THESIS DECLARATION

The undersigned

SURNAME: Mongelluzzo

NAME: Silvia

PhD Registration Number: 1370097

Thesis title: Bayesian Semiparametric Inference for Longitudinal Data with Applications

PhD in Statistics

Cycle XXIV

Candidates tutor: Sonia Petrone

Year of discussion: 2013

DECLARES Under his responsibility:

1. that, according to the Presidents decree of 28.12.2000, No. 445, mendacious declarations, falsifying records and the use of false records are punishable under the penal code and special laws, should any of these hypotheses prove true, all benefits included in this declaration and those of the temporary embargo are automatically forfeited from the beginning;

2. that the University has the obligation, according to art. 6, par. 11, Ministerial Decree of 30th April 1999 protocol no. 224/1999, to keep copy of the thesis on deposit at the Biblioteche Nazionali Centrali di Roma e Firenze, where consultation is permitted, unless there is a temporary embargo in order to protect the rights of external bodies and industrial/commercial exploitation of the thesis;

3. that the Servizio Biblioteca Bocconi will file the thesis in its Archivio istituzionale ad accesso aperto and will permit on-line consultation of the complete text (except in cases of a temporary embargo);

4. that in order keep the thesis on file at Biblioteca Bocconi, the University requires that the thesis be delivered by the candidate to

Societ NORMADEC (acting on behalf of the University) by online procedure the contents of which must be unalterable and that NORMADEC will indicate in each footnote the following information: - thesis Bayesian Semiparametric Inference for Longitudinal Data with Applications; - by Mongelluzzo Silvia; - discussed at Università Commerciale Luigi Bocconi Milano in 2013; - the thesis is protected by the regulations governing copyright (law of 22 April 1941, no. 633 and successive modifications). The exception is the right of Universit Commerciale Luigi Bocconi to reproduce the same for research and teaching purposes, quoting the source; 5) that the copy of the thesis deposited with NORMADEC by online procedure is identical to those handed in/sent to the Examiners and to any other copy deposited in the University offices on paper or electronic copy and, as a consequence, the University is absolved from any responsibility regarding errors, inaccuracy or omissions in the contents of the thesis; 6) that the contents and organization of the thesis is an original work carried out by the undersigned and does not in any way compromise the rights of third parties (law of 22 April 1941, no. 633 and successive integrations and modifications), including those regarding security of personal details; therefore the University is in any case absolved from any responsibility whatsoever, civil, administrative or penal and shall be exempt from any requests or claims from third parties; 7) choose hypothesis 7a or 7b indicated below: 7a) that the PhD thesis is not the result of work included in the regulations governing industrial property, it was not produced as part of projects financed by public or private bodies with restrictions on the diffusion of the results; it is not subject to patent or protection registrations, and therefore not subject to an embargo;

Date 20th October 2012

SURNAME Mongelluzzo NAME Silvia

Dedication

I would like to dedicate this PhD thesis to my blond nephews and nieces -*Antea, Paride and Maila*- who make me happily tired during my free time; to my dark boyfriend -*Massimiliano*- who distracts me from any source of stress (although sometimes by causing me other stress); and to my white father -*Enrico*- who has always given me his unconditional and full support.

Acknoledgement

I would like to thank my advisor Prof. Sonia Petrone for her support to my Ph.D study and research, for her patience, motivation, enthusiasm and knowledge.

Besides my advisor, I would like to thank the rest of my thesis committee: Prof. Raffaella Piccarreta and Dr. Fabio Rigat, for their insightful comments.

My thanks also goes to Dr. Marina Savelieva and Dr. Celestino Giron, for offering me the opportunities to work temporary in their groups and allowing me to deal with diverse exciting projects.

I would like to thank my PhD classmates too. In particular, I thank Sara Wade for her enthusiasm, her immense talent and her support. I have learned a lot from her. I want also to thank Agnese Vitali for helpful and motivating discussions.

Last but not the least, I would like to thank all my family: my parents, Enrico and Maurizia, my sister, Chiara, my brother, Alessio, and my boyfriend, Massimiliano, for their patience and support.

Contents

Introduction

This thesis deals with Bayesian inference for longitudinal data. Most of this thesis follows an application-driven model-building approach, with areas of applications ranging from economics to pharmacology.

Three common features underlying the whole thesis:

1. *Longitudinal data*: all the variables of interest have a *time* dimension and another dimension, called *unit* dimension, representing countries of the euro area (**Chapter 3**); sectors of an economy (**Chapter 4**); or patients enrolled (sequentially) in a clinical trial (**Chapter 5** and **Chapter 6**).

2. *Latent variable models*: all the proposals are based on latent variables models, which can be regarded -within a Bayesian framework- as a hierarchical model. Throughout this thesis, the first level of the hierarchy always specifies a conditional Gaussian model. The second level defines the model for the latent process under two alternative assumptions, i.e. with Markovian evolution or without any temporal dynamics. The population distribution of the latent process across units is mainly assumed to be unknown and random. When this is the case, a nonparametric prior is then assigned on it, representing the nonparametric feature of the models. The further stages of the hierarchy define priors and hyper-priors and are application-specific.

3. *Practical motivations and foundational arguments*: the Bayesian approach underlines the whole thesis. The practical motivation is to provide a framework where making an explicit and efficient use of the available context-specific prior information, which provides a complementary source of information to the available data. The foundational arguments of the Bayesian approach rely on the Bayes' rule, which connects the likelihood function with the posterior distributions, which, together with the prior distribution, define the procedure of acquisition of the information.

The thesis consists of three parts: theoretical aspects, balance sheet analysis and pharmacokinetics-related problems. Each part is made up of two chapters.

- **Part I** covers the theoretical aspects representing the skeleton for the other two parts. It includes a review of the literature, in **Chapter 1**, and the proposal of a novel prior for Bayesian nonparametric inference, in **Chapter 2**.
  - **Chapter 1** introduces the general problem and briefly reviews the main existing procedures related with the next chapters.
  - **Chapter 2** proposes a new nonparametric process, called Enriched Dirichlet Process (EDP). The EDP is a methodological contribution carried out with Sara Wade and Sonia Petrone. Its development began during the second year of the PhD as an assignment for the course of Mixture Model, taught by Sonia. One of the main drawbacks of the Dirichlet Process (DP) model is that it implicitly assume a global clustering. The EDP has been developed to overcome this and other weaknesses of the DP model in a multivariate environment, moving from the notion of enriched conjugate priors for parametric models. Starting with this idea of Sonia, Sara and I had daily worked together (for roughly a semester), under the

supervision of Sonia, to define this process and its properties. The main results have been written in a paper, whose content is basically reported in the **Chapter 2**. It was presented as a poster at the Ninth Valencia International Meeting on Bayesian Statistics on 2011. One year later, it was published on *Bayesian Analysis* with the title *An enriched conjugate prior for Bayesian nonparametric inference*, receiving the Lindley Prize 2010 (ISBA).

This process is used as a nonparametric prior for Bayesian inference in **Chapter 3** and **Chapter 5**.

- **Part II** deals with two applications related with balance sheet analysis. Both chapters make use of the financial accounts for the euro area and have been substantially carried out during my stay at the European Central Bank (ECB), Division of Macroeconomics Statistics. These two chapters are based on two working papers which have been written with Celestino Giron[1], and they are going to be submitted to the ECB working papers (hopefully) before the end of 2012. Moreover, **Chapter 3** has already been submitted to the *Proceedings of the Flow-of-Funds expert meeting*, held at the ECB on November 25, 2011.

  In any case, the views expressed in these chapters are those of myself and do not necessarily represent the views of, and should not be attributed to, the ECB.

  - **Chapter 3** focuses on the analysis of the leverage-ratio, defined as liabilities over assets ratio, across ten euro area countries. Most of the economic issues have been covered by Celestino Giron, whose contributions range from the definition of the set of covariates that could have an impact on the leverage, proposing to "split" the assets growth rate into the

[1] Principal economist-statistician at the European Central Bank.

notional assets growth rate and the price index, to the economic interpretation of the results. I was mainly responsible for the statistical definition of the models, i.e. the choice of using a Bayesian nonparametric random coefficients model, the inclusion of the break associated with Lehman-Brothers bankruptcy and the numerical implementation of the inference routine. In particular, the original (statistical) contributions are the following: first, the Bayesian nonparametric extension of the dynamic panel model with heterogeneity for the autoregressive coefficient, based on the DP prior on the mixing distribution of the country-specific coefficients; second, the inclusion of a structural break (associated with the Lehman-Brothers bankruptcy) for the distribution of a country-specific coefficient and the choice of including this break in the probabilistic prior using the EDP prior.

– **Chapter 4** focuses on making inference on the unknown portfolio composition and fitting the other volume changes (OVCs) of financial instruments on the asset side, i.e. securities other than shares, for institutional sectors. As for **Chapter 3**, Celestino Giron was mainly involved into the economic issues of the models, e.g. the definitions of the potential issuers for each sectors and the definition of the observational equation, whose dynamic has an economic justification, and the definition of the problem. I was mainly responsible for the statistical construction of the models, i.e. the choice of the model (a conditional state-space model), the proposal of the different approaches for making inference and the numerical implementation. Up to now, the actual reporting of the data results, as is in **Chapter 4**, has been carried out by myself, but the final version of the paper could bring significative improvements in this regards by the future contributions of Celestino

Giron. In particular, the (statistical) original contributions are the following. First, a procedure for estimating the unknown compositional data based on a conditional state-space model, independently for each sector. Its implementation is based on a particle filter. Second, the joint estimation of all the institutional sectors by including a resampling scheme within the particle filter algorithm that restricts the support of the filtering distributions of the portfolio structures whenever a constraint over the whole economy is not satisfied. Third, the use of a sub-series of available data on the portfolio structure for estimating of the parameters of the conditional state-space model by minimizing the conditional aggregate expected loss associated with the deviations between the estimated portfolio structures and the sub-series of available data. Fourth, a 3-step procedure for estimating the OVCs. This is the only chapter without any nonparametric feature, since, in our opinion, the parametric restrictions do not seem to be too stringent for this application.

- **Part III** is made up of two chapters related with pharmacokinetics (PK), a branch of the pharmacology dealing with the effects of administered doses of a new compound during the early phases of a clinical trial. PK focuses on the link between an administered dose and its concentration at the effective site. Both chapters propose novel procedures for making better use of all available PK information, together with another pharmacological source of information, i.e. pharmacodynamics (PD) and toxicology respectively. In **Chapter 6**, the PK model -together with a toxicity model- is used as a part of a decisional framework for sequential searching for an optimal dose.The unit dimension represents the patients enrolled in a clinical trials.

I was introduced to PK-PD problems during a 3-month summer internship at the Novartis, at the department of Modeling & Simulation (Basel, CH). Both chapter has been carried out by myself under the supervision of Sonia Petrone, who has helped me to identify the correct final aim of the analysis and, especially in **Chapter 6**, contributed to the definition of the model.

- **Chapter 5** focuses on studying population PK and PD problems. Two are the original contributions. First, a simple prior specification for the covariance matrix is proposed, based on the scaled Inverse-Wishart but replacing the Inverse-Wishart distribution with a d-Inverse-Gamma distribution. Second, we propose a nonparametric extension using the EDP prior on the random probability measure associated with the patient-specific random coefficients of population PK-PD models.
- **Chapter 6** focuses on the use of PK information within dose-finding problems. The proposal is a new Bayesian predictive approach for solving the problem of finding the MTD that allows for controlling the safety in terms of predictive toxicity. The novelty is the incorporation of the temporal evolution of the toxicity associated with the administration of a dose, modelled by allowing the toxicity to depend on the PK. Several possible scenario for the distribution of the heterogeneity across patients are taken into consideration. Analytical and computational solutions are provided.

Part I

Theoretical aspects: review and new proposal

Chapter 1

Bayesian semiparametric inference for longitudinal data: review of selected topics

Introduction

This chapter aims at giving a brief review of the literature and introducing the general model which is the basis of the developments in the thesis. It is organized as follows. Section 1.2 defines the common framework underlying the whole thesis in terms of the features of the data taken into consideration and the general model. More specifically, the general model is a 3-stage hierarchical model and it is discussed in Subsection 1.2.1. The semiparametric nature of the most general problem studied in the thesis is underlined in Subsection 1.2.2. Subsection 1.2.3 focuses on the traditional nonparametric approach for inference on the distribution

of the random coefficients, based on the Dirichlet Process prior. It follows a discussion on the main specific models treated throughout this thesis. First, models with randomly-distributed heterogeneity are discussed in Section 1.3. Then, Section 1.4 covers models with a Markovian latent process, i.e. state-space models. The chapter concludes with a brief review of some standard computational techniques, heavy used throughout the thesis.

1.1 A note on the notation

All the random quantities are denoted with bold font letters, e.g. $\boldsymbol{y}$ is random. Deterministic values, both observations and known coefficients, are denoted with non-bold font letters, e.g. y is known.

The focus of the thesis is on longitudinal data of the kind $y_{i,t}$, where i is the label associated with the unit index, $i = 1, .., N$, and t is the label associated with the time dimension, $t = 1, .., T_i$ for the ith unit. The random variable $\boldsymbol{y}_{i,t}$ refers to the ith unit and the tth time. The notation $\boldsymbol{y}_{i,1:t}$ stands for the set $\left\{\boldsymbol{y}_{i,1}, .., \boldsymbol{y}_{i,t}\right\}$. Similarly, $\boldsymbol{y}_{1:N,t}$ stands for the set $\left\{\boldsymbol{y}_{1,t}, .., \boldsymbol{y}_{N,t}\right\}$.

Due to the use of two subscripts for the indices, widely used, and in order to avoid notational confusion, in **Chapter 3** and **Chapter 4**, indices and level-variable associated with the indices will have a different notation. In particular, the unit-dimension i will not be the first subscript but the superscript. Hence, $\overset{\bullet}{p}^{i}_{t-1,t}$ denotes a price index for the time interval $[t-1, t)$ for the unit i. Similarly, p^i_t denotes the level of price at time t for the unit i.

1.2 General aspects of the thesis

Two elements that underlying the whole thesis are the use longitudinal data and the general model. Longitudinal data are common in many fields, and, for this reason, there are several names usually associated with the same kind of data sets, e.g. panel, or longitudinal, or temporal

cross-sectional, or repeated measures, or clustered, data sets[1]. In the following, the term *longitudinal data* is used for denoting a data set when an observable variable is measured at several time points for each unit, and the term *panel data* is reserved for longitudinal economic data. All the data set used in this thesis have the same number of observations over time per each unit, i.e. $T_i \equiv T$, which are called *balance* panel data in economics. Moreover, T is relatively small, implying that unit-by-unit inference is associated with not reliable estimation. At the other extreme, each unit has its own peculiarity, hence its own parameters, which would not be captured by a pooled analysis. Such data therefore require a procedure that lies in the middle between a pooled analysis and a unit-by-unit analysis. We propose the inclusion of some dependence across units by using the general model is explained in the next section.

1.2.1 General model

Let us consider a dataset made up of observations, $y_{i,t}$, of a variable of interest, e.g. the leverage-ratio, where $i = 1, .., N$ are the available units, e.g. labels associated with euro area countries, and the time index $t = 1, .., T$, e.g. from year 1999 to 2011. Let us regard $y_{i,t}$, $i = 1, .., N$, $t = 1, .., T$ as realizations from the following model for $\boldsymbol{y}_{i,t}$:

$$\boldsymbol{y}_{i,t} = f\left(\boldsymbol{\theta}_{i,t}, \boldsymbol{y}_{i,1:t-1}, \boldsymbol{X}_{i,t}, \boldsymbol{\mu}\right) + \boldsymbol{\epsilon}_{i,t} \qquad i \geq 1, t \geq 1 \tag{1.1}$$

where:

[1]Some of these names have some little differences, i.e. *repeated measures data* means that the observable variable is measured more than once for each unit whereas *longitudinal data* means the observable variable is measured at several time points for each unit, often over a relatively long period of time (West *et al.*, 2007). The term *panel data* comes instead from the terminology used in surveys of individuals, where a *panel* is a group of individuals surveyed repeatedly over time and, although historically reserved to labor economics applications, it nowadays refers to a general use of longitudinal data techniques applied to any branch of economics. The term *multilevel data* is used to refer to a wider category of data sets, i.e. data with possibly more than two dimensions and the time dimension is not necessarily one of them. It is mainly used in the social science literature.

- $\boldsymbol{y}_{i,t}$ is the univariate observable random variable.
- f is a deterministic function. It can be a linear or nonlinear function. In the latter case, it can have an analytical expression or be the numerical solution of ordinary differential equations.
- $\boldsymbol{\theta}_{i,t}$, for each fixed i and t, is a k-dimensional latent vector, expressing some latent random features of $\boldsymbol{y}_{i,t}$.
- $\boldsymbol{X}_{i,t}$ is a $m \times 1$ vector of covariates. They can represent a set of stochastic or deterministic regressors. In all the applications in the following chapters, the regressors are predetermined.
- $(\boldsymbol{\epsilon}_{i,t}, t \geq 1)$ is a Gaussian white noise, i.e. $\boldsymbol{\epsilon}_{i,t} \mid \boldsymbol{\sigma}_i^2 \overset{\text{iid}}{\sim} N\left(0, \boldsymbol{\sigma}_i^2\right)$, and it is independent from all the other random quantities.
- $\boldsymbol{\mu}$ is a p-dimensional vector of population parameters.

Model (1.1) is often called a *latent variables model* because it expresses the observable variables, $\boldsymbol{y}_{i,t}, i \geq 1, t \geq 1$, in terms of the latent process $(\boldsymbol{\theta}_{i,t}, i \geq 1, t \geq 1)$. Equation (1.1) hides the dependence structure across $\boldsymbol{y}_{i,t}$. The standard assumption, both in a classical and in a Bayesian approach, is that the observable $\boldsymbol{y}_{i,t}$ are independent conditionally on the process $(\boldsymbol{\theta}_{i,t}, i \geq 1, t \geq 1)$. The main difference between the Bayesian and the classical approach concerns the probability assumptions on the process $(\boldsymbol{\theta}_{i,t})$. In a Bayesian approach, one defines a model for $(\boldsymbol{\theta}_{i,t})$ conditionally on its unknown and random parameters. Then, one assigns on the latter a prior distribution at the next level of the hierarchy, allowing the incorporation of context-specific prior information, such as expert opinions. In a classical approach, the probability law of the process $(\boldsymbol{\theta}_{i,t})$ can have some unknown parameters too. However, these unknown parameters are regarded as unknown constants and therefore one defines a model for $(\boldsymbol{\theta}_{i,t})$ without conditioning on these unknown parameters. The Bayesian model resulting by combining (1.1), the model for $(\boldsymbol{\theta}_{i,t})$ and

the model for all the unknown parameters is often called a *hierarchical model*, because each of these models can be seen as a subsequent level of a hierarchy, specifying a probabilistic assumption conditionally on all the subsequently-modelled random quantities.

Let us consider a 3-stage hierarchical model. At the **first level** of the hierarchy, model (1.1) is defined, conditionally on the latent process and on all the random parameters. More precisely, we assume:

$$\boldsymbol{y}_{i,t} \mid \boldsymbol{\theta}_{i,t}, \boldsymbol{y}_{i,1:t}, \boldsymbol{X}_{i,t}, \boldsymbol{\mu}, \boldsymbol{\sigma}_i^2 \overset{indep}{\sim} N\left(f\left(\boldsymbol{\theta}_{i,t}, \boldsymbol{y}_{i,1:t-1}, \boldsymbol{X}_{i,t}, \boldsymbol{\mu}\right), \boldsymbol{\sigma}_i^2\right) \quad (1.2)$$

The Gaussian assumption will be used throughout the whole thesis. At the **second level** of the hierarchy, a model for the latent process $(\boldsymbol{\theta}_{i,t})$, conditionally on the random parameters, is defined. Two alternative options are taken here into consideration.

1. In some applications, it is reasonable to assume that $\boldsymbol{\theta}_{i,t}$ does not vary with time, i.e. $\boldsymbol{\theta}_{i,t} \equiv \boldsymbol{\theta}_i$. In this case, a common assumption is to regard the $\boldsymbol{\theta}_i, i = 1, 2, ..$ as a random sample from an unknown population distribution, denoted by $\boldsymbol{P}_\theta$. Due to the wide use of this approach in many different fields, several names are associated with this distribution: *population distribution*, *latent distribution* or *mixing distribution*. We will mainly call it *population distribution*. This level of the hierarchy accounts for the heterogeneity across the units (mainly) caused by unknown (unobserved or unobservable) covariates, assigning a model for the unit-specific latent vectors $\boldsymbol{\theta}_i$. A misspecified model for $\boldsymbol{\theta}_i$ can produce unreliable results. For instance, whenever the population is not homogeneous across units, imposing $\boldsymbol{\theta}_i \equiv \boldsymbol{\theta}$, for every i, can even produce inconsistent results (see e.g. Pesaran and Smith, 1995). The standard treatment of the cross-unit heterogeneity depends on the field of application. In some areas, the cross-unit heterogeneity expresses some specific and desirable feature and it deserves a particular term. For instance,

the unobserved heterogeneity in hazard models is called *frailty*. The basic idea is that individuals have different frailties. The most frail will die earlier than the least frail (Vaupel *et al.*, 1979). In pharmacokinetic models, the unobserved heterogeneity is sometimes called *individual pharmacokinetics* (Riddell *et al.*, 2006). Instead, in other fields, such as in econometrics, unobservable heterogeneity is mainly a cause of concern due to the related inconsistency issues for estimation. In the following, two alternative assumptions for modelling the cross-unit heterogeneity are considered: a parametric population distribution and a nonparametric prior on the unknown and random population distribution.

1.a) A parametric model for the population distribution assumes:

$$\boldsymbol{\theta}_i \mid \boldsymbol{\beta} \overset{\text{iid}}{\sim} P_\theta\left(\cdot \mid \boldsymbol{\beta}\right) \tag{1.3}$$

where $\boldsymbol{\beta}$ is a $R \times 1$ vector of parameters. A common parametric assumption is that $P_\theta\left(\cdot \mid \boldsymbol{\beta}\right)$ is a Gaussian distribution. In the following chapters will often be the starting point of the analysis, but we will find the need for nonparametric extensions, as in **Chapter 3** and **Chapter 5**. In other applications, the inclusion of a set of covariates, $\boldsymbol{Z}_i$, can lead to a model that is accurate enough, without requiring a nonparametric extension, as discussed in **Chapter 6**. Model (1.3) can then be replaced by:

$$\boldsymbol{\theta}_i \mid \boldsymbol{\beta}, \boldsymbol{Z}_i \overset{indep}{\sim} P_\theta\left(\cdot \mid \boldsymbol{\beta}, \boldsymbol{Z}_i\right) \tag{1.4}$$

where, for instance, $P_\theta\left(\cdot \mid \boldsymbol{\beta}, \boldsymbol{Z}_i\right) = N\left(\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \boldsymbol{Z}_i, \tau^2\right)$, $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \tau^2\right)$.

1.b) However, parametric assumptions may be restrictive; for example, they would not capture the possible presence of outliers

or subgroups among the $\boldsymbol{\theta}_i$'s. A Bayesian nonparametric approach assumes that:

$$\boldsymbol{\theta}_i \mid \boldsymbol{P}_\theta \overset{\text{iid}}{\sim} \boldsymbol{P}_\theta, \tag{1.5}$$

and assigning a prior on the random population distribution $\boldsymbol{P}_\theta$, which is not restricted to a parametroc form. We will develop a Bayesian nonparametric approach in **Chapter 3**, **Chapter 5** and (shortly) in **Chapter 6**.

Models (1.3) and (1.5) will be further discussed in Section 1.3.

2. In other applications, it is appropriate to allow for a temporal evolution of the unit-specific parameters. A wide class of latent variables models in this context is provided by state-space models, where the process $(\boldsymbol{\theta}_{i,t}, t \geq 1)$ is assumed to be Markovian (conditionally on the random transition matrix) with:

$$\boldsymbol{\theta}_{i,t} \mid \boldsymbol{\theta}_{i,1:t-1}, \boldsymbol{\beta} \sim P_\theta\left(\boldsymbol{\theta}_{i,t} \mid \boldsymbol{\theta}_{i,t-1}, \boldsymbol{\beta}\right) \quad \text{for each } i = 1, .., N. \tag{1.6}$$

where $\boldsymbol{\beta}$ are random parameters of the transition density, P_θ. Model (1.6), with a Gaussian P_θ, is the basis of the study in **Chapter 4**, where also some dependence across i is allowed.

Bayesian and classical inference depart in the estimation of the parameters involved in the first two levels of the hierarchy. Within the classical setting, depending on the focus of the analysis, i.e. whether estimating only the population parameters or all the unit-specific parameters, different procedures are available. For instance, under assumptions, one can use least square estimators or can compute the maximum likelihood estimates by combining these two levels and exploiting independence assumptions. A review of classical methods will be given in Section 1.3.1. In the thesis, instead, we follow a Bayesian approach, according to which the unknown parameters are assumed to be random and are given a prior

distribution. This constitutes the **third level** of the hierarchy, which is therefore the aspect of the model which is specifically Bayesian. Usually, conjugate priors are assigned on the population parameters, $\boldsymbol{\mu}$ (and $\boldsymbol{\sigma}^2$) and, for model (1.3), on $\boldsymbol{\beta}$. For models of the form (1.5), the unknown and random population distribution of the cross-unit latent vectors is given a nonparametric prior, either a Dirichlet Process prior or an Enriched Dirichlet Process prior in our applications. This is the nonparametric feature of the models used throughout this thesis. For models of the form (1.6), a prior is also assigned on the parameters of the Gaussian transition distribution, i.e. $\boldsymbol{\beta} \sim h(\boldsymbol{\beta})$.

Two important categories of distributions are therefore taken into account: *prior distributions* and *latent population distributions*. The *prior distribution* expresses subjective beliefs about a fixed coefficient. The *population distribution* represents the distribution of the state parameters across units and expresses how the unit-specific behavior deviates from the population mean.

1.2.2 Bayesian semiparametric inference

As outlined in the previous section, most problems studied in the thesis have a semiparametric nature, involving a parameter of the form $(\boldsymbol{\mu}, \boldsymbol{P}_\theta)$, where $\boldsymbol{\mu}$ is a *finite-dimensional vector*, e.g. $\boldsymbol{\mu}$ contains the coefficients of the model, let us say $\boldsymbol{\beta}$, and the observational variances, let us say $\boldsymbol{\sigma}^2$, and $\boldsymbol{P}_\theta$ is an *infinite-dimensional parameter*, corresponding to the latent population distribution. The need of regarding the unit-specific parameters as a sample from a population distribution, possibly not restricted to a parametric form, is well established in the literature, also in the maximum likelihood estimation approach. Let us consider models with randomly-distributed heterogeneity with no dynamics, i.e. $\boldsymbol{\theta}_{i,t} = \boldsymbol{\theta}_i$. Maximum likelihood estimation for models of the form (1.2), with no probability law on $\boldsymbol{\theta}_i$, would lead to inconsistent results due to the infinitely many *nuisance* parameters. The nuisance

parameters problem was first discussed by Neyman and Scott (1948). They pointed out that, in models of the form $\boldsymbol{y}_{i,t} \sim f\left(\boldsymbol{y}_{i,t} \mid \boldsymbol{\theta}_i, \sigma^2\right)$, $i = 1, .., N$, $t = 1, .., T$, the maximum likelihood estimates (MLEs) of the finite-dimensional parameter, σ^2, can be inconsistent or inefficient as $N \to +\infty$. Kiefer and Wolfowitz (1956) showed that regarding the nuisance parameters $\boldsymbol{\theta}_i$ as random variables with a common unknown distribution function, P_θ, one can consistently estimate not only the finite-dimensional parameter but also the distribution function P_θ. In particular, they considered maximum likelihood estimation for σ^2 and P_θ and showed that, under regularity conditions, the semiparametric MLEs for both σ^2 and P_θ are strongly consistent. The Bayesian approach fits particularly well in this context, allowing to take advantage of the regularization properties ensured by assigning a prior distribution on σ^2 and P_θ. Consider, for instance, a random variable $\boldsymbol{y}_i$ distributed as a mixture of two Gaussian distributions: $\boldsymbol{y}_i \sim wN(\mu, 1) + (1-w) N\left(\mu, \sigma^2\right)$. Given a sample $y_1, .., y_N$, assuming w fixed and μ and σ^2 unknown, Kiefer and Wolfowititz (1956) showed that the contribution to the likelihood of each observation, i.e. $w\phi(y_i \mid \mu, 1) + (1-w)\phi\left(y_i \mid \mu, \sigma^2\right)$, is unbounded and has many local maximum. In particular, setting $\mu = y_i$, then $\lim_{\sigma^2 \to 0}\left(w\phi(y_i \mid y_i, 1) + (1-w)\phi\left(y_i \mid y_i, \sigma^2\right)\right) = +\infty$. Now, if the model for the data is combined with a prior for the parameters, e.g. $p\left(\mu, \sigma^2\right) \propto p\left(\sigma^2\right)$ where $\sigma^2 \sim IG(a, b)$, a, b known, the variance can be kept sufficiently far from zero, solving the unbounded problem. Looking the parameters as random variables with some priori distributions is one of the foundational principles of the Bayesian approach. The notion of a random sample from P_θ is then replaced by the assumption of exchangeability or, equivalently, $\boldsymbol{\theta}_i$ are conditionally i.i.d. given P_θ, with common distribution P_θ. In the Bayesian approach, inference in semiparametric models requires a prior on σ^2 and P_θ, and it is solved by the conditional distribution of $\left(\sigma^2, P_\theta\right)$ given the data. In some applications, there is enough information to specify a parametric model for the population dis-

tribution, say $\boldsymbol{\theta}_i \mid \boldsymbol{\beta} \overset{\text{iid}}{\sim} P_\theta(\cdot \mid \boldsymbol{\beta})$, so that the prior on P_θ reduces to a prior on the finite-dimensional hyperparameter $\boldsymbol{\beta}$. Although this approach is very appealing, because it allows both for random variability across units and for borrowing information, parametric assumptions can be quite restrictive. For example, a commonly-used Gaussian population distribution has light tails and is unimodal, and it does not allow for outliers or clustering. Even a heavy-tailed distribution, such as the t-distribution, has a restrictive unimodal and symmetric shape. This motivates a semiparametric approach where P_θ is not restricted to a parametric class, but instead it is the infinite-dimensional unknown "parameter" to be estimated. As previously said, a Bayesian approach requires a prior on P_θ. The most popular nonparametric prior is the Dirichlet Process (Ferguson, 1973). Early applications in the context of random effects models are in Antoniak (1974), Berry and Christensen (1979), Cifarelli, Muliere and Scarsini (1981), Lo (1984). These Bayesian nonparametric models incorporate infinitely-many parameters within a prior that is centered on a base parametric model. Therefore, the approach adds flexibility over the population distribution, particularly important at the initial exploratory data analysis stage (see e.g. Walker and Wakefield, 1997). Obviously, the pro's and con's of the nonparametric approach are the opposite of those for parametric modelling. On one hand, nonparametric modelling typically imposes a few restrictions on the form of the distribution of the random coefficients, e.g. smoothness or monotonicity. On the other hand, the price one has to pay to relax the parametric assumptions is basically given by the computational (intensive) estimation procedure, which can bring to less precision estimates than the ones obtained with (well-suitable) parametric assumptions, and the possible reduction of parameters, which can bring to the impossibility of interpreting some of the relevant parameters of the model.

1.2.3 The Dirichlet process

The Dirichlet process (DP) is the most commonly used stochastic process in Bayesian nonparametric models and it has been introduced by Ferguson (1973). It is called Dirichlet process because it has Dirichlet distributed finite-dimensional distributions. It is characterized by measure-valued parameter, say α, which is usually split into two components: a positive real scalar, say α_0, called the precision parameter, and a distribution function P_0, called baseline distribution and representing the center of the process.

Definition Let α be a finite measure on a given Polish space $\boldsymbol{\Omega}$. A random measure $\boldsymbol{P}_\theta$ on $\boldsymbol{\Omega}$ is called a Dirichlet Process if for every finite measurable partition $C_1, ..., C_k$ of the space $\boldsymbol{\Omega}$, the joint distribution of $(\boldsymbol{P}_\theta(C_1), \boldsymbol{P}_\theta(C_2), ..., \boldsymbol{P}_\theta(C_k))$ is a finite-dimensional Dirichlet distribution with parameters $(\alpha(C_1), ..., \alpha(C_1))$.

Some important properties of the Dirichlet Process are briefly reported.

- If $\boldsymbol{P}_\theta \sim DP(\alpha_0 P_0)$, then, for any measurable set A,
  - $E(\boldsymbol{P}_\theta(A)) = P_0(A)$.
  - $V(\boldsymbol{P}_\theta(A)) = \dfrac{P_0(A)(1 - P_0(A))}{\alpha_0 + 1}$.
- As the finite-dimensional Dirichlet distribution is conjugate to the multinomial likelihood, the DP prior is conjugate for estimating an unknown distribution from exchangeable data. More precisely, if $\boldsymbol{\theta}_1$,..,$\boldsymbol{\theta}_n$, given $\boldsymbol{P}_\theta = P_\theta$, are i.i.d with distribution P_θ and $\boldsymbol{P}_\theta \sim DP(\alpha_0, P_0)$, then the posterior distribution of $\boldsymbol{P}_\theta \mid \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_n$ is a DP with updated parameter $\alpha_0 P_0 + \sum_{i=1}^{n} \delta_{\theta_i}$, $DP\left(\alpha_0 P_0 + \sum_{i=1}^{n} \delta_{\theta_i}\right)$. In particular, its posterior mean is given by:

$$E(\boldsymbol{P}_\theta \mid \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_n) = \frac{\alpha_0}{\alpha_0 + n} P_0 + \frac{n}{\alpha + n} P_n,$$

where $P_n = \frac{\sum_{i=1}^{n} \delta_{\theta_i}}{n}$ is the empirical distribution. Thus the posterior mean shrinks the empirical distribution towards the prior mean.

- A very important property of the DP is that it selects discrete distributions with probability one (Blackwell and MacQueen, 1973), even if P_0 is an absolutely continuous distribution. The discreteness property is a consequence of the presence of the atomic component, given by P_n, in the base measure of the posterior DP. It leads to ties in a random sample from $\boldsymbol{P}_\theta$.

The last-mentioned property, often referred as the *clustering property*, reveals to be quite powerful in applications of the DP in Bayesian mixture models. Let us consider the simplest model for the observations $\boldsymbol{y}_{i,t}$, i.e. an intercept-only model, described by equation (1.7) below for repeated measurements. Let us assign a DP prior on the random population distribution of the $\boldsymbol{\theta}_i$ through the following hierarchical specification:

$$\begin{pmatrix} \boldsymbol{y}_{i,1} \\ ... \\ \boldsymbol{y}_{i,T} \end{pmatrix} \mid \boldsymbol{\theta}_i \overset{indep}{\sim} N\left(\boldsymbol{\theta}_i, \sigma^2 I_T\right) \tag{1.7}$$

$$\boldsymbol{\theta}_i | \mathbf{P}_\theta = P_\theta \sim P_\theta \tag{1.8}$$

$$\mathbf{P}_\theta \sim DP(\alpha_0 P_0) \tag{1.9}$$

where I_T is the $T \times T$ identity matrix. Integrating out $\mathbf{P}_\theta$, the parameter $\boldsymbol{\theta}_i$ follows a Pòlya urn distribution (Blackwell and MacQueen, 1973):

$$\boldsymbol{\theta}_1 \sim P_0, \tag{1.10}$$

$$\text{for } n > 1, \quad \boldsymbol{\theta}_{n+1} | \boldsymbol{\theta}_{1:n} \sim \sum_{k=1}^{d_n} \frac{n_k}{i + \alpha_0} \delta_{\boldsymbol{\theta}_k^*} + \frac{\alpha_0}{n + \alpha_0} P_0, \tag{1.11}$$

where $\boldsymbol{\theta}_k^*$ denotes the kth distinct value among $\boldsymbol{\theta}_1$,..,$\boldsymbol{\theta}_n$, $k = 1, 2, .., d_n$ and n_k is the number of parameters $\boldsymbol{\theta}_1$,..,$\boldsymbol{\theta}_n$ having value $\boldsymbol{\theta}_k^*$.

The distribution of Blackwell and MacQueen corresponds to the law defining the random partition implied by the Chinese Restaurant Process (Aldous, 1985; Pitman, 1996). The Chinese Restaurant Process (CRP) is a random process in which customers sit down in a Chinese restaurant with an infinite number of tables, accordingly to the following rule:

- The first customer sits at the first table with probability one.
- The $(n+1)$th customer sits at a table according to the following distribution:

$$Prob\left((n+1)\text{th customers sits at the previously occupied table } k \mid \mathcal{F}_n\right) = \frac{n_k}{n+\alpha_0},$$

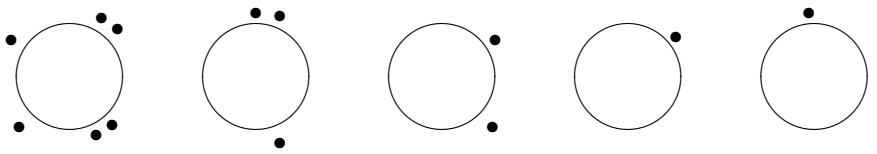
$$\text{for } k = 1, .., d_n;$$

$$Prob\left((n+1)\text{th customers sits at the next unoccupied table } d_n + 1 \mid \mathcal{F}_n\right) \propto \alpha_0$$

  where n_k is the number of customers currently sitting at table k, $k = 1, .., d_n$; d_n is the number of tables occupied by the first n customers; $\mathcal{F}_n$ denotes the situation after n customers have been seated.

After n customers sit down, the seating plan gives a partition of the n customers among the (random number of) tables. Picture 1 shows a possible partition of the customers, where each data point represents a customer and each table represents a cluster.

Picture 1: Dirichlet Process: Clustering - Chinese Restaurant Process



If one colors the tables with colors $\boldsymbol{\theta}_k^*$ drawn at random from a nonatomic P_0, and let $\boldsymbol{\theta}_i$ be the color of the table where the ith customer

sits in, then the resulting vector $(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_n)$ corresponds to a sample from the Pòlya urn distribution (Blackwell and MacQueen, 1973) in equation (1.11). Thus, the CRP defines a random partition, with an exchangeable distribution, which is exactly the same partition structure associated with draws from a Dirichlet process.

The nonparametric model (1.8)-(1.9) is appealing for at least three reasons. Two of them are also desirable features of the parametric model defined in (1.3). First, random effects and random coefficients introduce dependence across the observations, allowing to borrow strength across units. The resulting model lies therefore in the middle between the two extremes of independent unit-by-unit analysis and pooled analysis. Second, the Bayesian approach allows to incorporate prior information within the general framework. Due to the small T, prior information can have a crucial role to make reliable inference. The third property of the model is specific to the nonparametric setting defined by model (1.8)-(1.9), which allows to model clustering among the unit-specific coefficients while relaxing standard parametric assumptions.

One drawback of this approach is that, if $\boldsymbol{\theta}_i \in \mathbb{R}^k$, $k > 1$, a DP prior on its random probability measure implies the same clustering structure for all the k elements of $\boldsymbol{\theta}_i$. Let us consider $\boldsymbol{\theta}_i \equiv \left(\boldsymbol{\theta}_i^1, \boldsymbol{\theta}_i^2\right)'$, where $\boldsymbol{\theta}_i^1$ is a $k_1 \times 1$ and $\boldsymbol{\theta}_i^2$ is a $k_2 \times 1$, $k_2 \equiv k - k_1$ and call $\mathbf{P}_{\theta^1}, \mathbf{P}_{\theta^2}$ their corresponding random probability measures. Then, an alternative to assigning a single multivariate DP prior to $\mathbf{P}_\theta$ is the following set of priors:

$$\boldsymbol{P}_{\theta^1} \perp\!\!\!\perp \boldsymbol{P}_{\theta^2} \text{ with } \boldsymbol{P}_{\theta^i} \sim DP\left(\alpha_i\right), \; i = 1, 2. \tag{1.12}$$

Model (1.12) implies two independent clustering structures for $\boldsymbol{\theta}_i^1$ and $\boldsymbol{\theta}_i^2$. Sometimes, having only one global clustering structure for all the k elements of $\boldsymbol{\theta}_i$, as implied by model (1.9), or two (or more) independent clustering structures for sub-groups of $\boldsymbol{\theta}_i$, as implied by model (1.12), can be adequate. However, one may often expect to have sub-groups of

$\boldsymbol{\theta}_i$ with different but *dependent* clustering structures. **Chapter 2** goes in this direction and it defines an alternative nonparametric prior on $\mathbf{P}_\theta$, called the Enriched Dirichlet Process (EDP) prior. The EDP implies a hierarchical clustering structure between $\boldsymbol{\theta}_i^1$ and $\boldsymbol{\theta}_i^2$. The idea is to assign independent DP priors to the marginal, $\mathbf{P}_{\theta^1}$, and to the family of conditionals, $\mathbf{P}_{\theta^2|\theta^1}(\cdot|\theta^1)$ indexed by θ^1. The resulting clustering structure can be interpreted in terms of the CRP as follows. There is an infinite number of Chinese restaurants. First, each customer chooses a restaurant according to the CRP rule. Restaurants are painted with colors $\boldsymbol{\theta}_j^{1*}$. This gives the clustering of the subgroup of parameters $\boldsymbol{\theta}_i^1$. Then, within each restaurant, each customer sits down to a table according to the CRP. Tables are painted with colors $\boldsymbol{\theta}_i^{2*}$. This gives the clustering of the subgroup of parameters $\boldsymbol{\theta}_i^1$.

1.3 Heterogeneity among units

This section covers more in details models with randomly-distributed heterogeneity without dynamics, that is to say $\boldsymbol{\theta}_{i,t} \equiv \boldsymbol{\theta}_i$, introduced in Subsection 1.2.1. This topic, already mentioned in Subsection 1.2.1, is here deeper discussed, providing a brief overview within both the classical and the Bayesian setting.

The basic framework for the current discussion can be defined using a linear regression model of the form:

$$\boldsymbol{y}_{i,t} = \boldsymbol{X}_{i,t}^{\prime(1)}\boldsymbol{\theta}_i + \boldsymbol{X}_i^{\prime(2)}\boldsymbol{\alpha} + \boldsymbol{\epsilon}_{i,t} \tag{1.13}$$

where $\boldsymbol{y}_{i,t}$ is the univariate observable random variable, $\boldsymbol{X}_{i,t}^{\prime(1)}$ is a $k \times 1$ vector of random regressors, not including a constant term, and $\boldsymbol{X}_i^{\prime(2)}\boldsymbol{\alpha}$ is the heterogeneity or unit-specific effect with $\boldsymbol{X}_i^{\prime(2)}$ a $m \times 1$ vector containing a constant term and a set of random unit-specific variables (observed or unobserved) and constant over time t. The parameters $\boldsymbol{\theta}_i$ and $\boldsymbol{\alpha}$ are assumed to be either deterministic (within a classical approach,

and denoted by θ_i and α) or random (mainly within Bayesian approach). The error terms $\boldsymbol{\epsilon}_{i,t}$, when not otherwise specified, are assumed to be independent and identically distributed with:

$$E\left(\boldsymbol{\epsilon}_{i,t} \mid \boldsymbol{X}_{i,t}^{\prime(1)}, \boldsymbol{\theta}_i, \boldsymbol{X}_i^{\prime(2)}, \boldsymbol{\alpha}\right) = 0 \tag{1.14}$$

which implies that $\boldsymbol{\epsilon}_{i,t}$ is uncorrelated to $\left(\boldsymbol{X}_{i,t}^{\prime(1)}, \boldsymbol{\theta}_i, \boldsymbol{X}_i^{\prime(2)}, \boldsymbol{\alpha}\right)$. If $\boldsymbol{\theta}_i = \theta_i$ and $\boldsymbol{\alpha} = \alpha$ (i.e. they are deterministic), then equation (1.14) reduces to:

$$Corr\left(\boldsymbol{\epsilon}_{i,t}, \left(\boldsymbol{X}_{i,t}^{\prime(1)}, \boldsymbol{X}_i^{\prime(2)}\right)\right) = 0 \tag{1.15}$$

which is the basic assumption for obtaining consistent estimation within the classical approach. The first next subsection is about the classical approach and is mainly based on Greene (2003) and Wooldridge (2001), to which we refer the reader for a deeper discussion.

1.3.1 Classical approach

In the classical statistical treatment of heterogeneity, the coefficients (both common and unit-specific) are usually treated as unknown constants instead of random variables. Consequently, no probabilistic dependence among the observations is induced through them. The classical assumption among the observations $\boldsymbol{y}_{i,t}$ is therefore independence (conditionally on the random covariates) inherited by the independence of the residual series.

A wide range of estimation techniques are available and they differ in the choice of having or not having common coefficients across units and in the estimation procedure, e.g. running separate regressions for each unit or running a single pooled regression for all the units. In the following, some of most popular available options are briefly discussed.

Fixed Effects models

Fixing $\boldsymbol{\theta}_i = \theta$ for every i in equation (1.13) and assuming that $\boldsymbol{X}_i^{\prime(2)}$ is unobservable and correlated with $\boldsymbol{X}_{i,t}^{\prime(1)}$, it is well-known that the least

squares estimator of θ is biased and inconsistent as a consequence of an omitted variable. For instance, consider the special case in which $\boldsymbol{X}_i^{\prime(2)}\alpha = \alpha_i$:

$$\boldsymbol{y}_{i,t} = \boldsymbol{X}_{i,t}^{\prime(1)}\theta + \alpha_i + \boldsymbol{\epsilon}_{i,t} \qquad i \geq 1, t \geq 1 \tag{1.16}$$

The unit-specific parameter α_i is called *fixed effects*. The term *fixed effects* is potentially misleading and, moreover, there is not a unique definition in the econometrical literature, where α_i is both defined as a deterministic coefficient or as a random variable. In any case, the peculiarity of the fixed effects model is that the coefficient α_i is unit-specific that are not themselves modeled but it can be correlated with the regressors. The term *fixed effects* is used in contrast to *random effects*, where the unit-specific effects follow a specific model and is randomly distributed across the units.

The *Least Squares Dummy Variable (LSDV)* represents the main estimation technique for *fixed effects* models when the estimation of the fixed effects is required. Rearranging the observations in vectors, first across times, with $\mathbf{y}_i$, $\boldsymbol{X}_i^{\prime(1)}$ and $\boldsymbol{\epsilon}_i$ are the $T \times 1$ vectors collecting, respectively, $\mathbf{y}_{i,t}$, $\boldsymbol{X}_{i,t}^{\prime(1)}$ and $\boldsymbol{\epsilon}_{i,t}$, $t = 1, .., T$, for the unit i and defining ι a $T \times 1$ column of ones, and then across units, model (1.16) becomes

$$\boldsymbol{y} = \boldsymbol{X}^{\prime(1)}\boldsymbol{\theta} + D\alpha + \boldsymbol{\epsilon}. \tag{1.17}$$

where $D = [d_1 \ ... \ d_N]$ is a $NT \times N$ matrix and d_i is the dummy variable indicating the ith unit. Model (1.17) can then be estimated with Ordinary Least Squares (OLS).

Whenever the estimation of the unit-specific effects, α_i, is not required, alternative methods are available for consistently estimating the common θ. For instance, one can use the *first differencing method*, by taking the first differences of the cross-sectional model over time periods to eliminate the unobserved heterogeneity term α from the model. The

resulting model for equation (1.16) is given by:

$$\Delta \boldsymbol{y}_{i,t} = \Delta \boldsymbol{X}_{i,t}^{\prime(1)} \theta + \Delta \boldsymbol{\epsilon}_{i,t} \qquad t = 2, ..., T \tag{1.18}$$

where $\Delta \boldsymbol{y}_{i,t} = \boldsymbol{y}_{i,t} - \boldsymbol{y}_{i,t-1}$, $\Delta \boldsymbol{X}_{i,t}^{\prime(1)} = \boldsymbol{X}_{i,t}^{\prime(1)} - \boldsymbol{X}_{i,t-1}^{\prime(1)}$ and $\Delta \boldsymbol{\epsilon}_{i,t} = \boldsymbol{\epsilon}_{i,t} - \boldsymbol{\epsilon}_{i,t-1}$. In order to overcome the problem of induced correlation among the error terms, the first-difference estimator of θ is a Generalized Least Squares (GLS) estimator for the regression of $\Delta \boldsymbol{y}_{i,t}$ on $\Delta \boldsymbol{x}_{i,t}$ for $t = 2, .., T$ and $i = 1, .., N$. Call Q the $(T-1) \times T$ transformation matrix:

$$Q = \begin{pmatrix} -1 & 1 & 0.. & 0 & \\ 0 & -1 & 1 & .. & 0 \\ .. & .. & .. & .. & ... \\ 0 & 0 & ... & -1 & 1 \end{pmatrix}$$

The first-differences estimator is the GLS estimator for the regression of $Q\boldsymbol{y}$ on $Q\boldsymbol{x}$.

Alternatively, one can use a different transformation of the original model (1.16), e.g. the orthogonal deviation transformation based on the transformation matrix $A = (QQ')^{-1/2} Q$, and obtain the same estimator using an OLS on the transformed model.

Random Effects and Random Coefficients models

Random effects and random coefficients models are a category of models commonly used (also) within the classical approach. This class of models constitute the analogous, within the classical setting, of the Bayesian hierarchical models where the unobserved heterogeneity is randomly distributed across units discussed in Subsection 1.2.1. They assume a model for the heterogeneity of the form (1.3) but unknown, deterministic parameters to estimate. The main requirement is that regressors and unit-specific coefficients have to be uncorrelated, i.e. $Cov\left(\left(\boldsymbol{\alpha}_i, \boldsymbol{\theta}_i\right), \boldsymbol{X}_{j,t}^{\prime(1)}\right) = 0$ for every i and every j. Assume that, in equation (1.13), $\boldsymbol{\theta}_i = \theta$ and

$\boldsymbol{\alpha}_i$ is the random heterogeneity specific to all the observations associated with the unit i and constant over time, with $\boldsymbol{\alpha}_i = \boldsymbol{X}_i^{\prime(2)}\boldsymbol{\alpha}$. Then, equation (1.13) reduces to the following model:

$$\boldsymbol{y}_{i,t} = \boldsymbol{X}_{i,t}^{\prime(1)}\theta + \boldsymbol{\alpha}_i + \boldsymbol{\epsilon}_{i,t} \tag{1.19}$$

Equation (1.19) is called *random effects* model and it can be regarded as a linear regression model with a compound disturbance. Due to the assumption of not correlation, the parameters can be consistently estimated by OLS (albeit inefficiently because of the compound disturbance) . We will use the most general term *random coefficients models* for any sub-category of the general model with both unit-specific random intercept, $\boldsymbol{\alpha}_i$, and slope, $\boldsymbol{\theta}_i$, e.g. for random effects models and for models with only unit specific $\boldsymbol{\theta}_i$.

To illustrate some of the available estimation techniques for the *random effects* model, let us consider the simplest random effects model of the form of equation (1.19), where:

1. $E\left(\boldsymbol{\epsilon}_{i,t} \mid \boldsymbol{X}^{(1)}\right) = E\left(\boldsymbol{\alpha}_i \mid \boldsymbol{X}^{\prime(1)}\right) = 0$;
2. $E\left(\boldsymbol{\epsilon}_{i,t}^2 \mid \boldsymbol{X}^{\prime(1)}\right) = \sigma_\epsilon^2$ and $E\left(\boldsymbol{\alpha}_i^2 \mid \boldsymbol{X}^{\prime(1)}\right) = \sigma_\alpha^2$;
3. $E\left(\boldsymbol{\epsilon}_{i,t}\boldsymbol{\epsilon}_{j,s} \mid \boldsymbol{X}^{\prime(1)}\right) = 0$ for $t \neq s$ or $i \neq j$. $E\left(\boldsymbol{\alpha}_i\boldsymbol{\alpha}_j \mid \boldsymbol{X}^{\prime(1)}\right) = 0$ for $i \neq j$;
4. $E\left(\boldsymbol{\epsilon}_{i,t}\boldsymbol{\alpha}_j \mid \boldsymbol{X}^{\prime(1)}\right) = 0$ for all i, t and j.

Calling $\boldsymbol{\eta}_{i,t} = \boldsymbol{\epsilon}_{i,t} + \boldsymbol{\alpha}_i$ and $\underline{\boldsymbol{\eta}}_i = \left[\boldsymbol{\eta}_{i,1} \ldots, \boldsymbol{\eta}_{i,t}\right]'$, it follows that $E\left(\boldsymbol{\eta}_{i,t}^2 \mid \boldsymbol{X}^{(1)}\right) = \sigma_\epsilon^2 + \sigma_\alpha^2$ and

$$E\left(\boldsymbol{\eta}_{i,t}\boldsymbol{\eta}_{j,s} \mid \boldsymbol{X}^{(1)}\right) = \begin{cases} \sigma_\alpha^2, & \text{for } t \neq s \text{ and } i = j \\ 0, & \text{for all t and s if } i \neq j \end{cases}$$

For the T observations for unit i, $\Sigma = E\left(\underline{\boldsymbol{\eta}}_i \underline{\boldsymbol{\eta}}_i' \mid \boldsymbol{X}^{(1)}\right) = \sigma_\epsilon^2 I_T + \sigma_\alpha^2 \iota_T \iota_T'$, where ι_T is a $T \times 1$ column vector of one. Call the disturbance covariance matrix for the full NT observations $\Omega = I_N \otimes \Sigma$. Under this assumptions, some of the possible estimation techniques for $\boldsymbol{\theta}$ are:

- *Generalized Least Squares (GLS)*:

$$\hat{\boldsymbol{\theta}}^{RE} = \left(\boldsymbol{X}'^{(1)} \Omega^{-1} \boldsymbol{X}^{(1)}\right)^{-1} \left(\boldsymbol{X}'^{(1)} \Omega^{-1} \boldsymbol{y}\right) \tag{1.20}$$

  which is consistent and efficient, since the random $\boldsymbol{\alpha}_i$ imposes serial correlation across the error terms. This estimator is also called the *random effect estimator* for the parameter θ, and it can become feasible by estimating the disturbance variances using, for instance, the Balestra-Nerlove estimator (Balestra and Nerlove, 1966).

- *Instrument variables estimators*: Whenever the assumption of not correlation between the unobserved unit-specific effects, $\boldsymbol{\alpha}_i$, and the covariates, $\boldsymbol{X}_{i,t}^{(1)}$ is violated, instrument variable methods can be used. Hausman and Taylor's estimator (Hausman and Taylor, 1981) is an general solution for models of the form:

$$\boldsymbol{y}_{i,t} = \boldsymbol{X}_{1;i,t}'^{(1)} \theta_1 + \boldsymbol{X}_{2;i,t}'^{(1)} \theta_2 + \boldsymbol{X}_{1;i}'^{(1)} \alpha_1 + \boldsymbol{X}_{2;i}'^{(1)} \alpha_2 + \boldsymbol{\eta}_{i,t} \tag{1.21}$$

  where $\boldsymbol{\eta}_{i,t} = \boldsymbol{\epsilon}_{i,t} + \boldsymbol{\alpha}_i$. The basic idea is that there are two groups of regressors: the one uncorrelated to $\boldsymbol{\alpha}_i$, let us say $\boldsymbol{X}_{1;i,t}^{(1)}$ and $\boldsymbol{X}_{1;i}^{(1)}$, and the ones correlated to $\boldsymbol{\alpha}_i$, let us say $\boldsymbol{X}_{2;i,t}^{(1)}$ and $\boldsymbol{X}_{2;i}^{(1)}$. The procedure proposed by Hausman and Taylor (1981) consists in the following main steps:

  1. Obtain the LSDV (fixed effects) estimator for θ_1 and θ_2 based on $\mathbf{X}_1^{(1)}$ and $\mathbf{X}_2^{(1)}$.
  2. Regress $\mathbf{X}_{1;i}^{(1)}$ and $\mathbf{X}_{2;i}^{(1)}$ using $\mathbf{X}_{1;i}^{(1)}$ and $\mathbf{X}_{1;i,t}^{(1)}$ as instruments on the unit means of the residuals obtained from the first regres-

sion. The residual variance of this regression is a consistent estimator for the variance.

3. Use the residual variances in step 1 and step 2 to obtain a weight for the (Feasible) GLS and perform GLS transformation for all the variables.

4. Implement a weighted instrument variable estimator for estimating the coefficients of interest by instrumental variable regression.

- *Generalized Methods of Moments estimator*: Generalized Methods of Moments (GMM) are usually adopted with dynamic panel data, and more in general in econometric literature. Imposing $\boldsymbol{X}_{i,t} \equiv \boldsymbol{y}_{i,t-1}$ in equation (1.19), the model becomes:

$$\boldsymbol{y}_{i,t} = \boldsymbol{y}_{i,t-1}\theta + \boldsymbol{\alpha}_i + \boldsymbol{\epsilon}_{i,t} \tag{1.22}$$

By construction, $\boldsymbol{y}_{i,t-1}$ is correlated to the unobserved heterogeneity $\boldsymbol{\alpha}_i$. The estimator proposed by Arellano and Bond (1991) is based on the idea of, first, taking the first difference of both sides to eliminate the $\boldsymbol{\alpha}_i$ and, then, look for GMM estimators. The first difference of model (1.22) yields:

$$\Delta\boldsymbol{y}_{i,t} = \Delta\boldsymbol{y}_{i,t-1}\theta + \Delta\boldsymbol{\epsilon}_{i,t} \tag{1.23}$$

The $\boldsymbol{y}_{i,t-1}$ in $\Delta\boldsymbol{y}_{i,t-1}$ is a function of the $\boldsymbol{\epsilon}_{i,t}$ which is also in $\Delta\boldsymbol{\epsilon}_{i,t}$, and, consequently, there is an endogeneity problem. Several solutions have been proposed. Anderson and Hsiao (1981) suggested a Two-Stage Least Squares (TSLS) estimator based on further lags of $\Delta\boldsymbol{y}_{i,t}$ as instruments for $\Delta\boldsymbol{y}_{i,t-1}$. Arellano and Bond (1991) proposed a GMM approach based on moment equations constructed from further lagged levels of $\boldsymbol{y}_{i,t}$ and the first difference of the errors. Their GMM estimator is based on the extended formulation of

Hausman and Taylor (1981), discussed in the previous subsection, by including $\boldsymbol{y}_{i,t-1}$ in $\mathbf{X}_2^{(1)}$.

How to choose between fixed and random effects models: First considerations

From a conceptual point of view, unit-specific coefficients are random if one can think of the values of each unit-specific coefficient as realizations from a larger population. A typical and common example is to consider the school coefficients as a random coefficients when surveying students on different schools (**Chapter 2**). From a practical point of view, it is not always easy to choose between fixed and random effects. The choice is mainly based on some statistical tests, e.g. the Hausman test (Wooldridge, 2001). The Hausman test is based on the idea that if there is no endogeneity, both fixed and random effects estimators are consistent but the random effects estimator is efficient. Let us consider a model of the form (1.19). The null and the alternative hypothesis are:

$$H_0 : Cov\left(\boldsymbol{\alpha}_i, \boldsymbol{X}_{i,t}^{\prime(1)}\right) = 0 \quad \text{for every } \; i \; \text{ and every } \; j$$

$$H_1 : Cov\left(\boldsymbol{\alpha}_i, \boldsymbol{X}_{i,t}^{\prime(1)}\right) \neq 0 \qquad \text{for some } \; i \; \text{ and } \; j$$

The test statistics is given by:

$$H = \left(\hat{\boldsymbol{\theta}}^{RE} - \hat{\boldsymbol{\theta}}^{FE}\right) \left(Var\left(\hat{\boldsymbol{\theta}}^{RE} - \hat{\boldsymbol{\theta}}^{FE}\right)\right)^{-1} \left(\hat{\boldsymbol{\theta}}^{RE} - \hat{\boldsymbol{\theta}}^{FE}\right) \qquad (1.24)$$

which is distributed asymptotically, under H_0, according to a χ^2 distribution with k degrees of freedom, where k is the number of coefficients in θ.

Maximum likelihood approach

In all the approaches discussed up to here, no distributional assumption is required. First, for the residual series, $\boldsymbol{\epsilon}_{i,t}$, a specific distribution function is not necessary. Second, even when parameters are assumed

to be random, the classical approach mainly does not care about these random terms: they are mainly ignored and not estimated. Instead, maximum likelihood approach is based on the distribution assumption on the random components, in primis the residual series, $\boldsymbol{\epsilon}_{i,t}$. Maximum likelihood estimation for a univariate Gaussian random intercept models have been widely used (Bock and Aitken, 1981; Mislevy, 1985; Anderson and Aitken, 1985; Im and Gianola, 1988). The estimation is usually implemented by maximizing the marginal likelihood using the EM algorithm with numerical integration at each iteration. For multivariate Gaussian random effects models, several methods have been proposed (Korn and Whittemore, 1979; Stiratelli, Laird and Ware, 1984). For the applications investigated in this thesis, it is more interesting the discussion of the Maximum Likelihood Estimator (MLE) for dynamic panel models of the form of equation (1.22), re-written in the following:

$$\boldsymbol{y}_{i,t} = \boldsymbol{y}_{i,t-1}\theta + \boldsymbol{\alpha}_i + \boldsymbol{\epsilon}_{i,t} \tag{1.25}$$

The residual series are assumed to be $\boldsymbol{\epsilon}_{i,t} \overset{\text{iid}}{\sim} N\left(0, \sigma^2\right)$. It is also assumed[2] that $|\theta| < 1$. Under this model, the MLE of θ is inconsistent due to the classical incidental parameters problem, with an increasing number of parameters in the number of observations (Hsiao *et al.*, 2002). Consistent estimators based on instrument variables or generalized methods of moments have been already discussed in Subsubsection 1.3.1. Hsiao *et al.* (2002) proposed a likelihood-based estimator by taking the first-difference of model (1.22) to eliminate $\boldsymbol{\alpha}_i$, which is expressed by equation (1.23), and re-written in the following:

$$\Delta\boldsymbol{y}_{i,t} = \Delta\boldsymbol{y}_{i,t-1}\theta + \Delta\boldsymbol{\epsilon}_{i,t} \quad \text{for} \quad t = 2, .., T \tag{1.26}$$

[2] If $|\theta| < 1$, then $\lim_{T\to\infty} \Delta\boldsymbol{y}_{it-T+1}\theta^T = 0$.

By continuous substitution, equation (1.23) is equivalent to the following:

$$\Delta \boldsymbol{y}_{i,t} = \Delta \boldsymbol{y}_{it-T+1}\theta^T + \sum_{j=0}^{T-1} \theta^j \Delta \boldsymbol{\epsilon}_{it-j+1} \quad \text{for} \quad t = 2, .., T \tag{1.27}$$

The likelihood function is the following:

$$L\left(\theta, \sigma^2 \mid \underline{y}\right) = -\frac{NT}{2}\log\left(2\pi\right) - \frac{N}{2}\log\left(|\Omega|\right) - \frac{1}{2}\sum_{i=1}^{N} \Delta u_i' \Omega^{-1} \Delta u_i. \tag{1.28}$$

where

- $\underline{y} = \{y_{i,t}, i = 1, .., N, t = 1, .., T\}$.
- $\Delta u_i = \left(\Delta y_{i,1}, \Delta y_{i,2} - \theta \Delta y_{i,1}, ..., \Delta y_{i,t} - \theta \Delta y_{i,t-1}\right)'$
- The covariance matrix of Δu_i has the following form:

$$\Omega = \sigma^2 \begin{pmatrix} \frac{2}{1+\theta} & -1 & 0 & .. & 0 \\ -1 & 2 & -1 & .. & 0 \\ .. & ... & ... & ... & ... \\ 0 & ... & ... & -1 & 2 \end{pmatrix}$$

The MLE is highly nonlinear. To compute the maximum likelihood estimators, one can use a grid search procedure or an iterative procedure such as Newton-Raphson method. See Hsiao *et al.* (2002) for details.

Common Correlated Effects Estimator

Inference for panel data (with large N and T) is growing in interest over the past five years and robust estimation[3] procedures have been proposed. One of the most growing approaches is given by the multi-factor

[3]The common correlated effects estimator -as well as the whole classical approach- is typically studied in econometrics where a robust estimation is an estimator which is both unbiased and consistent. In statistical terms, *robust statistics* is instead the stability theory of statistical procedures. It is usually associated with distributional robustness. Therefore, in a statistical meaning, an estimator is robust if the mapping

approach, which assumes that the cross-dependence among units can be summarized by a finite number of unobserved common factors. Inference for such models is carried out either using maximum likelihood approach (Robertson and Symons, 2000; 2007) or the principal components procedures (Coakley *et al.*, 2002; Bai, 2009). The latter has been developed by Pesaran (2006) and Pesaran and Tosetti (2011). Their method is based on *Common Correlated Effects Estimator* (Pesaran, 2006), for models of the form:

$$\boldsymbol{y}_{i,t} = \boldsymbol{X}_t^{\prime(1)}\boldsymbol{\theta}_{i,1} + \boldsymbol{X}_{i,t}^{\prime(1)}\boldsymbol{\theta}_{i,2} + \boldsymbol{X}_t^{\prime(2)}\boldsymbol{\alpha}_i + \boldsymbol{\epsilon}_{i,t} \tag{1.29}$$

where $\boldsymbol{X}_t^{(1)}$ is a k_1-dimensional vector of observable common factors; $\boldsymbol{X}_{i,t}^{(1)}$ is a $k_2\,(\equiv k - k_1)$-dimensional vector of unit-specific covariates; $\boldsymbol{X}_t^{\prime(2)}$ is a m-dimensional vector of unobservable common factors; $\boldsymbol{\alpha}_i$ is a m-dimensional vector of factor loadings for unit i and $\boldsymbol{\epsilon}_{i,t}$ are the unit-specific errors, with possible spatially and temporally correlation.

Pesaran and Tosetti (2011) also assumes a specific dynamics for the unit-specific covariates, $\boldsymbol{X}_{i,t}^{\prime(1)}$, expressed as follows:

$$\boldsymbol{X}_{i,t}^{\prime(1)} = \boldsymbol{X}_t^{\prime(1)}\boldsymbol{A}_i + \boldsymbol{X}_t^{\prime(2)}\boldsymbol{\Gamma}_i + \boldsymbol{\nu}_{i,t} \tag{1.30}$$

Moreover, they assumed that the unit-specific random coefficients $\boldsymbol{\theta}_{i,2}$ are randomly distributed among units, with the following Gaussian distribution:

$$\boldsymbol{\theta}_{i,2} \mid \boldsymbol{\theta}_2, \Omega \overset{\text{iid}}{\sim} N\left(\boldsymbol{\theta}_2, \Omega\right) \tag{1.31}$$

Their idea is similar to the instrumental variable approach with multiple instruments, i.e. they use cross-section averages of the dependent and explanatory variables as instruments for the unobserved factors. Then, the standard panel regressions are augmented with these instruments. The corresponding *Common Correlated Effects* (CCE) mean group estimator

from distributions on the space of the random variable to distributions on the space of the estimates is continuous (Huber and Ronchetti, 2009). In this section, robustness is referred to its econometric meaning.

(CEEMG) is given by:

$$\hat{\boldsymbol{\theta}}^{\prime(2)}_{CCEMG} = \frac{1}{N}\sum_{i=1}^{N} \hat{\boldsymbol{\theta}}^{\prime(2)}_{CEE,i} \tag{1.32}$$

with $\hat{\boldsymbol{\theta}}^{\prime(2)}_{CEE,i} = \left(\boldsymbol{X}^{\prime(1i)}\bar{M}\boldsymbol{X}^{\prime(1i)}\right)^{-1}\boldsymbol{X}^{\prime(1i)}\bar{M}\boldsymbol{y}_i$, where

- $\boldsymbol{X}^{\prime(1i)}$ is the matrix collecting all the $\boldsymbol{X}^{\prime(1)}_{i,t}$.
- $\bar{M} = I_T - \bar{H}\left(\bar{\boldsymbol{H}}'\bar{\boldsymbol{H}}\right)^{-1}\bar{\boldsymbol{H}}'$.
- $\bar{\boldsymbol{H}} = \left(\boldsymbol{X}^{(1)}, \boldsymbol{Y}, \boldsymbol{X}^{(1i)}\right)$ with $\boldsymbol{X}^{(1)}, \boldsymbol{Y}, \boldsymbol{X}^{(1i)}$ the matrix collecting the $\boldsymbol{X}^{(1)}_t$, $\sum_{i=1}^{N}\boldsymbol{y}_{i,t}$, $\sum_{i=1}^{N}\boldsymbol{X}^{(1)}_{i,t}$.

The CEE pooled estimator (CEEP) is the following:

$$\hat{\boldsymbol{\theta}}^{\prime(2)}_{CCEEP} = \left(\sum_{i=1}^{N}\boldsymbol{X}^{\prime(1i)}\bar{M}\boldsymbol{X}^{\prime(1i)}\right)^{-1}\sum_{i=1}^{N}\boldsymbol{X}^{\prime(1i)}\bar{M}\boldsymbol{y}_i \tag{1.33}$$

Both these two estimators are robust, in the econometric sense (i.e. unbiased and consistent), and asymptotically Gaussian (Pesaran and Tosetti, 2011).

In **Chapter 3**, the application under investigation will be based on a dynamic panel model similar to model (1.22). The aims of the analysis will require the estimation of the unit-specific coefficients for the lagged variable and for other exogenous covariates. In particular, the variable of interest is the first-difference of a variable $\boldsymbol{y}_{i,t}$ and the model will have the following form[4]:

$$\Delta\boldsymbol{y}_{i,t} = \boldsymbol{\alpha}_i + \boldsymbol{y}_{i,t-1}\boldsymbol{\theta}_{i,1} + \boldsymbol{X}^{\prime(1)}_{1;i,t}\boldsymbol{\theta}_{i,2} + \boldsymbol{X}^{\prime(1)}_{2;i,t}\boldsymbol{\mu} + \boldsymbol{\epsilon}_{i,t} \tag{1.34}$$

[4]The lagged variable $\boldsymbol{y}_{i,t-1}$ is included in the set of regressors to avoid spurious correlation due to the stationarity of $\boldsymbol{y}_{i,t}$.

where $\boldsymbol{X}'^{(1)}_{1;i,t}$ is a uni-dimension exogenous regressor whose impact is different for each unit i; the regressor $\boldsymbol{X}'^{(1)}_{2;i,t}$ is a bi-dimension exogenous regressor with the same impact for all the units. The GMM estimation procedure -discussed in Subsubsection 1.3.1- would not allow the estimation of all the unit-specific coefficients involved in equation (1.34). Instead, conventional fixed effects (within) estimators -discussed in Subsubsection 1.3.1- are biased and inconsistent (Pesaran and Smith, 1995; Phillips and Sul, 2003). The *Common Correlated Effects* (CCE) estimator (Pesaran and Tosetti, 2011) -just discussed- is a possible solution for the estimation of the unit-specific coefficients within a frequentist framework. **Chapter 3** will propose the estimation of model (1.34) within a Bayesian framework, by extending previous Bayesian approaches (Hsiao *et al.*, 1999; Hirano, 2002).

1.3.2 Bayesian approach

As seen, the classical approach usually does not strictly require any distributional assumption. They are required only within the MLE approach. Instead, the Bayesian approach usually needs to specify the distribution functions: not only for residual terms but also for unknown coefficients. The main feature of any Bayesian model is that any unknown coefficient is regarded as a random variable, on which a (prior) distribution is specified. This prior distribution does not have a frequentistic interpretation, as a description of the variability of the coefficient in repeated trials, but is a description of the subjective uncertainty about the unknown value.

Dealing with *longitudinal* Bayesian models requires not only to express the typical elements of *non-longitudinal* Bayesian models, but also to define what kind of dependence exists among units. The typical elements for *non-longitudinal* Bayesian models are the definition of a model that expresses qualitative aspects of the knowledge (e.g. functional form for the regression function, forms of distribution function for the residu-

als and conditional independence across observations). The simplest way for specifying the probabilistic dependence across-units is to assume that the unit-specific parameters come from some common population distribution, as in model (1.8). As discussed in Subsection 1.2.1, the resulting model is usually called a *hierarchical model* since, to simplify the specification of the model, one can work in subsequent steps with variables that are independent conditionally on the other related random variables.

Random Effects and Random Coefficients models

In this and in the following subsections, some important features of the Bayesian models, introduced in Subsection 1.2.1, are discussed. In particular, this discussion aims to rethinking about the choice between fixed and random effects within a Bayesian setting and discussing about shrinkage effects.

Let us consider a linear regression model of the form (1.13), where the only temporal evolution for the variable of interest, $\boldsymbol{y}_{i,t}$, comes from the temporal evolution of the covariates, $\boldsymbol{X}_{i,t}^{(1)}$. Moreover, let us assume that the regressor $\boldsymbol{X}_{i}^{(2)}$ only contains the constant term. The random coefficients models, or hierarchical models, can then be written as:

$$\boldsymbol{y}_{i,t} \mid \boldsymbol{\alpha}_i, \boldsymbol{\theta}_i, \boldsymbol{X}_{i,t}^{\prime(1)} \overset{\text{iid}}{\sim} N\left(\boldsymbol{\alpha}_i + \boldsymbol{X}_{i,t}^{\prime(1)}\boldsymbol{\theta}_i, \sigma^2\right) \tag{1.35}$$

$$(\boldsymbol{\alpha}_i, \boldsymbol{\theta}_i) \mid \boldsymbol{P} = P \overset{\text{iid}}{\sim} P \tag{1.36}$$

where $\boldsymbol{\alpha}_i$ and $\boldsymbol{\theta}_i$ are unit-specific, at least for one of the elements of each vector and $\boldsymbol{P}$ is called latent population distribution. Model (1.35)-(1.36) can equivalently be expressed by the following parametrization in terms of the mixing distribution:

$$\boldsymbol{y}_{i,t} \mid \boldsymbol{P}, \boldsymbol{X}_{i,t}^{\prime(1)} \sim \int N\left(\boldsymbol{y}_{i,t} \mid \boldsymbol{\alpha}_i + \boldsymbol{X}_{i,t}^{\prime(1)}\boldsymbol{\theta}_i, \sigma^2\right) d\boldsymbol{P}\left(\boldsymbol{\alpha}_i, \boldsymbol{\theta}_i\right) \tag{1.37}$$

The expression (1.37) represents the distribution of the random variables $\boldsymbol{y}_{i,t}$, given the population distribution and the covariates, as a Gaussian

location mixture distribution.

As already said in Subsection 1.2.1, in some problems one can specify a parametric model for $\boldsymbol{P}$, such as a Gaussian distribution (Lindley and Smith, 1972). In a nonparametric approach, instead, $\boldsymbol{P}$ is a random distribution, with no parametric assumptions, and a prior G is assumed on the entire space of distribution functions on $(\boldsymbol{\alpha}, \boldsymbol{\theta})$. Constraining $\boldsymbol{P}$ to a parametric family has the main drawback of restricting the range of situations where the model can be applied. A nonparametric approach instead uses a prior with a large support, possibly the space of all distributions. Working with such a large space makes particularly important requiring the tractability of the posterior distribution (Ferguson, 1973). For this reason, one popular choice is to assume that $\boldsymbol{P}$ has a Dirichlet process prior (Ferguson, 1973), which is a conjugate prior for exchangeable data in the form of an i.i.d. sample of exact observations. Usually, in models like model (1.37), computations of the posterior distributions of interest are analytically difficult. In fact, as discussed in Section 1.3.1, even MLE of the mixing distribution and of the unit-specific coefficient is analytically hard or impossible. The Bayesian nonparametric approach based on the DP prior, thanks to the simple form of the predictive distribution (1.11), allows for an easy implementation through MCMC methods.

How to choose between fixed and random effects models. Further considerations

We can now reformulate the problem of the choice between fixed and random effects model from a Bayesian point of view. A Bayesian interpretation of the fixed effects is to assign independent priors without any hierarchical structure, i.e. $(\boldsymbol{\alpha}_i, \boldsymbol{\theta}_i) \mid \boldsymbol{P}_i = P_i \sim P_i$ without any probabilistic dependence across the $\boldsymbol{P}_i$. Therefore, the choice between random effects and fixed effects reduces to the choice between exchangeable and independent observations and it needs to be carefully considered based on whether borrowing strength makes sense or not.

Population distribution: Shrinkage

The presence of the distribution that ties together the exchangeable $(\boldsymbol{\alpha}_i, \boldsymbol{\theta}_i)$ has mainly two consequences: on one hand, it allows for borrowing strength across units; on the other hand, it causes shrinkage of the estimates towards a common mean. The idea of shrinkage is quite old and it was originally developed outside the random effects or Bayesian modeling literature (Stein, 1956; James and Stein, 1961). Lindley and Smith (1972) developed Bayesian shrinkage estimation in random effects linear models. Further discussion on the relationship between the Bayesian estimation and the James-Stein estimations in Box and Tiao (1973).

For the sake of simplicity, let us consider repeated measures data, $y_{i,t}$, $i = 1, .., N$, $t = 1, .., T_i$, where the time dimension, t, does not imply any dependence across observations but it is just an indicator for the recorded observation for each unit i. No covariates are included and that the observations $\boldsymbol{y}_{i,t}$ are independent across units. Let us assume that:

$$\begin{pmatrix} \boldsymbol{y}_{i,1} \\ ... \\ \boldsymbol{y}_{iT_i} \end{pmatrix} \mid \boldsymbol{\theta}_i \overset{indep}{\sim} N_{T_i}\left(\boldsymbol{\theta}_i, \sigma^2 I_{T_i}\right) \tag{1.38}$$

Then, the MLE of $\boldsymbol{\theta}_i$ is the sample mean:

$$\bar{\boldsymbol{y}}_i = \frac{\sum_{t=1}^{T_i} \boldsymbol{y}_{i,t}}{T_i} \tag{1.39}$$

which is also a sufficient statistics whenever the variance σ^2 is known. A fundamental result of Stein (1956) shows that the MLE of the N (unknown constant) population means, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, .., \boldsymbol{\theta}_N)$, given by the vector of sample means, $\bar{y} = (\bar{y}_1, ..., \bar{y}_N)$, is inadmissible, if $N \geq 3$. Stein's results assumes Gaussian distributions with a common variance σ^2, let us say $\sigma^2 = 1$, and a sum of mean squared component errors risk func-

tion. James and Stein (1961) provided a specific shrinkage estimator, the James-Stein estimator, which dominates the sample mean vector substantially with respect to the quadratic loss function. The James-Stein estimator is given by:

$$\hat{\boldsymbol{\theta}}_{JS,i} = \left(1 - \hat{\boldsymbol{B}}_{j,s}\right) \bar{\boldsymbol{y}}_i + \hat{\boldsymbol{B}}_{j,s}\mu_0 \tag{1.40}$$

where $\hat{\boldsymbol{B}}_{j,s} = \dfrac{N-2}{\boldsymbol{S}}$, $\boldsymbol{S} = \dfrac{\sum_{i=1}^{N} \boldsymbol{y}_i^2}{\sigma^2}$ is the James - Stein shrinkage coefficient and measures the shrinkage of the usual unbiased estimates y_i towards $\theta_0 \equiv 0$.

Starting from their seminal work, shrinkage estimation has become one of the major focuses of hierarchical models. Its justification can be easier understood using an example. Let us assume that the focus of the analysis is to evaluate the procyclical reaction of a sector across countries, as studied in **Chapter 3**. The longer the series of available data is, the better the assessment of their procyclical reaction will be. Let us consider two countries: one country, let us say country A, with few data points, and another country, e.g. B, with many data points. On one hand, the evaluation for country A cannot be based only on its few data, but it should be shrunk towards the overall mean of the procyclical reaction of all the countries. On the other hand, the country B has enough data points, and the assessment of its procyclical behavior should be close to the estimate obtained separately.

The Bayesian interpretation of the Stein paradox is that, although the experiment are physically independent, the unit-specific means, $\boldsymbol{\theta}_i$, $i = 1, .., N$, are probabilistically dependent and they can be regarded as a random sample from a population distribution. The shrinkage factors impact the posterior means. In particular, let us consider the following simple hierarchical model, with known and common variance σ^2 and with

$T_i = T$ for every i:

$$\begin{pmatrix} \boldsymbol{y}_{i,1} \\ ... \\ \boldsymbol{y}_{i,t} \end{pmatrix} \mid \boldsymbol{\theta}_i \overset{indep}{\sim} N\left(\boldsymbol{\theta}_i, \sigma^2 I_T\right) \tag{1.41}$$

$$\boldsymbol{\theta}_i \mid \boldsymbol{m}, \boldsymbol{\tau}^2 \overset{\text{iid}}{\sim} N\left(\boldsymbol{m}, \boldsymbol{\tau}^2\right) \tag{1.42}$$

Then,

$$\boldsymbol{\theta}_i \mid \boldsymbol{m}, \boldsymbol{\tau}^2, y_{1:N,1:T} \overset{\text{iid}}{\sim} N\left(\boldsymbol{B}_i \bar{\boldsymbol{y}}_i + (1 - \boldsymbol{B}_i)\, \boldsymbol{m}, \nu^2 (1 - \boldsymbol{B}_i)\right) \tag{1.43}$$

where $\boldsymbol{B}_i = \dfrac{\nu^2 T}{\nu^2 T + \sigma^2}$. The factors $(1 - \boldsymbol{B}_i)$ are known as shrinkage factors. In particular, the larger $(1 - \boldsymbol{B}_i)$ is (equivalently, the smaller $\boldsymbol{B}_i$ is), the more the population mean is shrunk back to the a priori population mean $\boldsymbol{m}$. The magnitude of the shrinkage factor depends on the sample size T, the population variance $\boldsymbol{\tau}^2$, and the observational variance σ^2. Higher values of σ^2 give stronger shrinkage. Therefore, using an absolutely continuous parametric distribution for the $\boldsymbol{\theta}_i$, the posterior means of the $\boldsymbol{\theta}_i$ are distinct but they are shrunk towards the prior mean $\boldsymbol{m}$. Instead, as previously discussed, assigning a discrete prior, such as the DP prior, allows for ties and clustering among he $\boldsymbol{\theta}_i$, accordingly the predictive distribution of the form (1.11).

