G}_0$: $\mathcal{G}$ converges to $\mathcal{G}_0$ pointwise in the limit as $\lambda \to \infty$, whereas $\mathcal{G}$ collapses to a point distribution in the limit as $\lambda \to 0$. Figure 2.2 illustrates this, where realisations from a DP with $\mathcal{G}_0$ being a standard normal distribution and different values of $\lambda$ are shown. For small $\lambda$, $\mathcal{G}$ has large masses concentrated on a few atoms, and its empirical cumulative distribution function (CDF) is highly variable. As $\lambda$ increases, more atoms have non-negligible weights, and the empirical CDF of $\mathcal{G}$ converges to the CDF of the base distribution (shown by the solid black line).

Suppose that $\mathcal{G} \sim \mathcal{DP}(\lambda, \mathcal{G}_0)$, and that $\boldsymbol{\theta}_{1:N} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N\}$ are independent and identically distributed samples generated from the atomic distribution $\mathcal{G}$. Define $n_j := |i : \boldsymbol{\theta}_i \in S_j|$ to be the number of samples observed in the measurable set $S_j, j = 1, \ldots, J$ such that $\sum_{j=1}^{J} n_j = N$. Using the conjugacy property of a Dirichlet

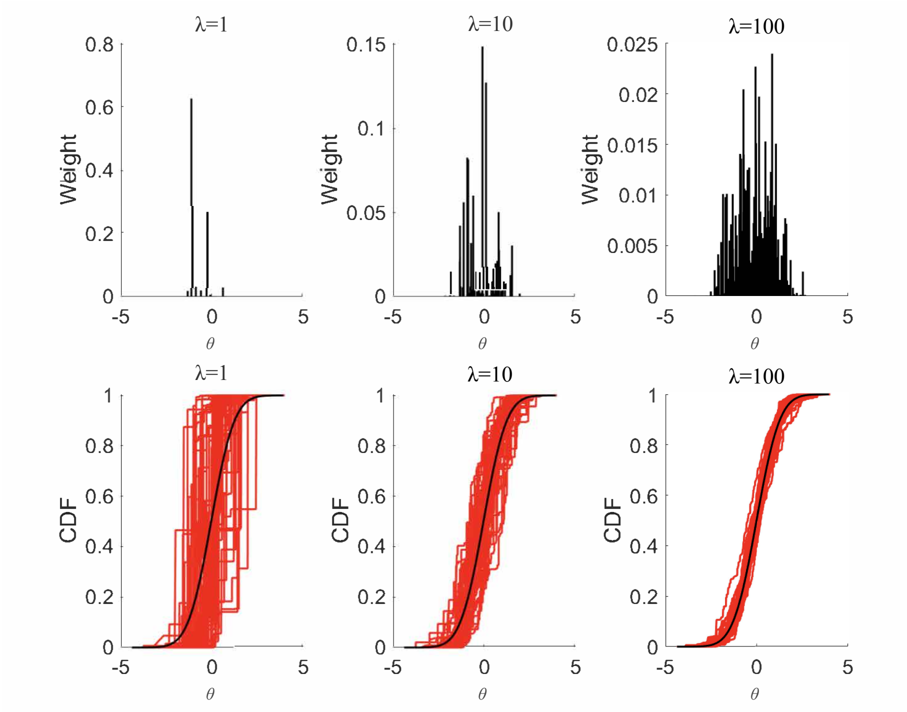



Figure 2.2: Realisations (top panel) resulting from a random draw from the DP prior $\mathcal{DP}(\lambda, \mathcal{G}_0)$ with $\lambda = 1, 10, 100$ and $\mathcal{G}_0$ is a standard normal distribution. Empirical CDFs (bottom panel) of 50 samples generated from $\mathcal{DP}(\lambda, \mathcal{G}_0)$ for each value of $\lambda$ are plotted against the CDF of $\mathcal{G}_0$ (black curve).

prior distribution to a multinomial likelihood function (Wade et al., 2011), we obtain

$$(\mathcal{G}(S_1), \ldots, \mathcal{G}(S_J))|\boldsymbol{\theta}_{1:N} \sim \text{Dirichlet}(\lambda \mathcal{G}_0(S_1) + n_1, \ldots, \lambda \mathcal{G}_0(S_J) + n_J). \tag{2.21}$$

Equation (2.21) holds for any finite measurable partition $\{S_1, \ldots, S_J\}$ of $\Theta$, and thus by definition the posterior distribution of $\mathcal{G}$ is itself another DP with concentration parameter $\widetilde{\lambda} = \lambda + N$ and base distribution

$$\widetilde{\mathcal{G}}_0 = \frac{\lambda}{\lambda + N}\mathcal{G}_0 + \frac{N}{\lambda + N}\sum_{i=1}^{N} \frac{\delta_{\boldsymbol{\theta}_i}}{N}.$$

This shows that the DP is a conjugate prior for the random distribution $\mathcal{G}$, and that the posterior base distribution $\widetilde{\mathcal{G}}_0$ is a weighted average of the prior base distribution (whose informativity is given by $\lambda$) and the empirical distribution resulting from the samples $\boldsymbol{\theta}_{1:N}$.

The DP induces a clustering effect and this makes it a powerful tool in classification problems when the actual number of classes $G$ is unknown. To illustrate this property, consider the following DP mixture model (Antoniak, 1974) in which

$$\boldsymbol{y}_i|\tilde{\boldsymbol{\theta}}_i \sim p_{\mathcal{K}}(\boldsymbol{y}|\tilde{\boldsymbol{\theta}}_i), \quad \tilde{\boldsymbol{\theta}}_i|\mathcal{G} \sim \mathcal{G}, \quad \mathcal{G} \sim \mathcal{DP}(\lambda, \mathcal{G}_0), \tag{2.22}$$

for $i = 1, \ldots, N$, where $\tilde{\boldsymbol{\theta}}_i$ is the parameter of the density $p_{\mathcal{K}}(\boldsymbol{y}|\tilde{\boldsymbol{\theta}}_i)$ responsible for generating the observation $\boldsymbol{y}_i$. Integrating out the discrete distribution $\mathcal{G}$ from (2.22), Blackwell and MacQueen (1973) show that the conditional prior distribution induced on $\tilde{\boldsymbol{\theta}}_i$ follows a Pólya urn scheme where

$$\tilde{\boldsymbol{\theta}}_i|\tilde{\boldsymbol{\theta}}_1, \ldots, \tilde{\boldsymbol{\theta}}_{i-1} \sim \frac{1}{\lambda + i - 1}\sum_{j=1}^{i-1} \delta_{\tilde{\boldsymbol{\theta}}_j} + \frac{\lambda}{\lambda + i - 1}\mathcal{G}_0. \tag{2.23}$$

Equation (2.23) suggests that the first parameter $\tilde{\boldsymbol{\theta}}_1$ is initialised by drawing a random sample from the base distribution $\mathcal{G}_0$ of the DP. Conditional on the set of previous samples drawn $\{\tilde{\boldsymbol{\theta}}_1, \ldots, \tilde{\boldsymbol{\theta}}_{i-1}\}$, the following parameter, $\tilde{\boldsymbol{\theta}}_i$, is determined probabilistically. With probability proportional to the concentration parameter $\lambda$, $\tilde{\boldsymbol{\theta}}_i$ is set to be a new random draw from $\mathcal{G}_0$, and with probability proportional to $i-1$, $\tilde{\boldsymbol{\theta}}_i$ is sampled uniformly from $\{\tilde{\boldsymbol{\theta}}_1, \ldots, \tilde{\boldsymbol{\theta}}_{i-1}\}$. The former introduces a new distinct value, while the latter induces clustering effect so that the generated parameters $\{\tilde{\boldsymbol{\theta}}_1, \ldots, \tilde{\boldsymbol{\theta}}_N\}$ concentrate on a set of unique values $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G\}$, with $G \ll N$. A larger value of $\lambda$ increases the probability of generating $\tilde{\boldsymbol{\theta}}_i$ from $\mathcal{G}_0$, which in turn gives rise to a larger value of $G$. In fact, Teh (2011) show that for $N, \lambda \gg 0$,

$$\mathbb{E}[G] \simeq \lambda \log\left(1 + \frac{N}{\lambda}\right),$$

indicating that the mean of $G$ scales logarithmically with the size of the dataset, $N$. Note that the value of $G$ is bounded above by $N$, and its (random) value can be determined via posterior inference.

Bayesian inference for mixture models often suffers from the label switching problem (Celeux et al., 2000; Jasra et al., 2005) due to the invariance of the likelihood function in (2.18) to permutations of the labels of the mixture components. This makes identification of and inference for each component (i.e. the clusters of individuals) challenging. Identification of each component is further complicated in a DP mixture model since $G$ is variable. Fritsch and Ickstadt (2009) proposed addressing this issue using Bayesian decision theory, where the best decision rule (for component membership) satisfies certain optimality conditions. A popular measure used for comparing competing membership clusterings $\boldsymbol{s}$ and $\tilde{\boldsymbol{s}}$, with $G$ and $\tilde{G}$ clusters respectively, is the adjusted Rand index (ARI; Hubert and Arabie, 1985) defined as

$$ARI(\boldsymbol{s}, \tilde{\boldsymbol{s}}) = \frac{\sum_{g=1}^{G} \sum_{h=1}^{\tilde{G}} \binom{N_{gh}}{2} - \sum_{g=1}^{G} \binom{N_{g+}}{2} \sum_{h=1}^{\tilde{G}} \binom{N_{+h}}{2} / \binom{N}{2}}{\frac{1}{2}\left(\sum_{g=1}^{G} \binom{N_{g+}}{2} + \sum_{h=1}^{\tilde{G}} \binom{N_{+h}}{2}\right) - \sum_{g=1}^{G} \binom{N_{g+}}{2} \sum_{h=1}^{\tilde{G}} \binom{N_{+h}}{2} / \binom{N}{2}}.$$

Here $N_{gh}$ is the number of individuals in group $g$ of membership clustering $\boldsymbol{s}$ that are also in group $h$ of membership clustering $\tilde{\boldsymbol{s}}$, $N_{g+} = \sum_{h=1}^{\tilde{G}} N_{gh}$, $N_{+h} = \sum_{g=1}^{G} N_{gh}$ and $\binom{n}{2} = n(n-1)/2$ is the Binomial coefficient. The ARI measures the similarity between the two clusterings, mostly taking values between 0 (for completely random clustering) and 1 (for identical clusterings). Thus, values closer to 1 indicate clusterings that are more consistent. Negative values of the ARI are possible but they have no substantive use. In the current context, the optimal clustering $\hat{\boldsymbol{s}}$ maximises the posterior expected adjusted Rand (PEAR) index. That is,

$$\hat{\boldsymbol{s}} = \arg\max_{\tilde{\boldsymbol{s}}} \mathbb{E}_{\boldsymbol{s}}[ARI(\boldsymbol{s}, \tilde{\boldsymbol{s}})],$$

where the expectation $\mathbb{E}_{\boldsymbol{s}}$ is taken with respect to the posterior distribution of $\boldsymbol{s}$.

# Chapter 3

# Efficient data augmentation for multivariate probit models with panel data: An application to general practitioner decision-making about contraceptives

## 3.1 Introduction

Bayesian inference for the multivariate probit (MVP) model is usually performed using the data augmentation representation of Chib and Greenberg (1998), whereby the latent variables indicating the observed outcomes are normally distributed. For unique identification of the regression parameters, the covariance matrix of these latent normal random variates is assumed to be a correlation matrix $\boldsymbol{R}_{\boldsymbol{\epsilon}}$. However, Monte Carlo sampling for $\boldsymbol{R}_{\boldsymbol{\epsilon}}$ in a Bayesian context is difficult due to the restrictions on the diagonal entries and the requirement that the matrix $\boldsymbol{R}_{\boldsymbol{\epsilon}}$ must be positive definite (Chib and Greenberg, 1998; Edwards and Allenby, 2003; Smith, 2013).

This chapter presents three contributions, two methodological and the third a subject matter one. The first methodological contribution provides an improved method for sampling the potentially high dimensional correlation matrix $\boldsymbol{R}_{\boldsymbol{\epsilon}}$ within a Markov chain Monte Carlo (MCMC) algorithm. Chib and Greenberg (1998) suggest sampling the correlation coefficients in blocks using a random walk Metropolis-Hastings (RWMH) algorithm with a multivariate $t$ proposal density whose degree of freedom is specified arbitrarily as 10, or some similar values. However, the resulting matrix obtained after each proposal is not guaranteed to be a valid correlation matrix and such proposals are rejected with certainty in the MH algorithm, in addition to the RWMH algorithm being notorious for its slow exploration of the parameter space (Sherlock et al., 2010; Neal, 2011). Tuning the parameters of this proposal distribution also requires finding an approximate mode of the log posterior distribution and the observed Fisher information for every iteration, resulting in high computational overheads. On the other hand, Barnard et al. (2000) adopt the Griddy-Gibbs sampler of Ritter and Tanner (1992) to sample $\boldsymbol{R}_{\boldsymbol{\epsilon}}$ in hierarchical regression models. Here, prior to the Gibbs step, one needs to solve a quadratic equation to determine the support for a single correlation coefficient (while keeping the rest fixed) which results in a valid correlation matrix. The design of drawing one correlation coefficient at a time becomes computationally prohibitive when the dimension of $\boldsymbol{R}_{\boldsymbol{\epsilon}}$ is large. In order to circumvent the positive definiteness restriction imposed on a correlation matrix, we adopt the reparameterisation strategy of Smith (2013) which re-expresses $\boldsymbol{R}_{\boldsymbol{\epsilon}}$ as an unconstrained Cholesky factor $\boldsymbol{\mathcal{L}}_{\boldsymbol{\epsilon}}$. This maps the manifold space of a correlation matrix to a Euclidean space, which improves the efficiency of posterior simulation while keeping the number of unknown parameters the same. We employ the Hamiltonian Monte Carlo (HMC) algorithm (Neal, 2011) to sample the high dimensional $\boldsymbol{\mathcal{L}}_{\boldsymbol{\epsilon}}$ efficiently, thereby avoiding the slow exploration of parameter space by random walk updates.

The second methodological contribution is to introduce antithetic sampling, based on the work of Hammersley and Morton (1956), into the Metropolis-Hastings (MH) literature in the context of MVP. In order to implement this idea, we specify the proposal distribution of parameter update as a deterministic function. Here, the generated samples will be super-efficient in terms of the reduction in variance of the Monte Carlo estimates compared to the same estimates constructed from uncorrelated samples. Although the proposal chain update is deterministic, the convergence properties are not compromised when this is embedded within a larger system of MCMC sampling based on our empirical results. Our proposed methodology is motivated by the over-relaxation algorithm (Adler, 1981; Barone and Frigessi, 1990), and is similar to the idea built within the framework of HMC in Pakman and Paninski (2014). However, our proposed sampler is different from these methods in two main aspects. First, there is no randomness in the proposal distribution for parameter updates in our method, whereas theirs still retain a certain degree of stochasticity. Second, we introduce perfect negative correlation between successive MCMC samples via the deterministic proposal, while they suggest partial or zero dependence between the samples. Results based on our real data application investigating the decision-making of general practitioners about contraceptives document a significant improvement of up to a 16 times performance gain in the mixing behaviour of the Markov chain, thereby lowering the autocorrelation between the iterates. The computing time of the algorithm is also marginally reduced due to the deterministic sampling.

Our third contribution is a methodological development is motivated by the staged stated preference panel data collection described in Fiebig et al. (2017), which is used to study the decision-making of Australian general practitioners (GPs) about female contraceptive products. Here, the authors used the data from the final stage of the three-stage decision process, whereas we explore outcomes from the second stage. This second stage relates to the question of which particular contraceptive products GPs would discuss with a female patient, defined by a vignette that is part of the experimental design. Separate univariate analyses on each product would ignore possible complex dependence structures that are useful in exploring which particular bundles of products are discussed with patients. This is important here because in any correlated choice problem there may be multiple close substitutes, which makes joint rather than marginal probabilities more relevant. Therefore, we model the GPs' choices by an MVP model, and employ our proposed methods to improve the sampling efficiency of the correlation matrix $\boldsymbol{R}_{\boldsymbol{\epsilon}}$ and the regression coefficients of covariates in the model. Inspection of the resulting graphical model describing this interaction between products lends support to the suitability of a multivariate approach. By using the MVP model, we are able to compute the joint probability of specific product bundles being discussed with a patient. Posterior estimation of this probability, based on a patient with certain socio-economic and clinical characteristics for which there is strong clinical evidence on the suitability of long acting contraceptive choices, reveals differing views among the GPs in the sample. This variability is known as medical practice variation in the health industry; see for example Wennberg et al. (1982), Scott and Shiell (1997) and Davis et al. (2000), whereby the decision making of GPs is influenced by both their personal characteristics such as gender, age and qualifications, as well as other unobservables that we model as random effects.

The rest of the chapter is organised as follows. Section 3.2 describes the MVP model with random effects. Section 3.3 presents our proposed methodology of sampling $\boldsymbol{R}_{\boldsymbol{\epsilon}}$, and Section 3.4 outlines the antithetic sampling technique whose efficiency is illustrated via simulation studies in Section 3.5. Section 3.6 provides our analysis of the discussion preference data of contraceptive products by Australian GPs, and Section 3.7 concludes. Appendices 3.8.1–3.8.5 provide further details on the contraceptive product data analysis.

## 3.2 Multivariate probit model with random effects

The MVP model has been used extensively to model correlated binary data (Gibbons and Wilcox-Gök, 1998; Buchmueller et al., 2013). Let $\boldsymbol{y}_{it} = (y_{1,it}, \ldots, y_{D,it})^\top$ be a vector of $D$ correlated binary outcomes for individual $i = 1, \ldots, N$ at time period $t$, for $t = 1, \ldots, T$. The latent variable representation of the MVP model, using the data augmentation approach of Albert and Chib (1993), is given by

$$\boldsymbol{y}_{it}^* = \boldsymbol{\alpha}_i + \boldsymbol{B}\boldsymbol{x}_{it} + \boldsymbol{\epsilon}_{it}, \tag{3.1}$$

$$\boldsymbol{\alpha}_i = (\alpha_{1,i}, \ldots, \alpha_{D,i})^\top \overset{iid}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\alpha}}), \tag{3.2}$$

$$\boldsymbol{\epsilon}_{it} = (\epsilon_{1,it}, \ldots, \epsilon_{D,it})^\top \overset{iid}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{R}_{\boldsymbol{\epsilon}}), \tag{3.3}$$

for $i = 1, \ldots, N, t = 1, \ldots, T$, where $\boldsymbol{y}_{it}^* = (y_{1,it}^*, \ldots, y_{D,it}^*)^\top$ is a continuous latent variable, $\boldsymbol{\alpha}_i$ is a $D$-vector of outcome-specific random intercepts for individual $i$ allowing for heterogeneity between individuals, $\boldsymbol{x}_{it} = (1, x_{1,it}, \ldots, x_{K-1,it})^\top$ is a set of exogenous variables, $\boldsymbol{B}$ is a $D \times K$ matrix of regression coefficients and $\boldsymbol{\epsilon}_{it}$ is a $D$-vector correlated error term which models the dependence structure between outcomes. The variable $\boldsymbol{x}_{it}$ is assumed to be uncorrelated with both $\boldsymbol{\alpha}_i$ and $\boldsymbol{\epsilon}_{it}$. This is entirely appropriate in the stated preference case that is our motivating analysis but relaxing the assumption of exogenous $\boldsymbol{x}_{it}$ represents a useful extension. In order for $\boldsymbol{B}$ to be uniquely identified (Chib and Greenberg, 1998), $\boldsymbol{R}_{\boldsymbol{\epsilon}}$ is set to be a correlation matrix. The observed outcome $\boldsymbol{y}_{it}$ is defined to be dependent on the latent variable $\boldsymbol{y}_{it}^*$ via the relationship

$$y_{d,it} = \mathbb{1}(y_{d,it}^* > 0), \quad d = 1, \ldots, D, \tag{3.4}$$

where $\mathbb{1}(\boldsymbol{E})$ is an indicator function which takes value 1 if the event $\boldsymbol{E}$ occurs and 0 otherwise. Let $\boldsymbol{y} = \{\boldsymbol{y}_{it}; i = 1, \ldots, N, t = 1, \ldots, T\}$ be the set of observed discrete outcomes. The density of the latent continuous variables $\boldsymbol{y}^*$ conditional on the random effects $\boldsymbol{\alpha}_{1:N} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_N)$ is given by

$$p(\boldsymbol{y}^* | \boldsymbol{\alpha}_{1:N}, \boldsymbol{\theta}) = \prod_{i=1}^{N} \prod_{t=1}^{T} \phi(\boldsymbol{y}_{it}^*; \boldsymbol{\mu}_{it}, \boldsymbol{R}_{\boldsymbol{\epsilon}}),$$

where $\boldsymbol{\theta} := (\boldsymbol{B}, \boldsymbol{R}_{\boldsymbol{\epsilon}}, \boldsymbol{\Sigma}_{\boldsymbol{\alpha}})$ denotes the vector of model parameters, $\boldsymbol{\mu}_{it} = \boldsymbol{\alpha}_i + \boldsymbol{B}\boldsymbol{x}_{it}$ and $\phi$ is the multivariate normal density function in $D$ dimensions.

Following the specification of the MVP model in (3.1)–(3.4), the posterior density is

$$\pi(\boldsymbol{y}^*, \boldsymbol{\alpha}_{1:N}, \boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{y}^*, \boldsymbol{\alpha}_{1:N}, \boldsymbol{\theta})p(\boldsymbol{y}^*|\boldsymbol{\alpha}_{1:N}, \boldsymbol{\theta})p(\boldsymbol{\alpha}_{1:N}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{y})}, \tag{3.5}$$

where $p(\boldsymbol{y}) = \int p(\boldsymbol{y}|\boldsymbol{y}^*, \boldsymbol{\alpha}_{1:N}, \boldsymbol{\theta})p(\boldsymbol{y}^*|\boldsymbol{\alpha}_{1:N}, \boldsymbol{\theta})p(\boldsymbol{\alpha}_{1:N}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$ is the marginal likelihood, $p(\boldsymbol{\theta})$ is the prior on all of the model parameters $\boldsymbol{\theta}$ and

$$p(\boldsymbol{y}|\boldsymbol{y}^*, \boldsymbol{\alpha}_{1:N}, \boldsymbol{\theta}) = \prod_{i=1}^{N}\prod_{t=1}^{T}\prod_{d=1}^{D}\Big(\mathbb{1}(y_{d,it} = 0)\mathbb{1}(y^*_{d,it} \leq 0) + \mathbb{1}(y_{d,it} = 1)\mathbb{1}(y^*_{d,it} > 0)\Big),$$

is the distribution of the data conditional on the unobserved latent variables $\boldsymbol{y}^*$.

## 3.3 Efficient sampling for $\boldsymbol{R}_{\boldsymbol{\epsilon}}$ when using a marginally uniform prior

This section describes an efficient way of sampling $\boldsymbol{R}_{\boldsymbol{\epsilon}}$ by utilising Hamiltonian dynamics (Duane et al., 1987). This involves reparameterising $\boldsymbol{R}_{\boldsymbol{\epsilon}}$ to enable sampling of parameters in an unconstrained space rather than $\boldsymbol{\mathcal{R}}^D$. Due to the attractive properties of the marginally uniform prior in (2.4) with $\nu = D + 1$ being invariant to different ordering of the outcome variables $\boldsymbol{y}$ and induces a shrinkage prior on the partial correlation (detailed discussion in Section 2.1.1), we will use this prior hereafter. Inference for the posterior distribution of $(\boldsymbol{B}, \boldsymbol{\alpha}_{1:N}, \boldsymbol{\Sigma_\alpha})$ in (3.5) can be performed using a Gibbs sampler (see Chapter 10 of Greenberg (2012) for details). Our focus here is on the following non-standard conditional posterior distribution

$$\pi(\boldsymbol{R}_{\boldsymbol{\epsilon}}|\boldsymbol{y}, \boldsymbol{y}^*, \boldsymbol{\alpha}_{1:N}, \boldsymbol{\theta}_{-\boldsymbol{R}_{\boldsymbol{\epsilon}}}) \propto \prod_{i=1}^{N}\prod_{t=1}^{T}\phi(\boldsymbol{y}^*_{it}; \boldsymbol{\mu}_{it}, \boldsymbol{R}_{\boldsymbol{\epsilon}}) \cdot p(\boldsymbol{R}_{\boldsymbol{\epsilon}}), \tag{3.6}$$

where $\boldsymbol{\theta}_{-\boldsymbol{R}_{\boldsymbol{\epsilon}}}$ is defined as $\boldsymbol{\theta}$, but excluding the parameters $\boldsymbol{R}_{\boldsymbol{\epsilon}}$.

### 3.3.1 An unconstrained parameterisation

Because of the restrictions on sampling correlation coefficients on a confined space, we adopt the reparameterisation strategy in Smith (2013) which re-expresses $\boldsymbol{R}_{\boldsymbol{\epsilon}}$ via a positive definite matrix $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ as

$$\boldsymbol{R}_{\boldsymbol{\epsilon}} = \boldsymbol{\Lambda}_{\boldsymbol{\epsilon}}^{-1/2}\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}\boldsymbol{\Lambda}_{\boldsymbol{\epsilon}}^{-1/2}, \tag{3.7}$$

where $\boldsymbol{\Lambda}_{\boldsymbol{\epsilon}}$ is a diagonal matrix made up from the diagonal elements of $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$. The covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ can then be written in terms of its Cholesky factorisation $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = \boldsymbol{\mathcal{L}}_{\boldsymbol{\epsilon}}\boldsymbol{\mathcal{L}}_{\boldsymbol{\epsilon}}^{\top}$, where $\boldsymbol{\mathcal{L}}_{\boldsymbol{\epsilon}}$ is a lower triangular matrix. The diagonal elements of $\boldsymbol{\mathcal{L}}_{\boldsymbol{\epsilon}}$ are set to 1 so that the transformation of $\boldsymbol{R}_{\boldsymbol{\epsilon}}$ to $\boldsymbol{\mathcal{L}}_{\boldsymbol{\epsilon}}$ is one-to-one (Smith, 2013). We define an operator vechL which vectorises the strict lower triangle of a matrix by row. The unknown parameter $\mathrm{vechL}(\boldsymbol{\mathcal{L}}_{\boldsymbol{\epsilon}}) = \{L_{ij}; i = 2, \ldots, D, j < i\}$ lies in $\mathbb{R}^{D(D-1)/2}$ and is therefore unconstrained. Lindstrom and Bates (1988) also implement the Cholesky factorisation on a covariance matrix to optimise the log-likelihood function of a linear mixed effects model. Other possible reparameterisation methods for $\boldsymbol{R}_{\boldsymbol{\epsilon}}$ include using polar coordinates (Rapisarda et al., 2007) and partial autocorrelations (Daniels and Pourahmadi, 2009), but we adopt the representation in (3.7) due to its computational tractability.

By using a change of variables, we can rewrite (3.6) in terms of $\boldsymbol{\mathcal{L}}_{\boldsymbol{\epsilon}}$ as

$$\pi(\boldsymbol{\mathcal{L}}_{\boldsymbol{\epsilon}}|\boldsymbol{y}, \boldsymbol{y}^*, \boldsymbol{\alpha}_{1:N}, \boldsymbol{\theta}_{-\boldsymbol{\mathcal{L}}_{\boldsymbol{\epsilon}}}) \propto \pi(\boldsymbol{R}_{\boldsymbol{\epsilon}}|\boldsymbol{y}, \boldsymbol{y}^*, \boldsymbol{\alpha}_{1:N}, \boldsymbol{\theta}_{-\boldsymbol{R}_{\boldsymbol{\epsilon}}}) \cdot |\mathbf{J}|, \tag{3.8}$$

where $|\mathbf{J}| = |\partial \mathrm{vechL}(\boldsymbol{R}_{\boldsymbol{\epsilon}})/\partial \mathrm{vechL}(\boldsymbol{\mathcal{L}}_{\boldsymbol{\epsilon}})^{\top}|$ is the determinant of the Jacobian for the transformation. We now note that for the transformation from $\boldsymbol{R}_{\boldsymbol{\epsilon}}$ to $\boldsymbol{\mathcal{L}}_{\boldsymbol{\epsilon}}$, the prior on lower triangular Cholesky factor $\boldsymbol{\mathcal{L}}_{\boldsymbol{\epsilon}}$ whose diagonal entries are all fixed as ones, given by

$$p(\boldsymbol{\mathcal{L}}_{\boldsymbol{\epsilon}}) \propto p(\boldsymbol{R}_{\boldsymbol{\epsilon}}) \cdot |\mathbf{J}|, \tag{3.9}$$

induces a marginally uniform prior on all correlation coefficients for $\nu = D + 1$.

### 3.3.2 Sampling the Cholesky factor using HMC

HMC, popularised by Neal (2011), has enjoyed considerable recent interest within the statistical literature due to its ability to generate credible but distant candidate parameters for the MH algorithm, thereby reducing autocorrelation in the posterior samples. It does so by exploiting gradient information of the log posterior density to simulate a trajectory according to physical dynamics.

HMC has an intuitive physical interpretation whereby the state of the system consists of the position of the variable of interest $\boldsymbol{\theta}$ and its momentum $\boldsymbol{u}$, which is assumed to follow a $\mathcal{N}(\boldsymbol{0}, \boldsymbol{M})$ pseudo-prior with mass matrix $\boldsymbol{M}$. The momentum of $\boldsymbol{\theta}$ allows it to move over surfaces of varying gradients in the physical system, and hence changing the position of $\boldsymbol{\theta}$. As a result of the introduction of the auxiliary momentum variable $\boldsymbol{u}$, the HMC targets the augmented distribution

$$\pi(\boldsymbol{\theta}, \boldsymbol{u}) \propto \exp(-\mathcal{H}(\boldsymbol{\theta}, \boldsymbol{u})),$$

where $\mathcal{H}(\boldsymbol{\theta}, \boldsymbol{u}) = -\log \pi(\boldsymbol{\theta}) + \frac{1}{2}\boldsymbol{u}^\top \boldsymbol{M}^{-1}\boldsymbol{u}$ is termed the Hamiltonian. As shown by Neal (2011), the invariant distribution of the Markov chain generated from the HMC algorithm is $\pi(\boldsymbol{\theta}, \boldsymbol{u})$ and samples from $\pi(\boldsymbol{\theta})$ can be obtained by marginalising out the momentum $\boldsymbol{u}$.

In order to implement the HMC algorithm as described in Section 2.3.3, computation of the derivatives of (3.8) with respect to the $L_{ij}$ is required for the leapfrog update. Lemma 1 derives the expressions for these gradients for the marginally uniform prior in (2.4) based on the Cholesky reparameterisation in (3.7).

**Lemma 1.** *Let $\boldsymbol{E}_d$ denote the matrix obtained by removing column $d$ from an identity matrix $\boldsymbol{I}$. For the parameterisation of $\boldsymbol{R}_\epsilon$ in (3.7),*

(i) $$\frac{\partial \boldsymbol{R}_\epsilon^{-1}}{\partial L_{ij}} = -\boldsymbol{\Lambda}_\epsilon^{1/2}\left(\boldsymbol{\Sigma}_\epsilon^{-1}\frac{\partial \boldsymbol{\Sigma}_\epsilon}{\partial L_{ij}}\boldsymbol{\Sigma}_\epsilon^{-1} + \frac{\partial \boldsymbol{\Lambda}_\epsilon^{-1/2}}{\partial L_{ij}}\boldsymbol{\Lambda}_\epsilon^{1/2}\boldsymbol{\Sigma}_\epsilon^{-1} + \boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{\Lambda}_\epsilon^{1/2}\frac{\partial \boldsymbol{\Lambda}_\epsilon^{-1/2}}{\partial L_{ij}}\right)\boldsymbol{\Lambda}_\epsilon^{1/2}.$$

(ii) $$\frac{\partial \log |\boldsymbol{R}_\epsilon(-d;-d)|}{\partial L_{ij}} = tr\left(\boldsymbol{R}_\epsilon^{-1}(-d;-d)\boldsymbol{E}_d^\top \frac{\partial \boldsymbol{R}_\epsilon}{\partial L_{ij}}\boldsymbol{E}_d\right).$$

(iii) $\dfrac{\partial \log |\boldsymbol{R}_{\boldsymbol{\epsilon}}|}{\partial L_{ij}} = -\dfrac{2L_{ij}}{\sum_{k=1}^{i} L_{ik}^2}.$

*Proof.*

(i) Using the chain rule and the reparameterisation in (3.7), we obtain

$$\frac{\partial \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}}{\partial L_{ij}} = \frac{\partial \boldsymbol{\Lambda}_{\boldsymbol{\epsilon}}^{-1/2}}{\partial L_{ij}} \boldsymbol{\Lambda}_{\boldsymbol{\epsilon}}^{1/2} \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} + \boldsymbol{\Lambda}_{\boldsymbol{\epsilon}}^{-1/2} \frac{\partial \boldsymbol{R}_{\boldsymbol{\epsilon}}^{-1}}{\partial L_{ij}} \boldsymbol{\Lambda}_{\boldsymbol{\epsilon}}^{-1/2} + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} \boldsymbol{\Lambda}_{\boldsymbol{\epsilon}}^{1/2} \frac{\partial \boldsymbol{\Lambda}_{\boldsymbol{\epsilon}}^{1/2}}{\partial L_{ij}}$$

and the result follows from Theorem 3 in Chapter 8 of Magnus and Neudecker (1999).

