Bayesian Nonparametrics and Applications

Christos Merkatas

PhD Thesis

University of the Aegean

Department of Mathematics

Division of Statistics and Actuarial science

Supervisor: Associate Prof. Spyridon J. Hatjispyros

22 June 2018

Bayesian Nonparametrics and Applications

Christos Merkatas

PhD Thesis

Committee:

Spyridon J. Hatjispyros (Supervisor–Participant of the 3 and 7 member committee)
Associate Professor, University of the Aegean.
Dept. of Mathematics.

Stelios Georgiou (Participant of the 3 and 7 member committee)
Senior Lecturer, RMIT University.
Dept. of Mathematical Sciences.

Stella Stylianou (Participant of the 3 and 7 member committee)
Senior Lecturer, RMIT University.
Dept. of Mathematical Sciences.

Nikos I. Karachalios (Participant of the 7 member committee)
Professor, University of the Aegean.
Dept. of Mathematics.

Alex Karagrigoriou (Participant of the 7 member committee)
Professor, University of the Aegean.
Dept. of Mathematics.

Theodoros Nicoleris (Participant of the 3 and 7 member committee)
Assistant Professor, University of Athens.
Dept. of Economics.

John Tsimikas (Participant of the 7 member committee)
Associate Professor, University of University of the Aegean.
Dept. of Mathematics.

22 June 2018

I, Christos Merkatas, declare that the research presented in this thesis is my own unless otherwise stated.

This thesis has been prepared using the program LaTeX and the style MasterDoctoralThesis with some modifications in a MacTeX distribution. Writing was made using the program "TeXShop" on the OS X operating system. The programming language used for the development of the presented methods is Julia. The figures were created using the programs R and Matlab.

This thesis contains original work published or submitted for publication to international journals. The following papers are included:

1. Merkatas C., Kaloudis K. and Hatjispyros S.J. (2017). A Bayesian nonparametric approach to reconstruction and prediction of random dynamical systems. *Chaos: An Interdisciplinary Journal of Nonlinear Science*. 27(6) 063116
2. Hatjispyros S.J., Merkatas C., Nicoleris T., Walker S.G. (2017). Dependent mixtures of geometric weights priors. *Computational Statistics and Data Analysis*. 119(3), 1–18.
3. Hatjispyros S.J., Merkatas C. (2018). Reconstruction and prediction of random dynamical systems under borrowing of strength. *Submitted for publication*.

Acknowledgements

I would like to thank my supervisor Spyridon Hatjispyros for his guidance and his patience with me. Without him this thesis would be impossible. Spyros was near to me in every problem I may had in those five wonderful years spent at University of the Aegean.

I would also like to thank the other members of my advising committee Stelios Georgiou and Stella Stylianou.

I would also like to acknowledge my friends for their continued support, particularly Aliki, Artemis, Dimitris N, Dimitris P, Kostas, Michalis, Nikos, Panagiota and Thanasis.

Finally, words alone cannot express the thanks I owe to my parents, Yannis and Georgia as also to Katerina, for their unconditional love and for making everything possible for me.

Abstract

In this thesis we use a Bayesian nonparametric prior with simple weights, namely the Geometric Stick–Breaking (GSB) random probability measure to deal with the problem of reconstruction and prediction of stochastic discretized nonlinear dynamical systems.

In the first half of the thesis we propose a Bayesian nonparametric mixture model for the reconstruction and prediction from observed time series data, of discretized stochastic dynamical systems, based on Markov Chain Monte Carlo (MCMC) methods. Our approach is nonparametric in the sense that we model the noise component with a highly flexible family of density functions. While the common assumption is the normality of the noise process, here we model the noise component as an infinite mixture of Normal kernels with the mixing weights driven by a random probability measure sampled from a GSB process.

In the second half we present a new approach on the joint estimation of partially exchangeable observations by constructing pairwise dependence between a finite collection of random density functions, each of which is modeled as a mixture of GSB processes. This approach is based on a new random central masses version of the Pairwise Dependent Dirichlet Process prior mixture model. We show that modelling with Pairwise Dependent Geometric Stick–Breaking Processes (PDGSBP) is sufficient for estimation and prediction purposes.

We also propose a Bayesian nonparametric mixture model for the full reconstruction of a finite collection of dynamical equations, given observed dynamically–noisy–corrupted chaotic time series based on PDGSBP mixture priors. Under the assumption that the each set of dynamical equations has a deterministic part with known functional form and that the noise processes are independent and identically distributed from some unknown zero mean process which may have common characteristics, we jointly estimate the parameters of the dynamical systems and perform density estimation of noise components. We show that if there is at least one sufficiently large data set, using borrowing–of–strength prior specifications we are able to reconstruct those dynamical processes that are responsible for the generation of time series with small sample sizes which are inadequate for an independent reconstruction.

Our contention is that modeling with GSB random probability measures is sufficient for estimation and prediction purposes. The proposed MCMC algorithms are faster and easier to implement than their Dirichlet process based counterparts. The advantages of the use of such a simple random probability measure in Bayesian nonparametric inference in terms of sufficiency and time complexity are illustrated in both synthetic and real data sets.

Περίληψη

Στην παρούσα Διδακτορική Διατριβή προτείνονται μέθοδοι μή παραμετρικής Μπεϋζιανής στατιστικής για την εκτίμηση παραμέτρων στοχαστικών δυναμικών συστημάτων διακριτού χρόνου κάνοντας χρήση τυχαίων μέτρων πιθανότητας με γεωμετρικά βάρη–Geometric stick breaking process (GSB).

Στο Κεφάλαιο 1, γίνεται μια εισαγωγή στις βασικές έννοιες της μή παραμετρικής Bayesian στατιστικής και τις βασικές έννοιες των Στοχαστικών Δυναμικών Συστημάτων. Επιπλέον, γίνεται ανασκόπηση της βιβλιογραφίας που είναι σχετική με το πρόβλημα της ανακατασκευής δυναμικών εξισώσεων.

Στο Κεφάλαιο 2, παρουσιάζονται αναλυτικά οι πιο δημοφιλείς a–priori κατανομές της μή παραμετρικής στατιστικής κατά Bayes. Συγκεκριμένα, παρουσιάζεται το τυχαίο μέτρο Dirichlet και οι ιδιότητες του (posterior κατανομή, posterior κατανομή πρόβλεψης). Έπειτα, γίνεται ανασκόπηση των δημοφιλέστερων τρόπων αναπαράστασης του τυχαίου μέτρου Dirichlet. Συγκεκριμένα, παρουσιάζονται οι αναπαραστάσεις stick–breaking, generalized Polya urn καθώς και η αναπαράστασή του ως, κανονικοποιημένου, εντελώς τυχαίου μέτρου πιθανότητας. Στη συνέχεια, παρουσιάζεται το τυχαίο μέτρο GSB και αποδεικνύονται βασικές ιδιότητες του. Λόγω της διακριτής φύσης των παραπάνω μέτρων, για την μοντελοποίηση απολύτως συνεχών κατανομών, εισάγονται οι μίξεις τυχαίων μέτρων ως μίξεις πυρήνων παραμετρικής οικογένειας κατανομών χρησιμοποιώντας ως μέτρα μίξης διακριτά τυχαία μέτρα Dirichlet ή GSB. Έπειτα, παρουσιάζονται τα εξαρτημένα τυχαία μέτρα πιθανότητας για την μοντελοποίηση δεδομένων τα οποία παραβιάζουν τη συνθήκη της ανταλλαξιμότητας. Στο κεφάλαιο αυτό, παρουσιάζονται και τα βασικά στοιχεία της μεθοδολογίας Markov Chain Monte Carlo (MCMC), απαραίτητης για posterior συμπερασματολογία με τα μοντέλα αυτά. Συγκεκριμένα παρατίθονται οι μεθοδολογίες δειγματοληψίας κατά Gibbs και η δειγματοληψία με χρήση βοηθητικών μεταβλητών (slice sampling). Με βάση αυτές τις δύο μεθόδους, παρουσιάζονται οι MCMC αλγόριθμοι για το πρόβλημα εκτίμησης πυκνότητας χρησιμοποιώντας τυχαία μέτρα Dirichlet και τυχαία μέτρα GSB.

Στο Κεφάλαιο 3, αρχικά γίνεται ανασκόπηση ενός μή παραμετρικού Bayesian μοντέλου για την ανακατασκευή δυναμικών εξισώσεων που βασίζεται στο τυχαίο μέτρο Dirichlet. Έπειτα εισάγεται ένα μοντέλο ανακατασκευής δυναμικών εξισώσεων, από παρατηρηθείσες χαοτικές χρονοσειρές, που βασίζεται στο τυχαίο μέτρο GSB και αναπτύσσεται ένας MCMC αλγόριθμος για posterior συμπερασματολογία. Η προτεινόμενη μεθοδολογία Geometric stick breaking reconstruction–GSBR επιτυγχάνει σωστή εκτίμηση των παραμέτρων των δυναμικών εξισώσεων ακόμη και από μικρό αριθμό παρατηρήσεων, ακόμη και σε περιπτώσεις που η κατανομή του θορύβου αποκλίνει από την Κανονική. Η μέθοδος μοντελοποιεί την κατανομή του θορύβου ως μια απειροδιάστατη μίξη κανονικών πυρήνων, όπου εκ των προτέρων, ο αριθμός των συνιστωσών καθώς και οι διακυμάνσεις

των πυρήνων είναι άγνωστα. Η συμπερασματολογία γίνεται με μεθόδους MCMC όπου εκτιμάται ο αριθμός των συνιστωσών και οι αντίστοιχες διακυμάνσεις τους· δηλαδή εκτιμάται η πυκνότητα της διαδικασίας του θορύβου από τα διαθέσιμα δεδομένα. Η μεθοδολογία συγκρίνεται με τη μεθοδολογία που βασίζεται στο τυχαίο μέτρο DIrichlet χρησιμοποιώντας χαοτικές χρονοσειρές που έχουν παραχθεί από πολυωνυμικά δυναμικά συστήματα. Τέλος, προκύπτει οτι με την προτεινόμενη μεθοδολογία, το ημι–αναλλοίωτο μέτρο του Στοχαστικού Δυναμικού Συστήματος προκύπτει ως μια a–posteriori περιθώρια κατανομή πρόβλεψης, δημιουργώντας φράγμα στον ορίζοντα πρόβλεψης.

Στο Κεφάλαιο 4, παρουσιάζεται μια νέα μέθοδος για την από κοινού εκτίμηση πυκνότητας μερικώς ανταλλάξιμων παρατηρήσεων, εισάγωντας εξάρτηση μεταξύ m τυχαίων πυκνοτήτων κατά ζεύγη, που μοντελοποιούνται σαν μίξεις από τυχαία μέτρα GSB. Οι πυκνότητες θεωρείται ότι έχουν κοινά χαρακτηριστικά και ο σκοπός είναι να επιτευχθεί σωστή εκτίμηση ακόμη και για τις πυκνότητες που υπάρχει μικρός αριθμός διαθέσιμων παρατηρήσεων. Η ιδέα αυτή βασίζεται στην πλήρως στοχαστικοποιημένη γενίκευση του μοντέλου Pairwise Dependent Dirichlet Prior mixture model (PDDP). Η βασική ιδέα είναι η εξάρτηση αυτή να εισαχθεί μέσω τυχαίων μέτρων, τα βάρη των οποίων είναι αναμενώμενες τιμές των βαρών των τυχαίων μέτρων Dirichlet. Η προτεινόμενη μέθοδος, Pairwise Dependent Geometric Stick Breaking Prior mixture model (PDGSBP) συγκρίνεται με την στοχαστικοποιημένη έκδοση της PDDP μεθόδου τόσο σε προσομοιωμένα όσο και σε πραγματικά δεδομένα. Συγκεκριμένα γίνεται σύγκριση των μεθόδων σε δεδομένα που έχουν παραχθεί από μίξεις κανονικών κατανομών καθώς και από μίξεις Γάμμα κατανομών. Η εγκυρότητα των εκτιμήσεων μετράται με την Hellinger μετρική. Η μέθοδος επίσης εφαρμόζεται σε πραγματικά δεδομένα που αφορούν την εκτίμηση πυκνότητας της κατανομής των μετρήσεων του ενζύμου SGOT από τρεις ομάδες ασθενών στις οποίες ο ασθενής είτε ζεί χωρίς μεταμόσχευση, είτε έκανε μεταμόσχευση είτε απεβίωσε χωρίς μεταμόσχευση. Τα αποτελέσματα στα πειράματα αυτά δείχνουν ότι η μοντελοποίηση με PDGSBP priors είναι επαρκής για εκτίμηση πυκνότητας και πρόβλεψη. Ο προτεινόμενος αλγόριθμος MCMC για posterior συμπερασματολογία με PDGSBP priors είναι ευκολότερος στην υλοποίηση και ταχύτερος στην εκτέλεση από τον αντίστοιχο MCMC αλγόριθμο για το PDDP μοντέλο.

Στο Κεφάλαιο 5, αναπτύσσεται αλγόριθμος MCMC βασισμένος στα a–priori πολυδιάστατα μέτρα PDGSBP για το πρόβλημα της από κοινού αναδόμησης των δυναμικών εξισώσεων από παρατηρηθείσες χρονοσειρές οι οποίες περιέχουν δυναμικό θόρυβο, οι οποίες παράγονται από μη–γραμμικές εξισώσεις διαφορών πρώτης τάξης. Ιδιαίτερη έμφαση δίνεται στην περίπτωση στην οποία υπάρχει μια χρονοσειρά μικρού μεγέθους όπου είναι αδύνατη η επιτυχής αναδόμηση της δυναμικής της εξίσωσης, ενώ υπάρχει τουλάχιστον μία χρονοσειρά επαρκούς μεγέθους της οποίας η αναδόμηση της δυναμικής της εξίσωσης είναι εφικτή. Η προτεινόμενη μεθοδολογία εφαρμόζεται σε προσομοιωμένες χαοτικές χρονοσειρές που παράγονται από πολυωνυμικές απεικονίσεις που περιέχουν μη–Κανονικό θόρυβο. Υπό την υπόθεση ότι οι κατανομές των διαταραχών έχουν κοινά χαρακτηριστικά, χρησιμοποιώντας πληροφοριακές εκ των προτέρων κατανομές, είναι εφικτή η αναδόμηση των δυναμικών εξισώσεων που είναι υπεύθυνες για την παραγωγή των δειγμάτων μικρού μεγέθους με ποσοστιαία σχετικά σφάλματα μικρότερα του 1%.

Τέλος, στο Κεφάλαιο 6, γίνεται σύντομη επισκόπηση της διδακτορικής διατριβής, παρουσιάζονται

τα συμπεράσματα και προτείνονται θέματα για μελλοντική έρευνα. Συγκεκριμένα, στο πεδίο έρευνας των στοχαστικών δυναμικών συστημάτων, προτείνεται η κατασκευή ενός μοντέλου για την ανακατασκευή ενός στοχαστικού δυναμικού συστήματος χωρίς να υπάρχει κάποια υπόθεση για τη συναρτησιακή μορφή, θέτοντας ως prior στη συναρτησιακή μορφή μια Gaussian διαδικασία επεκτείνοντας έτσι το GSBR μοντέλο σε ένα πλήρως μη παραμετρικό Bayesian μοντέλο. Επιπλέον προτείνεται να μελετηθεί η μοντελοποίηση των κατανομών των θορύβων σε ένα state–space μοντέλο με GSB priors. Στην περιοχή της μή παραμετρικής Μπεϋζιανής στατιστικής προτείνεται η γενίκευση του PDGSBP μοντέλουν να συμπεριλαμβάνει όλες τις δυνατές αλληλεπιδράσεις μεταξύ των τυχαίων πυκνοτήτων. Τέλος προτείνεται η κατασκευή ενός μη παραμετρικού prior με σκοπό την επίλυση του προβλήματος ταυτοποίησης κατανομών ώστε να επιτυγχάνεται ταυτοποίηση των κοινών χαρακτηριστικών από μία συλλογή τυχαίων πυκνοτήτων.

Ακολούθως παρατίθενται η βιβλιογραφία και τρία παραρτήματα. Το Παράρτημα Α παρέχει πληροφορίες για την δειγματοληψία από τις άγνωστες κατανομές που προκύπτουν στους MCMC αλγορίθμους που παρουσιάζονται στα Κεφάλαια 3 και 5. Στο Παράρτημα Β αναλύεται η δυναμική συμπεριφορά των πολυωνυμικών απεικονίσεων που χρησιμοποιούνται στα Κεφάλαια 3 και 5. Τέλος, το Παράρτημα C παρέχει πληροφορίες για την υλοποίηση των αλγορίθμων στη γλώσσα προγραμματισμού `Julia` καθώς και ένα σύνδεσμο (URL) για τη μεταφόρτωση των προγραμμάτων.

Contents

List of Figures

List of Tables

List of Abbreviations

Term	Description
AnDe	Analysis of densities
ApEn	Approximate Entropy
AR	Auto–Regressive
BNP	Baeysian nonparametric
CRM	Completely random measure
DDP	Dependent Dirichlet Process
DP	Dirichlet Process
DPM	Dirichlet Process Mixture
DPR	Dirichlet Process Reconstruction
ForeCa	Forecastable Component Analysis
GP	Gaussian Process
GSB	Geometric stick–breaking
GSBM	Geometric stick–breaking mixture
GSBR	Geometric Stick–Breaking Reconstruction
HDP	Hierarchical Dirichlet Process
IRP	Informative Reconstruction–Prediction
KDE	Kernel Density Estimator
MAP	Maximum a–posteriori
MC	Monte Carlo
MCMC	Markov Chain Monte Carlo
MET	Mean Execution Time
NRMI	Normalized random measure with independent increments
NRP	Noninformative Reconstruction–Prediction
PAR	Polynomial Autoregressive Process
Param	Parametric Gibbs sampler assuming Gaussian noise
PARE	Percentage Absolute Relative Error
PD–GSBR	Pairwise Dependent Geometric Stick Breaking Reconstruction
PDDP	Pairwise Dependent Dirichlet Processes
PDGSBP	Pairwise Dependent Geometric Stick–Breaking Processes
PPM	Posterior Predictive Marginal
rDPR	randomized Dirichlet Process Reconstruction model
RDS	Random Dynamical System

Term	Description
rPDDP	randomized Pairwise Dependent Dirichlet Processes
SDIC	Sensitive Dependence In Initial Conditions
SM	Sampling Mean
WOSA	Weighted Overlapping Segment Averaging

To my parents Yanni and Georgia
and to Katerina.

Chapter 1

Introduction

Let $X^{(\infty)} = (X_i)_{i\geq 1}$ be an infinite sequence of observations defined on a probability space $(\Omega, \mathcal{F}, P)$ and taking values on a measurable space $(\mathbb{X}, \mathcal{X})$, with $\mathbb{X}$ a Polish space and $\mathcal{X}$ the Borel σ–algebra of subsets of $\mathbb{X}$. In addition let $\mathcal{P}_{\mathbb{X}}$, denote the space of all probability measures supported on $\mathbb{X}$.

The basic idea of Bayesian inference is that all uncertainty must be expressed in terms of probability thus any parameter of interest is modeled as a random variable having its own distribution Π which is called the *prior* distribution. The Bayesian approach to statistical analysis can be justified through the concept of *exchangeability* and *de Finetti's representation theorem*. A sequence of random variables $X^{(n)} := (X_i)_{1\leq i\leq n}$ is said to be exchangeable if

$$(X_1, \ldots, X_n) \stackrel{\mathrm{d}}{=} (X_{\sigma(1)}, \ldots, X_{\sigma(n)}), \tag{1.1}$$

for any permutation σ of $\{1, \ldots, n\}$. Consequently an infinite sequence of random variables is exchangeable if every finite collection of it is exchangeable. Exchangeabiltty is the basic modeling assumption in Bayesian inference. Intuitively the concept of exchangeability indicates that the order that we collect a sample does not affect the joint distribution.

The next theorem is the so called de Finetti's representation theorem and it states that a collection of random variables is exchangeable if and only if it is a mixture of sequences of independent and identically distributed random variables.

Theorem 1.1 (De Finetti (1937))**.** *The sequence $X^{(n)}$ is exchangeable if and only if there exists a probability measure Π on $\mathcal{P}_{\mathbb{X}}$ such that, for any $n \geq 1$ and $A = A_1 \times \cdots \times A_n \times \mathbb{X}^{\infty}$,*

$$P(X^{(n)} \in A) = \int_{\mathbb{Q}\in\mathcal{P}_{\mathbb{X}}} P(X^{(n)} \in A \,|\, \mathbb{Q})\Pi(\mathrm{d}\mathbb{Q}) = \int_{\mathbb{Q}\in\mathcal{P}_{\mathbb{X}}} \prod_{i=1}^{n} \mathbb{Q}(X_i \in A_i)\Pi(\mathrm{d}\mathbb{Q}).$$

where $A_i \in \mathcal{X}$ for $1 \leq i \leq n$ and $\mathbb{X}^{\infty} = \mathbb{X} \times \mathbb{X} \times \cdots$.

The measure Π is called the *de Finetti* measure and is uniquely determined for a given exchangeable sequence. It is the de Finetti measure that takes the rôle of a prior distribution in Bayesian inference. By virtue of de Finetti's theorem, the data generating process in a Bayesian

model is a two–stage process

$$\begin{aligned} X_i \,|\, \mathbb{Q} &\overset{\text{iid}}{\sim} \mathbb{Q} \quad i = 1, \ldots, n \\ \mathbb{Q} &\sim \Pi, \end{aligned}$$

and the objective is to determine the *posterior* distribution

$$\Pi(\mathrm{d}\mathbb{Q} \,|\, X_1, \ldots, X_n) \propto \mathcal{L}(\mathbb{Q}\,; X_1, \ldots, X_n)\Pi(\mathrm{d}\mathbb{Q}),$$

where $\mathcal{L}(\mathbb{Q}\,; X_1, \ldots, X_n)$ is the likelihood function. The posterior distribution measures the uncertainty for the random variable of interest after seeing the observations.

Whenever Π is degenerate on a subset of $\mathcal{P}_{\mathbb{X}}$ that can be indexed by a finite dimensional parameter $\theta \in \Theta \subset \mathbb{R}^k$ the Bayesian model is *parametric* and Π is a prior probability measure over the parameter space Θ. Instead when we allow inference to be made to infinite dimensional subspaces of $\mathcal{P}_{\mathbb{X}}$, the model is called *nonparametric* and Π is a *random probability measure* that acts as a prior distribution over an infinite dimensional space. Distributions over infinite dimensional spaces are stochastic processes so the term process will be thrown a lot in the following when we consider the distribution of a random probability measure.

Nevertheless, restricting inference to parametric models may limit the scope and type of inferences that can be drawn from such models. In this thesis we aim to use and construct Bayesian nonparametric models for reconstruction and prediction of random dynamical systems. Bayesian nonparametric models assume the distribution of the observations to be unknown and assign the prior on the space of probability measures $\mathcal{P}_{\mathbb{X}}$ which now is the parameter space. Clearly this space is infinite dimensional and thus the justification as nonparametric models.

We proceed in this chapter with some fundamentals of the theory of dynamical systems and explain the need for Bayesian nonparametric modeling of the density of the noise components. It would be worth noting here that our methods are generally applicable in a time series setting. While we are interested in nonlinear random dynamical systems, the models in this thesis can be applied, to a similar manner, in many popular statistical time series models such as *autoregressive processes* (AR). The general theory on Bayesian nonparametic models and computational methods for posterior inference will be discussed thoroughly in Chapter 2.

1.1 Dynamical systems

It is common in science to model a physical process that changes over time. Such a process is called a *dynamical system*. Dynamical systems can be classified in two categories namely *deterministic* and *stochastic-random* dynamical systems. In the case of a deterministic dynamical system, its controlling mechanism is completely understood, and the states of the system are described by some mathematical model, involving previous states, completely describing the

evolution of the system. In contrast, a dynamical system that involves randomness in its mechanism is called stochastic. Subsequently, each of the previous classes contains (stochastic) dynamical systems of *discrete time* which are usually described by some difference equation, or of *continuous time* which are described by some differential equation or by its solution flow. It is worth noting that one can obtain a discrete time dynamical system by discretizing flows. In the following subsections we describe dynamical systems, both deterministic and stochastic evolving in discrete time, in more depth. We do not intend to give an extensive introduction to the theory of deterministic dynamical systems but only the basic notions that provide the necessary background for the methods developed in the thesis. More details on dynamical systems can be found in Alligood et al. (1996); Broer & Takens (2010); Chan & Tong (2013) and Galor (2007).

1.1.1 Deterministic dynamical systems

A dynamical system in discrete time defined on a state–space $\mathbb{X}$ can be described by a difference equation of the form

$$x_i = g(\vartheta, x_{i-1}, x_{i-2}, \ldots, x_{i-d}), \quad i \geq 1, \tag{1.2}$$

for some initial conditions $(x_0, x_{-1} \ldots, x_{-d+1}) \in \mathbb{X}^d$, and the function $g : \mathbb{X}^d \to \mathbb{X}$ in eq. (1.2) is continuous in $x_{i-1}, \ldots, x_{i-d}$ parametrized by some vector of *control parameters* $\vartheta \in \Theta$, where Θ is the parameter space. For the sake of simplicity, in the following we will assume that the transition in a state x_i depends solely on x_{i-1}. Thus we will consider dynamical systems in the form

$$x_i = g(\vartheta, x_{i-1}), \quad i \geq 1. \tag{1.3}$$

The evolution of the system after n iterations can be observed with the form of a time series $x^{(n)}$ of length n, where each point x_i is the i–fold functional iteration of the initial condition x_0 i.e.

$$x_i = g^i(\vartheta, x_{i-1}) := \underbrace{g\left(g\left(\cdots g\left(\vartheta, x_{i-1}\right)\right)\right)}_{i\text{–times}}, \quad \text{for } 1 \leq i \leq n. \tag{1.4}$$

Let us summarize the above descriptions with a definition that will be useful for future references.

Definition 1.1. *A function whose domain space and range space are the same is called a map. Let x be a point and let g be a map. The orbit of x under g is the set of points $\mathcal{O}_g(x) = \{g^k(x) : k \geq 0\}$. The starting point x for the orbit is called the initial value or initial condition of the orbit.*

The system may exhibit *chaotic* behavior if the function g is nonlinear. This means that if we observe the system by means of a time series we will see a complex and irregular behavior which resembles the behavior of a *stochastic process*. In Figure 1.1 (upper panel) we display two orbits x, y generated from the chaotic logistic map[1]

$$x_i = 1 - \vartheta x_{i-1}^2, \quad i = 1, \ldots, 200,$$

[1]We use the representation $x = 1 - \vartheta x^2,\ \vartheta \in [0, 2]$, in order to be able to identify the coefficients of a polynomial autoregressive process.

for $\vartheta = 1.71$ and initial conditions $x_0, x_0' \in \{1, 1.001\}$. Even though the generating mechanism is purely deterministic, the two time series have complex behavior and it is evident that it is not possible to make predictions for the state of the system after a few iterations. This is clear for $i \geq 70$, where the behavior of the two systems is significantly different.

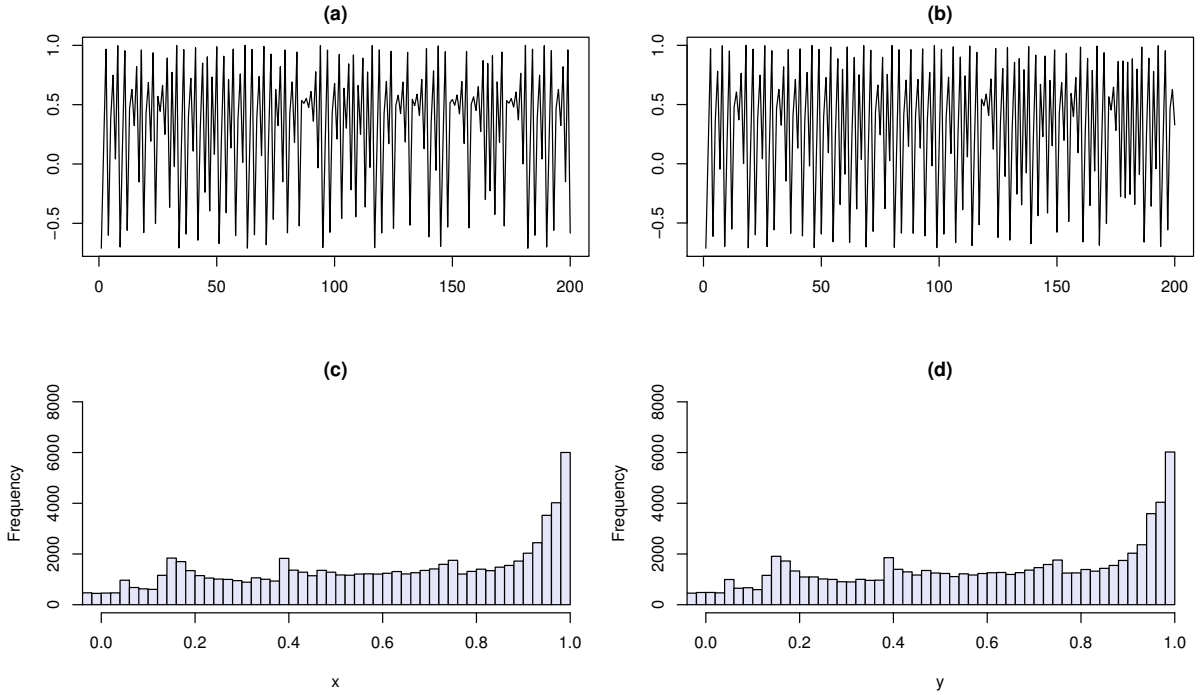

Figure 1.1: Two orbits x, y (upper panel) generated from the logistic equation for initial conditions $x_0 = 1$ and $y_0 = 1.001$. It is evident that purely deterministic mechanisms generate time series that will lose predictability soon. In the lower panel are depicted the histograms of the two orbits for $100,000$ iterations of the logistic map. Note how similar these histograms are even though the two orbits are significantly different.

A dynamical system might evolve under the presence of noise. In particular, there are two types of noise that a dynamical system may be subjected to; *measurement* or *observational* noise and *dynamical* or *system* noise which can be *additive* or *multiplicative*. Observational noise is usually present in laboratory or real world time series data where we have often inaccurate measurements of the underlying process making the true states of the system unobservable and does not affect the future evolution of the system. In contrast, dynamical noise is incorporated in such models as model error and can drastically affect the future evolution of the system. Below we provide some examples of dynamical systems contaminated by different types of noise.

Example 1.1 (Dynamical system with additive observational noise)**.** *When observational noise is present usually a state-space model is useful for the analysis of the system. If this is the case, the modeling assumption is that the true states of the system s_i are generated by the dynamical*

system g but we observe noise-contaminated quantities x_i, $1 \leq i \leq n$

$$s_i = g(\vartheta, s_{i-1}),$$
$$x_i = s_i + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2).$$

Example 1.2 (Dynamical system with additive observational and dynamical noise)**.** *An example of dynamical system that has additive observational and dynamical noise is a nonlinear Gaussian state-space model which formally is given by*

$$s_i = f(\varphi, s_{i-1}) + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_s^2)$$
$$x_i = g(\vartheta, s_i) + \epsilon_i', \quad \epsilon_i' \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_x^2).$$

Example 1.3 (Dynamical system with multiplicative dynamical and additive observational noise)**.** *A popular state-space model used in time series analysis is the unobserved ARCH (Giakoumatos et al., 2005) model which formally is given by*

$$s_i = (\alpha + \beta s_{i-1}^2)^{1/2} \epsilon_i'$$
$$x_i = s_i + \sigma \epsilon_i,$$

where $\epsilon_i, \epsilon_i' \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$. In the unobserved ARCH model we observe a realization of the process $x^{(n)}$ and s_i is the unobserved ARCH component at time i. Note that the observational noise is additive while the dynamical noise of the process $(s_i)_{i \geq 1}$ is multiplicative. For identifiability reasons, some constraints must be imposed on the parameters α and β but in this point it is irrelevant.

Dynamical systems that are contaminated with dynamical noise result to what is known as *stochastic* or *random dynamical systems* (RDS) (Arnold, 2013; Bhattacharya & Majumdar, 2007; Hatjispyros & Yannacopoulos, 2005; McGoff et al., 2015; Schenk-Hoppe, 1997; Lasota & Mackey, 1994). In this thesis we will be concerned with random dynamical systems that are subjected to additive dynamical noise. Before we proceed to the theory of random dynamical systems, we will provide some elements of chaos theory useful for understanding the material appearing in subsequent chapters.

1.1.2 Chaos in dynamical systems

We are interested in the analysis of chaotic time series so let us describe here the defining properties of a time series in order to be chaotic. We will give the properties for the deterministic case only; we remark that in all cases we study time series that originate under additive i.i.d noise, when the deterministic part g is in a chaotic state. In order to have chaotic behavior in observed time series (trajectories) the following three defining properties;*boundedness, aperiodicity* and *sensitive dependence on initial conditions* (SDIC) must exist. The definitions of these notions are given below.

Definition 1.2. *Let g be a map on $\mathbb{X}$. The set $B \subseteq \mathbb{X}$ is invariant with respect to g whenever $g(\vartheta, B) \subseteq B$.*

Definition 1.3 (Periodic trajectories)**.** *Let g be a map on $\mathbb{X}$. The point p is called periodic point of period k if $g^k(p) = p$, and k is the smallest such positive integer. In the special case where $k = 1$ then p is called a fixed point. The orbit with initial point p, consisting of k points is called periodic orbit period k.*

A fixed point p may either attract points that are near it or nearby points may spread far from the fixed point p under the dynamical system. In the first case the fixed point is called a *sink* while in the second case the fixed point is called *repelling* or *source*. The following theorem is useful to identify the stability of the fixed point i.e. if it is a sink or a source.

Theorem 1.2 (Stability of periodic orbits)**.** *Let $g : \mathbb{X} \to \mathbb{X}$ be a map and let $\{x_1, \ldots, x_k\}$ be a periodic orbit of length k. Then*

1. *If $|(g^k)'(x_1)| < 1$, the orbit is a sink (attracting k-cycle).*
2. *If $|(g^k)'(x_1)| > 1$, the orbit is a source (repelling k-cycle).*

Definition 1.4 (Asymptotical periodicity)**.** *Let g be a map on $\mathbb{X}$ and $g \in C^\infty(\mathbb{X})$ (the class of infinitely differentiable with continuous derivative functions). An orbit $\mathcal{O}_g(x_1)$ is called asymptotically periodic if it converges to a k-periodic orbit. That is, there exists a periodic orbit $\mathcal{O}_g(y_1) = \{y_1, \ldots, y_k\}$ such that*

$$\lim_{n\to\infty} |x_n - y_n| = 0.$$

Definition 1.5 (Aperiodicity)**.** *A time series (orbit) generated from a discrete time dynamical system is said to be aperiodic if it has no periodic points.*

Definition 1.6 (Sensitive dependence on initial conditions)**.** *Let g be a map on $\mathbb{R}$. A point x_0 has sensitive dependence on initial conditions if there exists $\epsilon > 0$ such that any neighborhood N of x_0 contains a point x such that $|g^k(x) - g^k(x_0)| \geq \epsilon$ for some nonnegative integer k. The point x_0 is sometimes called sensitive point.*

Intuitively, sensitive dependence on initial conditions, implies that two orbits from the same system starting from two infinitesimally small different initial conditions will diverge with exponential rate. A quantitative characteristic of a system to determine whether or not the system has sensitive dependence on initial condition is the *Lyapunov exponent*.

Suppose $x' = x + \epsilon$, then

$$|g^n(x') - g^n(x)| \approx \epsilon |(g^n)'(x)| = \epsilon \prod_{i=0}^{n-1} |g'(g^i(x))| \approx e^{nL}\epsilon$$

$$\Rightarrow L \approx \lim_{n\to\infty} \sum_{i=0}^{n-1} \log |g'(g^i(x))|.$$

Definition 1.7. *The Lyapunov exponent L is defined by*

$$L = \lim_{n\to\infty} \frac{1}{n} \left(\log |g'(x_1)| + \ldots + \log |g'(x_n)| \right), \tag{1.5}$$

where the derivative is taken with respect to x and $x_1, \ldots, x_n$ are successive iterates.

We are now ready to give the definition of a chaotic orbit based on the above definitions.

Definition 1.8 (Chaotic orbit)**.** *Let g be a map on $\mathbb{X}$ and let $\mathcal{O}_g(x_1)$ be a bounded orbit of g. The orbit is chaotic if*

1. *The orbit $\mathcal{O}_g(x_1)$ is not asymptotical periodic, and*
2. *The orbit has at least one positive Lyapunov exponent.*

Before we proceed to the description of random dynamical systems, it will be helpful for the following to present here some qualitative characteristics of dynamical systems such as *attractors, basin of atttraction* and *invariant measures*.

Attractors and the basin of attraction. In this paragraph we intend to describe some qualitative characteristics of discrete time dynamical systems described by a map $g : \mathbb{X} \to \mathbb{X}$. An *attractor* of a dynamical system is a subset of points $S \subseteq \mathbb{X}$ such that orbits starting from a variety of initial conditions $x_0 \in B \subseteq \mathbb{X}$ will fall, and remain into S. The attractor of a dynamical system can be a single point (e.g. fixed points constitute an attractor), a limit cycle, or more complicated sets with fractal structure called *strange attractors*. In contrast if the orbits starting from the initial condition are falling out of the set S then S is called a *repellor*.

The set B is called the *basin of attraction* of the system. The basin of attraction can be a single subset of $\mathbb{X}$ or a union of subsets of $\mathbb{X}$. In the former case the system has only one stable attractor while in the later, there coexist more than one strange attractors a phenomenon known as *multistability* (Kraut et al., 1999). In this thesis we deal with polynomial maps. In terms of stability we note that quadratic polynomial maps can exhibit for each parameter value one stable attractor at most. Multistability and coexistence of more than one strange attractors can be achieved under higher degree polynomials. An example of a dynamical system with stable attractor is the logistic map defined via

$$x_i = 1 - 1.71x_{i-1}^2, \quad i \geq 1, \tag{1.6}$$

while a bistable dynamical system is the cubic map

$$x_i = 0.05 + 2.53x_{i-1} - 0.99x_{i-1}^3, \quad i \geq 1 \tag{1.7}$$

for which the dynamical behavior will be extensively described in Chapter 3.

Chaotic attractor. We have given above a brief description of an attractor. It is the set of points that will be visited from the map. Since we are interested in chaotic orbits, in this paragraph we provide a formal definition of a *chaotic attractor*. Intuitevely, a chaotic attractor is the set of points that will be visited arbitrarily close and infinitely often by a chaotic orbit. Formally, a chaotic attractor is defined via the *forward limit set*.

Definition 1.9. *Let g be a map and let x_0 be an initial condition. The forward limit set of the orbit $\{g^n(x_0)\}$ is the set*

$$\omega(x_0) = \{x : \textit{ for all } N \textit{ and } \epsilon \textit{ there exists } n > N \textit{ such that } |g^n(x_0) - x| < \epsilon\}.$$

Definition 1.10 (Chaotic attractor)**.** *Let $\{g^k(x)\}$ be a chaotic orbit. If $x_0 \in \omega(x_0)$ then $\omega(x_0)$ is called a chaotic set. An attractor is a forward limit set which attracts a set of initial values that has nonzero measure and is called basin of attraction. A chaotic attractor is a chaotic set that is also an attractor.*

Invariant measures. In the lower panel of Figure 1.1, the histograms of the time series generated from the logistic map for the corresponding initial conditions x_0 based on $100,000$ iterations of the logistic map are displayed. This histogram represents the frequency with which a region of the state space $\mathbb{X}$ is visited. Even the two time series are completely different, it is obvious that the general form of the histograms is quite similar, implicating that all orbits generated from the logistic map have the same marginal distribution. In fact, the frequency that an orbit visits a specific value can be measured by means of a probability measure which is called the *invariant* measure. More formaly, the invariant measure generated from the map g is defined by

$$\mu_g(S) = \lim_{r\to 0} F(x_0, N(r,S)), \text{ where } F(x_0,S) = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathcal{I}(g^i(\vartheta, x_0) \in S),$$

and $N(r,S) = \{x : \mathrm{dist}(x,S) \leq r\}$. The formation of the invariant measure depicts the fact that long term predictions in a deterministic dynamical system exhibiting chaos is not possible. Having at our disposal the invariant measure $\mu_g(\mathrm{d}x)$ of the chaotic map $x_i = g(x_{i-1})$, it is possible to make probabilistic prediction arguments for the long term behavior of the system in the sense that $P(x_i \in A) = \mu_g(A)$ for arbitrary large i and for all measurable subsets A of $\mathbb{X}$.

Equivalently one can describe the marginal distribution with a density but this is not always possible especially for multidimensional dynamical systems. The evolution of the density of a system forward in time is given from the *Frobenius-Perron* operator but this goes beyond the scope of this thesis.

Bifurcation diagram. The dynamical behavior of a dynamical system can be described via a *bifurcation diagram* which shows the birth, evolution, and death of attracting sets. Equivalently, it shows the limiting behavior of orbits for different values of the control parameters ϑ. In Figure 1.2 we display the bifurcation diagram of the logistic map $x_i = 1 - \vartheta x_{i-1}$. To construct such a diagram, we choose an initial point $x_0 \in \mathbb{X}$ and an initial value for ϑ, calculate the orbit of x_0 under the map of interest for a large number of iterations M and discard the first (lets say) 100 iterates and plot the orbit of length $M - 100$. Then we increase ϑ and do the same procedure for a big range of values of ϑ.

For small values of ϑ ranging from 0 to 0.75 all orbits are attracted to a single point indicated by the x–axis. A period–two orbits arises at the *bifurcation point* $\vartheta = 0.75$, which in turn leads to period–four orbits and then more complicated orbits for larger values of ϑ. When the period–two orbit appears, the fixed point is no longer plotted because it does not attract orbits. This behavior is called *period-doubling route to chaos*.

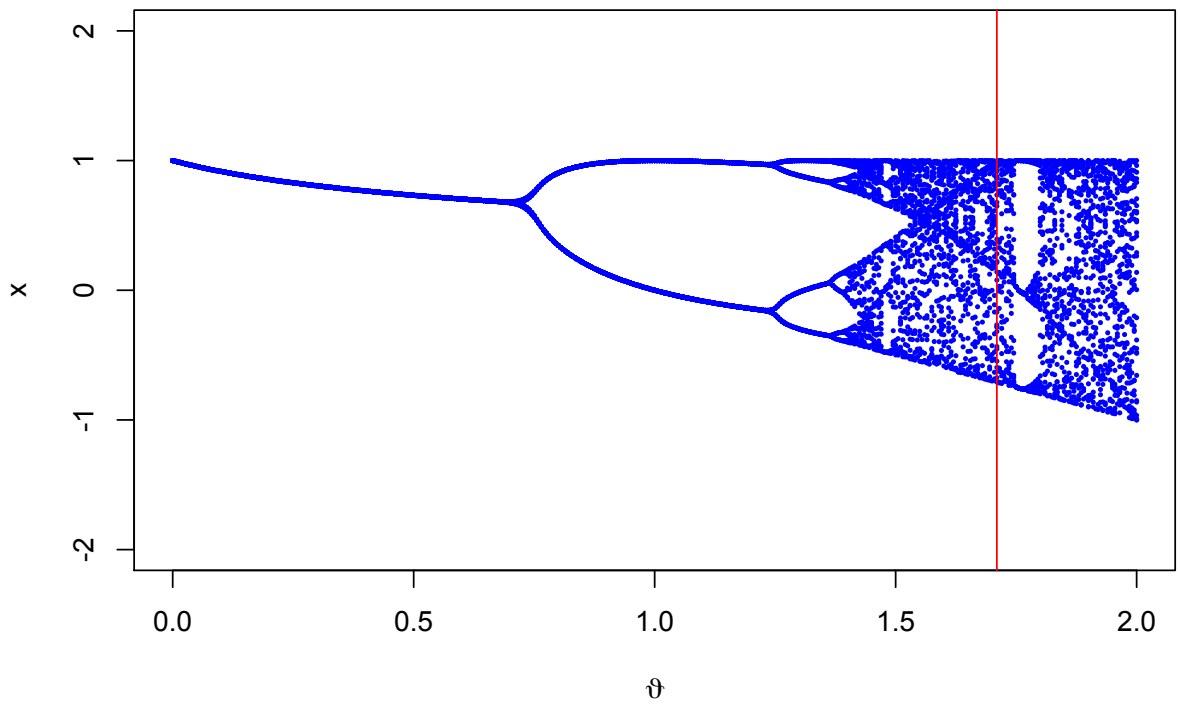

Figure 1.2: Bifurcation diagram of the logistic map. The red line indicates the control parameter $\vartheta = 1.71$ which we have used in our examples by now. For this value of ϑ the logistic map exhibits chaotic behavior.

1.1.3 Random dynamical systems

In section 1.1.1 we saw that a dynamical system generates a sequence $x^{(n)}$ of observations according to the map g for fixed values of the control parameters ϑ. However we mentioned that a dynamical system may be subjected to some source of noise. In this section we will consider the general case where the map g is perturbed by an ergodic process i.e.

$$x_i = g(\vartheta, x_{i-1}, \ldots, x_{i-d}) + z_i, \quad i = 1, \ldots, n \tag{1.8}$$

where z_i is an ergodic process. In this case the move from a state x_{i-1}, to its successor x_i is stochastic and the orbit $x^{(n)}$ depends on the initial values $x_0, x_{-1}, \ldots, x_{-d+1}$, as also the particular realization of the stochastic process $(z_i)_{i\geq 1}$. Thus noise in the system may enter in two ways. Either it disturbes the parameters ϑ, or the deterministic part g by some additive noise z_i. For simplicity we will assume that the transition to x_i depends only on the previous state x_{i-1}. Thus we consider the case

$$x_i = g(\vartheta, x_{i-1}) + z_i, \quad i \geq 1. \tag{1.9}$$

Before we proceed with the definition of random dynamical system let us introduce some definitions (Klenke, 2013) from *ergodic* theory. Intuitively, ergodic theory studies the long term behavior of a dynamical system. In the following $(\Omega, \mathcal{F}, P)$ will be our probability space and $\tau : \Omega \to \Omega$ will denote a measurable map.

Definition 1.11 (Invariant events)**.** *An event $A \in \mathcal{F}$ is called invariant if $\tau^{-1}(A) = A$. The σ-algebra of invariant events*

$$\mathcal{I}_{\mathcal{F}} = \{A \in \mathcal{F} : \tau^{-1}(A) = A\},$$

is called trivial if for all $A \in \mathcal{I}_{\mathcal{F}}$ it is that $P(A) \in \{0, 1\}$.

Definition 1.12 (Measure–preserving and ergodic dynamical systems)**.** *The map τ is measure-preserving if*

$$P[\tau^{-1}(A)] = P(A), \quad \textit{for all } A \in \mathcal{F}.$$

In this case the quadraple $(\Omega, \mathcal{F}, P, \tau)$ is called a measure-preserving dynamical system. If τ is measure-preserving and $\mathcal{I}_{\mathcal{F}}$ is P-trivial then $(\Omega, \mathcal{F}, P, \tau)$ is called ergodic.

There is a connection between measure preserving dynamical systems and stationary stochastic processes. If $(X_n)_{n\in\mathbb{N}}$ is a stochastic process on the probability space $(\mathbb{X}^{\mathbb{N}}, \mathcal{X}^{\otimes\mathbb{N}}, P)$ and τ is defined to be the shift operator, that is

$$\tau : \Omega \to \Omega \quad (\omega_n)_{n\in\mathbb{N}} \to (\omega_{n+1})_{n\in\mathbb{N}},$$

then the process $X_n(\omega) = X_0(\tau^n(\omega))$ is stationay if and only if $(\Omega, \mathcal{F}, P, \tau)$ is a measure preserving dynamical system. The stochastic process defined as above is ergodic if $(\Omega, \mathcal{F}, P, \tau)$ is ergodic.

For an ergodic transformation τ we have the following:

Theorem 1.3 (Birkhoff (1931) ergodic theorem)**.** *Let $X_0 \in L^1(P)$ i.e. $\int_\Omega |X_0(\omega)| \mathrm{d}P(\omega) < +\infty$. If τ is ergodic then*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} X_k = \mathbb{E}[X_0], \quad P - a.s.$$

Random dynamical systems. Having given the above definitions we are ready to define a random dynamical system. We can reformulate the system defined in eq. (1.9) as

$$x_i = g(\vartheta, x_{i-1}) + z(\tau^i\omega), \quad i \geq 1, \tag{1.10}$$

where $z(\tau^i\omega)$ is an ergodic process. A *random dynamical system* is defined by a measurable mapping φ as

$$\varphi(\omega, x) := g(\vartheta, x_{i-1}) + z(\omega).$$

Then $\varphi(n, \omega, x) := \varphi(\tau^{n-1}\omega) \circ \cdots \circ \varphi(\omega)x = x_n(\omega)$ with $x_0 = x$, and for $n \geq 0$, i.e. $\varphi(n, \omega, x)$ is the n–th iterate of the map g perturbed by the noise $z(\tau^i\omega))_{i\geq 0}$. It follows from the general theory that eq. (1.10) defines a random dynamical system with state space $\mathbb{X}$. More formally we have the following definition:

Definition 1.13. *A random dynamical system is a measurable mapping $\varphi : \mathbb{N} \times \Omega \times \mathbb{X} \to \mathbb{X}$, $(n, \omega, x) \to \varphi(n, \omega, x)$ such that for all $\omega \in \Omega$*

1. *$\varphi(0, \omega) = \mathrm{id}_{\mathbb{X}}$ and $\varphi(m + n, \omega) = \varphi(n, \tau^m\omega) \circ \varphi(m, \omega)$ for all $n, m \geq 0$.*
2. *$\varphi(n, \omega) : \mathbb{X} \to \mathbb{X}$ is smooth.*

3. *For i.i.d. noise,* $(\varphi(n,\cdot,x))_{n\in\mathbb{N}}$ *is a Markov process for all* $x \in \mathbb{X}$. *Then* φ *is called Markovian.*

In what follows, we will consider only the case of additive dynamical noise in which the random dynamical system can be described via

$$x_i = g(\vartheta, x_{i-1}) + z_i, \quad i \geq 1, \tag{1.11}$$

where $g : \Theta \times \mathbb{X} \to \mathbb{X}$, with $(x_i)_{i\geq 0}$ and $(z_i)_{i\geq 1}$ are real random variables defined over $(\Omega, \mathcal{F}, P)$. That is, for our purposes $z(\tau^i\omega) = z_i$, and we assume that the additive perturbations z_i are identically distributed from a zero mean distribution with unknown density f defined over the real line. Now clearly the dynamical system consists of a deterministic and a random part; that is the functional form described by the map g and the noise components respectively.

Recall that in the case of a deterministic dynamical system, the dynamics are described by the functional iterations of the map applied to the initial condition. In the case of RDS since every state is a random variable, the dynamics are described by the transition probability kernel $Q(\,\cdot\,,\,\cdot\,)$ of the homogeneous Markov chain defined by eq. (1.11). Formally it is that

$$Q(x, A) = P(x_n \in A \,|\, x_{n-1} = x).$$

The system in this case is observed via time series data which we will assume is not contaminated with observation noise so we have in our disposal a time series $x^{(n)}$ generated directly by the Markovian stochastic process given in eq. (1.11).

Quasi–invariant measures. In analogy with the deterministic systems, in a random dynamical system, there exists an associated *quasi–invariant measure* $\mu_{g,z}(\mathrm{d}x)$ which is the signature of the underlying interplay of the chaotic dynamics and dynamical noise perturbations. The quasi–invariant measure is a smoothed–out deformation of the associated invariant measure $\mu_g(\mathrm{d}x)$ of the deterministic part of the system. In analogy we can make long term probabilistic prediction arguments for random chaotic dynamical systems in the sense that now $P(x_i \in A) = \mu_{g,z}(A)$ for an arbitrary large i and for all measurable subsets A of $\mathbb{X}$. In fact the deterministic invariant measure is the limit of the invariant measures of a random dynamical system as in (1.11) with infinitesimal random disturbances i.e. $z_i \to 0$.

It is worth noting that the estimation of the quasi–invariant measure is difficult and not a straight–forward procedure. In Chapter 3 we will show that the associated quasi–invariant measure of a random dynamical system naturally arises as posterior predictive marginal (PPM) of the out–of–sample variables forming a prediction barrier. More detailes and an extensive treatment of quasi invariant measures can be found in Collet et al. (2012).

In Figure 1.3 we plot the time series generated from the random logistic map given by

$$x_i = 1 - 1.71x_{i-1}^2 + z_i, \quad z_i \sim \mathcal{N}(0, \sigma^2) \quad \text{for } i = 1, 2, \ldots \tag{1.12}$$

for twe initial conditions $x_0 \in \{1, 1.001\}$ and $\sigma = 0.01$. The marginal distributions of the two orbits are depicted in the lower panel in histogram representation.