Chapter 3 and **Chapter 5** propose the use of hierarchical models for longitudinal data within a Bayesian semiparametric framework, also based on the EDP prior. Both chapters are based on some extension of the model described by equations (1.7)-(1.9). In **Chapter 5**, Pharmacokinetics and Pharmacodynamics are taken into consideration and the model is nonlinear in its random parameters. In particular, the available data are $\left(y_{i,t}, z_{i,t}, X^{(1)}_{1;i}, X^{(1)}_{2;i}\right)$, $i = 1, .., N$; $t = 1, .., T$. The joint model for random variables $\left(\boldsymbol{y}_{i,t}, \boldsymbol{z}_{i,t}\right)$ can be expressed in terms of a marginal and a conditional model with the two following models for the data, which

substitute equation (1.7).

$$\begin{pmatrix} \boldsymbol{y}_{i,1} \\ ... \\ \boldsymbol{y}_{i,t} \\ ... \\ \boldsymbol{y}_{i,t} \end{pmatrix} \mid \boldsymbol{\theta}_i^1, \boldsymbol{\Sigma}_1 \overset{indep}{\sim} N \left(\begin{pmatrix} f_1\left(1, X_{1;i}^{(1)}, \boldsymbol{\theta}_i^1\right) \\ ... \\ f_1\left(t, X_{1;i}^{(1)}, \boldsymbol{\theta}_i^1\right) \\ ... \\ f_1\left(T, X_{1;i}^{(1)}, \boldsymbol{\theta}_i^1\right) \end{pmatrix}, \boldsymbol{\Sigma}_1 \right), \; i = 1, .., N \tag{1.44}$$

$$\begin{pmatrix} \boldsymbol{z}_{i,1} \\ ... \\ \boldsymbol{z}_{i,t} \\ ... \\ \boldsymbol{z}_{i,t} \end{pmatrix} \mid \boldsymbol{\theta}_i^1, \boldsymbol{\theta}_i^2, \boldsymbol{\Sigma}_2 \overset{indep}{\sim} N \left(\begin{pmatrix} f_2\left(1, X_{2;i}^{(1)}, \boldsymbol{\theta}_i^2, f_1\left(1, X_{1;i}^{(1)}, \boldsymbol{\theta}_i^1\right)\right) \\ ... \\ f_2\left(t, X_{2;i}^{(1)}, \boldsymbol{\theta}_i^2, f_1\left(t, X_{1;i}^{(1)}, \boldsymbol{\theta}_i^1\right)\right) \\ ... \\ f_2\left(T, X_{2;i}^{(1)}, \boldsymbol{\theta}_i^2, f_1\left(T, X_{1;i}^{(1)}, \boldsymbol{\theta}_i^1\right)\right) \end{pmatrix}, \boldsymbol{\Sigma} \right), \; i = 1, .., N \tag{1.45}$$

where both $f_1(\cdot, \cdot, \cdot)$ and $f_2(\cdot, \cdot, \cdot, \cdot)$ are deterministic function of the time, t, the deterministic univariate covariate x_i and the unit-specific random coefficients. The function $f_2(\cdot, \cdot, \cdot, \cdot)$ is also a function of the $f_1(\cdot, \cdot, \cdot)$.

Although $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ could introduce some temporal dependence, they will be assumed $\boldsymbol{\Sigma}_1 = \boldsymbol{\sigma}_1^2 I_T$ and $\boldsymbol{\Sigma}_2 = \boldsymbol{\sigma}_2^2 I_T$, meaning that the observations are conditionally independent also across time. The time dimension here enters in the means, which are deterministic function of -among the others- the time and the unit-specific random coefficients. Equation (1.8) is maintained for the vector $\boldsymbol{\theta}_i$ collecting $\boldsymbol{\theta}_i^1$ and $\boldsymbol{\theta}_i^2$. Equation (1.9) will be substituted with the EDP prior. The model can be completed with the priors for $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, and the other parameters.

In **Chapter 3**, the application taken into consideration is the analysis of the leverage-ratio and the model is an autoregressive with deterministic covariates, $X_{1;i,t}^{(1)}$ and $X_{2;i,t}^{(1)}$. This means that equation (1.7) will be substitute with the following one.

$$\begin{pmatrix} \boldsymbol{y}_{i,2} \\ ... \\ \boldsymbol{y}_{i,t} \\ ... \\ \boldsymbol{y}_{i,t} \end{pmatrix} \mid \boldsymbol{\theta}_i^1, \boldsymbol{\theta}_i^2, \boldsymbol{\mu}, \boldsymbol{\Sigma} \overset{indep}{\sim} N \left(\begin{pmatrix} \boldsymbol{\theta}_i^1 \boldsymbol{y}_{i,1} + \boldsymbol{\theta}_i^2 X_{1;i,t}^{(1)} + \boldsymbol{\mu} X_{2;i,t}^{(1)} \\ ... \\ \boldsymbol{\theta}_i^1 \boldsymbol{y}_{i,t-1} + \boldsymbol{\theta}_i^2 X_{1;i,t}^{(1)} + \boldsymbol{\mu} X_{2;i,t}^{(1)} \\ ... \\ \boldsymbol{\theta}_i^1 \boldsymbol{y}_{i,t-1} + \boldsymbol{\theta}_i^2 X_{1;i,t}^{(1)} + \boldsymbol{\mu} X_{2;i,t}^{(1)} \end{pmatrix}, \boldsymbol{\Sigma} \right), \quad i = 1, .., N \tag{1.46}$$

where $\boldsymbol{\theta}_i^1$ is a unidimensional vector, $\boldsymbol{\theta}_i^2$ is a k_2 -dimensional vector associated with the k_2 -dimensional regressors $x_{i,t}$ and $\boldsymbol{\mu}$ is a k_3 -dimensional vector common coefficient. The vector of the unit-specific random coefficients $\left(\boldsymbol{\theta}_i^1, \boldsymbol{\theta}_i^2\right)$ is assumed to have an unknown random population distribution with a nonparametric prior. First, a DP prior is assigned. Then, a break into the coefficients $\boldsymbol{\theta}_i^2$ is included and the EDP prior is used.

1.4 State-space models

Section 1.3 relies on the assumption that the temporal evolution of the observable $\boldsymbol{y}_{i,t}, i = 1, .., N, t = 1, .., T$ can be completely explained throughout the model for the data, i.e. at the first level of the hierarchical model. As discussed in Section 1.2.1, for some applications this approach cannot represent adequately the dynamics of the observable variables. In this section, we briefly discuss a class of latent variables models where the latent process has a Markovian temporal evolution, conditionally on the transitional matrix, taking Petris *et al.* (2009) as main reference. Within this class of models, an important category is represented by the state-space models.

Definition A *state-space model* consists of a unobserved state process, $(\boldsymbol{\theta}_t,\ t = 0, 1, ..)$ and an observation process, $(\boldsymbol{y}_t,\ t = 1, 2, ..)$, satisfying the following assumptions:

- $(\boldsymbol{\theta}_t)_t$ is a Markov Chain.
- Conditionally on $(\boldsymbol{\theta}_t)$, the $\boldsymbol{y}_t$'s are independent and the conditional distribution of $\boldsymbol{y}_t$ depends on $\boldsymbol{\theta}_t$ only.

A state-space model can then be specified by a prior distribution for $\boldsymbol{\theta}_0$,

together with the observation and evolution equations:

$$\begin{cases} \boldsymbol{y}_t = h_t\left(\boldsymbol{\theta}_t, \boldsymbol{v}_t\right) \\ \boldsymbol{\theta}_t = g_t\left(\boldsymbol{\theta}_{t-1}, \boldsymbol{w}_t\right) \end{cases} \tag{1.47}$$

for arbitrary functions g_t and h_t; where the noise terms, $(\boldsymbol{w}_t)$ and $(\boldsymbol{v}_t)$ are two independent sequences of serially independent random variables, usually with known distribution,and are independent from $\boldsymbol{\theta}_0$.

For simplicity of exposition and notation, let us assume that all the parameters of the functions h_t and g_t are known. Define $\boldsymbol{\theta}_{0:t} = \{\boldsymbol{\theta}_0, ..., \boldsymbol{\theta}_t\}$ and $y_{1:t} = \{y_1, ..., y_t\}$., and call the target distribution $\pi\left(\boldsymbol{\theta}_t | y_{1:t}\right)$. For a generic $n \in \{1, ..., T\}$, the joint distribution is given by:

$$\begin{aligned} \left(\boldsymbol{\theta}_{t_0:n}, y_{1:n}\right) &\sim \pi\left(\boldsymbol{\theta}_{t_0}\right) \prod_{t=1}^{n} p\left(\boldsymbol{\theta}_t, y_t | \boldsymbol{\theta}_{t_0:t-1}, y_{1:t-1}\right) \\ &= \pi\left(\boldsymbol{\theta}_{t_0}\right) \prod_{t=1}^{n} f\left(y_t | \boldsymbol{\theta}_{t_0:t}, y_{1:t-1}\right) p\left(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{t_0:t-1}, y_{1:t-1}\right) \\ &= \pi\left(\boldsymbol{\theta}_{t_0}\right) \prod_{t=1}^{n} f\left(y_t \mid y_{t-1}, \boldsymbol{\theta}_t\right) p\left(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{t-1}\right) \end{aligned} \tag{1.48}$$

The density of $(y_1, ..., y_n)$ can then be obtained by marginalizing this distribution with respect to the $\boldsymbol{\theta}_{0:t}$. In the filtering problem, the data are supposed to arrive sequentially in time and one usually is interested in obtaining the estimation for the current value of the state vector, based on the observations up to time t, and updating the estimation as a new data point becomes available at time $t + 1$. In particular, three general steps are usually implemented:

- State prediction:

$$p(\boldsymbol{\theta}_t | y_1, ..., y_{t-1}) = \int p(\boldsymbol{\theta}_t | \boldsymbol{\theta}_{t-1}) p(\boldsymbol{\theta}_{t-1} | y_1, ..., y_{t-1}) d\boldsymbol{\theta}_{t-1}. \tag{1.49}$$

- Observational prediction:

$$f(y_t|y_1,...,y_{t-1}) = \int f\left(y_t|\boldsymbol{\theta}_t\right) p(\boldsymbol{\theta}_t|y_1,...,y_{t-1})\, d\boldsymbol{\theta}_t. \tag{1.50}$$

- Filtering:

$$p(\boldsymbol{\theta}_t|y_1,...,y_t) = \frac{f\left(y_t|\boldsymbol{\theta}_t\right) p(\boldsymbol{\theta}_t|y_1,...,y_{t-1})}{f\left(y_t|y_1,...,y_{t-1}\right)}. \tag{1.51}$$

These computations are usually analytically solvable if the models are linear and Gaussian. In this case, the joint distribution, all the marginal and the conditionals distributions are Gaussian and the Kalman filter (Kalman, 1960) provides a sequential updating for the estimation of the latent vector. If the model is Gaussian but not linear, one can use the Extended Kalman Filter, which consists in linearizing , at each subsequent time t, the non-linear dynamics around the last consecutive predicted and filtered estimates of the state, and applies the Kalman Filter on the linearized dynamics. If the model is neither linear nor Gaussian, computational or numerical methods are usually required, e.g. Sequential Monte Carlo, discussed in the next section. See West and Harrison (1997), Doucet *et al.*(2001) and Petris *et al.* (2009) for further discussions.

The above model is stated for a time series. Several alternatives are possible for dealing with longitudinal data, $\boldsymbol{y}_{i,t}$, $i = 1,..,N$ and $t = 1,..,T$. In **Chapter 4**, the aim will be to model unobservable compositional data, representing unknown portfolio structures of sectors of the economy. We will assume a state-space model for each sector i. The cross-sectors feature could then be taken into consideration by assuming probabilistic dependence across the state-processes, $(\boldsymbol{\theta}_{i,t})$. Traditional approaches include Seemingly Unrelated Time Series Equations or SUTSE (Zellner, 1963), dynamic hierarchical models (Gamerman and Migon, 1993), dynamic factor models (Harvey, 1989; Forni *et al.*, 2000).

In the application taken into consideration in **Chapter 4**, we will introduce a constraint across the sectors, derived by a flows-of-funds equality that must hold at the economy-level, by constraining the support of filtering distribution of the sector-specific latent vectors.

1.5 Computational aspects

In Bayesian inference, it is very common that the posterior distribution, say π, of the parameters, say $\boldsymbol{\psi}$, and its summary, such as its mean and variance, cannot be evaluated analytically. Markov chain Monte Carlo (MCMC) algorithms and Sequential Monte Carlo (SMC) algorithms are computational methods that help solve this problem. Sequential Monte Carlo are used for on-line analysis. In the next subsections, we will briefly review some of the main characteristics of MCMC and SMC techniques, taking as main reference Petris *et al.* (2009). For a deeper discussion, refer to Gelman *et al.* (2009), Robert and Casella (2004), Tierney (1994) and Doucet *et al.* (2001).

1.5.1 Monte Carlo and importance sampling

Monte Carlo: Monte Carlo method allows to approximate an expectation by using the sample mean of a function of simulated random variables. Call ψ_1,..,ψ_n an i.i.d. sample from the distribution π. The standard *Monte Carlo* method for approximating the mean of any function $g(\boldsymbol{\psi})$ having finite posterior expectation is to compute the sample average:

$$E_\pi\left(g\left(\boldsymbol{\psi}\right)\right) = \int g\left(\psi\right)\pi\left(\psi\right)d\psi \approx \frac{1}{n}\sum_{j=1}^{n} g\left(\psi_j\right) \tag{1.52}$$

Importance sampling: Call f the importance density and let f be absolutely continuous with respect to π, i.e. $f(x) = 0$ implies that $\pi(x) = 0$.

Then one can write:

$$E_\pi\left(g\left(\boldsymbol{\psi}\right)\right)=\int g\left(\psi\right)\frac{\pi\left(\psi\right)}{f\left(\psi\right)}f\left(\psi\right)d\psi=E_f\left(g\left(\boldsymbol{\psi}\right)w^*\left(\boldsymbol{\psi}\right)\right). \tag{1.53}$$

where $w^*\left(\boldsymbol{\psi}\right)=\dfrac{\pi\left(\psi\right)}{f\left(\psi\right)}$ is called *importance function*. It follows that one can approximate the expected value of interest by generating a random sample of size n from f and computing

$$E_\pi\left(g\left(\boldsymbol{\psi}\right)\right)\approx\frac{1}{n}\sum_{j=1}^{n}g\left(\psi_j\right)w^*\left(\psi_j\right) \tag{1.54}$$

Since the normalization factor, say C, is unknown, one can let $\tilde{w}_i = Cw\left(\psi_i\right)$, take $g\left(\psi\right)\equiv C$, and re-write the expression (1.54) as follows:

$$E_\pi\left(g\left(\boldsymbol{\psi}\right)\right)\approx\frac{1}{n}\sum_{j=1}^{n}g\left(\psi_j\right)w^*\left(\psi_j\right)\approx\frac{1}{n}\sum_{j=1}^{n}g\left(\psi_j\right)w\left(\psi_j\right)\quad\text{with } w_i=\frac{\tilde{w}_i}{\sum_{j=1}^{n}\tilde{w}_i} \tag{1.55}$$

1.5.2 Monte Carlo Markov Chain

The equality (1.52) holds not only for independent samples but also for Markov chains. The corresponding approaches are called *Monte Carlo Markov Chain* (MCMC) methods, since are Monte Carlo methods based on simulating random variables from a Markov chain.

For an irreducible, aperiodic and recurrent Markov chain $\left(\psi_t\right)_{t\geq 1}$, having invariant distribution π, it can be shown that for every initial value ψ_1, the distribution of $\psi_t \to \pi$ as $t \to \infty$. Therefore, for an b that is sufficiently large, ψ_{b+1},...,ψ_{b+n} are approximately distributed as π. Since the law of large numbers, expressed by (1.52) , holds, one can

obtain the approximation:

$$E_\pi\left(g\left(\boldsymbol{\psi}\right)\right) \approx \frac{1}{n}\sum_{j=1}^{n} g\left(\psi_{b+j}\right) \tag{1.56}$$

In order to assess the accuracy of an ergodic average as an estimator of the corresponding expected value, one can compute the variance of the estimated function $g\left(\boldsymbol{\psi}\right)$.

Gibbs sampler algorithm

Consider the k-dimensional parameter of interest $\boldsymbol{\psi} = (\boldsymbol{\psi}_1, .., \boldsymbol{\psi}_k)$ with $k > 1$ and call the target distribution $\pi\left(\boldsymbol{\psi}_1, .., \boldsymbol{\psi}_k\right)$. The idea of the Gibbs sampler is to consider the univariate conditional distributions, called *full conditionals*, expressing the distributions of one random parameter when all the others are assigned fixed values. After initializing all the parameters, the Gibbs sampler is then the iterative method requiring to sample sequentially from the full conditional distributions, as follows:

- Generate $\psi_1^{(s)}$ from $\pi\left(\boldsymbol{\psi}_1^{(s)} \mid \psi_2^{(s-1)}, .., \psi_k^{(s-1)}\right)$;
- Generate $\psi_2^{(s)}$ from $\pi\left(\boldsymbol{\psi}_2^{(s)} \mid \psi_1^{(s)}, \psi_2^{(s-1)}, .., \psi_k^{(s-1)}\right)$;
- ...
- Generate $\psi_k^{(s)}$ from $\pi\left(\boldsymbol{\psi}_k^{(s)} \mid \psi_1^{(s)}, \psi_2^{(s)}, .., \psi_{k-1}^{(s)}, \psi_k^{(s-1)}\right)$;

See Casella and George (1992) for further details.

Metropolis Hastings algorithm

Metropolis - Hastings algorithm allows to generate the next state of the chain from an essentially arbitrary distribution, including an accept/reject step to ensure the invariance of the target distribution π. Calling ψ the current chain, the proposal $\tilde{\psi}$ is generated from a density $q\left(\psi, \cdot\right)$, called candidate generating density, where q is a density in

its second argument and is parametrized by the first argument, required to be irreducibility and aperiodicity (Chib and Greenberg, 1995). The proposal $\tilde{\psi}$ represents the next state of the chain with the probability:

$$\alpha\left(\psi,\tilde{\psi}\right)=\min\left\{1,\frac{\pi\left(\tilde{\psi}\right)q\left(\tilde{\psi},\psi\right)}{\pi\left(q\right)q\left(\psi,\tilde{\psi}\right)}\right\} \tag{1.57}$$

If the proposal is rejected, the chain stays in the current state ψ.

1.5.3 Sequential Monte Carlo

Sequential Monte Carlo (SMC) for state-space models are known as Particle Filters (PF). Focusing on state-space models of the form (1.47), in this subsection, we brief overview the basic particle filter. In filtering applications, the target distribution changes every time that a new observation is available, requiring a completely new MCMC every time a new observation becomes available. PF provides an alternative and efficient solution to estimating the filtering distribution by updating the discrete approximation of $\pi(\boldsymbol{\theta}_{0:t-1}|y_{1:t-1})$ when the observation y_t becomes available to obtain a discrete approximation of $\pi(\boldsymbol{\theta}_{0:t}|y_{1:t})$. For every s, let us denote by $\pi(\hat{\boldsymbol{\theta}_{0:s}}|y_{1:s})$ the approximation of $\pi(\boldsymbol{\theta}_{0:s}|y_{1:s})$. The underlying idea is the same as in the importance sampling algorithm and it requires two steps:

- For each point $\theta_{0:t-1}^{(i)}$ in the support $\pi(\hat{\boldsymbol{\theta}_{0:t-1}}|y_{1:t-1})$, draw an additional component $\theta_t^{(i)}$ to obtain $\theta_{0:t}^{(i)}$.
- Update its weight $v_{t-1}^{(i)}$ to obtain an appropriate $v_t^{(i)}$.

The weighted points $\left(\theta_t^{(i)},v_t^{(i)}\right)$, i = 1,...,n, provide the new discrete approximation, $\pi(\hat{\boldsymbol{\theta}_{0:t}}|y_{1:t})$. Calling $g_t\left(\theta_{0:t}|y_{0:t}\right)$ the importance density used to generate $\theta_{0:t}$, let us assume that g_t can be expressed in the following form:

$$g_t(\theta_{0:t}\mid y_{0:t})=g_{t|t-1}\left(\theta_t\mid\theta_{0:t-1},y_{0:t}\right)g_{t-1}\left(\theta_{0:t-1}\mid y_{0:t-1}\right).$$

It follows that $\theta_{0:t}$ can be obtained sequentially by combining $\theta_{0:t-1}$, drawn from g_{t-1}, with θ_t, drawn from $g_{t|t-1}\left(\theta_t|\theta_{0:t-1}, y_{0:t}\right)$. The weights are then updated as follows:

$$v_t \propto v_{t-1} \frac{\pi\left(y_t \mid \boldsymbol{\theta}_t^{(i)}\right) \pi\left(\boldsymbol{\theta}_t^{(i)} \mid \boldsymbol{\theta}_{t-1}^{(i)}\right)}{g_{t|t-1}\left(\boldsymbol{\theta}_t^{(i)} \mid \boldsymbol{\theta}_{0:t-1}^{(i)}, y_{1:t}\right)} \tag{1.58}$$

Once that $\theta_t^{(i)}$ has been drawn from $g_{t|t-1}\left(\theta_t \mid \theta_{0:t-1}^{(i)}, y_{1:t}\right)$ one can compute the normalized weights $\tilde{v}_t^{(i)}$ as

$$\tilde{v}_t = v_{t-1}^{(i)} \frac{\pi\left(y_t \mid \theta_t^{(i)}\right) \pi\left(\theta_t^{(i)} \mid \theta_{t-1}^{(i)}\right)}{g_{t|t-1}\left(\theta_t^{(i)} \mid \theta_{0:t-1}^{(i)}, y_{1:t}\right)} \tag{1.59}$$

The weights are then normalized as follows:

$$v_t^{(i)} = \frac{\tilde{v}_t^{(i)}}{\sum_{j=1}^{n} \tilde{v}_t^{(j)}} \tag{1.60}$$

Finally, in order to avoid the deterioration in the Monte Carlo approximation due to the practice selection, the effective sample size, defined as $N_{eff} = \left(\sum_{i=1}^{n} \left(v_t^{(i)}\right)^2\right)^{-1}$, is monitored and when N_{eff} is too small, then a resampling step is implemented.

Within this general framework, many alternative specifications are provided. For instance, the most used PF are Liu and West filter (Liu and West, 2001), Storvik filter (Storvik, 2002) and Particle learning (Carvalho *et al.*, 2010; Lopes and Carvalho, 2012). See Doucet *et al.* (2001) for further and in depth discussions.

Bibliography

[1] Aldous D.J. (1985). Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XII* (P.L. Hennequin, ed.) (1985) 1-198, Berlin: Springer-Verlag. Volume **1117**.

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

[3] Arellano M., Bond S. (1991). Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations. *The Review of Economic Studies*, **58**, 277-297.

[4] Anderson D.A., Aitken M. (1985). Variance Component Models with Binary Response: Interviewer Variability. *Journal of Royal Statistical Society, Series B*, **47**, 203-210.

[5] Anderson T.W., Hsiao C. (1981). Estimation of Dynamic Models with Error Components. *Journal of the American Statistical Association*, **76**, 598-606.

[6] Balestra P., Nerlove M. (1966). Pooling Cross-Section and Time Series Data in the Estimation of Dynamic Models: The Demand for Natural Gas. *Econometrica*, **34**, 585-612.

[7] Berry D.A, Christensen R. (1979). Empirical Bayes Estimation of a Binomial Parameter Via Mixtures of Dirichlet Processes. *Annals of Statistics*, **7**, 558-568.

[3] Blackwell D., MacQueen J. B. (1973). Ferguson Distributions Via Pólya Urn Schemes. *Annals of Statistics*, **1**, 353-355.

[9] Box G., Tiao G. (1973). *Bayesian Inference in Statistical Analysis*. New York: Addison-Wesley, reprinted by New York: John Wiley & Son (1992).

[10] Bock R., Aitken M. (1981). Marginal Maximum Likelihood Estimation of Item Parameters: Applications of an EM algorithm. *Psychometrika*, **46**, 443-459.

[11] Casella G. (1992). Explaining the Gibbs sampler. *The American Statistician*, **46**, 167-174.

[12] Chib S., Greenberg E. (1995). Understanding the Metropolis-Hastings Algorithm. *Journal of American Statistical Association*, **49**, 327-335.

[13] Cifarelli D. M., Muliere P., Scarsini M. (1981). Modello lineare nell'approccio Bayesiano nonparametrico. *Quaderni dell'Istituto Matematico "G.Castelnuovo"*. Roma: University of Rome, **15**.

[14] Doucet A., De Freitas N., Gordon N.J. (eds.) (2001). *Sequential Monte Carlo Methods in Practice*. New York: Springer.

[15] Forni M., Hallin M., Lippi M., Reichlin L. (2000). The generalized dynamic factor model: Identification and estimation. *Review of Economics and Statistics*, **82**, 540-552.

[16] Gamerman D., Migon H. (1993). Dynamic hierarchical models. *Journal of the Royal Statistical Society, Series B*, **55**, 629-642.

[17] Gelman A., Carlin J.B, Stern H.S., Rubin D.B. (eds.) (2009). *Bayesian Data Analysis* (2nd edition). Boca Raton, FL: Chapman & Hall/CRC Press.

[18] Greene W. H. (2003). *Econometric Analysis* (5th edition). New Jersey: Prentice Hall.

[19] Harvey A. (1989) *Forecasting, Structural Time Series Models and the Kalman filter*. Cambridge, U.K: Cambridge University Press.

[20] Hausman J., Taylor W. (1981). Panel data and Unobservable Individual Effects. *Econometrica*, **49**, 1377-1398.

[21] Hirano K. (2002). Semiparametric Bayesian Inference in Autoregressive Panel Data Models. *Econometrica*, **70**, 781-799.

[22] Hsiao C., Pesaran M. H., Tahmiscioglu K. (1999). Bayes Estimation of Short-Run Coefficients in Dynamic Panel Data Models. In *Analysis of Panels and Limited Dependent Variables: A Volume in Honour of G. S. Maddala* (Hsiao C., Lahiri K., Lee L.F., Pesaran M.H., eds.) 268-296, Cambridge, UK: Cambridge University Press.

[23] Hsiao C., Pesaran M. H., Tahmiscioglu K. (2002). Maximum likelihood estimation of fixed effects dynamic panel data models covering short time periods. *Journal of Econometrics*, **109**, 107-150.

[24] Huber P.J., Ronchetti E.M. (2009). *Robust Statistics* (2nd edition). New York: John Wiley & Son.

[25] Im S., Gianola D. (1988). Mixed models for binomial data with an application to lamb mortality. *Applied Statistics*, **37**, 196-204.

[26] James W., Stein C. (1961). Estimation with quadratic loss. In *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability* 361-379. Berkeley: University of California Press.

[27] Korn E.L., Whittemore A.S. (1979). Methods for analyzing panel studies of acute health effects of air pollution. *Biometrics*, **73**, 805-811.

[28] Kalman R.E. (1960). A New Approach to Linear Filtering and Prediction Problems. *Transactions of the ASME, Journal of Basic Engineering, Series D*, **82**, 35-45.

[29] Kiefer J., Wolfowitz J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics*, **27**, 886 -906.

[30] Lindley D.V., Smith A.F. (1972). Bayes Estimates for the Linear Model. *Journal of the Royal Statistical Society, Series B*, **34**, 1-41.

[31] Liu J., West M. (2001). Combined parameters and state estimation in simulation- based filtering, in [6].

[32] Lo A.Y. (1984), On a class of Bayesian nonparametric estimates: I. Density estimates. *Annals of Statistics*, **12**, 351-357.

[33] Mislevy R. (1985), Estimation of latent group effects. *Journal of American Statistical Association*, **80**, 993-997.

[34] Neyman J., Scott E.L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, **16**, 1-32.

[35] Pesaran M.H. (2006). Estimation and inference in large heterogenous panels with multifactor error structure. *Econometrica*, **74** 967-1012.

[36] Pesaran M.H., Smith R.P. (1995). Estimating Long-Run Relationships from Dynamic Heterogeneous Panels. *Journal of Econometrics*, **68**, 79-113.

[37] Pesaran M.H., Tosetti E. (2011). Large panels with common factors and spatial correlations. *Journal of Econometrics*, **161**, 182-202.

[38] Petris G., Petrone S., Campagnoli P. (2009). Dynamic linear models with R. New York: Springer.

[39] Phillips P.C.B., Sul D., 2003. Dynamic panel estimation and homogeneity testing under cross section dependence. *Journal of Econometrics*, **6** , 217-259.

[40] Pitman J. (1996). Some developments of the Blackwell-MacQueen urn scheme. In *Statistics, Probability and Game Theory* (Ferguson T.S., Shapley L.S., MacQueen J.B., eds.) 245-267, Hayward, California: Institute of Mathematical Statistics.

[41] Riddell C., Rosemarie R. (2006). Welfare Checks, Drug Consumption, and Health. *The Journal of Human Resources*, **41**, 138-161.

[42] Robert C., Casella G. (2004). *Monte Carlo statistical methods* (2nd edition). New York: Springer.

[44] Robertson D., Symons J. (2000). Factor residuals in SUR regressions: estimating panels allowing for cross sectional correlation. *Centre for Economic Performance discussion paper*. London: Centre for Economic Performance.

[44] Robertson D., Symons J. (2000). Maximum likelihood factor analysis with rank- deficient sample covariance matrices. *Journal of Multivariate Analysis*, **98**, 813-828.

[45] Stein C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the 3rd Berkeley Symposium on Mathematical Statistics and Probability* 197-206. Berkeley: University of California Press.

[46] Storvik G. (2002). Particle filters for state-space models with the presence of unknown static parameters. In *IEEE Transactions on Signal Processing*, **50**, 281-89.

[47] Tierney L. (1994). Markov Chains for Exploring Posterior Distributions. *The Annals of Statistics*, **22**, 1701-1728.

[48] Vaupel J. W., Manton K. G., Stallard E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, **16**, 439-454.

[49] Walker S., Wakefield J. (1997).Bayesian Nonparametric Population Models: Formulation and Comparison with Likelihood Approaches. *Journal of Pharmacokinetics and Biopharmaceutics*, **25**, 235-253.

[50] West B., Welch K., Galecki A. (2007). *Linear Mixed Models. A Practical Guide Using Statistical Software*. Boca Raton, FL: Chapman & Hall/CRC Press.

[51] West M., Harrison J. (1997). *Bayesian Forecasting and Dynamic Models*. New York: Springer.

[52] Wooldridge J. (2001). *Econometric Analysis of cross section and panel data*. Cambridge, MA: The MIT Press.

[53] Zellner A. (1963). Estimation for Seemingly Unrelated Regression Equations: Some Exact Finite Sample Results. *Journal of the American Statistical Association*, **58**, 977- 92.

Chapter 2

An Enriched Conjugate Prior for Bayesian Nonparametric Inference

Abstract :

The precision parameter α plays an important role in the Dirichlet Process. When assigning a Dirichlet Process prior to the set of probability measures on $\mathbb{R}^k$, $k > 1$, this can be restrictive in the sense that the variability is determined by a single parameter. The aim of this chapter is to construct an enrichment of the Dirichlet Process that is more flexible with respect to the precision parameter yet still conjugate, starting from the notion of *enriched conjugate priors*, which have been proposed to address an analogous lack of flexibility of standard conjugate priors in a parametric setting. The resulting enriched conjugate prior allows more flexibility in modeling uncertainty on the marginal and conditionals. We describe an enriched urn scheme which characterizes this process and show that it can also be obtained from the stick-breaking representation of the marginal and conditionals. For non atomic base measures, this allows global clustering of the marginal variables and local clustering of the conditional variables. Finally, we consider an application to mixture models that allows for uncertainty between homoskedasticity and heteroskedasticity.

2.1 Motivation

Conjugacy is a desirable property because the posterior distribution remains analytically tractable; this is especially true in nonparametric inference where the posterior distribution of non-conjugate priors can be very complex. The most popular prior in Bayesian nonparametric inference is the Dirichlet Process. It is conjugate: if $Z_i \mid \mathbf{P} = P$ are independent and identically distributed (i.i.d.) according to P, and $\mathbf{P}$ is a Dirichlet process, $DP(\alpha P_0)$, with precision parameter α and base measure P_0 on the sample space $\mathcal{Z}$, then $\mathbf{P} \mid Z_1 = z_1, \ldots, Z_n = z_n \sim DP(\alpha P_0 + \sum_{i=1}^{n} \delta_{z_i})$. However, when Z is a random vector and $\mathbf{P}$ is a random probability measure on $\mathbb{R}^k$, $k > 1$, as in many applications, the choice of a Dirichlet process prior implies that the variability is determined by a single parameter, α. Indeed, the precision parameter α plays an important role; it not only reflects the strength of belief in the prior guess of P_0, but also controls the ties configuration in a random sample from $\mathbf{P}$. Thus, having only one degree of freedom, α, in the prior can be quite restrictive.

In fact, a similar lack of flexibility arises in a parametric setting; standard conjugate priors for the natural exponential family have only one parameter to control variability. To overcome this issue, a general class of *enriched conjugate priors* (Consonni and Veronese, 2011) have been proposed. A Dirichlet Process, $DP(\alpha P_0)$, is characterized by the fact that the finite dimensional distributions of the probability over any measurable partition,$(C_1, \ldots, C_k)$, of $\mathcal{Z}$, are Dirichlet with parameters $(\alpha P_0(C_1), \ldots, \alpha P_0(C_k))$. The Dirichlet Process inherits conjugacy from the property of conjugacy of the standard Dirichlet distribution prior for multinomial sampling, but also inflexibility from the fact that the Dirichlet distribution, as all standard conjugate priors, has only one parameter to control variability. The question addressed in this chapter is whether one can extend the notion of enriched conjugate priors to nonparametric inference and construct a prior on a random probability measure over $\mathbb{R}^k$,

that is more flexible than the DP in allowing more parameters to control the variability, yet is still conjugate.

Actually, Doksum's Neutral to the Right Process (Doksum, 1974) is an extension of the enriched conjugate Generalized Dirichlet distribution to a process, providing a more flexible, conjugate prior for *univariate* random distribution functions. The Generalized Dirichlet distribution is defined for a specific ordering of the random probabilities; thus, extension to a multivariate random distribution is not obvious, since there is no natural ordering in $\mathbb{R}^k$.

Therefore, we start our analysis by constructing an enriched Dirichlet prior for a multivariate random distribution when the sample space is finite. To convey the main ideas, we will focus on the case when the random vector Z can be partitioned into two groups, $Z = (X, Y)$, and the sample space can be written as the product of two finite spaces (or in the more general case, the product of two complete separable metric spaces, $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$). In the finite case, the enriched Dirichlet distribution is obtained based on the re-parametrization of the joint probabilities in terms of the marginal and the conditionals.

Then, we extend this construction to a process by re-parametrizing the joint random probability measure in terms of the marginal and conditionals and assigning independent Dirichlet Process priors to each of these terms. The parameters of the resulting *enriched* Dirichlet process again include a base measure controlling the location, but there are now many more parameters to control the variability. We show that the Dirichlet Process is in fact a special case, which consequently, characterizes the distribution of the random conditionals. Although many desirably properties are maintained, some are necessarily weakened, including a clear asymmetry in the two (groups of) variables, that however may be reasonable in several applications.

The chapter is organized as follows. In Section 2, we give a brief overview of enriched conjugate priors for the natural exponential family.

In Section 3, we discuss the *enriched* Dirichlet distribution in the finite case as a particular enriched conjugate prior for multinomial sampling and provide a Pólya urn characterization. These notions are extended to a process in Section 4. Finally, a simple application to mixture models is illustrated using data on national test scores to compare schools in Section 5. Proofs are given in the Appendix.

2.2 Preliminaries: Enriched conjugate priors

For a Natural Exponential Family (NEF) $\mathcal{F}$ on $\mathbb{R}^d$, where d represents the dimension of the sufficient statistics, the likelihood for the natural parameter θ is given by:

$$L_\theta(\theta|\underline{s}, n) = exp(\theta^T \underline{s} - nM(\theta)) \quad \text{for } \theta \in \Theta,$$

where $\underline{s}$ is a d-dimensional vector of the sufficient statistics, $M(\theta) = \log \int exp(\theta^T x)\eta(dx)$, and η is a σ-finite measure on the Borel sets of $\mathbb{R}^d$.

The parameter space Θ is the interior of the set $\mathcal{N} = \{\theta \in \mathbb{R}^d : M(\theta) < \infty\}$. More generally, we have a Standard Exponential Family (SEF) if $\Theta \subseteq \mathcal{N}$ and it is non-empty and open.

A family of measures on the Borel sets of Θ whose densities with respect to the Lebesgue measure are of the form $\pi_\theta(\theta|\underline{s}', n') \propto L_\theta(\theta|\underline{s}', n')$ is called the *standard conjugate family of priors* of $\mathcal{F}$ relative to the parametrization θ, where the sufficient statistics, $\underline{s}$, are replaced by parameters, $\underline{s}'$, which control the location of the prior, and the sample size, n, is replaced by a single parameter, n', which controls the precision; see Diaconis and Ylvisaker (1979).

Consonni and Veronese (2001) discuss *enriched* conjugate priors for the NEF, moving from the notion of conditional reducibility.

A d-dimensional NEF is called k *conditionally reducible* if the density

can be decomposed as the product of k standard exponential families, each depending on their own parameters.

The notion of *enriched standard conjugate priors* involves replacing the sufficient statistics and the sample size with different hyperparameters within each SEF. This means giving independent standard conjugate priors to the parameters of the conditional densities and induces a prior on the original parameter of the NEF which enriches the standard conjugate prior by allowing for k precision parameters. For further discussions, see Consonni and Veronese (2001). One important example is given by the Generalized Dirichlet distribution of Connor and Mosiman (1969), which provides an enriched conjugate prior for the parameters of a multinomial distribution; see Consonni and Veronese (2001), Example 4. Briefly, if $(N_1, \ldots, N_k)$ is multinomial given $(\mathbf{p}_1 = p_1, \ldots, \mathbf{p}_k = p_k)$, one can decompose the multinomial probability function as:

$$p(N_1 = n_1, .., N_k = n_k \mid p_1, .., p_k) =$$
$$p(N_1 = n_1 \mid V_1)p(N_2 = n_2 \mid N_1 = n_1, V_2) \cdots p(N_k = n_k \mid N_1 = n_1, .., N_{k-1} = n_{k-1}, V_k)$$

where each factor in the product is a NEF (namely, binomial), depending on its own parameter, $\mathbf{V}_1 = \mathbf{p}_1, \mathbf{V}_i = \mathbf{p}_i/(1-\sum_{j=1}^{i-1} \mathbf{p}_j), i = 2, \ldots, k-1$, and $\mathbf{V}_k$ is degenerate at 1. The standard, Dirichlet$(\alpha_1, \ldots, \alpha_k)$ conjugate prior corresponds to assuming $\mathbf{V}_i \overset{indep}{\sim}$ beta$(\alpha_i, \sum_{j=i+1}^{k} \alpha_j)$, $i = 1, \ldots, k-1$. The enriched, or Generalized, Dirichlet conjugate prior allows a more flexible choice of the beta hyperparameters: $\mathbf{V}_i \overset{indep}{\sim}$ beta(α_i, β_i), $i = 1, \ldots, k-1$. It is worth underlining that some properties of the Dirichlet distribution are necessarily weakened. In particular, the Dirichlet prior implies that *any permutation* of $(\mathbf{p}_1, \ldots, \mathbf{p}_k)$ is completely neutral (the vector $(\mathbf{p}_1, \ldots, \mathbf{p}_k)$ is completely neutral iff $(\mathbf{p}_1, \mathbf{p}_2/(1-\mathbf{p}_1), ...\mathbf{p}_k/(1-\sum_{j=1}^{k-1} \mathbf{p}_j))$ are independent). The Generalized Dirichlet only assumes that *one* ordered vector $(\mathbf{p}_1, \ldots, \mathbf{p}_k)$ is completely neutral. This makes applications to the bivariate case of contingency

tables $\mathbf{p}_{i,j}$ not obvious, since there is no natural ordering in two dimensions. The enriched conjugate prior that we propose in the next section is a simple proposal in this direction.

2.3 Finite case: Enriched Dirichlet distribution

Let $\{(X_n, Y_n)\}_{n\in\mathbb{N}}$ be a sequence of discrete random vectors with values in $\mathcal{X} \times \mathcal{Y} = \{1, \ldots, k\} \times \{1, \ldots, m\}$, such that $(X_i, Y_i) \mid \mathbf{p} = p \overset{iid}{\sim} p$, where $\mathbf{p}$ is a random probability function with mass $\mathbf{p}_{i,j}$ on $(i, j), i = 1, \ldots, k; j = 1, \ldots, m$. Then, given $\mathbf{p} = p$, the vector of counts $(N_{1,1}, \ldots, N_{k,m})$, where $N_{i,j}$ is the number of times the pair (i, j) is observed in a sample $((X_1, Y_1), \ldots, (X_n, Y_n))$, has a multinomial probability function

$$p(n_{1,1}, .., n_{k,m-1} \mid p_{1,1}, .., p_{k,m-1}) =$$
$$\frac{n!}{n_{1,1}!..n_{k,m-1}!(n - \sum_{(i,j)\neq(k,m)} n_{i,j})!} p_{1,1}^{n_{1,1}} \cdots p_{k,m-1}^{n_{k,m-1}} (1 - \sum_{(i,j)\neq(k,m)} p_{i,j})^{n - \sum_{(i,j)\neq(k,m)} n_{i,j}}, \tag{2.1}$$

for $n_{i,j} \geq 0; \sum_{i=1}^{k} \sum_{j=1}^{m} n_{i,j} = n$. The standard conjugate prior for $(\mathbf{p}_{1,1}, \ldots, \mathbf{p}_{k,m})$ is the Dirichlet distribution, which involves replacing the $km - 1$ sufficient statistics in (2.1) with hyperparameters, $\underline{s'} = (s'_{1,1}, ..., s'_{k,m-1})$, that control the location of the prior, and the sample size with a single hyperparameter, n', that controls the precision of the prior.

As discussed in Section 2, a generalized Dirichlet prior is problematic in this case, since there is no natural ordering of the probabilities $\mathbf{p}_{i,j}$. However, a fairly natural and simple enrichment can be obtained by first

applying the linear transformation:

$$N_{i+} = \sum_{j=1}^{m} N_{i,j} \quad \text{for } i = 1, ..., k-1,$$
$$N_{i,j} = N_{i,j} \quad \text{for } i = 1, ..., k \quad j = 1, ..., m-1,$$

followed by the reparametrization:

$$\mathbf{p}_{i+} = \sum_{j=1}^{m} \mathbf{p}_{i,j} \quad \text{for } i = 1, ..., k-1,$$
$$\mathbf{p}_{j|i} = \frac{\mathbf{p}_{i,j}}{\mathbf{p}_{i+}} \quad \text{for } i = 1, ..., k-1 \quad j = 1, ..., m-1,$$
$$\mathbf{p}_{j|k} = \frac{\mathbf{p}_{k,j}}{1 - \sum_{i=1}^{k-1} \mathbf{p}_{i+}} \quad \text{for } j = 1, ..., m-1.$$

Define: $\underline{N}_+ = (N_{1+}, ..., N_{k-1+})$; $\underline{N}_i = (N_{i,1}, ..., N_{i,m-1})$; $\underline{\mathbf{p}}_+ = (\mathbf{p}_{1+}, ..., \mathbf{p}_{k-1+})$, and $\underline{\mathbf{p}}_i = (\mathbf{p}_{1|i}, ..., \mathbf{p}_{m-1|i})$, for $i = 1, ., k$. Under this linear transformation and re-parametrization, the multinomial is a $k+1$ conditionally reducible NEF:

$$p(\underline{n}_+, \underline{n}_1, ..., \underline{n}_k \mid \underline{p}_+, \underline{p}_1, ..., \underline{p}_k) = p(\underline{n}_+ \mid \underline{p}_+) \prod_{i=1}^{k} p(\underline{n}_i \mid \underline{p}_i, \underline{n}_+), \quad (2.2)$$
$$(N_{i,1}, ..., N_{i,m} \mid n_{i+}, p_{1|i}, ..., p_{m|i}) \sim \text{Mult}\left(n_{i+}, p_{1|i}, ..., p_{m|i}\right) \quad \text{for } i = 1, ..., k,$$
$$(N_{1+}, ..., N_{k+} \mid p_{1+}, ..., p_{k+}) \sim \text{Mult}(n, p_{1+}, ..., p_{k+}).$$

By replacing the sufficient statistics and sample size with different parameters within each SEF on the right hand side of (2.2), one can create a more flexible conjugate prior. In particular, letting $(\underline{s}'_{(+)}, \underline{s}'_{(1)}, ..., \underline{s}'_{(k)})$ denote the $km-1$ location parameters and $(n'_+, n'_1, ..., n'_k)$ denote the precision parameters, in terms of $(\underline{\mathbf{p}}_+, \underline{\mathbf{p}}_1, ..., \underline{\mathbf{p}}_k)$, the Enriched Dirichlet

conjugate prior is:

$$\mathbf{p}_{1+},...,\mathbf{p}_{k-1+} \sim \text{Dir}(s'_{1+},...,s'_{k-1+}, n'_{+} - \sum_{i=1}^{k-1} s'_{i+}), \tag{2.3}$$
$$\mathbf{p}_{1|i},...,\mathbf{p}_{m-1|i} \sim \text{Dir}(s'_{i,1},...,s'_{i,m-1}, n'_{i} - \sum_{j=1}^{m-1} s'_{i,j}),$$

where $(\mathbf{p}_{1+},...,\mathbf{p}_{k-1+}), (\mathbf{p}_{1|1},...,\mathbf{p}_{m-1|1}),...,(\mathbf{p}_{1|k},...,\mathbf{p}_{m-1|k})$ are independent. We get back to the Dirichlet distribution if $n'_i = s'_{i+}$ for $i = 1,...k-1$ and $n'_{+} = \sum_{i=1}^{k} n'_i$.

Remark 1. The Dirichlet distribution on the vector $\underline{\mathbf{p}} = (\mathbf{p}_{1,1},...,\mathbf{p}_{k,m})$ defining the random marginal, $\mathbf{p}_x$, $\mathbf{p}_y$, and conditional, $\mathbf{p}_{y|x}$, $\mathbf{p}_{x|y}$, probability functions is characterized by the properties
(i) $\mathbf{p}_x(\cdot)$ and $\mathbf{p}_{y|x}(\cdot|i), i = 1,\ldots,k$ are independent, and
(ii) $\mathbf{p}_y(\cdot)$ and $\mathbf{p}_{x|y}(\cdot|j), j = 1,\ldots,m$ are independent;
see Geiger and Heckerman (1997). The Enriched Dirichlet relaxes that the independence properties holds in both directions. We maintain (i) and allow more degrees of freedom in the distributions of $\mathbf{p}_x$ and $\mathbf{p}_{y|x}$.

Remark 2. Under the linear transformation discussed here, the multinomial could also be viewed as a $km-1$ conditionally reducible NEF; it can be written as the product of $km-1$ SEF (namely, binomial) each depending on its own parameters. The resulting enriched conjugate prior has $km-1$ parameters to control the precision and can be seen as nested version of Generalized Dirichlet distribution of Connor and Mosimann (1969).

In the rest of the chapter, we will use the following parametrization of the distributions (2.3). Let $\alpha(\cdot)$ be a finite measure on $\mathcal{X}$ and $\mu(\cdot,\cdot)$ be a mapping from $2^{\mathcal{Y}} \times \mathcal{X}$ to $\mathbb{R}_+$ such that for every $x \in \mathcal{X}$, $\mu(\cdot, x)$ is a

finite measure on $(\mathcal{Y}, 2^{\mathcal{Y}})$. Then we assume that the parameters in (2.3) are chosen in terms of $\alpha(\cdot)$ and and $\mu(\cdot,\cdot)$:

$$\begin{aligned} &\mathbf{p}_{1+},...,\mathbf{p}_{k+} \sim \text{Dir}(\alpha(1),...,\alpha(k)), \\ &\mathbf{p}_{1|i},...,\mathbf{p}_{m|i} \sim \text{Dir}(\mu(1,i),...,\mu(m,i)) \quad i=1,...,k, \end{aligned} \tag{2.4}$$

with the convention that if $\alpha(i) = 0$ then $\mathbf{p}_{i+}$ is degenerate at 0 and if $\mu(j,i) = 0$ then $\mathbf{p}_{j|i}$ is degenerate at 0. If $\alpha(i) > 0$ and $\mu(j,i) > 0$ for all i, j, then the enriched Dirichlet conjugate prior induced on $(\mathbf{p}_{1,1},\ldots,\mathbf{p}_{k,m})$ is:

$$f(p_{1,1},...,p_{k,m-1}) = \frac{\Gamma(\alpha(\mathcal{X}))}{\prod_{i=1}^{k}\Gamma(\alpha(i))}\prod_{i=1}^{k-1}(\sum_{j=1}^{m}p_{i,j})^{\alpha(i)-\mu(\mathcal{Y},i)}(1-\sum_{i=1}^{k-1}\sum_{j=1}^{m}p_{i,j})^{\alpha(k)-\mu(\mathcal{Y},k)}$$
$$\prod_{i=1}^{k}\frac{\Gamma(\mu(\mathcal{Y},i))}{\prod_{j=1}^{m}\Gamma(\mu(j,i))}\prod_{j=1}^{m-1}p_{i,j}^{\mu(j,i)-1}\prod_{i=1}^{k-1}p_{i,m}^{\mu(m,i)-1}(1-\sum_{i=1}^{k}\sum_{j=1}^{m-1}p_{i,j}-\sum_{i=1}^{k-1}p_{i,m})^{\mu(m,k)-1}.$$

To avoid making the notation heavier, we assume α and μ known if not otherwise specified.

Clearly, the prior of the marginal probabilities $(\mathbf{p}_{+1},\ldots,\mathbf{p}_{+m})$ on $\mathcal{Y}$ is no longer a Dirichlet distribution, and in fact, the density may not be available in closed form. But, we can give the following representation in terms of G-Meijer variables (Springer and Thompson, 1970). First, remembering the Gamma representation of the Dirichlet distribution and defining $\mathbf{U}_i \overset{indep}{\sim} \text{Gamma}(\alpha(i),1)$ and $\mathbf{V}_{ij} \overset{indep}{\sim} \text{Gamma}(\mu(j,i),1)$, we have the following G-Meijer representation of the vector $(\mathbf{p}_{1,1},...,\mathbf{p}_{k,m})$:

$$(\mathbf{p}_{1,1},...,\mathbf{p}_{k,m}) =^{d} \left(\frac{\mathbf{U}_1\mathbf{V}_{1,1}}{\sum_{i=1}^{k}\mathbf{U}_i\sum_{j=1}^{m}\mathbf{V}_{1j}},...,\frac{\mathbf{U}_k\mathbf{V}_{km}}{\sum_{i=1}^{k}\mathbf{U}_i\sum_{j=1}^{m}\mathbf{V}_{k,j}}\right),$$

which is independent of $\sum_{i=1}^{k}\mathbf{U}_i\sum_{j=1}^{m}\mathbf{V}_{1j},...,\sum_{i=1}^{k}\mathbf{U}_i\sum_{j=1}^{m}\mathbf{V}_{k,j}$; where the sym-

bol $=^d$ denotes equality in distribution. Therefore, the marginal probabilities over $\mathcal{Y}$ can be represented as the sum of G-Meijer random variables:

$$(\mathbf{p}_{+1}, ..., \mathbf{p}_{+m}) =^d \left(\sum_{i=1}^{k} \frac{\mathbf{U}_i \mathbf{V}_{i,1}}{\sum_{h=1}^{k} \mathbf{U}_h \sum_{j=1}^{m} \mathbf{V}_{ij}}, ..., \sum_{i=1}^{k} \frac{\mathbf{U}_i \mathbf{V}_{im}}{\sum_{h=1}^{k} \mathbf{U}_h \sum_{j=1}^{m} \mathbf{V}_{ij}} \right).$$

2.3.1 Enriched Pólya Urn

An alternative way to define the Enriched Dirichlet distribution is based on a Pólya urn scheme, which will be useful in extending the distribution to a process. In the bivariate setting, the standard Pólya urn scheme describes the predictive distribution of a sequence of random vectors. An urn contains pairs of balls of color $(i, j) \in \mathcal{X} \times \mathcal{Y}$. A pair of balls is drawn from the urn and replaced along with another pair of balls of the same colors. The random vector, (X_n, Y_n), is equal to (i, j) if the nth pair drawn is of color (i, j).

Alternatively, we can consider one urn containing just X-balls and k urns, say $Y|i$ urns, containing only Y-balls. We first draw an X-ball from the X-urn and replace it along with another ball of the same color, and then, depending on color of the X-ball, draw a Y-ball from urn associated with X-ball drawn, and replace it along with another ball of the same color. In this case, the random vector, (X_n, Y_n), is equal to (i, j) if the nth X-ball drawn is of color i and the Y ball associated with it is of color j. If the number of Y-balls in the $Y|i$ urn is equal to the number balls of color i in the X-urn, the two urn schemes are equivalent.

The Enriched Pólya Urn scheme enriches this urn scheme by relaxing the constraint that the number of Y-balls in the $Y|i$ urn has to equal the number of X-balls of color i in the X-urn. More precisely, the number of balls in each urn is specified as follows:

- $\alpha(i)$ is the number of X-balls of color i
- $\mu(j,i)$ is the number of Y-balls of color j in the $Y|i$ urn

where $\alpha(\mathcal{X}) = \sum_{i=1}^{k} \alpha(i)$ is the total number of balls in the X-urn and $\mu(\mathcal{Y},i) = \sum_{j=1}^{m} \mu(j,i)$ is the total number of balls in the $Y|i$ urn for $i = 1,...,k$. This urn scheme implies the following predictive distribution:

$$Pr(X_1 = i, Y_1 = j) = \frac{\alpha(i)}{\alpha(\mathcal{X})}\frac{\mu(j,i)}{\mu(\mathcal{Y},i)},$$

$$Pr(X_{n+1} = i, Y_{n+1} = j | X_1 = i_1, Y_1 = j_1, .., X_n = i_n, Y_n = j_n)$$
$$= \frac{\alpha(i) + \sum_{h=1}^{n} \delta_{i_h}(i)}{\alpha(\mathcal{X}) + n} \; \frac{\mu(j,i) + \sum_{h=1}^{n} \delta_{j_h,i_h}(j,i)}{\mu(\mathcal{Y},i) + \sum_{h=1}^{n} \delta_{i_h}(i)}.$$

Theorem 2.3.1 *Let $\{(X_n,Y_n)\}_{n\in\mathbb{N}}$ be a sequence of random vectors taking values in $\{1,...,k\}\times\{1,...,m\}$ with predictive distributions characterized by an Enriched Pólya urn scheme with parameters $\alpha(\cdot)$ and $\mu(\cdot,\cdot)$. Then,*

1. *the sequence of random vectors $\{(X_n,Y_n)\}_{n\in\mathbb{N}}$ is exchangeable, and its de Finetti measure is an Enriched Dirichlet distribution with parameters $\alpha(\cdot)$ and $\mu(\cdot,\cdot)$.*
2. *as $n \to \infty$, the sequence of the predictive distributions $p_n(i,j) = Pr(X_{n+1} = i, Y_{n+1} = j | X_1 = i_1, Y_1 = j_1, .., X_n = i_n, Y_n = j_n)$ converges a.s with respect to the exchangeable law to a random probability function, $\mathbf{p}$; and $\mathbf{p}$ is distributed according to the Enriched Dirichlet de Finetti measure.*

The proof is an extension of that used for the standard Pólya urn (see Ghosh and Ramamoorthi, 2003: 94-95)). The first step is to show the

sequence of random vectors is exchangeable. Next, computing their finite dimensional distributions and using de Finetti Representation Theorem, the random vectors are shown to be i.i.d given the random variables $(\mathbf{p}_{1+}, ...\mathbf{p}_{k+}, \mathbf{p}_{1|1}, ..., \mathbf{p}_{m|k}) = (p_{1+}, ...p_{k+}, p_{1|1}, ..., p_{m|k})$ which are distributed according to an Enriched Dirichlet distribution with parameters α and μ. A detailed proof is given in the Appendix.

2.4 Enriched Dirichlet Process

Assume $\mathcal{X}$ and $\mathcal{Y}$ are complete and separable metric spaces and $\mathcal{B}_X$ and $\mathcal{B}_Y$ are their Borel σ-algebras. Let $\mathcal{B}$ be the σ-algebra generated by the product of the σ-algebras of $\mathcal{X}$ and $\mathcal{Y}$ and $\mathcal{P}(\mathcal{B})$ be the set of probability measures on the measurable product space $(\mathcal{X} \times \mathcal{Y}, \mathcal{B})$ where $\mathcal{P}(\mathcal{B}_X)$, $\mathcal{P}(\mathcal{B}_Y)$ are similarly defined. For any $P \in \mathcal{P}(\mathcal{B})$, let P_X denote the marginal probability measure, $P_{Y|X}(\cdot|x)$ for $x \in \mathcal{X}$ denote a version of the conditional, and $P_{Y|X}$ denote the entire version of the conditional as an element of $\mathcal{P}(\mathcal{B}_Y)^{\mathcal{X}}$. Here, we consider the Borel σ-algebra under weak convergence on $\mathcal{P}(\mathcal{B})$, $\mathcal{P}(\mathcal{B}_X)$, and $\mathcal{P}(\mathcal{B}_Y)$ and the product σ-algebra on $\mathcal{P}(\mathcal{B}_Y)^{\mathcal{X}}$. We will define a probability measure on $\mathcal{P}(\mathcal{B})$ that is more flexible than the Dirichlet Process with respect to the precision parameter and still retains conjugacy by extending the ideas of the Enriched Dirichlet distribution.

Note that trying to enrich the DP by using the Enriched Dirichlet in place of the Dirichlet as the finite dimensional distributions, i.e., defining a random $\mathbf{P}$ such that $(\mathbf{P}(A_1 \times B_1), \ldots, \mathbf{P}(A_k \times B_m)) \sim$ Enriched Dirichlet distribution, would not succeed because finite additivity holds only with a specification of the parameters that is equivalent to the Dirichlet distribution.

Instead, we use directly the idea of the Enriched Dirichlet distribution, which defines a prior for the joint by decomposing it in terms of the marginal and conditionals and giving them independent conjugate priors. If $\mathcal{X}, \mathcal{Y}$ are general spaces, it is a delicate issue to establish that such an

approach induces a prior on the joint. In particular, given a prior on $\mathcal{P}(\mathcal{B}_X) \times \mathcal{P}(\mathcal{B}_Y)^{\mathcal{X}}$, the map $(P_X, P_{Y|X}) \to \int_{(\cdot)} P_{Y|X}(\cdot|x)dP_X(x)$ induces a prior on $\mathcal{P}(\mathcal{B})$ if it is jointly measurable in $(P_X, P_{Y|X})$, which is not generally true. Fortunately, if the prior for the marginal concentrates on the set of discrete probability measures and independence assumptions hold, the prior on the marginal and conditionals can be restricted to a subspace of $\mathcal{P}(\mathcal{B}_X)\times\mathcal{P}(\mathcal{B}_Y)^{\mathcal{X}}$ that has measure one, and on this subspace, the mapping is measurable, which is shown after the following definition.

Definition Let α be a finite measure on $(\mathcal{X}, \mathcal{B}_X)$ and μ be a mapping from $(\mathcal{B}_Y \times \mathcal{X})$ to $\mathbb{R}_+$ such that as a function of $B \in \mathcal{B}_Y$ it is a finite measure on $(\mathcal{Y}, \mathcal{B}_Y)$ and as a function of $x \in \mathcal{X}$ it is α-integrable. Assume:

1. Law of Marginal, Q^X: $\mathbf{P}_X$ is a random probability measure on $(\mathcal{X}, \mathcal{B}_X)$ where $\mathbf{P}_X \sim DP(\alpha)$.

2. Law of Conditionals, $Q_x^{Y|X}$: $\forall x \in \mathcal{X}$, $\mathbf{P}_{Y|X}(\cdot|x)$ is a random probability measure on $(\mathcal{Y}, \mathcal{B}_Y)$ where $\mathbf{P}_{Y|X}(\cdot|x) \sim DP(\mu(\cdot, x))$.

3. Joint Law of Conditionals, $Q^{Y|X} = \prod_{x\in\mathcal{X}} Q_x^{Y|X}$: $\mathbf{P}_{Y|X}(\cdot|x), x \in \mathcal{X}$ are independent among themselves.

4. Joint Law of Marginal and Conditionals, $Q = Q^X \times Q^{Y|X}$: $\mathbf{P}_X$ is independent of $\left\{\mathbf{P}_{Y|X}(\cdot|x)\right\}_{x\in\mathcal{X}}$.

The joint law of the marginal and conditionals, Q, induces the law, $\tilde{Q}$, of the stochastic process $\{\mathbf{P}(C)\}_{C\in\mathcal{B}}$ through the following re-parametrization:

$$\mathbf{P}(A \times B) =^d \int_A \mathbf{P}_{Y|X}(B \mid x)d\mathbf{P}_X(x), \quad \text{for any set } A \times B \in \mathcal{B}_X \times \mathcal{B}_Y. \tag{2.5}$$

This process is called an *Enriched Dirichlet Process* (EDP) with parameters α and μ, and is denoted $\mathbf{P} \sim EDP(\alpha, \mu)$.

The following arguments verify that the four conditions in definition (2.4) induce a distribution for the random joint probability measure. In particular, we define a subspace of $\mathcal{P}(\mathcal{B}_X) \times \mathcal{P}(\mathcal{B}_Y)^{\mathcal{X}}$ that has measure one, such that on this subspace, the mapping $(P_X, P_{Y|X}) \to \int_{(\cdot)} P_{Y|X}(\cdot|x)dP_X(x)$ is measurable.

First note that in order for $\left\{\mathbf{P}_{Y|X}(\cdot|x), x \in \mathcal{X}\right\}$ to be a set of conditional random probability measures, the following two properties need to be satisfied:

1. $\forall x \in \mathcal{X}$, $\mathbf{P}_{Y|X}(\cdot|x)$ is a probability measure on $(\mathcal{Y}, \mathcal{B}_Y)$ a.s $Q_x^{Y|X}$.
2. $\forall B \in \mathcal{B}_Y$, as a function of x, $\mathbf{P}_{Y|X}(B|x)$ is $\mathcal{B}_X$ measurable a.s $Q^{Y|X}$.

The first item is satisfied since $\mathbf{P}_{Y|X}(\cdot|x) \sim DP(\mu(\cdot, x))$ implies $\mathbf{P}_{Y|X}(\cdot|x) \in \mathcal{P}(\mathcal{B}_Y)$ with probability one. The second property follows from results of Ramamoorthi and Sangalli (2006). In particular, letting Δ be the subset of $\mathcal{P}(\mathcal{B}_Y)^{\mathcal{X}}$ such that $P_{Y|X}$ is measurable as a function of x, they show that if $\mathbf{P}_{Y|X}(\cdot|x)$ are independent among $x \in \mathcal{X}$, then the product measure, $Q^{Y|X} = \prod_{x \in \mathcal{X}} Q_x^{Y|X}$, given by Kolmogorov's Extension Theorem, assigns outer measure one to Δ.

Let $\mathcal{P}_D(\mathcal{B}_X)$ denote the set of discrete probability measures on the measurable space $(\mathcal{X}, \mathcal{B}_X)$. From properties of the DP, $Q^X(\mathcal{P}_D(\mathcal{B}_X)) = 1$. Therefore, by independence of $\mathbf{P}_X$ and $\mathbf{P}_{Y|X}$, the set $\mathcal{P}_D(\mathcal{B}_X) \times \Delta$ has measure one. Again, by results of Ramamoorthi and Sangalli (2006), on $\mathcal{P}_D(\mathcal{B}_X) \times \Delta$, for $A \in \mathcal{B}_X$ and $B \in \mathcal{B}_Y$, the function $(P_X, P_{Y|X}) \to \int_A P_{Y|X}(B|x)dP_X(x)$ is jointly measurable in $(P_X, P_{Y|X})$. These results imply that we can define a prior, $\tilde{Q}$, on $\mathcal{P}(\mathcal{B})$ induced from Q restricted to $\mathcal{P}_D(\mathcal{B}_X) \times \Delta$ via the map $(P_X, P_{Y|X}) \to \int_{(\cdot)} P_{Y|X}(\cdot|x)dP_X(x)$.

Obviously, this map is not $1-1$. In fact, the definition of the EDP states that the four conditions hold for the joint distribution of $(\mathbf{P}_X, \mathbf{P}_{Y|X})$ for some version of the conditional, and this induces a prior

on the joint. However, from the induced prior on the random joint probability measure, we can obtain the joint distribution of $\mathbf{P}_X$ and $\mathbf{P}_{Y|X}$ through the mapping $\mathbf{P} \to (\mathbf{P}_X, \mathbf{P}_{Y|X})$ defined from any version of the conditional. In the next section, we show that although the mapping is not 1-1, the joint law of $\mathbf{P}_X$ and $\mathbf{P}_{Y|X}$ defined from any version of the conditional and the induced law of the joint probability measure still satisfies the conditions in definition (2.4) through an extension of the enriched Pólya urn scheme to the infinte case.

2.4.1 Enriched Pólya Sequence

Similar to Blackwell and MacQueen (1973), we define an Enriched Pólya sequence which extends the enriched Pólya Urn scheme to the case when $\mathcal{X}$ and $\mathcal{Y}$ are complete separable metric spaces.

Definition The sequence of random vectors $\{(X_n, Y_n)\}_{n\in\mathbb{N}}$ taking values in $\mathcal{X} \times \mathcal{Y}$ is an *Enriched Pólya sequence* with parameters α and μ if:

1. For $A \in \mathcal{B}_X$ and for all $n \geq 1$,

$$Pr(X_1 \in A) = \frac{\alpha(A)}{\alpha(\mathcal{X})},$$

$$Pr(X_{n+1} \in A \mid X_1 = x_1, ..., X_n = x_n) = \frac{\alpha(A) + \sum_{i=1}^{n} \delta_{x_i}(A)}{\alpha(\mathcal{X}) + n}.$$

2. For $B \in \mathcal{B}_Y$ and for all $n \geq 1$,

$$Pr(Y_1 \in B \mid X_1 = x) = \frac{\mu(B, x)}{\mu(\mathcal{Y}, x)},$$

$$Pr(Y_{n+1} \in B \mid Y_1 = y_1, ..., Y_n = y_n, X_1 = x_1, ..., X_n = x_n, X_{n+1} = x)$$

$$= \frac{\mu(B, x) + \sum_{j=1}^{n_x} \delta_{y_{x,j}}(B)}{\mu(\mathcal{Y}, x) + n_x},$$

$$\text{where } n_x = \sum_{i=1}^{n} \delta_{x_i}(x) \text{ and } \{y_{x,j}\}_{j=1}^{n_x} = \{y_i : x_i = x, i = 1, ..., n\}.$$

In words, the predictive distributions characterizing the Enriched Pólya sequence can be interpreted in terms of draws from urns as follows; initially, there is an X-urn containing $\alpha(\mathcal{X})$ balls of color 0. A ball is first drawn from the X-urn, and once drawn, its true color, x_1, is revealed (where x_1 is the realization of a draw from $P_{0X} = \frac{\alpha}{\alpha(\mathcal{X})}$). A ball of color x_1 is added to the urn along with a ball of color 0, so that the urn is now composed of $\alpha(\mathcal{X})$ balls of color 0 and one ball of color x_1. Once the true color x_1 of the X-ball is revealed, a $Y|x_1$-urn is created with $\mu(\mathcal{Y}, x_1)$ balls of color 0. Next, a ball is drawn from the $Y|x_1$-urn and similarly, once drawn its true color is revealed to be y_1 (where y_1 is the realization of a draw from $P_{0Y|X}(\cdot|x_1) = \frac{\mu(\cdot, x_1)}{\mu(\mathcal{Y}, x_1)}$). This ball is then added to the $Y|x_1$-urn along with a ball of color 0, so that the urn contains $\mu(\mathcal{Y}, x_1)$ balls of color 0 and one ball of color y_1.

At the next stage, we again first draw a ball from the X-urn. We can either draw a 0 ball or an x_1 ball. If an x_1 ball is drawn, we replace it along with another ball of the same color and then draw a Y-ball from the $Y|x_1$ urn. If X-ball drawn is of color 0, then once drawn its true color is revealed, x_2, we add a ball of color x_2 to the X-urn and create a $Y|x_2$ urn with $\mu(\mathcal{Y}, x_2)$ balls of color 0. This process is repeated, so that a new $Y|x$ urn is created for each new value of X that is observed.

Note that if $\mathbf{P} \sim EDP(\alpha, \mu)$ and the random vectors $(X_1, Y_1), ...(X_n, Y_n)$ given $\mathbf{P} = P$ are i.i.d and distributed according to P, then $\{(X_n, Y_n)\}_{n\in\mathbb{N}}$ is an enriched Pólya sequence. Conversely, the following theorem proves that if $\{(X_n, Y_n)\}_{n\in\mathbb{N}}$ is an Enriched Pólya sequence, then given a random probability measure $\mathbf{P} = P$, the random vectors $(X_1, Y_1), ...(X_n, Y_n)$ are i.i.d and distributed according to P where the joint distribution of $(\mathbf{P}_X, \mathbf{P}_{Y|X})$ defined from any fixed version of the conditional satisfies the four conditions in definition (2.4). Therefore, in addition to the fact that the de Finetti measure of an Enriched Pólya sequence is an Enriched

Dirichlet Process, this theorem also shows that the induced law of the random joint from the four conditions in definition (2.4) still maintains those properties even though the mapping is not $1-1$.

Theorem 2.4.1 *If $\{(X_n, Y_n)\}_{n\in\mathbb{N}}$ is an Enriched Pólya sequence with parameters α and μ, then $\{(X_n, Y_n)\}_{n\in\mathbb{N}}$ is an exchangeable sequence and its de Finetti measure is an Enriched Dirichlet Process with parameters (α, μ).*

For a quick sketch of the proof, we start by showing that the sequence $\{(X_n, Y_n)\}_{n\in\mathbb{N}}$ is exchangeable, and then apply de Finetti's Theorem. Finally, after reparametrizing in terms of the marginal and the conditionals, we verify the de Finetti measure satisfies the four conditions in the definition of the EDP. A detailed proof is given in the Appendix.

2.4.2 Properties

Define $P_{0X} = \frac{\alpha}{\alpha(\mathcal{X})}$ and for every $x \in \mathcal{X}$, $P_{0Y|X}(\cdot|x) = \frac{\mu(\cdot,x)}{\mu(\mathcal{Y},x)}$. From well known properties of the Dirichlet process, we have:

Proposition 2.4.2 *If $\mathbf{P} \sim EDP(\alpha, \mu)$, for $A \in \mathcal{B}_X, B \in \mathcal{B}_Y$,*

- $E[\mathbf{P}_X(A)] = P_{0X}(A)$; $\quad Var(\mathbf{P}_X(A)) = \frac{P_{0X}(A)(1-P_{0X}(A))}{\alpha(\mathcal{X})+1}$.
- $\forall x \in \mathcal{X}$, $E[\mathbf{P}_{Y|X}(B \mid x)] = P_{0Y|X}(B|x)$; $\quad Var(\mathbf{P}_{Y|X}(B|x)) = \frac{P_{0Y|X}(B|x)(1-P_{0Y|X}(B|x))}{\mu(\mathcal{Y},x)+1}$.
- $E[\mathbf{P}(A \times B)] = \int_A P_{0Y|X}(B|x) dP_{0X}(x) := P_0(A \times B)$.

Therefore, similar to the DP, the location of the EDP is determined by the base measure P_0, but there are now many more parameters to control the precision, namely $\alpha(\mathcal{X})$ and $\mu(\mathcal{Y}, x)$ for every $x \in \mathcal{X}$. The following proposition states that the DP is in fact a special case of the EDP.

Proposition 2.4.3 $\mathbf{P} \sim EDP(\alpha, \mu)$ *with* $\mu(\mathcal{Y}, x) = \alpha(x), \forall x \in \mathcal{X}$ *is equivalent to* $\mathbf{P} \sim DP(\alpha(\mathcal{X})P_0)$.

The proof relies on the urn characterization of both processes; we show that an Enriched Pólya sequence is equivalent to a Pólya sequence with parameter $\alpha(\mathcal{X})\, P_0(\cdot)$, if $\mu(\mathcal{Y}, x) = \alpha(x)$, $\forall x \in \mathcal{X}$. A more detailed proof is given in the Appendix.

As a by-product of this proposition, if $\mathbf{P} \sim DP(\alpha(\mathcal{X})P_0)$, the law of the random conditionals is $\mathbf{P}_{Y|X}(\cdot|x) \sim DP(\alpha(x)P_{0Y|X}(\cdot|x))$, where $\mathbf{P}_{Y|X}(\cdot|x)$ is independent among $x \in \mathcal{X}$. In general, the marginal base measure can assign positive mass to countably many locations. All random conditional probability measures associated with x that has positive mass under the marginal base measure will be a DP with precision parameter equivalent to the mass under the marginal base measure times α. Since a DP with precision parameter 0 is degenerate a random location with probability one, the random conditional probability measures associated with all the other x will be degenerate at some $y \in \mathcal{Y}$ with probability one. Thus, in the case when P_0 is non-atomic, a DP implies assuming the conditionals are independent and degenerate a.s. which is consistent with results in Ramamoorthi and Sangalli (2006).

As noted by Ferguson (1973), a prior for nonparametric problems should have large topological support. The following theorem shows that the EDP has full weak support. Here, $\mathcal{X} = \mathbb{R}^{k_1}$ and $\mathcal{Y} = \mathbb{R}^{k_2}$, implying $\mathcal{X} \times \mathcal{Y} = \mathbb{R}^k$ where $k = k_1 + k_2$.

Theorem 2.4.4 *Let S_0 denote the topological support of P_0. If $\mathbf{P} \sim EDP(\alpha, \mu)$, then the topological support of $\mathbf{P}$ is*

$$M_0 = \{P \in \mathcal{P}(\mathcal{B}) : \textit{topological support}(P) \in S_0\}\,.$$

2.4.3 Posterior

Just as the finite dimensional Enriched Dirichlet distribution is conjugate to the multinomial likelihood, the Enriched Dirichlet Process is also conjugate for estimating a completely unknown distribution from exchangeable data. More precisely,

Proposition 2.4.5 *If* $(X_i, Y_i) \mid \mathbf{P} = P \overset{\text{iid}}{\sim} P$, *where* $\mathbf{P} \sim EDP(\alpha, \mu)$, *then*

$$\mathbf{P} \mid x_1, y_1, ..., x_n, y_n \sim EDP(\alpha_n, \mu_n),$$

where

$$\alpha_n = \alpha + \sum_{i=1}^{n} \delta_{x_i},$$

$$\forall x \in \mathcal{X} \quad \mu_n(\cdot, x) = \mu(\cdot, x) + \sum_{j=1}^{n_x} \delta_{y_{x,j}}; \quad n_x = \sum_{i=1}^{n} \delta_{x_i}(x), \quad \{y_{x,j}\}_{j=1}^{n_x} = \{y_j : x_j = x\}.$$

The proof of conjugacy is straightforward; one simply has to demonstrate that given the random sample the four conditions in the definition of EDP hold with the updated parameters specified above. The first two conditions, the fact that the marginal and conditionals are DP's with updated parameters, follow from conjugacy of the DP. The last two conditions, independence of the marginal and conditionals and independence among the conditionals, follow by combining the fact that a priori independence holds with independence of the random vectors $(X_1, ..., X_n)$ and $(Y_1, ..., Y_n | X_1 = x_1, ..., X_n = x_n)$ and independence of the random vectors $\{Y_{x,j}\}_{j=1}^{n_x}$ among $x \in \mathcal{X}$.

Posterior consistency is a frequentist validation tool that is useful in Bayesian nonparametric inference where the infinite dimension of the parameter space can make specification of a prior challenging and cause the prior to strongly influence the posterior even with large amounts of data. One of the reasons that makes the Dirichlet Process so appealing is that the posterior is weakly consistent for any probability measure, P^*, on the product space under the assumption that the sequence of random vectors are distributed according to the i.i.d. product measure $P^{*\infty}$. Another important property that the EDP maintains is posterior consistency. The proof requires that for a set $A \times B \in \mathcal{B}_X \times \mathcal{B}_Y$, the posterior expectation

of $\mathbf{P}(A \times B)$ converges to $P^*(A \times B)$ and its posterior variance goes to zero. In the following lemma, the variance of the probability over a set $A \times B \in \mathcal{B}_X \times \mathcal{B}_Y$ is specified.

Lemma 2.4.6 *If* $\mathbf{P} \sim EDP(\alpha, \mu)$*, for* $A \in \mathcal{B}_X, B \in \mathcal{B}_Y$*,*

$$Var(\mathbf{P}(A \times B)) = \frac{1}{\alpha(\mathcal{X}) + 1} \int_A \frac{P_{0Y|X}(B|x)(1 + \mu(\mathcal{Y}, x)P_{0Y|X}(B|x))}{\mu(\mathcal{Y}, x) + 1} dP_{0X}(x) \tag{2.6}$$

$$+ \frac{\alpha(\mathcal{X})}{\alpha(\mathcal{X}) + 1} \int_A \int_{\{x\}} \frac{P_{0Y|X}(B|x)(1 - P_{0Y|X}(B|x))}{\mu(\mathcal{Y}, x) + 1} dP_{0X}(x') dP_{0X}(x) \tag{2.7}$$

$$- \frac{1}{\alpha(\mathcal{X}) + 1} \int_A \int_{\{x\}} P_{0Y|X}(B|x)^2 dP_{0X}(x') dP_{0X}(x) \tag{2.8}$$

$$- \frac{1}{\alpha(\mathcal{X}) + 1} \int_A \int_{A \setminus \{x\}} P_{0Y|X}(B|x') P_{0Y|X}(B|x) dP_{0X}(x') dP_{0X}(x). \tag{2.9}$$

Theorem 2.4.7 *If* $\mathbf{P} \sim EDP(\alpha, \mu)$*, then, for* $P^* \in \mathcal{P}(\mathcal{B})$*, the posterior distribution,* Q_n*, of* $\mathbf{P}$ *converges weakly to* δ_{P^*} *for* $n \to \infty$*, a.s.* $P^{*\infty}$*.*

The proofs are given in the Appendix.

2.4.4 Square-Breaking construction

The following square-breaking representation of the EDP is a direct result of stick-breaking representation of the DP (Sethuraman, 1994).

Proposition 2.4.8 *If* $\mathbf{P} \sim EDP(\alpha, \mu)$*, it has the following square-breaking a.s. representation:*

$$\mathbf{P} = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \pi_i^X \pi_{j|i}^Y \delta_{X_i^*, Y_{j|i}^*},$$

where: $\pi_1^X = V_1^X$; $\pi_i^X = V_i^X \prod_{h=1}^{i-1}(1 - V_h^X)$, *with*

$$V_i^X \overset{iid}{\sim} beta(1, \alpha(\mathcal{X})), \quad X_i^* \overset{iid}{\sim} P_{0X},$$

and for $i = 1, 2, \ldots :$ $\pi_{1|i}^Y = V_{1|i}^Y$; $\pi_{j|i}^Y = V_{j|i}^Y \prod_{h=1}^{j-1}(1 - V_{h|i}^Y)$, *with*

$$V_{j|i}^Y | X_i^* = x_i^* \overset{iid}{\sim} beta(1, \mu(\mathcal{Y}, x_i^*)), \quad Y_{j|i}^* | X_i^* = x_i^* \overset{iid}{\sim} P_{0Y|X}(\cdot | x_i^*),$$

and the sequences $\left\{V_i^X\right\}_{i=1}^{\infty}$, $\left\{X_i^*\right\}_{i=1}^{\infty}$, $\left\{V_{j|1}^Y | X_1^* = x_1^*\right\}_{j=1}^{\infty}$, $\left\{V_{j|2}^Y | X_2^* = x_2^*\right\}_{j=1}^{\infty}$, .. *and* $\left\{Y_{j|1}^* | X_1^* = x_1^*\right\}_{j=1}^{\infty}, \left\{Y_{j|2}^* | X_2^* = x_2^*\right\}_{j=1}^{\infty}, \ldots$ *are independent.*

For an alternative view of this proposition, consider a square of area one; we break of rectangles off the square defined by a width of π_i^X and length of $\pi_{j|i}^Y$ and we assign the area of that rectangle, $\pi_i^X \pi_{j|i}^Y$, to a random location $(X_i^*, Y_{j|i}^*)$.

Note that while a closed form for the finite dimensional distributions of $\mathbf{P}_Y$ may not be available, we can obtain a square-breaking construction for the random marginal probability measure on $(\mathcal{Y}, \mathcal{B}_Y)$,

$$\mathbf{P}_Y = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \pi_i^X \pi_{j|i}^Y \delta_{Y_{j|i}^*},$$

where the distribution of $\left\{\pi_i^X\right\}_{i=1}^{\infty}$, $\left\{\pi_{j|i}^Y\right\}_{i,j=1}^{\infty}$, $\left\{Y_{j|i}^*\right\}_{i,j=1}^{\infty}$ is specified above.

2.4.5 Clustering structure

The clustering structure in a sample from $\mathbf{P} \sim EDP$ is characterized by the predictive rule. In particular, the predictive rule states that if P_0

is non-atomic, for $A \in \mathcal{B}_X, B \in \mathcal{B}_Y$:

$$Pr(X_{n+1} \in A, Y_{n+1} \in B | x_1, y_1, ..., x_n, y_n)$$
$$= \frac{\alpha(\mathcal{X})}{\alpha(\mathcal{X}) + n} P_0(A \times B) + \sum_{x_i^* \in A} \frac{n_i}{\alpha(\mathcal{X}) + n} \left(\frac{\mu(B, x_i^*) + \sum_{j=1}^{n_i} \delta_{y_{ij}}(B)}{\mu(\mathcal{Y}, x_i^*) + n_i} \right).$$

Thus, the pair (X_{n+1}, Y_{n+1}) is either a "new-new" , "old-new" , or "old-old" pair with probabilities obtained by replacing the set $A \times B$ with the sets $(\mathcal{X} \setminus \{x_1, ..., x_n\}) \times (\mathcal{Y} \setminus \{y_1, ..., y_n\})$, $\{x_1, ..., x_n\} \times (\mathcal{Y} \setminus \{y_1, ..., y_n\})$, or $\{x_1, ..., x_n\} \times \{y_1, ..., y_n\}$ respectively. Let $(x_1^*, ..., x_{d_n}^*)$ be the unique values of $(x_1, ..., x_n)$ where d_n is the number of unique values and $(y_{i,1}^*, ..., y_{i,d_{n_i}}^*)$ be the unique values of $(y_{i,1}, ..., y_{i,n_i})$ where d_{n_i} is the number of unique values in this set. Succinctly, the clustering structure is described as follows:

$$X_{n+1}, Y_{n+1} = \begin{cases} \text{new-new}, (X_{n+1}, Y_{n+1}) \sim P_0 & \text{wp } \frac{\alpha(\mathcal{X})}{\alpha(\mathcal{X})+n}; \\ \text{old-new}, x_i^*, i = 1, ..., d_n, Y_{n+1} \sim P_{0Y|X}(\cdot | x_i^*) & \text{wp } \frac{n_i}{\alpha(\mathcal{X})+n} \frac{\mu(\mathcal{Y}, x_i^*)}{\mu(\mathcal{Y}, x_i^*)+n_i}; \\ \text{old-old}, x_i^*, i = 1, ..., d_n, y_{i,j}^*, j = 1, ..., d_{n_i} & \text{wp } \frac{n_i}{\alpha(\mathcal{X})+n} \frac{n_{i,j}}{\mu(\mathcal{Y}, x_i^*)+n_i}. \end{cases}$$

This gives a "two-level" clustering which reduces to the global clustering of the DP if $\mu(\mathcal{Y}, x) = 0$ for all $x \in \mathcal{X}$.