(ii) We express $\boldsymbol{R}_{\boldsymbol{\epsilon}}(-d;-d)$ as $\boldsymbol{E}_d^\top \boldsymbol{R}_{\boldsymbol{\epsilon}} \boldsymbol{E}_d$ and the result follows from Theorem 2 in Chapter 8 of Magnus and Neudecker (1999).

(iii) We note that $|\boldsymbol{R}_{\boldsymbol{\epsilon}}| = |\boldsymbol{\Lambda}_{\boldsymbol{\epsilon}}|^{-1}$ since $|\boldsymbol{\Sigma}_{\epsilon}| = 1$ from the reparameterisation in (3.7) and that $|\boldsymbol{\Lambda}_{\boldsymbol{\epsilon}}| = \prod_{i=1}^{D} \sum_{k=1}^{i} L_{ik}^2$. The result is straightforward from simple calculus.

□

## 3.4 A deterministic proposal distribution

Various strategies have been proposed to reduce the variability in the Monte Carlo estimate of the expectation $\mathbb{E}_\pi[f(\boldsymbol{\theta})]$ of a scalar function $f$ of parameter $\boldsymbol{\theta}$ with respect to some posterior distribution $\pi(\boldsymbol{\theta})$, including the Rao-Blackwellisation (Robert and Casella, 2004) and the control variates (Dellaportas and Kontoyiannis, 2012; Oates et al., 2017). These techniques produce an efficient estimator of $\mathbb{E}_\pi[f(\boldsymbol{\theta})]$ based on sampled $\boldsymbol{\theta}$ generated from an MCMC sampler.

Here, we focus on a particular class of methods which integrate variance reduction techniques dynamically within an MCMC sampling algorithm. Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_P)^\top$ be a parameter vector that has $P$ univariate parameter with normal full conditional distributions $\theta_p|\boldsymbol{\theta}_{-p} \sim \mathcal{N}(\mu_p, \sigma_p^2)$, where the conditional mean $\mu_p$ and the conditional variance $\sigma_p^2$ may depend on $\boldsymbol{\theta}_{-p} = \{\theta_q : q = 1, \ldots, P, q \neq p\}$. Adler (1981) and

Barone and Frigessi (1990) introduce an over-relaxation method where the update on $\boldsymbol{\theta}$ is performed using Gibbs sampling, and where the new value $\theta_p'$ for each margin of $\boldsymbol{\theta}$ is generated as

$$\theta_p' = (1+\kappa)\mu_p - \kappa\theta_p + \delta\sigma_p\sqrt{1-\kappa^2}, \quad p = 1, \dots, P, \tag{3.10}$$

with $\delta \sim \mathcal{N}(0,1)$ being a standard normal random variable. Equation (3.10) allows for the introduction of dependence between successive samples via the constant antithetic parameter $\kappa$, which is required to be in the open interval $(-1,1)$ so that the Markov chain is ergodic and produces $\pi(\boldsymbol{\theta})$ as its invariant distribution. This scheme is exactly the conventional Gibbs sampler when $\kappa = 0$. Variance reduction in estimating $\mathbb{E}_\pi[f(\boldsymbol{\theta})]$ is achieved through the antithetic variable method (Hammersley and Morton, 1956) by setting $\kappa > 0$ so that the estimation bias in the previous sample is corrected in the opposite direction. The rate of convergence for the over-relaxation method in (3.10) is studied in Barone and Frigessi (1990), while Green and Han (1992) establish that the asymptotic variance of the estimator for $\mathbb{E}_\pi[f(\boldsymbol{\theta})]$ using this strategy for linear $f$ is proportional to $(1-\kappa)/(1+\kappa)$.

Motivated by the over-relaxation sampler (Adler, 1981; Barone and Frigessi, 1990) discussed previously and noting that the IACT can be less than 1 if some of the autocorrelations are negative, in which case a Monte Carlo estimator constructed is super-efficient in terms of the reduction in variance of the Monte Carlo estimates compared to the same estimates constructed from uncorrelated samples, we introduce into the MH literature a deterministic design of the proposal distribution for $\boldsymbol{\theta}$

$$q(\boldsymbol{\theta}'|\boldsymbol{\theta}) = \delta_{\psi(\boldsymbol{\theta})}(\boldsymbol{\theta}'),$$

where $\psi$ is a mapping function which introduces negative correlation between samples and $\delta_{\psi(\boldsymbol{\theta})}$ is the Dirac delta function at $\psi(\boldsymbol{\theta})$ so that $\theta' = \psi(\boldsymbol{\theta})$. In this case, the MH acceptance probability reduces to the ratio of $\pi(\boldsymbol{\theta})$ evaluated at $\boldsymbol{\theta}'$ and $\boldsymbol{\theta}$.

When $\pi(\boldsymbol{\theta})$ is a normal distribution, we propose setting

$$\psi(\boldsymbol{\theta}) = 2\boldsymbol{\mu}_{\boldsymbol{\theta}} - \boldsymbol{\theta}, \tag{3.11}$$

where $\boldsymbol{\mu}_{\boldsymbol{\theta}}$ is the mean of $\pi(\boldsymbol{\theta})$. It is clear that (3.11) represents an example of the antithetic variable with perfect negative correlation, and also an instance of the over-relaxation method in (3.10) with $\kappa = 1$, which is outside the range of values for which the Markov chain is ergodic. Symmetry of the normal density gives $\pi(\boldsymbol{\theta}') = \pi(\boldsymbol{\theta})$, which in turn translates to an acceptance probability of one. Clearly, our proposed antithetic sampling will only yield an ergodic Markov chain when it is coupled with stochastic simulation of additional parameters that affect the value of the deterministic proposal $\psi(\boldsymbol{\theta})$, in particular $\boldsymbol{\mu}_{\boldsymbol{\theta}}$. Under this condition, the value of $\boldsymbol{\mu}_{\boldsymbol{\theta}}$ changes in every iteration of the update and this drives the exploration of $\boldsymbol{\theta}$ in the parameter space. Furthermore, the dependence between $\boldsymbol{\theta}$ and other model parameters prevents exact periodicity from occurring, and thus the Markov chain is aperiodic.

The conditional posterior distribution of the random effects $\boldsymbol{\alpha}_{1:N}$ in our MVP model is normal and likewise for the regression parameters $\boldsymbol{\beta} = \text{vec}(\boldsymbol{B})$ when using a conjugate normal prior. Therefore, we can employ the antithetic sampling method in (3.11) to improve the IACTs of both the random effects $\boldsymbol{\alpha}_{1:N}$ and regression coefficients $\boldsymbol{\beta}$. In fact, antithetic sampling of normal random variables can also be understood in terms of a HMC update. Suppose that $\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$, and the prior on the momentum variable $\boldsymbol{u}$ is chosen as $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1})$. Pakman and Paninski (2014) show that the resulting Hamiltonian system can be solved analytically, with solution given by

$$\boldsymbol{\theta}(t) = \boldsymbol{\mu}_{\boldsymbol{\theta}} + \boldsymbol{\Sigma}_{\boldsymbol{\theta}}\boldsymbol{u}(0)\sin(t) + (\boldsymbol{\theta}(0) - \boldsymbol{\mu}_{\boldsymbol{\theta}})\cos(t), \tag{3.12}$$

which is a linear combination of $\boldsymbol{\mu}_{\boldsymbol{\theta}}$, the initial value $\boldsymbol{\theta}(0)$ of $\boldsymbol{\theta}$ and the initial momentum $\boldsymbol{u}(0)$. Note that (3.12) is a multivariate generalisation of (3.10) with $t = \cos^{-1}(-\kappa)$. Equation (3.12) is thus equivalent to the antithetic sampler in (3.11) when setting $t = \pi$ radians. Since there is no approximation error in the

Hamiltonian dynamics for a normal distribution, an MH accept-reject step is not required in the HMC sampler, and the proposed value of $\boldsymbol{\theta}$ will always be accepted. This equivalence relation was first observed by Pakman and Paninski (2014), but was not particularly useful in their framework of sampling from a truncated multivariate normal distribution. Our proposal for antithetic sampling is different from theirs in the sense that it is entirely deterministic, and we choose $t = \pi$ radians to induce a perfect negative proposal correlation. Pakman and Paninski (2014), on the other hand, suggest setting $t = \frac{\pi}{2}$ radians, which is equivalent to drawing a fresh sample from a random number generator when it is applied to the setting of a normal distribution. We refer to this approach as the independent sampler hereafter.

So far, our discussion has mainly focused on normal $\pi(\boldsymbol{\theta})$. This is because an analytic solution to the Hamiltonian system is only available for a normal distribution. It is possible to extend the proposed antithetic sampler to more general distributions by obtaining an approximation of $\boldsymbol{\mu_\theta}$ in order to propose a new value of $\boldsymbol{\theta}$, and then accept or reject the proposal in an MH algorithm to target the true $\pi(\boldsymbol{\theta})$, as suggested in Green and Han (1992). However, the application of this generalisation and its variants (e.g. Creutz (1987)) is somewhat limited due to high rejection rates in the accept-reject step (Neal, 1998). In this case, we consider the HMC algorithm as it provides a way to overcome this shortcoming.

## 3.5 Simulation studies

We now study the efficiency of the antithetic variable technique described in Section 3.4. Two examples are presented. The first examines the antithetic sampler in a more general setting, while the second is specific to the application in Section 3.6. Reported IACT values of the parameters are computed using the `coda` package version 0.19–3 (Plummer et al., 2006) in `R` version 3.6.2 (R Core Team, 2019).

**Example 1.** The invariant distribution $\pi(\boldsymbol{\theta})$ is specified as a bivariate normal distribution with high correlation (0.99) between the variables. For such a strong dependence between the variables, the Gibbs sampler is known to converge slowly to $\pi(\boldsymbol{\theta})$ (Gilks et al., 1994). We investigate the performance of three sampling schemes – the independent sampler, the over-relaxation algorithm with $\kappa = 0.9$, and a coupling of the over-relaxation algorithm (on the first margin) with the antithetic sampler (on the second margin). Note that this coupling strategy introduces stochasticity into the antithetic sampler, which is essential to produce an ergodic Markov chain. The three samplers are each run for 10 000 iterations from the same initialised value $(2, 2)$, and the update on each margin is performed conditional on the other. Figure 3.1 illustrates the trajectories of the first 50 samples generated for each of the three samplers. Exploration of the target space is reduced to a random walk under the independent sampler (Figure 3.1 (left)). In contrast, the other two samplers move between different contours of the density and explore the full support of the distribution in an elliptical manner, thereby reducing the IACT significantly (Figure 3.1 (middle)). The IACT decreases further (Figure 3.1 (right)) when the over-relaxation algorithm on the second margin is replaced by antithetic sampling. In this analysis, the mixing of both margins is improved by a factor of 1.75.

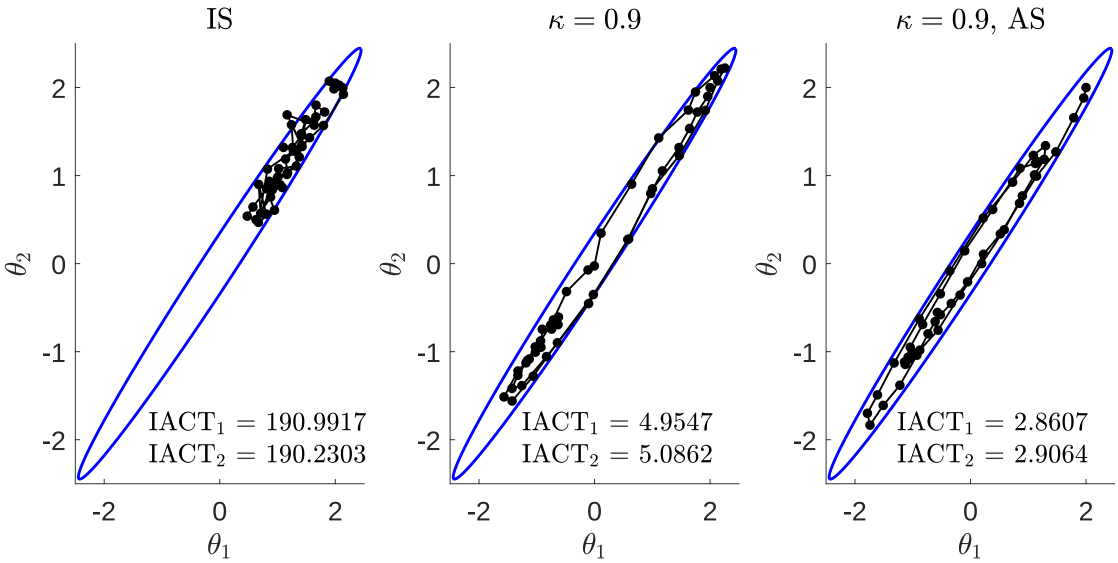



Figure 3.1: Trajectories of the first 50 samples generated from the independent sampler (left), the over-relaxation algorithm with $\kappa = 0.9$ (middle), and the over-relaxation algorithm coupled with the antithetic sampler (right). The blue solid lines represent the 95% confidence region of the bivariate normal distribution.

**Example 2.** A simulated dataset is generated following the MVP model given in (3.1)–(3.4), with $D = 8, N = 162, T = 16$ and values of the parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{R}_{\boldsymbol{\epsilon}}, \boldsymbol{\Sigma}_{\boldsymbol{\alpha}})$ set to be the posterior mean estimates of the parameters in Model 1 of the female contraceptive product analysis of Section 3.6. To avoid hand-tuning the stepsize $\varepsilon$ and the trajectory length $\mathcal{T}$ for the HMC update of $\boldsymbol{\mathcal{L}}_{\boldsymbol{\epsilon}}$, we utilise the No-U-Turn Sampler (NUTS) with the dual averaging scheme of Hoffman and Gelman (2014). We use the following weakly informative prior distributions: $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, 100\boldsymbol{I})$, $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}} \sim \mathcal{IW}(9, \boldsymbol{I})$ and the prior distribution on the lower triangular Cholesky factor $\boldsymbol{\mathcal{L}}_{\boldsymbol{\epsilon}}$ given in (3.9). The sampling scheme is run for 30 000 iterations, with the first 5 000 samples discarded as burn-in. Appendix 3.8.1 details the Gibbs sampling scheme.

Figure 3.2 compares graphically the marginal posterior densities and sample autocorrelations of randomly sampled random effects $\boldsymbol{\alpha}_{1:N}$ and the regression parameter $\boldsymbol{\beta}$ between independent and antithetic sampling. Despite the absence of

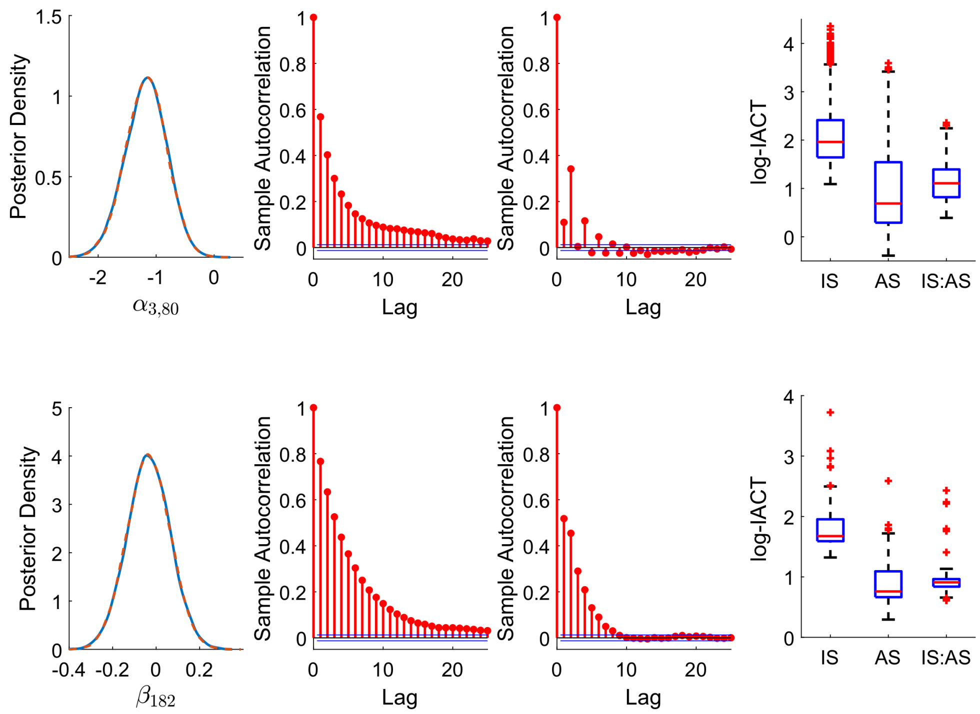



Figure 3.2: Marginal posterior densities of a randomly selected random effects term $\alpha_{3,80}$ (top panel) and regression coefficient $\beta_{182}$ (bottom panel), and their sample autocorrelation plots under independent sampling (IS) and antithetic sampling (AS). Rightmost column gives the distributions of the log IACT values and the element-wise log IACT ratios of IS to AS for all random effects $\boldsymbol{\alpha}_{1:N}$ (1 296 parameters) and regression coefficients $\boldsymbol{\beta}$ (216 parameters).

a stochastic component in the updates of $\boldsymbol{\alpha}_{1:N}$ and $\boldsymbol{\beta}$, the kernel density plots of these parameters indicate that the coupling of a stochastic MCMC scheme for the remaining parameters with the antithetic variable technique gives the same posterior distributions as those under independent sampling. The autocorrelation plots show that the samples generated from antithetic sampling have positive dependence that has a higher rate of decay over the number of lags, thereby reflecting the superior mixing of the Markov chain. The IACT values of the randomly sampled parameters are significantly lower, with improvement factors of 3.72 and 2.10 observed for $\alpha_{3,80}$ and $\beta_{182}$ respectively. The box plot showing the distribution of the IACT values across all $\boldsymbol{\alpha}_{1:N}$ also indicates that some of these parameters are super-efficient (i.e. IACT less than 1). Furthermore, the log IACT ratios of the independent sampler compared to the antithetic sampler are well above 0, suggesting that all $\boldsymbol{\alpha}_{1:N}$ and $\boldsymbol{\beta}$ parameters experience efficiency gains (see rightmost boxplots in Figure 3.2. Although perfect negative correlation is induced between successive samples by the deterministic proposal, this does not necessarily translate to an equivalent autocorrelation in the posterior samples. Rather, the negative relationship is used to reduce the magnitude of positive autocorrelation present in the MCMC samples. Note that convergence to the posterior distribution might be slow for poorly initialised values under antithetic sampling so we suggest using independent sampling during the burn-in period and later switching to the deterministic proposal.

The remaining simulation experiments investigate the performance of the MVP model in the context of recovering the true parameters of the data generating process under different specifications of prior distribution on $\boldsymbol{\theta}$. We use the posterior root-mean-square error (RMSE) defined by

$$\mathrm{RMSE}(\theta_p) = \sqrt{\frac{1}{n}\sum_{j=1}^{n}({\theta_p}^{[j]} - \theta_{p,\mathrm{true}})^2}, \tag{3.13}$$

as the performance measure, where ${\theta_p}^{[j]}$ is the $j$-th iterate from the $n$ posterior samples and $\theta_{p,\mathrm{true}}$ is the true value of $\theta_p$. The measure in (3.13) is defined for univariate $\theta_p$. For a multivariate $\boldsymbol{\theta}$, the posterior RMSE is calculated for each margin of $\boldsymbol{\theta}$. All the results shown are based on 1 000 different replicate sets of simulated data with the same true parameter values. The use of RMSE is valid here as a measure of fit since weakly informative priors are used.

We first consider the conditionally conjugate hierarchical inverse-Wishart $\mathcal{HIW}(\nu_0, \boldsymbol{A})$ prior of Huang and Wand (2013) with degrees of freedom $\nu_0$ and positive scale parameter $\boldsymbol{A} = (A_1, \ldots, A_D)^\top$ as an alternative to the inverse-Wishart prior on the $D \times D$ covariance matrix $\boldsymbol{\Sigma_\alpha}$,

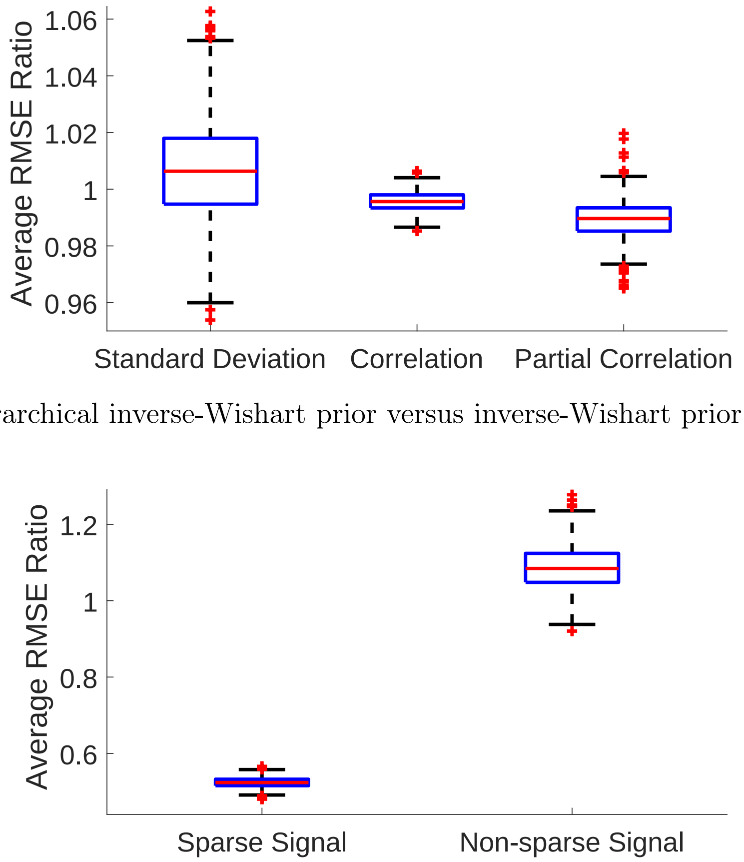
$$\boldsymbol{\Sigma_\alpha} | a_1, \ldots, a_D \sim \mathcal{IW}\left(\nu_0 + D - 1, 2\nu_0 \operatorname{diag}\left(\frac{1}{a_1}, \ldots, \frac{1}{a_D}\right)\right),$$
$$a_d \overset{iid}{\sim} \mathcal{IG}(0.5, A_d^{-2}), \quad d = 1, \ldots, D,$$

where $\mathcal{IG}(a, b)$ is an inverse-Gamma distribution with shape $a$ and scale $b$. The marginal prior of the standard deviation in $\boldsymbol{\Sigma_\alpha}$ is a half-$t(\nu_0, A_d)$ distribution, as suggested in Gelman (2006). In the simulation, we select $\nu_0 = 2$ and choose a weakly informative scale parameter whereby $A_1 = A_2 = 0.23$ and $A_3 = \cdots = A_8 = 0.46$ so that approximately 95% of the half-$t$ density is below 1 and 2 respectively. This specification is relevant to the real data application in Section 3.6, where our prior belief is that the variability in the tendency of GPs to discuss pill contraceptives is lower compared to non-pill alternatives. In contrast, the inverse-Wishart prior assumes the same variability for all variance parameters $\sigma^2_{\alpha_i}$ in $\boldsymbol{\Sigma_\alpha}$.

Figure 3.3a shows the distribution of the average RMSE ratio of each type of parameter in $\boldsymbol{\Sigma_\alpha}$, i.e.

$$\frac{1}{P} \sum_{p=1}^{P} \frac{\text{RMSE}_{\mathcal{HIW}}(\theta_p)}{\text{RMSE}_{\mathcal{IW}}(\theta_p)}$$

based on 1 000 replicate simulations, for the hierarchical inverse-Wishart prior versus the inverse-Wishart prior. Although the hierarchical inverse-Wishart prior is flexible enough to specify different strengths of prior on each $\sigma^2_{\alpha_i}$, Figure 3.3a shows that in this case its performance is similar to the more restrictive inverse-Wishart prior. This result is somewhat unsurprising considering that the estimated $\sigma^2_{\alpha_i}$ in the application




(a) Hierarchical inverse-Wishart prior versus inverse-Wishart prior on $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}$.



(b) Horseshoe shrinkage prior versus normal prior on $\boldsymbol{\beta}$.

Figure 3.3: Distributions of the average posterior RMSE ratio of all parameters in (a) $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}$ or (b) $\boldsymbol{\beta}$, based on 1 000 replicate analyses, under different prior choices. (a) Standard deviations, correlations and partial correlations for parameters in $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}$ for the hierarchical inverse-Wishart prior versus the inverse-Wishart prior on $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}$. (b) Sparse regression coefficients $\beta_k = 0$ and non-sparse coefficients $\beta_k \neq 0$ for the horseshoe prior versus the $\mathcal{N}(\mathbf{0}, 100\boldsymbol{I})$ prior on $\boldsymbol{\beta}$.

example are more or less similar across the different contraceptive products (see Appendix 3.8.5). The distributions for the posterior RMSE ratio of the correlation coefficients and the partial correlations are concentrated around 1 since both the hierarchical inverse-Wishart prior with $\nu_0 = 2$ and the inverse-Wishart prior with $D + 1$ degrees of freedom and scale matrix $\boldsymbol{I}$ induce the same marginally uniform prior, i.e. (2.4) with $\nu = D + 1$, on the resulting correlation matrix $\boldsymbol{R}_{\boldsymbol{\alpha}}$, which in turn gives the same implied LKJ distribution on the partial correlations.

To identify sparse signals (coefficients which are significant) in the regression parameter $\boldsymbol{\beta}$, we employ the horseshoe shrinkage prior (Carvalho et al., 2010) given by

$$\beta_k|\lambda_k, \tau \sim \mathcal{N}(0, \tau^2\lambda_k^2), \quad \lambda_k \sim \mathcal{C}^+(0,1), \quad \tau \sim \mathcal{C}^+(0,1),$$

where $\mathcal{C}^+(0,1)$ is a half-Cauchy distribution with location 0 and scale 1 restricted to positive support. The simulation is carried out by setting 75% of the smallest non-intercept posterior mean regression coefficients (in absolute value) in $\boldsymbol{\beta}$ to 0, from which we generate the simulated datasets. We model the prior on each intercept separately by a weakly informative $\mathcal{N}(0, 100)$ distribution to avoid heavily penalising these parameters. Gibbs sampling from the posterior distribution of $\boldsymbol{\beta}$ is implemented by adopting the latent variable formulation in Makalic and Schmidt (2016). Figure 3.3b displays the results of comparing this prior specification for $\boldsymbol{\beta}$ to a $\mathcal{N}(\mathbf{0}, 100\boldsymbol{I})$ prior, again in terms of the average RMSE ratio over all regression parameters. The horseshoe prior performs as well as the $\mathcal{N}(\mathbf{0}, 100\boldsymbol{I})$ prior on non-zero entries of $\boldsymbol{\beta}$, although the variability in the RMSE ratio is large. On the other hand, the horseshoe prior outperforms the normal prior for those parameters whose true values are zero, reducing the RMSE by half. This occurs as the horseshoe prior places a greater density around zero, which results in a more concentrated posterior distribution for parameters which are truly zero. Therefore, it is an attractive default option when we expect sparsity in the regression parameters, as is the case for our analysis of the characteristics affecting the decision-making behaviour of GPs in the next section.

## 3.6 Application to female contraceptive products by Australian GPs

### 3.6.1 Background and aims of study

In order to study the decision-making behaviour of Australian GPs, we obtain data from Fiebig et al. (2017) who design a stated preference experiment in which GPs are asked to select the contraceptive products that they would consider discussing with hypothetical female patients. The GPs evaluate a sequence of vignettes where patients are defined in terms of socio-economic and clinical characteristics that are varied as part of the experimental design to cover a range of different life cycle and fertility stages. Table 3.3 in Appendix 3.8.2 contains the attributes of the patients with a description for each level of the categorical variables. The GPs choose from a set of 9 products that they would discuss with the patient before deciding upon their most preferred product to be subsequently prescribed to the patient. A sample of 162 GPs from a list of 14 816 GPs from all states and territories of Australia volunteered to participate in the experiment between December 2008 and June 2009 where each subject makes choices for 16 different patients, resulting in 2 592 observations. The following covariate information is collected on the GPs themselves: age, gender, whether they are registered as a Fellow of the Royal Australian College of GPs, whether they have a certificate in family planning, whether they are an Australian medical graduate, whether their location of practice is in an urban area and whether they bulk-bill patients. Analysis of this panel data is based on the set of binary outcomes as to whether or not to discuss each of the contraceptive products. Due to low occurrences for the prescription of the hormonal patch which was yet to be released in the Australian market, we removed this product from the dataset leaving observations on the 8 remaining products.

The experiment is designed to mimic the choice problem faced by GPs in a consultation where they need to match a product with a particular patient. In characterising such a decision problem, Frank and Zeckhauser (2007) distinguish between “custom-made” and “ready-to-wear” (or norm-based) choices. A custom-made choice involves the GP undertaking a careful evaluation of the patient and then matching her to an appropriate product. However, as new products are introduced, GPs face considerable costs in the process of gaining the knowledge and expertise required to discuss and prescribe these products (Hauser and Wernerfelt, 1990; Roberts and Lattin, 1991; Gilbride and Allenby, 2004). This is particularly the case when more familiar products are available even though they may be somewhat inferior to the new products (Wellings et al., 2007); an especially salient situation in the market for contraceptive products. In such cases, some GPs will tend to adopt norms (here particular products) that work well for a broad class of patients and to place less weight on certain patient attributes that would indicate a different product that is potentially a better match.

Particular interest is in the dependence in recommendations among the products. That is, which products tend to be discussed together and which tend to form distinct clusters. If GPs pursue custom-made strategies, then a considerable portion of the dependence between products will be explained by the attributes of the patient. Conditional on the observable features of the patient and characteristics of the GPs, remaining dependencies will reflect the relationship between unobservables related to evaluations of the suitability of certain products for a particular patient, and how for individual GP's their product effects are correlated across products. The proposed model is designed to capture these forms of heterogeneity and will permit a detailed analysis of the choices.

The prevalence of ready-to-wear choices is one possible explanation for the relatively low uptake of long acting reversible contraceptive (LARC) methods in Australia (Black et al., 2013). LARC methods are contraceptives that are administered less frequently than monthly and include hormonal implants, intrauterine contraception (IUC), both hormonal and copper-bearing, and contraceptive injections. There is increasing support for the greater use of these more effective methods to reduce unintended pregnancies and abortion rates (Stevens-Simon et al., 2001; Blumenthal et al., 2011; Secura, 2013). In our analysis below, we will use the model to explore a case where there is no clinical reason why at least one of these LARC methods should not be considered for discussion by GPs. For ease of presentation, we will use the subscripts in Table 3.1 to denote the products.

| Subscript | Product |
|---|---|
| 1 | Combined pill |
| 2 | Mini-pill |
| 3 | Hormonal injection |
| 4 | Hormonal implant |
| 5 | Hormonal IUD |
| 6 | Vaginal ring |
| 7 | Copper IUD |
| 8 | Condom |

Table 3.1: Correspondence of parameter subscripts to each female contraceptive product. Long acting reversible contraceptive methods are shown in grey.

### 3.6.2 Analysis and results

We consider two different models for the data:

$$\text{Model 1: } \boldsymbol{y}_{it}^* = \boldsymbol{\alpha}_i + \boldsymbol{B}\boldsymbol{x}_{it} + \boldsymbol{\epsilon}_{it}, \tag{3.14}$$

$$\text{Model 2: } \boldsymbol{y}_{it}^* = \boldsymbol{\alpha}_i + \boldsymbol{B}\boldsymbol{x}_{it} + \boldsymbol{C}\boldsymbol{z}_i + \boldsymbol{\epsilon}_{it}, \tag{3.15}$$

for $i = 1, \ldots, N = 162$ GPs and $t = 1, \ldots, T = 16$ patients, where $\boldsymbol{y}^*$ is the latent normal random variables underlying the observed binary outcomes $\boldsymbol{y}$ described in (3.4). Here $\boldsymbol{\alpha}_i$ and $\boldsymbol{C}\boldsymbol{z}_i$ respectively represent GP-specific random and fixed effects with $\boldsymbol{z}_i$ being a vector of GP characteristics, and $\boldsymbol{B}\boldsymbol{x}_{it}$ represents fixed effects of the patient. We select a horseshoe prior on $\boldsymbol{\beta} = \text{vec}(\boldsymbol{B})$ and model the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}$ of the random effects by the $\mathcal{HIW}(2, \boldsymbol{A})$ prior in Section 3.5 where $\boldsymbol{A} = (0.23, 0.23, 0.46, \ldots, 0.46)^\top$. The scale is chosen to express the prior information that the variances of the random effects are expected to be small, with those for the pill products being less variable compared to the non-pill alternatives. The difference between these two models is the presence of the GP-specific fixed effects in Model 2, which explain some of the relationships in the random effects of Model 1. Let $\boldsymbol{X} = (X_1, \ldots, X_K)^\top$ be a vector of normal random variables with covariance matrix given by $\boldsymbol{\Sigma}_{\boldsymbol{X}}$. Recall that $X_i$ and $X_j$ are conditionally independent given the other random variables if the $(i, j)$-th entry of the precision matrix $\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}$ is zero.