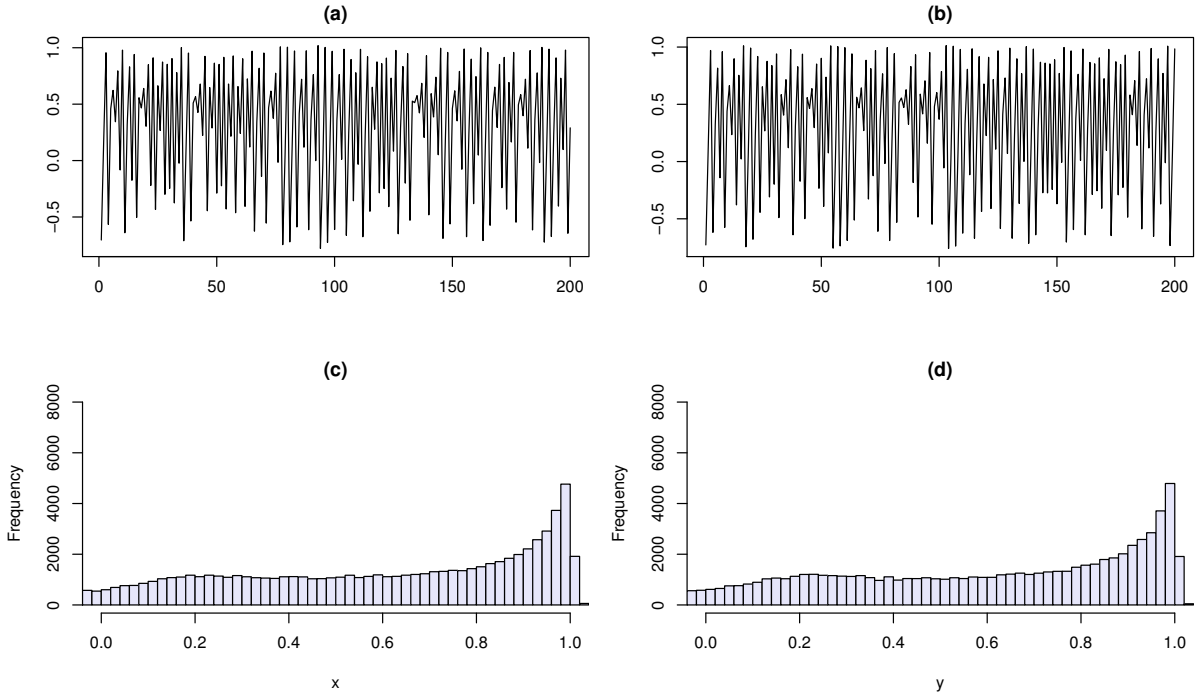

Figure 1.3: Two orbits x, y generated from the random logistic equation for initial conditions $x_0 = 1$ and $y_0 = 1.001$ for the value $\sigma = 0.01$. presented in the upper panel. The lower panel depicts the histogram of the quasi–invariant measure based on $100,000$ iterations. Note that the differences are indistinguishable. Comparing the lower panel with the lower panel of Figure 1.1, we see that the quasi–invariant distribution is a smoothed–out deformation of the invariant distribution given in Figure 1.1(c) and (d).

1.2 Reconstruction of random dynamical systems

Reconstruction of nonlinear dynamical systems which may exhibit chaotic behavior is of great interest in the communities of mathematics, physics, statistics and signal processing. The purpose of reconstructing the model of the dynamical system from measured time series data is to estimate the state space parameters of the system comprised of the (vector of) control parameters of the deterministic part and the characteristics of the dynamic noise components. Under the assumption that the dynamic noise components are independent and identically distributed from some distribution (typically the Normal distribution) the model becomes Markovian and the full reconstruction of the system is achieved using some statistical methodology that involves the likelihood function; that is the joint distribution of the observations conditioned on all the unknown variables.

In the *frequentist* framework the state space parameters are considered fixed and uknown and the researcher seeks the value of the parameters that maximize some objective function which is usually the likelihood function. In contrast in the *Bayesian* setting all unknown quantities are treated as random variables and any prior knowledge is incorporated in the model with

the form of a prior distribution π over the parameters of interest. Combining the available information given in the prior and the likelihood function, Bayes theorem provides the *posterior* density, i.e. the conditional density of the parameters given the data.

Thus, a Bayesian model consists of a likelihood function $\mathcal{L}(\theta\,;x_1,\ldots,x_n) = \prod_{i=1}^{n}\ell(x_i\,|\,\theta)$ for a sample of size n of observations $x^{(n)}$ which are considered to be realizations of the random variables $X^{(n)}$ taking values on a state space $\mathbb{X}$ with density ℓ; and some prior distribution with density π for the parameters of interest $\theta \in \Theta$. Bayesian inference is then carried out based on the posterior distribution given by Bayes' theorem

$$\pi(\theta\,|\,x_1,\ldots,x_n) = \frac{\mathcal{L}(\theta\,;x_1,\ldots,x_n)\pi(\theta)}{\int_\Theta \mathcal{L}(\theta\,;x_1,\ldots,x_n)\pi(\theta)\mathrm{d}\theta} \overset{\theta}{\propto} \mathcal{L}(\theta\,;x_1,\ldots,x_n)\pi(\theta).$$

An estimate for the parameter of interest is then taken from some statistic, such as the mean, the mode or the median of a sample taken from the posterior distribution.

Bayesian formulation (Robert, 2007) has been of great use in the general field of noise perturbed dynamical systems. It was initially demonstrated in this context by Davies (1998), where Markov Chain Monte Carlo (MCMC) methods were used for nonlinear noise reduction. In Meyer & Christensen (2000, 2001) MCMC methods were applied for the parameter estimation of state–space nonlinear models, extending maximum likelihood-based existing methods (McSharry & Smith, 1999). Later, in Smelyanskiy et al. (2005) a path integral representation was proposed for the likelihood function, in order to make inference in stochastic nonlinear dynamics, extended for nonstationary systems in Luchinsky et al. (2008). In Matsumoto et al. (2001) and Nakada et al. (2005) Bayesian methods were suggested for reconstruction and prediction of nonlinear dynamical systems. Recently in Molkov et al. (2012), a Bayesian technique was proposed for the prognosis of the qualitative behavior of random dynamical systems under different forms of dynamical noise.

The methods introduced from the above researchers rely on the common assumption of the normality of the noise process. Such an assumption cannot always be justified and can cause inferential problems when the noise process departs from normality, for example when it produces outlying errors. Then the estimated variance of the normal errors is artificially enlarged causing poor inference for the system parameters of interest. For this reason, it is obvious that more flexible models must be constructed which will lead to accurate estimations even when the noise process departs from normality. In this thesis we aim to provide a Bayesian nonparametric formulation for the estimation of the parameters of the dynamical equations in the signle and multiple time series setting.

1.3 Aim and scope of the thesis

The aim of this thesis is to provide a Bayesian nonparametric framework for the full reconstruction of random dynamical systems. We start with the problem of reconstruction of a single dynamical system using *Geometric stick breaking* (GSB) processes and sequentially we

construct *Dependent Bayesian nonparametric priors* to extend the methodology in the context of multiple time series. The running examples are based on random *polynomial* maps which exhibit chaotic behavior. A very similar family of processes which are of great interest in the communities of statistics and signal processing are the *Polynomial Autoregressive Processes* (PAR) (Karakuș et al., 2015). Such a process can be represented in the notation $P^{(p)}AR(k)$.

$$x_n = \sum_{i=1}^{k} a_i^{(1)} x_{n-i} + \sum_{i=1}^{k}\sum_{j=1}^{k} a_{i,j}^{(2)} x_{n-i}x_{n-j} + \cdots + \sum_{i_1=1}^{k}\cdots\sum_{i_p=1}^{k} a_{i_1,\ldots,i_p}^{(p)} x_{n-i_1}\cdots x_{n-i_p} + \epsilon_n,$$

where ϵ_n is an excitation sequence. These processes have been used a lot in the context of time series and signal processing because they are linear in the parameters and thus many mathematical applications developed for linear models can be employed easily.

In the first half of this thesis, we will take a Bayesian nonparametric approach to reconstruct and predict random dynamical systems. We relax the common assumption of Normality of the noise process and we model the dynamical noise using a highly flexible family of density functions, providing a Bayesian nonparametric formulation (Ferguson, 1973; Fuentes-García et al., 2010). We are confident that, contrary to the assumption of normality, our Bayesian modeling will be able to capture the right shape of the true underlying noise density hence leading to an improved and reliable statistical inference for the system even in cases where the size of the observed time series is small. Some recent applications of Bayesian nonparametric methods in nonlinear dynamical systems include Dirichlet Process (DP) based reconstruction (Hatjispyros et al., 2009) and joint state–measurement noise density estimation with non–Gaussian and Gaussian observational and dynamical noise components respectively (Jaoua et al., 2013).

The problems and methods discussed so far assume that we are interested in the reconstruction and prediction of a single time series. In the second half of the thesis we propose a Bayesian nonparametric mixture model for the joint full reconstruction of a finite collection of dynamical equations, given observed dynamically–noise–corrupted chaotic time series. The method of reconstruction is based on the Pairwise Dependent Geometric Stick–Breaking Process (PDGSBP) mixture priors. Based on the PDGSBP prior we are able to extend the inferential procedure provided by the GSBR model in a multiple time series setting.

CHAPTER 2 In the next chapter we provide the necessary background on Bayesian nonparametrics. We start with the description of the Dirichlet process (DP) and the Geometric stick–breaking process (GSB) priors which are the two random probability measures used in our methods. Due to their discrete nature, it is not possible to model densities. To model overcome this obstacle and in order to model densities, we discuss Bayesian nonparametric mixture models and their use as a prior on the space of densities. Since a Bayesian nonparametric model is a Bayesian model with the prior defined on an infinite dimensional space, we review MCMC methods that we will use in order to perform posterior inference in such models.

CHAPTER 3 The thesis then proceeds with a review of the Bayesian nonparametric reconstruction model DPR based on the Dirichlet process DP proposed by Hatjispyros et al. (2009) and then we move to the *Geometric Stick-Breaking reconstruction model* (GSBR) introduced in Merkatas et al. (2017). We

propose a Bayesian nonparametric framework for the estimation and prediction, from observed time series data, of discretized random dynamical systems. The size of the observed time series can be small and the additive noise may not be Gaussian distributed. We show that when the dynamical noise departs from normality, simple parametric MCMC models are inefficient. Our models, assume an unknown error process in the form of a countable mixture of zero mean normals, where a-priori the number of the countable normal components and their variances is unknown. Our method infers the number of unknown components and their variances i.e. it infers the density of the error process directly from the observed data. We demonstrate numerically that the associated quasi invariant measure of the system appears naturally as posterior predictive marginal of the out–of–sample variables forming a prediction barrier.

Here we introduce our second contribution. We present a new approach to the joint estimation of partially exchangeable observations. This is achieved by constructing a model with pairwise dependence between random density functions, each of which is modeled as a mixture of GSB processes. We demonstrate numerically that mixture modeling with *Pairwise Dependent Geometric Breaking process* (PDGSBP) priors introduced by Hatjispyros et al. (2017a) is sufficient for prediction and estimation purposes. Moreover the corresponding Gibbs sampler for estimation is faster and easier to implement than the DP counterpart. For a fair comparison between the proposed PDGSBP model and the PDDP model of Hatjispyros et al. (2011) we adopt synchronized prior specification. To this end we randomize the concentration masses of the PDDP model leading to a more efficient model which we refer to as *randomized Pairwise Dependent Dirichlet Process* prior. We provide a modified MCMC scheme for the update of the individual concentration masses. CHAPTER 4

Finally we provide a Bayesian nonparametric model for the joint reconstruction of dynamical equations from dynamically–noisy–corrupted chaotic time series data. The main idea is to apply a PDGSBP prior to the space of densities of the additive errors. Under the assumption that the zero–mean processes responsible for the generation of the time series have common characteristics, for example they have same tail behavior, it is possible under carefully selected *borrowing-of-strength* prior specifications, to reconstruct the dynamical equations of the processes responsible for the generation of the time series especially for those systems for which the corresponding sample size is small; i.e. inefficient for an independent estimation with an rDPR or GSBR model. CHAPTER 5

Chapter 2

Bayesian nonparametric models

In this chapter we review the Bayesian nonparametric (BNP) priors used for the construction of our models. We start with the definition of the most common BNP prior that is the Dirichlet process (DP) and its extensions and then we review the Geometric stick breaking (GSB) prior; a random probability measure with simple weight structure. A class of MCMC algorithms that update the components of the random probability measures in the inferential procedure is also presented for these models. For an extensive study of the theory and applications of BNP models in statistics and machine learning problems we refer to Hjort et al. (2010) and Müller et al. (2015).

To give an intuition for the need of BNP models consider the problem of density estimation which is of great importance in Statistics. That is, given a collection of observations $(x_1, \ldots, x_n)$ on some measurable space $(\mathbb{X}, \mathcal{X})$, where $\mathbb{X}$ is a Polish space and $\mathcal{X}$ its associated Borel σ–algebra, we want to estimate their distribution. Formally, we have a collection of random variables

$$x_1, \ldots, x_n \,|\, F \overset{\text{iid}}{\sim} F$$

and the aim is to infer the uknown distribution F.

In the Bayesian approach, one should define a prior over the parameter which is now the unknown distribution F. This prior is defined over $\mathcal{P}_{\mathbb{X}}$, the space of all probability measures on $\mathbb{X}$ which now acts as the parameter space. Such priors are *random probability measures* defined on $\mathbb{X}$ that is, measurable functions $\mathbb{G} : \Omega \times \mathcal{X} \to \mathcal{P}_{\mathbb{X}}$ such that

1. $\omega \mapsto \mathbb{G}(\omega, A)$ is a probability measure on $(\mathbb{X}, \mathcal{X})$ for each $A \in \mathcal{X}$.
2. $A \mapsto \mathbb{G}(\omega, A)$ is a random variable for each $\omega \in \Omega$.

Reviewing the *Dirichlet distribution* will reinforce our intuition and clarify certain types of representation and properties of the *Dirichlet Process*. For this reason the definition of the Dirichlet distribution is provided.

Definition 2.1 (Dirichlet Distribution)**.** *Let $Z_1, \ldots, Z_K$, be independent $\mathcal{G}(a_j, 1), 1 \leq j \leq K$, random variables and let $Z = Z_1 + \cdots + Z_K$. The vector $w = (w_1, \ldots, w_K)$ where*

$$w_j = \frac{Z_j}{Z},$$

has the Dirichlet distribution with parameters $\alpha = (\alpha_1, \ldots, \alpha_K)$ *with probability density function*

$$\text{Dirichlet}(w \,|\, \alpha) := \frac{\Gamma(\sum_{j=1}^{K} \alpha_j)}{\prod_{i=1}^{K} \Gamma(\alpha_j)} w_1^{\alpha_1 - 1} \cdots w_K^{\alpha_K - 1} \mathcal{I}((w_1, \ldots, w_K) \in \triangle^K),$$

with

$$\triangle^K = \left\{ (w_1, \ldots, w_K) \in \mathbb{R}_+^K : \sum_{i=1}^{K} w_i = 1 \right\}.$$

2.1 Dirichlet Process

The Dirichlet process (Ferguson, 1973) is essentially a distribution over all discrete distributions. Intutitively this means that if we draw a sample from a DP we get a probability distribution on $\mathbb{X}$ and we write $\mathbb{G} \sim \text{DP}$ to denote that $\mathbb{G}$ is a sample from a DP.

Definition 2.2 (Dirichlet process)**.** *Let* $c > 0$ *and* H *be a probability measure defined on* $\mathbb{X}$*. We say* $\mathbb{G}$ *is a draw from a Dirichlet process with concentration parameter* c *and base distribution* H*, or* $\mathbb{G} \sim \text{DP}(c, H)$*, if and only if for each finite measurable partition* $\{A_1, \ldots, A_n\}$ *of* $\mathbb{X}$*, the vector of random probabilities* $(\mathbb{G}(A_1), \ldots, \mathbb{G}(A_k))$ *is distributed according to the Dirichlet distribution with parameters* $(cH(A_1), \cdots, cH(A_k))$*.*

From the definition above, it is clear that $\mathbb{G}$ is a draw from a DP if all its finite marginal distributions are Dirichlet distributions. Ferguson (1973) has shown that such a construction meets the conditions of *Kolmogorov consistency theorem* guaranteeing the existence of the Dirichlet process on a state space $\mathbb{X}$.

To see the rôle of c and H consider the measurable partition of $\mathbb{X} = \{A, A'\}$. Then $\mathbb{G}(A)$ is distributed as a $\mathcal{B}e(cH(A), cH(A'))$ random variable so we have that $\mathbb{E}[\mathbb{G}(A)] = H(A)$. Thus H is specifying where the mass of $\mathbb{G}$ is distributed on average. From the properties of the Beta distribution we have that $\text{Var}[\mathbb{G}(A)] = (1 + c)^{-1}[H(A)(1 - H(A))]$, so the parameter c controls the variability around the mean and can be regarded as an inverse variance parameter. So as $c \to \infty$ the prior is more tightly concentrated around the mean. In Figure 2.1, we represent the effect of the parameter c on 20 random distributions sampled from a DP prior with mean distribution a standard normal distribution $\mathcal{N}(0, 1)$.

2.1.1 Properties of DP

Conjugacy of Dirichlet process

An attractive property of the DP is its conjugacy meaning that given a sample of observations drawn from a DP that is $x_1, \ldots, x_n \,|\, \mathbb{G} \overset{\text{iid}}{\sim} \mathbb{G}$, the posterior random measure $\mathbb{G} \,|\, x_1, \ldots, x_n$ is also distributed according to a DP. Of course the concentration parameter and the base distribution of the posterior random measure will be updated in the light of "data". Now $\mathbb{G}$ is itself a random distribution and thus we can draw samples from $\mathbb{G}$ which are regarded as "data".

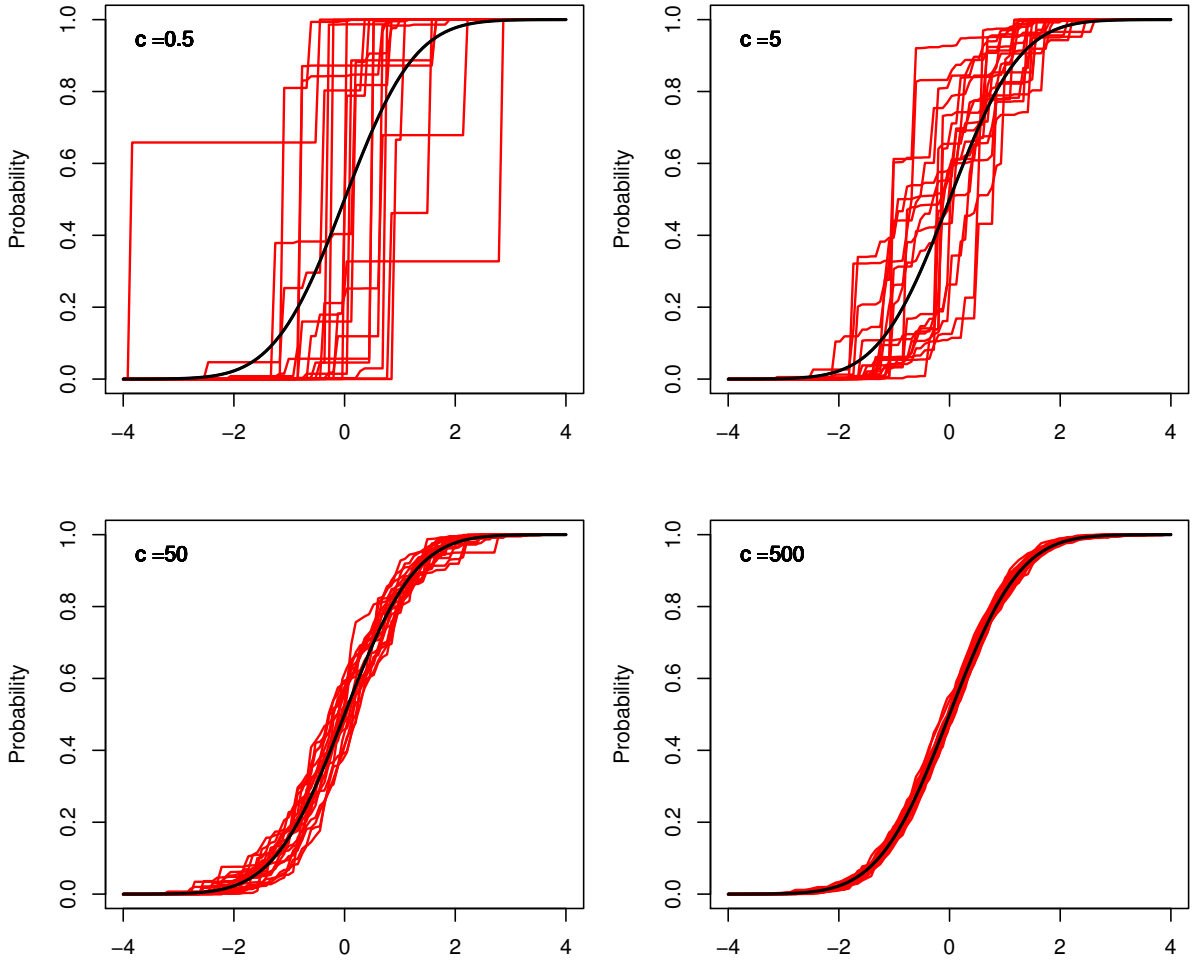

Figure 2.1: Random cdf's resulting from 20 draws from the prior $\mathrm{DP}(c, H)$ for $c = 0.5, 5, 50, 500$, and $H(\mathrm{d}x) = \mathcal{N}(x \,|\, 0, 1)\mathrm{d}x$. As the parameter c increases the prior concentrates around the mean H.

So, suppose that we have a sample $x_1, \ldots, x_n \,|\, \mathbb{G} \overset{\text{iid}}{\sim} \mathbb{G}$ and $\mathbb{G} \sim \mathrm{DP}(c, H)$. Let $A_1, \ldots, A_k$ be a finite measurable partition of $\mathbb{X}$ and let $n_j = \sum_{i=1}^{n} \mathcal{I}(x_i \in A_j)$ for $j = 1, \ldots, k$. The likelihood model for $(n_1, \ldots, n_k)$ is multinomial and from the conjugacy of the finite dimensional Dirichlet prior to the multinomial likelihood we have that

$$(\mathbb{G}(A_1), \ldots, \mathbb{G}(A_k)) \,|\, x_1, \ldots, x_n \sim \mathrm{Dirichlet}(cH(A_1) + n_1, \ldots, cH(A_k) + n_k).$$

The above relation holds for every finite measurable partition so we conclude that the posterior distribution must also be a DP. Manipulating the parameters of the finite dimensional posterior we can write the posterior of a DP as

$$\mathbb{G} \,|\, x_1, \ldots, x_n \sim \mathrm{DP}\left(c + n, \frac{c}{c+n} H + \frac{n}{c+n} \frac{\sum_{i=1}^{n} \delta_{x_i}}{n}\right). \tag{2.1}$$

It is clear from the above that the posterior base distribution is a weighted average of the prior base distribution H and the empirical distribution. We will show below that the posterior base

distribution is also the predictive distribution for a new sample drawn from the DP.

Predictive distributions

Now consider again a sample $x_1, \ldots, x_n$ from a random distribution $\mathbb{G}$ which is distributed according to a DP denoted by $\mathbb{G} \sim \mathrm{DP}(c, H)$. We are interested in the predictive distribution for a new observation $x_{n+1} \,|\, x_1, \ldots, x_n$. Integrating the random distribution $\mathbb{G}$ the predictive distribution of x_{n+1} for any $A \subset \mathbb{X}$ is given by

$$\begin{aligned} P(x_{n+1} \in A \,|\, x_1, \ldots, x_n) &= \int_{\mathbb{G} \in \mathcal{P}_{\mathbb{X}}} P(x_{n+1} \in A \,|\, \mathbb{G})\, P(\mathbb{G} \,|\, x_1, \ldots, x_n)\, \mathrm{d}\mathbb{G} \\ &= \int_{\mathbb{G} \in \mathcal{P}_{\mathbb{X}}} \mathbb{G}(A)\, P(\mathbb{G} \,|\, x_1, \ldots, x_n)\, \mathrm{d}\mathbb{G} \\ &= \mathbb{E}[\mathbb{G}(A) \,|\, x_1, \ldots, x_n] \\ &= \frac{c}{c+n} H(A) + \frac{n}{c+n} \frac{\sum_{i=1}^{n} \delta_{x_i}(A)}{n}. \end{aligned} \tag{2.2}$$

Thus with $\mathbb{G}$ integrated out the predictive distribution is given by

$$x_{n+1} \,|\, x_1, \ldots, x_n \sim \frac{1}{c+n} \left(cH + \sum_{i=1}^{n} \delta_{x_i} \right). \tag{2.3}$$

2.1.2 Representations of a DP

Many different representations have been proposed for the DP each of them giving nice properties useful for the construction of MCMC algorithms. Below we provide the most popular representations found in the literature.

Stick–breaking represenation

A random probability measure $\mathbb{G}$ sampled from a DP admits a stick–breaking representation. Sethuraman (1994) has shown that if $\mathbb{G} \sim \mathrm{DP}(c, H)$ then

$$\mathbb{G} = \sum_{k=1}^{\infty} w_k \delta_{x_k}, \tag{2.4}$$

where $(x_k)_{k \geq 1}$ is a sequence of independent and identical distributed random variables with distribution H and the weights $(w_k)_{k \geq 1}$ are stick–breaking, that is for a sequence $(z_k)_{k \geq 1}$ with $z_k \sim \mathcal{B}e(1, c)$

$$w_1 = z_1, \quad w_k = z_k \prod_{l=1}^{k-1} (1 - z_k),\ k \geq 2 \tag{2.5}$$

The name stick breaking comes from the definition of the weights which can be thought as the length of a piece of a unit–length stick assigned to the k–th value. From the stick breaking

represenation it is clear that random probabilitiy measures sampled from a DP are almost surely discrete.

The Dirichlet process belongs to the general class of stick–breaking priors (Ishwaran & James, 2001) where the Beta random variables are allowed to have different parameters for each k. That is different stick breaking priors can be obtained for a BNP inference if we let $z_k \sim \mathcal{B}e(a_k, b_k)$ for each k. Although to ensure that the weights will add up to 1 it must be verified that

$$\sum_{k=1}^{\infty} \log\left(1 + \frac{a_k}{b_k}\right) = +\infty.$$

We will see later that the stick breaking representation is extremely useful for planning Gibbs samplers imputing the random probability measure in the inferential procedure.

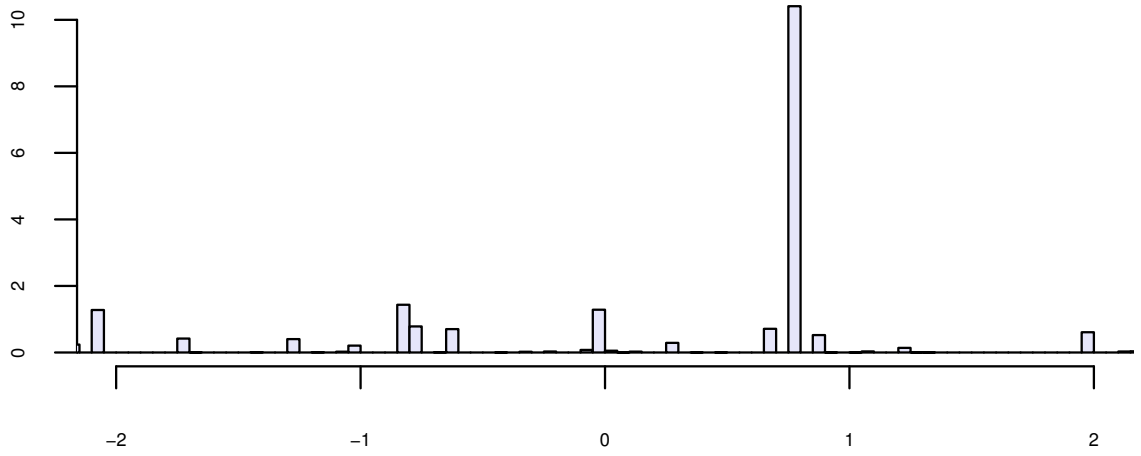

Figure 2.2: A draw from a Dirichlet process prior $\mathrm{DP}(c, H)$ with $c = 10$ and $H \sim \mathcal{N}(0, 1)$, using the stick–breaking representation.

Generalized Pólya urn

Beside the stick breaking representaion, the DP can be represented by a generalized Pólya urn (Blackwell & MacQueen, 1973). In this representation it becomes clear that the DP exhibits a clustering property.

Let $x_1, \ldots, x_n \,|\, \mathbb{G} \overset{\text{iid}}{\sim} \mathbb{G}$ with $\mathbb{G} \sim \mathrm{DP}(c, H)$. Then the distribution of $x_{n+1} \,|\, x_1, \ldots, x_n$ can be written in terms of successive conditional distributions as

$$x_{n+1} \,|\, x_1, \ldots, x_n \sim \frac{c}{c+n} H + \frac{1}{c+n} \sum_{i=1}^{n} \delta_{x_i}.$$

Thus, a new sample will be with probability $c(c+n)^{-1}$ a new draw from H or a previously seen sample x_i with probability $(n+c)^{-1}$. Due to the discrete nature of this distribution, ties will occur. If we denote by $(\tilde{x}_1, \ldots, \tilde{x}_k)$ the unique values of $(x_1, \ldots, x_n)$ and with $n_j, 1 \leq j \leq k$,

their corresponding frequencies such that $\sum_j n_j = n$, we can write the conditional distribution as

$$x_{n+1} \,|\, x_1, \ldots, x_n \sim \begin{cases} \delta_{\tilde{x}_j}, & \text{with probability } \frac{n_j}{n+c} \text{ for j = 1,. . . ,k.} \\ H, & \text{with probability } \frac{c}{c+n}. \end{cases}$$

The sequence of $(x_1, \ldots, x_n)$ constructed this way is infinitely exchangeable because $x^{(n)} \,|\, \mathbb{G}, n \geq 1$, are i.i.d samples from $\mathbb{G}$ and thus by de Finetti's representation theorem there exists a random distribution Π such that

$$P(x_1, \ldots, x_n) = \int_{\mathcal{P}_{\mathbb{X}}} \prod_{i=1}^{n} \mathbb{G}(x_i) \Pi(\mathrm{d}\mathbb{G}).$$

The random distribution Π of $\mathbb{G}$ is shown to be the Dirichlet process (Blackwell & MacQueen, 1973).

It is worth noting that this particular representation of the Dirichlet process is useful in planning *marginal samplers* for the DP. These samplers are MCMC algorithms which rely on the integration of the random distribution $\mathbb{G}$ to avoid infinite number of updates. The celebrated Algorithm 8 introduced by Neal (2000) belongs to this class and is used as a reference algorithm in the class of marginal samplers.

Representation as an NRMI

The DP is in the general class of random probability measures called *normalized random measures with independent increments* (NRMI). To define a NRMI the definition of a *completely random measure* (CRM) is required (Kingman, 1967). In the following we let $(\mathbf{M}_{\mathbb{X}}, \mathcal{M}_{\mathbb{X}})$ denote the measurable space of all finitely bounded measures on $\mathbb{X}$.

Definition 2.3 (Completely random measure)**.** *Let $\tilde{\mu}$ be a measurable mapping from $(\Omega, \mathcal{F}, P)$ into $(\mathbf{M}_{\mathbb{X}}, \mathcal{M}_{\mathbb{X}})$ such that for any collection $A_1, \ldots, A_k$ in $\mathcal{X}$ with $A_i \cap A_j = \emptyset$ for $i \neq j$, the collection of random variables $\tilde{\mu}(A_i), \ldots, \tilde{\mu}(A_k)$ are mutually independent. Then $\tilde{\mu}$ is called a completely random measure (CRM).*

In the following we will consider CRM's that can be represented as

$$\tilde{\mu} = \sum_{k=1}^{\infty} \tilde{J}_k \delta_{x_k}, \tag{2.6}$$

where the jumps $(\tilde{J})_{k \geq 1}$ and the $\mathbb{X}$–valued locations $(x_k)_{k \geq 1}$ are random.

The distribution of $\tilde{\mu}$ is characterized from the Lévy–Khintchine representation which states

$$\mathbb{E}\left[e^{-\int_{\mathbb{X}} f(x)\tilde{\mu}(\mathrm{d}x)}\right] = \exp\left\{-\int_{\mathbb{R}^+ \times \mathbb{X}} [1 - e^{-yf(x)}] \nu(\mathrm{d}s, \mathrm{d}x)\right\}.$$

The measure ν is called Lévy intensity and describes the distribution of the random points $(\tilde{J}_k, x_k)_{k \geq 1}$ as a Poisson random measure with mean ν. For our purposes we will assume that

the measure ν factorizes as $\nu(\mathrm{d}s, \mathrm{d}x) = \rho(\mathrm{d}s)H(\mathrm{d}x)$ implying independence between the random masses and the random locations. If this is the case, the CRM is called *homogeneous*.

A normalized random measure with independent increments can be defined through the normalization of a CRM.

Definition 2.4 (Normalized random measure)**.** *Let $\tilde{\mu}$ be a homogeneous CRM with $\nu(\mathrm{d}s, \mathrm{d}x) = \rho(\mathrm{d}s)H(\mathrm{d}x)$ and $T = \tilde{\mu}(\mathbb{X})$ be the total mass. An almost surely discrete random probability measure defined via*

$$\mu = \frac{\tilde{\mu}}{T} = \sum_{k=1}^{\infty} w_k \delta_{x_k},$$

where $(w_k)_{k\geq 1}$ is a sequence of random probabilities defined by normalizing $(\tilde{J}_k)_{k\geq 1}$ with respect to T, is called normalized random measure with independent increments (NRMI).

In order for the normalization to be well defined, it must be ensured that $T < +\infty$ almost surely. If the following two conditions for the Lévy measure are satisfied

$$\int_{\mathbb{R}^+} \rho(\mathrm{d}s) = +\infty \text{ and } \int_{\mathbb{R}^+} (1 - e^{-s})\rho(\mathrm{d}s) < +\infty.$$

then $T < +\infty$ almost surely.

An alternative definition of the DP as a normalized Gamma CRM was introduced in Ferguson (1973).

Definition 2.5 (Dirichlet Process (Ferguson, 1973))**.** *Let $\tilde{\mu}_g$ be a Gamma CRM that is a homogeneous CRM with Lévy intensity measure*

$$\rho(\mathrm{d}s)H(\mathrm{d}x) = c\, s^{-1}\, e^{-s}\mathrm{d}s\, H(\mathrm{d}x),$$

where $c > 0$ and let T be its total mass. The random probability measure

$$\mu = \frac{\tilde{\mu}_g}{T},$$

is a Dirichlet process with parameter c.

2.2 Geometric stick breaking process

An interesting random probability measure can be obtained from the general class of stick breaking priors (Ishwaran & James, 2001) by using only one Beta random variable for the construction of the weights. Recall that a stick breaking prior is a random probability measure that admits the representation

$$\mathbb{G} = \sum_{k=1}^{\infty} w_k \delta_{x_k},$$

where $(x_k)_{k\geq 1} \overset{\text{iid}}{\sim} H$ and the weights w_k are constructed via a stick breaking process as

$$w_1 = z_1, \quad w_k = z_k \prod_{l=1}^{k-1}(1 - z_l),\ k \geq 2,$$

for $(z_k)_{k\geq 1} \sim \mathcal{B}e(a_k, b_k)$.

If we replace the sequence $(z_k)_{k\geq 1}$ with a single random variable $\lambda \sim \mathcal{B}e(a, c)$ and construct the weights with a geometric structure as

$$w_k = \lambda(1 - \lambda)^{k-1}, \tag{2.7}$$

the resulting probability measure

$$\mathbb{G} = \lambda \sum_{k=1}^{\infty}(1 - \lambda)^{k-1}\delta_{x_k}, \tag{2.8}$$

is known as a *geometric stick breaking* (GSB) process or *geometric weights prior* (Fuentes-García et al., 2010). From now on, we will denote a random probability measure drawn from a GSB process as $\mathbb{G} \sim \text{GSB}(\lambda, H)$.

The GSB prior can be seen as a removal of a level of randomness in a nonparametric model based on the DP by replacing the stick breaking weights with their expectations. The expectated value of the stick breaking weights of the DP is given by

$$\mathbb{E}[w_k] = \frac{1}{1+c}\prod_{i=1}^{k-1}\frac{c}{1+c} = \frac{1}{1+c}\left(\frac{c}{1+c}\right)^{k-1},$$

which is a reparametrization of Equation (2.7) with $\lambda = (1 + c)^{-1}$.

The advantage of using a GSB as a BNP prior is its reduced variability of the weights thus leading to improved estimation. Despite the fact that a level of randomness is removed, standard results of Ongaro & Cattaneo (2004) can be used to prove that a GSB random probability measure still has full support on the space of discrete probability measures of $\mathcal{P}_{\mathbb{X}}$. For the GSB process we prove the following proposition.

Proposition 2.1. *Let* $\mathbb{G} \sim \text{GSB}(\lambda, H)$. *Then for every* $A \in \mathcal{X}$ *we have*

1. $\mathbb{E}[\mathbb{G}(A)] = H(A)$,
2. $\text{Var}[\mathbb{G}(A)] = \frac{\lambda}{2-\lambda}H(A)(1 - H(A))$.

Proof. PART 1. From the definition of the GSB random probability measure we have that $\mathbb{G} = \sum_{k=1}^{\infty} w_k\delta_{x_k}$. The expectation of the random variable $\mathbb{G}(A)$ is given by

$$\mathbb{E}[\mathbb{G}(A)] = \mathbb{E}\left[\sum_{k=1}^{\infty} w_k\delta_{x_k}\right] = \sum_{k=1}^{\infty} w_k\mathbb{E}[\delta_{x_k}(A)] = H(A).$$

PART 2. It is that

$$\begin{aligned}
\text{Var}[\mathbb{G}(A)] &= \mathbb{E}[\mathbb{G}^2(A)] - \mathbb{E}^2[\mathbb{G}(A)] \\
&= \mathbb{E}\left[\sum_{k=1}^{\infty} w_k^2 \delta_{x_k}^2(A) + \sum_{k\neq j} w_k w_j \delta_{x_k}(A)\delta_{x_j}(A)\right] - \left[\mathbb{E}\sum_{k=1}^{\infty} w_k \delta_{x_k}(A)\right]^2 \\
&= \mathbb{E}\sum_{k=1}^{\infty} w_k^2 H(A) + H^2(A)\mathbb{E}\sum_{k\neq j} w_k w_j - \left[\mathbb{E}\sum_{k=1}^{\infty} w_k H(A)\right]^2 \\
&= H(A)\mathbb{E}\sum_{k=1}^{\infty} w_k^2 + H^2(A)\mathbb{E}\sum_{k\neq j} w_k w_j - H^2(A)\left[\mathbb{E}\sum_{k=1}^{\infty} w_k^2\right] H^2(A) \\
&\quad - \mathbb{E}\sum_{k\neq j} w_k w_j H^2(A),
\end{aligned}$$

from which we conclude that $\text{Var}[\mathbb{G}(A)] = H(A)(1-H(A))\mathbb{E}\left(\sum_{k=1}^{\infty} w_k^2\right)$. Because $w_k = \lambda(1-\lambda)^{k-1}$ it follows that

$$\text{Var}[\mathbb{G}(A)] = \frac{\lambda}{2-\lambda} H(A)(1-H(A)).$$

□

2.3 Bayesian nonparametric mixtures

In the previous sections we have seen that stick breaking random probability measures are almost surely discrete even when the base distribution H is continuous. This implies that these BNP priors are not suitable to model densities. An approach for the construction of a prior process whose realizations are absolutely continuous random distribution functions was first proposed by Antoniak (1974) and followed by Lo et al. (1984). The main idea of the Dirichlet Process Mixture (DPM) model is to convolute a kernel of some parametric family with a random probability measure sampled from a BNP prior.

More formally, consider the parametric family of kernels

$$\mathcal{K}_\theta = \{K(\,\cdot\,|\,\theta)\,|\,\theta \in \Theta \subseteq \mathbb{R}^k\}. \tag{2.9}$$

A mixture distribution is a convex combination of the members of $\mathcal{K}_\theta$ with the representation

$$f(x) = \sum_{k=1}^{\infty} p_k K(x\,|\,\theta_k) \quad \text{with} \quad \sum_{k=1}^{\infty} p_k = 1. \tag{2.10}$$

where the sequence p_k belongs to the infinite dimensional *simplex*

$$\triangle^{\infty} = \left\{p_1, p_2, \ldots \,|\, \sum_{k=1}^{\infty} p_k = 1\right\}. \tag{2.11}$$

Any discrete probability measure over the parameter space Θ can be written as

$$\mathbb{G} = \sum_{k=1}^{\infty} p_k \delta_{\theta_k}. \tag{2.12}$$

Using measure–theoretic notation, the mixture distribution of eq. (2.10) can be written as

$$f(x) = \int_{\Theta} K(x \,|\, \theta) \mathbb{G}(\mathrm{d}\theta), \tag{2.13}$$

where $\mathbb{G}$ is called the *mixing distribution*.

A Bayesian mixture model can be defined if we let the parameters of the mixture distribution that is, $(p_k, \theta_k)_{k \geq 1}$ be random. In order to do so, the problem reduces to define a random probability measure over the measurable space $(\Theta, \mathcal{B}(\Theta))$. This is the main idea behind the DPM model proposed by Lo et al. (1984) who considered the distribution of the random distribution $\mathbb{G}$ to be a Dirichlet process. In a hierarchical fashion, the model is represented as

$$\begin{aligned} x_i \,|\, \theta_i &\overset{\text{ind}}{\sim} K(x_i \,|\, \theta_i) \\ \theta_i \,|\, \mathbb{G} &\overset{\text{iid}}{\sim} \mathbb{G} \\ \mathbb{G} &\sim \mathrm{DP}(c, H). \end{aligned}$$

The interesting property of the DPM is that the posterior is also a DPM. More formally we have the following

Proposition 2.2. *If* $x_i \,|\, \theta_i \overset{\text{ind}}{\sim} f(x_i \,|\, \theta_i), 1 \leq i \leq n$ *and* $\theta_i \overset{\text{iid}}{\sim} \mathbb{G}$ *where* $\mathbb{G} \sim \mathrm{DP}(c, H)$ *then*

$$\mathbb{G} \,|\, x_1 \ldots, x_n \sim \int_{\Theta} \mathrm{DP}(c^*, H^*) P(\mathrm{d}\theta \,|\, x_1, \ldots, x_n),$$

with $P(\theta \,|\, \cdots)$ *being the posterior density and* c^*, H^* *the posterior concentration and the posterior base measure given in eq.* (2.1).

From the formulation of the Bayesian mixture, it is obvious that the distribution of the mixing measure can be replaced by any discrete random probability measure resulting the different Bayesian nonparametric mixtures. In this thesis we will use Bayesian nonparametric mixture models that use as a prior over the parameters of the mixing distribution a DP or a GSB process. For completeness we provide the hierarchical representation of a Geometric stick breaking mixture (GSBM) model below. It is that

$$\begin{aligned} x_i \,|\, \theta_i &\overset{\text{ind}}{\sim} K(x_i \,|\, \theta_i) \\ \theta_i &\overset{\text{iid}}{\sim} \mathbb{G} \\ \mathbb{G} &\sim \mathrm{GSB}(\lambda, H) \end{aligned}$$

2.4 Dependent processes

In Bayesian nonparametrics, the use of nonparametric priors such as the Dirichlet process (Ferguson, 1973) is justified by the assumption that the observations are exchangeable. However exchangeability is not a valid assumption for all kinds of data. For example in time series data there may be correlation between observations in proximate times resulting in non–exchangeable data sets.

When the exchangeability assumption fails one needs to use non–exchangeable priors. These priors are dependent stochastic processes i.e. distributions over a collection of measures indexed by values in some covariate space, such that the marginal distribution is described by a known nonparametric prior.

Below we will review two types of dependent Bayesian nonparametric priors. At first, we review *covariate* dependent priors introduced in MacEachern (1999). These priors include additional information in a model conditioning on a non–random variable taking values in some covariate space. Then we focus to dependent processes which are distributions over exchangeable collections of measures. The construction of nonexchangeable priors is a fresh and very active field of research especially for the statistics and machine learning communities. In Foti & Williamson (2015), a survey of the common approaches in the construction of dependent nonparametric processes is given.

2.4.1 Covariate–dependent models

In many applications there are datasets which may contain temporal, spatial or categorical information for which we are not interested in making inference. Instead we would like to condition upon it in order to improve inference. This additional information is introduced in the model by considering a variable z taking values in some *covariate* space $\mathcal{Z}$ so now the effect of interest X is a function of z that is $X(z)$. Since X is a function of z this means that the parameters θ must also be a function of $\theta(z)$ of z. The aim now is to construct a flexible *dependent* Bayesian nonparametric prior which accounts for the information given by the covariate z.

Generalizing the stick breaking represenation of the DP, MacEachern (1999) showed that a *dependent Dirichlet process* (DDP) can be defined via

$$\mathbb{G}_z(\,\cdot\,) = \sum_{k=1}^{\infty} w_k(z)\delta_{\theta_k(z)}(\,\cdot\,),$$

where $(w_k(z), \theta_k(z))$ are stochastic processes indexed in $\mathcal{Z}$. A classical example of the use of dependent DP's is the Bayesian nonparametric regression problem where a random probability measure $\mathbb{G}_z$ is constructed for each covariate z. Extensions to dependent DP models can be found in De Iorio et al. (2004), Griffin & Steel (2006), and Dunson & Park (2008).

Recently there has been growing interest for the use of simpler random probability measures which while simpler are yet sufficient for a Bayesian nonparametric density estimation. The geometric stick breaking (GSB) random probability measure (Fuentes-García et al., 2010) has been used for density estimation and has been shown to provide an efficient alternative to DP mixture models. Some recent papers extend this nonparametric prior to a dependent nonparametric prior.

In the construction of covariate dependent processes, GSB processes have been seen to provide an adequate model to the traditional dependent DP model. For example, for Bayesian regression, Fuentes-García et al. (2009) propose a covariate dependent process based on random probability measures drawn from a GSB process. Mena et al. (2011) used GSB random probability measures in order to construct a purely atomic continuous time measure–valued process, useful for the analysis of time series data. In this case, the covariate $z \geq 0$ denotes the time that each observation is (discretely) recorded and conditionally on each observation is drawn from a time–dependent nonparametric mixture model based on GSB processes.

2.4.2 Distributions over exchangeable measures

We have seen before that the assumption of exchangeability may be violated when the data are observed with covariates. This though, is not the only case where the exchangeability assumption fails. In real life applications data are often *partially exchangeable*. For example the data may consist of independent observations sampled from m populations, or may be sampled from an experiment conducted in m different geographical places. This means that the joint law is invariant under permutations within m subgroups of observations $(X_{j,1}, \ldots, X_{j,n_j})$, $j = 1, \ldots, m$, then for all $\pi_j \in S(n_j)$

$$((X_{1,i_1})_{1\leq i_1\leq n_1}, \ldots, (X_{m,i_m})_{1\leq i_m\leq n_m}) \stackrel{\mathrm{d}}{=} ((X_{1,\pi_1(i_1)})_{1\leq i_1\leq n_1}, \ldots, (X_{m,\pi_m(i_m)})_{1\leq i_m\leq n_m}) \quad (2.14)$$

Most of the proposals rely on the notion of partial exchangeability as set forth by De Finetti (1938) who formalizes the above idea. In simple words, partial exchangeability means that although not valid across the whole set of observations, exchangeability can hold true within m different groups of observations.

More formally, in analogy with Theorem 1.1, an infinite $\mathbb{X}$–valued process $X_j^{(\infty)}$,$1 \leq j \leq m$, defined over a probability space $(\Omega, \mathcal{F}, P)$, is partially exchangeable as in eq. (2.14) if and only if there exists a probability distribution Π over $\mathcal{P}_{\mathbb{X}}^m$, that satisfies

$$\begin{aligned} &P\{X_{ji} \in A_{ji} : 1 \leq j \leq m, 1 \leq i \leq n_j\} \\ &= \int_{\mathcal{P}^m} P\{X_{ji} \in A_{ji} : 1 \leq j \leq m, 1 \leq i \leq n_j \,|\, \mathbb{Q}_1, \ldots, \mathbb{Q}_m\} \, \Pi(\mathrm{d}\mathbb{Q}_1, \ldots, \mathrm{d}\mathbb{Q}_m) \end{aligned}$$

$$= \int_{\mathcal{P}^m} \prod_{j=1}^{m} P\{X_{ji} \in A_{ji} : 1 \leq i \leq n_j \,|\, \mathbb{Q}_j\} \, \Pi(\mathrm{d}\mathbb{Q}_1, \ldots, \mathrm{d}\mathbb{Q}_m)$$
$$= \int_{\mathcal{P}^m} \prod_{j=1}^{m} \left\{ \prod_{i=1}^{n_j} \mathbb{Q}_j(A_{ji}) \right\} \Pi(\mathrm{d}\mathbb{Q}_1, \ldots, \mathrm{d}\mathbb{Q}_m).$$

The de Finetti measure Π represents a prior distribution over partially exchangeable observations and is the distribution of some vector $(\mathbb{Q}_1, \ldots, \mathbb{Q}_m)$ of random probability measures on $\mathbb{X}$.

We devote the following few paragraphs to describe some common dependent processes which create distributions over exchangeable collections of measures. A typical scenario in which the measure Π is employed is with mixture models to generate random densities.

Hierarchical Dirichlet Process

A very popular dependent process with applications in ad–mixture models is the Hierarchical Dirichlet Process (HDP) proposed by Teh et al. (2006) which induces dependence among a collection of random probability measures by setting a hierarchical model over the locations of the random distributions of the groups.

More specifically, in the HDP model a random measure $\mathbb{G}_0$ is sampled from a $\mathrm{DP}(\gamma, H)$, and then for each group of data x_j a random measure $\mathbb{G}_j$ is sampled from a $\mathrm{DP}(c, \mathbb{G}_0)$. The distributions $\mathbb{G}_j$ can be used as mixing measures to generate the random densities f_j for the observations. Formally the model can be summarized as

$$x_{ji} \,|\, \theta_{ji} \overset{\text{ind}}{\sim} K(\,\cdot \,|\, \theta_{ji}), \;\; \theta_{ji} \,|\, \mathbb{G}_j \overset{\text{iid}}{\sim} \mathbb{G}_j$$
$$\mathbb{G}_j = \sum_{k=1}^{\infty} w_{jk} \delta_{\theta_k} \overset{\text{iid}}{\sim} \mathrm{DP}(c, G_0), \;\; \mathbb{G}_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\theta_k} \sim \mathrm{DP}(\gamma, H).$$

Note that in the HDP model that all the random measures share the same set of "atoms" (locations). This is because the base measure of the group specific DPs is the random distribution $\mathbb{G}_0$ and thus all $\mathbb{G}_j$ have the same support as $\mathbb{G}_0$. This way, different observations in the same group can share the same parameters, but also observations across different groups. Shared characteristics would correspond to large stick breaking weights.

A closely related approach in the modeling of densities, which are thought to be related, is the *Analysis of Densities* (AnDe) model, proposed by Tomlinson & Escobar (1999). The difference is that in the AnDe model, the "global" distribution $\mathbb{G}_0$ is itself a DPM. In this case, since $\mathbb{G}_j$ are independent draws from $\mathrm{DP}(c, \mathbb{G}_0)$ they have no atoms in common. Thus clusters arise within a group via the discreteness of $\mathbb{G}_j$ but atoms in different groups are different and there is no sharing of clusters between groups.

Dependence through a shared component

Another way to introduce dependence between a finite collection of probability measures $\mathbb{G}_j$, $1 \leq j \leq m$ is to model each random measure $\mathbb{G}_j$ as a convex combination of a *common* and an index specific *idiosyncratic* part. So for $0 < p_j < 1$ it is that

$$\mathbb{G}_j = p_j \mathbb{G}_0 + (1 - p_j)\mathbb{G}_j^*,$$

where $\mathbb{G}_0$ is the common component of all other measures and $\mathbb{G}_j^*$, $1 \leq j \leq m$ are the idiosyncratic parts to each $\mathbb{G}_j$.

This approach has been adopted by Müller et al. (2004); Bulla et al. (2009); Kolossiatis et al. (2013) under the assumption that $\mathbb{G}_0, \mathbb{G}_j^* \overset{\text{iid}}{\sim} DP(c, H)$. The resulting $\mathbb{G}_j$ measures have been used in the context of DPM models generating a collection of dependent random densities $f_j(x)$, $1 \leq j \leq m$ given by

$$f_j(x) = p_j \int_\Theta K(x \,|\, \theta)\mathbb{G}_0(\mathrm{d}\theta) + (1 - p_j) \int_\Theta K(x \,|\, \theta)\mathbb{G}_j^*(\mathrm{d}\theta).$$

The generative process of the model can be summarized as

$$\begin{aligned}
&x_{ji} \,|\, \theta_{ji} \overset{\text{ind}}{\sim} K(\,\cdot \,|\, \theta_{ji}),\ \theta_{ji} \,|\, \mathbb{Q}_j \overset{\text{iid}}{\sim} \mathbb{Q}_j \\
&\mathbb{Q}_j = p_j \mathbb{G}_0 + (1 - p_j)\mathbb{G}_j^*,\ p_j + (1 - p_j) = 1 \\
&\mathbb{G}_0,\, \mathbb{G}_j^* \overset{\text{iid}}{\sim} \mathrm{DP}(c, H).
\end{aligned}$$

A similar approach was taken by Griffin et al. (2013); Lijoi et al. (2014) who have replaced the Dirichlet random measures $\mathbb{G}_0, \mathbb{G}_j^*$ with a normalized random probability measure based on the normalized generalized gamma process (NGGP) (Brix, 1999) and σ–stable process (Kingman, 1975) repsectively. The NGGP process is a completely random measure whose Lévy intensity is given by

$$\nu_{c,\sigma,\zeta}(\mathrm{d}s, \mathrm{d}x) = \frac{c s^{-1-\sigma}}{\Gamma(1 - \sigma)} e^{-\zeta s} \mathrm{d}s H(\mathrm{d}x),$$

with $\sigma \in (0, 1)$, $c > 0$ and $\zeta \geq 0$. The NGGP includes as special case the DP when $\sigma = 0, \zeta = 1$ and the σ–stable process when $\zeta = 0$.

Pairwise dependent random probability measures

A more general dependence structure between a collection of measure $\mathbb{G}_j$ has been proposed by Hatjispyros et al. (2011). They have modeled the random distributions $\mathbb{Q}_j$ to be pairwise dependent that is

$$\mathbb{Q}_j = \sum_{l=1}^{m} p_{jl} \mathbb{G}_{jl}, \quad \sum_{l=1}^{m} p_{jl} = 1 \,\text{a.s.},$$

with $\mathbb{G}_{jl} = \mathbb{G}_{lj}$ being iid from the $\mathrm{DP}(c, H)$. Equivalently the proposed model can be written in matrix notation form as

$$\mathbb{Q} = (p \otimes \mathbb{G})\mathbf{1}.$$

where $p = (p_{jl})$ is the matrix of random selection probabilities, $\mathbb{G} = (\mathbb{G}_{jl})$ is the symmetric matrix of the independent Dirichlet measures and $p \otimes \mathbb{G}$ is the Hadamard product of two matrices defined as $(p \otimes \mathbb{G})_{jl} = p_{jl}\mathbb{G}_{jl}$. Letting $\mathbf{1}$ denote the $m \times 1$ matrix of ones it is that the jth element of the vector $\mathbb{Q}$ is given by $\mathbb{Q}_j$ defined above.