2.4.6 Comparison with different approaches

In recent literature, there have been many proposals of generalizations of the Dirichlet process, particularly, dependent Dirichlet Processes. Such an approach exploits marginal conditional independence. One considers a collection of random variables $\{Y_j, j \in \mathcal{J}\}$ and assumes that they are conditionally independent, that is, for any $j_1, \ldots, j_m \in \mathcal{J}$, $Y_{j_1}, \ldots, Y_{j_m} \mid F_{j_1}, \ldots, F_{j_m} \sim \prod_{i=1}^{m} F_{j_i}(\cdot)$. Then, a prior is given on the family of random distributions $\{\mathbf{F}_j, j \in \mathcal{J}\}$, such that the $\mathbf{F}_j$'s are dependent.

A first proposal along these lines was given by Cifarelli and Regazzini (1978), who assumed that $\mathbf{F}_j \mid \lambda \overset{iid}{\sim} DP(\alpha F_0(\cdot; \lambda))$, with $\lambda \sim H(\lambda)$. A very interesting development is the Hierarchical Dirichlet Process pro-

posed by (Teh *et al.*, 2006), who model the random base measure $\mathbf{F}_0$ nonparametrically, assuming $\mathbf{F}_0 \sim DP(\gamma H)$. A further development is the Nested Dirichlet Process (Rodriguez *et al.*, 2006) where the model is given as $\mathbf{F}_j \mid G \overset{iid}{\sim} G$ and $G \sim DP(\alpha DP(\gamma H))$. The general and clever scheme given by the Dependent Dirichlet Processes (DDP; MacEachern, 1999), induces dependence across the $\mathbf{F}_j$'s by exploiting the stick breaking representation of the Dirichlet process and by assuming dependent weights and atoms along j.

Dependent DPs define the law of a collection of marginals $\{\mathbf{F}_j, j \in \mathcal{J}\}$ indexed by a non random covariate. If we simply replace $\mathcal{J}$ with $\mathcal{X}$, this does not necessarily define the law of the conditionals. In particular, since the covariate is non random, no σ-algebra on $\mathcal{X}$ is considered, and thus, measurability with respect to $\mathcal{B}_X$ a.s. is not required. If measurability with respect $\mathcal{B}_X$ a.s. is satisfied, this is a model on the random conditionals and does not induce a prior on the random joint distribution of (X, Y).

Instead, our approach gives a prior on the marginal-conditional pair and induces a prior on the joint. For a Dirichlet Process with non atomic base measure, the random conditionals are independent and degenerate a.s. We are extending this by allowing for non degenerate conditionals, but we will assume independence. A further extension would allow for dependence among the random conditionals through a dependent Dirichlet Process if measurability with respect to $\mathcal{B}_X$ a.s. is satisfied. However, some properties will be lost. For example, for a DDP, we would lose conjugacy, and the model would become much more complex, and using the Hierarchical DP or the Nested DP would remove dependence on x in the base measures for the conditionals.

Notice that the distribution of the conditional also as a random function of X is $\mathbf{P}_{Y|X}(\cdot|X) \sim \sum_{i=1}^{\infty} \pi_i^X \delta_{P^*_{Y|X}(\cdot|X_i^*)}$. This resembles the prior for the Nested Dirichlet Process, but is not directly comparable since $\mathbf{P}_{Y|X}(\cdot|X)$ is a different object then $\{\mathbf{F}_j, j \in \mathcal{J}\}$.

2.5 Example

We provide an illustration of the properties of the EDP prior in an application to mixture models. The problem we consider is comparing different schools based on national test scores. The dataset we analyze contains two different test scores for students in 65 inner-London schools. The first score is based on the London Reading Test (LRT), taken at age 11, and the second is a score derived from the Graduate Certificate of Secondary Education (GCSE) exams in a number of different subjects, taken at age 16. Taking into account earlier LRT scores can give a sense of the "value added" for each school. To answer the question of which schools are most effective, we consider modeling the relationship between LRT and GCSE for all the schools. The data are available at `http:// www.stata-press.com/data/mlmus.html`. School number 48 is dropped from the dataset since only 2 students were observed.

Rabe-Hesketh and Skrondal (2005) (Chapter 4) study the following multilevel parametric model where Y_{ij} and X_{ij} represent the GCSE and LRT score, respectively, for student i in school j:

$$Y_{ij} \mid \beta_{0j}, \beta_{1j}, x_{ij} \overset{indep}{\sim} N(\beta_{0j} + \beta_{1j} x_{ij}, \sigma^2),$$
$$\begin{bmatrix} \beta_{0j} \\ \beta_{1j} \end{bmatrix} \overset{iid}{\sim} N_2 \left(\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \Sigma_\beta \right).$$

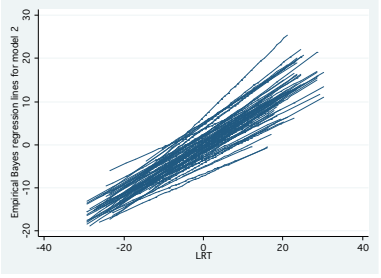
where β_{0j} and β_{1j} are independent of X_{ij}. The interest is in estimating the school specific coefficients $\beta_j = (\beta_{0j}, \beta_{1j})$. The intercept is interpreted as the school mean of GCSE scores for the students with the average LRT score of 0. The competitiveness of the school is captured by the school specific slope. Schools with greater slopes are competitive; more "value" is added for students with higher LRT scores. Schools with a slope of 0 are non-competitive; the performance of students is homogeneous regardless of how the students scored on the LRT. If parents are to choose the best school for their children, both average "value added" and competitiveness

are important.

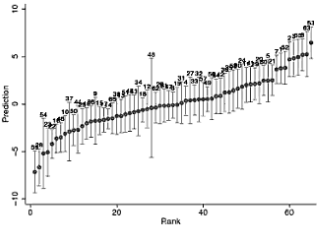
Maximum likelihood estimates of the parameters of the mixing distribution (Rabe-Hesketh and Skrondal, 2005) give $\hat{\beta}_0 = -.115$, with standard error $SE(\hat{\beta}_0) = .0199$, and $\hat{\beta}_1 = .55$, with $SE(\hat{\beta}_1) = .3978$, and estimated covariance matrix:

$$\hat{\Sigma}_\beta = \begin{bmatrix} 9.04 & .18 \\ .18 & .0145 \end{bmatrix}.$$

Empirical Bayes predictions of school specific intercept and slopes were then obtained; Figures 2.1a and 2.1b show the plots of estimated regression lines for each school and ranking of schools based on the intercept.



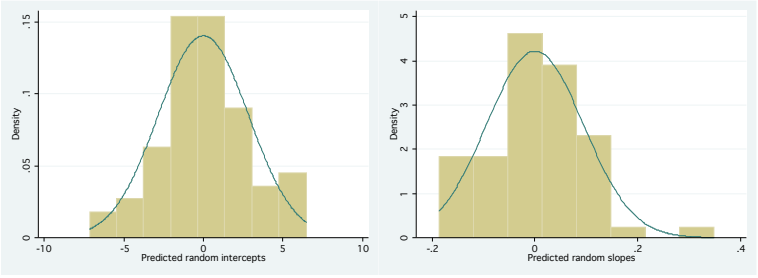
(a) Empirical Bayes Predictions of school-specific regression line



(b) Ranking of Schools

Figure 2.1: Results of Linear Mixed Effect model

By visual inspection of the histograms of the empirical Bayes estimates in Figures 2.2a and 2.2b, for the intercept and especially the slope, a normal distribution does not fit well. This may be due to the fact that there are only 65 schools, that the Gaussian assumption does not hold or a combination of the two. To enlarge the class of models, we can consider modelling the mixing distribution of the intercept and slope nonparametrically. A pitfall of this model is that it assumes the same variability for all the schools. In fact, the wide range of the naive OLS estimates of within school variance (not shown) supports a model which



(a) Predicted random intercept (b) Predicted random slope

Figure 2.2: Assessing the model

allows for school-specific variance.

Bayesian nonparametric extensions of this model would assign a DP prior on the mixing distribution of the (β_{0j}, β_{1j}) (a DP-location mixture), assuming the same variance σ^2 for each school, or model school specific variances σ_j^2, with a DP prior for the latent distribution of $(\beta_{0j}, \beta_{1j}, \sigma_j^2)$ (DP scale-location mixture). The EDP is an intermediate choice. It may model clusters of schools that share the same variance, with different β inside each cluster. We assume that

- $Y_{ij}|x_{ij}, \beta_{0j}, \beta_{1j}, \sigma_j^2 \overset{indep}{\sim} N(\beta_{0j} + \beta_{1j}x_{ij}, \sigma_j^2)$,
- $\beta_j, \sigma_j^2|\mathbf{P}_{\beta,\sigma^2} = P_{\beta,\sigma^2} \overset{iid}{\sim} P_{\beta,\sigma^2}$ where $\beta_j = (\beta_{0j}, \beta_{1j})$,
- $\mathbf{P}_{\beta,\sigma^2} \mid \big(\alpha_{\sigma^2}, \mu_\beta(\sigma^2)\big) \sim EDP(\alpha, \mu)$ where $\alpha = \alpha_{\sigma^2} P_{0,\sigma^2}$ and, for all $\sigma^2 \in \mathbb{R}_+$, $\mu(\cdot, \sigma^2) = \mu_\beta(\sigma^2) P_{0,\beta|\sigma^2}(\cdot|\sigma^2)$.

In the analysis reported below, we fixed the baseline measures $P_{0\sigma}$ as an Inverse-Gamma, with rate and shape parameters, respectively, 8 and 385, and $P_{0,\beta|\sigma^2}(\cdot|\sigma^2)$ as a bivariate Normal, $N_2(\mu_0, k_0\,\sigma^2\,\Sigma_0)$, with $\mu_0 = [0, .5]'$, $k_0 = 1/20$ and $\Sigma_0 = \begin{bmatrix} 9 & 3/16 \\ 3/16 & 1/64 \end{bmatrix}$. In particular, the parameters μ_0 and Σ_0 are set up equals to the values that round the

ones obtained using the Gaussian model. The parameters k_0 and σ^2 are chosen so that their product gives an a priori average of 2. Notice that if the precision parameter $\alpha_{\sigma^2} \approx 0$, we get back to a DP location mixture, and as the precision parameters $\mu_\beta(\sigma^2) \approx 0$ for all $\sigma^2 \in \mathbb{R}_+$, we get a DP scale-location mixture. Thus, with an EDP prior we can express uncertainty between homoskedasticity and heteroskedasticity.

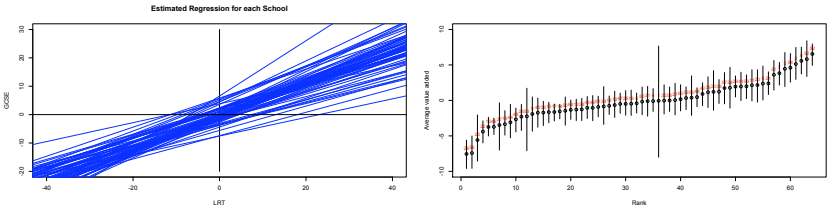
We assume that on average both their a priori expected values are equal to 2 and we model uncertainty about α_{σ^2} and $\mu_\beta(\sigma^2)$ through Gamma hyper-priors:
$\alpha_{\sigma^2} \sim Ga(u_\alpha, v_\alpha)$, where we choose $u_\alpha = 2$ and $v_\alpha = 1$, and for all $\sigma^2 \in \mathbb{R}_+$ $\mu_\beta(\sigma^2) \overset{iid}{\sim} Ga(u_{\mu_\beta}, v_{\mu_\beta})$, with $u_{\mu_\beta} = 2$ and $v_{\mu_\beta} = 1$.

The MCMC scheme to compute posterior distributions is based on the algorithm 6 described in Neal (2000)[1], which is a Metropolis-Hastings algorithm with candidates drawn from the prior. Resampling the precision parameters is done by introducing a latent beta-distributed variable, as described in Escobar and West (1995). The number of iterations is set up to 20,000 with 10% of burn-in. Looking at the trace and autocorrelation plots, convergence appears reached for the β in all the schools and for σ^2 in most schools. The results are summarized in Figures 2.3, which display the estimated regression line for each school and the ranking of schools based on average " value added " with empirical quantiles.

The MCMC posterior expectation of α_{σ^2} is 2.5, and Figures 2.4 and 2.5 depicts the estimated posterior values of $\mu_\beta(\sigma^2)$ for different values of σ^2, without and with logarithmic scale for the $\mu_\beta(\sigma^2)$ respectively.

Neither $\alpha_{\sigma^2} \approx 0$ nor $\mu_\beta(\sigma^2) \approx 0$ for all σ^2, and interestingly, the estimated values of $\mu_\beta(\sigma^2)$ are high for values of σ^2 which are more likely a posteriori, and close to zero for unlikely values of σ^2. Thus, the results favor a model which allows for homoskedacity among some schools with a more likely value σ^2 and some outlying schools with abnormally large

[1]We have implemented the algorithm in R, using the library coda for diagnostic checks.



(a) Estimated regression line for each school

(b) Ranking of Schools based on average value added with empirical quantile

Figure 2.3: Results of EDP model

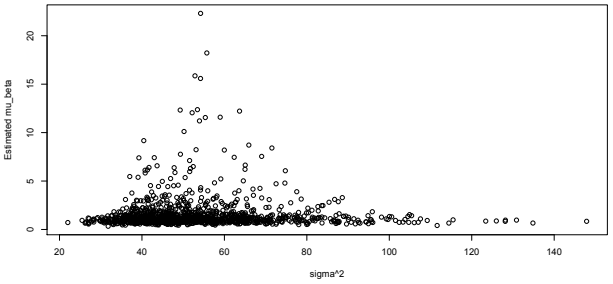


Figure 2.4: Estimated posterior values of $\mu_\beta(\sigma^2)$ for different values of σ^2

or small variances.

2.6 Final remarks

We have proposed an *enrichment* of the DP starting from the idea of enriched conjugate priors. The advantages of this process are that it allows for more flexible specification of prior information, includes the DP as a special case, and retains some desirable properties including conjugacy and the fact that it can be constructed from an enriched urn scheme. The disadvantages include the difficulty in obtaining a closed

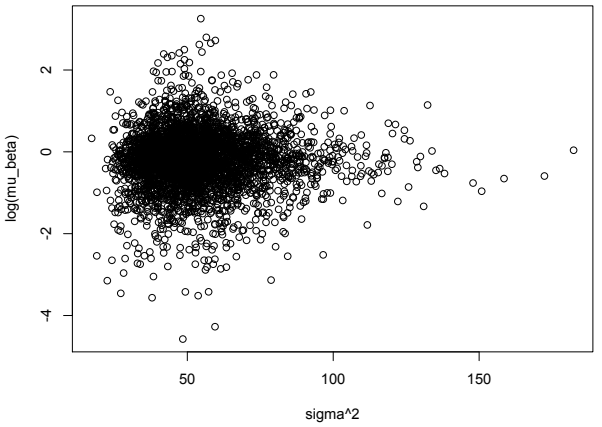


Figure 2.5: Estimated posterior values of $\mu_\beta(\sigma^2)$ for different values of σ^2 in logarithmic scale

form for the distribution of joint probability over a given set and for the distribution of the marginal probability over a measurable subset of $\mathcal{Y}$. There is a clear asymmetry introduced, and for a multivariate random probability measure, how to determine the partition of the random vector into two groups and the ordering depends on the application. There may be a natural ordering or partition and/or computational reasons, including decomposition of the base measure, for choosing the partition and ordering. In our example, we partitioned the random vector $(\beta_0, \beta_1, \sigma^2)$ into the two groups (σ^2) and (β_0, β_1) with σ^2 chosen first due to uncertainty in homoskedasticity and decomposition of the conjugate normal-Inverse-Gamma base measure. One may also examine all the plausible and interesting partitions and orderings.

We have focused on the partition of the random vector into two groups, but most results could be extended to any finite partition of the random vector, although this would of course this would imply a

further nested structure. Other future work includes examining the implied clustering structure in regression settings when the joint model is an EDP mixture and exploring if other conjugate nonparametric priors whose finite dimensionals are standard conjugate priors can be generalized starting from enriched conjugate priors, such as extension of the enriched distribution, mentioned in the Remark 2, to an enriched bivariate Neutral to the Right Processes.

We hope that having explored these features can shed light on potentialities and limitations and encourage further developments in constructing more flexible priors for a random probability measure on $\mathbb{R}^k$.

2.7 Appendix

Proof of Theorem 2.3.1 From the predictive distribution, it follows that the joint distribution can be expressed as:

$$Pr(X_1 = i_1, Y_1 = j_1, ..., X_n = i_n, Y_n = j_n) = \prod_{l=1}^{n} \frac{\alpha(i_l) + \sum_{h=1}^{l-1} \delta_{i_h}(i_l)}{\alpha(\mathcal{X}) + l - 1} \frac{\mu(j_l, i_l) + \sum_{h=1}^{l-1} \delta_{j_h, i_h}(j_l, i_l)}{\mu(\mathcal{Y}, i_l) + \sum_{h=1}^{l-1} \delta_{i_h}(i_l)},$$

which can be equivalently expressed as:

$$\frac{\Gamma(\alpha(\mathcal{X}))}{\prod_{i=1}^{k} \Gamma(\alpha(i))} \frac{\prod_{i=1}^{k} \Gamma(\alpha(i) + n_{i+})}{\Gamma(\alpha(\mathcal{X}) + n)} \prod_{i=1}^{k} \frac{\Gamma(\mu(\mathcal{Y}, i))}{\prod_{j=1}^{m} \Gamma(\mu(j, i))} \prod_{i=1}^{k} \frac{\prod_{j=1}^{m} \Gamma(\mu(j, i) + n_{ij})}{\Gamma(\mu(\mathcal{Y}, i) + n_{i+})}. \tag{2.10}$$

The joint distribution only depends on the number of unique pairs seen, not on the order in which they are observed. Thus, the pairs $\{X_n, Y_n\}_{n \in \mathbb{N}}$ form an exchangeable sequence. By de Finetti's Representation Theorem, there exists a probability measure $\tilde{Q}$ on the simplex

$$S_{k,m} = \left\{ p_{1,1}, ..., p_{k,m} : p_{i,j} \geq 0 \text{ and } \sum_{i=1}^{k} \sum_{j=1}^{m} p_{i,j} = 1 \right\} \text{ such that:}$$

$$Pr(X_1 = i_1, Y_1 = j_1, ..., X_n = i_n, Y_n = j_n) = \int_{[0,1]^{km}} \prod_{i=1}^{k} \prod_{j=1}^{m} p_{i,j}^{n_{i,j}} \tilde{Q}(dp_{1,1}, ..., dp_{k,m}).$$

Define the simplexes $S_k = \left\{ p_{1+}, ..., p_{k+} : p_{i+} \geq 0 \text{ and } \sum_{i=1}^{k} p_{i+} = 1 \right\}$ and $S_m^{(i)} = \left\{ p_{i|1}, ..., p_{i|k} : p_{j|i} \geq 0 \text{ and } \sum_{j=1}^{m} p_{j|i} = 1 \right\}$ for $i = 1, ...k$. Let Q be the probability measure on the product of the simplexes $S_k \times \prod_{i=1}^{k} S_m^{(i)}$ obtained from $\tilde{Q}$ via a re-parametrization in terms of $(\mathbf{p}_{1+}, ..., \mathbf{p}_{k+}, \mathbf{p}_{1|1}, ..., \mathbf{p}_{m|k})$. Then,

$$Pr(X_1 = i_1, Y_1 = j_1, ..., X_n = i_n, Y_n = j_n) = \int_{[0,1]^k \times [0,1]^{km}} \prod_{i=1}^{k} p_{i+}^{n_{i+}} \prod_{j=1}^{m} p_{j|i}^{n_{ij}} Q(dp_{1+}, ..., dp_{m|k}). \tag{2.11}$$

Since the Dirichlet distribution is determined by its moments, combining equations (2.11) and (2.10) implies that

$$\begin{aligned} &\mathbf{p}_{1+}, ..., \mathbf{p}_{k+} \sim \text{Dir}(\alpha(1), ..., \alpha(k)), \\ &\mathbf{p}_{1|i}, ..., \mathbf{p}_{m|i} \sim \text{Dir}(\mu(1,i), ..., \mu(m,i)) \quad i = 1, ..., k, \end{aligned}$$

where $(\mathbf{p}_{1+}, ..., \mathbf{p}_{k+})$, $(\mathbf{p}_{1|1}, ..., \mathbf{p}_{m|1})$,..., and $(\mathbf{p}_{1|k}, ..., \mathbf{p}_{m|k})$ are independent.

The second part of the theorem follows from results of de Finetti on the asymptotic behavior of the predictive distributions for exchangeable sequences; see Cifarelli and Regazzini (1996). ∎

Proof of Theorem 2.4.1.

We start by noting that the sequence $\{X_n\}_{n \in \mathbb{N}}$ is a Pólya sequence with parameter α. Recall that the predictive distribution of a Pólya

sequence converges to a discrete random probability measure with positive mass at the countable number of unique values of the sequence almost surely with respect to the exchangeable law. Therefore, given $X_1 = x_1, ..., X_n = x_n$ and letting $U(x_1, ..., x_n)$ denote the set of the unique values of $\{x_1, ..., x_n\}$, we have that for $x^* \in U(x_1, ..., x_n)$, $n_{x^*} = \sum_{i=1}^{n} \delta_{x^*}(x_i) \to \infty$ as $n \to \infty$ almost surely with respect to the exchangeable law. This implies that given $\{X_n = x_n\}_{n\in\mathbb{N}}$, for any $x^* \in U(\{x_n\}_{n\in\mathbb{N}})$, the set of random variables, $\{Y_{x^*,j}\} = \{Y_i : X_i = x^*, i \in \mathbb{N} | \{X_n = x_n\}_{n\in\mathbb{N}}\}$ is a countable sequence. Furthermore, by assumption, for $x_1^* \neq x_2^* \in U(\{x_n\}_{n\in\mathbb{N}})$, the sequences $\{Y_{x_1^*,j}\}_{j\in\mathbb{N}}$ and $\{Y_{x_2^*,j}\}_{j\in\mathbb{N}}$ are independent Pólya sequences with parameters $\mu(\cdot, x_1^*)$ and $\mu(\cdot, x_2^*)$ respectively. These observations imply exchangeability of the sequence $\{X_n, Y_n\}_{n\in\mathcal{N}}$, as shown in the following argument.

$$Pr(X_1 \in A_1, Y_1 \in B_1, ., X_n \in A_n, Y_n \in B_n) \tag{2.12}$$
$$= \int_{A_1\times...\times A_n} Pr(Y_1 \in B_1, ., Y_n \in B_n | x_1, ., x_n) dPr(x_1, ., x_n).$$

By independence of $\{Y_{x_1^*,j}\}_{j=1}^{n_{x_1^*}}$ and $\{Y_{x_2^*,j}\}_{j=1}^{n_{x_2^*}}$ for $x_1^* \neq x_2^* \in U(x_1, ..., x_n)$, we have that (2.12) is equal to:

$$\int_{A_1\times...\times A_n} \prod_{x^*\in U(x_1,...,x_n)} Pr(Y_{x^*,1} \in B_{x^*,1}, .., Y_{x^*,n_{x^*}} \in B_{x^*,n_{x^*}}) dPr(x_1, ., x_n). \tag{2.13}$$

A permutation, π, of the sets $(x_1 \times B_1), ..., (x_n \times B_n)$, is equivalent to the same permutation, π, of $(x_1, ..., x_n)$ and for $x^* \in U(x_1, .., x_n)$, a permutation, γ_{x^*}, of $(B_{x^*,1}, ..., B_{x^*,n_{x^*}})$. The term inside the integral is invariant to the permutation, π, of $(x_1, ..., x_n)$, and due to exchangeability of Pólya sequences, the laws of the random vectors $\{X_i\}_{i=1}^{n}$ and $\{Y_{x^*,j}\}_{j=1}^{n_x^*}$ are invariant to the permutations π and γ_{x^*} respectively. Thus, (2.13) is equal

to:

$$\int_{A_{\pi(1)}\times .\times A_{\pi(n)}} \prod_{x^*\in U(x_{\pi(1)},.,x_{\pi(n)})} Pr(Y_{x,1}\in B_{\gamma_{x^*}(1)},..,Y_{x,n_x}\in B_{\gamma_{x^*}(n_{x^*})})dPr(x_{\pi(1)},.,x_{\pi(n)})$$
$$= \int_{A_{\pi(1)}\times\cdots\times A_{\pi(n)}} Pr(Y_1\in B_{\pi(1)},.,Y_n\in B_{\pi(n)}|x_{\pi(1)},.,x_{\pi(n)})dPr(x_{\pi(1)},.,x_{\pi(n)}),$$
$$= Pr(X_1\in A_{\pi(1)},Y_1\in B_{\pi(1)},.,X_n\in A_{\pi(n)},Y_n\in B_{\pi(n)}).$$

De Finetti Representation theorem states that there exists a random probability measure, $\mathbf{P}$, with distribution $\tilde{Q}$ on $\mathcal{P}(\mathcal{B})$ such that:

$$Pr(X_1\in A_1,Y_1\in B_1,.,X_n\in A_n,Y_n\in B_n)=\int_{\mathcal{P}(\mathcal{B})}\prod_{h=1}^{n}P(A_h\times B_h)d\tilde{Q}(P), \tag{2.14}$$

and $\frac{1}{n}\sum_{h=1}^{n}\delta_{A\times B}(X_h,Y_h)\to^d \mathbf{P}(A\times B)$ a.s. with respect to the exchangeable law as $n\to\infty$ where $\mathbf{P}\sim\tilde{Q}$. The distribution $\tilde{Q}$ determines the joint distribution, Q, of the marginal and a fixed version of the conditionals. Re-parametrizing in terms of the marginal and conditionals implies:

$$Pr(X_1\in A_1,Y_1\in B_1,.,X_n\in A_n,Y_n\in B_n)$$
$$=\int_{\mathcal{P}(\mathcal{B}_X)\times\mathcal{P}(\mathcal{B}_Y)^{\mathcal{X}}}\prod_{h=1}^{n}\int_{A_h}P_{Y|X}(B_h|x)dP_X(x)dQ(P_X,\prod_{x\in\mathcal{X}}P_{Y|X}(\cdot|x)). \tag{2.15}$$

A simple application of the results of Blackwell and MacQueen (1973) for Pólya urn sequences, verifies that the first two conditions in the definition of the EDP hold. In particular, for any finite partition $A_1,...,A_k\subseteq\mathcal{B}_{\mathcal{X}}$, define the simple measurable function, $\phi(x)=i$ if $x\in A_i$ for $i=1,...,k$. Noting that $\{\phi(X_n)\}_{n\in\mathbb{N}}$, is a Pólya sequence with parameter $\alpha\circ(\phi)^{-1}$

taking values in the finite space $\{1, ..., k\}$, implies:

$$\mathbf{P}_X(\phi^{-1}(1), ..., \mathbf{P}_X(\phi^{-1}(k))) \sim Dir(\alpha(\phi^{-1}(1)), ..., \alpha(\phi^{-1}(k))),$$
$$\Leftrightarrow \mathbf{P}_X(A_1), .., \mathbf{P}_X(A_k) \sim Dir(\alpha(A_1), .., \alpha(A_k)).$$

Similarly, for any finite partition $B_1, .., B_m \subseteq \mathcal{B}_{\mathcal{Y}}$, define the simple measurable function $\varphi(y) = j$ if $y \in B_j$. For any $x^* \in U(\{x_n\}_{n\in\mathbb{N}})$, the sequence $\{\varphi(Y_{x^*,j})\}_{j\in\mathcal{N}}$ is a Pólya sequence taking values in the finite space $\{1, ..., m\}$ with parameter $\mu(\varphi^{-1}(\cdot), x^*)$. Again, it follows that:

$$\mathbf{P}_{Y|X}(\varphi^{-1}(1)|x^*), ..., \mathbf{P}_{Y|X}(\varphi^{-1}(m)|x^*) \sim Dir(\mu(\varphi^{-1}(1), x^*), ..., \mu(\varphi^{-1}(m), x^*))$$
$$\Leftrightarrow \mathbf{P}_{Y|X}(B_1|x^*), ..., \mathbf{P}_{Y|X}(B_m|x^*) \sim Dir(\mu(B_1, x^*), ..., \mu(B_m, x^*)). \tag{2.16}$$

The unique values of the Pólya sequence are actually draws from $P_{0X} = \frac{\alpha}{\alpha(\mathcal{X})}$ and can therefore take any value in $\mathcal{X}$. Thus, (2.16) holds for any $x \in \mathcal{X}$. Finally, we need to show the last two conditions in the definition of the EDP hold. Exchangeability of the pairs implies exchangeability of the sequence $\{Y_i|X_i = x_i\}_{i\in\mathbb{N}}$. Therefore, by de Finetti's theorem:

$$Pr(Y_1 \in B_1, ..., Y_n \in B_n|x_1, ., x_n) \tag{2.17}$$
$$= \int_{\mathcal{P}(\mathcal{B}_Y)^{U(x_1,.x_n)}} \prod_{x^*\in U(x_1,...,x_n)} \prod_{j=1}^{n_{x^*}} P_{Y|X}(B_{x^*,j}|x^*) dQ^{Y|X}_{U(x_1,...x_n)}(\prod_{x^*\in U(x_1,.,x_n)} P_{Y|X}(\cdot|x^*)). \tag{2.18}$$

Independence of the exchangeable sequences $\left\{Y_{x_1^*,j}\right\}_{j\in\mathcal{N}}$ and $\left\{Y_{x_2^*,j}\right\}_{j\in\mathcal{N}}$ for $x_1^* \neq x_2^*$ implies:

$$Pr(Y_1 \in B_1, ., Y_n \in B_n|x_1, x_n) = \prod_{x^*\in U(x_1,...,x_n)} Pr(Y_{x^*,1} \in B_{x^*,1}, ..., Y_{x^*,n_{x^*}} \in B_{x^*,n_{x^*}}),$$
$$= \prod_{x^*\in U(x_1,.,x_n)} \int_{\mathcal{P}(\mathcal{B}_Y)} \prod_{j=1}^{n_{x^*}} P_{Y|X}(B_{x^*,j}|x^*) dQ^{Y|X}_{x^*}(P_{Y|X}(\cdot|x^*)). \tag{2.19}$$

Comparing (2.18) and (2.19) shows that $Q^{Y|X}_{U(x_1,...x_n)} = \prod_{x^*\in U(x_1,...,x_n)} Q^{Y|X}_{x^*}$.

Since the unique values of $\{x_1, ...x_n\}$ are realizations of P_{0X} and can take any value in $\mathcal{X}$, independence of $\left\{\mathbf{P}_{Y|X}(\cdot|x)\right\}_{x\in\mathcal{X}}$ among $x \in \mathcal{X}$ follows. Therefore, (2.17) can be equivalently written as:

$$Pr(Y_1 \in B_1, ., Y_n \in B_n|x_1, ., x_n) = \int\limits_{\mathcal{P}(\mathcal{B}_Y)^{\mathcal{X}}} \prod_{h=1}^{n} P_{Y|X}(B_h|x_h) d\left(\prod_{x\in\mathcal{X}} Q_x^{Y|X}(P_{Y|X}(\cdot|x))\right).$$

Now combining this result with the fact that $\{X_n\}_{n\in\mathbb{N}}$ is an exchangeable sequence implies:

$$\begin{aligned}
&Pr(X_1 \in A_1, Y_1 \in B_1, ., X_n \in A_n, Y_n \in B_n) = \int\limits_{A_1\times\cdot\times A_n} Pr(Y_1 \in B_1, ., Y_n \in B_n|x_1, ., x_n) dPr(x_1, ., x_n),\\
&= \int\limits_{\mathcal{P}(\mathcal{B}_X)} \int\limits_{A_1\times\cdot\times A_n} \int\limits_{\mathcal{P}(\mathcal{B}_Y)^{\mathcal{X}}} \prod_{h=1}^{n} P_{Y|X}(B_h|x_h) d\left(\prod_{x\in\mathcal{X}} Q_x^{Y|X}(P_{Y|X}(\cdot|x))\right) d\left(\prod_{h=1}^{n} P_X(x_h)\right) dQ^X(P_X),\\
&= \int_{\mathcal{P}(\mathcal{B}_X)} \int_{\mathcal{P}(\mathcal{B}_Y)^{\mathcal{X}}} \prod_{h=1}^{n} \int_{A_h} P_{Y|X}(B_h \mid x_h) dP_X(x_h) d\left(\prod_{x\in\mathcal{X}} Q_x^{Y|X}(P_{Y|X}(\cdot \mid x))\right) dQ^X(P_X).
\end{aligned} \tag{2.20}$$

Comparing (2.15) with (2.20) implies that $Q = Q^X \times \prod_{x\in\mathcal{X}} Q_x^{Y|X}$, i.e independence of $\mathbf{P}_X$ and $\left\{\mathbf{P}_{Y|X}(\cdot|x)\right\}_{x\in\mathcal{X}}$. ∎

Proof of Proposition 2.4.3

We show that an Enriched Pólya sequence is equivalent to a Pólya sequence with parameter $\alpha(\mathcal{X})\, P_0(\cdot)$, if $\mu(\mathcal{Y}, x) = \alpha(x)$, $\forall x \in \mathcal{X}$. For an Enriched Pólya sequence with parameters α, μ and for $A \in \mathcal{B}_X, B \in \mathcal{B}_Y$, since $\lim\limits_{\mu(\mathcal{Y},x)\to\alpha(x)} Pr(Y_1 \in B \mid X_1 = x) = P_{0Y|X}(B|x)$, then if $\mu(\mathcal{Y}, x) = \alpha(x)$, $\forall x \in \mathcal{X}$, $Pr(X_1 \in A, Y_1 \in B) = P_0(A \times B)$. The joint predictive distribution is given by,

$$\begin{aligned}
&Pr(X_{n+1} \in A, Y_{n+1} \in B|X_1 = x_1, Y_1 = y_1, ., X_n = x_n, Y_n = y_n)\\
&= \int_A \frac{\mu(B, x) + \sum\limits_{j=1}^{n_x} \delta_{y_{x,j}}(B)}{\mu(\mathcal{Y}, x) + n_x} d\left(\frac{\alpha(x) + \sum\limits_{i=1}^{n} \delta_{x_i}(x)}{\alpha(\mathcal{X}) + n}\right).
\end{aligned} \tag{2.21}$$

Rewriting this as the sum of the integrals over the sets $A \setminus \{x_1, ..., x_n\}$ and $A \cap \{x_1, ..., x_n\}$ and replacing $\mu(\mathcal{Y}, x)$ with $\alpha(x)$, we get (2.21) is equal to,

$$\frac{\alpha(\mathcal{X})}{\alpha(\mathcal{X})+n} P_0(A \setminus \{x_1, ., x_n\} \times B) + \sum_{x \in \{A \cap \{x_1, ., x_n\}\}} \frac{\alpha(x) P_{0Y|X}(B|x) + \sum_{j=1}^{n_x} \delta_{y_{x,j}}(B)}{\alpha(x)+n_x} \frac{\alpha(x)+n_x}{\alpha(\mathcal{X})+n},$$

$$= \frac{\alpha(\mathcal{X})}{\alpha(\mathcal{X})+n} P_0(A \times B) + \frac{n}{\alpha(\mathcal{X})+n} \sum_{i=1}^{n} \frac{\delta_{x_i, y_i}(A, B)}{n}.$$

∎

Proof of Theorem 2.4.4 This proof is based on the proof of Theorem 3.2.4 in Ghosh and Ramamoorthi (2003). To show M_0 is the topological support - the smallest closed set of measure one - it is enough to show that M_0 is a closed set of measure one, such that for every $P^* \in M_0$, $Q(U) > 0$ for any neighborhood U of P^*.

First, we show M_0 is closed. If $P_n \in M_0$, then $P_n(S_0) = 1$ for all n and if $P_n \rightarrow^{weakly} P$, then for any closed set $C \in \mathcal{B}$, $\limsup_n P_n(C) \leq P(C)$. Together these imply $P(S_0) = 1$, or equivalently, $P \in M_0$.

Secondly, the set M_0 has measure one. This follows from the square breaking construction of $\mathbf{P}$. Since $X_i^*, Y_{j|i}^* \sim P_0$ implies $\delta_{X_i^*, Y_{j|i}^*}(S_0) = 1$ almost surely (a.s.), $\sum_{i=1}^{\infty} \pi_i^X = 1$ a.s., and for all i, $\sum_{j=1}^{\infty} \pi_{j|i}^Y = 1$ a.s, then $\mathbf{P}(S_0) = 1$ a.s. ($\Leftrightarrow Q(M_0) = 1$).

Lastly, our theorem will be proved if we show that for any $P^* \in M_0$ and any neighborhood U of P^*, $Q(U) > 0$. By extension of Proposition 2.5.2 in Ghosh and Ramamoorthi (2003), there exists points $q_{1,j} < ... < q_{n_j,j}$ in $\mathbb{R}$ for $j = 1, .., k$, and $\delta > 0$, such that

$$U^* = \{P \in \mathcal{P}(\mathcal{B}) : |P\left(\prod_{j=1}^{k} [q_{i_j,j}, q_{i_j+1,j})\right) - P^*\left(\prod_{j=1}^{k} [q_{i_j,j}, q_{i_j+1,j})\right)| < \delta \text{ and}$$

$$P^*(\partial \prod_{j=1}^{k} [q_{i_j,j}, q_{i_j+1,j})) = 0 \text{ for } i = 1, ..., n_j, j = 1, ..., k\} \subseteq U.$$

Define $A_{i_1,..,i_{k_1}} = \prod_{j=1}^{k_1}[q_{i_j,j}, q_{i_j+1,j})$ and $B_{i_1,..,i_{k_2}} = \prod_{j=k_1+1}^{k_2}[q_{i_j,j}, q_{i_j+1,j})$ and without loss of generality, we denote these sets as $A_1, ..., A_N$ and $B_1, ..., B_M$. If $P_0(A_n \times B_m) = 0$, then $\delta_{X_i^*, Y_{j|i}^*}(S_0) = 0$ almost surely (a.s.) and $\mathbf{P}(A_n \times B_m)$ is degenerate 0. In addition, $P_0(A_n \times B_m) = 0$ combined with the facts that $P^*(\partial A_n \times B_m) = 0$ and $P^*(S_0) = 1$, imply that $P^*(A_n \times B_n) = 0$. Therefore, $|\mathbf{P}(A_n \times B_m) - P^*(A_n \times B_m)| = 0$ a.s. If $P_0(A_n \times B_m) > 0$, then $\delta_{X_i^*, Y_{j|i}^*}(A_n \times B_m) = 1$ with positive probability. Thus, the square breaking construction implies that $Q(U^*) > 0$. ∎

Proof of Lemma 5.3

Proof

$$E[\mathbf{P}(A\times B)^2] = E[\sum_{i=1}^{\infty}\pi_i^2\mathbf{P}_{Y|X}(B|X_i^*)^2\delta_{X_i^*}(A)] \tag{2.22}$$

$$+ E[\sum_{i=1}^{\infty}\sum_{j\neq i}\pi_i\pi_j\mathbf{P}_{Y|X}(B|X_i^*)^2\delta_{X_i^*}(A)\delta_{X_j^*}(\{X_i^*\})] \tag{2.23}$$

$$+ E[\sum_{i=1}^{\infty}\sum_{j\neq i}\pi_i\pi_j\mathbf{P}_{Y|X}(B|X_i^*)\mathbf{P}_{Y|X}(B|X_j^*)\delta_{X_i^*}(A)\delta_{X_j^*}(A\setminus\{X_i^*\})] \tag{2.24}$$

Using the fact that $E_\pi[\sum_{i=1}^{\infty}\pi_i^2] = \frac{1}{\alpha(\mathcal{X})+1}$ and properties of the Dirichlet distribution,

$$\begin{aligned}(2.22) &= E_\pi[\sum_{i=1}^{\infty}\pi_i^2 E_{X^*}[E_{Q^{Y|X}}[\mathbf{P}_{Y|X}(B|X_i^*)^2|X_i^*]\delta_{X_i^*}(A)]]\\ &= \frac{1}{\alpha(\mathcal{X})+1}\int_A \frac{P_{0Y|X}(B|x)(1+\mu(\mathcal{Y},x)P_{0Y|X}(B|x))}{\mu(\mathcal{Y},x)+1}dP_{0X}(x)\end{aligned}$$

Now, using the fact that $E_\pi[\sum_{i=1}^{\infty}\sum_{i\neq j}\pi_i\pi_j] = \frac{\alpha(\mathcal{X})}{\alpha(\mathcal{X})+1}$ and, again, properties of the Dirichlet distribution,

$$\begin{aligned}(2.23) &= E_\pi[\sum_{i=1}^{\infty}\sum_{i\neq j}\pi_i\pi_j E_{X^*}[E_{Q^{Y|X}}[\mathbf{P}_{Y|X}(B|X_i^*)^2|X_i^*]\delta_{X_i^*}(A)\delta_{X_j^*}(\{X_i^*\})]]\\ &= \frac{\alpha(\mathcal{X})}{\alpha(\mathcal{X})+1}\int_A\int_{\{x\}}\frac{P_{0Y|X}(B|x)(1+\mu(\mathcal{Y},x)P_{0Y|X}(B|x))}{\mu(\mathcal{Y},x)+1}dP_{0X}(x')dP_{0X}(x)\\ (2.24) &= E_\pi[\sum_{i=1}^{\infty}\sum_{i\neq j}\pi_i\pi_j E_{X^*}[E_{Q^{Y|X}}[\mathbf{P}_{Y|X}(B|X_i^*)\mathbf{P}_{Y|X}(B|X_j^*)|X_i^*]\delta_{X_i^*}(A)\delta_{X_j^*}(A\setminus\{X_i^*\})]]\\ &= \frac{\alpha(\mathcal{X})}{\alpha(\mathcal{X})+1}\int_A\int_{A\setminus\{x\}}P_{0Y|X}(B|x')P_{0Y|X}(B|x)dP_{0X}(x')dP_{0X}(x)\end{aligned}$$

The result is obtained following some algebra. ∎

Proof of Theorem 2.4.7

First, we show that $E[\mathbf{P}(A\times B)|X_1=x_1,Y_1=y_1,...,X_n=X_n,Y_n=y_n]\to P^*(A\times B)$ a.s. $P^{*\infty}$.

$$E[\mathbf{P}(A\times B)|X_1=x_1,Y_1=y_1,...,X_n=X_n,Y_n=y_n]$$
$$=\frac{\alpha(\mathcal{X})}{\alpha(\mathcal{X})+n}P_0(A\setminus\{x_1,...,x_n\}\times B)+\sum_{x\in A\cap\{x_1,...,x_n\}}\frac{\mu(\mathcal{Y},x)+\sum_{j=1}^{n_x}\delta_{y_{x,j}}(B)}{\alpha(\mathcal{X})+n}\frac{\alpha(x)+n_x}{\mu(\mathcal{Y},x)+n_x},$$
$$\sim\frac{1}{n}\sum_{x\in A\cap\{x_1,...,x_n\}}\sum_{j=1}^{n_x}\delta_{y_{x,j}}(B)=\frac{1}{n}\sum_{i=1}^{n}\delta_{x_i,y_i}(A,B)$$
$$\to P^*(A\times B)\text{ a.s }P^{*\infty}$$

Using lemma (5.3), we show the posterior variance of $\mathbf{P}(A\times B)$ goes to 0, by showing the four terms in (5.3) each go to 0. Since $\frac{\alpha_n(A)}{\alpha_n(\mathcal{X})}\sim\frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}(A)$ and for, $x\in\{x_1,...,x_n\}$, $\frac{\mu_n(B,x)}{\mu_n(\mathcal{Y},x)}\sim\frac{1}{n_x}\sum_{i=1}^{n_x}\delta_{y_{x,j}}(B)$,

$$(2.6)\sim\frac{1}{n}\int_A(\frac{1}{n_x}\sum_{i=1}^{n_x}\delta_{y_{x,j}}(B))(\frac{1}{n_x}+\frac{1}{n_x}\sum_{i=1}^{n_x}\delta_{y_{x,j}}(B))d\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}(x)\right),$$
$$\to 0,$$
$$(2.7)\sim\int_A\int_{\{x\}}\frac{1}{n_x}(\frac{1}{n_x}\sum_{i=1}^{n_x}\delta_{y_{x,j}}(B))(\frac{1}{n_x}\sum_{i=1}^{n_x}\delta_{y_{x,j}}(B^c))d\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}(x')\right)d\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}(x)\right),$$
$$\to 0,$$
$$(2.8)\sim-\frac{1}{n}\int_A\int_{\{x\}}(\frac{1}{n_x}\sum_{i=1}^{n_x}\delta_{y_{x,j}}(B))^2d\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}(x')\right)d\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}(x)\right),$$
$$\to 0,$$
$$(2.9)\sim-\frac{1}{n}\int_A\int_{A\setminus\{x\}}(\frac{1}{n_x}\sum_{i=1}^{n_x}\delta_{y_{x,j}}(B))(\frac{1}{n_{x'}}\sum_{i=1}^{n_{x'}}\delta_{y_{x',j}}(B))d\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}(x')\right)d\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}(x)\right),$$
$$\to 0.$$

This holds for any finite collection of sets. By a straightforward extension of Theorem 2.5.2 of [11], this implies weak convergence of Q_n to δ_{P^*} a.s. $P^{*\infty}$. ∎

Bibliography

[3] Blackwell D., MacQueen J. B. (1973). Ferguson Distributions Via Pólya Urn Schemes. *Annals of Statistics*, **1**, 353-355.

[2] Cifarelli D.M., Regazzini E. (1978). Problemi statistici nonparametrici in condizioni di scambiabilità parziale e impiego di medie associative. *Quaderni Istituto di Matematica Finanziaria*, Turin: Università di Turin, **12**, 1-36.

[3] Cifarelli D.M., Regazzini E. (1996). De Finetti's contribution to probability and statistics. *Statistical Science*, **11**, 253-282.

[4] Connor R.J., Mosimann J. E. (1969). Concepts of Independence for Proportions with a Generalization of the Dirichlet Distribution. *Journal of the American Statistical Association*, **64**, 194-206.

[5] Consonni G., Veronese P. (2001). Conditionally reducible natural exponential families and enriched conjugate priors. *Scandinavian Journal of Statistics*, **28**, 377-406.

[6] Diaconis P., Ylvisaker D. (1979). Conjugate priors for exponential families. *Annals of Statistics*, **7**, 269-281.

[7] Doksum K. A. (1974). Tailfree and Neutral random probabilities and their posterior distributions. *Annals of Probability*, **2**, 183-201.

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

[9] Ferguson T.S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, **1**, 209-230.

[10] Geiger D., Heckerman D. (1997). A characterization of the Dirichlet distribution through global and local parameter independence. *Annals of Statistics*, **25**, 1344-1369.

[11] Ghosh J.K., Ramamoorthi R.V. (eds.) (2003). *Bayesian nonparametrics*. New York: Springer.

[15] MacEachern S.N. (1999). Dependent nonparametric processes. In *ASA Proceedings of the Section on Bayesian Statistical Science*. Alexandria, VA: American Statistical Association, 50-55.

[21] Neal R.M. (2000). Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics*, **9**, 249-265.

[14] Rabe-Hesketh S., Skrondal A. (2005). *Multilevel and longitudinal modeling using Stata* (2nd edition). College Station, Texas: Stata Press.

[15] Ramamoorthi R.V., Sangalli, L. (2006). On a Characterization of Dirichlet Distribution. In *Bayesian Statistics and its Applications* (Upadhyay S.K., Singh U., Dey D.K., eds.) (2006) 385-397. Varanasi: Anshan.

[16] Rodriguez A., Dunson D., Gelfand A. (2006). The nested Dirichlet Process. *Journal of the American Statistical Association*, **103**, 1131-1154

[25] Springer M.D., Thompson W.E. (1970). The distribution of Products of Beta, Gamma and normal Random Variables. *Journal on Applied Mathematics*, **18**, 721-737.

[18] Sethuraman J. (1994). A Constructive Definition of Dirichlet Priors. *Statistica Sinica*, **4**, 639-650.

[19] Teh Y., Jordan M., Beal M., Blei D. (2006). Hierarchical Dirichlet Process. *Journal of the American Statistical Association*, **101**, 1566-1581.

Part II

Balance sheet analysis

Chapter 3

Bayesian semiparametric panel inference for autoregressive leverage

Abstract

In this chapter we propose a Bayesian semiparametric panel model for analyzing the leverage-ratio in the euro area. For the institutional sector of *Monetary and Financial Institutions* and for ten euro area countries, we study the relationship between changes in the leverage-ratio and the growth of the balance-sheet and a possible break in such relationship after the bankruptcy of Lehman Brothers. We propose a Bayesian linear semiparametric approach assigning a Dirichlet Process (DP) prior to the mixing distribution. To study the structural break, we extend the model by assigning an Enriched Dirichlet Process (EDP) prior to the mixing distribution. We assume that the distribution of the post-break country-specific coefficients is centered on the pre-break country-specific coefficients, so that the results are conservative -that is to say that if a change in a distribution is observed, it cannot be due to the chosen prior- and, moreover, allowing to use more information for fitting the distribution of the post-break coefficients, which otherwise would rely on a small number of observations. We find *credit procyclicality* in the leverage-ratio (positive changes in leverage asso-

ciated with a positive balance-sheet growth) for some of the countries. We also find that the response of the leverage-ratio to credit developments has changed for most of the countries after the events in Autumn 2008, generally towards a less procyclical behaviour which suggests a precautionary reaction in those countries.

Introduction

The *leverage-ratio*, here defined as the liabilities-to-assets ratio, measures debt relative to equity in the balance-sheet of the economic agents. Excess leverage has been pointed out as a major contributor to financial crises, several empirical analysis also indicating that average leverage often rises prior to such episodes.

In this chapter, we study the behaviour of the leverage-ratio of euro area banks for the period 1999 to 2012. This encompasses two credit cycles marked respectively by the mild slowdown in 2001-2003 and the severe downturn which started in 2008. We analyze the impact of the rate of growth of the bank balance-sheet (which we identify with the credit cycle) on its leverage-ratio, also looking into possible changes in the relationship after the bankruptcy of Lehman Brothers in September 2008. The work has been inspired by the study of Adrian and Shin (2010) who found strong credit procyclicality in the leverage behavior (a positive relationship between balance-sheet growth and leverage increases) for the US investment banks.

Our contributions are the following:

- We analyze leverage behaviors in the framework of sector accounts. This methodological framework enables a comprehensive and consistent approach to the economic developments on the basis of the interactions between institutional sectors. For leverage analysis, the approach enables looking at the implications of individual sector leverage on the leverage of other sectors and the overall economy. In this work we take particularly advantage of the fact that marked-

to-market valuation of assets, liabilities and equity is required by the system. This allows a more insightful analysis of the behavioral links between leverage and balance-sheet developments. In the literature, there are studies for the procyclical behavior of leverage for US investment banks and US commercial banks (Adrian and Shin, 2010) both using micro and macro data, although the latter not necessarily adhering to the strict market valuation rules of the international standards for sector accounts (System of National Accounts, SNA). At the European level, only micro data of European banks have been analyzed, Baglioni *et al.* (2010; 2012). We perform here a macro analysis for the sector Monetary Financial institutions -MFIs (which covers banks) using the sector accounts for euro area countries, which are fully compliant with the sector accounts standards.

- To better understand the leverage behavior, we investigate separately the effects of asset price changes and of asset transactions. In particular, this distinction enables insulating the mechanical effects of asset price changes on leverage, which would otherwise blur the interplay of leverage with balance-sheet growth.

- We propose a Bayesian nonparametric approach (based on the Dirichlet Process) for estimating posterior distributions for country-specific random coefficients in dynamic panel model data, by extending the parametric approach of Hsiao, Pesaran and Tahmiscioglu (1999) and Hirano (1999; 2002). We also propose a further extension by allowing for variation in the coefficients capturing the relationship between leverage and the credit cycle between before and after the bankruptcy of Lehman Brothers. We do this by incorporating sequential dependence between their distributions (using the Enriched Dirichlet Process).

Our results show that procyclicality in bank leverage is clearly present

before 2008 in four of the countries examined: Belgium, France, Germany and the Netherlands. Relaxing somewhat our evaluation requirements procyclicality would also be present for Italy and Spain. Presence of procyclicality suggests that episodes of strong credit and balance-sheet growth are accompanied by insufficient built-up of precautionary capital buffers, while severe downturns in the credit cycle are linked to fast accumulations of capital. This behavior might contribute to the amplification of the credit cycle itself.

At the same time, no correlation between leverage and balance-sheet dynamics is found for Austria, Greece and Portugal, whose banks would then had behaved as if they had a fixed target leverage-ratio broadly independent of the amount of intermediated funds. Prudent bank behavior is clearly found for Finland which presents a negative coefficient indicating that balance-sheet growth is accompanied by faster capital built-up.

After 2008, most of the countries present a decrease in the cyclicality coefficient suggesting a precautionary reaction to the crisis. This was particularly acute in Portugal where the coefficient becomes negative. However, increases in the coefficient are found in Belgium and Germany suggesting a further deceleration in capital built-up relative to balance-sheet growth than might have contributed to alleviate the severity of the credit downturn in those countries.

The chapter is organized as follows. In Section 3.1, the problem is contextualized in its economic framework and motivated. In Section 3.2, the structure of the data is described and, in Section 3.3, the available alternative statistical approaches for such data are discussed. In Section 3.4, the proposed model is discussed, first without and then with allowing for a structural break in the country-specific coefficients. In Section 3.5, the results obtained by applying the approach to the leverage for the institutional sector of MFI in euro area countries are shown and discussed. A final section concludes this chapter.

3.1 Economic background and motivation

The financial crisis has brought leverage to the center of the economic discussion. Many have argued that excessive leverage is a main contributor to financial crises. The eventual emergence of difficulties to roll over debt that lay at high levels would lead to sizeable asset sales. A positive feedback loop would then be set in motion as asset sales result in asset price falls, reducing the value of the collateral for the remaining debt and forcing debtors to further disposals of assets via margin calls or similar mechanisms. In this process, leverage increases as a result of the asset price reductions (for assets decreases relative to debt, i.e. equity decreases relative to debt), while agents desperately try to prevent such increases by liquidating assets and debt. These efforts are self-defeating via plummeting asset prices. In short, scenarios of asset quality and price deterioration during crisis would prompt intermediaries to reduce leverage exerting a downward pressure on credit that adds to the crisis scenario.

The obvious question is whether leverage tends to reach high levels that might then result in such vulnerabilities and, if so, what the mechanism is that leads to that. Upward trends in leverage have been argued to be associated with credit expansions. Periods of increase in debt would be accompanied with relaxation in leverage objectives, which in turn would fuel credit expansions. To obtain such outcome, investors and borrowers should have a low aversion to the risks associated with leverage which would be seen as more than offset by the higher yield (for investors) or lower cost of finance (for borrowers) that could be derived from increasing leverage. Such effects would be characterized by situations of low interest rates and high equity prices as was the case during the Great Moderation. Moreover, the period previous to the financial turmoil of 2008 which was prone to financial innovation that facilitated

engaging in leverage in new, sophisticated ways that helped circumvent regulatory requirements that could have limited the leverage appetite. At the same time, a positive link between leverage and debt, which we refer to here -following the relevant literature (Adrian and Shin, 2010)- as procyclicality of leverage, does not need to occur. Prudent banking practices and effective supervision would ensure that increases in debt are accompanied by comparable built-up of capital. If so, no correlation -or even negative correlation if the regulatory framework is designed to be anticyclical- would be observed between leverage changes and credit growth. Starting with the pioneer works of Adrian and Shin(2010) for *US investment banks*, the literature studying the *procyclical* behavior of the leverage has grown in interest. Adrian *et al.* (2011) showed that leverage is procyclical also for *US commercial banks*. Baglioni *et al.* (2010; 2012) replicated the analysis of Adrian and Shin (2010), first for a sample of 13 *European major commercial banks* over 1999-2009 (Baglioni *et al.*, 2010), and then for a bigger sample of 77 European banks over the period 2000-2009 (Baglioni *et al.*, 2012). They conclude that there is procyclical leverage for European banks for which investment banking prevails over the traditional commercial banking activity.

The aim of in this chapter is to clarify the nature of the link between credit and leverage in the euro area. In particular, the purpose is to determine whether the notional component of the credit growth rate is correlated with the leverage ratio. In studying the leverage behavior, the role of asset prices is particularly relevant. On the one hand, changes in asset prices affect leverage heavily by changing the value of assets relative to debt. This effect might mask the nature of the link between the leverage and the credit cycle: correlation analysis would tend to find procyclicality just because asset prices are strongly procyclical. Behavioral links, like the one suggested to be at work during the Great Moderation, would therefore not be easily identifiable. On the other hand, asset prices changes interplay with leverage behavior. As described above, financial

crisis often see self-defeating deleveraging processes, where increases in leverage caused by asset price declines are followed by efforts to restore them. Moreover, Adrian and Shin (2010) identified a symmetric mechanism that would link asset prices and credit through leverage in economic upturns. Asset price increases result in reductions in leverage and therefore in the cost of debt financing (as collateral relative to debt increases), encouraging the incurrence in debt and the acquisition of assets. Agents would so act as if defending their leverage-ratio similarly as they do in downturns. Again the reaction is self-defeating as acquisitions of assets would push asset prices up contributing to positive feedback between asset prices and debts similar to those described in downturns. We acknowledge the relevance of asset prices in leverage issues by developing a special treatment for them as explained below, aiming in particular to facilitate the identification of the behavioral links with credit developments. Hence, the leverage-ratio can change not only for lending and borrowing choices but also for differences in price dynamics of liabilities and assets. Due to continuous market-(re)assessment of balance-sheets, assets and liabilities values are instantaneously affected by movements in prices, as it does their net worth. We call this mechanism an *automatic reaction* to the market developments. The other change mechanism -the one that excludes the arithmetical effect of asset price changes and reflects behavioral patterns- is here called an *active leverage* and results in increases in financial intermediation activities and the decision on how to finance these new activities. Our proposal is to control the price effects (both of assets and liabilities) and to focus on the active reaction of leverage to the changes in balance-sheet only due to the transactions. This gives us what we call the *cyclicality coefficient* describing the reaction of leverage-ratio to the credit cycle.

3.2 Preliminary data analysis

In this chapter, we focus on the leverage bahavior of the *Monetary and Financial Institutions* (MFIs) sector for a panel of euro area countries using the national sector accounts provided by the ECB Statistical Data Warehouse. The MFIs sector is the aggregate that covers banks in the sector accounts framework. It encompasses credit institutions and money market funds.

The panel under investigation consists of ten euro area countries: Austria (AT), Belgium (BE), Finland (FI), France (FR), Germany (DE), Greece (GR), Italy (IT), Netherlands (NL), Portugal (PT), and Spain (ES). The data are quarterly and balanced[1], consisting of observations for the period between the first quarter of 1999, denoted by $1999q1$, and the first quarter of 2012, denoted by $2012q1$.

Figure 3.1 shows the series of the leverage-ratio for each country. We also show, in Figure 3.2, the dynamic correlation between the notional asset growth rate and the logarithm of leverage-ratio[2] for each country. Dynamic correlation is a measure of dynamic comovement defined in the frequency domain, which allows to distinguish between correlation across long-run and short-run fluctuations (Croux, Forni and Reichlin, 2001). This measure is based on standard spectral analysis techniques, which decompose series into the sum of cyclical components at different frequencies. The x-axis of Figure 3.2 expresses the frequency domain and it can easily be converted into the time domain. For example, the conventional limits of the business cycle frequencies, i.e. 1.5 years and 1.8 years, correspond to the frequencies $\pi/3$ and $\pi/6$ respectively[3]. Bootstrap

[1]The last two quarters of DE and NL are estimated based on their single series.

[2]The choice of showing the plots of the dynamic correlation between the notional asset growth rate and the logarithm of leverage-ratio is because of the need of having stationary variables.

[3]The conversion from frequency- domain to time-domain is done as follows. Since data are quarterly, the period is 4. The frequency, say w, is then $w = 2\pi/T$ where $T = 4 * t$ and t is the time index in the time-domain. For example, if $t = 1.5$ years, then $w = 2\pi/(4 * 1.5) = \pi/3 = 1.0472$. Hence, the highest frequency is associated

confidence bands are also included into the plots. By visual inspection, three groups can be recognized:

- a first group of countries with positive correlation in the long-run: AT, BE, FI;
- a second group of countries with negative correlation in the lung-run: FR, DE, NL, PT;
- and another group with non-significative correlation in the long-run: GR, IT, ES.

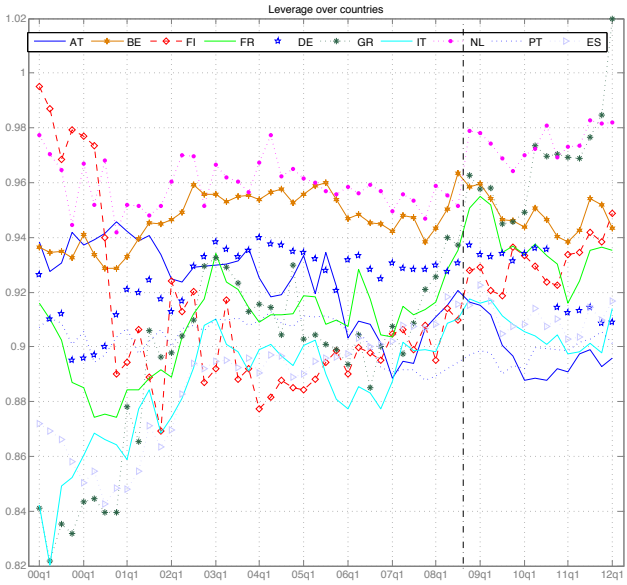


Figure 3.1: Leverage-ratio across countries for 1999q1 - 2012q1

with $t = 0.5$ year (2 quarters).

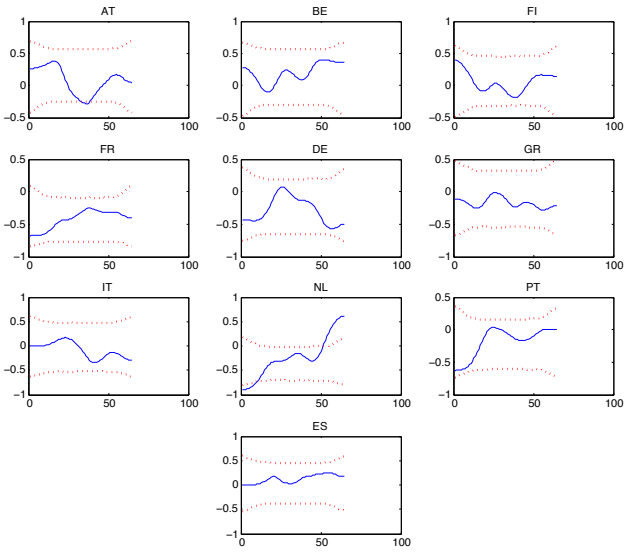


Figure 3.2: Dynamic Correlation between notional assets growth rate and logarithm of leverage (with 95 % bootstrap confidence bands).

3.3 Statistical framework and background

Let i be the label associated with the country i and let t be the discrete time index, with $i, t \leq \infty$. With the data described in Section 3.2, $t \in \{1999q1, 1999q2, ..., 2012q1\}$ and $i \in \{1, ..., 10\}$, where each of these labels is associated with country AT, BE, FI, FR, DE, GR, IT, NL, PT, ES respectively.

We are interested in isolating the impact of changes in the balance sheet, not imputable to revaluations, on the changes in the leverage-ratio. Let us therefore consider the following model for the leverage-ratio data:

$$\overset{\bullet}{\boldsymbol{y}}{}^{i}_{t-1,t} = \boldsymbol{\beta}_{0;i} + \boldsymbol{\beta}_{1;i} \log\left(\boldsymbol{y}^{i}_{t-1}\right) + \boldsymbol{\beta}_{2;i} \overset{\bullet}{x}{}^{i}_{t-1,t} + \boldsymbol{\beta}_{3}\, \overset{\bullet}{a}{}^{i}_{t-1,t} + \boldsymbol{\beta}_{4}\, \overset{\bullet}{l}{}^{i}_{t-1,t} + \boldsymbol{e}_{i,t} \tag{3.1}$$

where:

- $\overset{\bullet}{\boldsymbol{y}}{}^{i}_{t-1,t}$ indicates the leverage-ratio growth rate for the country i, i.e. the logarithmic difference between the leverage-ratio in quarter t, denoted by $\boldsymbol{y}^{i}_{t}$, and $\boldsymbol{y}^{i}_{t-1}$ for country i.
- $\left(\overset{\bullet}{x}{}^{i}_{t-1,t}, \overset{\bullet}{a}{}^{i}_{t-1,t}, \overset{\bullet}{l}{}^{i}_{t-1,t}\right)$ is the set of regressors, where
  - $\overset{\bullet}{x}{}^{i}_{t-1,t}$ is the growth rate of the notional assets between quarter $t-1$ and quarter t for country i;
  - $\overset{\bullet}{a}{}^{i}_{t-1,t}$ is the assets price index between quarter $t-1$ and quarter t for country i;
  - $\overset{\bullet}{l}{}^{i}_{t-1,t}$ is the liabilities price index between quarter $t-1$ and quarter t for country i.
- $\boldsymbol{e}_{i,t} \mid \boldsymbol{\sigma}^{2}_{i} \overset{indep}{\sim} N\left(0, \boldsymbol{\sigma}^{2}_{i}\right)$, i.e. is independent and non identically distributed over i where $\boldsymbol{\sigma}^{2}_{i}$ is the country-specific observational variance.

The above model has therefore cross-country heterogeneity for the coefficient reflecting cyclicality, $\boldsymbol{\beta}_{2;i}$, called *cyclicality coefficient*, the long-run leverage-ratio, $\boldsymbol{\beta}_{0;i}$, and the speed of adjustment to the long-run target leverage-ratio, $\boldsymbol{\beta}_{1;i}$. Let us call $\underline{\boldsymbol{\beta}}_{i}$ the vector which collects the country-specific coefficients, namely $\underline{\boldsymbol{\beta}}_{i} = \left(\boldsymbol{\beta}_{0;i}, \boldsymbol{\beta}_{1;i}, \boldsymbol{\beta}_{2;i}\right)$. The primary objective of this chapter is on making inference on all the $\boldsymbol{\beta}_{2;i}$. The model (3.1) is often called a *latent vector model* because of the latent vector $\underline{\boldsymbol{\beta}}_{i}$. The model must then be completed by specifying a model for $\underline{\boldsymbol{\beta}}_{i}$.

Heterogeneity across countries, here expressed through the $\underline{\boldsymbol{\beta}}_i$ and the $\boldsymbol{\sigma}_i^2$, is the central focus of our analysis and how to model them plays a crucial role. Most of the frequentist approaches available for the dynamic panel data with heterogeneity, e.g. GMM estimation, are mainly based on first differences to eliminate country-specific random intercepts (Arellano and Bond, 1991), without neither eliminating the other country-specific coefficients nor estimating them. The Bayesian approaches provide the general framework for a complete analysis also for panel autoregressive models (Hsiao *et al.*, 1999; Hsiao and Pesaran, 2004; Hirano, 1999; 2002; Koop, 2003; Zhang, 2006). Within a Bayesian framework, the main assumption is then that $\underline{\boldsymbol{\beta}}_i$ is independently distributed across i, meaning that one could express $\underline{\boldsymbol{\beta}}_i = \underline{\boldsymbol{\beta}} + \underline{\boldsymbol{\nu}}_i$ where $\underline{\boldsymbol{\nu}}_i$ is independently distributed over i. Within a dynamic panel setting, like in ours, it is not possible to assume that:

$$E\left(\boldsymbol{\nu}_{j;i}\boldsymbol{y}_{t-1}^i\right) = 0 \qquad \text{for } j = 1, 2, 3, \tag{3.2}$$

where $\boldsymbol{\nu}_{j;i}$ is the jth element of $\underline{\boldsymbol{\nu}}_i$. Expression (3.2) is the basic assumption required for applying standard approaches. It is here violated because the model is characterized by two sources of persistence over time: autocorrelation due to the presence of the lagged dependent variable across the regressors and country-specific coefficients characterizing the heterogeneity across the countries (Baldagi, 2000). Combined together, they imply that the regressors set and the random coefficients are not independent. Therefore, the fixed effects estimators not only would require the estimation of too many parameters compared with the number of observations, which violates the principle of parsimony, but also would not be consistent anymore. Pesaran and Smith (1995) have derived the asymptotic bias of the conventional fixed effects (within) estimator. Under the assumption that all the y_0^i are known and $\boldsymbol{\nu}_i$ and $\boldsymbol{e}_{i,t}$ are independent normally distributed, the Bayesian approach is an appealing solution (Hsiao and Pesaran, 2004, Hsiao *et al.*, 1999). Hsiao *et al.* (1999) compared the performance in finite samples of estimators that

attempt to correct for the finite T bias of country-specific estimates and Bayesian type estimators. They concluded suggesting that the Bayesian approach seems to perform reasonably well and that the mean group estimator is asymptotically Gaussian. Over the past five years, the estimation of cross sectionally dependent panels has intensive been studied and robust estimation procedures have been advanced also outside the Bayesian framework, e.g. the *Common Correlated Effects* (CCE) estimator (Pesaran and Tosetti, 2011) discussed in **Chapter 1**. Our focus will be here on Bayesian approaches and, in particular, to models that will relax the usual rigid parametric assumptions of the traditional Bayesian panel models.

3.3.1 Bayesian parametric approach

The Bayesian framework requires to specify a *hierarchical model*, which can be defined by combing (at least) the model (3.1) (first level of the hierarchy), a model for the $\underline{\boldsymbol{\beta}}_i$ and the $\boldsymbol{\sigma}_i^2$ (second level of the hierarchy), and models for all the unknown parameters (third level). For the sake of simplicity, we will focus on assigning a model only on the parameters $\underline{\boldsymbol{\beta}}_i$, whereas we will model the $\boldsymbol{\sigma}_i^2$ independently at the third level of the hierarchy without any structure among themselves. This choice is preferred because working with variances is not easy and finding a good model that links together all the variances is definitively an hard task. Moreover, any possible model for the variances could remarkably affect the results of the analysis. We therefore prefer to assume that the variances are all independent each other with uninformative priors on each of them. This prior specification for the standard deviation follows the Lindley and Smith (1972) approach.

The commonly-used hierarchical model is a Bayesian parametric model, which can be summarized as follows:

$$
\begin{aligned}
&I.\ \overset{\bullet}{\boldsymbol{y}}{}^{i}_{t-1,t} \mid \underline{\boldsymbol{\beta}}_{i,3:4}, \boldsymbol{\sigma}_i^2, \boldsymbol{y}^i_{t-1} \overset{indep}{\sim} N\left(m_{i,t}, \boldsymbol{\sigma}_i^2\right) \\
&\text{where } m_{i,t} = \boldsymbol{\beta}_{0;i} + \boldsymbol{\beta}_{1;i} \log\left(\boldsymbol{y}^i_{t-1}\right) + \boldsymbol{\beta}_{2;i} \overset{\bullet}{x}{}^{i}_{t-1,t} + \boldsymbol{\beta}_3 \overset{\bullet}{a}{}^{i}_{t-1,t} + \boldsymbol{\beta}_4 \overset{\bullet}{l}{}^{i}_{t-1,t} \\
&II.\ \underline{\boldsymbol{\beta}}_i \mid \boldsymbol{\mu_0}, \boldsymbol{\Sigma_0} \overset{\text{iid}}{\sim} N\left(\boldsymbol{\mu_0}, \boldsymbol{\Sigma_0}\right) \\
&III.\ \left(\boldsymbol{\mu_0}, \boldsymbol{\Sigma_0}, \boldsymbol{\beta}_3, \boldsymbol{\beta}_4\right) \mid \boldsymbol{\gamma} \sim p\left(\boldsymbol{\gamma}\right),\ \ \boldsymbol{\sigma}_i \sim \frac{1}{\boldsymbol{\sigma}_i}, i = 1, .., N
\end{aligned}
\tag{3.3}
$$

where $\underline{\boldsymbol{\beta}}_{i,3:4} = \left(\underline{\boldsymbol{\beta}}_i, \boldsymbol{\beta}_3, \boldsymbol{\beta}_4\right)$. Then, one can continue the hierarchy by modelling $\boldsymbol{\gamma} \sim \pi\left(\gamma_0\right)$ and so on. Although this Bayesian parametric approach is very appealing, since it allows both for random variability among countries and for borrowing information, a Gaussian distribution for the country-specific $\underline{\boldsymbol{\beta}}_i$ is a quite restrictive assumption. A Gaussian distribution has light tails and a symmetric and unimodal shape. Moreover, a Gaussian distribution does allow neither for outliers nor to have ties. For the application under investigation, there is no a priori information to trust in such well-behaving distribution of the country-specific coefficients across countries. As Figure 3.1 shows, some countries, e.g. Belgium and Finland, have instead very similar correlation behavior. This can suggest the possible appropriateness of some discrete distribution that allows to have groups of countries that cluster close together.

The Bayesian solution for the absence of parametric knowledge on the distribution of $\underline{\boldsymbol{\beta}}_i$, let us say $\boldsymbol{P}_\beta$, is to assign a prior on it with a support over the set of all the distributions on the real line. Moreover, in order to allows for ties across countries, Bayesian nonparametric statistics advices to choose a prior which selects almost surely a discrete distribution. In the following subsection, we will briefly discuss two processes which will be later used for assigning the prior on $\boldsymbol{P}_\beta$.

3.3.2 Bayesian semiparametric approaches

As most of the chapters of these thesis, the approach of this chapter is semiparametric in the sense that the parameters of interest include both an infinite-dimensional parameter, i.e. the latent population distribution of the country-specific coefficients, and some finite-dimensional parameters, e.g. the observational variances. The latent population distribution will be modelled using a Bayesian nonparametric model. Thanks to the massive development of computationally-based estimation procedures and to its robustness to model misspecifications, Bayesian nonparametric statistics has become a very popular solution to many statistical problems over the last ten years. The areas of applications range, e.g., from biostatistical applications (Dunson, 2010) to natural language processing (Sharif-Razavian and Andreas Zollmann, 2009). Instead, these approaches are only seldom used in the econometric literature. We will here propose their use within dynamic panel model for the analysis of the leverage-ratio.

Bayesian nonparametrics and dynamic panel models

Hirano (1999; 2002) developed a nonparametric generalization of the Bayesian parametric approach for random effect dynamic model for panel data. Although he also discussed (Hirano, 1999) a possible semiparametric extension for panel autoregressive models with common autoregressive coefficient across countries and assigning a DP prior on the random intercept, he then focused on the making the residual distribution more flexible because his aim was to characterize the entire joint distribution rather than some specific parameters.

By contrast, the main interest for the application under investigation here is on the country-specific coefficients and, in particular, on the coefficient of the notional assets growth rate, called the *cyclicality coefficient*. This coefficient together with the intercept and the autoregressive coefficient have to be different for each country because each banking sys-

tem behaves differently. Our approach therefore generalizes the Bayesian parametric model regarding the distributions of the heterogeneity. A nonparametric approach on the heterogeneities has many advantages. First, it ensures the robustness of the statistical results by avoiding rigid parametric assumptions on the shape of the latent mixing distribution. Second, it allows for ties across the country-specific coefficients. Third, this extension will be the starting point for introducing some time variation in the coefficients by allowing the distribution of the heterogeneity to change after a specific time. This approach is used to include the possible changes in the cyclicality coefficient after the Autumn 2008.

3.4 Our proposal

In this section, the two proposed approaches are discussed. Both proposals are Bayesian semiparametric dynamic panel models with model for the data defined by equation (3.1).

Let us make the following assumptions for model (3.1):

1. The country-specific coefficients, $\left(\boldsymbol{\beta}_{0;i}, \boldsymbol{\beta}_{1;i}, \boldsymbol{\beta}_{2;i}\right)$, given $\boldsymbol{P}_{\beta}$, are independently distributed across countries, with an unknown random distribution $\boldsymbol{P}_{\beta}$. They are also not correlated with the covariates.

2. All the regressors but the lagged variable, $\left(\overset{\bullet}{x}{}^{i}_{t-1,t}, \overset{\bullet}{a}{}^{i}_{t-1,t}, \overset{\bullet}{l}{}^{i}_{t-1,t}\right)$ are strictly exogenous, i.e. $E\left(\boldsymbol{e}_t^i \mid \left(\overset{\bullet}{x}{}^{i}_{t-1,t}, \overset{\bullet}{a}{}^{i}_{t-1,t}, \overset{\bullet}{l}{}^{i}_{t-1,t}\right)\right) = 0$.

3. The matrix collecting the regressors $\left(\overset{\bullet}{x}{}^{i}_{t-1,t}, \overset{\bullet}{a}{}^{i}_{t-1,t}, \overset{\bullet}{l}{}^{i}_{t-1,t}\right)$, $t = 1999q2, ..., 2012q1$, has full rank.

4. The disturbances are $\boldsymbol{e}_{i,t} \mid \boldsymbol{\sigma}_i^2 \overset{indep}{\sim} N\left(0, \boldsymbol{\sigma}_i^2\right)$, i.e. independent and non identically distributed over i where $\boldsymbol{\sigma}_i^2$ is the country-specific observational variance.

5. Assume that the initial values, y_0^i, $i = 1, .., N$, are all fixed.

The first assumption means that we can think of the coefficients that we observe for a country to be sampled from a larger population. We require this assumption because we want to borrow strength across the countries.

The second assumption means that if we observe a change is the leverage-ratio, this will not automatically imply a specific change in the $\left(\overset{\bullet}{x}^{i}_{t-1,t}, \overset{\bullet}{a}^{i}_{t-1,t}, \overset{\bullet}{l}^{i}_{t-1,t}\right)$. That is to say, if the leverage-ratio changes, this change will not imply a further change one of the covariates.

The third assumption means that each of the three regressors explains a different component of the changes in leverage.

The fourth assumption requires the residuals to be white noise within each country but with a country-specific variance.

The latter assumption translates to setting the initial value of leverage-ratio, i.e. 1999q1[4], to be fixed and given. As a consequence of these assumptions, leverage-ratio growth rate across countries are dependent among themselves and are conditionally independent given all the random coefficients (and the random covariates). Similarly, the country-specific random coefficients are conditionally independent given the random distribution function $\boldsymbol{P}_{\beta}$.

Depending on the model for the country-specific random coefficients, namely the second level of the hierarchy, two different hierarchical models are proposed in the following. The first proposal is to assign a DP prior to $\boldsymbol{P}_{\beta}$. The second proposal allows also for some time variation in the country-specific coefficients. In particular, it imposes a structural break in the *cyclicality coefficient* associated with the bankruptcy of Lehman Brothers, allowing also the distribution of the heterogeneity after the break to change. Moreover, it seems reasonable to assume that the behavior of the MFIs sector regarding the cyclicality reaction

[4]In 1998, eleven European Union member states had met the convergence criteria and the eurozone came into existence with the official launch of the euro on the first of January 1999. Across the countries we are studying, only Greece joined the eurozone later, i.e. on the first of January, 2001.

could have been more different across countries after the bankruptcy of Lehman Brothers than before. We therefore assume the nested clustering structure implied by the EDP, previously discussed .

Then, a third (and a forth) level is necessarily for making feasible the Bayesian estimator (Hsiao, 2003), since the variances and the other hyper-parameters are seldom known. Alternatively, one can substitute the variances, and potentially the other (hyper)-parameters, with some consistent estimators, e.g. working with empirical Bayes approach (Robbins, 1956). However, to take properly into account the variability of the variance parameters, the Lindley and Smith (1972) approach is preferable and it is here followed by assuming independent prior distributions for the variances.

Let us re-write equation (3.1), defining first-level of the hierarchy, as follows:

$$\left(\overset{\bullet}{\boldsymbol{y}}{}^{i}_{t-1,t} \mid \underline{\boldsymbol{\beta}}_{i,3:4}, \boldsymbol{\sigma}_i^2, \boldsymbol{y}^i_{t-1}\right) \overset{indep}{\sim} N\left(m^*_{i,t}, \boldsymbol{\sigma}_i^2\right) \tag{3.4}$$

where $m^*_{i,t} = \boldsymbol{\beta}_{0;i} + \boldsymbol{\beta}_{1;i} \log\left(\boldsymbol{y}^i_{t-1}\right) + \boldsymbol{\beta}^*_{2;i} \overset{\bullet}{x}{}^{i}_{t-1,t} + \boldsymbol{\beta}_3 \overset{\bullet}{a}{}^{i}_{t-1,t} + \boldsymbol{\beta}_4 \overset{\bullet}{l}{}^{i}_{t-1,t}$ and the value of $\boldsymbol{\beta}^*_{2;i,t}$ will be specified in the next subsections, since it will be different depending on whether the break is included or not. In the next subsection, first, the model without breaks in coefficients is discussed and, then, it is extended to allow for the break.

3.4.1 Our proposal with no temporal breaks: DP prior

As said, model (3.4) defines the first level of the hierarchy. We now introduce the second level of the hierarchy assuming no breaks, i.e. $\boldsymbol{\beta}^*_{2i,t} = \boldsymbol{\beta}_{2;i}$ for every t. This means that we believe that, even if the bankruptcy of Lehman Brothers has had a strong impact on the leverage-ratio of *Monetary and Financial institutions* sector, the changes in the leverage-ratio still agree with the structure of the model before the bankruptcy of Lehman Brothers occurred. We therefore propose the

following model for the country-specific random coefficients, $\underline{\boldsymbol{\beta}}_i$:

$$\underline{\boldsymbol{\beta}}_i \mid \mathbf{P}_\beta = P_\beta \sim P_\beta \tag{3.5}$$

$$\mathbf{P}_\beta \mid \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0 \sim DP(\alpha\, P_{0\beta}) \tag{3.6}$$

where $P_{0\beta} = N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ and the precision parameter α is fixed. The parameters of the base measure have the following form:

$$\boldsymbol{\mu}_0 = \begin{pmatrix} \boldsymbol{\mu}_{10} \\ \boldsymbol{\mu}_{lag} \\ \boldsymbol{\mu}_{13} \end{pmatrix},\ \boldsymbol{\Sigma}_0 = \begin{pmatrix} \boldsymbol{\sigma}^2_{1,1} & 0 & \boldsymbol{\sigma}_{13} \\ 0 & \boldsymbol{\sigma}^2_{lag} & 0 \\ \boldsymbol{\sigma}_{31} & 0 & \boldsymbol{\sigma}^2_{33} \end{pmatrix},\ \text{and call } \boldsymbol{\Sigma}_1 \equiv \begin{pmatrix} \boldsymbol{\sigma}^2_{1,1} & \boldsymbol{\sigma}_{13} \\ \boldsymbol{\sigma}_{31} & \boldsymbol{\sigma}^2_{33} \end{pmatrix}$$

Then, at the third level, the following partially-informative and independent priors are chosen:

- Independent uninformative priors on the residual variances, i.e. $\boldsymbol{\sigma}_i \sim \dfrac{1}{\boldsymbol{\sigma}_i}$.
- $\begin{pmatrix} \boldsymbol{\beta}_3 \\ \boldsymbol{\beta}_4 \end{pmatrix} \mid \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2 \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$
- $\boldsymbol{\Sigma}_1 \sim IW(c_1, D_1)$ and $(\boldsymbol{\mu}_{10}, \boldsymbol{\mu}_{13}) \mid \boldsymbol{\Sigma}_1 \sim N\left(\bar{\mu_1}, \dfrac{\boldsymbol{\Sigma}_1}{k_{01}}\right)$
- $\boldsymbol{\sigma}^2_{lag} \sim IG(c_3, d_3)$ and $\boldsymbol{\mu}_{lag} | \boldsymbol{\sigma}_{lag} \sim N\left(\bar{\mu}_{lag}, \dfrac{\boldsymbol{\sigma}^2_{lag}}{k_{03}}\right)$

The parameters $\bar{\mu_1}, \bar{\mu}_{lag}, k_{01}, k_{03}, c_1, c_3, D_1, d_3$ are known. Instead, the following hyper-priors are assigned on $\boldsymbol{\Sigma}_2$ and $\boldsymbol{\mu}_2$:

- $\boldsymbol{\Sigma}_2 \sim IW(c_2, D_2)$; and $\boldsymbol{\mu}_2|\boldsymbol{\Sigma}_2 \sim N\left(\bar{\mu_2}, \frac{\boldsymbol{\Sigma}_2}{k_{02}}\right)$

where $\bar{\mu_2}, k_{02}, c_2, D_2$ are known. This prior specification for the variances follows the Lindley and Smith (1972) approach, i.e Jeffreys prior on the standard deviation and an Inverse-Wishart for the covariance matrix of the random coefficients.