Figures 3.4 and 3.5 give graphical summaries of the posterior distribution of the dependence structures of the latent variable $\boldsymbol{y}_{it}^*$ conditional on $\boldsymbol{\alpha}_i$ and $\boldsymbol{x}_{it}$ (as well as $\boldsymbol{z}_i$ for Model 2), and the random effects $\boldsymbol{\alpha}_i$ respectively. All graphs are obtained by computing the 95% credible interval of the posterior distribution for each entry of $\boldsymbol{R}_{\boldsymbol{\epsilon}}^{-1}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}^{-1}$, where an edge is formed between two nodes if the credible interval does not include 0. The absence of an edge between any two nodes indicates a plausible conditional independence between the two variables given the rest. The dependence structures associated with the latent variables are the same for both Model 1 and 2. This supports the use of the MVP model in order to capture the complex dependencies between different products that would otherwise be ignored in separate univariate analyses on each product.

Figure 3.4 is also instrumental in explaining the suitability of the contraceptive products for a patient in terms of substitute goods, which are products with similar functions that can be used in place of each other. For conciseness, we only focus on some important relationships illustrated in the graphical model. The propensity to discuss pill products $(y_1^*, y_2^*)$ is independent of each other given whether the hormonal IUD and the vaginal ring $(y_5^*, y_6^*)$ are discussed, by the Markov property since all paths from $y_1^*$ to $y_2^*$ pass through $(y_5^*, y_6^*)$, reflecting the use of these non-pill contraceptives as pill alternatives dictated by particular clinical conditions. The clique formed between $(y_5^*, y_7^*, y_8^*)$ suggests dependence in the propensity to discuss the hormonal IUD, the copper IUD and condoms. In fact, the posterior correlation between the propensity scores for both the IUD methods $(y_5^*, y_7^*)$ is around 0.52 on average (see Appendix 3.8.4), suggesting a high tendency for these products to be discussed together. This also suggests that these IUD methods are substitutes. Noticeably, the propensity to discuss the hormonal injection and the hormonal implant $(y_4^*, y_5^*)$ exhibit the highest level of association, as indicated by our model, with a mean posterior correlation of 0.59. This indicates the likelihood of these two prominent LARC products being included together in discussions, and it is consistent with them being moderately close substitutes for many patients.

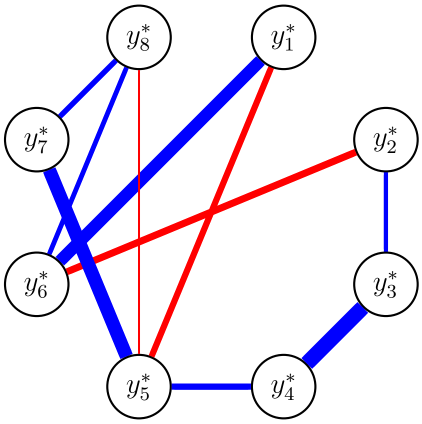
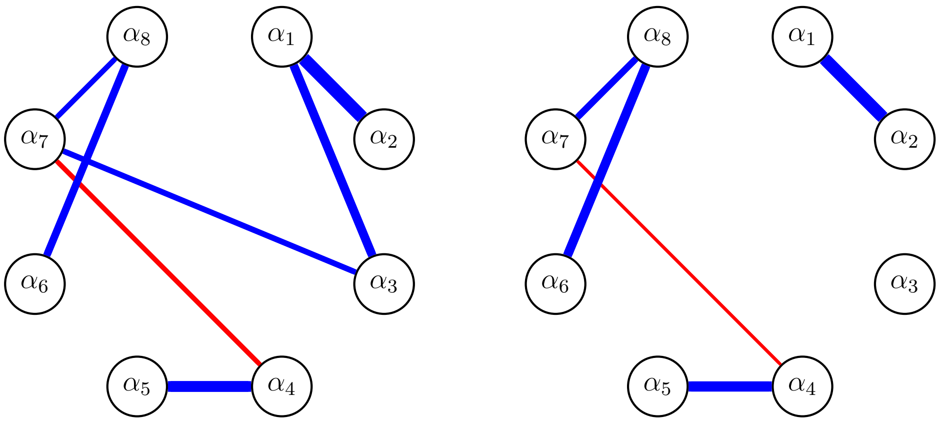



Figure 3.4: Graphical model illustrating estimated dependence structure of the latent variables $\boldsymbol{y}^*$ conditional on the random effects and the covariates in both Model 1 and 2. Edges between $y_i^*$ and $y_j^*$ are included if the 95% credible interval of the marginal posterior distribution of the $(i, j)$-th entry of $\boldsymbol{R}_{\boldsymbol{\epsilon}}^{-1}$ does not contain 0. Blue edges represent positive dependence while red edges represent negative dependence. The thickness of the edges is proportional to the strength of the dependence.




(a) Model 1. (b) Model 2.

Figure 3.5: Graphical models illustrating estimated dependence structure of the GP-specific random effects $\boldsymbol{\alpha}$ in each model. Edges between $\alpha_i$ and $\alpha_j$ are included if the 95% credible interval of the marginal posterior distribution of the $(i, j)$-th entry of $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}^{-1}$ does not contain 0. Blue edges represent positive dependence while red edges represent negative dependence. The thickness of the edges is proportional to the strength of the dependence.

Figure 3.5 can be interpreted in the same way as Figure 3.4, regarding the substitution of different products but in the context of ready-to-wear choices. This is because the random effects in (3.14) characterise the consistency of GPs in discussing a particular product after observing the patient's attributes. There are clear differences in the graphical structure when comparing Models 1 and 2 (Figures 3.5a and 3.5b). The changes in the dependence structure of the GP random effects arise because some of the consistency in product choices can be explained by GP characteristics. For example, the tendency of GPs to include both the hormonal injection and the copper IUD ($\alpha_3, \alpha_7$) as ready-to-wear choices is due to their age (see significance of GP characteristics in Appendix 3.8.3). The posterior structure also provides some confidence that the random effects specification is useful in capturing important GP characteristics that are not directly observed. Three clusters of products with substantial dependence in ready-to-wear choices are identified from the model after accounting for the observed GP characteristics. Particularly relevant is the dependence between the hormonal IUD and the implant ($\alpha_4, \alpha_5$). There is positive correlation between these two LARCs, indicating the tendency for GP attitudes (either positive or negative) to be aligned. A second cluster includes both of the pills ($\alpha_1, \alpha_2$) which is consistent with these products being used as a ready-to-wear choice. GPs who are more likely to discuss the combined pill after conditioning on the patient's attributes behave similarly when considering the mini-pill. Contraceptives that are not pill- nor hormone-based form the final bundle.

Our models allow us to examine posterior predictions for a range of patients. Since we are interested in the uptake of LARC products, we specify a particular female patient where there is no clinical reason why a LARC should not be considered for discussion. Table 3.3 of Appendix 3.8.2 gives the attributes of this base-case patient. Figure 3.6 summarises the estimate of the predictive probability of a GP discussing a particular product, where the range of predictions shown is generated for all GPs in the sample based on Model 2. For this particular base-case patient, there is considerable agreement amongst all GPs in the sample that the combined pill (product 1) is one of the most suitable products to be discussed, but they have much more variable views on the other products. Amongst the LARCs (products 3, 4, 5 and 7), the hormonal injection (product 3) and the implant (product 4) are the products which are the most likely to be discussed, with the variability across GPs perhaps simply reflecting a view that they are good substitutes to each other, which is in fact what we find in Figure 3.4. GPs could indeed have consistent views

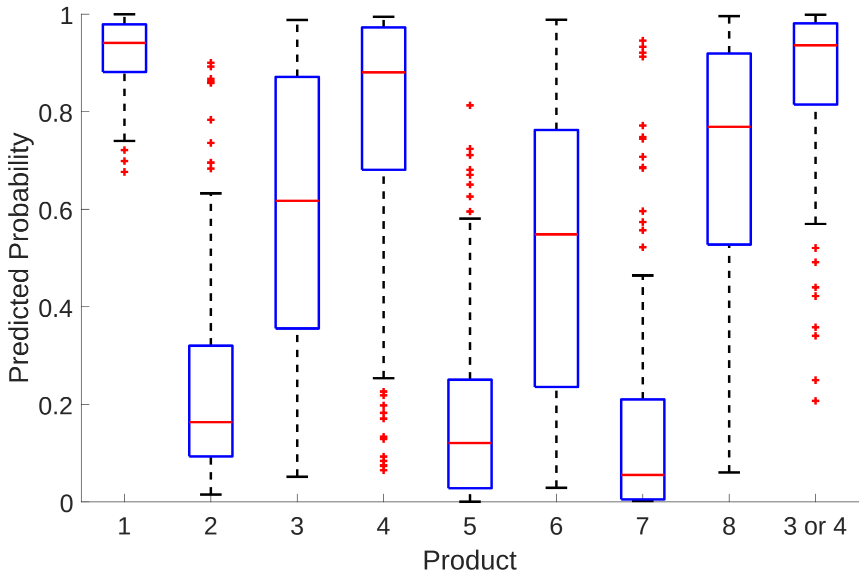
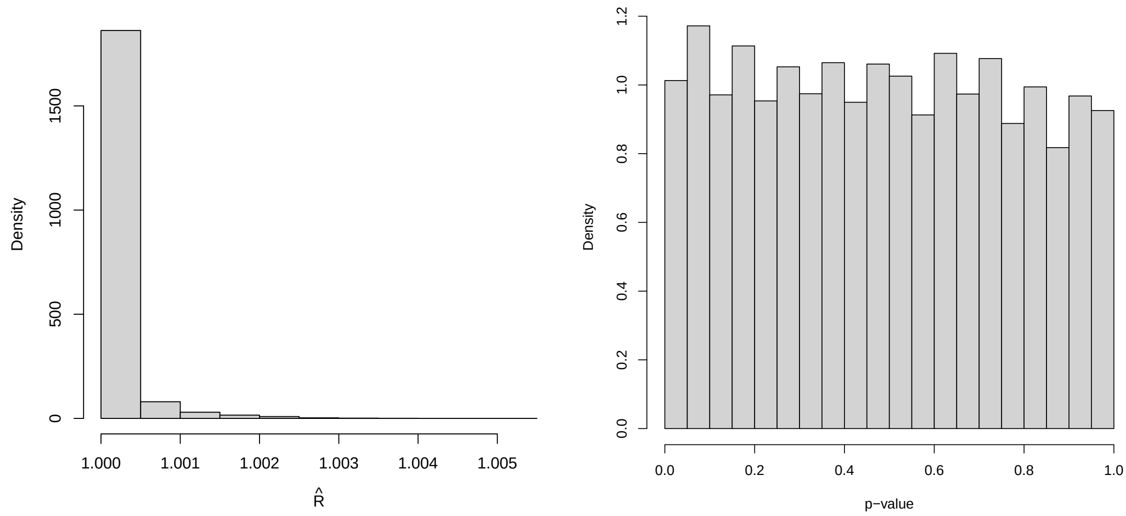



Figure 3.6: Predicted probability of a GP discussing each product for a base-case patient for each of the 162 Australian GPs.

about the need to discuss LARCs, as they do with the combined pill, but they are divided on which of the LARC products to discuss. To explore this possibility, the final column in Figure 3.6 shows the predicted probability of the GPs discussing at least one of these two products, that is $\mathbb{P}(y_3 + y_4 \geq 1)$. The results suggest that the GPs will discuss either product 3 or 4 (or both) with similar probability to the combined pill. While this joint probability does indicate a median that is similar to that of discussing the combined pill, the variability across GPs remains much larger than that associated with the combined pill. This evidence is consistent with the hypothesised resistance amongst some GPs to even discuss LARCs, let alone recommend them (Sundstrom et al., 2015).

To assess convergence of the posterior samples generated, we compute the $\hat{R}$ statistics defined in (2.16). Figure 3.7a shows the distribution of the $\hat{R}$ of all model parameters in Model 2. Since all values of $\hat{R}$ are close to 1, the model parameters have converged to their respective posterior distributions. We also perform posterior predictive checks (Rubin, 1984; Gelman and Rubin, 1995), i.e. comparing predictions drawn from the posterior predictive distribution to the observed data, to evaluate the model fit. Following Gelman et al. (2000), we compare the distribution of the realised continuous residuals $\boldsymbol{\epsilon}_{it}$ for each posterior sample to a multivariate normal




(a) $\hat{R}$. (b) p-value.

Figure 3.7: Distributions of the $\hat{R}$ of all model parameters (left) and the p-value of the non-parametric multi-sample E-statistic test (right) comparing the distribution of realised continuous residuals to a multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{R_\epsilon}$.

distribution with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{R_\epsilon}$ using the non-parametric multi-sample E-statistic test (Székely and Rizzo, 2004) implemented in the `mvn` package in `R` version 3.6.2 (R Core Team, 2019). Figure 3.7b shows the distribution of the p-value obtained. The p-value is less than 0.05 in approximately 5% of the samples, indicating that the model fits well.

### 3.6.3 Comparing sampling schemes

In order to investigate the performance of the antithetic sampler, Figure 3.8 illustrates marginal posterior distributions of those Model 2 parameters whose densities demonstrate the greatest visual differences between independent and antithetic sampling of the random effects $\boldsymbol{\alpha}_{1:N}$ and regression parameters $\boldsymbol{\beta}$. The marginal posterior distributions of $\alpha_{7,110}$ and $\beta_{236}$ are effectively the same under both updating approaches. This occurs because the mean of the conditional posterior distribution, which is a key ingredient in the deterministic antithetic sampler proposal, changes between iterations; a change largely driven by the stochastic update of the latent

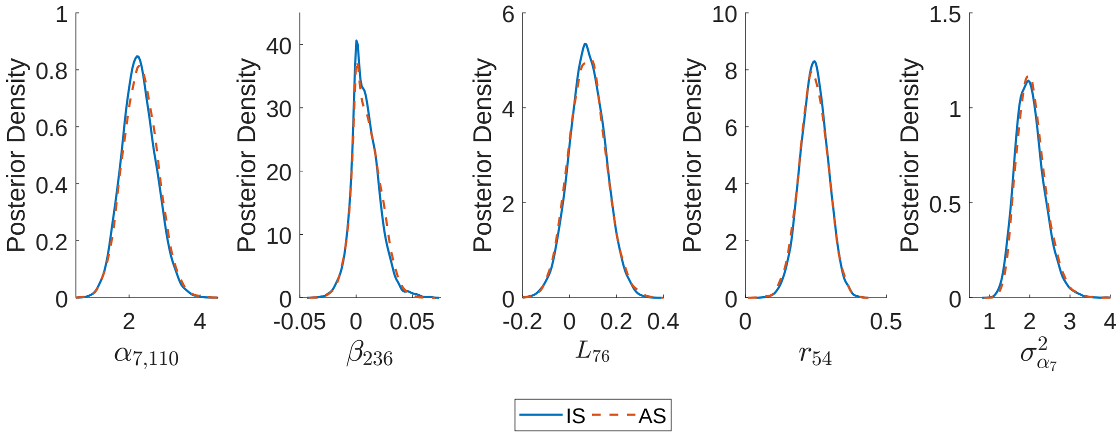



Figure 3.8: Marginal posterior density estimates of those Model 2 parameters with the greatest visual differences between using independent sampling (IS) and antithetic sampling (AS) for $\boldsymbol{\alpha}_{1:N}$ and $\boldsymbol{\beta}$.

variable $\boldsymbol{y}^*$. This outcome suggests that the posterior distribution of the other parameters remains adequately explored by the antithetic sampler.

Table 3.2 compares the performance between independent and antithetic sampling schemes when estimating Model 2. The antithetic variable method generates samples marginally faster than independent sampling because it is deterministic. Based on the results shown, we observe an improvement of 4.86 and 3.31 times performance gain on average in the mixing of $\boldsymbol{\alpha}_{1:N}$ and $\boldsymbol{\beta}$ respectively. As a result of this, the mean IACT of $\boldsymbol{y}^*$ is also improved.

| Parameter | Mean IACT | | IACT Ratio | | |
|---|---|---|---|---|---|
| | IS | AS | Min | Max | Mean |
| $\boldsymbol{y}^*$ | 3.6387 | 2.6686 | 0.8242 | 3.1419 | 1.2127 |
| $\boldsymbol{\alpha}_{1:N}$ | 16.8872 | 4.6456 | 1.4857 | 13.3424 | 4.8632 |
| $\boldsymbol{\beta}$ | 15.0446 | 4.0105 | 1.4566 | 16.0173 | 3.3111 |
| vechL($\boldsymbol{\mathcal{L}}_{\boldsymbol{\epsilon}}$) | 14.8292 | 14.5422 | 0.9338 | 1.1737 | 1.0191 |
| vechL($\boldsymbol{R}_{\boldsymbol{\epsilon}}$) | 12.7311 | 12.5170 | 0.9147 | 1.1509 | 1.0180 |
| diag($\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}$) | 24.8056 | 14.6929 | 1.3130 | 2.0651 | 1.7222 |
| vechL($\boldsymbol{R}_{\boldsymbol{\alpha}}$) | 9.5025 | 5.1716 | 1.4599 | 2.3336 | 1.8424 |
| Time per iteration | 0.0243 | 0.0239 | - | - | - |

Table 3.2: Comparison of the performance between independent sampling (IS) and antithetic sampling (AS) in the contraceptive products preference data in terms of the speed (seconds per iteration), the mean IACT and the IACT ratio for each block of parameter.

## 3.7 Conclusion

Many methods exist for fitting a multinomial logit model with random effects, such as simulated maximum likelihood (Gong et al., 2004), quadrature (Hartzel et al., 2001; Hedeker, 2003), multinomial-Poisson transformation (Lee et al., 2017), and moment-based estimation (Perry, 2017), among others. Computational strategies for the MVP model, in contrast, are less well studied and computationally expensive to implement. For instance, the random walk Metropolis-Hastings algorithm used in Chib and Greenberg (1998) does not guarantee positive definiteness of the correlation matrix in each proposal step, while at the same time not being able to explore the parameter space efficiently (Sherlock et al., 2010). In this chapter, we introduce a Hamiltonian Monte Carlo (Neal, 2011) sampling approach to generate the posterior samples of correlation matrix $\boldsymbol{R}_{\boldsymbol{\epsilon}}$, which requires reparameterising $\boldsymbol{R}_{\boldsymbol{\epsilon}}$ into an unconstrained Cholesky factor to circumvent the restrictive properties of a correlation matrix having diagonal entries of 1 and being positive definite. Credible but distant candidate parameters for the Cholesky factor can be generated from the Hamiltonian dynamics by exploiting gradient information of the posterior density, thereby reducing autocorrelation in the posterior samples. Furthermore, we propose a novel antithetic variable technique, motivated by the over-relaxation algorithm (Adler, 1981; Barone and Frigessi, 1990), to accelerate the mixing of the random effects and the regression parameters, where significant gains in efficiency are observed in our application. Although our antithetic sampling regime deterministically specifies the proposal distribution within the Metropolis-Hastings update, the ergodicity of the Markov chain is unaffected when it is embedded within a larger system of stochastic updates based on our empirical results.

Our methodology is applied to a longitudinal study of the discussion of female contraceptive products by Australian GPs, where the (binary) outcomes are obtained from the second stage of the stated preference data in Fiebig et al. (2017). An examination of the correlation matrix underlying the choices revealed a complex dependence structure between the products, hence indicating the plausibility of our formulation to model these choices in a multivariate setting. By using a multivariate model, we are able to make inference about conditional independence of the products using a graphical model, as well as computing the probability of the GPs discussing a set of contraceptive products which is not possible in univariate analyses. Our empirical study provided evidence of medical practice variation among the GPs, which is in line with the findings of earlier studies (Wennberg et al., 1982; Scott and Shiell, 1997; Davis et al., 2000). We found that the disparities in medical practice was especially more pronounced with regard to the inclusion of LARCs in the discussion with patients, and this result is also reported in Sundstrom et al. (2015). Our analysis indicated that the combined pill was the most popular contraceptive choice among the patients, and it represented a likely ready-to-wear option for many GPs. Without GPs even discussing LARCs, their uptake was likely to remain relatively constrained in such a context (Wellings et al., 2007).

## 3.8 Appendices

### 3.8.1 Sampling scheme for the MVP model with random effects

Suppose that we choose the following prior distributions: $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}_{\boldsymbol{\beta}}), \boldsymbol{\Sigma}_{\boldsymbol{\alpha}} \sim \mathcal{IW}(\nu_0, \boldsymbol{\Psi}_{\boldsymbol{\Sigma}})$ and the prior distribution on the lower triangular Cholesky factor $\boldsymbol{\mathcal{L}}_{\boldsymbol{\epsilon}}$ in (3.9) with $\nu = D + 1$. Let $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\mathcal{L}}_{\boldsymbol{\epsilon}}, \boldsymbol{\Sigma}_{\boldsymbol{\alpha}})$. Equation (3.5) gives the posterior distribution of interest under the data augmentation approach where we update $\boldsymbol{y}^*, \boldsymbol{\alpha}_{1:N}$ and each component of $\boldsymbol{\theta}$ using Gibbs sampling. For notational clarity, we will drop the superscript which indicates the sequence of the samples in a Markov chain where necessary.

**Step 1**: Updating $\boldsymbol{y}^*$

For $d = 1, \dots, D$, sample $\boldsymbol{y}^*$ conditionally one-at-a-time following Geweke (1991), i.e.

$$y^*_{d,it}|\boldsymbol{\alpha}_{1:N}, \boldsymbol{\theta}, \boldsymbol{y}^*_{-d,it}, y_{d,it} \sim \begin{cases} \mathcal{TN}_{(-\infty,0]}(\mu^{(d|-d)}_{d,it}, \sigma^{(d|-d)}_{d,it}) & \text{if } y_{d,it} = 0 \\ \mathcal{TN}_{(0,\infty)}(\mu^{(d|-d)}_{d,it}, \sigma^{(d|-d)}_{d,it}) & \text{if } y_{d,it} = 1 \end{cases}$$

where $\boldsymbol{y}^*_{-d,it} = (y_{1,it}, \dots, y_{d-1,it}, y_{d+1,it}, \dots, y_{D,it})^\top$, $\mu^{(d|-d)}_{d,it}$ and $\sigma^{(d|-d)}_{d,it}$ are the univariate $d$-th dimension conditional mean and conditional standard deviation respectively for the $\mathcal{N}(\boldsymbol{\mu}_{it}, \boldsymbol{R}_{\boldsymbol{\epsilon}})$ distribution and $\mathcal{TN}_{(a,b)}$ is a univariate normal distribution truncated to the interval $(a, b)$.

**Step 2**: Updating $\boldsymbol{\beta}$

Compute the posterior mean $\boldsymbol{\mu}_{\boldsymbol{\beta}}$ and the posterior covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ as

$$\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \left( \sum_{i=1}^{N} \sum_{t=1}^{T} (\boldsymbol{I} \otimes \boldsymbol{x}_{it}) \boldsymbol{R}_{\boldsymbol{\epsilon}}^{-1} (\boldsymbol{I} \otimes \boldsymbol{x}_{it})^\top + \boldsymbol{\Psi}_{\boldsymbol{\beta}}^{-1} \right)^{-1},$$
$$\boldsymbol{\mu}_{\boldsymbol{\beta}} = \boldsymbol{\Sigma}_{\boldsymbol{\beta}} \left( \sum_{i=1}^{N} \sum_{t=1}^{T} (\boldsymbol{I} \otimes \boldsymbol{x}_{it}) \boldsymbol{R}_{\boldsymbol{\epsilon}}^{-1} (\boldsymbol{y}^*_{it} - \boldsymbol{\alpha}_i) \right),$$

where $\otimes$ denotes the Kronecker product and set $\boldsymbol{\beta}^{[j+1]} = 2\boldsymbol{\mu}_{\boldsymbol{\beta}} - \boldsymbol{\beta}^{[j]}$ deterministically. If a horseshoe prior is specified on $\boldsymbol{\beta}$ instead, its update is the same by first sampling $\text{diag}(\boldsymbol{\Psi}_{\boldsymbol{\beta}})$ conditional on the local shrinkage parameters $\lambda_k$ and global shrinkage parameter $\tau$ (see Makalic and Schmidt (2016) for details).

**Step 3**: Updating $\boldsymbol{\mathcal{L}}_{\boldsymbol{\epsilon}}$

Sample $\boldsymbol{\mathcal{L}}_{\boldsymbol{\epsilon}}$ using the NUTS algorithm and obtain the correlation matrix $\boldsymbol{R}_{\boldsymbol{\epsilon}}$ from the relationship in (3.7).

**Step 4**: Updating $\boldsymbol{\alpha}_{1:N}$

For $i = 1, \dots, N$, compute the posterior mean $\boldsymbol{\mu}_{\boldsymbol{\alpha}_i}$ and the posterior covariance matrix $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{\alpha}}$ for the random effects $\boldsymbol{\alpha}_i$ as

$$\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{\alpha}} = \left(T\boldsymbol{R}_{\boldsymbol{\epsilon}}^{-1} + \boldsymbol{\Sigma}_{\boldsymbol{\alpha}}^{-1}\right)^{-1},$$

$$\boldsymbol{\mu}_{\boldsymbol{\alpha}_i} = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{\alpha}}\left(\boldsymbol{R}_{\boldsymbol{\epsilon}}^{-1}\sum_{t=1}^{T}\boldsymbol{y}_{it}^{*} - \boldsymbol{B}\boldsymbol{x}_{it}\right),$$

and set $\boldsymbol{\alpha}_i^{[j+1]} = 2\boldsymbol{\mu}_{\boldsymbol{\alpha}_i} - \boldsymbol{\alpha}_i^{[j]}$ deterministically.

<u>**Step 5**: Updating $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}$</u>

Sample

$$\boldsymbol{\Sigma}_{\boldsymbol{\alpha}} \sim \mathcal{IW}\left(\nu_0 + N, \sum_{i=1}^{N}\boldsymbol{\alpha}_i\boldsymbol{\alpha}_i^{\top} + \boldsymbol{\Psi}_{\boldsymbol{\Sigma}}\right).$$

Suppose that a $\mathcal{HIW}(\nu_0, \boldsymbol{A})$ prior with scales $\boldsymbol{A}$ is used for $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}$. Sample

$$a_d \sim \mathcal{IG}\left(\frac{\nu_0 + D}{2}, \nu_0\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}^{-1}(d;d) + \frac{1}{A_d^2}\right),$$

$$\boldsymbol{\Sigma}_{\boldsymbol{\alpha}} \sim \mathcal{IW}\left(\nu_0 + N + D - 1, \sum_{i=1}^{N}\boldsymbol{\alpha}_i\boldsymbol{\alpha}_i^{\top} + 2\nu_0\mathrm{diag}\left(\frac{1}{a_1}, \ldots, \frac{1}{a_D}\right)\right),$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}^{-1}(d;d)$ is the $d$-th diagonal entry of the precision matrix $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}^{-1}$.

### 3.8.2 Attributes of the patient in the Australian GP data

| Attribute | Variable | Description |
|---|---|---|
| Age | dagegp1 | Aged 16-19 years |
| | dagegp2 | Aged 20-29 years |
| | dagegp3 | Aged 30-39 years |
| | dagegp4 | Aged 40 years or more |
| Reason for encounter | drfe1 | Starting prescribed contraception for first time |
| | drfe2 | Recommencing prescribed contraception |
| | drfe3 | On pill but dissatisfied |
| | drfe4 | Using non-pill method but dissatisfied |
| Periods | dbleed1 | Heavy and/or painful periods |
| | dbleed2 | Irregular periods |
| | dbleed3 | No problems with periods |
| Blood pressure | dbp1 | Has low blood pressure |
| | dbp2 | Has normal blood pressure |
| | dbp3 | Elevated blood pressure |
| Relationship | drel1 | In long-standing relationship |
| | drel2 | In new relationship |
| | drel3 | Has no steady relationship |
| | drel4 | No information about relationship |
| Children | dchild1 | Is currently breastfeeding |
| | dchild2 | Has children but is not breastfeeding |
| | dchild3 | Has no children |
| Fertility plans | dfut1 | Does not want to have children in future |
| | dfut2 | Plans to have children in next 2 years |
| | dfut3 | Plans to have children but not in next 2 years |
| | dfut4 | Unsure about future fertility plans |
| Pill preference | dpil1 | Prefer pill to other methods |
| | dpil2 | Has no strong opinion about pill |
| | dpil3 | Prefers methods other than pill |
| Weight concern | dwt1 | Is concerned about gaining weight |
| | dwt2 | Is not concerned about gaining weight |
| Compliance | dcomp1 | Has no difficulty with compliance |
| | dcomp2 | Has difficulty with compliance |
| Income | dpay1 | Has a low to middle household income |
| | dpay2 | Has a health care card |
| | dpay3 | Has a high household income |
| Smoking | dsmk1 | Is a non-smoker |
| | dsmk2 | Smokes less than 10 cigarettes per day |
| | dsmk3 | Smokes 10 or more cigarettes per day |

Table 3.3: Categorical variables in the contraceptive discussion data with a text description for each level of attribute. Levels in grey define the attributes of a base-case patient.

### 3.8.3 Posterior means of the patient and GP fixed effects in the Australian GP data based on Model 2

| | Variable | Product | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Patient | Intercept | 1.4161 | −1.2576 | −0.3964 | 1.0991 | −2.3943 | −0.1142 | −1.7657 | 0.6918 |
| | dagegp1 | 0.1949 | −0.1329 | 0.0104 | 0.0744 | −0.5063 | −0.0205 | −0.2880 | 0.0637 |
| | dagegp3 | −0.1326 | 0.0621 | −0.0624 | −0.0002 | 0.3173 | −0.0037 | 0.0906 | 0.0108 |
| | dagegp4 | −0.3936 | 0.1851 | −0.2406 | −0.1041 | 0.8095 | −0.0270 | 0.3849 | 0.0013 |
| | drfe2 | −0.0426 | 0.0008 | −0.0388 | −0.0144 | 0.0441 | −0.0188 | −0.0449 | 0.0068 |
| | drfe3 | −0.2464 | −0.0541 | 0.0270 | 0.0788 | 0.0940 | 0.1069 | −0.0248 | 0.1364 |
| | drfe4 | −0.0206 | 0.1042 | −0.0099 | 0.0516 | 0.0678 | 0.0719 | −0.0702 | 0.0056 |
| | dbleed1 | 0.0493 | −0.1363 | 0.0615 | −0.0869 | 0.4000 | −0.0256 | −0.5274 | −0.2311 |
| | dbleed2 | 0.0160 | −0.0763 | 0.0213 | −0.0222 | 0.0070 | 0.0408 | −0.0869 | −0.0254 |
| | dbp1 | −0.0599 | −0.0011 | −0.0300 | 0.0292 | 0.0040 | 0.0317 | −0.0221 | −0.1433 |
| | dbp3 | −0.9956 | 0.2444 | 0.0070 | 0.0135 | 0.2375 | −0.2959 | 0.2561 | 0.0347 |
| | drel1 | 0.0436 | −0.0102 | −0.0963 | −0.0020 | 0.1570 | 0.0314 | 0.0282 | −0.3971 |
| | drel3 | −0.0141 | 0.0269 | −0.0208 | 0.0002 | −0.0271 | 0.0090 | −0.0186 | 0.0198 |
| | drel4 | −0.0914 | 0.0879 | 0.0667 | −0.0009 | −0.0101 | 0.0294 | 0.0029 | −0.2035 |
| | dchild1 | −1.7437 | 1.3074 | −0.0082 | −0.0889 | 0.9236 | −0.9909 | 0.5354 | −0.0371 |
| | dchild2 | −0.0458 | 0.0344 | −0.0632 | −0.0403 | 0.9850 | −0.0498 | 0.6007 | −0.0543 |
| | dfut1 | −0.3206 | −0.0043 | 0.1978 | 0.0245 | 0.6323 | −0.0786 | 0.2120 | −0.1143 |
| | dfut2 | −0.2861 | 0.1936 | −0.2169 | −0.1996 | −0.0068 | 0.0359 | −0.1438 | 0.0116 |
| | dfut4 | −0.3591 | 0.0485 | 0.0470 | 0.0099 | 0.2882 | 0.0067 | 0.0150 | 0.0323 |
| | dpil1 | 0.4724 | 0.3662 | −0.0948 | −0.2629 | −0.0120 | −0.0331 | −0.0430 | −0.0287 |
| | dpil3 | −0.1878 | −0.2417 | 0.0289 | 0.0618 | 0.0538 | 0.0329 | 0.0457 | 0.0814 |
| | dwt1 | 0.0831 | 0.0374 | −0.2582 | −0.0624 | 0.0318 | 0.0652 | −0.0130 | 0.0815 |
| | dcomp2 | −0.3401 | −0.1988 | 0.2152 | 0.0642 | 0.2321 | −0.0033 | 0.3133 | −0.0162 |
| | dpay2 | −0.0253 | −0.0558 | −0.0204 | −0.0026 | 0.0084 | 0.0595 | 0.0082 | 0.0074 |
| | dpay3 | 0.0317 | −0.0639 | −0.0697 | −0.0177 | −0.0373 | 0.2896 | −0.0177 | −0.0044 |
| | dsmk2 | −0.2665 | −0.0117 | −0.0266 | −0.0126 | −0.0038 | 0.0444 | 0.0892 | 0.0320 |
| | dsmk3 | −0.5218 | −0.0133 | 0.0132 | 0.0255 | 0.0148 | −0.0546 | 0.0467 | 0.0333 |
| GP | Female | −0.0662 | 0.0248 | −0.4417 | 0.0732 | 0.0368 | 0.5999 | −0.4474 | −0.0260 |
| | Fellow | −0.0183 | −0.0958 | 0.0709 | 0.0418 | 0.2067 | 0.1019 | −0.1456 | −0.0108 |
| | Family planning | −0.0002 | −0.0154 | −0.1203 | 0.2229 | 0.0434 | 0.0360 | −0.0324 | −0.0118 |
| | Bulk-bill | −0.0210 | −0.0349 | 0.0416 | −0.0372 | −0.0617 | 0.0036 | 0.0509 | 0.0038 |
| | Age | 0.0086 | 0.0080 | 0.0207 | −0.0061 | 0.0175 | −0.0044 | 0.0093 | −0.0100 |
| | Australian graduate | 0.0839 | 0.0564 | −0.0087 | 0.3466 | 0.0911 | −0.2385 | −0.0965 | 0.5515 |
| | Urban | −0.0888 | 0.0065 | 0.0706 | −0.0078 | 0.0099 | 0.0048 | −0.0222 | 0.1774 |

Table 3.4: Regression coefficient posterior mean estimates for the attributes of a female patient and the characteristics of a GP based on Model 2 for various products in the contraceptive discussion data. Parameters whose 90% credible interval does not include 0 are shown in grey.