For the observations, the generative process is summarized

$$\begin{aligned}
&x_{ji} \,|\, \theta_{ji} \overset{\text{ind}}{\sim} K(\,\cdot \,|\, \theta_{ji}),\ \theta_{ji} \,|\, \mathbb{Q}_j \overset{\text{iid}}{\sim} \mathbb{Q}_j \\
&\mathbb{Q}_j = \sum_{l=1}^{m} p_{jl}\mathbb{G}_{jl},\ \sum_{l=1}^{m} p_{jl} = 1,\ \mathbb{G}_{jl} = \mathbb{G}_{lj} \\
&\mathbb{G}_{jl} \overset{\text{iid}}{\sim} \mathrm{DP}(c, H).
\end{aligned}$$

The same dependence structure was adopted in Hatjispyros et al. (2016) but now the iid DPs are forced to have the same atoms. The authors showed that adopting common atoms to the involved Dirichlet processes is sufficient for prediction and density estimation purposes within the concept of borrowing of strength.

In Chapter 4 we are going to describe a dependent process based on the GSB process. Although these measures have been used in covariate–dependent models, they haven't been used for modeling related density functions when samples from each density function are available.

2.5 Markov Chain Monte Carlo methods

In a Bayesian model the prior is combined with the likelihood i.e. the joint density of the observations given any parameters and the objective is to determine the posterior distribution, that is the conditional distribution of parameters given the data. Formally we have the following Bayesian model

$$\begin{aligned}
X_1, \ldots, X_n \,|\, \theta &\overset{\text{iid}}{\sim} \ell(\cdot \,|\, \theta) \\
\theta &\sim \pi.
\end{aligned}$$

Letting $\mathcal{L}(\theta\,; x_1, \ldots, x_n) = \prod_{i=1}^{n} \ell(x_i \,|\, \theta)$, inference is based on a sample from the posterior distribution given by

$$\pi(\theta \,|\, x_1, \ldots, x_n) = \frac{\mathcal{L}(\theta\,; x_1, \ldots, x_n)\pi(\theta)}{\int_\Theta \mathcal{L}(\theta\,; x_1, \ldots, x_n)\pi(\theta)\mathrm{d}\theta}. \tag{2.15}$$

If the prior and the likelihood do not form a conjugate pair the posterior distribution given in eq. (2.15) does not have a closed form. This is because the integral that appears in the denominator is intractable. Although the problem of approximating an integral can be dealt

with methods from numerical analysis, Monte Carlo (MC) methods are based on properties of random variables and Laws of Large Numbers providing a solution to the problem of integration.

Monte Carlo (MC) integration. Suppose that we want to evaluate an integral of a function h of a random variable $\theta \in \Theta$ which we assume it has density π. That is

$$I(h) = \int_{\Theta} h(\theta)\pi(\theta)\mathrm{d}\theta = \mathbb{E}_{\pi}[h(\theta)]. \tag{2.16}$$

Monte Carlo (MC) methods assume that we have an i.i.d. sample $\theta^{(N)} = (\theta_i)_{1\leq i\leq N}$ from the target density $\pi(\theta)$ and that the function h can be evaluated point–wise. The integral is then approximated from the *Monte Carlo estimator* defined as

$$I_N(h) = \frac{1}{N}\sum_{i=1}^{N} h(\theta_i), \tag{2.17}$$

which from the strong LLN converges to the desired integral. That is,

$$I_N(h) = \frac{1}{N}\sum_{i=1}^{N} h(\theta^{(i)}) \xrightarrow{\text{a.s.}} I = \int_{\Theta} h(\theta)\pi(\theta)\mathrm{d}\theta \text{ as } N \to \infty. \tag{2.18}$$

This MC estimator is unbiased, and by the strong Law of Large numbers will a.s. converge to $I(h)$. If the variance of $h(\theta)$ satisfies

$$\sigma_h^2 = \mathbb{E}_{\pi}[h^2(\theta)] - \mathbb{E}_{\pi}^2[h(\theta)] = \mathbb{E}_{\pi}[h^2(\theta)] - I^2(h) < +\infty, \tag{2.19}$$

the MC estimator $I_N(h)$ satisfies the following Central Limit Theorem, yielding convergence in distribution of the error

$$\sqrt{N}(I_N(h) - I(h)) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma_h^2). \tag{2.20}$$

In practice however, it is not always possible to generate samples from π so Markov Chain Monte Carlo (MCMC) methods provide a framework to obtain samples from the desired density. The basic idea of MCMC methods is to construct a stationary Markov chain $(\theta_i)_{i\geq 1}$ with stationary density which is the desired target density π. Independently of the starting point of the chain, after a long enough period, in terms of samples, the Markov chain will converge to its stationary distribution and samples from it can be considered as samples from the target density.

There is an extensive literature in MCMC methods establishing theoretical results and output diagnostics that is not possible to review it extensively. In the following we will review only the Gibbs sampler, one of the most common MCMC algorithms which is the main tool for inference in our models. More details on MCMC methods and theory can be found in Robert (2004); Brooks et al. (2011); Liang et al. (2011); Besag & Green (1993).

2.5.1 The Gibbs sampler

The Gibbs sampler is the simpler and most popular MCMC algorithm for Bayesian inference when it comes to the sampling of multidimensional distributions. It is a special case of the *Metropolis-Hastings* algorithm (Metropolis et al., 1953; Hastings, 1970) initially used in statistical physics. The Gibbs sampler updates each component θ_j of the vector of variables $\theta = (\theta_1, \ldots, \theta_d)$ using as a proposal distribution the associated full conditional distribution $\pi(\theta_j \,|\, \theta_{-j})$, where θ_{-j} is the vector of parameters with the j-th component removed. This makes the acceptance probability at each step equal to 1. In the statistics community, Gibbs sampling has been used only after the work of Geman & Geman (1984) for analyzing Gibbs distributions on lattices in the context of image processing.

More clearly, suppose that the unknown parameter is multidimensional $\theta = (\theta_1, \ldots, \theta_d)$, so the target distribution is multivariate. The vector of parameters can be partitioned and written as $\theta = (\theta_1, \ldots, \theta_k)$ where each $\theta_j, j = 1, \ldots, k$ may be unidimensional or multidimensional so that $\dim(\theta_1) + \cdots \dim(\theta_k) = d$. Consequently the target density can be written as $\pi(\theta_1, \ldots, \theta_k)$. The Gibbs sampler starts from an aribtrary point $\theta^0 = (\theta_1^0, \ldots, \theta_k^0)$ and alternates updating the comonents of θ by drawing from the relevant conditional distributions $\pi(\theta_j \,|\, \theta_{-j})$, according to the scheme presented in Algorithm 1 until the number of desired samples N.

Algorithm 1 : GIBBS sampling for multidimensional parameter.

1: **procedure** SAMPLE $\theta = (\theta_1, \ldots, \theta_k)$.
2: Initiliaze the chain $\theta^0 = (\theta_1^0, \ldots, \theta_k^0)$.
3: **for** $i = 1$ to N **do**
4: **for** $j = 1$ to k **do**
5: Sample $\theta_j^i \sim \pi(\theta_j \,|\, \theta_1^i, \ldots, \theta_{j-1}^i, \theta_{j+1}^{i-1}, \ldots, \theta_k^{i-1})$.
6: **end for**
7: **end for**
8: **end procedure**

In its simplest form, it is assumed that the conditional distributions $\pi(\theta_j \,|\, \theta_{-j})$ are of standard form. Nevertheless if for some components the conditional distribution is unknown one can use a *Metropolis* step to sample from the conditional of the particular component leading to a *Metropolis within Gibbs* sampler. For an adaptive rejection Metropolis within Gibbs sampler we refer to Gilks et al. (1995). Alternatively one can use *slice sampling*; an *auxiliary variable* method to sample the components with nonstandard full conditionals, constructing embedded Gibbs samplers and thus circumventing the Metropolis step. Auxiliary variable methods will be analyzed later on in the text.

If some θ_j, $1 \leq j \leq k$ has $\dim(\theta_j) \geq 2$ then the elements of θ_j are sampled simultaneously as a *block*. If this is the case the Gibbs sampler is called a *blocked Gibbs* sampler. Note here that having blocks in a Gibbs sampler, the Markov chain reaches the stationary distribution faster but this comes at the expense of sampling from multivariate distributions.

2.5.2 Auxiliary variable methods–Slice sampling

Whenever π, the target density or the full conditional of θ_j, in the Gibbs sampler is of a non standard form, auxiliary variable methods (Damien et al., 1999) can be used to result in a Gibbs sampler having a set of easily sampled standard full conditionals.

These methods augment the target density with a positive latent variable u constructing the joint density of u and θ. This way, the marginal density for θ is given by π and the Gibbs sampler is extended to include an extra full conditional for u.

Suppose that we wish to sample from a density π given by

$$\pi(\theta) \propto q(\theta) f(\theta), \tag{2.21}$$

where q is a density of known form and f is a non–negative invertible function. With the introduction of a latent variable $u : \Omega \to \mathbb{R}_+$, the joint density can be written as

$$\pi(\theta, u) \propto q(\theta) \mathcal{I}(u < f(\theta)). \tag{2.22}$$

Marginalizing u from eq. (2.22) we get $\pi(\theta)$, thus, the augmentation is valid. The full conditional for u is uniform $\mathcal{U}(0, f(\theta))$. The full conditional for θ is now a truncated version of q restricted to the set

$$A_u = \{\theta : u < f(\theta)\}. \tag{2.23}$$

Below we provide a simple example to sample from a density with gaussian functional form.

Example 2.1. *Suppose we want to sample from the density given by* $\pi(\theta) \propto \exp\{-\frac{\tau}{2}\theta^2\}$. *We will introduce an auxiliary random variable* u *such that the joint density is*

$$\pi(\theta, u) \propto \exp\{-\frac{\tau}{2}u\}\mathcal{I}(u > \theta^2).$$

Clearly, integrating out u *leads to original density* $\pi(\theta)$. *Now we have that the full condtionals are given by*

$$\pi(u \,|\, \theta) \propto \exp\{-\frac{\tau}{2}u\}\mathcal{I}(u > \theta^2) \tag{2.24}$$

$$\pi(\theta \,|\, u) \propto \mathcal{I}(u > \theta^2). \tag{2.25}$$

where (2.24) is a truncated exponential density with rate $\tau/2$ *over the set* (θ^2, ∞) *easily sampled and (2.25) is a uniform density over the interval* $(-\sqrt{u}, \sqrt{u})$.

More examples for the usage of this method in fancy densities and applications in Bayesian hierarchical models can be found in the work of Damien et al. (1999). Some algorithmic improvements and convergence results are presented in Mira (1998) and Neal (2003).

2.6 MCMC for Bayesian nonparametric mixture models

In this section we review MCMC algorithm for posterior inference in Bayesian nonparametric mixture models. The algorithms are based on slice sampling with auxiliary variables and belong to the class of *conditional* samplers. In a conditional sampler, the random probability measure that acts as a prior on the mixture parameters is imputed in the inferential procedure. Retaining the random distribution is useful since it removes the dependence between the parameters which exist in the marginal samplers (Neal, 2000) based on the Pólya–urn scheme.

In what follows we will consider a Bayesian nonparametric mixture model where the prior on the mixing distribution is either a DP or a GSB process. That is, we have a Bayesian mixture model

$$\begin{aligned} x_i \,|\, \theta_i &\overset{\text{ind}}{\sim} K(x_i \,|\, \theta_i) \\ \theta_i \,|\, \mathbb{G} &\overset{\text{iid}}{\sim} \mathbb{G} \\ \mathbb{G} &\sim \Pi, \end{aligned}$$

where $\mathbb{G}$ is a random probability measure whose distribution Π is a DP or a GSB process. The density for one observation x_i reads

$$f(x_i) := f(x_i \,|\, \mathbb{G}) = \int_{\Theta} K(x_i \,|\, \theta)\mathbb{G}(\mathrm{d}\theta) = \sum_{k=1}^{\infty} w_k K(x_i \,|\, \theta_k). \tag{2.26}$$

Clearly, due to the infinite sum appearing in eq. (2.26) it is impossible to construct a Gibbs sampler with a finite number of updates. However, with the introduction of auxiliary variables a Gibbs sampler that needs only a finite number of summands can be devised. The main idea is to augment the state space of the random densities appearing in eq. (2.26), associating with each observation a clustering variable d_i and an almost surely finite random set $\mathbb{A}_i$, the slice set, such that the conditional distribution of x_i given the slice set attains a discrete uniform distribution

$$d_i \,|\, \mathbb{A}_i \sim \mathcal{DU}(\mathbb{A}_i).$$

The conditional density then becomes $f(d_i = k \,|\, \mathbb{A}_i) = |\mathbb{A}_i|^{-1}\mathcal{I}(k \in \mathbb{A}_i)$. For the clustering variables, the marginal distribution is $\sum_{k\geq 1} w_k \delta_k$. Conditionally on the slice set $\mathbb{A}_i$ the density of the observations becomes

$$\begin{aligned} f(x_i \,|\, \mathbb{A}_i) &= \sum_{k=1}^{\infty} f(x_i, d_i = k \,|\, \mathbb{A}_i) \\ &= \sum_{k=1}^{\infty} f(d_i = k \,|\, \mathbb{A}_i) f(x_i \,|\, d_i = k) = \sum_{k\in\mathbb{A}_i} \frac{1}{|\mathbb{A}_i|} K(x_i \,|\, \theta_k). \end{aligned} \tag{2.27}$$

Thus, with the introduction of strategic auxiliary random variables, the infinite sum becomes an almost surely finite and equally weighted mixture as is shown in eq. (2.27). In the following we present slice samplers in the case where Π is a DP or a GSB process.

2.6.1 Slice sampling DPM models

In the DPM model the weights w_k of the random distribution $\mathbb{G}$ are defined via the stick breaking representation of Sethuraman (1994). So letting w_k as in eq. (2.5) the DPM model in hierarchical representation is

$$\begin{aligned} K(x_i \,|\, \theta_i) &\overset{\text{ind}}{\sim} K(x_i \,|\, \theta_i) \\ \theta_i \,|\, \mathbb{G} &\overset{\text{iid}}{\sim} \mathbb{G} \\ \mathbb{G} &\sim \text{DP}(c, H), \end{aligned}$$

where $\mathbb{G} = \sum_{k\geq 1} w_k \delta_{\theta_j}$. The slice set depends on the sequence of weights $w^{(\infty)}$ through a (random) slice variable u (Walker, 2007) such that

$$\mathbb{A}_i = \{k \in \mathbb{N} : 0 < u_i < w_k\}.$$

Conditionally on the slice set $\mathbb{A}_i$ it is that

$$f(d_i \,|\, \mathbb{A}_i) = f(d_i = k \,|\, u_i) = \frac{\mathcal{I}(u_i < w_k)}{\sum_{s=1}^{\infty} \mathcal{I}(u_i < w_s)} = \frac{w_k \,\mathcal{U}(0, w_k)}{\sum_{s=1}^{\infty} w_s \,\mathcal{U}(0, w_s)}.$$

Therefore

$$u_i \,|\, w \sim \sum_{k=1}^{\infty} w_k \,\mathcal{U}(0, w_k) \text{ and } u_i \,|\, w, d_i = k \sim w_k \,\mathcal{U}(0, w_k).$$

From the joint density

$$f(u_i, d_i = k) = f(d_i = k) f(u_i \,|\, d_i = k) = w_k \,\mathcal{U}(u_i \,|\, 0, w_k),$$

and the fact that $f(x_i \,|\, d_i = k) = K(x_i \,|\, \theta_k)$, the (u_i, d_i)-augmented density is

$$\begin{aligned} f(x_i, u_i, d_i) &= f(d_i = k) f(u_i \,|\, d_i = k) f(x_i \,|\, d_i = k) \\ &= w_k \,\mathcal{U}(u_i \,|\, 0, w_k) K(x_i \,|\, \theta_k) \\ &= \mathcal{I}(u_i < w_k) K(x_i \,|\, \theta_k). \end{aligned}$$

The full likelihood based on a sample of size n is given by

$$f(x^{(n)}, u^{(n)}, d^{(n)} \,|\, w^{(\infty)}, \theta^{(\infty)}) = \prod_{i=1}^{n} \mathcal{I}(u_i < w_{d_i}) K(x_i \,|\, \theta_{d_i}). \tag{2.28}$$

The prior distribution on the mixture parameters is the distribution of the random probability measure G that is $\Pi(w^{(\infty)}, \theta^{(\infty)})$. Multiplied with the likelihood, the posterior model for inference is given by

$$\Pi(w^{(\infty)}, \theta^{(\infty)} \,|\, x^{(n)}, u^{(n)}, d^{(n)}) \propto \Pi(w^{(\infty)}, \theta^{(\infty)}) \prod_{i=1}^{n} \mathcal{I}(u_i < w_{d_i}) K(x_i \,|\, \theta_{d_i}). \tag{2.29}$$

Given the auxiliary variables $u^{(n)}, d^{(n)}$ it is possible to construct a Gibbs sampler with finite number of updates. At each sweep the latent variables (u_i, d_i) will be updated as also the parameters $w^{(\infty)}, \theta^{(\infty)}$ which are the quantities characterizing the density of interest. The algorithmic steps of the Gibbs sampler are given below. Having initialized the d_i's we sample the parameters of interest for $k = 1, \ldots, d^*$ where $d^* = \max_i\{d_i\}$.

1. The full conditional distribution of the locations is given by

$$\Pi(\theta_k \,|\, \cdots) \propto H(\theta_k) \prod_{d_i=k} K(x_i \,|\, \theta_k). \tag{2.30}$$

If there is no $d_i = k$ the θ_k's are sampled from the prior H.

2. The sampling of the z_k's and the slice variables u_i, $1 \leq i \leq n$ leading to a more efficient implementation (Kalli et al., 2011) can be done as a block. The full conditional distribution for the z_k's is given by

$$\Pi(z_k \,|\, \cdots, \text{exclude } u) = \mathcal{B}e\left(1 + \sum_{i=1}^{n} \mathcal{I}(d_i = k), c + \sum_{i=1}^{n} \mathcal{I}(d_i > k)\right). \tag{2.31}$$

Having sampled the z_k's the stick breaking weights are constructed via eq. (2.5).

3. Proceed by sampling the $\{u_i\}$ which are uniform on the interval $(0, w_{d_i})$

$$\Pi(u_i \,|\, \cdots) \propto \mathcal{I}(u_i < w_{d_i}). \tag{2.32}$$

4. The sampling of the clustering variables is from the discrete distribution

$$\Pi(d_i \,|\, \cdots) \propto K(x_i \,|\, \theta_{d_i}) \mathcal{I}(u_i < w_{d_i}). \tag{2.33}$$

In order to sample the d_i's exactly the explicit construction of the sets $\mathbb{A}_i = \{k \in \mathbb{N} : 0 < u_i < w_k\}$ is required. Let $N_i = \max_i \mathbb{A}_i$. In order to be sure that we have all the weights and locations for the algorithm to proceed we have to find for each i the smallest integer N such that

$$\sum_{k=1}^{N} w_k > 1 - u_i.$$

In fact we can be sure that we can sample all the d_i's when there is no $w_k > u_i$. So if we let $u^* = \min_i\{u_i\}$ we have to compute the smallest integer N^* such that

$$\sum_{k=1}^{N^*} w_k > 1 - u^*.$$

The additional weights $\{w_{d^*+1}, \ldots, w_{N^*}\}$ and locations $\{\theta_{d^*+1}, \ldots, \theta_{N^*}\}$ are sampled from their priors that is $\mathcal{B}e(1, c)$ and H respectively.

It is worth noting here that the number N^* is a random variable distributed as $1+\mathcal{P}oi(-c\log u^*)$ (Muliere & Tardella, 1998). To see this, note that N^* is defined as

$$\begin{aligned} N^* &= \inf\{n \in \mathbb{N} : \sum_{k=1}^{n} w_k > 1 - u^*\} \\ &= \inf\{n \in \mathbb{N} : 1 - \sum_{k=1}^{n} w_k < u^*\} \\ &= \inf\{n \in \mathbb{N} : \prod_{k=1}^{n}(1 - z_k) < u^*\}. \end{aligned}$$

Since $z_k \sim \mathcal{B}e(1, c)$ it follows that $1-z_k \sim \mathcal{B}e(c, 1)$ which implies that $-\log(1-z_k) \sim \mathcal{E}(c)$ where $\mathcal{E}(c)$ stands for the exponential distribution with rate c. Taking the quantity $-\log\prod_{k=1}^{n}(1-z_k)$, it is that

$$-\log\prod_{k=1}^{n}(1-z_k) = -\sum_{k=1}^{n}\log(1-z_k),$$

which is the sum of n exponential random variables with rate c. Thus $N^* - 1$ is the number of events of a Poisson process with mean c arriving at time $-\log u^*$.

5. Having updated the mixture allocation variables we proceed to the sampling of the concentration parameter c of the Dirichlet process. Following West (1992), we let κ to denote the number of unique labels of the clustering variables, that is $\kappa \in \{1, \ldots, n\}$. Then a sample for c can be obtained as follows

i. Sample $s \sim \mathcal{B}e(n+1, c)$ and then

ii. $c \,|\, s, \kappa \sim \rho_c\, \mathcal{G}(\alpha+\kappa, \beta - \log s) + (1-\rho_c)\, \mathcal{G}(\alpha+\kappa-1, \beta-\log s)$,

whith the weights ρ_c satisfying $\frac{\rho_c}{1-\rho_c} = \frac{\alpha+\kappa-1}{n(\beta-\log s)}$.

6. For density estimation purposes we have to sample from the predictive distributions given by

$$\Pi(\mathrm{d}x_{n+1} \,|\, \cdots) = \int_{\mathcal{P}_{\mathbb{X}}} \Pi(\mathrm{d}x_{n+1} \,|\, \mathbb{G})\Pi(\mathrm{d}\mathbb{G} \,|\, x_1, \ldots, x_n). \tag{2.34}$$

At each iteration of the Gibbs sampler we have points generated by the posterior random measure $G \,|\, x_1, \ldots, x_n$. These points are represented, at each iteration, by the posterior weights and locations (w^*, θ^*). Given those points we have to sample x_{n+1} from

$$x_{n+1} \sim \sum_{k=1}^{\infty} w_k^* K(\,\cdot \,|\, \theta_k^*). \tag{2.35}$$

We can estimate the density f by sampling a x_{n+1} given the current selection of parameters at each iteration of the Gibbs sampler. We sample the location $\theta_{n+1}^* = \theta_k^*$ using the weights. Generating a uniform u over the unit interval we take that θ_k^* for which $\sum_{j=1}^{k-1} w_j < u < \sum_{j=1}^{k}$. Even though we have not sampled all the weights, if we "run out" of weights we merely take the θ_{n+1}^* from the prior. Finally the predictive x value comes from $K(\,\cdot \,|\, \theta_{n+1}^*)$.

2.6.2 Geometric slice sampling GSBM models

The idea behind the definiton of the geometric stick breaking mixture model (Fuentes-García et al., 2010) which can be represented in a hierarchical fashion as

$$\begin{aligned} K(x_i \,|\, \theta_i) &\overset{\text{ind}}{\sim} K(x_i \,|\, \theta_i) \\ \theta_i \,|\, \mathbb{G} &\overset{\text{iid}}{\sim} \mathbb{G} \\ \mathbb{G} &\sim \text{GSB}(\lambda, H), \end{aligned}$$

where $\mathbb{G} = \sum_{k\geq 1} w_k \delta_{\theta_j}$ and $w_k = \lambda(1-\lambda)^{k-1}$ i.e. the geometric weights, is to construct a simple Gibbs sampler such that the slice sets don't need to have gaps. In contrast they are sequential.

In GSBM mixture models, to overcome the difficulties with the infinite mixture an auxiliary discrete random variable N_i is introduced for each observation, such that conditionally on N_i the clustering variable d_i will be a choice of the geometric slice set

$$\mathbb{B}_i = \{1, \ldots, N_i\}.$$

The random variable N_i is almost surely finite with distribution f_N that possibly depends on parameters. Then, given N_i the clustering variable attains a discrete uniform distribution over the elements of $\mathbb{B}_i$

$$f(d_i = k \,|\, N_i = l) = f(d_i = k \,|\, \mathbb{B}_i) = \frac{\mathcal{I}(k \in \mathbb{B}_i)}{\sum_{s=1}^{\infty} \mathcal{I}(s \in \mathbb{B}_i)} = l^{-1}\mathcal{I}(k \leq l).$$

The (d_i, N_i)-augmented density becomes

$$\begin{aligned} f(x_i, d_i = k, N_i = l) &= f_N(N_i = l) f(d_i = k \,|\, N_i = l) f(x_i \,|\, d_i = k) \\ &= f_N(N_i = l) l^{-1} \mathcal{I}(k \leq l) K(x_i \,|\, \theta_k). \end{aligned} \tag{2.36}$$

Now it is the weights that depend on the choice of the masses f_N. Marginalizing the random density in eq. (2.36) with respect to (N_i, d_i), we obtain

$$f(x_i) = \sum_{k=1}^{\infty} \sum_{l=k}^{\infty} f_N(N_i = l) l^{-1} K(x_i \,|\, \theta_k) = \sum_{k=1}^{\infty} w_k K(x_i \,|\, \theta_k),$$

with $w_k = \sum_{l=k}^{\infty} l^{-1} f_N(N_i = l)$. It is known (Fuentes-García et al., 2010) that in the particular case where the masses of N_i's are coming from the negative binomial distribution

$$\mathcal{NB}(l \,|\, 2, \lambda) = l\lambda^2(1-\lambda)^{l-1}\mathcal{I}(l \geq 1),$$

the weights w_k for $k \geq 1$ have the form:

$$w_k = \mathcal{NB}(k \,|\, 1, \lambda) = \lambda(1-\lambda)^{k-1}. \tag{2.37}$$

Thus we recover a geometric stick breaking mixture model meaning that the augmentation is valid. Substituting in eq. (2.36) $f_N(N_i = l) = \mathcal{NB}(2, \lambda)$ the likelihood based on a sample of size n coming from f is

$$f(x^{(n)}, N^{(n)}, d^{(n)}) = \prod_{i=1}^{n} \lambda^2 (1-\lambda)^{N_i - 1} \mathcal{I}(d_i \leq N_i) K(x_i \,|\, \theta_{d_i}). \tag{2.38}$$

Multiplying the likelihood with the prior $\Pi(w^{(\infty)}, \theta^{(\infty)})$ the posterior distribution is

$$\Pi(w^{(\infty)}, \theta^{(\infty)}, N^{(n)}, d^{(n)} \,|\, x^{(n)}) \propto \Pi(w^{(\infty)}, \theta^{(\infty)}) \prod_{i=1}^{n} \lambda^2 (1-\lambda)^{N_i - 1} \mathcal{I}(d_i \leq N_i) K(x_i \,|\, \theta_{d_i}). \tag{2.39}$$

Below we provide the Gibbs sampling algorithmic steps for posterior inference. Having initialized the $d^{(n)}, N^{(n)}$ we sample the parameters of interest for $k = 1, \ldots, N^*$ where $N^* = \max_i\{N_i\}$.

1. The full conditional for the geometric probability λ is under the Beta conjugate prior

$$\Pi(\lambda \,|\, \cdots) = \left\{ \prod_{i=1}^{n} \lambda^2 (1-\lambda)^{N_i - 1} \right\} \lambda^{\alpha - 1} (1-\lambda)^{\beta - 1} = \mathcal{B}e\left(\alpha + 2\, n, \beta + \sum_{i=1}^{n} N_i - n \right). \tag{2.40}$$

Having updated λ, we construct the geometric weights w_k for $1 \leq j \leq N^*$ via eq. (2.37).

2. The full conditional distribution of the locations is given by

$$\Pi(\theta_k \,|\, \cdots) \propto H(\theta_k) \prod_{d_i = k} K(x_i \,|\, \theta_k). \tag{2.41}$$

If there is no $d_i = k$ the θ_k's are sampled from the prior H.

3. We then sample the infinite mixture allocation variables d_i for $i = 1, \ldots, n$. It is that

$$\Pi(d_i = k \,|\, \cdots) \;\propto\; K(x_i \,|\, \theta_k) \mathcal{I}(k \leq N_i). \tag{2.42}$$

4. Next, to construct the sequential slice sets $\mathbb{A}_i$ for $1 \leq i \leq n_T$ we have to sample N_i from

$$\Pi(N_i = l \,|\, d_i = k, \cdots) \;\propto\; (1-p)^l \mathcal{I}(l \geq k), \tag{2.43}$$

which is a truncated geometric distribution over the set $\{k, k+1, \ldots\}$.

5. The density estimation step can be done in a similar manner with the estimation step **6.** for the DPM model. We sample x_{n+1} from eqs. (2.34) and (2.35).

As we will see in Chapter 4 these algorithms can be adopted for posterior inference in the case of dependent nonparametric priors when DP or GSB measures are used for the modeling of dependent random density functions. For the HDP and its extension the basic method for inference relies on the Polya urn representation i.e. marginal samplers but recently, distributed slice sampling algorithms (Ge et al., 2015) have been also proposed.

Chapter 3

Bayesian Nonparametric Reconstruction Models

3.1 Introduction

This chapter is devoted to Bayesian nonparametric models for reconstruction and prediction of random dynamical systems. In section 3.2 we formulate the model under the assumption that the noise density is a mixture of a parametric family with mixing measure, a general discrete random distribution.

In section 3.3 we show how the augmentation of the densities with non–sequential slice sets enables the *Dirichlet Process Reconstruction* (DPR) model first introduced in Hatjispyros et al. (2009). We extend the DPR model to a fully stochastic version namely the randomized Dirichlet Process Reconstruction (rDPR) model randomizing the concentration parameter of the associated DP measure. We also propose an antelrnative augmentation scheme for the nonstandard part, avoiding the representation of normal transition density of the observations as a Gamma mixture of uniforms.

Augmenting with auxiliary random variables which force the associated slice sets to have no–gaps, in section 3.4 we introduce the *Geometric Stick-Breaking Reconstruction* (GSBR) model proposed in Merkatas et al. (2017). In section 3.5 the perfomance of our GSBR model against the DPR model in simulated examples in reconstruction and prediction problems of chaotic time series generated by a cubic map is illustrated. To demonstrate the need for nonparametric models, we compare the nonparametric models with a simple parametric Gibbs sampler that assumes Gaussian noise. Finally, the chapter ends with some comments on the methods and directions for future research.

3.2 Building the inferential models

We consider the following random dynamical model given by

$$x_i = T(\vartheta, x_{i-1}, z_i) = g(\vartheta, x_{i-1}) + z_i, \quad i \geq 1, \tag{3.1}$$

where $g : \Theta \times \mathbb{X} \to \mathbb{X}$, for some compact subset $\mathbb{X}$ of $\mathbb{R}$, $(x_i)_{i\geq 0}$ and $(z_i)_{i\geq 1}$ are real random variables over some probability space $(\Omega, \mathcal{F}, P)$; the set Θ denotes the parameter space and g is nonlinear, and for simplicity, continuous in x_{i-1}. We assume that the random variables z_i are independent to each other, and independent of the states x_i.

In addition we assume that the additive perturbations z_i are identically distributed from a zero mean distribution with unknown density f defined over the real line, so that $T : \Theta \times \mathbb{X} \times \mathbb{R} \to \mathbb{R}$. We assume that there is no observational noise, so that we have at our disposal a time series $x^{(n)} = (x_1, \ldots, x_n)$ generated by the Markovian process defined in eq. (3.1). The time series $x^{(n)}$ depends solely on the initial distribution of x_0, the vector of parameters ϑ, and the particular realization of the noise process.

We model the errors in recurrence eq. (3.1) as a mixture of normal kernels of the form $\mathcal{N}(x \,|\, 0, \tau^{-1})$ with mean zero and precision τ and mixing measure, a general discrete random distribution $\mathbb{G} = \sum_{j\geq 1} \pi_j \, \delta_{\tau_j}$; then letting $\tau = (\tau_j)_{j\geq 1}$ and $\pi = (\pi_j)_{j\geq 1}$ we have

$$f_{\pi,\tau}(x) = \int_{\tau>0} \mathcal{N}(x \,|\, 0, \tau^{-1}) \mathbb{G}(\mathrm{d}\tau) = \sum_{j=1}^{\infty} \pi_j \, \mathcal{N}(x \,|\, 0, \tau_j^{-1}).$$

For the observations $(x^{(n)} \,|\, x_0)$ and for $1 \leq i \leq n$ we have the transition kernel

$$f_{\pi,\tau}(x_i \,|\, x_{i-1}, \vartheta) = \sum_{j=1}^{\infty} \pi_j \, \mathcal{N}(x_i \,|\, g(\vartheta, x_{i-1}), \tau_j^{-1}), \quad 1 \leq i \leq n, \tag{3.2}$$

and associated data likelihood

$$f_{\pi,\tau}\left(x_1, \ldots, x_n \,|\, x_0, \vartheta\right) = \prod_{i=1}^{n} \sum_{j=1}^{\infty} \pi_j \, \mathcal{N}(x_i \,|\, g(\vartheta, x_{i-1}), \tau_j^{-1}). \tag{3.3}$$

As it has been pointed out in Hatjispyros et al. (2009), a straightforward application of Gibbs sampling ideas, for sampling from the posterior distribution $f(\vartheta, x_0 \,|\, x_1, \ldots, x_n)$, is not possible due to the following two facts:

1. We have to sample from a mixture with an infinite number of components.
2. Full conditionals are of non-standard form.

For example, after assigning to the initial condition x_0 a uniform prior over the compact set $\mathbb{X}$, the full conditional for x_0 reads

$$f_{\pi,\tau}(x_0 \,|\, \cdots) \propto \sum_{j=1}^{\infty} \pi_j \, \left\{ \mathcal{I}(x_0 \in \mathbb{X}) \, \mathcal{N}(x_1 \,|\, g(\vartheta, x_0), \tau_j^{-1}) \right\},$$

where $\mathcal{I}(x_0 \in \mathbb{X})$ is the indicator function which equals 1 whenever x_0 is in the set $\mathbb{X}$ and 0 otherwise. Then the full conditional for x_0, whenever g is nonlinear in x_0, is an infinite mixture of truncated non-standard densities.

3.2.1 Dynamical Slice Sets

Due to the infinite mixture appearing in the product of the likelihood in the equation above, we are not able to construct Gibbs samplers of finite dimensions.

To make the number of variables that we have to sample finite, we use slice sampling techniques for infinite mixtures. For each observation x_i, we introduce the pair $(d_i, \mathbb{A}_i)$. The d_i are the clustering variables and indicate the component of the infinite mixture the observation x_i came from. The set $\mathbb{A}_i$ is the associated random slice set and is an almost surely finite set of indices. Notice, that, marginally, $d_i \,|\, \pi \sim \sum_{j\geq 1} \pi_j \delta_j$ and the variables d_i have an infinite state space.

Our aim is to have $x_i \,|\, \tau, \mathbb{A}_i$ coming from a finite mixture of normal kernels. Letting the random variable d_i conditionally on the event $\{d_i \in \mathbb{A}_i\}$ attain a discrete uniform distribution, over $\mathbb{A}_i$; that is

$$f(d_i \,|\, \mathbb{A}_i) = |\mathbb{A}_i|^{-1} \mathcal{I}(j \in \mathbb{A}_i),$$

we obtain

$$\begin{aligned} f_\tau(x_i \,|\, \mathbb{A}_i) &= \sum_{j=1}^{\infty} f(x_i, d_i = j \,|\, \mathbb{A}_i) \\ &= \sum_{j=1}^{\infty} f(d_i = j \,|\, \mathbb{A}_i) f_\tau(x_i \,|\, d_i = j) = \sum_{j \in \mathbb{A}_i} |\mathbb{A}_i|^{-1} \mathcal{N}(x_i \,|\, 0, \tau_j^{-1}). \end{aligned}$$

where $|A_i|$ denotes the cardinality of the set A_i. Thus, given the precisions τ and the slice set $\mathbb{A}_i$, the observation x_i comes from an equally weighted almost surely finite mixture of normal kernels.

Selecting specific forms for the slice sets, we can obtain different reconstruction models. In the following two sections we select $\mathbb{A}_i$ in such way that allows us to recover the DPR and the GSBR models respectively.

3.3 Dirichlet process reconstruction model

The DPR model is obtained as a special case of the general reconstruction model if we define the slice sets to be non–sequential. That is, we assign to each observation x_i a slice set that depends on the weights π via a random variable u_i such that

$$f_\pi(d_i = j \,|\, u_i) = f(d_i = j \,|\, \mathbb{A}_i) \;\text{ with }\; \mathbb{A}_i = \{j \in \mathbb{N} : 0 < u_i < \pi_j\},$$

as proposed in the slice sampler for the DPM model by Walker (2007) and

$$f_\pi(d_i = j \,|\, u_i) = \frac{\mathcal{I}(j \in \mathbb{A}_i)}{\sum_{s=1}^{\infty} \mathcal{I}(s \in \mathbb{A}_i)} = \frac{\mathcal{I}(u_i < \pi_j)}{\sum_{s=1}^{\infty} \mathcal{I}(u_i < \pi_s)} = \frac{\pi_j \mathcal{U}(u_i \,|\, 0, \pi_j)}{\sum_{s=1}^{\infty} \pi_s \mathcal{U}(u_i \,|\, 0, \pi_s)},$$

where $\mathcal{U}(x \,|\, a, b)$ is the uniform density over the interval (a, b). Therefore

$$u_i \,|\, \pi \sim \sum_{j=1}^{\infty} \pi_j \, \mathcal{U}(0, \pi_j) \;\text{ and }\; u_i \,|\, \pi, d_i = j \sim \mathcal{U}(0, \pi_j),$$

and from the joint $f_\pi(u_i, d_i = j) = \pi_j \mathcal{U}(u_i \,|\, 0, \pi_j)$ and the fact that given a particular value for d_i we have $f(x_i \,|\, d_i = j) = \mathcal{N}(x_i \,|\, 0, \tau_j^{-1})$ we obtain the augmented random densities

$$f_{\pi,\tau}(x_i, u_i, d_i = j) \;=\; \pi_j \, \mathcal{U}(u_i \,|\, 0, \pi_j) \, \mathcal{N}(x_i \,|\, 0, \tau_j^{-1}). \tag{3.4}$$

From eqs. (3.2) and (3.4) and letting $\pi_j = w_j$, where w_j are the weights in the stick breaking representation of the Dirichlet process, that is $w_1 = z_1$ and for $j > 1$:

$$w_j = z_j \prod_{s<j} (1 - z_s), \tag{3.5}$$

with z_j drawn i.i.d. from the beta distribution $\mathcal{B}e(1, c)$ for some $c > 0$, we have

$$f_{w,\tau}(x_i, u_i, d_i = j \,|\, x_{i-1}, \vartheta) = w_j \, \mathcal{U}(u_i \,|\, 0, w_j) \, \mathcal{N}(x_i \,|\, g(\vartheta, x_{i-1}), \tau_j^{-1}). \tag{3.6}$$

In a hierarchical fashion using the slice variables u_i and the stick-breaking representation we have for $i = 1, \ldots, n$ and $j \geq 1$:

$$\begin{aligned}
&(x_i \,|\, x_{i-1}, d_i = j, \theta, \tau) \overset{\text{ind}}{\sim} \mathcal{N}(x_i \,|\, g(\vartheta, x_{i-1}), \tau_j^{-1}) \\
&(u_i \,|\, d_i = j, w) \overset{\text{ind}}{\sim} \mathcal{U}(0, w_j) \\
&\Pr(d_i = j \,|\, w) = w_j \\
&w_j = z_j \textstyle\prod_{s<j}(1 - z_s), \;\; z_j \overset{\text{iid}}{\sim} \mathcal{B}e(1, c) \\
&c \sim \mathcal{G}(\alpha, \beta), \;\; \tau_j \overset{\text{iid}}{\sim} P_0.
\end{aligned}$$

Then given ϑ, x_0 and c the data likelihood based on a sample of size n is given by

$$\begin{aligned}
f_{w,\tau}(x_i, u_i, d_i;\, 1 \leq i \leq n \,|\, \vartheta, x_0, c) \propto \prod_{i=1}^{n} \mathcal{I}(u_i < w_{d_i}) \, \tau_{d_i}^{1/2} \\
\times \exp\left\{ -\frac{\tau_{d_i}}{2} h_\vartheta(x_i, x_{i-1}) \right\},
\end{aligned} \tag{3.7}$$

where $h_\vartheta(x_i, x_{i-1}) = (x_i - g(\vartheta, x_{i-1}))^2$.

Note that in eq. (3.7) the problem with the infinite mixture has been eliminated. The DPR model described here, slightly differs from the model introduced by Hatjispyros et al. (2009). First of all, we let the concentration parameter to be random in contrast with Hatjispyros et al. (2009). The second, and more important, modification is that we do not make the same effort with the nonlinear form of the means of the normal distributions appearing in eq. (3.7).

In their approach the problem with nonlinear map has been dealt with the introduction of an auxiliary variable v_i for each observation x_i for $1 \leq i \leq n_T$ defined as

$$\begin{aligned} v_j \,|\, \tau_j &\overset{\text{ind}}{\sim} \mathcal{G}(3/2, \tau_j/2), \\ x_i \,|\, x_{i-1}, v_i, \theta &\overset{\text{ind}}{\sim} \mathcal{U}(g(\vartheta, x_{i-1}) - \sqrt{v_i}, g(\vartheta, x_{i-1}) + \sqrt{v_i}), \end{aligned}$$

they wrote the normal distribution as a gamma mixture of uniforms, resulting to the following likelihood for the DPR model:

$$\begin{aligned} f_{w,\tau}(x_i, u_i, d_i, v_i;\, 1 \leq i \leq n \,|\, \vartheta, x_0, c) \propto \prod_{i=1}^{n} \mathcal{I}(u_i < w_{d_i})\, \tau_{d_i}^{3/2} \\ \times\, e^{-\frac{v_i \tau_{d_i}}{2}} \mathcal{I}(v_i > h_\vartheta(x_i, x_{i-1})). \end{aligned}$$

This approach has the advantage that all the distributions, which have to be sampled, are essentially mixtures of uniforms. This may lead though to bigger execution times if the sample size of the time series is large because the sampler has to sweep at each iteration over all the auxiliary variables v_i. As we will see later, the problem with nonlinear map can be solved with embedded Gibbs samplers augmenting the state space with a number of variables equal to the length of the vector parameter ϑ which usually, in applications, is much smaller than the sample size.

3.3.1 Extending the DPR model for prediction

In this section we describe how the DPR model can be extended for prediction purposes. In this case the problem is defined as follows. Given an observed time series $x^{(n)} = (x_1, \ldots, x_n)$, and a prediction horizon $T > n$, the aim of prediction is to obtain an estimate of the future unobserved values $(x_{n+1}, \ldots, x_{n+T})$.

Letting $n_T = n + T$, we can extend the DPR model for prediction with the introduction of the random variables $(x_{n+1}, \ldots, x_{n+T})$, and obtain the likelihood

$$\begin{aligned} f(x_i, u_i, d_i\,;\, 1 \leq i \leq n_T \,|\, \vartheta, x_0, c) \propto \prod_{i=1}^{n_T} \mathcal{I}(u_i < w_{d_i}) \tau_{d_i}^{1/2} \\ \times \exp\left\{ -\frac{\tau_{d_i}}{2} h_\vartheta(x_i, x_{i-1}) \right\}. \end{aligned} \tag{3.8}$$

In the Bayesian setting, prior distributions for these parameters must be assigned and the estimators are taken from their posterior distribution. In the next section we describe an MCMC based algorithm for the randomized DPR model. The corresponding prior distributions for the future unoberved values will be set constant, that is

$$\pi(x_{n+i}) \propto 1, \quad i = 1, \ldots, T.$$

3.3.2 Slice sampler for the rDPR model

In this section we describe an MCMC algorithm for estimating the model based on slice sampling. Specifically we are interested in the variables (x_0, ϑ) and the future unobserved values of $x_{n+1}, \ldots, x_{n+T}$. We complete the model by assigning uniform priors on the parameters of interest.

In particular for the initial condition x_0 we assign a uniform prior distribution over the set $\tilde{\mathbb{X}} \subseteq \mathbb{R}$, which represents our prior knowledge for the state space of the dynamical model given in eq. (3.1). Over the vector control parameters of the system ϑ we assume a uniform prior over the set $\tilde{\Theta}$ of the parameter space $\mathbb{R}^k$. For the Dirichlet random measure $\mathbb{P} \sim \mathrm{DP}(c, P_0)$, we assume for the base measure a Gamma distribution, namely $P_0(\mathrm{d}\tau) = \mathcal{G}(\tau \,|\, a, b)\mathrm{d}\tau$. Finally, the concentration parameter c attains a Gamma prior $\mathcal{G}(\alpha, \beta)$, and will be updated with the standard sampling scheme proposed by West (1992).

After initializing the variables d_i for $i = 1, \ldots, n_T$ and the variables c, x_0 and ϑ, at each iteration, we will sample the variables:

$$(\tau_j),\ 1 \leq j \leq N^*, \quad d_i,\ 1 \leq i \leq n_T,$$

and

$$(\vartheta,\ x_0,\ c,\ z_{n_T+1}),$$

with $N = \max_{1 \leq i \leq n_T} d_i$.

1. At first, given the clustering variables d_i, $i = 1, \ldots, n_T$, we update the stick–breaking weights. We update the z_j–s from

$$f(z_j \,|\, \cdots) = \mathcal{B}e\left(1 + \sum_{i=1}^{n_T} \mathcal{I}(d_i = j), c + \sum_{i=1}^{n_T} \mathcal{I}(d_i > j)\right), \tag{3.9}$$

for $1 \leq j \leq N$. Then the updated weights $(w_j)_{j \geq 1}$ are constructed via the stick–breaking representation.

2. Having the updated weights we can proceed to the sampling of the slice variables u_i, for $i = 1, \ldots, n_T$ which are uniform distributions on the interval $(0, w_{d_i})$, namely

$$f(u_i \,|\, \cdots) \propto \mathcal{I}(u_i < w_{d_i}). \tag{3.10}$$

3. We then sample the precisions τ_j for $j = 1, \ldots, N$ and $N = \max_{1 \leq i \leq n_T} d_i$. We have that

$$f(\tau_j \,|\, \cdots) = \mathcal{G}\left(a + \frac{1}{2}\sum_{i=1}^{n_T} \mathcal{I}(d_i = j),\ b + \frac{1}{2}\sum_{i=1}^{n_T} \mathcal{I}(d_i = j)\, h_\vartheta(x_i, x_{i-1})\right), \tag{3.11}$$

If $j > N$ we sample the additional τ_j's from the prior $\mathcal{G}(a, b)$. In the next step of the algorithm, the additional number of weights and precisions is obtained.

4. The additional number of weights and precision can be found by letting $u^* = \min_{1\leq i\leq n_T}\{u_i\}$ and find the smallest integer N^* for which

$$\sum_{j=1}^{N^*} w_j > 1 - u^*. \tag{3.12}$$

This is necessary in order to sample the mixture allocation variables exactly (See step **(4)** in section 2.6.1).

5. We then sample the infinite mixture allocation variables d_i for $i = 1, \ldots, n_T$. It is that

$$\Pr(d_i = j \,|\, \cdots) \propto \tau_j^{1/2} \exp\left\{-\frac{\tau_j}{2} h_\vartheta(x_i, x_{i-1})\right\} \mathcal{I}(j \in \mathbb{A}_i). \tag{3.13}$$

6. Having updated the mixture allocation variables we proceed to the sampling of the concentration parameter c of the Dirichlet process. Following West (1992), we let κ to denote the number of unique labels of the clustering variables, that is $\kappa \in \{1, \ldots, n_T\}$. Then a sample for c can be obtained as follows

i. Sample $s \sim \mathcal{B}e(n_T + 1, c)$ and then

ii. $c \,|\, s, \kappa \sim \rho_c \, \mathcal{G}(\alpha + \kappa, \beta - \log c) + (1 - \rho_c)\, \mathcal{G}(\alpha + \kappa - 1, \beta - \log c)$,

whith the weights ρ_c satisfying $\frac{\rho_c}{1-\rho_c} = \frac{\alpha+\kappa-1}{n_T(\beta-\log c)}$.

7. We are now ready to sample z_{n+1} from the noise predictive $f(z_{n+1} \,|\, x_1, \ldots, x_n)$. At each iteration of the Gibbs sampler we have updated weights $(\pi_j)_{1\leq j\leq N^*}$ and precisions $(\tau_j)_{1\leq j\leq N^*}$ and we sample independently $\rho \sim \mathcal{U}(0, 1)$. Then we take the τ_j with $1 \leq j \leq N^*$ satisfying

$$\sum_{i=0}^{j-1} \pi_i < \rho \leq \sum_{i=0}^{j} \pi_i, \;\; \pi_0 = 0.$$

If $\rho > \sum_{i=0}^{N^*} \pi_i$, we sample τ_j from the prior $\mathcal{G}(a, b)$. In any case we sample z_{n+1} from the normal kernel $\mathcal{N}(0, \tau_j^{-1})$.

8. The full conditional for x_0, will be

$$f(x_0 \,|\, \cdots) \propto \mathcal{I}(x_0 \in \tilde{\mathbb{X}}) \exp\left\{-\frac{\tau_{d_1}}{2} h_\vartheta(x_1, x_0)\right\}. \tag{3.14}$$

9. For the vector of parameters ϑ, the full conditional becomes

$$f(\vartheta \,|\, \cdots) \propto \mathcal{I}(\vartheta \in \tilde{\Theta}) \exp\left\{-\frac{1}{2} \sum_{i=1}^{n_T} \tau_{d_i} h_\vartheta(x_i, x_{i-1})\right\}. \tag{3.15}$$

10. The full conditional densities for the future unobserved observations, when $T \geq 2$ and for $j = 1, \ldots, T-1$, are given by

$$f(x_{n+j} \,|\, \cdots) \propto \exp\left\{-\frac{1}{2}\left[\tau_{d_{n+j}} h_\vartheta(x_{n+j}, x_{n+j-1}) + \tau_{d_{n+j+1}} h_\vartheta(x_{n+j+1}, x_{n+j})\right]\right\}. \tag{3.16}$$

For $j = T$ the full conditional is normal with mean $g(\vartheta, x_{n+T-1})$ and variance $\tau^{-1}_{d_{n+T}}$, that is

$$f(x_{n+T} \,|\, \cdots) = \mathcal{N}\left(x_{n+T} \,|\, g(\vartheta, x_{n+T-1}), \tau^{-1}_{d_{n+T}}\right). \tag{3.17}$$

3.4 Geometric stick–breaking reconstruction model

The GSBR model is derived from the generic reconstruction model if we use sequential slice sets of the form $\mathbb{A}_i = \{1, \ldots, N_i\}$, as proposed in Fuentes-García et al. (2010). Then the cluster allocation variables given the N_i attain a discrete uniform distribution over the elements of $\mathbb{A}_i$, that is

$$f(d_i = j \,|\, N_i) = f(d_i = j \,|\, \mathbb{A}_i) \;\text{ with }\; \mathbb{A}_i = \{1, \ldots, N_i\},$$

and

$$f(d_i = j \,|\, N_i = r) = \frac{\mathcal{I}(j \in \mathbb{A}_i)}{\sum_{s=1}^{\infty} \mathcal{I}(s \in \mathbb{A}_i)} = \frac{1}{l}\,\mathcal{I}(j \leq l),$$

and N_i is an almost surely finite discrete random variable of mass f_N, that possibly depends on parameters. In this case, from the joint

$$\begin{aligned} f_\pi(d_i = j, N_i = r) &= f_N(N_i = r) f(d_i = j \,|\, N_i = r) \\ &= f_N(N_i = r)\, r^{-1} \mathcal{I}(j \leq r) \end{aligned} \tag{3.18}$$

and the fact that $f(x_i \,|\, d_i = j) = \mathcal{N}(x_i \,|\, 0, \tau_j^{-1})$, the (d_i, N_i) augmented densities become

$$f_\tau(x_i, N_i = r, d_i = j) = f_N(N_i = r)\, r^{-1}\, \mathcal{I}(j \leq r)\, \mathcal{N}(x_i \,|\, 0, \tau_j^{-1}). \tag{3.19}$$

Now it is the weights that depend on the choice of the masses f_N. Marginalizing the random density in eq. (3.19) with respect to (N_i, d_i), we obtain

$$f_\tau(x_i) = \sum_{j=1}^{\infty} \pi_j\, \mathcal{N}(x_i \,|\, 0, \tau_j^{-1}), \;\text{ with, }\; \pi_j = \sum_{r=j}^{\infty} r^{-1} f_N(N_i = r).$$

It is known (Fuentes-García et al., 2010) that, in the particular case where the masses of N_i's are coming from the negative binomial distribution with state space $\{1, 2, \ldots\}$, namely

$$\mathcal{NB}(r \,|\, 2, \lambda) = r\lambda^2 (1-\lambda)^{r-1} \mathcal{I}(r \geq 1),$$

the weights π_j for $j \geq 1$ have the form:

$$\pi_j = \mathcal{NB}(j \,|\, 1, \lambda) = \lambda(1-\lambda)^{j-1}. \tag{3.20}$$

Note that the randomness included in the infinite number of weigths in the DPR model is now replaced by only one random variable $\lambda \sim \mathcal{B}e(a,b)$. It is the decreasing nature of the geometric weights that will lead to simpler Gibbs samplers than the associated sampler of the DP counterpart model described in the previous section. Having the weights ordered, it is not necessary to perform a complete search in the vector where the weights are stored and thus the execution time of the GSBR model is, as we will see, smaller.