3.4.2 Our proposal with a temporal break: EDP prior

We now assume that the bankruptcy of Lehman Brothers was such an important event for MFIs to impose a structural change in their behavior. This means that the bankruptcy of Lehman Brothers implied a complete change of view, a change of prospective, by reviewing their strategic reaction to notional asset growth rate. If the bankruptcy of Lehman Brothers instead did not impacted so greatly the cyclicality behavior of MFIs, no substantial changes in the posterior distribution of the country-specific *cyclicality coefficients* would be observed. If instead MFIs started to behave differently, this change and its magnitude would be inferred. For the sake of simplicity, the break is allowed only for the coefficient of major interest, the *cyclicality coefficient*, although structural changes can be included potentially in all the coefficients.

The *cyclicality coefficient* is now specified as follows:

$$\boldsymbol{\beta}^*_{2;i,t} = \begin{cases} \boldsymbol{\beta}^{PRE}_{2;i}, & \text{if } t < 2008q3 \\ \boldsymbol{\beta}^{POST}_{2;i}, & \text{if } t \text{ i} \geq 2008q3 \end{cases} \tag{3.7}$$

Let us now assume two groups of independent country-specific coefficients. The first groups is made up of $\left(\boldsymbol{\beta}_{0;i}, \boldsymbol{\beta}_{1;i}\right)$ and the second group by $\left(\boldsymbol{\beta}^{PRE}_{2;i}, \boldsymbol{\beta}^{POST}_{2;i}\right)$. Let us consider the following model:

$$\left(\boldsymbol{\beta}_{0;i}, \boldsymbol{\beta}_{1;i}\right) \mid \mathbf{P}_{\beta_{01}} = P_{\beta_{01}} \sim P_{\beta_{01}}; \quad \left(\boldsymbol{\beta}_{2;i}^{PRE}, \boldsymbol{\beta}_{2;i}^{POST}\right) \mid \mathbf{P}_{\beta_2} = P_{\beta_2} \sim P_{\beta_2} \tag{3.8}$$

$$\mathbf{P}_{\beta_{01}} \sim DP\left(\alpha_0 \;\; P_0\right), \quad \mathbf{P}_{\beta_2} \sim EDP\left(\alpha_1 \;\; P_{0;PRE}, \alpha_2 \;\; P_{0;POST}\left(\cdot \mid \boldsymbol{\beta}_{2;i}^{PRE}\right)\right) \tag{3.9}$$

where P_0 is the Gaussian distribution defined in Subsection 3.4.1 but restricted on the bi-dimensional space, i.e. only the first two elements of $\boldsymbol{\mu}_0$ and the first 2×2 submatrix of $\boldsymbol{\Sigma}_0$; $P_{0;PRE} = N\left(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1\right)$ and $P_{0;POST}\left(\boldsymbol{\beta}_{2;i}^{PRE}\right) = N\left(\boldsymbol{\beta}_{2;i}^{PRE}, \boldsymbol{\Sigma}_2\right)$. As discussed in Chapter 2, this EDP prior is equivalent to the "sequential" prior specification defined by the following expressions

$$\mathbf{P}_{\boldsymbol{\beta}_2^{PRE}} \sim DP\left(\alpha_1 P_{0;PRE}\right) \tag{3.10}$$

$$\mathbf{P}_{\boldsymbol{\beta}_2^{POST} \mid \boldsymbol{\beta}_2^{PRE}}(\cdot \mid \boldsymbol{\beta}_{2;i}^{PRE}) \sim DP\left(\alpha_2\left(\boldsymbol{\beta}_{2;i}^{PRE}\right) P_{0;POST}\left(\cdot \mid \beta^{PRE}\right)\right) \tag{3.11}$$

where

- $\mathbf{P}_{\boldsymbol{\beta}_2^{POST} \mid \boldsymbol{\beta}_2^{PRE}}(\cdot \mid \beta_{2;i}^{PRE})$ is independent from $\mathbf{P}_{\boldsymbol{\beta}_2^{POST} \mid \boldsymbol{\beta}_2^{PRE}}(\cdot \mid \beta_{2;i}^{PRE,1})$ for all $\beta_{2;i}^{PRE} \neq \beta_{2;i}^{PRE,1}$;

- $\mathbf{P}_{\boldsymbol{\beta}_2^{POST} \mid \boldsymbol{\beta}_2^{PRE}}(\cdot \mid \boldsymbol{\beta}_{2;i}^{PRE})$ is independent from $\mathbf{P}_{\boldsymbol{\beta}_2^{PRE}}$.

The base measures are given by $P_{0;PRE} = N\left(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1\right)$ and $P_{0;POST}\left(\cdot \mid \boldsymbol{\beta}_{2;i}^{PRE}\right) = N\left(\boldsymbol{\beta}_{2;i}^{PRE}, \boldsymbol{\Sigma}_2\right)$). This means that, first, a DP prior is assigned to the distribution of the pre-break country-specific random coefficients, $\mathbf{P}_{\boldsymbol{\beta}_2^{PRE}}$, and the pre-break groups are estimated. Then,

given the clustering structure for the $\beta_{2;i}^{PRE}$, an independent DP prior is assigned on the conditional distribution, $\mathbf{P}_{\boldsymbol{\beta}_2^{POST}|\boldsymbol{\beta}_2^{PRE}}(\cdot \mid \boldsymbol{\beta}_{2;i}^{PRE})$. This implies a nested structure where, within each pre-break cluster, one or more post-break clusters can be defined. It should be pointed out that, although the verse of sequentiality between pre- and post- break coefficients is given by the structure of the real problem (first the coefficient associated with the observations before the Autumn 2008, then the coefficient for observations after that), the nested structure of the clustering is a modelling choice. This hierarchical clustering structure is adequate whenever it is reasonable to assume that *cyclicality coefficients* before the bankruptcy of Lehman Brothers can have larger subgroups of countries, and after the bankruptcy of Lehman Brothers these subgroups starts to differentiate more, making some further subgroups within the existent clustering structure.

The priors and hyperpriors assigned at the third (and fourth) level of the hierarchy are the same as in Subsection 3.4.1.

3.4.3 Inference

In this subsection, we will briefly discuss how to make inference on the proposed models. Although marginal posterior distributions have no closed-form analytical expression, they can be computed by implementing a Gibbs sampling algorithm (Gelfand and Smith, 1990) since the choices for priors, hyper-priors and mixing distributions bring to a conditionally conjugate model. Hsiao *et al.* (1999) discussed the implementation of the Gibbs sampling for a similar model in a parametric setting using a Gaussian distribution. We derive the complete set of full conditional distributions for the parametric model (3.3), completed with the common chosen prior distributions used also for the DP and the EDP model, in Appendix.

We discuss here the nonparametric extensions, starting from the DP model to ending with the EDP model. Posterior MCMC simulation for

DP mixture models is developed, for example, in Escobar and West (1995) and Neal (2000).

Call $\boldsymbol{d}^i = \left(\overset{\bullet}{\boldsymbol{y}}^i_{1999q1,1999q2}, .., \overset{\bullet}{\boldsymbol{y}}^i_{2011q4,2012q1}\right)$ and $f\left(d_i \mid \underline{\boldsymbol{\beta}}_i, \boldsymbol{\beta}_3, \boldsymbol{\beta}_4, \boldsymbol{\sigma}_i^2\right)$ the $T-1$ dimensional density associated with the vector of observations d_i. The model for the data is indeed a mixture with an infinite number of parameters:

$$\overset{\bullet}{\boldsymbol{y}}^i \mid \boldsymbol{P}_\beta, \boldsymbol{\beta}_3, \boldsymbol{\beta}_4, \boldsymbol{\sigma}_i^2 \sim \int f\left(d_i \mid \underline{\boldsymbol{\beta}}, \boldsymbol{\beta}_3, \boldsymbol{\beta}_4, \boldsymbol{\sigma}_i^2\right) dP_\beta\left(\underline{\boldsymbol{\beta}}\right)$$

In order to avoid dealing with an infinite number of parameters, P_β is integrated out and the following model is considered:

$$\overset{\bullet}{\boldsymbol{y}}^i \mid \underline{\boldsymbol{\beta}}_i, \phi \sim f\left(d_i \mid \underline{\boldsymbol{\beta}}_i, \phi\right)$$

where, for simplicity of notation, $\phi = \left(\boldsymbol{\beta}_3, \boldsymbol{\beta}_4, \boldsymbol{\sigma}_1^2, ...\boldsymbol{\sigma}_N^2\right)$[5]. The model is then completed with equations (3.5) and (3.6), which can be summarized by the Pòlya urn scheme (Blackwell and MacQueen, 1973), defined in Chapter 1, and with the priors for the other parameters and hyperparameters defined in Subsection 3.4.1. The posterior distribution of interest is then the following:

$$p\left(\underline{\boldsymbol{\beta}}_1, .., \underline{\boldsymbol{\beta}}_N, \boldsymbol{\phi}, \boldsymbol{\alpha}_\beta, \boldsymbol{\eta} \mid \overset{\bullet}{\boldsymbol{y}}^1, .., \overset{\bullet}{\boldsymbol{y}}^N\right) \propto$$
$$\left(\prod_{i=1}^{N} f\left(d_i \mid \underline{\boldsymbol{\beta}}_i, \phi\right)\right) p\left(\underline{\boldsymbol{\beta}}_1, .., \underline{\boldsymbol{\beta}}_N \mid \boldsymbol{\alpha}_\beta, \boldsymbol{\eta}\right) p\left(\boldsymbol{\alpha}_\beta, \boldsymbol{\eta}, \boldsymbol{\phi}\right)$$

where $\boldsymbol{\eta}$ collects all the parameters contained in $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$. In order to make inference, four full conditional distributions are required:

1. $p\left(\underline{\boldsymbol{\beta}}_i \mid \underline{\boldsymbol{\beta}}_{(-i)}, \boldsymbol{\alpha}_\beta, \boldsymbol{\eta}, \boldsymbol{\phi}, D\right)$
2. $p\left(\boldsymbol{\alpha}_\beta \mid \underline{\boldsymbol{\beta}}_1, .., \underline{\boldsymbol{\beta}}_N\right)$

[5]The inclusion of $\boldsymbol{\sigma}_1^2, ...\boldsymbol{\sigma}_N^2$ instead of just $\boldsymbol{\sigma}_i^2$ is preferred to keep the subscript i for the coefficients $\boldsymbol{\beta}_i$. However, the $\boldsymbol{\sigma}_i^2$ are independent among themselves.

3. $p\left(\boldsymbol{\eta} \mid \underline{\boldsymbol{\beta}}_1, .., \underline{\boldsymbol{\beta}}_N\right)$

4. $p\left(\boldsymbol{\phi} \mid \underline{\boldsymbol{\beta}}_1, .., \underline{\boldsymbol{\beta}}_N, D\right)$

where $\underline{\boldsymbol{\beta}}_{(-i)} = \left(\underline{\boldsymbol{\beta}}_1, .., \underline{\boldsymbol{\beta}}_{i-1}, \underline{\boldsymbol{\beta}}_{i+1}, .., \underline{\boldsymbol{\beta}}_N\right)$;and D is a $T-1 \times N$ matrix collecting d_1,..,d_N.

We focus here on $p\left(\underline{\boldsymbol{\beta}}_i \mid \underline{\boldsymbol{\beta}}_{(-i)}, \boldsymbol{\alpha}_\beta, \boldsymbol{\eta}, \boldsymbol{\phi}, D\right)$, which is given by:

$$p\left(\underline{\boldsymbol{\beta}}_i \mid \underline{\boldsymbol{\beta}}_{(-i)}, \boldsymbol{\alpha}_\beta, \boldsymbol{\eta}, \boldsymbol{\phi}, D\right) \propto f\left(d_i \mid \underline{\boldsymbol{\beta}}_i, \phi\right) \times p\left(\underline{\boldsymbol{\beta}}_1, .., \underline{\boldsymbol{\beta}}_N \mid \boldsymbol{\alpha}_\beta, \boldsymbol{\eta}\right)$$

The other full conditional distributions, relevant for the applications, are discussed in Appendix.

Inference for the DP model

For the DP model, using the the Pòlya urn scheme (Blackwell and MacQueen, 1973) for the distribution of the random coefficients, it follows that:

$$\begin{aligned} & p\left(\underline{\boldsymbol{\beta}}_i \mid \underline{\boldsymbol{\beta}}_{(-i)}, \boldsymbol{\alpha}_\beta, \boldsymbol{\eta}, \boldsymbol{\phi}, D\right) \propto f\left(d_i \mid \underline{\boldsymbol{\beta}}_i, \phi\right) \times \left(\alpha_\beta p_{0\beta}\left(\underline{\boldsymbol{\beta}}_i \mid \boldsymbol{\eta}\right) + \sum_{j \neq i} \delta_{\underline{\boldsymbol{\beta}}_j}\left(\underline{\boldsymbol{\beta}}_i\right)\right) \\ & = \alpha_\beta p\left(d_i \mid \boldsymbol{\phi}, \boldsymbol{\eta}\right) \frac{p_{0\beta}\left(\underline{\boldsymbol{\beta}}_i \mid \boldsymbol{\eta}\right) f\left(d_i \mid \underline{\boldsymbol{\beta}}_i, \boldsymbol{\phi}\right)}{p\left(d_i \mid \boldsymbol{\phi}, \boldsymbol{\eta}\right)} + \sum_{j \neq i} \delta_{\underline{\boldsymbol{\beta}}_j}\left(\underline{\boldsymbol{\beta}}_i\right) f\left(d_i \mid \underline{\boldsymbol{\beta}}_j, \boldsymbol{\phi}\right) \end{aligned}$$

where $p\left(d_i \mid \boldsymbol{\phi}, \boldsymbol{\eta}\right) = \int f\left(d_i \mid \underline{\boldsymbol{\beta}}_i, \boldsymbol{\phi}\right) p_{0\underline{\beta}}\left(\underline{\beta}_i \mid \boldsymbol{\eta}\right) d\beta_i$ and $p_{0\beta}$ is the density associated with $P_{0\beta}$. In a more compact form:

$$p\left(\underline{\boldsymbol{\beta}}_i \mid \underline{\boldsymbol{\beta}}_{(-i)}, \boldsymbol{\alpha}_\beta, \boldsymbol{\eta}, \boldsymbol{\phi}, D\right) \propto q_0 f\left(d_i \mid \underline{\boldsymbol{\beta}}_i, \boldsymbol{\phi}\right) p_{0\beta}\left(\underline{\boldsymbol{\beta}}_i\right) + \sum_{j \neq i} q_j \delta_{\underline{\boldsymbol{\beta}}_j}\left(\underline{\boldsymbol{\beta}}_i\right)$$

where $q_0 = \alpha_\beta p\left(d_i \mid \boldsymbol{\phi}, \boldsymbol{\eta}\right) d\beta_i$; $q_j = f\left(d_i \mid \underline{\boldsymbol{\beta}}_j, \boldsymbol{\phi}\right)$; and $q_0 + \sum_{j=1, j \neq 1}^{N} q_j = 1$.

Since $f\left(\cdot \mid \underline{\boldsymbol{\beta}}_j, \boldsymbol{\phi}\right)$ is the joint distribution of $\left(\overset{\bullet}{\boldsymbol{y}}{}^i_{1999q1,1999q2}, .., \overset{\bullet}{\boldsymbol{y}}{}^i_{2011q4,2012q1}\right)$, with $\boldsymbol{y}^i_{1999q1}$ fixed, given the assumption of conditional independence over

times, it follows that:

$$f\left(d_i \mid \underline{\boldsymbol{\beta}}_j, \boldsymbol{\phi}\right) = \prod_{t=1999q2}^{2012q1} p\left(\overset{\bullet}{\boldsymbol{y}}{}^{i}_{t-1,t} \mid \underline{\boldsymbol{\beta}}_j, \boldsymbol{\phi}, \boldsymbol{y}^i_{1999q1,\ldots,t-1}\right)$$

where $p\left(\overset{\bullet}{\boldsymbol{y}}{}^{i}_{t-1,t} \mid \underline{\boldsymbol{\beta}}_j, \boldsymbol{\phi}, \boldsymbol{y}^i_{1999q1,\ldots,t-1}\right)$ is an univariate Gaussian density, specified by equation (3.4). Under these assumptions, and with the chosen priors, q_0 can be easily evaluated analytically.

Inference for the EDP model

Let us now consider the model defined by expressions (3.8) and (3.9), with the focus on the EDP model. Repeating the same steps just illustrated for the DP and using the Enriched Pòlya urn scheme introduced in **Chapter 2**, the full conditional distribution for $\underline{\boldsymbol{\beta}}_{2;i} = \left(\boldsymbol{\beta}^{PRE}_{2;i}, \boldsymbol{\beta}^{POST}_{2;i}\right)$ for the Gibbs sampling is defined by the following two equations:

$$\begin{aligned} &p\left(\boldsymbol{\beta}^{PRE}_{2;i} \mid \boldsymbol{\beta}^{PRE}_{2;(-i)}, \boldsymbol{\alpha}_\beta, \boldsymbol{\eta}, \boldsymbol{\phi}, D\right) \\ &\propto q_0^{PRE} f\left(d_i^{PRE} \mid \boldsymbol{\beta}_i, \boldsymbol{\phi}\right) p_{0;PRE}\left(\boldsymbol{\beta}^{PRE}_{2;i} \mid \boldsymbol{\eta}\right) + \sum_{j \neq i} q_j^{PRE} \delta_{\boldsymbol{\beta}^{PRE}_{2;j}}\left(\boldsymbol{\beta}^{PRE}_{2;i}\right) \end{aligned}$$

where

- $\boldsymbol{\beta}^{PRE}_{2;(-i)} = \left(\boldsymbol{\beta}^{PRE}_1, .., \boldsymbol{\beta}^{PRE}_{2;i-1}, \boldsymbol{\beta}^{PRE}_{2;i+1}, .., \boldsymbol{\beta}^{PRE}_{2;N}\right)$;
- $d_i^{PRE} = \left(\overset{\bullet}{\boldsymbol{y}}{}^{i}_{1999q1,1999q2}, .., \overset{\bullet}{\boldsymbol{y}}{}^{i}_{2008q1,2008q2}\right)$.
- $q_0^{PRE} \propto \alpha_1 \int f\left(d_i^{PRE} \mid \beta^{PRE}_{2;i}, \boldsymbol{\phi}\right) p_{0;PRE}\left(\beta^{PRE}_{2;i} \mid \boldsymbol{\eta}\right) d\beta^{PRE}_{2;i}$; with $p_{0;PRE}$ representing the density associated with $P_{0;PRE}$.
- $q_j^{PRE} \propto f\left(d_i^{PRE} \mid \beta^{PRE}_{2;j}, \boldsymbol{\phi}\right)$;
- and $q_0^{PRE} + \sum_{j=1, j\neq 1}^{N} q_j^{PRE} = 1$.

Call $\boldsymbol{\beta}^{PRE**}_{2;1}, .., \boldsymbol{\beta}^{PRE**}_{2;n}$ the unique values of $\boldsymbol{\beta}^{PRE}_{2;1}$,...,$\boldsymbol{\beta}^{PRE}_{2;N}$ and $d_i^{POST} = \left(\overset{\bullet}{\boldsymbol{y}}{}^{i}_{2008q2,2008q3}, .., \overset{\bullet}{\boldsymbol{y}}{}^{i}_{2011q4,2012q1}\right)$. For each $s = 1, .., n$, the full conditional distribution for $\beta^{POST}_{2;i}$ for the Gibbs sampling is given by:

$$p\left(\boldsymbol{\beta}_{2;i}^{POST} \mid \boldsymbol{\beta}_{2;s}^{PRE**}, \boldsymbol{\beta}_{2;(-i)}^{POST}, \boldsymbol{\alpha}_{\beta}, \boldsymbol{\eta}, \boldsymbol{\phi}, D\right) \propto$$

$$q_0^{POST} f\left(d_i^{POST} \mid \boldsymbol{\beta}_{2;i}^{POST}, \boldsymbol{\phi}\right) p_{0;POST}\left(\boldsymbol{\beta}_{2;i}^{POST} \mid \boldsymbol{\beta}_{2;s}^{PRE**}\right)$$
$$+ \sum_{j \neq i} q_j^{POST} \delta_{\boldsymbol{\beta}_j^{POST} \mid \boldsymbol{\beta}_{2;s}^{PRE**}}\left(\boldsymbol{\beta}_{2;i}^{POST} \mid \boldsymbol{\beta}_{2;s}^{PRE**}\right)$$

where

- $\boldsymbol{\beta}_{2;(-i)}^{POST} = \left(\boldsymbol{\beta}_{2;1}^{POST}, .., \boldsymbol{\beta}_{2;i-1}^{POST}, \boldsymbol{\beta}_{2;i+1}^{POST}, .., \boldsymbol{\beta}_{2;N}^{POST}\right)$;
- $q_0^{POST} \propto \alpha_2 \int f\left(d_i^{POST} \mid \boldsymbol{\beta}_{2;i}^{POST}, \boldsymbol{\phi}\right) p_{0;POST}\left(\boldsymbol{\beta}_{2;i}^{POST} \mid \boldsymbol{\beta}_{2;i}^{PRE}\right) d\boldsymbol{\beta}_{2;i}^{POST}$;
- $q_j^{POST} \propto f\left(d_i^{POST} \mid \boldsymbol{\beta}_{2;j}^{POST}, \boldsymbol{\beta}_{2;i}^{PRE}, \boldsymbol{\phi}\right)$; and $q_0^{POST} + \sum_{j=1, j\neq 1}^{N} q_j^{POST} = 1$.

As before, q_0^{PRE} and q_0^{POST} can be evaluated analytically. Notice that in **Chapter 5** the procedure for making inference is similar, but in that case the function linking the dependent variable, here $\boldsymbol{y}_{t-1,t}^i$, with its random coefficients, here $\underline{\boldsymbol{\beta}}_i$, conditionally on all the other random coefficients, is a non-linear and non-analytically computable function. The corresponding q_0 will be not anymore analytically evaluable and computational methods will be required.

Implementation

All the algorithms are implemented setting the number of iterations to 100, 000 samples. Then, the first 20,000 are discarded as burn-in. For the posterior inference, the 2,000 subsamples from the remaining 80,000 samples, with a thinning equal to 40, are taken into considerations.

3.5 Results and conclusions

The posterior distributions of the unknown random coefficients are shown in Figures 3.3-3.10. In order to assess the impact of a covariate, the interquartile range (IQR) of the posterior distribution is analyzed. In particular, we say that there is a procyclical reaction if more than 75

% of the posterior distribution of the *cyclicality coefficient* has positive support, a threshold that is considered appropriate given the small sample size. Conversely, we say that there is a anticyclical reaction if this IQR entirely lies below zero.

3.5.1 Parametric prior

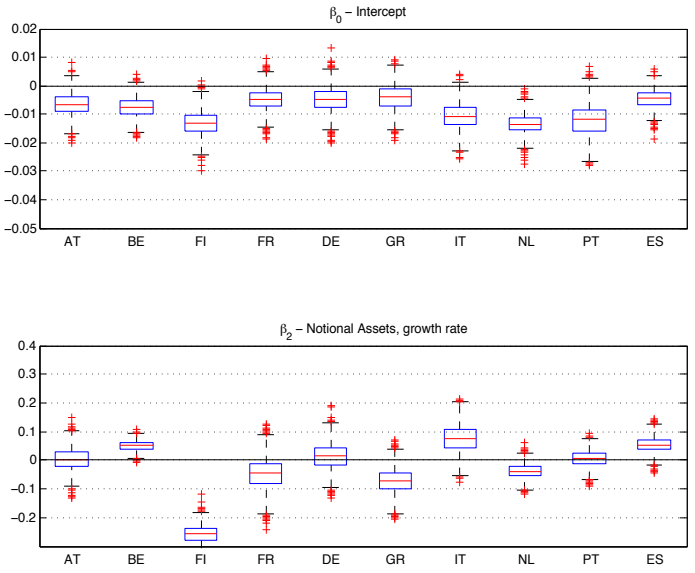


Figure 3.3: Posterior distributions: Parametric Gaussian mixing distribution

Figure 3.3 shows the parametric results for both the time-constant intercept and the *cyclicality coefficient*. We obtain procyclical leverage for some countries: Belgium, Italy and Spain; and negative cyclicality coefficient for others, i.e. Finland, France, Greece and Netherlands. At the same time, for Portugal and Austria, the posterior distribution of the cyclicality coefficient is mainly concentrated around zero. We would like

to point out that the lack of procyclicality for Germany is mainly due to the impact of the last quarters, when the leverage-ratio of MFIs has declined significantly[6], as shown in Figure 3.1. However, these results are to a certain extent induced by our model for the use of a well-behaving mixing distribution like the Gaussian one which imposes a symmetric allocation of the country-specific cyclicality coefficients around their mean. This explains the rather balanced distribution of coefficient values.

In order to evaluate the impact of the chosen prior, let us focus on the Italy. In Figure 3.4, the prior and the posterior distributions are shown. The posterior distribution moves towards right and the posterior variance of $\boldsymbol{\beta}_{2,IT}$ is half of its prior variance. Figure 3.5 shows the prior and posterior predictive distribution for $\overset{\bullet}{\boldsymbol{y}}^{IT}_{2011q4,2012q1}$. It appears that, although the prior predictive distribution is vague (mainly due to the high prior variances for the parameters), the posterior distribution is highly concentrated around its mean. Therefore, although the temporal range is made up of only 51 observations, the use of a panel of 10 countries allows to reduce substantially the variability of the priors. However, this could be related to the imposition of a specific parametric functional form.

3.5.2 DP prior

In Figure 3.6, we show the results of relaxing this rigid parametric assumption and assuming a DP prior on the mixing distribution. For all the countries, the mean of the cyclicality coefficient has still the same sign as in the parametric case, but in many cases the effect is not anymore significative. Indeed, no country has a positive and significative cyclicality coefficient except for Germany. Finland and Greece show weakly an anti-cyclical behavior. Figure 3.7 and 3.8 are the analogous of Figure 3.4 and 3.5, using the DP prior. Again, the posterior predictive distribution for the growth rate of the leverage-rate in Italy for the quarter 2011q4-2012q1, the posterior distribution is still more concentrated than

[6]Without including the last two quarters, we obtain procyclicality also for Germany, using this parametric model.

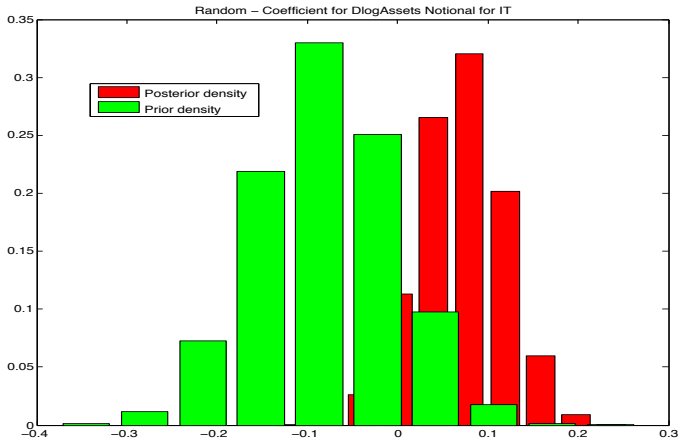


Figure 3.4: Parametric prior: Prior and Posterior distribution for $\boldsymbol{\beta}_{2,IT}$ (Italy)

the prior distribution. Moreover, looking the histograms of the prior and posterior distributions, it appears that the shape and the mean of the posterior remarkably changed, i.e. the posterior mean is much higher than the prior mean.

3.5.3 EDP prior

Figures 3.9 and 3.10 show the results of the nonparametric extension with the inclusion of the break for the cyclicality coefficient. Figures 3.11 and 3.12 show the posterior distribution for $\boldsymbol{\beta}_{2,IT}^{POST}$ for Italy and the posterior predictive distribution for the leverage-ratio growth rate of Italy, 2012q1.

Procyclicality before the Autumn 2008 is found for Belgium, France, Germany, the Netherlands, Italy and Spain according to the criterion, more than 75% of the (non-cumulative) probability distribution above zero. After the Autumn 2008, the cyclicality coefficients concentrate most

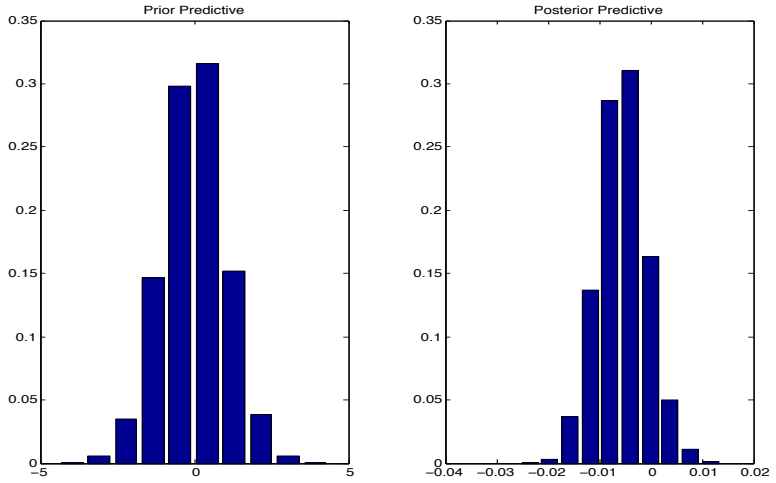


Figure 3.5: Parametric prior: Prior Predictive and Posterior Predictive distribution for the leverage-ratio growth rate of Italy, 2012q1.

of their distribution on zero for all the countries but Belgium, which instead is associated with an increase in the coefficient. As we said, the Belgium is the only country with negative assets growth rate over the period 2008q3-2012q1. Therefore, the positive and significative cyclicality coefficient for Belgium must still be considered as a prudent behaviour.

The positive cyclicality coefficient can be interpreted as signalling a banking system where episodes of strong credit and balance-sheet growth are accompanied by an insufficient build-up of precautionary capital buffers, while severe downturns where credit growth becomes negative are linked to fast accumulation of capital and leverage reductions. This behavior might contribute to the amplification of the credit cycle itself.

Therefore, for most of these countries, the support of the posterior distribution of the *cyclicality coefficient* moved downwards after Autumn 2008. We interpret this as decline in procyclicality and it would indicate an anticyclical and precautionary reaction of banks which would have

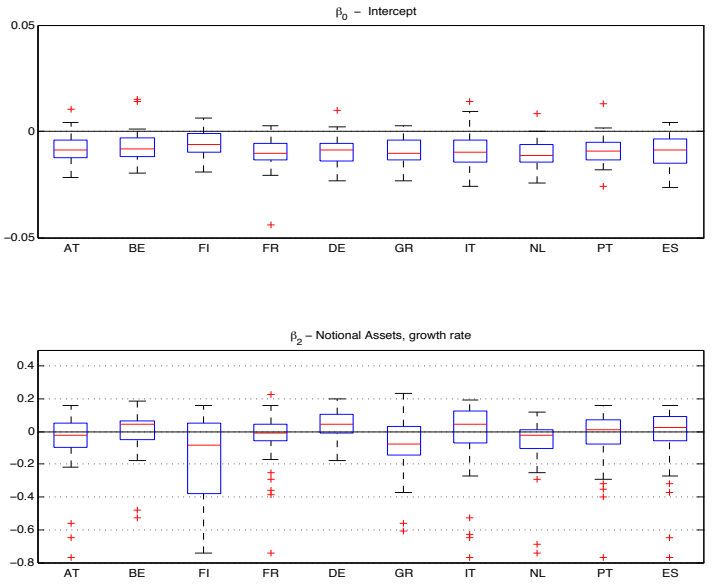


Figure 3.6: Posterior distributions: Dirichlet Process

started to reduce the mismatch between growth in funds intermediated and built-up of capital.

In the cases of the France, Germany, Netherlands, Italy and Spain, the support of the posterior distribution moves downwards from being above zero to end up being concentrated around zero after Lehman, i.e. they lose their pro-cyclical behavior after the onset of the crisis. Note that this powerful result also shows the relevance of introducing the break in our analysis: without the possibility of a change in the cyclicality coefficient, even without rigid parametric assumptions, as in the DP model, we obtained no correlation between credit and leverage for most of these countries; including the break, we instead observe a small but significative procyclical behavior before the bankruptcy of Lehman Brothers, which disappears after the crisis.

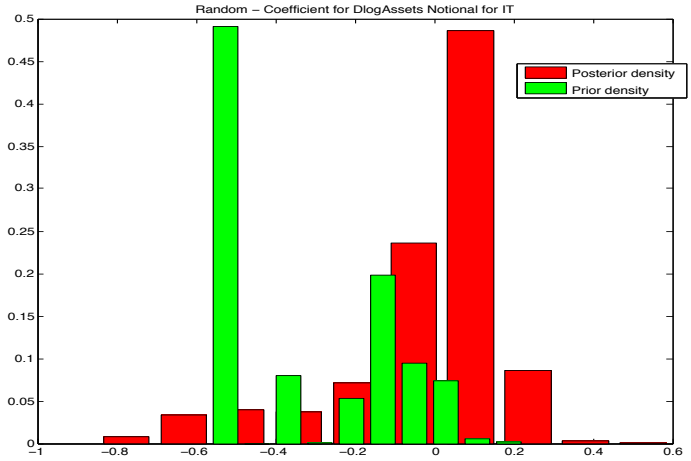


Figure 3.7: DP prior: Prior and Posterior distribution for $\beta_{2,IT}$ (Italy)

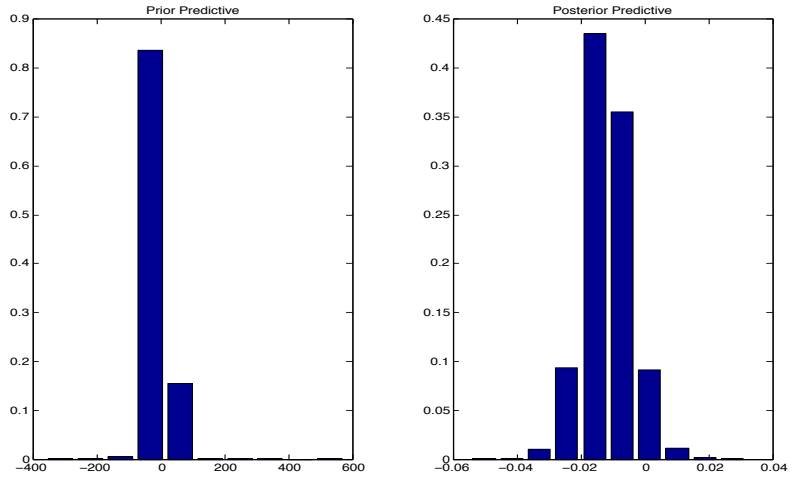


Figure 3.8: DP prior: Prior Predictive and Posterior Predictive distribution for the leverage-ratio growth rate of Italy, 2012q1.

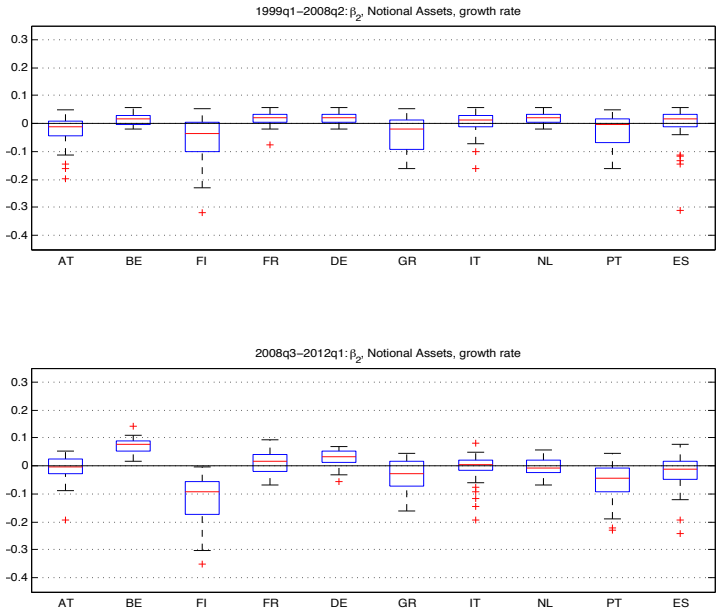


Figure 3.9: Posterior distributions: Enriched Dirichlet Process

As for the DP model, **no correlation** is found for Austria and Greece following the decision criterion based on 75% of the probability distribution over the whole of the period studied. Absence of cyclicality suggests that MFIs behave as if they had a fixed target leverage-ratio which is independent of the size of their balance-sheet. This can be also interpreted as a signal of prudent bank behavior as banks would accumulate sufficient capital during credit expansions so that their leverage-ratio remains unaffected. Similarly episodes of credit downturn would not be accompanied by excessive deleveraging. Only for Finland the support of the procyclical coefficient is negative based on a less requiring criterion, i.e. 60

Finally, Portugal shows a posterior distribution for the cyclicality coefficient that is concentrated around zero before 2008, with a slightly

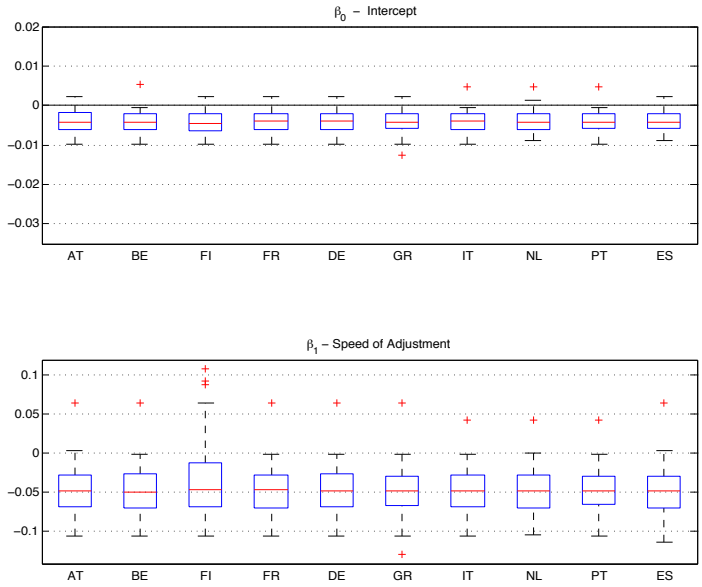


Figure 3.10: Posterior distributions: Enriched Dirichlet Process

negative median. For the second part of our sample, although the posterior distribution is still mainly concentrated on zero, the IQR moves towards the negative territory, showing a precautionary reaction similar to that of other countries.

It is important to reiterate the relevance of our modelling choices. Using a Gaussian parametric specification for the mixing distribution imposes a specific allocation of the country-specific coefficients within the available set of countries, forcing them to show as having, by construction, even a significative (although potentially unreliable) effect on the variable of interest. Thus, for example Italy and Spain display a significative procyclicality behavior using a parametric mixing distribution but no significative effect (according to the 75 % rule) is found when relaxing such a rigid parametric assumption. At the same time, results

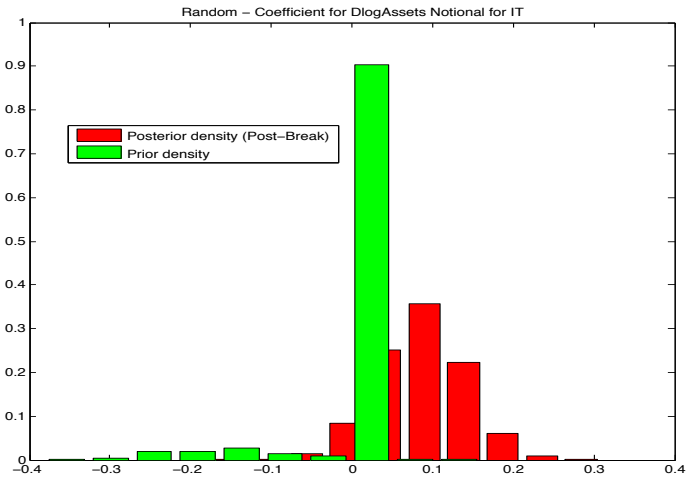


Figure 3.11: EDP prior: Prior and Posterior distribution for $\boldsymbol{\beta}_{2,IT}^{POST}$ (Italy)

obtained for the overall period do not necessarily hold for relevant subperiods. For most of the countries, e.g. Germany, France, Spain and Italy, we have found a significative and positive effect before the Autumn 2008 although the overall cyclicality coefficients of the DP model without time break are not significative different from zero.

There are reasons to believe that the effective change after the break could have had be even stronger than the one we obtain. Our sample is small, especially after the bankruptcy of Lehman Brothers. This could imply that the prior (in our case centered on the pre-bankruptcy-of-Lehman Brothers coefficients) has had a very high impact on the posterior distributions and the effective changes could be stronger.

We conclude by stating some caveats and suggesting further research. Our major difficulty is that the sample size is quite small, making the results quite sensitive to small changes of the data. Furthermore, the conclusions are based on the position of the support of the posterior

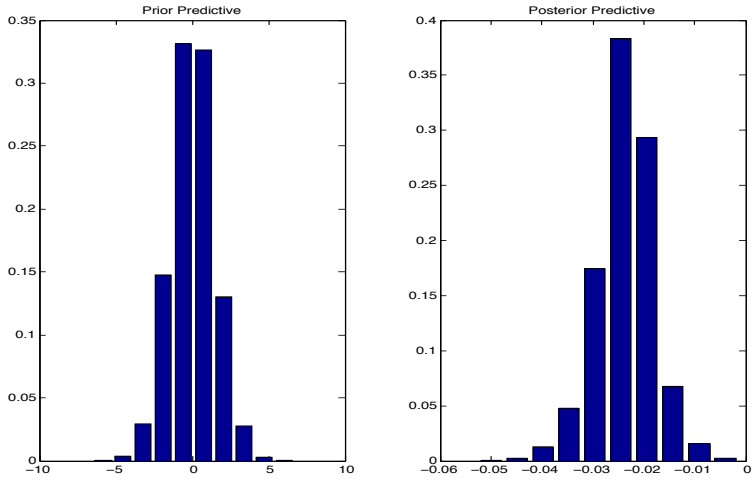


Figure 3.12: EDP prior: Prior Predictive and Posterior Predictive distribution for the leverage-ratio growth rate of Italy, 2012q1.

distributions, regardless of their shapes. Further investigation could require the implementation of some test for studying the equality of the whole pre- and post bankruptcy of Lehman Brothers posterior distributions. Moreover, it could also be required a more formal evaluation of the assumptions underlying the various models. The proposed approach is fully explorative, in the sense that we aim to understand the cyclicality of leverage. If one wants to make predictions for leverage, the co-movements across countries could be better exploited, e.g. using factor models. Finally, we are currently investigating a more "robust" approach, in the sense the results have not be affected by (possible) endogeneity bias, due to the potential correlation between notional assets growth rate and (the omitted) notional debts growth rate[7]. We are now looking for some *valid* instruments for the notional assets growth rate. Till now, only weak

[7]Notice that if we had included also the notional debts growth rate in the regression, it would have become a perfect equality instead of a true regression. Moreover, the data shows that there is such correlation, although of small entity.

instruments have been found and we think that using an instrument variable approach with weak instruments can be even worse than a (small) endogeneity problem. Finding valid instruments here is not easy due to the hard task of finding variables which are correlated with the "real" growth but without being correlated with the leverage (since the recent crisis had an impact on both of them).

Appendix

Gibbs Sampler algorithm and full conditional distributions

The Gibbs sampler for the parametric case is based on Hsiao *et al.* (1999), with the inclusion of the common-across countries coefficients for the two country-specific regressors of price indices.

Gibbs sampler

In this subsection, we derive the full conditional densities from the joint posterior density. We describe the parametric case. The changes required for the nonparametric extensions have been already discussed in Subsection 3.4.3. For simplicity of notation, denote $t = 1999q1$ with $t = 0$ and $T = 2012q1$, so that $t = 0, 1, .., T$. We write $\Delta log\left(y_t^i\right) \equiv log\left(y_t^i\right) - log\left(y_{t-1}^i\right)$ instead of $\overset{\bullet}{y}{}_{t-1,t}^{i}$ (t=1,..,T). Let $\Delta log\left(y^i\right)$ be the vector collecting $\Delta log\left(y_t^i\right)$ for t=1,...,T. Let y_{-1}^i the vector collecting y_t^i for t=0,..,T-1. We assume that $\beta_{1;i}$ is independent from $(\beta_{0;i}, \beta_{2;i})$. Call $\gamma_{1;i} = (\beta_{0;i}, \beta_{2;i})$ and $\gamma_2 = (\beta_3, \beta_4)$. Call 1_T a $T \times 1$ vector of ones and I_T the $T \times T$ identity matrix.

Given $y_0^1, ..., y_0^N$, the joint distribution of data and unobservable (parameters and latent variables) is given by:

$$f\left(\Delta y^1,...,\Delta y^N,\beta_{1,1},...\beta_{1N},\gamma_{1,1},...,\gamma_{1N},\gamma_2,\sigma_1^2,...,\sigma_N^2,\mu_1,\Sigma_1,\mu_2,\Sigma_2,\mu_{lag},\sigma_{lag}^2\right)=$$
$$\prod_{i=1}^{N}\left[N_T\left(\Delta log\left(y^i\right);\beta_{0;i}1_T+\beta_{1;i}\log\left(y_{-1}^i\right)+\beta_{2;i}\overset{\bullet}{A}+\beta_3\overset{\bullet}{p}^{A,i}+\beta_4\overset{\bullet}{p}^{L,i},\sigma_i^2I_T\right)\right]\cdot$$
$$\prod_{i=1}^{N}\left[N_1\left(\beta_{1;i};\mu_{lag},\sigma_{lag}^2\right)\cdot N_2\left(\gamma_{1;i};\mu_1,\Sigma_1\right)\cdot\frac{1}{\sigma_i^2}\right]\cdot$$
$$N_2\left(\gamma_2;\mu_2,\Sigma_2\right)\cdot IW\left(\Sigma_2;c_2,D_2\right)\cdot N_2\left(\mu_2;\bar{\mu}_2,\frac{\Sigma_2}{k_{02}}\right)\cdot IW\left(\Sigma_1;c_1,D_1\right)\cdot N_2\left(\mu_1;\bar{\mu}_1,\frac{\Sigma_1}{k_{01}}\right)\cdot$$
$$IG\left(\sigma_{lag};c_3,D_3\right)\cdot N_1\left(\mu_{lag};\bar{\mu}_{lag},\frac{\sigma_{lag}^2}{k_{03}}\right)$$

The sampler consists of the following blocks:

⋆ *Heterogeneities:*

$$p\left(\beta_{1;i}|y^i,\gamma_{1;i},\gamma_2,\sigma_i^2,\mu_1,\Sigma_1,\mu_2,\Sigma_2,\mu_{lag},\sigma_{lag}^2\right)=N\left(b_{1T},\frac{\sigma_i^2\sigma_{lag}^2}{\sigma_i^2+\sigma_{lag}^2\sum_{t=1}^{T}\log\left(y_{t-1}^i\right)}\right)$$

with

$$b_{1T}=\frac{\sigma_i^2\mu_{lag}+\sigma_{lag}^2\sum_{t=1}^{T}\log\left(y_{t-1}^i\right)\left(\log\left(y_t^i\right)-\log\left(y_{t-1}^i\right)-\beta_{0;i}-\beta_{2;i}\overset{\bullet}{x}_{t-1,t}^i-\beta_3\overset{\bullet}{a}_{t-1,t}^i-\beta_4\overset{\bullet}{l}_{t-1,t}^i\right)}{\sigma_i^2+\sigma_{lag}^2\sum_{t=1}^{T}\log\left(y_{t-1}^i\right)}$$

$$p\left(\gamma_{1;i}|y^i,\beta_{1;i},\gamma_2,\sigma_i^2,\mu_1,\Sigma_1,\mu_2,\Sigma_2,\mu_{lag},\sigma_{lag}^2\right)=$$
$$N\left(s_n\left(\Sigma_1^{-1}\mu_1+\frac{\sum_{t=1}^{T}[1\ \ \beta_{2;i}]'}{\sigma_i^2}\left(\log\left(y_t^i\right)-\left(\beta_{1;i}+1\right)\log\left(y_{t-1}^i\right)-\beta_3\overset{\bullet}{a}_{t-1,t}^i-\beta_4\overset{\bullet}{l}_{t-1,t}^i\right)\right),s_n\right)$$

where $s_n=\left(\Sigma_1^{-1}+\frac{\sum_{t=1}^{T}X_{1;i}'X_{1;i}}{\sigma_i^2}\right)^{-1}$

⋆*(Hyper-)Parameters:*

$$p\left(\sigma_i^2|y^i,\beta_{1;i},\gamma_{1;i},\gamma_2,\mu_1,\Sigma_1,\mu_2,\Sigma_2,\mu_{lag},\sigma_{lag}^2\right)=$$
$$IG\left(\frac{T}{2},\frac{1}{2}\sum_{t=1}^{T}\left(\log\left(y_t^i\right)-\beta_{0;i}-\left(\beta_{1;i}+1\right)\log\left(y_{t-1}^i\right)-\beta_{2;i}\overset{\bullet}{x}{}_{t-1,t}^{i}-\beta_3\overset{\bullet}{a}{}_{t-1,t}^{i}-\beta_4\overset{\bullet}{l}{}_{t-1,t}^{i}\right)^2\right)$$

$$p\left(\gamma_2|y^i,\beta_{1;i},\gamma_{1;i},\sigma_i^2,\mu_1,\Sigma_1,\mu_2,\Sigma_2,\mu_{lag},\sigma_{lag}^2\right)=N\left(f_n\cdot m_n,f_n\right)$$

where $f_n=\left(\Sigma_2^{-1}+\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{1}{\sigma_i^2}[\overset{\bullet}{a}{}_{t-1,t}^{i}\ \ \overset{\bullet}{l}{}_{t-1,t}^{i}]'\cdot[\overset{\bullet}{a}{}_{t-1,t}^{i}\ \ \overset{\bullet}{l}{}_{t-1,t}^{i}]\right)^{-1}$
and
$m_n=\Sigma_2^{-1}\mu_2+\frac{1}{\sigma_i^2}\Sigma_{t=1}^{T}\Sigma_{i=1}^{N}[\overset{\bullet}{a}{}_{t-1,t}^{i}\ \ \overset{\bullet}{l}{}_{t-1,t}^{i}]'\sigma_i^2\left(\log\left(y_t^i\right)-\beta_{0;i}-\beta_{1;i}\log\left(y_{t-1}^i\right)-\beta_{2;i}\overset{\bullet}{x}{}_{t-1,t}^{i}\right)$

Since γ_2 is common for all the countries, we decide to include also another hyper-prior on the parameters of its Gaussian distribution.

$$p\left(\mu_1|y^i,\beta_{1;i},\gamma_{1;i},\gamma_2,\sigma_i^2,\Sigma_1,\mu_2,\Sigma_2,\mu_{lag},\sigma_{lag}^2\right)=N\left(\frac{k_{01}}{k_{01}+N}\mu_{01}+\frac{N}{k_{01}+N}\bar{\gamma_1},\frac{\Sigma_1}{k_{01}}\right)$$

where $\bar{\gamma_1}=\dfrac{1}{N}\sum_{i=1}^{N}\gamma_{1;i}$.

$$p\left(\Sigma_1|y^i,\beta_{1;i},\gamma_{1;i},\gamma_2,\sigma_i^2,\mu_1,\mu_2,\Sigma_2,\mu_{lag},\sigma_{lag}^2\right)=$$
$$IW\left(c_1+N,D_1^{-1}+\sum_{i=1}^{N}\left(\gamma_{1;i}-\bar{\gamma_1}\right)\left(\gamma_{1;i}-\bar{\gamma_1}\right)'+\frac{k_{01}N}{k_{01}+N}\left(\bar{\gamma_1}-\mu_{01}\right)\left(\bar{\gamma_1}-\mu_{01}\right)'\right)$$

Similarly for (μ_2,Σ_2).

$$p\left(\mu_{lag}|y^i,\beta_{1;i},\gamma_{1;i},\gamma_2,\sigma_i^2,\Sigma_1,\mu_2,\Sigma_2,\mu_1,\sigma_{lag}^2\right)=N\left(\frac{k_{03}}{k_{03}+N}\mu_{0,lag}+\frac{N}{k_{03}+N}\mu_{\bar{lag}},\frac{\sigma_{lag}^2}{N+k_{03}}\right)$$

$$p\left(\sigma_{lag}^2|y^i,\beta_{1;i},\gamma_{1;i},\gamma_2,\mu_1,\Sigma_1,\mu_2,\Sigma_2,\mu_{lag},\sigma_i^2\right)=$$

$$IG\left(c_3+\frac{N}{2}, D_3^{-1}+\frac{1}{2}\sum_{i=1}^{N}\left(\beta_{1;i}-\bar{\beta_{1;i}}\right)^2+\frac{Nk_{03}}{2(N+k_{03})}\left(\bar{\beta_{1;i}}-\mu_{0,lag}\right)^2\right)$$

where $\bar{\beta_1}=\frac{1}{N}\sum_{i=1}^{N}\beta_{1;i}$.

Bibliography

[1] Adrian T., Shin H.S. (2010). Liquidity and Leverage. *Journal of Financial Intermediation*, **19**, 418-437.

[2] Adrian T., Colla P., Shin H.S. (2011). Which Financial Frictions? Parsing the Evidence from the Financial Crisis of 2007-09. *Federal Reserve Bank of New York Staff Report* no. 528.

[3] Arellano M., Bond S. (1991). Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations. *The Review of Economic Studies*, **58**, 277-297.

[4] Baglioni A., Boitani A., Liberatore M., Monticini A. (2010). Is the Leverage of European Commercial Banks Pro-Cyclical? *Quaderni dell' Istituto di Economia e Finanza*. Milan: Università Cattolica del Sacro Cuore.

[5] Baglioni A., Boitani A., Liberatore M., Monticini A. (2012). Is the Leverage of European Commercial Banks Pro-Cyclical? Technical Report available at `http://monticini.altervista.org/bbbm_r1.pdf`.

[6] Baltagi B. H. (2000). *Nonstationary Panels, Panel Cointegration, and Dynamic Panels*. New York: Elsevier, Advances in Econometrics. Volume **15**.

[7] Blackwell D., MacQueen J.B. (1973). Ferguson Distributions Via Pòlya Urn Schemes. *Annals of Statistics*, **1**, 353-355.

[8] Croux C., Forni M., Reichlin L. (2001). A Measure of Comovement for Economic Variables: Theory and Empirics. *Review of Economics and Statistics*, **83**, 232-241.

[9] Dunson D.B. (2010). Nonparametric Bayes applications to biostatistics. In [14].

[10] Ferguson T.S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, **1**, 209-230.

[11] Gelfand A., Smith A. (1990). Sampling-Based Approaches to Calculating Marginal Densities. *Journal of American Statistical Association*, **85**, 398-409.

[12] Hirano K. (1999). Semiparametric Bayesian models for Dynamic Earnings Data. Manuscript, available online on his website: `http://www.u.arizona.edu/~hirano/research.html`.

[13] Hirano K. (2002). Semiparametric Bayesian Inference in Autoregressive Panel Data Models. *Econometrica*, **70**, 781-799.

[14] Hjort N.L., Holmes C., Müller P., Walker S. (2010). *Bayesian Nonparametrics*. Cambridge, UK: Cambridge University Press.

[15] Hsiao C., Pesaran M. H., Tahmiscioglu K. (1999). Bayes Estimation of Short-Run Coefficients in Dynamic Panel Data Models. In *Analysis of Panels and Limited Dependent Variables: A Volume in Honour of G. S. Maddala* (Hsiao C., Lahiri K., Lee L.F., Pesaran M.H., eds.) 268-296, Cambridge, UK: Cambridge University Press.

[16] Hsiao C. (2003). *Analysis of Panel Data*. Cambridge, UK: Cambridge University Press.

[17] Hsiao C., Pesaran M. H. (2004). Random Coefficient Panel Data Models. *CESifo working paper*, 1233.

[18] Koop G. (2003). *Bayesian Econometrics*. Chichester, UK: John Wiley & Son.

[19] Lindley D.V., Smith F.M. (1972). Bayes Estimates for the Linear Model.*Journal of the Royal Statistical Society, Series B*, **34**, 1-41.

[20] Müller P., Quintana F. (2004). Nonparametric Bayesian Data Analysis. *Statistical Science*, **19**, 95-110.

[21] Neal R. M. (2000). Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics*, **9**, 249-265.

[22] Pesaran M. H., Smith R.P. (1995). Estimating Long-Run Relationships from Dynamic Heterogeneous Panels. *Journal of Econometrics*, **68**, 79-113.

[23] Pesaran M. H., Tosetti E. (2011). Large panels with common factors and spatial correlations. *Journal of Econometrics*, **161**, 182-202.

[24] Sharif-Razavian N., Zollmann A. (2009). An Overview of Nonparametric Bayesian Models and Applications to Natural Language Processing. Manuscript, available online at `http://www.cs.cmu.edu/simzollmann/publications/nonparametric.pdf`.

[25] Zhang P., Small D. (2006). Bayesian Inference for random coefficient dynamic panel data models. Manuscript, available online at `http://repository.upenn.edu/dissertations/AAI3225572/`.

Chapter 4

Non-linear filtering for estimating assets portfolio composition with application to securities in euro area accounts

Abstract :

This chapter proposes a novel approach for making inference on unknown compositional data. It is based on sector accounts for the euro area and these unknown compositional data represent the unknown portfolio structure of securities for each of the seven institutional sectors of the economy. The proposed approach combines a Markovian assumption on the portfolio structure with the observed dynamics of the total holdings of securities at market-value, their corresponding transactions and price indices representing the set of potential issuer sectors within a conditional sector-specific state-space model. For each sector, two series are then estimated. One is the series of the means of the filtering distributions associated with the unknown portfolio structure by counterpart sector, also known as *who-to-whom matrix*. The other is the series of the *other volume changes*, identified as the outliers of the model. The inference is based on a sector-specific particle filter algorithm and the parameters are mainly estimated based on drawing values from a continuous importance density within the algorithm (Liu and West, 2001). We also propose some further developments of this sector-specific approach. First, the inference is implemented jointly for all the sector and a vertical consistency constraint is

included to restrict the support of the sector filtering distributions. Second, it is also proposed to estimate the parameters of the sector-specific model by minimizing the conditional aggregate expected loss associated with the deviations between the corresponding filtered means in some periods and some data available for a few sub-periods.

4.1 Introduction

The incapacity of forecasting the recent crisis and its propagation have pointed out how difficult it is to properly understand the structure and the dynamics of our complex economies. These failures have motivated researchers to look for (relatively) new directions of research. In order to have a deeper understanding of our economies, many aspects should be taken into consideration. This chapter focuses on two of them. First, economic systems are increasingly built on interdependencies, e.g. among countries and among sectors. The undervaluation of these networks entails the incorrect estimation of the systemic implications associated with the crisis of a single sector, like the dramatic consequences followed by the crisis of the investment banks in the Autumn 2008. Second, many of the standard models do not allow for sharp changes in the dynamics of the series under modeling, implying the impossibility of properly estimating and forecasting crisis by construction.

This chapter proposes a procedure for making online inference on the sectoral *inter-dependencies*, represented by a matrix, namely the *who-to-whom matrix*, which collects, for each quarter, all the portfolio structure of institutional sectors by counterpart sectors. The knowledge of this matrix can then allow to properly evaluate the dynamic propagation of recessive or expansive shocks and, more generally, for financial stability assessment. The proposed methodology also identifies the series of *other volume changes* such as the outliers in the framework of the underlying data generation process, describing events like bankruptcies, crisis, defaults, et cetera.

In the particular application taken into consideration, the portfolio

structures are *compositional data* expressing the composition of securities held by a sector. At macro level, their values are seldom known. In literature, several models for (observable) compositional data are available. Most of them are based on a state-space model where the observational process is associated with the sequence of compositional data. The reasons for analyzing observed compositional data include their forecasting, the estimation of the underlying trend and of the effect of covariates and interventions and the estimation of some unobservable signal. Among the available approaches for known compositional data, there is a model based on a Dirichlet distribution on the conditional observational data (Gary *et al.*, 1993), a logistic transformations (da-Silva *et al.*, 2001) and a Bayesian beta model (da-Silva *et al.*, 2010). The estimation techniques are mainly based on the Kalman filters, when possible, or on Markov Chain Monte Carlo (MCMC) techniques. Our approach is different from the ones existing in literature with respect to both the data availability -the compositional data are unobserved- and the estimation procedure, which is based on a (much more efficient) sequential Monte Carlo technique. Consequently, the aims are different too. The primary aims of our analysis is to make inference on these unknown compositional data.

The chapter is organized as follows. In Section 4.2, the basic notions on the sector accounts are introduced. Section 4.3 briefly discusses a possible procedure for estimating *other volume changes* when portfolio structure is known, i.e. for liabilities. In Section 4.4, first the proposed model for the unknown compositional data is introduced, i.e. a neither linear nor Gaussian conditional state-space model. Then, the proposed procedure for estimating the *other volume changes* is described, namely a sequential, three-step procedure which makes use of the fitted portfolio structure. In Section 4.5, we briefly describe the estimation technique, i.e. an *auxiliary particle filter* based on the algorithm of Liu and West (2001). We also discuss the inclusion of a constraint on the support of the filtering distribution. In Section 4.6, the proposed procedure is applied

to *securities other than shares* for the euro area and the results are discussed. In Section 4.7, a possible extension of the model is then proposed by incorporating the available sub-series of portfolio structure. Finally, in Section 4.8, the discussion of possible extensions, improvements and conclusions. In Section 4.9, some technical details are provided.

4.2 Basic sector accounts notions and accounting identity

In this chapter, we consider quarterly data associated with the market value of an aggregate financial instrument across the seven institutional sectors of the whole euro area economy. In this section, we briefly define the main notions involved in this problem and introduce the concrete objective of the analysis.

4.2.1 Framework for the euro area accounts

The ESA 1995 (European Commission, 1996) and, consequently, the euro area account (European Central Bank, 2010) define the framework for producing the euro area accounts, e.g. defining the categories of financial asset, namely *financial instruments*, and specifying institutional sectors .

The financial instruments are classified on the basis of liquidity factors and legal characteristics, including categories like securities, loans or shares. For the sake of expositional simplicity, the focus of this chapter will be only on the *securities other than shares*, although the proposed procedure can be applied to any instrument on the assets side.

The seven institutional sectors all together represent the whole economy. Six of them are domestic sectors: 1) *non-financial corporations*, 2) *monetary financial institutions*, 3) *other financial intermediaries different from insurance corporations and pension funds*, 4) *insurance corporations and pension funds*, 5) *government*, and 6) *household*. The seventh one is the so-called *rest-of-the-world* and it includes all the non-residential

agents, regardless of their activities.

Let us denote the time with t, $t = t_0, .., t_T$, and label the above-mentioned sectors with $i = 1, .., 7$ respectively. The variables taken into consideration, for each sector, are roughly either quantities (at market value) or indices. The indices taken into consideration are price indices, denoted with $\overset{\bullet}{p}^{i}_{t-1,t}$, and exchange rates. For the sake of simplicity, let us assume that the exchange rate has a negligible impact[1]. The quantities used for the analysis are instead three:

1. *stocks*, denoted by $S_{i,t}$, which represents the outstanding amounts accumulated over time;

2. *transactions*, denoted by $T_{i,t}$, which are quarterly flows representing the net amount invested or disinvested in the specific quarter realized with mutual agreement;

3. *Other Volume Changes* (OVCs), denoted by $O_{i,t}$, which are quarterly flows realized without mutual agreement and not imputable to monetary changes;

4. *flows*, let us say $F_{i,t}$, defined as the sum of transactions and *other volume changes*, i.e. $F_{i,t} = T_{i,t} + O_{i,t}$. Properly speaking, the so-defined $F_{i,t}$ identifies all the *flows different from revaluations*, but we prefer to keep separate the pure quantitative flows and the revaluations.

Indices can be used to link all of these quantities together. In particular, the current stock is the sum of the previous stock and flows, *once appropriate revaluations are computed*. The main objectives of this chapter,

[1] We ignore the exchange rate because we believe that its inclusion in the model would basically only add complexity. In particular, for all the sectors but rest-of-the world, the exchange rate effect is indirectly included in the price (or revaluation) index effect since the price index for the *rest-of-the-world* is converted in euro. For the *rest-of-the-world*, all the variables are expressed in euro, therefore an explicit exchange rate component is not necessary because its effect is already included in the economic value of the quantities.

i.e. to estimate the portfolio structure and the series of *other volume changes*, reduce to compute appropriate revaluations. If the composition of the stocks were homogeneous or if the dynamic series of portfolio composition were known, then computing appropriate revaluations would be relatively simple.

The seven main sectors are usually not explicitly modelled all together. First, each sector is studied alone and a model expressing the relationship among the quantities of each sector is specified. Then, missing quantities are estimated on the basis of this model, e.g. missing OVCs are estimated. Finally, the sectors are all integrated using an appropriate *consistency techniques*. The resulting values associated with each of the three above-mentioned quantities are usually collected within a square matrix, called *who-to-whom matrix* or *Social Accounting Matrix* (SAM), whose entries represent changes of the quantities in a specific quarter by counterpart sector. Each row of this matrix represents the decomposition of the quantities for a specific sector among all the sectors on the assets side and each column the decomposition of the flows of a sector among all the sectors on the liabilities side. The problem of ensuring *consistency* mainly requires adjusting the entries of the *who-to-whom matrix* so that their sum on the assets side (horizontal sum) and on the liabilities side (vertical sum) equal the two given margins for the total sector flows on the assets side and for the liabilities side). This problem is also known as *balancing of a matrix*. The usual approach consists in implementing some algorithms for matrix balancing, e.g. a scaling algorithm or an optimization algorithm. The former multiplies the rows and columns of the matrix by positive constants until the matrix is balanced. The latter minimizes a penalty function that measure the deviation of a candidate balanced matrix from some given matrix (Schneider *et al.*, 1990). In Section 4.5.2, we propose a different way for stimulating consistency. It will not necessary ensure exact consistency, but it will avoid big discrepancy with respect to the balanced matrix. However, before Section 4.5.2, all

the sectors are modelled separately.

4.2.2 Relationship among variables

If $\overset{\bullet}{p}^{i}_{t-1,t}$ were representative indices of *both* stock, then, for each sector i, it would be possible to write:

$$S_{i,t} = S_{i,t-1}(1 + \overset{\bullet}{p}^{i}_{t-1,t}) + (T_{i,t} + O_{i,t})\,(1 + \overset{\bullet}{p}^{i}_{t-1,t}) \tag{4.1}$$

That is, the stock at time t for the sector i, $S_{i,t}$, is equal to the previous stock, $S_{i,t-1}$, revaluated accordingly to the representative price index for the sector i, $\overset{\bullet}{p}^{i}_{t-1,t}$, plus all the flows happened within the interval $(t-1, t]$, $(T_{i,t} + O_{i,t})$, also revaluated accordingly to $\overset{\bullet}{p}^{i}_{t-1,t}$. But having a $\overset{\bullet}{p}^{i}_{t-1,t}$ representative of *both* stock and flows requires the same composition for both stock and flows, which implies homogeneity of the stock and requires that all the flows happen at the beginning of the quarter. In the real world, flows are temporally distributed over the quarter and, assuming that such representative indices are available, the general accounting equality (4.1) for each sector i should be written as follows:

$$S_{i,t} = S_{i,t-1}(1+\overset{\bullet}{p}^{i}_{t-1,t}) + \sum_{\tau=t-1}^{t} T_{i,\tau}(1+\overset{\bullet}{p}^{i}_{\tau,t}) + \sum_{\tau'=t-1}^{t} O_{i,\tau'}(1+\overset{\bullet}{p}^{i}_{\tau',t}) \tag{4.2}$$

where $\overset{\bullet}{p}^{i}_{\tau,t}$ is the revaluation index from the time τ till the time t; $T_{i,\tau}$ and $O_{i,\tau'}$ are, respectively, the intra-quarter transactions and the intra-quarter other volume changes realized at time i, with $t-1 \leq \tau, \tau' < t$, and

$$T_{i,t} = \sum_{\tau=t-1}^{t} T_{i,\tau}, \qquad O_{i,t} = \sum_{\tau'=t-1}^{t} O_{i,\tau'}$$

where τ and τ' represent the counter for each transaction and each *other volume changes*. However, the intra-quarter flows are usually not observed and, hence, assumptions are required for revaluating $T_{i,t}$ and $O_{i,t}$

within their quarter. Moreover, the portfolio composition is usually heterogenous and it is often unknown for instruments on the assets side. Hence, representative price indices are usually not available.

4.2.3 Objectives of the analysis

The final main objectives of this chapter are basically two:

- to find the portfolio composition of both the stock and the flows, or, equivalently, to find the appropriate revaluation components;
- to improve the quality of the series of *other volume changes*, $O_{i,t}$, when some preliminary series is available or to produce them when no preliminary value is available.

Another objective is to control for the cyclical problem for the unknown variables. In particular, the *other volume changes* are generally defined as the *outliers* of equation (4.2), and therefore they depend on revaluations, which in turn depend on the estimated portfolio composition. This portfolio composition is estimated conditionally on the flows, $F_{i,t} = (T_{i,t} + O_{i,t})$. Hence, there is a cyclical and recursive problem in their definition.

4.3 Knowing the portfolio composition: Liabilities

Let us consider a generic sector i and let us start by assuming that the composition of the stock is homogeneous. For example, let us consider *securities other than shares* on the liabilities side of the government-sector balance sheet.

If all the involved variables in equation (4.2) were known, then we would have just an equality and no need of modeling. As said in the previous section, even when both stock and flows compositions are known, in order to properly revaluate transactions, the knowledge of their temporal

distribution within the quarter is required because each single transaction within the quarter should be revaluated using a different index. Unfortunately, this information is usually unknown. Hence, even in the simplest case of liabilities, some assumptions for computing revaluations of the flows $F_{i,t}$ are necessary. The usual procedure is to revaluate the total quarterly transactions with a unique quarterly index and assume that the revaluation index for transactions is exactly half of the revaluation index for stocks. This corresponds to the assumption that all the transactions are recorded in the mid-point of the quarter and the revaluation index from the mid-quarter to the end-quarter point is exactly half of the total revaluation index for the whole quarter[2]. Instead, the *other volume changes* are usually not revaluated. The model can then be expressed, for each sector i, as follows:

$$\boldsymbol{S}_{i,t} = \boldsymbol{S}_{i,t-1}(1 + \beta_{\mathrm{p};i}\overset{\bullet}{p}{}^{i}_{t-1,t}) + T_{i,t}\left(1 + \frac{\beta_{\mathrm{p};i}\overset{\bullet}{p}{}^{i}_{t-1,t}}{2}\right) + O^{P}_{i,t} + \boldsymbol{u}_{i,t} \quad (4.3)$$

where, for each sector i, $\boldsymbol{u}_{i,t}$ is the Gaussian-distributed noise, with potentially correlation over time, $Corr\left(\boldsymbol{u}_{i,s}, \boldsymbol{u}_{i,t}\right) \neq 0, s \neq t$, but uncorrelated among sectors, i.e. $Corr\left(\boldsymbol{u}_{i's}, \boldsymbol{u}_{i,t}\right) = 0, i \neq i'$; the $O^{P}_{i,t}$ represents preliminary other volume changes for the sector i, which is identically equal to zero when no preliminary values are provided; and the coefficient $\beta_{\mathrm{p};i}$ is associated with the sector i and it represents the possible difference in speed and magnitude of the revaluations from those of the corresponding aggregate market, represented by $\overset{\bullet}{p}{}^{i}_{t-1,t}$.

The corresponding likelihood can be written with standard statistical techniques (e.g. Kalman Filter) and, depending on the assumptions on the error term, different kind of temporal correlation can be introduced (Diz, 2009). Once the parameters of the above model are fitted, the fitted

[2]This assumption have several potential interpretations; for example, it is like assuming that the prices move monotonically and deterministically over the quarter, or they move uniformly between the end-quarter prices

residuals, let us say $\hat{u}_{i,t}$, are computed for each sector i. The common approach for updating the preliminary series or computing the series of *other volume changes* is to compute the least likelihood values under standard model assumptions (i.e. Gaussian errors). Then, the OVCs are defined as the part of the residuals greater than some a priori specified threshold.

4.4 Our proposed model for assets: sector-by-sector approach

The above approach is simple, clear and, therefore, attractive. Given last-quarter stock, flows of the quarter and revaluation indices, it is possible to estimate the current-quarter *other volume changes* with standard outliers detecting techniques. Knowing which revaluation index should be used requires knowing the portfolio composition of the stocks and of the flows; and, for most of sectors of the economy, instrument or aggregate assets portfolio structures are not available. This section presents our proposal for filtering the portfolio structure from the market-value stock dynamics, flows and the dynamics of a set of indices representative of the potential issuers for the specific instrument. In Subsection 4.4.1, the proposal is discussed by assuming that the *other volume changes* are given. Then, in Subsection 4.4.2, this assumption is removed and a three-step procedure is proposed for updating or estimating *other volume changes*.

4.4.1 Model:

Let us start by considering the *securities other than shares* held by the *rest-of-the-world* sector ($i = 7$) in the whole Euro Area (EA). Studying the *rest-of-the-world* is simpler than other sectors because the *rest-of-the-world* sector can hold in Euro area either EA corporation securities[3] or

[3]For sick of simplicity, financial and non financial corporations are clustered together in the category *corporations*. In particular, *corporations* is defined as the

EA government securities. Consequently, the latent variable representing the portfolio structure is univariate, because of the sum-to-one constraint.

Defining the variables

The sector price $\overset{\bullet}{p}{}^{7}_{t-1,t}$ is the representative price index of the changes in prices of holding of the *rest-of-the-world* sector between $t-1$ and t. As said, it depends on the unknown portfolio structure and it can be derived as a combination of the price indices, each representative of the price dynamics of an issuer. Let us denote by $\overset{\bullet}{p}{}^{C}_{t-1,t}$ and $\overset{\bullet}{p}{}^{G}_{t-1,t}$ the revaluation indices representing revaluation dynamics for corporation securities and revaluation dynamics for EA government securities respectively. Let us call $\underset{\sim}{\overset{\bullet}{p}}_{t-1,t} = \left(\overset{\bullet}{p}{}^{C}_{t-1,t}, \overset{\bullet}{p}{}^{G}_{t-1,t}\right)$; $\underset{\sim}{w}_{7,t-1} = \left(w^{C}_{7,t-1}, w^{G}_{7,t-1}\right)$ the unknown portfolio structure at time t-1 and $S_{7,t-1}$ the stock of securities for the *rest-of-the-world* in EA securities at time $t-1$ for the *rest-of-the-world* sector. Conditionally on knowing $w_{7,t-1}$, $\left(100 * w^{C}_{7,t-1}\right)\%$ of the total stock $S_{7,t-1}$ should be revaluated using $\overset{\bullet}{p}{}^{C}_{t-1,t}$ and $\left(100 * w^{G}_{7,t-1}\right)\%$ of the total stock $S_{7,t-1}$ using $\overset{\bullet}{p}{}^{G}_{t-1,t}$. Therefore, the economic value of $S_{7,t-1}$ one-quarter ahead becomes $S_{7,t-1}\left(1+\beta^{S}_{\mathrm{p};7}\,\underset{\sim}{\overset{\bullet}{p}}_{t-1,t}\,\underset{\sim}{w}'_{7,t-1}\right)$. Within the quarter, some flows, $F_{7,t} = T_{7,t} + O_{7,t}$, on the assets side of the balance sheet of the *rest-of-the-world* are usually observed. Their composition is different from the stock composition and indeed it should reflect the corresponding changes in the composition of stock by issuers. These flows can also be issued by either EA corporations or EA governments. Let us call their respective proportions $\left(\tilde{w}{}^{C}_{7,t}, \tilde{w}{}^{G}_{7,t}\right) = \underset{\sim}{\tilde{w}}_{7,t}$. Saying that the element $\tilde{w}{}^{C}_{7,t}$ should reflect the change of corporation securities between two consecutive quarters means that:

$$\tilde{w}{}^{C}_{7,t} \propto \left| w^{C}_{7,t}S_{7,t} - w^{C}_{7,t-1}S_{7,t-1}\left(1+\beta^{S}_{\mathrm{p};7}\overset{\bullet}{p}{}^{C}_{t-1,t}w^{C}_{7,t-1}\right)\right|$$

sum of *non-financial corporations*, *monetary financial institutions*, *other financial intermediaries different from insurance corporations and pension funds* and *insurance corporations and pension funds*.

where $\propto$ indicates proportionality. Similarly, for the element $\widetilde{w}^{G}_{7,t}$, it holds that

$$\widetilde{w}^{G}_{7,t} \propto | \, w^{G}_{7,t} S_{7,t} - w^{G}_{7,t-1} S_{7,t-1} \left(1 + \beta^{S}_{\mathrm{p};7} \overset{\bullet}{p}{}^{G}_{t-1,t} w^{G}_{7,t-1} \right) |$$

Moreover, $\widetilde{w}^{C}_{7,t}$ and $\widetilde{w}^{G}_{7,t}$ have to sum to one. Let us now assume the following transformations, let us say *flows transformations*, for the sector *rest-of-the-world*:

$$\widetilde{w}^{C}_{7,t} = \frac{|w^{C}_{7,t} S_{7,t} - w^{C}_{7,t-1} S_{7,t-1} \left(1 + \beta^{S}_{\mathrm{p};7} \overset{\bullet}{p}{}^{C}_{t-1,t} \right) |}{|S_{7,t} - S_{7,t-1} \left(1 + \beta^{S}_{\mathrm{p};7} \, \underline{\overset{\bullet}{p}}_{t-1,t} \, \underline{w}'_{7,t-1} \right) |};$$

$$\widetilde{w}^{G}_{7,t} = \frac{|w^{G}_{7,t} S_{7,t} - w^{G}_{7,t-1} S_{7,t-1} \left(1 + \beta^{S}_{\mathrm{p};7} \overset{\bullet}{p}{}^{G}_{t-1,t} \right) |}{|S_{7,t} - S_{7,t-1} \left(1 + \beta^{S}_{\mathrm{p};7} \, \underline{\overset{\bullet}{p}}_{t-1,t} \, \underline{w}'_{7,t-1} \right) |}.$$

Notice that the total-quarter flows can be the result of positive and negative within-quarter flows. If the sign of the flows is the same as the signs of

$$\left(w^{issuer}_{7,t} S_{7,t} - w^{issuer}_{7,t-1} S_{7,t-1} \left(1 + \beta^{S}_{\mathrm{p};7} \overset{\bullet}{p}{}^{issuer}_{t-1,t} \right) \right) \quad \text{with } issuer = \{G, C\},$$

then the *flows transformations* are true by constructions. If the signs are different, it instead means that compensative effects between positive and negative flows happen within the quarter and we assume that the *flows transformations* holds.

The flows need to be revaluated too. The assumption that all the within-quarter flows are concentrated at a single time is maintained but the imposition that this single time coincides with the mid-quarter point

is avoided by allowing a different parameter for the revaluation of the flows[4]. Consequently, the economic value of the flows, $(T_{7,t} + O_{7,t})$, at the end of the quarter is:

$$(T_{t;7} + O_{t;7}) \left(1 + \beta^F_{\mathrm{p};7} \overset{\bullet}{\underline{p}}_{t-1,t} \underline{\widetilde{w}}'_{7,t}\right)$$

The model for the economic value of the securities for the *rest-of-the-world* can then be expressed as follows:

$$\boldsymbol{S}_{7,t} = \boldsymbol{S}_{7,t-1} \left(1 + \beta^S_{\mathrm{p};i} \overset{\bullet}{\underline{p}}_{t-1,t} \underline{w}'_{7,t-1}\right) + (T_{7,t} + O_{7,t}) \left(1 + \beta^F_{\mathrm{p};i} \overset{\bullet}{\underline{p}}_{t-1,t} \underline{\widetilde{w}}'_{7,t}\right) + \boldsymbol{u}_{7,t} \tag{4.4}$$

where $\boldsymbol{u}_{7,t}$ is a Gaussian white noise, i.e. $\boldsymbol{u}_{7,t} \overset{\text{iid}}{\sim} N\left(0, \sigma^2_{obs,7}\right)$. Equation (4.4) describes the dynamics of the economic value of securities as a function of some unknown parameters, $\left(\beta^S_{\mathrm{p};i}, \beta^F_{\mathrm{p};i}\right)$, and some other latent variables, $\underline{w}_{7,\tau}$ and $\underline{\widetilde{w}}_{7,\tau}$ with $\tau = 1, .., t$, representing the portfolio composition of the *rest-of-the-world* and evolving over time. Let us assume that the processes $\left(\underline{w}_{7,t}\right)$ and $\left(\underline{\widetilde{w}}_{7,t}\right)$ are Markovian processes and that the unknown parameters are random variables.