### 3.8.4 Posterior mean of $\boldsymbol{R}_\epsilon$ in the Australian GP data based on Model 2

$$\boldsymbol{R}_\epsilon = \begin{bmatrix} 1.0000 & -0.1126 & -0.0515 & -0.0450 & -0.2349 & 0.4712 & -0.2065 & -0.0204 \\ -0.1126 & 1.0000 & 0.1625 & 0.0449 & -0.0263 & -0.2679 & -0.0537 & -0.0494 \\ -0.0515 & 0.1625 & 1.0000 & 0.5873 & 0.1779 & 0.0153 & 0.1836 & 0.0189 \\ -0.0450 & 0.0449 & 0.5873 & 1.0000 & 0.2414 & 0.0379 & 0.1889 & 0.1048 \\ -0.2349 & -0.0263 & 0.1779 & 0.2414 & 1.0000 & -0.0696 & 0.5177 & -0.0771 \\ 0.4712 & -0.2679 & 0.0153 & 0.0379 & -0.0696 & 1.0000 & -0.0055 & 0.1831 \\ -0.2065 & -0.0537 & 0.1836 & 0.1889 & 0.5177 & -0.0055 & 1.0000 & 0.2058 \\ -0.0204 & -0.0494 & 0.0189 & 0.1048 & -0.0771 & 0.1831 & 0.2058 & 1.0000 \end{bmatrix}$$

### 3.8.5 Posterior mean of $\boldsymbol{\Sigma}_\alpha$ in the Australian GP data based on Model 2

$$\boldsymbol{\Sigma}_\alpha = \begin{bmatrix} 0.5574 & 0.3005 & 0.2760 & 0.2490 & 0.0795 & 0.2056 & 0.0634 & 0.2592 \\ 0.3005 & 0.6923 & 0.3040 & 0.2679 & 0.1875 & 0.2199 & 0.2418 & 0.3358 \\ 0.2760 & 0.3040 & 1.3574 & 0.2590 & -0.0188 & 0.1065 & 0.2586 & 0.0751 \\ 0.2490 & 0.2679 & 0.2590 & 1.6084 & 0.5244 & 0.2538 & -0.2229 & 0.2383 \\ 0.0795 & 0.1875 & -0.0188 & 0.5244 & 1.1040 & 0.2911 & 0.0135 & 0.2612 \\ 0.2056 & 0.2199 & 0.1065 & 0.2538 & 0.2911 & 1.5142 & 0.2950 & 0.4906 \\ 0.0634 & 0.2418 & 0.2586 & -0.2229 & 0.0135 & 0.2950 & 2.0530 & 0.4144 \\ 0.2592 & 0.3358 & 0.0751 & 0.2383 & 0.2612 & 0.4906 & 0.4144 & 1.2942 \end{bmatrix}$$

# Chapter 4

# Multiclass classification of growth curves using random change points and heterogeneous random effects

## 4.1 Introduction

According to the latest joint malnutrition estimates by the United Nations Children's Fund, World Health Organization (WHO), and World Bank Group (2018), it is estimated that in 2017, stunted growth is prevalent in 22.2% of the global population under the age of 5, or over 150 million children worldwide. This is particularly serious in low to medium income countries where the rate of stunting is 35.0%. A major contributor to stunted growth is prolonged faltering, defined as a slower rate of growth compared to a reference healthy population of the same age and gender, which comes with adverse consequences such as increased susceptibility of individuals to diarrhoea and respiratory infections (Kossmann et al., 2000), abnormal neurointegrative development (Benítez-Bribiesca et al., 1999) and at a population scale, a capital loss to the labour market (Hoddinott et al., 2013). Therefore, it is imperative to take early preventive measures to minimise these impacts. In order to implement preventive measures, faltered children must first be identified in the population. It is additionally important to distinguish between different growth patterns, as each type represents a particular growth behaviour and so merits a different response (Collins et al., 2006). For example, children who caught up on growth after having faltered may have benefited from the intake of better diets or nutritional supplements. Such strategies can then be extended to other children in the cohort to improve their growth.

In the functional data analysis (Ramsay and Silverman, 2005) literature, several clustering methods are proposed. One of these methods first reduces the dimension of the inherently infinite-dimensional functional data using a finite basis expansion (see e.g. Abraham et al. (2003) and Rossi et al. (2004) for such approximation using a B-spline) or a functional principal component analysis, for example in Peng and Müller (2008) and Xiao et al. (2016), and then clusters the basis coefficients or the functional principal component scores using classical classification algorithms (see Chapter 14 in Hastie et al., 2009). Another related classification approach that is proposed in Tarpey and Kinateder (2003), Ferraty and Vieu (2006) and Tokushige et al. (2007) computes the pairwise Euclidean norm between curves, which is used as a similarity measure for $k$-means clustering. On the other hand, model-based clustering methods are presented in Shi et al. (1996), James and Sugar (2003), Heard et al. (2006) and Giacofci et al. (2013), whereby the basis coefficients are typically assumed to be distributed according to a mixture of normal distributions, with each component determining the classification memberships. Model-based techniques using principal components modelling are considered in Bouveyron et al. (2007) and Bouveyron and Jacques (2011), whereby the data are fitted in group-specific functional subspaces using the notion of functional probability density defined in Delaigle and Hall (2010). However, functional methods are not suitable for sparse observations. This is particularly relevant when studying the growth development of children in low to medium income countries, where the number of height-related measurements per child is small and they are taken at irregular intervals (see e.g. Anderson et al. (2019)

for a summary of these statistics for the datasets used in their study). This has made regression-based methods a more popular approach in modelling growth curves.

Similarly, both multi-stage and model-based methods are proposed for growth curve regression models. Leung et al. (2017) suggests classifying children with the lowest 10% of values of random velocity estimates extracted from a linear mixed effects model (Laird and Ware, 1982) as having experienced an "abnormal" growth, whereas Lee et al. (2018) classifies children into two groups based on their minimum random random velocity estimates obtained from a broken stick model (Ruppert et al., 2003), which is a piecewise linear model with breaks at several knots. Here, the velocity refers to the rate of change in the measurements used for growth modelling (Cole, 1998), which in practice includes the raw height and other suitable metrics such as z-scores. Z-scores, such as the height-for-age z-score (HAZ), is a measure defined by the WHO Multicentre Growth Reference Study Group (2006) that compares the anthropometric measurements of a child matched against a reference population of healthy children of the same age and gender. These staged methods rely on an arbitrary threshold and the resulting classifications are not necessarily comparable between different populations. As such, growth mixture models (Muthén and Shedden, 1999; Nagin, 1999; Muthén and Muthén, 2000; Li et al., 2001; Muthén, 2008) are considered for analysing the change of growth patterns in longitudinal measurement data. These models assume that the population is heterogeneous and comprises of multiple smaller subgroups (Berlin et al., 2014). Unobserved heterogeneity between children is modelled using random effects and finite mixture distribution (McLachlan and Peel, 2000), thereby allowing different sets of velocities to capture group-specific growth trajectories. Despite its extensive use, one unresolved issue in the application of a growth mixture model is how to determine the "correct" number of mixture components $G$ (Nylund et al., 2007). Most approaches choose $G$ based on information criterion (Dasgupta and Raftery, 1998; Magidson and Vermunt, 2004), likelihood ratio test (Titterington et al., 1985), goodness-of-fit test (Verbeke and Lesaffre, 1996), among others, which requires fitting multiple models with different values of $G$.

This chapter adopts a different strategy in choosing $G$: modelling the mixture distribution of growth velocities within a Bayesian framework using the Dirichlet process (DP; Ferguson, 1973) prior, with $G$ being inferred as part of the posterior sampling algorithm. We note that the concept of using a DP prior for classification is not new. In the context of functional data, Scarpa and Dunson (2014) cluster body temperature curves of women during menstrual cycle using a functional DP prior, Ray and Mallick (2006) and Suarez and Ghosal (2016) use a wavelet representation of the functional data with a DP prior on the coefficients to classify precipitation curves, while Rodriguez et al. (2008) propose a nested DP model to study differences in health care quality. A DP prior is particularly convenient in our present framework of growth curve modelling as it circumvents model selection procedure by modelling the mixture distribution non-parametrically and adapting the complexity of the model, i.e. the value of $G$, to the amount of data available. Furthermore, Teh (2011) show that $G$ scales logarithmically in the number of observations so that the number of subgroups is bounded above.

Anderson et al. (2019) compare the most common growth modelling approaches and find that the broken stick model, when used in conjunction with the z-scores has superior performance in terms of out-of-sample prediction. As such, we use the broken stick model as the underlying growth curves model. The locations of the knots in the broken stick model are generally assumed to be equally spaced (Lee et al., 2018; Anderson et al., 2019). A further contribution of this chapter is to introduce random change points for the knots into the broken stick model, rather than their locations being arbitrarily fixed, and simultaneously modelling the growth velocities of each child as random slopes. These change points are modelled as random effects so that the difference in the timing of growth phases between children can be accommodated within the model and the classification process. Probabilistic inference for these change points is straightforward and can be implemented using Markov chain Monte Carlo (MCMC) algorithms. Our simulation studies demonstrate the superior performance of the random change points model compared to the broken stick model with fixed knots in terms of the agreement to the true component membership. Due to the limited flexibility of the broken stick model with fixed knot locations in capturing change point heterogeneity, the number of clusters tends to be overestimated. This occurs as a result of the bias in the estimation of growth velocity.

This chapter is organised as follows: Section 4.2 describes the extension of the broken stick model to include mixture distributed random effects using a DP prior. It also introduces individual-specific random change points into the model, and provides implementation details. Section 4.3 investigates the performance of the proposed model via simulation studies. Section 4.4 provides an analysis of the Vellore growth curve dataset, and Section 4.5 concludes.

## 4.2 Methods

### 4.2.1 A broken stick model with mixture distributed random slopes

A popular method for modelling longitudinal growth data in the epidemiological literature is the broken stick model (Ruppert et al., 2003). This may be defined as

$$y_{it} = \alpha_i + \beta_{0i}(\omega_{it} - (\omega_{it} - \xi_1)_+) + \beta_{Ki}(\omega_{it} - \xi_K)_+ + \sum_{k=1}^{K-1} \beta_{ki}((\omega_{it} - \xi_k)_+ - (\omega_{it} - \xi_{k+1})_+) + \epsilon_{it}, \tag{4.1}$$

$$\alpha_i \sim \mathcal{N}(\mu_\alpha, \sigma_\alpha^2), \quad \epsilon_{it} \sim \mathcal{N}(0, \sigma_\epsilon^2), \tag{4.2}$$

for $i = 1, \ldots, N, t = 1, \ldots, T_i$, where $y_{it} \in \mathbb{R}$ denotes the height-for-age z-score (HAZ) for child $i$ on the $t$-th measurement occasion at age $\omega_{it}$, $(a)_+ = \max\{0, a\}$ is the positive part of $a$ and $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_K)^\top$ is an ordered vector of $K$ predetermined knots, or change points, such that $\xi_1 < \cdots < \xi_K$. The individual-specific random intercept $\alpha_i$ and error $\epsilon_{it}$ are both assumed to be independent and normally distributed with parameter vectors given by $(\mu_\alpha, \sigma_\alpha^2)$ and $(0, \sigma_\epsilon^2)$ respectively. The child-specific and time invariant random intercept $\alpha_i$ controls for heterogeneity in the HAZ at birth, centred around the population mean $\mu_\alpha$, and is assumed to be uncorrelated with the error term $\epsilon_{it}$. The broken stick model fits $K+1$ piecewise linear segments with breaks at $\boldsymbol{\xi}$ to model the growth trajectory calibrated in terms of the HAZ. The formulation in (4.1) enables an individual child's growth velocity to be obtained directly from the regression coefficients since $\beta_{ki}$ represents the rate of change in the HAZ between years $\xi_k$ and $\xi_{k+1}$. This model could be extended to higher order polynomials but this would complicate direct interpretation of the parameters as a measure of change.

We now consider distributional assumptions on the individual-specific growth velocity vector $\boldsymbol{\beta}_i = (\beta_{0i}, \ldots \beta_{Ki})^\top$. Anderson et al. (2019) and Lee et al. (2018) model $\boldsymbol{\beta}_i$ as realisations from a multivariate normal distribution with mean vector $\boldsymbol{\mu}_{\boldsymbol{\beta}}$ and covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$, i.e.

$$\boldsymbol{\beta}_i \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}).$$

This signifies a homogeneous population model where individual growth profiles largely follow the trend of a global trajectory, with the variability of deviations across individuals from this mean curve determined by $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$. Under this assumption the growth rate is, on average, the same for all children in the population at each knot. However, this is rarely the case in practice. For example, Goode et al. (2014) find that higher socio-economic status has a positive impact on the HAZ through greater health consciousness and better household sanitation systems in their analysis of child health data from the China Health and Nutrition Survey, which is a household survey conducted in nine Chinese provinces: Guangxi, Guizhou, Heilongjiang, Henan, Hubei, Hunan, Jiangsu, Liaoning, and Shandong. Studies have also found evidence of correlation between growth velocity during childhood and biological factors such as maternal height (see e.g. Ramakrishnan et al. (1999) for a related study in rural eastern Guatemala). Therefore, we alternatively consider a more structured normal mixture distribution

$$\boldsymbol{\beta}_i \sim \sum_{g=1}^{G} w_g \mathcal{N}(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g), \tag{4.3}$$

for positive weights $w_g > 0$ with $\sum_{g=1}^{G} w_g = 1$, in order to accommodate a more complex composition of the population. Each mixture component $g$ in (4.3) therefore corresponds to a particular type of growth pattern as characterised by the set of mean growth velocities $\boldsymbol{\mu}_g$, and each child belongs (probabilistically) to one of these $G$ subgroups. By clustering the children into different subgroups, subsequent analyses can then identify risk factors which are associated with the manifestation of certain growth behaviours. Equation (4.3) requires specifying the number of subgroups $G$, which is typically unknown *a priori* in practice. Common methods for choosing $G$ are discussed in Section 2.1.2. In the next section, we describe a Bayesian approach that incorporates the estimation of $G$ via a Dirichlet process prior.

### 4.2.2 Bayesian non-parametric mixture modelling

Choosing a suitable value for the number of components $G$ in a mixture distribution is a non-trivial problem. Most of the methods described in Section 2.1.2 are ad-hoc, requiring the need to fit multiple models of differing complexity, and selecting the "best" model based on certain criteria. In order to circumvent this kind of model selection procedure, we employ a Bayesian non-parametric approach to fit a model which allows its complexity to be parameterized within the model. In general, such flexibility is achieved by assuming an infinite dimensional parameter space $\Theta$, on which a prior distribution is then developed. In our present context, subscribing to this framework leads to an infinite mixture model.

Specifically, we consider a DP mixture model (Antoniak, 1974) for $\boldsymbol{\beta}_i$ for which

$$\boldsymbol{\beta}_i|(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \quad (\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)|\mathcal{G} \sim \mathcal{G}, \quad \mathcal{G} \sim \mathcal{DP}(\lambda, \mathcal{G}_0), \tag{4.4}$$

for $i = 1, \ldots, N$, where $\tilde{\boldsymbol{\theta}}_i = (\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ are the parameters of a normal distribution specifying the mixture component from which the growth velocity $\boldsymbol{\beta}_i$ of child $i$ is generated and $\mathcal{DP}(\lambda, \mathcal{G}_0)$ denotes a Dirichlet process with concentration parameter $\lambda > 0$ and base distribution $\mathcal{G}_0$. Since the parameter of interest is a mean vector and covariance matrix pair, one common choice of $\mathcal{G}_0$ is the normal-inverse-Wishart distribution with parameters $(\boldsymbol{m}, c, \nu, \boldsymbol{\Psi})$, having density function

$$p(\tilde{\boldsymbol{\theta}}) \propto |\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{c}{2}(\boldsymbol{\mu}-\boldsymbol{m})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\boldsymbol{m})\right) \times |\boldsymbol{\Sigma}|^{-(\nu+K+2)/2} \exp\left(-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Psi}\boldsymbol{\Sigma}^{-1})\right).$$

Integrating out $\mathcal{G}$ from (4.4), Blackwell and MacQueen (1973) show that the conditional prior distribution induced on $\tilde{\boldsymbol{\theta}}_i$ follows a Pólya urn scheme, constructed as

$$\tilde{\boldsymbol{\theta}}_i | \tilde{\boldsymbol{\theta}}_1, \ldots, \tilde{\boldsymbol{\theta}}_{i-1} \sim \frac{1}{\lambda + i - 1} \sum_{j=1}^{i-1} \delta_{\tilde{\boldsymbol{\theta}}_j} + \frac{\lambda}{\lambda + i - 1} \mathcal{G}_0. \tag{4.5}$$

From (4.5), the generating mechanism of the first parameter $\tilde{\boldsymbol{\theta}}_1$ involves drawing an independent sample from the base distribution $\mathcal{G}_0$. Subsequent samples, $\tilde{\boldsymbol{\theta}}_i$, are then obtained by setting $\tilde{\boldsymbol{\theta}}_i$ to be a random draw from the previous samples $\{\tilde{\boldsymbol{\theta}}_1, \ldots, \tilde{\boldsymbol{\theta}}_{i-1}\}$ with probability proportional to $i - 1$ (thereby directly introducing a clustering effect within the mixture model) or a new sample from $\mathcal{G}_0$ (i.e. a new mixture component) with probability proportional to $\lambda$. Accordingly, the generated samples $\{\tilde{\boldsymbol{\theta}}_1, \ldots, \tilde{\boldsymbol{\theta}}_N\}$ concentrate on a set of unique values $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G\}$, with a larger value of $\lambda$ giving rise to a larger (random) value of $G$.

### 4.2.3 Knot locations as random effects

So far, the knot location vector $\boldsymbol{\xi}$ in (4.1) has been treated as predetermined and fixed across all children in the population. However, this is unrealistic in the current context as individual children react differently to treatment interventions such as the administration of vitamins or to negative experiences such as infections, which will likely occur at individual-specific time points. The heterogeneity in the timing of such events is likely to cause individual trajectories to change course at different time points. Furthermore, erroneously fixing $\boldsymbol{\xi}$ in the broken stick model, will result in a biased estimate of the growth velocity $\boldsymbol{\beta}_i$ as the regression lines between two neighbouring segments are connected at the knot. This then affects the classification of each child because their growth patterns are summarised by $\boldsymbol{\beta}_i$. Therefore, a sensible approach is to model the knot locations within the interval of $[0, \mathcal{I}]$ as a child-specific ordered vector of knot random effects $\boldsymbol{\xi}_i = (\xi_{i1}, \ldots, \xi_{iK})^\top$. We construct the prior distribution of each $\boldsymbol{\xi}_i$ as

$$p(\boldsymbol{\xi}_i) \propto \prod_{k=1}^{K+1} (\xi_{ik} - \xi_{i,k-1}) \times \prod_{k=1}^{K} \mathbb{1}\left(\xi_{ik} \in \left(\frac{(k-1)\mathcal{I}}{K}, \frac{k\mathcal{I}}{K}\right)\right), \tag{4.6}$$

for $i = 1, \ldots, N$, where $\xi_{i0} = 0$ and $\xi_{i,K+1} = \mathcal{I}$ for convenience and $\mathbb{1}(E)$ is an indicator function which takes value 1 if the event $E$ occurs and 0 otherwise. The first product term in (4.6) is the distribution of the even-numbered order statistics from $2K + 1$ points uniformly distributed on $[0, \mathcal{I}]$, as used in Green (1995), which probabilistically encourages consecutive knot points to be uniformly spaced such that each random knot $\xi_{ik}$ is centered around $(k - 0.5)\mathcal{I}/K$ *a priori*. Although the even-numbered order statistics distribution usefully penalises short subintervals, it would still be possible for the knots $\boldsymbol{\xi}_i$ to be concentrated in regions where there is an abundance of informative data. As such, we additionally impose a hard constraint via the second product term in (4.6), which ensures that there is exactly one knot within each of the $K$ subintervals of equal length on $[0, \mathcal{I}]$ (see e.g. Fan et al. (2010) for a similar construction).

### 4.2.4 Posterior inference and cluster analysis

Posterior simulation for the DP mixture model defined in (4.4) is straightforward to implement using MCMC methods (Gelman et al., 2013). However, naive sampling schemes based on the Pólya urn construction of the DP prior in (4.5) can be highly inefficient due to the resulting numerical approximations of high dimensional integrals when the dimension of $\boldsymbol{\beta}_i$ is large. Let $\boldsymbol{s} = (s_1, \ldots, s_N)^\top, s_i \in \{1, 2, \ldots\}$ be the vector of cluster allocation variables determining which subgroup each child belongs to. Here we focus on the MCMC sampling of $\boldsymbol{s}$, the weight of each mixture component $w_g$, the concentration parameter $\lambda$ of the DP prior, and the knot random effects $\boldsymbol{\xi}_i$, as MCMC updates for other model parameters are straightforward. We implement the slice sampler proposed by Walker (2007) which is based on the stick-breaking representation. The slice sampling algorithm introduces auxiliary variables $\mathfrak{u}_i, i = 1, \ldots, N$, whose distribution conditional on the label $s_i$ is uniform on $[0, w_{s_i}]$. This parameter augmentation strategy gives the conditional posterior distribution $\mathbb{P}(s_i = g|\cdots)$ of $s_i$ as

$$\mathbb{P}(s_i = g|\cdots) = \frac{\prod_{t=1}^{T_i} p(y_{it}|\boldsymbol{\theta}_g, \alpha_i, \boldsymbol{x}_{it}(\boldsymbol{\xi}_i), \sigma_\epsilon^2)\mathbb{1}(w_g > \mathfrak{u}_i)}{\sum_{h:w_h>\mathfrak{u}_i} \prod_{t=1}^{T_i} p(y_{it}|\boldsymbol{\theta}_h, \alpha_i, \boldsymbol{x}_{it}(\boldsymbol{\xi}_i), \sigma_\epsilon^2)}, \tag{4.7}$$

where $\boldsymbol{x}_{it}(\boldsymbol{\xi}_i)$ is the regressor in (4.1) by replacing the fixed knots $\boldsymbol{\xi}$ with individual-specific random knots $\boldsymbol{\xi}_i$, and given the univariate normal density $\phi$

$$p(y_{it}|\boldsymbol{\theta}_g, \alpha_i, \boldsymbol{x}_{it}(\boldsymbol{\xi}_i), \sigma_\epsilon^2) = \phi(y_{it}; \alpha_i + \boldsymbol{x}_{it}^\top(\boldsymbol{\xi}_i)\boldsymbol{\mu}_g, \boldsymbol{x}_{it}^\top(\boldsymbol{\xi}_i)\boldsymbol{\Sigma}_g\boldsymbol{x}_{it}(\boldsymbol{\xi}_i) + \sigma_\epsilon^2),$$

is the likelihood function for the $t$-th measurement from child $i$ under mixture group $g$. Conditional on the other model parameters, (4.7) indicates that the possible subgroups to which any child belongs are restricted to a finite set of components in the infinite dimensional parameter space $\Theta$ whose weights are greater than $\mathfrak{u}_i$. Given this, the probability of child $i$ belonging to any of these subgroups is then proportional to the appropriate likelihood term for each group.

Denoting the number of children in the $g$-th occupied mixture component by $N_g, g = 1, \ldots, G$, where each child is assigned to one of the mixture components probabilistically according to (4.7), the conditional posterior distribution of the weights can be shown to be Dirichlet distributed (Ge et al., 2015), i.e.

$$(w_1, \ldots, w_G, w')|\cdots \sim \text{Dirichlet}(N_1, \ldots, N_G, \lambda),$$

where $w' = 1 - \sum_{g=1}^{G} w_g$ is the weight on $\Theta' = \Theta\backslash\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G\}$. The stick-breaking process in (2.20) is then applied to $w'$ until the length of the stick is less than $\min\{\mathfrak{u}_1, \ldots, \mathfrak{u}_N\}$. For each additional break of the stick with initial length $w'$, a new sample $\boldsymbol{\theta}' \in \Theta'$ is drawn from the base distribution, $\mathcal{G}_0$. The rationale behind this is to ensure that $\Theta'$ has zero probability of being sampled in (4.7). The generation of additional empty mixture components and the removal of unoccupied components after sampling $\boldsymbol{s}$ changes the value of $G$ between MCMC iterations.

Escobar and West (1995) show that likelihood function of the hyperparameter $\lambda$ is given by

$$p(\lambda|\cdots) \propto \lambda^G \frac{\Gamma(\lambda)}{\Gamma(\lambda+N)} \propto \lambda^G \int_0^1 c^{\lambda-1}(1-c)^{N-1}dc,$$

where $\Gamma(\lambda) = \int_0^\infty x^{\lambda-1}\exp(-x)dx$ denotes the gamma function and $c$ is a latent variable. Under a Gamma$(a_\lambda, b_\lambda)$ prior for $\lambda$, the joint posterior distribution of $(\lambda, c)$ can be written as

$$\pi(\lambda, c|\cdots) \propto \lambda^{a_\lambda+G-1}\exp(-b_\lambda\lambda)\, c^{\lambda-1}(1-c)^{N-1}. \tag{4.8}$$

Equation (4.8) simplifies posterior sampling of the hyperparameter $\lambda$ since the conditional posterior distributions of $\lambda$ and $c$ are now standard probability distributions. Therefore, the MCMC update for $(\lambda, c)$ can be performed by Gibbs sampling, whereby $c$ is generated from a Beta$(\lambda, N)$ distribution and then $\lambda$ from a Gamma$(a_\lambda + G, b_\lambda - \log c)$ distribution.

The individual-specific knots, $\boldsymbol{\xi}_i$, can be updated one knot component, $k = 1, \ldots, K$, at a time using a Metropolis-Hastings update. Writing $\tilde{\boldsymbol{\xi}}_i^{(k)} = (\xi_{i1}, \ldots, \xi_{i,k-1}, \tilde{\xi}_{ik}, \xi_{i,k+1}, \ldots, \xi_{iK})^\top$ as the proposed vector of knot locations with $\tilde{\xi}_{ik}$ sampled uniformly from the subinterval $((k-1)\mathcal{I}/K, k\mathcal{I}/K)$ and all other $\boldsymbol{\xi}_i$ components set at their previous values, the probability of accepting the proposal is given by

$$\min\left\{1, \prod_{t=1}^{T_i} \frac{p(y_{it}|\boldsymbol{\beta}_i, \alpha_i, \boldsymbol{x}_{it}(\tilde{\boldsymbol{\xi}}_i^{(k)}), \sigma_\epsilon^2)}{p(y_{it}|\boldsymbol{\beta}_i, \alpha_i, \boldsymbol{x}_{it}(\boldsymbol{\xi}_i), \sigma_\epsilon^2)} \times \frac{(\xi_{i,k+1} - \tilde{\xi}_{ik})(\tilde{\xi}_{ik} - \xi_{i,k-1})}{(\xi_{i,k+1} - \xi_{ik})(\xi_{ik} - \xi_{i,k-1})}\right\},$$

where $p(y_{it}|\boldsymbol{\beta}_i, \alpha_i, \boldsymbol{x}_{it}(\cdot), \sigma_\epsilon^2) = \phi(y_{it}; \alpha_i + \boldsymbol{x}_{it}^\top(\cdot)\boldsymbol{\beta}_i, \sigma_\epsilon^2)$. For improved efficiency, the update for $\boldsymbol{\xi}_i$ can be performed in parallel for each child.

## 4.3 Simulation study

We now examine how the above model and inferential procedure performs in a controlled setting. A population of $N = 400$ children was generated under the broken stick model in (4.1). The child-specific random intercepts $\alpha_i$ were chosen to follow a $\mathcal{N}(0.75, 0.5)$ distribution and the error variance $\sigma_\epsilon^2$ is set as 0.15. The individual growth velocity vectors $\boldsymbol{\beta}_i$ are generated from a normal mixture distribution with $G = 4$ components with equal weights. The number of knots is specified as $K = 2$ so that the growth trajectories are constructed from three piecewise linear segments. The mean velocities for each subgroup, $\boldsymbol{\mu}_g$, $g = 1, \ldots, 4$, are given by

$$\begin{bmatrix} \boldsymbol{\mu}_1 & \boldsymbol{\mu}_2 & \boldsymbol{\mu}_3 & \boldsymbol{\mu}_4 \end{bmatrix} = \begin{bmatrix} -3.0 & -7.5 & -3.0 & 4.0 \\ -3.0 & -5.0 & -1.0 & 1.0 \\ -3.0 & 0.0 & 3.0 & -3.0 \end{bmatrix},$$

and the covariance matrix for each subgroup, $\boldsymbol{\Sigma}_g = 0.2\boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix. Figure 4.1 shows three scatterplots, with each plot illustrating the bivariate marginal distributions of the generated $\boldsymbol{\beta}_i$. Using the same $\boldsymbol{\beta}_i$, we construct two sets of data to examine different designs on the knot locations. The first dataset ($D_{fixed}$)

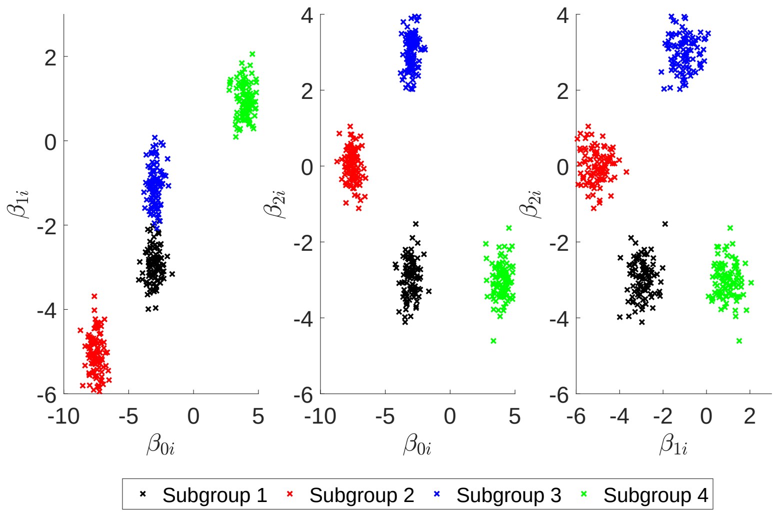



Figure 4.1: Scatterplots illustrating the bivariate marginal distributions of the generated $N = 400$ growth trajectory vectors $\boldsymbol{\beta}_i$, $i = 1, \ldots, N$. The realised vectors are coloured by subgroup membership.

has fixed and equally spaced knots within the interval $[0, 1]$ so that $(\xi_1, \xi_2) = (\frac{1}{3}, \frac{2}{3})^\top$, while the second dataset ($D_{random}$) generates random knots for each child, with the first and second knots drawn uniformly from the intervals $[0, 0.5]$ and $[0.5, 1]$ respectively. Figure 4.2 illustrates growth profiles for one representative individual from each subgroup (columns) and also compares their differences between $D_{fixed}$ (top panels) and $D_{random}$ (bottom panels) for the same $\boldsymbol{\beta}_i$. The first three subgroups exhibit a faltering pattern with different rates during the first two time periods. This faltering then either continues (subgroup 1), plateaus (subgroup 2) or growth improves (subgroup 3) in the third time period. In contrast, subgroup 4 is qualitatively different, whereby the children experience accelerated growth over time before a decline in the HAZ score is observed closer to age 1.