In order to make the geometric slice sampling steps described in the next section clearer, as well as the dependencies, we write the model in a hierarchical fashion. Using the slice variables N_i we have for $i = 1, \ldots, n$ and $j \geq 1$:

$$\begin{aligned}
&(x_i \,|\, x_{i-1}, \theta, d_i = j, \tau) \overset{\text{ind}}{\sim} \mathcal{N}(x_i \,|\, g(\vartheta, x_{i-1}), \tau_j^{-1}) \\
&(d_i \,|\, N_i = r) \overset{\text{ind}}{\sim} \mathcal{DU}\{1, \ldots, r\} \\
&\pi_j = \mathcal{NB}(j \,|\, 1, \lambda),\ N_i \overset{\text{iid}}{\sim} \mathcal{NB}(2, \lambda) \\
&\tau_j \overset{\text{iid}}{\sim} P_0,
\end{aligned}$$

where $\mathcal{DU}\{1, \ldots, r\}$ denotes the discrete uniform mass over the set $\{1, \ldots, r\}$. Therefore, the data likelihood based on a sample of size n, given ϑ, x_0 and λ is seen to be

$$\begin{aligned}
f_\tau(x_i, N_i, d_i;\, 1 \leq i \leq n \,|\, \vartheta, x_0, \lambda) \propto \prod_{i=1}^{n} \lambda^2 (1-\lambda)^{N_i - 1} \mathcal{I}(d_i \leq N_i)\, \tau_{d_i}^{1/2} \\
\times \exp\left\{ -\frac{\tau_{d_i}}{2} h_\vartheta(x_i, x_{i-1}) \right\}. \qquad (3.21)
\end{aligned}$$

3.4.1 Extending the GSBR model for prediction

In a similar way, we can extend the GSBR model as we have done with the DPR model. After the introduction of the additional random variables $(x_{n+1}, \ldots, x_{n+T})$, the likelihood of the GSBR model for prediction becomes

$$\begin{aligned}
f(x_i, d_i, N_i\,;\, 1 \leq i \leq n_T \,|\, \vartheta, x_0, \lambda) \propto \prod_{i=1}^{n_T} \lambda^2 (1-\lambda)^{N_i - 1} \mathcal{I}(d_i \leq N_i)\, \tau_{d_i}^{1/2} \\
\times \exp\left\{ -\frac{\tau_{d_i}}{2} h_\vartheta(x_i, x_{i-1}) \right\}. \qquad (3.22)
\end{aligned}$$

For the associated slice sampler which is now described in section 3.4.2 we set the prior distributions for the future unobserved values to be constant

$$\pi(x_{n+i}) \propto 1, \quad i = 1, \ldots, T.$$

3.4.2 Slice sampler for the GSBR model

In this section, we describe the MCMC algorithm based on slice sampling for inference with the GSBR model. We set as a base measure of the GSB process $\mathbb{P}$, a Gamma distribution, that

is $P_0(\mathrm{d}\tau) = \mathcal{G}(\tau \,|\, a, b)\mathrm{d}\tau$, and the geometric probability attains a Beta conjugate prior, that is $\lambda \sim \mathcal{B}e(\alpha, \beta)$.

Having completed the model, we are now ready to describe the Gibbs sampler and the full conditional densities for estimating the GSBR model. After initializing the variables d_i, N_i for $i = 1, \ldots, n_T$ and the variables λ, x_0 and ϑ, at each iteration we will sample the variables:

$$(\tau_j),\ 1 \leq j \leq N^*, \quad (d_i, N_i),\ 1 \leq i \leq n_T,$$

and

$$(\vartheta,\ x_0,\ \lambda,\ z_{n_T+1}),$$

with $N^* = \max_{1 \leq i \leq n_T} N_i$.

1. The full conditional for the geometric probability λ is under the Beta conjugate prior

$$f(\lambda \,|\, \cdots) = \mathcal{B}e\left(\alpha + 2\, n_T, \beta + \sum_{i=1}^{n_T} N_i - n_T\right), \tag{3.23}$$

Having updated λ, we construct the geometric weights π_j for $1 \leq j \leq N^*$ via eq. (3.20).

3. We then sample the precisions τ_j for $j = 1, \ldots, N^*$ and $N^* = \max_{1 \leq i \leq n_T} N_i$. We have that

$$f(\tau_j \,|\, \cdots) = \mathcal{G}\left(a + \frac{1}{2}\sum_{i=1}^{n_T} \mathcal{I}(d_i = j),\ b + \frac{1}{2}\sum_{i=1}^{n_T} \mathcal{I}(d_i = j)\, h_\vartheta(x_i, x_{i-1})\right), \tag{3.24}$$

where the expression $f(\tau_j \,|\, \cdots)$ denotes the density of τ_j conditional on the rest of the variables.

4. We then sample the infinite mixture allocation variables d_i for $i = 1, \ldots, n_T$. It is that

$$\Pr(d_i = j \,|\, N_i, \cdots) \ \propto\ \tau_j^{1/2} \exp\left\{-\frac{\tau_j}{2} h_\vartheta(x_i, x_{i-1})\right\} \mathcal{I}(j \leq N_i). \tag{3.25}$$

5. Next, to construct the sequential slice sets $\mathbb{A}_i$ for $1 \leq i \leq n_T$ we have to sample N_i from

$$\Pr(N_i = r \,|\, d_i = j, \cdots) \ \propto\ (1 - \lambda)^r\, \mathcal{I}(j \leq r), \tag{3.26}$$

which is a truncated geometric distribution over the set $\{j, j+1, \ldots\}$.

6. In this step sample z_{n+1} from the noise predictive $f(z_{n+1} \,|\, x_1, \ldots, x_n)$. At each iteration of the Gibbs sampler, we have updated weights $(\pi_j)_{1 \leq j \leq N^*}$ and precisions $(\tau_j)_{1 \leq j \leq N^*}$ and we sample independently $\rho \sim \mathcal{U}(0, 1)$. Then we take the τ_j with $1 \leq j \leq N^*$ satisfying

$$\sum_{i=0}^{j-1} \pi_i < \rho \leq \sum_{i=0}^{j} \pi_i, \ \ \pi_0 = 0.$$

If $\rho > \sum_{i=0}^{N^*} \pi_i$, we sample τ_j from the prior $\mathcal{G}(a, b)$. In any case we sample z_{n+1} from the normal kernel $\mathcal{N}(0, \tau_j^{-1})$.

7. The full conditional for x_0, with a uniform prior over the set $\tilde{\mathbb{X}} \subseteq \mathbb{R}$ that represents our prior knowledge for the state space of the dynamical system in eq. (3.1) will be

$$f(x_0 \,|\, \cdots) \;\propto\; \mathcal{I}(x_0 \in \tilde{\mathbb{X}}) \exp\left\{-\frac{\tau_{d_1}}{2} h_\vartheta(x_1, x_0)\right\}. \tag{3.27}$$

8. For the vector of parameters ϑ, and assuming a uniform prior over the subset $\tilde{\Theta}$ of the parameter space $\mathbb{R}^k$, the full conditional becomes

$$f(\vartheta \,|\, \cdots) \;\propto\; \mathcal{I}(\theta \in \tilde{\Theta}) \exp\left\{-\frac{1}{2} \sum_{i=1}^{n_T} \tau_{d_i} h_\vartheta(x_i, x_{i-1})\right\}. \tag{3.28}$$

9. The full conditional densities for the future unobserved observations, when $T \geq 2$ and for $j = 1, \ldots, T-1$, are given by

$$f(x_{n+j} \,|\, \cdots) \;\propto\; \exp\left\{-\frac{1}{2}\left[\tau_{d_{n+j}} h_\vartheta(x_{n+j}, x_{n+j-1}) + \tau_{d_{n+j+1}} h_\vartheta(x_{n+j+1}, x_{n+j})\right]\right\}. \tag{3.29}$$

For $j = T$ the full conditional is normal with mean $g(\vartheta, x_{n+T-1})$ and variance $\tau^{-1}_{d_{n+T}}$, that is

$$f(x_{n+T} \,|\, \cdots) \;=\; \mathcal{N}\left(x_{n+T} \,|\, g(\vartheta, x_{n+T-1}), \tau^{-1}_{d_{n+T}}\right). \tag{3.30}$$

Note that, here, we have set a Beta prior over the geometric probability, that is $\lambda \sim \mathcal{B}e(\alpha, \beta)$, leading to a conjugate posterior full conditional for λ. We refer to the above sampler as *conjugate GSBR sampler*. In the next section where we attempt to compare the performance of the two models, it is reasonable to "synchronize" their prior specifications.

For the purposes of prior synchronization we will not use the conjugate GSBR sampler but a slight modification of it. Instead of setting a Beta prior on the geometric probability, we assign a transformed gamma prior over the geometric probability λ via $\lambda = (1+c)^{-1}$. So as a prior over λ we set

$$f(\lambda) = \mathcal{TG}(\lambda \,|\, \alpha, \beta) = \frac{\beta^\alpha e^\beta}{\Gamma(\alpha)} \lambda^{-(\alpha+1)} e^{-\beta/\lambda} (1-\lambda)^{\alpha-1}, \tag{3.31}$$

with $\lambda \in (0, 1)$.

Taking into consideration relation (3.31), the full conditional for the geometric probability λ in step **1.** of the conjugate GSBR sampler is now

$$f(\lambda \,|\, \cdots) \;\propto\; \lambda^{2n_T - \alpha - 1} \, (1-\lambda)^{L_{n_T}} \, e^{-\beta/\lambda} \, \mathcal{I}(0 < \lambda < 1), \tag{3.32}$$

where $L_{n_T} = \alpha + \sum_{i=1}^{n_T} N_i - n_T - 1$.

Details on sampling efficiently, via embedded Gibbs samplers, the nonstandard densities arising in eqs. (3.14) to (3.17), eqs. (3.27) to (3.30) as well as for the transformed posterior of the geometric propability λ given in eq. (3.32), are provided in Appendices A.1 and A.2. We thus circumvent Metropolis–within–Gibbs implementations.

3.5 Simulation results

Having described the two dynamical reconstruction models, in this section we compare the performance of the proposed GSBR model using as a benchmark the rDPR model. Our findings are that the GSBR models are more amenable to dynamical reconstruction purposes; they are as accurate as the rDPR models, they give smaller execution times and are less complicated and thus easier to implement.

In all the examples, we also compare the results with the results obtained from a parametric reconstruction and prediction Gibbs sampler, that is, assuming just Gaussian noise. We refer to this model as Param in the tables. As a measure for the accuracy, we use the Percentage Absolute Relative Error (PARE) given by the quantity $\mathrm{PARE} = 100 \times |x - x^*|/|x|$, where x and x^* are the true and estimated values of the quantities of interest respectively.

3.5.1 Experimental setup

Dynamical behavior of the cubic map: Quadratic polynomial maps, can exhibit for each parameter value at most one stable attractor. Multistability and coexistence of more than one strange attractors can be achieved under higher degree polynomial maps (Kraut et al., 1999). We will generate observations from a cubic random map with a deterministic part given by

$$\tilde{g}(\vartheta^*, x) = 0.05 + \vartheta^* x - 0.99x^3. \tag{3.33}$$

When $\vartheta^* \in [\underline{\vartheta}, \overline{\vartheta}]$ with $\underline{\vartheta} = -0.04$ and $\overline{\vartheta} = 2.81$ the dynamics of $\tilde{g}$, starting from $x_0 = 1$, are bounded. The map becomes bistable in the regions under the extrema of (3.33) when $\vartheta^* \in \Theta_{\mathrm{bi}} = [\underline{\vartheta}_{\mathrm{bi}}, \overline{\vartheta}_{\mathrm{bi}}]$ with $\underline{\vartheta}_{\mathrm{bi}} = 1.27$ and $\overline{\vartheta}_{\mathrm{bi}} = 2.54$. In the phase space of the map we can identify two mutually exclusive period-doubling cascades together with two mutually exclusive basins of attraction. The dynamical behavior of the cubic map in eq. (3.33) can be depicted via the bifurcation diagram given in Figure 3.1. The two coexisting attracting sets for $\vartheta^* \in \Theta_{\mathrm{bi}}$ are $\mathcal{O}^+$ (in blue) and $\mathcal{O}^-$ (in green).

For values of ϑ^* slightly larger than 2.54, the set $\mathcal{O}^+$ undergoes a sudden change. It becomes repelling, and all orbits are attracted by the "lower" set $\mathcal{O}^-$. The same behavior can be observed for all $\vartheta^* \in (2.54, 2.65]$. Nevertheless, orbits in the presence of dynamical noise of sufficient intensity, visit the vicinity of the repelling set $\mathcal{O}^+$, ad infinitum. For values of ϑ^* greater than 2.65, there is only one stable attractor.

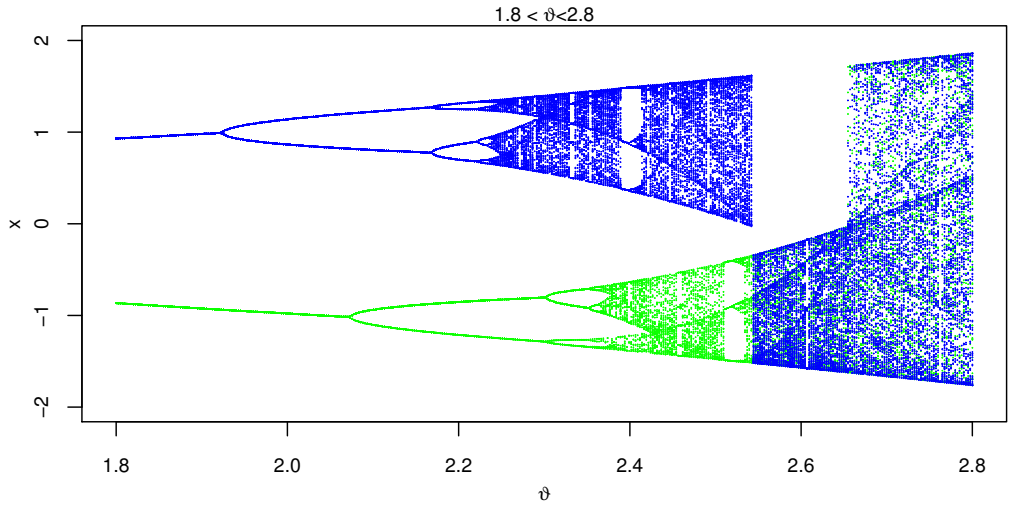

Figure 3.1: The bifurcation diagram for the deterministic map $x_i = g(\vartheta^*, x_{i-1})$.

In Figure 3.2, we set the value of the control parameter to $\vartheta^* = 2.55$ (the value of the control parameter we have used in our numerical experiments) and we superimpose two deterministic and one $f_{2,4}$-perturbed stochastic orbit. The two deterministic orbits, starting from $x_0 = 1$ and $x_0 = -1$, are depicted in blue and green respectively, whereas the stochastic, starting from $x_0 = 1$, in red.

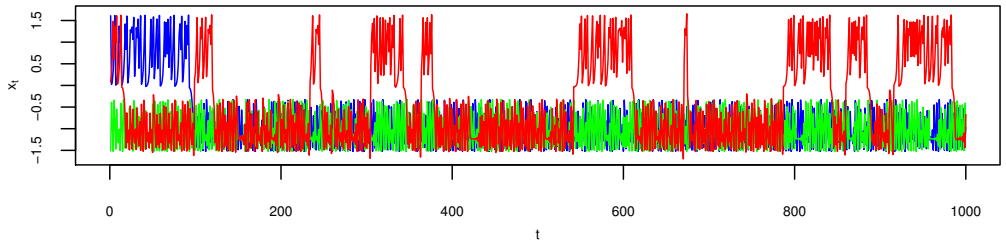

Figure 3.2: The orbits of the the deterministic map $x_i = g(\vartheta^*, x_{i-1})$, with $\vartheta^* = 2.55$, starting from $x_0 = 1$ and $x_0 = -1$ are depicted in blue and green respectively. A dynamically $f_{2,4}$-perturbed orbit, starting from $x_0 = 1$, is given in red.

Noise processes: We illustrate the GSBR and rDPR models with simulated data sets, consisting of observations generated from the cubic random recurrence $x_i = \tilde{g}(\vartheta^*, x_{i-1}) + z_i$, for the specific parameter value $\vartheta^* = 2.55$ and initial condition $x_0 = 1$. The dynamical noise z_i was sampled from:

1. The equally weighted normal 4-mixture

$$f_1 = \sum_{r=0}^{3} \frac{1}{4} \mathcal{N}\left(0, (5r+1)\sigma^2\right), \ \sigma = 10^{-2}. \tag{3.34}$$

2. The normal 2-mixtures, which exhibit progressively heavier tails for $1 \leq l \leq 4$

$$f_{2,l} = \frac{5+l}{10} \mathcal{N}(0, \sigma^2) + \frac{5-l}{10} \mathcal{N}\left(0, (200\sigma)^2\right), \ \sigma = 10^{-3}. \tag{3.35}$$

As a measure of the tail fatness of the density $z \sim f$, we use the mean absolute deviation from the mean normalized by the standard deviation, for a zero mean z it is that $TF_f = \mathbb{E}|z|/\sqrt{\mathbb{E}|z|^2}$. The closer TF_f is to 1, the thinner the tails are. It can be verified numerically that

$$TF_{f_1} > TF_{f_{2,1}} > \cdots > TF_{f_{2,4}}.$$

We model the deterministic part $g(\vartheta, x)$ of the map in eq. (3.1) with a polynomial in x of degree $m = 5$.

Prior specifications: Here we define the synchronized prior specifications of the GSBR and rDPR Gibbs samplers. We use the following general prior set up:

$$\begin{aligned} &c \sim \mathcal{G}(\alpha, \beta), \quad \lambda \sim \mathcal{TG}(\alpha, \beta), \quad \{\tau_j \sim \mathcal{G}(a, b) : j \geq 1\} \\ &\vartheta \sim \mathcal{U}((-M, M)^{k+1}), \quad x_0 \sim \mathcal{U}(-M_0, M_0), \end{aligned}$$

where k is the degree of the modeling polynomial.

A. Noninformative reconstruction and prediction – NRP: In the absence of any prior knowledge, we propose a noninformative prior specification for simultaneous reconstruction and prediction, namely

$$\mathcal{PS}_{\text{NRP}} : \alpha = \beta \geq 10^{-1}, \ a = b \geq 10^{-4}, \ M \gg 1, \ M_0 \gg 1.$$

B. Informative reconstruction and prediction – IRP: When a–priori we believe that the dynamical noise resembles a finite mixture of zero mean Gaussians with variances that are close to each other, we set:

$$\mathcal{PS}_{\text{IRP}} : \alpha > \beta \geq 10^{-1}, \ a > b \geq 10^{-4}, \ M \gg 1, \ M_0 \gg 1.$$

Such prior specifications induce a small average GSB probability λ (and consequently a large average DP concentration mass c), forcing the Gibbs samplers to activate a large number of normal kernels. Thus, generating a more detailed Gaussian mixture representation of the unknown dynamical noise.

Data sets and invariant sets: In Figure 3.3(a), we display the deterministic orbit of length 280 of the deterministic map $y_i = \tilde{g}(\vartheta^*, y_{i-1})$, with starting point at $y_0 = 1$. We have approximated the interval $\mathbb{X}$ that remains invariant under the action of $\tilde{g}(\vartheta^*, \cdot)$ by $[-1.8881, 1, 8991]$ (see Appendix B), and the associated average characteristic Liapunov exponent by 0.4625. Realizations of the random recurrence $x_i = \tilde{g}(\vartheta^*, x_{i-1}) + z_i, \ x_0 = 1$ under different types of noise are given in Figure 3.3(b) and (c) respectively.

Our observations for reconstruction and out–of–sample prediction will be the data sets $x_{f_1}^{(200)}$ and $\{x_{f_{2,l}}^{(200)} : 1 \leq l \leq 4\}$. The latter data sets, have been generated in `R` under the random number generator seeds $\texttt{RNG}_{f_1} = 1$ and $\texttt{RNG}_{f_{2,l}:1\leq l\leq 4} = \{10, 15, 13, 38\}$.

Approximations of the deterministic and noisy invariant measures are given in Figure 3.3(d)-(f). The deterministic invariant measure $\mu_{\tilde{g},0}(\mathrm{d}y)$ is approximated in Figure 3.3(d). The z-noisy measures $\mu_{\tilde{g},z}(\mathrm{d}x)$ approximated in Figure 3.3(e) and (f), are quasi-invariant in the sense that for all measurable subsets B of $\mathbb{R}$ it is that $\mu_{\tilde{g},z}(B) = \lim_{t\to\infty} P(x_t \in B \,|\, \tau_{\mathbb{X}'} > t)$, where $\tau_{\mathbb{X}'}$ is a random time denoting the first time the system enters the trapping set $\mathbb{X}'$ (see Appendix B).

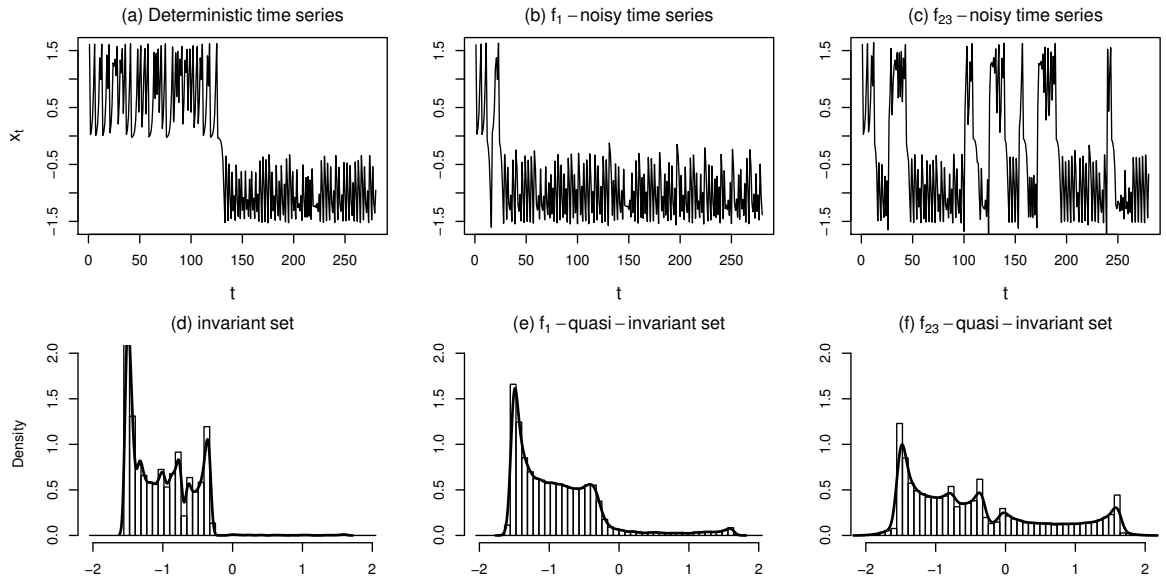

Figure 3.3: In Figure 3.3(a)-(c) we display the deterministic orbit and f_1 and $f_{2,3}$ data-realizations with initial condition $x_0 = 1$. In Figure 3.3(d)-(f) we display the deterministic and the f_1 and $f_{2,3}$ quasi-invariant set approximations respectively.

Complexity measures and prior specifications: The occurrence of an informative structure in the available data sets may help the practitioner to decide between an informative and a noninformative prior set up.

Approximate entropy (ApEn) (Borchers, 2015; Pincus, 1991) can be used to assess the complexity of the available set $x_f^{(n)}$ of observations. Large ApEn values indicate irregular and unpredictable time series data. Nevertheless, it is known that ApEn values are heavily dependent on sample size (lower than expected for small sample sizes).

A recently developed complexity measure that is less dependent on the sample size is the forecastable component analysis Ω (ForeCa) (Goerg, 2013, 2016), which is based on the entropy of the spectral density of the time series, and is normalized between zero and one. Large Ω values characterize more predictable time series.

In Figure 3.4, we display the Ω curves as functions of the sample size n, for the time series $x_{f_1}^{(n)}$ and $\{x_{f_{2,l}}^{(n)} : 1 \leq l \leq 4\}$. For the computation of the Ω curves we have used the weighted overlapping segment averaging (WOSA) method Goerg (2016). The data sets $\{x_{f_{2,l}}^{(n)} : 1 \leq l \leq 4\}$

have the more informative structure as for $n > 80$ and $1 \leq l \leq 4$ it is that

$$\Omega(x_{f_{2l}}^{(n)}) > \Omega(x_{f_1}^{(n)}).$$

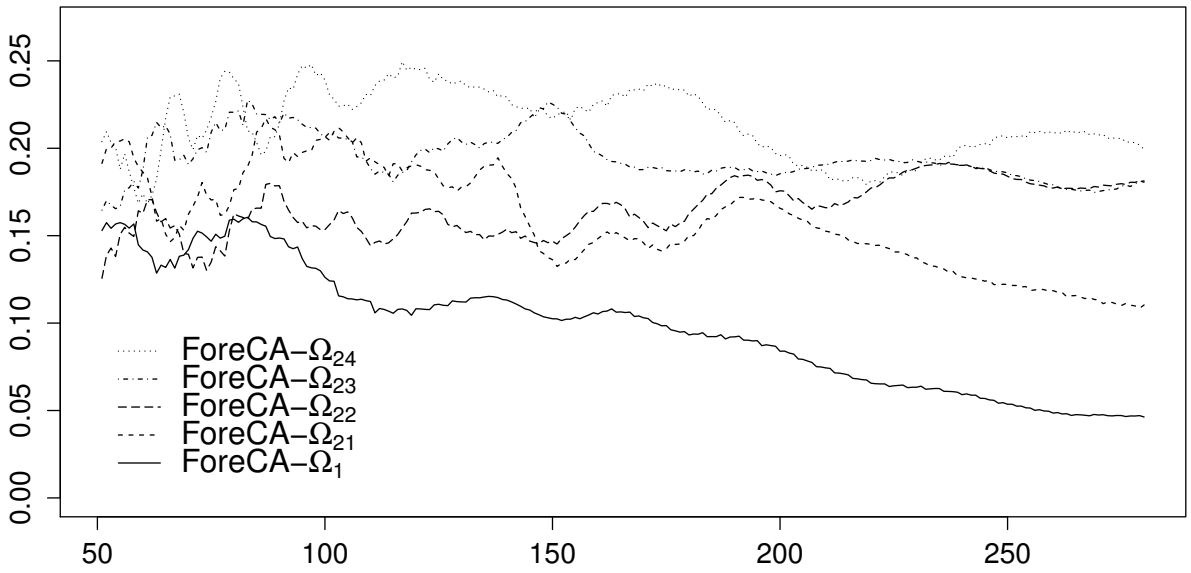

Figure 3.4: Here we display the Ω curves relating to the data sets $x_{f_1}^{(n)}$ and $\{x_{f_{2,l}}^{(n)} : 1 \leq l \leq 4\}$ for n between 50 and 280.

3.5.2 Informative reconstruction and prediction under the f_1 dynamic noise

We ran the Param, rDPR and GSBR Gibbs samplers for $T = 20$ in a synchronized mode, for 5×10^5 iterations and a burn-in period of $10,000$, using data set $x_{f_1}^{(200)}$ under the informative prior specification (IRP) $\mathcal{PS}_{\mathrm{IRP}}$ with $\alpha = 3$, $\beta = 0.3$, $a = 1$, $b = 10^{-3}$ and $M = M_0 = 10$.

We remark that under noninformative prior (NRP) specifications of the form $\alpha = \beta \leq 0.3$, and $a = b \leq 10^{-3}$, the average number of active normals for both nonparametric samplers is lesser than four, leading to less accurate estimations. The following provide a summary and some brief comments.

Initial condition and dynamical noise density estimations: In Figure 3.5(a) we display kernel density estimations (KDE's) based on the predictive samples of the marginal posterior (PPM) for the initial condition x_0. The differences between the two predictives coming from the GSBR and rDPR samplers are indistinguishable.

The three modes of the predictive density of x_0 are very close to the three real roots of the polynomial equation $\tilde{g}(\vartheta^*, x) - \tilde{g}(\vartheta^*, 1) = 0$ which are the preimages of $\tilde{g}(\vartheta^*, 1)$. Note that for $\vartheta \in (0.74, 2.97)$, it is that $\tilde{g}^{-1}(\vartheta, \tilde{g}(\vartheta, 1)) \in \{\rho, -1-\rho, 1\}$ with $\rho = -(1 + \sqrt{4\vartheta/0.99 - 3})/2$. We refer to the three preimages of $\tilde{g}(\vartheta, 1)$ by $x_{\mathrm{L}} = \rho$ (left), $x_{\mathrm{M}} = -1 - \rho$ (middle) and $x_{\mathrm{R}} = 1$ (right).

In Figure 3.5(b), we give superimposed the noise predictives coming from the two models together with the true density of the noise component given in eq. (3.34). We note how the

synchronized execution produces almost identical dynamical noise density estimations, which are very close to the true noise density f_1 (solid line in red).

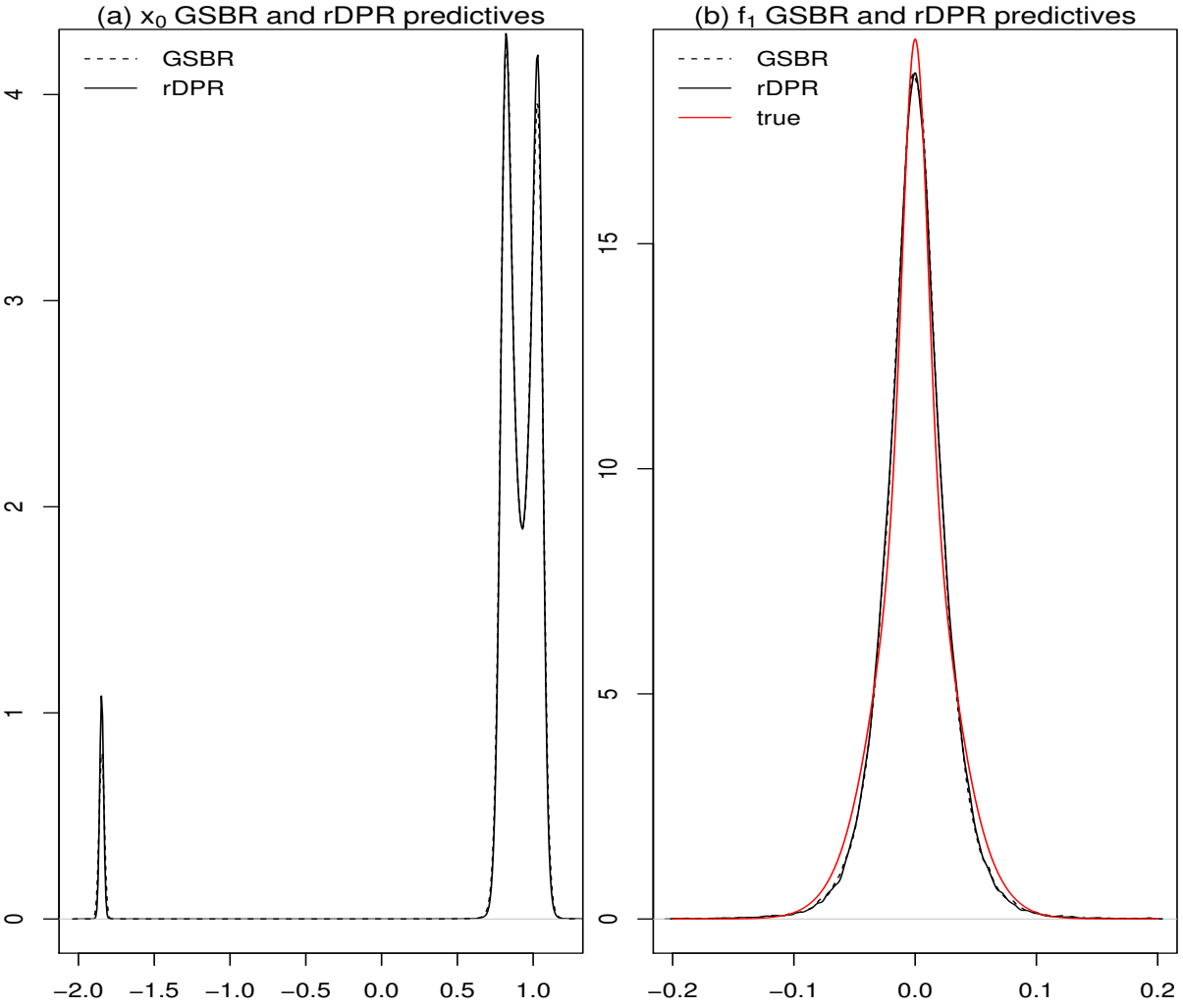

Figure 3.5: In Figure 3.5(a) we give superimposed the KDE's based on the posterior marginal predictive samples of the initial condition variable x_0. In Figure 3.5(b) we superimpose the GSBR and the rDPR noise density estimations together with the true dynamical error density.

In Figure 3.6(a)-(f), we plot the running ergodic averages for the θ_j variables of the first $80,000$ iterations after burn-in. We observe that the θ_j chains have converged after the first $10,000$ iterations, and that the chains are mixing well. In Table 3.1 we display the percentage absolute relative errors (PARE's) of the synchronized estimations. For each j, we have created $K = 47$ approximately independent samples of size $N = 10^4$, each sample separated by $s = 500$ observations

$$\{\theta_j^{(i_r)} : M_r + 1 \leq i_r \leq M_r + N\} \quad \text{with} \quad M_k = (r-1)(N+s),$$

for $r = 1, \ldots, K$. Then we created K realizations of the sampling mean (SM) estimator. Finally we took

$$\hat{\theta}_j = \frac{1}{K}\sum_{r=1}^{K}\frac{1}{N}\sum_{i=M_r+1}^{M_r+N}\theta_j^{(i)}, \quad 0 \leq j \leq 5.$$

We estimate x_0 by the maximum a–posteriori (MAP) of the x_0 predictive sample, by dividing the interval $[-2, 2]$ into 300 bins. We remark the accuracy and the closeness of the estimated ϑ values.

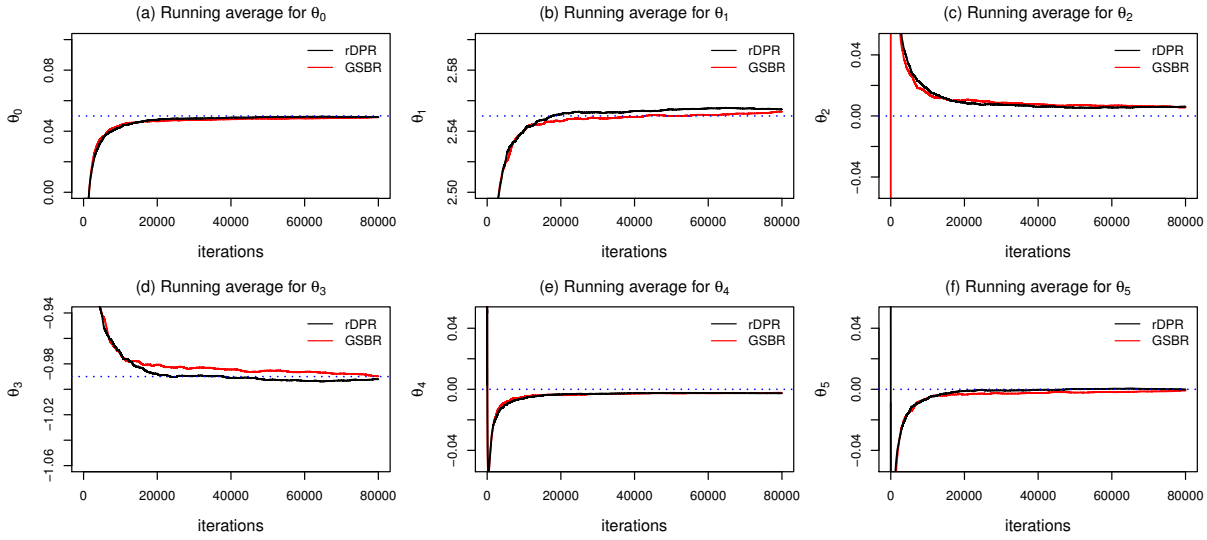

Figure 3.6: Chain ergodic averages for the θ_j variables based on the data set $x_{f_1}^{(200)}$, under prior specification $\mathcal{PS}_{\mathrm{IR}}$, are superimposed in Figure 3.6.

Table 3.1: (ϑ, x_0) reconstruction PAREs ($T = 0$) under the informative prior configuration.

Model	θ_0	θ_1	θ_2	θ_3	θ_4	θ_5	x_0
Param.	1.98	0.37	0.03	0.58	0.00	0.04	x_{M} : 3.87
rDPR	0.81	0.29	0.01	0.09	0.04	0.14	x_{M} : 0.80
GSBR	0.19	0.27	0.05	0.04	0.02	0.18	x_{R} : 0.60
Estim.	x_{201}	x_{202}	x_{203}	x_{204}	x_{205}	GSBR-Av	Par-Av
SM	6.43	7.35	29.70	5.48	13.68	12.53	53.49
MAP	3.84	11.48	19.16	2.15	149.06	37.14	53.25

Out–of–sample posterior predictive marginals and the prediction barrier: In Figure 3.7(a)-(j) we display the KDEs of the marginal posterior predictive samples of the variables $x_{201}, \ldots, x_{205}$ and $x_{216}, \ldots, x_{220}$ coming from the GSBR (solid red line) and rDPR (dashed black line) superimposed. Together, we superimpose the f_1 quasi-invariant measure approximation (solid black line). We note how the synchronized execution produces almost identical posterior predictive marginals (PPM's).

As the prediction horizon increases, the PPM densities are starting to resemble to the f_1 quasi-invariant density approximation, which naturally forms a prediction barrier. As such, any attempt to predict beyond this time horizon will replicate the quasi-invariant measure approximation. From this point on, we can make only probabilistic prediction arguments for the long term behavior of the system that involve the quasi-invariant measure i.e. $P(x_{n+i} \in A) = \mu_{\tilde{g},z}(A)$ for all $i \geq T$ and for all measurable subsets A of $\mathbb{R}$.

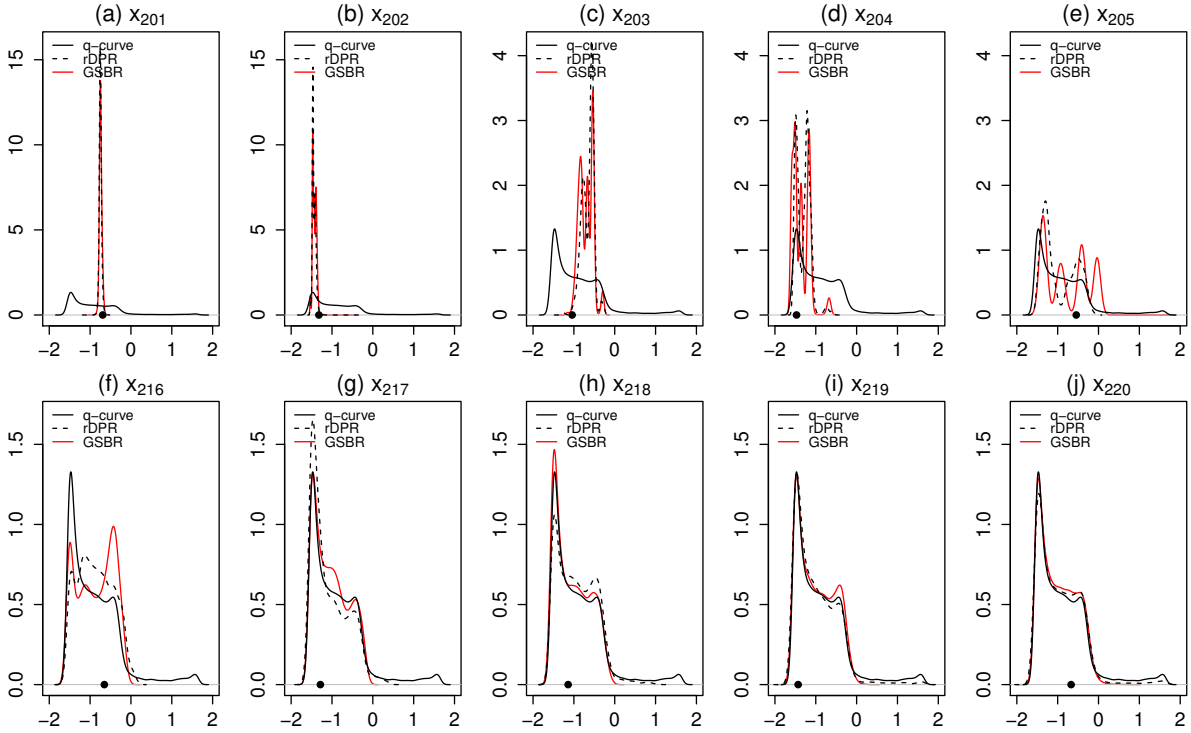

Figure 3.7: In Figure 3.7(a)-(j) we display superimposed the first five and the last five KDE's of the out–of–sample posterior marginal predictive based on data set $x_{f_1}^{(200)}$ under the informative specification $\mathcal{PS}_{\text{IRP}}$. Together we superimpose the KDE of the f_1 quasi invariant density (solid black line). In all Figures, the bullet point represents the corresponding true future value.

In Table 3.2, we give the mean computational time per 10^3 iterations relating to the synchronized execution of the rDPR and GSBR samplers under prior set up $\mathcal{PS}_{\text{IRP}}$ for a simple reconstruction ($T = 0$) and prediction ($T = 20$). In both cases, the GSBR sampler has the fastest execution times. In the last two rows of Table 3.1 we give the PARE's of the first five GSBR out–of–sample predictions using the SM and MAP estimators. The last two columns exhibit the mean PARE's under a GSBR and a parametric (Param) prediction.

Table 3.2: Mean execution times in seconds per 10^3 iterations for $x_{f_1}^{(200)}$.

Data set $x_{f_1}^{(200)}$			
Prior spec.	Algorithm	$T = 0$	$T = 20$
$\mathcal{PS}_{\text{IRP}}$	rDPR	5.44	11.76
$\mathcal{PS}_{\text{IRP}}$	GSBR	2.24	8.65

3.5.3 Noninformative reconstruction and prediction under the $f_{2,l}$ heavy tailed dynamic noise

Here we simultaneously reconstruct and predict using the noninformative prior set up. More specifically for $T = 20$ we set $\alpha = \beta = 0.3, a = b = 10^{-3}, M = M_0 = 10$; we iterated the GSBR sampler 5×10^5 after a burn-in period of $10,000$.

In Figure 3.8 we display the KDE's based on the PPM samples of the out–of–sample variables $\{x_{201}, \ldots, x_{205}\}$ and $\{x_{216}, \ldots, x_{220}\}$ (solid lines in red) under data sets $x_{f_{2,l}}^{(200)} : 1 \leq l \leq 4\}$ (rows (a) to (d)). Together we superimpose the KDE of the associated quasi-invariant densities for $1 \leq l \leq 4$ (solid lines in black).

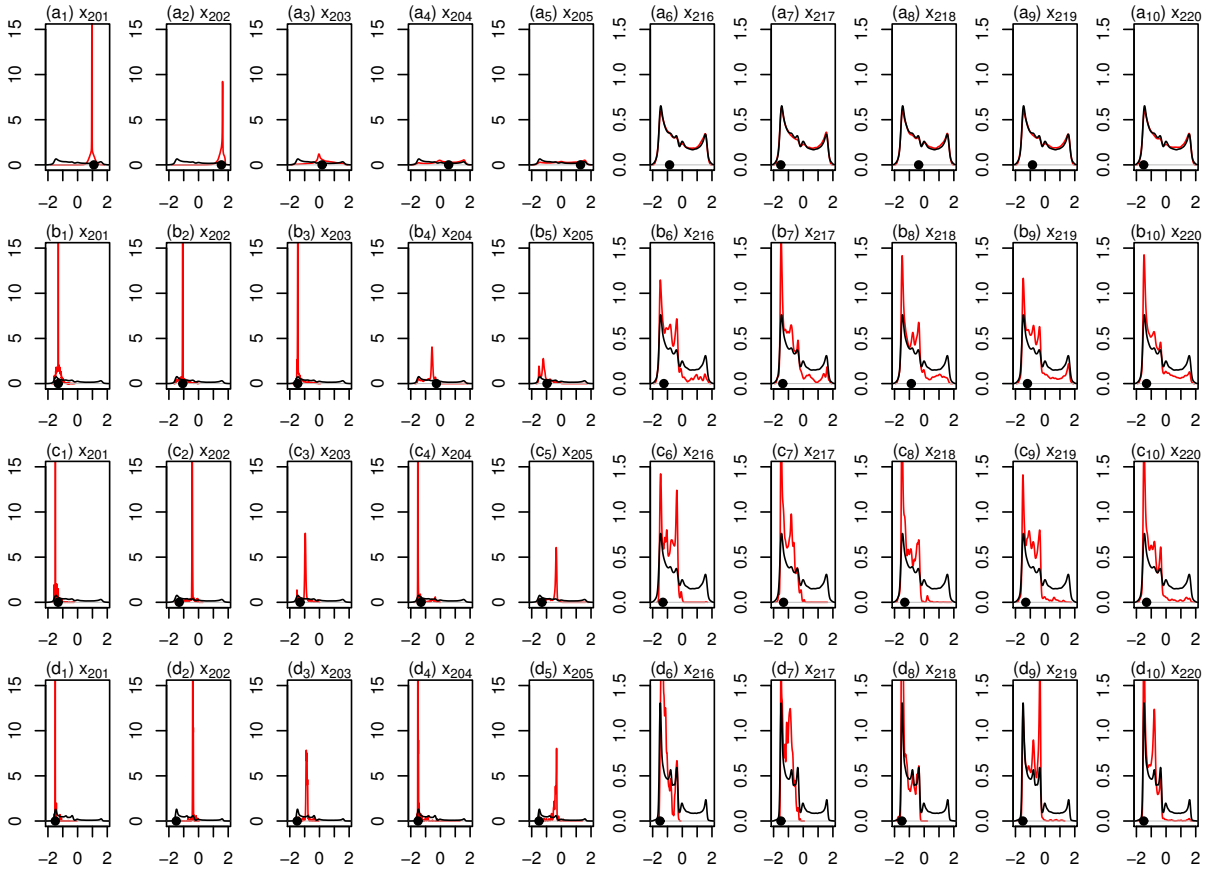

Figure 3.8: In Figure 3.8 we display the GSBR KDE's of the PPM sample of the out–of–sample variables $\{x_{201}, \ldots, x_{205}\}$ and $\{x_{216}, \ldots, x_{220}\}$(solid lines in red) based on samples $x_{f_{2,l}}^{(200)} : 1 \leq l \leq 4\}$ (rows (a) to (d)) under the noninformative prior specification. Together we superimpose the KDE of the $f_{2,l}$ quasi-invariant densities for $1 \leq l \leq 4$ (solid lines in black).

In Tables 3.3 and 3.4 we display a PARE summary of (ϑ, x_0) estimations and out-of-sample prediction respectively, based on data sets $\{x_{f_{2,l}}^{(200)} : 1 \leq l \leq 4\}$.

In Table 3.3 we compare horizontally the PARE results coming from the GSBR and the parametric sampler (Param); we notice that in all cases, the accuracy of the GSBR model is considerably higher than its parametric counterpart. In all cases, the parametric algorithm predicts a quintic polynomial deterministic part. Also, the GSBR model precision improves as the noise model becomes more heavy tailed.

Table 3.3: Simultaneous reconstruction-prediction under the noninformative prior specification. The (ϑ, x_0) PARE's are based on the data sets $\{x_{f_{2,l}}^{(200)} : 1 \leq l \leq 4\}$ for $T = 20$.

Noise	Model	θ_0	θ_1	θ_2	θ_3	θ_4	θ_5	x_0
$f_{2,1}$	Param.	19.95	1.54	4.83	4.39	2.52	1.01	7.27
	GSBR	0.51	0.01	0.06	0.02	0.02	0.00	x_{R} : 0.03
$f_{2,2}$	Param.	2.89	0.94	4.07	2.37	2.07	0.76	7.49
	GSBR	0.54	0.05	0.06	0.12	0.03	0.03	x_{R} : 0.03
$f_{2,3}$	Param.	29.97	0.40	4.97	1.25	1.88	0.41	7.55
	GSBR	0.20	0.04	0.04	0.13	0.02	0.04	x_{R} : 0.03
$f_{2,4}$	Param.	15.57	1.07	1.33	3.71	0.43	1.03	6.40
	GSBR	0.10	0.01	0.05	0.03	0.01	0.00	x_{R} : 0.03

In Table 3.4 when we compare the average PARE results coming from the GSBR and the parametric sampler (the last two columns) we notice that in all cases for both the SM and the MAP estimators, the prediction of the GSBR model is considerably better. We also notice, that as we move to a more heavy tailed noise model, the GSBR prediction gradually improves and the MAP–GSBR estimator becomes more efficient. This is due to the multimodal nature of the PPM's generated by GSBR.

Table 3.4: Simultaneous reconstruction-prediction under the noninformative prior specification. The out–of–sample PARE's are based on data sets $\{x_{f_{2,l}}^{(200)} : 1 \leq l \leq 4\}$ for $T = 20$. The GSBR-Av and Par-Av columns are the PARE means of the first five out–of–sample estimations using the GSBR and the parametric Gibbs (Param) samplers respectively.

Noise	Estim.	x_{201}	x_{202}	x_{203}	x_{204}	x_{205}	GSBR-Av	Par-Av
$f_{2,1}$	SM	12.50	0.86	12.57	44.04	82.11	30.42	58.72
	MAP	12.86	2.10	77.13	25.89	39.99	31.59	69.62
$f_{2,2}$	SM	0.52	0.70	8.07	167.16	15.17	38.32	65.08
	MAP	0.29	1.72	0.50	103.00	20.96	25.29	65.57
$f_{2,3}$	SM	0.72	7.99	0.01	9.74	49.94	13.68	233.53
	MAP	0.14	0.47	2.34	0.39	1.38	0.93	234.80
$f_{2,4}$	SM	0.24	1.01	2.95	3.79	40.25	9.65	60.69
	MAP	0.07	0.86	4.78	0.13	21.00	5.37	109.23

3.6 Conclusions

We have described a Bayesian nonparametric approach for dynamical reconstruction and prediction from observed time series data. The key insight is to use the GSB process, developed by Fuentes-García et al. (2010), as a prior (over the space of densities) on the noise component.

The GSBR model removes a level from the hierarchy of the rDPR model as it replaces the weights of the stick breaking representation of the DP with their expected values, leading to a simpler model with only one infinite dimensional parameter, the locations of the atoms (τ_j) of

the random measure. GSB mixture dynamical modeling is as accurate as DP based modeling but it gives smaller execution times, and is easier to implement.

We have also shown that in a joint prediction of future values of a low dimensional noisy chaotic time series, the quasi-invariant set appears as a "prediction barrier". Also, our numerical experiments indicate that when the sample size of the time series is small, the forecastable component analysis Ω measure can group the available sets of observations in terms of their complexity. A larger Ω index suggests a less informative prior set up.

Chapter 4

Pairwise Dependent Random Mixtures

4.1 Introduction

In this chapter we focus on the construction of *Pairwise Dependent Geometric Stick Breaking Processes* (PDGSBP), a dependent Bayesian nonparametric prior for partially exchangeable observations based on the GSB process (Hatjispyros et al., 2017a).

That is, we are going to model a finite collection of m random distribution functions $(\mathbb{G}_j)_{1\leq j\leq m}$, where each $\mathbb{G}_j$ is a GSB random probability measure, such that there is a unique common component for each pair $(\mathbb{G}_j, \mathbb{G}_{j'})$ with $j \neq j'$. We are going to use these measures in the context of GSB mixture models, generating a collection of m GSB pairwise dependent random densities $(f_j(x))_{1\leq j\leq m}$. Hence we obtain a set of random densities $(f_1, \ldots, f_m)$, where marginally each f_j is a random density function

$$f_j(x) = \int_\Theta K(x \,|\, \theta)\, \mathbb{G}_j(\mathrm{d}\theta),$$

thus generalizing the GSB priors to a multivariate setting for partially exchangeable observations.

In the problem considered here, these random density functions $(f_j)_{1\leq j\leq m}$ are thought to be related or similar, e.g. pertubations of each other, and so we aim to share information between groups to improve estimation of each density, especially for those densities f_j for which the corresponding sample size n_j is small.

We are going to provide evidence through numerical experiments, that dependent GSB mixture models provide an efficient alternative to pairwise dependent DP (PDDP) priors; that is making the weights more exotic does not actually enlarge the support of the prior. At first, we will randomize the existing PDDP model of Hatjispyros et al. (2011, 2016), by imposing gamma priors on its concentration masses, and then we will conduct a–priori synchronized density estimation comparison studies between the randomized PDDP model (rPDDP) and the pairwise dependent GSB process (PDGSBP) model using synthetic and real data examples.

4.2 Randomized pairwise dependent Dirichlet process

To introduce pairwise dependence between m random density functions, Hatjispyros et al. (2011), introduced the following hierarchical model. For the m subgroups of observations $\{(x_{ji})_{1\leq i\leq n_j} : 1 \leq j \leq m\}$,

$$\begin{aligned}
&x_{ji} \,|\, \theta_{ji} \overset{\text{ind}}{\sim} K(\,\cdot\,|\,\theta_{ji})\\
&\theta_{ji} \,|\, \mathbb{Q}_j \overset{\text{iid}}{\sim} \mathbb{Q}_j(\,\cdot\,)\\
&\mathbb{Q}_j = \sum_{l=1}^{m} p_{jl}\mathbb{P}_{jl},\ \sum_{l=1}^{m} p_{jl} = 1,\ \mathbb{P}_{jl} = \mathbb{P}_{lj}\\
&\mathbb{P}_{jl} \overset{\text{iid}}{\sim}\ \text{DP}(c, P_0),\ 1 \leq j \leq l \leq m,
\end{aligned}$$

for some kernel density $K(\,\cdot\,|\,\cdot\,)$, concentration parameter $c > 0$ and parametric central measure P_0 for which $\mathbb{E}[\mathbb{P}_{jl}(\mathrm{d}\theta)] = P_0(\mathrm{d}\theta)$.

So, the random densities $f_j(x)$ are dependent mixtures of the dependent random measures $\mathbb{Q}_j$ via $f_j(x\,|\,\mathbb{Q}_j) = \int_\Theta K(x\,|\,\theta)\mathbb{Q}_j(\mathrm{d}\theta)$, or equivalently, dependent mixtures of the m independent mixtures $g_{jl}(x\,|\,\mathbb{P}_{jl}) = \int_\Theta K(x\,|\,\theta)\,\mathbb{P}_{jl}(\mathrm{d}\theta)$, $l = 1, \ldots, m$. The density function for an observation x_{ji} then becomes

$$f_j(x_{ji}\,|\,\mathbb{Q}_j) := f_j(x_{ji}) = \sum_{l=1}^{m} p_{jl} g_{jl}(x_{ji}). \tag{4.1}$$

The $g_{jl}(x_{ji})$ are random density functions defined by a DPM, that is

$$g_{jl}(x_{ji}) = \int_\Theta K(x_{ji}\,|\,\theta)\mathbb{P}(\mathrm{d}\theta) = \sum_{k=1}^{\infty} w_{jlk} K(x_{ji}\,|\,\theta_{jlk}), \tag{4.2}$$

where $(w_{jlk})_{k\geq 1}$ are the stick breaking weights of the stick breaking representation of the Dirichlet process. Then the random density of the observation x_{ji} can be written explicitly as

$$f_j(x_{ji}\,|\,\mathbb{P}_{jl}, 1 \leq l \leq m) = \sum_{l=1}^{m} p_{jl} \left\{ \sum_{k=1}^{\infty} w_{jlk} K(x_{ji}\,|\,\theta_{jlk}) \right\} \tag{4.3}$$

To introduce the rPDDP model, we randomize the PDDP model by sampling the $\mathbb{P}_{jl}$ measures from the independent Dirichlet processes $\text{DP}(c_{jl}, P_0)$ and then impose gamma priors on the concentration masses, i.e. $\mathbb{P}_{jl} \overset{\text{ind}}{\sim}\ \text{DP}(c_{jl}, P_0),\ c_{jl} \overset{\text{ind}}{\sim} \mathcal{G}(a_{jl}, b_{jl}),\ 1 \leq j \leq l \leq m$.