In order to impose the sum-one and the 0-1 bound constraints for each element of the vector representing the portfolio structure for each quarter, we introduce a latent, auxiliary, **U**nconstrained random variable, say $\boldsymbol{u}^U_{7,t}$. Then, we transform it through a logistic or inverse-logit function. Let us assume that this unconstrained vector has the simplest dynamics with the Markovian property, i.e. a random walk $\boldsymbol{w}^U_{7,t} = \boldsymbol{w}^U_{7,t-1} + \boldsymbol{v}_{7,t}$ where $\boldsymbol{w}^U_{7,t}$ and $\boldsymbol{v}_{7,t}$ are vectors of dimension equal to the number of issuer sectors minus one, e.g. 2-1=1 for the *rest-of-the-world*. Notice that the random walk assumption imposes sharp and persistence changes in their temporal evolution, which seems to be a reasonable assumption for modelling the evolution of the portfolio structure avoiding that all the

[4]That is to say, we are still assuming that the revaluation index of the flows is a certain percentage of the stock revaluation index, but we don't impose what percentage should be.

big variabilities are imputed to OVCs. However, the assumption could be easily relaxed. Moreover, $\boldsymbol{v}_{7,t}$ is a Gaussian white noise, i.e uncorrelated over time, and is independent from $\boldsymbol{u}_{7,t}$. Then, and conditionally on $\boldsymbol{w}^U_{7,t}$, we can obtain the portfolio composition as follows:

$$\boldsymbol{w}^C_{7,t} = \frac{\exp\left(\boldsymbol{w}^U_{7,t}\right)}{1+\exp\left(\boldsymbol{w}^U_{7,t}\right)} \quad \text{and} \quad \boldsymbol{w}^G_{7,t} = \frac{1}{1+\exp\left(\boldsymbol{w}^U_{7,t}\right)}.$$

Defining the model

The above considerations for the *rest-of-the-world* ($i = 7$) can be summarized in a two-equations system. However, similar considerations can be carried out for all the other sectors, just changing the *flows transformations*. Hence, for all the sectors, $i = 1, .., 7$, we can assume the following conditional state-space model:

$$\begin{cases} \boldsymbol{S}_{i,t} = \boldsymbol{S}_{i,t-1}\left(1+\boldsymbol{\beta}^S_{\mathrm{p};i}\underset{\bullet}{\underline{p}}_{t-1,t}\underline{\boldsymbol{w}}_{i,t-1}\right) + (T_{i,t}+O_{i,t})\left(1+\boldsymbol{\beta}^F_{\mathrm{p};i}\underset{\bullet}{\underline{p}}_{t-1,t}\underline{\widetilde{\boldsymbol{w}}}_{i,t}\right) + \boldsymbol{u}_{i,t} \\ \underline{\boldsymbol{w}}^U_{i,t} = \underline{\boldsymbol{w}}^U_{i,t-1} + \underline{\boldsymbol{v}}_{i,t} \end{cases} \tag{4.5}$$

The state equation, i.e. $\underline{\boldsymbol{w}}^U_{i,t} = \underline{\boldsymbol{w}}^U_{i,t-1} + \underline{\boldsymbol{v}}_{i,t}$, is written with vectorial notation for homogeneity with the other sectors, but for the rest-of-the-world all the involved vector are indeed random variables. In the observational equation, we write $\underline{w}_{i,t-1}$ to denote the logistic transformation of $w^U_{i,t-1}$ and we write $\underline{\widetilde{w}}_{i,t}$ to denote the *flows transformations*. In particular, the involved variables are defined as follows.

- For the *rest-of-the-world* sector ($i = 7$), $\underset{\bullet}{\underline{p}}_{t-1,t} = \left(\overset{\bullet}{p}^C_{t-1,t}, \overset{\bullet}{p}^G_{t-1,t}\right)$ and the following relationships are defined:

$$\boldsymbol{w}^C_{7,t} = \frac{\exp\left(\boldsymbol{w}^U_{7,t}\right)}{1+\exp\left(\boldsymbol{w}^U_{7,t}\right)} \text{ and } \boldsymbol{w}^G_{7,t} = \frac{1}{1+\exp\left(\boldsymbol{w}^U_{7,t}\right)}$$

$$\underset{\sim}{\boldsymbol{w}}^{C}_{7,t} = \frac{|\boldsymbol{w}^{C}_{7,t}\boldsymbol{S}_{7,t} - \boldsymbol{w}^{C}_{7,t-1}\boldsymbol{S}_{7,t-1}\left(1+\boldsymbol{\beta}^{S}_{\mathrm{p};i}\overset{\bullet}{p}^{C}_{t-1,t}\right)|}{|\boldsymbol{S}_{7,t} - \boldsymbol{S}_{7,t-1}\left(1+\boldsymbol{\beta}^{S}_{\mathrm{p};i}(\boldsymbol{w}^{C}_{7,t-1}\overset{\bullet}{p}^{C}_{t-1,t} + \boldsymbol{w}^{G}_{7,t-1}\overset{\bullet}{p}^{G}_{t-1,t})\right)|}$$

and

$$\underset{\sim}{\boldsymbol{w}}^{G}_{7,t} = \frac{|\boldsymbol{w}^{G}_{7,t}S_{7,t} - \boldsymbol{w}^{G}_{7,t-1}S_{7,t-1}\left(1+\boldsymbol{\beta}^{S}_{\mathrm{p};i}\overset{\bullet}{p}^{C}_{t-1,t}\right)|}{|\boldsymbol{S}_{7,t} - \boldsymbol{S}_{7,t-1}\left(1+\boldsymbol{\beta}^{S}_{\mathrm{p};i}(w^{C}_{7,t-1}\overset{\bullet}{p}^{C}_{t-1,t} + \boldsymbol{w}^{G}_{7,t-1}\overset{\bullet}{p}^{G}_{t-1,t})\right)|}$$

- For all the other sectors (i.e. $i \neq 7$), it is assumed that there are three potential issuers: not only EA corporations and EA government, but they have also the possibility of investing on assets issued by *rest-of-the-world*. Consequently, $\underset{\bullet}{p}_{t-1,t} = \left(\overset{\bullet}{p}^{C}_{t-1,t}, \overset{\bullet}{p}^{G}_{t-1,t}, \overset{\bullet}{p}^{RoW}_{t-1,t}\right)$ and, for each $i = 1, .., 6$, the following relationships are defined:

$$\boldsymbol{w}^{C}_{i,t} = \frac{\exp\left(\boldsymbol{w}^{U\ C}_{i,t}\right)}{1+\exp\left(\boldsymbol{w}^{U\ C}_{i,t}\right)+\exp\left(\boldsymbol{w}^{U\ G}_{i,t}\right)}; \quad \boldsymbol{w}^{G}_{i,t} = \frac{\exp\left(\boldsymbol{w}^{U\ G}_{i,t}\right)}{1+\exp\left(\boldsymbol{w}^{U\ C}_{i,t}\right)+\exp\left(\boldsymbol{w}^{U\ G}_{i,t}\right)}$$

$$\text{and} \quad \boldsymbol{w}^{RoW}_{i,t} = \frac{1}{1+\exp\left(\boldsymbol{w}^{U\ C}_{i,t}\right)+\exp\left(\boldsymbol{w}^{U\ G}_{i,t}\right)}$$

$$\underset{\sim}{\boldsymbol{w}}^{C}_{i,t} = \frac{|\boldsymbol{w}^{C}_{i,t}S_{i,t} - \boldsymbol{w}^{C}_{i,t-1}S_{i,t-1}\left(1+\boldsymbol{\beta}^{S}_{\mathrm{p};i}\overset{\bullet}{p}^{C}_{t-1,t}\right)|}{|\boldsymbol{S}_{i,t} - \boldsymbol{S}_{i,t-1}\left(1+\boldsymbol{\beta}^{S}_{\mathrm{p};i}(\boldsymbol{w}^{C}_{i,t-1}\overset{\bullet}{p}^{C}_{t-1,t} + \boldsymbol{w}^{G}_{i,t-1}\overset{\bullet}{p}^{G}_{t-1,t} + \boldsymbol{w}^{RoW}_{i,t-1}\overset{\bullet}{p}^{RoW}_{t-1,t})\right)|}$$

$$\underset{\sim}{\boldsymbol{w}}_{i,t}^{G} = \frac{|\boldsymbol{w}_{i,t}^{G}\boldsymbol{S}_{i,t} - \boldsymbol{w}_{i,t-1}^{G}\boldsymbol{S}_{i,t-1}\left(1+\boldsymbol{\beta}_{\mathrm{p};i}^{S}\overset{\bullet}{p}_{t-1,t}^{C}\right)|}{|\boldsymbol{S}_{i,t} - \boldsymbol{S}_{i,t-1}\left(1+\boldsymbol{\beta}_{\mathrm{p};i}^{S}(\boldsymbol{w}_{i,t-1}^{C}\overset{\bullet}{p}_{t-1,t}^{C} + \boldsymbol{w}_{i,t-1}^{G}\overset{\bullet}{p}_{t-1,t}^{G} + \boldsymbol{w}_{i,t-1}^{RoW}\overset{\bullet}{p}_{t-1,t}^{RoW})\right)|}$$

$$\underset{\sim}{\boldsymbol{w}}_{i,t}^{RoW} = \frac{|\boldsymbol{w}_{i,t}^{RoW}\boldsymbol{S}_{i,t} - \boldsymbol{w}_{i,t-1}^{RoW}\boldsymbol{S}_{i,t-1}\left(1+\boldsymbol{\beta}_{\mathrm{p};i}^{S}\overset{\bullet}{p}_{t-1,t}^{RoW}\right)|}{|\boldsymbol{S}_{i,t} - \boldsymbol{S}_{i,t-1}\left(1+\boldsymbol{\beta}_{\mathrm{p};i}^{S}(\boldsymbol{w}_{i,t-1}^{C}\overset{\bullet}{p}_{t-1,t}^{C} + \boldsymbol{w}_{i,t-1}^{G}\overset{\bullet}{p}_{t-1,t}^{G} + \boldsymbol{w}_{i,t-1}^{RoW}\overset{\bullet}{p}_{t-1,t}^{RoW})\right)|}$$

Call $\underline{\boldsymbol{w}}_{i,t} = \left(\boldsymbol{w}_{i,t}^{C}, \boldsymbol{w}_{i,t}^{G}, \boldsymbol{w}_{i,t}^{RoW}\right)$, $\underline{\boldsymbol{w}}_{i,t}^{U} = \left(\boldsymbol{w}_{i,t}^{C\,U}, \boldsymbol{w}_{i,t}^{G\,U}\right)$, $\underset{\sim}{\underline{\boldsymbol{w}}}_{i,t} = \left(\underset{\sim}{\boldsymbol{w}}_{i,t}^{C}, \underset{\sim}{\boldsymbol{w}}_{i,t}^{G}, \underset{\sim}{\boldsymbol{w}}_{i,t}^{RoW}\right)$.

We assume that, for each i, $\underline{\boldsymbol{v}}_{i,t} \mid \boldsymbol{\Sigma}_{0;i} \overset{\text{iid}}{\sim} N(0, \boldsymbol{\Sigma}_{0;i})$, and $\boldsymbol{u}_{i,t} \mid \boldsymbol{\sigma}_{obs,i}^{2} \overset{\text{iid}}{\sim} N\left(0, \boldsymbol{\sigma}_{obs,i}^{2}\right)$, both independent across t. The processes $\left\{\underline{\boldsymbol{v}}_{i,s}\right\}_{s}$ and $\left\{\boldsymbol{u}_{i,t}\right\}_{t}$ are independent each other and they are also independent among sectors. Notice that $\boldsymbol{w}_{i,t}^{C\,U}$ and $\boldsymbol{w}_{i,t}^{G\,U}$ are unobservable and auxiliary variables, not associated with a single element of the portfolio composition, but used together to define all the three $\boldsymbol{w}_{i,t}^{C}, \boldsymbol{w}_{i,t}^{G}$, and $\boldsymbol{w}_{i,t}^{RoW}$.

In Section 1.4, we have discussed state-space models. The model (4.5) is not state-space model, because the observations are not conditional independent, conditionally on the latent process. However, they are conditionally independent given their most recent value. The estimation procedure will be discussed in Section 4.5 and further details are provided in the appendix.

4.4.2 Method for estimating OVCs: 3-step procedure

The starting point of the previous subsection is that *other volume changes* are given. In particular, *given the other volume changes*, *portfolio composition can be estimated* based on model (4.5). As discussed in Section 4.3, the reverse holds too: conditionally on the portfolio struc-

ture, the series of *other volume changes* can be defined using standard outliers detecting techniques. Unfortunately, the complete series of OVCs will be available for the euro area accounts on the ECB website only after November 2014. Currently, only a preliminary subseries of *other volume changes* is available and only for a few sectors. Therefore, neither the portfolio structure nor the *other volume changes* are now known. The problem can be dealt with in different ways. In principle, if the variances were properly specified or easily inferable, the above model would allow for radical changes in the portfolio structure and it leaves unexplained variability both in the observational and in the state equation. One can then define the big observational residuals as OVCs. Here, the signal-to-noise ratio plays a crucial role. It measures the strength of the portfolio structure signal relative to the observational noise and it cannot be properly estimated over a such short period of time. Therefore, the peaks of variability in the observational errors needs to be controlled for. In practice, *preliminary other volume changes* are necessarily and, depending on them, different portfolio structure will be estimated and then different final estimation of *other volume changes*. Unfortunately, it will not be possible to known if peaks in the noises are peaks of measurement errors -and consequently other volume changes- or if they are really associated with sharp changes in the portfolio composition. In this sense, the problem is intrinsically undetermined. The chosen random walk model for the unobservable process, defining then the portfolio structure, is in favor of sharp changes in the portfolio structure, because we prefer to be conservative[5] in the estimation of the OVCs. The chosen approach in the following 3-step procedure:

1. **Compute preliminary *other volume changes*, when not available**. Several options are available for computing these preliminary *other volume changes*. We opt for the following: for each given in-

[5]The initial OVCs are usually all zeros. Therefore, to be conservative here means to impute only the strictly necessary OVCs.

strument, assume that the portfolio composition of assets is the same as the portfolio composition of the total liabilities[6]. Moreover, if some data are available for the portfolio structure of some sectors for some quarters, they are incorporated. Once this preliminary portfolio structure has been estimated, the model is estimated accordingly to the same procedure described in Section 4.3, computing the preliminary *other volume changes* as part of the big residuals.

2. **Use preliminary other volume changes for obtaining the portfolio structure.** These preliminary other volume changes are used as part of the flows and the dynamics of the economic value of securities is filtered to obtain the estimated portfolio structure. Then, these filtered values are used for computing the residuals and **updating the series of other volume changes.**

3. **Use the updated other volume changes for estimating the portfolio structure.**

Notice that the imputation of only a part of the residual, instead of the whole, is required for maintaining a *good* signal-to-noise ratio.

4.5 Estimation

In this section, we provide general description of the filtering problem whereas the technical details are provided in the appendix.

Let us define $\underline{\boldsymbol{w}}^U_{i,t_0:t} = \{\underline{\boldsymbol{w}}^U_{i,t_0}, ..., \underline{\boldsymbol{w}}^U_{i,t}\}$ and $\boldsymbol{S}_{it_0:t} = \{\boldsymbol{S}_{i,t_0}, ..., \boldsymbol{S}_{i,t}\}$. Call $\underline{\boldsymbol{\psi}}$ the vector collecting the unknown parameters, $\underline{\boldsymbol{\psi}} = \left(\boldsymbol{\beta}^S_{\mathrm{p};i}, \boldsymbol{\beta}^F_{\mathrm{p};i}, \boldsymbol{\sigma}^2_{obs,i}, \boldsymbol{\Sigma}_{0;i}\right)$.

For estimating the state vector (i.e. filtering) and estimating the parameters, the conditional densities $p(\underline{\boldsymbol{w}}^U_{i,t}, \underline{\boldsymbol{\psi}} | \boldsymbol{S}_{i,t_0}, ..., \boldsymbol{S}_{i,t})$ need to be computed. As discussed in Subsection 1.5.3, in filtering applications the data are supposed to arrive sequentially in time and, hence, the estimation procedure for the state vector at time t must be based on the

[6]Notice that this implies the same assumed portfolio structure for all the sectors.

observations up to time t, and it must be further updated as a new data point becomes available at time $t+1$. Due to its neither linear nor gaussian feature, even in the simplest case of known parameters, the posterior distributions for the problem under investigation cannot be obtained analytically. It is however possible to make use of the recursive nature of the filter approach by exploiting the samples generated from $p(\underline{w}^U_{i,0}, ..., \underline{w}^U_{i,t}, \underline{\psi}|S_{i,t_0}, ..., S_{i,t})$ in simulating from:

$$p(\underline{w}^U_{i,0}, ..., \underline{w}^U_{i,t}, \underline{w}^U_{i,t+1}, \underline{\psi}|S_{i,t_0}, ..., S_{i,t}, S_{i,t+1}).$$

This is done using *Sequential Monte Carlo* techniques, discussed in Subsection 1.5.3. See Doucet *et al.* (2001) for a general discussion of Sequential Monte Carlo methods. The idea is to draw, for each quarter, and independently for each sector, a large number of candidate portfolio structures, called particles, from some suitable distribution. Then, a weight is associated with each of these particles, depending on their "likelihood". The corresponding weighted average is the filtered mean, representing the portfolio structure in a specific quarter. This filtering problem is here solved by using the Auxiliary particle filter (Liu and West, 2001), assigning Inverse-Wishart (IW) priors to the variances and Gaussian priors to the revaluation coefficients. Moreover, fixing the parameters does not avoid the implementation of the particle filter algorithm, because even if the parameters were known, there is no analytic expression for the filtering distributions. The algorithm is provided in Appendix.

4.5.1 Simulation study

We now illustrate the performance of the chosen algorithm by simulating data based on the true data for the whole euro area, *long-term securities other than shares*, held by the rest of the world. Given the true sequence of transactions, the true series of government and corporation price indices, let us assume that the ratio of the total liabilities of each sector with respect to the whole economy liabilities as portfolio structure.

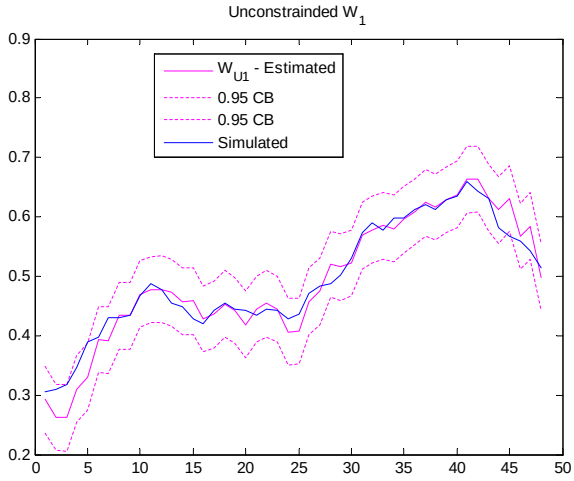


Figure 4.1: Unconstrained portfolio structure. Results from one simulation

Stocks are accordingly simulated starting from 1999q3 and then cumulating revaluated transactions. For the sake of simplicity, let us assume that the *other volume changes* are all zeros -or equivalently that the OVCs are known and recorded together with transactions. Then simulated noises, with variance equals to 100, are added to each quarterly stock. In Figures 4.1 and 4.2, the filtered means with 95 % confidence interval and the simulated series are plotted. It follows that, if the variances are small and the *other volume changes* series is known, the chosen filtering algorithm works well.

4.5.2 Adding (vertical) consistency for the whole economy

As said, the basic idea of the particle filter algorithm is to estimate the portfolio structure using a *weighted* average of **some** *possible* portfolio structures. Up to now, these *possible* portfolio structures have been ob-

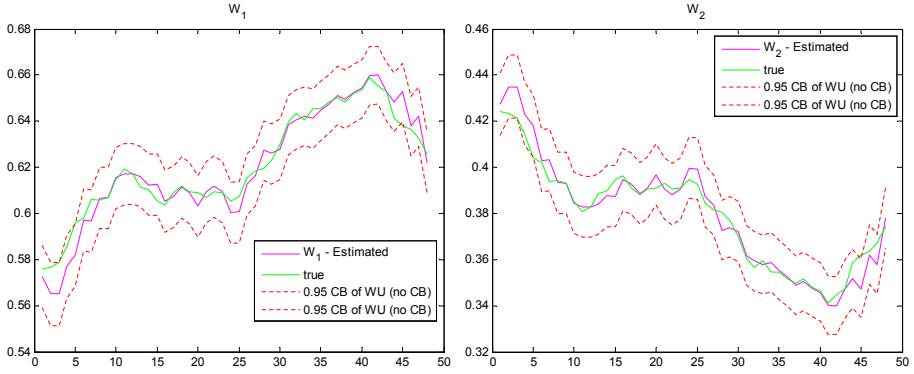


Figure 4.2: Portfolio structure. Results from one simulation.

tained by purely sectoral considerations. Each sector has been analyzed separately and the set of predicted portfolio structures for the next quarter are based only on their most recent value. In this section, we propose a "whole-economy" approach for the estimation procedure by considering all the seven main sectors of the economy together. Our proposal is to impose a resampling scheme when a (vertical) consistency constraint for the whole-economy does not hold. This allows an improvement of the choice of the possible portfolio structures for each of the seven sectors of the economy. This means constraining the supports of the filtered portfolio structures to the ones that satisfy the specific consistency constraint.

Let us call $\boldsymbol{w}_{i,t}^{Government}$ the percentage of assets of a given instrument held by the sector i and issued by the sector *Government*. The sum

$$\sum_{i=1}^{7} \tilde{\boldsymbol{w}}_{i,t}^{Government} \; T_{i,t}$$

represents the total transactions of assets issued by Government and held by the total economy. It is well-known that the total transactions of assets issued by Government must be equal to the total transactions of

liabilities issued by Government, let us say $T_t^{Liability,Government}$. The vertical consistency for government requires that

$$T_t^{Liability,Government} \stackrel{!}{=} \sum_{i=1}^{7} \tilde{\boldsymbol{w}}_{i,t}^{Government} T_{i,t}$$

In order to better select potential portfolio structures, we decide to introduce the following resampling mechanism:

For some fix τ^*, for a set of particles $\tilde{w}_{i,t}^{Government,(s)}$ at iteration s,

if $|T_t^{Liability,Government} - \sum_{i=1}^{7} \tilde{w}_{i,t}^{Government,(s)} T_{i,t}| > \tau^*$, then resample.

The resampling scheme is done for all the seven possible portfolio structures and so that the discrepancy between these two quantities decreases.

Notice that, although the approach imposes a cross-section portfolio composition restriction, we are not imposing the exact balance of the matrix, but just stimulating the balancing when the discrepancy with respect to a balanced matrix would be big. Therefore, the model (4.5) remains the same, but now a consistency constraint enters as a resampling scheme, i.e if a set of particles does not satisfy the consistency, then they are re-sampled. Consequently, this implies that the supports of the seven sector portfolio structures are constrained to not deviate too much from the specific consistency constraint. On one side, this approach has the advantage to stimulate consistency. On the other side, it imposes a deviation from what the data propose as the best option for the next quarter.

4.6 Application to EA data: *Securities other than shares*

In this section, the application of the proposed methodology to *securities other than shares* for the euro area is discussed. Each institutional sector holds some *securities other than shares*. They are issued by three macro aggregates: *corporations* (both financial and non-financial), *government* and *rest-of-the-world*, for all the sectors except for the *rest-of-the-world*. The *rest-of-the-world* can hold *securities other than shares* issued only by *corporations* (both financial and non-financial) and *government*. No preliminary series for the *other volume changes* is provided. The available series for quantities are the stocks and the transactions for the period t_0 =1999q2 till t_1 =2011q3. Hence, the estimation period is 1999q3-2011q3.

The generic vector of price indices is called $\underset{\bullet}{\underline{p}}_{t-1,t} = \left(\overset{\bullet}{p}^{C}_{t-1,t}, \overset{\bullet}{p}^{G}_{t-1,t}, \overset{\bullet}{p}^{RoW}_{t-1,t}\right)$, where:

- $\overset{\bullet}{p}^{C}_{t-1,t}$ is the 5-year euro corporate bond price index, provided by the ECB Statistical Data Warehouse.
- $\overset{\bullet}{p}^{G}_{t-1,t}$ is a weighted average of the 5-year government benchmark bond yield, provided by Reuters for the following countries: AT, BE, DE, ES, FR, IT, NL, PT. The weights are given by the ratio of the government securities issued by a given country over the total government securities issued by sum of them.
- $\overset{\bullet}{p}^{RoW}_{t-1,t}$ is the USA 5-year government benchmark bond yield, provided by the ECB Statistical Data Warehouse and converted from dollar to euro.

The threshold for the definition of the *other volume changes* is defined by a combination of two times the standard deviation of the residuals and a percentage of the stock, which depends on the sector. We fix the

number of iterations, J, at $15,000$. For the sake of simplicity, when implementing the algorithm with consistency resampling, we fix the sectoral parameters, using the estimated values obtained in the sector-by-sector approach discussed in Section 4.4.

In Figure 4.3, we show the fitted portfolio structure for the sector *non financial corporation* (and 95 % confidence interval) with and without consistency resampling. We plot in the same graph the available sub-series of data, denominated SHS. We also display the fitted series of *other volume changes*. In Appendix, the results for the other six sectors are shown in Figures 4.5-4.10.

Notice that the results between the analysis of sector-by-sector and the analysis done with the consistency resampling over the whole economy can be different. If the two results are close (like for the *non financial corporation* sector), this suggests that the results obtained without consistency resampling are already good. That is to say, the obtained estimation considering the single sector satisfy the consistency constraint. If they are different or with sharp changes in the estimation with consistency resampling, i.e. for *insurance corporation and pension funds* sector, this means that the best estimations for the single sector do not respect the vertical constraint over the whole economy.

4.7 Further developments: incorporation of available data

Our focus has been on estimating the whole portfolio structure and we have used the shorter series of available data to make comparison with estimated series, when these shorter series are available. In particular, the estimation sample is for $t = 1999q3, .., 2011q3$ and we have data for the sector i on the portfolio structure, denoted by $w_{i,t}^{DATA}$, for $t = 2009q1, .., 2011q3$. The next step of the analysis is the explicit incorporation of the available short series of portfolio structure in a

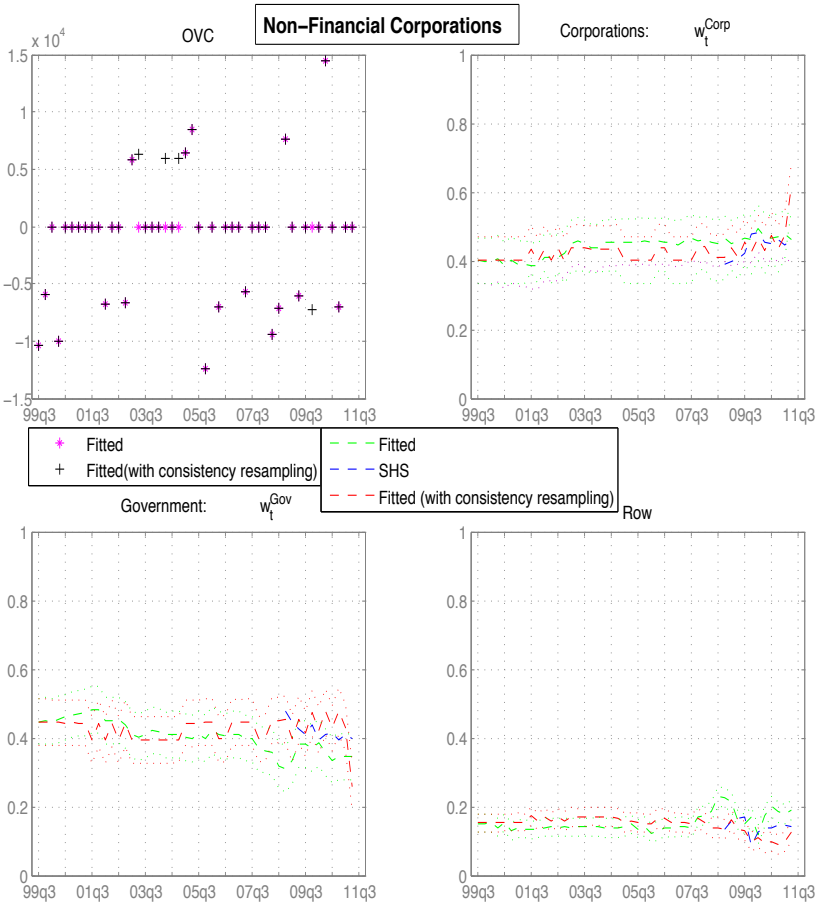


Figure 4.3: Results for holdings of F33 (Securities other than shares) for the institutional sector *Non Financial Corporations* in the euro area: OVC (Other Volume Changes) and the composition of the portfolio decomposed in assets issued by corporations, governments and RoW (Rest-of-the-World): Fitted portfolio structure (and 95 % confidence interval) with and without consistency resampling scheme and sub-series of data available from SHS (Security holdings statistics).

general framework. One way is to include the conditional state-space model within a general decision theory framework[7]. Let us introduce the quadratic loss function:

$$L_\theta\left(\underline{\boldsymbol{w}}_{i,t}, \underline{w}_{i,t}^{DATA}\right) = \left(\underline{\boldsymbol{w}}_{i,t}\left(\theta\right) - \underline{w}_{i,t}^{DATA}\right)^2,$$

where θ are (some of) the parameters of the whole procedure. These parameters can include not only the parameters of the model but also the parameters of the particle filter. The problem can be formulated as follows:

Find the parameters θ so that the conditional aggregate expected loss associated with the deviations from the available data is minimized:

$$\min_{\theta} \sum_{t=2009q1}^{2011q3} E\left(\left(\underline{\hat{\boldsymbol{w}}}_{i,t|t}\left(\theta\right) - \underline{w}_{i,t}^{DATA}\right)^2 \mid \boldsymbol{S}_{i1999q3}, .., \boldsymbol{S}_{i,t}\right)$$

so that $\underline{\hat{\boldsymbol{w}}}_{i,t|t}\left(\theta\right)$ is the filtering mean, using all the data available till time t, based on model (4.5)*, re-written in the following for reader's convenience:*

$$\begin{cases} \boldsymbol{S}_{i,t} = \boldsymbol{S}_{i,t-1}\left(1 + \beta_{\mathrm{p};i}^{S} \overset{\bullet}{\underline{p}}_{t-1,t} \underline{\boldsymbol{w}}_{i,t-1}\right) + \left(T_{i,t} + O_{i,t}\right)\left(1 + \beta_{\mathrm{p};i}^{F} \overset{\bullet}{\underline{p}}_{t-1,t} \widetilde{\underline{\boldsymbol{w}}}_{i,t}\right) + \boldsymbol{u}_{i,t} \\ \boldsymbol{w}_{i,t}^{U} = \boldsymbol{w}_{i,t-1}^{U} + \boldsymbol{v}_{i,t} \end{cases}$$

Choosing a quadratic loss function implies that the corresponding expected loss function can be written in terms of the variance and the expected value of the argument. In any case, under model (4.5), the expected loss function cannot be derived analytically, because both the filtered means are not available in closed form and the logit-Gaussian

[7] See e.g. Parmigiani and Inoue (2009) for a general description on decision theory.

distribution has no analytic form for its moments. Hence, it will be computed numerically.

For the sake of simplicity, let us focus on $i = 7$, namely the *rest-of-the-world*. It would be possible to proceed similarly for the other sectors. However, the other sectors are a bit more complicated because they require the minimization of the aggregate expected loss also with respect to the a 2×2 semi-positive covariance matrix of the unconstrained weights instead of a scalar variance. In Figure 4.4, we show the estimated portfolio structure when the parameters are the ones that minimize the aggregate expected loss of deviating from the available data. This approach

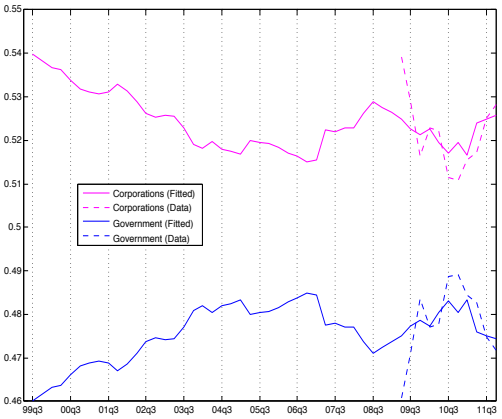


Figure 4.4: F33 - Rest-of-the-world: Portfolio structure using available data

has the advantage of allowing for explicit incorporation of the data, but the disadvantage that the parameters are defined considering only the subset of observations from $2009q1$ to $2011q3$. A better estimation strategy could be to estimate the parameters based both on the posterior estimation means obtained based on drawing values from a continuous importance density over the whole time period and the parameters that solve the aggregate expected loss minimization problem.

4.8 Discussion and conclusions

This chapter proposes a general framework for estimating the dynamics of the portfolio composition of a market-value dynamic assets and for computing the series of *other volume changes*. We have proposed three approaches: sector-by-sector analysis, whole economy analysis and the incorporation of available data. All these three approaches are based on a conditional state-space model and on a three-step procedure. The three-step procedure allows to estimate the portfolio structure and the OVCs by fixing rules that overcome the indeterminacy between OVCs and portfolio structure (i.e. OVCs depend on the portfolio structure and the portfolio structure depends on the OVCs).

In the sector-by-sector analysis, we assume that the parameters and the estimated portfolio structures are independent among sectors. In the whole-economy analysis, we add a consistency resampling scheme in the algorithm that imposes a constraint on the support of the candidate portfolio structures to the set of portfolio structures that satisfy the consistency constraint over the whole economy. The last approach, based on the incorporation of available data, is still under development and its aim is the explicit incorporation of the available sub-series of portfolio structures in the general framework. We propose to choose the parameters of the procedure so that the aggregate posterior expected loss of choosing some filtered series that deviates from the available data is minimized. The filtered series are chosen among all the possible filtered series of portfolio structure coming from the conditional state-space model.

Finally, we have applied our proposed procedure to euro area data of *securities other than shares*. Due to the small sample size and / or to the noisy data, there is high sensitivity to the prior, in particular the parameters for the variances matters. Hence, it is advisable to incorporate all the available partial information on the portfolio structure. Also, one can move towards some multi-sector approach, i.e. not only constraining

the support of the sector portfolio structure but also introducing some interdependence structure among sectors.

Obviously, the worst quarters for the estimation of the portfolio structure are often associated with higher correlated price indices. Although this approach exploits the dynamics of the sectoral price indices, i.e. each issuer is identified through a different price index, the OVCs are robust in this respect. In principle, if they are equal, there is no way to estimate portfolio structure. But the partition of the economic value between them does not matter for computing the residuals and, consequently, the OVCs.

4.9 Appendix

4.9.1 Algorithm: Auxiliary Particle Filter

We assume that a set of J state samples (particles) and the corresponding weights (weights of the particles) can represent the filtering distribution $p(w^U_{i,t-1}|S_{i,1:t-1})$ at time $t-1$. The procedure proceeds updating the particle set to represent the filtering distribution $p(w^U_{i,t}|S_{i,1:t})$ for current time t according to the following iterations. The vectors are not underlined for avoiding heavy notation. Moreover, again for simplicity of notation, $w^U_{1;i,t}$ stands for $w^{C^U}_{i,t}$ and $w^U_{2;i,t}$ stands for $w^{G^U}_{i,t}$. Let us call J the total number of iterations.

Algorithm 4.9.1 *For each sector i,*

- *Initialize:*

  Draw $(w^{U(1)}_{1;i,0}, w^{U(1)}_{2;i,0}), ..., (w^{U(J)}_{1;i,0}, w^{U(J)}_{2;i,0})$ *independently from* $\pi\left(w^U_{i,0}\right)$.

  Set $w^{PF,(j)}_{i,0} = J^{-1}$ *j=1,...,J.*

  Draw $\psi^{(1)}, ..., \psi^{(J)}$ *independently from* $N\left(\hat{\psi}, \hat{\Sigma}\right)$

Compute $\hat{\pi}_0 = \sum_{j=1}^{J} w_{i,0}^{PF,(j)} \delta_{(w_0^{U(j)})}$

- *For t=1,...,T*

  Compute $\bar{\psi} = E_{\hat{\pi}_{i,t-1}}(\psi)$ *and* $\Sigma = Var_{\hat{\pi}_{i,t-1}}(\psi)$

  For j=1,...,J

  Set $m^{(j)} = a\psi^{(j)} + (1-a)\bar{\psi}$ *and*

  $\hat{w}_t^{(j)} = E\left(w_{i,t} \mid w_{i,t-1} = w_{i,t-1}^{(j)}, \psi = m^{(j)}\right) = w_{i,t-1}^{(j)}$.

  Draw a classification variable, I_j*, with probability:*

$$P(I_j = r) \propto w_{i,t-1}^{PF,(r)} f\left(S_{i,t} | \hat{w}_{i,t}^{U(r)}, \psi = m^{(j)}\right)$$

  where $\hat{w}_{i,t}^{U(r)}, \hat{S}_{i,t}^{(r)}$ *are generally central values (we choose the mean) of* $p\left(w_{i,t}^{U} | w_{i,t-1}^{U} = w_{i,t-1}^{U(r)}, \psi = m^{(j)}\right)$

  Draw $\psi^{(j)}$ *from* $N\left(m^{(I_j)}, h^2\Sigma\right)$

  Draw $w_{i,t}^{U(j)}$ *(and then transform* $w_{i,t}^{U}$ *in* $w_{i,t}$ *using the specified logistic function) from:*

$$p\left(w_{i,t}^{U} | w_{i,t-1}^{U} = w_{i,t-1}^{U(I_j)}, \psi = \psi^{(I_j)}\right)$$

  Set $\tilde{w}_{i,t}^{PF,(j)} = \dfrac{f\left(S_{i,t} | w_{i,t}^{U(j)}, \psi = \psi^{(j)}\right)}{f\left(S_{i,t} | \hat{w}_{i,t}^{U(j)}, , \psi = m^{(j)}\right)}$ [8]

- *Normalize the weights:*

$$w_{i,t}^{PF(j)} = \frac{\tilde{w}_{i,t}^{PF(j)}}{\sum_{k=1}^{J} \tilde{w}_{i,t}^{PF(k)}}$$

[8] If $f\left(S_{i,t} | \hat{w}_{i,t}^{U(j)}, \hat{S}_{i,t}^{(j)}, \psi = m^{(j)}\right) = 0$, it means that means of current states given the previous-quarter have 0 likelihood, hence every new value is preferred (or at least not worst) than those. If $f\left(S_{i,t} | \hat{w}_{i,t}^{U(j)}, , \psi = m^{(j)}\right) = 0$, we assume that $w_{i,t}^{PF,(j)} = \frac{1}{J}$

- *Compute* $N_{eff,i} = \left(\sum_{k=1}^{J} \left(w_{i,t}^{PF(k)} \right)^2 \right)^{-1}$ *(e.g. 0.6 * J)*
- *If* $N_{eff,i} < N_0$*, resample: draw a sample of size J from a discrete distribution so that*

$$P\left((w_{U,i0:t}) = (w_{U;i,0:t}^{(j)}) \right) = w_{i,t}^{PF,(j)}$$

for j=1,..,J and relabel this sample $(w_{i,0:t}^{U(1)}), ..., (w_{i,0:t}^{U(J)})$*. Reset the weights* $w_{i,t}^{PF,(j)} = J^{-1}$ *j=1,...,J.*

- *Set* $P\left(w_{i,0:t}, \psi | S_{i,1:t}\right) = \sum_{j=1}^{J} w_{i,t}^{PF(j} \delta_{w_{i,0:t}^{(j)}, \psi^{(j)}}$

4.9.2 Algorithm with (vertical) consistency: whole economy

Algorithm 4.9.2 *1. Initialize the variables for all the seven sectors, i.e. for* $i = 1, .., 7$:

Draw $(w_{1;i,0}^{U(1)}, w_{2;i,0}^{U(1)}), ..., (w_{1;i,0}^{U(J)}, w_{2;i,0}^{U(J)})$ *independently from* $\pi\left(w_{i,0}^{U}\right)$.

Set $w_{i,0}^{PF,(j)} = J^{-1}$ *j=1,...,J*

Draw $\psi_i^{(1)}, ..., \psi_i^{(J)}$ *independently from* $N\left(\hat{\psi}_i, \hat{\Sigma}_i\right)$ *for i=1,..,7.*

Compute $\hat{\pi}_{i,0} = \sum_{j=1}^{J} w_{i,0}^{PF,(j)} \delta_{\left(w_{i,0}^{U(j)}\right)}$

2. For $t = 1, ..., T$ *and* $i = 1, .., 7$*:*

Compute $\bar{\psi}_i = E_{\hat{\pi}_{i,t-1}}(\psi_i)$ *and* $\Sigma_i = Var_{\hat{\pi}_{i,t-1}}(\psi_i)$

For $j = 1, ..., J$*, for* $i = 1, .., 7$ *::*

Set $m_r^{(j)} = a\psi_r^{(j)} + (1-a)\bar{\psi}_i$ *and*

$\hat{w}_{i,t}^{(j)} = E\left(w_{i,t} | w_{i,t-1} = w_{i,t-1}^{(j)}, \psi_i = m_i^{(j)}\right) = w_{i,t-1}^{(j)}$.

Draw a classification variable, $I_{j,i}$*, with probability* $P(I_{j,i} = r) \propto w_{i,t-1}^{PF,(r)} f\left(S_{i,t} | \hat{w}_{i,t}^{U(r)}, \psi = m^{(j)}\right)$ *where* $\hat{w}_{i,t}^{U(r)}, \hat{S}_{i,t}^{(r)}$ *are generally central values (we choose the mean) of* $p\left(w_{i,t}^{U} | w_{i,t-1}^{U} = w_{i,t-1}^{U(r)}, \psi_i = m^{(j)}\right)$

Draw $\psi_i^{(j)}$ *from* $N\left(m_i^{(I_{j,i})}, h^2\Sigma_i\right)$

Draw $w_{i,t}^{U(j)}$ *(and then transform* $w_{i,t}^{U}$ *in* $w_{i,t}$ *using the specified logistic function) from*

$$p\left(w_{i,t}^{U} | w_{i,t-1}^{U} = w_{i,t}^{U(I_{j,i})}, \psi_i = \psi_i^{(I_{j,i})}\right)$$

$$Set\ \tilde{w}_{i,t}^{PF,(j)} = \frac{f\left(S_{i,t}|w_{i,t}^{U(j)}, \psi_i = \psi_i^{(j)}\right)}{f\left(S_{i,t}|\hat{w}_{i,t}^{U(j)},, \psi_i = m^{(j)}\right)}$$ [9]

3. ***Check vertical consistency and possible resampling:***

   - Check consistency: *We say that the consistency is satisfied for the aim of the resampling if*

     $$|T_t^{Liability,Government} - \sum_{i=1}^{7} \tilde{w}_{i,t}^{Government} T_{i,t}| > T_t^{Liability,Government}$$

     If consistency is satisfied, go to the normalization step, point (4) . If consistency is not satisfied, do the consistency resampling .

   - Consistency resampling: *If consistency is not satisfied, then sample a new value for the percentage of assets issued by Government for that sector. The new sample depends on the sign of the transactions of that sector and on the sign of the discrepancy. Call* sign of the discrepancy *the sign of* $T_t^{Liability,Government} - \sum_{i=1}^{7} \tilde{w}_{i,t}^{Government} T_{i,t}$.

     If, for a given sector, the sign of the transactions is the same as the sign of the discrepancy, then sample a new value for the percentage of assets issued by Government for that sector as follows: perturb the old particle $w_{i,t}^{U,Government}$*, and call the new particle* $w_{i,t}^{U,Government,**}$.

     If the two sign are the same, this perturbation is summed to the old particle and is assumed to be a small percentage of the prod-

[9] If $f\left(S_{i,t}|\hat{w}_{i,t}^{U(j)}, \hat{S}_{i,t}^{(j)}, \psi_i = m_i^{(j)}\right) = 0$, it means that means of current states given the previous-quarter have 0 likelihood, hence every new value is preferred (or at least not worst) than those. If $f\left(S_{i,t}|\hat{w}_{i,t}^{U(j)},, \psi_i = m_i^{(j)}\right) = 0$, we assume that $w_{i,t}^{PF,(j)} = \frac{1}{J}$.

uct of this common sign and a realization from a log-normal distribution with parameters $\left(0, \sqrt{2\varsigma w_{i,t}^{Government}}\right)$ *where* ς *is fixed.*

If the sign are different, this perturbation is subtracted.

4. *Normalize the weights for* $i = i, .., 7$:

$$w_{i,t}^{PF(j)} = \frac{\tilde{w}_{i,t}^{PF(j)}}{\sum_{j=1}^{J} \tilde{w}_{i,t}^{PF(j)}}$$

5. *Compute* $N_{eff,i} = \left(\sum_{j=1}^{J} \left(w_{i,t}^{PF(j)}\right)^2\right)^{-1}$ *for i=1,..,7 (e.g. 0.6 * J)*

6. *If, for some i,* $N_{eff,i} < N_{0,i}$*, then resample the particles related to this i: draw a sample of size J from a discrete distribution so that* $P\left((w_{U,i0:t}) = (w_{U;i,0:t}^{(j)})\right) = w_{i,t}^{PF,(j)}$ *for j=1,..,J and relabel this sample* $(w_{i,0:t}^{U(1)}), ..., (w_{i,0:t}^{U(J)})$*. Reset the weights* $w_{i,t}^{PF,(j)} = J^{-1}$ *j=1,...,J.*

7. *Set* $P\left(w_{i,0:t}, \psi | S_{i,1:t}\right) = \sum_{j=1}^{J} w_{i,t}^{PF(j)} \delta_{w_{i,0:t}^{(j)}, \psi_i^{(j)}}$ *for* $i = 1, .., 7$.

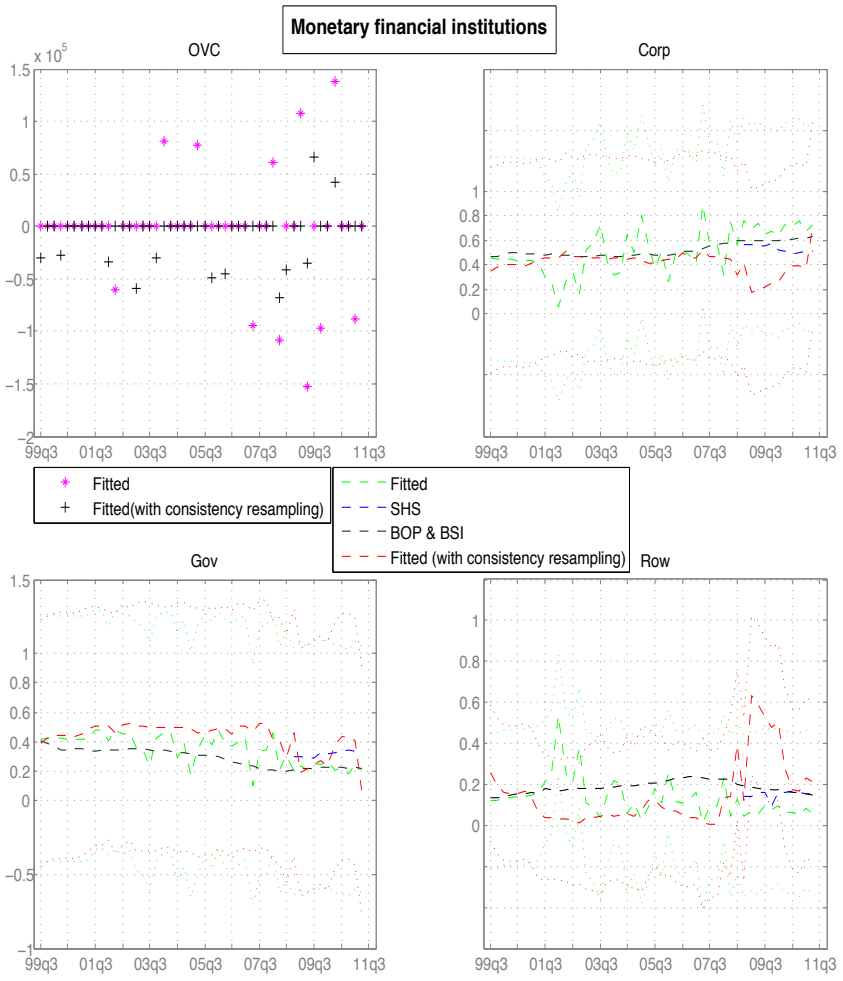


Figure 4.5: Results for holdings of F33 (Securities other than shares) for the institutional sector *Monetary and Financial Institutions* in the euro area: OVC (Other Volume Changes) and the composition of the portfolio decomposed in assets issued by corporations, governments and RoW (Rest-of-the-World): Fitted portfolio structure (and 95 % confidence interval) with and without consistency resampling scheme and sub-series of data available from SHS (Security holdings statistics).

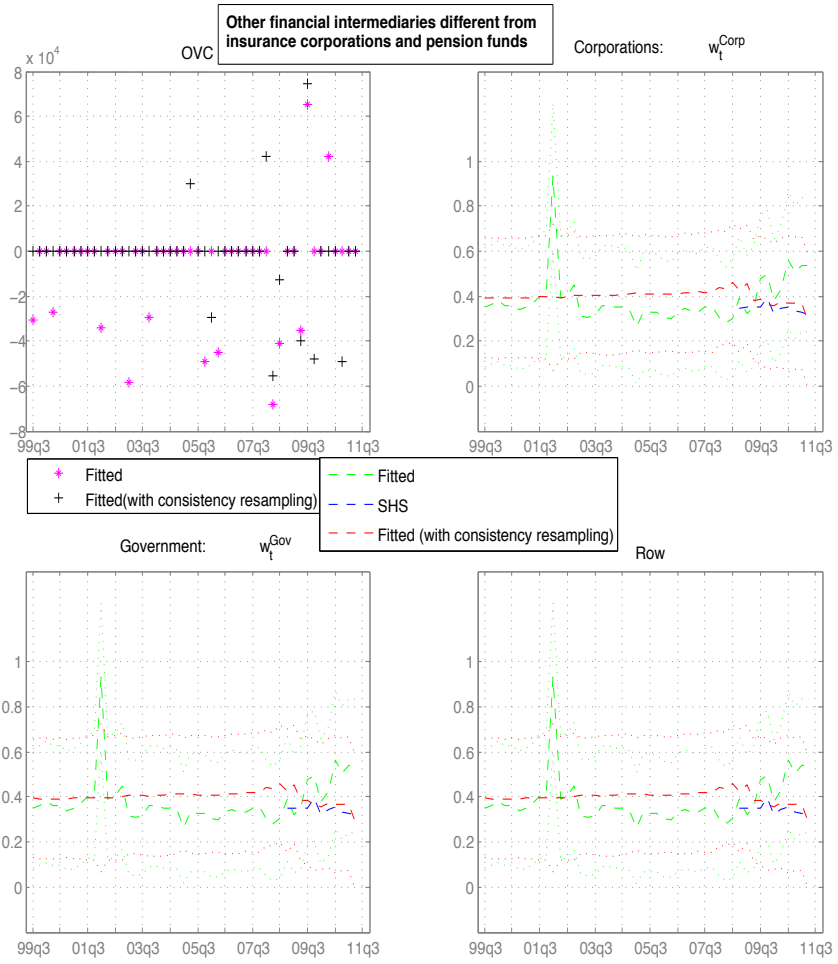


Figure 4.6: Results for holdings of F33 (Securities other than shares) for the institutional sector *Other Financial Intermediaries* in the euro area: OVC (Other Volume Changes) and the composition of the portfolio decomposed in assets issued by corporations, governments and RoW (Rest-of-the-World): Fitted portfolio structure (and 95 % confidence interval) with and without consistency resampling scheme and sub-series of data available from SHS (Security holdings statistics).

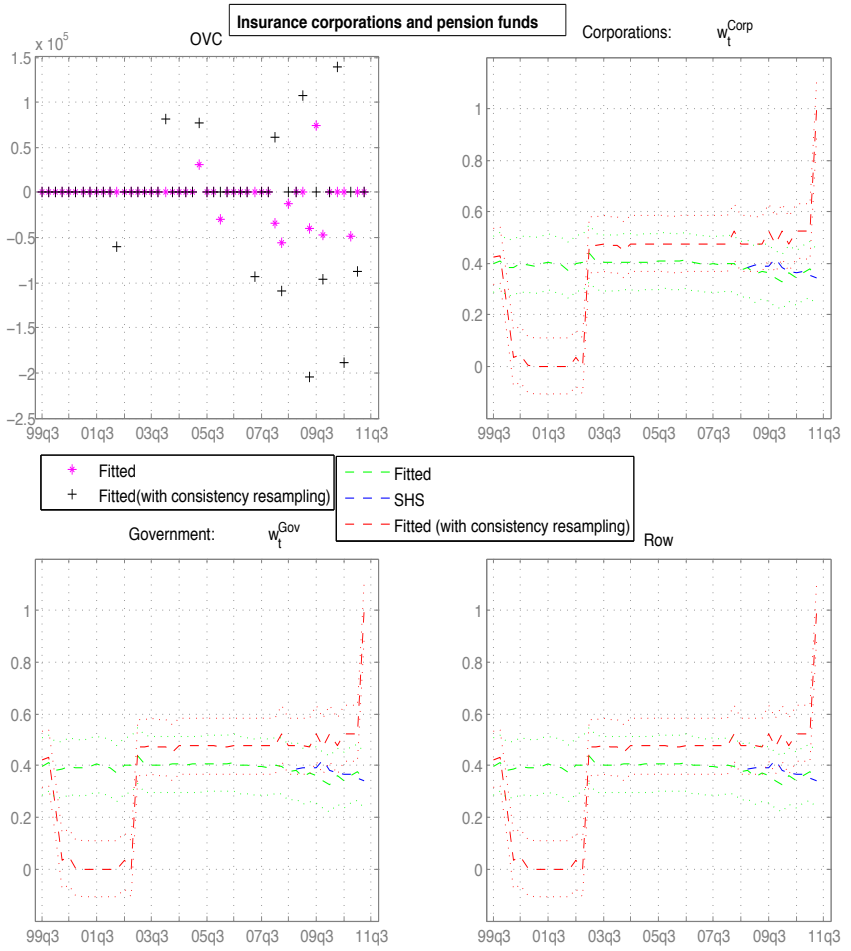


Figure 4.7: Results for holdings of F33 (Securities other than shares) for the institutional sector *Insurance Corporations and Pension Funds* in the euro area: OVC (Other Volume Changes) and the composition of the portfolio decomposed in assets issued by corporations, governments and RoW (Rest-of-the-World): Fitted portfolio structure (and 95 % confidence interval) with and without consistency resampling scheme and sub-series of data available from SHS (Security holdings statistics).

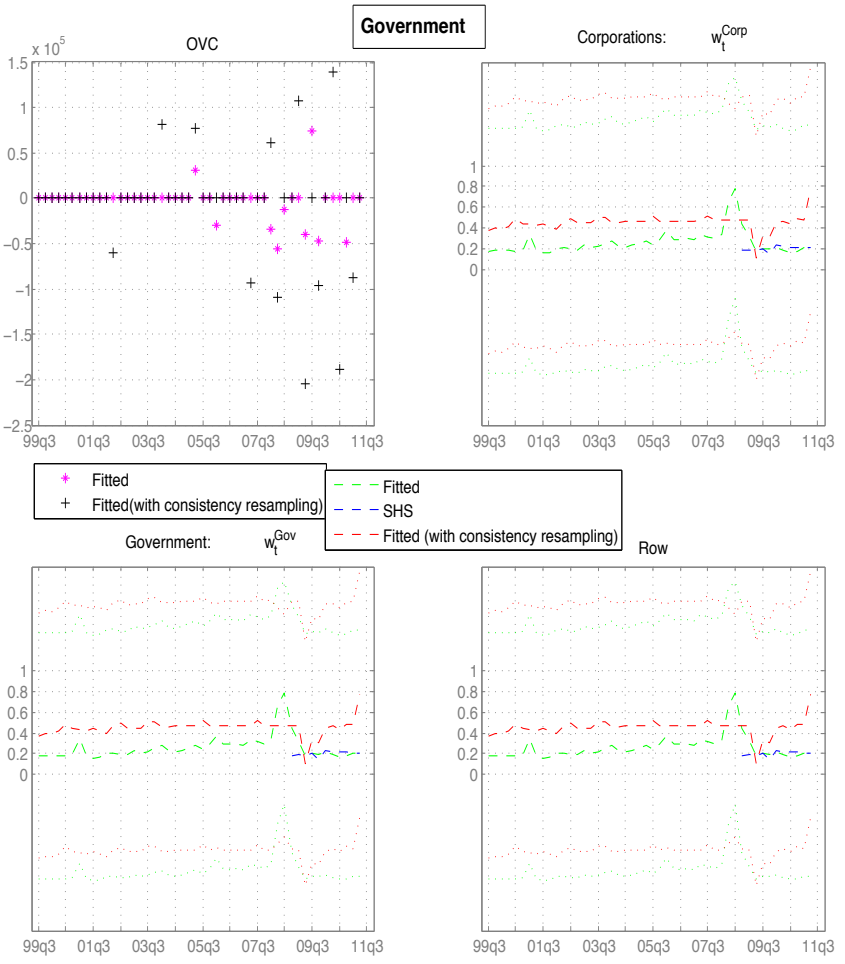


Figure 4.8: Results for holdings of F33 (Securities other than shares) for the institutional sector *Government* in the euro area: OVC (Other Volume Changes) and the composition of the portfolio decomposed in assets issued by corporations, governments and RoW (Rest-of-the-World): Fitted portfolio structure (and 95 % confidence interval) with and without consistency resampling scheme and sub-series of data available from SHS (Security holdings statistics).

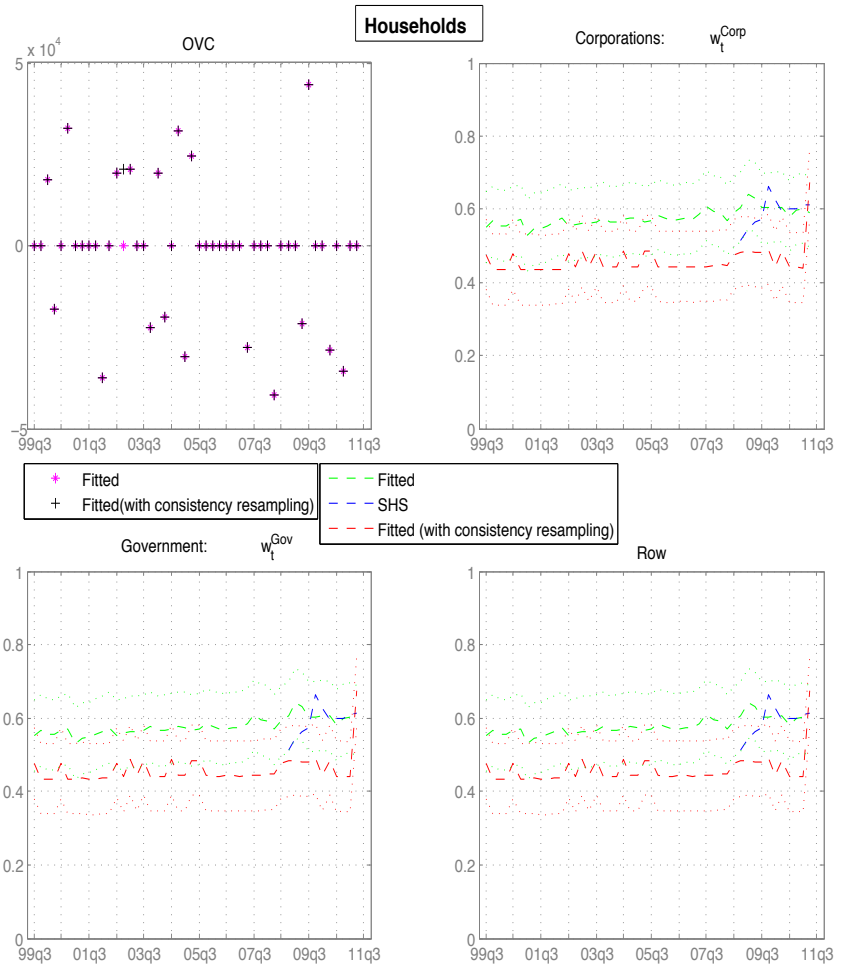


Figure 4.9: Results for holdings of F33 (Securities other than shares) for the institutional sector *Households* in the euro area: OVC (Other Volume Changes) and the composition of the portfolio decomposed in assets issued by corporations, governments and RoW (Rest-of-the-World): Fitted portfolio structure (and 95 % confidence interval) with and without consistency resampling scheme and sub-series of data available from SHS (Security holdings statistics).

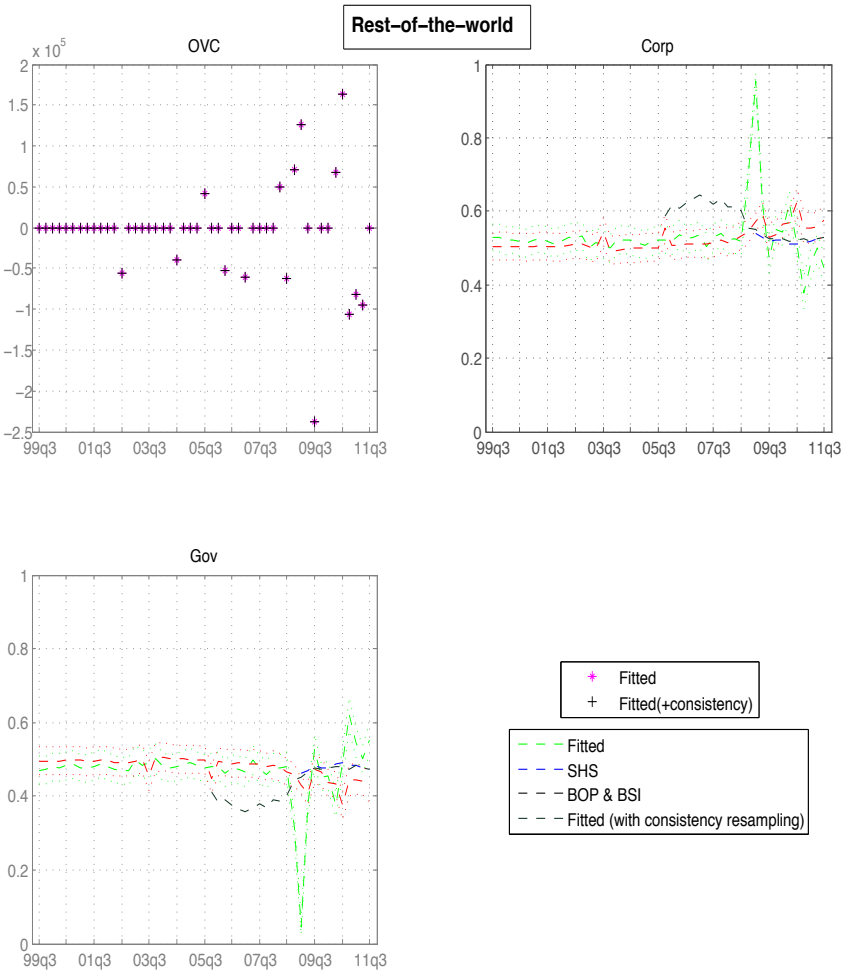


Figure 4.10: Results for holdings of F33 (Securities other than shares) for the institutional sector *Rest-of-the-World* in the euro area: OVC (Other Volume Changes) and the composition of the portfolio decomposed in assets issued by corporations, governments and RoW (Rest-of-the-World): Fitted portfolio structure (and 95 % confidence interval) with and without consistency resampling scheme and sub-series of data available from SHS (Security holdings statistics).

Bibliography

[1] European Central Bank (2010)Euro Area Statistics methodological notes. *ECB Monthly Bulletin*, Chapter 3, available online.

[2] da Silva D.B.N., Smith T.M.F. (2001). Modelling compositional time series from repeated surveys. *Survey Methodology. Statistics Canada*, **27**, 205-215.

[3] da Silva D.B.N., Migon H.S. ,Correia L.T. (2011). Dynamic Bayesian Beta models. *Computational statistics and data analysis*, **55**, 2074-2089.

[4] Diz J.D. (2009). Stock-Flow model for liabilities, internal ECB report.

[5] Doucet A., De Freitas N., Gordon N.J. (eds.) (2001). Sequential Monte Carlo Methods in Practice. *New York: Springer-Verlag.*

[6] European Commission (1996). European System of Accounts 1995 (ESA), Eurostat, Brussels-Luxembourg.

[7] Grunwald G., Raftery A., Guttorp P. (1993). Time series of continuous proportions. *Journal of the royal statistical society, Series B*, **55**, 103-116.

[8] Liu J., M. West (2011).Combined parameter and state estimation in simulation-based filtering. In *Sequential Monte Carlo Methods in Practice* (eds: A. Doucet *et al.*), Springer-Verlag.

[9] Parmigiani G., Inoue L. (2009). *Decision Theory, Principles and applications*, Chichester, UK: John Wiley & Son.

[10] Schneider M.H., Zenios S.A. (1990). A comparative study of algorithms for matrix balancing . *Operations research*, **38**, 439-455.

Part III

Applications in Pharmacokinetics and Clinical Trials

Chapter 5

Bayesian semiparametric inference for population pharmacokinetics-pharmacodynamics

Abstract :

Part III deals with pharmacokinetic problems. This chapter studies population pharmacokinetics - pharmacodynamics (PKPD) inference within a Bayesian setting. The general aim is an explanatory analysis of heterogeneity of PKPD across patients. We first consider a parametric (Gaussian) population distribution and examine with care the specification of the prior on the unknown covariance matrix, which characterizes the degree of heterogeneity of the population and the connections among coefficients. We propose a simple but flexible prior based on a parsimonious characterization of the covariance matrix, related to the scaled Inverse-Wishart. The proposed specification is based on a parsimonious characterization of the this covariance matrix and it is developed from the scaled Inverse-Wishart. The proposed prior allows the patient-specific random coefficients to have two clustering mechanisms with no hierarchical structure, one for the PK and the other for the PD parameters. The second contribution

is within a Bayesian semiparametric setting. We propose to incorporate the possible sequentiality between the PK and the PD model through the use of an appropriate nonparametric prior on the population distribution. The proposed Enriched Dirichlet Process (EDP) prior allows a hierarchical clustering of the PK and PD patient-specific random coefficients.

5.1 Introduction

Early phases of a clinical trial arise challenging statistical problems due both to the small sample size and to the need of providing precise answers. Some of the main goals are to study, *under pre-specified conditions*, what the body does to the drug -area of analysis of the PK- and what the drug in turn does to the body -area of analysis of the PD. Concentration profiles are the focus of PK studies. The aim is to describe, usually through a system of ordinary differential equations (ODEs), the kinetic behavior of the administered drugs, i.e. the absorption, the distribution and the elimination of the drug throughout a time interval in which (or prior to which) many doses have been received. Finding the link between these concentration profiles and the effects of the treatment is the focus of the PD analysis. PD describes the physiological effects of the drugs on the body and the mechanisms of drug action, usually using nonlinear differential equations. In order to properly evaluate all these mechanisms, it is necessarily to understand what the *pre-specified conditions* are, by correctly defining the framework of intervention, and properly incorporating them in the statistical model. *Pre-specified conditions* include not only the general observable health conditions and the disease status of the patients, but also patient-specific (unobservable or unobserved) features of each patient of the trial, highly characterized by heterogeneity.

Heterogeneity plays an important role in every clinical trial: heterogeneity among sub-types of diseases; heterogeneity among disease status when the patient starts the treatment; heterogeneity among the patients; and so on. Even when several covariates are included in the model, the

patients can still be different: they have different reactions, different frailties and there is still room for unexplained heterogeneity across patients. If heterogeneity is not properly taken into account, assuming that the whole population of patients is homogeneous, this source of variability is misleadingly imputed to the model or to the data quality. The first step for properly evaluating the patient-specific features is to include patient-specific coefficients, possibly with a flexible latent distribution. Since PKPD data are usually a collection of a few data points, it is helpful to borrow strength from the other patients for inferring each patient-specific coefficient instead of modelling each patient separately (Wakefield *et al.*, 1994; Wakefield and Racine-Poon, 1995). For the same reason, it can be useful to add some further information in the model, such as some information a priori available on some parameters. This motivates the researcher to perform the analysis within a Bayesian framework.

The first Bayesian models proposed for PKPD applications were based on a parametric distribution for the patient-specific random coefficients (Racine-Poon, 1985; Wakefield *et al.*, 1994; Wakefield and Racine-Poon, 1995; Lunn *et al.* 2002). Bayesian parametric population PKPD models allow to exploit the information collected in the whole data set. The resulting borrowing strength across patients improves the estimation of each patient-specific coefficient. In this parametric setting, the choice of the prior on the parameters of the latent distribution (also called *population distribution*) matters significantly. The covariance matrix of the population distribution contains information about the degree of heterogeneity of the PK and PD patient-specific parameters (e.g. the variability of the elimination rate of a drug across the population) and the correlations (e.g. patients with higher elimination rate have a smaller volume of distribution). These are clearly crucial aspects of the analysis. An aspect that does not appear to be sufficiently acknowledged is that the population of patients can intrinsically have different degrees of heterogeneity in the PK mechanisms and in the PD mechanisms. Furthermore,

"indirect reasons", i.e. related to the measurements errors or to misspecifications, can affect differently the PK and the PD stages. Therefore, we propose a careful specification of the prior on the population covariance matrix, that allows to model a different clustering behaviour of the PK and the PD parameters. Namely, we model the clustering structure of the coefficients (i.e. within each vector of patient-specific random coefficients, and not for clusters of patients) through two independent blocks of dependent patient-specific random coefficients: one for the PK and another for the PD. The proposed prior on the population covariance matrix will be named *scaled d- Inverse-Gamma* (sDIG), since it is based on the scaled Inverse-Wishart distribution (O'Malley and Zaslavsky, 2008; Gelman *et al.*, 2009). The sDIG has similar features, with a block-diagonal covariance matrix where the block elements have Inverse-Gamma and Inverse-Wishart distributions respectively, but it is even more parsimonious. The idea is to break up the covariance matrix into a diagonal matrix of scale parameters and an unscaled covariance matrix which is given a d-Inverse-Gamma (DIG) distribution. The clustering structure is obtained by imposing common scale parameters for all the elements belonging to the same block, which means in our case to have only two scale parameters, i.e. one for the PK and the other for the PD cluster.

In other situations, sequentiality between the PK and PD groups of patient-specific random coefficients is instead more realistic than independence. Indeed, sequentiality can be important for at least three reasons.

The first reason is that the analysis of population PKPD data is usually computational intensive and it requires techniques that simplify, in some sense, the problem, to obtain results quickly. This problem is important even in the easiest parametric case and it becomes worse in the nonparametric setting. As a possible solution, some techniques based on the sequentiality between the PK and the PD models have been proposed within a parametric framework (Zhang *et al.*, 2003; Lunn *et al.*, 2009; Lacroix *et al.*, 2012). Zhang *et al.* (2003) discussed three possible

sequential approaches for parametric population PKPD models: Population PK Parameters and Data (PPP&D), Population PK Parameters (PPP) and Individual PK Parameters (IPP). All of them first fit the PK sub-model is estimated using the PK data alone; then, the PD sub-model, using the PD data and plugging-in point estimates of the PK patient-specific random coefficients. The three approaches discussed by Zhang *et al.* (2003) are characterized by different estimation procedure for the PK model and different use of the PK data for the PD model. PPP&D and PPP make use of maximum likelihood estimates for the PK patient-specific random coefficients. They differ in the inclusion of the PK data into the likelihood: PPP&D includes them whereas PPP does not. IPP instead plugs the modal posterior Bayes estimates of the PK patient-specific random coefficients in the PD likelihood, with a prior derived only from the PK data. In contrast with these approaches, that do not allow the uncertainty on the PK parameters to propagate through the inference on the PD parameters, Lunn *et al.* (2009) proposed a multiple imputation approach which simulates several possible values for the concentrations based on the PK analysis and then averages the resulting PD estimates associated with each simulated concentrations. The second reason for which sequentiality can be important is that it expresses the physiological asymmetry between PK and PD mechanisms. First, a drug is administered, absorbed, distributed within the body. Then, and only after the absorption starts, the drug begins (hopefully) to produce its therapeutic benefits. The third reason is that it is often realistic to have more information on the PK mechanisms than on the PD ones. Taking into account this sequentiality can help to prevent unwanted feedbacks in estimation procedure, i.e. it prevents modification of the fitted population PK parameters by the PD model (Lunn *et al.*, 2009).

In this chapter, we also propose an enrichment of the traditional Bayesian semiparametric PKPD model, which includes the asymmetry between the PK and the PD groups of random coefficients through the

Enriched Dirichlet Process (EDP) prior (see **Chapter 2**) on the distribution of the patient-specific random coefficients. Instead of imposing the sequentiality in the estimation procedure (Lunn *et al.*, 2002; Zhang *et al.*, 2003), this proposal allows for sequentiality within the probabilistic model underlying the random coefficients.

The chapter is organized as follows. In Section 5.2, a brief introduction on population PKPD problems is provided and the standard Bayesian PKPD modelling approaches are discussed. Section 5.3 includes the discussion on the covariance matrix and the proposed sDIG prior specification. In Section 5.4, a short review of Bayesian nonparametric PKPD approaches is provided. Section 5.5 introduces the *Enriched* Bayesian nonparametric PKPD model based on the use of the EDP is introduced. Subsection 5.5.1 shortly describes the procedure for making inference on it. In Section 5.6, the proposals are illustrated through a simulation study. A final discussion section concludes the chapter.

5.2 Population PKPD models

Population PKPD data usually consist of dose histories, covariates, concentration (or PK) measurements and response (or PD) measurements, for a pool of patients. Let i, $i = 1, .., N$, be the label associated with each patient and t, $t = 1, .., T_i$, be the measurement time index for patient i. In a typical clinical trial, one or more doses, say $d_{i,t}$, are administered to each patient i at different times t. Concentrations, say $y_{i,t}$, and responses, say $z_{i,t}$, are recorded for each patient i at times $t = 1, .., T_i$. For the sake of simplicity, let us focus on the administration of a single dose, d_i, at time 1, for each patient i. Let us assume that the measurement time indices are the same for all the patients, i.e. $T_i \equiv T$, and no covariates are included in the model. For each patient i, a single dose is associated with a whole sequence of concentrations or PK measurements, $y_{i,t}$, $t = 1, .., T$, because the dose concentration within the body does

not vanish immediately after the administration of the dose but it persists and evolves over time. Concentrations are usually modelled using PK compartment models (Bauer, 2008; Ette and Williams, 2007). These models describe the kinetic behavior of the administered drugs making use of a fixed number, let us say p, of hypothetical compartments within the body. This kinetics is usually expressed by a system of p ODEs, one equation for each hypothetical compartment, representing the network among these compartments that the dose has to go through to arrive to the *site of action* or *effective site*. The hypothetical concentrations of the dose for patient i in the p compartments at time t, denoted by $\underline{\boldsymbol{G}}_{i,t} = (\boldsymbol{G}_{i,t;1}, ..., \boldsymbol{G}_{i,t;p})'$, can then be modelled as follows:

$$d\boldsymbol{G}_{i,t;j} = \mu_{PK}\left(\underline{\boldsymbol{G}}_{i,t}, d_i, \boldsymbol{\beta}_i^{PK}\right) dt, \quad j = 1, .., p. \tag{5.1}$$

where $\boldsymbol{\beta}_i^{PK}$ are the unknown coefficients for patient i describing the whole system of ODEs. Solving equation (5.1), one can find the p-dimensional curve $\underline{\boldsymbol{G}}_{i,t}$.

The sequence of PK measurements, $y_{i,t}$, $t = 1, .., T$, is usually well-described by the evolution of a specific equation or compartment, let us say the jth equation, from the general multivariate vector $\underline{\boldsymbol{G}}_{i,t}$. This compartment is called *effective compartment* and it usually corresponds to the central compartment. Let us call the evolution of the concentration in this jth compartment $\boldsymbol{G}_{i,t;j} \equiv f_{PK}\left(t, d_i, \boldsymbol{\theta}_i^{PK}\right)$, where $\boldsymbol{\theta}_i^{PK}$ are the K_1 random coefficients associated with the jth compartment. The function $f_{PK}\left(t, d_i, \boldsymbol{\theta}_i^{PK}\right)$ is usually called the *deterministic* (or *logical*) *part of the PK measurements*, since the observations $y_{i,t}$ are some noisy realizations around it. For instance, the simplest PK model is the one associated with only one compartment, with $\frac{d\boldsymbol{G}_i}{dt} = -\boldsymbol{k}_i\boldsymbol{G}_i$, where $\boldsymbol{k}_i$ is the rate of elimination of the drug for patient i and the initial condition depends on the administered dose[1] as follows: $\boldsymbol{G}_{i,0} = \frac{d_i}{\boldsymbol{V}_i}$, where $\boldsymbol{V}_i$ is the apparent

[1] An alternative initial condition is $\boldsymbol{G}_{i,0} = d_i$. This would imply that $\boldsymbol{G}_{i,t}$ repre-

volume of distribution. For each given d_i, this differential equation admits the following explicit solution:

$$\boldsymbol{G}_{i,t} = \frac{d_i}{\boldsymbol{V}_i} e^{-\boldsymbol{k}_i t} \tag{5.2}$$

with $\boldsymbol{\theta}_i^{PK} = (\log(\boldsymbol{V}_i), \log(\boldsymbol{k}_i))$. This model can be used, for instance, for drugs that rapidly equilibrate in the tissue compartment. It uses only one volume term for each patient i, $\boldsymbol{V}_i$, and it assumes that the amount of drug is decreasing at a rate that is proportional to the amount of drug remaining in the body. In the following sections, we will assume $p = 1$ with pharmacokinetics described by equation (5.2).

A first source of randomness taken into consideration in population PKPD models is the *intra-patient* variation, i.e. the deviation between each particular observation and its mean value, $y_{i,t} - f_{PK}\left(t, d_i, \boldsymbol{\theta}_i^{PK}\right)$ (Wakefield and Racine-Poon, 1995). It is usually more convenient to model the logarithmic transformation of the concentrations, as well as for the response variables, to stabilize the variances. One common and convenient choice for the log-transformed variables is the Gaussian distribution. In making this choice, we are implicitly assuming that the error of measurement is symmetrically distributed around the log-mean obtained by the PK deterministic component. That is to say, there is no bias in the available measurement tools and symmetric errors are equally likelihood. If the number of observations were enough, then this assumption could be easily verified using specific statistical tests, such as the Kolmogorov-Smirnov test or the Kuiper test, or, to obtain a rough indication of the goodness of the fit, simple graphical tools, e.g. QQ plot. If this assumption is reasonable, the PK model for the PK data, conditionally on the unknown PK parameters, can be formulated as follows:

sents the *amount* of dose instead of the concentration of dose. This latter would then be derived by dividing the amount $\boldsymbol{G}_{i,t}$ by the volume $\boldsymbol{V}_i$.

$$\log\left(\boldsymbol{y}_{i,t}\right) = \log\left(f_{PK}\left(t, d_i, \boldsymbol{\theta}_i^{PK}\right)\right) + \boldsymbol{\epsilon}_{i,t}^{PK},$$

$$\text{with} \quad \boldsymbol{\epsilon}_{i,t}^{PK} \mid \boldsymbol{\sigma}_{PK}^2 \overset{\text{iid}}{\sim} N\left(0, \boldsymbol{\sigma}_{PK}^2\right).$$

or, equivalently:

$$\log\left(\boldsymbol{y}_{i,t}\right) \mid \boldsymbol{\theta}_i^{PK}, \boldsymbol{\sigma}_{PK}^2 \overset{indep}{\sim} N\left(\log\left(f_{PK}\left(t, d_i, \boldsymbol{\theta}_i^{PK}\right)\right), \boldsymbol{\sigma}_{PK}^2\right). \quad (5.3)$$

Let us now consider the PD measurements, $z_{i,t}$. The PD model links the mean value of the PK measurement, $\boldsymbol{G}_{i,t;j} \equiv f_{PK}\left(t, d_i, \boldsymbol{\theta}_i^{PK}\right)$, with the observed response, $z_{i,t}$, for each t and i. This link depends on the nature of response, on the kind of administered drug and on the reaction of the response variable to the drug. It can be expressed by some functional form, say $f_{PD}\left(t, d_i, \boldsymbol{\theta}_i^{PD}, f_{PK}\left(t, d_i, \boldsymbol{\theta}_i^{PK}\right)\right)$, which depends on the measurement time, t, the administered dose, d_i, the mean value of the PK measurement, $f_{PK}\left(t, d_i, \boldsymbol{\theta}_i^{PK}\right)$, and the $K_2\,(\equiv K - K_1)$ PD patient-specific random-coefficients, $\boldsymbol{\theta}_i^{PD}$. This functional form describes the predictable part of the PD measurements and it can be the (analytical or numerical) solution of some differential equation, usually a nonlinear differential equation. It is called the *deterministic (or logical) part of the PD measurements*. For example, within tumor growth analysis as well as for many other treatments, a common choice for an inhibitory drug is to use a semi-mechanistic, indirect response model (Ette and Williams, 2007; Bueno *et al.*, 2008), e.g. a treatment for inhibition of water reabsorption by loop diuretics such as furosemide. Let $\boldsymbol{R}_{i,t}$ represent the platelet count and $\boldsymbol{Y}_{\text{eff};i,0}$ its value at the baseline. The variable of interest can be $\boldsymbol{Y}_{\text{eff};i,t} = \boldsymbol{R}_{i,t} - \boldsymbol{Y}_{\text{eff};i,0}$. The semi-mechanistic, indirect response model, as modelled by Ette and Williams (2007) and Bueno *et al.*, (2008), assumes that $\boldsymbol{R}_{i,t}$ is governed by a zero-order production

process, with rate $\boldsymbol{k}_{\text{in};i}$, and a first-order loss-of-response process, with rate $\boldsymbol{k}_{\text{out};i}$, that is to say, the dynamics of $\boldsymbol{R}_{i,t}$ is the following:

$$\frac{\boldsymbol{R}_{i,t}}{dt} = \boldsymbol{k}_{\text{in};i} - \boldsymbol{k}_{\text{out};i}\boldsymbol{R}_{i,t}.$$

The inhibitory effects of the drug is included into the model by including a decelerating factor on the production process, expressed by the $\boldsymbol{k}_{\text{in};i}$. The model becomes (Ette and Williams, 2007; Bueno *et al.*, (2008):

$$\frac{\boldsymbol{R}_{i,t}}{dt} = \boldsymbol{k}_{\text{in};i}\left(1 - \frac{\boldsymbol{I}_{\max;i}\boldsymbol{G}_{i,t;j}}{\boldsymbol{IC}_{50;i} + \boldsymbol{G}_{i,t;j}}\right) - \boldsymbol{k}_{\text{out};i}\boldsymbol{R}_{i,t} \tag{5.4}$$

where $\boldsymbol{G}_{i,t;j} \equiv f_{PK}\left(t, d_i, \boldsymbol{\theta}_i^{PK}\right)$ is the mean value of the concentration coming from the PK model and $\boldsymbol{\theta}_i^{PD}$ is defined as follows:

$$\boldsymbol{\theta}_i^{PD} = (\log\left(\boldsymbol{Y}_{\text{eff};0,i}\right), \log(\boldsymbol{I}_{\max;i}), \log(\boldsymbol{IC}_{50;i}), \log(\boldsymbol{k}_{\text{out};i})).$$

To ensure stationary, it is common to assume that $\boldsymbol{k}_{\text{in};i} = \boldsymbol{k}_{\text{out};i}\boldsymbol{IC}_{50;i}$. The parameter $\boldsymbol{I}_{\max;i}$ is the maximum inhibition of $\boldsymbol{k}_{\text{in};i}$, constrained between 0 (no inhibition) and 1 (maximal inhibition); and $\boldsymbol{IC}_{50;i}$ is the drug concentration that produces 50 % of the maximum inhibition. It follows that the factor $\left(1 - \frac{\boldsymbol{I}_{\max;i}\boldsymbol{G}_{i,t;j}}{\boldsymbol{IC}_{50;i} + \boldsymbol{G}_{i,t;j}}\right)$ moves between 0 and 1, acting indeed as inhibitory factor for the production of $\boldsymbol{R}_{i,t}$.