For each child, a random number (uniformly between 10 and 20) of HAZ observations was generated, with the measurement time of each observation being uniformly

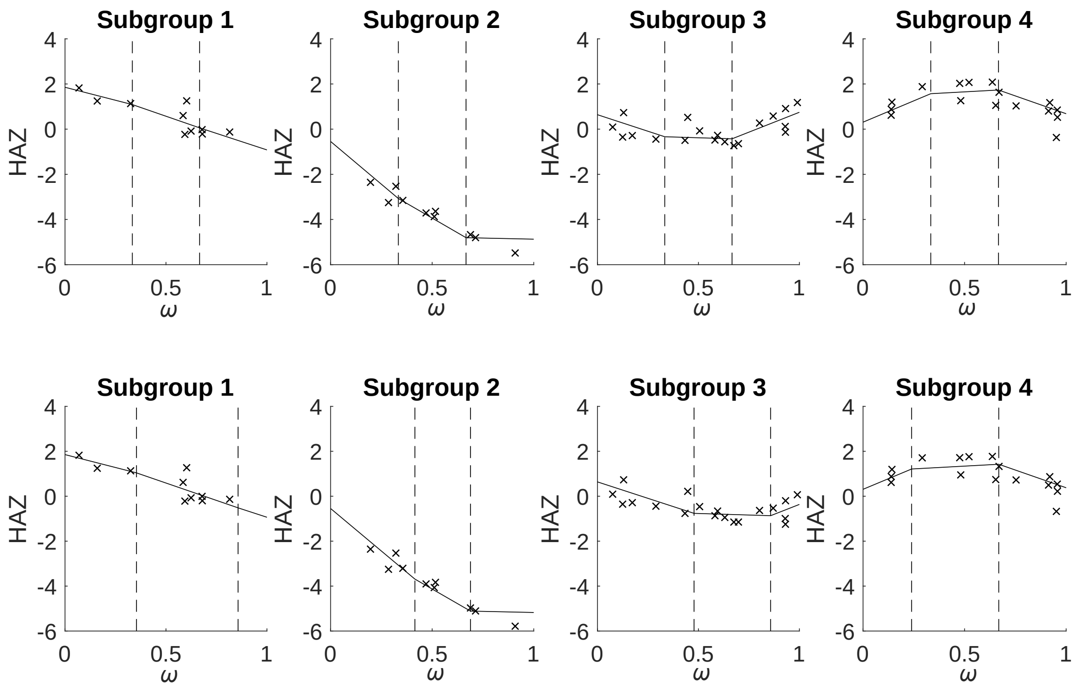



Figure 4.2: HAZ score versus age (from birth until year 1) of representative simulated individuals from each of the four distinct growth trajectory groups (columns) under the broken stick model. Growth curve knot points are indicated by vertical dashed lines; the top panels showing equally spaced fixed knots ($D_{fixed}$) and the bottom panels showing the same individuals but with random knot points ($D_{random}$). The observed data ($\times$) is generated using the same random errors around each growth curve for each individual (top versus bottom panel in each column).

distributed between birth and age 1. To reduce variability in the comparison between $D_{fixed}$ and $D_{random}$, the two datasets were generated using the same number, measurement time of observations and random deviations around the two growth trajectories for each individual child. In this manner, the empirical residuals around each growth trajectory are identical between the two datasets for each child, and the knot locations are the only component which varies between $D_{fixed}$ and $D_{random}$.

For inference we adopted weakly informative conjugate prior distributions. In particular we specified $\mu_\alpha \sim \mathcal{N}(0, 25)$, $\sigma_\alpha, \sigma_\epsilon \sim$ half-Cauchy(5), $\lambda \sim$ Gamma(2, 4), and the base distribution $\mathcal{G}_0$ has a normal-inverse-Wishart distribution with parameters $(\boldsymbol{m}, c, \nu, \boldsymbol{\Psi}) = (\boldsymbol{0}, 10^{-3}, K + 2, \boldsymbol{I})$. Posterior sampling is achieved using MCMC following the details in Section 4.2.4 with a chain of 100 000 iterations, with the first 50 000 iterations discarded as burn in and retaining every 20th of the remaining samples, yielding 2 500 thinned iterates for analysis. We fit two model variants to each dataset: $M_{fixed}$ is the model with $K = 2$ fixed and equally spaced knots at $\omega = \frac{1}{3}$ and $\omega = \frac{2}{3}$, whereas $M_{random}$ is the model where the two knot points are allowed to vary for each individual. The optimal clustering $\hat{\boldsymbol{s}}$ is obtained by optimising the PEAR (or $\mathbb{E}_{\boldsymbol{s}}\left[ARI(\hat{\boldsymbol{s}}, \boldsymbol{s})\right]$) described in Section 2.4.1 using the `mcclust` package version 1.0 (Fritsch, 2012) in `R` version 3.6.2 (R Core Team, 2019).

Table 4.1 presents a summary of the performance of models $M_{fixed}$ and $M_{random}$ for both datasets, in terms of the final group classification outcome. For the fixed knot dataset $D_{fixed}$ both models perform similarly well by correctly identifying the true number of groups. This largely occurs as the $M_{fixed}$ model is contained within the $M_{random}$ model, and so the latter has the capacity to achieve the same performance as the former when the data have the knot structure assumed in $M_{fixed}$. Of course, here model $M_{fixed}$ is slightly outperforming $M_{random}$ in terms of the $\mathbb{E}_{\boldsymbol{s}}\left[ARI(\hat{\boldsymbol{s}}, \boldsymbol{s})\right]$ because the latter needs to estimate the knot locations based on a small number of observed datapoints, which introduces some variability into the final classification. As the number of observations per individual increases, we can expect these two models to perform similarly. Although model $M_{random}$ produces

| Dataset | Model | $G_{min}$ | $G_{max}$ | $G_{mode}$ | $\hat{G}$ | $\mathbb{E}_{\boldsymbol{s}}\left[ARI(\hat{\boldsymbol{s}}, \boldsymbol{s})\right]$ | $ARI(\hat{\boldsymbol{s}}, \boldsymbol{s}_{true})$ |
|---|---|---|---|---|---|---|---|
| $D_{fixed}$ | $M_{fixed}$ | 4 | 6 | 4 | 4 | 0.9674 | 0.9734 |
| | $M_{random}$ | 4 | 5 | 4 | 4 | 0.9096 | 0.9606 |
| $D_{random}$ | $M_{fixed}$ | 6 | 9 | 7 | 7 | 0.7096 | 0.6102 |
| | $M_{random}$ | 4 | 5 | 4 | 4 | 0.8245 | 0.7756 |

Table 4.1: Performance summary when fitting fixed and random knot location models ($M_{fixed}$ and $M_{random}$) to fixed and random knot location datasets ($D_{fixed}$ and $D_{random}$). For each dataset/model pair, columns indicate minimum, maximum and mode of the posterior of the number of mixture components ($G_{min}$, $G_{max}$, $G_{mode}$); the number of groups $\hat{G}$ in the optimal clustering $\hat{s}$; the value of the posterior expectation $\mathbb{E}_{\boldsymbol{s}}\left[ARI(\hat{\boldsymbol{s}}, \boldsymbol{s})\right]$) evaluated at $\hat{\boldsymbol{s}}$; and the ARI score comparing the estimated $\hat{\boldsymbol{s}}$ to the true group structure $\boldsymbol{s}_{true}$.

lower agreement with the true clustering $\boldsymbol{s}_{true}$, the realised classification obtained from the optimisation of PEAR is very much comparable to that of model $M_{fixed}$ as shown in Table 4.2: most children are allocated to their respective true groups, with a 1.5% misclassification rate.

In contrast, when modelling the more heterogeneous (and realistic) dataset $D_{random}$, which is more realistic in practice, the fixed knot model performs significantly worse (measured by PEAR) than the random knot model. The $M_{fixed}$ model gives an ARI score which indicates poor concurrence with $\boldsymbol{s}_{true}$. This arises as the fixed knots lead to biased estimates of each child's growth velocities $\boldsymbol{\beta}_i$, which then results in a much larger estimated number of groups as the estimated growth curves are forced to be more dissimilar. Figure 4.3 shows the discrepancy between the estimated $\boldsymbol{\beta}_i$ and their true values. True subgroup 1 has the same average velocity ($-3.0$) across

| | $D_{fixed}$ | | | | | | | | $D_{random}$ | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $M_{fixed}$ | | | | $M_{random}$ | | | | $M_{fixed}$ | | | | | | | $M_{random}$ | | | |
| $\boldsymbol{s}_{true}\backslash\hat{\boldsymbol{s}}$ | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 1 | 2 | 3 | 4 |
| 1 | 99 | 0 | 1 | 0 | 96 | 0 | 4 | 0 | 97 | 0 | 1 | 2 | 0 | 0 | 0 | 96 | 0 | 4 | 0 |
| 2 | 1 | 99 | 0 | 0 | 0 | 100 | 0 | 0 | 10 | 50 | 40 | 0 | 0 | 0 | 0 | 12 | 88 | 0 | 0 |
| 3 | 2 | 0 | 98 | 0 | 0 | 0 | 98 | 2 | 7 | 0 | 0 | 51 | 42 | 0 | 0 | 14 | 0 | 82 | 4 |
| 4 | 0 | 0 | 0 | 100 | 0 | 0 | 0 | 100 | 0 | 0 | 0 | 0 | 0 | 56 | 44 | 0 | 0 | 2 | 98 |

Table 4.2: Contingency table comparing the true group allocations $\boldsymbol{s}_{true}$ to those in the estimated optimal clusterings $\hat{\boldsymbol{s}}$. Results are shown when fitting fixed and random knot location models ($M_{fixed}$ and $M_{random}$) to fixed and random knot location datasets ($D_{fixed}$ and $D_{random}$).

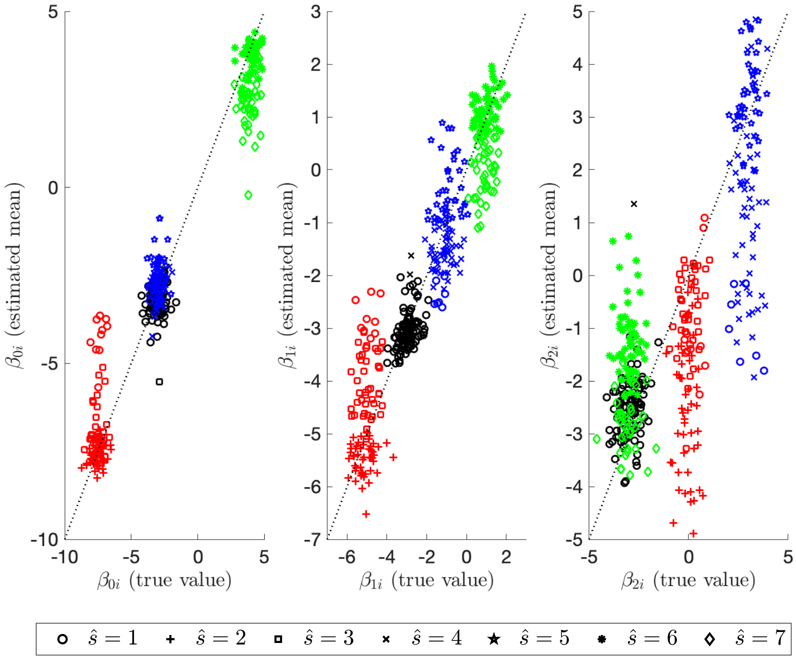



Figure 4.3: Scatterplots comparing the true and estimated mean values of $\boldsymbol{\beta}_i$ when fitting fixed knot location model $M_{fixed}$ to random knot location dataset $D_{random}$. The true group allocations $\boldsymbol{s}_{true}$ are represented by different colours, i.e. $\boldsymbol{s}_{true} = 1$ (black), $\boldsymbol{s}_{true} = 2$ (red), $\boldsymbol{s}_{true} = 3$ (blue) and $\boldsymbol{s}_{true} = 4$ (green), while the estimated optimal clusterings $\hat{\boldsymbol{s}}$ are represented by different symbols.

all broken stick segments. As a result, the location of the knots is not critical, and so any bias in $\boldsymbol{\beta}_i$ for members in this group is relatively small. Therefore, this group can largely be identified correctly under the fixed knot model. This is not the case for the other true clusters: the individuals in these clusters tend to be split into smaller subgroups, and this is largely due to the biases in the estimation of $\beta_{1i}$ and $\beta_{2i}$. Fitting the random knot model $M_{random}$ naturally performs well, as expected. Overall, it is clear that unless the true knot points for any growth curve are known (which will not be the case in practice) and so can be fixed in the model, the random knot location model, which allows for the heterogeneity between each individual child's growth stages, will outperform the fixed knot location model.

## 4.4 Application: Longitudinal birth cohort in India

The Healthy Birth, Growth and Development knowledge integration (HBGDki) project is an initiative supported by the Bill and Melinda Gates Foundation to combine and standardise information from various epidemiological studies into a single knowledge base (Jumbe et al., 2016). The principal objective of this project is to facilitate interdisciplinary collaboration among experts across different fields to gain insights into global child growth and development issues. The life quality of children, particularly those in low to medium income countries, can be greatly improved by the development of appropriate and timely health solutions. To date the project has amassed data sets from 192 studies, involving close to 11.5 million children and spanning 36 countries.

Our focus is on the classification of growth curves for a longitudinal study from the HBGDki project, examining the prevalence of rotavirus infections in a birth cohort in Vellore, India (Paul et al., 2014). The sample population of $N = 373$ children were followed up for three years from birth and had their anthropometric measurements recorded. For the present analysis, we only analyse the HAZ scores from birth to year 1 as this is the period of fastest growth in mental development (Olusanya and Renner, 2013). We remove outliers (HAZ $< -6$ or HAZ $> 6$) based on WHO recommendations (WHO Multicentre Growth Reference Study Group, 2006). This results in 5 to 15 observations for each child, with the first measurement taken between days 1 and 225. The time scale is represented as age in years (between 0 and 1) whereby age 1 is equivalent to day 365, and the number of knot points on the growth curve is specified as $K = 3$. Here, we fix the value of $K = 3$ which provides sufficient flexibility in the shape of the individual growth curves as the number of observed measurements for each child is relatively small. Two broken stick models with mixture distributed random slopes are fitted, one has random change points while the other has fixed knot locations. Further sensitivity analyses using $K > 3$ did not result in noticeably different analysis outcomes (results not shown). Prior distributions on model parameters follow those in Section 4.3. The MCMC algorithm is run for 300 000 iterations, with the first 100 000 iterations discarded as burn in, retaining every 50th of the remaining samples, yielding 4 000 MCMC iterates for cluster analysis.

Figure 4.4 shows empirical growth curves obtained from the estimated groupings of children using the broken stick model with random change points $M_{random}$, and Figure 4.5 shows the associated fitted posterior mean growth curves. The optimal clustering $\hat{\boldsymbol{s}}_{random}$ identifies 9 unique subgroups, which coincides with the posterior mode for the number of groups in the posterior distribution for $G$. Children in the two largest subgroups (1 and 2) show faltering growth where the distinction between them is the slight improvement in the HAZ score from birth in subgroup 1. Subgroup 3 (51 children) experiences severe early stage faltering which persists for approximately 6 months, after which the HAZ scores subsequently improve. This growth pattern is also observed in subgroup 7 (only 4 children), but the changes are milder. Subgroups 4 and 6 (43 and 5 children respectively) each alternate between growth and faltering, with the amplitude of each change differentiating between the two groups. Children in subgroups 5 and 8 exhibit a steep decline in the HAZ score before a short interval of significant recovery is observed, and which is then followed by another onset of faltered growth. The difference between these two subgroups lies in the times at which catch-up growth occurs ($\omega \in [0.25, 0.5]$ for subgroup 5, $\omega \in [0.5, 0.8]$ for subgroup 8). Subgroup 9 consists of a small number of children with severe and continued faltering growth between birth and age 1.

Figure 4.6 shows a sample of empirical growth curves obtained from the estimated groupings of children using the broken stick model with fixed knot locations $M_{fixed}$, and Figure 4.7 shows the corresponding fitted posterior mean growth curves for these children. The optimal clustering $\hat{\boldsymbol{s}}_{fixed}$ now identifies 11 subgroups of children from the birth cohort, but the patterns of growth trajectories obtained are relatively similar to those in $\hat{\boldsymbol{s}}_{random}$. It is unsurprising that the number of distinct subgroups in $\hat{\boldsymbol{s}}_{fixed}$ is higher compared to $\hat{\boldsymbol{s}}_{random}$, which corroborates with the results presented in Section 4.3. The limited flexibility of $M_{fixed}$ in capturing the heterogeneity in the timing of growth phases between children leads to biased estimates of the growth velocities – a key metric that determines the membership classification. This in turn causes the splitting of a larger subgroup into multiple smaller subgroups. Table 4.3

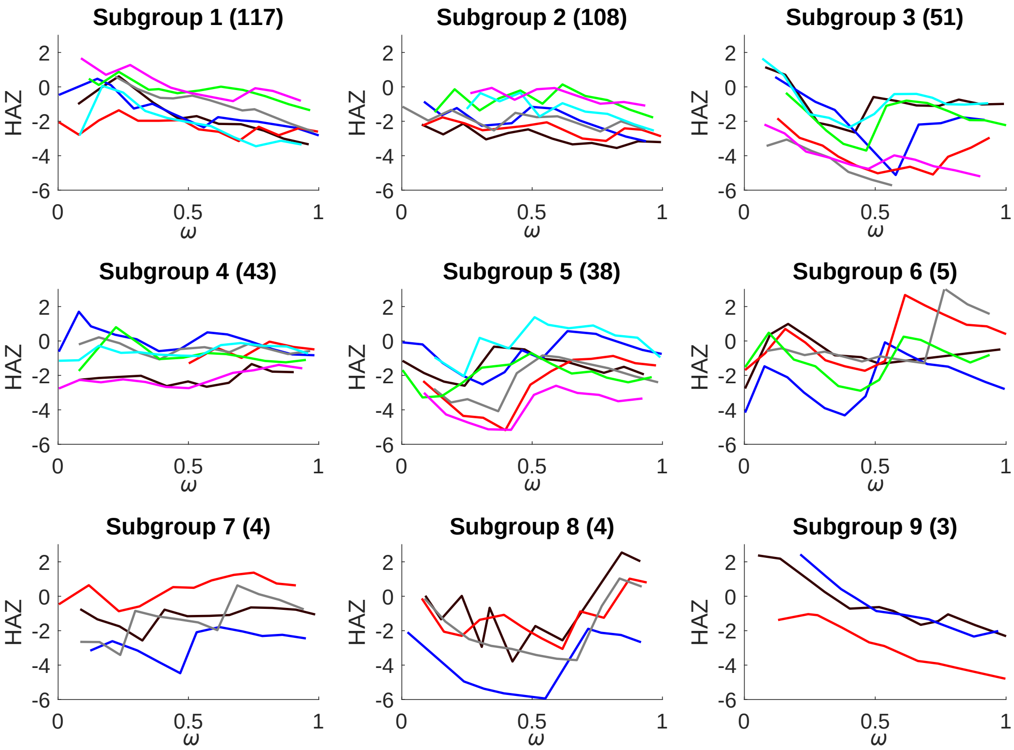



Figure 4.4: Subgroups of children from the Vellore cohort based on the broken stick model with random change points. Individual raw trajectories, obtained by connecting the observations with straight lines, are shown for a sample of children from each subgroup. The number of children in each subgroup is given in parentheses.

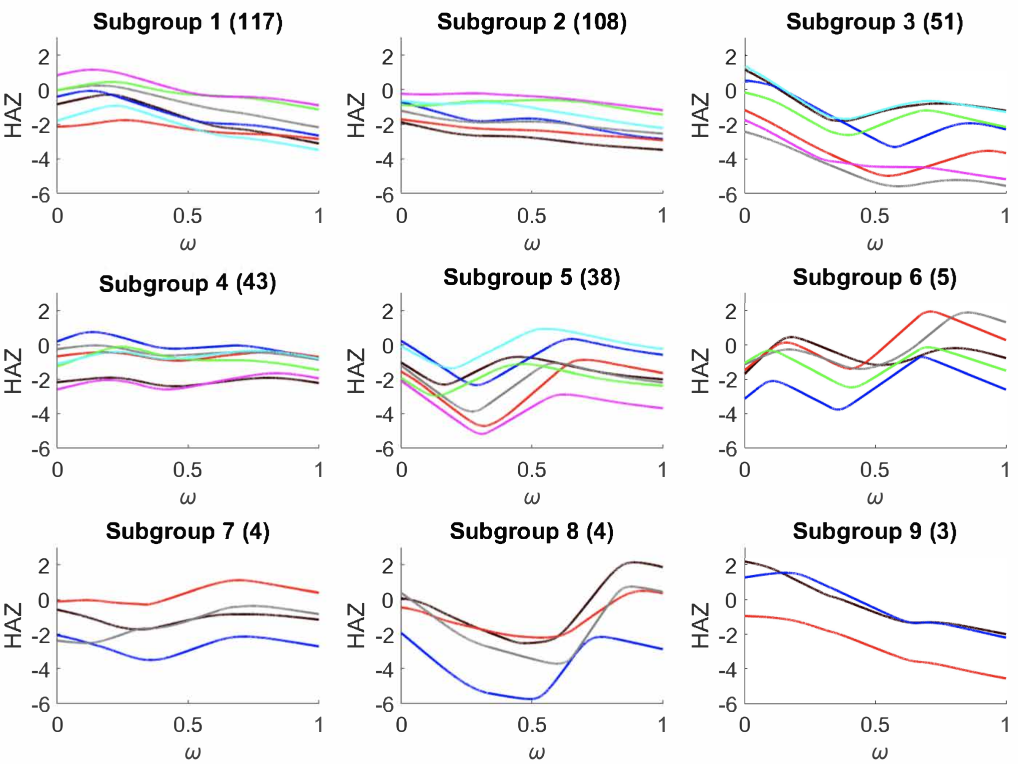



Figure 4.5: Estimated posterior mean trajectories for the same sample and groupings of children in Figure 4.4.

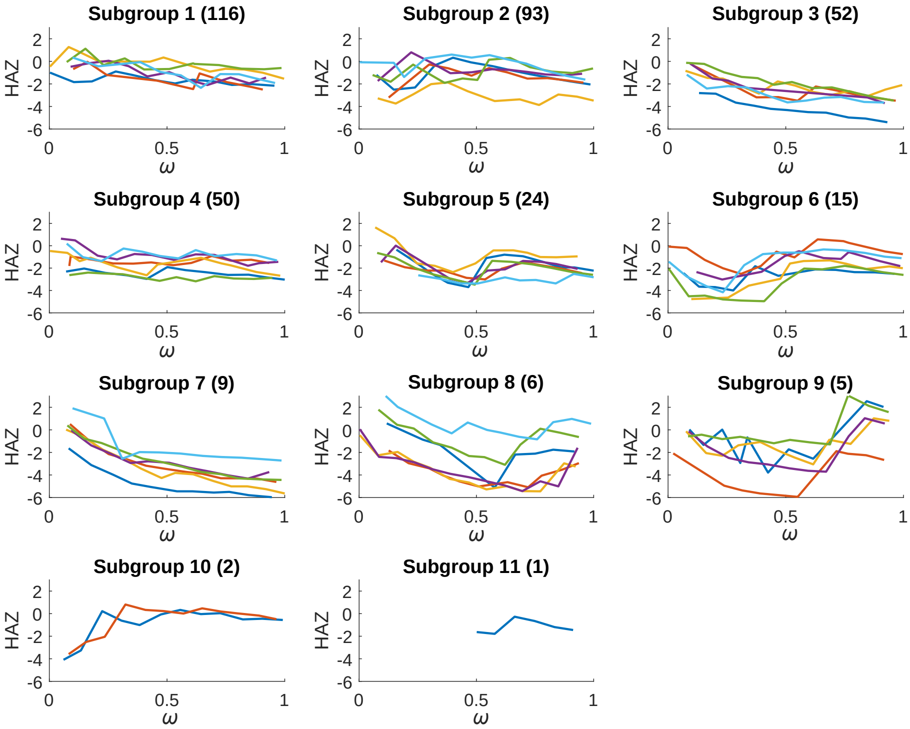



Figure 4.6: Subgroups of children from the Vellore cohort based on the broken stick model with fixed change points. Individual raw trajectories, obtained by connecting the observations with straight lines, are shown for a sample of children from each subgroup. The number of children in each subgroup is given in parentheses.

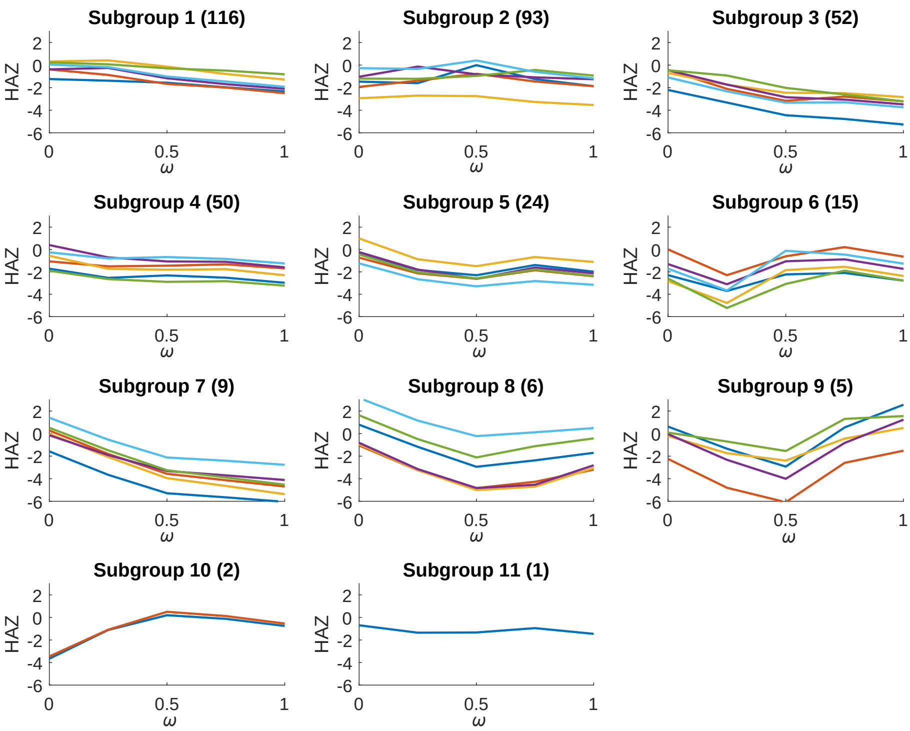



Figure 4.7: Estimated posterior mean trajectories for the same sample and groupings of children in Figure 4.6.

compares the allocated subgroup membership between $\hat{\boldsymbol{s}}_{random}$ and $\hat{\boldsymbol{s}}_{fixed}$. Children from subgroup 2 in $\hat{\boldsymbol{s}}_{random}$, which exhibit a faltering trend with a constant negative slope, are assigned almost equally in numbers between three different subgroups (1, 2 and 4) in $\hat{\boldsymbol{s}}_{fixed}$. Similarly, children from subgroup 5 in $\hat{\boldsymbol{s}}_{random}$ are separated into two distinct subgroups (2 and 6) in $\hat{\boldsymbol{s}}_{fixed}$. While subgroup 3 in $\hat{\boldsymbol{s}}_{random}$ has a more heterogeneous composition of children in terms of their growth curves in the second half of the first year, these children are classified into two different subgroups in $\hat{\boldsymbol{s}}_{fixed}$, i.e. subgroup 3 which shows a steep decline in the HAZ and subgroup 5 which also shows a similar faltering pattern but with a recovery period between $\omega = 0.5$ and $\omega = 0.75$.

Using $\hat{\boldsymbol{s}}_{random}$ as the benchmark due to its parsimony, we conduct a further analysis to explore whether there is any relationship between various covariates recorded on each individual and the classification of children into the subgroups illustrated in Figure 4.4. Figure 4.8 shows the results. In terms of gender composition, subgroup 1 comprises mostly females (59.0%), whereas subgroup 3 has a disproportionately large number of males (68.6%). These subgroups deviate significantly from the composition of the full sample which has approximately equal proportions for each

| $\hat{\boldsymbol{s}}_{fixed} \backslash \hat{\boldsymbol{s}}_{random}$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 78 | 33 | 0 | 1 | 2 | 0 | 0 | 0 | 2 |
| 2 | 10 | 32 | 0 | 27 | 18 | 4 | 2 | 0 | 0 |
| 3 | 27 | 4 | 18 | 3 | 0 | 0 | 0 | 0 | 0 |
| 4 | 2 | 38 | 1 | 5 | 2 | 0 | 2 | 0 | 0 |
| 5 | 0 | 0 | 18 | 6 | 0 | 0 | 0 | 0 | 0 |
| 6 | 0 | 1 | 0 | 0 | 14 | 0 | 0 | 0 | 0 |
| 7 | 0 | 0 | 8 | 0 | 0 | 0 | 0 | 0 | 1 |
| 8 | 0 | 0 | 6 | 0 | 0 | 0 | 0 | 0 | 0 |
| 9 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 4 | 0 |
| 10 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 |
| 11 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |

Table 4.3: Comparison of the estimated optimal clusterings ($\hat{\boldsymbol{s}}_{random}$ and $\hat{\boldsymbol{s}}_{fixed}$) based on the broken stick model with random change points $M_{random}$ and fixed knot location $M_{fixed}$.

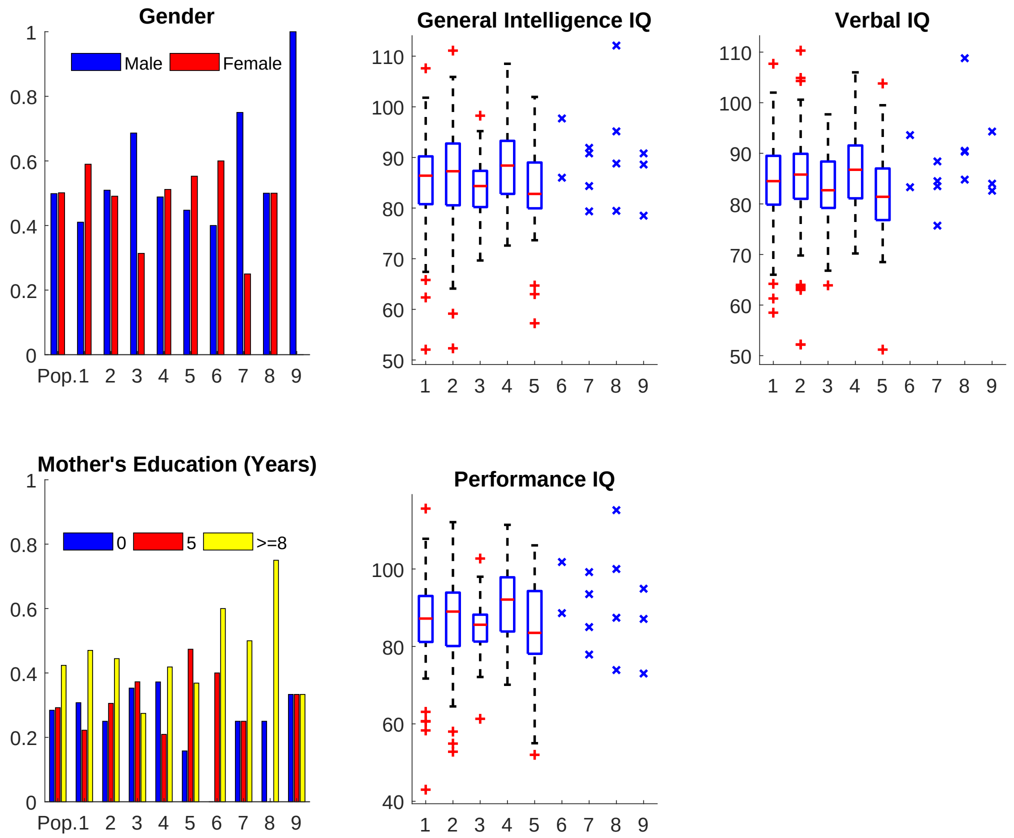



Figure 4.8: Bar charts illustrating the proportion of children in terms of gender and maternal education levels in different subgroups (left panels), and boxplots showing the distributions of IQ scores types (general intelligence, verbal and performance) for children in different subgroups (center and right panels). Raw data ($\times$) for IQ scores are shown for subgroups 6–9 which have a small number of observations. Not all children are represented in each boxplot due to missing data.

gender. Mothers who received no formal education are more likely to give birth to children that exhibit the growth patterns in subgroups 3 and 4. Children are also more likely to experience severe faltering in their early childhood (subgroups 3 and 5) if they are birthed by mothers who completed 5 years (a moderate amount) of education. Moreover, these children have lower IQ scores (general intelligence, performance and verbal) compared to their peers, as indicated by the lower median scores for these tests in subgroups 3 and 5. On the other hand, children in subgroup 4, which exhibits the mildest faltering of all subgroups, have the highest median IQ scores for all tests and this is in line with the results in Emond et al. (2007). For children in the smaller subgroups (i.e. subgroups 6–9), subgroups 7 and 9 are dominated by male children (75% and 100% respectively), while those in subgroups

6 and 8 are mostly borne by mothers who are highly educated. There are no obvious covariate patterns to account for those children who experienced severe faltering in the first year (subgroup 9). However, these conclusions are unreliable due to the small number of children in these subgroups.

## 4.5 Conclusion

This chapter proposes a new model to classify growth patterns in longitudinal child growth studies where the number of classes is not known in advance. We model the evolution of growth in terms of the HAZ scores by piecewise linear segments (i.e. the broken stick model) whereby an individual child's rates of growth are characterised by the slopes of these segments. Accordingly, it is plausible to use these slope parameters as a proxy for the growth pattern and so model their similarities via a mixture distribution. A mixture distribution requires specifying a suitable number of components to prevent over- and under-fitting. It is a common practice in the growth mixture modelling literature (Muthén and Shedden, 1999; Muthén and Muthén, 2000; Magidson and Vermunt, 2004; Nylund et al., 2007) to select the number of mixture components based on model selection information criteria and likelihood ratio test, which requires fitting multiple models of differing complexity. To overcome this issue, a Bayesian non-parametric approach is adopted using the DP prior (Ferguson, 1973), so that the number of mixture components is driven by the complexity of the data and inferred as part of the posterior sampling algorithm.