According to Hatjispyros et al. (2011, 2016) the augmentation of each f_j in eq. (4.1) with positive auxiliary random variables will make the number of updates of the Gibbs sampler finite almost surely. To this end we introduce:

1. The DP mixture selection variables $\boldsymbol{\delta} = (\delta_{ji})$; for an observation x_{ji} that comes from f_j, δ_{ji} selects the DP mixture $g_{jl}(x)$ that the observation came from. In particular we have that $\Pr(\delta_{ji} = l) = p_{jl}$.

2. The clustering variables $\boldsymbol{d} = (d_{ji})$; for an observation x_{ji} that comes from f_j, given δ_{ji}, d_{ji} allocates the component of the DP mixture $g_{j\delta_{ji}}(x)$ that x_{ji} came from.

Finally, we define the stochastic variables $\mathbf{u} = (u_{ji})$ for $1 \leq i \leq n_j$ and $1 \leq j \leq m$, associated with a non–sequential slice set $A_{w_{jl}}(u_{ji}) = \{k \in \mathbb{N} : 0 < u_{ji} < w_{jlk}\}$.

For the clustering variables we have that

$$\Pr(d_{ji} = k) = \sum_{l=1}^{m} \Pr(d_{ji} = k, \delta_{ji} = l) = \sum_{l=1}^{m} \Pr(\delta_{ji} = l)\Pr(d_{ji} = k \,|\, \delta_{ji} = l) = \sum_{l=1}^{m} p_{jl} w_{jlk}.$$

Conditionally on the event $\{\delta_{ji} = l\}$ the clustering variables have an infinite state space, that is

$$(d_{ji} \,|\, \delta_{ji} = l) \sim \sum_{k=1}^{\infty} w_{jlk} \delta_k,$$

from which we deduce that $\Pr(d_{ji} = k \,|\, \delta_{ji} = l) = w_{jlk}$.

Proposition 4.1. *Suppose that the clustering variables (d_{ji}) conditionally on the slice variables u_{ji} are having a discrete uniform distribution over the elements of the slice sets $A_{w_{jl}}(u_{ji})$ that is $d_{ji} \,|\, u_{ji} \sim \mathcal{DU}(A_{w_{jl}}(u_{ji}))$, then*

$$f_j(x_{ji}, u_{ji}) = \sum_{l=1}^{m} p_{jl} \sum_{k \in A_{w_{jl}}(u_{ji})} K(x_{ji} \,|\, \theta_{jlk}). \tag{4.4}$$

and

$$f_j(x_{ji}, u_{ji}, d_{ji} = k \,|\, \delta_{ji} = l) = w_{jlk}\, \mathcal{U}(u_{ji} \,|\, 0, w_{jlk})\, K(x_{ji} \,|\, \theta_{jlk}). \tag{4.5}$$

Proof. Starting from the u_{ji}–augmented random densities we have

$$\begin{aligned}
f_j(x_{ji}, u_{ji}) &= \sum_{l=1}^{m} f_j(x_{ji}, u_{ji}, \delta_{ji} = l) \\
&= \sum_{l=1}^{m} \Pr(\delta_{ji} = l)\, f_j(x_{ji}, u_{ji} \,|\, \delta_{ji} = l) \\
&= \sum_{l=1}^{m} p_{jl} \sum_{k=1}^{\infty} f_j(x_{ji}, u_{ji}, d_{ji} = k \,|\, \delta_{ji} = l) \\
&= \sum_{l=1}^{m} p_{jl} \sum_{k=1}^{\infty} f_j(d_{ji} = k \,|\, \delta_{ji} = l)\, f_j(u_{ji} \,|\, d_{ji} = k, \delta_{ji} = l)\, f_j(x_{ji} \,|\, d_{ji} = k, \delta_{ji} = l) \\
&= \sum_{l=1}^{m} p_{jl} \sum_{k=1}^{\infty} w_{jlk}\, \mathcal{U}(u_{ji} \,|\, 0, w_{jlk})\, K(x_{ji} \,|\, \theta_{jlk}) \\
&= \sum_{l=1}^{m} \sum_{k=1}^{\infty} \mathcal{I}(u_{ji} < w_{jlk})\, K(x_{ji} \theta_{jlk}) \\
&= \sum_{l=1}^{m} p_{jl} \sum_{k \in A_{w_{jl}}(u_{ji})} K(x_{ji} \,|\, \theta_{jlk}).
\end{aligned}$$

Augmenting further with variables d_{ji} and δ_{ji} yields

$$f_j(x_{ji}, u_{ji}, d_{ji} = k, \delta_{ji} = l) = p_{jl}\, w_{jlk}\, \mathcal{U}(u_{ji} \,|\, 0, w_{jlk})\, K(x_{ji} \,|\, \theta_{jlk}).$$

Because $\Pr(\delta_{ji} = l) = p_{jl}$ the last equation leads to eq. (4.5) and the desired result follows. □

The following result is the main property of the slice variables (u_{ji}) that allows us to create Gibbs samplers with a finite number of updates for the PDDP model. Letting $|S|$ stand for the cardinality of a set S, we show that

Proposition 4.2. *Given the random sets $A_{w_{jl}}(u_{ji})$ the random functions in eq.* (4.3) *become finite mixtures of a.s. finite equally weighted mixtures of the $K(\,\cdot\,|\,\cdot\,)$ probability kernels, that is*

$$f_j(x_{ji} \,|\, u_{ji}) = \sum_{l=1}^{m} \mathcal{W}_{jl} \frac{1}{|A_{w_{jl}}(u_{ji})|} \sum_{k \in A_{w_{jl}}(u_{ji})} K(x_{ji} \,|\, \theta_{jlk}), \tag{4.6}$$

with

$$\mathcal{W}_{jl} = \frac{p_{jl}|A_{w_{jl}}(u_{ji})|}{\sum_{r=1}^{m} p_{jr}|A_{w_{jr}}(u_{ji})|}.$$

Proof. First note that marginally, for the slice variables (u_{ji}) it is that

$$\begin{aligned}
f(u_{ji}) &= \sum_{l=1}^{m} \sum_{k=1}^{\infty} f_j(u_{ji}, d_{ji} = k, \delta_{ji} = l) \\
&= \sum_{l=1}^{m} \sum_{k=1}^{\infty} \Pr(\delta_{ji} = l) f_j(u_{ji} \,|\, d_{ji} = k, \delta_{ji} = l) \\
&= \sum_{l=1}^{m} \sum_{k=1}^{\infty} p_{jl}\, \mathcal{U}(u_{ji} \,|\, 0, w_{jlk}) \\
&= \sum_{l=1}^{m} p_{jl} \mathcal{I}(u_{ji} < w_{jlk}) \\
&= \sum_{l=1}^{m} p_{jl}|A_{w_{jl}}(u_{ji})|.
\end{aligned} \tag{4.7}$$

Having the marginal of (u_{ji}) the conditional density of x_{ji} given the slice variable u_{ji} is given by

$$f_j(x_{ji} \,|\, u_{ji}) = \frac{f_j(x_{ji}, u_{ji})}{f_j(u_{ji})} = \frac{\sum_{l=1}^{m} \sum_{k=1}^{\infty} f_j(x_{ji}, u_{ji}, d_{ji} = k, \delta_{ji} = l)}{\sum_{r=1}^{m} \sum_{s=1}^{\infty} f_j(u_{ji}, d_{ji} = s, \delta_{ji} = r)}$$

$$= \frac{\sum_{l=1}^{m} p_{jl} \sum_{k=1}^{\infty} \mathcal{I}(u_{ji} < w_{jlk}) K(x_{ji} \,|\, \theta_{jlk})}{\sum_{r=1}^{m} p_{jr} \sum_{s=1}^{\infty} \mathcal{I}(u_{ji} < w_{jls})}$$

$$= \frac{\sum_{l=1}^{m} p_{jl} \sum_{k \in A_{w_{jl}}(u_{ji})} K(x_{ji} \,|\, \theta_{jlk})}{\sum_{r=1}^{m} p_{jr}|A_{w_{jr}}(u_{ji})|}$$

Letting

$$\mathcal{W}_{jl} = \frac{p_{jl}|A_{w_{jl}}(u_{ji})|}{\sum_{r=1}^{m} p_{jr}|A_{w_{jr}}(u_{ji})|},$$

the proposition follows. □

From now on we leave the auxiliary variables unspecified; especially for δ_{ji} we use the notation

$$\delta_{ji} = (\delta_{ji}^1, \ldots, \delta_{ji}^m) \in \{\mathbf{e}_1, \ldots, \mathbf{e}_m\} \;\text{ with }\; \Pr(\delta_{ji} = \mathbf{e}_l) = p_{jl},$$

where $\mathbf{e}_l$ denotes the usual basis vector having its only nonzero component equal to 1 at position l. Hence, for a sample of size n_1 from f_1, a sample of size n_2 from f_2, etc., a sample of size n_m from f_m we can write the full likelihood as the triple product:

$$\begin{aligned} f(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{d} \,|\, \boldsymbol{\delta}) &= \prod_{j=1}^{m}\prod_{i=1}^{n_j} f_j\left(x_{ji}, u_{ji}, d_{ji} \,|\, \delta_{ji} = (\delta_{ji}^1, \ldots, \delta_{ji}^m)\right), \quad (\delta_{ji}^1, \ldots, \delta_{ji}^m) \in \{\mathbf{e}_1, \ldots, \mathbf{e}_m\} \\ &= \prod_{j=1}^{m}\prod_{i=1}^{n_j}\prod_{l=1}^{m} \left\{\mathcal{I}(u_{ji} < w_{jld_{ji}})\, K(x_{ji} \,|\, \theta_{jld_{ji}})\right\}^{\delta_{ji}^l}. \end{aligned} \tag{4.8}$$

Equivalently using a hierarchical representation it is that

$$\begin{aligned} &x_{ji}, u_{ji} \,|\, d_{ji}, \delta_{ji}, (\theta_{jr\delta_{ji}})_{1\le r\le m}, (w_{jr\delta_{ji}})_{1\le r\le m} \overset{\text{ind}}{\sim} \prod_{r=1}^{m} \left\{\mathcal{U}(u_{ji} \,|\, 0, w_{jrd_{ji}})\, K(x_{ji} \,|\, \theta_{jrd_{ji}})\right\}^{\delta_{ji}^r} \\ &\Pr(d_{ji} = k \,|\, w_{ji}, \delta_{ji} = \mathbf{e}_l) = w_{jlk}, \;\; \Pr(\delta_{ji} = \mathbf{e}_l) = p_{jl} \\ &w_{jik} = z_{jlk}\prod_{s<k}(1 - z_{jls}), \;\; z_{jlk} \overset{\text{iid}}{\sim} \mathcal{B}e(1, c), \;\; \theta_{jik} \overset{\text{iid}}{\sim} P_0, \;\; k \in \mathbb{N}. \end{aligned}$$

4.2.1 The rPDDP Gibbs sampler

In this section, we describe the algorithmic Gibbs sampling steps for estimating the rPDDP model. The algorithm is an extended version of the algorithm described in Hatjispyros et al. (2011), including an additional step for the sampling of the m concentration parameters of the independent DP's. First, let us complete the model assigning a Dirichlet prior with parameters (over the selection probabilities $\boldsymbol{p}_j = (p_{j1}, \ldots, p_{jm})$ for $1 \le j \le m$, that is

$$f(p_j \,|\, \boldsymbol{a}_j) \propto \prod_{l=1}^{m} p_{jl}^{\alpha_{jl}-1}.$$

Having initialized (d_{ji}, δ_{ji}) for $1 \le j \le m$ and for $1 \le i \le n_j$ we will sample at each iteration of the Gibbs sampler the following variables

$$\begin{aligned} w_{jlk}, \theta_{jlk}, &\quad 1 \le j \le l \le m,\, 1 \le k \le N^*, \\ u_{ji}, d_{ji}, \delta_{ji}, &\quad 1 \le j \le m,\, 1 \le i \le n_j, \\ p_{jl}, &\quad 1 \le j \le m,\, 1 \le l \le m, \end{aligned}$$

with N^* is almost surely finite. Later, we will see that the computation of N^* is a generalization of the computation of N^* in the DPM model in many dimensions.

1. The first step is to sample the sequence of the stick–breaking weights and their associated locations, that is (w_{jlk}, θ_{jlk}). Following standard results of Kalli et al. (2011), we will sample the variables from their conditional, having the slice variables u_{ji}, integrated out. Then for $1 \leq j \leq m$, $1 \leq i \leq n_j$ it is that

$$f(z_{jjk} \,|\, \cdots) = \mathcal{B}e\left(z_{jjk} \,|\, 1 + \sum_{i=1}^{n_j} \mathcal{I}(d_{ji} = k, \delta_{ji} = \mathbf{e}_j), c_{jj} + \sum_{i=1}^{n_j} \mathcal{I}(d_{ji} > k, \delta_{ji} = \mathbf{e}_j)\right),$$

while for $j \neq l$ we have that

$$f(z_{jlk} \,|\, \cdots) = \mathcal{B}e\Bigg(z_{jlk} \,|\, 1 + \sum_{i=1}^{n_j} \mathcal{I}(d_{ji} = k, \delta_{ji} = \mathbf{e}_l) + \sum_{i=1}^{n_l} \mathcal{I}(d_{li} = k, \delta_{li} = \mathbf{e}_j),$$
$$c_{jl} + \sum_{i=1}^{n_j} \mathcal{I}(d_{ji} > k, \delta_{ji} = \mathbf{e}_l) + \sum_{i=1}^{n_l} \mathcal{I}(d_{li} > k, \delta_{li} = \mathbf{e}_j)\Bigg).$$

The z_{jlk} and θ_{jlk} will only be sampled for $k \leq d^* = \max_{i,j} d_{ji}$. If there are any $k > d^*$ we sample them independently from the $\mathcal{B}e(1, c_{jl})$ and take the θ_{jlk} independently from p_0. Having sampled the sequence of z_{jlk} we construct the w_{jlk} weights via the stick–breaking formula.

2. For the locations of the random measures for $k = 1, \ldots, d^*$ where $d^* = \max_{j,i} d_{ji}$, it is that

$$f(\theta_{jlk} | \cdots) \propto f(\theta_{jlk}) \begin{cases} \prod_{i=1}^{n_j} K(x_{ji}|\theta_{jlk})^{\mathcal{I}(\delta_{ji}=\mathbf{e}_l,\, d_{ji}=k)} \prod_{i=1}^{n_l} K(x_{li}|\theta_{jlk})^{\mathcal{I}(\delta_{li}=\mathbf{e}_j,\, d_{li}=k)} & l > j, \\ \prod_{i=1}^{n_j} K(x_{ji}|\theta_{jjk})^{\mathcal{I}(\delta_{ji}=\mathbf{e}_j,\, d_{ji}=k)} & l = j. \end{cases}$$

3. In this step we sample the u_{ji}'s. This will enable us to sample the additional number of weights and locations for $k > d^*$. From the likelihood, it follows that

$$f(u_{ji} \,|\, \cdots) \propto \prod_{i=1}^{m} \mathcal{I}(u_{ji} < w_{jld_{ji}})^{\delta^l_{ji}}.$$

When the δ_{ji} is specified, for example if $\delta_{ji} = \mathbf{e}_l$ we have

$$f(u_{ji} \,|\, \delta_{ji} = \mathbf{e}_j, \cdots) = \mathcal{U}(u_{ji} \,|\, 0, w_{jld_{ji}}).$$

4. Here we find the additional number of weights (w_{jlk}), and locations (θ_{jlk}), we have to sample beyond d^* in order for the chain to proceed. To this end, let N_{ji} be the smallest integer N^* for which

$$\sum_{k=1}^{N^*} w_{jlk} > 1 - u^*_{jl},$$

where for $j = l$ it is that

$$u^*_{jj} = \min_i\{u_{ji}\}, \quad 1 \le j \le m,$$

and for $1 \le j \le l \le m$ we have

$$u^*_{jl} = \min\{\min_i\{u_{ji}\}, \min_i\{u_{li}\}\}.$$

This implies that we must sample (w_{jlk}, θ_{jlk}) from the prior for $k = d^* + 1, \ldots, N^*$, with $N^* = \max_{jl} N_{jl}$.

5. Here we concentrate on the sampling of (d_{ji}, δ_{ji}). By construction, the clustering variables belong to the union of the slice sets, that is

$$\Pr\left\{ d_{ji} \in \bigcup_{l=1}^{m} A_{w_{jl}}(u_{ji}) \,|\, u_{ji},\, 1 \le i \le n_j \right\} = 1.$$

Then conditionally on the δ_{ji} variables it is that

$$\Pr\left\{ d_{ji} \in A_{w_{jl}}(u_{ji}) \,|\, \delta_{ji} = \mathbf{e}_l, u_{ji},\, 1 \le i \le n_j \right\} = 1.$$

Then it follows that

$$\Pr(d_{ji} = k, d_{ji} = \mathbf{e}_l \,|\, \cdots) \propto p_{jl}\, K(x_{ji} \,|\, \theta_{jlk})\, \mathcal{I}(k \in A_{w_{jl}}(u_{ji}))\, \mathcal{I}(1 \le l \le m),$$

a bivariate discrete distribution. Thus, we sample (d_{ji}, δ_{ji}) as a block.

6. The full conditional for $j = 1, \ldots, m$ for the selection probabilities $\boldsymbol{p}_j = (p_{j1}, \ldots, p_{jm})$, under the Dirichlet prior $f(\boldsymbol{p}_j \,|\, \boldsymbol{a}_j) \propto \prod_{l=1}^{m} p_{jl}^{a_{jl}-1}$, with hyperparameter $\boldsymbol{a}_j = (a_{j1}, \ldots, a_{jm})$, is Dirichlet

$$f(\boldsymbol{p}_j \,|\, \cdots) \propto \prod_{l=1}^{m} p_{jl}^{a_{jl} + \sum_{i=1}^{n_l} \mathcal{I}(\delta_{ji} = \mathbf{e}_l) - 1}.$$

7. Here, we describe the associated Gibbs sampling steps for the updates of the concentration parameters of the independent Dirichlet processes appearing in the random measures $\mathbb{Q}_j$. In our model, the random densities (f_j) are represented as finite mixtures of the DP mixtures $g_{jl}(x \,|\, \mathbb{P}_{jl})$ with $\mathbb{P}_{jl} \overset{\text{ind}}{\sim} \text{DP}(c_{jl}, P_0)$. We let $c_{jl} \sim \mathcal{G}(a_{jl}, b_{jl})$. Then, following West (1992), we have the following two special cases:

A. For $j = l$, the posterior c_{jj}'s will be affected only by the size of the data set $\boldsymbol{x}_j$ and the number of unique clusters for which $\delta_{ji} = \mathbf{e}_j$. Letting

$$\rho_{jj} = \#\{d_{jj} : \delta_{ji} = \mathbf{e}_j, 1 \le i \le n_j\},$$

we have

$$\begin{aligned}&\beta \sim \mathcal{B}e(c_{jj}+1, n_j)\\ &c_{jj}\,|\,\beta, \rho_{jj} \;\sim\; \pi_\beta\,\mathcal{G}(a_{jj}+\rho_{jj}, b_{jj}-\log\beta) + (1-\pi_\beta)\,\mathcal{G}(a_{jj}+\rho_{jj}-1, b_{jj}-\log\beta)\end{aligned}$$

with the weights π_β satisfying $\frac{\pi_\beta}{1-\pi_\beta} = \frac{a_{jj}+\rho_{jj}-1}{n_j(b_{jj}-\log\beta)}$.

B. For $j \neq l$, the posterior c_{jl}'s will be affected by the size of the data sets $\boldsymbol{x}_j$ and $\boldsymbol{x}_l$ and the cumulative number of unique clusters d_{ji} for which $\delta_{ji} = \mathbf{e}_l$ and the unique clusters d_{li} for which $\delta_{li} = \mathbf{e}_j$. Letting

$$\rho_{jl} = \#\{d_{ji} : \delta_{ji} = \mathbf{e}_l, 1 \leq i \leq n_j\} + \#\{d_{li} : \delta_{li} = \mathbf{e}_j, 1 \leq i \leq n_l\},$$

it is that

$$\begin{aligned}&\beta \sim \mathcal{B}e(c_{jl}+1, n_j+n_l)\\ &c_{jl}\,|\,\beta, \rho_{jl} \;\sim\; \pi_\beta\,\mathcal{G}(a_{jl}+\rho_{jl}, b_{jl}-\log\beta) + (1-\pi_\beta)\,\mathcal{G}(a_{jl}+\rho_{jl}-1, b_{jl}-\log\beta),\end{aligned}$$

with the weights π_β satisfying $\frac{\pi_\beta}{1-\pi_\beta} = \frac{a_{jl}+\rho_{jl}-1}{(n_j+n_l)(b_{jl}-\log\beta)}$.

Due to the fact that $\rho_{jl} = 0$ is always a possibility, so that we impose $a_{jl} > 1$.

4.2.2 Superiority of rPDDP against PDDP

In this subsection, we demonstrate the superiority of the proposed rPDDP algorithm against the existent PDDP algorithm on a complex simulated data set. Specifically, we will generate datasets x_1 of sample size $n_1 = 200$ and x_2 of sample size $n_2 = 200$ from the mixture densities given by

$$\begin{aligned}f_1(x) &= \frac{1}{5}\sum_{k=4}^{8}\mathcal{N}(-70+10(k-1),1) = \frac{1}{5}\sum_{k\in B}\mathcal{N}(-70+10(k-1),1)\\ f_2(x) &= \frac{1}{12}\sum_{k=1}^{12}\mathcal{N}(-70+10(k-1),1) = \frac{1}{12}\sum_{k\in A}\mathcal{N}(-70+10(k-1),1)\\ &= \frac{5}{12}\left\{\frac{1}{5}\sum_{k\in B}\mathcal{N}(-70+10(k-1),1)\right\} + \frac{7}{12}\left\{\frac{1}{7}\sum_{k\in A\setminus B}\mathcal{N}(-70+10(k-1),1)\right\}.\end{aligned}$$

Equivalently in $g_{jl}(x)$ notation, we have

$$\begin{aligned}f_1(x) &= 0\cdot g_{11}(x) + 1\cdot g_{12}(x)\\ f_2(x) &= \frac{5}{12}\cdot g_{12}(x) + \frac{7}{12}\cdot g_{22}(x).\end{aligned}$$

Prior specifications. For the comparison we choose normal kernels $K(x\,|\,\theta) = \mathcal{N}(x\,|\,\theta)$ where $\theta = (\mu, \tau^{-1})$ and $\tau = \sigma^{-2}$ is the precision. The prior over the means and precisions of the PDDP

and the rPDDP model (P_0) is the independent normal-gamma measure, given by

$$P_0(\mathrm{d}\mu, \mathrm{d}\tau) = G_0(\mathrm{d}\mu, \mathrm{d}\tau) = \mathcal{N}(\mu \,|\, \mu_0, \tau_0^{-1})\, \mathcal{G}(\tau \,|\, \epsilon_1, \epsilon_2)\, \mathrm{d}\mu \mathrm{d}\tau.$$

Attempting a noninformative prior specification we took $\mu_0 = 0$ and $\tau_0 = \epsilon_1 = \epsilon_2 = 10^{-3}$. For the concentration masses of the rPDDP model, a–priori, we set $c_{jl} \sim \mathcal{G}(a_{jl}, b_{jl})$, with $a_{jl} = b_{jl} = 0.5$.

In Section 4.2.1 we have shown that such prior specifications are valid for $a_{jl} > 1$. However for the special case where $m = 2$ it is allowed to have any $a_{jl} > 0$ because $\rho_{jl} \neq 0$ always. This is because ρ_{jl} is defined as

$$\rho_{jl} = \#\{d_{ji} : \delta_{ji} = \mathbf{e}_l, 1 \leq i \leq n_j\} + \#\{d_{li} : \delta_{li} = \mathbf{e}_j, 1 \leq i \leq n_l\},$$

and now, for $1 \leq j \leq l \leq 2$, the events $\#\{d_{ji} : \delta_{ji} = \mathbf{e}_l, 1 \leq i \leq n_j\}$ and $\#\{d_{li} : \delta_{li} = \mathbf{e}_j, 1 \leq i \leq n_l\}$ are complementary. The hyperparameters (α_{jl}) of the Dirichlet priors over the matrix of the selection probabilities $p = (p_{jl})$ has been set to $\alpha_{jl} = 1$.

As a measure of superiority of the proposed methodology, we will measure the similarity between the true and the estimated probability distributions with the Hellinger distance. For two density functions f, g the Hellinger distance is defined as

$$\mathcal{H}(f, g) = \frac{1}{2} \int_{\mathbb{R}} \left(\sqrt{f(x)} - \sqrt{g(x)} \right)^2 \mathrm{d}x.$$

In this example, $\mathcal{H}(f, \hat{f})$ and $\mathcal{H}_{\mathcal{R}}(f, \hat{f})$, will denote the Hellinger distance between the true density f and the predictive density $\hat{f}$ of the PDDP and rPDDP algorithms, respectively. The Gibbs samplers run for 11×10^4 iterations leaving the first 10^4 samples as a burn-in period.

In Figure 4.1, we give the histograms of the data sets generated from f_1, f_2 which are overladed with the kernel density estimations (KDE's) based on the predictive samples coming from the PDDP (dashed line) and the rPDDP (solid line) models. The differences between the two models in the quality of the f_1 estimation (panel (a)) are nearly indistinguishable. This is due to the simple form of the mixture f_1. However, on the more complicated mixture density f_2, the randomization of the concentration parameters c_{jl} of the independent DPs, provides us with accurate density estimations (panel (b)). When c_{jl} are kept constant, the PDDP algorithm fails to capture the modes located at $x = -60, -40$ and recognizes a single mode in $x = -50$. The same holds for the modes located at $x = 40, 50$. The Hellinger distances between the true and the estimated densities are given in table 4.1.

Table 4.1: Hellinger distance between the true and the estimated densities obtained from the PDDP ($\mathcal{H}$) and rPDDP ($\mathcal{H}_{\mathcal{R}}$) models.

i	$\mathcal{H}(f_i, \hat{f}_i)$	$\mathcal{H}_{\mathcal{R}}(f_i, \hat{f}_i)$
1	0.11	0.11
2	0.27	0.20

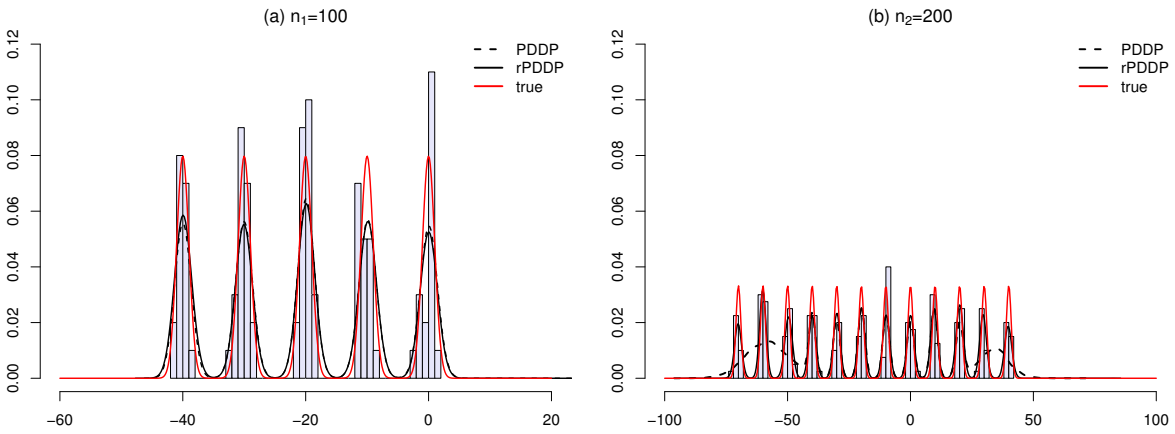

Figure 4.1: Density estimation with the PDDP model (curves in dashed–black) and the rPDDP model (solid black curve) for the $5-12$ mixture based on the samples from the predictive. The true density has been superimposed in red.

4.3 Pairwise dependent geometric stick–breaking process

In this section, we develop the Pairwise Dependent Geometric Stick Breaking Process prior. In order to do so, we let the random density functions $f_j(x)$ to be generated via

$$f_j(x) := f_j(x \,|\, \mathbb{Q}_j) = \sum_{l=1}^{m} p_{jl} g_{jl}(x \,|\, \mathbb{G}_{jl}), \quad \mathbb{Q}_j = \sum_{l=1}^{m} p_{jl} \mathbb{G}_{jl}, \quad 1 \leq j \leq m. \tag{4.9}$$

Now, the $g_{jl}(x) := g_{jl}(x \,|\, \mathbb{G}_{jl}) = \int_\Theta K(x \,|\, \theta) \mathbb{G}_{jl}(\mathrm{d}\theta)$ random densities are independent mixtures of GSB processes, satisfying $g_{jl} = g_{lj}$, under the slightly altered definition

$$\mathbb{G}_{jl} = \sum_{k=1}^{\infty} w_{jlk} \delta_{\theta_{jlk}} \quad \text{with} \quad w_{jlk} = \lambda_{jl}(1-\lambda_{jl})^{k-1},\ \lambda_{jl} \sim h(\cdot \,|\, \xi_{jl}),\ \theta_{jlk} \overset{\text{iid}}{\sim} G_0. \tag{4.10}$$

Following a univariate construction of geometric slice sets (Fuentes-García et al., 2010), we define the stochastic variables $\mathbf{N} = (N_{ji})$ for $1 \leq i \leq n_j$ and $1 \leq j \leq m$, where N_{ji} is an almost surely finite random variable of mass f_N, possibly depending on parameters, associated with the sequential slice set $\mathcal{S}_{ji} = \{1, \ldots, N_{ji}\}$. Following Hatjispyros et al. (2011, 2016), we introduce:

1. The GSB mixture selection variables $\boldsymbol{\delta} = (\delta_{ji})$; for an observation x_{ji} that comes from f_j, δ_{ji} selects the GSB mixture $g_{ji}(x)$ that the observation came from.
2. The clustering variables $\boldsymbol{d} = (d_{ji})$; for an observation x_{ji} that comes from f_j, given δ_{ji}, d_{ji} allocates the component of the GSB mixture $g_{j\delta_{ji}}(x)$ that x_{ji} came from.

Proposition 4.3. *Suppose that the clustering variables (d_{ji}) conditionally on the slice variables (N_{ji}) are having the discrete uniform distribution over the sets $(\mathcal{S}_{ji})$ that is $d_{ji} \,|\, N_{ji} \sim \mathcal{DU}(\mathcal{S}_{ji})$,*

then

$$f_j(x_{ji}, N_{ji} = r) = \frac{1}{r} \sum_{l=1}^{m} p_{jl} f_N(r \mid \lambda_{jl}) \sum_{k=1}^{r} K(x_{ji} \mid \theta_{jlk}), \tag{4.11}$$

and

$$f_j(x_{ji}, N_{ji} = r, d_{ji} = k \mid \delta_{ji} = l) = \frac{1}{r} f_N(r \mid \lambda_{jl})\, \mathcal{I}(k \leq r)\, K(x_{ji} \mid \theta_{jlk}). \tag{4.12}$$

Proof. Starting from the N_{ji}–augmented random densities we have

$$\begin{aligned}
f_j(x_{ji}, N_{ji} = r) &= \sum_{l=1}^{m} f_j(x_{ji}, N_{ji} = r, \delta_{ji} = l) \\
&= \sum_{l=1}^{m} p_{jl}\, f_j(x_{ji}, N_{ji} = r \mid \delta_{ji} = l) \\
&= \sum_{l=1}^{m} p_{jl} \sum_{k=1}^{\infty} f_j(x_{ji}, N_{ji} = r, d_{ji} = k \mid \delta_{ji} = l) \\
&= \sum_{l=1}^{m} p_{jl} f_j(N_{ji} = r \mid \delta_{ji} = l) \\
&\quad \times \sum_{k=1}^{\infty} f_j(d_{ji} = k \mid N_{ji} = r) f_j(x_{ji} \mid d_{ji} = k, \delta_{ji} = l).
\end{aligned}$$

Because $f_j(N_{ji} = r \mid \delta_{ji} = l) = f_N(r \mid \lambda_{jl})$ and $f_j(x_{ji} \mid d_{ji} = k, \delta_{ji} = l) = K(x_{ji} \mid \theta_{jlk})$, the last equation gives

$$\begin{aligned}
f_j(x_{ji}, N_{ji} = r) &= \sum_{l=1}^{m} p_{jl} f_N(r \mid \lambda_{jl}) \sum_{k=1}^{\infty} \frac{1}{r} \mathcal{I}(k \leq r) K(x_{ji} \mid \theta_{jlk}) \\
&= \frac{1}{r} \sum_{l=1}^{m} p_{jl} f_N(r \mid \lambda_{jl}) \sum_{k=1}^{r} K(x_{ji} \mid \theta_{jlk}).
\end{aligned}$$

Augmenting further with the variables d_{ji} and δ_{ji} yields

$$f_j(x_{ji}, N_{ji} = r, d_{ji} = k, \delta_{ji} = l) = \frac{1}{r}\, p_{jl}\, f_N(r \mid \lambda_{jl})\, \mathcal{I}(k \leq r)\, K(x_{ji} \mid \theta_{jlk}).$$

Because $\Pr(\delta_{ji} = l) = p_{jl}$, the last equation leads to eq. (4.12) and the proposition follows. □

The following proposition gives a multivariate analogue of equation (2) in Fuentes-García et al. (2010):

Proposition 4.4. *Given the random set $\mathcal{S}_{ji}$, the random functions in eq.* (4.9) *become finite mixtures of a.s. finite equally weighted mixtures of the $K(\,\cdot \mid \cdot\,)$ probability kernels, that is*

$$f_j(x_{ji} \mid N_{ji} = r) = \sum_{l=1}^{m} \mathcal{W}(r; \lambda_{jl}) \sum_{k=1}^{r} \frac{1}{r} K(x_{ji} \mid \theta_{jlk}), \tag{4.13}$$

with

$$\mathcal{W}(r; \lambda_{jl}) = \frac{p_{jl} f_N(r \mid \lambda_{jl})}{\sum_{l'=1}^{m} p_{jl'} f_N(r; \lambda_{jl'})}.$$

Proof. Marginalizing the joint of x_{ji} and N_{ji} with respect to x_{ji} we obtain

$$f_j(N_{ji} = r) = \sum_{l=1}^{m} p_{jl} f_N(r \,|\, \lambda_{jl}).$$

Then dividing eq. (4.11) with the probability that N_{ji} equals r we obtain eq. (4.13). □

We note that the one–dimensional model introduced in Fuentes-García et al. (2010) under our notation has the representation

$$f_j(x_{ji} \,|\, N_{ji} = r, \delta_{ji} = l) = \sum_{k=1}^{r} \frac{1}{r} K(x_{ji} \,|\, \theta_{jlk}).$$

Marginalizing eq. (4.12) with respect to (N_{ji}, d_{ji}) we obtain

$$f_j(x_{ji} \,|\, \delta_{ji} = l) = \sum_{k=1}^{\infty} \left(\sum_{r=k}^{\infty} \frac{1}{r} f_N(r \,|\, \lambda_{jl}) \right) K(x_{ji} | \theta_{jlk}). \tag{4.14}$$

The quantity inside the parenthesis on the right-hand side of the previous equation is $f_j(d_{ji} \,|\, \delta_{ji} = l)$. Following Fuentes-García et al. (2010), we substitute $f_N(r \,|\, \lambda_{jl})$ with the negative binomial distribution $\mathcal{NB}(r \,|\, 2, \lambda_{jl})$, i.e.

$$f_N(r \,|\, \lambda_{jl}) = r \lambda_{jl}^2 (1 - \lambda_{jl})^{r-1} \mathcal{I}(r \geq 1), \tag{4.15}$$

then eq. (4.14) becomes

$$f_j(x_{ji} \,|\, \delta_{ji} = l) = \sum_{k=1}^{\infty} q_{jlk} K(x_{ji} \,|\, \theta_{jlk}) \;\text{ with }\; q_{jlk} = \lambda_{jl}(1 - \lambda_{jl})^{k-1},$$

and the f_j random densities take the form of a finite mixture of GSB mixtures

$$f_j(x_{ji}) = \sum_{l=1}^{m} p_{jl} \sum_{k=1}^{\infty} q_{jlk} K(x_{ji} \,|\, \theta_{jlk}).$$

We denote the set of observations along the m groups as $\boldsymbol{x} = (x_{ji})$ and with $\boldsymbol{x}_j$ the set of observations in the jth group. The three sets of latent variables in the jth group will be denoted as $\boldsymbol{N}_j$ for the slice variables, $\boldsymbol{d}_j$ for the clustering variables, and finally $\boldsymbol{\delta}_j$ for the set of GSB mixture allocation variables. From now on, we are going to leave the auxiliary variables unspecified; especially for δ_{ji} we use the notation

$$\delta_{ji} = (\delta_{ji}^1, \ldots, \delta_{ji}^m) \in \{\mathbf{e}_1, \ldots, \mathbf{e}_m\} \;\text{ with }\; \Pr(\delta_{ji} = \mathbf{e}_l) = p_{jl},$$

where $\mathbf{e}_l$ denotes the usual basis vector having its only nonzero component equal to 1 at position l. Hence, for a sample of size n_1 from f_1, a sample of size n_2 from f_2, etc., a sample of size n_m

from f_m we can write the full likelihood as a multiple product:

$$f(\boldsymbol{x}, \boldsymbol{N}, \boldsymbol{d} \,|\, \boldsymbol{\delta}) = \prod_{j=1}^{m} f(\boldsymbol{x}_j, \boldsymbol{N}_j, \boldsymbol{d}_j \,|\, \boldsymbol{\delta}_j)$$
$$= \prod_{j=1}^{m} \prod_{i=1}^{n_j} \mathcal{I}(d_{ji} \leq N_{ji}) \prod_{l=1}^{m} \left\{ \lambda_{jl}^2 (1-\lambda_{jl})^{N_{ji}-1} K(x_{ji} \,|\, \theta_{jld_{ji}}) \right\}^{\delta_{ji}^l}.$$

In a hierarchical fashion, using the auxiliary variables, we have for $j = 1, \ldots, m$ and $i = 1, \ldots, n_j$,

$$x_{ji}, N_{ji} \,|\, d_{ji}, \delta_{ji}, (\theta_{jr\delta_{ji}})_{1\leq r\leq m}, \lambda_{j\delta_{ji}} \overset{\text{ind}}{\sim} \prod_{r=1}^{m} \left\{ \lambda_{jl}^2 (1-\lambda_{jl})^{N_{ji}-1} K(x_{ji} \,|\, \theta_{jrd_{ji}}) \right\}^{\delta_{ji}^r} \mathcal{I}(N_{ji} \geq d_{ji})$$
$$d_{ji} \,|\, N_{ji} \overset{\text{ind}}{\sim} \mathcal{DU}(\mathcal{S}_{ji}), \;\; \Pr(\delta_{ji} = \mathbf{e}_l) = p_{jl}$$
$$q_{jik} = \lambda_{ji}(1-\lambda_{ji})^{k-1}, \;\; \theta_{jik} \overset{\text{iid}}{\sim} P_0, \;\; k \in \mathbb{N}.$$

4.3.1 The PDGSBP covariance and correlation

In this subsection, we find the covariance and the correlation between $f_j(x)$ and $f_i(x)$. First, we provide the following lemma.

Lemma 4.1. *Let $g_{\mathbb{G}}(x) = \int_\Theta K(x \,|\, \theta)\mathbb{G}(\mathrm{d}\theta)$ be a random density, with $\mathbb{G} = \lambda \sum_{j=1}^{\infty} (1-\lambda)^{j-1} \delta_{\theta_j}$ and $\theta_j \overset{\text{iid}}{\sim} G_0$, then*

$$\mathbb{E}\left\{ g_{\mathbb{G}}(x)^2 \right\} = \left(\frac{1}{2-\lambda} \right) \left\{ \lambda \int_\Theta K(x \,|\, \theta)^2 G_0(\mathrm{d}\theta) + 2(1-\lambda) \left(\int_\Theta K(x \,|\, \theta) G_0(\mathrm{d}\theta) \right)^2 \right\}.$$

Proof. Because $g_{\mathbb{G}}(x) = \lambda \sum_{j=1}^{\infty} (1-\lambda)^{j-1} K(x \,|\, \theta_j)$, we have

$$\mathbb{E}\left\{ g_{\mathbb{G}}(x)^2 \right\} = \lambda^2 \, \mathbb{E}\left\{ \left(\sum_{j=1}^{\infty} (1-\lambda)^{j-1} K(x \,|\, \theta_j) \right)^2 \right\}$$
$$= \lambda^2 \left\{ \sum_{j=1}^{\infty} (1-\lambda)^{2j-2} \, \mathbb{E}\left[K(x \,|\, \theta_j)^2 \right] + 2 \sum_{k=2}^{\infty} \sum_{j=1}^{k-1} (1-\lambda)^{j+k-2} \, \mathbb{E}[K(x \,|\, \theta_j) K(x \,|\, \theta_k)] \right\}$$
$$= \lambda^2 \left\{ \sum_{j=1}^{\infty} (1-\lambda)^{2j-2} \mathbb{E}\left[K(x \,|\, \theta)^2 \right] + 2 \sum_{k=2}^{\infty} \sum_{j=1}^{k-1} (1-\lambda)^{j+k-2} \mathbb{E}[K(x \,|\, \theta)]^2 \right\}$$
$$= \lambda^2 \left\{ \frac{1}{\lambda(2-\lambda)} \mathbb{E}\left[K(x \,|\, \theta)^2 \right] + 2 \, \frac{1-\lambda}{\lambda^2(2-\lambda)} \mathbb{E}[K(x \,|\, \theta)]^2 \right\},$$

which gives the desired result. □

Proposition 4.5. *It is that*

$$\mathrm{Cov}(f_j(x), f_i(x)) \;=\; p_{ji}\, p_{ij} \mathrm{Var}\left(\int_\Theta K(x \,|\, \theta) \mathbb{G}_{ji}(\mathrm{d}\theta) \right), \tag{4.16}$$

with

$$\mathrm{Var}\left(\int_\Theta K(x \,|\, \theta) \mathbb{G}_{ji}(\mathrm{d}\theta) \right) = \frac{\lambda_{ji}}{2 - \lambda_{ji}} \mathrm{Var}(K(x \,|\, \theta)). \tag{4.17}$$

Proof. The random densities $f_i(x) = \sum_{l=1}^m p_{il}\, g_{il}(x)$ and $f_j(x) = \sum_{l=1}^m p_{jl}\, g_{jl}(x)$ depend to each other through the random measure $\mathbb{G}_{ji}$, therefore

$$\mathbb{E}[f_i(x) f_j(x)] = \mathbb{E}[\,\mathbb{E}(f_i(x) f_j(x) | \mathbb{G}_{ji})\,] = \mathbb{E}\{\,\mathbb{E}[f_i(x) \,|\, \mathbb{G}_{ji}]\, \mathbb{E}[f_j(x) \,|\, \mathbb{G}_{ji}]\,\}, \tag{4.18}$$

and

$$\begin{aligned}
\mathbb{E}[f_j(x) \,|\, \mathbb{G}_{ji}] &= \sum_{l \neq i} p_{jl}\, \mathbb{E}[g_{jl}(x)] + p_{ji} g_{ji}(x) = (1 - p_{ji})\, \mathbb{E}[K(x \,|\, \theta)] + p_{ji} g_{ji}(x) \\
\mathbb{E}[f_i(x) \,|\, \mathbb{G}_{ji}] &= \sum_{l \neq j} p_{il}\, \mathbb{E}[g_{il}(x)] + p_{ij} g_{ji}(x) = (1 - p_{ij})\, \mathbb{E}[K(x \,|\, \theta)] + p_{ij} g_{ji}(x)\,.
\end{aligned}$$

Substituting back to equation (4.18) one obtains

$$\mathbb{E}[f_i(x) f_j(x)] = (1 - p_{ij} p_{ji})\, \mathbb{E}[K(x \,|\, \theta)]^2 + p_{ij} p_{ji}\, \mathbb{E}\left[g_{ji}(x)^2 \right].$$

Using lemma 4.1, the last equation becomes

$$\mathbb{E}[f_i(x) f_j(x)] = \frac{\lambda_{ji} p_{ji} p_{ij}}{2 - \lambda_{ji}} \left\{ \mathbb{E}[K(x \,|\, \theta)^2] - \mathbb{E}[K(x \,|\, \theta)]^2 \right\} + \mathbb{E}[K(x \,|\, \theta)]^2,$$

or that

$$\mathrm{Cov}(f_j(x), f_i(x)) \;=\; \frac{\lambda_{ji} p_{ji}\, p_{ij}}{2 - \lambda_{ji}} \mathrm{Var}(K(x \,|\, \theta)).$$

The desired result, comes from the fact that

$$\begin{aligned}
\mathrm{Var}\left(\int_\Theta K(x \,|\, \theta) \mathbb{G}_{ji}(\mathrm{d}\theta) \right) &= \left\{ \frac{\lambda_{ji}}{2 - \lambda_{ji}} \mathbb{E}[K(x \,|\, \theta)^2] + \frac{2(1 - \lambda_{ji})}{2 - \lambda_{ji}} \mathbb{E}[K(x \,|\, \theta)]^2 \right\} - \mathbb{E}[K(x \,|\, \theta)]^2 \\
&= \frac{\lambda_{ji}}{2 - \lambda_{ji}} \left(\mathbb{E}[K(x \,|\, \theta)^2] - \mathbb{E}[K(x \,|\, \theta)]^2 \right).
\end{aligned}$$

□

Suppose now that $(f_j^{\mathcal{D}}(x))_{1 \leq j \leq m}$ and $(f_j^{\mathcal{G}}(x))_{1 \leq j \leq m}$, are two collections of m DP and m GSB pairwise dependent random densities respectively, i.e. $f_j^{\mathcal{D}}(x) = \sum_{l=1}^m p_{jl} g_{jl}^{\mathcal{D}}(x)$ with $g_{jl}^{\mathcal{D}}(x) = g_{jl}(x \,|\, \mathbb{P}_{jl})$, and $f_j^{\mathcal{G}}(x) = \sum_{l=1}^m p_{jl} g_{jl}^{\mathcal{G}}(x)$ with $g_{jl}^{\mathcal{G}}(x) = g_{jl}(x \,|\, \mathbb{G}_{jl})$. Then we have the following proposition:

Proposition 4.6. *For given parameters* (λ_{ji}), (c_{ji}), *and matrix of selection probabilities* (p_{ji}) *it is that*

1. *The PDGSBP and rPDDP correlations are given by*

$$\operatorname{Corr}(f_j^{\mathcal{G}}(x), f_i^{\mathcal{G}}(x)) = \frac{\lambda_{ji} p_{ji} p_{ij}}{2-\lambda_{ji}} \left(\sum_{l=1}^{m} \sum_{r=1}^{m} \frac{p_{jl}^2 p_{ir}^2 \lambda_{jl} \lambda_{ir}}{(2-\lambda_{jl})(2-\lambda_{ir})} \right)^{-1/2}, \tag{4.19}$$

and

$$\operatorname{Corr}(f_j^{\mathcal{D}}(x), f_i^{\mathcal{D}}(x)) = \frac{p_{ji} p_{ij}}{1+c_{ji}} \left(\sum_{l=1}^{m} \sum_{r=1}^{m} \frac{p_{jl}^2 p_{ir}^2}{(1+c_{jl})(1+c_{ir})} \right)^{-1/2}. \tag{4.20}$$

2. *When* $\lambda_{ji} = \lambda$ *and* $c_{ji} = c$ *for all* $1 \leq j \leq i \leq m$, *the expressions for the rPDDP and PDGSBP correlations simplify to*

$$\operatorname{Corr}(f_j^{\mathcal{G}}(x), f_i^{\mathcal{G}}(x)) = \operatorname{Corr}(f_j^{\mathcal{D}}(x), f_i^{\mathcal{D}}(x)) = p_{ji} p_{ij} \left(\sum_{l=1}^{m} \sum_{r=1}^{m} p_{jl}^2 p_{ir}^2 \right)^{-1/2}.$$

Proof. *1.* From eq. (4.17) and proposition 4.5, we have that

$$\operatorname{Var}(f_j^{\mathcal{G}}(x)) = \operatorname{Var}\left(\sum_{l=1}^{m} p_{jl} g_{jl}^{\mathcal{G}}(x) \right) = \sum_{l=1}^{m} \frac{p_{ji}^2 \lambda_{ji}}{2-\lambda_{ji}} \operatorname{Var}(K(x \,|\, \theta)).$$

Normalizing the covariance in eq. (4.16) with the associated standard deviations, yields

$$\operatorname{Corr}(f_j^{\mathcal{G}}(x), f_i^{\mathcal{G}}(x)) = \frac{\lambda_{ji} p_{ji} p_{ij}}{2-\lambda_{ji}} \left(\sum_{l=1}^{m} \sum_{r=1}^{m} \frac{p_{jl}^2 p_{ir}^2 \lambda_{jl} \lambda_{ir}}{(2-\lambda_{jl})(2-\lambda_{ir})} \right)^{-1/2}. \tag{4.21}$$

Similarly, from proposition 1 in Hatjispyros et al. (2011), it is that

$$\operatorname{Var}(f_j^{\mathcal{D}}(x)) = \sum_{l=1}^{m} \frac{p_{ji}^2}{1+c_{ji}} \operatorname{Var}(K(x \,|\, \theta)),$$

and

$$\operatorname{Corr}(f_j^{\mathcal{D}}(x), f_i^{\mathcal{D}}(x)) = \frac{p_{ji} p_{ij}}{1+c_{ji}} \left(\sum_{l=1}^{m} \sum_{r=1}^{m} \frac{p_{jl}^2 p_{ir}^2}{(1+c_{jl})(1+c_{ir})} \right)^{-1/2}. \tag{4.22}$$

2. When $\lambda_{ji} = \lambda$ and $c_{ji} = c$ for all $1 \leq j \leq i \leq m$, from eqs. (4.21) and (4.22), it is clear that

$$\operatorname{Corr}(f_j^{\mathcal{G}}(x), f_i^{\mathcal{G}}(x)) = \operatorname{Corr}(f_j^{\mathcal{D}}(x), f_i^{\mathcal{D}}(x)) = p_{ji} p_{ij} \left(\sum_{l=1}^{m} \sum_{r=1}^{m} p_{jl}^2 p_{ir}^2 \right)^{-1/2}.$$

□

It is clear that, irrespective of the model, the random densities $f_j(x)$ and $f_i(x)$ are positively correlated whenever $p_{ji} = p_{ij} = 1$. Similarly, the random densities $f_j(x)$ and $f_i(x)$ are independent (have no common part) whenever $p_{ji} = p_{ij} = 0$. Another, less obvious feature, upon synchronization, is the ability of controlling the correlation among the models. For example, suppose that for $m = 2$, the random densities $f_1(x)$ and $f_2(x)$ are dependent, and that $\lambda_{ji} = (1 + c_{ji})^{-1}$; then consider the expression

$$D_{12} := \lambda_{12}^2 \, p_{12}^2 \, p_{21}^2 \, \left\{ \mathrm{Corr}(f_1^{\mathcal{G}}(x), f_2^{\mathcal{G}}(x))^{-2} - \mathrm{Corr}(f_1^{\mathcal{D}}(x), f_2^{\mathcal{D}}(x))^{-2} \right\}.$$

Since correlations are positive, $D_{12} \geq 0$ whenever $\mathrm{Corr}(f_1^{\mathcal{G}}(x), f_2^{\mathcal{G}}(x)) \leq \mathrm{Corr}(f_1^{\mathcal{D}}(x), f_2^{\mathcal{D}}(x))$, and that $D_{12} < 0$ whenever $\mathrm{Corr}(f_1^{\mathcal{G}}(x), f_2^{\mathcal{G}}(x)) > \mathrm{Corr}(f_1^{\mathcal{D}}(x), f_2^{\mathcal{D}}(x))$. Then, it not difficult to see that

$$D_{12} = \left(p_{12}^2 \lambda_{12} + r_1 p_{11}^2 \lambda_{11}\right) \left(p_{21}^2 \lambda_{12} + r_2 p_{22}^2 \lambda_{22}\right) - \left(p_{12}^2 \lambda_{12} + p_{11}^2 \lambda_{11}\right) \left(p_{21}^2 \lambda_{12} + p_{22}^2 \lambda_{22}\right)$$

with $r_k = (2 - \lambda_{12})/(2 - \lambda_{kk})$, $k = 1, 2$. We have the following cases:

1. $\lambda_{12} > \max\{\lambda_{11}, \lambda_{22}\} \Leftrightarrow r_1 < 1, r_2 < 1 \Leftrightarrow \mathrm{Corr}(f_1^{\mathcal{G}}(x), f_2^{\mathcal{G}}(x)) > \mathrm{Corr}(f_1^{\mathcal{D}}(x), f_2^{\mathcal{D}}(x))$.
2. $\lambda_{12} < \min\{\lambda_{11}, \lambda_{22}\} \Leftrightarrow r_1 > 1, r_2 > 1 \Leftrightarrow \mathrm{Corr}(f_1^{\mathcal{G}}(x), f_2^{\mathcal{G}}(x)) < \mathrm{Corr}(f_1^{\mathcal{D}}(x), f_2^{\mathcal{D}}(x))$.
3. $\lambda_{12} = \lambda_{11} = \lambda_{22} \Leftrightarrow r_1 = r_2 = 1 \Leftrightarrow \mathrm{Corr}(f_1^{\mathcal{G}}(x), f_2^{\mathcal{G}}(x)) = \mathrm{Corr}(f_1^{\mathcal{D}}(x), f_2^{\mathcal{D}}(x))$.