Equation (5.4) has no analytical solution since it is nonlinear in its parameters. Therefore, numerical methods are required for solving it, we used the Runge-Kutta 4th order method to produce the plot in Figure 5.1.

Figure 5.1 shows the deterministic part of the PK and PD measurements defined by models (5.2) and (5.4), associated with different initial doses. The figures are drawn for the fixed parameters: V_i=25, k=0.2,

$Y_{\text{eff};0}$=50, $I_{\text{max};i} = 1$, $IC_{50;i}$=5, and $k_{\text{out};i} = 0.1$.

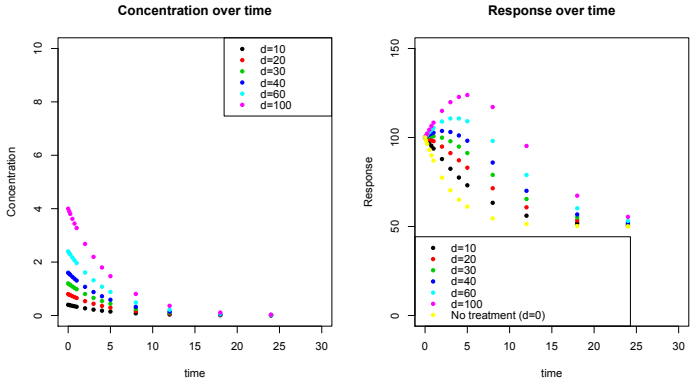


Figure 5.1: One-compartmental PK model with semi-mechanistic, indirect response model associated with the administration of different doses at time 1.

Again, a first source of randomness for the PD model is the *intra-patient* variation, i.e. $z_{i,t} - f_{PD}\left(t, d_i, \boldsymbol{\theta}_i^{PD}, f_{PK}\left(t, d_i, \boldsymbol{\theta}_i^{PK}\right)\right), i = 1, .., N$ and $t = 1, .., T$. As for the PK model, provided that the assumption of symmetric measurement errors is reasonable, the model for the PD measurements usually has the following form:

$$\log(\boldsymbol{z}_{i,t}) = log\left(f_{PD}\left(t, d_i, \boldsymbol{\theta}_i^{PD}, f_{PK}\left(t, d_i, \boldsymbol{\theta}_i^{PK}\right)\right)\right) + \boldsymbol{\epsilon}_{i,t}^{PD},$$

$$\text{with} \quad \boldsymbol{\epsilon}_{i,t}^{PD} \mid \boldsymbol{\sigma}_{PD}^2 \overset{\text{iid}}{\sim} N\left(0, \boldsymbol{\sigma}_{PD}^2\right),$$

or, equivalently,

$$\log(\boldsymbol{z}_{i,t}) | \boldsymbol{\theta}_i^{PD}, \boldsymbol{\theta}_i^{PK}, \boldsymbol{\sigma}_{PD}^2 \overset{\text{iid}}{\sim} N\left(log\left(f_{PD}\left(t, d_i, \boldsymbol{\theta}_i^{PD}, f_{PK}\left(t, d_i, \boldsymbol{\theta}_i^{PK}\right)\right)\right), \boldsymbol{\sigma}_{PD}^2\right). \quad (5.5)$$

The PD variance $\boldsymbol{\sigma}^2_{PD}$ expresses the measurement variability of the response variable. As for the $\boldsymbol{\sigma}^2_{PK}$, it is closely related with the precision of the measurement tools, which have not to be necessarily the same.

The PD error terms $\boldsymbol{\epsilon}^{PK}_{i,t}$ are assumed to be independent from $\boldsymbol{\epsilon}^{PD}_{i,t}$ and this assumption will be maintained in the following.

Equation (5.5) points out the sequentiality and the asymmetry between the PK and the PD models: first a PK model is required, then, conditionally on the PK mean value, a PD model is defined. Indeed, models (5.3) and (5.5) are, respectively, the marginal model of the PK observations, $\boldsymbol{y}_{i,t}$, and the conditional model of the PD observations, given the PK model, $\boldsymbol{z}_{i,t} \mid \boldsymbol{y}_{i,t}$. The joint PKPD model, *given the patient-specific random coefficients and the unknown parameters*, is thus obtained by combining equations (5.3) and (5.5). Within a Bayesian framework, they represent the first level of a hierarchical model. For the sake of simplicity, the observational variances, $\boldsymbol{\sigma}^2_{PK}$ and $\boldsymbol{\sigma}^2_{PD}$, are here assumed to be common for all the patients. More generally, one could assume patient specific variances, $\boldsymbol{\sigma}^2_{PK,i}$ and $\boldsymbol{\sigma}^2_{PD,i}$.

As said, it is helpful to make use of a Bayesian framework to borrow strength across patients. A random-coefficients model is therefore desirable. Hence, a second source of randomness is given by the *inter-patient* variation, i.e. the variability of the patient-specific random coefficients, $\boldsymbol{\theta}^{PK}_i$ and $\boldsymbol{\theta}^{PD}_i$, within the population of patients. This source of randomness is modelled at the second level of the hierarchy through a model for the patient-specific random coefficients, $\boldsymbol{\theta}_i \equiv \left(\boldsymbol{\theta}^{PK}_i, \boldsymbol{\theta}^{PD}_i\right)$. In the next section, we discuss the Bayesian parametric approach for $\boldsymbol{\theta}_i$. Then, starting from Section 5.4, we will take into consideration Bayesian nonparametric models.

5.3 Bayesian parametric approach

A Bayesian parametric framework could be adequate whenever the researcher is confident enough on the form of the latent population distribu-

tion. Usually, it is assumed that the population distribution is Gaussian (Wakefield and Racine-Poon, 1995; Lunn *et al.* 2002):

$$\left(\boldsymbol{\theta}_i^{PK}, \boldsymbol{\theta}_i^{PD}\right) \mid \boldsymbol{\mu}, \boldsymbol{\Sigma} \overset{\text{iid}}{\sim} N_K\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \tag{5.6}$$

where $N_K(\cdot, \cdot)$ denotes the K- dimensional Gaussian distribution.

Due to the typical small sample size of early clinical trials, robustness of the results, both to model assumptions and to prior specification, plays a crucial role. The distributional robustness is one of main features under investigation in the model validation stage and it can be addressed by using nonparametric inference, which will be discussed later within this chapter. Robustness to different prior specifications on $\boldsymbol{\Sigma}$ is instead the focus of this subsection.

Choice of the prior on $\boldsymbol{\Sigma}$

Particular attention is required for the covariance matrix, $\boldsymbol{\Sigma}$, of the latent population distribution. The covariance matrix $\boldsymbol{\Sigma}$ contains information about the degree of heterogeneity of the PK and PD patient-specific parameters across the population and about their correlations. Consequently, its role is to represent the strength of the link among the patient-specific coefficients and the nature of the shrinkage of the posterior patient-specific coefficients towards their common means.

The most popular prior on a full covariance matrix is the Inverse-Wishart (IW) distribution. According to the parametrization used in Barnard *et al.* (2000), if $\boldsymbol{\Sigma} \sim IW(\nu, A)$, its density is equal to:

$$f_K(\Sigma \mid \nu) \propto \mid \Sigma \mid^{-\frac{\nu + K + 1}{2}} e^{-\frac{1}{2}\text{tr}\left(A\Sigma^{-1}\right)} \tag{5.7}$$

where $\nu > K - 1$ is the degree of freedom, expressing the uncertainty on the elements of $\boldsymbol{\Sigma}$, and A is a $K \times K$ positive defined matrix. The

expected value is:

$$E(\boldsymbol{\Sigma}) = \frac{A}{\nu - K - 1}.$$

However, an IW prior can be quite restrictive, since it has only one precision parameter, ν. At the same time, a IW prior is not a parsimonious choice because it involves a full covariance matrix and, especially when its dimension is big, this choice can slow significantly the speed of convergence of the MCMC.

Different specifications for the covariance matrix may represent better options. Possible alternative choices can be the scaled IW distribution (O'Malley and Zaslavsky, 2008; Gelman *et al.*, 2009), the generalized Inverse-Wishart conjugate prior (Brown *et al.*, 1994; Consonni and Veronese, 2003), the d-Inverse-Gamma (Frühwirth-Schnatter, 1992) or a block-diagonal matrix with an IW prior on each block.

The scaled IW (sIW) distribution has been introduced by O'Malley and Zaslavsky (2008), although its current name has been assigned by Gelman *et al.* (2009). It allows more flexibility with respect to the scale parameter than the IW. It breaks up the covariance matrix $\boldsymbol{\Sigma}$ into a diagonal matrix of scale parameters and an unscaled covariance matrix which is given the IW distribution. In particular, $\boldsymbol{\Sigma}$ is specified as follows:

$$\boldsymbol{\Sigma} = diag(\underline{\boldsymbol{\xi}})\ \boldsymbol{Q}\ diag(\underline{\boldsymbol{\xi}}) \tag{5.8}$$

where $\boldsymbol{Q}$ is a full IW distributed matrix, called unscaled covariance matrix, and $\underline{\boldsymbol{\xi}}$ is a K-dimensional vectors containing the scale parameters $\boldsymbol{\xi}_k$, $k = 1, .., K$. The model is completed with a prior on $\boldsymbol{\xi}_k$, $k = 1, .., K$. Let $\boldsymbol{\sigma}^2_{j,k}$ be the element in the jth row and kth column of the matrix $\boldsymbol{\Sigma}$ and, similarly, $\boldsymbol{Q}_{j,k}$ be the (j,k)th element of the matrix $\boldsymbol{Q}$. It follows that:

$$\boldsymbol{\sigma}^2_{k,k} = \boldsymbol{\xi}^2_k\ \boldsymbol{Q}_{k,k} \quad \text{for} \quad k = 1, .., K \tag{5.9}$$

$$\boldsymbol{\sigma}^2_{k,j} = \boldsymbol{\xi}_k \boldsymbol{\xi}_j \, \boldsymbol{Q}_{k,j} \quad \text{for} \quad k,j = 1,..,K \tag{5.10}$$

In defining such a prior, O'Malley and Zaslavsky (2008) move from the *separation strategy prior* of Barnard, McCulloch and Meng (2000). Barnard *et al.* (2000) express the covariance matrix as in equation (5.8), but $\boldsymbol{Q}$ is the correlation matrix and $\underline{\boldsymbol{\xi}}$ the vector of standard deviations. Consequently, expression (5.10) remains the same, but expression (5.9) reduces to $\boldsymbol{\sigma}^2_{k,k} = \boldsymbol{\xi}^2_k$ for $k = 1,..,K$. In the sIW specification, the parameters in $\underline{\boldsymbol{\xi}}$ and $\boldsymbol{Q}$ cannot instead be interpreted separately. They are used for setting up conveniently the model, but the interest is on $\boldsymbol{\sigma}^2_{k,k}$ and $\boldsymbol{\sigma}^2_{k,j}$, $k,j = 1,..,K$. Moreover, as with the unscaled Wishart, setting the degrees-of-freedom parameter to $K+1$ has the effect of setting a uniform distribution on the individual correlation parameters (Gelman *et al.*, 2009). The sIW is an appealing specification because it introduces multiple scale parameters. However, it is still not parsimonious.

There are many other priors for the covariance matrix in the literature, but that do not solve our needs. For instance, another popular extension of the conjugate IW is the generalized Inverse-Wishart conjugate prior, which is also not parsimonious. Many other prior specifications for the covariance matrix have been proposed for K large (Dempster,1972; Dawid and Lauritzen, 1993). Making use of graph theory, these priors impose zeros in the precision matrix in correspondence with the absence of edges. These priors are very useful and appealing. However, the dimension of the patient-specific random-coefficients vector is usually not so high to justify such approaches.

The simplest parsimonious prior is the d-Inverse-Gamma distribution (DIG), which assumes $\boldsymbol{\Sigma}$ diagonal with independent Inverse Gamma random variables on the diagonal:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\sigma}^2_{1,1} & 0 & .. & 0 & 0 \\ 0 & \boldsymbol{\sigma}^2_{2,2} & ... & 0 & 0 \\ .. & .. & ... & ... & ... \\ 0 & 0 & ... & 0 & \boldsymbol{\sigma}^2_{k,k} \end{pmatrix}, \tag{5.11}$$

with $\boldsymbol{\sigma}^2_{k,k} \overset{indep}{\sim} IG\,(a_k, b_k),\ k = 1, .., K$. Clearly, this specification implies the strong assumption of independence of the K patient-specific random coefficients.

Two desired features motivates our proposal of another prior for the covariance matrix, which we call the *scaled d-Inverse-Gamma* distribution (sDIG): being parsimonious and allowing for dependence. We assume that $\boldsymbol{\Sigma}$ can be represented as in expression (5.8). However, the peculiarity is that, in place of a full IW matrix Q, we use a DIG matrix. More specifically, we assume that:

$$\boldsymbol{\Sigma} = diag(\underline{\boldsymbol{\xi}})\ \boldsymbol{D}\ diag(\underline{\boldsymbol{\xi}}) \tag{5.12}$$

where

- the unscaled covariance matrix $\boldsymbol{D}$ has a DIG prior, i.e. $\boldsymbol{D}_{k,k} \overset{indep}{\sim} IG(a_k, b_k),\ k = 1, .., K$
- The jth element of the K dimensional scaling factor, $\underline{\boldsymbol{\xi}}$, has an IG distribution, i.e. $\boldsymbol{\xi}^2_k \overset{indep}{\sim} IG(c_k, d_k),\ k = 1, .., K$.

This specification allows for the introduction of some dependence across the patient-specific random coefficients by simply imposing some of the scale elements of the vector $\underline{\boldsymbol{\xi}}$ to be common. For example, imposing $\boldsymbol{\xi}_1 \equiv \boldsymbol{\xi}_2$, the first and the second element of the vector $\left(\boldsymbol{\theta}^{PK}_i, \boldsymbol{\theta}^{PD}_i\right)$ are conditional independent given $\boldsymbol{\xi}_1$ but marginally dependent. Indeed, the underlying idea of the sDIG prior is similar to the block-diagonal IG and block-diagonal IW distributions, where block diagonal structures are imposed and full sub-matrices are assumed to have independent IG and

IW distributions, respectively. However, the number of parameters to be estimated with the sDIG is smaller. This approach is very simple yet the result is a quite flexible prior with a few parameters. For PKPD applications, each $\boldsymbol{\theta}_i$ is a K-dimensional vector collecting both the PK (the first K_1 elements of $\boldsymbol{\theta}_i$) and the PD (the remaining $K_2 \equiv K - K_1$) random coefficients. Therefore, a convenient sDIG specification for $\boldsymbol{\Sigma}$ imposes that $\boldsymbol{\xi}_k = \boldsymbol{\xi}_{PK}$ for every $k = 1, .., K_1$ and $\boldsymbol{\xi}_j = \boldsymbol{\xi}_{PD}$ for $j = K_1 + 1, .., K$. This expresses the idea that two independent blocks of dependent patient-specific random coefficients exist: one for the PK patient-specific random coefficients and another for the PD patient-specific random coefficients.

Some analytical properties can also be derived. In particular, it is possible to show that the implied distribution for each non-zero variance element is an inverse G-Meier distribution (Springer and Thompson, 1970).

To carry out the Bayesian inference, the hierarchical model made up of the models for the data, i.e. (5.3) and (5.5), and the model for the patient-specific coefficients, (5.6), is completed with the priors on all the unknown parameters and hyper-parameters of the model, including the prior on the covariance matrix. Statistical inference for the unknown parameters and the random coefficients is then based on their posterior distributions, whose evaluation cannot be carried out in a closed-form solution. However, as indicated in Appendix, it is relatively straightforward to derive the full conditional distributions for most of the parameters and one can easily combine the Gibbs sampler and the MetropolisHastings (MH) algorithm to carry out the MCMC procedure. See Lunn *et al.* (2002) for more detailed discussion of MCMC algorithms for PKPD models.

5.4 Bayesian nonparametric approach

We now move towards a Bayesian nonparametric settings where the latent population distributions of the PK and PD patient-specific random

coefficients is is not restricted to a parametric family.

Indeed, the assumption of a Gaussian population distribution can be too rigid. First, it implies a global shrinkage of the patient-specific coefficients towards their common mean; second, it does not allow to properly take into account outliers, skewness, multimodality and asymmetry. For relaxing the assumption of a Gaussian population distribution, we take a Bayesian nonparametric approach. Other solutions can of course be envisaged. In the context of PKPD application, Huang and Dagne (2012) have proposed the use of a Skew-Normal distribution for the residual terms. A flexible Skew-Normal distribution could be also used for the latent distribution. We prefer a Bayesian nonparametric approach, with a discrete nonparametric prior, since this induces an implicit clustering of the patient-specific random coefficients. In particular, if the population distribution were absolutely continuous, then all the $\boldsymbol{\theta}_i$ would be different almost surely but locally shrink towards the modes. Instead, assuming a discrete population distribution allows $\boldsymbol{\theta}_i$ to have ties, i.e. to be clustered together. The model for the patient-specific random coefficients becomes:

$$\left(\boldsymbol{\theta}_i^{PK}, \boldsymbol{\theta}_i^{PD}\right) | \mathbf{P} = P \overset{\text{iid}}{\sim} P. \tag{5.13}$$

A common choice as a prior on $\mathbf{P}$ is the Dirichlet Process (DP) (Ferguson, 1973):

$$\mathbf{P} \sim DP\left(\alpha_\theta P_{0\theta}\right) \tag{5.14}$$

where α_θ is the precision parameter and $P_{0\theta}$ is the base measure[2]. The model is completed by the prior on the variances $\boldsymbol{\sigma}^2_{PK}$ and $\boldsymbol{\sigma}^2_{PD}$, usually assumed independent Inverse-Gamma.

DP random coefficients models are widely used; see Dunson (2010) for a recent review. In PKPD studies, there have been used, among

[2]Notice that we will use the same symbol for denoting both a probability measure and a distribution function.

the others, by Walker and Wakefield (1997), Müller and Rosner (1997) and Müller *et al.* (2007). As discussed in **Chapter 2**, the choice of a DP prior can be quite restrictive in a multivariate setting: here, it implies that both the PK and the PD groups of random coefficients have the same clustering structure. Alternatively, assuming independent DP priors for the distribution of the PK and PD patient-specific random coefficients implies having two independent clustering structures. In both cases, the asymmetry between PK and PD is not included. The model in the next section is proposed to overcome this weakness and it is based on the Enriched Dirichlet Process (EDP). For a discussion on other possible alternatives to the DP model, still within a Bayesian nonparametric setting, see Müller and Quintana (2004).

5.5 Our proposal: The Enriched Bayesian semiparametric population PKPD model

Our proposal is an "enriched" Bayesian population model where the DP prior (5.14) is replaced by the Enriched Dirichlet Process (EDP). More specifically, we consider the observational model defined by (5.3) and (5.5), with population distribution as in (5.13), but now assuming:

$$\mathbf{P} \sim EDP\left(\alpha_{\theta^{PK}} P_{0\theta^{PK}}, \alpha_{\theta^{PD}} P_{0\theta^{PD}}\left(\cdot|\theta^{PK}\right)\right); \tag{5.15}$$

see **Chapter 2** for the definition of the EDP. As before, the model is completed by the prior on the variances,

$$\boldsymbol{\sigma}^2_{PK} \sim IG\left(a_{PK}, b_{PK}\right) \text{ and } \boldsymbol{\sigma}^2_{PD} \sim IG\left(a_{PD}, b_{PD}\right), \tag{5.16}$$

with $\boldsymbol{\sigma}^2_{PK}$ independent from $\boldsymbol{\sigma}^2_{PD}$. The EDP prior is equivalent to express $\mathbf{P}$ in terms of the marginal probability of $\boldsymbol{\theta}_i^{PK}$, say $\mathbf{P}_{PK}$, and

the conditional of $\boldsymbol{\theta}_i^{PD}$ given $\boldsymbol{\theta}_i^{PK} = \theta_{PK}$, say $\mathbf{P}_{PD}(\cdot \mid \theta_{PK})$. Then, the prior is assigned "sequentially" as follows:

$$\mathbf{P}_{PK} \sim DP\left(\alpha_{\theta^{PK}} P_{0\theta^{PK}}\right),$$

$$\mathbf{P}_{PD}(\cdot \mid \theta_{PK}) \sim DP\left(\alpha_{\theta^{PD}} P_{0\theta^{PD}}\left(\cdot|\theta^{PK}\right)\right)$$

where $\mathbf{P}_{PD}(\cdot \mid \theta_{PK})$ is independent from $\mathbf{P}_{PD}(\cdot \mid \theta'_{PK})$, $\forall \theta'_{PK} \neq \theta_{PK}$ and $\mathbf{P}_{PD}(\cdot \mid \theta_{PK})$ is independent from $\mathbf{P}_{PK}$. This means that, first, a DP prior is assigned to the distribution of the PK patient-specific random coefficients, $\mathbf{P}_{PK}$, and the PK groups are estimated. Then, an independent DP prior is assigned on the conditional distribution, $\mathbf{P}_{PD}(\cdot \mid \theta_{PK})$. It is possible to assign priors to the hyper-parameters of the EDP, i.e. priors to the precision parameters, $\alpha_{\theta^{PK}}$ and $\alpha_{\theta^{PD}}$, and priors to the parameters of $P_{0\theta^{PK}}$, and $P_{0\theta^{PD}}\left(\cdot|\theta^{PK}\right)$. For the sake of simplicity, these parameters here are assumed to be fixed.

The EDP prior allows for an asymmetric and sequential treatment of the PK and PD patient-specific coefficients. It implies a "nested" clustering structure, where, within each PK cluster, one or more PD clusters can be defined. It should be pointed out that, although the verse of sequentiality between PK and PD is given by the structure of the real problem (first PK mechanisms, then PD reactions), the nested structure of the clustering is a modeling choice. The appropriateness of this nested clustering structure is problem-specific and it can be even reversed for some applications. The preference for the clustering structure implied by equation (5.15) expresses the fact that there is usually more confidence in the PK model specification than in the PD model specification. The mechanisms of absorption of the drug can be well explained by controlling for available covariates, leaving less unexplained heterogeneity. Consequently, a few and clearly-defined PK clusters can be expected (e.g. one group with higher elimination rate and another

with smaller elimination rate). The PD mechanisms instead have higher unexplained heterogeneity. It is therefore possible that, within each of the estimated PK groups, the underlying PD parameters differentiate more among the patients belonging to the same PK group, in order to capture unobserved or unobservable covariates relevant for the PD mechanisms. The PD sub-clusters will then be estimated, grouping patients with similar mechanisms of response.

5.5.1 Inference

In this section, we will briefly discuss how to make inference on the hierarchical model defined by expressions (5.3)-(5.5) for the data; model (5.13) for the random coefficients; and priors (5.15) and (5.16) for the unknown parameters. The main steps for making inference has already been discussed in Subsection 3.4.3, to which we refer together with the appendix of this chapter, where details are provided. As further reference, see, for example, Escobar and West (1995), Neal (2000) and, for PKPD models, Walker and Wakefield (1998).

Integrating out the unknown P and using the Pòlya urn scheme (Blackwell and MacQueen, 1973), described in **Chapter 1** by expression (1.11), we can obtain:

$$p\left(\boldsymbol{\theta}_1,..,\boldsymbol{\theta}_N\right) = p_{0\theta}\left(\theta_1\right)\prod_{i=2}^{N}\left(\alpha_\theta p_{0\theta}\left(\theta_i\right) + \sum_{j<i}\delta_{\theta_j}\left(\theta_i\right)\right)$$

where $p_{0\theta}$ is the density associated with $P_{0\theta}$. With the marginalized model associated with a DP prior, the full conditional distribution for θ_i for the Gibbs sampling is given by:

$$\begin{aligned} &p\left(\boldsymbol{\theta}_i \mid \boldsymbol{\theta}_{(-i)}, \boldsymbol{\sigma}^2_{PK}, \boldsymbol{\sigma}^2_{PD}, w\right) \propto \\ &q_0 N_{2T}\left(w_i \mid f\left(\boldsymbol{\theta}_i\right), V\left(\boldsymbol{\sigma}^2_{PK}, \boldsymbol{\sigma}^2_{PD}\right)\right) p_{0\theta}\left(\boldsymbol{\theta}_i\right) + \sum_{j\neq i} q_j \delta_{\theta_j}\left(\theta_i\right) \end{aligned}$$

where

- $\boldsymbol{\theta}_{(-i)} = (\boldsymbol{\theta}_1,..,\boldsymbol{\theta}_{i-1},\boldsymbol{\theta}_{i+1},..,\boldsymbol{\theta}_N)$;

- w_i is the $2\,T \times 1$ vector collecting
  $(\log(y_{i,1}), .., \log(y_{i,T}), \log(z_{i,1}), .., \log(z_{i,T}))$;
- w collects all the w_i, $i = 1, .., 2N$;
- $V\left(\boldsymbol{\sigma}^2_{PK}, \boldsymbol{\sigma}^2_{PD}\right)$ is a $2\ T \times 2\ T$ diagonal matrix with $\boldsymbol{\sigma}^2_{PK}$ on the first T elements of the diagonal and $\boldsymbol{\sigma}^2_{PD}$ on the last T elements of the diagonal;
- $f(\boldsymbol{\theta}_i)$ is a $2\ T \times 1$ vector collecting $f_{PK}\left(t, d_i\boldsymbol{\theta}^{PK}_i\right)$ first, for $t = 1, .., T$, and, then,
  $f_{PD}\left(t, d_i\boldsymbol{\theta}^{PD}_i, f_{PK}\left(t, d_i\boldsymbol{\theta}^{PK}_i\right)\right)$, $t = 1, .., T$;
- $q_0 \propto \alpha_{0\theta} \int p(w_i \mid \theta_i)\, p_{0\theta}(\theta_i)\, d\theta_i$ where $p(w_i \mid \theta_i)$ is the density associated with the vector of observations, defined by the models for the data, given the parameters θ_i, and $\left(\boldsymbol{\sigma}^2_{PK}, \boldsymbol{\sigma}^2_{PD}\right)$; $q_j \propto p(w_i \mid \theta_j)$, whose dependence is omitted for simplicity of notation; and $q_0 + \sum_{j=1, j\neq 1}^{N} q_j = 1$.

However, it is not possible here to evaluate analytically q_0. Consequently, a Metropolis-Hastings algorithm is required. It is described in Appendix.

Let us now consider the EDP model, defined by (5.15). The joint distribution of the random coefficients is:

$$p(\boldsymbol{\theta}_1, .., \boldsymbol{\theta}_N) \propto p_{0\theta^{PK}}\left(\boldsymbol{\theta}^{PK}_1\right) \prod_{i=2}^{N} \left(\alpha_{\theta^{PK}} p_{0\theta^{PK}}\left(\boldsymbol{\theta}^{PK}_i\right) + \sum_{j<i} \delta_{\boldsymbol{\theta}^{PK}_j}\left(\boldsymbol{\theta}^{PK}_i\right)\right) \times \left(p_{0\theta^{PD}}\left(\boldsymbol{\theta}^{PD}_1 \mid \boldsymbol{\theta}^{PK}_i\right) \prod_{s=2}^{n_i} \left(\alpha_{\theta^{PD}} p_{0\theta^{PD}}\left(\boldsymbol{\theta}^{PD}_s \mid \boldsymbol{\theta}^{PK}_i\right) + \sum_{r<s} \delta_{\boldsymbol{\theta}^{PD}_r \mid \boldsymbol{\theta}^{PK}_i}\left(\boldsymbol{\theta}^{PD}_s \mid \boldsymbol{\theta}^{PK}_i\right)\right)\right)$$

where $n_i = \sum_{j<i} \delta_{\boldsymbol{\theta}^{PK}_j}\left(\boldsymbol{\theta}^{PK}_i\right)$. The full conditional distribution for $\boldsymbol{\theta}^{PK}_i$ for the Gibbs sampling is given by:

$$p\left(\boldsymbol{\theta}^{PK}_i \mid \boldsymbol{\theta}^{PK}_{(-i)}, \boldsymbol{\sigma}^2_{PK}, y_i\right) \propto q_0^{PK} N_T\left(\log(y_i) \mid f_{PK}\left(\boldsymbol{\theta}^{PK}_i\right), \boldsymbol{\sigma}^2_{PK} I_T\right) p_{0\theta^{PK}}\left(\boldsymbol{\theta}^{PK}_i\right) + \sum_{j\neq i} q_j^{PK} \delta_{\boldsymbol{\theta}^{PK}_j}\left(\boldsymbol{\theta}^{PK}_i\right)$$

where

- $\boldsymbol{\theta}_{(-i)}^{PK} = \left(\boldsymbol{\theta}_1^{PK}, .., \boldsymbol{\theta}_{i-1}^{PK}, \boldsymbol{\theta}_{i+1}^{PK}, .., \boldsymbol{\theta}_N^{PK}\right)$;
- $\log(y_i)$ is the $T \times 1$ vector collecting $(\log(y_{i,1}), .., \log(y_{i,T}))$;
- I_T is a $T \times T$ identity matrix;
- $f_{PK}\left(\boldsymbol{\theta}_i^{PK}\right)$ is a $T \times 1$ vector collecting $f_{PK}\left(t, d_i, \boldsymbol{\theta}_i^{PK}\right)$, $t = 1, .., T$
- $q_0^{PK} \propto \alpha_{0\theta^{PK}} \int p\left(\log(y_i) \mid \theta_i^{PK}\right) p_{0\theta^{PK}}\left(\theta_i^{PK}\right) d\theta_i^{PK}$; $q_j^{PK} \propto p\left(\log(y_i) \mid \theta_j^{PK}\right)$; and $q_0^{PK} + \sum_{j=1, j\neq 1}^{N} q_j^{PK} = 1$.

Call $\boldsymbol{\theta}_1^{PK**}, .., \boldsymbol{\theta}_n^{PK**}$ the unique values of $\boldsymbol{\theta}_1^{PK}$,...,$\boldsymbol{\theta}_N^{PK}$. For each $s = 1, .., n$, the full conditional distribution for $\boldsymbol{\theta}_i^{PD}$ for the Gibbs sampling is given by:

$$p\left(\boldsymbol{\theta}_i^{PD} \mid \boldsymbol{\theta}_s^{PK**}, \boldsymbol{\theta}_{(-i)}^{PD}, \boldsymbol{\sigma}_{PD}^2, z_i\right) \propto$$

$$q_0^{PD} N_T\left(\log(z_i) \mid f_{PD}\left(\boldsymbol{\theta}_i^{PD}\right), \boldsymbol{\sigma}_{PD}^2 I_T\right) p_{0\theta^{PD}}\left(\boldsymbol{\theta}_i^{PD} \mid \boldsymbol{\theta}_s^{PK**}\right) +$$

$$+ \sum_{j\neq i} q_j^{PD} \delta_{\boldsymbol{\theta}_j^{PD} \mid \boldsymbol{\theta}_s^{PK**}}\left(\boldsymbol{\theta}_i^{PD} \mid \boldsymbol{\theta}_s^{PK**}\right)$$

where

- $\boldsymbol{\theta}_{(-i)}^{PD} = \left(\boldsymbol{\theta}_1^{PD}, .., \boldsymbol{\theta}_{i-1}^{PD}, \boldsymbol{\theta}_{i+1}^{PD}, .., \boldsymbol{\theta}_N^{PD}\right)$;
- $\log(z_i)$ is the $T \times 1$ vector collecting $(\log(z_{i,1}), .., \log(z_{i,T}))$;
- $f_{PD}\left(\boldsymbol{\theta}_i^{PD}\right)$ is a $T \times 1$ vector collecting $f_{PD}\left(t, d_i, \boldsymbol{\theta}_i^{PD}, f_{PK}\left(t, d_i \boldsymbol{\theta}_i^{PK}\right)\right)$, $t = 1, .., T$
- $q_0^{PD} \propto \alpha_{0\theta^{PD}} \int p\left(\log(z_i) \mid \theta_i^{PD}\right) p_{0\theta^{PD}}\left(\theta_i^{PD} \mid \theta_i^{PK}\right) d\theta_i^{PD}$; $q_j^{PD} \propto p\left(\log(z_i) \mid \theta_j^{PD}, \theta_i^{PK}\right)$; and $q_0^{PD} + \sum_{j=1, j\neq 1}^{N} q_j^{PD} = 1$.

Again, with two nonlinear models, f_{PK} and f_{PD}, for the data, it is not possible to evaluate q_0^{PK} and q_0^{PD} analytically. Consequently, a Metropolis-Hastings algorithm is required, which is described in the appendix.

5.6 Simulation study

In this section, we present two separate simulation studies that illustrate behaviour of the proposed parametric, called *simulation study I*, and nonparametric models, called *simulation study II*. The data are simulated from the one-compartmental pharmacokinetic model described by equation (5.1) with the indirect and semi-mechanistic pharmacodynamic model described by equation (5.4). Two data sets are simulated. Both data sets are simulated assuming $N = 60$ patients with the same $T = 14$ measurement times (in hours), in particular:

$$t \in \{0, 0.125, 0.25, 0.5, 0.75, 1, 2, 3, 4, 6, 8, 12, 18, 24\}.$$

The parameter $\boldsymbol{I}_{\max;i}$ is the maximum fractional inhibition that can be produced by the drug. We assume that $\boldsymbol{I}_{\max;i} \equiv 1$ for every i, which means that high drug concentrations completely inhibit $\boldsymbol{k}_i$. We also assume that $\boldsymbol{\sigma}_{PK}^2$ and $\boldsymbol{\sigma}_{PD}^2$ are known.

The numerical solution of the non-linear PD differential equation is always based on Runge-Kutta 4th order method[3].

5.6.1 Simulation Study I: Parametric analysis

This subsection illustrates the impact of the prior specification on $\boldsymbol{\Sigma}$ of equation (5.6). We will take into considerations four different alternatives: IW prior, Block-diagonal IW prior, DIG prior and scale DIG prior.

[3]Implementation is done in *R*, where the numerical solution is found using the *odesolve* package (Woodrow, 2012).

The PK and PD patient-specific random coefficients are collected in the two following vectors:

$$\boldsymbol{\theta}_i^{PK} = \left(\log\left(\boldsymbol{V}_i\right), \log\left(\boldsymbol{k}_i\right)\right) \quad \text{and}$$

$$\boldsymbol{\theta}_i^{PD} = \left(\log\left(\boldsymbol{Y}_{\text{eff};0,i}\right), \log\left(\boldsymbol{IC}_{50;i}\right), \log\left(\boldsymbol{k}_{\text{out};i}\right)\right)$$

Let us call $\boldsymbol{\theta}_i$ the vector collecting $\boldsymbol{\theta}_i^{PK}$ and $\boldsymbol{\theta}_i^{PD}$:

$$\boldsymbol{\theta}_i = \left(\log\left(\boldsymbol{V}_i\right), \log\left(\boldsymbol{k}_i\right), \log\left(\boldsymbol{Y}_{\text{eff};0,i}\right), \log\left(\boldsymbol{IC}_{50;i}\right), \log\left(\boldsymbol{k}_{\text{out};i}\right)\right)$$

Patient-specific coefficients are simulated from the Gaussian population distribution:

$$\boldsymbol{\theta}_i \mid \boldsymbol{\Sigma} \sim N\left(\log\left(\bar{\mu}\right), \bar{\Sigma}\right)$$

$$\text{where } \bar{\mu} = \begin{pmatrix} 35 \\ 0.2 \\ 50 \\ 0.5 \\ 0.3 \end{pmatrix} \text{ and } \bar{\Sigma} = \begin{pmatrix} 0.22 & 0.01 & -0.04 & 0.05 & -0.01 \\ 0.01 & 0.00 & -0.02 & 0.00 & -0.00 \\ -0.04 & -0.02 & 0.16 & 0.01 & 0.01 \\ 0.05 & 0.00 & 0.01 & 0.02 & -0.00 \\ -0.01 & -0.00 & 0.01 & -0.00 & 0.00 \end{pmatrix}.$$

The value of $\bar{\Sigma}$ has been simulated from

$$IW\left(7, \begin{pmatrix} 0.75 & 0 & 0 & 0.25 & 0 \\ 0 & 0.0075 & 0 & 0 & 0 \\ 0 & 0 & 0.75 & 0 & 0 \\ 0.25 & 0 & 0 & 0.0015 & 0 \\ 0 & 0 & 0 & 0 & 0.0075 \end{pmatrix}\right).$$

The PK and PD observations are then simulated accordingly. Four different priors are considered. The population mean $\boldsymbol{\mu}$ has always the following Gaussian prior:

$$\boldsymbol{\mu} \mid \boldsymbol{\Sigma} \sim N\left(\bar{\mu}, \frac{\boldsymbol{\Sigma}}{k_0}\right), \text{ with } k_0 = 5.$$

On $\boldsymbol{\Sigma}$, we consider the following four different specifications:

- IW prior: $\boldsymbol{\Sigma} \sim IW\left(10, 4 * \bar{\Sigma}\right)$
- Block-diagonal IW (BIW) prior:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_1 & \underline{0}_1 \\ \underline{0}_2 & \boldsymbol{\Sigma}_2 \end{pmatrix}$$

  where $\underline{0}_1$ and $\underline{0}_2$ are a 2×3 and a 3×2 matrix of zeros respectively; $\boldsymbol{\Sigma_1} \sim IW\left(7, 2 * \bar{\Sigma}_1\right)$ and independently $\boldsymbol{\Sigma_2} \sim IW\left(8, 2 * \bar{\Sigma}_1\right)$, where $\bar{\Sigma}_1$ is the top-left 2×2 sub-matrix of $\bar{\Sigma}$ and $\bar{\Sigma}_2$ is the bottom-right 3×3 sub-matrix of $\bar{\Sigma}$.

- DIG prior:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\sigma}_1^2 & 0 & 0 & 0 & 0 \\ 0 & \boldsymbol{\sigma}_2^2 & 0 & 0 & 0 \\ 0 & 0 & \boldsymbol{\sigma}_3^2 & 0 & 0 \\ 0 & 0 & 0 & \boldsymbol{\sigma}_4^2 & 0 \\ 0 & 0 & 0 & 0 & \boldsymbol{\sigma}_5^2 \end{pmatrix}$$

  where[4] $\boldsymbol{\sigma}_k^2 \overset{indep}{\sim} IW\left(6, \bar{\sigma}_k^2\right), k = 1, .., 5.$

- sDIG prior:

[4] Univariate IW distributions are equivalent to IG distributions. We prefer to state here the DIG prior in terms of IW distributions instead of IG distributions to avoid the need of introducing the chosen parametrization for the IG distribution.

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\xi}_1^2\boldsymbol{\sigma}_1^2 & 0 & 0 & 0 & 0 \\ 0 & \boldsymbol{\xi}_1^2\boldsymbol{\sigma}_2^2 & 0 & 0 & 0 \\ 0 & 0 & \boldsymbol{\xi}_2^2\boldsymbol{\sigma}_3^2 & 0 & 0 \\ 0 & 0 & 0 & \boldsymbol{\xi}_2^2\boldsymbol{\sigma}_4^2 & 0 \\ 0 & 0 & 0 & 0 & \boldsymbol{\xi}_2^2\boldsymbol{\sigma}_5^2 \end{pmatrix}$$

where $\boldsymbol{\sigma}_k^2 \overset{indep}{\sim} IW\left(6, \bar{\sigma}_k^2\right), k = 1, .., 5$ and, independently, $\boldsymbol{\xi}_r^2 \overset{indep}{\sim} IW\left(3, 1\right), r = 1, 2$.

The estimation algorithms are based on Metropolis Hastings-within-Gibbs sampling algorithm provided in Appendix. For all the four priors, the number of iterations is set up to $10,000$ with 10% of burn-in and the thinning is set up to 2. The corresponding three MCMC are all initialized with the true population means for the patient-specific random coefficients.

The impact of the variance specification is here done by considering the MCMC means and their MCMC errors. The MCMC errors are computed accordingly to Sokal (1989). They are calculated by truncating the infinite sum of theoretical correlations and plugging-in empirical correlations. Tables 5.1-5.5 show the MCMC estimates of the posterior means $E\left(\boldsymbol{\theta}_{i,j} \mid y_{1:N,1:T}, z_{1:N,1:T}\right)$ together with their MCMC errors (in parenthesis) for the first 10 patients ($i = 1, .., 10$).

Discussion of the results

As said, the patient-specific coefficients have been simulated using a full covariance matrix, Σ, simulated by an IW distribution. Therefore, it would be reasonable to expect that the model with the IW prior performs better than the others. Indeed, for the particular simulation that we have carried out, the best results -in terms of closeness of the fitted means to their true means- are mainly obtained either with an IW prior or with a sDIG prior. In particular, the estimated means obtained with the sDIG prior are closer to the true than the one obtained with the IW prior in

Patient	IW	BIW	DIG	sDIG	True value
1	35.03	32.09	33.90	31.62	34.53
	(0.02)	(0.02)	(0.01)	(0.01)	
2	35.23	33.52	33.97	32.94	45.05
	(0.02)	(0.00)	(0.01)	(0.00)	
3	35.06	32.18	33.95	31.67	40.16
	(0.04)	(0.06)	(0.01)	(0.01)	
4	34.48	31.84	33.64	20.95	15.67
	(0.00)	(0.01)	(0.01)	(0.00)	
5	56.85	32.86	40.86	32.27	55.62
	(0.00)	(0.00)	(0.00)	(0.00)	
6	35.22	32.94	33.98	32.94	38.94
	(0.01)	(0.00)	(0.01)	(0.00)	
7	35.20	31.97	33.98	31.67	57.74
	(0.02)	(0.00)	(0.04)	(0.00)	
8	34.97	32.08	33.84	31.55	27.14
	(0.02)	(0.01)	(0.03)	(0.01)	
9	23.01	32.21	33.84	31.77	26.48
	(0.00)	(0.02)	(0.01)	(0.00)	
10	35.13	32.20	33.94	31.64	28.66
	(0.02)	(0.02)	(0.00)	(0.01)	

Table 5.1: MCMC approximation of the posterior mean of V_i, for the patients $i = 1, .., 10$, for the simulation study I, under different priors, i.e. Inverse Wishart (IW), Blocked-Inverse Wishart (BIW), Diagonal Inverse Gamma (DIG) and scaled-Diagonal Inverse Gamma (sDIG). In parenthesis, the MCMC standard error.

Patient	IW	BIW	DIG	sDIG	True value
1	0.21	0.22	0.21	0.22	0.21
	(0.00)	(0.00)	(0.00)	(0.00)	
2	0.19	0.19	0.20	0.19	0.18
	(0.00)	(0.00)	(0.00)	(0.00)	
3	0.21	0.22	0.21	0.22	0.20
	(0.00)	(0.00)	(0.00)	(0.00)	
4	0.14	0.14	0.14	0.19	0.19
	(0.00)	(0.00)	(0.00)	(0.00)	
5	0.20	0.22	0.20	0.22	0.19
	(0.00)	(0.00)	(0.00)	(0.00)	
6	0.19	0.19	0.19	0.19	0.18
	(0.00)	(0.00)	(0.00)	(0.00)	
7	0.25	0.26	0.25	0.26	0.22
	(0.00)	(0.00)	(0.00)	(0.00)	
8	0.16	0.16	0.16	0.16	0.18
	(0.00)	(0.00)	(0.00)	(0.00)	
9	0.19	0.16	0.16	0.16	0.18
	(0.00)	(0.00)	(0.00)	(0.00)	
10	0.17	0.18	0.17	0.18	0.19
	(0.00)	(0.00)	(0.00)	(0.00)	

Table 5.2: MCMC approximation of the posterior mean of k_i, for the patients $i = 1, .., 10$, for the simulation study I, under different priors, i.e. Inverse Wishart (IW), Blocked-Inverse Wishart (BIW), Diagonal Inverse Gamma (DIG) and scaled-Diagonal Inverse Gamma (sDIG). In parenthesis, the MCMC standard error.

Patient	IW	BIW	DIG	sDIG	True value
1	54.20	58.57	53.36	57.99	41.21
	(0.03)	(0.15)	(0.25)	(0.02)	
2	54.42	56.58	53.44	57.22	71.02
	(0.04)	(0.00)	(0.27)	(0.00)	
3	54.33	57.81	53.47	57.99	35.58
	(0.03)	(0.10)	(0.27)	(0.03)	
4	54.39	60.48	53.83	93.43	48.83
	(0.00)	(0.13)	(0.26)	(0.01)	
5	38.23	60.06	45.23	58.79	62.77
	(0.00)	(0.00)	(0.00)	(0.00)	
6	54.54	57.22	54.55	57.22	70.01
	(0.01)	(0.00)	(0.26)	(0.00)	
7	54.69	57.83	53.70	57.96	28.88
	(0.06)	(0.03)	(0.29)	(0.00)	
8	54.19	61.45	53.64	58.14	73.56
	(0.02)	(0.05)	(0.25)	(0.01)	
9	74.97	62.09	53.97	58.17	98.18
	(0.00)	(0.10)	(0.37)	(0.00)	
10	54.28	62.04	56.43	57.98	87.82
	(0.03)	(0.01)	(0.00)	(0.03)	

Table 5.3: MCMC approximation of the posterior mean of $Y_{\text{eff};0,i}$, for the patients $i = 1, .., 10$, for the simulation study I, under different priors, i.e. Inverse Wishart (IW), Blocked-Inverse Wishart (BIW), Diagonal Inverse Gamma (DIG) and scaled-Diagonal Inverse Gamma (sDIG). In parenthesis, the MCMC standard error.

Patient	IW	BIW	DIG	sDIG	True value
1	0.53	0.85	0.57	0.67	0.40
	(0.00)	(0.01)	(0.00)	(0.01)	
2	0.42	0.79	0.44	0.48	0.57
	(0.00)	(0.00)	(0.00)	(0.00)	
3	0.69	1.17	0.66	0.74	0.56
	(0.02)	(0.06)	(0.00)	(0.01)	
4	0.42	0.49	0.42	0.68	0.41
	(0.00)	(0.01)	(0.00)	(0.00)	
5	0.46	0.74	0.83	0.51	0.52
	(0.00)	(0.00)	(0.00)	(0.00)	
6	0.35	0.48	0.39	0.48	0.54
	(0.00)	(0.00)	(0.00)	(0.00)	
7	0.68	1.45	0.69	0.77	0.49
	(0.00)	(0.00)	(0.01)	(0.00)	
8	0.37	0.44	0.39	0.47	0.61
	(0.00)	(0.00)	(0.00)	(0.00)	
9	0.52	0.34	0.36	0.41	0.49
	(0.00)	(0.00)	(0.01)	(0.00)	
10	0.33	0.32	0.41	0.44	0.60
	(0.00)	(0.00)	(0.00)	(0.00)	

Table 5.4: MCMC approximation of the posterior mean of $IC_{50;i}$, for the patients $i = 1, .., 10$, for the simulation study I, under different priors, i.e. Inverse Wishart (IW), Blocked-Inverse Wishart (BIW), Diagonal Inverse Gamma (DIG) and scaled-Diagonal Inverse Gamma (sDIG). In parenthesis, the MCMC standard error.

Patient	IW	BIW	DIG	sDIG	True value
1	0.33	0.66	0.41	0.45	0.29
	(0.00)	(0.02)	(0.01)	(0.01)	
2	0.25	0.02	0.30	0.30	0.29
	(0.00)	(0.00)	(0.01)	(0.00)	
3	0.37	0.60	0.44	0.49	0.30
	(0.00)	(0.02)	(0.00)	(0.01)	
4	0.31	0.63	0.32	0.30	0.30
	(0.00)	(0.01)	(0.00)	(0.00)	
5	0.29	0.32	0.28	0.38	0.29
	(0.00)	(0.00)	(0.00)	(0.00)	
6	0.21	0.30	0.31	0.30	0.31
	(0.00)	(0.00)	(0.00)	(0.00)	
7	0.39	0.72	0.43	0.50	0.28
	(0.01)	(0.01)	(0.01)	(0.00)	
8	0.29	0.42	0.33	0.33	0.30
	(0.00)	(0.00)	(0.00)	(0.00)	
9	0.32	0.56	0.32	0.26	0.29
	(0.00)	(0.01)	(0.00)	(0.00)	
10	0.35	0.64	0.40	0.37	0.30
	(0.01)	(0.01)	(0.00)	(0.01)	

Table 5.5: MCMC approximation of the posterior mean of $k_{\text{out};i}$, for the patients $i = 1, .., 10$, for the simulation study I, under different priors, i.e. Inverse Wishart (IW), Blocked-Inverse Wishart (BIW), Diagonal Inverse Gamma (DIG) and scaled-Diagonal Inverse Gamma (sDIG). In parenthesis, the MCMC standard error.

presence of "outliers" among the $\boldsymbol{\theta}_i$, e.g. see the V_4 in Table 5.1. This is because the sDIG prior allows the elements of the vector $\boldsymbol{\theta}_i$ to be more "free" than with the IW prior, which can be more appropriate if there is just one parameter of the patient-specific 5-dimensional vector $\boldsymbol{\theta}_i$ which is "extreme". The sDIG prior imposes some dependence but not through a full covariance matrix. With the sDIG, there is only one parameter defining the correlation among the PK coefficients and one for the PD coefficients. With a full covariance matrix, like with an IW prior, each element of $\boldsymbol{\theta}_i$, $\boldsymbol{\theta}_{i;j}$, is linked with all the other $\boldsymbol{\theta}_{i;s}, s \neq j$ through the non-zero covariances.

The results also show that it is harder to obtain good results with a DIG prior. This is because of the assumed independence across all the elements of $\boldsymbol{\theta}_i$. This preserves each element of $\boldsymbol{\theta}_i$ to use the information contained in the other elements. Finally, the worst results are obtained for $\boldsymbol{Y}_{\text{eff};0,i}$, whose true values are quite heterogeneous, e.g. in the Table 5.3 it ranges from 18.2154 (patient 17) to 98.1812 (patient 9).

5.6.2 Simulation Study II: Nonparametric EDP population distribution

We now present a simulation study to illustrate the behaviour of the nonparametric model proposed in Section 5.5, aiming to underline some of the features of the EDP prior. For simplicity, we assume that the only patient-specific unknown parameters are $\boldsymbol{V}_i$ and $\boldsymbol{IC}_{50;i}$. The other parameters are supposed to be constant across patients and their value is known, e.g. fixed from some external sources like previous studies (Cheung *et al.*, 2008; Gillespie, 2009). Thus, $\boldsymbol{\theta}_i^{PK} = \log(\boldsymbol{V}_i)$ and $\boldsymbol{\theta}_i^{PD} = \log(\boldsymbol{IC}_{50;i})$. The parameters for simulating this second data set are defined as follows:

- The common values for all the patients are fixed to: $k_{\text{in};i} = 0.2$, $Y_{\text{eff};i} = 50$ and $k_{\text{out};i} = 0.3$, for all $i = 1, .., N$.

- The patient-specific parameters are defined with the following structure. The data are simulated with two PK groups: $V_1^* = 45$, and $V_2^* = 25$, which corresponds to:
  - Two PD groups within the first PK group, $V_1^* = 45$, with the following chosen parameters: $\left(IC^*_{50;1} \mid V_1^*\right) = 0.7$ and $\left(IC^*_{50;2} \mid V_1^*\right) = 20$.
  - Two PD groups for the second PK group, $V_2^* = 25$: with the following chosen parameters: $\left(IC^*_{50;1} \mid V_2^*\right) = 0.3$ and $\left(IC^*_{50;1} \mid V_2^*\right) = 0.5$.

This structure of the simulated data is represented in Figure 5.2. The two PK groups correspond to the two "global" clusters and, inside each of them, we have two PD sub-groups, representing the "local" clustering structure.

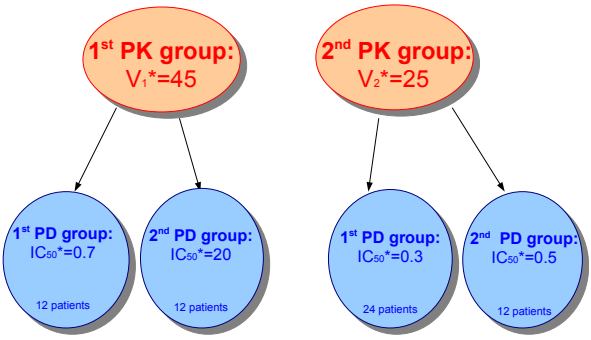


Figure 5.2: Simulated population structure

This choice for the parameters implies that the drug is very effective for most of the patients except for one group, namely the one with parameter $IC_{50} = 20$ associated with a much higher value of the concentration

required to obtain half of the maximum effect. Consequently, for this group, a higher dose is required to obtain the same effect. As previously discussed, this peculiar value could be just a signal of the lack of adequacy of the chosen PD model for this sub-group. Each group includes 12 patients, except for the group with $V_2^* = 25$ and $IC^*_{50;1} = 0.3$, which is of 24 patients. Finally, the standard deviations are fixed to $\sigma_{PK} = 0.1$ and $\sigma_{PD} = 0.1$. Observations on concentrations and for the response variables are then simulated accordingly.

Three models are implemented using these data: the parametric[5] PKPD model with Gaussian population distribution; the nonparametric PKPD based on the DP and the nonparametric PKPD based on the EDP. In all of these three models, all the parameters are known and fixed equal to the true values except from the $\boldsymbol{V}_i, \boldsymbol{IC}_{50;i}$, $i = 1, .., N$, which constitute the objects of the inference. The base measure of the DP and EDP are based on the multivariate Gaussian distribution with the true overall population means and with $Cov\left(\boldsymbol{V}_i, \boldsymbol{IC}_{50;i}\right) = 0.25$, $Var\left(\boldsymbol{V}_i\right) = 100$ and $Var\left(\boldsymbol{IC}_{50}\right) = 500$, which is also the distribution used for the parametric case. In particular, the base measure of the DP is a multivariate Gaussian distribution and for the EDP the two base measures are the corresponding marginal and conditional distributions. The EDP precision parameters are fixed to $\alpha_{PK} = 50$ and $\alpha_{PD} = 1$, and the DP precision parameter is the mean of the two precision parameters of the EDP.

Notice that the "mean" population distribution of the patient-specific random coefficients is the same for all the three models. They *just* differ in the uncertainty around the "mean" distribution. For the parametric model, the population distribution corresponds to this "mean" population distribution. In the DP and the EDP models, there is instead uncertainty around the "mean" population distribution. Within the DP

[5]Notice that the parametric model here implemented is slightly different from the one of the previous illustrative example, i.e. all the variances and covariances are here fixed.

model, the uncertainty is equal both for the distributions associated with the PK and for the PD patient-specific random coefficients. Within the EDP model, it is also possible to diversify the degree of uncertainty on the latent population distributions of the PK and PD patient-specific random coefficients.

Inference is carried out based on posterior MCMC simulation. In particular, the MCMC scheme to compute posterior distributions is based on the algorithm 6 described in Neal (2000), which is a Metropolis-Hastings algorithm[6] with candidates drawn from the prior. Details are provided in the Appendix. The number of iterations is set up to $100,000$ with 10% of burn-in. The corresponding three MCMC are initialized with the overall population means for the patient-specific random coefficients. The results are summarized in Figures $5.3, 5.14, 5.17$, which display the posterior distributions for 5 representative patients (out of the total of 60 patients) for the three models: respectively, using the EDP prior, using the DP prior and finally using the Gaussian population distribution. In particular, Figures $5.3, 5.14, 5.17$ show the posterior distributions of $\exp(\theta_{PK,i}) \equiv \boldsymbol{V}_i$ (on the left-hand-side) and $\exp(\theta_{PD,i}) \equiv \boldsymbol{IC}_{50}$ (on the right-hand-side) for the five representative patients. Each of these representative patient represents a sub-group of 12 patients. In particular, the first one represents the first PD group within the first PK group; the second one represents the second PD group within the first PK group; the third and the forth ones represent the first PD group within the second PK group; and the last one represents the second PD group within the second PK group.

Discussion of the results

As previously said, the data have been simulated so that, for each PK cluster, there are two PD clusters of random-coefficients. Therefore, if

[6] To include also the observational variances in the set of random parameters to be inferred, one could use a Metropolis-Hastings-within-Gibbs-sampling and drawing the variances from their full conditionals. See Appendix for similar algorithm.

PK and PD clusters cannot be separated, as when using the DP prior, it is reasonable to expect either a proliferation of the PK clusters to accommodate the PD clusters or the correct inference on the PK cluster but the shrinkage of the PD patient-specific random coefficients within each PK group. With the chosen precision parameter ($\alpha = 25$ for the DP model), the first case is observed and four clusters are found. This result depends on the chosen value for α: decreasing α, the expected number of clusters decreases and a proliferation of the PK clusters to accommodate the PD clusters could be obtained. For example, we have repeated the procedure setting $\alpha = 6$ and we have observed a smaller number of clusters, 3, and a shrinkage of some of the PD patient-specific random coefficients. See Figure 5.20 in Appendix. Therefore, since a DP prior cannot give a flexible and asymmetric clustering structure for PK and PD, depending on the particular α, at least one of the two PK and PD clustering structures for patient-specific random coefficients needs to adequate to the other. Instead, the parametric case does not allow, in principle, to clusters the patient-specific random coefficients, since their values in each draw of the MCMC cannot be exactly the same, due the continuity of the Gaussian distribution. In practice, the patient-specific random coefficients can be just slightly different among patients. As a results, the clustering structure for the PK patient-specific random coefficients is well represented. Moreover, although the Gaussian distribution does not allow for outliers, allowing for higher variances of the patient-specific random coefficients than the true ones, it is still possible to recognize the "extreme" value in the PD clustering structure. Again, the role of $\boldsymbol{\Sigma}$ is critical. Thanks to its flexible clustering structure, the EDP is instead able to capture easily the hierarchical structure, in the sense that it is not so much depending on the chosen precision parameter as for the DP model and is not depending on the variances of the patient-specific random coefficients as for in the parametric case.

In Figure 5.6, prior and posterior distributions for $\boldsymbol{V}_i$ and $\boldsymbol{IC}_{50;i}$

Figure 5.3: Posterior distributions with EDP prior

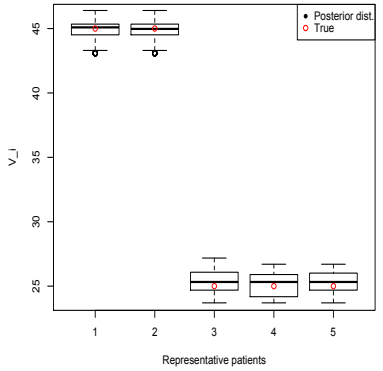


Figure 5.4: $\boldsymbol{V}_i$

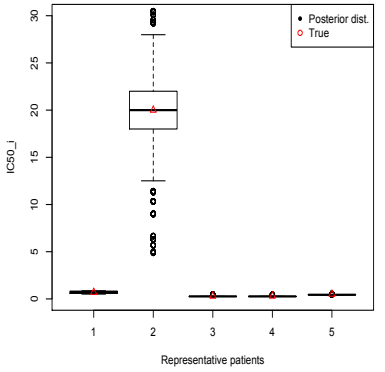


Figure 5.5: $\boldsymbol{IC}_{50;i}$

for a specific patient, i.e. patient 1, are shown. It is possible to see that the posterior distributions are much more concentrated on the true values. Figure 5.9 shows the prior and posterior predictive distribution for $log(y_{i,t})$ and $log(z_{i,t})$ for the same patient and time 1.

Figure 5.12 and 5.13 shows the clustering structure for the PK and PD parameters, respectively. True clustering structure and estimated clustering structures are reported. A darker color (i.e. red) means higher probability for two patients to be clustered together according the values of their PK or PD parameters. A lighter color (i.e. light yellow) means lower probability to be clustered together.

5.7 Discussion and conclusion

The Bayesian parametric approach allows us to naturally treat random coefficients models and to borrow strength across patients. This chapter has discussed the primary role of $\boldsymbol{\Sigma}$, especially within the para-

Figure 5.6: EDP: Prior and Posterior distribution for V_i and $IC_{50,i}$, patient 1

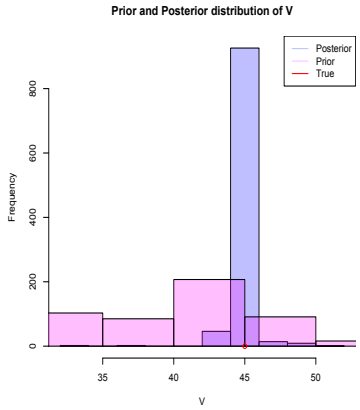


Figure 5.7: $\boldsymbol{V}_i$

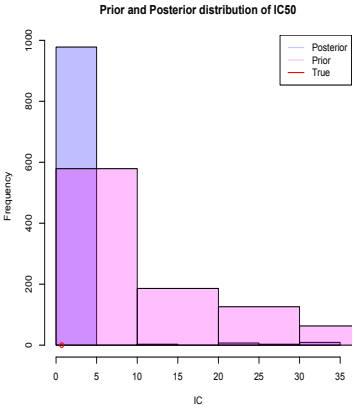


Figure 5.8: $\boldsymbol{IC}_{50;i}$

metric setting. We proposed a simple and parsimonious prior specification for the covariance matrix covariance $\boldsymbol{\Sigma}$ which allows for local clusters.

Whenever the data have outliers, skewness or multimodality, the parametric approach is too restrictive. A nonparametric prior can favor flexible clustering structure and relax rigid parametric assumptions. We proposed to use the EDP instead of the commonly-used DP as prior on the latent population distribution. The EDP is particularly appropriate for the PKPD applications since it allows for asymmetric and sequential treatment of the PK and PD patient-specific random coefficients. It implies a global clustering structure for the PK patient-specific random coefficients and, conditionally on this, a local clustering structure within each PK group. The adequacy of this hierarchical clustering structure is problem specific but it is reasonable whenever one has different degrees of confidence between the PK and the PD model.

The main limitation of the proposed nonparametric model is a com-

Figure 5.9: EDP: Prior and Posterior predictive distribution for $log(y_{1,1})$ and $log(z_{1,1})$ (patient 1, time 1)

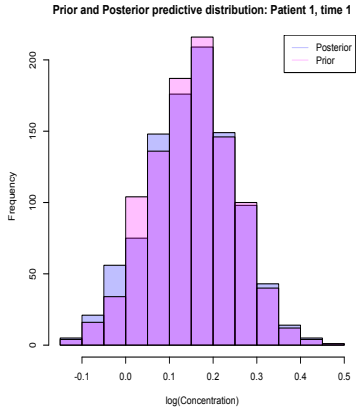


Figure 5.10: $log(y_{1,1})$

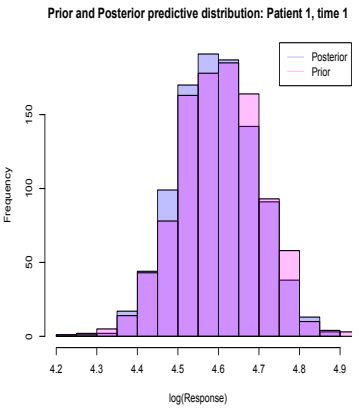


Figure 5.11: $log(z_{1,1})$

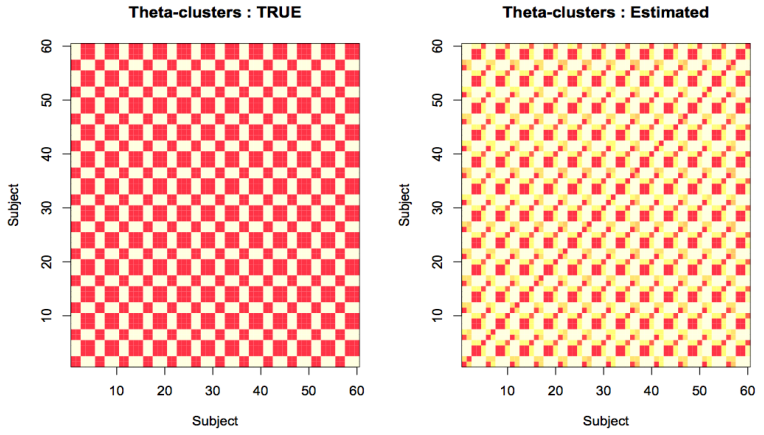


Figure 5.12: Clustering structure: θ clusters (true vs estimated with EDP prior)

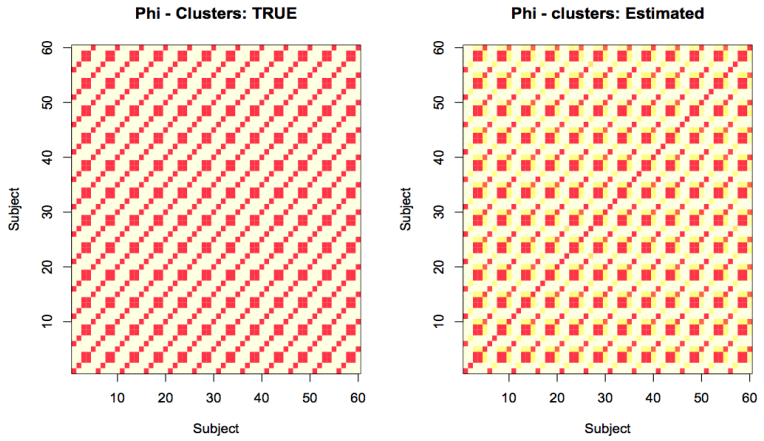


Figure 5.13: Clustering structure: ϕ clusters (true vs estimated with EDP prior)

Figure 5.14: Posterior distributions with DP prior

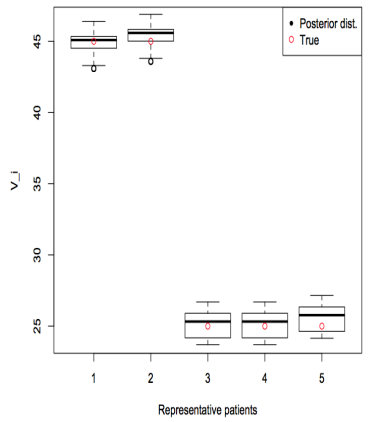


Figure 5.15: $\boldsymbol{V}_i$

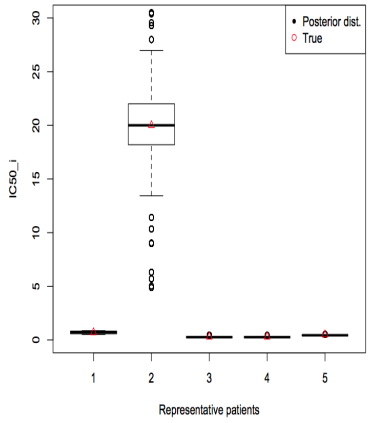


Figure 5.16: $\boldsymbol{IC}_{50;i}$

mon drawback of all Bayesian nonparametric PKPD models: they are computation-intensive.

Figure 5.17: Posterior distributions with Gaussian population distribution

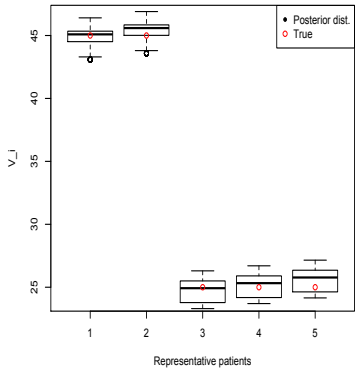


Figure 5.18: $\boldsymbol{V}_i$

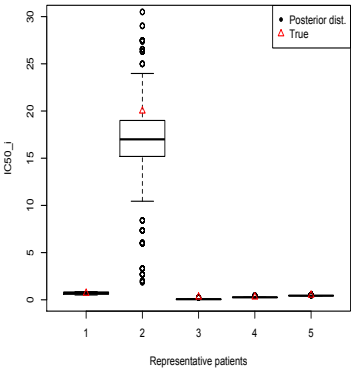


Figure 5.19: $\boldsymbol{IC}_{50;i}$

The proposed simulation studies had a pure illustrative purpose. More simulation studies, with more flexibility over the hyper-parameters, should be carried out. Moreover, it would be particularly interesting to analyze the proposed models on some real data sets, which I had the opportunity to study during my stay at Novartis, but cannot be reproduced due to confidentiality reasons.

5.8 Appendix: Computational details

5.8.1 Parametric models

Let us consider the deterministic models (5.2) and (5.4) for the first level of the hierarchy represented by models (5.3) and (5.5). Let us consider the following Gaussian population distribution of the patient-specific random coefficients, $\boldsymbol{\theta}_i$:

$$\boldsymbol{\theta}_i \equiv \left(\log\left(\boldsymbol{V}_i\right), \log\left(\boldsymbol{k}_i\right), \log\left(\boldsymbol{Y}_{\text{eff};0,i}\right), \log\left(\boldsymbol{IC}_{50;i}\right), \log\left(\boldsymbol{k}_{\text{out};i}\right)\right) \mid \boldsymbol{\mu}, \boldsymbol{\Sigma} \overset{\text{iid}}{\sim} N_5\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$$

$$\boldsymbol{\mu} \mid \boldsymbol{\Sigma} \sim N_5\left(\mu_0, \frac{\boldsymbol{\Sigma}}{k_0}\right)$$

Depending on the specification on $\boldsymbol{\Sigma}$ and its prior, the algorithm will be different. All of them are based on a Metropolis-Hastings-within-Gibbs sampling algorithm. The values of the hyper-parameters, e.g.k_0 and μ_0, have been specified in Subsection 5.6.1.

Metropolis-Hastings-within-Gibbs sampling algorithm for the model with IW prior on $\boldsymbol{\Sigma}$

Let us assume that:

$$\boldsymbol{\Sigma} \sim IW\left(\nu_0, \Sigma_0\right)$$

The following Metropolis-Hastings-within-Gibbs sampling algorithm is used for the posterior inference:

Algorithm 5.8.1 *For each iteration, s, $s = 2, .., S$:*

1. *For $i = 1, .., N$, draw $\theta_i^{(s)}$ from $p\left(\theta_i \mid \mu^{(s-1)}, \Sigma^{(s-1)}, \theta_{1:i-1}^{(s)}, \theta_{i+1:N}^{(s-1)}\right)$ as follows:*

   - *Draw a candidate θ_i^* from $N_5\left(\mu^{(s-1)}, \Sigma^{(s-1)}\right)$, where $\theta_i^* = \left(\theta_i^{PK*}, \theta_i^{PD*}\right)$.*
   - *Compute the acceptance probability:*

$$\alpha\left(\theta_i^*, \theta_i^{s-1}\right) = \min\left(1, \frac{\prod_{i=1}^{N} N_T\left(y_i \mid f_{PK}^*, \sigma_{PK}^2 I_T\right) \prod_{i=1}^{N} N_T\left(z_i \mid f_{PD}^*, \sigma_{PD}^2 I_T\right)}{\prod_{i=1}^{N} N_T\left(y_i \mid f_{PK}^{(s-1)}, \sigma_{PK}^2 I_T\right) \prod_{i=1}^{N} N_T\left(z_i \mid f_{PD}^{(s-1)}, \sigma_{PD}^2 I_T\right)}\right)$$

   where

- y_i *and* z_i *are the* T*-dimensional vectors collecting* $y_{i,t}$ *and* $z_{i,t}$ *respectively;*
- $f^*_{PK} \equiv f_{PK}\left((1,..,t,..,T), d_i, \theta_i^{PK*}\right)$ *is the* T*-dimensional vector of estimated concentrations for the patient* i *with parameters* θ_i^{PK*}*;*
- *similarly,* $f_{PK}^{(s-1)} \equiv f_{PK}\left((1,..,T), d_i, \theta_i^{PK(s-1)}\right)$
- $f^*_{PD} \equiv f_{PK}\left(\begin{pmatrix} 1 & .. & t & .. & T \end{pmatrix}, d_i, \theta_i^{PD*}, f^*_{PK}\right)$ *is the* T*-dimensional vector of estimated responses for the patient* i *with parameters* θ_i^{PD*}*;*
- *similarly,* $f_{PD}^{(s-1)} \equiv f_{PK}\left((1,...,T), d_i, \theta_i^{PD(s-1)}, f^*_{PK}\right)$
- *The symbol* N *is used to denote both the distribution function and the density function of a Gaussian distribution.*

- *With probability* $\alpha\left(\theta_i^*, \theta_i^{s-1}\right)$*, set* $\theta_i^{(s)} = \theta_i^*$*, otherwise set* $\theta_i^{(s)} = \theta_i^{s-1}$

2. *Sample* $\Sigma^{(s)}$ *from* $p\left(\Sigma \mid \mu^{(s-1)}, \theta_i^{(s)}, i = 1,..,N\right)$*, where*

$$p\left(\Sigma \mid \mu^{(s-1)}, \theta_i^{(s)}, i = 1,..,N\right) = IW\left(\nu_n, \Sigma_n\right)$$

where

- $\nu_n = \nu_0 + N$
- $\Sigma_n = \left(\Sigma_0 + Q + \dfrac{k_0 N}{k_0 + N}\left(\bar{\theta}^{(s)} - \mu_0\right)\left(\bar{\theta}^{(s)} - \mu_0\right)'\right)^{-1}$

  where $Q = \sum_{i=1}^{N}\left(\theta_i^{(s)} - \bar{\theta}^{(s)}\right)\left(\theta_i^{(s)} - \bar{\theta}^{(s)}\right)'$ *and* $\bar{\theta}^{(s)} = \dfrac{1}{N}\sum_{j=1}^{N}\theta_j^{(s)}$

3. *Sample* $\mu^{(s)}$ *from* $p\left(\mu \mid \Sigma^{(s)}, \theta_i^{(s)}, i = 1,..,N\right)$*, where*

$$p\left(\mu \mid \Sigma^{(s)}, \theta_i^{(s)}, i = 1,..,N\right) = N\left(\mu_n, \frac{\Sigma^{(s)}}{k_n}\right)$$

where

- $\mu_n = \frac{k_0}{k_0+N}\mu_0 + \frac{N}{k_0+N}\left(\bar{\theta}^{(s)} - \mu_0\right)'$
- $k_n = k_0 + N$

Metropolis-Hastings-within-Gibbs sampling algorithm for the model with Block-diagonal IW prior on $\boldsymbol{\Sigma}$

Let us assume that:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_1 & \underline{0}_1 \\ \underline{0}_2 & \boldsymbol{\Sigma}_2 \end{pmatrix} \text{ with } \boldsymbol{\Sigma}_j \overset{indep}{\sim} IW\left(\nu_j, \Sigma_j\right)$$

The Metropolis-Hastings-within-Gibbs sampling algorithm is basically the same as the Algorithm 5.8.1, except for the point 2, which changes as follows

For each iteration, s, $s = 2, .., S$:

2. Sample $\Sigma_j^{(s)}$ from $p\left(\Sigma_j \mid \mu^{(s-1)}, \theta_i^{(s)}, i = 1, .., N\right)$, for $j = 1, 2$, and independently across j, where

$$p\left(\Sigma_j \mid \mu^{(s-1)}, \theta_i^{(s)}, i = 1, .., N\right) = IW\left(\nu_{j,n}, \Sigma_{j,n}\right)$$

where

- $\nu_{j,n} = \nu_j + N$
- $\Sigma_{j,n} = \left(\Sigma_j + Q_j + \frac{k_0 N}{k_0+N}\left(\bar{\theta}_j^{(s)} - \mu_{j,0}\right)\left(\bar{\theta}_j^{(s)} - \mu_{0,j}\right)'\right)^{-1}$

where $Q_j = \sum_{i=1}^{N}\left(\theta_{i,j}^{(s)} - \bar{\theta}_j^{(s)}\right)\left(\theta_{j,i}^{(s)} - \bar{\theta}_j^{(s)}\right)'$ and $\bar{\theta}_j^{(s)} = \frac{1}{N}\sum_{g=1}^{N}\theta_{j,g}^{(s)}$, and $\theta_{i,j}$ is the vector collecting the elements of θ_i which refers to the matrix j. In particular, $\theta_{i,1}$ collects the first two elements of θ_i and $\theta_{i,2}$ the remaining three elements. Similarly for $\mu_{j,0}$.

Metropolis-Hastings-within-Gibbs sampling algorithm for the model with DIG prior on $\boldsymbol{\Sigma}$

Let us assume that:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\sigma}_1^2 & 0 & 0 & 0 & 0 \\ 0 & \boldsymbol{\sigma}_2^2 & 0 & 0 & 0 \\ 0 & 0 & \boldsymbol{\sigma}_3^2 & 0 & 0 \\ 0 & 0 & 0 & \boldsymbol{\sigma}_4^2 & 0 \\ 0 & 0 & 0 & 0 & \boldsymbol{\sigma}_5^2 \end{pmatrix} \text{ with } \boldsymbol{\sigma}_j^2 \overset{indep}{\sim} IW\left(\nu_j, \sigma_{j,0}\right) \equiv IG\left(a_j, b_j\right)$$

Again, the Metropolis-Hastings-within-Gibbs sampling algorithm is basically the same as the Algorithm 5.8.1, except for the point 2, which changes as follows

For each iteration, s, $s = 2, .., S$:

2. Sample $\sigma_j^{2(s)}$ from $p\left(\sigma_j^2 \mid \mu^{(s-1)}, \theta_i^{(s)}, i = 1, .., N\right)$, for $j = 1, 2$, and independently across j, where

$$p\left(\sigma_j^2 \mid \mu^{(s-1)}, \theta_i^{(s)}, i = 1, .., N\right) = IW\left(\nu_{j,n}, \sigma_{j,n}\right)$$

where

- $\nu_{j,n} = \nu_j + N$
- $\sigma_{j,n}^2 = \left(\sigma_{j,0}^2 + Q_j + \dfrac{k_0 N}{k_0 + N}\left(\bar{\theta}_j^{(s)} - \mu_{j,0}\right)^2\right)^{-1}$

where $Q_j = \sum_{i=1}^{N}\left(\theta_{i,j}^{(s)} - \bar{\theta}_j^{(s)}\right)^2$ and $\bar{\theta}_j^{(s)} = \dfrac{1}{N}\sum_{g=1}^{N}\theta_{j,g}^{(s)}$, and $\theta_{i,j}$ is the jth element of θ_i.

Metropolis-Hastings-within-Gibbs sampling algorithm for the model with sDIG prior on $\boldsymbol{\Sigma}$

Let us assume that:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\xi}_1^2\boldsymbol{\sigma}_1^2 & 0 & 0 & 0 & 0 \\ 0 & \boldsymbol{\xi}_1^2\boldsymbol{\sigma}_2^2 & 0 & 0 & 0 \\ 0 & 0 & \boldsymbol{\xi}_2^2\boldsymbol{\sigma}_3^2 & 0 & 0 \\ 0 & 0 & 0 & \boldsymbol{\xi}_2^2\boldsymbol{\sigma}_4^2 & 0 \\ 0 & 0 & 0 & 0 & \boldsymbol{\xi}_2^2\boldsymbol{\sigma}_5^2 \end{pmatrix}$$

where $\boldsymbol{\sigma}_j^2 \overset{indep}{\sim} IW\left(\nu_j, \sigma_{j,0}^2\right), j = 1, .., 5$ and, independently, $\boldsymbol{\xi}_r^2 \overset{indep}{\sim} IW\left(\nu_{\xi;1}, \sigma_{\xi;1}^2\right), r = 1, 2.$

As for the previously discussed cases, the Metropolis-Hastings-within-Gibbs sampling algorithm is basically the same as the Algorithm 5.8.1, except for the point 2. In particular, the point 2 now includes the point 2 of the algorithm with a DIG prior, but also another sub-point for the draws of ξ_r. Therefore, the whole point 2 of the Algorithm 5.8.1 is substituted with the following:

For each iteration, s, $s = 2, .., S$:

2.a) Sample $\sigma_j^{2(s)}$ from $p\left(\sigma_j^2 \mid \mu^{(s-1)}, \theta_i^{(s)}, i = 1, .., N\right)$, for $j = 1, 2$, and independently across j, where

$$p\left(\sigma_j^2 \mid \mu^{(s-1)}, \theta_i^{(s)}, i = 1, .., N\right) = IW\left(\nu_{j,n}, \sigma_{j,n}\right)$$

where

- $\nu_{j,n} = \nu_j + N$
- $\sigma_{j,n}^2 = \left(\sigma_j^2 + Q_j + \dfrac{k_0 N}{k_0 + N}\left(\bar{\theta}_j^{(s)} - \mu_{j,0}\right)^2\right)^{-1}$

  where $Q_j = \sum_{i=1}^{N}\left(\theta_{i,j}^{(s)} - \bar{\theta}_j^{(s)}\right)^2$ and $\bar{\theta}_j^{(s)} = \dfrac{1}{N}\sum_{g=1}^{N}\theta_{j,g}^{(s)}$, and $\theta_{i,j}$ is the jth element of θ_i.

2.b) For $r = 1, 2$, sample $\xi_r^{2(s)}$ from

$$p\left(\xi_r^2 \mid \mu^{(s-1)}, \theta_i^{(s)}, i = 1, .., N, \sigma_j^{2(s)}, j = 1, ., 5\right) :$$

- Draw a candidate ξ_r^{2*} from $IW\left(\nu_{\xi;1}, \sigma_{\xi;1}^2\right)$.
- Compute the acceptance probability:

$$\alpha\left(\xi_i^{2*}, \xi^{2(s-1)}\right) = \min\left(1, \frac{\prod_{i=1}^{N} N_T\left(\theta_i^{(s)} \mid \mu^{(s-1)}, \Sigma^*\right)}{\prod_{i=1}^{N} N_T\left(\theta_i^{(s)} \mid \mu^{(s-1)}, \Sigma^{(s-1)}\right)}\right)$$

where

$$\Sigma^* = \begin{pmatrix} \boldsymbol{\xi}_1^{2*}\boldsymbol{\sigma}_1^{2(s)} & 0 & 0 & 0 & 0 \\ 0 & \boldsymbol{\xi}_1^{2*}\boldsymbol{\sigma}_2^{2(s)} & 0 & 0 & 0 \\ 0 & 0 & \boldsymbol{\xi}_2^{2*}\boldsymbol{\sigma}_3^{2(s)} & 0 & 0 \\ 0 & 0 & 0 & \boldsymbol{\xi}_2^{2*}\boldsymbol{\sigma}_4^{2(s)} & 0 \\ 0 & 0 & 0 & 0 & \boldsymbol{\xi}_2^{2*}\boldsymbol{\sigma}_5^{2(s)} \end{pmatrix}.$$

$$\Sigma^{2,(s-1)} = \begin{pmatrix} \boldsymbol{\xi}_1^{2(s-1)}\boldsymbol{\sigma}_1^{2(s)} & 0 & 0 & 0 & 0 \\ 0 & \boldsymbol{\xi}_1^{2(s-1)}\boldsymbol{\sigma}_2^{2(s)} & 0 & 0 & 0 \\ 0 & 0 & \boldsymbol{\xi}_2^{2(s-1)}\boldsymbol{\sigma}_3^{2(s)} & 0 & 0 \\ 0 & 0 & 0 & \boldsymbol{\xi}_2^{2(s-1)}\boldsymbol{\sigma}_4^{2(s)} & 0 \\ 0 & 0 & 0 & 0 & \boldsymbol{\xi}_2^{2(s-1)}\boldsymbol{\sigma}_5^{2(s)} \end{pmatrix}.$$

- With probability $\alpha\left(\xi_i^*, \xi^{s-1}\right)$, set $\xi_r^{(s)} = \xi_i^*$, otherwise set $\xi_i^{(s)} = \xi_i^{s-1}$

5.8.2 Nonparametric models

Let us consider again the deterministic models (5.2) and (5.4) for the first level of the hierarchy represented by models (5.3) and (5.5). Let us now consider the following nonparametric priors for the population distribution of the patient-specific random coefficients, $\boldsymbol{\theta}_i$:

$$\left(\boldsymbol{\theta}_i^{PK}, \boldsymbol{\theta}_i^{PD}\right) | \mathbf{P} = P \overset{\text{iid}}{\sim} P.$$

For the sake of simplicity, let us assume that all the other parameters are fixed and known. The Metropolis-Hastings algorithm is different depending the nonparametric assumption on $\mathbf{P}$.