In order to extend the flexibility of the broken stick model, and ensure that it can be a viable model in practice given the heterogeneity inherent in observed datasets, we incorporated random knot locations into the model. The location of the knots varies between children and follows the even-numbered order statistics distribution in Green (1995) *a priori*. In addition, we impose a structural restriction which ensures that there is a knot within each of several equally divided segments of the observational period. This is because we regard two growth curves, where one has the same shape as the other but lags by one period, as being different.

Our simulation studies suggest that overall the random knot point model performs well: the fixed knot points model overestimates the number of components when the true data generating process has random change points due to the resulting biased estimation of the velocity vectors.

Our methodology is applied to a longitudinal study of birth cohort in Vellore, India from the Healthy Birth, Growth and Development knowledge integration (HBGDki) project funded by the Bill and Melinda Gates Foundation. Analysis of the posterior distribution indicated that there are 9 different types of growth profiles in the population, with a majority exhibiting improved growth followed by a faltering trend. We note that the granularity of the classification can be increased if we are willing to impose stronger assumptions in the model, for example by having a shared covariance matrix across all subgroups, or by restricting the covariance matrices to be diagonal. More sophisticated analyses could treat the number of knots in the broken stick model as unknown with some prior specification, which could be implemented via reversible-jump style MCMC algorithms (Green, 1995; Sisson, 2005).

# Chapter 5

# Modelling age-related changes in executive functions of soccer players

## 5.1 Introduction

Sports activity has long been an integral part of everyday life, and is synonymous with a healthy lifestyle due to well-established relationships between an adequate level of physical exercise and physiological benefits such as reduced risk of developing coronary heart disease (Bassuk and Manson, 2005; Sofi et al., 2008) and improved cognitive performance (Kramer and Erickson, 2007; Hillman et al., 2008). One of the most popular sports in the world is soccer (also known as football), which is evident from the over 3.5 billion total viewerships (Fédération Internationale de Football Association (FIFA), 2018) it garnered for the quadrennial World Cup tournament in 2018. Due to its ubiquitous global presence, the sport has grown into a multi-billion dollar industry over the last few decades, with soccer clubs investing heavily in talent identification programmes guided by measures of executive functioning. Formally, executive functions are cognitive processes that facilitate decision-making (Best and Miller, 2010; Furley and Wood, 2016) and goal-oriented behaviours (Zelazo et al., 2004) based on information within the context of the task (Alvarez and Emory, 2006).

Some examples of these processes are problem-solving, sustained attention, resistance to interference, utilisation of feedback and multi-tasking (Grafman and Litvan, 1999; Burgess et al., 2000; Chan et al., 2008).

The proposition advocating for the use of executive functions testing within the talent identification process is based on considerable amount of evidence indicating that neuropsychological evaluations provide a useful indicator for performance success in young soccer player. For example, Verburgh et al. (2014) revealed that elite players show superior motor inhibition as measured by a stop signal task compared to amateur players of the same age, and this finding has been further reaffirmed in other similar experiments (Huijgen et al., 2015; Sakamoto et al., 2018). However, longitudinal studies of the developmental trajectories of executive functions across different stages of life in an athlete population are lacking in the literature as previous examinations focus on a general population (Zelazo et al., 2004; Huizinga and Smidts, 2010; Zelazo and Carlson, 2012). Since Jacobson and Matthaeus (2014) argued that executive functions can be improved by active participation in sports, the generalisation of existing results to an athlete population is therefore limited. Nevertheless, general studies conducted so far establish that executive functions are attributed to the frontal lobes of the brain (Stuss and Alexander, 2000) which undergo protracted maturation from childhood to early adulthood (Lebel et al., 2008; Taylor et al., 2013), and this is followed by cognitive declines (Dempster, 1992; Jurado and Rosselli, 2007) due to the attrition of dendrites during the ageing process. These developments give rise to an inverted U-shaped executive functions trajectory across the lifespan of an individual (Kail and Salthouse, 1994; Cepeda et al., 2001; Zelazo et al., 2004). Huizinga et al. (2006) and Zelazo and Müller (2011) documented that improvements in executive functions occur the most rapidly from late childhood into adolescence. In fact, Diamond (2002) reported that children between the age of 12 and 15 years old attain adult levels of performance in neuropsychological assessments.

This chapter analyses the age-related architecture of executive functions in a sample of male elite soccer players representing a professional German club, and compares our findings with existing theories. A gender- and sport-specific study is necessary here as evidence has shown that the impact of athletic participation varies significantly between gender (Habacha et al., 2014) and sport type (Krenn et al., 2018). Particular emphasis is given to examining the developmental changes in executive functions trajectories between late childhood (10–12 years old) and early adulthood (18–21 years old) since these periods are of relevance to talent identification and development. Contrary to previous investigations in the sport science literature which are based on cross-sectional data (see e.g. Verburgh et al., 2014; Huijgen et al., 2015; Sakamoto et al., 2018), our analysis uses longitudinal cognitive data collected from a test battery of soccer and non-soccer related neuropsychological assessments performed by the players over a period of three years. These outcomes are modelled using a latent variable model (Dunson, 2000; Muthén, 2002; Proust et al., 2006), which relates speed and accuracy observables to latent variables representing executive functions. Furthermore, we make a distinction between domain-generic and domain-specific executive functions, in accordance with the two-component intellectual development model introduced by Li et al. (2004). Indeed, Furley and Wood (2016) recommended longitudinal studies that assess both domain-generic and domain-specific executive functions jointly, rather than examining them independently. This is done by modelling parameters of the underlying latent variables using a multivariate formulation.

This chapter is organised as follows. Section 5.2 describes the data and the neuropsychological assessments performed by the players. Section 5.3 describes the latent variable model used in our analysis. Section 5.4 presents the results, and Section 5.5 concludes.

## 5.2 Background of study

The study collects the outcome variables from a test battery of neuropsychological assessments undergone by elite soccer players. The assessments measuring executive functions used in the study include a determination test to measure general perceptual abilities, a response inhibition test to measure motor impulsivity, a pre-cued choice response time task to measure vigilance under interference, a Helix test to measure soccer-specific perceptual abilities and a Footbonaut test to measure soccer-specific technical abilities. Figure 5.1 graphically illustrates these assessments, while the following subsections give further details on each assessment.

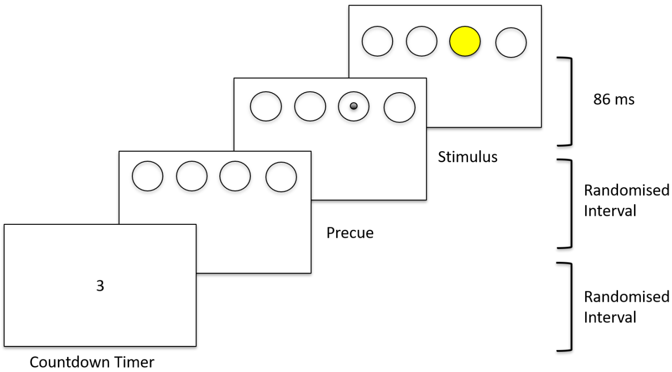
### 5.2.1 Determination test

The determination test (Schuhfried GmbH, Mödling, Austria) is a multi-stimuli assessment used in sports such as motor racing (Baur et al., 2006) to analyse the perceptual-motor abilities of a participant. The participant is presented with a combination of two audio tones (2 000 Hz and 100 Hz) and five coloured signals on a computer screen in the assessment, to which the participant must react by choosing the appropriate buttons on the keyboard panels via hand and foot responses. The number of correct answers to the stimuli within the four-minute assessment session and the median response time based on correct responses are recorded. Since the participants are required to respond correctly to as many stimuli as possible within the stipulated time, the assessment provides a good evaluation of the reactive stress tolerance and receptivity of the participants. The validity and reliability of the determination test in measuring executive functions have been confirmed by several studies (Whiteside et al., 2003; Ljac et al., 2012; Beavan et al., 2019).

### 5.2.2 Response inhibition test

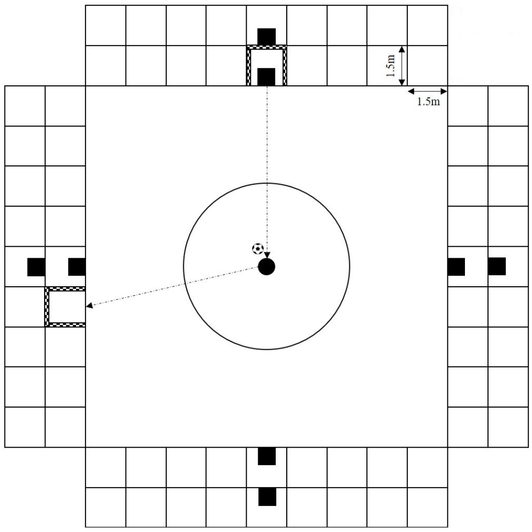
The response inhibition test (Schuhfried GmbH, Mödling, Austria) uses a stop signal paradigm, and has been widely established to provide a good assessment of impulse control (see e.g. Alderson et al., 2007; Zhong et al., 2014; Cox et al., 2016). The assessment is made up of 100 trials, with each presenting a left- or right-pointing arrow to which the participant must respond by pressing the corresponding button on the keyboard panel. Each signal is displayed on the computer screen for a period of one second, followed by a one-second lapse before the subsequent signal appears.

(a) Determination test.




(b) Pre-cued choice response time task.


(c) Helix test.




(d) Footbonaut.

Figure 5.1: Graphical illustrations for some of the neuropsychological assessments used in the study. (a) The test equipment used for the determination test in which a participant is required to respond to different types of stimuli by pressing the appropriate buttons on the keyboard panel and foot pedal. The response inhibition test uses the same equipment, but with a simpler keyboard design and without the foot pedal. (b) A congruent trial (the pre-cue appears in the same circle as the stimulus) in the pre-cued choice reaction time task. (c) A participant in the midst of identifying the players whom he is assigned to track in the Helix test while being monitored by a staff member. (d) A design plan of the Footbonaut in which a participant (a solid circle) is required to pass the soccer ball from one of the dispenser gates (square panels with a solid square within) to the illuminated target gate (an empty square panel partially surrounded by a striped band.)

There are two types of signals: go signal (around three quarters of the total trials) which requires instant reaction from the participant, and stop signal (usually appears after a go signal) to which the participant must refrain themselves from responding. The stop signal is indicated by a tone at a pitch of 1 000 Hz for 100 milliseconds. The difficulty of the assessment is adaptive, i.e. a correct response to a stop signal will increase the visual-audio delay between the appearance of the arrow and the tone in the next stop trial by 50 milliseconds (up to a maximum of 350 milliseconds), and vice versa. The variable of interest that reflects the ability to override prepotent actions (Logan, 1994) is the stop signal reaction time (SSRT), which is calculated by deducting the mean stop signal delay from the mean reaction time. Measurements recorded are the SSRT, the mean reaction time and the number of correct responses.

### 5.2.3 Pre-cued choice response time task

In a pre-cued choice response time task, the participant is required to press the correct button on a joystick panel as fast as possible in response to a visual stimulus. Four blank circles are arranged side by side and presented on the screen after a three-second countdown timer is shown. One of them turns yellow after a randomised interval of between 2 to 4 seconds. Prior to the colour change, a small dot appears in the center of one circle. A total of 24 trials are conducted in the assessment, half of which has the dot in the same circle that turns yellow (congruent trials; see Figure 5.1b), and the rest in a different circle (incongruent trials). The mean response time for correct answers is recorded. This task has been previously used to assess psycho-motor vigilance under interference in both general population (Barela et al., 2019) and professional athletes (Beavan et al., 2019), thereby validating its use as a measure of executive functions.

### 5.2.4 Helix test

The Helix test (SAP SE, Walldorf, Germany) is designed to train various aspects of a participant's perceptual abilities including sustained attention, decision-making, multiple object tracking, and peripheral vision using a soccer-relevant stimulus. A participant stands facing a 180-degree curved screen where a soccer stadium is reproduced. Eight soccer players, with equal representation from two teams, are depicted as human-size avatars differing in the jersey colour, but otherwise having the same physical features and jersey number. During the assessment, the animated avatars run randomly across the virtual pitch away from the participant for eight seconds before lining up across the screen in a different order to the initial line-up. The participant must then identify four of the avatars that he is assigned to track before the start of the trial (see Figure 5.1c). Ten trials are conducted and each correct identification earns a point to a maximum score of 40 points. Although the Helix test is a newly developed assessment tool, Beavan et al. (2020) found that it distinguishes executive functions of professional soccer players based on their playing experience.

### 5.2.5 Footbonaut test

The Footbonaut (CGoal GmbH, Berlin, Germany) is an innovative tool which aims to measure a participant's soccer-specific skill performance such as dribbling, passing and shooting (Beavan et al., 2019), as well as perceptual-motor abilities. The assessment system is made up of a square artificial turf surrounded by four walls. Each wall contains 18 square panels arranged side by side in two rows, where the two panels in the middle serve as the ball dispenser gates while the rest are the target gates. All gates measure 1.5 meters $\times$ 1.5 meters in dimension and are fitted with light barriers and light-emitting diodes (LEDs). During the assessment, a ball is dispensed from one of the eight possible dispenser gates at a speed of 50 kilometers per hour. Immediately before the dispense of the ball, the LEDs along the perimeter of the gate light up and an audio signal is given to the participant. This is followed by the same stimuli 0.8 seconds later from the target gate, to which the participant is required to pass the ball. Thirty-two trials are conducted and the mean reaction time for successful passes that enter the target gate is measured using the light barriers.

Saal et al. (2018) investigated the validity of the passing test in the Footbonaut, and concluded that it offers a reliable method to differentiate between skilled and less-skilled soccer players.

### 5.2.6 Description of data

Table 5.1 summarises the measurement variables for each of the neuropsychological assessments that were collected on 304 male soccer players, aged between 10 and 21 years old, who represented a professional German club in the Bundesliga. Repeated measurements on the players were recorded over a study period of three years from the 2016–17 season to the 2018–19 season, whereby an assessment session was conducted twice in a year – pre-season (between July and August) and post-season (between January and February). The number of assessment sessions that an individual participated in varied, due to player mobility between soccer clubs through the transfer market or player dropout from the soccer academy. Additionally, the pre-cued choice response time task was only integrated into the test battery from the start of the 2017–18 season, whereas the Helix test was excluded from the test battery throughout the entire 2018–19 season. These changes in the test battery setup resulted in an increased proportion of missing data, as well as a reduction in the mean number of observations per player for measurements under both of these

| Neuropsychological assessment | Variable | Outcome | Mean number of observations per player | Proportion of missing observations* |
|---|---|---|---|---|
| Determination | $y_1$ | Number of correct answers | 2.98 | 0.03 |
| | $y_2$ | log median response time | 2.98 | 0.03 |
| Response inhibition | $y_3$ | log SSRT | 2.98 | 0.03 |
| | $y_4$ | log mean response time | 2.87 | 0.07 |
| | $y_5$ | Number of correct answers | 2.98 | 0.03 |
| Choice response | $y_6$ | log mean response time (congruent) | 1.61 | 0.48 |
| | $y_7$ | log mean response time (incongruent) | 1.61 | 0.48 |
| Helix | $y_8$ | Number of correct answers | 1.73 | 0.44 |
| Footbonaut | $y_9$ | Number of correct answers | 2.39 | 0.22 |
| | $y_{10}$ | log mean response time | 2.39 | 0.22 |

* Complete data have one observation per participated assessment session.

Table 5.1: The mean number of observations per player and the proportion of missing observations for each outcome variable of the neuropsychological assessments in the executive functions test battery.

assessments. More missing data were also present in the measured outcomes as a result of data mismanagement.

Table 5.1 shows the variables collected, which measure players' performance in accuracy and speed. Figure 5.2 illustrates the pairwise dependence structure between these variables by computing their Spearman's rho correlation coefficients using complete observations from the 2017–18 pre-season assessment session (the only test battery that contains all five assessments). Although the assessments are very dissimilar in terms of their design and domain of cognitive abilities evaluated, the speed components $(y_2, y_3, y_4, y_6, y_7, y_{10})$ are noticeably strongly positively correlated among themselves. This suggests that players' speed is measured relatively consistently across different assessments. In contrast, the negative dependence observed in the speed-accuracy pairs from the determination test $(y_1, y_2)$ and the Footbonaut test $(y_9, y_{10})$ implies that higher-scoring players also tend to have shorter response time. This trend is in contrast to the speed-accuracy trade-off (longer response time for

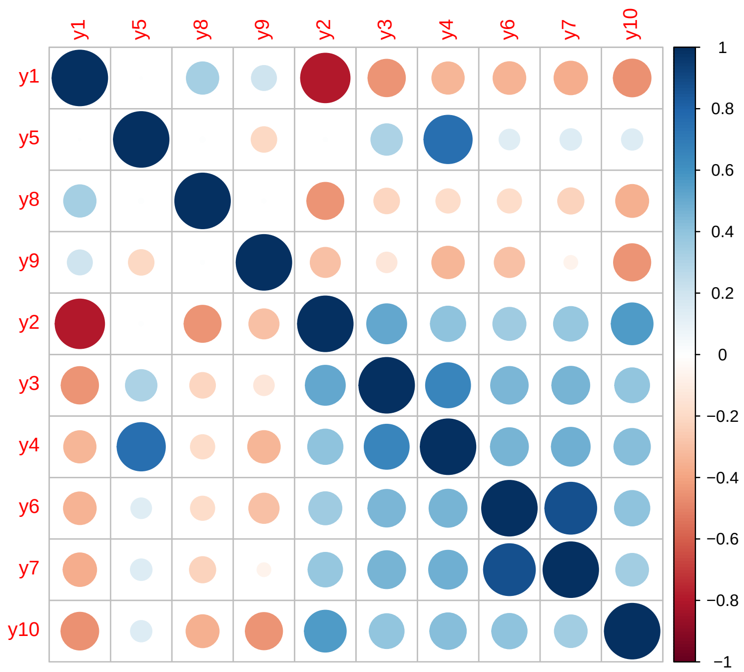



Figure 5.2: Spearman's rho correlation coefficients between the measurement variables collected from the 2017–18 pre-season assessment session. Circle size is proportional to correlation magnitude, with darker blue/red indicating stronger positive/negative correlation. Variables are ordered such that the first four $(y_1, y_5, y_8, y_9)$ report accuracy components while the remainder $(y_2, y_3, y_4, y_6, y_7, y_{10})$ report speed components.

increased accuracy) documented among soccer players in the experimental study conducted in Andersen and Dörge (2011).

Each participating player can be grouped according to their playing position in the team, either as forward, midfielder, defender or goalkeeper. According to the developmental model of sport participation in Côté et al. (2007), players' age range can be divided into four age groups that represent different developmental stages of executive functions: 10–12 years old (late childhood), 12–15 years old (pre-adolescence), 15–18 years old (adolescence) and 18–21 years old (early adulthood). Figure 5.3 shows the distribution of players by age group and playing position in each assessment session. More than two-thirds of players are aged between 12 and 18 years old when they participate in the test battery, whereas there are only at most 20 players in the youngest age group (10–12 years old). In terms of playing position, most players are either midfielders or defenders. Unsurprisingly, the goalkeeper category has the fewest number of players due to the composition of a soccer team which comprises ten field players and only one goalkeeper.

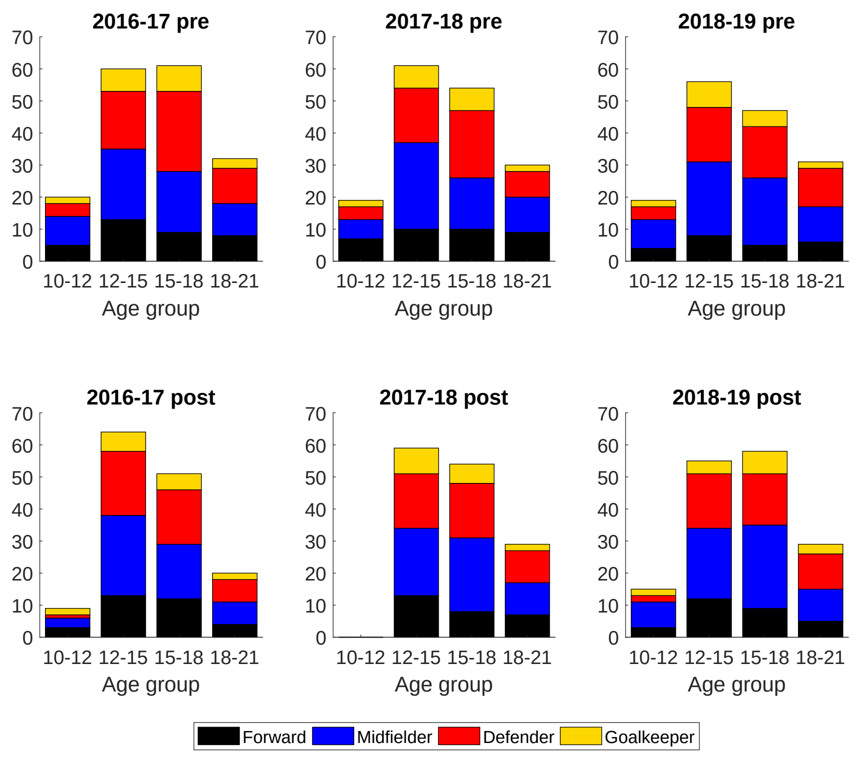
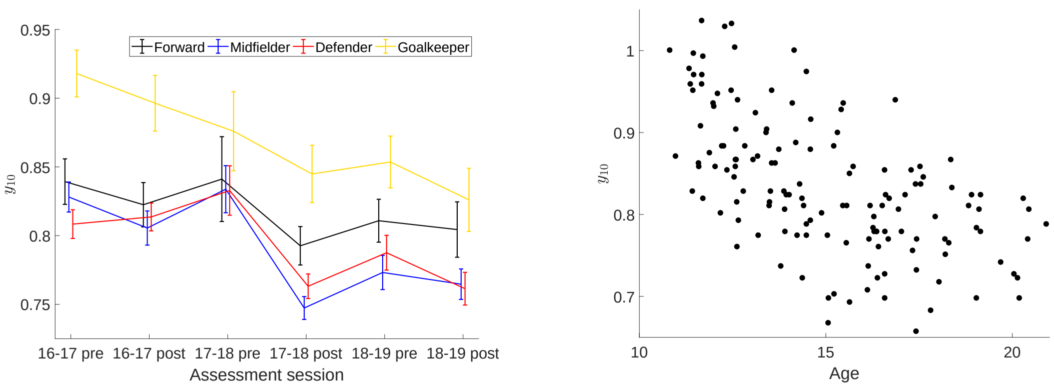



Figure 5.3: Bar charts showing the distribution of players by age group and playing position across the 3-year, pre- and post-season study period.

Using the covariates available (assessment session and playing position), we perform an exploratory analysis to examine whether there are systematic differences in cognitive abilities between different groups of players. We demonstrate the general underlying pattern with the example of log response time measurement $y_{10}$ from the Footbonaut test. Figure 5.4a shows the changes in mean value of $y_{10}$ grouped by playing position across the study period. The goalkeepers consistently improve on their response time in the first four sessions, whereas the other field players' performance fluctuates considerably. Goalkeepers also take the longest time to react to the stimuli in the Footbonaut test. Visually, and on average, players react faster in post-season assessment sessions within the same season, with a possible exception of defenders in the 2016–17 season. This improvement in post-season cognitive functioning can be attributed to the effect of active athletic participation (Jacobson and Matthaeus, 2014). Overall, there is a general improvement in the response time as the study progresses. However, it must be recognised that this positive shift in performance level is likely to be confounded by the age of the players.

Figure 5.4b explores this age effect by showing differences in log response time as a function of age, based on observations from the 2016–17 pre-season assessment session (observations from other sessions produced similar graphs). The clear




(a) Changes in the grouped mean log response time measurement from the Footbonaut test across the study period. Error bars indicate two standard errors of the mean.

(b) Scatterplot showing the relationship between the log response time measurement from the Footbonaut test and the player age, in the 2016–17 pre-season assessment session.

Figure 5.4: Exploratory data analysis to examine performance variation between players that is due to assessment session, playing position and age.

negative correlation underlines that performance heterogeneity between players is age-dependent. Response time decreases significantly between ages 10 and 15, with a reduced rate of improvement as players age further. The age effect also offers possible explanation of two anomalies observed in Figure 5.4a: (i) post-season deterioration in the defenders performance in the 2016–17 season and, (ii) a sharp reduction in post-season response time for all players in the 2017–18 season. For the former, the session effect is offset because there are fewer defenders in the older 15–18 and 18–21 age groups (see Figure 5.3) who tend to perform better and are able to improve the group mean. For the latter, the age effect is amplified as there is no player below 12 years old (see Figure 5.3), resulting in the mean response time values being significantly lower.

## 5.3 Methods

Formal statistical modelling of the evolution of the players' executive functions requires consideration of the longitudinal nature of the study, as well as the fact that cognitive performance is determined by a battery of different but correlated neuropsychological assessments. Analysis of test outcomes is typically carried out using a multivariate linear mixed effects model (e.g. Hall et al., 2001; Sliwinski et al., 2003) whereby a unique mean latent growth trajectory that is representative of the entire population is estimated for each measured outcome. However, Salthouse et al. (1996) establish in an experimental study that age-related differences in the observed psychometric outcomes are not exclusively attributable to assessment-specific cognitive processes, but instead are the manifestation of a common cognitive factor. Furthermore, the two-component characterisation of lifespan intellectual development (Li et al., 2004) distinguishes between two distinct facets of cognitive processes, namely fluid and crystallised abilities. The fluid abilities refer to biological and genetically pre-disposed intelligence in information processing, whereas the crystallised abilities relate to the normative and pragmatic aspects of expertise acquired through contextualised personal experiences and socio-cultural influences.

In the current context, the determination test, the response inhibition test and the pre-cued choice response time task examine general intelligence, and thus can be categorised as assessments that measure fluid (domain-generic) executive functions. In contrast, the Helix and the Footbonaut are specialised test systems designed to evaluate soccer-related skills that are developed through active participation in the sport, thereby demonstrating their roles in measuring crystallised (domain-specific) executive functions. Using this two-class grouping of neuropsychological assessments, we introduce two latent curves, one shared between domain-generic variables $(y_1, \ldots, y_7)$ and another shared between domain-specific variables $(y_8, y_9, y_{10})$, within a latent variable model (Dunson, 2000; Muthén, 2002; Proust et al., 2006). In the following we describe a latent process representing age-related changes in executive functions in a structural model. This model is then related to recorded outcomes of the corresponding neuropsychological assessments through a measurement model.

### 5.3.1 The structural model

We model the latent curve underlying either of the two facets of executive functions (domain-specific and domain-generic) according to a piecewise linear spline model, which is a commonly used method in the epidemiological literature to model longitudinal growth curves (e.g. Werner and Bodin, 2006; De Kroon et al., 2011; Anderson et al., 2019; Chin et al., 2019). In the current setting, this models $\zeta_i \in \mathbb{R}$, the unobserved executive functions for player $i$ at age $\omega_{it}$, as

$$\zeta_i(\omega_{it}) = \beta_{0i}(\omega_{it} - (\omega_{it} - \xi_1)_+) + \beta_{Ki}(\omega_{it} - \xi_K)_+ + \sum_{k=1}^{K-1} \beta_{ki}((\omega_{it} - \xi_k)_+ - (\omega_{it} - \xi_{k+1})_+) + e_{it}, \tag{5.1}$$

$$\boldsymbol{\beta}_i = (\beta_{0i}, \ldots, \beta_{Ki})^\top \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}), \tag{5.2}$$

for $i = 1, \ldots, 304$, where $(a)_+ = \max\{0, a\}$ is the positive part of $a$, and $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_K)^\top$ is an ordered vector of $K$ knots such that $\xi_1 < \ldots < \xi_K$. Here, $t$ measures time on the scale of the age of each individual, so that e.g. $\zeta_i(\omega_{it})$ with

$\omega_{it} = 10$ refers to executive functions of individual $i$ when the individual is 10 years old. In this way, (5.1) models the evolution of a player's executive functions over time (age) by $K+1$ piecewise linear segments with breaks at $\boldsymbol{\xi}$. Greater flexibility can be achieved by introducing player-specific random change points, but here we fix $\boldsymbol{\xi}$ due to the sparsity of the data. The vector of random slope coefficients $\boldsymbol{\beta}_i = (\beta_{0i}, \ldots, \beta_{Ki})^\top$, where $\beta_{ki}$ represents the rate of change in $\zeta_i$ between knots $\xi_k$ and $\xi_{k+1}$, allows for the heterogeneity between players that is due to unobservables such as the number of training hours and familiarity with the test. By assuming a Gaussian distribution on the vector of random effects $\boldsymbol{\beta}_i$ in (5.2), we postulate that the mean rate of developmental growth in executive functions in the population is $\boldsymbol{\mu}_{\boldsymbol{\beta}}$, and that the variability of player-specific deviations from this global trend is characterised by $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$.

### 5.3.2 The measurement model

Write $y_{d,it}$ as the observed test outcome $d$ for player $i$, on the $t$-th measurement occasion, where $i = 1, \ldots, 304$, $t = 1, \ldots, T_i$ and $d = 1, \ldots, 10$. Assume for the moment that each $y_{d,it}$ is a continuous measurement (this is relaxed below). Using a latent variable model (e.g. Dunson, 2000; Muthén, 2002; Proust et al., 2006), we link the unobserved executive functions to the outcome variable by

$$y_{d,it} = \alpha_d + \boldsymbol{x}_{it}^\top \boldsymbol{\gamma}_d + c_{d\ell}\zeta_{i\ell}(\omega_{it}) + \epsilon_{d,it}, \quad \boldsymbol{\epsilon}_{it} = (\epsilon_{1,it}, \ldots, \epsilon_{D,it})^\top \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}), \quad (5.3)$$

where $\alpha_d$ is an intercept for test outcome $d$ and $\boldsymbol{x}_{it}$ is a vector of time-dependent player-specific covariates associated with a vector of fixed effects $\boldsymbol{\gamma}_d$ for test $d$. The recorded covariates for this study include a player's playing position (forward, midfield, defence or goalkeeper) and an indicator for post-season assessment measurements. In this way, the covariates can vary in their effects on each test outcome $d$. The error term $\epsilon_{d,it}$ is assumed to be uncorrelated with the exogenous variables $\boldsymbol{x}_{it}$ and the latent executive functions $\zeta_i$. To incorporate the distinction between the two facets of cognitive processes, we introduce an additional index $\ell \in \{1, 2\}$ on executive functions in (5.3) such that $\ell = 1$ when $d = 1, \ldots, 7$ refers to domain-generic functions, and $\ell = 2$ when $d = 8, \ldots, 10$ refers to domain-specific functions. As a result, the term $\alpha_d + c_{d\ell}\zeta_{i\ell}(\omega_{it})$ stipulates a test-level linear rescaling (to account for scale differences in the outcome variables) of either the domain-generic or domain-specific executive functions, which is itself an individual-specific random effect around a population mean as detailed in Section 5.3.1. In order for all model parameters to be identifiable, one of the $c_{d\ell} = 1$ for each value of $\ell$, i.e. $c_{d1} = 1$ for one $d \in \{1, \ldots, 7\}$ and $c_{d2} = 1$ for one $d \in \{8, \ldots, 10\}$, and $e_{it}$ is restricted to a standard normal distribution $\mathcal{N}(0, 1)$. Typically, the value of $d$ for which $c_{d\ell} = 1$ is chosen to be the measurement with the largest scale so that the magnitude of $c_{d\ell}$ is less than 1 for other measurements.

Some of the measured outcomes are count variables, e.g. the number of correct answers in the determination test. As a result, the assumption of normality on the errors for these outcomes in (5.3) may be unsuitable. To account for this, we follow Gelman et al. (2013) and transform the count variables into continuous outcomes using the Gaussian kernel. In particular,

$$y_{d,it} = h(y^*_{d,it}), \quad d \in \{1, 5, 8, 9\},$$

where $h(\cdot)$ is a rounding function such that $h(y^*) = p$ if $y^* \in (a_p, a_{p+1}]$ for $p = 0, \ldots, \infty$, $a_0 = -\infty$, and $a_p = p - 1, p = 1, \ldots, \infty$. The latent continuous variable $y^*_{d,it}$ is then modelled following the measurement model in (5.3). As mentioned in Section 5.2.6, the non-availability of certain neuropsychological assessments and data management practices have resulted in the outcome vectors $\boldsymbol{y}_{it} = (y_{1,it}, \ldots, y_{D,it})^\top$ being partially observed. We overcome the missingness by fitting the model assuming full data, and the missing values are sampled from their full conditional distributions (which are the posterior predictive distributions) in the Bayesian sampling scheme.

## 5.4 Analysis and results

We analyse the elite soccer player performance data through the above model in the Bayesian framework, implemented via Markov chain Monte Carlo (Robert and Casella, 2004). Following Côté et al. (2007), we consider the development of executive functions in the four different stages of growth: late childhood (10–12 years old), pre-adolescence (12–15 years old), adolescence (15–18 years old) and early adulthood (18–21 years old). Accordingly, the three knot locations ($K = 3$) in the piecewise linear spline model in (5.1) are specified as $\boldsymbol{\xi} = (12, 15, 18)^\top$. For prior distributions, we specify a horseshoe prior (Makalic and Schmidt, 2016) on $\boldsymbol{\mu}_{\boldsymbol{\beta}}$ and $\boldsymbol{\gamma}_d, d = 1, \ldots, 10$, which is designed to have concentration at zero and to shrink small coefficients towards zero, while having heavy tails to avoid over-shrinkage of larger coefficients, a hierarchical inverse-Wishart prior (Huang and Wand, 2013) with 2 degrees of freedom and scale parameter 25 on $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ to induce a sparse structure on the partial correlation matrices (Chin et al., 2020), and a standard diffuse $\mathcal{N}(0, 10^3)$ prior on each $\alpha_d$.