4.3.2 The PDGSBP Gibbs Sampler

In this section, we are going to describe the PDGSBP Gibbs sampler for estimating the model. At each iteration we will sample variables,

$$\begin{aligned} &\theta_{jlk}, 1 \leq j \leq l \leq m, \, 1 \leq k \leq N^*, \\ &d_{ji}, N_{ji}, \delta_{ji}, 1 \leq j \leq m, \, 1 \leq i \leq n_j, \\ &p_{jl}, 1 \leq j \leq m, 1 \leq l \leq m, \end{aligned}$$

with $N^* = \max_{j,i} N_{ji}$ almost surely finite.

1. For the locations of the random measures for $k = 1, \ldots, d^*$ where $d^* = \max_{j,i} d_{ji}$, it is that

$$f(\theta_{jlk} \mid \cdots) \propto f(\theta_{jlk}) \begin{cases} \prod_{i=1}^{n_j} K(x_{ji} \mid \theta_{jlk})^{\mathcal{I}(\delta_{ji} = \mathbf{e}_l, \, d_{ji} = k)} \prod_{i=1}^{n_l} K(x_{li} \mid \theta_{jlk})^{\mathcal{I}(\delta_{li} = \mathbf{e}_j, \, d_{li} = k)} & l > j, \\ \prod_{i=1}^{n_j} K(x_{ji} \mid \theta_{jjk})^{\mathcal{I}(\delta_{ji} = \mathbf{e}_j, \, d_{ji} = k)} & l = j. \end{cases}$$

2. Here, we sample the allocation variables d_{ji} and the mixture component indicator variables δ_{ji} as a block. For $j = 1, \ldots, m$ and $i = 1, \ldots, n_j$, we have

$$\Pr(d_{ji} = k, \delta_{ji} = \mathbf{e}_l \mid N_{ji} = r, \cdots) \propto p_{jl} \, K(x_{ji} \mid \theta_{jlk}) \, \mathcal{I}(l \leq m) \, \mathcal{I}(k \leq r).$$

3. The slice variables N_{ji} have full conditional distributions given by

$$\Pr(N_{ji} = r \,|\, \delta_{ji} = \mathbf{e}_l, d_{ji} = l, \cdots) \propto (1 - \lambda_{jl})^r \, \mathcal{I}(r \geq l),$$

which are truncated geometric distributions over the set $\{l, l+1, \ldots\}$.

4. The full conditional, for $j = 1, \ldots, m$, for the selection probabilities $\boldsymbol{p}_j = (p_{j1}, \ldots, p_{jm})$, under a Dirichlet prior $f(\boldsymbol{p}_j \,|\, \boldsymbol{a}_j) \propto \prod_{l=1}^m p_{jl}^{a_{jl}-1}$, with hyperparameter $\boldsymbol{a}_j = (a_{j1}, \ldots, a_{jm})$, is Dirichlet

$$f(\boldsymbol{p}_j \,|\, \cdots) \propto \prod_{l=1}^m p_{jl}^{a_{jl} + \sum_{i=1}^{n_l} \mathcal{I}(\delta_{ji} = \mathbf{e}_l) - 1}.$$

5. Here, we update the geometric probabilities (λ_{jl}) of the GSB measures. For $1 \leq j \leq l \leq m$, it is that

$$f(\lambda_{jl} \,|\, \cdots) \propto f(\lambda_{jl}) \begin{cases} \prod_{i=1}^{n_j} \left\{\lambda_{jl}^2 (1-\lambda_{jl})^{N_{ji}-1}\right\}^{\mathcal{I}(\delta_{ji} = \mathbf{e}_l)} \prod_{i=1}^{n_l} \left\{\lambda_{jl}^2 (1-\lambda_{jl})^{N_{li}-1}\right\}^{\mathcal{I}(\delta_{li} = \mathbf{e}_j)} & l > j, \\ \prod_{i=1}^{n_j} \left\{\lambda_{jj}^2 (1-\lambda_{jj})^{N_{ji}-1}\right\}^{\mathcal{I}(\delta_{ji} = \mathbf{e}_j)} & l = j. \end{cases}$$

To complete the model, we assign priors to the geometric probabilities. For a fair comparison between the two models, we apply $\lambda_{jl} = (1 + c_{jl})^{-1}$ transformed priors. So, by placing gamma priors $c_{jl} \sim \mathcal{G}(a_{jl}, b_{jl})$ over the concentration masses c_{jl} of the PDDP model, we have

$$f(\lambda_{jl}) = \mathcal{TG}(\lambda_{jl} \,|\, a_{jl}, b_{jl}) \propto \lambda_{jl}^{-(a_{jl}+1)} e^{-b_{jl}/\lambda_{jl}} (1 - \lambda_{jl})^{a_{jl}-1} \, \mathcal{I}(0 < \lambda_{jl} < 1). \tag{4.23}$$

Conditionally on the mixture allocation variable δ_{jl}, the geometric probability λ_{jl} can be sampled with the auxiliary variable method described in Appendix A.2

4.4 Experiments

In this section, we illustrate the efficiency of the PDGSBP model. For the choice of a normal kernel (unless otherwise specified) $K(x|\theta) = \mathcal{N}(x|\theta)$ where $\theta = (\mu, \tau^{-1})$ and $\tau = \sigma^{-2}$ is the precision. The prior over the means and precisions of the PDGSBP (G_0) and the rPDDP model (P_0) is the independent normal-gamma measure, given by

$$P_0(\mathrm{d}\mu, \mathrm{d}\tau) = G_0(\mathrm{d}\mu, \mathrm{d}\tau) = \mathcal{N}(\mu \,|\, \mu_0, \tau_0^{-1}) \, \mathcal{G}(\tau \,|\, \epsilon_1, \epsilon_2) \, \mathrm{d}\mu \mathrm{d}\tau.$$

Attempting a noninformative prior specification (unless otherwise specified), we took $\mu_0 = 0$ and $\tau_0 = \epsilon_1 = \epsilon_2 = 10^{-3}$. For the concentration masses of the rPDDP model, a–priori, we set $c_{jl} \sim \mathcal{G}(a_{jl}, b_{jl})$. For an objective evaluation of the execution time, of the two algorithms under different scenarios, we choose a synchronized prior specification, namely, for the geometric probabilities, we set $\lambda_{jl} \sim \mathcal{TG}(a_{jl}, b_{jl})$ - the transformed gamma density given in eq. (4.23). In section 4.2.1, we have shown that such prior specifications are valid for $a_{jl} > 1$. In all our numerical examples, we took $a_{jl} = b_{jl} = 1.1$. For our numerical experiments (unless otherwise specified), the hyperparameters (α_{jl}) of the Dirichlet priors over the matrix of the selection probabilities $p = (p_{jl})$ has been set to $\alpha_{jl} = 1$.

As a measure of accuracy of the proposed methodologies, we measure the similarity between probability distributions with the Hellinger distance. So for example, $\mathcal{H}_{\mathcal{G}}(f, \hat{f})$ and $\mathcal{H}_{\mathcal{D}}(f, \hat{f})$, will denote the Hellinger distance between the true density f and the predictive density $\hat{f}$ of the PDGSBP and rPDDP algorithms, respectively. The Gibbs samplers run for 11×10^4 iterations leaving the first 10^4 samples as a burn-in period.

4.4.1 Time execution efficiency of the PDGSBP model

Nested normal mixtures with a unimodal common and idiosyncratic part: Here, we choose to include all pairwise and idiosyncratic dependences in the form of unimodal equally weighted normal mixture components. The mixture components are well separated with unit variance. We define each data model $\mathcal{M}_m = \{f_j^{(m)} : 1 \leq j \leq m\}$ of dimension $m \in \{2, 3, 4\}$, based on a 4×10 matrix $M = (M_{jk})$, with entries in the set $\{0, 1\}$, having at most two ones in each column and exactly four ones in each row. When there is exactly one entry of one, the column defines an idiosyncratic part. The appearance of exactly two ones in a column defines a common component. We let the matrix M given by

$$M = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix},$$

and for $m \in \{2, 3, 4\}$, we define

$$\mathcal{M}_m : f_j^{(m)}(x) \propto \sum_{k=5-m}^{2(m+1)} M_{jk} \mathcal{N}(x|10(k-6), 1), \ 1 \leq j \leq m,$$

We are taking independently samples of sizes $n_j^{(2)} = 60$ from the $f_j^{(2)}$'s, $n_j^{(3)} = 120$ from the $f_j^{(3)}$'s, and, $n_j^{(4)} = 200$ from the $f_j^{(4)}$'s. In all cases, the PDGSBP and the rPDDP density estimations are of the same quality.

In Figure 4.2 (a)–(d), we give the histograms of the data sets for the specific case $m = 4$, which are overladed with the kernel density estimations (KDE's) based on the predictive samples of the $f_j^{(4)}$'s coming from the PDGSBP (solid line) and the rPDDP (dashed line) models. The differences between the two models are nearly indistinguishable. The Hellinger distances between the true and the estimated densities for the case $m = 4$ is given in table 4.2.

In table 4.3, we summarize the mean execution times (MET's) per 10^3 iterations in seconds. The PDGSBP sampler is about three times faster than the rPDDP sampler. The corresponding MET ratios for $m = 2, 3$ and 4 are $2.96, 3.04$ and 3.37 respectively. We can see that the PDGSBP Gibbs sampler gives slightly faster execution times with increasing m. This will become more clear in our next simulated data example, where the average sample size per mode is being kept constant.

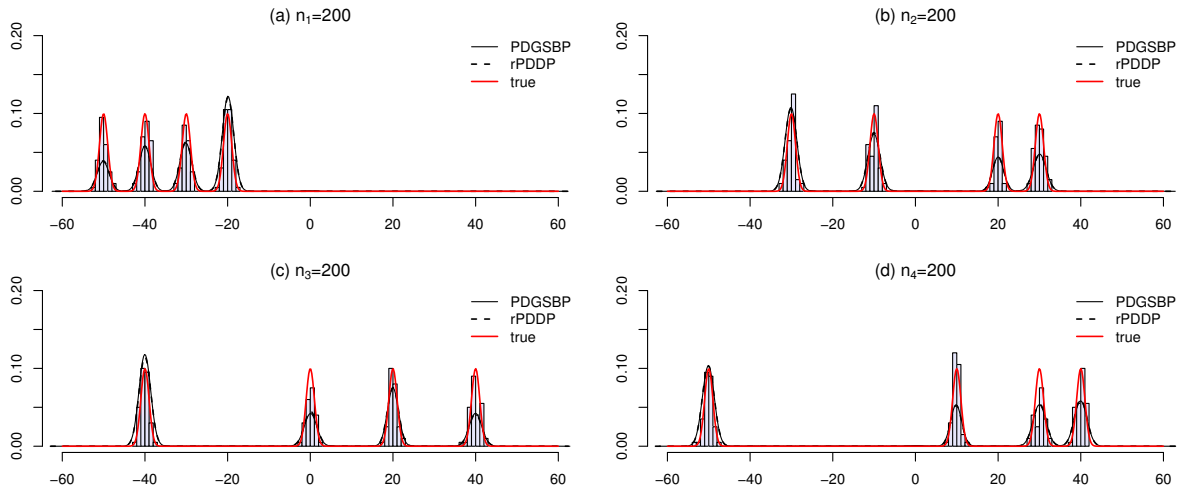

Figure 4.2: Histograms of data sets coming for the case $m = 4$. The superimposed KDE's are based on the predictive samples obtained from the PDGSBP and the rPDDP models.

Table 4.2: Hellinger distances for the case $m = 4$.

i	$\mathcal{H}_{\mathcal{G}}(f_i^{(4)}, \hat{f}_i^{(4)})$	$\mathcal{H}_{\mathcal{D}}(f_i^{(4)}, \hat{f}_i^{(4)})$
1	0.17	0.17
2	0.19	0.18
3	0.22	0.22
4	0.20	0.20

Table 4.3: Mean execution times in seconds per 10^3 iterations.

m	Model	Sample size	MET
2	PDGSBP	$n_j^{(2)} = 60$	0.57
	rPDDP		1.68
3	PDGSBP	$n_j^{(3)} = 120$	2.16
	rPDDP		6.57
4	PDGSBP	$n_j^{(4)} = 200$	5.30
	rPDDP		17.87

Sparse m–scalable data set models: In this example, we attempt to create m-scalable normal mixture data sets of the lowest possible sample size. To this respect, we sample independently m groups of data sets from the densities

$$f_j^{(m)}(x) \propto \mathcal{N}(x \,|\, (j-1)\xi, 1)\, \mathcal{I}(1 \leq j < m) + \sum_{k=1}^{m-1} \mathcal{N}(x \,|\, (k-1)\,\xi, 1)\, \mathcal{I}(j = m),$$

with sample sizes $n_j^{(m)} = n\{\mathcal{I}(1 \leq j < m) + (m-1)\,\mathcal{I}(j = m)\}$. We have chosen $\xi = 10$ and an average sample size per mode of $n = 20$, for $m \in \{2, \ldots, 10\}$.

In Figure 4.3, we depict the average execution times as functions of the dimension m. We can see how fast the two MET-curves diverge with increasing m. In Figure 4.4(a)–(j), for the specific case $m = 10$, we give the histograms of the data sets, overladed with the KDE's based on the predictive samples of the $f_j^{(10)}$'s coming from the PDGSBP (solid line) and the rPDDP (dashed line) models. We can see that the PDGSBP and the rPDDP density estimations are of the same quality.

The Hellinger distances, between the true and the estimated densities for the specific case $m = 10$, are given in Table 4.4. The large values of the Hellinger distances $\mathcal{H}_{\mathcal{G}}(f_{10}^{(10)}, \hat{f}_{10}^{(10)}) \approx \mathcal{H}_{\mathcal{D}}(f_{10}^{(10)}, \hat{f}_{10}^{(10)}) \approx 0.22$, are caused by the enlargement of the variances of the underrepresented modes due to the small sample size.

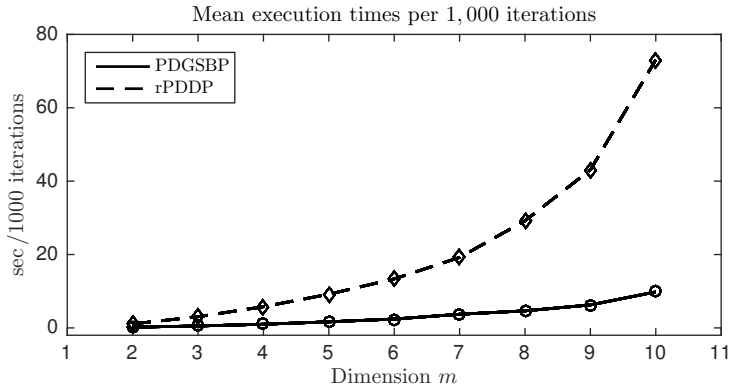

Figure 4.3: Mean execution times for the two models, based on the sparse m-scalable data sets.

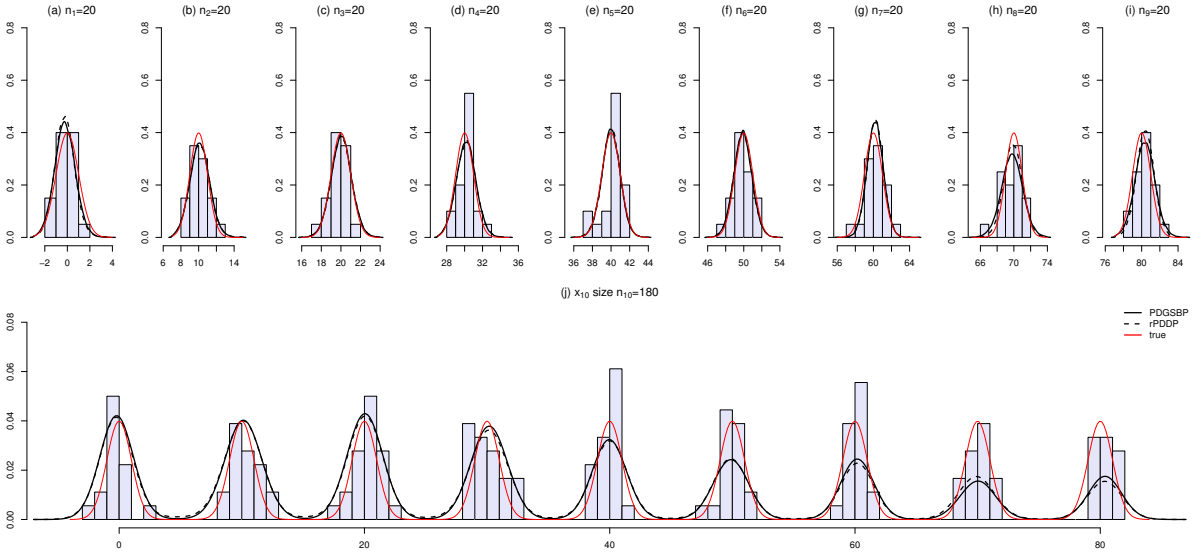

Figure 4.4: Histograms of sparse m-scalable data sets for the case $m = 10$. The superimposed KDE's are based on the predictive samples of the PDGSBP and the rPDDP models.

Table 4.4: Hellinger distances between true and estimated densities for the case $m = 10$ of the sparse scalable data example.

i	1	2	3	4	5	6	7	8	9	10
$\mathcal{H}_{\mathcal{G}}(f_i^{(10)}, \hat{f}_i^{(10)})$	0.08	0.10	0.09	0.14	0.14	0.13	0.14	0.09	0.11	0.22
$\mathcal{H}_{\mathcal{D}}(f_i^{(10)}, \hat{f}_i^{(10)})$	0.09	0.11	0.10	0.15	0.12	0.10	0.14	0.09	0.09	0.22

4.4.2 Normal and gamma mixture models that are not well separated

The normal mixture example: We will first consider a normal model for $m = 2$, first appeared in Lijoi et al. (2014). The data models for f_1 and f_2 are 7-mixtures. Their common part is a 4-mixture that is weighed differently between the two mixtures. More specifically, we sample two data sets of sample size $n_1 = n_2 = 200$, independently from

$$(f_1, f_2) = \left(\frac{1}{2}\, g_{11} + \frac{1}{2}\, g_{12},\ \frac{4}{7}\, g_{21} + \frac{3}{7}\, g_{22} \right),$$

with

$$\begin{aligned}
g_{11} &= \frac{2}{7}\mathcal{N}(-8, 0.25^2) + \frac{3}{7}\mathcal{N}(1, 0.5^2) + \frac{2}{7}\mathcal{N}(10, 1) \\
g_{12} &= \frac{1}{7}\mathcal{N}(-10, 0.5^2) + \frac{3}{7}\mathcal{N}(-3, 0.75^2) + \frac{1}{7}\mathcal{N}(3, 0.25^2) + \frac{2}{7}\mathcal{N}(7, 0.25^2) \\
g_{21} &= \frac{2}{8}\mathcal{N}(-10, 0.5^2) + \frac{3}{8}\mathcal{N}(-3, 0.75^2) + \frac{2}{8}\mathcal{N}(3, 0.25^2) + \frac{1}{8}\mathcal{N}(7, 0.25^2) \\
g_{22} &= \frac{1}{3}\mathcal{N}(-6, 0.5^2) + \frac{1}{3}\mathcal{N}(-1, 0.25^2) + \frac{1}{3}\mathcal{N}(5, 0.5^2).
\end{aligned}$$

For this case, a–priori we took $(\mu_0, \tau_0, \epsilon_1, \epsilon_2) = (0, 10^{-3}, 1, 10^{-2})$. In Figure 4.5(a)–(b), we give the histograms of the data sets, with the predictive densities of the PDGSBP and rPDDP models superimposed in black solid and black dashed curves, respectively. We can see that the PDGSBP and the rPDDP density estimations are of the same quality.

In Table 4.5, we give the Hellinger distance between the true and the estimated densities

Table 4.5: Hellinger distance between the true and the estimated densities.

i	$\mathcal{H}_{\mathcal{G}}(f_i, \hat{f}_i)$	$\mathcal{H}_{\mathcal{D}}(f_i, \hat{f}_i)$
1	0.19	0.18
2	0.18	0.15

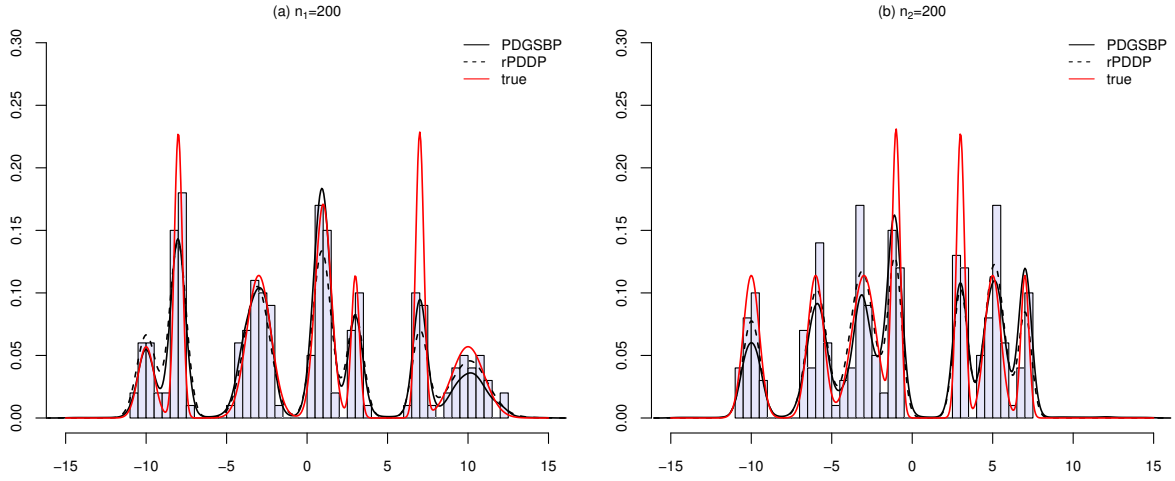

Figure 4.5: Density estimations of the 7-mixtures data sets, under the PDGSBP and the rPDDP models. The true densities have been superimposed in red.

The gamma mixture example: In this example we took $m = 2$. The data models for f_1 and f_2 are gamma 4-mixtures. The common part is a gamma 2-mixture, weighted identically among the two mixtures. More specifically, we sample two data sets of sample size $n_1 = n_2 = 160$, independently from

$$(f_1, f_2) = \left(\frac{2}{5}\, g_{11} + \frac{3}{5}\, g_{12},\ \frac{7}{10}\, g_{12} + \frac{3}{10}\, g_{22}\right),$$

with

$$\begin{aligned} g_{11} &= \frac{2}{3}\mathcal{G}(2, 1.1) + \frac{1}{3}\mathcal{G}(80, 2) \\ g_{12} &= \frac{8}{14}\mathcal{G}(10, 0.9) + \frac{6}{14}\mathcal{G}(200, 8.1) \\ g_{22} &= \frac{2}{3}\mathcal{G}(105, 3) + \frac{1}{3}\mathcal{G}(500, 10), \end{aligned}$$

Because we want to estimate the density of non negative observations, we find it more appropriate to take the kernel to be a log-normal distribution (Hatjispyros et al., 2017b). That is $K(x|\theta) = \mathcal{LN}(x|\theta)$ with $\theta = (\mu, \sigma^2)$, is the log-normal density with mean $\exp(\mu + \sigma^2/2)$. For this case, a-priori we set

$$(\mu_0, \tau_0, \epsilon_1, \epsilon_2) = (\bar{S}, 0.5, 2, 0.01), \quad \bar{S} = \frac{1}{n_1 + n_2}\left(\sum_{j=1}^{n_1} \log x_{1j} + \sum_{j=1}^{n_2} \log x_{2j}\right).$$

In Figure 4.6(a)–(b), we display the KDE's based on the predictive samples of the two models. We can see that the PDGSBP and the rPDDP density estimations are of the same quality. In Table 4.6 we give the Hellinger distances.

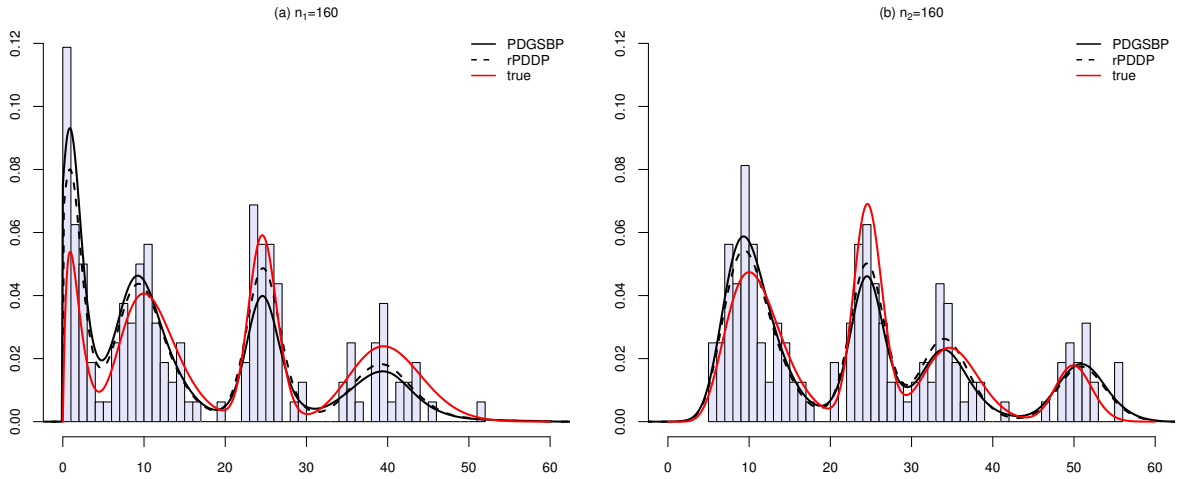

Figure 4.6: The KDE's are based on the predictive sample of the PDGSBP model (solid curve in black) and the predictive sample of the rPDDP model (dashed curve in black).

Table 4.6: Hellinger distances for the gamma mixture data model.

i	$\mathcal{H}_{\mathcal{G}}(f_i, \hat{f}_i)$	$\mathcal{H}_{\mathcal{D}}(f_i, \hat{f}_i)$
1	0.13	0.11
2	0.19	0.18

Because the common part is equally weighted among f_1 and f_2, it makes sense to display the estimations of the selection probability matrices under the two models

$$\mathbb{E}_{\mathcal{G}}(p \,|\, (x_{ji})) = \begin{pmatrix} 0.42 & 0.58 \\ 0.64 & 0.36 \end{pmatrix}, \quad \mathbb{E}_{\mathcal{D}}(p \,|\, (x_{ji})) = \begin{pmatrix} 0.42 & 0.58 \\ 0.69 & 0.31 \end{pmatrix}, \quad p_{\text{true}} = \begin{pmatrix} 0.4 & 0.6 \\ 0.7 & 0.3 \end{pmatrix}.$$

4.4.3 Borrowing of strength of the PDGSBP model

In this example we consider three populations $\{D_j^{(s)} : j = 1, 2, 3\}$, under three different scenarios $s \in \{1, 2, 3\}$. The sample sizes are always the same, namely, $n_1 = 200$, $n_2 = 50$ and $n_3 = 200$ – the second population is sampled only once. The three data sets $D_1^{(s)}$, $D_2^{(s)}$ and $D_3^{(s)}$, are sampled independently from the normal mixtures

$$(f_1^{(s)}, f_2^{(s)}, f_3^{(s)}) = \left((1 - q^{(s)})f + q^{(s)}g_1,\ f,\ (1 - q^{(s)})f + q^{(s)}g_2\right),$$

where

$$\begin{aligned} f &= \frac{3}{10}\mathcal{N}(-10, 1) + \frac{2}{10}\mathcal{N}(-6, 1) + \frac{2}{10}\mathcal{N}(6, 1) + \frac{3}{10}\mathcal{N}(10, 1) \\ g_1 &= \frac{1}{2}\mathcal{N}(-4, 1) + \frac{1}{2}\mathcal{N}(4, 1) \\ g_2 &= \frac{1}{2}\mathcal{N}(-12, 1) + \frac{1}{2}\mathcal{N}(12, 1). \end{aligned}$$

More specifically, the three scenarios are:

1. For $s = 1$, we set, $q^{(1)} = 0$. This is the case where the three populations are coming from the same 4–mixture f. We depict the density estimations under the first scenario in Figures 7(a)–(c). This is the case where the small data set, benefits the most in terms of borrowing of strength.

2. For $s = 2$, we set, $q^{(2)} = 1/2$. The 2-mixtures g_1 and g_2 are the the idiosyncratic parts of the 6-mixtures $f_1^{(2)}$ and $f_3^{(2)}$, respectively. The density estimations under the second scenario are given in Figures 7(d)–(f). In this case, the strength of borrowing between the small data set and the two large data sets weakens.

3. For $s = 3$ we set $q^{(3)} = 1$. In this case the three populations have no common parts. The density estimations are given in Figure 4.7(g)–(i). This is the worst case scenario, where there is no borrowing of strength between the small and the two large data sets.

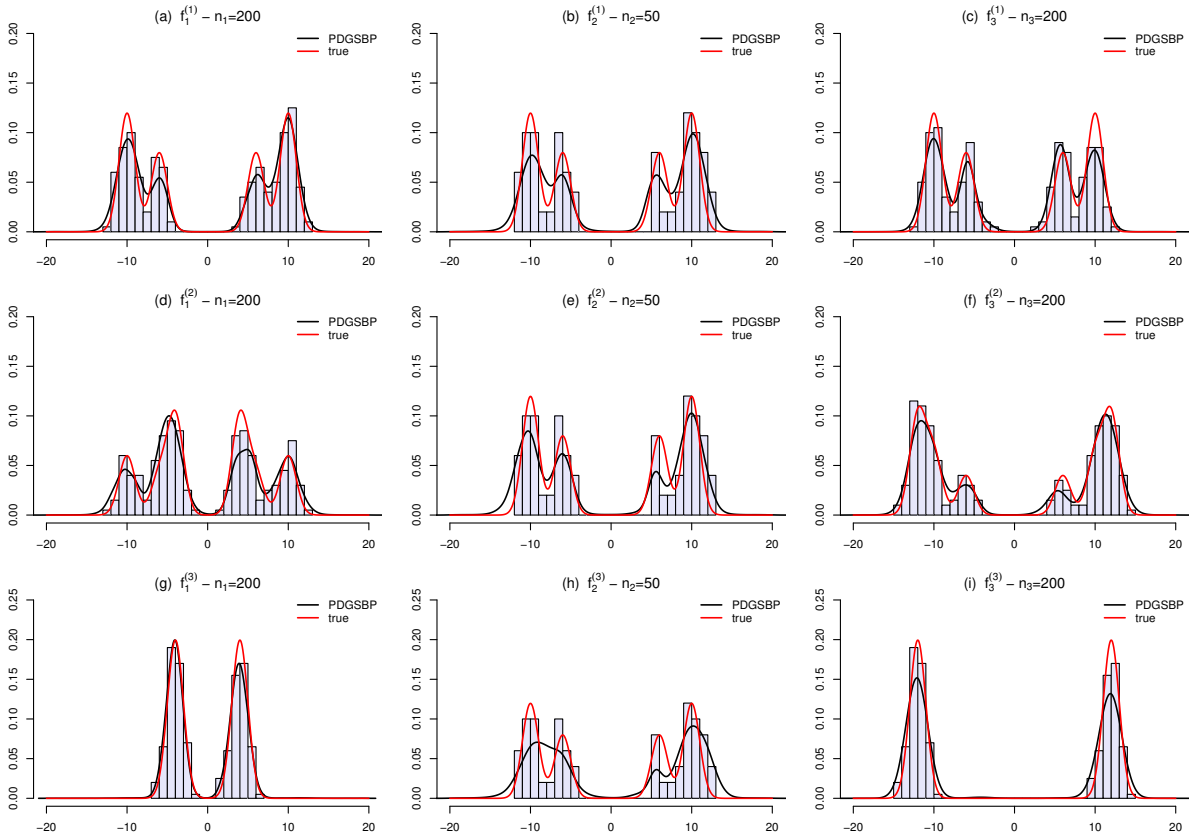

Figure 4.7: Density estimation with the PDGSBP model (curves in black) under the three different scenarios. The true density has been superimposed in red.

The Hellinger distances between the true and the estimated densities, for the three scenarios, are given in Table 4.7. In the second column of Table 4.7, we can see how the Hellinger distance of the estimation $\hat{f}_2^{(s)}$ and the true density $f_2^{(s)}$ increases as the borrowing of strength weakens, it is that $\mathcal{H}_{\mathcal{G}}(f_2^{(1)}, \hat{f}_2^{(1)}) < \mathcal{H}_{\mathcal{G}}(f_2^{(2)}, \hat{f}_2^{(2)}) < \mathcal{H}_{\mathcal{G}}(f_2^{(3)}, \hat{f}_2^{(3)})$.

Table 4.7: Hellinger distances between the true and the estimated densities for the three scenario example.

s	$\mathcal{H}_{\mathcal{G}}(f_1^{(s)}, \hat{f}_1^{(s)})$	$\mathcal{H}_{\mathcal{G}}(f_2^{(s)}, \hat{f}_2^{(s)})$	$\mathcal{H}_{\mathcal{G}}(f_3^{(s)}, \hat{f}_3^{(s)})$
1	0.14	0.19	0.13
2	0.15	0.22	0.15
3	0.12	0.26	0.12

4.4.4 Real data example

The data set is to be found at `http://lib.stat.cmu.edu/datasets/pbcseq` and involves data from 310 individuals. We take the observation as SGOT (serum glutamic-oxaloacetic transaminase) level, just prior to liver transplant or death or the last observation recorded, under three conditions on the individual

1. The individual is dead without transplantation.
2. The individual had a transplant.
3. The individual is alive without transplantation.

We normalize the means of all three data sets to zero. Since it is reasonable to assume the densities for the observations are similar for the three categories (especially for the last two), we adopt the models proposed in this paper with $m = 3$. The number of transplanted individuals is small (sample size of 28) so it is reasonable to borrow strength for this density from the other two. In this example, we set the hyperparameters of the Dirichlet priors for the selection probabilities to

$$\alpha_{jl} = \begin{cases} 10, & \text{if } j = l = 1 \text{ or } j = l = 3 \\ 1, & \text{otherwise.} \end{cases}$$

1. In Figure 4.8(a)–(c), we provide histograms of the real data sets and superimpose the KDE's based on the predictive samples of the PDGSBP and rPDDP samplers. The two models give nearly identical density estimations.
2. The estimated a–posteriori selection probabilities are given below

$$\mathbb{E}_{\mathcal{G}}(p \,|\, (x_{ji})) = \begin{pmatrix} 0.61 & 0.23 & 0.16 \\ 0.34 & 0.10 & 0.56 \\ 0.08 & 0.12 & 0.80 \end{pmatrix}, \quad \mathbb{E}_{\mathcal{D}}(p \,|\, (x_{ji})) = \begin{pmatrix} 0.67 & 0.16 & 0.17 \\ 0.29 & 0.15 & 0.56 \\ 0.10 & 0.12 & 0.78 \end{pmatrix}.$$

By comparing the second rows of the selection matrices, we conclude that the the strength of borrowing is slightly larger in the case of PDGSBP model .

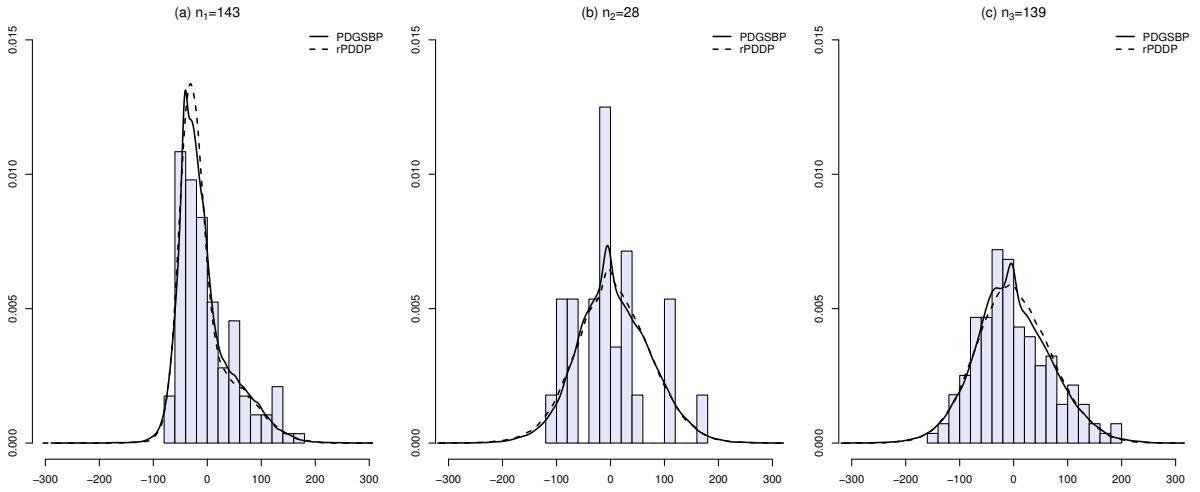

Figure 4.8: Histograms of the real data sets with superimposed KDE curves based on the predictive samples of the PDGSBP and rPDDP models.

4.5 Time–efficiency of the PDGSBP model

In the previous section, we have seen that, there are great differences in the mean execution time of the two methods with the PDGSBP model being faster than the rPDDP. This is due to the ordered nature of the geometric stick breaking weights. Having the weights ordered will lead (in most cases) in faster sampling times from the joint distribution of the clustering variables d_{ji}. In the following sections, we analyze the computational complexity of the two models when it comes to the sampling of d_{ji}.

4.5.1 Sampling d_{ji} in the rPDDP model

The state space of the variable (d_{ji}, δ_{ji}) conditionally on the slice variable u_{ji} is $(d_{ji}, \delta_{ji})(\Omega) = \cup_{l=1}^{m} \left(A_{w_{jl}}(u_{ji}) \times \{\mathbf{e}_l\}\right)$, where $A_{w_{jl}}(u_{ji}) = \{r \in \mathbb{N} : u_{ji} < w_{jlr}\}$ is the a.s. finite slice set corresponding to the observation x_{ji} (Walker, 2007).

At each iteration of the Gibbs sampler, we have $m(m+1)/2$ vectors of stick-breaking weights $\mathbf{w}_{jl}$, each of length N^*_{jl}; where $N^*_{jl} \sim 1 + \text{Poisson}(-c_{jl} \log u^*_{jl})$ with c_{jl} being the concentration parameter of the Dirichlet process $\mathbb{P}_{jl}$ and u^*_{jl} being the minimum of the slice variables in densities f_j and f_l.

In Algorithm 2, the procedure for the blocked sampling of the clustering and mixture indicator variables is presented. To give an intuition about how the slice sets are created we provide an illustration of the effect of the slice variable u_{ji} in Figure 4.9.

Algorithm 2 : rPDDP

1: **procedure** SAMPLE (d_{ji}, δ_{ji})
2: **for** random densities f_j, $j = 1$ to m **do**
3: **for** each data point $x_{ji} \in f_j$ $i = 1$ to n_j **do**
4: **for** each mixture component $K(x_{ji}|\theta_{jl})$, $l = 1$ to m **do**
5: Construct slice sets $A_{w_{jl}}(u_{ji})$
6: **end for**
7: Sample $(d_{ji} = k, \delta_{ji} = r|\cdots) \propto K(x_{ji}|\theta_{jrk})\mathcal{I}\left((k,r) \in \cup_{l=1}^{m}\left(A_{w_{jl}}(u_{ji}) \times \{\mathbf{e}_l\}\right)\right)$
8: **end for**
9: **end for**
10: **end procedure**

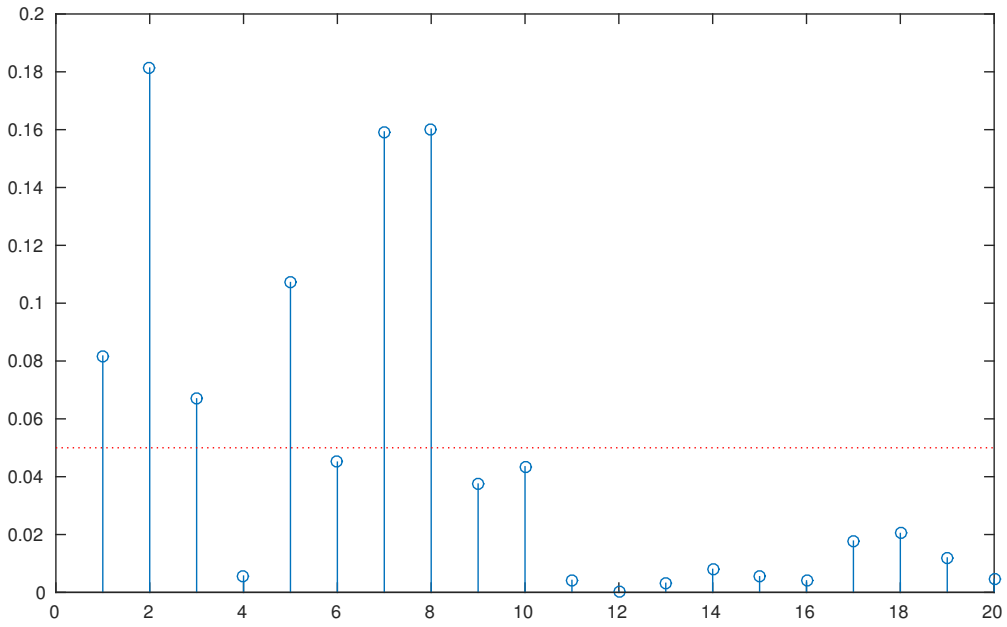

Figure 4.9: Stick-breaking weights for some $N_{jl}^* = 20$. The red dashed line represents the slice variable $u_{ji} = 0.05$. The algorithm must check all the N_{jl}^* values to accept those that they satisfy $u_{ji} < w_{jlk}$. After a complete search, the slice set is $A_{w_{jl}}(u_{ji}) = \{1, 2, 3, 5, 7, 8\}$.

Since the weights forming the stick-breaking representation are not in an ordered form, the construction of the slice sets in step 5 of Algorithm 2 requires a complete search in the array where the weights are stored. This operation is done in $\mathcal{O}(N_{jl}^*)$ time. For the sampling of the d_{ji} and δ_{ji} variables in step 6, the choice of their value is an element from the union $\cup_{l=1}^{m}\left(A_{w_{jl}}(u_{ji}) \times \{\mathbf{e}_l\}\right)$. This means, that the rPDDP algorithm, must create for each j, m slice sets which require N_{jl}^* comparisons each. The worst case scenario is that the sampled (d_{ji}, δ_{ji}) will be the last element of $\cup_{l=1}^{m}\left(A_{w_{jl}}(u_{ji}) \times \{\mathbf{e}_l\}\right)$. Thus, the DP based procedure of sampling

(d_{ji}, δ_{ji}) is of order

$$\mathcal{O}\left(m^2 n_j N_{jl}^* \sum_{l=1}^{m} |A_{w_{jl}}(u_{ji})|\right) = \mathcal{O}\left(N_{jl}^* \sum_{l=1}^{m} |A_{w_{jl}}(u_{ji})|\right).$$

4.5.2 Sampling d_{ji} in the PDGSBP model

The state space of the variable (d_{ji}, δ_{ji}) conditionally on the slice variable N_{ji} is $(d_{ji}, \delta_{ji})(\Omega) = \cup_{l=1}^{m} (\mathcal{S}_{ji} \times \{\mathbf{e}_l\})$. In the GSB case, the slice variable has a different rôle. It indicates at which random point the search for the appropriate d_{ji} will stop. In Figure 4.10, we illustrate this argument. In Algorithm 3, the worst case scenario is that the sampled (d_{ji}, δ_{ji}) will be the last element of $\cup_{l=1}^{m} (\mathcal{S}_{ji} \times \{\mathbf{e}_l\})$. Thus, the GSB based procedure of sampling (d_{ji}, δ_{ji}) is of order $\mathcal{O}\left(m^2 n_j N_{jl}\right) = \mathcal{O}\left(N_{jl}\right)$.

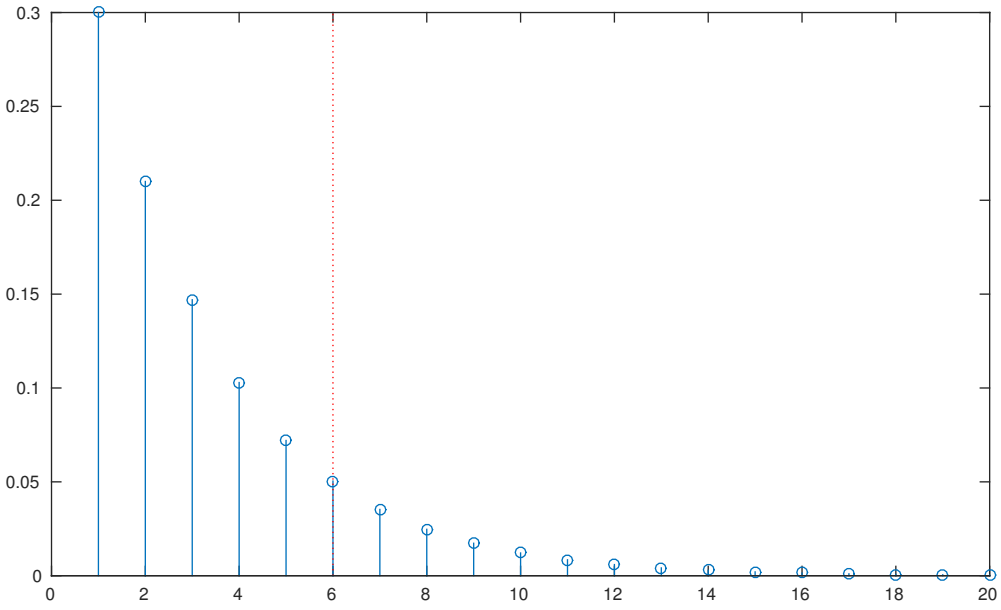

Figure 4.10: Geometric stick-breaking weights for $N_{jl}^* = 20$. The red dashed line represents the slice variable $N_{ji} = 6$. The slice set is simply $\mathcal{S}_{ji} = \{1, 2, 3, 4, 5, 6\}$.

Algorithm 3 : PDGSBP

procedure SAMPLE (d_{ji}, δ_{ji})
  for random densities $f_j,\ \ j = 1$ to m **do**
    for each data point $x_{ji} \in f_j\ \ i = 1$ to n_j **do**
      for each mixture component $K(x_{ji}|\theta_{jl}),\ \ l = 1$ to m **do**
        Sample $(d_{ji} = k, \delta_{ji} = r|\cdots) \propto K(x_{ji}|\theta_{jrk})\,\mathcal{I}(k \leq N_{ji})\,\mathcal{I}(r \leq m)$
      end for
    end for
  end for
end procedure

4.6 Conclusions

We have generalized the GSB process to a multidimensional dependent stochastic process which can be used as a Bayesian nonparametric prior for density estimation in the case of partially exchangeable data sets. The resulting Gibbs sampler is as accurate as its DP based counterpart, yet faster and far less complicated in terms of computational complexity and ease of implentation. The main reason for this is that the GSB sampled value of the allocation variable d_{ji} will be a choice from the sequential slice set $\mathcal{S}_{ji} = \{1, \ldots, N_{ji}\}$. Thus, there is no need to search the arrays of the weights (see Section 4.5).

Also, for an objective comparison of the execution times of the two models, we have run the two samples in an a–priori synchronized mode. This, involves the placing of $\mathcal{G}(a_{jl}, b_{jl})$ priors over the DP c_{jl} concentration masses, leading to a more efficient version of the PDDP model introduced in Hatjispyros et al. (2011, 2016).

Finally, we have shown that when the PDGSBP and rPDDP models are synchronized, i.e. their parameters satisfy $\lambda_{ji} = (1 + c_{ji})^{-1}$, the correlation between the models can be controlled by imposing further restrictions among the λ_{ji} parameters.

Chapter 5

Joint reconstruction of RDS with pairwise dependent GSBR priors

5.1 Introduction

A number of approaches for modeling time series in a Bayesian nonparametric context have been proposed in the literature. For example, an infinite mixture of time series models has been proposed in Rodriguez & Ter Horst (2008). A Markov–switching finite mixture of independent DPM's has been proposed by Taddy & Kottas (2009). Recenlty, Jensen & Maheu (2010) and Griffin (2010) considered DPM for stochastic volatility models in discrete and continuous time respectively. An approach for continuous time series modeling based on time dependent GSB process mixtures can be found in Mena et al. (2011).

Recently there has been a growing research interest for Bayesian nonparametric modeling in the context of multiple time series. In a recent work of Fox et al. (2009) a Bayesian nonparametric model based on the Beta process was introduced In order to model dynamical behavior shared among a number of time series. They represented the behavioral set with an *attribute list* encoded by an $n \times k$ binary matrix, with n the number of time series and k the number of features. Their approach allowed for potentially an infinite number of behaviors k. This was an improvement of a similar approach of a previous work of Fox et al. (2008) where the time series shared exactly the same set of behaviors.

In Nieto-Barajas & Quintana (2016) a Bayesian nonparametric dynamic autoregressive model for the analysis of multiple time series was introduced. They considered an autoregressive model of order p for each of the time series in the collection, and a Bayesian nonparametric prior based on dependent Polya trees. The dependent prior, with its median fixed at zero, was used for the modeling of the errors. Such models rely on the concept of partial exchangeability meaning that the order that the samples have been collected, over groups, does not affect their distribution.

In Chapter 3 we have dealt with the problem of reconstruction of the dynamical equation consisting the deterministic part of a stochastic dynamical system, and modeling the density of the noise process with a GSB mixture process. In this chapter we wish to generalize the so called

GSBR model in a multivariate setting in order to reconstruct jointly a finite collection of dynamical equations. We propose a Bayesian nonparametric mixture model for the joint reconstruction of m dynamical equations, given m observed dynamically–noise–corrupted chaotic time series. The method of reconstruction is based on the Pairwise Dependent Geometric Stick Breaking Processes (PDGSBP) mixture priors (Hatjispyros et al., 2017a) already described in Chapter 4. We assume that the dynamical equations have deterministic parts g_j belong to known families of functions; for example they can be polynomial or (and) rational functions. A–priori we assume that we have the knowledge that the noise processes have common characteristics, for example they could reveal a similar tail behavior or (and) have common variances, or simply they come from the same noise process which is (perhaps) non Gaussian. Our contention is that whenever there is at least one sufficiently large data set, using informative borrowing–of–strength prior specifications we will be able to reconstruct the dynamical processes for which we have insufficient information, namely, their sample sizes are inadequate for an independent GSBR reconstruction and prediction.

This Chapter is organized as follows. In section 5.2, we derive the *Pairwise Dependent Geometric Stick Breaking Reconstruction* (PD–GSBR) model, a Bayesian nonparametric mixture model for the reconstruction and prediction of multiple dynamical equations from observed time series data, by applying a PDGSBP mixture prior. In section 5.3, the associated MCMC algorithm is presented. In section 5.4, we resort to simulation. We apply the PD–GSBR model on the reconstruction and prediction of random polynomial maps of arbitrary degree that are dynamically perturbed by noise processes which are (perhaps) non Gaussian. Finally, conclusions and some directions for future research are discussed.

5.2 The Pairwise Depedent GSBR model

We will consider initially, the general case of a finite collection of m dynamic nonlinear models. Letting $x_{j,i:l_j} := (x_{j,i-1}, \cdots, x_{j,i-l_j})$ we have

$$x_{ji} = g_j(\vartheta_j, x_{j,i:l_j}) + z_{ji}, \;\; 1 \leq j \leq m, \; 1 \leq i \leq n_j, \tag{5.1}$$

where $g_j : \Theta_j \times \mathbb{X}_j^{l_j} \to \mathbb{X}_j$ for some compact subsets $\mathbb{X}_j$ of $\mathbb{R}$, and $\Theta_j \subseteq \mathbb{R}^{q_j}$. l_j is the lag of the j-th dynamical model and g_j is a nonlinear map continuous in the variable $x_{j,i:l_j}$, with additive errors $z_{ji} \overset{\text{iid}}{\sim} f_j$ for $1 \leq i \leq n_j$, for all $1 \leq j \leq m$ with f_j some unknown symmetric zero mean density with support over $\mathbb{R}$. Additionally, we assume that there is no observational noise so that we have at our disposal m groups of observations $x_j^{(n_j)} := (x_{j1}, \ldots, x_{j,n_j}), \;\; 1 \leq j \leq m$, associated with the unknown initial conditions $\{x_{j,1:l_j} : 1 \leq j \leq m\}$.

We wish to estimate the control parameters ϑ_j the intial condition $x_{j,1:l_j}$ and the distribution of the dynamical error processes f_j, for all $j = 1, ..., m$.

The z_{ji} are independent and identically distributed random variables with density function f_j for which we do not assume that they belong to a particular parametric family of densities. Instead we take f_j to be nonparametric densities based on the PDGSBP mixture model (Hatjispyros

et al., 2017a). The PDGSBP mixture implies the following hierarchical model for the errors. For $1 \leq j \leq m$ and $1 \leq i \leq n_j$

$$\begin{aligned}
&z_{ji} \,|\, \tau_{ji} \overset{\text{ind}}{\sim} \mathcal{N}(0, \tau_{ji}^{-1}) \\
&\tau_{ji} \,|\, \mathbb{Q}_j \overset{\text{iid}}{\sim} \mathbb{Q}_j \\
&\mathbb{Q}_j = \sum_{l=1}^{m} p_{jl} \mathbb{G}_{jl}, \; \sum_{l=1}^{m} p_{jl} = 1, \; \mathbb{G}_{jl} = \mathbb{G}_{lj} \\
&\mathbb{G}_{jl} \overset{\text{ind}}{\sim} \text{GSB}(\lambda_{jl}, G_0), \; 1 \leq j \leq l \leq m \\
&\lambda_{jl} \overset{\text{ind}}{\sim} \mathcal{B}e(\alpha_{jl}, \beta_{jl}),
\end{aligned}$$

and G_0 is the parametric base measure of the GSB process with $\mathbb{E}[\mathbb{G}_{jl}(A)] = G_0(A)$. Here the measure G_0 is assumed to have density g_0 which we will take to be $\mathcal{G}(a, b)$, a gamma density with mean a/b.