Metropolis-Hastings algorithm for DP prior

Let us assume that:

$$\mathbf{P} \sim DP\left(\alpha P_0\right)$$

where $P_0 \equiv N_5\left(\mu, \Sigma\right)$.

Algorithm 5.8.2 *For each iteration, s, $s = 2, .., S$:*
For $i = 1, .., N$, sample $\theta_i^{(s)}$ from $p\left(\theta_i \mid \theta_{-i}^{(s-1)}\right)$,
where $\theta_{-i}^{(s-1)} = \theta_1^{(s-1)}, ..., \theta_{i-1}^{(s-1)}, \theta_{i+1}^{(s-1)}, .., \theta_N^{(s-1)}$.

- *Draw a candidate θ_i^* accordingly to the Pòlya urn distribution (Blackwell and MacQueen, 1973) described in* **Chapter 1** *by expression (1.11).*
  In particular, call θ_1^{**},..,θ_n^{**} the unique values among $\theta_1^{(s-1)}, .., \theta_{i-1}^{(s-1)}, \theta_{i+1}^{(s-1)}, .., \theta_N^{(s-1)}$.

$$\theta_i^* = \begin{cases} \theta^{new} \sim N_5\left(\mu, \Sigma\right), & w.p. \ \dfrac{\alpha}{\alpha + N - 1} \\ \theta_r^{**}, r = 1, .., n & w.p. \ \dfrac{\sum_{z=1}^{N-1} \delta_{\theta_i^{(s-1)} = \theta_r^{**}}}{\alpha + N - 1} \end{cases}$$

- *Compute the acceptance probability:*

$$\alpha\left(\theta_i^*, \theta_i^{s-1}\right) = \min\left(1, \frac{\prod_{i=1}^{N} N_T\left(y_i \mid f_{PK}^*, \sigma_{PK}^2 I_T\right) \prod_{i=1}^{N} N_T\left(z_i \mid f_{PD}^*, \sigma_{PD}^2 I_T\right)}{\prod_{i=1}^{N} N_T\left(y_i \mid f_{PK}^{(s-1)}, \sigma_{PK}^2 I_T\right) \prod_{i=1}^{N} N_T\left(z_i \mid f_{PD}^{(s-1)}, \sigma_{PD}^2 I_T\right)}\right)$$

where

- y_i *and* z_i *are the* T*-dimensional vectors collecting, respectively,* $y_{i,t}$ *and* $z_{i,t}$*;*
- $f^*_{PK} \equiv f_{PK}\left((1,..,t,..,T), d_i, \theta_i^{PK*}\right)$ *is the* T*-dimensional vector of estimated concentrations for the patient* i *with parameters* θ_i^{PK*}*;*
- *similarly,* $f^{(s-1)}_{PK} \equiv f_{PK}\left((1,..,T), d_i, \theta_i^{PK(s-1)}\right)$
- $f^*_{PD} \equiv f_{PK}\left(\begin{pmatrix} 1 & .. & t & .. & T \end{pmatrix}, d_i, \theta_i^{PD*}, f^*_{PK}\right)$ *is the* T*-dimensional vector of estimated responses for the patient* i *with parameters* θ_i^{PD*}*;*
- *similarly,* $f^{(s-1)}_{PD} \equiv f_{PK}\left((1,...,T), d_i, \theta_i^{PD(s-1)}, f^*_{PK}\right)$
- *The symbol* N *is used to denote both the distribution function and the density function of a Gaussian distribution.*

With probability $\alpha\left(\theta_i^*, \theta_i^{s-1}\right)$*, set* $\theta_i^{(s)} = \theta_i^*$*, otherwise set* $\theta_i^{(s)} = \theta_i^{s-1}$

Metropolis-Hastings algorithm for EDP prior

Let us now assume that:

$$\mathbf{P} \sim EDP\left(\alpha_{\theta^{PK}} P_{0\theta^{PK}}, \alpha_{\theta^{PD}} P_{0\theta^{PD}}\left(\cdot|\theta^{PK}\right)\right)$$

where $P_{0\theta^{PK}}$ and $P_{0\theta^{PD}}\left(\cdot|\theta^{PK}\right)$ are the marginal and the conditional distribution of $P_0 \equiv N_5(\mu, \Sigma)$ respectively.

Algorithm 5.8.3 *For each iteration,* s*,* $s = 2,..,S$*:*
For $i = 1,..,N$*, sample* $\theta_i^{(s)}$ *from* $p\left(\theta_i \mid \theta_{-i}^{(s-1)}\right)$*,*
where $\theta_{-i}^{(s-1)} = \theta_1^{(s-1)},...,\theta_{i-1}^{(s-1)}, \theta_{i+1}^{(s-1)},..,\theta_N^{(s-1)}$.

- *Draw a candidate* θ_i^* *accordingly to the Enriched Pòlya urn distribution described in Section* 2.4.1*:*

  Let us call θ_1^{PK**},..,$\theta_{n_{PK}}^{PK**}$ *the unique values among* $\theta_1^{PK(s-1)},..,\theta_{i-1}^{PK(s-1)}, \theta_{i+1}^{PK(s-1)},..,\theta_N^{PK(s-1)}$*, with* θ_{i+1}^{PK} *denoting the first two random-coefficients associated with the PK model. For each of these unique PK patient-specific random coefficient, there is a set*

of unique PD patient-specific random coefficients, $\theta_1^{PD**},..,\theta_{n_{PD}|PK_r}^{PD**}$.

The PK and the PD elements of θ_i^* *are draw sequentially in two groups. First, draw the PK patient-specific coefficients:*

$$\theta_i^{PK*} = \begin{cases} \theta^{PK,new} \sim P_{0\theta^{PD}}\left(\cdot|\theta^{PK*}\right), & w.p. \ \dfrac{\alpha_{\theta^{PK}}}{\alpha_{\theta^{PK}}+N-1} \\ \theta_r^{**}, r=1,..,n_{PK} & w.p. \ \dfrac{\sum_{z=1}^{N-1}\delta_{\theta_z^{PK(s-1)}=\theta_r^{PK**}}}{\alpha_{\theta^{PK}}+N-1} \end{cases}$$

Then, and conditionally on the value of θ_i^{PK*}*, draw the PD patient-specific coefficients,* θ_i^{PD*}*, as follows:*

$$\theta_i^{PD*} = \begin{cases} \theta^{PD,new} \sim P_{0\theta^{PD}}\left(\cdot|\theta^{PK*}\right) \mid \theta_i^{PK*}=\theta^{PK,new} \\ \qquad wp \ \dfrac{\alpha_{\theta^{PK}}}{\alpha_{\theta^{PK}}+N-1}; \\ \theta^{PD,new} \sim P_{0\theta^{PD}}\left(\cdot|\theta^{PK*}\right) \mid \theta_i^{PK*}=\theta_r^{PK**} \\ \qquad wp \ \dfrac{\sum_{z=1}^{N-1}\delta_{\theta_i^{PK(s-1)}=\theta_r^{PK**}}}{\alpha_{\theta^{PK}}+N-1}\dfrac{\mu(\mathcal{Y},x_i^*)}{\mu(\mathcal{Y},x_i^*)+n_i}, \ r=1,.,n_{PK} \ ; \\ \theta_i^{PD*}=\theta_s^{PD**} \mid \theta_i^{PK*}=\theta_r^{PK**} \\ \qquad wp \ \dfrac{\sum_{z=1}^{N-1}\delta_{\theta_i^{PK(s-1)}=\theta_r^{PK**}}}{\alpha_{\theta^{PK}}+N-1}\dfrac{\sum_{z=1}^{N-1}\delta_{\theta_z^{PD(s-1)}=\theta_s^{PD**|\theta_r^{PK**}}}}{\alpha_{\theta^{PK}}+N-1}, \\ \qquad r=1,..,n_{PK}, \ s=1,..,n_{PD|PK_r}. \end{cases}$$

Let us call the resulting vector $\theta_i^* = \left(\theta_i^{PK*}, \theta_i^{PD*}\right)$.

- *Compute the acceptance probability:*

$$\alpha\left(\theta_i^*, \theta_i^{s-1}\right) = \min\left(1, \frac{\prod_{i=1}^{N} N_T\left(y_i \mid f_{PK}^*, \sigma_{PK}^2 I_T\right) \prod_{i=1}^{N} N_T\left(z_i \mid f_{PD}^*, \sigma_{PD}^2 I_T\right)}{\prod_{i=1}^{N} N_T\left(y_i \mid f_{PK}^{(s-1)}, \sigma_{PK}^2 I_T\right) \prod_{i=1}^{N} N_T\left(z_i \mid f_{PD}^{(s-1)}, \sigma_{PD}^2 I_T\right)}\right)$$

where

- y_i *and* z_i *are the* T*-dimensional vectors collecting* $y_{i,t}$ *and* $z_{i,t}$ *respectively;*
- $f_{PK}^* \equiv f_{PK}\left((1,..,t,..,T), d_i, \theta_i^{PK*}\right)$ *is the* T*-dimensional vector of estimated concentrations for the patient* i *with parameters* θ_i^{PK*}*;*
- *similarly,* $f_{PK}^{(s-1)} \equiv f_{PK}\left((1,..,T), d_i, \theta_i^{PK(s-1)}\right)$
- $f_{PD}^* \equiv f_{PK}\left(\begin{pmatrix} 1 & .. & t & .. & T \end{pmatrix}, d_i, \theta_i^{PD*}, f_{PK}^*\right)$ *is the* T*-dimensional vector of estimated responses for the patient* i *with parameters* θ_i^{PD*}*;*
- *similarly,* $f_{PD}^{(s-1)} \equiv f_{PK}\left((1,...,T), d_i, \theta_i^{PD(s-1)}, f_{PK}^*\right)$
- *The symbol* N *is used to denote both the distribution function and the density function of a Gaussian distribution.*

With probability $\alpha\left(\theta_i^*, \theta_i^{s-1}\right)$*, set* $\theta_i^{(s)} = \theta_i^*$*, otherwise set* $\theta_i^{(s)} = \theta_i^{s-1}$

Figure 5.20: Posterior distributions with DP prior with $\alpha = 6$

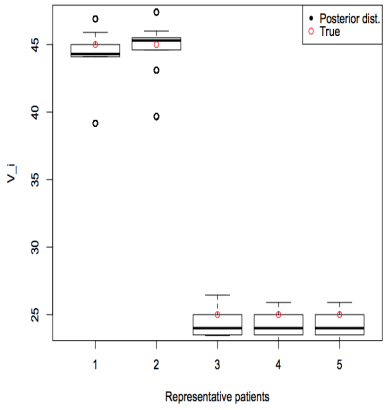


Figure 5.21: $\boldsymbol{V}_i$

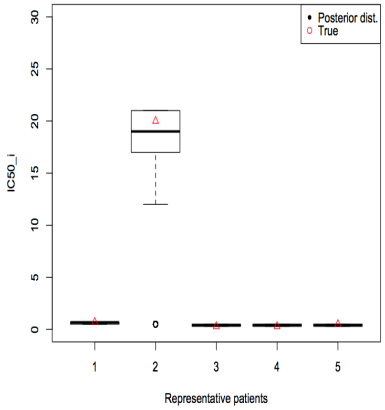


Figure 5.22: $\boldsymbol{IC}_{50;i}$

Bibliography

[1] Barnard J., McCulloch R., Meng X.L. (2000). Modeling Covariance Matrices in terms of standard deviations and correlations with application to Shrinkage. *Statistica Sinica*, **10**, 1281-1311.

[2] Bauer L.(2008). *Applied Clinical Pharmacokinetics* (2nd edition). New York: The McGraw-Hill Companies, Inc..

[3] Blackwell D., MacQueen, J. B. (1973). Ferguson Distributions Via Pólya Urn Schemes. *Annals of Statistics*, **1**, 353-355.

[4] Brown P.J., Le N.D., Zidek J.M. (1994). Inference for a covariance matrix. In *Aspects of Uncertainty: A Tribute to D. V. Lindley* (Freeman P.R., Smith A.F.M., eds.) 77-92. Chichester, UK: John Wiley & Son.

[5] Cheung S.Y., Evans N.D., Chappell M.J., Godfrey K.R., Smith P.J., Errington R.J. (2008). Exploration of the intercellular heterogeneity of topotecanuptake into human breast cancer cells through compartmental modelling. *Mathematical Biosciences*, **213**, 119-34.

[6] Consonni G., Veronese P. (2003). Enriched conjugate and reference priors for the Wishart family on symmetric cones. *Annals of Statistics*, **31**, 1491-1516.

[7] Dawid A.P., Lauritzen S.L. (1993). Hyper-Markov laws in the statistical analysis of decomposable graphical models. *Annals of Statistics*, **21**, 1272-1317.

[8] Dempster A. P. (1972). Covariance Selection. *Biometrics*, **28**, 157-175.

[9] Dunson D.B. (2010). Nonparametric Bayes applications to biostatistics. In *Nonparametric Bayes Statistical Modeling* (Hjort N.L., Holmes C., Müller P., Walker S., eds.). Cambridge, UK: Cambridge University Press

[10] Ette E.I., Williams P.J. (2007). *Pharmacometrics: the Science of Quantitative Pharmacology.* Hoboken: John Wiley & Son.

[11] Ferguson T.S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, **1**, 209-230.

[12] Frühwirth-Schnatter S. (1992). Data Augmentation and Dynamic Linear Models. *Working paper*, Department of Statistics and Mathematics, Vienna: WU Vienna University of Economics and Business.

[13] Gelman A., Carlin J.B, Stern H.S., Rubin D.B. (2009). *Bayesian Data Analysis* (2nd edition). London: Chapman and Hall.

[14] Gillespie W.R. (2009). Prototype PKPD Model Library for WinBUGS User Manual: Version 1.1. Augusta, ME: Metrum Institute.

[15] Huang Y., Dagne G. (2012). Simultaneous Bayesian Inference for Skew-Normal Semiparametric Nonlinear Mixed-Effects Models with Covariate Measurement Errors. *Bayesian Analysis*, **7**, 189-210.

[16] Lacroix B.D., Friberg L.E., Karlsson M.O. (2012). Evaluation of IPPSE, an alternative method for sequential population PKPD analysis. *Journal of Pharmacokinetics and Pharmacodynamics*, **39**, 177-193.

[17] Lunn D.J., Best N., Thomas A., Wakefield J., Spiegelhalter D. (2002). Bayesian Analysis of Population PK/PD Models: General Concepts and Software. *Journal of Pharmacokinetics and Pharmacodynamics*, **29**, 271-307.

[18] Lunn D.J., Best N., Thomas A., Wakefield J., Spiegelhalter D. (2009). Combining MCMC with sequential PKPD modelling. *Journal of Pharmacokinetics and Pharmacodynamics*, **36**, 19-38.

[19] Müller P., Quintana F. (2004). Nonparametric Bayesian Data Analysis. *Statistical Science*, **19**, 95-110.

[20] Müller P., Rosner G. (1997). A Bayesian Population Model with Hierarchical Mixture Priors applied to Blood Count data .*Journal of the American Statistical Association*, **92**, 1279-1292.

[21] Neal R.M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, **9**, 249-265.

[22] O'Malley A., Zaslavsky A. (2008). Domain-level covariance analysis for survey data with structured nonresponse. *Journal of the American Statistical Association*, **103**, 1405-1418.

[23] Racine-Poon A. (1985).A Bayesian approach to non-linear random effects models *Biometrics*, **41**, 1015-1024.

[24] Sokal A. (1989). Monte Carlo Methods in Statistical Mechanics: Foundations and New Algorithms.*Cours de Troìsieme Cycle de la Physique en Suisse Romande*, Lausanne. Manuscript, available online at `http://www.stat.unc.edu/faculty/cji/Sokal.pdf`.

[25] Springer M.D., Thompson W.E. (1970). The distribution of Products of Beta, Gamma and normal Random Variables. *Journal on Applied Mathematics*, **18**, 721-737.

[26] Wakefield J., Racine-Poon A. (1995), An application of Bayesian population pharmacokinetic/pharmacodynamic models to dose recommendation. *Statistics in Medicine*, **14**, 971-86.

[27] Wakefield J., Smith A., Racine-Poon A., Gelfand A. (1994), Bayesian Analysis of Linear and Non-Linear Population Models by Using the Gibbs Sampler. *Journal of the Royal Statistical Society, Series C*, **43**, 201-221.

[28] Walker S., Wakefield J. (1998). Population models with a nonparametric random coefficient distribution. *Sankhya, Series B*, **60**, 196-214.

[29] Woodrow R.S. (2012). odesolve: Solvers for Ordinary Differential Equations. *R package version 0.9-9*, available online at `http://CRAN.R-project.org/package=odesolve`.

[30] Zhang L., Beal S.L., Sheiner L.B. (2003). Simultaneous vs sequential analysis for population PK/PD data I: best-case performance. *Journal of Pharmacokinetics and Pharmacodynamics*, **30**, 387-404.

Chapter 6

A Bayesian predictive approach for finding the maximal tolerated dose using pharmacokinetic information

Abstract :

This chapter proposes a novel Bayesian predictive approach for solving the problem of finding the maximal tolerated dose, by providing a general framework that allows to control for safety in terms of predictive toxicity. The proposal is to include a toxicity model and a pharmacokinetic model within the same framework. Given the data for a first pool of patients, the maximum tolerated dose for the next patient is defined by minimizing the conditional aggregate posterior expected loss of the predictive sequence of toxicities associated with the selected dose. In the meantime, a safety constraint for the predictive probability of future toxicities has to be satisfied.

6.1 Introduction

All clinical trials are characterized by specific purposes related to test treatment mechanisms and peculiar features of the data, such as sequentiality and heterogeneity. The specific purposes depend on the phase of

the trials. In early development studies, the main aim is to understand the mechanisms of action of a new agent in the body and the reactions of the body, and to ensure proper safety conditions. Three general and related objectives for Phase I studies can be identified: safety and control of the toxicity; estimating an optimal safe dose; describing the pharmacokinetics (e.g., as we did in **Chapter 5**). Generally, these three objectives are treated separately and each of them translates into the need of estimating an important quantity of interest. In oncology, treatments typically produce serious side effects. The primary purpose is then to estimate an optimal dose, referred as the Maximum Tolerated Dose (MTD) and defined as the highest dose that does not produce unacceptable toxicity. Our proposal is to combine the three-above mentioned objectives into a general framework and to estimate the MTD by controlling for safety and for toxicity, incorporating pharmacokinetic information.

The sequentiality characterizing early phase trials has stimulated adaptive clinical trials, assigning dose levels to the next patients based on the observed side effects of the previously treated patients. The ideal approach is to design the trial so that the number of unacceptable toxic events is minimized and the number of patients treated at an optimal dose is maximized. This means that the design should control the probability of overdosing patients any time new data become available and produce a sequence of doses that converges to the MTD. This approach should take into account the heterogeneous nature of cancer phase I trial patients. The main feature and challenge of an adaptive clinical trial is how to define the continual updating of the design using the accumulated information from the sequentially-coming data. Faced with such adaptive design problem, Bayesian inference appears desirable. We propose a novel approach for defining the continual updating in sequential phase I clinical trials based on three salient features. First, we propose a sequential procedure for determining the MTD based on the minimization of a target function constrained to a Bayesian predictive safety condition.

Second, we include a PK model, in order to improve the explanation of toxicity and, consequently, the explanation and the prediction of the MTD. Third, we allow for heterogeneity among patients by using random coefficients in the pharmacokinetic and in the toxicity models. The main novelty of our approach concerns the use of the *whole PK profile*, instead of only one of its summaries, within the toxicity model. The predictive toxicity can then be better explained and the next MTD better defined.

The chapter is organized as follows. In Section 6.2, we give a description of traditional approaches for finding the MTD, related to our proposal, i.e. approaches for finding the MTD which include PK data and predictive approaches for finding the MTD. In Section 6.3, we introduce our general proposal, without imposing any specific PK model. We then focus on the one-compartmental PK model. In Sections 6.4 and 6.5, we assume homogeneity across patients. In particular, in Section 6.4, we consider a simple model with known variances and assume a single administered dose to each patient. These assumptions allow us to solve the problem analytically and focus on the proposed procedure. Section 6.5 removes these assumptions and solves, by stochastic approximation, a more general problem characterized by multiple doses that are administered to each patient in a homogeneous population. For a homogeneous population, we consider not only the MTD for the next patient but also the MTD for a sequence of patients. We illustrate the procedure comparing the results with a benchmark procedure. Sections 6.6 and 6.7 extend the analysis to a heterogeneous population, assuming a Gaussian population distribution of the patient-specific random coefficients. In Section 6.8, a nonparametric extension based on the Dirichlet Process prior is presented. The proposed approaches are illustrated in Section 6.9. A brief discussion concludes the chapter.

6.2 Review

Depending on how new enrolled patients are assigned dose levels, phase I trials can be classified into rule-based and model-based methods (Berry *et al.*, 2011). Standard rule-based designs assign new patients to dose levels according to pre-specified rules, without making any assumption on the dose-toxicity curve. An example are the *up-and-down designs*, where the dose is increased or decreased depending on the absence or the presence of severe toxicity in the previous cohort of treated patients. The traditional 3+3 design is a rule-based design which remains, in several variations, widely used in clinical practice. Hiemenz *et al.* (2005) and Meyerhardt *et al.* (2007) studied pharmacokinetic profiles and MTD within the traditional 3+3 design but without providing any formal connection between the MTD and the pharmacokinetic model. Collins *et al.* (1990) proposed the formal inclusion of PK information in the dose-finding problem. Their approach, called *Pharmacologically Guided Dose Escalation method*, consists in including a summary of the PK information, namely the area under the concentration-time curve (AUC), within the traditional 3+3 design. The idea is to continue the dose escalation until some pre-specified plasma exposure defined by the AUC is reached.

Model-based approaches usually assume a monotone dose-toxicity relationship so that, letting $Y(d)$ denote the toxicity at dose d, the probability:

$$p(d) \equiv Prob(\mathbf{Y}(d) \geq \eta) \tag{6.1}$$

is monotone increasing in d, where η is the specified target toxicity level, that is to say, the maximum value of tolerated toxicity to which getting close but without exceeding.

The two most used model-based methods are two Bayesian adaptive approaches: the continual reassessment method (O'Quigley *et al.*, 1990) and the escalation with overdose control (Babb *et al.*,1998).

In its basic form, the Continual Reassessment Method (CRM) char-

acterizes the dose-toxicity relationship by a simple one-parameter model, such as a one-parameter power model:

$$p\left(d \mid \boldsymbol{\theta}\right) \equiv Prob\left(\boldsymbol{Y}\left(d\right) \geq \eta \mid \boldsymbol{\theta}\right) = d^{\boldsymbol{\theta}}, \qquad \boldsymbol{\theta} > 0$$

The original CRM is carried out by starting with a vague prior on $\boldsymbol{\theta}$ and treat the first patient at the level closest to the current estimate of the MTD. Then, any time a new data-point is available, update the posterior distribution of $\boldsymbol{\theta}$ and treat the next patient at the level closest to the updated estimate of the MTD based on the posterior distribution of $\boldsymbol{\theta}$. Piantadosi and Liu (1996) proposed the incorporation of PK data into a CRM design by including the AUC in the model. In particular, they suggested a model of the form:

$$logit\left(p\left(d \mid \boldsymbol{\theta}\right)\right) = \boldsymbol{\theta}_0 + \boldsymbol{\theta}_1\, d - \boldsymbol{\theta}_2 \Delta AUC$$

where $\Delta AUC = AUC - \dfrac{d}{\gamma}$, $\boldsymbol{\theta} = (\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, and provided evidence of improved efficiency and accuracy.

Escalation With Overdose Control (EWOC) is the same as CRM except in the way of selecting the successive new dose. The EWOC method allows the clinician to pre-specify the dose levels that will be available in a phase I trial, so that the anticipated proportion of patients who experience toxicity is set equal to a specified value α. This is accomplished by computing the corresponding posterior distribution of the MTD, $\boldsymbol{\xi}_\alpha$, at the time of each dose assignment. After N dose assignments, the posterior distribution of the MTD is given by:

$$F_N\left(d\right) \equiv Prob\left(\boldsymbol{\xi}_\alpha \leq d \mid \mathcal{F}_N\right)$$

where $\mathcal{F}_N$ denotes the data at the time of treatment for the Nth patient. EWOC selects the dose level d_{N+1} for the $(N+1)$th patient such that:

$$F_N\left(d_{N+1}\right) = \alpha \tag{6.2}$$

CRM and EWOC produce, under some model assumptions and regularity conditions, *consistent* sequences of doses. The concept of *consistency* of a sequence of doses has been introduced by Eichhorn and Zacks (1973), together with the concept of *feasibility*. *Feasibility* controls the probability that a given dose in the sequences exceeds the MTD. *Consistency* ensures that the sequence of doses converges to the MTD. In terms of equation (6.2), any sequence of doses d_j, $j = 1, .., \nu$ so that $F_N\left(d_j\right) \leq \alpha$ is said *feasible*. For this reason, α is called the *feasibility bound*. Moreover, if $d_j \to \xi_\alpha$, then the procedure is said to be *consistent*. Later, Robinson (1978) introduced the concept of *safety* in substitution to the condition of feasibility. *Safety* requires controlling for the probability that the resulting toxicity does not exceed a threshold. The *safety* condition can be expressed by specifying the event corresponding to the DLT by using a model for the toxicity, as in equation (6.1). Let y_i denote the observed toxicity at dose d_i, for patients $i = 1, .., N$, $N \geq 1$. Eichhorn and Zacks (1981) assume a linear model for the random toxicity $\boldsymbol{Y}_i$ given the dose d_i:

$$\boldsymbol{Y}_i \mid d_i \overset{indep}{\sim} N\left(\boldsymbol{\beta}\left(d_i - d_0\right), \sigma^2\left(d_i - d_0\right)^2\right) \tag{6.3}$$

where d_0 is known, and fixed the dose d_{N+1} for the next patient, $N+1$, as the largest dose value such that the following *safety condition* is satisfied:

$$Prob\left(\boldsymbol{Y}_{N+1}\left(d_{N+1}\right) \leq \eta \mid \boldsymbol{\beta}, d_{N+1}\right) \geq \gamma \tag{6.4}$$

where η denotes the threshold of dangerous toxicity levels and γ is the target toxicity probability of MTD. The MTD, ξ_γ, is here defined as the dose d_i that satisfied the equation (6.4) with the equality and taking β and σ^2 as known. Standardizing the equation (6.4), it follows that ξ_γ

is the solution of :

$$\Phi\left(\frac{\eta-\beta\left(\xi_\gamma-d_0\right)}{\sigma\left(\xi_\gamma-d_0\right)}\right)=\gamma$$

where Φ is the standard distribution function, that is to say:

$$\xi_\gamma = d_0 + \frac{\eta}{\beta+\sigma\ z_\gamma} \tag{6.5}$$

where z_γ is the γth quantile of the standard Gaussian distribution. Unfortunately, β and σ are seldom known, requiring further efforts for estimating ξ_γ. In many papers (Eichhorn and Zacks, 1973; 1981; Robinson, 1978; Shih, 1989), the problem is solved by taking the relationship (6.5) as given, assuming σ known and giving a prior distribution on $\boldsymbol{\beta}$. Then, one can find the posterior distribution of $\boldsymbol{\beta}$, given the data on the first N patients, and compute the posterior of ξ_γ via the transformation (6.5). Analogous approaches based on the relationship between the dose and a dichotomic toxicity are described in Zacks *et al.* (1998); Babb and Rogatko (2001); and Tighiouart and Rogatko (2010). An approach that exploits better the probabilistic dependence across the observations $\boldsymbol{Y}_{i,t}$ is the Bayesian predictive distribution. Muliere and Petrone (1993) consider a model for discrete-value doses, d_i. Since the dose range is finite, each dose d_i can be repeated, $s = 1, .., S_i$:

$$\boldsymbol{Y}_{i,s} \mid \boldsymbol{\beta}, \boldsymbol{\sigma}^2, d_i \overset{indep}{\sim} N\left(\boldsymbol{\beta}\left(d_i-d_0\right), \boldsymbol{\sigma}^2\left(d_i-d_0\right)^2\right) \tag{6.6}$$

Consequently, the distribution of the toxicities for the next patient, namely the distribution of $\boldsymbol{Y}_{N+1,s}$, given $\mathcal{F}_N = \{y_{i,s}, i = 1 : N; s = 1, .., S_i\}$, is obtained as:

$$p\left(\boldsymbol{Y}_{N+1,s} \mid \mathcal{F}_N, d_{N+1}\right) = \int p\left(\boldsymbol{Y}_{N+1,s} \mid \beta, \sigma^2, d_{N+1}\right) p\left(\beta, \sigma^2 \mid \mathcal{F}_N\right) d\left(\beta, \sigma^2\right) \tag{6.7}$$

where $p\left(\beta, \sigma^2 \mid \mathcal{F}_N\right)$ denotes the posterior distribution of $\left(\beta, \sigma^2\right)$ given $\mathcal{F}_N$. Muliere and Petrone (1993) determined the MTD for the next patient as the value which minimizes a target function subject to the condition that the Bayesian predictive probability that the toxicity exceeds a tolerated threshold η is small:

$$Prob\left(\boldsymbol{Y}_{N+1,s}\left(d_{N+1}\right) > \eta \mid \mathcal{F}_N, d_{N+1}\right) \leq \gamma \tag{6.8}$$

Inequality (6.8) states the DLT event conditionally on the available data, namely in terms of the predictive distribution of the toxicity, whereas inequality (6.4) is stated conditionally on the unknown parameter $\boldsymbol{\beta}$. For this reason, the inequality (6.8) is called *Bayesian predictive safety condition.*

Developments of the predictive approach of Muliere and Petrone (1993) are in Muliere and Walker (1997) and Mezzetti *et al.* (2007), who determine the MTD using a Bayesian nonparametric approach. Whitehead *et al.* (2006) and Zhou *et al.* (2008) proposed Bayesian predictive approaches by controlling the predictive distribution of some linearly-modelled random variable, such as the natural logarithm of the area under the curve (log AUC) or of the maximum concentration. Their predictive safety conditions is again based only on a summary of the PK data. Bayesian predictive approaches in early phases of a clinical trial are discussed, among the others, in Lee and Liu (2008) and Berry *et al.* (2011).

As most recent proposal shown, it is important to use the concentration instead of the dose of an administered drug. However, in literature, only a summary of the PK information is used. Our proposal is to use the whole PK profile for finding the MTD within a Bayesian predictive approach.

6.3 Our proposal: general formulation

Our proposal extends the approach of Muliere and Petrone (1993) by modeling the dynamics of the toxicity over time and incorporating its dependence on the concentration profile. Therefore, for each patient i and for each administered drug, a sequence of toxicities is recorded. The data include the sequence of administered doses, $d_{i,t} \geq 0$, where at least one dose $d_{i,t}$ is different from zero (in t_0), the toxicities, $y_{i,t}$, and the sequence of corresponding concentrations, let us say $c_{i,t}$, with $t \in \{t_0, .., t_T\}$.

6.3.1 Toxicity model and PK

Our proposal is to provide a better model for the toxicities by allowing the toxicity curve to depend on the sequence of corresponding concentrations, or some function of them. Our proposed model for the random toxicity, $\boldsymbol{Y}_{i,t}$, has the following form:

$$\log\left(\boldsymbol{Y}_{i,t}\right) = \boldsymbol{\alpha}_i + \log\left(\sum_{r=max(t_{-Q},t_0)}^{t} \boldsymbol{\varphi}_{r;i}\boldsymbol{C}_{i,r}\right) + \boldsymbol{\xi}_{i,t} \quad \text{where} \quad \boldsymbol{\xi}_{i,t} \sim N\left(0, \sigma^2_{Tox}\right). \tag{6.9}$$

with α_i and $\boldsymbol{\varphi}_{r;i}$ potentially patient-specific and where t_{-Q} stands for "Q sampling times before t", e.g. if the sampling time t is equal to t_t, we mean t_{t-Q}.

Model (6.9) assumes that the logarithm of the toxicity, $\log\left(\boldsymbol{Y}_{i,t}\right)$, depends on the logarithm of the weighted sum of the past concentrations $\boldsymbol{C}_{i,r}$, $r = t_{-Q}, t_{-Q+1}, .., t$. Of course, alternative models could be considered. For instance, one could assume that the toxicity depends only on the contemporaneous concentration but it is autoregressive of order r:

$$\log(\boldsymbol{Y}_{i,t}) = \boldsymbol{\alpha}_i + \sum_{r=max(t_{-Q},t_0)}^{t-1} \boldsymbol{\varphi}_{r;i} \log(\boldsymbol{Y}_{i,r}) + \boldsymbol{\varphi}_{0;i} \log(\boldsymbol{C}_{i,r}) + \xi_{i,t} \quad (6.10)$$

We prefer model (6.9) because the toxicity is monotone increasing in the discounted sum of the Q concentrations. In model (6.10), the current toxicity depends directly only on the current concentration and indirectly on all the previous concentrations. Consequently, if the elimination rate is large enough, the toxicity could be even decreasing in the concentration. In any case, the toxicity model is problem-specific and should be carefully defined depending on the specific problem.

In order to identify $\boldsymbol{\varphi}_{r;i}$ of model (6.9), we assume the exponential Almon specification (Almon, 1965) for the parameters of each patient i:

$$\boldsymbol{\varphi}_{r;i} = \frac{e^{\boldsymbol{\theta}_{1;i} r + \boldsymbol{\theta}_{2;i} r^2}}{\sum_{r=max(t_{-Q},t_0)}^{t} e^{\boldsymbol{\theta}_{1;i} r + \boldsymbol{\theta}_{2;i} r^2}}. \quad (6.11)$$

This specification implies that more recent concentrations have higher impact on the current toxicity. Call $\boldsymbol{\theta}_i^{Tox}$ the vector collecting all the patient-specific coefficients, i.e. $\boldsymbol{\theta}_i^{Tox} = (\boldsymbol{\alpha}_i, \boldsymbol{\theta}_{1;i}, \boldsymbol{\theta}_{2;i})$.

Another crucial element of model (6.9) is represented by the concentrations $\boldsymbol{C}_{i,r}$. For each patient i, a whole sequence of PK measurements, $c_{i,t}$, $t = t_0, .., t_T$ is collected, associated with one or more doses and assuming that after t_T the concentration of the drug is negligible. Concentrations are the focus of pharmacokinetic studies and are usually modelled using compartment models. The kinetics of the administered dose is usually expressed by a system of ordinary differential equations (ODEs), one equation for each hypothetical compartment. Assuming a fix number p of compartments, the concentration of the dose for the patient i in each of these p compartments at time t, say $\underline{\boldsymbol{G}}_{i,t}$, can be modelled as follows:

$$d\underline{\boldsymbol{G}}_{i,t} = \boldsymbol{\mu}_{PK}\left(\underline{\boldsymbol{G}}_{i,t}, d_{i,t_0:t}, \boldsymbol{\theta}_i\right) dt \tag{6.12}$$

Solving equation (6.12), one can find the p-dimensional vector $\underline{\boldsymbol{G}}_{i,t}$ describing the hypothetical concentration at time t in each of these p compartments. At each time, a single PK measurement is recorded for each patient. This sequence of values is usually well-described by the evolution of the effective compartment from the general multivariate vector $\underline{\boldsymbol{G}}_{i,t}$. We call the evolution of the concentration in this specific compartment:

$$\boldsymbol{C}\left(t, d_{i,t_0:t}, \boldsymbol{\theta}_i^{PK}\right) \equiv f_{PK}\left(t, d_{i,1:t}, \boldsymbol{\theta}_i^{PK}\right) \tag{6.13}$$

For notational simplicity, $\boldsymbol{C}\left(t, d_{i,t_0:t}, \boldsymbol{\theta}_i^{PK}\right)$ is also denoted by $\bar{\boldsymbol{C}}_{i,t}$. The simplest PK model considers only one compartment and assumes $\frac{d\bar{\boldsymbol{C}}_i}{dt} = -k_i\bar{\boldsymbol{C}}_i$, where $\boldsymbol{k}_i$ is the rate of elimination of drug for patient i, and with the initial condition depending on all the previous administered doses. Assuming that the administered dose is the first one, the initial condition is $\bar{\boldsymbol{C}}_{i,0} = \frac{d_i}{\boldsymbol{V}_i}$, where $\boldsymbol{V}_i$ is the apparent volume of distribution. The differential equation admits the following explicit solution:

$$\bar{\boldsymbol{C}}_{i,t} = \frac{d_i}{\boldsymbol{V}_i} e^{-\boldsymbol{k}_i t} \tag{6.14}$$

We will use this simple PK model. Model (6.14) assumes that the amount of drug is decreasing at a rate that is proportional to the amount of the drug remaining in the body. Randomness comes in the model through *intra-patient* variation, i.e. the deviation between each particular observation and its predictable value, $\left(c_{i,t} - f_{PK}\left(t, d_i, \boldsymbol{\theta}_i^{PK}\right)\right)$, and through *inter-patient* variation, i.e. the dispersion of the patient-specific random coefficients, $\boldsymbol{\theta}_i^{PK}$. The general PK model usually assumes a multiplicative log-normal-distributed noise for (6.13). Taking its logarithmic transfor-

mation for convenience, the model becomes:

$$\log\left(\boldsymbol{C}_{i,t}\right) = \log\left(f_{PK}\left(t, d_i, \boldsymbol{\theta}_i^{PK}\right)\right) + \boldsymbol{\epsilon}_{i,t} \text{ where } \boldsymbol{\epsilon}_{i,t} \sim N\left(0, \boldsymbol{\sigma}_C^2\right). \tag{6.15}$$

6.3.2 Sequential search of the MTD for the next patient

Given the PK-toxicity model expressed by (6.15) and (6.9), the purpose of the analysis is to find the MTD for the next patient, $N+1$, using the data on toxicities that are associated with the administered doses for the first N patients. Efficiency of the treatment and safety are the two general goals. In general, higher doses are needed to improve the therapeutic effect. However, they are associated to higher toxicity effects. We assume that medical doctors suggest the *threshold of acceptable toxicity*, denoted by η, that allows the best trade-off between efficiency and safety. Thus, the aim is to choose the next dose, d, so that the corresponding predictive toxicities reach the specific balance between efficiency and safety, i.e. the (discounted) toxicities for the next patient are as close as possible to the fixed η.

Similarly to Muliere and Petrone (1993), we formalize this problem within a decisional framework, fixing a symmetric loss function for the deviation of the predictive log-toxicity[1] level $Y_{N+1,t}$, which depends on the chosen dose d, from the fixed threshold η. We choose the following quadratic loss function:

$$L_\eta\left(\boldsymbol{Y}_{N+1,t}\left(d\right), d\right) = \left(\log\left(\boldsymbol{Y}_{N+1,t}\left(d\right)\right) - \eta\right)^2 \tag{6.16}$$

where $\boldsymbol{Y}_{N+1,t}\left(d\right)$ denotes the random toxicity for patient $N+1$ at

[1] Concentrations and toxicities are usually modelled in logarithm to stabilize variances. The use of the logarithm will also facilitate the analytical resolution of the problem in Section 6.4 and 6.6.

time t, corresponding to the dose d. Although more appropriate, possibly asymmetric, loss function could be used, we use the quadratic loss function defined by (6.3.2) for the sake of simplicity.

The Bayesian solution of the decision problem is based on the posterior expected loss associated to the dose d, denoted by $R_\eta\left(\boldsymbol{Y}_{N+1,t}\left(d\right),d\right)$ and given by:

$$R_\eta\left(\boldsymbol{Y}_{N+1,t}\left(d\right),d\right)=E\left(\left(\log\left(\boldsymbol{Y}_{N+1,t}\left(d\right)\right)-\eta\right)^2\mid\mathcal{F}_N\right),$$

where $\mathcal{F}_N$ denotes, as before, the data available for the first N patients. With the specified quadratic loss function, the posterior expected loss can be decomposed as:

$$R_\eta\left(\boldsymbol{Y}_{N+1,t}\left(d\right),d\right)=V\left(\log\left(\boldsymbol{Y}_{N+1,t}\left(d\right)\right)\mid\mathcal{F}_N\right)+\left(E\left(\log\left(\boldsymbol{Y}_{N+1,t}\left(d\right)\right)\mid\mathcal{F}_N\right)-\eta\right)^2.$$

We can now define the conditional *aggregate* posterior expected loss as the sum of the expected posterior loss over the fix grid of time points, corresponding to the measurement times[2]:

$$\sum_{t=t_0}^{t_T}\delta_t R_\eta\left(\boldsymbol{Y}_{N+1,t}\left(d\right),d\right)$$

where δ_t represents a discount factor, for example fixed as $\delta_t=\dfrac{1}{1+\delta_0^t}$, with $\delta_0\geq 0$.

The decisional setting us completed by a safety constraint, called *overall Bayesian predictive safety constraint*, that we express as the following:

$$Pr\left(\bigcup_{t=t_0}^{t_T}\log\left(\boldsymbol{Y}_{N+1,t}(d)\right)\geqslant\eta|\mathcal{F}_N\right)\leqslant 1-\gamma, \tag{6.17}$$

[2]For the sake of simplicity, we are implicitly assuming that all the past and the future observations will be recorded after a fix number of hours after the administration of the new dose. In principle, it is also possible to consider a continuous interval of time over which computing conditional *aggregate* posterior expected loss. Then, once these data collection is carried out only that discrete grid would be available.

with $\gamma \in (0,1)$. The aim is to find the dose for the patient $N+1$, d_{N+1}, for $N = 1, 2, ...$ that minimizes the aggregate posterior expected loss and satisfies the overall Bayesian predictive safety constraint. The problem can be summarized as follows:

$$d_{N+1} = argmin_d \left(\sum_{t=t_0}^{t_T} \delta_t R_\eta \left(\boldsymbol{Y}_{N+1,t}(d), d \right) \right), \tag{6.18}$$

$$\text{with the constraint } Pr\left(\bigcup_{t=t_0}^{t_T} \log\left(\boldsymbol{Y}_{N+1,t}(d) \right) \geqslant \eta | \mathcal{F}_N \right) \leqslant 1-\gamma. \tag{6.19}$$

The value of the discount factor, δ_t, is problem specific and it can also be set equal to 1 if one believes that all the periods should have the same weight. However, since η is the target, in the periods right after the administration of a positive dose, the toxicities will be higher, because the toxicity is a positive function of the concentration, which decreases, at least from a certain time. Therefore, the discount factor can be chosen so to give more weights to the periods right after the administration of a dose and less to the following times. This pattern is similar to the one of the toxicity but it can be reinforced using a proper value of δ_t.

For all the following sections, we make the following assumptions:

- The two error terms in equations (6.15) and (6.9), $\boldsymbol{\epsilon}_{i,t}$ and $\boldsymbol{\xi}_{i,t}$, are independent among themselves, across t and across i.

- The PK patient-specific random coefficients, $\boldsymbol{\theta}_i^{PK}$, are independent from the toxicity patient-specific random coefficients, $\boldsymbol{\theta}_i^{Tox}$.

- We assume that the concentration profiles, recorded up to the fixed t_T, are always greater than some small value, i.e. greater that 1, and after t_T the concentration of the dose becomes negligible. The latter assumption is used in Section 6.4 and Section 6.6, to make computations easier.

6.4 One-compartmental model: single administered dose

In this section, we consider the simplest statistical model within the general framework discussed in Section 6.3, with the aim of having an analytical solution and focusing on the procedure. The restrictive assumptions here introduced will be later relaxed. Thus, here we assume that the data on concentration and toxicity correspond to a single administered dose for each patient and that the population of patients is homogeneous, i.e. $\boldsymbol{\theta}_i^{PK} = \boldsymbol{\theta}_{PK}$ and $\boldsymbol{\theta}_i^{Tox} = \boldsymbol{\theta}^{Tox}$. The available data for taking a decision on the next patient, $N+1$, are hence $(d_i, c_{i,t}, y_{i,t}), i = 1:N; t = t_0, .., t_T$, where $d_1, .., d_N$ are the doses already administered.

From here on, we assume that the one-compartmental PK model given by equation (6.14) describes the PK data well. Under these assumptions, model (6.15) reduces to:

$$\log(\boldsymbol{C}_{i,t}) = \log\left(\frac{d_i}{\boldsymbol{V}} e^{-\boldsymbol{k}t}\right) + \boldsymbol{\epsilon}_{i,t} \text{ where } \boldsymbol{\epsilon}_{i,t} \mid \boldsymbol{\sigma}_C^2 \overset{\text{iid}}{\sim} N\left(0, \boldsymbol{\sigma}_C^2\right). \quad (6.20)$$

In this section, we also consider the simplest toxicity model, that assumes only a simultaneous effect of the concentration on the toxicity and a multiplicative noise, i.e. $\boldsymbol{Y}_{i,t} = \tilde{\boldsymbol{\alpha}} \boldsymbol{C}_{i,t} \tilde{\boldsymbol{\xi}}_{i,t}$. In terms of model (6.9), this means assuming $Q = 1$, $\boldsymbol{\alpha} = \log(\tilde{\boldsymbol{\alpha}})$, $\boldsymbol{\xi} = \log\left(\tilde{\boldsymbol{\xi}}\right)$ and to fix the coefficient, $\boldsymbol{\varphi}_1$, equal to one. The resulting model is equivalent to:

$$\log(\boldsymbol{Y}_{i,t}) = \boldsymbol{\alpha} + \log(\boldsymbol{C}_{i,t}) + \boldsymbol{\xi}_{i,t} \text{ where } \boldsymbol{\xi}_{i,t} \sim N\left(0, \boldsymbol{\sigma}_{Tox}^2\right) \quad (6.21)$$

The problem admits an analytical expression for the posterior distribution choosing standard parametric priors, as will be discussed in

the following subsection. The assumption that $\boldsymbol{\varphi} = 1$ could seem too restrictive. Relaxing it, the resulting model would be: $\log(\boldsymbol{Y}_{i,t}) = \boldsymbol{\alpha} + \log(\boldsymbol{\varphi}\,\boldsymbol{C}_{i,t}) + \boldsymbol{\xi}_{i,t} = (\boldsymbol{\alpha} + \log(\boldsymbol{\varphi})) + \log(\boldsymbol{C}_{i,t}) + \boldsymbol{\xi}_{i,t}$, namely, $\boldsymbol{\alpha}$ and $\log(\boldsymbol{\varphi})$ would not be identifiable. A valid alternative to model (6.21) would be $\log(\boldsymbol{Y}_{i,t}) = \boldsymbol{\alpha} + \boldsymbol{\varphi}\log(\boldsymbol{C}_{i,t}) + \boldsymbol{\xi}_{i,t}$, but this is not a special case of model (6.9) but of model (6.10). Although it also admits analytical computations and with $Q = 1$ can be a valid alternative, for the sake of simplicity, we will start here by studying model (6.21). We will later relax these restrictive assumptions and we will study the general model defined by expression (6.9).

Note that the joint model for the PK and toxicity variables, i.e. for $(\log(\boldsymbol{C}_{i,t}), \log(\boldsymbol{Y}_{i,t})) \mid \boldsymbol{V}, \boldsymbol{k}, \boldsymbol{\alpha}, d_i$, is expressed in terms of the marginal model for the PK variable conditionally on its own coefficients, i.e. in model (6.20) for $\log(\boldsymbol{C}_{i,t}) \mid \boldsymbol{V}, \boldsymbol{k}, d_i$, and the conditional model for the toxicity variable conditionally on the PK variable and on its own coefficients, i.e. in model (6.21) for $\log(\boldsymbol{Y}_{i,t}) \mid \log(\boldsymbol{C}_{i,t}), \boldsymbol{\alpha}$. We can re-write models (6.20) and (6.9) as:

$$\log(\boldsymbol{C}_{i,t}) \mid \boldsymbol{V}, \boldsymbol{k}, d_i \sim N\left(\log\left(\frac{d_i}{\boldsymbol{V}} e^{-\boldsymbol{k}t}\right), \sigma_C^2\right) \tag{6.22}$$

$$\log(\boldsymbol{Y}_{i,t}) \mid \log(\boldsymbol{C}_{i,t}), \boldsymbol{\alpha} \sim N\left(\boldsymbol{\alpha} + \log(\boldsymbol{C}_{i,t}), \boldsymbol{\sigma}_{Tox}^2\right). \tag{6.23}$$

6.4.1 PK model

Letting

$$\boldsymbol{\omega} = \begin{pmatrix} log(\boldsymbol{V}) \\ \boldsymbol{k} \end{pmatrix}, \; X_t = \begin{pmatrix} -1 \\ -t \end{pmatrix}, \tag{6.24}$$

we can re-write the PK model (6.20) as:

$$\log(\boldsymbol{C}_{i,t}) = \log(d_i) + \boldsymbol{\omega}' X_t + \; \epsilon_{i,t} \text{ where } \boldsymbol{\epsilon}_{i,t} \mid \boldsymbol{\sigma}_C^2 \overset{\text{iid}}{\sim} N\left(0, \boldsymbol{\sigma}_C^2\right)$$

We assume a Gaussian prior on ω:

$$\boldsymbol{\omega} \sim N(m, \Omega)$$

with $m = [V_0 \quad k_0]$ and $\Omega = \begin{pmatrix} \sigma_V^2 & 0 \\ 0 & \sigma_k^2 \end{pmatrix}$, i.e. $\boldsymbol{V}$ and $\boldsymbol{k}$ independent, with a log-normal distribution for $\boldsymbol{V}$ and a Gaussian distribution for $\boldsymbol{k}$, where the variance of k is chosen so that the probability of observing negative k is negligible. These priors are motivated by analytical simplicity and will be kept throughout the whole chapter.

Once the observations $c_{i,t}$ are available, inference on the unknown PK parameters is solved by computing their posterior distributions. With the chosen conjugate priors, the posterior distribution for $\boldsymbol{\omega}$, given $\mathcal{F}_N^{\backslash y} = \{(d_i, c_{i,t}), i = 1, .., N; t = t_0, .., t_T\}$, is still Gaussian with updated parameters:

$$m_N = \left(\Omega^{-1} + \frac{N \sum_{t=0}^{T} X_t X_t'}{\sigma_C^2} \right)^{-1} \left(\Omega^{-1} m + \frac{\sum_{t=0}^{T} X_t \left(\sum_{i=1}^{N} (\log(c_{i,t}) - \log(d_i)) \right)}{\sigma_C^2} \right) \equiv \begin{pmatrix} m_{N;1} \\ m_{N;2} \end{pmatrix}$$

$$\Omega_N = \left(\Omega^{-1} + \frac{N \sum_{t=0}^{T} X_t X_t'}{\sigma_C^2} \right)^{-1} \equiv \begin{pmatrix} \omega_{1,1} & \omega_{1,2} \\ \omega_{1,2} & \omega_{2,2} \end{pmatrix}$$

The predictive distribution for the log-concentration corresponding to dose d_{N+1} for the next $N+1$ patient is obtained as follows:

$$p\left(\log(\boldsymbol{C}_{N+1,s}) \mid \mathcal{F}_N^{\backslash y}, d_{N+1}\right) = \int p(\log(\boldsymbol{C}_{N+1,s}) \mid \omega, d_{N+1})\, p\left(\omega \mid \mathcal{F}_N^{\backslash y}\right) d\omega \quad \text{for} \quad s = t_0, .., t_T$$

For each $s = t_0, .., t_T$, it results to be Gaussian with expected value:

$$u_{N+1,s} \equiv E\left(\log(\boldsymbol{C}_{N+1,s}) \mid \mathcal{F}_N^{\backslash y}, d_{N+1}\right) = \log(d_{N+1}) + m_N' X_s = \log(d_{N+1}) - m_{N;1} - m_{N;2} s$$

and variance:

$$r_{N+1,s} \equiv Var\left(\log\left(\boldsymbol{C}_{N+1,s}\right) \mid \mathcal{F}_N^{\setminus y}, d_{N+1}\right) = \sigma_C^2 + X_s' \Omega_N X_s = \sigma_C^2 + s^2\omega_{2,2} + 2s\omega_{1,2} + \omega_{1,1}$$

6.4.2 Toxicity model

For toxicity, we assume here the simplest model (6.21), where $\boldsymbol{Y}_{i,t}$ only depends on the contemporaneous drug concentration $\boldsymbol{C}_{i,t}$. This assumption will be relaxed in Section 6.8. We choose a Gaussian prior on the parameter $\boldsymbol{\alpha}$:

$$\boldsymbol{\alpha} \sim N\left(\alpha_0, \sigma_\alpha^2\right).$$

Consequently, the posterior distribution of $\boldsymbol{\alpha}$ given

$\mathcal{F}_N = \{(d_i, c_{i,t}, y_{i,t}), i = 1 : N; t = t_0 : t_T\}$ is Gaussian with expected value α_N and variance σ_{α_N} given by

$$\alpha_N = \frac{\alpha_0\sigma_{Tox}^2 + \sigma_\alpha^2 \sum_{i=1}^{N}\sum_{t=0}^{t_T}\left(\log\left(y_{i,t}\right) - \log\left(c_{i,t}\right)\right)}{\sigma_{Tox}^2 + \sigma_\alpha^2 TN} \text{ and}$$

$$\sigma_{\alpha_N} = \frac{\sigma_{Tox}^2\sigma_\alpha^2}{\sigma_{Tox}^2 + \sigma_\alpha^2(T+1)N}$$

The predictive distribution $p\left(\log\left(\boldsymbol{Y}_{N+1,s}\right) \mid \mathcal{F}_N, d_j\right)$ for the log-toxicity for the next patient $N+1$, for each $s = t_0, .., t_T$ is Gaussian with parameters:

$$\rho_{j,s} \equiv E\left(\log\left(\boldsymbol{Y}_{j,s}\right) \mid \mathcal{F}_N, d_j\right) = u_{j,s} + \alpha_N, \tag{6.25}$$

$$\sigma_s^2 \equiv Var\left(\log\left(\boldsymbol{Y}_{j,s}\right) \mid \mathcal{F}_N, d_j\right) = \sigma_{Tox}^2 + r_{j,s} + \sigma_{\alpha_N}^2. \tag{6.26}$$

6.4.3 Constrained minimization problem

The problem of finding the MTD is solved by choosing dose d_{N+1} for the next patient $N+1$ so that the problem defined by (6.18)-(6.19) is solved. Using the quadratic loss function (6.3.2) and the chosen priors, the aggregate posterior expected loss for a dose d is given by:

$$\sum_{s=t_0}^{t_T} \delta_s R_\eta \left(\boldsymbol{Y}_{N+1,s} \left(d \right), d \right) = \sum_{s=t_0}^{t_T} \delta_s \left(\sigma_s^2 + \rho_{N+1,s}^2 - 2\eta\rho_{N+1,s} + \eta^2 \right) \tag{6.27}$$

Substituting the expression (6.25) and (6.26) of the posterior parameters $\rho_{N+1,s}$ and σ_s^2, the sum (6.27) becomes:

$$\sum_{s=t_0}^{t_T} R\left(\eta, \boldsymbol{Y}_{N+1,t}\right) =$$

$$\left(\sum_{s=t_0}^{t_T} \delta_s \right) \log\left(d\right)^2 + \left(2 \left(\sum_{s=t_0}^{t_T} \delta_s \right) \left(\alpha_N - \eta - m_{N;1} \right) - 2 m_{N;2} \left(\sum_{s=t_0}^{t_T} \delta_s s \right) \right) \log\left(d\right) + a,$$

where a is a constant not relevant for the optimization procedure and more generally for the whole procedure.

For controlling the overall Bayesian predictive safety constraint (6.19), we impose

$$Pr\left(\log\left(\boldsymbol{Y}_{N+1,s}(d) \right) \geqslant \eta | \mathcal{F}_N \right) \leqslant 1 - \tilde{\gamma} \tag{6.28}$$

where $\tilde{\gamma} = \dfrac{T+\gamma}{T+1}$. By Boole's inequality[3], this implies (6.19).

Since $\log\left(\boldsymbol{Y}_{N+1,s}\right) \mid \mathcal{F}_N, d_N + 1 \sim N\left(\rho_{N+1,s}, \sigma_s^2\right)$, inequality (6.28) holds if:

$$\eta \geq \rho_{N+1,s} + \sqrt{\sigma_s^2} z_{\tilde{\gamma}} \tag{6.29}$$

where $z_{\tilde{\gamma}}$ is the $\tilde{\gamma} th$ quantile of the standard Gaussian distribution. Using the expressions (6.25) and (6.26) for $\rho_{N+1,s}$ and σ_s^2, the inequality

[3]Boole's inequality, or the union bound, says that for any finite or countable set of events, say $A_1, A_2, ...$, the probability that at least one of the events happens is no greater than the sum of the probabilities of the individual events: $P\left(\cup_i A_i\right) \leq \sum_i P\left(A_i\right)$.

(6.28) is satified for values of the dose d such that:

$$\log(d) \leq \eta - m_N' X_s - \alpha_N - \sqrt{\sigma^2_{\alpha_N} + r_{N+1,s} + \sigma^2_{Tox}} z_{\tilde{\gamma}} \quad \text{for every } s = t_0, ..., t_T$$

It follows that:

$$\begin{aligned} \log(d) &\leq \min_{s \in (t_0,..,t_T)} \left(\eta + m_{N;1} + m_{N;2}\, s - \alpha_N - \sqrt{\sigma^2_{\alpha_N} + r_{N+1,s^*} + \sigma^2_{Tox}} z_{\tilde{\gamma}} \right) \\ &= \eta + m_{N;1} + m_{N;2}\, s^* - \alpha_N - \sqrt{\sigma^2_{\alpha_N} + r_{N+1,s} + \sigma^2_{Tox}} z_{\tilde{\gamma}}, \text{ say.} \end{aligned} \tag{6.30}$$

The Lagrangian function for the resulting constrained minimization problem is:

$$\mathcal{L} = \left(\sum_{s=t_0}^{t_T} \delta_s \right) \log(d)^2 + \left(2 \left(\sum_{s=t_0}^{t_T} \delta_s \right) (\alpha_N - \eta - m_{N;1}) - 2 m_{N;2} \left(\sum_{s=t_0}^{t_T} \delta_s s \right) \right) \log(d) + a +$$

$$+ \lambda \left(\eta + m_{N;1} + m_{N;2} s^* - \alpha_N - \sqrt{\sigma^2_{\alpha_N} + r_{N+1,s^*} + \sigma^2_{Tox}} z_{\tilde{\gamma}} - \log(d) \right)$$

where a is a constant which does not depend on d, and therefore is irrelevant for the resolution of the problem. Necessary conditions requires:

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial \log(d)} \leq 0; \quad \log(d) \geq 0; \quad \dfrac{\partial \mathcal{L}}{\partial \log(d)} \log(d) = 0 \\ \log(d) \leq \eta + m_{N;1} + m_{N;2} s^* - \alpha_N - \sqrt{\sigma^2_{\alpha_N} + r_{N+1,s^*} + \sigma^2_{Tox}} z_{\tilde{\gamma}} \\ \lambda \left(\eta - m_{N;1} + m_{N;2} s^* - \sqrt{\sigma^2_{\alpha_N} + r_{N+1,s} + \sigma^2_{Tox}} z_{\gamma} - \log(d) \right) = 0, \\ \lambda \geq 0 \end{cases}$$

The unconstrained and boundary solutions are:

- Unconstrained solution ($\lambda = 0$):

$$d_1 = \exp\left\{ -\frac{\sum_{s=t_0}^{t_T} \delta_s \left(\alpha_N - \eta - m_{N;1} - m_{N;2} s\right)}{\sum_{s=t_0}^{t_T} \delta_s} \right\} \tag{6.31}$$

- Boundary solution ($\lambda \neq 0$):

$$d_2 = \exp\left\{ \eta + m_{N;1} + m_{N;2} s^* - \alpha_N - \sqrt{\sigma^2_{\alpha_N} + r_{N+1,s^*} + \sigma^2_{Tox}} z_{\tilde{\gamma}} \right\} \tag{6.32}$$

By concavity of the target function and linearity of the constraint in $\log(d)$, the sufficient conditions are always satisfied[4]. It follows that the dose for the next patient, $N+1$, is:

$$d_{N+1} = \min\left(d_1, d_2\right), \tag{6.33}$$

where d_1 and d_2 are defined by (6.31) and (6.32).

It appears clearly that the dose d_{N+1} for the next patient is increasing in $\eta, m_{N;1}, m_{N;2}$, decreasing in α_N and non-increasing in each variance terms and in γ. This means that:

- If the level of DoseLimiting Toxicity (DLT) increases, i.e. the threshold of acceptable toxicity level, η, increases, then d_{N+1} is higher, as expected.
- If the target toxicity probability of the MTD decreases, i.e. γ increases, this means that the safety constraint is more stringent, and d_{N+1} will be lower.
- If there is more variability at the observational level (i.e. for concentrations and toxicities) or more uncertainty on the parameters, then

[4] We choose to derivate the Lagrangian function with respect to $\log(d)$ so that the sufficient conditions are globally satisfied.

in order to preserve safety d_{N+1} decreases.

6.5 One-compartmental model: multiple administered doses

The simple assumptions used in Section 6.4 have allowed us to derive analytically the MTD for each new patient. In fact, four assumptions are quite restrictive: known observational variances, administration of a single dose, homogeneity among patients and imposing $\varphi = 1$ in the toxicity model (6.21). We now extend the model in Section 6.4 to include unknown observational variances and multiple administered doses, say D doses, potentially different for each patient $i = 1 : N$, say $d_{i,t} \in \left\{d^*_{i,\tau_1}, .., d^*_{i,\tau_D}\right\}$,with $t_0 \leq \tau_{i,\tau_j} \leq t_T$, $j = 1, .., D$, and $\tau_1 = t_0$. The assumptions of homogeneous patients and of $\varphi = 1$ are instead still maintained and will be relaxed in the next sections. The available data after N patients are: $\mathcal{F}_N = \{(d_{i,t}, c_{i,t}, y_{i,t}), i = 1, .., N; t = t_0, .., t_T\}$, where $d_{i,t} = d^*_{i,\tau}$ if $\tau = t$ and 0 otherwise. The proposed PK model becomes:

$$\log(\boldsymbol{C}_{i,t}) = \log\left(\sum_{j=1}^{D} \frac{d^*_{i,\tau_j}}{\boldsymbol{V}} e^{-\boldsymbol{k}(t-\tau_j)} 1_{t\geq\tau_j}\right) + \boldsymbol{\epsilon}_{i,t} \quad \boldsymbol{\epsilon}_t \mid \boldsymbol{\sigma}^2_C \overset{\text{iid}}{\sim} N\left(0, \boldsymbol{\sigma}^2_C\right). \tag{6.34}$$

where $1_{t\geq\tau_j}$ is the indicator of $t \geq \tau_j$.

The model can be re-written as:

$$\log(\boldsymbol{C}_{i,t}) = \log\left(\sum_{j=1}^{D} d^*_{i,j} e^{-\boldsymbol{k}(t-\tau_j)} 1_{t\geq\tau_j}\right) - \log(\boldsymbol{V}) + \boldsymbol{\epsilon}_{i,t} \text{ where } \boldsymbol{\epsilon}_{i,t} \sim N\left(0, \sigma^2_C\right) \tag{6.35}$$

The toxicity model remains the same as in equation (6.21). We keep the same set of independent prior distributions: $\log(\boldsymbol{V}) \sim N\left(\log(V_0), \sigma^2_V\right)$,

$\boldsymbol{k} \sim N\left(k_0, \sigma_k^2\right)$ and $\alpha \sim N\left(\alpha_0, \sigma_\alpha^2\right)$. We remove the assumption of known variances and we assume that the observational variances, $\boldsymbol{\sigma}_C^2$ and $\boldsymbol{\sigma}_Y^2$ have independent priors. In particular, in the next subsection we will assume independent Inverse-Gamma priors, i.e. $\boldsymbol{\sigma}_C^2 \sim IG\left(a_C, b_C\right)$ and $\boldsymbol{\sigma}_Y^2 \sim IG\left(a_Y, b_Y\right)$. Instead, in Subsection 6.5.3, we will assume independent log-Normal priors, i.e. $\boldsymbol{\sigma}_C^2 \sim logN\left(a_C, b_C\right)$ and $\boldsymbol{\sigma}_Y^2 \sim logN\left(a_Y, b_Y\right)$.

Computations cannot be done analytically anymore, since the PK model cannot be linearized with respect to the parameters using a simple logarithmic transformation. In the next subsection, we describe a MCMC algorithm for finding the MTD for the next patient. Then, in Subsection 6.5.3, we will consider a whole sequence of patients and the procedure will be implemented using a particle filter algorithm.

6.5.1 Finding the dose for the next patient: MCMC

Computation of the posterior and predictive distributions is implemented using MCMC methods, in particular by the Metropolis-Hastings within-Gibbs sampling algorithm. The joint posterior distribution is given by:

$$\pi\left(\log\left(\boldsymbol{V}\right), k, \alpha, \boldsymbol{\sigma}_C^2, \boldsymbol{\sigma}_Y^2 \mid \mathcal{F}_N\right) \propto$$

$$\prod_{i=1}^{N}\prod_{t=t_0}^{t_T} N\left(\log\left(c_{i,t}\right) \mid \log\left(\sum_{j=1}^{D} d_{i,j}^* e^{-\boldsymbol{k}(t-\tau_j)} 1_{t\geq\tau_j}\right) - \log\left(\boldsymbol{V}\right), \boldsymbol{\sigma}_C^2\right) \times$$

$$\prod_{i=1}^{N}\prod_{t=t_0}^{t_T} N\left(\log\left(y_{i,t}\right) \mid \alpha + \log\left(\boldsymbol{C}_{i,t}\right), \boldsymbol{\sigma}_Y^2\right) N\left(\log\left(\boldsymbol{V}\right), \mid \log\left(V_0\right), \sigma_V^2\right) \times$$

$$N\left(\boldsymbol{k} \mid k_0, \sigma_k^2\right) N\left(\boldsymbol{\alpha} \mid \alpha_0, \sigma_\alpha^2\right) \times$$

$$IG\left(\boldsymbol{\sigma}_C^2 \mid a_C, b_C\right)\ IG\left(\boldsymbol{\sigma}_Y^2 \mid a_Y, b_Y\right)$$

where $IG(\cdot \mid a, b)$ is the density of the Inverse-Gamma with parameters a and b.

The MCMC proceeds by sampling iteratively from the full conditionals. At iterations s, $s = 1, .., S$:

- Sample $\log\left(V^{(s)}\right) \mid ... \sim p\left(\log(V) \mid k^{(s-1)}, \sigma_C^{(s-1)}, \mathcal{F}_N\right) = N\left(V_n, \sigma_{V_n}^2\right)$ where

$$V_n = \frac{\sigma_C^{2,(s-1)} \log(V_0) + \sum_{i=1}^{N} \sum_{t=0}^{T} \left(\log\left(\sum_{j=1}^{D} d_{i,j}^* e^{-k(t-\tau_j)} 1_{t \geq \tau_j} \right) - \log(c_{i,t}) \right)}{\sigma_C^{2,(s-1)} + \sigma_V^2 N(T+1)}$$

and $\sigma_{V_n}^2 = \dfrac{\sigma_C^{2,(s-1)} \sigma_V^2}{\sigma_C^{2,(s-1)} + \sigma_V^2 N(T+1)}$.

- Sample $k^{(s)} \mid ... \sim p\left(k \mid \log\left(V^{(s)}\right), \sigma_C^{2,(s-1)}, \mathcal{F}_N\right)$ using a Metropolis - Hastings algorithm:

  - Draw a candidate k^* from $N\left(k_0, \sigma_k^2\right)$

  - Compute the acceptance probability:

    $a\left(k^* \mid k^{(s-1)}\right) = \min\left(1, \dfrac{f(k^*)\, p\left(k^{(s-1)} \mid k^*\right)}{f\left(k^{(s-1)}\right) p\left(k^* \mid k^{(s-1)}\right)}\right)$ where $f(k)$ is the density of the data associated with k, i.e.

$$f(k) = \prod_{i=1}^{N} \prod_{t=t_0}^{t_T} N\left(\log(c_{i,t}) \mid \log\left(\sum_{j=1}^{D} d_{i,j}^* e^{-k\left(t-\tau_j\right)} 1_{t \geq \tau_j} \right) - \log(V), \sigma_C^{2,(s-1)} \right)$$

    and $p(k_1 \mid k_2)$ is the density associated with k_1, given k_2, i.e. $p(k_1 \mid k_2) = N\left(k_1 \mid k_2, \sigma_k^2\right)$.

– With probability $a\left(k^{*} \mid k^{(s-1)}\right)$ set $k^{(s)}=k^{*}$, otherwise set $k^{(s)}=k^{(s-1)}$

- Sample

$$\alpha^{(s)} \mid . \sim N\left(\frac{\alpha_0\sigma_{Tox}^{2,(s-1)}+\sigma_\alpha^2\sum\limits_{i=1}^{N}\sum\limits_{t=t_0}^{t_T}\left(\log\left(y_{i,t}\right)-\log\left(c_{i,t}\right)\right)}{\sigma_{Tox}^{2,(s-1)}+\sigma_\alpha(T+1)N}, \frac{\sigma_{Tox}^{2,(s-1)}\sigma_\alpha^2}{\sigma_{Tox}^{2,(s-1)}+\sigma_\alpha(T+1)N}\right)$$

- Sample

$$\sigma_C^{2,(s)} \mid . \sim IG\left(f_n, b_C+N\,T\right) \text{ where}$$
$$f_n=a_C+\sum_{i=1}^{N}\sum_{t=0}^{T}\left(\log\left(c_{i,t}\right)-\log\left(\sum_{j=1}^{D} d_{i,j}^{*} e^{-k^{(s)}(t-\tau_j)}1_{t\geq\tau_j}\right)+\log\left(V^{(s)}\right)\right)^2.$$

- Sample

$$\sigma_{Tox}^{2,(s)} \mid .. \sim IG\left(a_Y+\sum_{i=1}^{N}\sum_{t=0}^{T}\left(\log\left(y_{i,t}\right)-\alpha_i^{(s)}-\log\left(c_{i,t}\right)\right)^2, b_Y+N\,T\right)$$

The MCMC sample $\Theta^{(s)}=\left(\log\left(V^{(s)}\right), k^{(s)}, \alpha^{(s)}\sigma_C^{2,(s)}, \sigma_{Tox}^{2,(s)}\right)$ can be used for deriving the predictive distribution of the concentration for the next patient associated to a dose d_{N+1} by sampling $\log\left(C^{(s)}\right)$, $s=1,..,S$, $t=t_0,..,t_T$ from

$$N\left(\log\left(C_{N+1,t}^{(s)}\right) \mid \log\left(d_{N+1}\right)-\log\left(V^{(s)}\right)-k^{(s)}t, \sigma_C^{2,(s-1)}\right).$$

The predictive distribution for the log-toxicity for the next patient, $N+1$, can be derived by integrating out the posterior distribution from the distribution of the log-toxicity given the parameters and the log-concentration observations with respect to the posterior distribution for α, i.e.:

$$p\left(\log\left(\boldsymbol{Y}_{N+1,s}\right) \mid \mathcal{F}_N, d_{N+1}\right)=$$

$$\int p\left(\log\left(\boldsymbol{Y}_{N+1s}\right) \mid \alpha, \log\left(C_{N+1,s}\right), d_j\right) p\left(\log\left(C_{N+1,s}\right), \alpha \mid \mathcal{F}_N\right) d\left(\log\left(C_{N+1,s}\right), \alpha\right)$$

$$\text{for } \; s = t_0, .., t_T$$

The resulting distribution $p\left(\log\left(\boldsymbol{Y}_{j,s}\right) \mid \mathcal{F}_N, d_j\right)$ is still Gaussian with parameters $\rho_{j,s}$ and σ_s^2, with

$$\rho_{j,s} = u_{j,s} + \alpha_N$$

$$\sigma_s^2 = \sigma_{Tox}^2 + r_{j,s} + \sigma_{\alpha_N}^2$$

where the expected value $E\left(\log\left(\boldsymbol{C}_{j,s}\right) \mid \mathcal{F}_N^{\setminus y}, d_j\right) \equiv u_{j,s}$ and the variance $Var\left(\log\left(\boldsymbol{C}_{j,s}\right) \mid \mathcal{F}_N^{\setminus y}, d_j\right) \equiv r_s$ are approximated using the MCMC samples.

6.5.2 Minimizing the aggregate posterior expected loss

As before, we want to find the dose d for the next patient N+1. The problem is still defined by expressions (6.18)-(6.19) and it can be re-written as:

$$\min_d \sum_{s=t_0}^{t_T} \delta_t R_\eta\left(\boldsymbol{Y}_{N+1,s}\left(d\right), d\right) = \sum_{s=t_0}^{t_T} \delta_t \left(\sigma_s^2 + \rho_{N+1,s}^2 - 2\eta\rho_{N+1,s} + \eta^2\right)$$

$$\text{so that } \rho_{j,s} + \sqrt{\sigma_s^2} z_{\tilde{\gamma}} - \eta \leq 0 \;\text{ for }\; s = t_0, .., t_T$$

where $z_{\tilde{\gamma}}$ is the $\tilde{\gamma}th$ quantile of a standard Gaussian distribution.

The above expression depends on the expected value $u_{j,s}$ and the variance r_s, which are approximated by MCMC. The overall Bayesian predictive safety inequality is still controlled by using the Boole's inequality. The unconstrained minimum is obtained using numerical optimization

methods, i.e. using the *simplex search method* of Lagarias *et al.*(1998)[5]. Then, the constraint is verified. If the optimal unconstrained minimum does not satisfied the constraint, then the optimal is the highest value of the dose that satisfy the constraint.

6.5.3 Searching the MTD: Particle filter

The analysis carried out till now has pointed out how sequential information should be used for obtaining a *safety* sequence of doses. Another important property of a dose-finding procedure is *consistency*, which ensures that the sequence of doses converges to the MTD. The MTD is indeed defined as the limiting dose of the sequence ($d_N = d$) for $N \to \infty$, where d_N is the solution of the decision problem (6.18) and (6.19). To address consistency, we need to describe how to find the MTD for models (6.35) and (6.21) and *a whole sequence of patients*, while in Subsection 6.5.1 we considered only one patient, $N + 1$. MCMC would be quite inefficient in such a sequential setting. In fact, the sequential nature of the decisional problem requires a sequential estimation procedure, such as particle filters (Doucet *et al.*, 2001). Although Bayesian adaptive trials have been proposed by many authors (Chaloner and Verdinelli, 1995; Müller *et al.*, 2007), there has been little application of these methods to sequential design problems (Gramacy and Polson, 2011; Drovandi *et al.*, 2013). Drovandi *et al.* (2013) discussed the use of particle filters for dose-finding problems, where the optimal dose d_N is found by maximizing a given utility function. We discuss here a similar problem, with some differences. We make use of the Auxiliary Particle Filter (Liu and West, 2001) and the "optimal" dose, here representing the MTD, is the

[5]This algorithm converges slower than others but it always converges to the global minimum. This is a direct search method that does not use numerical or analytic gradients: a simplex in 1-dimensional space is characterized by the 2 distinct vectors that are its vertices. At each step of the search, a new point in or near the current simplex is generated. The function value at the new point is compared with the function values at the vertices of the simplex and, usually, one of the vertices is replaced by the new point, giving a new simplex. This step is repeated until the diameter of the simplex is less than the specified tolerance.

solution of a constrained optimization problem.

For the sake of simplicity, we will here assume that the observational variances have independent log-Normal distributions as prior distributions.

Algorithm

Let assume that the length of the treatment is the same for all patients, say T and let the first dose for the patient 1 be given. Let ψ be the vector of unknown parameters, $\psi = \left(\log(V), k, \alpha, \log\left(\sigma^2_{PK}\right), \log\left(\sigma^2_{Tox}\right)\right)$.

The particle filter algorithm is described as follows. Let S be the total number of particles. For notational simplicity, let us define the algorithm for $j = 2, .., N * T$, where N is the (maximum) total number of (future) patients[6].

Algorithm 6.5.1 *Initialize:*
Draw $\psi^{(s)}$, $s = 1, .., S$, *independently from* $N(\psi \mid \psi_0, \Sigma_0)$. *Set* $w_{1,1}^{PF(s)} = \frac{1}{N}$, $s = 1, .., S$.
For each $j = 2, .., N * T$:

1. *Update parameters using data available till the last observation,* $(d_{j-1}, \log(C_{j-1}), \log(Y_{j-1}))$:
   - *Compute* $\bar{\psi} = E_{j-1}(\psi)$ *and* $\Sigma = Var_{j-1}(\psi)$.
   - *For* $s = 1, ..., S$:
     - *Set* $m^{(s)} = a\psi^{(s)} + (1-a)\bar{\psi}$
     - *Draw a classification variable,* I_s, *with probability* $P(I_s = r) \propto w_{j-1}^{PF,(r)} f\left(\log(C_{j-1}), \log(Y_{j-1}) \mid \psi = m^{(s)}\right)$.
     - *Draw* $\psi^{(s)}$ *from* $N\left(m^{(I_s)}, h^2\Sigma\right)$
     - *Set* $\tilde{w}_j^{PF(s)} = \dfrac{p\left(\log(C_{1:j-1}), \log(Y_{1:j-1}) \mid \psi^s\right)}{p\left(\log(C_{1:j-1}), \log(Y_{1:j-1}) \mid \psi^{I_s}\right)}$

[6] This notation simplifies the following: for each i, with $i = 1, ..., N$, iterate for $t = 2, .., T$ if $i = 1$ and $t = 1, .., T$ for $i > 1$.

– *Normalize the weights:*

$$w_j^{PF(s)} = \frac{\tilde{w}_j^{PF(s)}}{\sum_{s=1}^{S} \tilde{w}_j^{PF(s)}}$$

- *Compute* $N_{eff,j} = \left(\sum_{s=1}^{S} \left(w_j^{PF(s)} \right)^2 \right)^{-1}$. *If* $N_{eff,j} < N_0$ *(e.g. 0.6 * S), resample* $\psi^{(s)}$ *and reset the weights* $w_j^{PF,(s)} = S^{-1}$ *s=1,...,S.*
- *Set* $P\left(\psi | d_{1:j-1}, \log\left(C_{1:j-1}\right), \log\left(Y_{1:j-1}\right)\right) = \sum_{s=1}^{S} w_j^{PF(s)} \delta_{\psi^{(s)}}$.

2. *Select the MTD for the next* j*, i.e. either for the same patient* i *for the next time* $t+1$ *or for the next patient* $i+1$ *for time* 1.
3. *Collect the data, e.g. simulate a new data-point for concentration and toxicity from the true models.*

6.5.4 Simulation study

The aim of this section is to illustrate the proposed decision procedures discussed in Subsections 6.5.1 and 6.5.3.

Single administered dose

We first consider a homogeneous population with a single dose, as in Section 6.4. We illustrate the sequential procedure (6.33), comparing it with the one of Muliere and Petrone (1993), where PK information is not taken into account. The implementation is based on a single simulation scheme. In particular, we compare our proposal with model (6.6), with $d_0 = 0$ and with $\boldsymbol{\beta} \sim N\left(b_0, \sigma_b^2\right)$ and with $\boldsymbol{\sigma}^2$ equal to σ_0^2. In order to compute the two sequences of selected doses, two datasets are simulated according to the following common scheme: starting with an initial dose, d_1, simulate data accordingly to models (6.20) and (6.21). Then, determine the dose for the next patient according with one of the two procedure, d_2, and simulate new data accordingly to models (6.20) and

(6.21), and so on. The simulations are based on the following parameters: $\log(V) = \log(50)$, $k = 0.5$, $\alpha = 2$, $\sigma_C^2 = 0.1^2$, $\sigma_{Tox}^2 = 0.1^2$, $\sigma_V^2 = 0.25^2$, $\sigma_k^2 = 0.1^2$, $\sigma_\alpha^2 = 0.1^2$, $\gamma = 0.9$, $\delta_0 = 0$, $\eta = \log(1)$. The parameters of the prior distribution for the model of Muliere and Petrone (1993) are $b_0 = exp(\alpha)/V$ and $\sigma_b^2 = 0.3^2$, $\sigma_0^2 = \sigma_{Tox}^2 + \sigma_{PK}^2$.

As preliminary study, Figure 6.1 shows the aggregate posterior expected loss corresponding to different choices of the discount factor, δ_0, for a grid of dose values. Without discount factor (i.e. $\delta_0 = 0$), the aggregate posterior expected loss is very high and it decreases by increasing the dose. This is because, with a single dose and a one-compartmental model, it is more likely that the target threshold of toxicity will be reached only in the first period whereas in the next periods the distance between the toxicity and the target will increase. This can explain why the unconstrained optimal solution can be very high, compared to the one obtained using the overall Bayesian predictive safety constraint. The discount factor allows us to give less weight to these further distances and more to the contemporaneous distances. Notice that with $\delta_0 = 2$ the shape of the aggregate posterior expected loss is still the same but the magnitude is very different. Further increases in δ_0 changes the shape and the aggregate posterior expected loss increases for higher doses.

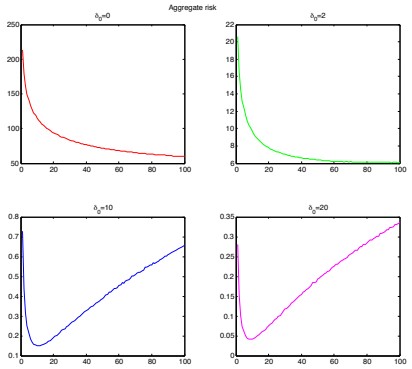


Figure 6.1: Aggregate posterior expected loss, with different choices of discount factor, δ_0, over a discrete grid of doses.

Table 6.1 shows the sequence of doses d_N determined by our procedure, with the associated corresponding observed toxicities, recorded at $0, 6, 12, 18$ hours after the administration of the dose, and the decisions taken using the procedure of Muliere and Petrone, referred as (MP, 1993). Although there is no formal rule defining when stopping the treatment, the illustration shows that both the sequences stabilize, roughly after 1,000 dose administered. The two limiting values represent the MTD associated with the two procedures. The MTD using our proposed procedure is higher than the one obtained using the procedure of Muliere and Petrone (1993). This could suggest that our approach provides a better explanation (and prediction) of the toxicities, allowing for an higher MTD.