For parameter identifiability, we set the scale coefficients for the number of correct answers in the determination test and the Helix test equal to one (i.e. $c_{11} = c_{81} = 1$). These were chosen as $y_1$ and $y_8$ have the largest scales among the variables measuring the two facets of executive functions, and so this ensures that each of the other $c_{d\ell}$ will typically scale around or less than 1 in magnitude. We expect a priori that assessment tasks associated with the same type of assessment type (speed or accuracy) are likely to be positively correlated, but that tasks are likely negatively correlated between these two groups. This is largely in evidence in Figure 5.2. Since both $y_1$ and $y_8$ are accuracy components, we specify an informative $\mathcal{N}(0.5, 0.25)$ prior on $c_{d\ell}$ if measurement $d$ also relates to an accuracy component of the assessment $(y_5, y_9)$ to express the prior belief that $c_{d\ell}$ is likely to be a value between 0 and 1. Conversely, a $\mathcal{N}(-0.5, 0.25)$ prior is used if the measurement relates to log speed $(y_2, y_3, y_4, y_6, y_7, y_{10})$.

Figure 5.5 shows the estimated mean trajectories of domain-generic and domain-specific executive functions for a chosen sample of players, and compares them to the population mean whose 95% highest posterior density (HPD) credible levels are represented by shaded regions. For simplicity, these trajectories are based on the number of correct answers scored in the determination test (domain-generic) and the Helix test (domain-specific) whereby the scale factor $c_{d\ell}$ is assumed a fixed value of 1. Our estimation results indicate that changes in executive functions of the elite soccer player population occur mainly between 10 to 15 years old since the magnitudes of the slopes within this age range are the largest. In particular, the most rapid increase happens during late childhood (10–12 years old) for domain-generic executive functions (an average rate of 28.59, which is marginally higher than the value of 24.86 in the next period), whereas the most rapid increase happens during pre-adolescence (12–15 years old) for domain-specific executive functions (an average rate of 0.95, which is nearly twice as large as the value of 0.48 in late childhood). Domain-generic executive functions continue to develop, albeit at a much slower

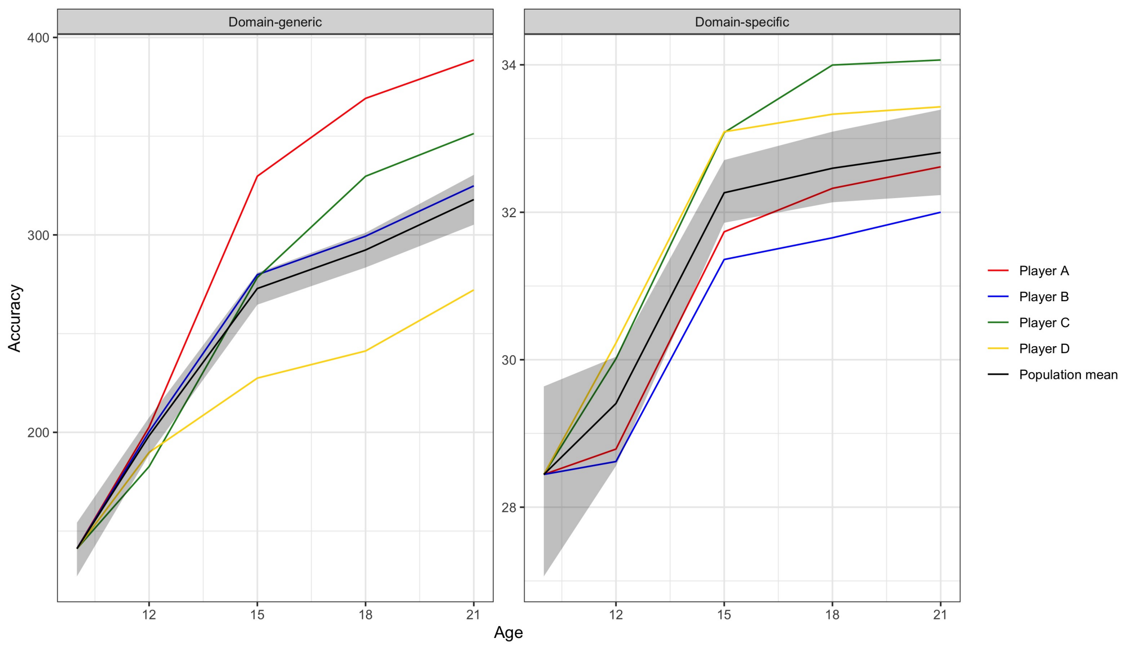



Figure 5.5: Domain-generic (left) and domain-specific (right) executive functions for a sample of players plotted against the posterior mean trajectories of the population, based on the accuracy of the determination test and the Helix test respectively. 95% HPD credible intervals of the population mean trajectories are given by the grey shaded regions.

pace, during adolescence (an average rate of 6.51) and early adulthood (an average rate of 8.52). This observation is consistent with the findings in Diamond (2002), which argues that performance in domain-generic executive functions reaches adult performance levels between 12 to 15 years old.

Given that maturation in domain-specific executive functions is largely conditioned by occupational expertise (Li et al., 2004), it could be hypothesised that improvements in soccer-related abilities will be reflected in the trajectory across all developmental stages as players are continuously challenged to refine their skills in order to remain competitive (Mann et al., 2007). However, our results show that the increase in domain-specific executive functions is almost negligible after 15 years old (average rates of 0.11 and 0.07 during adolescence and early adulthood respectively). A possible explanation for the observed plateau is that both domain-generic and domain-specific assessments used in our study do not necessarily represent the way in which perception and action of competition are coupled in soccer (Pinder et al., 2011). For example, while the Footbonaut test has some validity for measuring soccer skills, it requires players to use passing actions to react to visual and auditory stimuli that are unrelated to soccer (Beavan et al., 2019). Meanwhile, the Helix test measures high-level perceptual abilities specific to soccer but its design lacks an action component. Therefore, it is unsurprising that the expected positive association between domain-specific executive functions and age as a proxy for soccer experience is not observed.

When inspecting the posterior distribution of the dependence structures in $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ (results not shown), we find that the slopes of the piecewise linear spline model are independent of each other and across both facets of executive functions. This suggests that previous studies which are based on cross-sectional data (Verburgh et al., 2014; Huijgen et al., 2015; Vestberg et al., 2017; Sakamoto et al., 2018) may have overstated the usefulness of domain-generic executive functions in soccer talent identification. This is because a strong relationship between both types of executive functions should exist if domain-generic executive functions are a prognostic tool for soccer performance. A sample of four players are chosen and shown in Figure 5.5 to illustrate this argument. We observe that players A and B demonstrate an above-average level of domain-generic executive functions but fail to reproduce similar level of superiority in soccer-specific assessments. On the other hand, player D who has less developed domain-generic executive functions compared to the population mean outperforms his peers in terms of soccer expertise. Player D also achieve a comparable level of domain-specific executive functions to that of player C although the latter performs better in the generic abilities test battery.

We now examine the impact of the covariates on each outcome variable in the test battery. Table 5.2 shows the regression coefficient posterior mean estimates for the assessment session and playing position. The players tend to have better response times ($y_2, y_3, y_6, y_7$) in domain-generic tasks during post-season assessment sessions, indicating that there is an acute effect of soccer participation on performance in these assessments. The absence of positional effects in the Helix test further reinforces our previous argument that its design may not have adequately coupled perceptual information with soccer-specific actions. We also observe that goalkeepers generally perform the worst in the Footbonaut test in terms of the response time $y_{10}$ (i.e. players in the other positions respond much faster). This is because as part of their training,

| Variable | Intercept | Post-season | Forward | Midfielder | Defender |
|---|---|---|---|---|---|
| $y_1$ | 141.08 | 1.43 | 0.00 | 0.00 | −0.01 |
| $y_2$ | 0.01 | −0.03 | 0.00 | 0.00 | 0.00 |
| $y_3$ | −1.26 | −0.09 | 0.00 | −0.04 | 0.00 |
| $y_4$ | −0.64 | −0.02 | 0.00 | 0.00 | 0.00 |
| $y_5$ | 81.85 | −0.22 | 0.06 | 0.02 | −0.34 |
| $y_6$ | −0.49 | −0.04 | 0.00 | 0.00 | 0.00 |
| $y_7$ | −0.44 | −0.04 | 0.00 | 0.00 | 0.00 |
| $y_8$ | 28.44 | 0.01 | −0.02 | 0.02 | 0.00 |
| $y_9$ | 22.93 | 0.15 | −0.60 | 0.23 | 0.27 |
| $y_{10}$ | 1.03 | −0.01 | −0.05 | −0.06 | −0.04 |

Table 5.2: Estimated posterior means of regression coefficients $\boldsymbol{\gamma}_d$ for the covariates for each outcome variable. Parameters whose 95% HPD credible interval does not include 0 are highlighted in grey.

goalkeepers tend not to train receiving, control and passing of the ball to the same extent as players in other positions. As a result, we can conclude that the Footbonaut test represents a more useful measure of performance for field players rather than for goalkeepers.

## 5.5 Conclusion

This chapter has explored the relationship between age and executive functions in an athlete population by modelling the cognitive outcomes from a test battery of neuropsychological assessments performed by elite soccer players in a longitudinal study using a latent variable model (Dunson, 2000; Muthén, 2002; Proust et al., 2006). The findings of previous research on the developmental trajectories of executive functions were drawn from cross-sectional studies (Verburgh et al., 2014; Huijgen et al., 2015; Sakamoto et al., 2018), and to the best of our knowledge, this is the first study of its kind in the sport science literature that is based on longitudinal data. The latent growth curve representing the unobserved executive functions is modelled to evolve in a piecewise linear fashion across time using a random effects model, and is linked to the observed outcomes via a measurement model. Following the argument in Li et al. (2004), we differentiate between fluid (domain-generic) and crystallised (domain-specific) executive functions, where the former develops biologically while the latter is acquired through occupational experience. This allows us to make a comparison between their trajectories across different stages of growth development (Côté et al., 2007). Rather than examining both types of executive functions independently as what is commonly done in the literature (Furley and Wood, 2016), we model them jointly in a multivariate formulation to investigate if the claims made on the importance of domain-generic executive functions as a prognostic tool for excellence in soccer can be substantiated.

Our analysis shows that both facets of executive functions exhibit a rapid increase between 10 to 15 years old, and while domain-generic executive functions continue to develop at a much slower rate, domain-specific executive functions begin to plateau after that. The latter observation is in contrary to popular belief that soccer players who excel in competitive settings tend to possess more developed technical abilities shaped by their playing experience. However, the lack of evidence supporting this expectation in our study could possibly be due to the failure of the assessment design to reproduce the perception-action couplings experienced by players during an actual match (Pinder et al., 2011). We also find no substantial dependencies in the rate of developmental growth between domain-generic and domain-specific executive functions, thereby contradicting the findings of earlier studies (Vestberg et al., 2012; Verburgh et al., 2014; Sakamoto et al., 2018) and weakening the argument that domain-generic executive functions provide useful information for soccer talent identification. The longitudinal nature of the study allows our modelling approach to control for unobserved heterogeneity such as the number of training hours (Huijgen et al., 2015), and hence providing a closer representation of the underlying mechanistic development in cognitive abilities.

Considering the results that we have presented, it is clear to conclude that integrating neuropsychological test battery in soccer talent identification programmes is likely a debatable topic given that no interaction is established between domain-generic and domain-specific executive functions. Furthermore, a comprehensive study on the reliability of each neuropsychological assessment in the test battery should be undertaken to validate their use (Dicks et al., 2009), especially if the results of these tests are used pervasively.

# Chapter 6

# Summary and discussion

Chapter 3 presents an analysis of the decision-making of Australian general practitioners (GPs) in an experiment which is designed to mimic the choice problem faced in a medical consultation, where the GPs need to match a set of contraceptive products with a particular female patient. A graphical model representing the dependence structure of the latent variables indicating the observed binary outcomes, identifies products which are perceived to be substitutes and can be used in place of one another. Conditional on the observable characteristics of the patient, the remaining dependencies captured by the GP-specific random effects characterise the persistence of GPs in discussing a particular product in ready-to-wear choices (norms that work well for a broad class of patients). The latter dependence structure suggests evidence of medical practice variation (Wennberg et al., 1982; Scott and Shiell, 1997; Davis et al., 2000) among the GPs which is largely attributed to their age, gender and qualifications. This phenomenon is likely to be one of the contributing factors to the low uptake of long acting reversible contraceptive methods in Australia (Black et al., 2013), as demonstrated by the varied views among the GPs on the suitability of long acting contraceptive choices in a simulated case.

Motivated by the application example which is modelled using a multivariate probit model, we provide an efficient method for sampling the potentially high dimensional correlation matrix $\boldsymbol{R}_{\boldsymbol{\epsilon}}$ for the dependence structure of the variables.

The correlation matrix $\boldsymbol{R}_{\boldsymbol{\epsilon}}$ is reparameterised as an unconstrained lower triangular Cholesky factor which is then sampled using the Hamiltonian Monte Carlo algorithm (Duane et al., 1987; Neal, 2011). Bayesian inference often relies on Markov chain Monte Carlo algorithms to generate samples from the posterior distribution. However, these samples tend to be positively correlated and in turn increase the variability of Monte-Carlo based estimators. To address this issue, we propose an antithetic sampler, which generates proposals in the Metropolis-Hastings algorithm deterministically. In order to obtain an ergodic Markov chain, the antithetic sampler must be coupled with a stochastic update of the other parameters. Our results show that significant improvement is observed in the performance measure and some parameters achieve super-efficiency. While we illustrate the efficiency of the antithetic sampler on an example of a highly correlated bivariate normal distribution and the multivariate probit model, an extension to more general settings or models is worth future research. In addition, establishing the convergence properties of the antithetic sampler would certainly enhance our empirical results.

Chapter 4 investigates faltering growth among young children that is endemic in low to medium income countries. Faltered growth is generally defined as a slower rate of growth compared to a reference healthy population of the same age and gender. As faltering is closely associated with reduced physical, intellectual (Benítez-Bribiesca et al., 1999) and economic productivity potential (Hoddinott et al., 2013), it is important to identify faltered children and be able to characterise different growth patterns, as each type represents a particular growth behaviour and so merits a target-specific medical treatment (Collins et al., 2006). We use a multiclass classification approach by approximating the smooth growth curves by piecewise linear segments with random slopes using the broken stick model (Ruppert et al., 2003). The heterogeneity in the growth velocity between children is captured by allowing the random slopes of the broken stick model to be distributed according to a mixture distribution. Therefore, the mixture component from which the vector of random slopes is generated dictates the clustering of growth profiles into different classes, as in the formulation of a Gaussian mixture model. However, specifying the number of mixture components $G$ is non-trivial, and fitting separate models with different values of $G$ is ad-hoc and ignores uncertainty in the model. This is overcome by adopting a Bayesian non-parametric approach using the Dirichlet process (DP; Ferguson, 1973) prior. Subscribing to this framework allows the complexity of the model, i.e. the value of $G$, to be entirely data-driven since the DP prior exhibits clustering *a priori*. Furthermore, we flexibly extend the broken stick model to ensure that the model remains a sensible approach in practice where children are likely to react differently to treatment interventions. In this extension, we relax the fixed knot locations of the broken stick model to allow for child-specific random change points. The change points are modelled probabilistically using a modified even-numbered order statistics distribution (Green, 1995) so that there is exactly one knot in each subinterval of equal length over the observational period. Simulation results show that the broken stick model with fixed knots produces a biased estimate of the random slopes, which subsequently leads to an overestimation of $G$.

In our work, classification performed on 373 children aged between 0 and 1 from a longitudinal study from the Healthy Birth, Growth and Development knowledge integration (HBGDki) project suggests 9 different growth trajectory patterns. A majority of the children experience faltering growth between birth and age one. Further exploratory analysis suggests that certain growth patterns are more likely to be dominated by a particular gender or maternal education level. Children who experience severe faltering are also found to have lower IQ scores. It would be interesting to investigate whether such patterns are observed in other studies in the HBGDki project. Although the broken stick model provides a reasonable approximation of the growth curves, it requires specifying the number of (fixed or random) knots in advance, which is often unknown in practical applications. One possible extension is to consider an infinite mixture of Gaussian processes by unifying the DP and Gaussian processes. Conceptually, the Gaussian process models each curve non-parametrically without having to pre-specify the number of knots, while the

DP classifies the functionals generated from the Gaussian process. This extends the present framework of DP to functional data setting, whereby a probability measure is now defined on a function space so that a random draw from the functional DP is a smooth function that provides a better approximation to growth curves.

Finally, Chapter 5 aims to gain further insights into the developmental trajectories of executive functions over the playing time of a soccer player. Executive functions are complex cognitive abilities which allow an individual to reason, plan actions and execute strategies to achieve a goal (Grafman and Litvan, 1999; Burgess et al., 2000; Chan et al., 2008). Contemporary research based on cross-sectional data has generally supported the hypothesis that executive functions can be used as a variable in predicting the prospective performance of a soccer player (Verburgh et al., 2014; Huijgen et al., 2015; Sakamoto et al., 2018), and thus they serve as good measures in identifying young talented players. However, longitudinal studies of the developmental trajectories of executive functions across different stages of life in an athlete population are lacking in the literature and previous results were established for a general population (Zelazo et al., 2004; Huizinga and Smidts, 2010; Zelazo and Carlson, 2012). Recent research has shown that active participation in sports has a positive impact on executive functioning (Jacobson and Matthaeus, 2014), and therefore the generalisation of existing results to an athlete population is limited. We address this problem by analysing longitudinal data that examines executive functions of male soccer players representing a professional German Bundesliga club through a series of neuropsychological assessments, which to the best of our knowledge, is the most comprehensive set of data available in the sport science literature. These assessments can be broadly classified into two different categories depending on the facet of executive functions that they measure. Domain-generic assessments measure the level of general intelligence, whereas domain-specific assessments measure the skilfulness of the players acquired through active participation in the sport (Li et al., 2004). Using a latent variable model (Dunson, 2000; Proust et al., 2006), these two facets of executive functions are modelled as unobservable curves to reflect their latency, and their manifestation is related to the corresponding cognitive outcome through a measurement model.

The analysis results show that domain-specific executive functions of the players do not change significantly between the age of 15 and 21 years old. This suggests that the design of the domain-specific assessments may not be specialised enough to test soccer-related skills (Pinder et al., 2011) or that these skills may be irrelevant for soccer expertise (Beavan et al., 2019). Furthermore, no dependence structure is observed between both facets of executive functions, which disproves earlier findings on the usefulness of domain-generic executive functions in soccer talent identification (Verburgh et al., 2014; Huijgen et al., 2015; Vestberg et al., 2017; Sakamoto et al., 2018). The findings that we obtained can be further reinforced if more data is available or the study is carried out over a longer time period. It would also be interesting to investigate the relationship between executive functions and career attainment by having more delineated variables recorded. Additionally, an extensive examination on the reliability and accuracy of each neuropsychological assessment in the test battery should be undertaken to validate their use (Dicks et al., 2009), especially if the results of these tests are used in any major decision-making process.

# Bibliography


Abraham, C., P.-A. Cornillon, E. Matzner-Løber, and N. Molinari (2003). Unsupervised curve clustering using B-splines. *Scandinavian Journal of Statistics 30*(3), 581–595.

Adler, S. L. (1981). Over-relaxation method for the Monte Carlo evaluation of the partition function for multiquadratic actions. *Physical Review D 23*(12), 2901–2904.

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

Alderson, R. M., M. D. Rapport, and M. J. Kofler (2007). Attention-deficit/hyperactivity disorder and behavioral inhibition: a meta-analytic review of the stop-signal paradigm. *Journal of Abnormal Child Psychology 35*(5), 745–758.

Alvarez, J. A. and E. Emory (2006). Executive function and the frontal lobes: A meta-analytic review. *Neuropsychology Review 16*(1), 17–42.

Andersen, T. B. and H. C. Dörge (2011). The influence of speed of approach and accuracy constraint on the maximal speed of the ball in soccer kicking. *Scandinavian Journal of Medicine & Science in Sports 21*(1), 79–84.

Anderson, C., R. Hafen, O. Sofrygin, L. Ryan, and HBGDki Community (2019). Comparing predictive abilities of longitudinal child growth models. *Statistics in Medicine 38*(19), 3555–3570.

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics 2*(6), 1152–1174.

Ashford, J. and R. Sowden (1970). Multivariate probit analysis. *Biometrics 26*(3), 535–546.

Baltes, P. B. and U. Lindenberger (1997). Emergence of a powerful connection between sensory and cognitive functions across the adult life span: A new window to the study of cognitive aging? *Psychology and Aging 12*(1), 12–21.

Bardenet, R., A. Doucet, and C. Holmes (2017). On Markov chain Monte Carlo methods for tall data. *Journal of Machine Learning Research 18*(1), 1515–1557.

Barela, J. A., A. A. Rocha, A. R. Novak, J. Fransen, and G. A. Figueiredo (2019). Age differences in the use of implicit visual cues in a response time task. *Brazilian Journal of Motor Behavior 13*(2), 86–93.

Barnard, J., R. McCulloch, and X.-L. Meng (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica 10*(4), 1281–1311.

Barone, P. and A. Frigessi (1990). Improving stochastic relaxation for Gaussian random fields. *Probability in the Engineering and Informational Sciences 4*(3), 369–389.

Bassuk, S. S. and J. E. Manson (2005). Epidemiological evidence for the role of physical activity in reducing risk of type 2 diabetes and cardiovascular disease. *Journal of Applied Physiology 99*(3), 1193–1204.

Baur, H., S. Müller, A. Hirschmüller, G. Huber, and F. Mayer (2006). Reactivity, stability, and strength performance capacity in motor sports. *British Journal of Sports Medicine 40*(11), 906–911.

Beavan, A., J. Fransen, J. Spielmann, J. Mayer, S. Skorski, and T. Meyer (2019). The Footbonaut as a new football-specific skills test: Reproducibility and age-related differences in highly trained youth players. *Science and Medicine in Football 3*(3), 177–182.

Beavan, A., J. Spielmann, J. Mayer, S. Skorski, T. Meyer, and J. Fransen (2020). The rise and fall of executive functions in high-level football players. *Psychology of Sport and Exercise 49*, 101677.

Beavan, A. F., J. Spielmann, J. Mayer, S. Skorski, T. Meyer, and J. Fransen (2019). Age-related differences in executive functions within high-level youth soccer players. *Brazilian Journal of Motor Behavior 13*(2), 64–75.

Benítez-Bribiesca, L., I. De la Rosa-Alvarez, and A. Mansilla-Olivares (1999). Dendritic spine pathology in infants with severe protein-calorie malnutrition. *Pediatrics 104*(2), e21.

Berlin, K. S., G. R. Parra, and N. A. Williams (2014). An introduction to latent variable mixture modeling (part 2): Longitudinal latent class growth analysis and growth mixture models. *Journal of Pediatric Psychology 39*(2), 188–203.

Best, J. R. and P. H. Miller (2010). A developmental perspective on executive function. *Child Development 81*(6), 1641–1660.

Betancourt, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*.

Bierkens, J., P. Fearnhead, and G. Roberts (2019). The zig-zag process and super-efficient sampling for Bayesian analysis of big data. *Annals of Statistics 47*(3), 1288–1320.

Birren, J. E. and L. M. Fisher (1995). Aging and speed of behavior: Possible consequences for psychological functioning. *Annual Review of Psychology 46*(1), 329–353.

Black, K. I., D. Bateson, and C. Harvey (2013). Australian women need increased access to long-acting reversible contraception. *Medical Journal of Australia 199*(5), 317–318.

Blackwell, D. and J. B. MacQueen (1973). Ferguson distributions via Pólya urn schemes. *Annals of Statistics 1*(2), 353–355.

Blumenthal, P., A. Voedisch, and K. Gemzell-Danielsson (2011). Strategies to prevent unintended pregnancy: Increasing use of long-acting reversible contraception. *Human Reproduction Update 17*(1), 121–137.

Bollen, K. A. and P. J. Curran (2006). *Latent Curve Models: A Structural Equation Perspective*. John Wiley & Sons.

Bouveyron, C., S. Girard, and C. Schmid (2007). High-dimensional data clustering. *Computational Statistics & Data Analysis 52*(1), 502–519.

Bouveyron, C. and J. Jacques (2011). Model-based clustering of time series in group-specific functional subspaces. *Advances in Data Analysis and Classification 5*(4), 281–300.

Brooks, S. and A. Gelman (1998). Some issues in monitoring convergence of iterative simulations. In *Proceedings of the Section on Statistical Computing*, pp. 148–188.

Buchmueller, T. C., D. G. Fiebig, G. Jones, and E. Savage (2013). Preference heterogeneity and selection in private health insurance: The case of Australia. *Journal of Health Economics 32*(5), 757–767.

Burgess, P. W., E. Veitch, A. de Lacy Costello, and T. Shallice (2000). The cognitive and neuroanatomical correlates of multitasking. *Neuropsychologia 38*(6), 848–863.

Carvalho, C. M., N. G. Polson, and J. G. Scott (2010). The horseshoe estimator for sparse signals. *Biometrika 97*(2), 465–480.

Casella, G. and C. P. Robert (1996). Rao-Blackwellisation of Sampling Schemes. *Biometrika 83*(1), 81–94.

Celeux, G., M. Hurn, and C. P. Robert (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association 95*(451), 957–970.

Cepeda, N. J., A. F. Kramer, and J. Gonzalez de Sather (2001). Changes in executive control across the life span: Examination of task-switching performance. *Developmental Psychology 37*(5), 715–730.

Chan, R. C., D. Shum, T. Toulopoulou, and E. Y. Chen (2008). Assessment of executive functions: Review of instruments and identification of critical issues. *Archives of Clinical Neuropsychology 23*(2), 201–216.

Chib, S. and E. Greenberg (1998). Analysis of multivariate probit models. *Biometrika 85*(2), 347–361.

Chin, V., D. Gunawan, D. G. Fiebig, R. Kohn, and S. A. Sisson (2020). Efficient data augmentation for multivariate probit models with panel data: An application to general practitioner decision making about contraceptives. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 69*(2), 277–300.

Chin, V., J. Y. L. Lee, L. M. Ryan, R. Kohn, and S. A. Sisson (2019). Multiclass classification of growth curves using random change points and heterogeneous random effects. *arXiv preprint arXiv:1909.07550.*

Clyde, M. and E. I. George (2004). Model uncertainty. *Statistical Science 19*(1), 81–94.

Cole, T. (1998). Presenting information on growth distance and conditional velocity in one chart: Practical issues of chart design. *Statistics in Medicine 17*(23), 2697–2707.

Collins, S., N. Dent, P. Binns, P. Bahwere, K. Sadler, and A. Hallam (2006). Management of severe acute malnutrition in children. *The Lancet 368*(9551), 1992–2000.

Côté, J., J. Baker, and B. Abernethy (2007). Practice and play in the development of sport expertise. In G. Tenenbaum and R. C. Eklund (Eds.), *Handbook of Sport Psychology* (3rd ed.)., Chapter 8, pp. 184–202. John Wiley & Sons.

Cox, S. M., D. J. Cox, M. J. Kofler, M. A. Moncrief, R. J. Johnson, A. E. Lambert, S. A. Cain, and R. E. Reeve (2016). Driving simulator performance in novice drivers with autism spectrum disorder: The role of executive functions and basic motor skills. *Journal of Autism and Developmental Disorders 46*(4), 1379–1391.

Creutz, M. (1987). Overrelaxation and Monte Carlo simulation. *Physical Review D 36*(2), 515–519.

Crozier, S. R., W. Johnson, T. J. Cole, C. Macdonald-Wallis, G. Muniz-Terrera, H. M. Inskip, and K. Tilling (2019). A discussion of statistical methods to characterise early growth and its impact on bone mineral content later in childhood. *Annals of Human Biology 46*(1), 17–26.

da Silva, A. R. F. (2007). A Dirichlet process mixture model for brain MRI tissue classification. *Medical Image Analysis 11*(2), 169–182.

Daniels, M. J. and M. Pourahmadi (2009). Modeling covariance matrices via partial autocorrelations. *Journal of Multivariate Analysis 100*(10), 2352–2363.

Darroch, J. E., G. Sedgh, and H. Ball (2011). Contraceptive technologies: Responding to women's needs. *New York: Guttmacher Institute*.

Dasgupta, A. and A. E. Raftery (1998). Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American Statistical Association 93*(441), 294–302.

Davis, P., B. Gribben, A. Scott, and R. Lay-Yee (2000). The "supply hypothesis" and medical practice variation in primary care: Testing economic and clinical models of inter-practitioner variation. *Social Science & Medicine 50*(3), 407–418.

De Kroon, M. L., C. M. Renders, J. P. Van Wouwe, R. A. Hirasing, and S. Van Buuren (2011). Identifying young children without overweight at high risk for adult overweight: The Terneuzen Birth Cohort. *International Journal of Pediatric Obesity 6*, e187–e195.

Delaigle, A. and P. Hall (2010). Defining probability density for a distribution of random functions. *Annals of Statistics 38*(2), 1171–1193.

Dellaportas, P. and I. Kontoyiannis (2012). Control variates for estimation based on reversible Markov chain Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 74*(1), 133–161.

Dempster, F. N. (1992). The rise and fall of the inhibitory mechanism: Toward a unified theory of cognitive development and aging. *Developmental Review 12*(1), 45–75.

Diamond, A. (2002). Normal development of prefrontal cortex from birth to young adulthood: Cognitive functions, anatomy, and biochemistry. In D. T. Stuss and R. T. Knight (Eds.), *Principles of Frontal Lobe Function*, pp. 466–503. Oxford University Press.

Diamond, A. (2013). Executive functions. *Annual Review of Psychology 64*, 135–168.

Dicks, M., K. Davids, and C. Button (2009). Representative task design for the study of perception and action in sport. *International Journal of Sport Psychology 40*(4), 506.

Doucet, A., M. K. Pitt, G. Deligiannidis, and R. Kohn (2015). Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. *Biometrika 102*(2), 295–313.

Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 57*(1), 45–70.

Duane, S., A. D. Kennedy, B. J. Pendleton, and D. Roweth (1987). Hybrid Monte Carlo. *Physics Letters B 195*(2), 216–222.

Duncan, T. E., S. C. Duncan, and M. Stoolmiller (1994). Modeling developmental processes using latent growth structural equation methodology. *Applied Psychological Measurement 18*(4), 343–354.

Duncan, T. E., S. C. Duncan, and L. A. Strycker (2013). *An Introduction to Latent Variable Growth Curve Modeling: Concepts, Issues, and Application* (2nd ed.). Quantitative Methodology. Routledge Academic.

Dunson, D. B. (2000). Bayesian latent variable models for clustered mixed outcomes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 62*(2), 355–366.

Dunson, D. B. (2003). Dynamic latent trait models for multidimensional longitudinal data. *Journal of the American Statistical Association 98*(463), 555–563.

Edwards, Y. D. and G. M. Allenby (2003). Multivariate analysis of multiple response data. *Journal of Marketing Research 40*(3), 321–334.

Emond, A. M., P. S. Blair, P. M. Emmett, and R. F. Drewett (2007). Weight faltering in infancy and IQ levels at 8 years in the Avon Longitudinal Study of Parents and Children. *Pediatrics 120*(4), e1051–e1058.

Escobar, M. D. and M. West (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association 90*(430), 577–588.

Fan, Y., J.-L. Dortet-Bernadet, and S. Sisson (2010). On Bayesian curve fitting via auxiliary variables. *Journal of Computational and Graphical Statistics 19*(3), 626–644.

Fédération Internationale de Football Association (FIFA) (2018). FIFA Activity Report 2018. Retrieved from: `https://resources.fifa.com/image/upload/yjibhdqzfwwz5onqszo0.pdf`.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics 1*(2), 209–230.

Ferraty, F. and P. Vieu (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. Springer Series in Statistics. Springer Science & Business Media.

Fiebig, D. G., R. Viney, S. Knox, M. Haas, D. J. Street, A. R. Hole, E. Weisberg, and D. Bateson (2017). Consideration sets and their role in modelling doctor recommendations about contraceptives. *Health Economics 26*(1), 54–73.

Frank, R. G. and R. J. Zeckhauser (2007). Custom-made versus ready-to-wear treatments: Behavioral propensities in physicians' choices. *Journal of Health Economics 26*(6), 1101–1127.

Fritsch, A. (2012). *mcclust: Process an MCMC Sample of Clusterings*. R package version 1.0.

Fritsch, A. and K. Ickstadt (2009). Improved criteria for clustering based on the posterior similarity matrix. *Bayesian Analysis 4*(2), 367–391.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer Series in Statistics. Springer Science & Business Media.

Fúquene, J., M. Steel, and D. Rossell (2019). On choosing mixture components via non-local priors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 81*(5), 809–837.

Furley, P. and G. Wood (2016). Working memory, attentional control, and expertise in sports: A review of current literature and directions for future research. *Journal of Applied Research in Memory and Cognition 5*(4), 415–425.