While our method can be used to reconstruct dynamical systems where each state x_{ji} depends on the previous l_j states $x_{j,i-1}, \ldots, x_{j,i-l_j}$, for simplicity and ease of illustration we focus on the case $l_j = 1$. In this case, the dynamical system has a Markovian dependence structure and can be written as

$$x_{ji} = g_j(\vartheta_j, x_{j,i-1}) + z_{ji}, \quad j = 1, \ldots, m, \quad i = 1, \ldots, n_j. \tag{5.2}$$

The hierarchical model for the observations $x_{j1}, \ldots, x_{jn_j}$ conditional on the unknown initial condition x_{j0} for $1 \leq j \leq m$ becomes

$$\begin{aligned}
&x_{ji} \,|\, x_{j,i-1}, \vartheta_j, \tau_{ji} \overset{\text{ind}}{\sim} \mathcal{N}(g_j(\vartheta, x_{j,i-1}) \,|\, 0, \tau_{ji}^{-1}) \\
&\tau_{ji} \,|\, \mathbb{Q}_j \overset{\text{iid}}{\sim} \mathbb{Q}_j \\
&\mathbb{Q}_j = \sum_{l=1}^{m} p_{jl} \mathbb{G}_{jl}, \; \sum_{l=1}^{m} p_{jl} = 1, \; \mathbb{G}_{jl} = \mathbb{G}_{lj} \\
&\mathbb{G}_{jl} \overset{\text{ind}}{\sim} \text{GSB}(\lambda_{jl}, G_0), \; 1 \leq j \leq l \leq m \\
&\lambda_{jl} \overset{\text{ind}}{\sim} \mathcal{B}e(\alpha_{jl}, \beta_{jl}).
\end{aligned}$$

Using the fact that

$$\mathbb{G}_{jl} = \sum_{k=1}^{\infty} w_{jlk} \delta_{\tau_{jlk}} \text{ with } \; w_{jlk} = \lambda_{jl}(1 - \lambda_{jl})^{k-1}, \; \lambda_{jl} \sim \mathcal{B}e(\alpha_{jl}, \beta_{jl}), \; \tau_{jlk} \overset{\text{iid}}{\sim} \mathcal{G}(a, b), \tag{5.3}$$

the transition density of the i-th observation in the j-th group can be written as a random finite mixture of random infinite mixtures of normal kernels, that is

$$f_j(x_{ji} \,|\, x_{j,i-1}, \vartheta_j, (\tau_{jl})_{l \geq 1}) = \sum_{l=1}^{m} p_{jl} \left\{ \sum_{k=1}^{\infty} w_{jlk} \mathcal{N}(x_{ji} \,|\, g_j(\vartheta_j, x_{j,i-1}), \tau_{jlk}^{-1}) \right\}, \tag{5.4}$$

The density shown in eq. (5.4) is of a nonstandard form. We will deal with the nonlinear maps

g_j, appearing in the means of the normal kernels, with the introduction of auxiliary variables in a similar fashion as in Chapter 3. We will apply the same density augmentations as in Chapter 4. That is, we define the stochastic variables $\mathbf{N} = (N_{ji})$ for $1 \leq i \leq n_j$ and $1 \leq j \leq m$, where N_{ji} is an almost surely finite random variable following the specific negative binomial distribution $\mathcal{N}b(N_{ji} \,|\, 2, \lambda_{jl}) = N_{ji}\lambda_{jl}^2(1-\lambda_{jl})^{N_{ji}-1}$. Consequently we introduce

1. The GSB mixture selection variables $\boldsymbol{\delta} = (\delta_{ji})$; for an observation x_{ji} that comes from f_j, δ_{ji} selects the GSB mixture $g_{ji}(x)$ that the observation came from. It is that $\Pr\{\delta_{ji} = l\} = p_{jl}$.
2. The clustering variables $\boldsymbol{d} = (d_{ji})$; for an observation x_{ji} that comes from f_j, given δ_{ji}, d_{ji} allocates the component of the GSB mixture $g_{j\delta_{ji}}(x)$ that x_{ji} came from. Given N_{ji} the variables d_{ji} have a discrete uniform distribution over the integers $\{1, \ldots, N_{ji}\}$.

Augmenting the random densities given in (5.4) with N_{ji} we have that

$$\begin{aligned} f_j(x_{ji}, N_{ji} = r) &= \sum_{l=1}^{m} f_j(x_{ji}, N_{ji} = r, \delta_{ji} = l) \\ &= \sum_{l=1}^{m} p_{jl} \sum_{k=1}^{\infty} f_j(x_{ji}, N_{ji} = r, d_{ji} = k \,|\, \delta_{ji} = l) \\ &= \sum_{l=1}^{m} p_{jl} \sum_{k=1}^{\infty} \mathcal{N}b(N_{ji} = r \,|\, 2, \lambda_{jl}) \mathcal{DU}(k \,|\, 1, \ldots, r) \mathcal{N}(x_{ji} \,|\, g_j(\vartheta_j, x_{j,i-1}), \tau_k^{-1}) \\ &= \frac{1}{r} \sum_{l=1}^{m} p_{jl} \sum_{k=1}^{r} \mathcal{N}b(N_{ji} = r \,|\, 2, \lambda_{jl}) \mathcal{N}(x_{ji} \,|\, g_j(\vartheta_j, x_{j,i-1}), \tau_k^{-1}), \end{aligned}$$

leading to the d_{ji}, N_{ji} augmented density:

$$f_j(x_{ji}, N_{ji} = r, d_{ji} = k \,|\, \delta_{ji} = l) = \lambda_{jl}^2 (1-\lambda_{jl})^{k-1} \mathcal{N}(x_{ji} \,|\, g_j(\vartheta_j, x_{j,i-1}), \tau_k^{-1}). \tag{5.5}$$

We denote the set of observations along the m groups as $\boldsymbol{x} = \{x_{ji} : 1 \leq j \leq m, 1 \leq i \leq n_j\}$ and with $\boldsymbol{x}_j = \{x_{ji} : 1 \leq i \leq n_j\}$ the set of observations in the j–th group. The three sets of latent variables in the jth group will be denoted as $\boldsymbol{N}_j = \{N_{ji} : 1 \leq i \leq n_j\}$ for the slice variables, $\boldsymbol{d}_j = \{d_{ji} : 1 \leq i \leq n_j\}$ for the clustering variables, and finally $\boldsymbol{\delta}_j = \{\delta_{ji} : 1 \leq i \leq n_j\}$ for the set of GSB mixture allocation variables. From now on, we are going to leave the auxiliary variables unspecified; especially for δ_{ji} we use the notation

$$\delta_{ji} = (\delta_{ji}^1, \ldots, \delta_{ji}^m) \in \{\mathbf{e}_1, \ldots, \mathbf{e}_m\} \;\text{ with }\; \Pr(\delta_{ji} = \mathbf{e}_l) = p_{jl},$$

where $\mathbf{e}_l$ denotes the usual basis vector having its only nonzero component equal to 1 at position l. Hence, for a sample of size n_1 from f_1, a sample of size n_2 from f_2, etc., a sample of size n_m

from f_m we can write the full likelihood as a multiple product:

$$\begin{aligned} f(\boldsymbol{x}, \boldsymbol{N}, \boldsymbol{d} \,|\, \boldsymbol{\delta}) &= \prod_{j=1}^{m} f(\boldsymbol{x}_j, \boldsymbol{N}_j, \boldsymbol{d}_j \,|\, \boldsymbol{\delta}_j) \\ &= \prod_{j=1}^{m} \prod_{i=1}^{n_j} \mathcal{I}(d_{ji} \leq N_{ji}) \prod_{l=1}^{m} \left\{ \lambda_{jl}^2 (1-\lambda_{jl})^{N_{ji}-1} \mathcal{N}(x_{ji} \,|\, g_j(\vartheta_j,\, x_{j,i-1}),\, \tau_{jld_{ji}}) \right\}^{\delta_{ji}^l}. \end{aligned}$$

In a hierarchical fashion, using the auxiliary variables, we have for $j = 1, \ldots, m$ and $i = 1, \ldots, n_j$,

$$\begin{aligned} x_{ji}, N_{ji} \,|\, d_{ji}, \delta_{ji}, (\tau_{jr\delta_{ji}})_{1\leq r\leq m}, \lambda_{j\delta_{ji}} &\overset{\text{ind}}{\sim} \prod_{r=1}^{m} \left\{ \lambda_{jr}^2 (1-\lambda_{jr})^{N_{ji}-1} \mathcal{N}(x_{ji} \,|\, g_j(\vartheta_j,\, x_{j,i-1}),\, \tau_{jrd_{ji}}) \right\}^{\delta_{ji}^r} \\ &\quad \times \mathcal{I}(N_{ji} \geq d_{ji}) \\ d_{ji} \,|\, N_{ji} &\overset{\text{ind}}{\sim} \mathcal{DU}(\{1, \ldots, N_{ji}\}),\ \Pr(\delta_{ji} = \mathbf{e}_l) = p_{jl} \\ w_{jik} = \lambda_{ji}(1-\lambda_{ji})^{k-1},\ \tau_{jik} &\overset{\text{iid}}{\sim} P_0,\ \ k \in \mathbb{N}, \end{aligned}$$

5.3 The PD-GSBR Gibbs sampler

In this section, we are going to describe the PD-GSBR Gibbs sampler. At each iteration we will sample the variables,

$$\begin{aligned} &\tau_{jlk}, 1 \leq j \leq l \leq m,\, 1 \leq k \leq N^*, \\ &d_{ji}, N_{ji}, \delta_{ji}, 1 \leq j \leq m,\, 1 \leq i \leq n_j, \\ &p_{jl}, 1 \leq j \leq m, 1 \leq l \leq m, \\ &\vartheta_j, x_{j0}, 1 \leq j \leq m \end{aligned}$$

with $N^* = \max_{j,i} N_{ji}$ almost surely finite.

1. For the precisions of the random measures for $k = 1, \ldots, N^*$ where $N^* = \max_{j,i} N_{ji}$, it is that for $l > j$

$$\begin{aligned} f(\tau_{jlk} \,|\, \cdots) \propto g_0(\tau_{jlk}) &\prod_{i=1}^{n_j} \mathcal{N}(x_{ji} \,|\, g_j(\vartheta_j, x_{j,i-1}), \tau_{jlk}^{-1})^{\mathcal{I}(\delta_{ji}=\mathbf{e}_l,\, d_{ji}=k)} \\ &\times \prod_{i=1}^{n_l} \mathcal{N}(x_{ji} \,|\, g_j(\vartheta_j, x_{j,i-1}), \tau_{jlk}^{-1})^{\mathcal{I}(\delta_{li}=\mathbf{e}_j,\, d_{li}=k)}. \end{aligned}$$

For $l = j$, it is that

$$f(\tau_{jjk} \,|\, \cdots) \propto g_0(\tau_{jjk}) \prod_{i=1}^{n_j} \mathcal{N}(x_{ji} \,|\, g_j(\vartheta_j, x_{j,i-1}), \tau_{jjk}^{-1})^{\mathcal{I}(\delta_{ji}=\mathbf{e}_j,\, d_{ji}=k)}.$$

2. Here, we sample the allocation variables d_{ji} and the mixture component indicator variables δ_{ji} as a block. For $j = 1, \ldots, m$ and $i = 1, \ldots, n_j$, we have

$$\Pr(d_{ji} = k, \delta_{ji} = \mathbf{e}_l \,|\, N_{ji} = r, \cdots) \;\propto\; p_{jl}\, \mathcal{N}(x_{ji} \,|\, g_j(\vartheta_{jlk}, x_{j,i-1}), \tau_{jlk}^{-1})\, \mathcal{I}(l \leq m)\, \mathcal{I}(k \leq r).$$

3. The geometric slice variables N_{ji} have full conditional distributions given by

$$\Pr(N_{ji} = r \,|\, \delta_{ji} = \mathbf{e}_l, d_{ji} = l, \cdots) \propto (1 - \lambda_{jl})^r\, \mathcal{I}(l \leq r),$$

which are truncated geometric distributions over the set $\{l, l+1, \ldots\}$.

4. The full conditional, for $j = 1, \ldots, m$, for the selection probabilities $\boldsymbol{p}_j = (p_{j1}, \ldots, p_{jm})$, under a Dirichlet prior $f(\boldsymbol{p}_j \,|\, \boldsymbol{a}_j) \propto \prod_{l=1}^m p_{jl}^{a_{jl}-1}$, with hyperparameter $\boldsymbol{a}_j = (a_{j1}, \ldots, a_{jm})$, is a Dirichlet distribution. Namely, for $j = 1, \ldots, m$ we have

$$f(\boldsymbol{p}_j \,|\, \cdots) \;\propto\; \prod_{l=1}^{m} p_{jl}^{a_{jl} + \sum_{i=1}^{n_l} \mathcal{I}(\delta_{ji} = \mathbf{e}_l) - 1}.$$

5. The full conditionals for the geometric probabilities λ_{jl} under beta conjugate prior $\mathcal{B}e(a_{jl}, b_{jl})$ are Beta distributions. Letting

$$S_{jl} = \sum_{i=1}^{n_j} \mathcal{I}(\delta_{ji} = \mathbf{e}_l) \quad \text{and} \quad S'_{jl} = \sum_{i=1}^{n_j} \mathcal{I}(\delta_{ji} = \mathbf{e}_l)(N_{ji} - 1),$$

for $l = j$ it is that

$$f(\lambda_{jj} \,|\, \cdots) = \mathcal{B}e(\lambda_{jl} \,|\, a_{jj} + 2S_{jj}, b_{jj} + S'_{jj}),$$

also, for $l \neq j$ we have

$$f(\lambda_{jl} \,|\, \cdots) = \mathcal{B}e(\lambda_{jl} \,|\, a_{jl} + 2(S_{jl} + S_{lj}), b_{jl} + S'_{jl} + S'_{lj}).$$

6. For the vectors of parameters ϑ_j, $1 \leq j \leq m$, and assuming a uniform prior over the subset $\tilde{\Theta}_j$ of the parameter space $\mathbb{R}^k$, the full conditional becomes

$$f(\vartheta_j \,|\, \cdots) \;\propto\; \mathcal{I}(\vartheta_j \in \tilde{\Theta}_j) \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n_j} \tau_{jld_{ji}} h_{\vartheta_j}(x_{ji}, x_{j,i-1}) \right\}, \tag{5.6}$$

where $h_{\vartheta_j}(x_{ji}, x_{j,i-1}) := (x_{ji} - g_j(\vartheta_j, x_{j,i-1}))^2$.

7. The full conditional for x_{j0}, with a uniform prior over the sets $\tilde{\mathbb{X}}_j \subseteq \mathbb{R}$ that represents our prior knowledge for the state space of the dynamical systems in relation (5.1) will be

$$f(x_{j0} \,|\, \cdots) \;\propto\; \mathcal{I}(x_{j0} \in \tilde{\mathbb{X}}) \exp\left\{ -\frac{\tau_{jld_1}}{2} h_{\vartheta_j}(x_{j1}, x_{j0}) \right\}. \tag{5.7}$$

8. The full conditionals for sampling the next T future unobserved observations for $k = 1, \ldots, T-1$ are given by

$$f(x_{j,n_j+k} \mid \cdots) \propto \exp\left\{-\frac{1}{2}\left[\tau_{d_{j,n_j+k}} h_{\vartheta_j}(x_{n_j+k}, x_{n_j+k-1}) + \tau_{d_{j,n_j+k+1}} h_{\vartheta_j}(x_{n_j+k+1}, x_{n_j+k})\right]\right\} \tag{5.8}$$

For $k = T$, the full conditional is normal with mean $g_j(\vartheta_j, x_{j,n_j+T-1})$ and variance $\tau^{-1}_{j\delta_{j,n_j+T} d_{j,n_j+T}}$, that is

$$f_j(x_{j,n_j+T} \mid \cdots) = \mathcal{N}\left(x_{j,n_j+T} \mid g_j(\vartheta_j, x_{j,n_j+T-1}), \tau^{-1}_{j\delta_{j,n_j+T} d_{j,n_j+T}}\right). \tag{5.9}$$

For simplicity, in our numerical experiments, we will sample only the first future unobserved observation x_{j,n_j+1}.

Conditionally on the mixture allocation variable δ_{ji}, we can sample from eq. (5.6) through eq. (5.9) with the auxuliary variables method as described in Appendix A.1.

5.4 Numerical illustrations

In this section, we will demonstrate the efficiency of the proposed PD–GSBR model for synthetic time series, for the case $m = 2$ and $l_1 = l_2 = 1$. The deterministic parts of the dynamical systems for the data simulation, are given by polynomial autoregressive processes. We will use the following chaotic dynamical systems $x_i = \mathcal{C}_r(x_{i-1})$, and $x_i = \mathcal{Q}_r(x_{i-1})$, for $r = 1, 2$, with

$$\mathcal{C}_r(x) = 0.05 + c_r x_{i-1} - 0.99 x_{i-1}^3 \text{ with } c_1 = 2.55 \text{ and } c_2 = 2.65 \tag{5.10}$$

$$\mathcal{Q}_r(x) = 1 - q_r x_{i-1}^2 \text{ with } q_1 = 1.71 \text{ and } q_2 = 1.85. \tag{5.11}$$

The dynamical systems in (5.10) and (5.11) are chaotic. Moreover, both cubic maps $\mathcal{C}_1$ and $\mathcal{C}_2$, when perturbed by dynamical noise of sufficient intensity, they follow a scenario of noise induced jumps as seen in Chapter 3. The values q_1 and q_2 belong to the Pomeau-Manneville chaotic band (Pomeau & Manneville, 1980).

We will illustrate different scenarios for which the PD–GSBR reconstruction and prediction, is beneficial to one of the time series, for which an independent GSBR reconstruction and prediction, is problematic due to its small sample size. We will specify the sample sizes of the time series in each example separately.

Some non-Gaussian dynamical noise processes: In the sequel, we will illustrate the PD-GSBR sampler, with Gaussian and non-Gaussian noise processes. As non-Gaussian noise processes, we will use the $\mathcal{E}_1$ and $\mathcal{E}_2$ pairs of mixtures of normals densities given by

$$f_1 = \left(f_{11} = \frac{1}{4}\mathcal{N}(0, 10^{-6}) + \frac{3}{4} f_{12},\ f_{12} = \frac{6}{10}\mathcal{N}(0, \sigma_1^2) + \frac{4}{10}\mathcal{N}\left(0, (10\,\sigma_1)^2\right)\right), \tag{5.12}$$

with $\sigma_1^2 = 3 \times 10^{-3}$, and

$$f_2 = \left(f_{21} = \frac{3}{4}\mathcal{N}(0, 10^{-7}) + \frac{1}{4} f_{22},\ f_{22} = \frac{9}{10}\mathcal{N}(0, \sigma_2^2) + \frac{1}{10}\mathcal{N}\left(0, (200\,\sigma_2)^2\right) \right), \tag{5.13}$$

with $\sigma_2^2 = 10^{-6}$. Both pairs of noise processes are exhibiting a heavier tail behavior than the standard normal, approximately we have

$$\mathrm{TF}_{f_1} = (0.505, 0.576),\ \mathrm{TF}_{f_2} = (0.138, 0.264).$$

We remark that the tail fatness of the standard normal is $\sqrt{2/\pi} \approx 0.798$, and that the more the TF index gets closer to zero, the heavier the tails are.

For the reconstruction of the deterministic parts given in (5.10) and (5.11), as model polynomials, in all our illustrations, we have used the quintic polynomials

$$g_j(\vartheta_j, x) = \sum_{r=0}^{5} \vartheta_{jr} x^r,\ j = 1, 2.$$

Prior specifications: We first define the prior distributions for all PD-GSBR Gibbs sampler variables, except the selection probabilities. In all our numerical experiments, we will use the following noninformative prior set up:

$$\begin{aligned} &\lambda_{jl} \sim \mathcal{B}e(1,1) \equiv \mathcal{U}(0,1), \quad \{\tau_{jlk} \sim \mathcal{G}(10^{-3}, 10^{-3}) : k \geq 1\},\ 1 \leq j \leq l \leq 2 \\ &\vartheta_j \sim \mathcal{U}((-10, 10)^6), \quad x_{j0} \sim \mathcal{U}(-10, 10),\ j = 1, 2. \end{aligned}$$

Because the borrowing of information between the two dynamic models, can be quantified by the posterior mean of the selection probabilities

$$\mathbb{E}(p_{j1} | x_1^{(n_1)}, x_2^{(n_2)}),\ j = 1, 2,$$

the prior distribution over the selection probabilities (p_{ji}), plays a decisive role on the strength of borrowing of information between the two dynamic models. In the sequel we will make use of the following borrowing of strength configurations:

1. To force a *weak borrowing* scenario, we impose a-priori

   $$\mathcal{P}_{\mathrm{W}} :\ p_{11} \sim \mathcal{B}e(10, 1), \quad p_{21} \sim \mathcal{B}e(1, 10),$$

   and the prior mean matrix of the selection probabilities becomes

   $$\mathbb{E}\{(p_{ji})\} = \begin{pmatrix} 10/11 & 1/11 \\ 1/11 & 10/11 \end{pmatrix}.$$

2. Suppose now that n_1 is considerably greater than n_2. When a–priori we believe that there is some kind of similarity between the components of the noise process pairs, we increase

the prior probability of selection of the common part of the large data set dynamical system, and we use noninformative a–priori probabilities of selection over the small data set, that is

$$\mathcal{P}_{\rm SN}: \;\; p_{11} \sim \mathcal{B}e(1,10), \quad p_{21} \sim \mathcal{B}e(1,1).$$

We call such an a–priori configuration a *strong noninformative borrowing* scenario. Then, the prior mean matrix of the selection probabilities takes the form

$$\mathbb{E}\{(p_{ji})\} = \begin{pmatrix} 1/11 & 10/11 \\ 1/2 & 1/2 \end{pmatrix},$$

meaning that the common component, associated with the large sample size (first row of the matrix), becomes very influential. On the other hand, the uniform prior, points to the times series with the small sample size.

3. When our prior beliefs advocate that the components of the noise process pairs are about the same, for a *strong informative common noise borrowing of strength* scenario, we set

$$\mathcal{P}_{\rm SI}: \;\; p_{11} \sim \mathcal{B}e(1,10), \quad p_{21} \sim \mathcal{B}e(10,1).$$

Now, it is that

$$\mathbb{E}\{(p_{ji})\} = \begin{pmatrix} 1/11 & 10/11 \\ 10/11 & 1/11 \end{pmatrix},$$

and the main dynamic noise contribution comes from the common noise component between the two time series.

We remark that for $m = 2$ it is that $p_{j2} = 1 - p_{j1}$ for $j = 1, 2$, and, $p \sim \mathcal{B}e(a, b)$ if and only if $1 - p \sim \mathcal{B}e(b, a)$.

In the sequel, with

$$(g_1 + \eta, g_2 + \xi) \to (x_1^{(n_1)}, x_2^{(n_2)}),$$

we denote the fact that the pair of synthetic time series $(x_1^{(n_1)}, x_2^{(n_2)})$ of respective sample sizes n_1 and n_2, have been simulated via a pair of dynamical systems having deterministic parts g_1 and g_2, perturbed dynamically by the noise process pair $f = (\eta, \xi)$.

In all our numerical experiments, as a starting point we have chosen $x_{10} = x_{20} = 1$, and we have ran the PD-GSBR Gibbs sampler for $N = 60,000$ iterations, after a burn-in period of $20,000$ iterations.

A. Borrowing from a cubic to a quadratic map under the f_1 noise pair: In our first numerical example, we make use of the configuration

$$(\mathcal{C}_1 + f_{11}, \mathcal{Q}_1 + f_{12}) \to (x_1^{(200)}, x_2^{(50)}).$$

We will attempt to demonstrate numerically, that it is possible, the borrowing of information from the estimated noise process $\hat{f}_{11}$, based on the time series $x_1^{(200)}$ perturbing the cubic map,

to improve the overall estimation process that is based on the shorter times series $x_2^{(50)}$. This shorter time series, has been produced by the quadratic map, perturbed by the noise process f_{12} which is the non-Gaussian mixture component of the actual process f_{11}.

In Figure 5.1(a) we can see the trace of the stochastic trajectory $x_1^{(200)}$. The time series experiences noise induced jumps from the interval $I_1 = [-1.60, -0.10)$ (containing the chaotic attractor) to the interval $I_2 = [-0.10, 1.67]$ (containing the chaotic repellor). The second dynamical system has the deterministic invariant set $\mathbb{X} = [-1.11, 1.11]$, that is $\mathcal{Q}_1(\mathbb{X}) \subset \mathbb{X}$. Nevertheless, under the dynamical noise perturbation f_{12}, the quadratic trajectory *escapes its invariant set* after the first 46 iterations. In fact, it can be verified that for $x \notin \mathbb{X}$, $\mathcal{Q}_1^n(x) \to -\infty$ as $n \to \infty$. This situation is depicted in Figure 5.1(b).

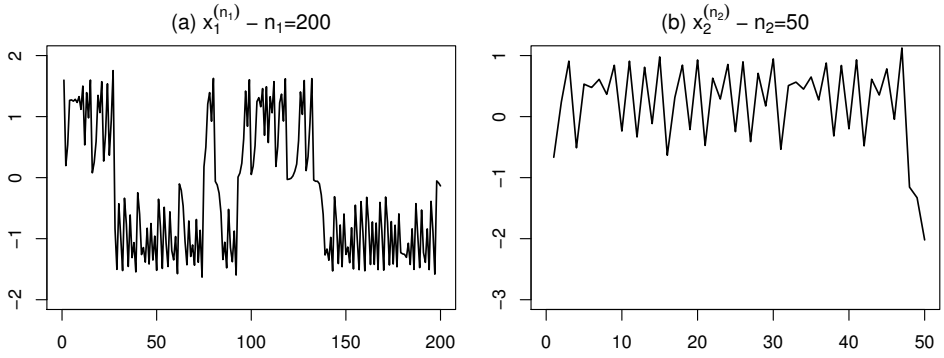

Figure 5.1: The f_1 noise pair perturbed time series corresponding to the cubic map $\mathcal{C}_1$ and the quadratic map $\mathcal{Q}_1$ are given in Figures (a) and (b), respectively.

The ergodic means, coming from the PD-GSBR sampler, for the coefficients of the deterministic parts, based on the $x_1^{(200)}$ and the $x_2^{(50)}$ time series, under the weak borrowing prior specification $\mathcal{P}_{\mathrm{W}}$ (black solid curves) and the strong borrowing noninformative prior specification $\mathcal{P}_{\mathrm{SN}}$ (red solid curves), are given in Figure 5.2(a)-(f), and Figure 5.2(g)-(l), respectively. It can be seen that for the large data set, $x_1^{(200)}$, the running averages based on the predictive samples $\{\hat{\vartheta}_{1i}^k : 1 \leq k \leq N, 0 \leq i \leq 5\}$, after burn-in, and under the $\mathcal{P}_{\mathrm{W}}$ prior, are converging fast to the true values of the ϑ_1-coefficients (represented by the dotted horizontal lines). In parallel, the estimation of the coefficients of the small data set (the ergodic averages in black in Figure 5.2(g)-(l)), is problematic. It leads to a *biased* estimation of the ϑ_2-coefficients.

In Table 5.1, we provide the Percentage Absolute Relative Errors (PAREs), of the the joint $(x_1^{(200)}, x_2^{(50)})$-coefficient estimation, with respect to the true values.

1. In the first two lines of Table 5.1, we can see the effect of the $\mathcal{P}_{\mathrm{W}}$ prior. The estimation associated with $x_1^{(200)}$ is very accurate, and enables the identification of the respective $\mathcal{C}_1$ dynamical system responsible for the observed time series. Nevertheless, the part of the joint estimation based on the time series $x_2^{(50)}$ exhibits large errors, hindering the identification of the second dynamical system.

2. In the last two lines of Table 5.1, we present the effect of borrowing on the estimation of the coefficients via the $\mathcal{P}_{\mathrm{SN}}$ prior. Strong borrowing reduces the average PARE associated

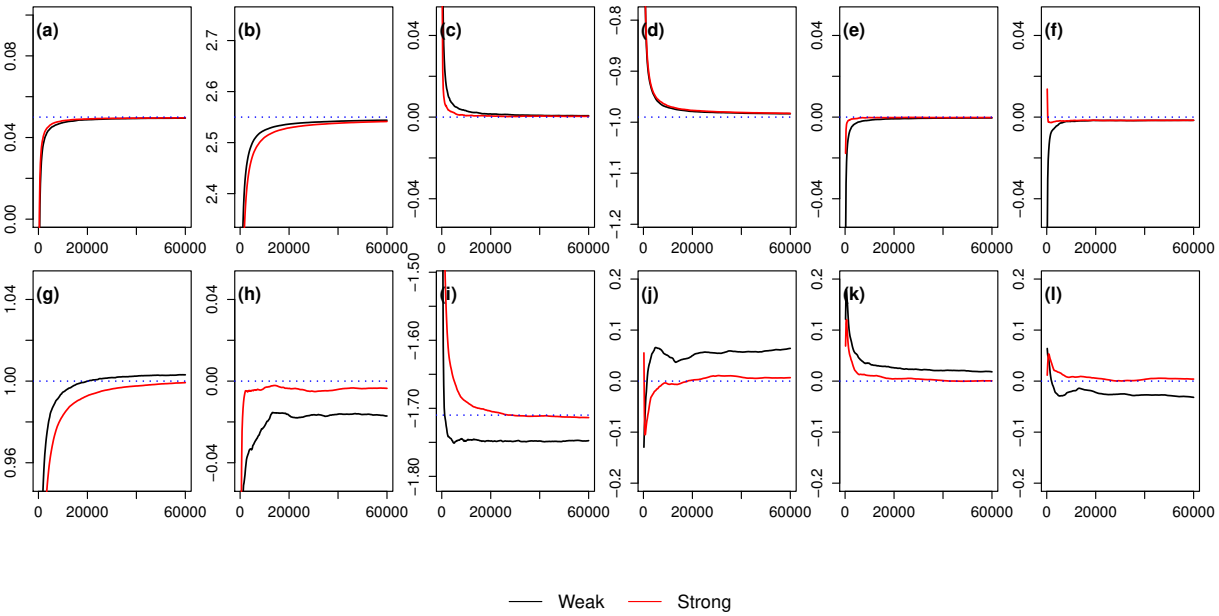

Figure 5.2: Ergodic averages for the $(\vartheta_1, \vartheta_2)$ pair of coefficients of the modeling polynomials under weak (solid curves in black) and strong (solid curves in red) borrowing. The averages associated with the cubic map $\mathcal{C}_1$ appear in Figures (a)-(f), and the averages associated with the quadratic map $\mathcal{Q}_1$ appear in Figures (g)-(l).

with the short time series from 2.67% to a mere 0.37% enabling the identification of the deterministic part $\mathcal{Q}_1$.

Table 5.1: PAREs of the joint GSBR coefficient estimation based on the pair of time series $(x_1^{(200)}, x_2^{(50)})$ under the f_1 noise pair. The estimation is based on a polynomial modeling of fifth degree, assuming the weak borrowing $\mathcal{P}_{\mathrm{W}}$, and the strong borrowing noninformative prior $\mathcal{P}_{\mathrm{SN}}$.

Prior	Time series	ϑ_{j0}	ϑ_{j1}	ϑ_{j2}	ϑ_{j3}	ϑ_{j4}	ϑ_{j5}	$\bar{\vartheta}$
$\mathcal{P}_{\mathrm{W}}$	$x_1^{(200)}$	0.44	0.09	0.04	0.40	0.02	0.14	0.19
	$x_2^{(50)}$	0.55	1.57	2.39	6.44	1.81	3.24	2.67
$\mathcal{P}_{\mathrm{SN}}$	$x_1^{(200)}$	0.50	0.11	0.06	0.50	0.03	0.17	0.23
	$x_2^{(50)}$	0.23	0.25	0.48	0.57	0.01	0.42	0.37

In Figure 5.3(a)-(b) we present the marginal noise densities $(\hat{f}_{21}, \hat{f}_{12})$, of the joint estimation, under the $\mathcal{P}_{\mathrm{W}}$ and $\mathcal{P}_{\mathrm{SN}}$ priors, in black and red, respectively. For the choice of $\mathcal{P}_{\mathrm{W}}$ and $\mathcal{P}_{\mathrm{SN}}$ priors, the posterior mean matrices of the selection probabilities are given by

$$\mathbb{E}_{\mathrm{W}}\{(p_{ji})|\, x_1^{(200)}, x_2^{(50)}\} = \begin{pmatrix} 0.724 & 0.276 \\ 0.142 & 0.858 \end{pmatrix}, \ \mathbb{E}_{\mathrm{SN}}\{(p_{ji})|\, x_1^{(200)}, x_2^{(50)}\} = \begin{pmatrix} 0.230 & 0.770 \\ 0.927 & 0.073 \end{pmatrix},$$

respectively. We can see that in the case of the short time series, under the weak borrowing prior, when sampling from the noise component, the samples come from the common component only 14.2% of the times. Under the strong borrowing prior, sampling from the common component increases to 92.7%. The predictive density of the marginal posterior pair of initial conditions (x_{10}, x_{20}) is depicted in Figure 5.3(c)-(d). We can see that the estimation is nearly

identical under the two prior configurations. In Figure 5.3(e)-(f), we exhibit the predictive density of the marginal posterior pair $(x_{1,201}, x_{2,51})$ for one future observation for each time series. The differences on the estimation of the density of the future observation of the first dynamical system, under the two priors $\mathcal{P}_{\mathrm{W}}$ and $\mathcal{P}_{\mathrm{SN}}$, are nearly indistinguishable. Nevertheless, the joint estimation of the density of the future observation of the second dynamical system under the strong borrowing prior makes a huge difference. The $\mathcal{P}_{\mathrm{SN}}$ prior enables an accurate prediction of the future value $x_{2,51}$ that lies outside the invariant set $\mathbb{X}$. The associated 95% highest posterior density intervals (HPDIs) under the weak and the strong borrowing priors, are given by

$$\mathrm{HPDI}(x_{2,51}; \mathcal{P}_{\mathrm{W}}) = [-6.112, -4.429] \quad \text{and} \quad \mathrm{HPDI}(x_{2,51}; \mathcal{P}_{\mathrm{SN}}) = [-7.067, -5.481],$$

respectively. We remark that the true future value of the quadratic trajectory is at about $x^*_{2,51} = -5.969$ (blue vertical dotted line in Figure 5.3(f)), outside the invariant set $\mathbb{X}$.

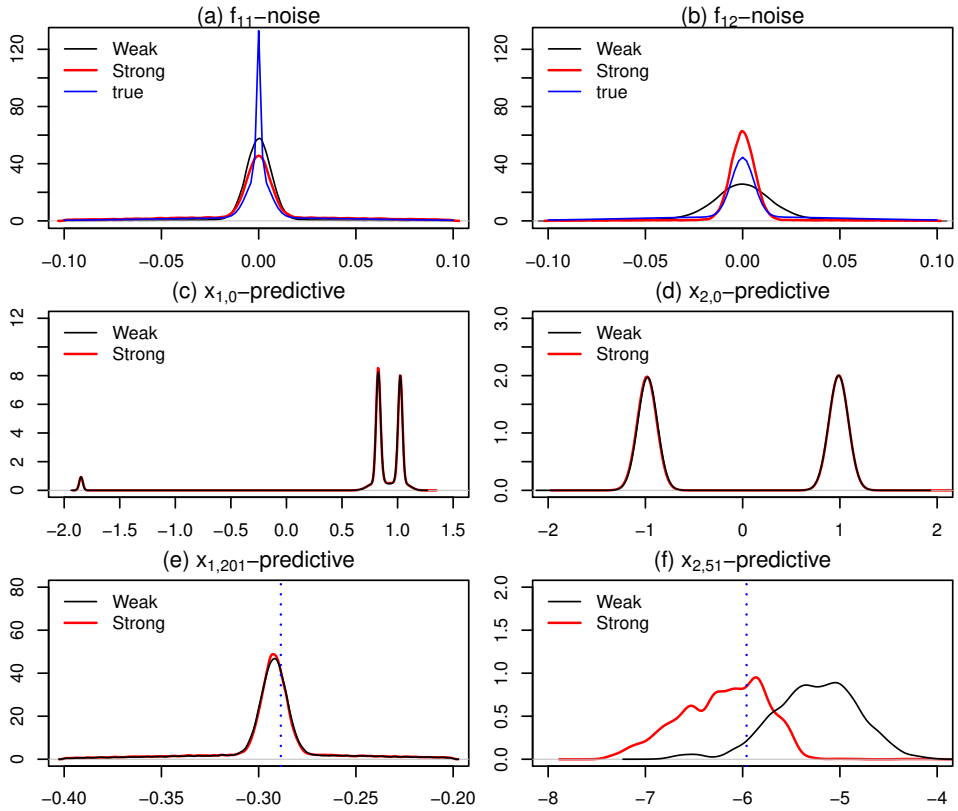

Figure 5.3: Kernel density estimations based on the predictive samples coming from the PD-GSBR Gibbs sampler. Weak borrowing corresponds to the densities in black, and strong borrowing to the densities in red. Figures (a), (c) and (e) correspond to the cubic map $\mathcal{C}_1$, and Figures (b), (d) and (f) correspond to the quadratic map $\mathcal{Q}_1$. The noise predictive densities are given in Figures (a) and (b). The initial conditions predictive densities are given in Figures (c) and (d). In Figures (e) and (f) we give the predictive densities of the first future observation.

B. Borrowing from a cubic to a quadratic map under the f_2 noise pair: In our second numerical example, we use the configuration

$$(\mathcal{C}_1 + f_{21}, \mathcal{Q}_1 + f_{22}) \to (x_1^{(200)}, x_2^{(20)}).$$

In this case the time series generated from the quadratic map is much shorter. At the same time the noise process pair f_2 has heavier tails and larger mixture variances than those of the f_1 pair. In this numerical example, we will show that when *serious mixing issues* occur, the situation can be corrected by applying a joint prior that induces strong borrowing.

In Figure 5.4(a) we give the trace of the stochastic trajectory $x_1^{(200)}$ of length 200, experiencing noise induced jumps due to the dynamic perturbations of the noise process f_{21}. In Figure 5.4(b) we display the quadratic trajectory $x_1^{(20)}$ which is perturbed by the noise process f_{22}.

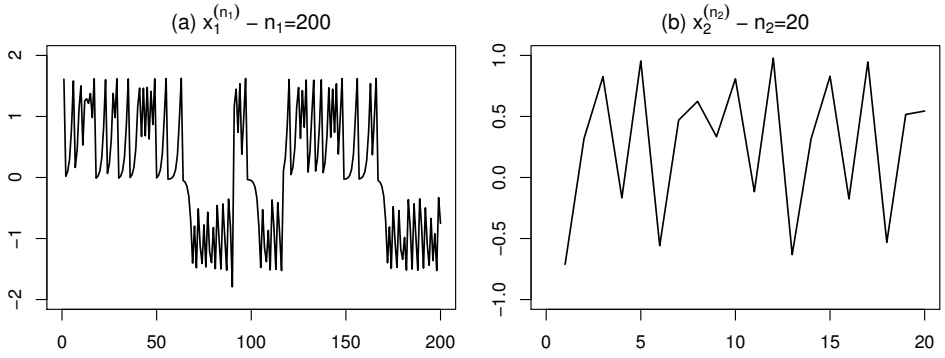

Figure 5.4: The f_2 noise pair perturbed time series corresponding to the cubic map $\mathcal{C}_1$ and the quadratic map $\mathcal{Q}_1$ are given in Figures (a) and (b), respectively.

The ergodic means of the ϑ-coefficients, coming from the PD-GSBR sampler, under the weak borrowing prior specification $\mathcal{P}_{\rm W}$ (black solid curves) and the strong borrowing noninformative prior specification $\mathcal{P}_{\rm SN}$ (red solid curves), are given in Figure 5.5(a)-(f), and Figure 5.5(g)-(l), respectively. For the large data set, the running averages based of the predictive samples under both prior configurations, are converging fast to the true values. On the other hand, the estimation of the coefficients associated with the small data set, under the joint weak borrowing prior $\mathcal{P}_{\rm W}$ is very problematic. For example, the chains for the variables $\vartheta_{21}, \vartheta_{22}, \vartheta_{24}$ and ϑ_{25}, are kept stuck to certain regions of the state space for a large number of iterations of the Gibbs sampler. The situation can be corrected by the introduction of strong borrowing via the $\mathcal{P}_{\rm SN}$ prior. The convergence of the ergodic means to the true values under the $\mathcal{P}_{\rm SN}$ prior are given in Figure 5.5(g)-(l) (solid curves in red).

In the first two lines of Table 5.2, we can see the effect of the weak borrowing of strength prior $\mathcal{P}_{\rm W}$. The estimation of the coefficients of the first dynamical system is very accurate, enabling the identification of the cubic map. The part of the joint estimation based on the short time series exhibits large errors, hindering identification. In the last two lines of Table 5.2, we present the effect of the strong borrowing prior. Borrowing gives the part of the estimation associated with the short time series nice mixing properties, and it reduces the average PARE from 7.51% to a mere 0.26%, thus, enabling the identification of the quadratic map.

In Figure 5.6(a)-(b) we present the predictive posterior marginal noise density pair $(\hat{f}_{21}, \hat{f}_{22})$, under the $\mathcal{P}_{\rm W}$ and $\mathcal{P}_{\rm SN}$ priors, in black and red solid curves, respectively. The posterior mean

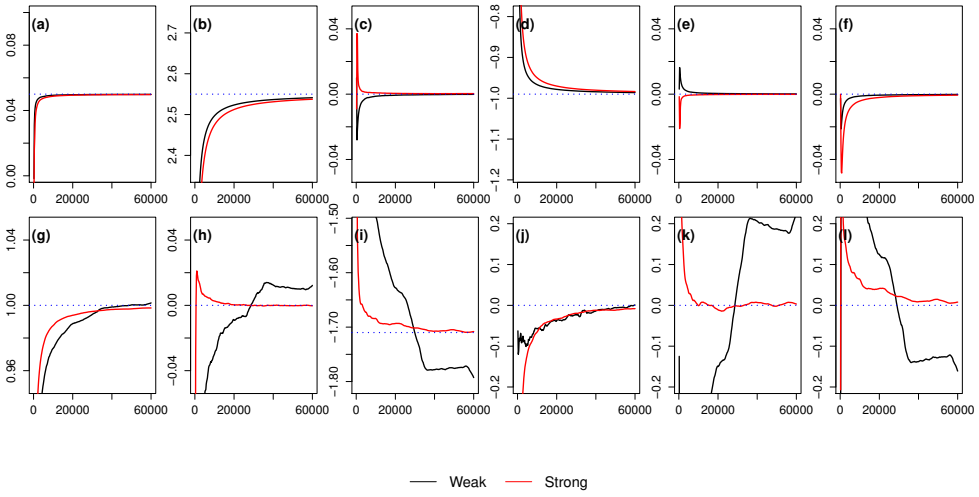

Figure 5.5: Ergodic averages for the $(\vartheta_1, \vartheta_2)$ pair of coefficients of the modeling polynomials under weak (solid curves in black) and strong (solid curves in red) borrowing. The averages associated with the cubic map $\mathcal{C}_1$ appear in Figures (a)-(f), and the averages associated with the quadratic map $\mathcal{Q}_1$ appear in Figures (g)-(l).

Table 5.2: PAREs of the joint GSBR coefficient estimation based on the pair of time series $(x_1^{(200)}, x_2^{(20)})$ under the f_2 noise pair. The estimation is based on a polynomial modeling of fifth degree, assuming weak borrowing and strong borrowing.

Prior	Time series	ϑ_{j0}	ϑ_{j1}	ϑ_{j2}	ϑ_{j3}	ϑ_{j4}	ϑ_{j5}	$\bar{\vartheta}$
$\mathcal{P}_{\mathrm{W}}$	$x_1^{(200)}$	0.18	0.15	0.03	0.20	0.01	0.02	0.10
	$x_2^{(20)}$	0.37	1.35	4.95	0.22	22.31	15.87	7.51
$\mathcal{P}_{\mathrm{SN}}$	$x_1^{(200)}$	0.00	0.29	0.02	0.46	0.01	0.06	0.14
	$x_2^{(20)}$	0.02	0.02	0.08	0.67	0.05	0.73	0.26

matrices of the selection probabilities under the $\mathcal{P}_{\mathrm{W}}$ and $\mathcal{P}_{\mathrm{SN}}$ priors, are

$$\mathbb{E}_{\mathrm{W}}\{(p_{ji})|\, x_1^{(200)}, x_2^{(20)}\} = \begin{pmatrix} 0.995 & 0.005 \\ 0.033 & 0.967 \end{pmatrix}, \quad \mathbb{E}_{\mathrm{SN}}\{(p_{ji})|\, x_1^{(200)}, x_2^{(20)}\} = \begin{pmatrix} 0.005 & 0.995 \\ 0.956 & 0.044 \end{pmatrix},$$

respectively. Under the weak borrowing prior, the joint estimation is nearly independent, as the off-diagonal elements of the first matrix are close to zero. We remark here, that the strong borrowing prior is very efficient as 95.6% of the noise samples are coming from the common component. The predictive density of the marginal posterior pair of initial conditions (x_{10}, x_{20}) is depicted in Figure 5.6(c)-(d). More specifically, in Figure 5.6(d) we can see that the estimation part corresponding to the short time series under the strong prior configuration is more accurate. In Figure 5.6(e)-(f), we exhibit the predictive density of the marginal posterior pair $(x_{1,201}, x_{2,21})$ for one future observation. The posterior mean estimations for the cubic dynamical system, under the weak and strong borrowing priors are of the same quality, yet, the joint estimation associated with the quadratic map, under the strong borrowing prior, shrinks the length of the corresponding 95%-HPDI by a factor of 0.13, namely

$$\mathrm{HPDI}(x_{2,21}; \mathcal{P}_{\mathrm{W}}) = [0.426, 0.554] \quad \text{and} \quad \mathrm{HPDI}(x_{2,21}; \mathcal{P}_{\mathrm{SN}}) = [0.485, 0.502].$$

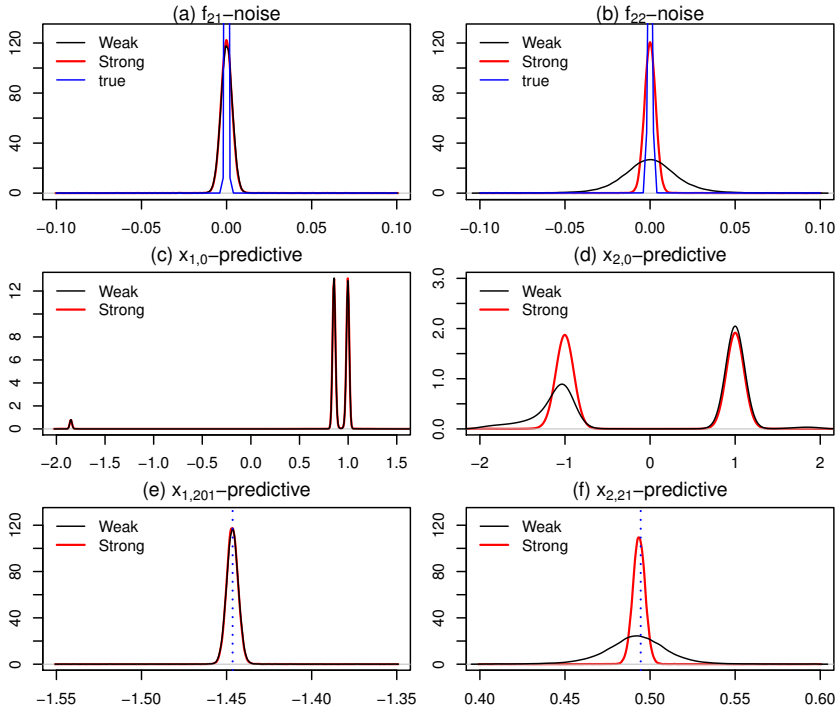

Figure 5.6: Kernel density estimations based on the predictive samples coming from the PD-GSBR Gibbs sampler. Weak borrowing corresponds to the densities in black, and strong borrowing to the densities in red. Figures (a), (c) and (e) correspond to the cubic map $\mathcal{C}_1$, and Figures (b), (d) and (f) correspond to the quadratic map $\mathcal{Q}_1$. The noise predictive densities are given in Figures (a) and (b). The initial conditions predictive densities are given in Figures (c) and (d). In Figures (e) and (f) we give the predictive densities of the first future observation.

C. Borrowing between cubic maps perturbed by an identical noise process: In our third numerical example, we have generated a pair of time series via the configuration

$$(\mathcal{C}_1 + f_{22}, \mathcal{C}_2 + f_{22}) \to (x_1^{(200)}, x_2^{(30)}).$$

In this example, both cubic maps are perturbed dynamically by the noise process f_{22} given in (5.13). Both dynamical trajectories experience noise induced jumps. The traces of the two time series $x_1^{(200)}$ and $x_2^{(30)}$ are given in Figure 5.7(a) and Figure 5.7(b) respectively. Here we will demonstrate numerically, that strong informative borrowing of strength between the estimated noise processes via the joint prior $\mathcal{P}_{\rm SI}$, accelerates the slow convergence of the ergodic averages corresponding to the short time series.

The ergodic averages for the ϑ-coefficients, coming from the PD-GSBR sampler, under the weak borrowing prior specification $\mathcal{P}_{\rm W}$ (black solid curves) and the strong borrowing noninformative prior specification $\mathcal{P}_{\rm SI}$ (red solid curves), are given in Figure 5.8(a)-(f), and Figure 5.8(g)-(l), respectively. The averages associated with the large data set, are converging fast irrespectively of the joint prior configuration, yet, convergence associated with the short time series, under weak borrowing is very slow. For example, the chains for the variables ϑ_{24} and ϑ_{25} and especially the chain for the variable ϑ_{22}, have a very slow convergence. This situation can be corrected by the introduction of the strong borrowing prior $\mathcal{P}_{\rm SI}$. The improved convergence of the ergodic means to the true values, are depicted in Figure 5.8(g)-(l) (solid curves in red).

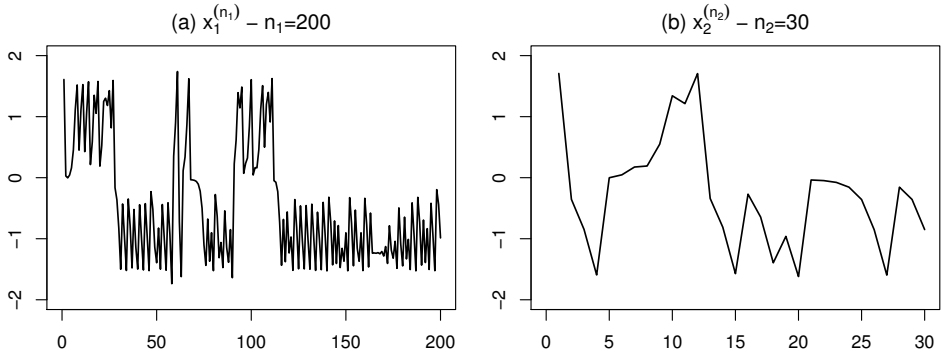

Figure 5.7: The f_1 noise pair perturbed time series corresponding to the cubic map $\mathcal{C}_1$ and the cubic map $\mathcal{C}_2$ are given in Figures (a) and (b), respectively.

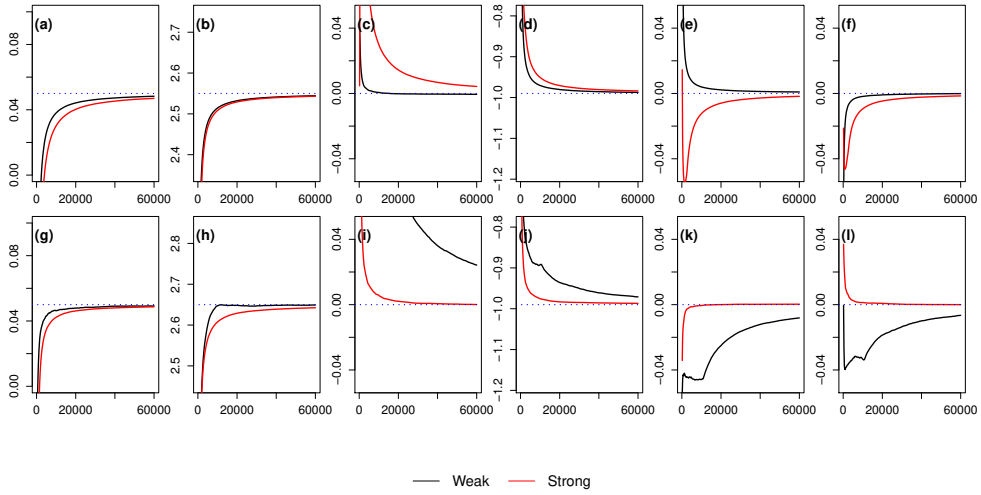

Figure 5.8: Ergodic averages for the $(\vartheta_1, \vartheta_2)$ pair of coefficients of the modeling polynomials under weak (solid curves in black) and strong (solid curves in red) borrowing. The averages associated with the cubic map $\mathcal{C}_1$ appear in Figures (a)-(f), and the averages associated with the cubic map $\mathcal{C}_2$ appear in Figures (g)-(l).

In the first two lines of Table 5.3, we can see the effect of weak borrowing of strength. The estimation for the coefficients of the first cubic map $\mathcal{C}_1$ is very accurate, attaining an average PARE of 0.08%, thus, enabling the identification of the map. The part of the joint estimation based on the short time series exhibits larger errors, hindering identification. In the last two lines of Table 5.3, we present the effect of the strong informative borrowing prior. Borrowing accelerates the part of the estimation associated with the short time series, and it reduces the average PARE from 1.14% to a mere 0.10%, thus, enabling the identification of the second cubic map.

In Figure 5.9(a)-(b) we present the predictive posterior marginal noise density pairs under under weak and strong borrowing prior specifications, in black and red solid curves, respectively. The posterior mean matrices of the selection probabilities, are

$$\mathbb{E}_{\mathrm{W}}\{(p_{ji})|\, x_1^{(200)}, x_2^{(30)}\} = \begin{pmatrix} 0.879 & 0.121 \\ 0.100 & 0.900 \end{pmatrix}, \; \mathbb{E}_{\mathrm{SI}}\{(p_{ji})|\, x_1^{(200)}, x_2^{(30)}\} = \begin{pmatrix} 0.005 & 0.995 \\ 0.976 & 0.024 \end{pmatrix},$$

Table 5.3: PAREs for the PD-GSBR estimation of the ϑ-coefficients, based on the pair of time series $(x_1^{(200)}, x_2^{(30)})$, under the identical noise process f_{22}, assuming weak and strong borrowing.