Multiple administered doses in the data set

With multiple administered doses, computations cannot be done analytically. In this subsection, we illustrate the procedure for finding the dose for the next patient, whereas in the next subsection we will analyze a whole sequence of doses. Assume $\delta_0 = 4$ and consider simulated data with 5 patients with parameters fixed as in Subsection 6.5.4. The

N	d_N	Y_{N,t_1}	Y_{N,t_2}	Y_{N,t_3}	Y_{N,t_4}	d_N (MP, 1993)
1	10.0000	1.8333	0.0785	0.0032	0.0002	10.0000
2	4.5006	0.2956	0.0225	0.0009	0.0000	2.5103
3	5.1562	0.5408	0.0201	0.0010	0.0001	2.7596
4	5.1837	0.3773	0.0193	0.0009	0.0000	2.7256
5	5.2825	0.4692	0.0246	0.0008	0.0000	2.8011
6	5.3686	0.4206	0.0209	0.0009	0.0000	2.8116
7	5.3108	0.4377	0.0254	0.0011	0.0000	2.8408
8	5.3343	0.3772	0.0243	0.0011	0.0001	2.8580
9	5.3383	0.4788	0.0205	0.0010	0.0001	2.8927
10	5.3357	0.4255	0.0269	0.0011	0.0000	2.8916
...	...	...	...	...	...	...
30	5.1651	0.2924	0.0208	0.0013	0.0001	2.9394
50	5.1547	0.4346	0.0234	0.0010	0.0001	2.9748
70	5.1213	0.4222	0.0223	0.0010	0.0001	2.9882
90	5.1614	0.4516	0.0187	0.0011	0.0001	3.0061
100	5.1550	0.4857	0.0217	0.0012	0.0001	3.0089
...	...	...	...	...	...	...
996	5.1465	0.3528	0.0229	0.0011	0.0001	3.0208
997	5.1465	0.4635	0.0200	0.0011	0.0001	3.0211
998	5.1463	0.4283	0.0208	0.0010	0.0001	3.0211
999	5.1466	0.4756	0.0190	0.0010	0.0001	3.0211
1000	5.1463	0.3878	0.0246	0.0011	0.0000	3.0211
...	...	...	...	...	...	...
9998	5.1463	0.5221	0.0176	0.0009	0.0001	3.0211
9999	5.1463	0.4723	0.0228	0.0011	0.0001	3.0211
10000	5.1463	0.3934	0.0261	0.0013	0.0001	3.0211

Table 6.1: Example of sequential dosage trial. Threshold of acceptable toxicity equals to 1. Having data on the first i patients, the table shows the decisions on the next dose according to the proposed procedure (6.33) and the procedure of Muliere and Petrone (1993) (column d_N (MP, 1993)).

sequence of doses is assumed as given and it is shown in Table 6.2. The data for the concentration and the toxicity are simulated fixing the pa-

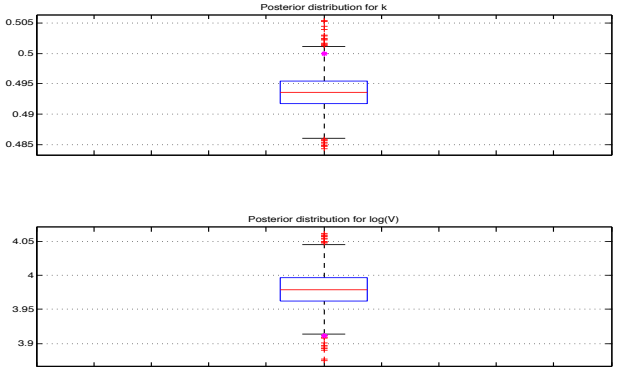
rameters as in the previous subsection. We generated 100, 000 samples

	Patient=1	Patient=2	Patient=3	Patient=4	Patient=5
t_0=1	10	20	30	40	45
t_1=7	0	0	0	0	0
t_2=13	0	0	0	0	0
t_3=19	0	0	0	0	0
t_4=25	15	25	25	35	45
t_5=31	0	0	0	0	0
t_6=37	0	0	0	0	0
t_7=43	0	0	0	0	0
t_8=49	20	25	30	35	45
t_9=55	0	0	0	0	0
t_{10}=61	0	0	0	0	0
t_{11}=67	0	0	0	0	0
t_{12}=73	25	30	25	35	45
t_{13}=79	0	0	0	0	0
t_{14}=85	0	0	0	0	0
t_{15}=91	0	0	0	0	0

Table 6.2: Data on $d_{i,t}$

from the MH-within Gibbs sampling and discarded the first 20,000 as burn-in. For posterior inference, we use 2,000 subsamples from the remaining 80,000 samples, with a thinning equal to 40. Figure 6.8*a* shows the two MCMC approximations of the posterior distributions of the PK parameters.

Due to the small sample size, the impact of the sample variability is quite high. Although not centered on the true (simulated) values (denoted with the pink dot in the figures), the true values are included in within the inter-quantile range (between quantile 0.25 and 0.75) of the posterior distributions.



(a) Posterior Distributions

Figure 6.2: Posterior distributions of the $\boldsymbol{k}$ obtained from the study 6.5.4.

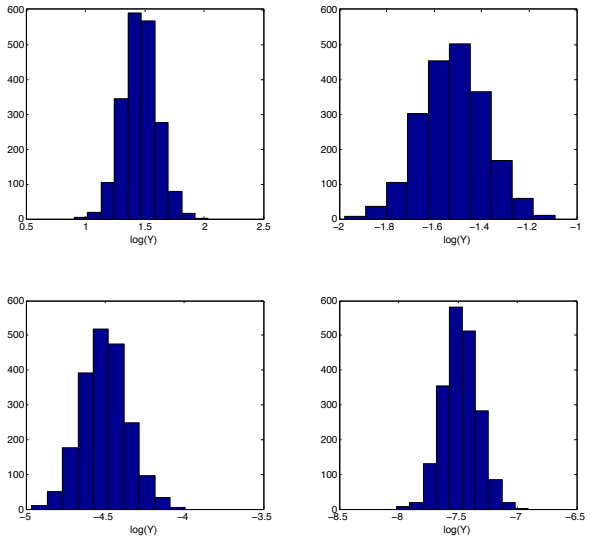
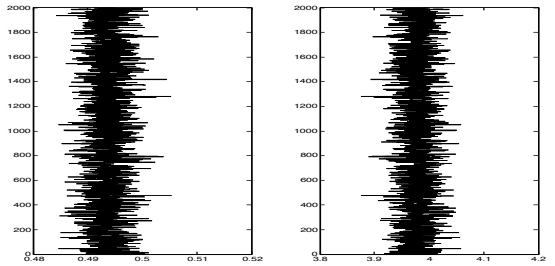


Figure 6.4: Predictive density for the $\log(Y_{N+1,t})$ associated with a dose of 30



(a) Convergence

Figure 6.3: Posterior distributions of the log $(\boldsymbol{V})$ obtained from the study 6.5.4.

The constrained minimum obtained by stochastic approximation is 5.5255.

Sequential decision with multiple doses

We now illustrate the sequential procedure discussed in Subsection 6.5.3. The design schedule is given, i.e. each patient receive 4 doses at times 1, 25, 49, and 73 (hours). Observations are collected at times $1, 7, 13, 19, 25, 31, 37, 43, 49, 55, 61, 67, 73, 79, 85$ and 91 (hours). The parameters used for the simulation are the same as in the previous subsections. The problem is to find the sequence of doses to administer to the sequence of patients. The results from one simulation of 1,000 patients are reported in Table 6.3. The number of particles is set up equal to 1,000.

After 1,000 patients and the collection of data associated with four administered doses to each of them, the MTD starts to stabilize, without reaching a full convergence. However, this limiting dose is just slightly higher than the one found with only one administered dose, although increasing the number of patients, it could decrease.

Patient	$d_{i,1}$	$d_{i,25}$	$d_{i,49}$	$d_{i,73}$
1	10.0000	21.6298	16.6951	9.8740
2	11.3551	12.4675	10.1710	7.4383
3	10.8051	9.2578	8.1640	9.1742
4	7.3142	5.8568	6.1194	5.8611
5	5.5073	5.6287	5.7594	5.2882
25	5.7618	5.7491	5.6166	5.7618
35	5.1733	5.5638	5.9209	5.1758
45	5.4422	5.6044	5.8021	5.3327
55	5.7125	5.5330	5.6778	5.1741
65	5.9444	5.9045	5.1913	5.1623
75	5.7394	5.9602	5.1449	5.2841
85	5.9072	5.5531	5.0146	5.1227
96	5.6545	5.4224	5.2996	5.3510
97	5.9347	5.6691	5.2314	5.3068
98	5.8928	5.8575	5.3116	5.9002
99	5.8901	5.6351	5.6103	5.4178
100	5.4831	5.4873	5.3604	5.4669
997	5.2319	5.2382	5.2315	5.2312
998	5.2418	5.2216	5.2405	5.2375
999	5.2489	5.2298	5.2382	5.2364
1000	5.2296	5.2323	5.2272	5.2401

Table 6.3: Sequential decision on $d_{i,t}$: results from one simulation

6.6 One-compartmental random coefficients model: single administered dose

In clinical trials, the heterogeneity across patients plays a crucial role. We could have heterogeneity among sub-types of diseases and among disease status when the patient starts the treatment. The remaining heterogeneity among the patients can still play a role. If this heterogeneity

is not properly taken into account, we implicitly assume that the whole population is homogeneous and we misleadingly impute this extra variability to the model or to the data quality. Therefore, we now introduce the most realistic assumption of heterogeneity across the patients and use the covariates M_i, $i = 1, .., N$, for partially explaining such heterogeneity. In this section, we assume that all the variances are known and derive the solution to the problem analytically working with data associated with a single administered dose. In the next section, these assumptions are relaxed by allowing unknown observational variances and multiple administered doses. We limit our analysis to the problem of determining the dose d_{N+1} for the next patient.

6.6.1 PK model

Let us consider the following hierarchical model:

I. Model for the data, for $i = 1, .., N$ and $t = t_0, .., t_T$,

$$(\log(\boldsymbol{C}_{i,t}) \mid \omega_i, X_t) = \log(d_i) + \boldsymbol{\omega}_i' X_t + \boldsymbol{\epsilon}_{i,t} \quad \text{with} \quad \boldsymbol{\epsilon}_{i,t} \sim N\left(0, \sigma_C^2\right)$$

II. Model for the Random Coefficients, for $i = 1, .., N$,

$$\boldsymbol{\omega}_i \mid \boldsymbol{\omega}_0, \boldsymbol{\beta} \sim N(\boldsymbol{\omega}_0 + M_i\boldsymbol{\beta}, \Sigma_\omega)$$

III. Priors

$$(\boldsymbol{\omega}_0, \boldsymbol{\beta}) \sim N(m, \Omega)$$

To fix the dimensions, we assume that there is only a patient-specific covariate for each $\boldsymbol{\omega}_{i,j}$, j=1,2, i.e. $M_i = \begin{pmatrix} M_{i;1} & 0 \\ 0 & M_{i;2} \end{pmatrix}$, and two different patient-specific coefficients, i.e. $\log(\boldsymbol{V})$ and $\boldsymbol{k}$, i.e. $\boldsymbol{\beta}$ a 2×1 vector.

We can re-write the above model in a more compact way as follows:

I. Model for the data:

$$\log\left(\underline{\boldsymbol{C}}\right) \mid \boldsymbol{\omega}, A_1^{PK}, \log\left(\underline{d}\right) \sim N\left(\log\left(\underline{d}\right) \otimes \underline{1}_{T+1} + A_1^{PK}\boldsymbol{\omega}, C_1^{PK}\right)$$

II. Model for the Random Coefficients:

$$\boldsymbol{\omega} \mid \boldsymbol{\theta}_2^{PK} \sim N\left(A_2^{PK}\boldsymbol{\theta}_2^{PK}, I_N \otimes \Sigma_\omega\right)$$

III. Priors:

$$\boldsymbol{\theta}_2^{PK} \sim N\left(m, C_3^{PK}\right)$$

where $\otimes$ denotes the Kronecker product,

$\boldsymbol{\omega} = \begin{pmatrix} \boldsymbol{\omega}_1 \\ ... \\ \boldsymbol{\omega}_i \\ ... \\ \boldsymbol{\omega}_N \end{pmatrix}$ with $\boldsymbol{\omega}_i$ 2×1, A_1^{PK} is the $N(T+1) \times 2N$: $A_1^{PK} = I_N \otimes X$, where

$$X = \begin{pmatrix} X_0' \\ ... \\ X_t' \\ ... \\ X_T' \end{pmatrix} \text{ where } X_t' \text{ is a } 1 \times 2 \text{ vector.}$$

$$\boldsymbol{\theta}_2^{PK} = \begin{pmatrix} \boldsymbol{\omega}_0 \\ \boldsymbol{\beta} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\omega}_{01} \\ \boldsymbol{\omega}_{02} \\ \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} \text{ is a } 4 \times 1 \text{ vector,}$$

$$A_2^{PK} = \begin{pmatrix} 1 & 0 & M_{1,1} & 0 \\ 0 & 1 & 0 & M_{1,2} \\ 1 & 0 & M_{2,1} & 0 \\ 0 & 1 & 0 & M_{2,2} \\ ... & ... & ... & ... \\ 1 & 0 & M_{N,1} & 0 \\ 0 & 1 & 0 & M_{N,2} \end{pmatrix}$$

$C_1^{PK} = \sigma_C^2 I_{N(T+1)}$, and now m is a 4×1 vector and C_3^{PK} is a 4×4 matrix.

Under these assumptions, we can derive the posterior distributions, which are:

$$\boldsymbol{\omega} \mid \mathcal{F}_N \sim N\left(m_N, \Omega_N\right)$$

with

$$\Omega_N = \left(A_1^{PK,\prime} C_1^{PK,-1} A_1^{PK} + \left(I_N \otimes \Sigma_\omega + A_2^{PK} C_3^{PK} A_2^{PK,\prime}\right)^{-1}\right)^{-1} \equiv$$

$$\begin{pmatrix} ... & ... & ... & ... \\ ... & \omega_{11i} & \omega_{12i} & ... \\ ... & \omega_{12i} & \omega_{22i} & ... \\ ... & ... & ... & ... \end{pmatrix}$$

$$m_N = \Omega_N \left(A_1^{PK,\prime} C_1^{PK,-1} \left(\log\left(\underline{C}\right) - \log\left(\underline{d}\right) \otimes \underline{1}_{T+1}\right) + \left(I_N \otimes \Sigma_\omega + A_2^{PK} C_3^{PK} A_2^{PK,\prime}\right)^{-1} A_2^{PK} m\right)$$

where $m_N \equiv \begin{pmatrix} m_{N,1} \\ ... \\ m_{N,N} \end{pmatrix}$ with $m_{N,i} \equiv \begin{pmatrix} m_{N,i;1} \\ m_{N,i;2} \end{pmatrix}$

and

$$\boldsymbol{\theta}_2^{PK} \mid \mathcal{F}_N \sim N\left(m_0, \Omega_0\right)$$

with

$$\Omega_0 = \left(A_2^{PK\prime} A_1^{PK,\prime} \left(C_1^{PK} + A_1^{PK} I_N \otimes \Sigma_\omega A_1^{PK\prime} \right)^{-1} A_1^{PK} A_2^{PK} + C_3^{PK} \right)^{-1} \equiv$$

$$\begin{pmatrix} \omega_{110} & \omega_{120} & \omega_{130} & \omega_{140} \\ \omega_{120} & \omega_{220} & \omega_{230} & \omega_{240} \\ \omega_{130} & \omega_{230} & \omega_{330} & \omega_{340} \\ \omega_{140} & \omega_{240} & \omega_{340} & \omega_{440} \end{pmatrix}$$

$$m_0 = \Omega_0 \left(A_2^{PK\prime} A_1^{PK,\prime} \left(C_1^{PK} + A_1^{PK} I_N \otimes \Sigma_\omega A_1^{PK\prime} \right)^{-1} \left(\log\left(\underline{c}\right) - \log\left(d\right) \otimes \underline{1}_{T+1} \right) + C_3^{PK\prime} m \right) \equiv$$

$$\begin{pmatrix} m_{10} \\ m_{20} \\ m_{30} \\ m_{40} \end{pmatrix}$$

The patient-specific coefficients for the new patient $N+1$, $\boldsymbol{\omega}_{N+1}$, have a Gaussian population distribution with mean $\boldsymbol{\omega}_0 + \boldsymbol{\beta} M_{N+1}$ and variance Σ_ω, and the distribution of $(\boldsymbol{\omega}_0, \boldsymbol{\beta})$ is the posterior distribution with parameters m_0, Ω_0. It follows that:

$$p\left(\log\left(\boldsymbol{C}_{N+1,s}\right) \mid \mathcal{F}_N, d_{N+1}\right) = \int p\left(\log\left(\boldsymbol{C}_{N+1,s}\right) \mid \boldsymbol{\theta}_2^{PK}, d_{N+1}\right) p\left(\theta_2^{PK} \mid \mathcal{F}_N\right) d\theta_2^{PK}$$

for $s = t_0, t_0+1, ...t_T$.

This distribution is Gaussian:

$$\log\left(\boldsymbol{C}_{N+1,s}\right) \mid \mathcal{F}_N, d_{N+1} \sim N\left(u_{N+1,s}, r_{N+1,s}\right)$$

with :

$$u_{N+1,s} = \log\left(d_{N+1}\right) + X_s' \begin{pmatrix} 1 & 0 & M_{N+1;1} & 0 \\ 0 & 1 & 0 & M_{N+1;2} \end{pmatrix} m_0$$

$$r_{N+1,s} = \sigma_C^2 +$$

$$X_s' \left(\Sigma_\omega + \begin{pmatrix} 1 & 0 & M_{N+1;1} & 0 \\ 0 & 1 & 0 & M_{N+1;2} \end{pmatrix} \Omega_0 \begin{pmatrix} 1 & 0 & M_{N+1;1} & 0 \\ 0 & 1 & 0 & M_{N+1;2} \end{pmatrix}' \right) X_s$$

6.6.2 Toxicity model

Concerning the toxicity, we maintan the simplest toxicity model but now allowing patient-specific coefficients, $\boldsymbol{\alpha}_i$, and covariates P_i, $i = 1, .., N$ for the random coefficients α_i. For the sake of simplicity, we assume that P_i is univariate. We consider the following hierarchical model:

I. The model for the data, for i=1,..,N; $t = t_0, .., t_T$ is given by (6.21), re-written as:

$$\log(\boldsymbol{Y}_{i,t}) \mid \log(\boldsymbol{C}_{i,t}) = \boldsymbol{\alpha}_i + \log(\boldsymbol{C}_{i,t}) + \boldsymbol{\xi}_{i,t} \quad \text{where} \quad \boldsymbol{\xi}_{i,t} \sim N\left(0, \sigma_{Tox}^2\right).$$

II. The model for random coefficients, for i=1,..,N, is:

$$\boldsymbol{\alpha}_i \sim N\left(\boldsymbol{\alpha}_0 + \boldsymbol{\kappa} P_i, \sigma_\alpha^2\right)$$

III. Prior:

$$\boldsymbol{\alpha}_0 \sim N\left(a_1, \sigma_{a;1}^2\right) \quad \text{and} \quad \boldsymbol{\kappa} \sim N\left(\kappa_0, \sigma_{a;2}^2\right)$$

The model can be re-written in a more compact way as follows:

I. Model for the data:

$$\log(\underline{\boldsymbol{Y}}) \mid \log(\underline{\boldsymbol{C}}) = A_1^{Tox} \boldsymbol{\alpha} + \log(\underline{\boldsymbol{C}}) + \boldsymbol{\xi} \quad \text{where} \quad \boldsymbol{\xi} \sim N\left(0, C_1^{Tox}\right).$$

II. Model for the random coefficients:

$$\boldsymbol{\alpha} \sim N\left(A_2^{Tox} \boldsymbol{\theta}_2^{Tox}, C_2^{Tox}\right)$$

III. Priors:

$$\boldsymbol{\theta}_2^{Tox} \sim N\left(a, \Sigma_a\right)$$

$$\text{where } \boldsymbol{\alpha} = \begin{pmatrix} \boldsymbol{\alpha}_1 \\ ... \\ \boldsymbol{\alpha}_N \end{pmatrix}, \boldsymbol{\theta}_2^{Tox} = \begin{pmatrix} \boldsymbol{\alpha}_0 \\ \boldsymbol{\kappa} \end{pmatrix}, a = \begin{pmatrix} a_1 \\ \kappa_0 \end{pmatrix}, \Sigma_a = \begin{pmatrix} \sigma_{a;1}^2 & 0 \\ 0 & \sigma_{a;2}^2 \end{pmatrix}$$

$$A_1^{Tox} = I_N \otimes \underline{1}_{T+1},\ C_1^{Tox} = \sigma_{Tox}^2 I_{N(T+1)},\ A_2^{Tox} = \begin{pmatrix} 1 & P_1 \\ ... & ... \\ 1 & P_i \\ ... & ... \\ 1 & P_N \end{pmatrix},$$

$$C_2^{Tox} = \sigma_\alpha^2 I_N$$

The posterior distribution is still Gaussian:

$$\boldsymbol{\alpha} \mid \mathcal{F}_N \sim N\left(\alpha_N, \sigma_{\alpha_N}^2\right)$$

with

$$\sigma_{\alpha_N}^2 = \left(A_1^{Tox,\prime} C_1^{Tox,-1} A_1^{Tox} + \left(C_2^{Tox} + A_2^{Tox} \Sigma_a A_2^{Tox,\prime}\right)^{-1}\right)^{-1}$$

$$\alpha_N = \sigma_{\alpha_N}^2 \left(A_1^{Tox,\prime} C_1^{Tox,-1} \left(\log\left(\underline{y}\right) - \log\left(\underline{c}\right)\right) + \left(C_2^{Tox} + A_2^{Tox} \Sigma_a A_2^{Tox,\prime}\right)^{-1} A_2^{Tox} a\right) =$$

$$=^{def} \begin{pmatrix} \alpha_{T1} \\ ... \\ \alpha_{TN} \end{pmatrix}$$

$$\boldsymbol{\theta}_2 \mid \mathcal{F}_N \sim N\left(a_n, v_n^2\right)$$

with

$$v_n^2 = \left(A_2^{Tox\prime} A_1^{Tox\prime} \left(C_1^{Tox} + A_1^{Tox} C_2^{Tox} A_1^{Tox,\prime} \right)^{-1} A_1^{Tox} A_2^{Tox} + \Sigma_a \right)^{-1}$$

$$a_n = v_n^2 \left(A_2^{Tox\prime} A_1^{Tox\prime} \left(C_1^{Tox} + A_1^{Tox} C_2^{Tox} A_1^{Tox,\prime} \right)^{-1} \left(\log\left(\underline{y}\right) - \log\left(\underline{c}\right) \right) + \Sigma_a' a \right)$$

As before, the predictive distribution for the log-toxicity for the next patient $N+1$ can be derived by integrating out the posterior distribution from the distribution of the log-toxicity given parameters and the log-concentration observations with respect to the posterior population mean $\boldsymbol{\alpha}_0$, i.e.:

$$p\left(\log\left(\boldsymbol{Y}_{N+1,s}\right) \mid \mathcal{F}_N, d_{N+1}\right) =$$

$$\int p\left(\log\left(\boldsymbol{Y}_{N+1,s}\right) \mid \theta_2^{Tox}, \log\left(C_{N+1,s}\right), d_{N+1}\right) p\left(\theta_2^{Tox}, \log\left(C_{N+1,s}\right) \mid \mathcal{F}_N\right) d\left(\theta_2^{Tox}, \log\left(C_{N+1,s}\right)\right)$$

for $s = t_0, .., t_T$. The resulting distribution $p\left(\log\left(\boldsymbol{Y}_{N+1,s}\right) \mid \mathcal{F}_N, d_{N+1}\right)$ is still Gaussian with parameters $\rho_{N+1,s}$ and $\sigma^2_{N+1,s}$, with

$$\rho_{N+1,s} = u_{N+1,s} + a_n' \begin{pmatrix} 1 \\ P_{N+1} \end{pmatrix}$$

$$\sigma_s^2 = \sigma_{Tox}^2 + r_{N+1,s} + \sigma_\alpha^2 + \begin{pmatrix} 1 & P_{N+1} \end{pmatrix} v_n^2 \begin{pmatrix} 1 \\ P_{N+1} \end{pmatrix}$$

6.6.3 Constrained minimization problem

As for the homogeneity cases, the problem is defined by expressions (6.18)-(6.19), and, under the quadratic loss assumptions and the chosen priors, the aggregate posterior expected loss is expressed by equation (6.27). Substituting the just-defined posterior parameters, the aggregate

posterior expected loss becomes:

$$\sum_{s=t_0}^{t_T} \delta_s R_\eta \left(\boldsymbol{Y}_{N+1,s}\left(d\right), d\right) = \log\left(d\right)^2 \sum_{s=t_0}^{t_T} \delta_s + \\ +2\log\left(d\right) \sum_{s=t_0}^{t_T} \delta_s \left(X_s' \begin{pmatrix} 1 & 0 & M_{N+1;1} & 0 \\ 0 & 1 & 0 & M_{N+2;2} \end{pmatrix} m_0 + a_n' \begin{pmatrix} 1 \\ P_{N+1} \end{pmatrix} - \eta \right) + a'$$

where a' is a constant irrelevant for the minimization. Similarly to Subsection 6.4.3, for satisfying the overall Bayesian predictive safety constraint (6.19), we still use the Boole's inequality and, substituting the values of $\rho_{j,s}$ and σ_s^2 in expression (6.29), we can obtain:

$$\log\left(d\right) \leq \eta - X_s' \begin{pmatrix} 1 & 0 & M_{N+1;1} & 0 \\ 0 & 1 & 0 & M_{N+1;2} \end{pmatrix} m_0 - a_n' \begin{pmatrix} 1 \\ P_{N+1} \end{pmatrix} + \\ - \sqrt{\sigma_{Tox}^2 + r_{N+1,s} + \sigma_\alpha^2 + \begin{pmatrix} 1 & P_{N+1} \end{pmatrix} v_n^2 \begin{pmatrix} 1 \\ P_{N+1} \end{pmatrix}} z_{\tilde{\gamma}}$$

for each s. It follows that:

$$\log\left(d\right) \leq \min_s \left(\eta - X_s' \begin{pmatrix} 1 & 0 & M_{N+1;1} & 0 \\ 0 & 1 & 0 & M_{N+1;2} \end{pmatrix} m_0 - a_n' \begin{pmatrix} 1 \\ P_{N+1} \end{pmatrix} - \sqrt{\sigma_s^2} z_{\tilde{\gamma}} \right)$$

Call s^* the one that satisfies the equality. The Lagrangian function of this problem is:

$$\mathcal{L} = \log\left(d\right)^2 \sum_{s=t_0}^{t_T} \delta_s + \\ +2\log\left(d\right) \sum_{s=t_0}^{t_T} \delta_s \left(X_s' \begin{pmatrix} 1 & 0 & M_{N+1;1} & 0 \\ 0 & 1 & 0 & M_{n+1;2} \end{pmatrix} m_0 + a_n' \begin{pmatrix} 1 \\ P_{N+1} \end{pmatrix} - \eta \right) + a' \\ +\lambda \left(\eta - X_s' \begin{pmatrix} 1 & 0 & M_{N+1;1} & 0 \\ 0 & 1 & 0 & M_{N+1;2} \end{pmatrix} m_0 - a_n' \begin{pmatrix} 1 \\ P_j \end{pmatrix} - \sqrt{\sigma_s^2} z_{\tilde{\gamma}} - \log\left(d\right) \right)$$

Necessary conditions requires:

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial \log(d)} \leq 0; \quad \log(d) \geq 0; \quad \dfrac{\partial \mathcal{L}}{\partial \log(d)} \log(d) = 0 \\ \log(d) \leq \eta - X'_{s*} \begin{pmatrix} 1 & 0 & M_{N+1;1} & 0 \\ 0 & 1 & 0 & M_{N+1;2} \end{pmatrix} m_0 - a'_n \begin{pmatrix} 1 \\ P_j \end{pmatrix} - \sqrt{\sigma_s^2} z_{\tilde{\gamma}} \\ \lambda \left(\eta - X'_{s*} \begin{pmatrix} 1 & 0 & M_{N+1;1} & 0 \\ 0 & 1 & 0 & M_{N+1;2} \end{pmatrix} m_0 - a'_n \begin{pmatrix} 1 \\ P_{N+1} \end{pmatrix} - \sqrt{\sigma_s^2} z_{\tilde{\gamma}} - \log(d) \right) = 0 \\ \lambda \geq 0 \end{cases}$$

The unconstrained and border solutions are:

- Unconstrained solution ($\lambda = 0$):

$$d_1 = e^{-\dfrac{\sum_{s=t_0}^{t_T} \delta_s \left(X'_s \begin{pmatrix} 1 & 0 & M_{N+1;1} & 0 \\ 0 & 1 & 0 & M_{N+1;2} \end{pmatrix} m_0 + a'_n \begin{pmatrix} 1 \\ P_{N+1} \end{pmatrix} - \eta \right)}{\sum_{s=t_0}^{t_T} \delta_s}} \tag{6.36}$$

- Boundary solution ($\lambda \neq 0$):

$$d_2 = e^{\eta - X'_{s*} \begin{pmatrix} 1 & 0 & M_{N+1;1} & 0 \\ 0 & 1 & 0 & M_{N+1;2} \end{pmatrix} m_0 - a'_n \begin{pmatrix} 1 \\ P_{N+1} \end{pmatrix}} \times$$

$$e^{-\sqrt{\sigma^2_{Tox} + r_{s*} + \sigma^2_\alpha + \begin{pmatrix} 1 & P_{N+1} \end{pmatrix} v_n^2 \begin{pmatrix} 1 \\ P_{N+1} \end{pmatrix}} z_{\tilde{\gamma}}}$$

By concavity of the function, the sufficient conditions are always sat-

isfied. It follows that the MTD for the next patient is:

$$d^* = \min(d_1, d_2) \tag{6.37}$$

6.7 One-compartmental random coefficients model: multiple administered doses

In each of the previous sections, we have gradually added some element of complexity, which brought us to a more realistic scenario. We now consider the most general case. Within the same framework of the last section, we now consider data associated with multiple administered doses and heterogeneity partially explained using covariates. The available data consist of data for patients with D administered doses potentially different for each patient $i = 1, .., N$, say $d^*_{i,\tau_1}, .., d^*_{i,\tau_D}$,with $0 \leq \tau_{i,\tau_j} \leq T$, $j = 1, .., D$. Without loss of generality, assume the first administration time is $\tau_1 = t_0$.

6.7.1 PK model

The corresponding one-compartment PK model is expressed by model (6.34) and it can be re-written as model (6.35). As before, we assume that $\log(\boldsymbol{V})$ and $\boldsymbol{k}$ are independent, although removing this assumption is straightforward. The hierarchical model becomes:

I. Model for the data, for $i = 1, .., N$ and $t = t_0, .., t_T$,

$$\log(\boldsymbol{C}_{i,t}) = \log\left(\sum_{j=1}^{D} d^*_{i,j} e^{-\boldsymbol{k}_i(t-\tau_j)} 1_{t\geq\tau_j}\right) - \log(\boldsymbol{V}_i) + \log(\boldsymbol{\epsilon}_{i,t})$$

where $\log(\boldsymbol{\epsilon}_{i,t}) \sim N\left(0, \boldsymbol{\sigma}^2_C\right)$.

II. Model for the random coefficients, for $i = 1, .., N$:

$$\begin{pmatrix} \log(\boldsymbol{V}_i) \\ \boldsymbol{k}_i \end{pmatrix} \mid \log(\boldsymbol{V}_0), \boldsymbol{k}_0, \boldsymbol{\beta} \sim N\left(\begin{pmatrix} \log(\boldsymbol{V}_0) \\ \boldsymbol{k}_0 \end{pmatrix} + \begin{pmatrix} M_{i,1} & 0 \\ 0 & M_{i,2} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}, \begin{pmatrix} \sigma_V^2 & 0 \\ 0 & \sigma_k^2 \end{pmatrix} \right)$$

III. Priors:

$$\begin{pmatrix} \log(\boldsymbol{V}_0) \\ \boldsymbol{k}_0 \\ \boldsymbol{\beta} \end{pmatrix} \sim N\left(\begin{pmatrix} m_V \\ m_k \\ m_{\beta_1} \\ m_{\beta_2} \end{pmatrix}, \begin{pmatrix} c_V^2 & 0 & 0 & 0 \\ 0 & c_k^2 & 0 & 0 \\ 0 & 0 & c_{\beta_1}^2 & 0 \\ 0 & 0 & 0 & c_{\beta_2}^2 \end{pmatrix} \right)$$

6.7.2 Toxicity model

We now consider the most general toxicity model defined by equation (6.9) with coefficients expressed by using the exponential Almon specification defined by (6.11) for each i. This specification implies that the most recent concentrations have the highest impact on the current toxicity. We assume that $Q = 8$. We assume that some covariates P_i, $i = 1, .., N$ can partially explain the random effects $\boldsymbol{\alpha}_i$. For the sake of simplicity, we still assume that P_i is univariate. Let us assume the following hierarchical model:

I. Model for the data, for i=1,..,N; $t = t_0, .., t_T$,

$$\log(\boldsymbol{Y}_{i,t}) = \boldsymbol{\alpha}_i + \log\left(\sum_{r=max(t_{-Q}, t_0)}^{t} \boldsymbol{\varphi}_{r;i} \boldsymbol{C}_{i,r} \right) + \boldsymbol{\xi}_{i,t} \quad \text{where} \quad \boldsymbol{\xi}_{i,t} \sim N\left(0, \boldsymbol{\sigma}_Y^2\right).$$

II. Model for the random coefficients, for i=1,..,N:

$$\begin{pmatrix} \boldsymbol{\alpha}_i \\ \boldsymbol{\theta}_{1;i} \\ \boldsymbol{\theta}_{2;i} \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{\alpha}_0 + \kappa P_i \\ \boldsymbol{\theta}_{01} \\ \boldsymbol{\theta}_{02} \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & 0 & 0 \\ 0 & \sigma_{\theta_1} & 0 \\ 0 & 0 & \sigma_{\theta_2} \end{pmatrix}\right)$$

III. Priors:

$$\begin{pmatrix} \boldsymbol{\alpha}_0 \\ \boldsymbol{\kappa} \\ \boldsymbol{\theta}_{01} \\ \boldsymbol{\theta}_{02} \end{pmatrix} \sim N\left(\begin{pmatrix} m_\alpha \\ m_\kappa \\ m_{\theta_1} \\ m_{\theta_2} \end{pmatrix}, \begin{pmatrix} c_\alpha^2 & 0 & 0 & 0 \\ 0 & c_\kappa^2 & 0 & 0 \\ 0 & 0 & c_{\theta_1}^2 & 0 \\ 0 & 0 & 0 & c_{\theta_2}^2 \end{pmatrix}\right)$$

We also assume that the observational variances are unknown and have independent Inverse-Gamma priors, i.e. $\boldsymbol{\sigma}_C^2 \sim IG(a_C, b_C)$ and $\boldsymbol{\sigma}_Y^2 \sim IG(a_Y, b_Y)$.

Again, computations cannot be done analytically anymore, since the PK model and the toxicity model cannot be linearized with respect to the parameters and in particular with regards of $\boldsymbol{k}_i$ and $\boldsymbol{\theta}_{1;i}, \boldsymbol{\theta}_{2;i}$ using a simple logarithmic transformation.

6.7.3 MTD for the next patient: MCMC

The computations of the posterior and predictive distributions are implemented using a Metropolis-Hastings algorithm within-Gibbs sampling. The joint posterior distribution is given by:

$$\pi\left(\log(\boldsymbol{V}_i), \boldsymbol{k}_i, \boldsymbol{\alpha}_i, \boldsymbol{\theta}_{1,i}, \boldsymbol{\theta}_{2,i}, i = 1, .., N, \log(\boldsymbol{V}_0), k_0, \boldsymbol{\alpha}_0, \boldsymbol{\theta}_{01}, \boldsymbol{\theta}_{02}, \boldsymbol{\sigma}_C^2, \boldsymbol{\sigma}_Y^2 \mid \mathcal{F}_N\right) \propto$$

$$\prod_{i=1}^{N} \prod_{t=t_0}^{t_T} N\left(\log(c_{i,t}) \mid \log\left(\sum_{j=1}^{D} d_{i,j}^* e^{-\boldsymbol{k}_i(t-\tau_j)} 1_{t \geq \tau_j}\right) - \log(\boldsymbol{V}_i), \boldsymbol{\sigma}_C^2\right) \times$$

$$\prod_{i=1}^{N}\prod_{t=t_0}^{t_T} N\left(\log\left(y_{i,t}\right) \mid \boldsymbol{\alpha}_i + \log\left(\sum_{r=max(t_{-Q},t_0)}^{t} \boldsymbol{\varphi}_{r;i}\boldsymbol{C}_{i,r}\right), \boldsymbol{\sigma}_Y^2\right) \times$$

$$\prod_{i=1}^{N} N\left(\begin{pmatrix} \log\left(\boldsymbol{\sigma}_i\right) \\ \boldsymbol{k}_i \end{pmatrix} \mid \left(\begin{pmatrix} \log\left(\boldsymbol{V}_0\right) \\ \boldsymbol{k}_0 \end{pmatrix} + \begin{pmatrix} M_{i,1} & 0 \\ 0 & M_{i,2} \end{pmatrix}\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \begin{pmatrix} \sigma_V^2 & 0 \\ 0 & \sigma_k^2 \end{pmatrix}\right)\right) \times$$

$$\prod_{i=1}^{N} N\left(\begin{pmatrix} \boldsymbol{\alpha}_i \\ \boldsymbol{\theta}_{1;i} \\ \boldsymbol{\theta}_{2;i} \end{pmatrix} \mid \begin{pmatrix} \boldsymbol{\alpha}_0 + \boldsymbol{\kappa}P_i \\ \boldsymbol{\theta}_{01} \\ \boldsymbol{\theta}_{02} \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & 0 & 0 \\ 0 & \sigma_{\theta_1} & 0 \\ 0 & 0 & \sigma_{\theta_2} \end{pmatrix}\right) \times$$

$$N\left(\begin{pmatrix} \log\left(\boldsymbol{V}_0\right) \\ \boldsymbol{k}_0 \\ \boldsymbol{\beta} \end{pmatrix} \mid \begin{pmatrix} m_V \\ m_k \\ m_{\beta_1} \\ m_{\beta_2} \end{pmatrix}, \begin{pmatrix} c_V^2 & 0 & 0 & 0 \\ 0 & c_k^2 & 0 & 0 \\ 0 & 0 & c_{\beta_1}^2 & 0 \\ 0 & 0 & 0 & c_{\beta_2}^2 \end{pmatrix}\right) \times$$

$$N\left(\begin{pmatrix} \boldsymbol{\alpha}_0 \\ \boldsymbol{\kappa} \\ \boldsymbol{\theta}_{01} \\ \boldsymbol{\theta}_{02} \end{pmatrix} \mid \begin{pmatrix} m_\alpha \\ m_\kappa \\ m_{\theta_1} \\ m_{\theta_2} \end{pmatrix}, \begin{pmatrix} c_\alpha^2 & 0 & 0 & 0 \\ 0 & c_\kappa^2 & 0 & 0 \\ 0 & 0 & c_{\theta_1}^2 & 0 \\ 0 & 0 & 0 & c_{\theta_2}^2 \end{pmatrix}\right) \times$$

$$IG\left(\boldsymbol{\sigma}_C^2 \mid a_C, b_C\right)\ IG\left(\boldsymbol{\sigma}_Y^2 \mid a_Y, b_Y\right)$$

The above posterior distribution cannot be written in closed form and as before a Metropolis-Hastings algorithm within Gibbs-sampling is implemented. Moreover, given the observations, $\log\left(c_{i,t}\right), i = 1, .., N; t = t_0, .., t_T$ and the independence assumptions, the PK parameters and the toxicity parameter are again independent.

The Metropolis-Hastings algorithm within Gibbs-sampling samples from the following full conditional distributions for $s = 1, .., S$:

- For $i = 1, .., N$:

  - Sample
    $\log\left(V_i^{(s)}\right) \mid ... \sim p\left(\log\left(V_i\right) \mid k_i^{(s-1)}, \log\left(V_0^{(s-1)}\right), \beta_1^{(s-1)}, \sigma_C^{(s-1)}, \mathcal{F}_N\right) = N\left(V_n, \sigma_{V_n}^2\right)$
    where V_n is given by:

$$\frac{\sigma_V^2 \sum_{t=t_0}^{t_T}\left(\log\left(c_{i,t}\right) - \log\left(\sum_{j=1}^{D} d_{i,j}^* e^{-k_i^{(s-1)}\left(t-\tau_j\right)} 1_{t\geq\tau_j}\right)\right) + \sigma_C^{(s-1),2}\left(\log\left(V_0^{(s-1)}\right) + M_{i,1}\beta_1^{(s-1)}\right)}{\sigma_V^2\left(T+1\right) + \sigma_C^{(s-1),2}}$$

and $\sigma_{V_n}^2 = \dfrac{\sigma_V^2 \sigma_C^{(s-1),2}}{\sigma_V^2\left(T+1\right) + \sigma_C^{(s-1),2}}$

  - Sample $k_i^{(s)} \mid ... \sim p\left(k_i \mid k_0^{(s-1)}, \beta_2^{(s-1)}, \log\left(V_i^{(s)}\right), \sigma_C^{(s-1),2}, \mathcal{F}_N\right)$ using a Metropolis - Hastings algorithm:

    * Draw a candidate k_i^* from $N\left(k_0^{(s-1)} + M_{i,2}\beta_2^{(s-1)}, \sigma_k^2\right)$

    * Compute the acceptance probability:
      $a\left(k_i^* \mid k_i^{(s-1)}\right) = \min\left(1, \dfrac{f\left(k_i^*\right) p\left(k_i^{(s-1)} \mid k_i^*\right)}{f\left(k_i^{(s-1)}\right) p\left(k_i^* \mid k_i^{(s-1)}\right)}\right)$ where $f\left(k_i\right)$ is the density of the data associated with k_i, i.e.

$$f\left(k_i\right) = \prod_{t=t_0}^{t_T} N\left(\log\left(c_{i,t}\right) \mid \log\left(\sum_{j=1}^{D} d_{i,j}^* e^{-k_i\left(t-\tau_j\right)} 1_{t\geq\tau_j}\right) - \log\left(V_i\right), \sigma_C^{(s-1),2}\right)$$

      and $p\left(k_1 \mid k_2\right)$ is the density associated with k_1, given k_2, i.e. $p\left(k_1 \mid k_2\right) = N\left(k_1 \mid k_2, \sigma_k^2\right)$.

    * With probability $a\left(k_i^* \mid k_i^{(s-1)}\right)$ set $k_i^{(s)} = k_i^*$, otherwise set $k_i^{(s)} = k_i^{(s-1)}$

– Sample

$\alpha_i^{(s)} \mid ... \sim p\left(\alpha_i \mid \theta_{1;i}^{(s-1)}, \theta_{2;i}^{(s-1)}, \alpha_0^{(s-1)}, \kappa^{(s-1)}, \sigma_{Tox}^{(s-1)}, \mathcal{F}_N\right) = N\left(\alpha_n, \sigma_{\alpha_n}^2\right)$

where

$$\alpha_n = \frac{\sigma_\alpha^2 \sum_{t=t_0}^{t_T} \left(\log\left(y_{i,t}\right) - \log\left(\sum_{r=max(t_{-Q},t_0)}^{t} \varphi_{r;i}^{(s-1)} c_{i,r}\right)\right) + \sigma_{Tox}^{(s-1),2}\left(\alpha_0^{(s-1)} + \kappa^{(s-1)} P_i\right)}{\sigma_\alpha^2 \left(T+1\right) + \sigma_{Tox}^{(s-1),Y}}$$

and

$$\sigma_{\alpha_n}^2 = \frac{\sigma_\alpha^2 \sigma_{Tox}^{(s-1),2}}{\sigma_\alpha^2 \left(T+1\right) + \sigma_{Tox}^{(s-1),2}}, \text{ with } \varphi_{r;i}^{(s-1)} = \frac{e^{\theta_{1;i}^{(s-1)} r + \theta_{2;i}^{(s-1)} r^2}}{\sum_{r=max(t_{-Q},t_0)}^{t} e^{\theta_{1;i}^{(s-1)} r + \theta_{2;i}^{(s-1)} r^2}}$$

– Sample $\varphi_{r;i}^{(s)} \mid ...$ by sampling $\theta_{i,1}^{(s)} \mid ... \sim p\left(\theta_{i,1} \mid \theta_{01}^{(s-1)}, \mathcal{F}_N\right)$ and $\theta_{i,2}^{(s)} \mid ... \sim p\left(\theta_{i,2} \mid \theta_{02}^{(s-1)}, \mathcal{F}_N\right)$ using a Metropolis - Hastings algorithm for each of them:

* Draw a candidate $\theta_{i,1}^*$ from $N\left(\theta_{01}^{(s-1)}, \sigma_{\theta_1}^2\right)$

* Compute the acceptance probability:

$$a\left(\theta_{i,1}^* \mid \theta_{i,1}^{(s-1)}\right) = \min\left(1, \frac{f\left(\theta_{i,1}^*\right) p\left(\theta_{i,1}^{(s-1)} \mid \theta_{i,1}^*\right)}{f\left(\theta_{i,1}^{(s-1)}\right) p\left(\theta_{i,1}^* \mid \theta_{i,1}^{(s-1)}\right)}\right) \text{ where}$$

$f\left(\theta_{i,1}\right)$ is the density of the data associated with $\theta_{i,1}$, i.e.

$$f\left(\theta_{i,1}\right) =$$

$$\prod_{t=t_0}^{t_T} N\left(\log\left(y_{i,t}\right) \mid \alpha_i^{(s)} + \log\left(\sum_{r=max(t_{-Q},t_0)}^{t} \frac{e^{\theta_{1;i} r + \theta_{2;i}^{(s-1)} r^2}}{\sum_{r=max(t_{-Q},t_0)}^{t} e^{\theta_{1;i} r + \theta_{2;i}^{(s-1)} r^2}} c_{i,r}\right), \sigma_{Tox}^{(s-1),2}\right)$$

and $p\left(\theta_{i,1,1} \mid \theta_{i,1,2}\right)$ is the density associated with $\theta_{i,1,1}$, given $\theta_{i,1,2}$, i.e. $p\left(\theta_{i,1,1} \mid \theta_{i,1,2}\right) = N\left(\theta_{i,1,1} \mid \theta_{i,1,2}, \sigma_{\theta_1}^2\right)$.

* With probability $a\left(\theta_{i,1}^{*} \mid \theta_{i,1}^{(s-1)}\right)$ set $\theta_{i,1}^{(s)} = \theta_{i,1}^{*}$, otherwise set $\theta_{i,1}^{(s)} = \theta_{i,1}^{(s-1)}$

* Draw a candidate $\theta_{i,2}^{*}$ from $N\left(\theta_{02}^{(s-1)}, \sigma_{\theta_2}^2\right)$

* Compute the acceptance probability:

$a\left(\theta_{i,2}^{*} \mid \theta_{i,2}^{(s-1)}\right) = \min\left(1, \frac{f\left(\theta_{i,2}^{*}\right) p\left(\theta_{i,2}^{(s-1)} \mid \theta_{i,2}^{*}\right)}{f\left(\theta_{i,2}^{(s-1)}\right) p\left(\theta_{i,2}^{*} \mid \theta_{i,2}^{(s-1)}\right)}\right)$ where $f\left(\theta_{i,2}\right)$ is the density of the data associated with $\theta_{i,2}$, i.e.

$$f\left(\theta_{i,2}\right) =$$

$$\prod_{t=t_0}^{t_T} N\left(\log\left(y_{i,t}\right) \mid \alpha_i^{(s)} + \log\left(\sum_{r=max(t_{-Q},t_0)}^{t} \frac{e^{\theta_{1;i}^{(s)} r + \theta_{2;i} r^2}}{\sum_{r=max(t_{-Q},t_0)}^{t} e^{\theta_{1;i}^{(s)} r + \theta_{2;i} r^2}} c_{i,r}\right), \sigma_{Tox}^{(s-1),2}\right)$$

and $p\left(\theta_{i2,1} \mid \theta_{i,2,2}\right)$ is the density associated with $\theta_{i2,1}$, given $\theta_{i,2,2}$, i.e. $p\left(\theta_{i2,1} \mid \theta_{i,2,2}\right) = N\left(\theta_{i2,1} \mid \theta_{i,2,2}, \sigma_{\theta_2}^2\right)$.

* With probability $a\left(\theta_{i,2}^{*} \mid \theta_{i,2}^{(s-1)}\right)$ set $\theta_{i,2}^{(s)} = \theta_{i,2}^{*}$, otherwise set $\theta_{i,2}^{(s)} = \theta_{i,2}^{(s-1)}$

Compute $\varphi_{r;i}^{(s)} = \frac{e^{\theta_{1;i}^{(s)} r + \theta_{2;i}^{(s)} r^2}}{\sum_{r=max(t_{-Q},t_0)}^{t} e^{\theta_{1;i}^{(s)} r + \theta_{2;i}^{(s)} r^2}}$

- Sample $\log\left(V_0^{(s)}\right) \mid ... \sim p\left(\log\left(V_0\right) \mid \log\left(V_i^{(s)}\right), i = 1:n, \beta_1^{(s-1)}, \mathcal{F}_N\right) = N\left(\log\left(V_{0n}\right), \sigma_{V_{0n}}^2\right)$

  where

$$\log\left(V_{0n}\right) = \frac{c_V^2 \sum_{i=1}^{N}\left(\log\left(V_i^{(s)}\right) - M_{i,1}\beta_1^{(s-1)}\right) + m_V\sigma_V^2}{Nc_V^2 + \sigma_V^2}$$

$$\sigma_{V_{0n}}^2 = \frac{c_V^2\sigma_V^2}{c_V^2 N + \sigma_V^2}$$

- Sample $k_0^{(s)} \mid ... \sim p\left(k_0 \mid k_i^{(s)}, i = 1:n, \beta_2^{(s-1)}, \mathcal{F}_N\right) = N\left(k_{0n}, \sigma_{k_{0n}}^2\right)$

  where

$$k_{0n} = \frac{c_k^2 \sum_{i=1}^{N}\left(k_i^{(s)} - M_{i,2}\beta_2^{(s-1)}\right) + m_k\sigma_k^2}{Nc_k^2 + \sigma_k^2}$$

$$\sigma_{k_{0n}}^2 = \frac{c_k^2\sigma_k^2}{c_k^2 N + \sigma_k^2}$$

- Sample $\alpha_0^{(s)} \mid ... \sim p\left(\alpha_0 \mid \log\left(\alpha_i^{(s)}\right), i = 1:n, \kappa^{(s-1)}, \mathcal{F}_N\right) = N\left(\log\left(\alpha_{0n}\right), \sigma_{\alpha_{0n}}^2\right)$

  where

$$\alpha_{0n} = \frac{c_\alpha^2 \sum_{i=1}^{N}\left(\alpha_i^{(s)} - \kappa^{(s-1)}P_i\right) + m_\alpha\sigma_\alpha^2}{Nc_\alpha^2 + \sigma_\alpha^2}$$

$$\sigma_{\alpha_{0n}}^2 = \frac{c_\alpha^2\sigma_\alpha^2}{c_\alpha^2 N + \sigma_\alpha^2}$$

- Sample $\kappa^{(s)} \mid ... \sim p\left(\kappa \mid \alpha_i^{(s)}, \alpha_0^{(s)}, i = 1:n, \mathcal{F}_N\right) = N\left(\kappa_n, \sigma_{\kappa_n}^2\right)$

  where

$$\kappa_n = \frac{c_\kappa^2 \sum_{i=1}^{N} \left(P_i \left(\alpha_i^{(s)} - \alpha_0^{(s)} \right) \right) + m_\kappa \sigma_\kappa^2}{c_\kappa^2 \left(\sum_{i=1}^{N} P_i \right)^2 + \sigma_\alpha^2}$$

$$\sigma_{\kappa_{0n}}^2 = \frac{c_\kappa^2 \sigma_\kappa^2}{c_\kappa^2 \left(\sum_{i=1}^{N} P_i \right)^2 + \sigma_\alpha^2}$$

- Sample $\beta_1^{(s)} \mid ... \sim p\left(\beta_1 \mid \log\left(V_i^{(s)}\right), \log\left(V_0^{(s)}\right), i = 1:n, \mathcal{F}_N\right) = N\left(\beta_{1n}, \sigma_{\beta_{1n}}^2\right)$

  where

$$\beta_{1n} = \frac{c_{\beta_1}^2 \sum_{i=1}^{N} \left(M_{i,1} \left(\log\left(V_i^{(s)}\right) - \log\left(V_0^{(s)}\right) \right) \right) + m_{\beta_1} \sigma_V^2}{c_{\beta_1}^2 \left(\sum_{i=1}^{N} M_{i,1} \right)^2 + \sigma_V^2}$$

$$\sigma_{\beta_{1n}}^2 = \frac{c_{\beta_1}^2 \sigma_V^2}{c_{\beta_1}^2 \left(\sum_{i=1}^{N} M_{i,1} \right)^2 + \sigma_V^2}$$

- Sample $\beta_2^{(s)} \mid ... \sim p\left(\beta_2 \mid k_i^{(s)}, k_0^{(s)}, i = 1:n, \mathcal{F}_N\right) = N\left(\beta_{2n}, \sigma_{\beta_{2n}}^2\right)$

  where

$$\beta_{2n} = \frac{c_{\beta_2}^2 \sum_{i=1}^{N} \left(M_{i,2} \left(k_i^{(s)} - k_0^{(s)} \right) \right) + m_{\beta_2} \sigma_k^2}{c_{\beta_2}^2 \left(\sum_{i=1}^{N} M_{i,2} \right)^2 + \sigma_k^2}$$

$$\sigma^2_{\beta_{2n}} = \frac{c^2_{\beta_2}\sigma^2_k}{c^2_{\beta_2}\left(\sum_{i=1}^{N} M_{i,2}\right)^2 + \sigma^2_k}$$

- Sample $\theta^{(s)}_{01} \mid ... \sim p\left(\theta_{01} \mid \theta^{(s)}_{1;i}, i = 1:n, \mathcal{F}_N\right) = N\left(\theta_{1n}, \sigma^2_{\theta_{1n}}\right)$

  where

$$\theta_{1n} = \frac{c^2_{\theta_1}\sum_{i=1}^{N}\theta^{(s)}_{1;i} + m_{\theta_1}\sigma^2_{\theta_1}}{Nc^2_{\theta_1} + \sigma^2_{\theta_1}}$$

$$\sigma^2_{\theta_{1n}} = \frac{c^2_{\theta_1}\sigma^2_{\theta_1}}{c^2_{\theta_1}N + \sigma^2_{\theta_1}}$$

- Sample $\theta^{(s)}_{02} \mid ... \sim p\left(\theta_{02} \mid \theta^{(s)}_{2;i}, i = 1:n, \mathcal{F}_N\right) = N\left(\theta_{2n}, \sigma^2_{\theta_{2n}}\right)$

  where

$$\theta_{2n} = \frac{c^2_{\theta_2}\sum_{i=1}^{N}\theta^{(s)}_{2;i} + m_{\theta_2}\sigma^2_{\theta_2}}{Nc^2_{\theta_2} + \sigma^2_{\theta_2}}$$

$$\sigma^2_{\theta_{2n}} = \frac{c^2_{\theta_2}\sigma^2_{\theta_2}}{c^2_{\theta_2}N + \sigma^2_{\theta_2}}$$

- Sample $\sigma^{2,(s)}_C \mid ... \sim$

$$IG\left(a_C + \sum_{i=1}^{N}\sum_{t=0}^{T}\left(\log\left(c_{i,t}\right) - \log\left(\sum_{j=1}^{D} d^*_{i,j}e^{-k^{(s)}(t-\tau_j)}1_{t\geq\tau_j}\right) + \log\left(V^{(s)}\right)\right)^2, b_C + N\,T\right)$$

- Sample $\sigma^{2,(s)}_{Tox} \mid ... \sim$

$$IG\left(a_Y + \sum_{i=1}^{N}\sum_{t=0}^{T}\left(\log\left(y_{i,t}\right) - \alpha^{(s)}_i - \log\left(\sum_{r=max(t_{-Q},t_0)}^{t}\varphi^{(s)}_{r;i}C_{i,t}\right)\right)^2, b_Y + N\,T\right)$$

Call $\Theta^{(s)}$ the vector collecting all the parameters estimated in the iteration s. The vectors $\Theta^{(0)}$,$\Theta^{(1)}$, ..,$\Theta^{(S)}$ can be used for deriving the predictive distribution by sampling $\log\left(C^{(s)}\right)$, $s = 1, .., S$ from

$$N\left(\log\left(C^{(s)}\right) \mid \log\left(\sum_{j=1}^{D} d_{i,j}^{*} e^{-k_i^{(s)}(t-\tau_j)} 1_{t\geq\tau_j}\right) - \log\left(V_i^{(s)}\right), \sigma_C^2\right)$$

Call $\psi_{N+1} = (\alpha_{N+1}, \theta_{1N+1}, \theta_{2N+1})$. The predictive distribution for the log-toxicity, for $s = t_0, .., t_T$, is:

$$p\left(\log\left(\boldsymbol{Y}_{N+1,s}\right) \mid \mathcal{F}_N, d_{N+1}\right) =$$

$$\int p\left(\log\left(\boldsymbol{Y}_{N+1,s}\right) \mid \psi_{N+1}, \log\left(\boldsymbol{C}_{N+1,s}\right), d_{N+1}\right) p\left(\psi_{N+1}, \log\left(\boldsymbol{C}_{N+1,s}\right) \mid \mathcal{F}_N\right) d\left(\psi_j, \log\left(\boldsymbol{C}_{N+1,s}\right)\right)$$

Call $\rho_{N+1,s}$, $\sigma^2_{N+1,s}$ the mean and the variance of this predictive distribution respectively.

6.7.4 Minimizing the aggregate posterior expected loss

As before, we want to find the dose d for the next patient N+1 by solving the problem defined by expressions (6.18)-(6.19). The problem reduces to find the dose d so that the aggregate posterior expected loss is minimized and the overall Bayesian predictive safety constraint satisfied:

$$\min_d \sum_{s=t_0}^{t_T} \delta_t R_\eta\left(\boldsymbol{Y}_{N+1,s}\left(d\right), d\right) = \sum_{s=0}^{T} \delta_t \left(\sigma_s^2 + \rho_{N+1,s}^2 - 2\eta\rho_{N+1,s} + \eta^2\right)$$

$$\rho_{j,s} + \sqrt{\sigma_s^2} z_{\tilde{\gamma}} - \eta \leq 0 \;\text{ for }\; s = t_0, .., t_T$$

where $z_{\tilde{\gamma}}$ is the $\tilde{\gamma}th$ quantile of the predictive distribution for the patient $N+1$.

All of these involved quantities can be computed using the samples from the posterior distributions obtained using the MCMC method. The overall Bayesian predictive safety constraint is still expressed using the

Boole's inequality. The constrained optimum is then computed as in Subsection 6.5.2.

6.8 One-compartmental random coefficients model: nonparametric distribution approach

Up to here, we have assumed that the population distribution, let us say F, for the random coefficients is known and in particular given by independent Gaussian distributions. As already widely discussed throughout the whole thesis, this parametric choice for the population distribution of the random coefficients can have important implications. First, a strong shrinkage of all the patient-specific random coefficients towards their common means, due to thin tails. Second, a symmetric allocation of the coefficients around their means. Finally, no rooms would be left for outliers, skewness, multimodality, irregular behaviour et cetera. In some cases, the Gaussian distribution can be a good choice - e.g. when the main covariates that explain the heterogeneity are available. But often no a priori information is available on the population distribution. In this section, we discuss the nonparametric extension of the previous model, i.e. we choose a Bayesian nonparametric prior on the mixing distribution of the heterogeneity by assuming that the distribution F is unknown and assigning it a Dirichlet Process prior. We focus on a semiparametric Bayesian model, meaning that we replace traditional Gaussian random coefficients distributions with nonparametric Bayesian models, having infinite-many parameters, whereas other finite-dimensional parameters, e.g. observational variances, are still estimated.

The Dirichlet process is the most used stochastic process in Bayesian nonparametric models and it has been discussed in **Chapter 1**. The nonparametric extension based on the DP prior can be easily imple-

mented within the Metropolis-Hasting-within-Gibbs-Sampler algorithm described before using the Polỳa urn representation (Blackwell and MacQueen, 1973). Integrating out F, the parameters $(\omega_i, \alpha_i, \theta_{1;i}, \theta_{2;i})$ follow a Pòlya urn distribution (Blackwell and MacQueen, 1973), in which the previously drawn values of $(\omega_i, \alpha_i, \theta_{1;i}, \theta_{2;i})$ have a strictly positive probability of being redrawn again, thus making the underlying probability measure P discrete with probability one. Calling $\phi_i = (\omega_i, \alpha_i, \theta_{1;i}, \theta_{2;i})$, the Pòlya urn distribution is given by the following:

$$\phi_i|\phi_{1:i-1}, F_0, a \sim \sum_k \frac{n_k}{i-1+a}\delta_{\phi_k^*} + \frac{a}{i-1+a}F_0 \tag{6.38}$$

where ϕ_k^* denotes the distinct values across the parameters $\phi_j, j = 1, .., i-1$ and n_k is the number of parameters $\phi_j, j = 1, .., i-1$ having value ϕ_k^*.

The specific choice for the nonparametric prior depends on the particular problem and on the given data. Due to time reasons, we restrict here the focus on assigning a Dirichlet Process prior jointly for all the random coefficients without including covariates, but many alternatives are available, among the others:

- If the group of random coefficients for the PK model and the one for the toxicity model are reasonably independent, they can be assumed to have two independent DP priors.

- If the group of random coefficients for the PK model and the one for the toxicity model are reasonably sequential, e.g. there is the belief that the clusters of the toxicity coefficients are within the PK clusters, then the DP prior can be substituted with other processes, e.g. the Enriched Dirichlet Process that allows for sequential and asymmetric clustering structure. See **Chapter 2** for theoretical background and **Chapter 3** and **Chapter 5** for applications of the EDP to two different problems.

- The inclusion of the covariates could requires some other nonparameteric processes, e.g. the Dependent Dirichlet Process (MacEachern, 1999).

For the sake of simplicity, we neglect the impact of the covariates on the random coefficients.

6.8.1 Model

We assume the following hierarchical model:

I. Model for the data, for $i = 1, .., N$ and $t = t_0, .., t_T$,

$$(\log(\boldsymbol{C}_{i,t}) \mid \boldsymbol{\omega}_i, X_t) = \log(d_i) + \boldsymbol{\omega}_i' X_t + \boldsymbol{\epsilon}_{i,t} \tag{6.39}$$

$$\log(\boldsymbol{Y}_{i,t}) = \boldsymbol{\alpha}_i + \log\left(\sum_{r=max(t_{-Q},t_0)}^{t} \boldsymbol{\varphi}_{r;i}\boldsymbol{C}_{i,r}\right) + \boldsymbol{\xi}_{i,t} \tag{6.40}$$

where $\boldsymbol{\epsilon}_{i,t} \sim N\left(0, \sigma_C^2\right)$, $\boldsymbol{\xi}_{i,t} \sim N\left(0, \sigma_{Tox}^2\right)$, $\boldsymbol{\epsilon}_{i,t}$ is independent from $\boldsymbol{\xi}_{jt'}$, for every t, t', i and j; and $\varphi_{r;i} = \dfrac{e^{\boldsymbol{\theta}_{1;i}r + \boldsymbol{\theta}_{2;i}r^2}}{\sum_{r=max(t_{-Q},t_0)}^{t} e^{\boldsymbol{\theta}_{1;i}r + \boldsymbol{\theta}_{2;i}r^2}}$.

II. Model for the random coefficients for $i = 1, .., N$,

$$(\boldsymbol{\omega}_i, \boldsymbol{\alpha}_i, \boldsymbol{\theta}_{1;i}, \boldsymbol{\theta}_{2;i}) \mid \boldsymbol{\omega}_0, \boldsymbol{\alpha}_0, \boldsymbol{\theta}_{01}, \boldsymbol{\theta}_{02}, \boldsymbol{P} \sim \boldsymbol{P} \tag{6.41}$$

III. Priors:

$$\boldsymbol{P} \sim DP(\boldsymbol{a}, N(m_0, \Sigma_0)) \tag{6.42}$$

$$\text{where } m_0 = (\omega_0, \alpha_0, \theta_{01}, \theta_{02})' \text{ and } \Sigma_0 = \begin{pmatrix} \sigma_V^2 & 0 & 0 & 0 & 0 \\ 0 & \sigma_k^2 & 0 & 0 & 0 \\ 0 & 0 & \sigma_\alpha^2 & 0 & 0 \\ 0 & 0 & 0 & \sigma_{\theta_1}^2 & 0 \\ 0 & 0 & 0 & 0 & \sigma_{\theta_1}^2 \end{pmatrix}$$

IV Hyper-priors:

$$\boldsymbol{a} \sim IG(a_1, b_1)\,;\ \boldsymbol{\sigma}_C^2 \sim IG(a_C, b_C)\,;\ \boldsymbol{\sigma}_Y^2 \sim IG(a_Y, b_Y)\,. \tag{6.43}$$

$$\begin{pmatrix} \log(\boldsymbol{V}_0) \\ \boldsymbol{k}_0 \\ \boldsymbol{\alpha}_0 \\ \boldsymbol{\theta}_{01} \\ \boldsymbol{\theta}_{02} \end{pmatrix} \sim N\left(\left| \begin{pmatrix} m_V \\ m_k \\ m_\alpha \\ m_{\theta_1} \\ m_{\theta_2} \end{pmatrix}, \begin{pmatrix} c_V^2 & 0 & 0 & 0 & 0 \\ 0 & c_k^2 & 0 & 0 & 0 \\ 0 & 0 & c_\alpha^2 & 0 & 0 \\ 0 & 0 & 0 & c_{\theta_1}^2 & 0 \\ 0 & 0 & 0 & 0 & c_{\theta_1}^2 \end{pmatrix} \right) \right.$$

Integrating out $\boldsymbol{P}_1$, the parameter $(\boldsymbol{\omega}_i, \boldsymbol{\alpha}_i, \boldsymbol{\theta}_{1;i}, \boldsymbol{\theta}_{2;i})$ follows a Polỳa urn distribution (Blackwell and MacQueen, 1973), in which the previously drawn values of $(\boldsymbol{\theta}_i, \boldsymbol{\alpha}_i, \boldsymbol{\theta}_{1;i}, \boldsymbol{\theta}_{2;i})$ have a strictly positive probability of being redrawn again, thus making the underlying probability measure $\boldsymbol{P}$ discrete with probability one. Calling $\boldsymbol{\phi}_i = (\boldsymbol{\omega}_i, \boldsymbol{\alpha}_i, \boldsymbol{\theta}_{1;i}, \boldsymbol{\theta}_{2;i})$, it follows that:

$$\boldsymbol{\phi}_i | \boldsymbol{\phi}_{1:i-1}, m_0, a \sim \sum_k \frac{n_k}{i-1+a} \delta_{\boldsymbol{\phi}_k^*} + \frac{a}{i-1+a} N(m_0, \Sigma_0) \tag{6.44}$$

where $\boldsymbol{\phi}_k^*$ denotes the distinct values across the parameters $\boldsymbol{\phi}_j, j = 1, .., i-1$ and n_k is the number of parameters $\boldsymbol{\phi}_j, j = 1, .., i-1$ having value $\boldsymbol{\phi}_k^*$.

6.8.2 Estimation

The above expressions (6.44) can be used as a full conditional distribution for the Metropolis-Hastings -within- Gibbs sampler algorithm.

They represent mixtures between previous draws and new draw from the Gaussian distribution obtained by multiplying the prior with the likelihood.

6.9 Simulation study: Comparing the different proposed approaches for finding the MTD for the next patient

It seems reasonable to assume that the heterogeneity can be explained by some covariates and, knowing these covariates, a parametric distribution can be used as population distribution. Unfortunately, these covariates are seldom known. In this section, we aim to illustrate the proposed models using the same simulated data set. We simulate data according to the model with parametric-distributed heterogeneity explained by covariates. Then, we compare the performance of the parametric heterogeneity model with the nonparametric heterogeneity model (without covariates explaining the heterogeneity) and we expect that the latter fits reasonably well. We also compare the results with the model without heterogeneity (homogeneous population).

The sequence of doses and the covariates for the first five patients are given as in Table 6.4. The data for the concentration and the toxicity are simulated with $\log(V_0) = \log(50)$, $k_0 = 0.25$, $\alpha_0 = 2$, $\sigma_C^2 = 0.25^2$, $\sigma_{Tox}^2 = 0.25^2$, $\sigma_V^2 = 0.25^2$, $\sigma_k^2 = 0.01^2$, $\sigma_\alpha^2 = 0.01^2$. We also assume β_1=0.25, β_2=0.35, κ=0.3, $Q = 0$ and $\delta_0 = 2$. Simulated data are shown in Figures 6.5*a* and 6.5*b*.

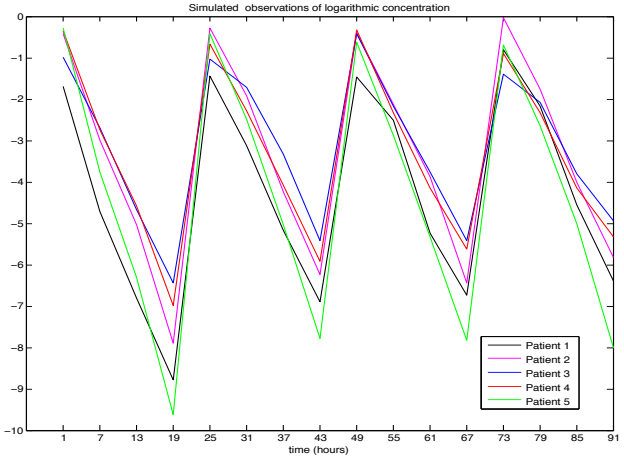
The covariates for the new patient, N+1, are $M_{1,N+1}$=0.40, $M_{2,N+1}$=0.34 and P_{N+1}=0.30.

In the model without heterogeneity, the parameters are common for all the patients. Therefore, the more heterogeneous the patients, the

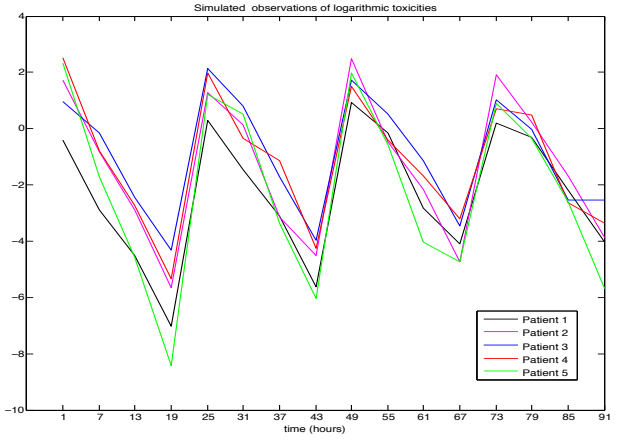
Doses	Patient=1	Patient=2	Patient=3	Patient=4	Patient=5
t_0=1	10	20	30	40	45
t_1=7	0	0	0	0	0
t_2=13	0	0	0	0	0
t_3=19	0	0	0	0	0
t_4=25	15	25	25	35	45
t_5=31	0	0	0	0	0
t_6=37	0	0	0	0	0
t_7=43	0	0	0	0	0
t_8=49	20	25	30	35	45
t_9=55	0	0	0	0	0
t_{10}=61	0	0	0	0	0
t_{11}=67	0	0	0	0	0
t_{12}=73	25	30	25	35	45
t_{13}=79	0	0	0	0	0
t_{14}=85	0	0	0	0	0
t_{15}=91	0	0	0	0	0
Covariates					
$M_{1;i}$	0.12	0.17	0.34	0.43	0.56
$M_{2;i}$	0.21	0.32	0.43	0.24	0.65
P_i	0.25	0.12	0.4	0.52	0.6912

Table 6.4: Data on $d_{i,t}$ and covariates for heterogeneity

worse the performance of the estimation procedure. The nonparametric heterogeneity model allows for different parameters, without knowing the covariates implying these heterogeneity. The boundary solutions are 3.5011 (with homogeneity), 6.1612 (with parametric heterogeneity), and 12.2084 (with nonparametric heterogeneity). The unconstrained minimum are 57.2656 (with homogeneity), 32.2969 (with parametric heterogeneity), and 58.5001 (with nonparametric heterogeneity). The better fit is associated with the model with parametric heterogeneity since the data are simulated from this model. In Figure 6.9*a*, the different aggregate posterior expected loss for the next patient are shown. Although the



(a) Log-Concentration



(b) Log-Toxicity

Figure 6.5: Log-concentration and log-Toxicity profiles (Simulated data)

aggregate posterior expected loss associated with a model with no heterogeneity and with parametric heterogeneity are very closed, their shape are different. Consequently we have the unconstrained minimum with-

out heterogeneity is much higher (almost double) than the unconstrained minimum with parametric heterogeneity. The unconstrained minimum estimated without heterogeneity and with nonparametric heterogeneity are very similar. Instead, the constrained solutions are all different. As regards to the MTDs for the next patient, all the models provide as optimum the constrained solution. The parametric heterogeneity model provides an intermediate solution between the highest solution coming from the nonparametric heterogeneity model and the lowest solution coming from model without heterogeneity.

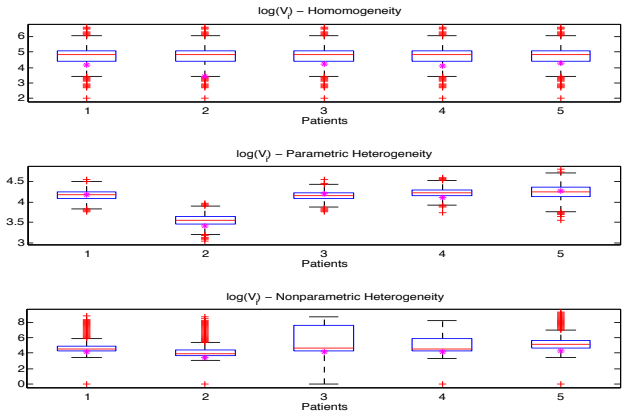


Figure 6.6: Posterior distributions for the patient-specific $log(V_i)$.

6.10 Discussion and Further Developments

In this chapter, we have worked with simulated data. Working with real data is obviously more challenging and the model should be adapted to the peculiarity of the specific problem. Several extension can be easily added by including some simple changes, e.g., in the full conditionals of the previously-discussed Metropolis-Hastings -within- Gibbs sampler

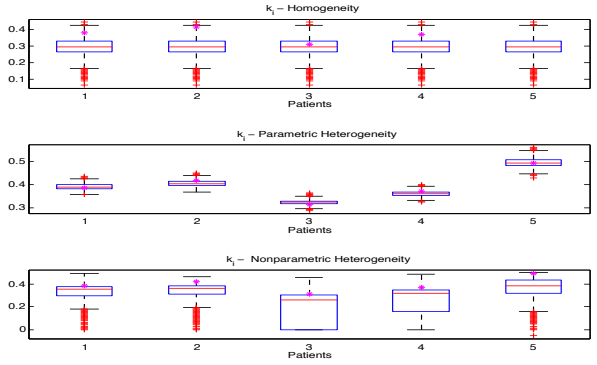


Figure 6.7: Posterior distributions for the patient-specific k_i.

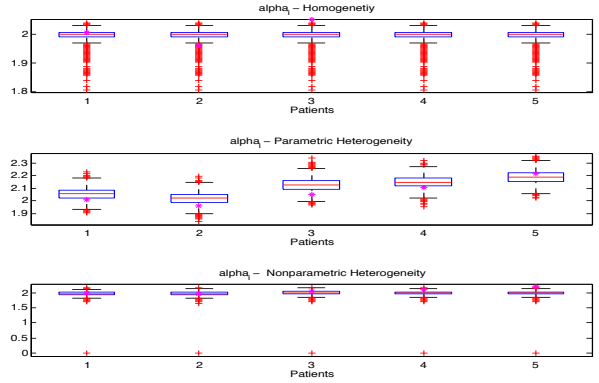


Figure 6.8: Posterior distributions for the patient-specific α_i.

algorithms. For instance, more variances can be estimated by assigning an Inverse-Gamma prior on scalar variances and an Inverse-Wishart distributions for matrix variances. More complex PK model or also more covariates can be included in the toxicity model or in the PK model. The use of more complex PK models allows the researcher to choose what is

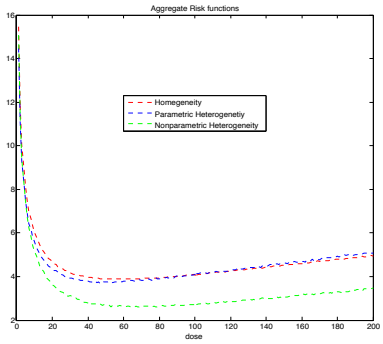


Figure 6.9: Aggregate posterior expected loss assuming homogeneity, parametric heterogeneity and nonparametric heterogeneity

the compartment that mostly matters for toxicity, which can be different from the effective site. These changes can be implemented by just making some small changes in the full conditional distributions.

More efforts are instead required for studying strictly related problems. We have proposed a general approach for including PK and toxicity information within a predictive Bayesian approach for choosing the MTD for the next patient. This procedure is just one of the elements of the general problem in the design of a clinical trial. The final aims are: first, find the MTD as the number of patients grows to infinity; second, defining an optimal stopping rule. The former requires a particle filter algorithm, which we have just started to consider for the homogeneous population. The latter would require more analytical efforts to formalize the concept that the trial should stop whenever the variation of the MTD is negligible. For time reasons, both these final aims are not properly discussed here and they need to be further investigated.

Bibliography

[1] Almon S. (1965). The distributed lag between capital appropriations and expenditures. *Econometrica*, **33**, 178-196.

[2] Berry S. M., Carlin B.P., Lee J.J., Müller P. (2011). *Bayesian Adaptive Methods for Clinical Trials*. Boca Raton, FL: Chapman & Hall/CRC Press.

[3] Babb J., Rogatko A., Zacks S. (1998). Cancer phase I clinical trials: efficient dose escalation with overdose control. *Statistics in Medicine*, **17**, 1103-1120.

[4] Chaloner K., Verdinelli I.(1995). Bayesian experimental design. *Statistical Science*, **10**, 273-304.

[5] Collins, J.M., Grieshaber, C.K., Chabner, B.A. (1990). Pharmacologically guided phase I clinical trials based upon preclinical drug development. *Journal of the National Cancer Institute*, **82**, 1321-1326.

[6] Doucet A., De Freitas N., Gordon N.J. (eds.) (2001). Sequential Monte Carlo Methods in Practice. *New York: Springer*.

[7] Drovandi C., McGree J., Pettitt A. (2013). Sequential Monte Carlo for Bayesian sequentially designed experiments for discrete data. *Computational Statistics and Data Analysis*, **57**, 320-335.

[8] Eichhorn B.H., Zacks S. (1973). Sequential search of an optimal dosage. *Journal of the American Statistical Association*, **68**, 594-598.

[9] Eichhorn B.H., Zacks S. (1981). Bayes sequential search of an optimal dosage: linear regression with both parameters unknown. *Communications in Statistics - Theory and Methods*, **10**, 931-953.

[10] Gramacy R., Polson N. (2011). Particle learning of Gaussian process models for sequential design and optimization. *Journal of Computational and Graphical Statistics*, **20**, 102-118.

[11] Hiemenz J., Cagnoni P., Simpson D., Devine S., Chao N., Keirns J., Lau W., Facklam D., Buell D. (2005), Pharmacokinetics and Maximum Tolerated Dose Study of Micafungin in Combination with Fluconazole versus Fluconazole Alone for Prophylaxis of Fungal Infections in Adult Patients Undergoing a Bone Marrow or Peripheral Stem Cell Transplant. *Antimicrobial Agents and Chemotherapy*, **49**, 1331-1336.

[12] Lagarias J., Reeds J., Wright M., Wright P. (1998). Convergence properties of the Nelder-Mead simplex method in low dimensions. *Society for Industrial and Applied Mathematics Journal of Optimization*, **9**, 112-147.

[13] Lee J., Liu D. (2008). A predictive probability design for phase II cancer clinical trials. *Clinical Trials*, **5**, 93-106.

[14] Liu J., West M. (2001). Combined parameters and state estimation in simulation- based filtering, in [6].

[15] MacEachern, S.N. (1999). Dependent nonparametric processes. *ASA Proceedings of the Section on Bayesian Statistical Science*. Alexandria, VA: American Statistical Association, 50-55.

[16] Meyerhardt J., Clark J., Supko J., Eder J., Ogino S., Stewart C., D'Amato F., Dancey J., Enzinger P., Zhu A., Ryan D., Earle C.,

Mayer R., Michelini A., Kinsella K., Fuchs C. (2007). Phase I study of geftinib, irinotecan, 5-fuorouracil and leucovorin in patients with metastatic colorectal cancer. *Cancer Chemother Pharmacol*, **60**, 661-670.

[17] Mezzetti M., Muliere P., Bulla P. (2007). An application of reinforced urn processes to determining maximum tolerated dose. *Statistics and Probability Letters*, **77**, 740-747.

[18] Muliere P., Petrone S. (1993). A Bayesian predictive approach to sequential search for an optimal dose: parametric and nonparametric models. *Journal of Italian Statistical Society*, **3**, 349-364.

[19] Muliere P., Walker S. (1997). A Bayesian nonparametric approach to determining a maximum tolerated dose. *Original Research Article, Journal of Statistical Planning and Inference*, **61**, 339-353.

[20] Müller P., Beryy D., Grieve A., Smith M., Kraps M. (2007). Simulation-based sequential Bayesian design. *Journal of Statistical Planning and Inference*, **137**, 3140-3150.

[21] Piantadosi S., Liu G.(1996). Improved design for dose escalation using pharmacokinetic measurement. *Statistics in Medicine*, **15**, 1605-1618.

[22] O' Quigley J., Pepe M., Fisher L. (1990). Continual reassessment method: a practical design for phase I clinical trials in cancer. *Biometrics*, **46**, 33-48.

[23] Robinson J.A. (1978). Sequential choice of an optimal dose: A predictive intervals approach. *Biometrika*, **65**, 75-78.

[24] Shih W.J. (1989). Prediction Approaches to Sequentially Searching for an Optimal Dose. *Biometrics*, **45**, 623-628.

[25] Tighiouart M., Rogatko A.(2010). Dose Finding with Escalation with Overdose Control (EWOC) in Cancer Clinical Trials. *Statistical Science*, **25**, 217-226.

[26] Whitehead J., Zhou Y., Mander A., Ritchie S., Sabin A., Wright A. (2006) An evaluation of Bayesian designs for dose-escalation studies in healthy volunteers. *Statistics in Medicine*, **25**, 433-445.

[27] Zacks S., Rogatko, Babb J. (1998). Optimal Bayesian-feasible dose escalation for cancer phase I trials. *Statistics and Probability Letters*, **38**, 215-220.

[28] Zhou Y., Whitehead J., Korhonen P., Mustonen M. (2008). Implementation of a Bayesian Design in a Dose-Escalation Study of an Experimental Agent in Healthy Volunteers. *Biometrics*, **64**, 299-308.