Ge, H., Y. Chen, M. Wan, and Z. Ghahramani (2015). Distributed inference for Dirichlet process mixture models. In F. Bach and D. Blei (Eds.), *International Conference on Machine Learning*, pp. 2276–2284.

Gelfand, A. E. and A. F. Smith (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association 85*(410), 398–409.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis 1*(3), 515–534.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2013). *Bayesian Data Analysis* (3rd ed.). Texts in Statistical Science Series. CRC Press.

Gelman, A., Y. Goegebeur, F. Tuerlinckx, and I. Van Mechelen (2000). Diagnostic checks for discrete data regression models using posterior predictive simulations. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 49*(2), 247–268.

Gelman, A. and D. B. Rubin (1992). Inference from iterative simulation using multiple sequences. *Statistical Science 7*(4), 457–472.

Gelman, A. and D. B. Rubin (1995). Avoiding model selection in Bayesian social research. *Sociological Methodology 25*, 165–173.

Geman, S. and D. Geman (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence 6*(6), 721–741.

Gershman, S. J. and D. M. Blei (2012). A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology 56*(1), 1–12.

Geweke, J. (1988). Antithetic acceleration of Monte Carlo integration in Bayesian inference. *Journal of Econometrics 38*(1-2), 73–89.

Geweke, J. (1991). Efficient simulation from the multivariate normal and Student-$t$ distributions subject to linear constraints and the evaluation of constraint probabilities. In E. M. Keramidas and S. M. Kaufman (Eds.), *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, pp. 571–578.

Ghahramani, Z. (2013). Bayesian non-parametrics and the probabilistic approach to modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 371*(1984), 20110553.

Ghisletta, P., O. Renaud, N. Jacot, and D. Courvoisier (2015). Linear mixed-effects and latent curve models for longitudinal life course analyses. In S. Cullati, A. Sacker, C. Burton-Jeangros, and D. Blane (Eds.), *A Life Course Perspective on Health Trajectories and Transitions*, pp. 155–178. Springer, Cham.

Ghosh, J. and M. A. Clyde (2011). Rao-Blackwellization for Bayesian variable selection and model averaging in linear and binary regression: A novel data augmentation approach. *Journal of the American Statistical Association 106*(495), 1041–1052.

Ghosh, J. K. and P. K. Sen (1985). On the asymptotic performance of the log likelihood ratio statistic for the mixture model and related results. In *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer*, pp. 789–806.

Giacofci, M., S. Lambert-Lacroix, G. Marot, and F. Picard (2013). Wavelet-based clustering for mixed-effects functional models in high dimension. *Biometrics 69*(1), 31–40.

Gibbons, R. D. and V. Wilcox-Gök (1998). Health service utilization and insurance coverage: A multivariate probit analysis. *Journal of the American Statistical Association 93*(441), 63–72.

Gilbride, T. J. and G. M. Allenby (2004). A choice model with conjunctive, disjunctive, and compensatory screening rules. *Marketing Science 23*(3), 391–406.

Giles, M., T. Nagapetyan, L. Szpruch, S. Vollmer, and K. Zygalakis (2016). Multilevel Monte Carlo for scalable Bayesian computations. *arXiv preprint arXiv:1609.06144*.

Gilks, W. R., G. O. Roberts, and E. I. George (1994). Adaptive direction sampling. *Journal of the Royal Statistical Society: Series D (The Statistician) 43*(1), 179–189.

Glynn, P. W. and R. Szechtman (2002). Some new perspectives on the method of control variates. In K.-T. Fang, F. J. Hickernell, and H. Niederreiter (Eds.), *Monte Carlo and Quasi-Monte Carlo Methods*, pp. 27–49. Springer.

Gong, X., A. van Soest, and E. Villagomez (2004). Mobility in the urban labor market: A panel data analysis for Mexico. *Economic Development and Cultural Change 53*(1), 1–36.

Goode, A., K. Mavromaras, and R. Zhu (2014). Family income and child health in China. *China Economic Review 29*, 152–165.

Grafman, J. and I. Litvan (1999). Importance of deficits in executive functions. *The Lancet 354*(9194), 1921–1923.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika 82*(4), 711–732.

Green, P. J. and X.-l. Han (1992). Metropolis methods, Gaussian proposals and antithetic variables. In P. Barone, A. Frigessi, and M. Piccioni (Eds.), *Stochastic Models, Statistical Methods, and Algorithms in Image Analysis*, Lecture Notes in Statistics, pp. 142–164. Springer-Verlag New York.

Greenberg, E. (2012). *Introduction to Bayesian Econometrics* (2nd ed.). Cambridge University Press.

Grimm, K. J., N. Ram, and F. Hamagami (2011). Nonlinear growth curves in developmental research. *Child Development 82*(5), 1357–1371.

Gunawan, D., D. Fiebig, R. Kohn, et al. (2017). Efficient Bayesian estimation for flexible panel models for multivariate outcomes: Impact of life events on mental health and excessive alcohol consumption. *arXiv preprint arXiv:1706.03953*.

Habacha, H., C. Molinaro, and F. Dosseville (2014). Effects of gender, imagery ability, and sports practice on the performance of a mental rotation task. *The American Journal of Psychology 127*(3), 313–323.

Hall, C. B., J. Ying, L. Kuo, M. Sliwinski, H. Buschke, M. Katz, and R. B. Lipton (2001). Estimation of bivariate measurements having different change points, with application to cognitive ageing. *Statistics in Medicine 20*(24), 3695–3714.

Hammersley, J. and K. Morton (1956). A new Monte Carlo technique: Antithetic variates. In *Mathematical Proceedings of the Cambridge Philosophical Society*, Volume 52, pp. 449–475.

Hartzel, J., A. Agresti, and B. Caffo (2001). Multinomial logit random effects models. *Statistical Modelling 1*(2), 81–102.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning* (2nd ed.). Springer Series in Statistics. Springer Science & Business Media.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika 57*(1), 97–109.

Hauser, J. R. and B. Wernerfelt (1990). An evaluation cost model of consideration sets. *Journal of Consumer Research 16*(4), 393–408.

Heard, N. A., C. C. Holmes, and D. A. Stephens (2006). A quantitative study of gene regulation involved in the immune response of anopheline mosquitoes: An application of Bayesian hierarchical clustering of curves. *Journal of the American Statistical Association 101*(473), 18–29.

Hedeker, D. (2003). A mixed-effects multinomial logistic regression model. *Statistics in Medicine 22*(9), 1433–1446.

Hillman, C. H., K. I. Erickson, and A. F. Kramer (2008). Be smart, exercise your heart: Exercise effects on brain and cognition. *Nature Reviews Neuroscience 9*(1), 58–65.

Hjort, N. L., C. Holmes, P. Müller, and S. G. Walker (2010). *Bayesian Nonparametrics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Hoddinott, J., H. Alderman, J. R. Behrman, L. Haddad, and S. Horton (2013). The economic rationale for investing in stunting reduction. *Maternal & Child Nutrition 9*(2), 69–82.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky (1999). Bayesian model averaging: A tutorial. *Statistical Science 14*(4), 382–401.

Hoffman, M. D. and A. Gelman (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research 15*(1), 1593–1623.

Huang, A. and M. P. Wand (2013). Simple marginally noninformative prior distributions for covariance matrices. *Bayesian Analysis 8*(2), 439–452.

Hubert, L. and P. Arabie (1985). Comparing partitions. *Journal of Classification 2*(1), 193–218.

Huijgen, B. C., S. Leemhuis, N. M. Kok, L. Verburgh, J. Oosterlaan, M. T. Elferink-Gemser, and C. Visscher (2015). Cognitive functions in elite and sub-elite youth soccer players aged 13 to 17 years. *PLoS ONE 10*(12), e0144580.

Huizinga, M., C. V. Dolan, and M. W. van der Molen (2006). Age-related change in executive function: Developmental trends and a latent variable analysis. *Neuropsychologia 44*(11), 2017–2036.

Huizinga, M. and D. P. Smidts (2010). Age-related changes in executive function: A normative study with the Dutch version of the Behavior Rating Inventory of Executive Function (BRIEF). *Child Neuropsychology 17*(1), 51–66.

Jacobson, J. and L. Matthaeus (2014). Athletics and executive functioning: How athletic participation and sport type correlate with cognitive performance. *Psychology of Sport and Exercise 15*(5), 521–527.

Jacqmin-Gadda, H., C. Proust-Lima, and H. Amiéva (2010). Semi-parametric latent process model for longitudinal ordinal data: Application to cognitive decline. *Statistics in Medicine 29*(26), 2723–2731.

James, G. M. and C. A. Sugar (2003). Clustering for sparsely sampled functional data. *Journal of the American Statistical Association 98*(462), 397–408.

Jasra, A., C. C. Holmes, and D. A. Stephens (2005). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science 20*(1), 50–67.

Johnson, A. A., G. L. Jones, and R. C. Neath (2013). Component-wise Markov chain Monte Carlo: Uniform and geometric ergodicity under mixing and composition. *Statistical Science 28*(3), 360–375.

Jumbe, N. L., J. C. Murray, and S. Kern (2016). Data sharing and inductive learning – Toward healthy birth, growth, and development. *New England Journal of Medicine 374*(25), 2415–2417.

Jurado, M. B. and M. Rosselli (2007). The elusive nature of executive functions: A review of our current understanding. *Neuropsychology Review 17*(3), 213–233.

Kail, R. and T. A. Salthouse (1994). Processing speed as a mental capacity. *Acta Psychologica 86*(2-3), 199–225.

Kass, R. E., B. P. Carlin, A. Gelman, and R. M. Neal (1998). Markov chain Monte Carlo in practice: A roundtable discussion. *American Statistician 52*(2), 93–100.

Keino, S., G. Plasqui, G. Ettyang, and B. van den Borne (2014). Determinants of stunting and overweight among young children and adolescents in sub-Saharan Africa. *Food and Nutrition Bulletin 35*(2), 167–178.

Kirby, M. and E. Danner (2009). Nutritional deficiencies in children on restricted diets. *Pediatric Clinics 56*(5), 1085–1103.

Kossmann, J., P. Nestel, M. Herrera, A. Amin, and W. Fawzi (2000). Undernutrition in relation to childhood infections: A prospective study in the Sudan. *European Journal of Clinical Nutrition 54*(6), 463–472.

Kramer, A. F. and K. I. Erickson (2007). Capitalizing on cortical plasticity: influence of physical activity on cognition and brain function. *Trends in Cognitive Sciences 11*(8), 342–348.

Krenn, B., T. Finkenzeller, S. Würth, and G. Amesberger (2018). Sport type determines differences in executive functions in elite athletes. *Psychology of Sport and Exercise 38*, 72–79.

Kroese, D. P., T. Taimre, and Z. I. Botev (2013). *Handbook of Monte Carlo Methods*. Wiley Series in Probability and Statistics. John Wiley & Sons.

Lai, D., H. Xu, D. Koller, T. Foroud, and S. Gao (2016). A multivariate finite mixture latent trajectory model with application to dementia studies. *Journal of Applied Statistics 43*(14), 2503–2523.

Laird, N. M. and J. H. Ware (1982). Random-effects models for longitudinal data. *Biometrics 38*(4), 963–974.

Lartey, A. (2008). Maternal and child nutrition in Sub-Saharan Africa: Challenges and interventions. *Proceedings of the Nutrition Society 67*(1), 105–108.

Lebel, C., L. Walker, A. Leemans, L. Phillips, and C. Beaulieu (2008). Microstructural maturation of the human brain from childhood to adulthood. *Neuroimage 40*(3), 1044–1055.

Lee, J. Y., P. J. Green, and L. M. Ryan (2017). On the "Poisson Trick" and its extensions for fitting multinomial regression models. *arXiv preprint arXiv:1707.08538*.

Lee, J. Y. L., C. Anderson, W. T. Hung, H. Hwang, and L. M. Ryan (2018). Detecting faltering growth in children via minimum random slopes. *arXiv preprint arXiv:1812.05903*.

Leung, M., D. G. Bassani, A. Racine-Poon, A. Goldenberg, S. A. Ali, G. Kang, P. S. Premkumar, and D. E. Roth (2017). Conditional random slope: A new approach for estimating individual child growth velocity in epidemiological research. *American Journal of Human Biology 29*(5), e23009.

Lewandowski, D., D. Kurowicka, and H. Joe (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis 100*(9), 1989–2001.

Li, F., T. E. Duncan, S. C. Duncan, and A. Acock (2001). Latent growth modeling of longitudinal data: A finite growth mixture modeling approach. *Structural Equation Modeling 8*(4), 493–530.

Li, S.-C., U. Lindenberger, B. Hommel, G. Aschersleben, W. Prinz, and P. B. Baltes (2004). Transformations in the couplings among intellectual abilities and constituent cognitive processes across the life span. *Psychological Science 15*(3), 155–163.

Liechty, J. C., M. W. Liechty, and P. Müller (2004). Bayesian correlation estimation. *Biometrika 91*(1), 1–14.

Lin, H., B. W. Turnbull, C. E. McCulloch, and E. H. Slate (2002). Latent class models for joint analysis of longitudinal biomarker and event process data: Application to longitudinal prostate-specific antigen readings and prostate cancer. *Journal of the American Statistical Association 97*(457), 53–65.

Lindenberger, U. and P. B. Baltes (1997). Intellectual functioning in old and very old age: Cross-sectional results from the Berlin Aging Study. *Psychology and Aging 12*(3), 410–432.

Lindstrom, M. J. and D. M. Bates (1988). Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association 83*(404), 1014–1022.

Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing.* Springer Series in Statistics. Springer-Verlag New York.

Ljac, V., Z. Witkowski, B. Gutni, A. Samovarov, and D. Nash (2012). Toward effective forecast of professionally important sensorimotor cognitive abilities of young soccer players. *Perceptual and Motor Skills 114*(2), 485–506.

Loève, M. (1977). *Probability Theory I* (4th ed.). Graduate Texts in Mathematics. Springer-Verlag New York.

Logan, G. D. (1994). On the ability to inhibit thought and action: A users' guide to the stop signal paradigm. In D. Dagenbach and T. Carr (Eds.), *Inhibitory Processes in Attention, Memory, and Language*, Chapter 5, pp. 189–239. Academic Press.

Long, J. and J. Ryoo (2010). Using fractional polynomials to model non-linear trends in longitudinal data. *British Journal of Mathematical and Statistical Psychology 63*(1), 177–203.

Magidson, J. and J. K. Vermunt (2004). Latent class models. In D. Kaplan (Ed.), *Handbook of Quantitative Methodology for the Social Sciences*, pp. 175–198. Sage Publications.

Magnus, J. R. and H. Neudecker (1999). *Matrix Differential Calculus with Applications in Statistics and Econometrics* (Revised ed.). Wiley Series in Probability and Statistics. John Wiley & Sons.

Makalic, E. and D. F. Schmidt (2016). A simple sampler for the horseshoe estimator. *IEEE Signal Processing Letters 23*(1), 179–182.

Mann, D. T., A. M. Williams, P. Ward, and C. M. Janelle (2007). Perceptual-cognitive expertise in sport: A meta-analysis. *Journal of Sport and Exercise Psychology 29*(4), 457–478.

Martorell, R. (1999). The nature of child malnutrition and its long-term implications. *Food and Nutrition Bulletin 20*(3), 288–292.

McArdle, J. J. (1986). Latent variable growth within behavior genetic models. *Behavior Genetics 16*(1), 163–200.

McArdle, J. J. and D. Epstein (1987). Latent growth curves within developmental structural equation models. *Child Development 58*(1), 110–133.

McElreath, R. (2020). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan.* CRC press.

McLachlan, G. J. and D. Peel (2000). *Finite Mixture Models.* Wiley Series in Probability and Statistics. John Wiley & Sons.

Meredith, W. and J. Tisak (1990). Latent curve analysis. *Psychometrika 55*(1), 107–122.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics 21*(6), 1087–1092.

Metropolis, N. and S. Ulam (1949). The Monte Carlo method. *Journal of the American Statistical Association 44*(247), 335–341.

Muthén, B. (2008). Latent variable hybrids: Overview of old and new models. In G. R. H. Karen M. Samuelsen (Ed.), *Advances in Latent Variable Mixture Models*, pp. 1–24. Information Age Publishing.

Muthén, B. and L. K. Muthén (2000). Integrating person-centered and variable-centered analyses: Growth mixture modeling with latent trajectory classes. *Alcoholism: Clinical and experimental Research 24*(6), 882–891.

Muthén, B. and K. Shedden (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics 55*(2), 463–469.

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement 28*(4), 338–354.

Muthén, B. O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika 29*(1), 81–117.

Muthén, B. O. (2004). Latent variable analysis: Growth mixture modeling and related techniques for longitudinal data. In D. Kaplan (Ed.), *Handbook of Quantitative Methodology for the Social Sciences*, pp. 345–368. Sage Publications.

Nagin, D. S. (1999). Analyzing developmental trajectories: A semiparametric, group-based approach. *Psychological Methods 4*(2), 139–157.

Neal, R. M. (1996). *Bayesian Learning for Neural Networks*. Lecture Notes in Statistics. Springer-Verlag New York.

Neal, R. M. (1998). Suppressing random walks in Markov chain Monte Carlo using ordered overrelaxation. In M. I. Jordan (Ed.), *Learning in Graphical Models*, Nato Science Series D: Behavioural and Social Sciences, pp. 205–228. Springer Netherlands.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In S. Brooks, A. Gelman, G. Jones, and X.-L. Meng (Eds.), *Handbook of Markov Chain Monte Carlo*, Chapter 5, pp. 113–162. Chapman & Hall.

Nesterov, Y. (2009). Primal-dual subgradient methods for convex problems. *Mathematical Programming 120*(1), 221–259.

Nylund, K. L., T. Asparouhov, and B. O. Muthén (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling 14*(4), 535–569.

Oates, C. J., M. Girolami, and N. Chopin (2017). Control functionals for Monte Carlo integration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 79*(3), 695–718.

Olsson, J. and T. Ryden (2011). Rao-Blackwellization of particle Markov chain Monte Carlo methods using forward filtering backward sampling. *IEEE Transactions on Signal Processing 59*(10), 4606–4619.

Olusanya, B. O. and J. K. Renner (2013). Pattern and characteristics of growth faltering in early infancy in an urban sub-Saharan African setting. *Pediatrics & Neonatology 54*(2), 119–127.

Onofiok, N. and D. Nnanyelugo (1998). Weaning foods in West Africa: Nutritional problems and possible solutions. *Food and Nutrition Bulletin 19*(1), 27–33.

Orbanz, P. and Y. W. Teh (2010). Bayesian nonparametric models. In C. Sammut and G. I. Webb (Eds.), *Encyclopedia of Machine Learning*, pp. 81–89. Springer-Verlag New York.

Pakman, A. and L. Paninski (2014). Exact Hamiltonian Monte Carlo for truncated multivariate Gaussians. *Journal of Computational and Graphical Statistics 23*(2), 518–542.

Pan, H. and H. Goldstein (1998). Multi-level repeated measures growth modelling using extended spline functions. *Statistics in Medicine 17*(23), 2755–2770.

Paul, A., B. P. Gladstone, I. Mukhopadhya, and G. Kang (2014). Rotavirus infections in a community based cohort in Vellore, India. *Vaccine 32*(11), A49–A54.

Peng, J. and H.-G. Müller (2008). Distance-based clustering of sparsely observed stochastic processes, with applications to online auctions. *Annals of Applied Statistics 2*(3), 1056–1077.

Perry, P. O. (2017). Fast moment-based estimation for hierarchical models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 79*(1), 267–291.

Philippe, A. and C. P. Robert (2001). Riemann sums for MCMC estimation and convergence monitoring. *Statistics and Computing 11*(2), 103–115.

Pinder, R. A., K. Davids, I. Renshaw, and D. Araújo (2011). Manipulating informational constraints shapes movement reorganization in interceptive actions. *Attention, Perception, & Psychophysics 73*(4), 1242–1254.

Pitt, M. K., R. dos Santos Silva, P. Giordani, and R. Kohn (2012). On some properties of Markov chain Monte Carlo simulation methods based on the particle filter. *Journal of Econometrics 171*(2), 134–151.

Plummer, M., N. Best, K. Cowles, and K. Vines (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R News 6*(1), 7–11.

Proust, C., H. Jacqmin-Gadda, J. M. Taylor, J. Ganiayre, and D. Commenges (2006). A nonlinear model with latent process for cognitive evolution using multivariate longitudinal data. *Biometrics 62*(4), 1014–1024.

Proust-Lima, C., P. Joly, J.-F. Dartigues, and H. Jacqmin-Gadda (2009). Joint modelling of multivariate longitudinal outcomes and a time-to-event: A nonlinear latent class approach. *Computational Statistics & Data Analysis 53*(4), 1142–1154.

R Core Team (2019). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing.

Ram, N. and K. J. Grimm (2009). Methods and measures: Growth mixture modeling: A method for identifying differences in longitudinal change among unobserved groups. *International Journal of Behavioral Development 33*(6), 565–576.

Ramakrishnan, U., R. Martorell, D. G. Schroeder, and R. Flores (1999). Role of intergenerational effects on linear growth. *Journal of Nutrition 129*(2), 544S–549S.

Ramsay, J. and B. Silverman (2005). *Functional Data Analysis* (2nd ed.). Springer Series in Statistics. Springer-Verlag New York.

Rapisarda, F., D. Brigo, and F. Mercurio (2007). Parameterizing correlations: A geometric interpretation. *IMA Journal of Management Mathematics 18*(1), 55–73.

Rasmussen, C. E. (2000). The infinite Gaussian mixture model. In S. Solla, T. Leen, and K. Müller (Eds.), *Advances in Neural Information Processing Systems*, pp. 554–560.

Rasmussen, C. E. and C. K. I. Williams (2006). *Gaussian Processes for Machine Learning*. MIT Press.

Ray, S. and B. Mallick (2006). Functional clustering by Bayesian wavelet methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68*(2), 305–332.

Richardson, S. and P. J. Green (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 59*(4), 731–792.

Ritter, C. and M. A. Tanner (1992). Facilitating the Gibbs sampler: The Gibbs stopper and the Griddy-Gibbs sampler. *Journal of the American Statistical Association 87*(419), 861–868.

Robert, C. P. and G. Casella (2004). *Monte Carlo Statistical Methods* (2nd ed.). Springer Texts in Statistics. Springer-Verlag New York.

Roberts, G. O. and J. S. Rosenthal (2009). Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics 18*(2), 349–367.

Roberts, J. H. and J. M. Lattin (1991). Development and testing of a model of consideration set composition. *Journal of Marketing Research 28*(4), 429–440.

Rodriguez, A., D. B. Dunson, and A. E. Gelfand (2008). The nested Dirichlet process. *Journal of the American Statistical Association 103*(483), 1131–1154.

Rolfe, M. I. (2010). *Bayesian models for longitudinal data.* Ph. D. thesis, Queensland University of Technology.

Rossi, F., B. Conan-Guez, and A. El Golli (2004). Clustering functional data with the SOM algorithm. In *Proceedings of the 12th European Symposium on Artificial Neural Networks*, pp. 305–312.

Roy, J. and X. Lin (2000). Latent variable models for longitudinal data with multiple continuous outcomes. *Biometrics 56*(4), 1047–1054.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applies statistician. *Annals of Statistics 12*(4), 1151–1172.

Rubinstein, R. Y. and D. P. Kroese (2017). *Simulation and the Monte Carlo Method.* Wiley Series in Probability and Statistics. John Wiley & Sons.

Ruppert, D., M. P. Wand, and R. J. Carroll (2003). *Semiparametric Regression.* Cambridge University Press.

Saal, C., J. Zinner, H. Fiedler, R. Lanwehr, and J. Krug (2018). Reliability and validity of a soccer passing test using the Footbonaut. *German Journal Of Exercise And Sport Research 48*(3), 334–340.

Sakamoto, S., H. Takeuchi, N. Ihara, B. Ligao, and K. Suzukawa (2018). Possible requirement of executive functions for high performance in soccer. *PLoS ONE 13*(8), e0201871.

Salthouse, T. A. (1996). The processing-speed theory of adult age differences in cognition. *Psychological Review 103*(3), 403–428.

Salthouse, T. A., D. Z. Hambrick, and K. E. McGuthry (1998). Shared age-related influences on cognitive and noncognitive variables. *Psychology and Aging 13*(3), 486–500.

Salthouse, T. A., H. E. Hancock, E. J. Meinz, and D. Z. Hambrick (1996). Interrelations of age, visual acuity, and cognitive functioning. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences 51*(6), 317–330.

Sammel, M. D. and L. M. Ryan (1996). Latent variable models with fixed effects. *Biometrics 52*(2), 650–663.

Scarpa, B. and D. B. Dunson (2014). Enriched stick-breaking processes for functional data. *Journal of the American Statistical Association 109*(506), 647–660.

Scott, A. and A. Shiell (1997). Analysing the effect of competition on general practitioners' behaviour using a multilevel modelling framework. *Health Economics 6*(6), 577–588.

Secura, G. (2013). Long-acting reversible contraception: A practical solution to reduce unintended pregnancy. *Minerva Ginecologica 65*(3), 271–277.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica 4*(2), 639–650.

Sherlock, C., P. Fearnhead, and G. O. Roberts (2010). The random walk Metropolis: Linking theory and practice through a case study. *Statistical Science 25*(2), 172–190.

Shi, M., R. E. Weiss, and J. M. Taylor (1996). An analysis of paediatric CD4 counts for acquired immune deficiency syndrome using flexible random curves. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 45*(2), 151–163.

Sisson, S. A. (2005). Trans-dimensional Markov chains: A decade of progress and future perspectives. *Journal of the American Statistical Association 100*(471), 1077–1089.

Sliwinski, M. J., S. M. Hofer, and C. Hall (2003). Correlated and coupled cognitive change in older adults with and without preclinical dementia. *Psychology and Aging 18*(4), 672.

Smith, M. S. (2013). Bayesian approaches to copula modelling. In P. Damien, P. Dellaportas, N. G. Polson, and D. A. Stephens (Eds.), *Bayesian Theory and Applications*, Chapter 17, pp. 336–358. Oxford University Press.

Sofi, F., A. Capalbo, F. Cesari, R. Abbate, and G. F. Gensini (2008). Physical activity during leisure time and primary prevention of coronary heart disease: An updated meta-analysis of cohort studies. *European Journal of Cardiovascular Prevention & Rehabilitation 15*(3), 247–257.

Stan Development Team (2017). *Stan Modeling Language Users Guide and Reference Manual.* Version 2.17.0.

Sterba, S. K. (2014). Fitting nonlinear latent growth curve models with individually varying time points. *Structural Equation Modeling: A Multidisciplinary Journal 21*(4), 630–647.

Stevens-Simon, C., L. Kelly, and R. Kulick (2001). A village would be nice but...: It takes a long-acting contraceptive to prevent repeat adolescent pregnancies. *American Journal of Preventive Medicine 21*(1), 60–65.

Stoolmiller, M. (1995). Using latent growth curve models to study developmental processes. In J. M. Gottman (Ed.), *The Analysis of Change*, pp. 103–138. Lawrence Erlbaum Associates.

Stuss, D. T. and M. P. Alexander (2000). Executive functions and the frontal lobes: a conceptual view. *Psychological Research 63*(3-4), 289–298.

Suarez, A. J. and S. Ghosal (2016). Bayesian clustering of functional data using local features. *Bayesian Analysis 11*(1), 71–98.

Sugar, C. A. and G. M. James (2003). Finding the number of clusters in a dataset: An information-theoretic approach. *Journal of the American Statistical Association 98*(463), 750–763.

Sundstrom, B., A. Baker-Whitcomb, and A. L. DeMaria (2015). A qualitative analysis of long-acting reversible contraception. *Maternal and Child Health Journal 19*(7), 1507–1514.

Sundstrom, B., A. L. DeMaria, M. Ferrara, S. Meier, and D. Billings (2019). "The Closer, the Better:" The role of telehealth in increasing contraceptive access among women in rural South Carolina. *Maternal and Child Health Journal 23*(9), 1196–1205.

Székely, G. J. and M. L. Rizzo (2004). Testing for equal distributions in high dimension. *InterStat 5*(16.10), 1249–1272.

Tan, Q., M. Thomassen, J. v. B. Hjelmborg, A. Clemmensen, K. E. Andersen, T. K. Petersen, M. McGue, K. Christensen, and T. A. Kruse (2011). A growth curve model with fractional polynomials for analysing incomplete time-course data in microarray gene expression studies. *Advances in Bioinformatics 2011*, 1–6.

Tarpey, T. and K. K. Kinateder (2003). Clustering functional data. *Journal of Classification 20*(1), 093–114.

Taylor, S. J., L. A. Barker, L. Heavey, and S. McHale (2013). The typical developmental trajectory of social and executive functions in late adolescence and early adulthood. *Developmental Psychology 49*(7), 1253.

Teh, Y. W. (2011). Dirichlet process. In C. Sammut and G. I. Webb (Eds.), *Encyclopedia of Machine Learning*, pp. 280–287. Springer-Verlag New York.

Tilling, K., C. Macdonald-Wallis, D. A. Lawlor, R. A. Hughes, and L. D. Howe (2014). Modelling childhood growth using fractional polynomials and linear splines. *Annals of Nutrition and Metabolism 65*(2-3), 129–138.

Titterington, D. M., A. F. Smith, and U. E. Makov (1985). *Statistical Analysis of Finite Mixture Distributions.* John Wiley & Son.

Tokuda, T., B. Goodrich, I. Van Mechelen, A. Gelman, and F. Tuerlinckx (2011). Visualizing distributions of covariance matrices. Technical report, Columbia University.

Tokushige, S., H. Yadohisa, and K. Inada (2007). Crisp and fuzzy $k$-means clustering algorithms for multivariate functional data. *Computational Statistics 22*(1), 1–16.

United Nations Children's Fund, World Health Organization (WHO), and World Bank Group (2018). *Levels and trends in child malnutrition: Key findings of the 2018 Edition of the Joint Child Malnutrition Estimates*. Geneva: World Health Organization.

Verbeke, G. and E. Lesaffre (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association 91*(433), 217–221.

Verburgh, L., E. J. Scherder, P. A. van Lange, and J. Oosterlaan (2014). Executive functioning in highly talented soccer players. *PLoS ONE 9*(3), e91254.

Vestberg, T., R. Gustafson, L. Maurex, M. Ingvar, and P. Petrovic (2012). Executive functions predict the success of top-soccer players. *PLoS ONE 7*(4), e34731.

Vestberg, T., G. Reinebo, L. Maurex, M. Ingvar, and P. Petrovic (2017). Core executive functions are associated with success in young elite soccer players. *PLoS ONE 12*(2), e0170845.

Wade, S., S. Mongelluzzo, and S. Petrone (2011). An enriched conjugate prior for Bayesian nonparametric inference. *Bayesian Analysis 6*(3), 359–385.

Walker, S. G. (2007). Sampling the Dirichlet mixture model with slices. *Communications in Statistics-Simulation and Computation 36*(1), 45–54.

Wang, Z., Y. Wu, and H. Chu (2018). On equivalence of the LKJ distribution and the restricted Wishart distribution. *arXiv preprint arXiv:1809.04746*.

Wellings, K., Z. Zhihong, A. Krentel, G. Barrett, and A. Glasier (2007). Attitudes towards long-acting reversible methods of contraception in general practice in the UK. *Contraception 76*(3), 208–214.

Wennberg, J. E., B. A. Barnes, and M. Zubkoff (1982). Professional uncertainty and the problem of supplier-induced demand. *Social Science & Medicine 16*(7), 811–824.

Werner, B. and L. Bodin (2006). Growth from birth to age 19 for children in Sweden born in 1981: Descriptive values. *Acta Paediatrica 95*(5), 600–613.

Whiteside, A., G. Parker, and R. Snodgrass (2003). A review of selected tests from the Vienna Test System. *Selection and Development Review 19*(4), 7–11.

WHO Multicentre Growth Reference Study Group (2006). *WHO Child Growth Standards: Length/height-for-age, weight-for-age, weight-for-length, weight-for-height and body mass index-for-age: Methods and development*. Geneva: World Health Organization.

Wolf, A. P. (2016). *Identification and prediction of inter-individual differences in cognitive training trajectories: A growth mixture modelling approach*. Ph. D. thesis, University of Tasmania.

Xiao, L., V. Zipunnikov, D. Ruppert, and C. Crainiceanu (2016). Fast covariance estimation for high-dimensional functional data. *Statistics and Computing 26*(1-2), 409–421.

Zelazo, P. D. and S. M. Carlson (2012). Hot and cool executive function in childhood and adolescence: Development and plasticity. *Child Development Perspectives 6*(4), 354–360.

Zelazo, P. D., F. I. Craik, and L. Booth (2004). Executive function across the life span. *Acta Psychologica 115*(2-3), 167–183.

Zelazo, P. D. and U. Müller (2011). Executive function in typical and atypical development. In U. Goswami (Ed.), *The Wiley-Blackwell Handbook of Childhood Cognitive Development* (2nd ed.)., Chapter 22, pp. 574–603. Wiley-Blackwell.

Zhang, J., Z. Ghahramani, and Y. Yang (2005). A probabilistic model for online document clustering with application to novelty detection. In *Advances in Neural Information Processing Systems*, pp. 1617–1624.

Zhong, J., A. Rifkin-Graboi, A. T. Ta, K. L. Yap, K.-H. Chuang, M. J. Meaney, and A. Qiu (2014). Functional networks in parallel with cortical development associate with executive functions in children. *Cerebral Cortex 24*(7), 1937–1947.