Prior	Time series	ϑ_{j0}	ϑ_{j1}	ϑ_{j2}	ϑ_{j3}	ϑ_{j4}	ϑ_{j5}	$\bar{\vartheta}$
$\mathcal{P}_{\mathrm{W}}$	$x_1^{(200)}$	0.36	0.01	0.06	0.00	0.02	0.00	0.08
	$x_2^{(30)}$	1.16	0.19	2.36	1.70	0.78	0.66	1.14
$\mathcal{P}_{\mathrm{SI}}$	$x_1^{(200)}$	0.60	0.02	0.50	0.34	0.21	0.16	0.31
	$x_2^{(30)}$	0.31	0.03	0.09	0.04	0.04	0.09	0.10

In the case of the short time series, under weak borrowing, when sampling from the noise component, the samples come from the common component only 10% of the times. Under strong borrowing, sampling from the common component increases to 97.6%.

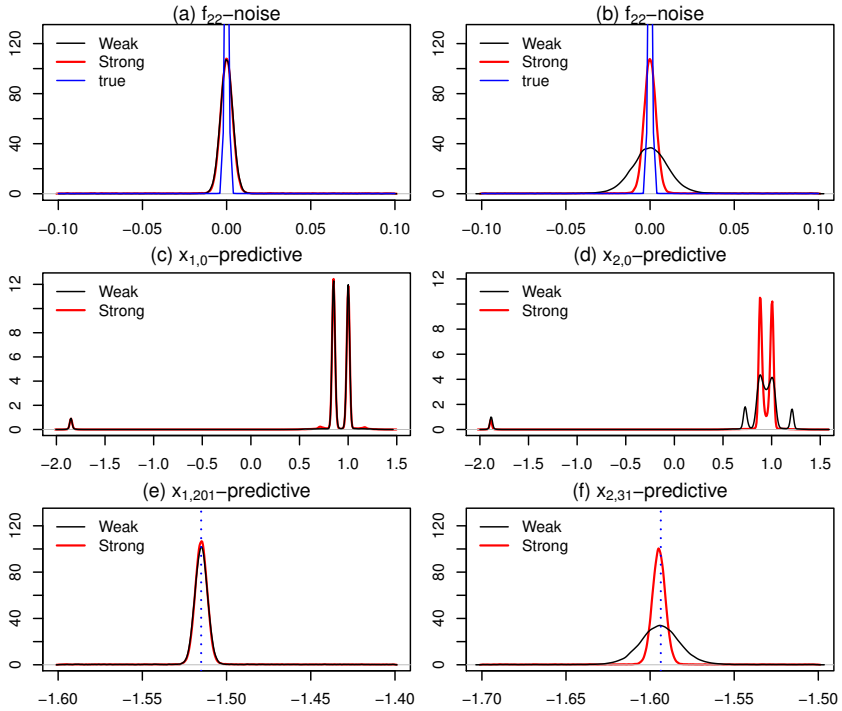

Figure 5.9: Kernel density estimations based on the predictive samples coming from the PD-GSBR Gibbs sampler. Weak borrowing corresponds to the densities in black, and strong borrowing to the densities in red. Figures (a), (c) and (e) correspond to the cubic map $\mathcal{C}_1$, and Figures (b), (d) and (f) correspond to the cubic map $\mathcal{C}_2$. The noise predictive densities are given in Figures (a) and (b). The initial conditions predictive densities are given in Figures (c) and (d). In Figures (e) and (f) we give the predictive densities of the first future observation.

The predictive density of the marginal posterior pair of initial conditions (x_{10}, x_{20}) is depicted in Figure 5.9(c)-(d). More specifically, in Figure 5.9(d) we can see that the predictive density associated with the short time series under weak borrowing, exhibits two more spurious modes at about 0.74 and 1.18 (solid black curve). The spurious modes disappear after the introduction of strong borrowing (solid red curve). In Figure 5.9(e)-(f), we exhibit the predictive density of the marginal posterior pair $(x_{1,201}, x_{2,31})$. The posterior mean estimations for the cubic dynamical systems, under the weak and strong borrowing priors are of the same quality, yet, under strong borrowing the predictive density associated with the short time series cubic map, exhibits a

95%-HPDI shrinkage factor of 0.45, namely

$$\mathrm{HPDI}(x_{2,31};\mathcal{P}_{\mathrm{W}}) = [-1.622, -1.566] \quad \text{and} \quad \mathrm{HPDI}(x_{2,31};\mathcal{P}_{\mathrm{SI}}) = [-1.607, -1.582].$$

5.5 A joint parametric Gibbs sampler

When there is evidence that the noise components are coming from the same gaussian distribution, i.e. $z_{ji} \overset{\text{iid}}{\sim} \mathcal{N}(0, \tau^{-1})$, where τ is the unknown precision of the normal component, borrowing of strength between the observed time series, can be achieved by a joint parametric Gibbs sampler that assumes only gaussian noise.

If this is the case, the following parametric hierarchical model for the time series observations $\{x_j^{(n_j)} : j = 1, \ldots, m\}$, conditional on the unknown initial conditions $\{x_{j,1:l_j} : j = 1, \ldots, m\}$ is sufficient for reconstruction and prediction, namely

$$\begin{aligned} &x_{ji} \,|\, x_{j,i:l_j}, \tau, \vartheta_j \overset{\text{ind}}{\sim} \mathcal{N}(g_j(\vartheta_j, x_{j,i:l_j}), \tau^{-1}),\ i = 1, \ldots, n_j,\ j = 1, \ldots, m \\ &\tau \sim \mathcal{G}(a, b). \end{aligned}$$

The likelihood based on a sample of size n_1 from the system g_1, n_2 from g_2 etc., n_m from g_m, is proportional to

$$\prod_{j=1}^{m} \prod_{i=1}^{n_j} \tau^{1/2} \exp\left\{ -\frac{\tau}{2} (x_{ji} - g_j(\vartheta_j, x_{j,i:l_j}))^2 \right\}.$$

To complete the model, and attempting a noninformative prior specification, we assign the translation invariant priors $f(x_{j,1:l_l}) \propto 1$ and $f(\vartheta_j) \propto 1$ to the initial conditions and the model coefficients, respectively, and a scale invariant prior $f(\tau) \propto \tau^{-1}$ to the precision variable. Then the posterior distribution for τ, $\vartheta = (\vartheta_1, \ldots, \vartheta_m)$ and $x_{j,1:l_j}$, attains the representation

$$f(\tau, \vartheta, x_{j,1:l_j} \,|\, x_1^{(n_1)}, \ldots, x_m^{(n_m)}) \propto \prod_{j=1}^{m} \prod_{i=1}^{n_j} \tau^{1/2} \exp\left\{ -\frac{\tau}{2} (x_{ji} - g_j(\vartheta_j, x_{j,i:l_j}))^2 \right\}.$$

The full conditional distributions can be derived in a similar way as in Section 5.3. More specifically, the full conditional for the common precision term λ, is given by

$$(\tau \,|\, \cdots) \sim \mathcal{G}\left(\frac{1}{2} \sum_{j=1}^{m} n_j,\ \frac{1}{2} \sum_{j=1}^{m} \sum_{i=1}^{n_j} (x_{ji} - g_j(\vartheta_j, x_{j,i:l_j}))^2 \right).$$

We remark the double sum appearing in the rate parameter of the full conditional of τ. This is how borrowing of strength is realized in a parametric setting.

Parametric borrowing between quadratic maps: To illustrate the joint parametric Gibbs sampler, for the specific case $m = 2$ and $l_1 = l_2 = 1$, we use a pair of time series realized via

the configuration

$$(\mathcal{Q}_1 + \mathcal{N}(0, \sigma^2), \mathcal{Q}_2 + \mathcal{N}(0, \sigma^2)) \to (x_1^{(200)}, x_2^{(30)}),\ \sigma^2 = 10^{-6}.$$

The time series $x_1^{(200)}$ and $x_2^{(30)}$ are presented in Figure 5.10(a) and (b), respectively.

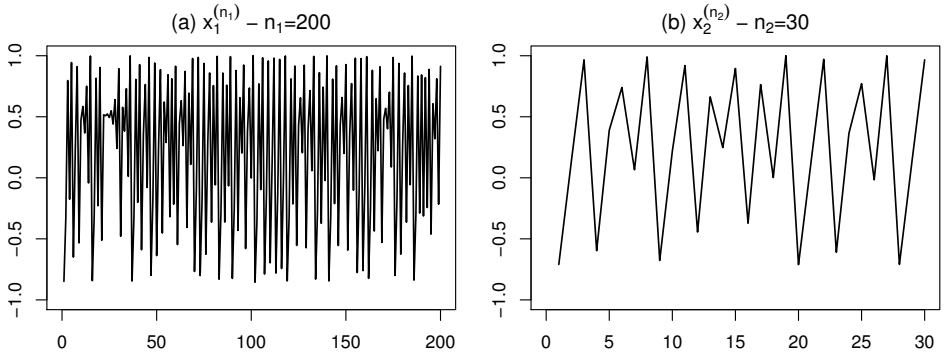

Figure 5.10: The gaussian noise perturbed time series corresponding to the quadratic map $\mathcal{Q}_1$ and the quadratic map $\mathcal{Q}_2$ are given in Figures (a) and (b), respectively.

The ergodic averages for the ϑ_1 and ϑ_2 coefficient vectors, coming from independent parametric Gibbs samplers and the joint parametric Gibbs sampler, are given in Figure 5.11. The averages associated with the large data set, are converging fast irrespectively of the parametric sampler, yet, convergence associated with the short time series, under the independent sampler exhibits mixing issues. This becomes apparent from the chains of the variables ϑ_2 in Figure 5.11(i) through (l) (solid curves in black). This situation is corrected by the introduction of the joint parametric sampler. The improved convergence of the ergodic means to the true values, are depicted in Figure 5.11(g) through 11(l) (solid curves in red).

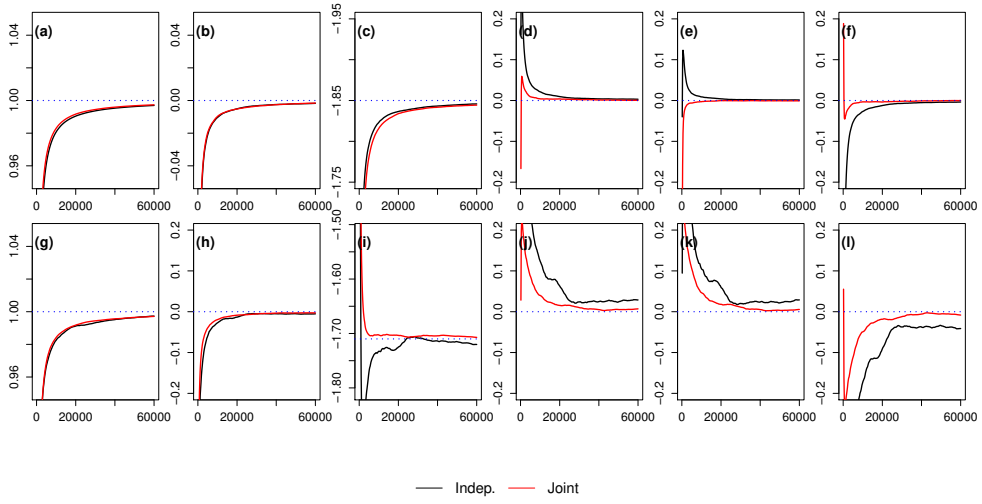

Figure 5.11: Ergodic averages for the $(\vartheta_1, \vartheta_2)$ pair of coefficients of the modeling polynomials under the independent parametric samplers (solid curves in black) and the joint parametric sampler (solid curves in red). The averages associated with the quadratic map $\mathcal{Q}_1$ appear in Figures (a)-(f), and the averages associated with the quadratic map $\mathcal{Q}_2$ appear in Figures (g)-(l).

In the first two lines of Table 5.4, we can see the effect reconstructing the coefficients of the quadratic dynamical systems with the independent parametric samplers. The reconstruction of the first quadratic map $\mathcal{Q}_1$ is very accurate, with an average PARE of 0.05%, thus, enabling the identification of the map. The independent estimation based on the short time series exhibits larger errors, hindering identification. In the last two lines of Table 5.4, we present the effect of reconstruction under the joint parametric sampler. The average PARE associated with the large time series decreases even further to 0.03%. At the same time it eliminates the mixing issues associated with the short time series, and it reduces the average PARE from 1.96% to a mere 0.36%, thus, enabling the identification of the second quadratic map.

Table 5.4: PAREs for the estimation of the ϑ-coefficients, based on the pair of time series $(x_1^{(200)}, x_2^{(30)})$, under the the independent and the joint parametric samplers.

Par. Sampler	Time series	ϑ_{j0}	ϑ_{j1}	ϑ_{j2}	ϑ_{j3}	ϑ_{j4}	ϑ_{j5}	$\bar{\vartheta}$
Independent	$x_1^{(200)}$	0.02	0.09	0.08	0.04	0.02	0.04	0.05
	$x_2^{(30)}$	0.09	0.50	0.98	3.08	3.05	4.08	1.96
Joint	$x_1^{(200)}$	0.01	0.04	0.01	0.01	0.07	0.03	0.03
	$x_2^{(30)}$	0.01	0.10	0.01	0.75	0.41	0.89	0.36

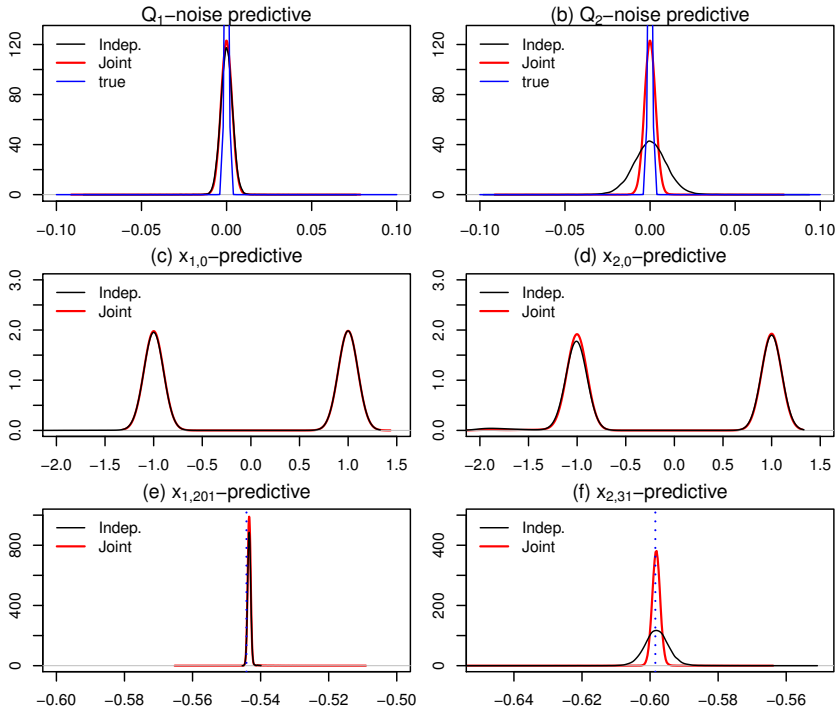

Figure 5.12: Kernel density estimations based on the predictive samples coming from the independent Gibbs samplers correspond to the densities in black, the joint Gibbs sampler predictives correspond to the densities in red. Figures (a), (c) and (e) correspond to the quadratic map $\mathcal{Q}_1$, and Figures (b), (d) and (f) correspond to the quadratic map $\mathcal{Q}_2$. The noise predictive densities are given in Figures (a) and (b). The initial conditions predictive densities are given in Figures (c) and (d). In Figures (e) and (f) we give the predictive densities of the first future observation.

In Figure 5.12(a)-(b) we present the normal noise densities based on the estimated precision

τ under the independent parametric samplers (solid black curves) and the joint parametric sampler (solid red curves). The predictive density of the posterior pair of initial conditions (x_{10}, x_{20}) under the independent and joint estimations are depicted in Figure 5.12(c)-(d). In Figure 5.12(e)-(f), we display the predictive densities of the marginal posterior pair $(x_{1,201}, x_{2,31})$ coming from the independent and joint estimations. The posterior mean estimations for the second quadratic map, are of of about the same quality, yet, the joint sampler shrinks the length of the corresponding 95%-HPDI by a factor of 0.29, namely

$$\text{HPDI}(x_{2,31}; \text{Indep.}) = [-0.605, -0.591] \text{ and } \text{HPDI}(x_{2,31}; \text{Joint}) = [-0.600, -0.596].$$

5.6 Conclusions

We have proposed a new Bayesian nonparametric model for the pairwise reconstruction of the dynamical equations based on observed dynamically noise perturbed chaotic time series data. The PD-GSBR model is based on the multivariate nonparametric prior model PDGSBP, here applied to the additive error processes of the dynamical equations. Experiments on simulated pairs of data sets are indicating that when the densities of the noise processes have common characteristics, we are able, by imposing certain informative prior specifications over the selection probabilities of the PD-GSBR model, to recover the dynamical equation corresponding to the short time series for which an independent identification is not possible.

Although in principle the model can estimate simultaneously more than two dynamical equations, in more than two dimensions, borrowing of strength coming from the prevalent data set tends to be more weak. This is because borrowing works as a two sided interplay between the short and the prevalent time series, also, in more than two dimensions, there will be borrowing between the short time series, thus, corrupting the overall effect of strength borrowing from the large data set.

We have also introduced a joint parametric Gibbs sampler. In this case the dynamical noise is assumed to be normal, coming approximately from the same noise source. In this case the borrowing of strength, between the pairs of data sets, comes from the full conditional of the common precision.

Chapter 6

Conclusions and future research

6.1 Conclusions

In this thesis, firstly we have developed a Bayesian nonparametric model based on the Geometric stick breaking process (Fuentes-García et al., 2010) for the reconstruction and prediction of random dynamical systems, dropping the assumption of Gaussian noise. We have shown that as the dynamical noise departs from normality, simple MCMC models are inefficient. Modeling the error process as an infinite mixture of zero mean normals, our proposed GSBR model is able to infer the number of unknown components and their variances, that is infers the density of the error process directly from observed data. We have shown through numerical examples that the associated quasi–invariant measure of the random dynamical system appears naturally as posterior predictive marginal of the out–of–sample variables forming a prediction barrier.

Next, we have constructed pairwise dependent random probability measures based on GSB process namely the PDGSBP prior to use them in mixture modeling to generate random densities which are thought to be related. That is, we have modeled the random densities to be generated via

$$f_j(x) = f_j(x \,|\, \mathbb{Q}_j) = \sum_{l=1}^{\infty} p_{jl} g_{jl}(x \,|\, \mathbb{G}_{jl}), \quad \mathbb{Q}_j = \sum_{l=1}^{m} p_{jl} \mathbb{G}_{jl}, \quad 1 \leq j \leq m,$$

where m is the number of different populations and $\mathbb{G}_{jl} = \mathbb{G}_{lj}$ are independent GSB processes. The $g_{jl}(x \,|\, \mathbb{G}_{jl}) = \int_{\Theta} K(x \,|\, \theta) \mathbb{G}_{jl}(\mathrm{d}\theta)$ random densities are independent mixtures of GSB processes. The aim is to share information among groups and improve estimation of each density especially for those whose the corresponding size is small.

Based on the PDGSBP prior we extended the GSBR model to PDGSBP reconstruction model, a Bayesian nonparametric mixture model for the joint full reconstruction of a finite collection of dynamical equations, given observed dynamically–noisy–corrupted chaotic time series. We have shown numerically that whenever there is at least on sufficiently large data set, using carefully selected informative borrowing–of–strength prior specifications we are able to reconstruct those dynamical processes that are responsible for the generation of time series with small samples sizes; namely sample sizes that are inadequete for an independent reconstruction, i.e. with GSBR model.

In all of the problems described in this thesis, we have shown that the GSB random probability measures are sufficient for estimation and prediction purposes; that is, making the weights more exotic does not actually enlarge the support of the nonparametric prior. Moreover, the corresponding Gibbs samplers for estimation with GSB random probability measures are faster and easier to implement than the Dirichlet process counterparts.

6.2 Directions for future research

Because of the interdisciplinary profile of the research presented in this thesis, a number of interesting research paths appeared during the development of our methods which we believe should be explored in the near future. Below we provide some of the research paths in the field of random dynamical systems as also in the field of Bayesian nonparametric inference.

6.2.1 Random dynamical systems

Modeling dynamical systems with Gaussian processes

Most of the methods that aim for the reconstruction of dynamical equations assume some known functional form for the deterministic part of the random dynamical system. It would worth to extend the GSBR model by assigning a prior over the space of functions, i.e. a Gaussian Process (GP) (Rasmussen & Williams, 2006) prior for the deterministic part extending the GSBR model to a full Bayesian nonparametric model.

Extension of the GSBR model to a state space model

When the available data are contaminated with dynamical and observational noise, the GSBR model could be extended to a q–lagged state space model as

$$\begin{aligned} X_i &= \varrho(\vartheta, X_{i-1}, \ldots, X_{i-q}) + Z_i, \quad i \geq q \\ Y_i &= h(\varphi, X_i) + W_i, \end{aligned}$$

for some function h. Here the assumption is that noisy measurements of the output occur at all times, making the sequence $X^{(n)}$ unobservable. The set of observations in this case is the $Y^{(n)}$ time series, which can be modeled via a GSB random measure $\mathbb{P}_Y$. Then the latent $X^{(n)}$ series can be modeled with a second independent GSB random measure $\mathbb{P}_X$, such that the random variables $[X_i \,|\, X_{i-1}, \ldots, X_{i-q}, \vartheta, \mathbb{P}_X]$ and $[Y_i \,|\, X_i, \varphi, \mathbb{P}_Y]$ are independent. In this case we have to estimate the initial condition $(X_0, \ldots, X_{q-1}, Y_0)$, the parameter (ϑ, φ), the density of the noise component (Z_i, W_i) as well as the hidden orbit $\{X_i : i = q, \ldots, n\}$.

6.2.2 Bayesian nonparametrics

Generalization of the PDGSBP model to include all possible interactions

An interesting research path would be the generalization of the pairwise dependent $\mathbb{Q}_j$ measures involved in the PDGSBP, to include all possible interactions, in the sence that

$$\mathbb{Q}_j(\,\cdot\,) = p_j\,\mathbb{G}_j(\,\cdot\,) + \sum_{l=2}^{m}\sum_{\eta\,\in\,\mathcal{C}_{j,l,m}} p_{j,\eta}\,\mathbb{G}_{\eta_{(j)}}(\,\cdot\,) \quad \text{with} \quad p_j + \sum_{l=2}^{m}\sum_{\eta\,\in\,\mathcal{C}_{j,l,m}} p_{j,\eta} = 1,$$

where the $\mathbb{G}_j$ and the $\mathbb{G}_{\eta_{(j)}}$'s are independent GSB processes, $\mathcal{C}_{j,l,m} = \{(k_1,\ldots,k_{l-1}) : 1 \leq k_1 < \cdots < k_{l-1} \leq m, k_r \neq j, 1 \leq r \leq m-1\}$ and $\eta_{(j)}$ is the ordered vector of the elements of the vector η and $\{j\}$. Now the f_j densities will be a mixture of 2^{m-1} GSB mixtures, and the total number of the independent GSB processes needed to model $(f_1,\ldots,f_m)$ will be $2^m - 1$.

Due to the exponential growth of the random measures in need as m gets large it would be interesting to develop parallel MCMC algorithms for the GSB process as also for its multivariate extensions.

Identification of common and idiosyncratic parts in dependent mixture models.

Due to identifiability issues it is not possible to perform density estimation for the random densities g_{jl}, that is, the common and idiosyncrating parts composing the random densities f_j of the PDGSBP model. It would worth to extend a univariate mixture model proposed by Mena & Walker (2015) in the multivariate case. Perhaps, generalizing this prior in the multivariate case will sovle the identifiability issues arising when one tries to estimate the idiosyncratic as also the common parts of the mixture densities. This generalization may lead to accurate estimation of the common and idiosyncratic parts of the random densities as also accurate estimation for their number of active clusters.

Bibliography

Alligood, K. T., Sauer, T. D., & Yorke, J. A. (1996). *Chaos.* Springer.

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

Arnold, L. (2013). *Random dynamical systems.* Springer Science & Business Media.

Besag, J., & Green, P. J. (1993). Spatial statistics and Bayesian computation. *Journal of the Royal Statistical Society. Series B (Methodological)*, *55*(1), 25–37.

Bezanson, J., Karpinski, S., Shah, V. B., & Edelman, A. (2012). Julia: A fast dynamic language for technical computing. *arXiv preprint arXiv:1209.5145*.

Bhattacharya, R., & Majumdar, M. (2007). *Random dynamical systems: theory and applications.* Cambridge University Press.

Birkhoff, G. D. (1931). Proof of the ergodic theorem. *Proceedings of the National Academy of Sciences*, *17*(12), 656–660.

Blackwell, D., & MacQueen, J. B. (1973). Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, (pp. 353–355).

Borchers, H. W. (2015). *pracma: Practical Numerical Math Functions,.* R package version 1.8.3.

Brix, A. (1999). Generalized gamma measures and shot-noise Cox processes. *Advances in Applied Probability*, *31*(4), 929–953.

Broer, H., & Takens, F. (2010). *Dynamical systems and chaos*, vol. 172. Springer Science & Business Media.

Brooks, S., Gelman, A., Jones, G., & Meng, X.-L. (2011). *Handbook of Markov Chain Monte Carlo.* CRC press.

Bulla, P., Muliere, P., & Walker, S. G. (2009). A Bayesian nonparametric estimator of a multivariate survival function. *Journal of Statistical Planning and Inference*, *139*(10), 3639–3648.

Chan, K. S., & Tong, H. (2013). *Chaos: a statistical perspective.* Springer Science & Business Media.

Collet, P., Martínez, S., & San Martín, J. (2012). *Quasi-stationary distributions: Markov chains, diffusions and dynamical systems.* Springer Science & Business Media.

DAMIEN, P., WAKEFIELD, J., & WALKER, S. (1999). Gibbs sampling for bayesian non-conjugate and hierarchical models by using auxiliary variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *61*(2), 331–344.

DAVIES, M. (1998). Nonlinear noise reduction through Monte Carlo sampling. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *8*(4), 775–781.

DE FINETTI, B. (1937). La prévision: ses lois logiques, ses sources subjectives. *Annales de l'institut Henri Poincaré*, *7*(1), 1–68.

DE FINETTI, B. (1938). Sur la condition d' equivalence partielle. *Actualités Scientifiques et Industrielles*, *739*, 5–18.

DE IORIO, M., MÜLLER, P., ROSNER, G. L., & MACEACHERN, S. N. (2004). An ANOVA model for dependent random measures. *Journal of the American Statistical Association*, *99*(465), 205–215.

DUNSON, D. B., & PARK, J. H. (2008). Kernel stick-breaking processes. *Biometrika*, *95*(2), 307–323.

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

FOTI, N. J., & WILLIAMSON, S. A. (2015). A survey of non-exchangeable priors for Bayesian nonparametric models. *IEEE transactions on pattern analysis and machine intelligence*, *37*(2), 359–371.

FOX, E., JORDAN, M. I., SUDDERTH, E. B., & WILLSKY, A. S. (2009). Sharing features among dynamical systems with Beta processes. *Advances in Neural Information Processing Systems*, (pp. 549–557).

FOX, E. B., SUDDERTH, E. B., JORDAN, M. I., & WILLSKY, A. S. (2008). An HDP–HMM for systems with state persistence. In *Proceedings of the 25th international conference on Machine learning*, (pp. 312–319). ACM.

FUENTES-GARCÍA, R., MENA, R. H., & WALKER, S. G. (2009). A nonparametric dependent process for Bayesian regression. *Statistics & Probability Letters*, *79*(8), 1112–1119.

FUENTES-GARCÍA, R., MENA, R. H., & WALKER, S. G. (2010). A new Bayesian nonparametric mixture model. *Communications in Statistics-Simulation and Computation*, *39*(4), 669–682.

GALOR, O. (2007). *Discrete dynamical systems*. Springer Science & Business Media.

GE, H., CHEN, Y., WAN, M., & GHAHRAMANI, Z. (2015). Distributed inference for Dirichlet process mixture models. In *International Conference on Machine Learning*, (pp. 2276–2284).

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

GIAKOUMATOS, S. G., DELLAPORTAS, P., & POLITIS, D. N. (2005). Bayesian analysis of the unobserved arch model. *Statistics and Computing*, *15*(2), 103–111.

GILKS, W. R., BEST, N., & TAN, K. (1995). Adaptive rejection Metropolis sampling within Gibbs sampling. *Applied Statistics*, *44*(4), 455–472.

GOERG, G. M. (2013). Forecastable component analysis. *Proceedings of the 30th International Conference on Machine Learning*, *28*, 64–72.

GOERG, G. M. (2016). *ForeCA: An R package for Forecastable Component Analysis*. R package version 0.2.4.

GRIFFIN, J. E. (2010). Inference in infinite superpositions of non–Gaussian Ornstein–Uhlenbeck processes using Bayesian nonparametic methods. *Journal of Financial Econometrics*, *9*(3), 519–549.

GRIFFIN, J. E., KOLOSSIATIS, M., & STEEL, M. F. (2013). Comparing distributions by using dependent normalized random-measure mixtures. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *75*(3), 499–529.

GRIFFIN, J. E., & STEEL, M. J. (2006). Order-based dependent Dirichlet processes. *Journal of the American statistical Association*, *101*(473), 179–194.

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

HATJISPYROS, S. J., MERKATAS, C., NICOLERIS, T., & WALKER, S. G. (2017a). Dependent mixtures of geometric weights priors. *Computational Statistics & Data Analysis*, *119*(3), 1–18.

HATJISPYROS, S. J., NICOLERIS, T., & WALKER, S. G. (2009). A Bayesian nonparametric study of a dynamic nonlinear model. *Computational Statistics & Data Analysis*, *53*(12), 3948–3956.

HATJISPYROS, S. J., NICOLERIS, T., & WALKER, S. G. (2011). Dependent mixtures of Dirichlet processes. *Computational Statistics & Data Analysis*, *55*(6), 2011–2025.

HATJISPYROS, S. J., NICOLERIS, T., & WALKER, S. G. (2016). Random density functions with common atoms and pairwise dependence. *Computational Statistics & Data Analysis*, *101*, 236–249.

HATJISPYROS, S. J., NICOLERIS, T., & WALKER, S. G. (2017b). Bayesian nonparametric density estimation under length bias. *Communications in Statistics-Simulation and Computation*, *46*(10), 8064–8076.

HATJISPYROS, S. J., & YANNACOPOULOS, A. N. (2005). A random dynamical system model of a stylized equity market. *Physica A: Statistical Mechanics and its Applications*, *347*, 583–612.

HJORT, N. L., HOLMES, C., MÜLLER, P., & WALKER, S. G. (2010). *Bayesian nonparametrics*, vol. 28. Cambridge University Press.

ISHWARAN, H., & JAMES, L. F. (2001). Gibbs sampling methods for stick–breaking priors. *Journal of the American Statistical Association*, *96*(453), 161–173.

JAOUA, N., DUFLOS, E., VANHEEGHE, P., & SEPTIER, F. (2013). Bayesian nonparametric state and impulsive measurement noise density estimation in nonlinear dynamic systems. *In Proceedings 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5755–5759.

JENSEN, M. J., & MAHEU, J. M. (2010). Bayesian semiparametric stochastic volatility modeling. *Journal of Econometrics*, *157*(2), 306–316.

KALLI, M., GRIFFIN, J. E., & WALKER, S. G. (2011). Slice sampling mixture models. *Statistics and computing*, *21*(1), 93–105.

KARAKUŞ, O., KURUOĞLU, E. E., & ALTINKAYA, M. A. (2015). Estimation of the nonlinearity degree for polynomial autoregressive processes with RJMCMC. *In Proceedings 23rd European Signal Processing Conference (EUSIPCO)*, 953–957.

KINGMAN, J. F. C. (1967). Completely random measures. *Pacific Journal of Mathematics*, *21*(1), 59–78.

KINGMAN, J. F. C. (1975). Random discrete distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, *37*(1), 1–22.

KLENKE, A. (2013). *Probability theory: a comprehensive course*. Springer Science & Business Media.

KOLOSSIATIS, M., GRIFFIN, J. E., & STEEL, M. F. (2013). On Bayesian nonparametric modelling of two correlated distributions. *Statistics and Computing*, *23*(1), 1–15.

KRAUT, S., FEUDEL, U., & GREBOGI, C. (1999). Preference of attractors in noisy multistable systems. *Physical Review E*, *59*(5), 5253–5260.

LASOTA, A., & MACKEY, M. C. (1994). *Chaos, fractals, and noise: stochastic aspects of dynamics*. Springer Science & Business Media.

LIANG, F., LIU, C., & CARROLL, R. (2011). *Advanced Markov Chain Monte Carlo methods: learning from past samples*, vol. 714. John Wiley & Sons.

LIJOI, A., NIPOTI, B., & PRÜNSTER, I. (2014). Dependent mixture models: clustering and borrowing information. *Computational Statistics & Data Analysis*, *71*, 417–433.

LO, A. Y., ET AL. (1984). On a class of Bayesian nonparametric estimates: I. density estimates. *The Annals of Statistics*, *12*(1), 351–357.

LUCHINSKY, D. G., SMELYANSKIY, V. N., DUGGENTO, A., & MCCLINTOCK, P. V. E. (2008). Inferential framework for nonstationary dynamics. I. Theory. *Physical Review E*, *77*(6), 061105.

MACEACHERN, S. N. (1999). Dependent nonparametric processes. In *ASA proceedings of the section on Bayesian statistical science*, (pp. 50–55). Alexandria, Virginia. Virginia: American Statistical Association; 1999.

MATSUMOTO, T., NAKAJIMA, Y., SAITO, M., SUGI, J., & HAMAGISHI, H. (2001). Reconstructions and predictions of nonlinear dynamical systems: A hierarchical Bayesian approach. *IEEE transactions on signal processing*, *49*(9), 2138–2155.

MCGOFF, K., MUKHERJEE, S., & PILLAI, N. (2015). Statistical inference for dynamical systems: A review. *Statistics Surveys*, *9*, 209–252.

MCSHARRY, P. E., & SMITH, L. A. (1999). Better nonlinear models from noisy data: Attractors with maximum likelihood. *Physical Review Letters*, *83*(21), 4285–4288.

MENA, R. H., RUGGIERO, M., & WALKER, S. G. (2011). Geometric stick–breaking processes for continuous–time Bayesian nonparametric modeling. *Journal of Statistical Planning and Inference*, *141*(9), 3217–3230.

MENA, R. H., & WALKER, S. G. (2015). On the Bayesian mixture model and identifiability. *Journal of Computational and Graphical Statistics*, *24*(4), 1155–1169.

MERKATAS, C., KALOUDIS, K., & HATJISPYROS, S. J. (2017). A Bayesian nonparametric approach to reconstruction and prediction of random dynamical systems. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *27*(6), 063116.

METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M. N., TELLER, A. H., & TELLER, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, *21*(6), 1087–1092.

MEYER, R., & CHRISTENSEN, N. (2000). Bayesian reconstruction of chaotic dynamical systems. *Physical Review E*, *62*(3), 3535–3542.

MEYER, R., & CHRISTENSEN, N. (2001). Fast Bayesian reconstruction of chaotic dynamical systems via extended Kalman filtering. *Physical Review E*, *65*(1), 016206.

MIRA, A. (1998). *Ordering, slicing and splitting Monte Carlo Markov chains*. Ph.D. thesis, University of Minnesota.

MOLKOV, Y. I., LOSKUTOV, E. M., MUKHIN, D. N., & FEIGIN, A. M. (2012). Random dynamical models from time series. *Physical Review E*, *85*(3), 036216.

MULIERE, P., & TARDELLA, L. (1998). Approximating distributions of random functionals of Ferguson–Dirichlet priors. *Canadian Journal of Statistics*, *26*(2), 283–297.

MÜLLER, P., QUINTANA, F., & ROSNER, G. (2004). A method for combining inference across related nonparametric bayesian models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *66*(3), 735–749.

MÜLLER, P., QUINTANA, F. A., JARA, A., & HANSON, T. (2015). *Bayesian nonparametric data analysis*. Springer.

NAKADA, Y., MATSUMOTO, T., KURIHARA, T., & YOSUI, K. (2005). Bayesian reconstructions and predictions of nonlinear dynamical systems via the Hybrid Monte Carlo scheme. *Signal processing*, *85*(1), 129–145.

NEAL, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics*, *9*(2), 249–265.

NEAL, R. M. (2003). Slice sampling. *The Annals of Statistics*, *31*(3), 705–741.

NIETO-BARAJAS, L. E., & QUINTANA, F. A. (2016). A Bayesian non-parametric dynamic AR model for multiple time series analysis. *Journal of Time Series Analysis*, *37*(5), 675–689.

ONGARO, A., & CATTANEO, C. (2004). Discrete random probability measures: a general framework for nonparametric Bayesian inference. *Statistics & Probability Letters*, *67*(1), 33–45.

PINCUS, S. M. (1991). Approximate entropy as a measure of system complexity. *In Proceedings of the National Academy of Sciences*, *88*(6), 2297–2301.

POMEAU, Y., & MANNEVILLE, P. (1980). Intermittent transition to turbulence in dissipative dynamical systems. *Comm. Math. Phys.*, *74*(2), 189–197.

RASMUSSEN, C. E., & WILLIAMS, C. K. (2006). *Gaussian processes for machine learning*. MIT press Cambridge.

ROBERT, C. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media.

ROBERT, C. P. (2004). *Monte Carlo methods*. Wiley Online Library.

RODRIGUEZ, A., & TER HORST, E. (2008). Bayesian dynamic density estimation. *Bayesian Analysis*, *3*(2), 339–365.

SCHENK-HOPPE, K. R. (1997). Bifurcations of the randomly perturbed logistic map.

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

SMELYANSKIY, V. N., LUC?HINSKY, D. G., TIMUCIN, D. A., & BANDRIVSKYY, A. (2005). Reconstruction of stochastic nonlinear dynamical models from trajectory measurements. *Physical Review E*, *72*(2), 026202.

TADDY, M. A., & KOTTAS, A. (2009). Markov switching Dirichlet process mixture regression. *Bayesian Analysis*, *4*(4), 793–816.

TEH, Y. W., JORDAN, M. I., BEAL, M. J., & BLEI, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, *101*(476), 1566–1581.

TOMLINSON, G., & ESCOBAR, M. (1999). *Analysis of densities.*. University of Toronto Technical report.

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

WEST, M. (1992). *Hyperparameter estimation in Dirichlet process mixture models*. Duke University ISDS Discussion Paper# 92-A03.

Appendix A

Sampling from nonstandard full conditionals

A.1 Sampling ϑ, x_0 and $x_{n+j}, 1 \leq j \leq T-1$

Here we adapt our calculations for the specific case where the deterministic part is a polynomial of degree m, namely $g(\theta, x) = \sum_{k=0}^{m} \theta_k \, x^k$.

A.1.1 Sampling the $\vartheta = (\theta)_{0 \leq j \leq m}$ coefficients

From eqs. (3.15) and (3.28) and for $j = 1, \ldots, m$ it is that

$$f(\theta_j | \cdots) \propto \mathcal{I}(\theta \in \tilde{\Theta}_j) \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} \lambda_{d_i} h_\theta(x_i, x_{i-1}) \right\}, \tag{A.1}$$

where $\tilde{\Theta}_j$ is the j–th projection interval of the set $\tilde{\Theta}$. Letting $\xi_{ji} := x_i - \sum_{\substack{k=0 \\ k \neq j}}^{m} \theta_k \, x_{i-1}^k$, we obtain the full conditional for θ_j, which is a normal truncated over the set $\tilde{\Theta}_j$ given by

$$f(\theta_j | \cdots) \propto \mathcal{I}(\theta \in \tilde{\Theta}_j) \mathcal{N}(\theta_j | \mu_j, \tau_j^{-1}) \tag{A.2}$$

with

$$\mu_j := \tau_j^{-1} \sum_{i=1}^{n} \lambda_{d_i} \xi_{ji} x_{i-1}^j, \;\; \tau_j := \sum_{i=1}^{n} \lambda_{d_i} x_{i-1}^{2j}.$$

To sample from this density, a-priori we set $\theta_j \in \tilde{\Theta}_j := (\theta_j^-, \theta_j^+)$ and we augment the θ_j full conditionals by the auxiliary variables θ_j' (Damien et al., 1999) such that jointly

$$f(\theta_j, \theta_j' | \cdots) \propto \mathcal{U}(\theta_j | \theta_j^-, \theta_j^+) \, \mathcal{I}\left(\theta_j' > (\theta_j - \mu_j)^2\right) e^{-\tau_j \theta_j'/2}. \tag{A.3}$$

Then we have the following Lemma:

Lemma A.1. *The augmentation of the full conditionals of θ_j for $j = 1, \ldots, m$ with the positive random variables θ_j' such that they jointly satisfy (A.3), leads to the following embedded Gibbs*

sampling scheme:

$$f(\theta_j'|\theta_j,\cdots) \propto \mathcal{E}(\theta_j'|\tau_j/2)\,\mathcal{I}(\theta_j' > (\theta_j-\mu_j)^2)$$
$$f(\theta_j|\theta_j',\cdots) = \mathcal{U}(\theta_j|\alpha_j,\beta_j),\ \alpha_j := \max\{\theta_j^-, \mu_j - \theta_j'^{1/2}\},\ \beta_j := \min\{\theta_j^+, \mu_j + \theta_j'^{1/2}\}.$$

where $\mathcal{E}(\theta_j'|\tau_j/2)$ denotes the exponential density with rate $\tau_j/2$.

Proof. These are the full conditionals of the bivariate density given in Equation (A.3). □

A.1.2 Sampling the initial condition x_0

Similarly, to sample from the full conditional of x_0 in eqs. (3.14) and (3.27), we introduce the variable x_0' such that

$$f(x_0, x_0'|\cdots) \propto \mathcal{I}(x_0 \in \tilde{X})\,\mathcal{I}\left(x_0' > h_\theta(x_1, x_0)\right) e^{-\lambda_{d_1} x_0'/2}.$$

Clearly, the full conditional of x_0' is an exponential of rate $\lambda_{d_1}/2$, truncated over the interval $(h_\theta(x_1,x_0),\infty)$. The new full conditional for x_0 is a mixture of at most m uniforms given by

$$f(x_0|x_0',\cdots) \propto \mathcal{I}(x_0 \in \tilde{X})\,\mathcal{I}(x_0 \in \mathcal{R}_g),\ \mathcal{R}_g := \{x : \underline{x}_0 < g(\theta,x) < \overline{x}_0\}, \tag{A.4}$$

where $\underline{x}_0 := x_1 - x_0'^{1/2}$ and $\overline{x}_0 := x_1 + x_0'^{1/2}$. The set $\mathcal{R}_g$ can be represented as the union of intervals, with boundaries defined by the real roots of the two polynomial equations

$$\underline{q}(x_0) := g(\theta,x_0) - \underline{x}_0 = 0, \quad \overline{q}(x_0) := g(\theta,x_0) - \overline{x}_0 = 0. \tag{A.5}$$

More specifically, we are going to show that there is $r \le m$ such that

$$\mathcal{R}_g = \textstyle\bigcup_{i=1}^{r} (\rho_{2i-1}, \rho_{2i}), \tag{A.6}$$

with $\{\rho_1,\ldots,\rho_{2r}\}$ the ordered set of the real roots of the two polynomial equations in (A.5). In the sequel we make use of the following notation

$$\{\overline{q} < 0\} := \{x_0 \in \mathbb{R} : \overline{q}(x_0) < 0\},$$
$$\{\underline{q} > 0\} := \{x_0 \in \mathbb{R} : \underline{q}(x_0) > 0\}.$$

First we will consider the two even degree cases. When the leading coefficient is positive, the equation $\overline{q} = 0$ has at least two real roots. If there are more than two real roots, their number will be a multiple of two. On the other hand, when $\underline{q} = 0$ has real solutions their number will be even. Then for $s' \ge 1$ and $t' \ge 0$ it is that

$$\{\overline{q} < 0\} = (\overline{\rho}_1, \overline{\rho}_2) \cup \cdots \cup (\overline{\rho}_{2s'-1}, \overline{\rho}_{2s'}) \tag{A.7}$$
$$\{\underline{q} > 0\} = (-\infty, \underline{\rho}_1) \cup \cdots \cup (\underline{\rho}_{2t'}, \infty). \tag{A.8}$$

When $t' \geq 1$ it is that $\overline{\rho}_1 < \underline{\rho}_1 < \underline{\rho}_{2t'} < \overline{\rho}_{2s'}$. Therefore $r = 2(s'+t')$ and the intersection of the two sets $\{\overline{q} < 0\}$ and $\{\underline{q} > 0\}$ is of the form (A.6). When the leading coefficient is negative the result is similar with the right hand sides of equations (A.7) and (A.8) interchanged.

When the degree is odd and the leading coefficient is positive, both equations $\overline{q} = 0$ and $\underline{q} = 0$ have at least one real solution $\overline{\rho}_1$ and $\underline{\rho}_1$ respectively, with $\underline{\rho}_1 < \overline{\rho}_1$. If some of the two equations have more than one real solution, the number of the additional roots will be a multiple of two. So for $s' \geq 0$ and $t' \geq 0$ it is that

$$\{\overline{q} < 0\} = (-\infty, \overline{\rho}_1) \cup (\overline{\rho}_2, \overline{\rho}_3) \cup \cdots \cup (\overline{\rho}_{2s'}, \overline{\rho}_{2s'+1}) \tag{A.9}$$

$$\{\underline{q} > 0\} = (\underline{\rho}_1, \underline{\rho}_2) \cup \cdots \cup (\underline{\rho}_{2t'-1}, \underline{\rho}_{2t'}) \cup (\underline{\rho}_{2t'+1}, \infty). \tag{A.10}$$

For $s' \geq 1$ and $t' \geq 1$ we have $\underline{\rho}_1 < \overline{\rho}_1 < \underline{\rho}_{2t'+1} < \overline{\rho}_{2s'+1}$, and $r = 2(s'+t'+1)$ which shows that the intersection of the two sets $\{\overline{q} < 0\}$ and $\{\underline{q} > 0\}$ is of the form (A.6). When the leading coefficient is negative the result is similar with the right hand sides of the equations (A.9) and (A.10) interchanged.

So we have proved the following lemma:

Lemma A.2. *The augmentation of the full conditional of x_0 with the positive random variable x_0' leads to the following embedded Gibbs sampling scheme:*

$$\begin{aligned} f(x_0'|x_0, \cdots) &\propto \mathcal{E}(x_0'|\lambda_{d_1}/2)\, \mathcal{I}(x_0' > h_\theta(x_1, x_0)) \\ f(x_0|x_0', \cdots) &\propto \mathcal{I}(x_0 \in \tilde{X})\, \mathcal{I}\left(x_0 \in \textstyle\bigcup_{i=1}^{r} (\rho_{2i-1}, \rho_{2i})\right), \end{aligned}$$

for some $r \leq m$, with $\{\rho_1, \ldots, \rho_{2r}\}$ being the ordered set of the real roots of the two polynomial equations in (A.5).

A.1.3 Sampling the first $T-1$ future observations

The full conditionals x_{n+j} for $1 \leq j \leq T-1$ in eqs. (3.16) and (3.29) given in the main text are nonstandard densities. We augment the conditional of x_{n+j} with the pair of variables (x'_{n+j}, x''_{n+j}) such that jointly

$$\begin{aligned} f(x_{n+j}, x'_{n+j}, x''_{n+j}|\cdots) \propto\ & e^{-\frac{1}{2}\lambda_{d_{n+j}} x'_{n+j}}\, \mathcal{I}(x'_{n+j} > h_\theta(x_{n+j}, x_{n+j-1})) \\ & \times e^{-\frac{1}{2}\lambda_{d_{n+j+1}} x''_{n+j}}\, \mathcal{I}(x''_{n+j} > h_\theta(x_{n+j+1}, x_{n+j})). \end{aligned}$$

The full conditionals of x'_{n+j} and x''_{n+j} are truncated exponentials with rates $\lambda_{d_{n+j}}/2$ and $\lambda_{d_{n+j+1}}/2$ over the intervals $(h_\theta(x_{n+j}, x_{n+j-1}), \infty)$ and $(h_\theta(x_{n+j+1}, x_{n+j}), \infty)$ respectively.

The full conditional of x_{n+j} is of the form (A.4) with the set $\tilde{X}$ replaced by the set (x^-_{n+j}, x^+_{n+j}) with $x^\pm_{n+j} := g(\theta, x_{n+j-1}) \pm x'^{1/2}_{n+j}$, and the set $\mathcal{R}_g$ replaced by the set $\{x : \underline{x}_{n+j} < g(\theta, x) < \overline{x}_{n+j}\}$ with $\underline{x}_{n+j} := x_{n+j+1} - x''^{1/2}_{n+j}$ and $\overline{x}_{n+j} := x_{n+j+1} + x''^{1/2}_{n+j}$.

A.2 Sampling the geometric probability λ

To sample from the density ineq. (3.32) in the main text we include the pair of positive auxiliary random variables p_1 and p_2 such that

$$f(\lambda, \lambda_1, \lambda_2 \,|\, \cdots) \propto \lambda^{2n_T-\alpha-1}\mathcal{I}(\lambda_1 < (1-\lambda)^{L_{n_T}})\mathcal{I}(\lambda_2 < e^{-\beta/\lambda}),$$

with $\lambda \in (0,1)$. The full conditionals for λ_1 and λ_2 are uniforms

$$f(\lambda_1 \,|\, \cdots) = \mathcal{U}(\lambda_1|\, 0, (1-\lambda)^{L_{n_T}}), \quad f(\lambda_2 \,|\, \cdots) = \mathcal{U}(\lambda_2|\, 0, e^{-\beta/\lambda}).$$

The new full conditional for λ becomes

$$f(\lambda \,|\, \lambda_1, \lambda_2, \cdots) \;\propto\; \lambda^{2n_T-\alpha-1} \begin{cases} \mathcal{I}\left(-\frac{\beta}{\log\lambda_2} < \lambda < 1-\lambda_1^{1/L_{n_T}}\right) & L_{n_T} \geq 0 \\ \mathcal{I}\left(\max\left\{-\frac{\beta}{\log\lambda_2}, 1-\lambda_1^{1/L_{n_T}}\right\} < \lambda < 1\right) & L_{n_T} < 0. \end{cases}$$

We can sample from this density using the inverse cumulative distibution function technique.

Appendix B

Invariant set of the map $x' = \tilde{g}(\vartheta^*, x)$

For $\vartheta = \vartheta^* = 2.55$ we let

$$\tilde{g}(x) \equiv \tilde{g}(\vartheta^*, x) = 0.05 + 2.55x - 0.99x^3,$$

and we define $\tilde{g}^{(n)}$ to be the n-fold composition of $\tilde{g}$ with itself. We let $\mathcal{R}^{(2)}$ to be the set of real roots of the polynomial equation $\tilde{g}^{(2)}(x) = x$, with $\underline{x} = \min \mathcal{R}^{(2)}$, $\overline{x} = \max \mathcal{R}^{(2)}$ and $\mathbb{X} = [\,\underline{x}, \overline{x}\,]$. We denote the complement of $\mathbb{X}$ by $\mathbb{X}' = \mathbb{X}'_- \cup \mathbb{X}'_+$, where $\mathbb{X}'_- = (-\infty, \underline{x}\,)$ and $\mathbb{X}'_+ = (\overline{x}, \infty)$. We will prove the following lemma:

Lemma B.1. *Let $\tilde{g}$ be the polynomial given in eq.* (3.33)*, then for all $x \in \mathbb{X}'$, it is that*

$$\liminf_{n\to\infty} \tilde{g}^{(n)}(x) = -\infty \text{ and } \limsup_{n\to\infty} \tilde{g}^{(n)}(x) = \infty.$$

Proof. It is not difficult to verify geometrically the following facts:

1. $\tilde{g}(\underline{x}\,) = \overline{x},\ \tilde{g}(\overline{x}\,) = \underline{x}$.
2. $\underline{x} \le x \le \overline{x} \Leftrightarrow \underline{x} \le \tilde{g}(x) \le \overline{x}$.
3. $\tilde{g}(x) > x,\ \tilde{g}^{(2)}(x) < x,\ \forall\, x \in \mathbb{X}'_-$.
4. $\tilde{g}(x) < x,\ \tilde{g}^{(2)}(x) > x,\ \forall\, x \in \mathbb{X}'_+$.
5. The restrictions of $\tilde{g}$ and $\tilde{g}^{(2)}$ to $\mathbb{X}'$, are decreasing and increasing functions respectively.

Then for all $x \in \mathbb{X}'_-$ we have the set of inequalities

$$\tilde{g}^{(2n+1)}(x) < \tilde{g}^{(2n-1)}(x) < \cdots < \tilde{g}(x) < \underline{x}.$$

Suppose that $\lim_{n\to\infty} \tilde{g}^{(2n+1)}(x) = x^*$ then $\lim_{n\to\infty} \tilde{g}^{(2n+3)}(x) = \tilde{g}^{(2)}(x^*) = x^*$, meaning that $x^* \in \mathcal{R}^{(2)}$ which is a contradiction. Therefore $\lim_{n\to\infty} \tilde{g}^{(2n+1)}(x) = -\infty$, for all $x \in \mathbb{X}'_-$. Similarly for all $x \in \mathbb{X}'_+$ we have the set of inequalities

$$\tilde{g}^{(2n)}(x) > \tilde{g}^{(2n-2)}(x) > \cdots > \tilde{g}^{(2)}(x) > \overline{x},$$

from which $\lim_{n\to\infty} \tilde{g}^{(2n)}(x) = \infty$, for all $x \in \mathbb{X}'_+$. □

Appendix C

Julia codes

The algorithms for all the models constructed in this thesis, that is the GSBR, the PDGSBP (as well as their Dirichlet process counterparts rDPR,rPDDP), and the PDGSBP reconstruction model have been developed in the `Julia` language (Bezanson et al., 2012).

The associated software is available and can be downloaded from the URL:

`Link to thesis codes`

or available upon request via e-mail:

`cmerkatas@aegean.gr` or `merxri@gmail.com`.